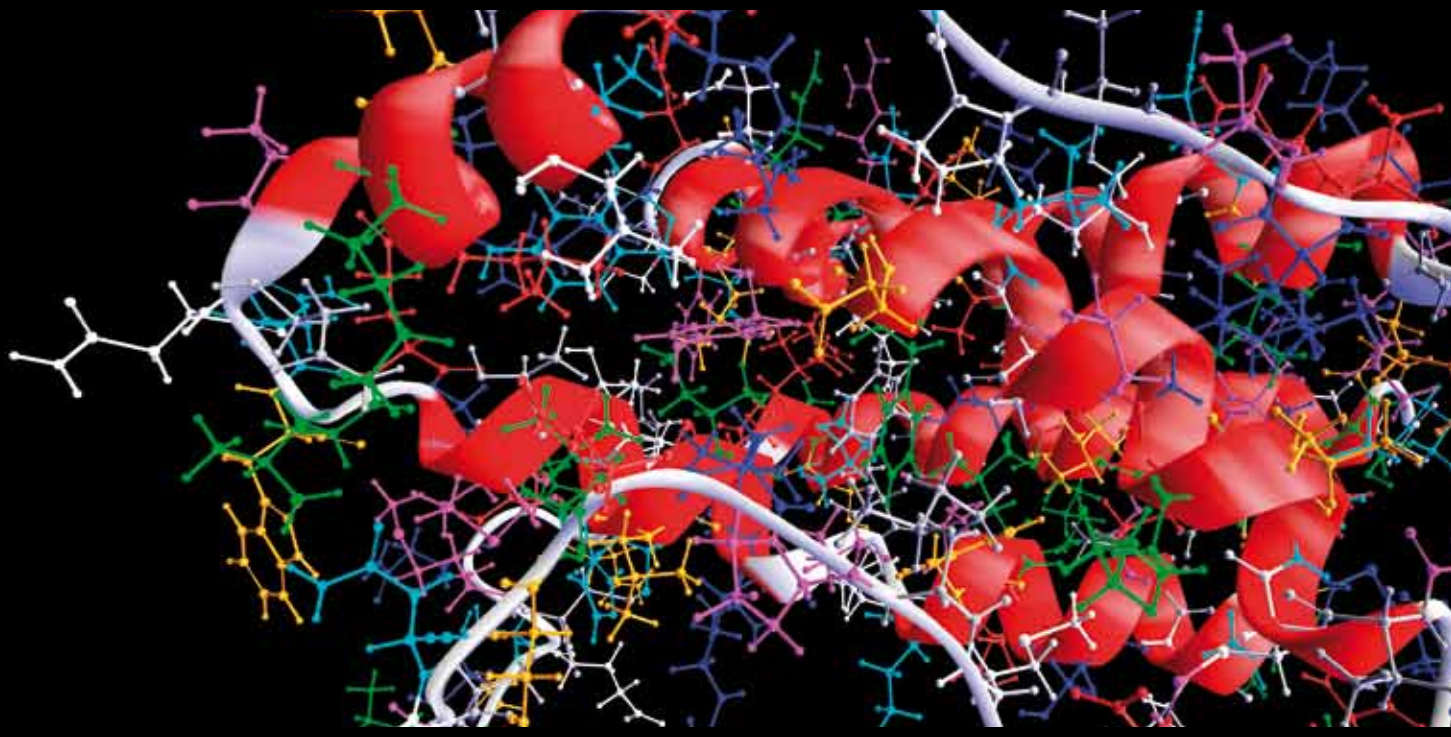
Computational and Mathematical Methods in Medicine

Statistical Genetics and Its Applications in Medical Studies

Guest Editors: Ao Yuan, Wenqing He, Gengsheng Qin, and Qizhai Li



Statistical Genetics and Its Applications in Medical Studies

Computational and Mathematical Methods in Medicine

Statistical Genetics and Its Applications in Medical Studies

Guest Editors: Ao Yuan, Wenqing He, Gengsheng Qin, and Qizhai Li

Editorial Board

Contents

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 712073, 3 pages
http://dx.doi.org/10.1155/2014/712073

Editorial
Statistical Genetics and Its Applications in Medical Studies

Ao Yuan,[1] Wenqing He,[2] Gengsheng Qin,[3] and Qizhai Li[4]

[1] *Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington, DC 20057, USA*
[2] *Department of Statistics and Actuarial Science, University of Western ontario, London, ON, Canada N6A 5B7*
[3] *Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, USA*
[4] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

Correspondence should be addressed to Ao Yuan; yuanao@hotmail.com

Received 8 December 2013; Accepted 9 December 2013; Published 6 March 2014

Statistical genetics can be viewed as a classical branch of applied probability and statistics, which has recently gained much momentum, due to the significant breakthroughs in genetics. With the availability of modern techniques, new methods, and significantly increased data information, it is imperative to study the relationship between gene traits such as diseases and genetic susceptibilities in an unprecedented manner. This area is among the hottest topics in applied statistics, applied mathematics, biological/medical studies, and other related sciences.

The main aim of this special issue focuses on the new development and applications of computational, mathematical, and statistical methods in genetic disease study. The special issue could become an international forum for researchers to exchange their new thoughts, most recent developments, and ideas in the field.

In this special issue, we selected seven articles within the above topics. Below is a brief summary of the selected articles in this special issue.

Recent advances in biotechnologies have led to the identification of an enormous number of genetic markers in disease association studies; how to select a smaller set of genes to explore the relation between genes and disease is a challenging task. Bayesian methods have the advantage of incorporating prior information into the model for such analysis. Article "*Applications of bayesian gene selection and classification with mixtures of generalized singular g-priors*" by W.-K. Chien and C. K. Hsiao addresses this problem using Bayesian method with a Gaussian prior and inverse gamma hyperprior. The proposed approach is applied to a colon and leukemia cancer study. Comparison with other existing methods was conducted. The authors find that classification accuracy of the proposed model is higher with a smaller set of selected genes and that the results not only replicated findings in several earlier studies, but also provided the strength of association with posterior probabilities.

Article "*Modified logistic regression models using gene coexpression and clinical features to predict prostate cancer progression*" by H. Zhao et al. proposed a new logistic regression model for predicting prostate cancer progression. They incorporated coexpressed gene profiles into the logistic model based on clinical features to improve the inference accuracy. Then they use the top-scoring pair method to select genes with significant association with the disease. The performance of the proposed method is compared with some commonly used methods for such problem, using data sets from such published studies. Their study suggests that the proposed method performs better than a commonly used one and that the top-scoring pair method is a useful tool for feature (and/or gene) selection to be used in prognostic models.

Resampling-based multiple testing procedures are widely used in genomic studies to identify differentially expressed genes and for genome-wide association studies. The power and stability of these popular procedures have not been extensively evaluated. Article "*Power and stability properties of resampling-based multiple testing procedures with applications to gene oncology studies*" by D. Li and T. D. Dye investigates the power and stability of seven commonly used resampling-based multiple testing procedures that are frequently used in high-throughput data analysis for small sample size data. Simulations and real data gene oncology

examples are employed in their investigation. Their study suggests that the bootstrap single-step minP procedure and the bootstrap step-down minP procedure perform the best, when sample size is as small as 3 in each group and either familywise error rate or false discovery rate control is desired. When sample size increases to 12 and false discovery rate control is desired, the permutation maxT procedure and the permutation minP procedure perform the best.

Article "*Transcriptional protein-protein cooperativity in POU/HMG/DNA complexes revealed by normal mode analysis*" by D. D. Wang and H. Yan investigates how proteins in POU/HMG/DNA ternary complexes interact cooperatively, which are crucial in transcriptional regulation of embryonic stem cells. They use the normal mode analysis to detect the most cooperative or collective motions (essential modes) of a large number of proteins, a commonly used tool to analyze the structural dynamics of biomolecules, which combines some techniques in engineering, mathematics, and statistics. Their work reveals how the two proteins Oct-1 and Sox-2 work together physically and structurally at two specific DNA biding sites, by analyzing the motion magnitude functions. A correlation measure is used to characterize the amount of cooperativity of pairs of proteins. The proposed methods provide useful information for understanding the complicated interaction mechanism in the POU/HMG/DNA complexes. The corresponding online computational tools are also provided.

In modern medical diagnosis or genetic studies, the receiver operating characteristic (ROC) curve is a popular tool to evaluate the discrimination performance of biomarkers on a disease status or a phenotype. With the presence of a number of covariates in the data, how to select the most relevant covariables, or to select the model with good overall properties, is a challenging problem. Article "*Variable selection in ROC regression*" by B. Wang addresses this problem with an interesting idea. There are a large number of criteria available for this problem. The author first rewrites the ROC regression into a grouped variable selection form so that current criteria can be applied and then proposes a general two-stage framework with a BIC selector for the group SCAD algorithm under the local model assumption. Basic asymptotic properties of the proposed methods are derived. Simulation studies and real data analysis show that the proposed grouped variable selection is superior to the traditional model selections. Furthermore, the author finds that the focused information criterion provides more accurate estimated area under the ROC curve compared with other criteria.

Two-stage design and analysis are often adopted in genome-wide association studies (GWASs). Considering the genetic model uncertainty, many robust procedures have been proposed and applied in GWASs. The existing approaches mostly focused on binary traits, and many of these methods analyze data based on two separate stages, and few work has been done on continuous (quantitative) traits. Article "*Robust joint analysis with data fusion in two-stage quantitative trait genome-wide association studies*" by D.-D. Pan et al. proposes a powerful F-statistic-based robust joint analysis method for quantitative traits using the combined raw data from both stages, in which the genetic effects are modeled as regression parameters. Variations of the MAX testing statistic are constructed to calculate the statistical significance and power. It is well known that critical values and power of the MAX type statistic are not easy to compute. The authors derived analytic expressions on the basis of the asymptotic distributions, so that these quantities can be easily obtained. They show using simulations that the proposed method is substantially more robust than the F-test based on the commonly used additive model when the underlying genetic model is unknown.

Multiple meta-analyses may use similar search criteria and focus on the same topic of interest, but they may yield different or sometimes discordant results. The lack of statistical methods for synthesizing these findings makes it challenging to properly interpret the results from multiple meta-analyses, especially when their results are conflicting. Article "*A statistical method for synthesizing meta-analyses*" by L. L. Tang et al. introduces a method to synthesize the meta-analytic results under two cases: (1) when multiple meta-analyses use the same type of summary effect estimates and (2) when meta-analyses use different types of effect sizes. In case 2, the meta-analysis results cannot be directly combined; therefore they propose a two-step frequentist procedure to first convert the effect size estimates to the same metric and then summarize them with a weighted mean estimate. The proposed method has the following advantages over some existing methods: different types of summary effect sizes can be considered; the same overall effect size can be provided by conducting a meta-analysis on all individual studies from multiple meta-analyses.

One of the main objectives of a genome-wide association study (GWAS) is to develop a prediction model for a binary clinical outcome using single-nucleotide polymorphisms (SNPs) which can be used for diagnostic and prognostic purposes and for better understanding of the relationship between the disease and SNPs. Penalized support vector machine (SVM) methods have been widely used toward this end. However, since investigators often ignore the genetic models of SNPs, a final model results in a loss of efficiency in prediction of the clinical outcome. Article "*SNP selection in genome-wide association studies via penalized support vector machine with MAX test*" by J. Kim et al. proposes a two-stage method such that the genetic models of each SNP are identified using the MAX test and then a prediction model is fitted using a penalized SVM method. They apply the proposed method to various penalized SVMs and compare their performances using various penalty functions. They show by simulations and real GWAS data analysis that the proposed method performs better than the prediction methods that ignore the genetic models, in terms of prediction power and selectivity.

Using DNA sequence data in the study of ancestral history of human population is an essential part in the understanding of human evolution. The existing methods for such coalescence inference using the method of either the rooted tree or unrooted tree constructed from the observed data, both of which use recursion formulae to compute the data probabilities. These methods are useful in

practical applications but computationally complicated. Article "*On coalescence analysis using genealogy rooted trees*" by A. Yuan et al. explores a new method for this problem. They first investigate the asymptotic behavior of such inference; their results indicate that, broadly, the estimated coalescent time will be consistent to a finite limit. Then they study a relatively simple computation method for this analysis and illustrate how to use it.

Acknowledgment

We thank all the authors who contributed to this special issue.

Ao Yuan
Wenqing He
Gengsheng Qin
Qizhai Li

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 194202, 8 pages
http://dx.doi.org/10.1155/2014/194202

Research Article

On Coalescence Analysis Using Genealogy Rooted Trees

Ao Yuan,[1] Gengsheng Qin,[2] Wenqing He,[3] and Qizhai Li[4]

[1] *Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington, DC 20057, USA*
[2] *Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, USA*
[3] *Department of Statistics and Actuarial Science, University of Western Ontario, London, ON, Canada N6A 5B7*
[4] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

Correspondence should be addressed to Ao Yuan; yuanao@hotmail.com

Received 10 August 2013; Accepted 10 January 2014; Published 23 February 2014

Academic Editor: Henggui Zhang

DNA sequence data are now being used to study the ancestral history of human population. The existing methods for such coalescence inference use recursion formula to compute the data probabilities. These methods are useful in practical applications, but computationally complicated. Here we first investigate the asymptotic behavior of such inference; results indicate that, broadly, the estimated coalescent time will be consistent to a finite limit. Then we study a relatively simple computation method for this analysis and illustrate how to use it.

1. Introduction

In the past decades, considerable progress has been made in the field of population genetics. One of the main goals is to infer the coalescence time of the population under study, that is, to infer the time since their most recent common ancestor (MRCA) and its distribution based on the observed data.

In genetics, coalescent theory is a retrospective of population genetics that traces all genes in a sample from a population to a single ancestral copy shared by all the members of the population. The coalescent time of a population is the time of their most recent common ancestor. The inheritance relationship among the genes is typically represented as a gene genealogy, similar to a phylogenetic tree. The goal of coalescent analysis is to infer the coalescent time of a sample of n individuals independently sampled from a population of size N, based on their observed DNA sequence diversity. Unlike parameter inference for independent and identically distributed (iid) data, for which asymptotic limit can be used conveniently to characterize the estimator when the data size is large, various existing studies indicate that the estimated MRCA, in unit of N generations, is unclear as whether it will concentrate as the data sample size increases without bound. In contrast, in the estimation of mutation rate in the same setting, the estimate is consistent and asymptotically normal [1], although at a much slower rate of $\log^{1/2}(n)$, compared to the rate of $n^{1/2}$ for i.i.d. data. Also, different from usual parameters, the MRCA changes with n, the number of sequences. This prompts us to the investigate the asymptotic behavior of the estimated coalescent time. We want to know whether such estimator will be asymptotically consistent and in what sense if it does. Conditioning on the total number of segregating sites, we find that such estimators converge or not to some nonnegative finite limits in posterior mean, depending on the behavior of the number of mutations on all the branches of the rooted trees constructed from the observed data. Also, analysis of this problem with this type of data is often computationally extensive and complicated; we study a relatively simple simulation method for this problem. We first study the asymptotic behavior of this method in Section 3, and then describe and illustrate our method for this problem in Section 4.

In coalescence inference, mitochondrial DNA (mtDNA) data plays an important role. Mitochondria is one of the few genes existing outside the cell nucleus, and for mammalian it is only maternally inherited. Human mtDNA is a double-stranded molecule sequence about 16,500 base pairs in length. It is outside the cell nuclear, and it is known that the mutation

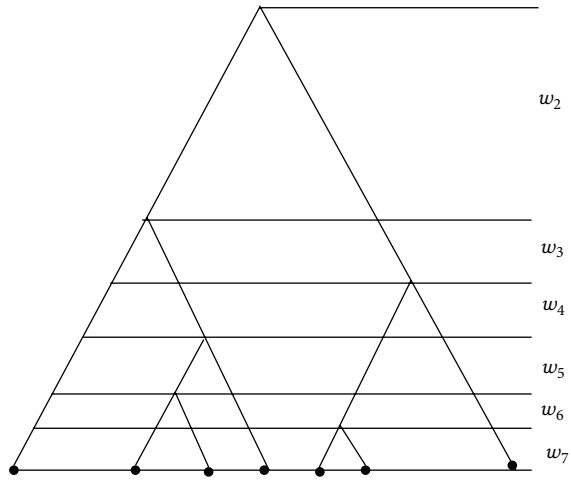


Figure 1: Coalescent tree for a sample of seven individuals.

rate in mtDNA is about 10 times that of the nuclear genes, and that on one section of the mitochondria, its control region, the mutation rate is even one order higher. The simple inheritance pattern and high variability make mtDNA an important source in the study of human evolutionary history. Each site on the DNA strand has one of the four bases A, C, G, or T. As the molecule evolves, mutations occur in the form of base substitutions. The change between purines (A,G) or pyrimidines (C,T) is called transition; that between a purine and pyrimidine is transversion. The former type of substitution is much more common than the latter.

We focus on the control region of the mitochondrial data in Griffiths and Tavaré [2], which is part of the data in Ward et al. [3]. They are from a segment of the control region, with 352 base pairs (sites), out of which 159 are purine sites and 193 are pyrimidine sites. This data contains 63 sequences sampled from a North American Indian tribe, the Nuu-Chah-Nulth, from Vancouver Island. After eliminating sequences with multiple mutations on some single sites, so that the assumption of at most one mutation each site is met, the remaining data has 55 sequences, with 14 distinct sequences (called lineages) in the data. Site at which not all the observed sequences have the same base is a *segregating site*. The whole sequences are long, but only the segregating sites are informative for the analysis; the other sites are ignored. The mentioned data has 18 segregating sites and is presented in Table 1, with the frequency (or multiplicity) of each lineage.

2. Brief Review of Background and Related Methods

The coalescent is a model for the genealogical tree of a random sample of n DNA sequences from a large population. An example of such a tree of sample size $n = 7$ is given in Figure 1.

For more detailed reviews of this topic, see Hudson [4] and Donnelly and Tavaré [5].

In coalescence inference one has the following.

Basic Assumptions. The population size N is large, remains unchanged for many generations into the past, and is known, or can be estimated from other sources; the data is a random sample from the population; the number of births in each generation follows the Wright-Fisher model (since the population is of constant size, the number of deaths also follows the similar model); mutation (substitution) at any nucleotide site can occur only once in the ancestry and is irreversible; mutations that occur in different time intervals are independent; the time point at which mutation occurs follows a Poison distribution with rate $\theta/2$ to be defined latter, independently in each branch of the genealogy tree, where θ is known, or can be estimated from other methods or sources.

The inference of coalescence time t_n of a sample population of size n has two steps. The first step is modeling the distribution of t_n without any data, the *predata* distribution; then in the second step, update the predata distribution, using the observed data, to the *postdata* distribution, based on which the formal inference is conducted. The predata distribution is pioneered by Kingman [6, 7]; he showed that, in time units of N generations,

$$t_n = \sum_{j=2}^{n} w_j, \tag{1}$$

where the w_j's are independent waiting times. w_j is the time from $j-1$ common ancestors of the sample to j common ancestors. A quick reference on this can be found in Tavaré [8]. Here w_j is distributed as exponential $\mathrm{Exp}(j(j-1)/2)$, with $E(w_j) = 2/(j(j-1))$. The w_js can be represented graphically as a coalescent tree as in Figure 1; then t_n is the height of the tree. Define the tree length as

$$l_n = \sum_{j=2}^{n} j w_j; \tag{2}$$

then (Kingman)

$$\begin{aligned} E(t_n) &= 2\left(1 - \frac{1}{n}\right), \\ \mathrm{Var}(t_n) &= 8\sum_{j=2}^{n} \frac{1}{j^2} - 4\left(1 - \frac{1}{n}\right)^2; \\ E(l_n) &= 2\sum_{j=1}^{n-1} \frac{1}{j}, \\ \mathrm{Var}(l_n) &= 4\sum_{j=1}^{n-1} \frac{1}{j^2}. \end{aligned} \tag{3}$$

The time unit is transformed to years by the relationship $t_n NY$, where Y is the average years of each generation, which is usually taken as 20–25. Here we see that, as an initial analysis without the observed data, the coalescent time of a random sample of size n from a population of size N is roughly $2N$ generations, as long as $n(\leq N)$ is moderately large.

Table 1: Nucleotide position in control region.

Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Lineage
			Purines									Pyrimidines							Freqs.
Lineage																			
a	A	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	2
b	A	G	G	A	A	T	C	C	T	T	T	T	C	T	C	T	T	C	2
c	G	A	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	1
d	G	G	A	G	A	C	C	C	C	C	T	T	C	C	C	T	T	C	3
e	G	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	19
f	G	G	G	A	G	T	C	C	T	C	T	T	C	T	C	T	T	C	1
g	G	G	G	G	A	C	C	C	T	C	C	C	C	C	C	T	T	T	1
h	G	G	G	G	A	C	C	C	T	C	C	C	T	C	C	T	T	T	1
i	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	C	T	4
j	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	T	T	8
k	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	C	5
l	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	4
m	G	G	G	G	A	C	C	T	T	C	T	T	C	C	C	T	T	C	3
n	G	G	G	G	A	C	T	C	T	C	T	T	C	C	T	T	T	C	1

Each row of the table represents a DNA sequence lineage. In this data, there are transitions but no transversion observed.

Thus, the coalescent time of a sample from a subpopulation is roughly the same as that of the population (as long as the sample size is moderately large). This phenomenon is further investigated by Watterson [9], who showed that

$$P\left(A_N\left(t_n\right)=1\right)=\frac{(n-1)(N+1)}{(n+1)(N-1)}, \tag{4}$$

where $A_N(t_n)$ is the number of ancestors, at t_n generations ago, of the population with size N from which the data sample of size n is drawn. Here the sample must be a random draw from the population; otherwise the result may not be reliable. For example, the sample of size n is drawn from a subpopulation of size $N_1 < N$ from a population of size N; then by (3), the predata estimated of the coalescent time t_n of this sample is roughly $2N_1$ generations, but also it is roughly $2N$ generations since the sample is also from the whole population. The paradox arises from the sampling scheme. If the sample is drawn from the subpopulation of size N_1, one can only use $2N_1$ as the time scale, not $2N$, since the samples drawn from the subpopulation are expected to have smaller genetic variation than from the whole population.

For mutation, the common assumption is that the times at which mutation occurs follow a Poison process with constant rate $\theta/2$, so that, in any branch of length l from the tree, the number of mutations on that branch has a Poison distribution with mean $l\theta/2$, independently of the mutations on the other branches. For the time scale mentioned before, usually $\theta = 2N\mu$, where μ is the probability of a mutation that occurs per sequence per generation. For DNA sequences, μ is the sequence length (number of bases) times the mutation rate per site per generation and is often available from other sources. Since the coalescent time of a sample with moderate size is approximately $2N$ generations, θ can be approximately interpreted as the cumulative (since the time of MRCA) mutation rate (number of mutations) per sequence. Also, since the population size is N, $\theta/2$ can also be interpreted as the mutation rate of the whole population per generation.

Thus, given the mutation rate θ and the tree length l_n, the number of mutations s_n in a sample of n individuals from the given population follows the Poison distribution $\mathrm{Po}(\theta l_n/2)$ [10]

$$\begin{aligned} P\left(s_n=k \mid l_n=l\right) &= e^{-\theta l/2}\frac{(\theta l/2)^k}{k!} \\ &:= \mathrm{Po}\left(k,\frac{\theta l}{2}\right) \quad k=0,1,2,\ldots \end{aligned} \tag{5}$$

Note that this probability does not depend on n, but on k, l, and θ. Why $\theta l_n/2$?. Take $n = 2$; then $\theta l_n/2 = \theta t_n \approx \theta$, which is the expected number of cumulative mutations since t_n generations ago in a sequence. So $\theta l_n/2$ is a reasonable choice of the parameter in the Poison distribution. But if we model the number k of cumulative mutations per sequence since t_n generations ago, for moderately large n, we should use $\mathrm{Po}(k, t_n N\mu) \approx \mathrm{Po}(k, 2N\mu) = \mathrm{Po}(k, \theta)$.

The key in the coalescence inference is to evaluate the postdata distribution of t_n, which is much more involved than its predata distribution, it depends heavily on the mutation distribution in the data. For example, if more mutations occur in the earlier stage of the genealogy tree, then the estimated t_n will be bigger. Although under the assumption that mutation can only occur at most once at each site and mutation is irreversible, the total number of mutations in the observed data is just the number of segregating sites. But how the mutations distribute in the branches of the genealogy tree is unknown. Such distribution is crucial in the inference of t_n, which depends on how much data information being used and on the actual methods. This is our focus from now on. Denoting by D_n the observed data, the estimated coalescent

time $\widehat{t}_n$ of the sample is given by the postdata distribution mean of t_n as

$$\widehat{t}_n = E\left(t_n \mid D_n\right). \tag{6}$$

The inference can be viewed as a Bayesian procedure, with the predata and postdata distributions that correspond to the prior and posterior distributions in a Bayesian framework. But unlike the common Bayes setting, here the parameter t_n varies with the sample size n, and the data cannot be modeled i.i.d. with this parameter. That is the reason the inference of t_n cannot be made arbitrarily accurate, in the sense that the variance of the postdata distribution cannot be arbitrarily small, as the sample size increases without bound. Also, generally the postdata distribution is not in closed form and has to be evaluated by sampling methods. Tavaré et al. [10] derived the postdata distribution based on only the number of segregating sites in the sample. This method is very convenient to use, but does not use the DNA sequences structural information. The well known method in Griffiths and Tavaré [2], hereafter GT, is based on the full data information represented by a set of rooted trees. This method is one of the basic tools in coalescent inference using full data information, but is computationally complicated.

To evaluate the postdata coalescent distribution, GT used the probabilities recursion formula, derived in Ethier and Griffiths [11]. The method is not easy to fully understand and correctly use for many geneticists. Also these probabilities are computationally prohibitive; the postdata distribution of t_n is computed by a Markov chain Monte Carlo sampling and is quite involved.

Here we study a relatively simple approximate method using the full data information; in this method, instead of computing the tree probabilities as in GT, we just set the postdata tree probabilities as uniform for the $s+1$ rooted trees and use a simulation method to compute the coalescent distribution; thus, getting round of the complicated evaluations of the tree probabilities, it is easy to understand and much simpler in computation.

The rooted tree plays an important role in the analysis, which is not uniquely determined from the data. The data is equivalent to an unrooted tree, which is equivalent to a set of unrooted trees. Each rooted tree has a 0-1 valued matrix representation which is convenient for some computations, but not any 0-1 valued matrix corresponds to a rooted tree. In the following, we give more details about them and their relationships.

Rooted Tree. A rooted tree consists of a system of branches, subbranches, and so forth. The tip of each branch or subbranch represents a known lineage. The observed mutations in the sample are represented as dots in the branches, subbranches, and so forth at specified positions. The observed multiplicity of each lineage is represented as leaves at the tip of each branch or subbranch, and so forth.

The presentation of a rooted tree is unique up to the relative positions of its branches, subbranches, and so forth. A rooted tree has several levels of randomness. If we only know the sample size n, then the rooted tree has a total of n leaves; apart from that, the shape of the tree, how to split, how to allocate the leaves, how many mutations, and the distribution of the mutations are all random. If the data and the number of mutations are given, then the tree can only take a few shapes. Different from GT and other related literatures, here we put the observed lineage frequencies (multiplicities) as leaves in the corresponding tips of branches, subbranches, and so forth of the rooted tree.

Different from a coalescent tree which has a complete time ordering of the splitting points of branches, a rooted tree has only partial time orderings of these splits and mutations. We only know that splits of branch(es) occurred before those of its subbranches, but do not know the ordering of splits of different branches. We know that mutation(s) on the branch occurred before those on its subbranch(es), but do not know the order of ones on the same branch, same subbranch(es), or on different subbranches. For a given sequence data, it may correspond to more than one different rooted tree. For the observed data in Table 1, all the columns are for segregating sites, and there is no transversion. Under the assumption that mutation can only occur at most once at each site and mutation is irreversible, at each segregating site, one and only one of the base types is mutant; the other type is ancestral. So if we know the mutation status at each segregating site, the mutation statuses are said to be labelled, and we can use a 0-1 valued matrix $\mathbf{X} = (x_{ij})$ to denote the observed data, where $x_{ij} = 1$, if the base type of lineage i at site j is mutant, and $x_{ij} = 0$ otherwise. Such 0-1 matrix representation of the data is convenient in the analysis. It is easy to see that each rooted tree uniquely determines a 0-1 valued matrix $\mathbf{X}$, but an arbitrary 0-1 valued matrix may not correspond to a rooted tree. It must satisfy some conditions to corresponds a rooted tree. There are abundant methods and algorithms on how to judge if a given 0-1 values matrix is a valid representation of a rooted tree, and if so how to build the rooted tree (e.g., [12–16]). We find the method that appeared in a number of articles and is stated as Lemma 1 in Gusfiled [16] is easy to use. Given a valid 0-1 valued matrix $\mathbf{X}$ (means it satisfies the condition for representing a tree), one can uniquely draw a rooted tree corresponding to it. Here, uniqueness means the genealogy relationships, including which lineages are in the same branch or subbranch, and so forth and which mutation sites are on which section of which branch or subbranch and so forth, are determined, but the particular shape of the tree, such as some branch put on the left or right side, the angle of branches, their lengths, and so forth, are irrelevant. Thus, there is a 1-1 correspondence between a rooted tree and a valid 0-1 valued matrix. Given the observed data, the mutation statuses at the sites are usually unknown. For data with s segregating sites, there are 2^s different ways to labelling the mutation statues, but most of the labeling matrices do not qualify to be representations of a rooted tree; it is known that there are only $s + 1$ different rooted trees, and hence $s + 1$ different labellings (matrices) correspond to the data, and there are existing algorithms to construct the rooted trees and their corresponding matrices (e.g., [16, 17]). However, we find that the method in GT is convenient. By this method, one first needs to construct one rooted tree from the data or its valid 0-1 valued matrix. For example, start from the least shared mutations labeling that, on each column (site)

of the data, label the less common base type as mutant (the other as ancestral). It is easy to check the conditions for its validity using Lemma 1 mentioned above. Construct the rooted tree corresponding to this matrix and convert it to an unrooted tree as in GT; that is, absorb those subbranches without mutations into their branch(es), and then straighten the branches, subbranches, and so forth. The unrooted tree is uniquely determined from any of the $s+1$ rooted trees.

Then, based on this unrooted tree, one can get all the other rooted trees as in Griffiths and Tavaré [18]; that is, alternatively put the tree root point near each of the vertexes that stretch out that vertex, then arrange the branches, subbranches, and so forth into the desired shapes; if there are more than one mutation between two adjacent vertexes, put the tree root point in the middle of two such adjacent mutations, alternatively for all such pairs of mutations, and shape the tree as above. This way we get all the rooted trees from the unrooted tree. In fact, given any rooted tree, all the other s rooted trees can be constructed in the same way above, without using the unrooted tree. Once the rooted trees are constructed, the corresponding matrix representations are at hand.

3. Asymptotic Behavior of MRCA Estimate

For parameter inference with independent and identically distributed data and sample size n, it is known that the estimator is asymptotically consistent and asymptotically normal with rate $\sqrt{n}$. But for inference of MRCA, the data D_n are not independent and identically distributed, and existing studies indicated that the distribution of the estimated MRCA $t_n \mid D_n$ will not concentrate, even if $n \to \infty$. In the case of estimating the mutation rate with the same data, the estimator is found to be consistent and asymptotically normal with rate $\log^{1/2}(n)$ [1]. This motivates us to investigate the asymptotic behavior of $\bar{t}_n = E(t_n \mid D_n)$ as a commonly used point estimator of the coalescent time. We want to know whether this estimator has similar asymptotic behavior as the mutation rate estimator. We find that such estimators are not consistent almost surely. To describe the result, we consider the data set in three different commonly used forms. The first type of data we consider is in the form of a coalescent tree as in Figure 1. This type of data is often not practical, as for most real data we do not have the information to construct such tree. But as a starting point it will provide us some guide on the result. There are $n-1$ nodes (splitting points) in the tree numbered 2 to n in their time order. Recall the definition of the ith coalescent time w_i. Between the $(i-1)$th and ith node there are exactly i segments, denote them as $w_{i1}, \ldots, w_{ii}$ from left to right, each has length w_i. Assume the number of mutations k_{ij} on segment w_{ij} is known. Let $\mathbf{w} = \{w_{ij} : i = 2, \ldots, n; j = 1, \ldots, i.\}$, $\mathbf{k} = \{k_{ij} : i = 2, \ldots, n; j = 1, \ldots, i.\}$ be the mutation distribution corresponding to $\mathbf{w}$ and $k_i = \sum_{j=1}^{i} k_{ij}$. Here this type of data is fully represented by $\mathbf{k}$. When we do not have $\mathbf{w}$, $\mathbf{k}$ is not uniquely determined. But given each rooted tree $\mathbf{T}_r$, $\mathbf{w}$ and the location information of the mutations, a mutation vector $k_r = \{k_{r,i} : i = 2, \ldots, n; k_{r,i} = \sum_{j=1}^{i} k_{r,ij}\}$ can be constructed by a random manner (to be detailed in Section 4) corresponding to $\mathbf{T}_r$. Denote $\pi(\mathbf{T}_r) = \pi(\mathbf{T}_r \mid D_n) = 1/(s+1)$ $(r = 1, \ldots, s+1)$ be our prior on the rooted tree $\mathbf{T}_r$'s, that is, without additional knowledge we treat each rooted tree as equally likely from the observed data. Here our $\pi(\mathbf{T}_r)$'s have different meaning from the probabilities $p^0(\mathbf{T}_r, \mathbf{n})$'s as in GT (the latter do not sum up to one, but to the probability of obtaining the unrooted tree from the observed data). We have (Appendix)

$$E(t_n \mid D_n, \theta) = \sum_{r=1}^{s+1} E\left[E(t_n \mid \mathbf{k}_r, D_n, \theta)\right] \pi(\mathbf{T}_r \mid D_n) = \frac{1}{s+1} \sum_{r=1}^{s+1} \sum_{i=2}^{n} \frac{E[k_{r,i}] + 1}{i(i+\theta-1)}. \tag{7}$$

The commonly available data is in the form of Table 1, which is equivalent to $s+1$ rooted trees; here $s = |\mathbf{k}| := \sum_{i,j} k_{ij} = |\mathbf{k}_r|$ $(r = 1, \ldots, s+1)$.

The last method is to estimate t_n only by the number of mutations s, without using the information in the rooted trees.

We have the following result (proof in Appendix).

Proposition 1. *(i) One has*

$$E(t_n \mid \mathbf{k}, s, \theta) = 2 \sum_{i=2}^{n} \frac{k_i + 1}{i(i+\theta-1)}; \tag{8}$$

consequently, the above estimator will diverge almost surely, if $\mathbf{k}$ *is treated as random.*

(ii) One has

$$E(t_n \mid D_n, s, \theta) = \frac{2}{s+1} \sum_{r=1}^{s+1} \sum_{i=2}^{n} \frac{E[k_{r,i}] + 1}{i(i+\theta-1)}, \tag{9}$$

and $\widehat{t}_n$ *will converge or not depending on that of the series above.*

(iii) One has

$$E(t_n \mid s, \theta) = \frac{2}{s+1} \sum_{|\mathbf{k}|=s} \sum_{i=2}^{n} \frac{E[k_i] + 1}{i(i+\theta-1)}, \tag{10}$$

and the asymptotic behavior of the above estimator depends on the series above.

Remark 2. The above result tells us that $\widehat{t}_n$ cannot be characterized by an asymptotic deterministic quantity, even for large data size. The estimator is dominated by the number of mutations in the first few coalescent times. Hence, the only practical way to infer the coalescent time is via numerical methods, as the postdata coalescent distribution has no closed form even asymptotically. In contrast, the predata mean $E(t_n) = 2(1 - 1/n) \to 2$ is convergent but is inaccurate as an estimator of the coalescence time for the population under study.

4. The Proposed Method

The method is to construct the mutation vector $k_r = (k_{r,2}, \ldots, k_{r,n})$ and compute the data probability directly from

the genealogy rooted tree $\mathbf{T}_r$'s. Suppose that there are s segregating sites in the sequence data, which is exactly the total number of mutations occurred in the history of the n sampled individuals, then there are $s+1$ different rooted trees $\mathbf{T}_r$'s compatible with the data. Each of the rooted trees is a fixed genealogy structure, with the multiplicities as the leaves, but the number of mutations among the tree segments is random, subject to the total number of mutations being s. The structure consists of the tree branches, subbranches within each branches, sub-subbranches, and so on,, and the leaves. These are the fixed features of a rooted tree. Given the data, the rooted tree is a display of how the s mutations are distributed along the lineages, but there is no time scale in the tree, so (5) cannot be used to compute the mutation probabilities. Each rooted tree tells us a partial ordering of the mutations. For example, in the rooted tree, we know mutations at sites 4, 6, and 14 occurred before the split of lineages a, b, e, and f, thus occurred before the mutations at sites 1, 5, and 10. But we do not know which of 4, 6, and 14 occurred first. We know mutation 1 occurred before 10, but we do not know the order of 1 and 5, and so forth. If we have the full data $(\mathbf{k}_r, \mathbf{w})$ corresponding to all the rooted trees, $\mathbf{T}_r$'s, we can compute $\widehat{t}_n = E(t_n \mid D_n, \theta)$ as in Proposition 1(ii). But $\mathbf{w}$ and the $\mathbf{k}_r$s are not directly available; however, $\mathbf{w}$ can be easily simulated by the prior exponential distribution, and each rooted tree $\mathbf{T}_r$ has an initial mutation distribution on its branch segments. Denote by $s_{ij\ldots}$ the $(i, j, \ldots)$th segment (the order is arbitrary, e.g., we can lable them from upper to lower and left to right locations), and let $|s_{ij\ldots}|$ be the number of mutations on it (many of them are zeros; we can concentrate on the segments with nonzero mutations). Denote $\mathbf{s} = \{s_{ij\ldots}\}$. Given $(\mathbf{w}, \mathbf{s})$, k_r can be sampled from $\mathbf{T}_r$ (to be detailed latter). Let $E_{(\mathbf{w},\mathbf{k}_r)}$ be the expectation with respect to $(\mathbf{w}, \mathbf{k}_r)$. The above motivates us to estimate t_n by

$$\widehat{t}_n = E\left(t_n \mid D_n, \theta\right) = \frac{2}{s+1} \sum_{r=1}^{s+1} \sum_{i=2}^{n} E_{(\mathbf{w},\mathbf{k}_r)} \left[\frac{k_{r,i}+1}{i\,(i+\theta-1)} \right]. \quad (11)$$

The above expectation is not easy to compute directly since we do not know the joint distribution of $(\mathbf{w}, \mathbf{k}_r)$. Instead we use simulation method. For this, we sample $\mathbf{w}^{(1)} \cdots \mathbf{w}^{(M)}$ independently and generate $k_r^{(m)} = \{k_{r,i}^{(m)}\}$ (see below) corresponding to $\mathbf{w}^{(m)}$ and $\mathbf{T}_r$ for each r then approximate $\widehat{t}_n$ as

$$\widehat{t}_n \approx \frac{2}{M} \sum_{m=1}^{M} \frac{1}{s+1} \sum_{r=1}^{s+1} \sum_{i=2}^{n} \frac{k_{r,i}^{(m)}+1}{i\,(i+\theta-1)}. \quad (12)$$

Now we consider generating $k_r^{(m)}$. After $w^{(m)} = (w_2^{(m)}, \ldots, w_n^{(m)})$ is allocated among the branches of $\mathbf{T}_r$, we only need to consider each segment $s_{ij\ldots}$ with nonzero number t_r of mutations in them. Each length of $s_{ij\ldots}0$ in $\mathbf{T}_r$ is the summation of some $d = d_{ij\ldots}$ of the $w_j^{(m)}$'s. For simplicity of exposition and notation, suppose that they are $w_2^{(m)}, \ldots, w_{d+1}^{(m)}$; then given the t_r mutations in $[0, w_2^{(m)} + \cdots + w_{d+1}^{(m)}]$ and using (5), it is easy to see that the number of mutations $\widetilde{k}_r = (\widetilde{k}_{r,1}, \ldots, \widetilde{k}_{r,d})$ in each of the d intervals $[0, w_2^{(m)}]$, $[w_2^{(m)}, w_2^{(m)} + w_3^{(m)}]$,..., $[w_2^{(m)} + \cdots + w_{d-1}^{(m)}, w_2^{(m)} + \cdots + w_d^{(m)}]$ follows the multinomial distribution

$$\begin{aligned} P\left(\widetilde{k}_r = (k_{r,1}, \ldots, k_{r,d}) \mid t_r\right) &= M\left(k_{r,1}, \ldots, k_{r,d}; t_r, q_1, \ldots, q_d\right) \\ &= \frac{t_r!}{k_{r,1}! \cdots k_{r,d}!} q_1^{k_1} \cdots q_d^{k_d}, \end{aligned} \quad (13)$$

where $q_j = w_{j+1}/(w_2 + \cdots + w_{d+1})$ $(j = 1, \ldots, d)$. After all the nonzero $t_r = |s_{ij\ldots}|$'s are allocated in the corresponding intervals, we have

$$k_{r,i}^{(m)} = \sum_{(i)} \widetilde{k}_{r,l}, \quad (i = 2, \ldots, n), \quad (14)$$

where the summation is for all $\widetilde{k}_{r,l}$'s that fall in $[w_2^{(m)} + \cdots + w_i^{(m)}, w_2^{(m)} + \cdots + w_{i+1}^{(m)}]$.

Specifically, the simulation method is as below. For $m = 1, \ldots, M$, do the following steps.

(i) Sample $w^{(m)} = (w_2^{(m)}, \ldots, w_n^{(m)})$ from the coalescent distribution as in (1); that is, the $w_i^{(m)}$'s are independent, with $w_i^{(m)} \sim \exp(i(i-1)/2)$. Or equivalently, sample $u \sim U(0,1)$ and set $w_i^{(m)} = -2/(i(i-1))\ln(1-u)$.

(ii) For each fixed $1 \le r \le s+1$, allocate $w^{(m)}$ to the $n-1$ coalescent events of the n sequences based on each rooted tree $\mathbf{T}_r$. See illustration below for details.

(iii) Allocate the $\widetilde{k}_r^{(m)}$ mutations in the corresponding segments according to (13). Then get the $k_{r,i}^{(m)}$'s as in (14).

After all the M iterations, evaluate (12) until convergence, which can be assessed by relative error, for example.

Illustration: Allocate $\mathbf{w}^{(m)}$ to the $n-1$ coalescent events of the n sequences based on rooted tree. We use the backward method; that is, first allocate $w_n^{(m)}$, then $w_{n-1}^{(m)}$, ..., and last $w_2^{(m)}$. Consider the rooted tree, for example. There are $n = 55$ sequences, with frequencies $(3, 1, 19, 2, 2, 1, 5, 1, 1, 1, 4, 8, 8, 3)$ for lineages $(m, n, e, b, a, f, k, c, h, g, i, j, l, d)$. Note that sequences (leaves) in each lineage (branch) only coalescence within each branch (if the branch has more than one leaves), and branch with a single leaf coalescences only at MRCA $w_2^{(m)}$. We first decide $w_{55}^{(m)}$ goes to which branch or pairs of single branches. Since it is the latest coalescent time, it can only go to a pair of leaves in some branch with multiple leaves. Since branch n has only 1 leaf; it is excluded at this step. The remaining branches $(m, \langle e, b, a, f\rangle, k, \langle c, h, g, i, j, l\rangle, d)$ all have multiple leaves with a total of 54. We assign $w_{55}^{(m)}$ to one of these branches with weights proportional to their number of leaves, that is, with probabilities $(3, 24, 5, 19, 3)/54$. Suppose that $w_{55}^{(m)}$ is assigned to $\langle e, b, a, f\rangle$; we need to decide which subbranch it goes to. We have three candidate subbranches (e, b, a) with number of leaves $(19, 2, 2)$. We randomly assign $w_{55}^{(m)}$ to them with weights $(19, 2, 2)/23$. Suppose it is assigned

to branch e; then $w_{55}^{(m)}$ will go to a pair within this branch, and which pair is irrelevant. But the pair will be treated as a single leaf in assigning the rest $w_j^{(m)}$'s. So after this step, we reassign the number of leaves in e as 18.

Now we assign $w_{54}^{(m)}$. The procedure is the same as above; the only difference is now e has 18 leaves. The candidate branches are still $(m, \langle e, b, a, f\rangle, k, \langle c, h, g, i, j, l\rangle, d)$ with weights $(3, 23, 5, 19, 3)/53$. Suppose that $w_{54}^{(m)}$ is also allocated to e of branch $\langle e, b, a, f\rangle$; then e has 17 leaves now.

We now allocate $w_{53}^{(m)}$ to candidates $(m, \langle e, b, a, f\rangle, k, \langle c, h, g, i, j, l\rangle, d)$ with weights $(3, 22, 5, 19, 3)/52$. Suppose that $w_{53}^{(m)}$ is allocated to $\langle c, h, g, i, j, l\rangle$; we need to decide which of the 4 subbranches it will go to. c has only 1 leaf and is excluded. So we allocate subbranches $(\langle h, g\rangle, \langle i, j\rangle, l)$ with weights $(2, 12, 4)/18$. Supposing it goes to $\langle h, g\rangle$, since it has only one pair of leaves, then h and g are merged as one leaf after this assignment.

Continue this way, until $w_2^{(m)}$ is allocated. Then all the branches in this rooted tree have lengths as the $w_i^{(m)}$'s allocated to them. After this step, the length of each segment $s_{ij\ldots}$ of $\mathbf{T}_r$ is a summation of some $w_i^{(m)}$'s. Since $|s_{ij\ldots}|$ is known from each $\mathbf{T}_r$, we can allocate each of the $\widetilde{k}^{(m)}$'s by (13), then get the $k_i^{(m)}$'s by the formula that follows it. Then compute (12).

The assumption that the population size N is constant can be relaxed the same way as in GS and Tavaré et al. [10].

Appendix

Proof of the Proposition

(i) Recall that the k_{ij}'s are independent, the k_i's are independent, and the w_i's are independent with $w_i \sim \exp(i(i-1)/2)$, $E(w_i) = 2/(i(i-1)/2)$, $k_{ij} \mid w_i \sim \mathrm{Po}(\cdot, w_i\theta/2)$, $k_i|w_i \sim \mathrm{Po}(\cdot, iw_i\theta/2)$, and so $E(k_i) = E(E(k_i \mid w_i)) = \theta/(i-1)$. Observe

$$\begin{aligned} P(w_i \mid D_n, \theta) &= P(w_i \mid \mathbf{k}, \theta) = P(w_i \mid k_i, \theta) \\ &\propto P(w_i)\, P(k_i \mid w_i, \theta) \\ &= \frac{i(i-1)}{2} \exp\left(-\frac{i(i-1)\, w_i}{2}\right) \\ &\quad \times \mathrm{Po}\left(k_i, \frac{iw_i\theta}{2}\right) \\ &\propto (w_i\theta)^{k_i} \exp\left(-\frac{i(i+\theta-1)\, w_i}{2}\right). \end{aligned} \tag{A.1}$$

The right-hand side above is the density of $\Gamma(k_i+1, 2/(i(i+\theta-1))$ distribution, up to an normalizing constant. It has mean $E(w_i \mid \mathbf{k}, \theta) = 2(k_i+1)/(i(i+\theta-1))$ and variance $\mathrm{Var}(w_i \mid \mathbf{k}, \theta) = 4(k_i+1)/(i^2(i+\theta-1)^2)$. Since $t_n = \sum_{i=2}^n w_i$, we have

$$\begin{aligned} E(t_n \mid \mathbf{k}, \theta) &= \sum_{i=1}^n E(w_i \mid \mathbf{k}, \theta) \\ &= 2\sum_{i=2}^n \frac{k_i+1}{i(i+\theta-1)} \\ &= S_n \sum_{i=2}^n a_{n,i} x_i + b_n, \end{aligned} \tag{A.2}$$

where $x_i = (i-1)k_i$, the x_i's are i.i.d with $E(x_i) = \theta$, $a_{n,i} = a_{n,i}(\theta) = S_n^{-1}(\theta) 2/[i(i-1)(i+\theta-1)]$, $S_n = S_n(\theta) = 2\sum_{i=2}^n 1/[i(i-1)(i+\theta-1)]$, and $b_n(\theta) = 2\sum_{i=2}^n 1/[i(i+\theta-1)]$. Note that $\{S_n(\theta)\}$ and $\{b_n(\theta)\}$ are convergent sequences with

$$\begin{aligned} S_n(\theta) &\longrightarrow S(\theta) := \sum_{i=2}^{\infty} \frac{2}{i(i-1)(i+\theta-1)} < \infty, \\ b_n(\theta) &\longrightarrow b(\theta) := \sum_{i=2}^{\infty} \frac{2}{i(i+\theta-1)} < \infty. \end{aligned} \tag{A.3}$$

Note also that $\sum_{i=2}^n a_{n,i} = 1$, that is, $\{a_{n,i}\}$ is a weight sequence for each fixed n. Since $\lim_n a_{n,i} > 0$ for fixed i, by the results for weighted sum of i.i.d. random variables (see [19], for a review of such results), a necessary condition for $\sum_{i=2}^n a_{n,i} x_i$ to converge (a.s.) is that $\lim_n a_{n,i} = 0$ for all fixed i, or no term will be dominant as $n \to \infty$. Since this condition is not satisfied, the first few terms are dominant and we have

$$\begin{aligned} &\sum_{i=2}^n a_{n,i} x_i \text{ diverges (a.s.)}, \\ &\text{or equivalently } E(t_n \mid \mathbf{k}, \theta) \text{ diverges (a.s.)}. \end{aligned} \tag{A.4}$$

The proofs of part (ii) and (iii) are similar to that of part (i) and omitted.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] A. Yuan, Z. Zhang, G. Liu, and G. Chen, "On the estimation of mutation rate based on coalescence genealogy," *Journal of Advanced Bioinformatics Applications and Research*. In press.

[2] R. C. Griffiths and S. Tavaré, "Ancestral inference in population genetics," *Statistical Science*, vol. 9, no. 3, pp. 307–319, 1994.

[3] R. H. Ward, B. L. Frazier, K. Dew-Jager, and S. Pääbo, "Extensive mitochondrial diversity within a single Amerindian tribe," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 19, pp. 8720–8724, 1991.

[4] R. R. Hudson, "Gene genealogies and the coalescent process," in *Oxford Surveys in Evolutionary Biology*, D. Futuyma and J. Antonovics, Eds., pp. 1–44, Oxford University Press, New York, NY, USA, 1991.

[5] P. Donnelly and S. Tavaré, "Coalescents and genealogical structure under neutrality," *Annual Review of Genetics*, vol. 29, pp. 401–421, 1995.

[6] J. F. C. Kingman, "On the genealogy of large populations," *Journal of Applied Probability*, vol. 19, pp. 27–43, 1982.

[7] J. F. C. Kingman, "Exchangeability and the evolution of large populations," in *Exchangeability in Probability and Statistics*, G. Koch and F. Spizzchino, Eds., pp. 97–112, North-Holland, Amsterdam, The Netherlands, 1982.

[8] S. Tavaré, "Ancestral inference from DNA sequence data," in *Case Studies in Mathematical Modeling: Ecology, Physiology and Cell Biology*, H. G. Othmer, F. R. Adler, M. A. Lewis, and J. Dallon, Eds., chapter 5, pp. 81–96, Prentice Hall, New York, NY, USA, 1997.

[9] G. A. Watterson, "Mutant substitutions at linked nucleotide sites," *Advances in Applied Probability*, vol. 14, no. 2, pp. 206–224, 1982.

[10] S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly, "Inferring coalescence times from DNA sequence data," *Genetics*, vol. 145, no. 2, pp. 505–518, 1997.

[11] S. N. Ethier and R. C. Griffiths, "The infinitely-many-sites model as a measure valued diffusion," *Annals of Probability*, vol. 15, no. 2, pp. 515–545, 1987.

[12] J. Camin and R. Sokal, "A method for deducing branching sequences in phylogeny," *Evolution*, vol. 19, no. 3, pp. 311–326, 1965.

[13] J. S. Farris, "Inferring phylogenetic trees from chromosome inversion data," *Systematic Zoology*, vol. 27, pp. 275–284, 1967.

[14] K. S. Booth and G. S. Lueker, "Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms," *Journal of Computer and System Sciences*, vol. 13, no. 3, pp. 335–379, 1976.

[15] J. Felsenstein, "Numerical methods for inferring evolutionary trees," *The Quarterly Review of Biology*, vol. 57, no. 4, pp. 379–404, 1982.

[16] D. Gusfield, "Efficient algorithms for inferring evolutionary trees," *Networks*, vol. 21, no. 1, pp. 19–28, 1991.

[17] R. C. Griffiths, "An algorithm for constructing genealogical trees," Statistics Research Report 163, Department of Mathematics, Monash University, Melbourne, Australia, 1987.

[18] R. C. Griffiths and S. Tavaré, "Unrooted genealogical tree probabilities in the infinitely-many-sites model," *Mathematical Biosciences*, vol. 127, no. 1, pp. 77–98, 1995.

[19] N. H. Bingham, "Extensions of the strong law," *Advances in Applied Probability*, pp. 27–36, 1986.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 420412, 11 pages
http://dx.doi.org/10.1155/2013/420412

Research Article

Applications of Bayesian Gene Selection and Classification with Mixtures of Generalized Singular g-Priors

Wen-Kuei Chien[1] and Chuhsing Kate Hsiao[2,3]

[1] *Biostatistics Center, Taipei Medical University, Taipei 11031, Taiwan*
[2] *Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei 10055, Taiwan*
[3] *Bioinformatics and Biostatistics Core, Division of Genomic Medicine, Research Center for Medical Excellence, National Taiwan University, Taipei 10055, Taiwan*

Correspondence should be addressed to Chuhsing Kate Hsiao; ckhsiao@ntu.edu.tw

Received 4 September 2013; Revised 10 November 2013; Accepted 10 November 2013

Academic Editor: Ao Yuan

Recent advancement in microarray technologies has led to a collection of an enormous number of genetic markers in disease association studies, and yet scientists are interested in selecting a smaller set of genes to explore the relation between genes and disease. Current approaches either adopt a single marker test which ignores the possible interaction among genes or consider a multistage procedure that reduces the large size of genes before evaluation of the association. Among the latter, Bayesian analysis can further accommodate the correlation between genes through the specification of a multivariate prior distribution and estimate the probabilities of association through latent variables. The covariance matrix, however, depends on an unknown parameter. In this research, we suggested a reference hyperprior distribution for such uncertainty, outlined the implementation of its computation, and illustrated this fully Bayesian approach with a colon and leukemia cancer study. Comparison with other existing methods was also conducted. The classification accuracy of our proposed model is higher with a smaller set of selected genes. The results not only replicated findings in several earlier studies, but also provided the strength of association with posterior probabilities.

1. Introduction

Recent advancement in oligonucleotide microarray technologies has resulted in production of thousands of gene expression levels in a single experiment. With such vast amount of data, one major task for researchers is to develop classification rules for prediction of cancers or cancer subtypes based on gene expression levels of tissue samples. The accuracy of such classification rules may be crucial for diagnosis and treatment, since different cancer subtypes may require different target-specific therapies. However, the development of good and efficient classification rules has not been straightforward, either because of the huge number of genes collected from a relatively small number of tissue samples or because of the model complexity associated with the biological mechanism. The identification of a smaller set of relevant genes to characterize different disease classes, therefore, has been a challenging task. Procedures which are efficient in gene selection as well as in classification do play an important role in cancer research.

Many approaches have been proposed for classes classification. For example, several analyses identified a subset of classifying genes with t-statistics, regression model approach, mixture model, Wilcoxon score test, or the between-within classes sum of squares (BSS/WSS) [1–7]. These methods are univariate in the sense that each gene is tested individually. Others started with an initial step of dimension reduction before classification procedures, such as the principle components analysis (PCA) [8–10] and the partial least squares algorithm (PLS algorithm) [11–15]. These methods may reduce dimension (the number of genes) effectively but may not be biologically interpretable. To capture the gene-gene correlations, researchers proposed the pair-based method [16], correlation-based feature selection [17], and the Markov

random field prior [18]. Although these methods can model the gene-gene interaction, they can be computationally time-consuming.

Bayesian approach can accommodate naturally the interplay between genes via prior distributions, under the setting of regression models. Examples included the Bayesian hierarchical mixture model [19–21] and a logistic or probit link with latent variables and stochastic search variable selection (SSVS) procedure for binary and multicategorical phenotypes [22–25]. To consider all genes simultaneously, most Bayesian approaches adopt a multivariate analysis with a natural conjugate prior $N(\mathbf{0}, c(\mathbf{X}^T\mathbf{X})^{-1})$, called g-prior, for the regression parameters $\boldsymbol{\beta}$ [26]. This *a priori* distribution utilizes the design matrix as the prior covariance matrix of $\boldsymbol{\beta}$ and can lead to a relatively simple posterior distribution. However, if the number of genes is much larger than the number of samples available, the dimension of $\mathbf{X}$ becomes large and a high degree of multicollinearity may occur. In that case, the covariance matrix of Zellner's g-prior becomes nearly singular. Modifications included the gsg-prior distribution with the Moore-Penrose generalized inverse matrix [27] and use of a ridge parameter [28, 29]. Alternatively, other researchers focused on the scalar c in $c(\mathbf{X}^T\mathbf{X})^{-1}$ which controls the expected size of the nonzero regression coefficients. For instance, it was reported that the final results are insensitive to the values of c between 10 and 100, and the value $c = 100$ has been suggested after extensive examinations [30]. Instead of fixing c at a constant, George and Foster [31] proposed an empirical Bayes estimate for c, while Liang and colleagues [32] suggested a hyper-g prior, a special case of the incomplete inverse-gamma prior in Cui and George [33].

The main purpose of this research is the application of fully Bayesian approaches with a hyperprior on c. Specifically we adopted an inverse-gamma prior IG$(1/2, n/2)$ which was commented earlier that it could lead to computational difficulty. Therefore, we outlined a MCMC algorithm and demonstrated its implementation. In this paper, we considered a probit regression model for classification with SSVS to identify the influential genes, augmented the response variables $Y_1, Y_2, \ldots, Y_n$ with latent variables $Z_1, Z_2, \ldots, Z_n$, and converted the probit model to a Gaussian regression problem with the generalized singular g-prior (gsg-prior). For the choice of c, we assigned a hyperprior for the uncertainty in c. This hyperprior is intuitive and differs from those in [32, 33]. Finally, we defined an indicator variable γ_j for the jth gene and perform MCMC methods to generate posterior samples for gene selection and class classification. The rest of the paper is arranged as follows. In Section 2, we briefly described the model specification including the data augmentation approach and SSVS methods. Under this hyperprior on c, we also demonstrated the implementation of the Bayesian inference. Applications of three cancer studies, acute leukemia, colon cancer, and large B-cell lymphoma (DLBCL), were presented in Section 3. Conclusion and discussion were given in Section 4.

2. Model and Notation

Let $(\mathbf{X}, \mathbf{Y})$ indicate the observed data,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \tag{1}$$

where x_{ij} denotes the expression level of the jth gene from the ith sample and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ denotes the response vector, where $Y_i = 1$ indicates that sample i is a cancer tissue and $Y_i = 0$ for normal tissue. Assume that $Y_1, Y_2, \ldots, Y_n$ are n independent random variables with $p_i = \Pr(Y_i = 1)$.

2.1. Probit Model with Latent Variable. The gene expression measurements can be linked to the response outcome with a probit regression model:

$$p_i = \Pr\left(Y_i = 1\right) = \Phi\left(\alpha + \mathbf{X}_i\boldsymbol{\beta}\right), \tag{2}$$

where α represents the intercept, $\mathbf{X}_i$ is the ith row in the $n \times p$ design matrix $\mathbf{X}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression coefficients, and Φ is the standard normal cumulative distribution function.

To perform statistical inference under this probit regression model, we first adopt n independent latent variables $Z_1, Z_2, \ldots, Z_n$, where

$$Z_i = \alpha + \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1), \ i = 1, \ldots, n, \tag{3}$$

and the Z_i corresponds to the disease status as

$$Y_i = \begin{cases} 1, & \text{if } Z_i > 0, \\ 0, & \text{if } Z_i \le 0. \end{cases} \tag{4}$$

The use of such latent variables helps to determine which category the ith sample is to be classified. Note that multiplying a constant on both sides in (3) does not change the model; thus a unit variance is considered for ε_i.

If a noninformative prior is assumed for $\boldsymbol{\beta}$, then the posterior covariance matrix of $\boldsymbol{\beta}$ given $\mathbf{Z} \equiv (Z_1, Z_2, \ldots, Z_n)$ becomes $(\mathbf{X}^T\mathbf{X})^{-1}$. However, due to the enormous size of microarray data, $(\mathbf{X}^T\mathbf{X})^{-1}$ may be nearly singular, and variable selection for dimension reduction is needed. We define for variable selection the vector $\boldsymbol{\gamma} \equiv (\gamma_1, \gamma_2, \ldots, \gamma_p)$ whose elements are all binary, where

$$\gamma_i = \begin{cases} 1, & \text{if } \beta_i \neq 0 \ (\text{the } i\text{th gene selected}), \\ 0, & \text{if } \beta_i = 0 \ (\text{the } i\text{th gene not selected}). \end{cases} \tag{5}$$

Given $\boldsymbol{\gamma}$, we denote p^γ as the number of 1's in $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}^\gamma$ a $p^\gamma \times 1$ reduced vector containing the regression coefficients β_j if its corresponding γ_j is 1. Accordingly, for all $\gamma_j = 1$, the corresponding columns in $\mathbf{X}$ are collected to build $\mathbf{X}^\gamma$, an $n \times p^\gamma$ reduced gene expression matrix. Given $\boldsymbol{\gamma}$, the probit regression model in (3) can be written as

$$Z_i = \alpha + \mathbf{X}_i^\gamma\boldsymbol{\beta}^\gamma + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1), \ i = 1, \ldots, n, \tag{6}$$

where $\mathbf{X}_i^\gamma$ is the ith row in $\mathbf{X}^\gamma$.

2.2. Choice of Prior Distributions. To complete the model specification, we assign a normal $N(0,h)$ prior for the intercept α with a large h indicating no *a priori* information. For the regression parameters, the commonly applied g-prior $\boldsymbol{\beta}^{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, c \sim N(\mathbf{0}, c(\mathbf{X}^{\boldsymbol{\gamma}T}\mathbf{X}^{\boldsymbol{\gamma}})^{-1})$ may not work if the sample size n is less than the number $p^{\boldsymbol{\gamma}}$, leading to the results that $\mathbf{X}^{\boldsymbol{\gamma}T}\mathbf{X}^{\boldsymbol{\gamma}}$ is not of full rank and $(\mathbf{X}^{\boldsymbol{\gamma}T}\mathbf{X}^{\boldsymbol{\gamma}})^{-1}$ does not exist. Therefore, we consider the *gsg*-prior distribution with $(\mathbf{X}^{\boldsymbol{\gamma}T}\mathbf{X}^{\boldsymbol{\gamma}})^{+}$ as the pseudoinverse of $\mathbf{X}^{\boldsymbol{\gamma}T}\mathbf{X}^{\boldsymbol{\gamma}}$ for $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ conditioning on $(\boldsymbol{\gamma}, c)$, $\boldsymbol{\beta}^{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, c \sim N(\mathbf{0}, c(\mathbf{X}^{\boldsymbol{\gamma}T}\mathbf{X}^{\boldsymbol{\gamma}})^{+})$. This would solve the singularity problem. Next, we assign for $\boldsymbol{\gamma}$ and c the priors

$$\begin{aligned} \pi(c) &= \frac{(n/2)^{1/2}}{\Gamma(1/2)} c^{-3/2} e^{-n/(2c)}, \\ \gamma_i &\sim \text{Ber}(\pi_i), \quad 0 \le \pi_i \le 1, \ i = 1, \ldots, p, \end{aligned} \tag{7}$$

and assume that γ_i are independent for $i = 1, \ldots, p$. Note that here the π_i's are of small values, implying a small set of influential genes.

We now complete the model specification:

$$\begin{aligned} &\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T, \quad \text{where} \\ &p_i = \Pr(Y_i = 1) = \Phi(\alpha + \mathbf{X}_i\boldsymbol{\beta}), \\ &Z_i = \alpha + \mathbf{X}_i^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}} + \varepsilon_i, \quad \text{where} \\ &Y_i = 1 \text{ if } Z_i > 0, \text{ and 0 otherwise} \\ &\boldsymbol{\beta}^{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, c \sim N\left(\mathbf{0}, c\left(\mathbf{X}^{\boldsymbol{\gamma}T}\mathbf{X}^{\boldsymbol{\gamma}}\right)^{+}\right), \\ &\pi(c) \sim \text{IG}\left(\frac{1}{2}, \frac{n}{2}\right), \\ &\gamma_i \sim \text{Ber}(\pi_i). \end{aligned} \tag{8}$$

Note that $Y_i = 1$ if the ith sample is a cancer tissue, α is the intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression coefficients, Φ is the standard normal cumulative distribution function, and $\mathbf{X}$ is the design matrix:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}. \tag{9}$$

And $\boldsymbol{\gamma} \equiv (\gamma_1, \gamma_2, \ldots, \gamma_p)$ contains the binary γ_i, where $\gamma_i = 1$ if the ith gene is selected ($\beta_i \neq 0$), $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ is a $p^{\boldsymbol{\gamma}} \times 1$ reduced vector containing the regression coefficients β_j if its corresponding γ_j is 1, $p^{\boldsymbol{\gamma}}$ is the number of 1's in $\boldsymbol{\gamma}$, and $\mathbf{X}_i^{\boldsymbol{\gamma}}$ is the ith row in $\mathbf{X}^{\boldsymbol{\gamma}}$.

2.3. Computation and Posterior Inference. Based on the prior distributions specified in previous sections, the joint posterior distribution can be derived as

$$\begin{aligned} &P(\mathbf{Z}, \alpha, \boldsymbol{\beta}^{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, c \mid \mathbf{Y}, \mathbf{X}) \\ &\propto \left[\exp\left\{-\frac{\sum_{i=1}^{n}(Z_i - \alpha - \mathbf{X}_i^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}})^2}{2}\right\}\prod_{i=1}^{n} I(A_i)\right] \\ &\cdot \exp\left(-\frac{\alpha^2}{2h}\right) \\ &\cdot \left[\exp\left(-\frac{\boldsymbol{\beta}^{\boldsymbol{\gamma}T}\mathbf{X}^{\boldsymbol{\gamma}T}\mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}}}{2c}\right)\prod_{i=1}^{m_{\boldsymbol{\gamma}}}\lambda_i^{-1/2}\right] \\ &\cdot \left[\prod_{i=1}^{p}\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}\right] \\ &\cdot \left[c^{-3/2}\exp\left(-\frac{n}{2c}\right)\right], \end{aligned} \tag{10}$$

where

$$A_i = \begin{cases} \{Z_i : Z_i > 0\} & \text{if } Y_i = 1, \\ \{Z_i : Z_i \le 0\} & \text{if } Y_i = 0, \end{cases} \tag{11}$$

and $\lambda_1, \lambda_2, \ldots, \lambda_{m_{\boldsymbol{\gamma}}}$ ($m_{\boldsymbol{\gamma}} \le p_{\boldsymbol{\gamma}}$) are the nonzero eigenvalues of $(\mathbf{X}^{\boldsymbol{\gamma}T}\mathbf{X}^{\boldsymbol{\gamma}})^{+}$. From (10), $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ given $(\mathbf{Z}, \alpha, \boldsymbol{\gamma}, c, \mathbf{Y}, \mathbf{X})$ is a multivariate normal distribution with a covariance matrix $c(\mathbf{X}^{\boldsymbol{\gamma}T}\mathbf{X}^{\boldsymbol{\gamma}})^{+}/(c+1)$. In the case where $\mathbf{X}^{\boldsymbol{\gamma}}$ is not of full column rank, the problem of convergence may occur in the MCMC algorithm because the covariance matrix is not positive definite and the multivariate normal distribution becomes degenerated. To avoid this problem and speed up the computations, we integrate out α and $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ in (10) following Yang and Song's [27] suggestion and derive

$$\begin{aligned} &p(\mathbf{Z}, \boldsymbol{\gamma}, c \mid \mathbf{Y}, \mathbf{X}) \\ &\propto \frac{1}{\left|\Sigma_{\boldsymbol{\gamma}}\right|^{1/2}}\exp\left(-\frac{\mathbf{Z}^T\Sigma_{\boldsymbol{\gamma}}^{-1}\mathbf{Z}}{2}\right)\prod_{i=1}^{n} I(A_i) \\ &\cdot \prod_{i=1}^{p}\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i} c^{-3/2} e^{-n/2c}, \end{aligned} \tag{12}$$

where $\Sigma_{\boldsymbol{\gamma}} = \mathbf{I}_n + h\mathbf{1}\mathbf{1}^T + c\mathbf{X}^{\boldsymbol{\gamma}}(\mathbf{X}^{\boldsymbol{\gamma}T}\mathbf{X}^{\boldsymbol{\gamma}})^{+}\mathbf{X}^{\boldsymbol{\gamma}T}$. As the posterior distribution is not available in an explicit form, we use the MCMC technique to obtain posterior sample observations. The computational sampling scheme is as follows.

(1) Draw $\mathbf{Z}$ from $p(\mathbf{Z} \mid \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}, c)$, where

$$p(\mathbf{Z} \mid \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}, c) \propto N\left(\mathbf{0}, \Sigma_{\boldsymbol{\gamma}}\right)\prod_{i=1}^{n} I(A_i). \tag{13}$$

The conditional distribution of $\mathbf{Z}$ given $(\mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}, c)$ is a multivariate truncated normal. Since it is difficult to directly sample $\mathbf{Z}$ from this distribution, we draw

samples Z_i, $i = 1, \ldots, n$, from $p(Z_i \mid \mathbf{Z}_{(-i)}, \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}, c)$, where $\mathbf{Z}_{(-i)}$ is the vector of $\mathbf{Z}$ without the ith element [34].

(2) Draw $\boldsymbol{\gamma}$ from $p(\boldsymbol{\gamma} \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, c)$, where

$$p(\boldsymbol{\gamma} \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, c) \propto \frac{1}{\left|\Sigma_{\boldsymbol{\gamma}}\right|^{1/2}} \exp\left(-\frac{\mathbf{Z}^T \Sigma_{\boldsymbol{\gamma}}^{-1} \mathbf{Z}}{2}\right) \prod_{i=1}^{p} \pi_i^{\gamma_i} (1-\pi_i)^{1-\gamma_i}. \tag{14}$$

Similar to the above procedure, we draw samples γ_i, $i = 1, \ldots, n$, from $p(\gamma_i \mid \boldsymbol{\gamma}_{(-i)}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, c)$. It can be shown that

$$\begin{aligned} & p\left(\gamma_i \mid \boldsymbol{\gamma}_{(-i)}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, c\right) \\ &= \frac{p\left(\gamma_i = 1 \mid \boldsymbol{\gamma}_{(-i)}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, c\right)}{p\left(\gamma_i = 1 \mid \boldsymbol{\gamma}_{(-i)}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, c\right) + p\left(\gamma_i = 0 \mid \boldsymbol{\gamma}_{(-i)}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, c\right)} \\ &= \left(1 + \frac{1-\pi_i}{\pi_i} \rho\right)^{-1}, \end{aligned} \tag{15}$$

where

$$\begin{aligned} \rho &= \left|\Sigma_{\boldsymbol{\gamma}^1} \Sigma_{\boldsymbol{\gamma}^0}^{-1}\right|^{1/2} \exp\left\{\frac{\mathbf{Z}^T \left(\Sigma_{\boldsymbol{\gamma}^1}^{-1} - \Sigma_{\boldsymbol{\gamma}^0}^{-1}\right) \mathbf{Z}}{2}\right\}, \\ \boldsymbol{\gamma}^1 &= \left(\gamma_1, \ldots, \gamma_{i-1}, \gamma_i = 1, \gamma_{i+1}, \ldots, \gamma_p\right), \\ \boldsymbol{\gamma}^0 &= \left(\gamma_1, \ldots, \gamma_{i-1}, \gamma_i = 0, \gamma_{i+1}, \ldots, \gamma_p\right), \end{aligned} \tag{16}$$

$\Sigma_{\boldsymbol{\gamma}}^1$ and $\Sigma_{\boldsymbol{\gamma}}^0$ are similar to $\Sigma_{\boldsymbol{\gamma}}$ with $\boldsymbol{\gamma}$ replaced by $\boldsymbol{\gamma}^1$ and $\boldsymbol{\gamma}^0$, respectively.

(3) Draw c from $p(c \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\gamma})$, where

$$\begin{aligned} p(c \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\gamma}) &\propto p(\mathbf{Z}, \boldsymbol{\gamma}, c \mid \mathbf{Y}, \mathbf{X}) \\ &\propto \frac{1}{\left|\Sigma_{\boldsymbol{\gamma}}\right|^{1/2}} \exp\left(-\frac{\mathbf{Z} \Sigma_{\boldsymbol{\gamma}}^{-1} \mathbf{Z}}{2}\right) \cdot c^{-3/2} e^{-n/2c}. \end{aligned} \tag{17}$$

The above distribution does not belong to any standard distribution, so we will use Metropolis-Hastings algorithm to sample c.

The iteration therefore starts with initial values of $\mathbf{Z}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$, and $c^{(0)}$, and our MCMC procedures at the tth iteration are as follows.

Step 1. Draw $Z_i^{(t)}$ from $p(Z_i \mid \mathbf{Z}_{(-i)}^{(t-1)}, \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}^{(t-1)}, c^{(t-1)})$, $i = 1, \ldots, n$.

Step 2. For $i = 1, \ldots, p$, calculate $p_i^{(t)} \equiv p(\gamma_i^{(t)} = 1 \mid \boldsymbol{\gamma}_{(-i)}^{(t-1)}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}^{(t)}, c^{(t-1)})$, generate a random number u_i from $U(0, 1)$, and let

$$\gamma_i^{(t)} = \begin{cases} 1, & u_i < p_i^{(t)}, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

Step 3. Draw c from (17) by the following steps:

(i) maximize (17) to obtain c_{opt};

(ii) generate the proposal value

$$c^{(t)} = c_{\text{opt}} + \varepsilon^{(t)}, \tag{19}$$

where $\varepsilon^{(t)}$ follows a normal $N(\mu, \sigma^2)$ truncated in a positive region (a,b) with a density q;

(iii) accept $c^{(t)}$ with the acceptance probability:

$$R = \min\left\{1, \frac{p\left(c^{(t)} \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\gamma}\right)}{p\left(c^{(t-1)} \mid \left|\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\gamma}\right.\right)} \cdot \frac{q\left(c^{(t-1)} - c_{\text{opt}}\right)}{q\left(c^{(t)} - c_{\text{opt}}\right)}\right\}. \tag{20}$$

After the initial burn-in period, we obtain the MCMC samples $\{(\mathbf{Z}^{(t)}, \boldsymbol{\gamma}^{(t)}, c^{(t)}),\ t = 1, \ldots, M\}$ which are next used to estimate the posterior gene inclusion probability by

$$\widehat{p}\left(\gamma_i = 1 \mid \mathbf{Y}, \mathbf{X}\right) = \frac{1}{M} \sum_{t=1}^{M} I\left(\gamma_i^{(t)} = 1\right), \tag{21}$$

and genes with higher posterior inclusion probabilities are considered more relevant to classification.

2.4. Classification. To assess the performance of our procedures, testing data sets are considered. For example, a testing set $(X_{\text{new}}, Y_{\text{new}})$ is available, and the predictive probability of Y_{new} given X_{new} is

$$p\left(Y_{\text{new}} \mid \mathbf{Y}, \mathbf{X}, X_{\text{new}}\right) = \int p\left(Y_{\text{new}} \mid \mathbf{Y}, \mathbf{X}, X_{\text{new}}, \mathbf{Z}, \boldsymbol{\gamma}, c\right) p\left(\mathbf{Z}, \boldsymbol{\gamma}, c \mid \mathbf{Y}, \mathbf{X}\right) d\left(\mathbf{Z}, \boldsymbol{\gamma}, c\right). \tag{22}$$

Based on the MCMC samples, we estimate the probability with

$$\widehat{p}\left(Y_{\text{new}} \mid \mathbf{Y}, \mathbf{X}, X_{\text{new}}\right) = \frac{1}{M} \sum_{t=1}^{M} p\left(Y_{\text{new}} \mid \mathbf{Y}, \mathbf{X}, X_{\text{new}}, \mathbf{Z}^{(t)}, \boldsymbol{\gamma}^{(t)}, c^{(t)}\right). \tag{23}$$

When there are no testing sets available, we adopt the leave-one-out cross-validation (LOOCV) method to evaluate

Table 1: The posterior inclusion probability and description of the leading 20 genes for the colon cancer study. Genes identified in other studies were also noted.

Gene	Probability	Description
Z50753	0.1519	*H. sapiens* mRNA for GCAP-II/uroguanylin precursor[abc]
D14812	0.1303	Human mRNA for ORF, complete cds[bc]
H06524	0.1163	Gelsolin precursor, plasma (*Homo sapiens*)[ac]
R87126	0.1081	Myosin heavy chain, nonmuscle (*Gallus gallus*)[abc]
H08393	0.1012	Collagen alpha-2(XI) chain (*Homo sapiens*)[abc]
T62947	0.0987	60S ribosomal protein L24 (*Arabidopsis thaliana*)[abc]
T57882	0.0881	Myosin heavy chain, nonmuscle type A (*Homo sapiens*)[b]
R88740	0.0594	Atp synthase coupling factor 6, mitochondrial precursor (*Homo sapiens*)[bc]
J02854	0.0527	Myosin regulatory light chain 2, smooth muscle isoform (*Homo sapiens*); contains TAR1 repetitive element[ab]
T94579	0.0494	Human chitotriosidase precursor mRNA, complete cds[b]
H64807	0.0490	Placental folate transporter (*Homo sapiens*)[bc]
M59040	0.0439	Human cell adhesion molecule (CD44) mRNA, complete cds[c]
R55310	0.0437	S36390 mitochondrial processing peptidase[c]
M82919	0.0333	Human gamma aminobutyric acid (GABAA) receptor beta-3 subunit mRNA, complete cds[bc]
H20709	0.0330	Myosin light chain alkali, smooth-muscle isoform (*Homo sapiens*)[bc]
T92451	0.0319	Tropomyosin, fibroblast, and epithelial muscle-type (*Homo sapiens*)[a]
R33481	0.0312	Transcription factors ATF-A and ATF-A-DELTA (*Homo sapiens*)[b]
L06175	0.0309	*Homo sapiens* P5-1 mRNA, complete cds
T64012	0.0309	Acetylcholine receptor protein, delta chain precursor (*xenopus laevis*)
H09719	0.0300	Tubulin alpha-6 chain (*Mus musculus*)

[a]Gene also identified in Ben-Dor et al. [38].
[b]Gene also identified in Furlanello et al. [39].
[c]Gene also identified in Chu et al. [40].

the performance with the training data. Because the predictive probability for Y_i is

$$p\left(Y_i \mid \mathbf{Y}_{(-i)}, \mathbf{X}\right) = \left(\iiint p\left(Y_i \mid \mathbf{Y}_{(-i)}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\gamma}, c\right)^{-1} \times p\left(\mathbf{Z}, \boldsymbol{\gamma}, c \mid \mathbf{Y}, \mathbf{X}\right) d\mathbf{Z} d\boldsymbol{\gamma} dc \right)^{-1}, \tag{24}$$

where $\mathbf{Y}_{(-i)}$ denotes the vector of $\mathbf{Y}$ without the ith element. We estimate this probability based on the generated MCMC samples,

$$\widehat{p}\left(Y_i \mid \mathbf{Y}_{(-i)}, \mathbf{X}\right) = \frac{M}{\sum_{t=1}^{M} p\left(Y_i \mid \mathbf{Y}_{(-i)}, \mathbf{X}, \mathbf{Z}^{(t)}, \boldsymbol{\gamma}^{(t)}, c^{(t)}\right)^{-1}}. \tag{25}$$

3. Applications

In this section, we applied the fully Bayesian approach and the reference prior to three cancer studies: colon cancer, leukemia, and a large B-cell lymphoma (DLBCL) study [35–37]. We also compared the performance of this approach with other existing gene selection and classification methods. These data have been extensively studied with various methods but we only included a limited set of them. Others can be found in the reference lists of the work cited here.

3.1. Colon Cancer Study. The data of the colon cancer study contained 2000 expression levels from 40 tumor and 22 normal colon tissues. These expression levels were first transformed with a base 10 logarithmic function and then standardized to zero mean and unit variance for each gene. We then performed the MCMC sampler fixing the h in Σ_γ at 100 and $\pi_i = \Pr(\gamma_i = 1) = 0.005$ for all $i = 1, \ldots, p$. We burned in the first 12000 iterations, collected every 30th sample, and obtained 6700 posterior points in total for further analysis. The leading 20 genes with the largest posterior inclusion probabilities were presented in Table 1. This list was compared with the findings in three other studies [38–40] and similar findings were denoted in Table 1. The first 19 genes were identified in at least one of the three studies. For reference, Figure 1 displays the 100 largest posterior probabilities of the 100 corresponding genes.

For classification, we adopted the external leave-one-out cross-validation (LOOCV) procedure to evaluate the performance of classification with the selected genes. The procedures were the following: (i) removing one sample from the training set; (ii) ranking the genes in terms of t-statistics using the remaining samples and retaining the top 50 genes as the starting set to reduce computational burden; (iii) selecting the p^* most influential genes from the 50 genes based on our Bayesian method; and (iv) using these p^* genes to classify the previously removed sample. The procedures were repeated for each sample in the dataset. With different choices of p^* like $p^* = 6$, $p^* = 10$, and $p^* = 14$, the error rates were 0.1452, 0.1452, and 0.1129, respectively. The performance of other

Table 2: Performance comparison of different procedures with LOOCV for the colon cancer study.

Methods	No. of genes	LOOCV error rate	LOOCV accuracy
Bayesian g-prior	6	0.1452 (9/62)	0.8548 (53/62)
Bayesian g-prior	10	0.1452 (9/62)	0.8548 (53/62)
Bayesian g-prior	14	**0.1129 (7/62)**	**0.8871 (55/62)**
SVM[a]	1000	**0.0968 (6/62)**	**0.9032 (56/62)**
Classification tree[b]	200	0.1452 (9/62)	0.8548 (53/62)
1-Nearest-neighbor[b]	25	0.1452 (9/62)	0.8548 (53/62)
LogitBoost, estimated[b]	25	0.1935 (12/62)	0.8065 (50/62)
LogitBoost, 100 iterations[b]	10	0.1452 (9/62)	0.8548 (53/62)
AdaBoost, 100 iterations[b]	10	0.1613 (10/62)	0.8387 (52/62)
MAVE-LD[c]	50	0.1613 (10/62)	0.8387 (52/62)
IRWPLS[d]	20	**0.1129 (7/62)**	**0.8871 (55/62)**
SGLasso[e]	19	0.1290 (8/62)	0.8710 (54/62)
MRMS + SVM + D1[f]	5	0.1290 (8/62)	0.8710 (54/62)
MRMS + SVM + D2[f]	33	0.1452 (9/62)	0.8548 (53/62)
t-test + probit regression	6	0.1452 (9/62)	0.8548 (53/62)
t-test + probit regression	10	0.1774 (11/62)	0.8226 (51/62)
t-test + probit regression	14	0.2258 (14/62)	0.7742 (48/62)

[a]Proposed by Furey et al. [41].
[b]Proposed by Dettling and Bühlmann [42].
[c]Proposed by Antoniadis et al. [43].
[d]Proposed by Ding and Gentleman [44].
[e]Proposed by Ma et al. [45].
[f]Proposed by Maji and Paul [46].

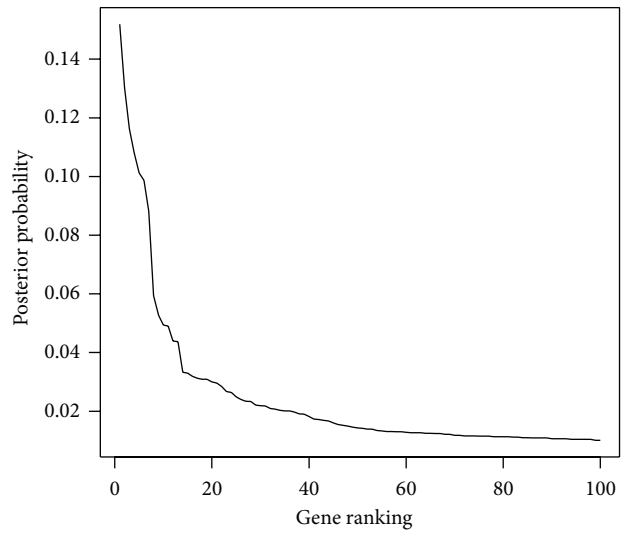


Figure 1: The largest 100 posterior probabilities of the genes for colon cancer study.

methods, including SVM [41]; classification tree followed by 1-Nearest-neighbor and LogitBoost with 100 iterations [42]; MAVE-LD [43]; IRWPLS [44]; supervised group Lasso (SGLasso, [45]) and MRMS [46]; and t-test for single markers in probit regression was summarized in Table 2. SVM had the smallest error rate, but it apparently included too many genes (1000 in this set). One other method MRMS+SVM+D1 performed better, with one more correct classification, than our proposed procedure when 6 or 10 genes were selected.

3.2. Leukemia Study. Next we considered the leukemia study with gene expression levels from 72 tissues including 47 acute lymphoblastic leukemia (ALL) patients and 25 acute myeloid leukemia (AML) subjects. These data contained 38 training and 34 testing samples. The training data contained 27 ALL cases and 11 AML cases, whereas the testing data were with 20 ALL cases and 14 AML cases. As described in other studies [2], the preprocessing steps such as thresholding and filtering were applied first and then followed by a base 10 logarithmic transformation. A total of 3571 genes were left for analysis. Next, we standardized the data across samples, and we ranked these genes by the same MCMC procedures described earlier. The top 20 genes with the largest posterior inclusion probabilities were presented in Table 3, and genes identified by other studies [36, 41, 47, 48] were also noted. For reference, Figure 2 displays the 100 largest posterior probabilities of the 100 corresponding genes.

For the classification procedure, similar to the procedures for colon cancer study, we selected p^* most influential genes from a starting set of 50 genes and next used them to examine the testing data. With $p^* = 6$, 10, or 14 genes, only the 61st and 66th observations were misclassified by our procedure. We also compared the results with weighted voting machine [36], MAVE-LD [43], two-step EBM [47], KIGP + PK [48], and t-test for single markers with probit regression, as summarized in Table 4. Note that although MAVE-LD and two-step EBM methods performed better than our proposed

Table 3: The posterior inclusion probability and description of the leading 20 genes for the leukemia study. Genes identified in other studies were also noted.

Gene	Probability	Description
X95735	0.0691	Zyxin[abc]
M27891	0.0519	CST3 cystatin C (amyloid angiopathy and cerebral hemorrhage)[abc]
M23197	0.0302	CD33 cD33 antigen (differentiation antigen)[abc]
Y12670	0.0251	LEPR leptin receptor[a]
X85116	0.0226	Epb72 gene exon 1[ab]
D88422	0.0196	CYSTATIN A[bc]
X62654	0.0196	ME491 gene extracted from *H. sapiens* gene for Me491/CD63 antigen[b]
X04085	0.0195	Catalase (EC 1.11.1.6) 5′ank and exon 1 mapping to chromosome 11, band p13 (and joined CDS)[a]
L09209	0.0195	APLP2 amyloid beta (A4) precursor-like protein 2[bc]
HG1612-HT1612	0.0186	Macmarcks[bc]
M16038	0.0186	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog[abc]
U50136	0.0181	Leukotriene C4 synthase (LTC4S) gene[ab]
M55150	0.0172	FAH fumarylacetoacetate[ab]
M92287	0.0172	CCND3 cyclin D3[bc]
M22960	0.0168	PPGB protective protein for beta-galactosidase (galactosialidosis)[bc]
X70297	0.0168	CHRNA7 cholinergic receptor, nicotinic, and alpha polypeptide 7[b]
X51521	0.0163	VIL2 Villin 2 (ezrin)[b]
M63138	0.0154	CTSD cathepsin D (lysosomal aspartyl protease)[ab]
M27783	0.0154	ELA2 elastase 2, neutrophil[c]
U81554	0.0137	CaM kinase II isoform mRNA

[a]Gene also identified in Golub et al. [36].
[b]Gene also identified in Ben-Dor et al. [38].
[c]Gene also identified in in Lee et al. [22].

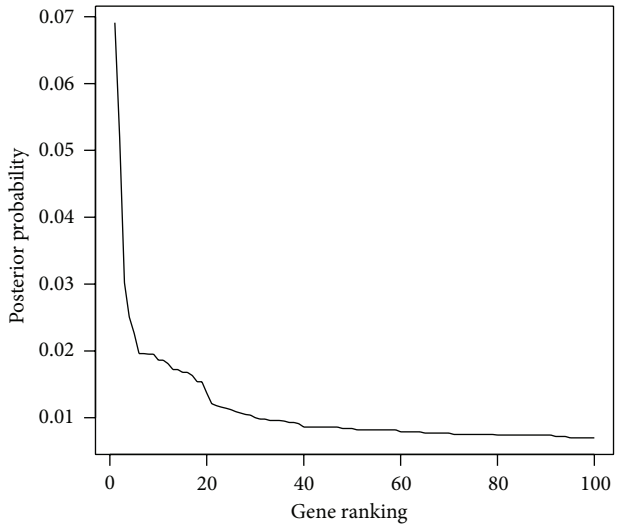


Figure 2: The largest 100 posterior probabilities of the genes for leukemia study.

procedure, both methods used more genes (50 and 512) and yet achieved only one less misclassification. Among this list, our procedure apparently considered a smaller set of genes with a satisfactory performance.

3.3. Diffuse Large B-Cell Lymphoma (DLBCL) Study. This study collected 58 samples from DLBCL patients and 19 samples from follicular lymphoma [37]. The original dataset contained 7129 genes. After the preprocessing steps such as thresholding and filtering were applied and a base 10 logarithmic transformation was conducted, a total of 6285 genes were left for analysis. Next, we standardized the data across samples and ranked these genes by the same MCMC procedures described in earlier sections. The error rates for $p^* = 6$, 10, or 14 under LOOCV were 0.0519, 0.0649, and 0.0779, and the accuracy was between 0.92 and 0.95, as listed in Table 5. To achieve a smaller error rate, we considered $p^* = 5$ and obtained a smaller rate 0.0390, the same rate achieved by the hyperbox enclosure (HBE) method [49]. Similar to the discussion in the previous two applications, our proposed model can achieve the same or smaller error rate with a smaller set of genes.

4. Conclusion and Discussion

In this Bayesian framework, we considered a mixture of g-prior to complete a fully Bayesian analysis for gene selection and cancer classification. Different from other existing methods that treated the c as a fixed value, we incorporated its uncertainty by assuming a reference inverse-gamma prior distribution. Earlier studies mentioned this prior, but considered it difficult to derive posterior inference. We therefore

Table 4: Performance comparison of different procedures for the leukemia study.

Methods	No. of genes	Testing error rate	Testing accuracy
Bayesian g-prior	6	**0.0588 (2/34)**	**0.9412 (32/34)**
Bayesian g-prior	10	0.0588 (2/34)	0.9412 (32/34)
Bayesian g-prior	14	0.0588 (2/34)	0.9412 (32/34)
Weighted voting machine[a]	50	0.1471 (5/34)	0.8529 (29/34)
MAVE-LD[b]	50	**0.0294 (1/34)**	**0.9706 (33/34)**
Two-step EBM[c]	32	0.1471 (5/34)	0.8529 (29/34)
Two-step EBM[c]	256	0.0588 (2/34)	0.9412 (32/34)
Two-step EBM[c]	512	**0.0294 (1/34)**	**0.9706 (33/34)**
KIGP + PK[d]	20	0.0588 (2/34)	0.9412 (32/34)
t-test + probit regression	6	0.1765 (6/34)	0.8235 (28/34)
t-test + probit regression	10	0.0882 (3/34)	0.9118 (31/34)
t-test + probit regression	14	0.1176 (4/34)	0.8824 (30/34)

[a]Proposed by Gloub et al. [36].
[b]Proposed by Antoniadis et al. [43].
[c]Proposed by Ji et al. [47].
[d]Proposed by Zhao and Cheung [48].

Table 5: Performance comparison of different procedures with LOOCV for the colon cancer study.

Methods	No. of genes	LOOCV error rate	LOOCV accuracy
Bayesian g-prior	5	0.0390 (3/77)	0.9610 (74/77)
Bayesian g-prior	6	0.0519 (4/77)	0.9481 (73/77)
Bayesian g-prior	10	0.0649 (5/77)	0.9351 (72/77)
Bayesian g-prior	14	0.0779 (6/77)	0.9221 (71/77)
Bayesian g-prior	20	0.0779 (6/77)	0.9221 (71/77)
HBE	6	0.0390 (3/77)	0.9610 (74/77)
t-test + probit regression	6	0.1169 (9/77)	0.8831 (68/77)
t-test + probit regression	10	0.1558 (12/77)	0.8442 (65/77)
t-test + probit regression	14	0.2208 (17/77)	0.7792 (60/77)

outlined the implementation for computation under this model setting for future applications. This approach is more flexible in the process of model building. This model is able to evaluate how influential a gene can be with posterior probabilities that can be used next for variable selection. Such an approach is useful in biomedical interpretations for the selection of relevant genes for disease of interest. When compared with other existing methods, our proposed procedure achieves a better or comparable accurate rate in classification with fewer genes. In the analyses of colon cancer and leukemia studies, we replicate several relevant genes identified by other research groups. The findings have accumulated evidence for further laboratory research.

In the application section, we listed only the results from $p^* = 6$, 10, and 14 selected genes. Other values for p^* have been tried and the performance remains good. For instance, the pink line in Figures 3 and 4 displays the accuracy of the proposed procedure when the number of selected genes p^* varies between 5 and 20 for the colon cancer and leukemia study, respectively. For the colon cancer study, the largest accuracy 0.8871 occurs at $p^* = 14$, while other values of p^* lead to the accuracy between 0.8387 and 0.8871. These

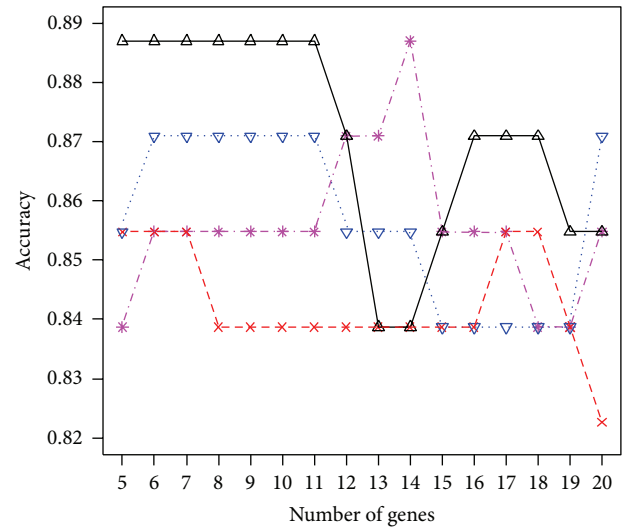


Figure 3: The accuracy of the proposed procedure at different numbers ($p^* = 5, \ldots, 20$) of selected genes with c following the generalized g-prior (pink line) or fixed at constant 5 (red line), 10 (blue), or 20 (black) for the colon cancer study.

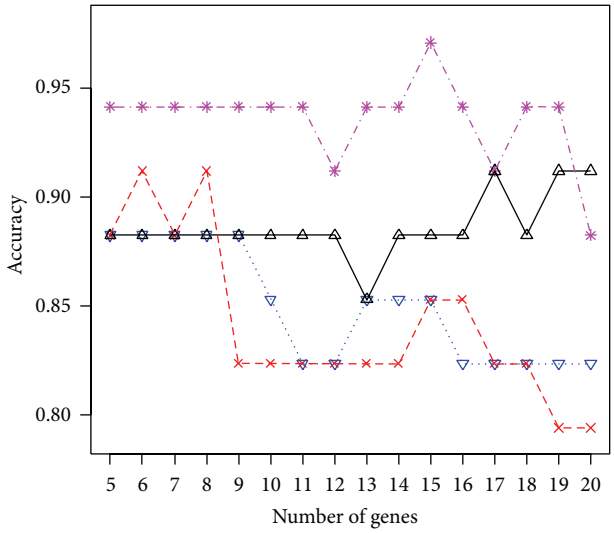


FIGURE 4: The accuracy of the proposed procedure at different numbers ($p^* = 5, \ldots, 20$) of selected genes with c following the generalized g-prior (pink line) or fixed at constant 5 (red line), 10 (blue), or 20 (black) for the leukemia study.

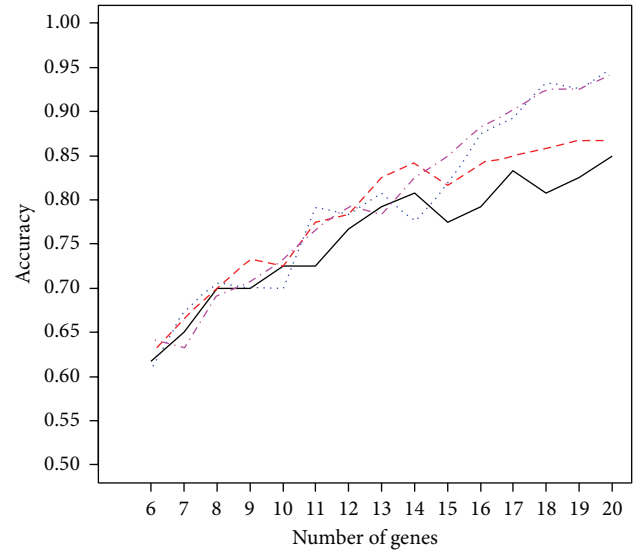


FIGURE 5: Average accuracy when the number of genes ranges from 1 to 15 under the mixtures of g-priors on c (pink line), c fixed at 5 (black), c at 50 (red), and c at 500 (blue).

correspond to at least 52 correctly identified subjects out of 62. For the leukemia study, the largest accuracy 0.9706 occurs at $p^* = 15$. Other values of p^* all lead to an accuracy larger than 90% except when $p^* = 20$ (accuracy is $0.8824 = 30/34$). In addition, we compared the results under the proposed generalized g-prior with c fixed at a constant. The colored lines in Figures 3 and 4 are for c fixed at 5 (red line), 10 (blue), or 20 (black), respectively. Again, results under the prior distribution assumption lead to a higher accuracy with a less number of selected genes. Another issue is related to the choice of the number of genes in the starting set. We have considered 50 in all three applications. This value can certainly be changed. However, the computational complexity increased as the value becomes larger. This cost in computation remains a research topic for future research.

To compare the performance of a stochastic c and a constant c, we also conducted a small simulation study to investigate the effect of assigning a prior on c versus fixing c at different constant values. We used the R package penalizedSVM [50, 51] to simulate three data sets; each contains 500 genes with 15 genes associated with the disease. The numbers of training and testing sample were 200 and 40, respectively. We then conducted the gene selection procedures with a prior on c, $c = 5$, $c = 50$, and $c = 500$ at $p^* = 1, 2, \ldots, 15$ and recorded the accuracy under each setting. Figure 5 plots the average accuracy with the pink line standing for the accuracy under the mixtures of g-priors on c, the black line for $c = 5$, the red line for $c = 50$, and the blue line for $c = 500$. It can be observed that only when c is assigned with a very large number like 500, the corresponding accuracy can be slightly better than that under a prior for the uncertainty in c. This again supports the use of the mixtures of g-priors for a better and robust result.

Here in this paper we have focused on the analysis of binary data. However, the probit regression model can be extended to a multinomial probit model to solve the multiclass problems, and the Bayesian inference can be carried out similarly. Such analysis will involve a larger computational load and further research in this direction is needed. Another point worth mentioning is the inclusion of interactions between genes. Further research can incorporate a power prior into the prior of $\boldsymbol{\gamma}$ [52] or include information on gene-gene network structure [18] to complete the procedure for variable selection.

Acknowledgment

Part of this research was supported by NSC 100-2314-B-002-107-MY3.

References

[1] V. T. Chu, R. Gottardo, A. E. Raftery, R. E. Bumgarner, and K. Y. Yeung, "MeV+R: using MeV as a graphical user interface for Bioconductor applications in microarray analysis," *Genome Biology*, vol. 9, no. 7, article R118, 2008.

[2] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–86, 2002.

[3] A. Hirakawa, Y. Sato, D. Hamada, and I. Yoshimura, "A new test statistic based on shrunken sample variance for identifying differentially expressed genes in small microarray experiments," *Bioinformatics and Biology Insights*, vol. 2, pp. 145–156, 2008.

[4] W. Pan, J. Lin, and C. T. Le, "A mixture model approach to detecting differentially expressed genes with microarray data," *Functional and Integrative Genomics*, vol. 3, no. 3, pp. 117–124, 2003.

[5] K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery, "Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data," *Bioinformatics*, vol. 21, no. 10, pp. 2394–2402, 2005.

[6] A. Gusnanto, A. Ploner, F. Shuweihdi, and Y. Pawitan, "Partial least squares and logistic regression random-effects estimates for gene selection in supervised classification of gene expression data," *Journal of Biomedical Informatics*, vol. 4, pp. 697–709, 2013.

[7] Y. Liang, C. Liu, X. Z. Luan et al., "Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification," *BMC Bioinformatics*, vol. 14, article 198, 2013.

[8] G.-Z. Li, H.-L. Bu, M. Q. Yang, X.-Q. Zeng, and J. Y. Yang, "Selecting subsets of newly extracted features from PCA and PLS in microarray data analysis," *BMC Genomics*, vol. 9, no. 2, article S24, 2008.

[9] A. Wang and E. A. Gehan, "Gene selection for microarray data analysis using principal component analysis," *Statistics in Medicine*, vol. 24, no. 13, pp. 2069–2087, 2005.

[10] S. Bicciato, A. Luchini, and C. Di Bello, "PCA disjoint models for multiclass cancer analysis using gene expression data," *Bioinformatics*, vol. 19, no. 5, pp. 571–578, 2003.

[11] X. Q. Zeng, G. Z. Li, M. Q. Yang, G. F. Wu, and J. Y. Yang, "Orthogonal projection weights in dimension reduction based on partial least squares," *International Journal of Computational Intelligence in Bioinformatics and Systems Biology*, vol. 1, pp. 100–115, 2009.

[12] A.-L. Boulesteix and K. Strimmer, "Partial least squares: a versatile tool for the analysis of high-dimensional genomic data," *Briefings in Bioinformatics*, vol. 8, no. 1, pp. 32–44, 2007.

[13] D. V. Nguyen and D. M. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, no. 1, pp. 39–50, 2002.

[14] J. X. Liu, Y. Xu, C. H. Zheng, Y. Wang, and J. Y. Yang, "Characteristic gene selection via weighting principal components by singular values," *PLoS ONE*, vol. 7, no. 7, Article ID e38873, 2012.

[15] S. Student and K. Fujarewicz, "Stable feature selection and classification algorithms for multiclass microarray data," *Biology Direct*, vol. 7, article 33, 2012.

[16] T. Bø and I. Jonassen, "New feature subset selection procedures for classification of expression profiles," *Genome Biology*, vol. 3, no. 4, pp. 1–17, 2002.

[17] Y. Wang, I. V. Tetko, M. A. Hall et al., "Gene selection from microarray data for cancer classification—a machine learning approach," *Computational Biology and Chemistry*, vol. 29, no. 1, pp. 37–46, 2005.

[18] F. C. Stingo and M. Vannucci, "Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data," *Bioinformatics*, vol. 27, no. 4, pp. 495–501, 2011.

[19] J. G. Ibrahim, M.-H. Chen, and R. J. Gray, "Bayesian models for gene expression with DNA microarray data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 88–99, 2002.

[20] Y.-C. Wei, S.-H. Wen, P.-C. Chen, C.-H. Wang, and C. K. Hsiao, "A simple Bayesian mixture model with a hybrid procedure for genome-wide association studies," *European Journal of Human Genetics*, vol. 18, no. 8, pp. 942–947, 2010.

[21] B. Peng, D. Zhu, and B. P. Ander, "An Integrative Framework for Bayesian variable selection with informative priors for identifying genes and pathways," *PLoS ONE*, vol. 8, no. 7, Article ID 0067672, 2013.

[22] K. E. Lee, N. Sha, E. R. Dougherty, M. Vannucci, and B. K. Mallick, "Gene selection: a Bayesian variable selection approach," *Bioinformatics*, vol. 19, no. 1, pp. 90–97, 2003.

[23] N. Sha, M. Vannucci, M. G. Tadesse et al., "Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage," *Biometrics*, vol. 60, no. 3, pp. 812–819, 2004.

[24] X. Zhou, K.-Y. Liu, and S. T. C. Wong, "Cancer classification and prediction using logistic regression with Bayesian gene selection," *Journal of Biomedical Informatics*, vol. 37, no. 4, pp. 249–259, 2004.

[25] J. G. Liao and K.-V. Chin, "Logistic regression for disease classification using microarray data: model selection in a large p and small n case," *Bioinformatics*, vol. 23, no. 15, pp. 1945–1951, 2007.

[26] A. Zellner, "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, pp. 233–243, North-Holland, Amsterdam, The Netherlands, 1986.

[27] A.-J. Yang and X.-Y. Song, "Bayesian variable selection for disease classification using gene expression data," *Bioinformatics*, vol. 26, no. 2, pp. 215–222, 2010.

[28] M. Baragatti and D. Pommeret, "A study of variable selection using g-prior distribution with ridge parameter," *Computational Statistics and Data Analysis*, vol. 56, no. 6, pp. 1920–1934, 2012.

[29] E. Leya and M. F. J. Steel, "Mixtures of *g*-priors for Bayesian model averaging with economic applications," *Journal of Econometrics*, vol. 171, no. 2, pp. 251–266, 2012.

[30] M. Smith and R. Kohn, "Nonparametric regression using Bayesian variable selection," *Journal of Econometrics*, vol. 75, no. 2, pp. 317–343, 1996.

[31] E. I. George and D. P. Foster, "Calibration and empirical bayes variable selection," *Biometrika*, vol. 87, no. 4, pp. 731–747, 2000.

[32] F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger, "Mixtures of g priors for Bayesian variable selection," *Journal of the American Statistical Association*, vol. 103, no. 481, pp. 410–423, 2008.

[33] W. Cui and E. I. George, "Empirical Bayes versus fully Bayes variable selection," *Journal of Statistical Planning and Inference*, vol. 138, no. 4, pp. 888–900, 2008.

[34] C. P. Robert, "Convergence control methods for Markov chain Monte Carlo algorithms," *Statistical Science*, vol. 10, pp. 231–253, 1995.

[35] U. Alon, N. Barka, D. A. Notterman et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, pp. 6745–6750, 1999.

[36] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[37] M. A. Shipp, K. N. Ross, P. Tamayo et al., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[38] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, "Tissue classification with gene expression profiles," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 559–583, 2000.

[39] C. Furlanello, M. Serafini, S. Merler, and G. Jurman, "Entropy-based gene ranking without selection bias for the predictive classification of microarray data," *BMC Bioinformatics*, vol. 4, article 54, 2003.

[40] W. Chu, Z. Ghahramani, F. Falciani, and D. L. Wild, "Biomarker discovery in microarray gene expression data with Gaussian processes," *Bioinformatics*, vol. 21, no. 16, pp. 3385–3393, 2005.

[41] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[42] M. Dettling and P. Bühlmann, "Boosting for tumor classification with gene expression data," *Bioinformatics*, vol. 19, no. 9, pp. 1061–1069, 2003.

[43] A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc, "Effective dimension reduction methods for tumor classification using gene expression data," *Bioinformatics*, vol. 19, no. 5, pp. 563–570, 2003.

[44] B. Ding and R. Gentleman, "Classification Using Generalized Partial Least Squares," Bioconductor Project Working Papers, 2004, http://www.bepress.com/bioconductor /paper5.

[45] S. Ma, X. Song, and J. Huang, "Supervised group Lasso with applications to microarray data analysis," *BMC Bioinformatics*, vol. 8, article 60, 2007.

[46] P. Maji and S. Paul, "Rough set based maximum relevance-maximum significance criterion and Gene selection from microarray data," *International Journal of Approximate Reasoning*, vol. 52, no. 3, pp. 408–426, 2011.

[47] Y. Ji, K.-W. Tsui, and K. Kim, "A novel means of using gene clusters in a two-step empirical Bayes method for predicting classes of samples," *Bioinformatics*, vol. 21, no. 7, pp. 1055–1061, 2005.

[48] X. Zhao and L. W.-K. Cheung, "Kernel-imbedded Gaussian processes for disease classification using microarray gene expression data," *BMC Bioinformatics*, vol. 8, article 67, 2007.

[49] O. Dagliyan, F. Uney-Yuksektepe, I. H. Kavakli, and M. Turkay, "Optimization based tumor classification from microarray gene expression data," *PLoS ONE*, vol. 6, no. 2, Article ID e14579, 2011.

[50] H. H. Zhang, J. Ahn, X. Lin, and C. Park, "Gene selection using support vector machines with non-convex penalty," *Bioinformatics*, vol. 22, no. 1, pp. 88–95, 2006.

[51] G. M. Fung and O. L. Mangasarian, "A feature selection Newton method for support vector machine classification," *Computational Optimization and Applications*, vol. 28, no. 2, pp. 185–202, 2004.

[52] A. Krishna, H. D. Bondell, and S. K. Ghosh, "Bayesian variable selection using an adaptive powered correlation prior," *Journal of Statistical Planning and Inference*, vol. 139, no. 8, pp. 2665–2674, 2009.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 917502, 7 pages
http://dx.doi.org/10.1155/2013/917502

Research Article

Modified Logistic Regression Models Using Gene Coexpression and Clinical Features to Predict Prostate Cancer Progression

Hongya Zhao,[1,2] Christopher J. Logothetis,[2] Ivan P. Gorlov,[2] Jia Zeng,[3] and Jianguo Dai[4]

[1] *Industrial Center, Shenzhen Polytechnic, Shenzhen, Guangdong 518055, China*
[2] *Department of Genitourinary Medical Oncology, Unit 1374, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030-4009, USA*
[3] *School of Computer Science and Technology, Soochow University, Suzhou 215006, China*
[4] *School of Applied Chemistry and Biotechnology, Shenzhen Polytechnic, Shenzhen, Guangdong 518055, China*

Correspondence should be addressed to Ivan P. Gorlov; gorurofu@mail.ru

Received 12 June 2013; Accepted 3 September 2013

Academic Editor: Qizhai Li

Predicting disease progression is one of the most challenging problems in prostate cancer research. Adding gene expression data to prediction models that are based on clinical features has been proposed to improve accuracy. In the current study, we applied a logistic regression (LR) model combining clinical features and gene co-expression data to improve the accuracy of the prediction of prostate cancer progression. The top-scoring pair (TSP) method was used to select genes for the model. The proposed models not only preserved the basic properties of the TSP algorithm but also incorporated the clinical features into the prognostic models. Based on the statistical inference with the iterative cross validation, we demonstrated that prediction LR models that included genes selected by the TSP method provided better predictions of prostate cancer progression than those using clinical variables only and/or those that included genes selected by the one-gene-at-a-time approach. Thus, we conclude that TSP selection is a useful tool for feature (and/or gene) selection to use in prognostic models and our model also provides an alternative for predicting prostate cancer progression.

1. Introduction

Prostate cancer (PCa) is the second leading cause of cancer-related deaths among men in the USA [1, 2]. Screening using serum prostate-specific antigen (PSA) has improved the early detection of PCa and has resulted in an increase in the proportion of patients with disease that is curable by prostatectomy [3, 4]. However, 20% to 30% of treated patients will develop a local or metastatic recurrence which reflects the most adverse clinical outcome [4]. Thus, from the clinical perspective, it is important to be able to predict which patients will experience a relapse.

Traditional PCa prognosis models are based on some clinical features, such as pretreatment PSA levels, biopsy Gleason score (GS), and clinical stage, but in practice, they are inadequate to accurately predict disease progression [5]. With the development of microarray technology in recent years, a number of studies have been conducted to characterize the dynamics of gene expression in PCa progression by using DNA microarrays. In some studies, tumor expression signatures associated with clinical parameters and outcomes have been identified [6–9]. As a result, it is possible to develop the clinical models with the variables of gene signatures identified from microarray data and some clinical features to predict which men would experience progression to the metastatic form of PCa.

However, it has been found that none of the predictive models using gene expression profiles are significantly better than models using clinical variables only in predicting PCa progression [10, 11]. In fact, only a limited number of genes are used to avoid overfitting in these models. The genes are usually selected through a gene-by-gene comparison. The results of recent studies, however, suggest that assessing the expression of more than one gene (i.e., coexpression analysis) yields a better prediction of tumor progression than the analysis of individual genes does [12–15].

In this study, we tried to propose such models by merging the coexpressed genes' profiles and some clinical features to predict the patients who would suffer from PCa progression. The genes used in our models are identified by a top-scoring pair (TSP) algorithm. The TSP method was initially introduced by Geman et al. as a classification technique for microarray data [16]. We applied the TSP-based LR model to published microarray experiments whose patients suffered from PCa progression. We analyzed the effects of the number of coexpressed genes included in the models and the selection of the clinical variables on the accuracy of the prediction. We also compared the performance of the most commonly used classification methods with our proposed method.

2. Materials and Methods

2.1. Logistic Regression Model for the Classification of Gene Microarrays. Genome-wide microarray data from different cells give insight into the gene expression variation of various genotypes and phenotypes. Classification of patients is an important aspect of cancer diagnosis and treatment. For example, microarray experiments can be employed to screen gene expression levels from cancerous and normal phenotypes so that proper prediction rules can be built from these gene expression data. In this section, we introduce a logistic regression (LR) model to classify the phenotypes of microarray data.

We denote a gene expression matrix by $D = \{x_{ij}\}_{M\times N}$, where there are M genes and N samples, and x_{ij} denotes the expression value of the ith gene, $i \in \{1, \ldots, M\}$, from the jth sample, $j \in \{1, \ldots, N\}$. The vector $G_{i\cdot} = (x_{i1}, \ldots, x_{iN})$ represents the i gene expression values over all N samples and $S_{\cdot j} = (x_{1j}, \ldots, x_{Mj})$ is the expression profile of all M genes for the jth sample. Let y_j be the binary phenotype of the jth profile: $y_j = 0$ indicates that the jth sample belongs to class 0 (e.g., normal tissues) and $y_j = 1$ indicates that the jth sample belongs to class 1 (e.g., tumor tissues).

The classification of microarray data has been intensively researched for years. But some limitations have stood out, such as the small-sample dilemma, "black box," and lack of prediction strength [16–18]. We used LR to build the prediction models for a binary outcome. Obviously, the underlying probability of labels and contribution of predictor variables can be explicitly provided in LR models, which is helpful for biologists in discovering the genes that interact and cause the occurrence of disease.

The goal of classification with LR was to find a formula that gives the probability p_j that the jth sample with all its measured expressions $S_{\cdot j}$ represents a class 1 case. Since only two classes are considered, the probability of the sample representing class 0 is consequently $1 - p_j$. We used the following normal LR model:

$$\eta_j = \log \frac{p_j}{1 - p_j} = \beta_0 + \sum_{i=1}^{M} \beta_i x_{ij}, \tag{1}$$

where $\beta_0, \beta_1, \ldots, \beta_M$ are parameters that can be estimated by maximizing the following likelihood:

$$L(\beta_0, \beta) = \sum_{j=1}^{N} y_j \log p_j + \sum_{j=1}^{N} \left(1 - y_j\right) \log\left(1 - p_j\right). \tag{2}$$

For microarray experiments of typical "large p, small n," the number of samples, N, is usually on the order of tens, but the number of genes, M, is usually on the order of thousands or even tens of thousands. So the number of samples is much less than the number of variables ($N \ll M$). This situation presents a number of problems when building a LR model, such as overfitting, multicollinearity of the gene expression profiles, and infinite solutions for β_i [17–19]. Feature selection can be used to identify the significant genes that contribute to most of the classification. Thus, some dimension-reduction techniques, such as support vector machines (SVMs), singular value decomposition, and partial least squares, are commonly used to tackle those problems and make the computation feasible [17–19]. However, the featured genes are usually selected one by one. According to the biological mechanism, genes do not work by themselves, so we employed coexpressed TSP genes in the model, as described in the following section.

2.2. Identification of Coexpressed Genes. Recent studies have suggested that assessing the expression of more than one gene (i.e., coexpression analysis) provides a better prediction of tumor progression than analyzing the expression of individual genes [20–22]. We identified the coexpressed gene with the paired-gene approach of the top-scoring pairs (TSP) algorithm as described by Geman et al. [16]. The TSP algorithm was originally developed for the binary classification of phenotypes according to the relative expression profiles of one-gene pair. The TSP classifier has the following advantages over the standard classifiers used in gene expression studies: (i) it is a parameter-free and data-driven machine-learning method that avoids overfitting by eliminating the need to perform specific parameter tuning, as in other machine-learning techniques, such as SVMs and neural networks; (ii) it involves only two genes, which leads to easily interpretable data and inexpensive diagnostic tests; (iii) the rank-based TSP classifiers are less affected by technical factors or normalization than classifiers which are based on expression levels of individual genes; and (iv) the simple and accurate results generated by TSP facilitate follow-up studies.

TSP gene pairs may be considered biomarker genes in a diagnostic test from microarray experiments [16, 20–22]. The methodology is being extended from one TSP gene pair to top-scoring pair of groups (TSPG) as gene signatures [20–22]. However, there are still some unresolved issues of biological explanation and the selection criteria related to the use of gene pairs instead of larger groups of significant genes. Most of the algorithms in gene selection are based on the distribution assumption of the gene expression data. However, the rank-based TSP algorithm is a parameter-free, data-driven machine-learning method. It is difficult to determine the number of gene pairs selected, but current

research indicates that only a few gene pairs with the top scores need to be considered [20, 21].

For simplicity, using the gene expression matrix $D = \{x_{ij}\}_{M\times N}$ with M genes and N samples, we assume that N_1 samples are labeled class 0, $y_j = 0$ ($j \in \{1, \ldots, N_1\}$), and N_2 samples are labeled class 1, $y_j = 1$ ($j \in \{N_1 + 1, \ldots, N\}$), and $N_1 + N_2 = N$. In this method, we focused on detecting "marker gene pairs" (u, v) because there is a significant difference in the probability of observing $G_u < G_v$ between class 1 and class 0, where G_u and G_v denote the uth and vth rows of D. The conditional probabilities of observing $G_u < G_v$ in each class are defined as

$$p_{uv}(0) = P(G_u < G_v \mid c = 0) = \frac{1}{N_1}\sum_{j=1}^{N_1} I\left(x_{uj} < x_{vj}\right),$$

$$p_{uv}(1) = P(G_u < G_v \mid c = 1) = \frac{1}{N_2}\sum_{j=N_1+1}^{N} I\left(x_{uj} < x_{vj}\right), \tag{3}$$

where $I(x_{uj} < x_{vj})$ is the indicator function defined as

$$I\left(x_{uj} < x_{vj}\right) = \begin{cases} 1, & x_{uj} < x_{vj}, \\ 0, & x_{uj} \geq x_{vj}, \end{cases} \quad j = 1, 2, \ldots, N. \tag{4}$$

The typical TSP method is based on maximizing the following score of (u, v) defined by German et al. [16]:

$$\Delta_{uv} = \left|p_{uv}(0) - p_{uv}(1)\right|. \tag{5}$$

This approach has been shown to be as accurate as SVMs and other more sophisticated methods [20–22]. Although maximizing delta identifies the best classifier with high accuracy, it may be associated with relatively low sensitivity or specificity, as pointed out by German et al. and Ummanni et al. [16, 23]. For example, in the classification of cancer versus normal samples, accuracy is defined as the ratio between the number of correctly predicted samples and the total number of samples, and sensitivity (resp., specificity) is the ratio between the number of correctly predicted cancer (resp., normal) samples and the total number of cancer (resp., normal) samples [16]. This low sensitivity or specificity restricts us to use the classifier of one TSP for making medical decisions. This issue was improved with the use of multiple gene pairs as the classifier, which can achieve similar scores with high accuracy, sensitivity, and specificity [20–22]. Thus, we considered not just one but multiple TSP gene pairs in our models.

2.3. Evaluation of the Model Using Published Datasets. To evaluate the efficiency of the TSP-based LR model, we applied our model to datasets with both clinical parameters and gene expression values. We selected a dataset with a large sample size because we could obtain more reliable estimates of the efficiency of the classifiers. The dataset was from the recently published study of Sboner et al. [5], who analyzed gene expression in patients with up to 30 years of clinical follow-up data. Men who died within 10 years of being diagnosed with PCa were considered to have "lethal" disease, and those who survived at least 10 years after diagnosis were considered to have "indolent" disease. There were 165 men with lethal and 116 with indolent disease. The GS, tumor percentage, and presence of an estrogen-regulated gene (ERG) rearrangement were provided for each patient in the study. The expression of 6,100 genes was assessed using a custom gene expression array (GSE 16560).

For our model, we first randomly separated the 281 samples into a learning set with 186 samples and a validation set with the other 95 samples, with an approximately equal proportion of men with lethal and indolent PCa in each group. The learning set was utilized to create the models whose performance was evaluated in the validation set by means of the area under the receiver operating characteristic (ROC) curve (AUC). To compare the performance of our model, we performed the statistical testing based on the null hypothesis that there is no difference between the AUCs of Sboner's models and ours. Similar to the estimation of AUCs in [5], the corresponding 95% confidence intervals of the AUCs were computed in 100 iterative 10-fold cross validation procedures that enabled an unbiased estimation of the model's performance since the evaluation was performed on an independent dataset. The model is inferred to be better only if its AUC is statistically larger than that of the other models. In the original study, the authors conducted an extensive comparative analysis of the most frequently used classification methods, including the k-nearest neighbor, the nearest template prediction, diagonal linear discriminate analysis, SVMs, and neural network analysis. Their results allowed us to compare the performance of the TSP-based LR classifiers with that of the other classifiers.

To optimize and select the best models, we adopted an iterative cross validation procedure within the learning set that was similar to the procedure used by Sboner et al. [5]. The stratified tenfold cross validation procedure split the learning set into 10 disjointed partitions, test_i ($i = 1, \ldots, 10$), with approximately equal proportions of lethal and indolent cases in each. For a given partition, test_i, the models were fitted using all the other cases in the learning set, that is, the training_i set and then were evaluated with AUC analysis of test_i. In the procedure of 10-fold cross validation, the model_i ($i = 1, \ldots, 10$) was first parameterized in the training_i sets and then the corresponding AUC ontest_i ($i = 1, \ldots, 10$) sets were calculated from model_i. To avoid potential biases in the selection of the 10 partitions, the entire procedure was repeated 100 times, for 1,000 different partitions. We identified the best model with the largest AUC by comparing them as obtained in the 100 iterations. Furthermore, the featured gene pairs and estimated parameters in the model were also considered as the best model in learning set. The rationale was that the results of this procedure enable the identification of the best model, which can then be used to build a classifier that was finally evaluated on the validation set.

During the iterations of our cross validation procedure, the feature-selection procedure was carried out to identify the subsets of genes that are expressed differently in the lethal and indolent samples. In the study by Sboner et al. [5], a two-sided

Table 1: Logistic regression models that included TSP-selected gene pairs and different combinations of clinical variables.

Model number	Patient's age	Gleason score	Tumor percentage	Fusion ERG arrangement	TSP genes
1.1					X
1.2	X				X
1.3		X			X
1.4			X		X
1.5				X	X
1.6	X	X			X
1.7	X		X		X
1.8	X			X	X
1.9		X	X		X
1.10		X		X	X
1.11			X	X	X
1.12	X	X	X		X
1.13	X	X		X	X
1.14	X		X	X	X
1.15		X	X	X	X
1.16	X	X	X	X	X

t-test was performed for each gene to identify the differently expressed genes. We then compared our models using TSP-selected coexpressed genes with the models described by Sboner et al. [5].

3. Results

We proposed the LR models by combining TSP-selected genes and clinical features to identify and predict the patients whose PCa will progress. The performance of the models was evaluated with dataset GSE 16560. Table 1 lists the 16 LR models that we tested. Our models include all possible combinations of the following variables: age, GS, tumor percentage, presence of ERG gene rearrangement, and TSP-selected genes. The AUCs of 1,000 different partitions were calculated to select the best models. Figure 1(a) shows the AUC boxplots for the 100 tenfold cross validations of the 16 models listed in Table 1, with one pair of TSP-selected genes per model. The red stars denote the AUC values from the validation datasets that correspond to the best LR models from the learning dataset. Figure 1(b) shows the AUC boxplots for the same 16 models but with two TSP-selected gene pairs per model.

We plotted the AUC values of the validation dataset to assess the effects of the variables on the models (Figure 2). The blue line represents AUC values from the models with one TSP-selected gene pair, and the black line represents those from the models with two TSP-selected gene pairs. Furthermore, we tested the statistical significance of the models based on the null hypothesis that there is no difference between the AUCs of Sboner's models. It is found that most of AUC values in Sboner's models were out of the 95% confidence intervals of AUCs in our models. So our models can provide an alternative in predicting prostate cancer progression. The addition of TSP-selected gene pairs can improve our models' prediction of PCa progression, which differed from Sboner's results.

What is the role of TSP-selected gene pairs in comparison with the fusion ERG and the other clinical features, especially GS? Obviously, the GS was the most statistically significant variable because all the top models included it. In Figure 2, the red circles label the 8 models that include the GS. The AUCs in those 8 models were much higher than they were in the others and were very similar in the one- and two-gene-pair models. The 8 AUCs were more than 0.8, as shown in Figure 2, so we can conclude that the models with TSP-selected gene pairs performed better than all of Sboner's models, for which the largest AUC was 0.79 in [5].

The model using only the GS yielded an AUC of 0.76; by adding fusion ERG, the largest AUC observed by Sboner et al. was 0.79 [5]. Similarly, the other models that used only GS and tumor percent (or age) without molecular profiles could yield a higher AUC if fusion ERG was added. Therefore, the addition of fusion ERG may improve the prediction capability of models that use only clinical features [5].

However, the effect of fusion ERG was a little different in our analysis. First, our models could perform better by replacing fusion ERG with TSP-selected genes. In comparison with the best model with GS and fusion ERG (AUC, 0.79) in [5], our model 1.3 with the GS and TSP-selected gene pairs performed better, with an AUC of 0.84 (95% CI = [0.81, 0.88]); our best model was model 1.9, which used GS, tumor percentage, and one TSP-selected gene pair (AUC, 0.86; 95% CI = [0.79, 0.92]), but the corresponding model reported by Sboner with fusion ERG for replacement yielded an AUC of 0.75 [5]. On the other hand, the addition of fusion ERG had little or no effect on our models that included TSP-selected gene pairs. For example, the same AUC was obtained with our Model 1.3 and 1.10, with GS, fusion ERG, and TSP-selected gene pairs. Thus, TSP-selected genes seem to have a more important effect than the fusion ERG does in predicting PCa progression.

The addition of genes other than fusion ERG could not improve the prediction capability because the best models in the Sboner study [5] lacked molecular profiles. However,

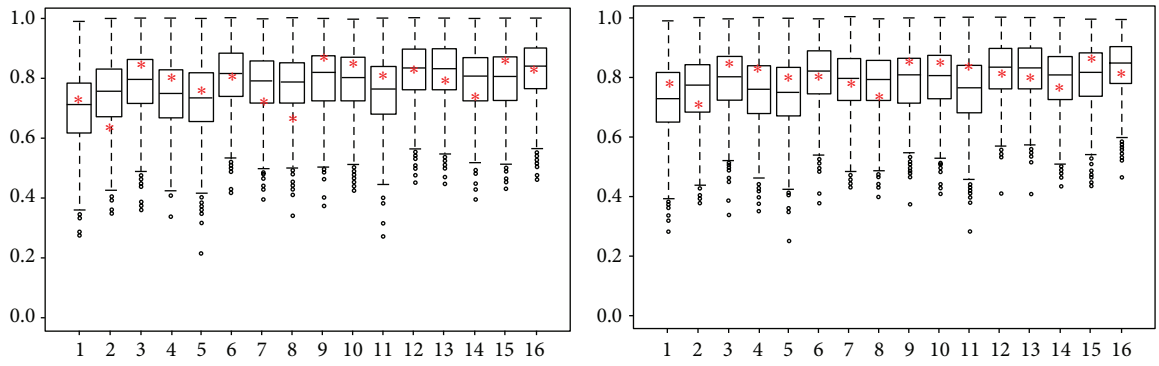


Figure 1: AUC boxplots for 100 tenfold cross validations of 16 models that include TSP-selected genes. The x-axis is the index of 16 models listed in Table 1, and the y-axis is the AUC values. The red star denotes the corresponding AUC values of the validation dataset that uses the best logistic regression models from the learning dataset. (a) Models that included one pair of TSP-selected genes and (b) those that included two such gene pairs.

Table 2: Comparison of the performance of our logistic regression models with that of the nine models evaluated by Sboner et al. [5], using the same number of genes.

Model number	Patient's age	Gleason score	Tumor percentage	Fusion ERG	Number of genes	AUC in ref. [5]	AUC of our model
2.1				X	18	0.672	0.769
2.2	X			X	9	0.708	0.732
2.3					18	0.713	0.736
2.4		X		X	21	0.726	0.793
2.5	X				11	0.730	0.712
2.6	X	X	X		3	0.738	0.806
2.7	X	X		X	12	0.745	0.804
2.8		X			16	0.749	0.813
2.9	X	X			12	0.750	0.788

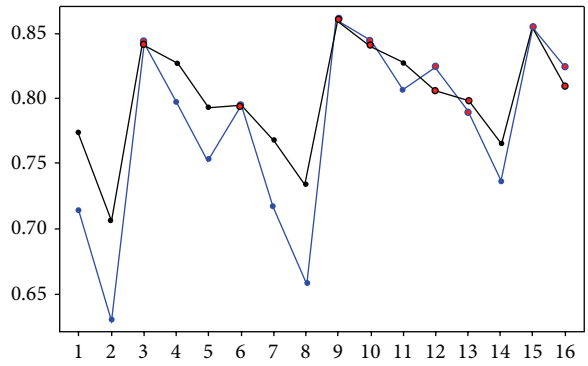


Figure 2: The AUCs of the 16 best models from the validation dataset. The x-axis is the index of the 16 models listed in Table 1, and the y-axis is the AUC values. The blue line shows the AUCs from the models with one TSP-selected gene pair, and the black line shows those from models with two TSP-selected gene pairs. The points circled in red are the AUCs in the 8 models that included the Gleason score as a variable.

some improvement was observed in our study: by replacing the molecular profiles in the Sboner models with one or two TSP-selected gene pairs, our models performed better than theirs did. For example, the best model with molecular profiles in the Sboner study used GS, age, and 12 genes and yielded an AUC of 0.75, whereas our model 1.6, which used GS, age, and TSP-selected gene pairs, yielded an AUC of 0.8 (95% CI = [0.76, 0.85]). Thus, although we added fewer genes to our model, its performance was better. Moreover, the prediction capability of Sboner's models was also improved when the same number of genes was replaced with TSP-selected gene pairs, as demonstrated in Table 2. Therefore, adding the TSP-selected genes had an important effect on the performance of the original models.

Models that use only clinical features may perform better if appropriate genes, such as those selected with the TSP algorithm, are added. To explore the effect of adding genes, we compared our approach with that used in the nine models in the original study by Sboner et al. [5] that included genes. For the comparison, we selected the same number of genes for our models. However, the featured genes in the models differed each time because the 1,000 random training and testing partitions were different in the iterative cross validation procedure.

The results of our comparison were presented in Table 2. As noted, the AUCs in our models were often higher than those in the study by Sboner et al. The prognostic models for

PCa can perform better if the featured genes are selected. In particular, TSP-selected genes may play an important role. First, the AUC of the model using only 18 genes increased from 0.71 in Sboner's study to 0.74 (95% CI = [0.71, 0.77]) in our model. Further, the AUCs of the models that used one and two TSP-selected gene pairs were 0.71 (95% CI = [0.67, 0.74]) and 0.77 (95% CI = [0.73, 0.81]), respectively. Thus, the TSP-based models performed better with a smaller number of genes.

The model from the Sboner study that included the GS and 16 genes did not perform any better than their model did that used the GS only, with AUCs of 0.75 and 0.75, respectively [5]. However, the AUC of our model that included the GS and 16 TSP-selected genes was 0.81 (95% CI = [0.76, 0.85]) in Table 2, and the models that used GS and one (or two) TSP gene pair(s) performed better, with an AUC of 0.84 in Figure 2.

Finally, of all the models tested in the original study, the one that included the GS and the ERG rearrangement (with no gene expression data) had the highest AUC value, 0.79 [5], whereas most of the AUC values for our models were higher than that. Therefore, in contrast to the conclusion reached by Sboner et al., we believe that adding the molecular profiles can improve the results obtained with the traditional prognostic models of PCa if the appropriate genes are selected.

From the results in Figure 2 and Table 2, we may conclude that the models' performance was not improved by the addition of large numbers of genes but was improved by the addition of significant clinical features and molecular profiles. For example, adding one TSP-selected gene pair is enough if the important clinical variables, such as GS, are included in the model. However, in the case of model 1.1, which included only one gene signature, and models 1.2 and 1.8, which also included age, the addition of more gene pairs can greatly improve the performance. Obviously, the gene selection strongly depends on the patient samples and so some statistical techniques such as bootstrap, repeated sampling, or cross validation were commonly used in the TSP-extended algorithms. In the current research, the computation cost of TSP-based algorithms is not the main concern, but the topics about the optimal number of gene pairs added to improve the clinical models are still interesting in further research.

4. Conclusion and Discussion

We have introduced an LR-based classification method that combines TSP-selected genes and clinical measurements. The empirical results of [19, 20] based on the datasets of prostate cancer progression show that the classification models using one or two TSP-selected gene pairs perform better than most commonly used one-gene-at-a-time approaches. With the combination of LR, our models not only preserved the basic advantages of the TSP algorithm but also incorporated the clinical features. Furthermore, the LR-TSP model provides the underlying probability of predictionand coexpressed genes that are used as biomarkers in the model. Thus, our proposed method provides explicit biologic interpretation of the clinical tests. Based on the statistical inference with the iterative cross validation, the better performance was shown in our models.

As mentioned in the report of Sboner et al. [5], many factors can influence the performance of the models, such as the definitions of lethal and indolent PCa, the use of samples contaminated with stromal tissue, the selection of genes assayed using a DASL (cDNA-mediated annealing, selection, extension, and ligation) array, and the effect of intertumor heterogeneity. Based on the study of GSE 16560, we explored the possible effect of genes used in the clinical models. The featured genes are often selected by using a one-gene-at-a-time approach. Sboner et al. performed a two-sided t-test for each gene within the training$_i$ partition, thereby avoiding overfitting because the selection of the genes was performed on only training sets [5]. They also implemented stepwise-like feature selection, sorted the genes according to their P values from the t-testing, and then added the genes to their models. Our study, on the other hand, demonstrates that coexpression analysis yields better prediction of tumor progression than the analysis of individual genes does. Therefore, we conclude that TSP selection is a useful tool for feature (and/or gene) selection to use in prognostic models.

Acknowledgments

This study was supported by the David Koch Center for Applied Research in Genitourinary Cancer, National Cancer Institute Grant CA16672, National Natural Science Funds of China (31100958), and GDHVPS (2012).

References

[1] C. S. Grasso, Y. M. Wu, D. R. Robinson et al., "The mutational landscape of lethal castration-resistant prostate cancer," *Nature*, vol. 487, pp. 239–243, 2012.

[2] F. H. Schröder, J. Hugosson, M. J. Roobol et al., "Prostate-cancer mortality at 11 years of follow-up," *The New England Journal of Medicine*, vol. 366, pp. 981–990, 2012.

[3] A. Qaseem, M. J. Barry, T. D. Denberg et al., "Screening for prostate cancer: a guidance statement from the clinical guidelines committee of the American college of physicians," *Annals of Internal Medicine*, vol. 158, no. 10, pp. 761–769, 2013.

[4] S. M. Dhanasekaran, A. Dash, J. Yu et al., "Molecular profiling of human prostate tissues: insights into gene expression patterns of prostate development during puberty," *The FASEB Journal*, vol. 19, no. 2, pp. 243–245, 2005.

[5] A. Sboner, F. Demichelis, S. Calza et al., "Molecular sampling of prostate cancer: a dilemma for predicting disease progression," *BMC Medical Genomics*, vol. 3, article 8, 2010.

[6] E. LaTulippe, J. Satagopan, A. Smith et al., "Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease," *Cancer Research*, vol. 62, no. 15, pp. 4499–4506, 2002.

[7] J.-H. Luo, Y. P. Yu, K. Cieply et al., "Gene expression analysis of prostate cancers," *Molecular Carcinogenesis*, vol. 33, no. 1, pp. 25–35, 2002.

[8] K. Tamura, M. Furihata, T. Tsunoda et al., "Molecular features of hormone-refractory prostate cancer cells by genome-wide gene expression profiles," *Cancer Research*, vol. 67, no. 11, pp. 5117–5125, 2007.

[9] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[10] C. Peterson and M. Ringner, "Analyzing tumor gene expression files," *Artificial Intelligence in Medicine*, vol. 28, pp. 59–74, 2003.

[11] P. Xu, G. N. Brock, and R. S. Parrish, "Modified linear discriminant analysis approaches for classification of high-dimensional microarray data," *Computational Statistics and Data Analysis*, vol. 53, no. 5, pp. 1674–1687, 2009.

[12] U. R. Chandran, C. Ma, R. Dhir et al., "Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process," *BMC Cancer*, vol. 7, article 64, 2007.

[13] R. S. Hudson, M. Yi, D. Esposito et al., "MicroRNA-106b-25 cluster expression is associated with early disease recurrence and targets caspase-7 and focal adhesion in human prostate cancer," *Oncogene*, vol. 32, no. 35, pp. 4139–4147, 2012.

[14] E. Martinez and V. Trevino, "Modelling gene expression profiles related to prostate tumor progression using binary states," *Theoretical Biology and Medical Modelling*, vol. 10, article 37, 2013.

[15] S. Feng, O. Dakhova, C. J. Creighton, and M. M. Ittmann, "The endocrine fibroblast growth factor FGF19 promotes prostate cancer progression," *Cancer Research*, vol. 73, no. 8, pp. 2551–2562, 2012.

[16] D. Geman, C. d'Avignon, D. Q. Naiman, and R. L. Winslow, "Classifying gene expression profiles from pairwise mRNA comparisons," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, article 19, 2004.

[17] J. Zhu and T. Hastie, "Classification of gene microarrays by penalized logistic regression," *Biostatistics*, vol. 5, no. 3, pp. 427–443, 2004.

[18] L. Shen and E. C. Tan, "Dimension reduction-based penalized logistic regression for cancer classification using microarray data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 2, pp. 166–175, 2005.

[19] J. G. Liao and K.-V. Chin, "Logistic regression for disease classification using microarray data: model selection in a large p and small n case," *Bioinformatics*, vol. 23, no. 15, pp. 1945–1951, 2007.

[20] A. C. Tan, D. Q. Naiman, L. Xu, R. L. Winslow, and D. Geman, "Simple decision rules for classifying human cancers from gene expression profiles," *Bioinformatics*, vol. 21, no. 20, pp. 3896–3904, 2005.

[21] L. Xu, A. C. Tan, D. Q. Naiman, D. Geman, and R. L. Winslow, "Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data," *Bioinformatics*, vol. 21, no. 20, pp. 3905–3911, 2005.

[22] Y. Zhang, J. Szustakowski, and M. Schinke, "Bioinformatics analysis of microarray data," *Methods in Molecular Biology*, vol. 573, pp. 259–284, 2009.

[23] R. Ummanni, S. Teller, H. Junker et al., "Altered expression of tumor protein D52 regulates apoptosis and migration of prostate cancer cells," *FEBS Journal*, vol. 275, no. 22, pp. 5703–5713, 2008.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 610297, 11 pages
http://dx.doi.org/10.1155/2013/610297

Research Article

Power and Stability Properties of Resampling-Based Multiple Testing Procedures with Applications to Gene Oncology Studies

Dongmei Li[1] and Timothy D. Dye[2]

[1] *Department of Public Health Sciences, Office of Public Health Studies, The University of Hawaii at Manoa, 1960 East-West Road, Honolulu, HI 96822, USA*
[2] *Department of Obstetrics, Gynecology, and Women's Health, John A. Burns School of Medicine, University of Hawaii, 651 Ilalo Street, Honolulu, HI 96813, USA*

Correspondence should be addressed to Dongmei Li; dongmeil@hawaii.edu

Received 10 August 2013; Revised 14 October 2013; Accepted 18 October 2013

Academic Editor: Ao Yuan

Resampling-based multiple testing procedures are widely used in genomic studies to identify differentially expressed genes and to conduct genome-wide association studies. However, the power and stability properties of these popular resampling-based multiple testing procedures have not been extensively evaluated. Our study focuses on investigating the power and stability of seven resampling-based multiple testing procedures frequently used in high-throughput data analysis for small sample size data through simulations and gene oncology examples. The bootstrap single-step min$P$ procedure and the bootstrap step-down min$P$ procedure perform the best among all tested procedures, when sample size is as small as 3 in each group and either familywise error rate or false discovery rate control is desired. When sample size increases to 12 and false discovery rate control is desired, the permutation max$T$ procedure and the permutation min$P$ procedure perform best. Our results provide guidance for high-throughput data analysis when sample size is small.

1. Introduction

With rapidly developing biotechnology, microarrays and next generation sequencing have been widely used in biomedical and biological fields for identifying differentially expressed genes, detecting transcription factor binding sites, and mapping complex traits using single nucleotide polymorphisms (SNPs) [1–7]. The multiple testing error rates associated with thousands, even millions of hypotheses testing, need to be taken into account. Common multiple testing error rates controlled in multiple hypotheses testing are the familywise error rate (FWER), which is the probability of at least one false rejection [8, 9] and the false discovery rate (FDR), which is the expected proportion of falsely rejected null hypotheses [10].

Resampling-based multiple testing procedures are widely used in high-throughput data analysis (e.g., microarray and next generation sequencing), especially when the sample size is small or the distribution of test statistic is nonnormally distributed or is unknown. Resampling-based multiple testing procedures can account for dependent structures among P values or test statistics, resulting in lower type II errors. The commonly used resampling techniques include permutation tests and bootstrap methods.

Permutation tests are nonparametric statistical significance tests, where the test statistics' distribution under the null hypothesis is constructed by calculating all possible values or a concrete number of test statistics (usually 1000 or above) from permuted observations under the null hypothesis. The theory of the permutation test is based on the work by Fisher and Pitman in the 1930s. Permutation tests are distribution-free, which can provide exact P values even when sample size is small.

The bootstrap method, first introduced by Efron [11] and further discussed by Efron and Tibshirani [12], is a way of approximating the sampling distribution from just one sample. Instead of taking many simple random samples from the population to find the sampling distribution of a sample

statistic, the bootstrap method repeatedly samples with replacement from one random sample. Efron [11] showed that the bootstrap method provides an asymptotically unbiased estimator for the variance of a sample median and for error rates in a linear discrimination problem (outperforming cross-validation). Freedman [13] conclusively showed that bootstrap approximation of the distribution of least square estimates is valid. Finally, Hall [14] showed that the bootstrap method's reduction of error coverage probability, from $o(n^{-1/2})$ to $o(n^{-1})$, makes the bootstrap method one order of magnitude more accurate than the delta method. The P values computed by the bootstrap method are less exact than P values obtained from the permutation method, and additionally, P values estimated by the bootstrap method are asymptotically convergent to the true P values [15].

Different resampling methods can draw different conclusions, however, when applied to the same data set. An investigation of multiple testing error rate control, power, and stability of those resampling methods under different situations is necessary to provide guidance for data analysis, so that optimal methods in different scenarios could be used to maximize power and minimize multiple testing error rates.

In this paper, we focus on investigating the power and stability properties of several commonly used resampling-based multiple testing procedures: (1) the permutation tests [16]; (2) the permutation-based significant analysis of microarray (SAM) procedure [17]; and (3) the bootstrap multiple testing procedures [15].

2. Materials and Methods

2.1. Permutation Test. To carry out a permutation test based on a test statistic that measures the size of an effect of interest, we proceed as follows.

(1) Compute the test statistics for the observed data set, such as two sample t-test statistics.

(2) Permute the original data in a way that matches the null hypothesis to get permuted resamples and construct the reference distribution using the test statistics calculated from permuted resamples.

(3) Calculate the critical value of a level α test based on the upper α percentile of the reference distribution, or obtain the raw P value by computing the proportion of permutation test statistics that are as extreme as or more extreme than the observed test statistic.

Westfall and Young [16] proposed two methods to adjust raw P values to control the multiple testing error rates. One is single-step min$P$ procedure and the other is single-step max$T$ procedure.

The single-step min$P$ adjusted P values are defined as [18]

$$\widetilde{p}_i = \Pr\left(\min_{1\le l\le m} P_l \le p_i \mid H_M\right). \quad (1)$$

The single-step max$T$ adjusted P values are defined in terms of test statistics T_i, namely, [18]

$$\widetilde{p}_i^T = \Pr\left(\max_{1\le l\le m} |T_l| \ge |t_i| \mid H_M\right), \quad (2)$$

where H_M is the complete null hypothesis. P_l is the raw P value for the lth hypothesis, and T_l is the observed test statistic for the lth hypothesis.

2.2. Significance Analysis of Microarrays (SAM) Procedure. The Significance Analysis of Microarrays (SAM) procedure proposed by Tusher et al. [17] identifies genes with significant changes in expression using a set of gene-specific t-tests. In SAM, genes are assigned with scores relative to change in gene expression and its standard deviation of repeated measurements. Scatter plots of the observed relative differences and the expected relative differences estimated through permutation identifies statistically significant genes based on a fixed threshold.

Based on the description of SAM in Tusher et al. [17], the SAM procedure can be summarized as follows.

(1) Compute a test statistic t_i for each gene i ($i = 1, \ldots, g$).

(2) Compute order statistics $t_{(i)}$ such that $t_{(1)} \le t_{(2)} \le \cdots \le t_{(g)}$.

(3) Perform B permutations of the responses/covariates $y_1, \ldots, y_n$. For each permutation b, compute the permuted test statistics $t_{i,b}$ and the corresponding order statistics $t_{(1),b} \le t_{(2),b} \le \cdots \le t_{(g),b}$.

(4) From the B permutations, estimate the expected values of order statistics by $\bar{t}_{(i)} = (1/B)\sum_{b=1}^{B} t_{(i),b}$.

(5) Form a quantile-quantile (Q-Q) plot (SAM plot) of the observed $t_{(i)}$ versus the expected $\bar{t}_{(i)}$.

(6) For a given threshold Δ, starting at the origin, and moving up to find the first $i = i_1$ such that $t_{(i)} - \bar{t}_{(i)} > \Delta$. All genes past i_1 are called significant positives. Similarly, starting at the origin and moving down to the left, find the first $i = i_2$ such that $\bar{t}_{(i)} - t_{(i)} > \Delta$. All genes past i_2 are called significant negatives. Define the upper cut point $\mathrm{Cut}_{\mathrm{up}}(\Delta) = \min\{t_{(i)} : i \le i_1\} = t_{(i_1)}$ and the lower cut point $\mathrm{Cut}_{\mathrm{low}}(\Delta) = \max\{t_{(i)} : i \ge i_2\} = t_{(i_2)}$.

(7) For a given threshold, the expected number of false rejections $E(V)$ is estimated by computing the number of genes with $t_{i,b}$ above $\mathrm{Cut}_{\mathrm{up}}(\Delta)$ or below $\mathrm{Cut}_{\mathrm{low}}(\Delta)$ for each of the B permutation and averaging the numbers over B permutations.

(8) A threshold Δ is chosen to control the Fdr (Fdr $= (E(V)/r)$) under the complete null hypothesis, at an acceptable nominal level.

2.3. Bootstrap Method. The bootstrap method based on estimated null distribution of test statistics was introduced by Pollard and van der Laan [15] and proceeds as follows:

(1) Compute the observed test statistic for the observed data set.

(2) Resample the data with replacement within each group to obtain bootstrap resamples, compute the resampled test statistics for each resampled data set,

and construct the reference distribution using the centered and/or scaled resampled test statistics.

(3) Calculate the critical value of a level α test based on the upper α percentile of the reference distribution, or obtain the raw P values by computing the proportion of bootstrapped test statistics that is as extreme as or more extreme than the observed test statistic.

The MTP function based on the bootstrap method includes single-step min$P$ and max$T$ adjusted P values, as well as step-down min$P$ and step-down max$T$ adjusted P values. The single-step max$T$ and min$P$ adjusted P values are defined as before.

The step-down min$P$ adjusted P values are defined as

$$\widetilde{p}_{r_i} = \max_{k=1,\ldots,i} \left\{ \Pr\left(\min_{l=k,\ldots,m} P_{r_l} \leq p_{r_k} \mid H_M \right) \right\}, \quad (3)$$

and the step-down max$T$ adjusted P values are defined as

$$\widetilde{p}_{S_i} = \max_{k=1,\ldots,i} \left\{ \Pr\left(\max_{l=k,\ldots,m} \left|T_{s_l}\right| \geq \left|t_{s_k}\right| \mid H_M \right) \right\}, \quad (4)$$

where $|t_{s_1}| \geq |t_{s_2}| \geq \cdots \geq |t_{s_m}|$ denote the ordered test statistics [18].

2.4. Simulation Setup. Simulation studies were conducted to compare the power and stability of the resampling-based multiple testing procedures for both independent test statistics and dependent test statistics. According to Rubin et al. [19], the power is defined as the expected proportion of true positives. The stability is measured as the variance of true discoveries and variance of total discoveries.

In our first simulation study, each set includes 100 independently generated groups of two samples with equal sample size of 3 or 12 in each group. 100 repetitions are chosen because computationally 100 is more efficient than 1000 or even higher repetitions. 1000 repetitions are also tried and similar results are obtained. Thus, 100 repetitions are chosen for computational efficiency. The total number of genes (m) is set to be 2000 with the fraction of true null hypotheses (m_0/m) at 50%. In the two-group comparison, the standardized logarithms of gene expression levels are generated from multivariate normal distribution. One group has 50% of genes with means at μ and the remaining with means at 0. All genes in the other group have means at 0. The mean expression level μ on log2 scale is set to be from 1 to 6 with step 0.50 for the first simulation study. The variances of the standardized logarithm of gene expression levels are equal to 1 in both groups. Thus, the mean differences of μ in gene expression between the two groups are the Cohen's d effect sizes. The pairwise correlation coefficients of test statistics are set to be 0 in our simulation study. The test statistics used are equal variance t-test throughout the simulation study. The FWER/FDR level is set at 5% ($\alpha = 0.05$).

We conducted another simulation study to examine the effect of fraction of true null hypotheses on power and stability. In our second simulation study, each data set includes 100 independently generated samples of two groups with equal sample size of 3. The total number of genes (m) is set to be 1000, with the fraction of differentially expressed genes ($(m - m_0)/m$) equal to 10%, 25%, 50%, 75%, and 90% to cover all possible scenarios. In the two-group comparison, the gene expression level on log2 scale is generated randomly from a multivariate normal distribution with $\mu = 0$ and $\sigma = 1$. The correlations between genes are randomly fluctuated between 0 and 1 to mimic the correlations in real microarray data. The mean differences are set between 1 and 2, with the step equaling the inverse of the number of differentially expressed gene ($m - m_0$). The variances are set to 1. Equal variance t-tests are used for this simulation study, and the FWER/FDR level is set at 5% ($\alpha = 0.05$).

The mt.max$T$ and mt.min$P$ functions in R were used to evaluate the Westfall and Young's permutation test. The sam function in R was used for the SAM procedure. The Bootstrap method proposed by Pollard and van der Laan [15] was executed using the MTP function in R. The MTP function includes the max$T$ method, the min$P$ method, the single step procedure, and the step-down procedure, which results into four different functions, including single-step max$T$ (ss.max$T$), single-step min$P$ (ss.min$P$), step-down max$T$ (sd.max$T$), and step-down min$P$ (sd.min$P$).

2.5. Cancer Microarray Example. Ovarian cancer is a common cause of cancer deaths in women [20]. Microarray experiments were conducted to identify differentially expressed genes between chemotherapy favorable patients and chemotherapy unfavorable patients [21]. Those differentially expressed genes could be used to develop optimal treatment for a new ovarian cancer patient by predicting possible response to chemotherapy. The gene expression data of 12,625 genes from 6 patients' mRNA samples, obtained from Moreno et al.'s ovarian cancer microarray study, were used to show the differences in the number of total discoveries among those resampling-based multiple testing procedures with FWER or FDR controlled at 5% (data accessible at NCBI GEO database [22], accession GSE7463). The preprocessing of the ovarian cancer data set was done using the RMA background correction, quantile normalization, and robust linear model summarization. The raw P values and the adjusted P values of comparisons between the chemotherapy favorable group (3 subjects) and chemotherapy unfavorable group (3 subjects) were calculated using the resampling-based multiple testing functions in the multitest package and the siggenes package in R.

3. Results

Simulation studies were conducted to compare the power and stability across all tested multiple testing procedures for normally distributed data with either independent or randomly correlated test statistics. The sample size is 3 or 12 in each group for independent test statistics and 3 in each group for randomly correlated test statistics.

3.1. Simulation Results for Independent Test Statistics. For independent test statistics with FWER controlled at 5%,

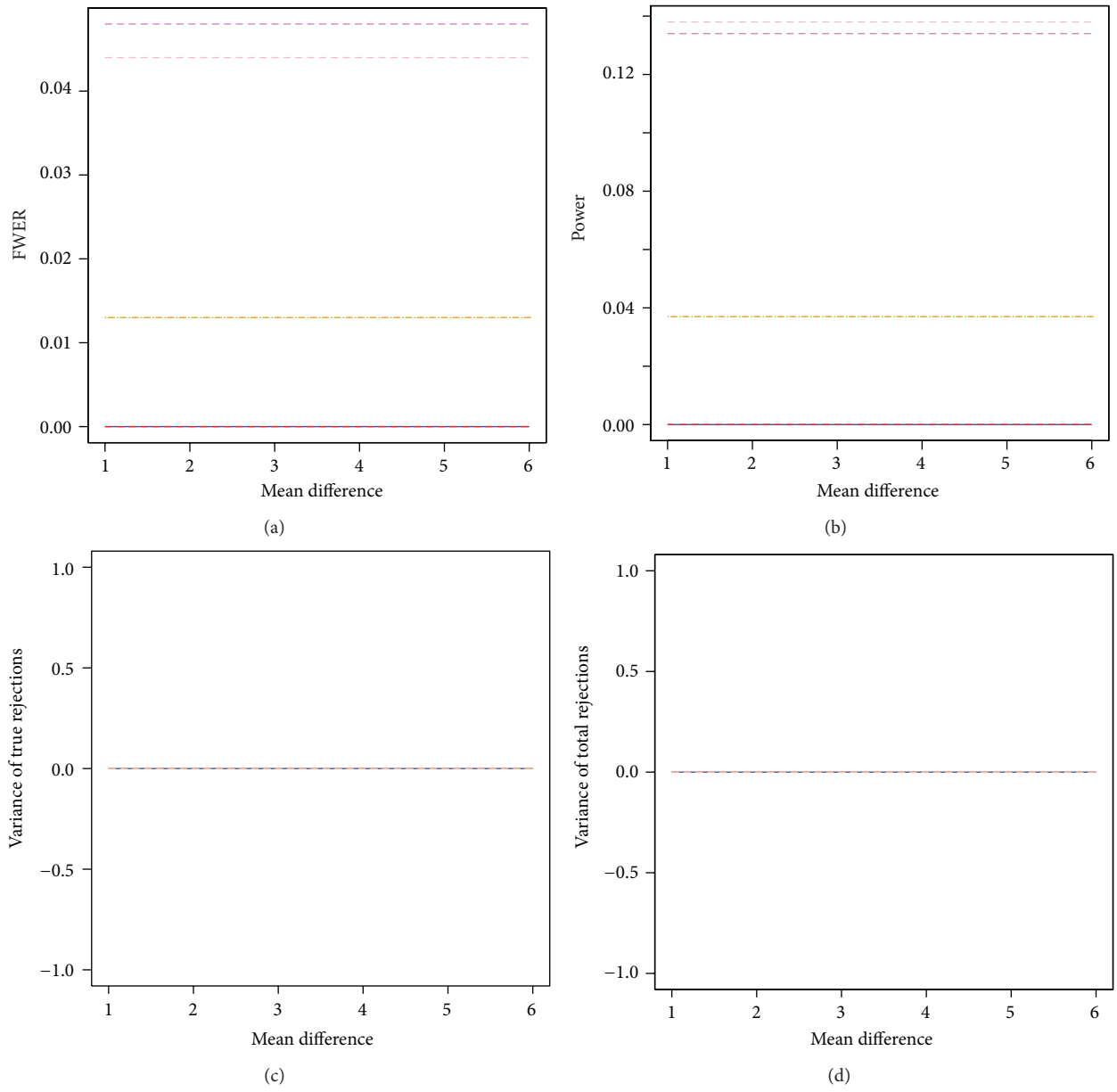


FIGURE 1: Power and stability properties of resampling-based multiple testing procedures for independent test statistics with FWER controlled at 5% and small sample size of 3 in each group ($m_0/m = 50\%$). Solid blue line: permutation single-step max$T$ procedure (mt.max$T$ function); red dashed line: permutation single-step min$P$ (mt.min$P$ function); green dotted line: bootstrap single-step max$T$ (MTP ss.max$T$ function); violet dashed line: bootstrap single-step min$P$ (MTP ss.min$P$ function); orange dashed line: bootstrap step-down max$T$ (MTP sd.max$T$ function); pink dashed line: bootstrap step-down min$P$ (MTP sd.min$P$ function).

the two bootstrap min$P$ procedures outperformed all other tested procedures when sample size is 3 in each group (Figure 1). Both the bootstrap single-step min$P$ and the bootstrap step-down min$P$ procedures were more powerful than all other tested procedures, and their FWER estimates were close to 5% nominal level. The two permutation-based procedures (mt.max$T$ and mt.min$P$) had no power to detect any significant difference between groups, and their FWER estimates were close to 0. The power of the bootstrap max$T$ procedures (ss.max$T$ and sd.max$T$) were between the permutation procedures and the bootstrap min$P$ procedures. The estimated variances of true discoveries and total number of discoveries were around 0 for all tested resampling-based multiple testing procedures. The estimated FWER, power, and stability were constant across effect sizes.

The bootstrap single-step and step-down min$P$ procedures remained to have the largest power among all tested procedures when FDR was controlled at 5% and sample size was 3 in each group (Figure 2). The FDR estimates from the bootstrap single-step and step-down min$P$ procedures also stayed around 5% nominal level. Both the SAM procedure and the two permutation-based max$T$ and min$P$ procedures

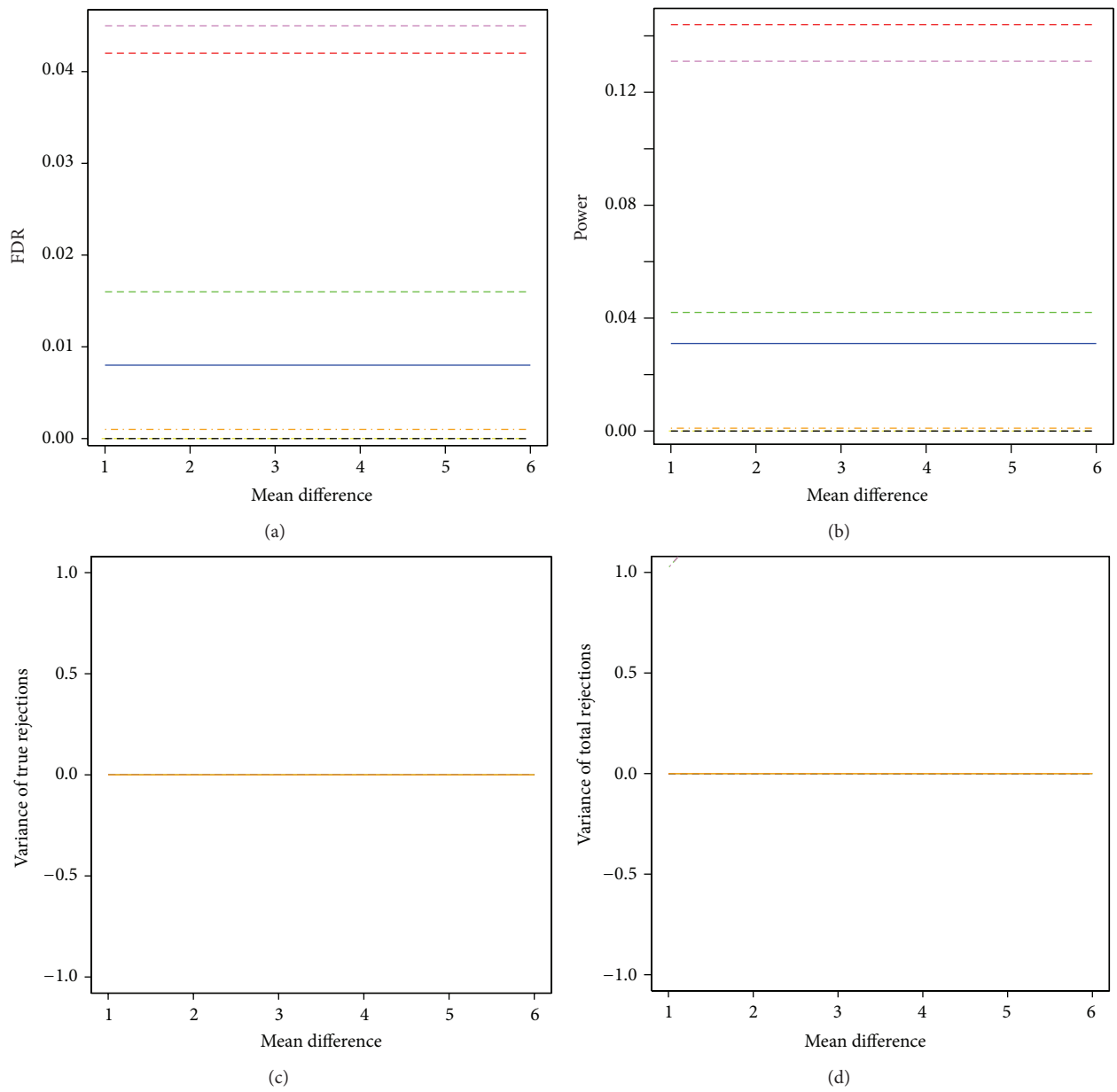


Figure 2: Power and stability properties of resampling-based multiple testing procedures for independent test statistics with FDR controlled at 5% and small sample size of 3 in each group (m_0/m = 50%). Yellow dashed line: permutation single-step max$T$ procedure (mt.max$T$ function); black dashed line: permutation single-step min$P$ (mt.min$P$ function); solid blue line: bootstrap single-step max$T$ (MTP ss.max$T$ function); red dashed line: bootstrap single-step min$P$ (MTP ss.min$P$ function); green dashed line: bootstrap step-down max$T$ (MTP sd.max$T$ function); violet dashed line: bootstrap step-down min$P$ (MTP sd.min$P$ function); orange dashed line: SAM procedure (sam function).

had no power to detect any significant difference, and their FDR estimates were also close to 0. Both the FDR estimates and power of the two bootstrap single-step and step-down max$T$ procedures were between the SAM procedure, the permutation procedures, and the bootstrap min$P$ procedures. All resampling-based multiple testing procedures had estimated variances of true discoveries and total number of discoveries around 0. The estimated FDR, power, and stability were constant across effect sizes.

The bootstrap step-down min$P$ procedure had the largest power across all tested procedures when sample size increased to 12 in each group (Figure 3). The bootstrap single-step max$T$ procedure, the bootstrap step-down max$T$ procedure, and the permutation single-step min$P$ procedure showed almost zero power for detecting any difference between groups. All tested procedures had FWER estimates around 0 and showed very small estimated variances of true rejections and variances of total rejections. The estimated FWER and power remained constant across effect sizes.

The permutation single-step max$T$ procedure and the permutation single-step min$P$ procedure performed the best when FDR is controlled at 5% and sample size is 12 in each

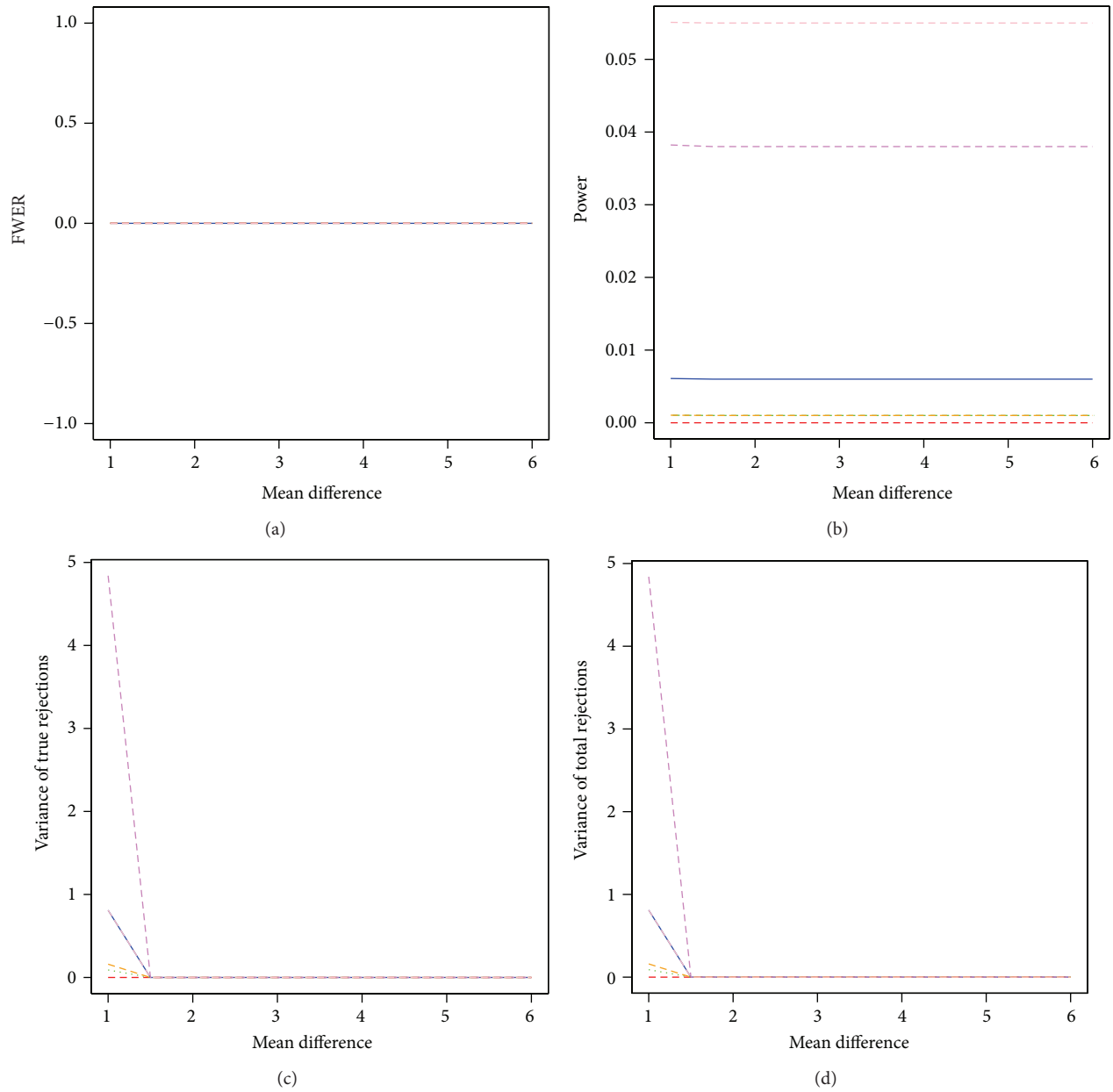


FIGURE 3: Power and stability properties of resampling-based multiple testing procedures for independent test statistics with FWER controlled at 5% and moderate sample size of 12 in each group ($m_0/m = 50\%$). Solid blue line: permutation single-step max$T$ procedure (mt.max$T$ function); red dashed line: permutation single-step min$P$ (mt.min$P$ function); green dotted line: bootstrap single-step max$T$ (MTP ss.max$T$ function); violet dashed line: bootstrap single-step min$P$ (MTP ss.min$P$ function); orange dashed line: bootstrap step-down max$T$ (MTP sd.max$T$ function); pink dashed line: bootstrap step-down min$P$ (MTP sd.min$P$ function).

group (Figure 4). The two permutation max$T$ and min$P$ procedures had much larger power than the four bootstrap MTP procedures and also had estimated FDR less than 5%. The SAM procedure failed to control the FDR at the desired level of 5%, although it had larger power than all other tested procedures. The estimated variances of total discoveries from the SAM procedure were much larger than all other procedures when the effect size is around 1. The permutation single-step max$T$ and min$P$ procedures had small variances of true discoveries and total discoveries. The four bootstrap MTP procedures had low power but similar stability as the permutation max$T$ and min$P$ procedures. The estimated FDR and power were also constant across effect sizes.

3.2. Simulation Results for Dependent Test Statistics. The two bootstrap min$P$ procedures (ss.min$P$ and sd.min$P$) showed higher power than all other tested procedures across various proportions of nontrue null hypotheses, when test statistics are dependent and FWER is controlled. The two bootstrap

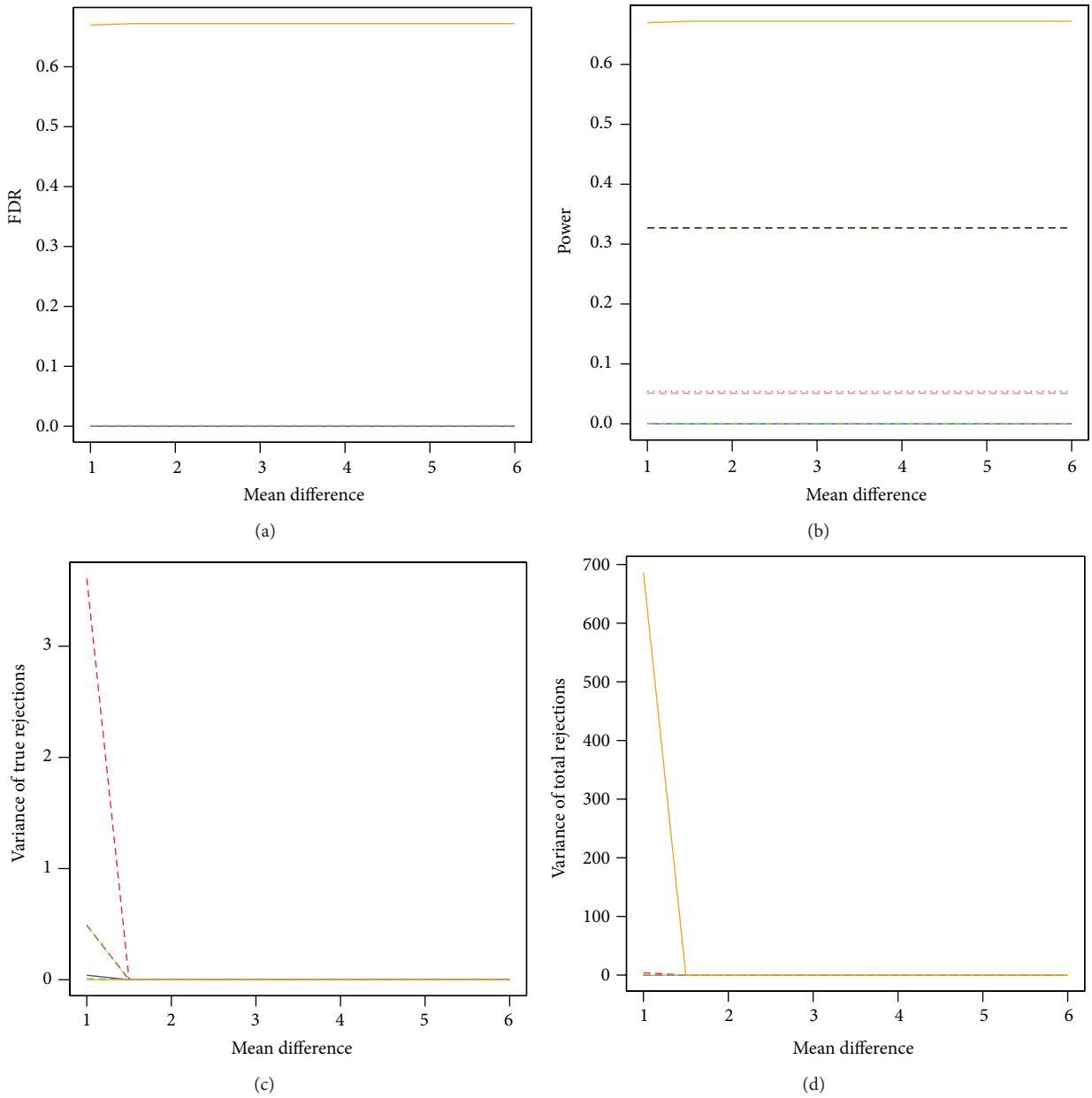


FIGURE 4: Power and stability properties of resampling-based multiple testing procedures for independent test statistics with FDR controlled at 5% and small sample size of 12 in each group (m_0/m = 50%). Yellow dashed line: permutation single-step max$T$ procedure (mt.max$T$ function); black dashed line: permutation single-step min$P$ (mt.min$P$ function); solid blue line: bootstrap single-step max$T$ (MTP ss.max$T$ function); red dashed line: bootstrap single-step min$P$ (MTP ss.min$P$ function); green dashed line: bootstrap step-down max$T$ (MTP sd.max$T$ function); violet dashed line: bootstrap step-down min$P$ (MTP sd.min$P$ function); orange dashed line: SAM procedure (sam function).

min$P$ procedures had desired FWER control as well, when the proportions of nontrue null hypotheses were greater than 50% (Table 1 and Figure 5). The two bootstrap max$T$ procedures (ss.max$T$ and sd.max$T$) had lower power than the two bootstrap min$P$ procedures. They showed desired FWER control, however, when the proportions of nontrue null hypotheses were over 25%. The permutation single-step max$T$ and min$P$ procedures had no power to detect any significant difference between groups. All resampling-based procedures had estimated variances of true discoveries and total discoveries around 0 across various proportions of nontrue null hypotheses, when sample size was as small as 3 in each group.

The power and stability of all four bootstrap methods (ss.min$P$, sd.min$P$, ss.max$T$, and sd.max$T$) and the two permutation methods (mt.max$T$ and mt.min$P$) showed similar results, when FDR were controlled, as that when FWER were controlled (Table 2 and Figure 6). The SAM procedure had decent FDR control, but very low power when the proportions of nontrue null hypotheses were less than 50%.

Table 1: Comparison of the estimated FWER and power for the resampling-based multiple testing procedures with FWER controlled at 5%.

	$(m - m_0)/m$	mt.max$T$	mt.min$P$	ss.max$T$	ss.min$P$	sd.max$T$	sd.min$P$
FWER	0.10	0.00	0.00	0.18	0.53	0.14	0.53
	0.25	0.00	0.00	0.07	0.19	0.05	0.19
	0.50	0.00	0.00	0.02	0.06	0.02	0.07
	0.75	0.00	0.00	0.01	0.02	0.01	0.02
	0.90	0.00	0.00	0.00	0.01	0.00	0.01
Power	0.10	0.00	0.00	0.11	0.27	0.09	0.30
	0.25	0.00	0.00	0.12	0.32	0.11	0.32
	0.50	0.00	0.00	0.10	0.28	0.09	0.29
	0.75	0.00	0.00	0.11	0.27	0.10	0.28
	0.90	0.00	0.00	0.11	0.28	0.11	0.28

mt.max$T$: permutation single-step max$T$ procedure; mt.min$P$: permutation single-step min$P$ procedure; ss.max$T$: bootstrap single-step max$T$ procedure; ss.min$P$: bootstrap single-step min$P$ procedure; sd.max$T$: bootstrap step-down max$T$ procedure; sd.min$P$: bootstrap step-down min$P$ procedure.

Table 2: Comparison of the estimated FDR and power for the resampling-based multiple testing procedures with FDR controlled at 5%.

	$(m - m_0)/m$	mt.max$T$	mt.min$P$	ss.max$T$	ss.min$P$	sd.max$T$	sd.min$P$	sam
FDR	0.10	0.00	0.00	0.13	0.53	0.13	0.54	0.04
	0.25	0.00	0.00	0.05	0.20	0.05	0.20	0.00
	0.50	0.00	0.00	0.02	0.07	0.02	0.07	0.00
	0.75	0.00	0.00	0.01	0.02	0.01	0.02	0.15
	0.90	0.00	0.00	0.00	0.01	0.00	0.01	0.55
Power	0.10	0.00	0.00	0.08	0.29	0.09	0.28	0.04
	0.25	0.00	0.00	0.10	0.31	0.10	0.33	0.00
	0.50	0.00	0.00	0.08	0.29	0.09	0.28	0.00
	0.75	0.00	0.00	0.09	0.29	0.10	0.28	0.15
	0.90	0.00	0.00	0.09	0.29	0.10	0.29	0.55

mt.max$T$: permutation single-step max$T$ procedure; mt.min$P$: permutation single-step min$P$ procedure; ss.max$T$: bootstrap single-step max$T$ procedure; ss.min$P$: bootstrap single-step min$P$ procedure; sd.max$T$: bootstrap step-down max$T$ procedure; sd.min$P$: bootstrap step-down min$P$ procedure; sam: the SAM procedure.

Both the estimated FDR and power increased when the proportions of nontrue null hypotheses were greater than 50% for the SAM procedure.

3.3. Real Data Example. The gene expression levels of 12625 genes from 6 subjects on log2 scale were used to compare total number of discoveries identified from all tested resampling-based multiple testing procedures (Table 3). The two bootstrap min$P$ procedures had more rejections than the two bootstrap max$T$ procedures, when FWER was controlled at 5%. The bootstrap step-down min$P$ and single-step min$P$ procedures remained higher number of rejections than the bootstrap step-down max$T$ and single-step max$T$ procedures, when FDR was controlled at 5%. The SAM procedure only rejected 2 genes. The permutation max$T$ and min$P$ procedures rejected none of those genes. The bootstrap multiple testing procedures has higher power than all other tested procedures and rejected much more null hypotheses compared to the permutation test procedures. The bootstrap min$P$ procedures rejected more hypotheses than the bootstrap max$T$ procedures. The total number of rejections from this real microarray data analysis is consistent with the results from the simulation studies.

Table 3: Comparisons of number of total discoveries for the resampling-based multiple testing procedures for the ovarian cancer example with 12,625 genes.

Resampling methods	Rejected number of hypothesis	
	FWER controlled at 5%	FDR controlled at 5%
mt.max$T$	0	0
mt.min$P$	0	0
ss.max$T$	250	385
sd.max$T$	407	397
ss.min$P$	1785	1649
sd.min$P$	1766	1706
sam		2

mt.max$T$: permutation single-step max$T$ procedure; mt.min$P$: permutation single-step min$P$ procedure; ss.max$T$: bootstrap single-step max$T$ procedure; ss.min$P$: bootstrap single-step min$P$ procedure; sd.max$T$: bootstrap step-down max$T$ procedure; sd.min$P$: bootstrap step-down min$P$ procedure; sam: the SAM procedure.

4. Discussion

This paper investigated the power and stability properties of several popular resampling-based multiple testing

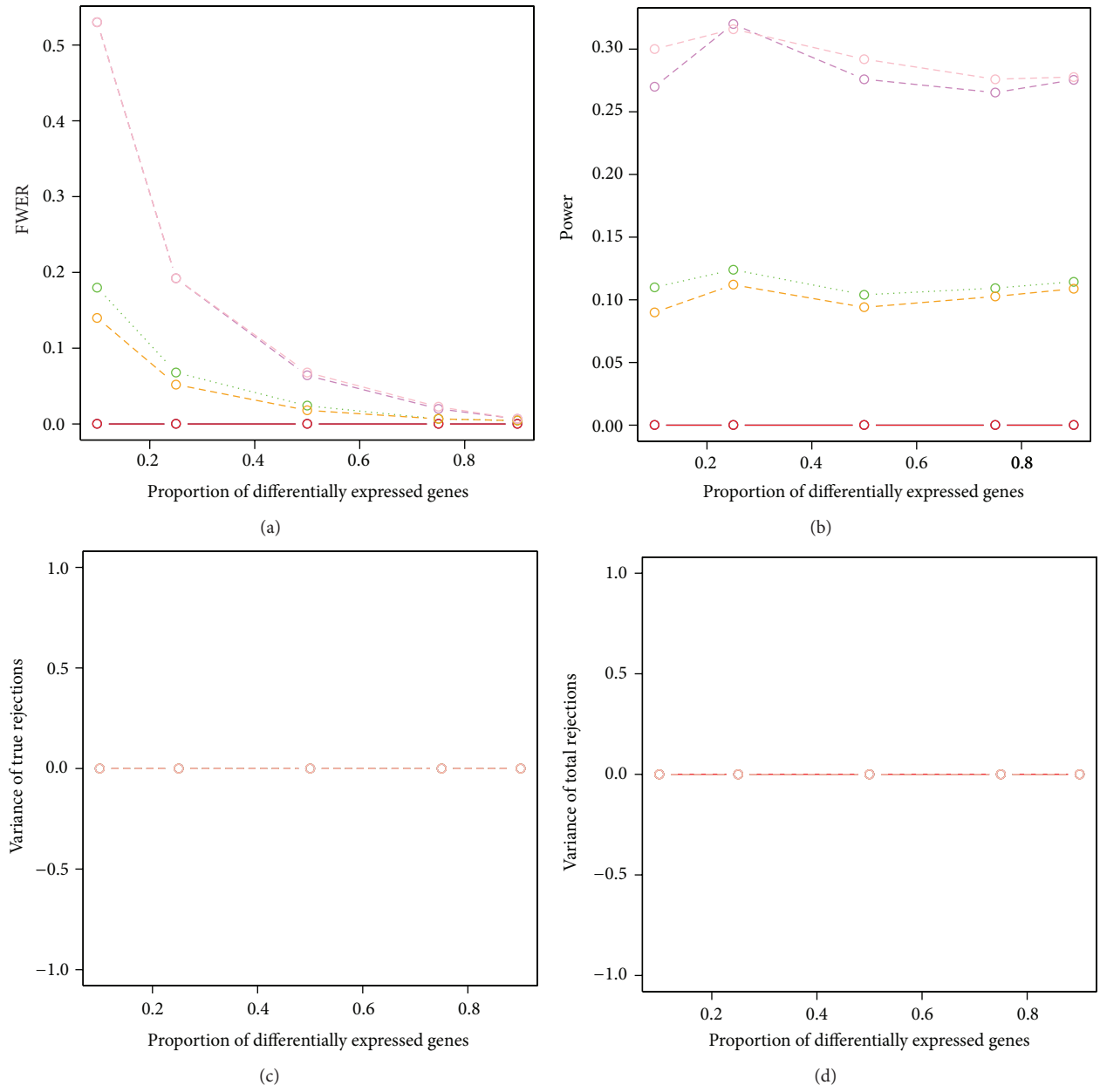


Figure 5: Power and stability properties of resampling-based multiple testing procedures for dependent test statistics with random correlations and the FWER is controlled at 5% ($n = 3$ in each group). Blue dashed line: permutation single-step max$T$ procedure (mt.max$T$ function); solid red line: permutation single-step min$P$ (mt.min$P$ function); green dotted line: bootstrap single-step max$T$ (MTP ss.max$T$ function); orange dashed line: bootstrap step-down max$T$ (MTP sd.max$T$ function); violet dashed line: bootstrap single-step min$P$ (MTP ss.min$P$ function); pink dashed line: bootstrap step-down min$P$ (MTP sd.min$P$ function).

procedures for both independent and dependent test statistics, when sample size is small or moderate, using available functions in *R*. Our simulation results and real data example show that the bootstrap single-step and step-down min$P$ procedures perform the best for both small sample size data (3 in each group) and moderate sample size data (12 in each group) when FWER control is desired. The bootstrap single-step and step-down min$P$ procedures are the best when FDR control is desired for data with small sample size (3 in each group). The permutation max$T$ and min$P$ procedures perform the best for data with moderate sample size when FDR control is desired. The SAM procedure overestimates FDR, although it has higher power than the permutation and bootstrap max$T$ and min$P$ procedures.

The simulation results also showed that the permutation test procedures have no power to detect any significant differences between groups when sample size is as small as 3 in each group; the permutation test procedures perform well when sample size increases to 12 in each group; the SAM procedure has no power for detecting significant differences when

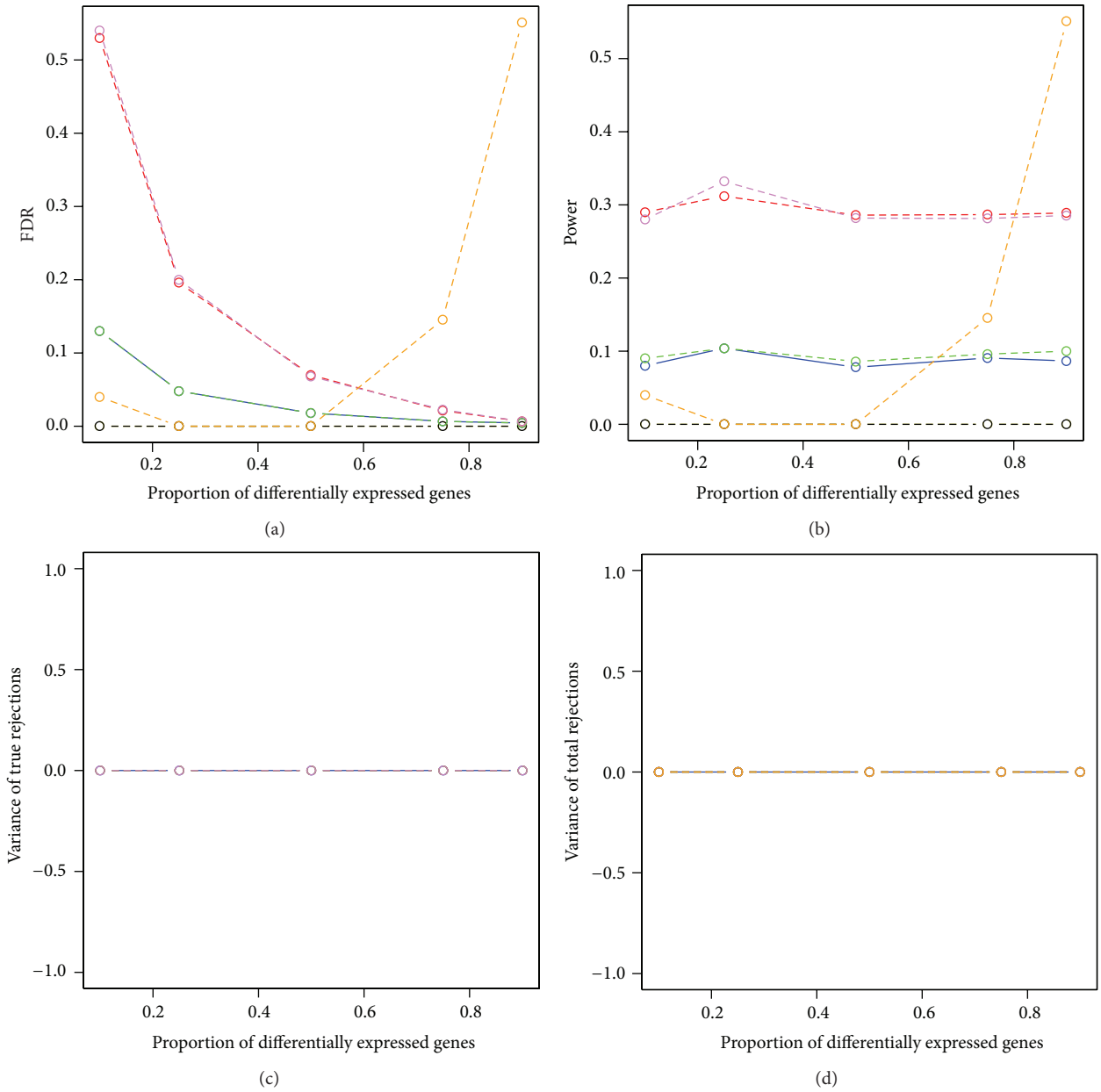


FIGURE 6: Power and stability properties of resampling-based multiple testing procedures for dependent test statistics with random correlations and the FDR is controlled at 5% ($n = 3$ in each group). Yellow dashed line: permutation single-step max$T$ procedure (mt.max$T$ function); black dashed line: permutation single-step min$P$ (mt.min$P$ function); solid blue line: bootstrap single-step max$T$ (MTP ss.max$T$ function); green dashed line: bootstrap step-down max$T$ (MTP sd.max$T$ function); red dashed line: bootstrap single-step min$P$ (MTP ss.min$P$ function); violet dashed line: bootstrap step-down min$P$ (MTP sd.min$P$ function); orange dashed line: SAM procedure (sam function).

the proportion of nontrue null hypotheses is less than 50% and sample size is 3; the bootstrap multiple testing procedures perform better than the permutation test procedures and the SAM procedure for small sample size data.

The zero power of the permutation test procedures is due to its limited number of permutated test statistics for data set with small sample sizes. For example, the complete number of enumeration is only 20 for both permutation single-step max$T$ procedure and permutation single-step min$P$ procedure when sample size is only 3 in each group. Thus, the smallest raw P value from the permutation procedures will be 0.05. After adjusting the raw P values to control FWER or FDR, all adjusted P values will be larger than 0.05, thus no hypotheses will be rejected. As such, the estimated FWER, FDR, and power will all be zero.

Our current investigation only focuses on normally distributed data. Further investigation is needed to extend the distribution in the simulations from multivariate normal to other distributions, such as lognormal and binomial distributions. To examine the power and stability properties

of those resampling-based multiple testing procedures, under nonnormal distributions, will be a focus for our future research.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank Professor Carlos S. Moreno for permission of using his microarray data. This work was supported in part by the National Institute of Minority Health and Health Disparities Awards U54MD007584 (J. Hedges, PI) and G12MD007601 (M. Berry, PI).

References

[1] D. A. Kulesh, D. R. Clive, D. S. Zarlenga, and J. J. Greene, "Identification of interferon-modulated proliferation-related cDNA sequences," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 84, no. 23, pp. 8453–8457, 1987.

[2] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, no. 5235, pp. 467–470, 1995.

[3] D. A. Lashkari, J. L. Derisi, J. H. Mccusker et al., "Yeast microarrays for genome wide parallel genetic and gene expression analysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 24, pp. 13057–13062, 1997.

[4] J. R. Pollack, C. M. Perou, A. A. Alizadeh et al., "Genome-wide analysis of DNA copy-number changes using cDNA microarrays," *Nature Genetics*, vol. 23, no. 1, pp. 41–46, 1999.

[5] M. J. Buck and J. D. Lieb, "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments," *Genomics*, vol. 83, no. 3, pp. 349–360, 2004.

[6] R. Mei, P. C. Galipeau, C. Prass et al., "Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays," *Genome Research*, vol. 10, no. 8, pp. 1126–1137, 2000.

[7] J. Y. Hehir-Kwa, M. Egmont-Petersen, I. M. Janssen, D. Smeets, A. G. van Kessel, and J. A. Veltman, "Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis," *DNA Research*, vol. 14, no. 1, pp. 1–11, 2007.

[8] Y. Hochberg and A. C. Tamhane, *Multiple Comparison Procedures*, John Wiley & Sons, New York, NY, USA, 1987.

[9] J. P. Shaffer, "Multiple hypothesis testing: a review," *Annual Review of Psychology*, vol. 46, pp. 561–584, 1995.

[10] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, vol. 57, no. 1, pp. 289–300, 1995.

[11] B. Efron, "Bootstrap methods: another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[12] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, CRC Press, New York, NY, USA, 1994.

[13] D. A. Freedman, "Bootstrapping regression models," *The Annals of Statistics*, vol. 9, no. 6, pp. 1218–1228, 1981.

[14] P. Hall, "On the bootstrap and confidence intervals," *The Annals of Statistics*, vol. 14, no. 4, pp. 1431–1452, 1986.

[15] K. S. Pollard and M. K. van der Laan, "Choice of a null distribution in resampling-based multiple testing," *Journal of Statistical Planning and Inference*, vol. 125, no. 1-2, pp. 85–100, 2004.

[16] P. H. Westfall and S. S. Young, *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*, John Wiley & Sons, New York, NY, USA, 1993.

[17] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[18] Y. Ge, S. Dudoit, and T. P. Speed, "Resampling-based multiple testing for microarray data analysis," *Test*, vol. 12, no. 1, pp. 1–77, 2003.

[19] D. Rubin, S. Dudoit, and M. van der Laan, "A method to increase the power of multiple testing procedures through sample splitting," *Statistical Applications in Genetics and Molecular Biology*, vol. 5, no. 1, article 19, 2006.

[20] A. Jemal, R. Siegel, E. Ward et al., "Cancer statistics, 2006," *CA: A Cancer Journal for Clinicians*, vol. 56, no. 2, pp. 106–130, 2006.

[21] C. S. Moreno, L. Matyunina, E. B. Dickerson et al., "Evidence that p53-mediated cell-cycle-arrest inhibits chemotherapeutic treatment of ovarian carcinomas," *PLoS ONE*, vol. 2, no. 5, article e441, 2007.

[22] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 854710, 10 pages
http://dx.doi.org/10.1155/2013/854710

Research Article

Transcriptional Protein-Protein Cooperativity in POU/HMG/DNA Complexes Revealed by Normal Mode Analysis

Debby D. Wang and Hong Yan

Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

Correspondence should be addressed to Debby D. Wang; debby.d.wang@gmail.com

Received 7 August 2013; Accepted 22 September 2013

Academic Editor: Ao Yuan

Biomolecular cooperativity is of great scientific interest due to its role in biological processes. Two transcription factors (TFs), Oct-4 and Sox-2, are crucial in transcriptional regulation of embryonic stem cells. In this paper, we analyze how Oct-1 (a similar POU factor) and Sox-2, interact cooperatively at their enhancer binding sites in collective motions. Normal mode analysis (NMA) is implemented to study the collective motions of two complexes with each involving these TFs and an enhancer. The special structure of Oct proteins is analyzed comprehensively, after which each Oct/Sox group is reassembled into two protein pairs. We subsequently propose a segmentation idea to extract the most correlated segments in each pair, using correlations of motion magnitude curves. The median analysis on these correlation values shows the intimacy of subunit POUS (Oct-1) and Sox-2. Using those larger-than-median correlation values, we conduct statistical studies and propose several protein-protein cooperative modes (*S* and *D*) coupled with their subtypes. Additional filters are applied and similar results are obtained. A supplementary study on the rotation angle curves reaches an agreement with these modes. Overall, these proposed cooperative modes provide useful information for us to understand the complicated interaction mechanism in the POU/HMG/DNA complexes.

1. Introduction

Embryonic stem cells (ES cells) possess the pluripotency of differentiating into all the three germ layers (endoderm, mesoderm, and ectoderm), which correspond to hundreds of cell types. These pluripotent stem cells are transcriptionally regulated by a number of transcription factors (TFs) [1]. A specific TF called Oct-4, belonging to the POU class of homeodomain proteins, is regarded as a necessity for maintaining the undifferentiated state of embryonic ES cells. Generally, Oct-4 interacts with other TFs as a group to affect the gene expression of mouse ES cells in early embryo development [2], and Oct-4 coupled with its cofactor Sox-2 (HMG-box domain) is at the center of this group. Botquin and Nishimoto have both proven the cooperative effects of Oct-4 and Sox-2 on the expression of several genes in mouse embryonic ES cells [3, 4]. Dailey and Basilico further bring forward the idea that the interaction within the POU/HMG group, especially for groups composed of Oct and Sox proteins, at DNA binding sites is a fundamental mechanism for transcriptional regulation in early embryo development [5].

At the early stage of transcription, TFs bind to specific regulatory DNA regions to cooperatively affect the transcription sites. Enhancers, which act as activators or stimulators for transcription [6], are a major type of regulatory DNA regions. Unlike promoters, enhancers may be located kilobases away from their target genes, but geometrically they are most probably close to the genes due to the supercoil structure of DNA molecules, and thus there can be direct contacts between the enhancer-TF complexes and the transcription sites. Studies on the enhancer-TF complexes are very important for understanding the complicated mechanism of transcriptional regulation.

On the other hand, molecular dynamics are involved in many biological processes [7, 8], such as reproduction, regulation of gene expression, and protein interaction. As an indispensable component of gene expression, transcription must undergo a series of dynamical changes of biomolecules. Therefore, studies on dynamics of the aforementioned enhancer-TF complexes would provide a deep insight into their properties and functions in the transcriptional regulation. Specifically, deciphering the roles of Oct and Sox in

the interaction mechanism of their enhancer-bounded complexes in the collective dynamics is of great scientific interest. Further, the cooperativity of the two proteins is a major research topic in these studies.

In our work, the dynamics of the POU/HMG group at its enhancer binding sites, referred to as POU/HMG/DNA complexes, are surveyed. Two POU/HMG/DNA complexes, which are DNA-binding portions of a POU factor Oct-1 and an HMG factor Sox-2 bound to an enhancer, are specifically studied from a structural and molecular dynamic view. Normal mode analysis (NMA) is implemented to study the collective or cooperative motions of these POU/HMG/DNA ternary complexes, after which the interaction of the POU and HMG factors at their DNA binding sites in these collective motions is explored. We propose a segmentation idea for the proteins to construct an equal-length-chain comparison and measure the correlation of each protein segment pair using the linear correlation. A statistical analysis on the significantly correlated pairs provides useful information on how these TFs have a synergistic control on enhancer DNAs in transcriptional regulation.

2. Materials and Method

2.1. Normal Mode Analysis (NMA)

2.1.1. Introduction. NMA is an efficient method to detect the most cooperative or collective motions (essential modes) of large harmonic oscillating systems. With the constraint that the studied conformations are in the vicinity of the systematic equilibrium, which exists in most harmonic oscillating systems [9], NMA is useful for studying large structural deformations or motions of these systems. The idea is to use harmonic potentials to approximate a multidimensional energy landscape around an energy minimum for a system and to detect the most easily accessible modes on this energy landscape. NMA is broadly used to analyze the structural dynamics of biomolecules.

Specifically, if we describe an N-site-system with a position vector q, in which the three-dimensional coordinates of each site $(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_N, y_N, z_N)$ are used, we can mathematically expand the potential energy V in a second-order Taylor series around the equilibrium conformation q^0 [9]. Finally, we obtain a quadratic approximation as follows:

$$V(q) = \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^0 \left(q_i - q_i^0\right)\left(q_j - q_j^0\right) = \frac{1}{2}\Delta q^T H \Delta q. \quad (1)$$

Here Δq stands for the systematic structural changes relative to q^0, and H is a $3N \times 3N$ Hessian matrix, whose elements have the following form:

$$H_{ij} = \frac{\partial^2 V}{\partial q_i \partial q_j}. \quad (2)$$

Subsequently, the kinetic energy is brought in to slightly modify the Hessian to a mass-weighted one. These Hessian matrices contain key structural information for our observed systems.

One broadly used construction method for the Hessian matrices is the elastic network models (ENMs) [9–12], which include the Gaussian network models (GNMs) [11] and anisotropic network models (ANMs) [12] as representatives. When ENM is applied, the equilibrium exploration can be skipped since the starting state is designed for this equilibrium. When constructing the ENM structure, the original system can be transformed into a network with nodes (CG-sites) and connecting springs, and a cutoff distance γ_c is used to define all the connecting springs γ_{ij} [9, 10]. Gaussian network model (GNM) selects representatives for substructures in the system, such as using C_α-atoms for amino acids [9, 11], to further lower the computational cost, leading to the potential form shown as (3) (R_i or R_j represents a CG-site):

$$V_{\mathrm{GNM}} = \frac{1}{2}\sum_{i,j}\gamma_{ij}\left(\Delta R_i - \Delta R_j\right)^2. \quad (3)$$

Similarly, ANM proposes the potential form in (4) and ignores some influences caused by the distance vectors:

$$V_{\mathrm{ANM}} = \frac{1}{2}\sum_{i,j}\gamma_{ij}\left(\left\|R_i^0 + \Delta R_i - R_j^0 - \Delta R_j\right\| - \left\|R_i^0 - R_j^0\right\|\right)^2. \quad (4)$$

Each eigenvalue of an above-constructed Hessian matrix denotes the associated systematic energy for the observed system, and its corresponding eigenvector represents the direction of a specific normal mode motion. Among the obtained $3N$ normal mode directions, the first six are trivial since they all correspond to zero eigenvalues, which means these structural changes have no effect on the systematic potential energy. For the remaining $3N - 6$ eigenvectors, we will select a small set that corresponds to small eigenvalues (essential modes) for analysis [9]. In previous research, the first 10~15 essential modes are chosen by many researchers for their work [13–15], and the first 10 are analyzed in our work.

2.1.2. Computational Platform. Several online tools are available for normal mode calculations. An online server called NOMAD-Ref at http://lorentz.immstr.pasteur.fr/nomad-ref.php [16] is utilized in our experiments. It is an ENM model-based method. The implementation of a rotation-translation block approach [16] and an ARPACK library for the sparse matrix data storage and decomposition [17] in the computations of Hessian matrices makes it possible to retain up to 100,000 atoms for each structure. In our work, when calculating the motions using NOMAD-Ref, all atoms in POU/HMG/DNA ternary complexes are used, while only motions of the POU and HMG proteins are analyzed since only protein-protein interactions in POU/HMG complexes at the DNA binding sites are of interest here.

2.2. Experimental Data and the Analysis on Their NMA Results

2.2.1. Experimental Data. Two POU/HMG/DNA ternary complexes, 1GT0 and 1O4X, are downloaded from the Protein Data Bank (PDB) [18] for analysis. Each structure is

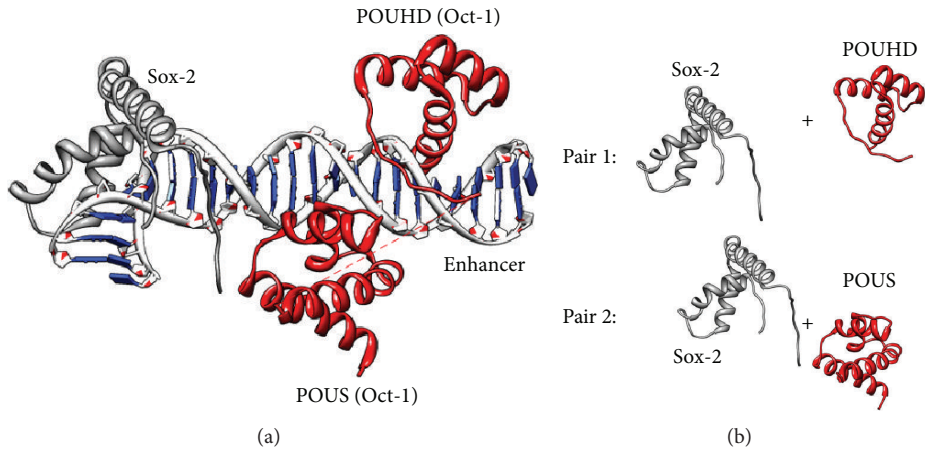


Figure 1: (a) The 3D structure of the POU/HMG/DNA ternary complex 1GT0. The gray protein represents an HMG factor Sox-2, and the red one is a POU factor Oct-1, which is composed of two subunits POUS and POUHD. (b) The two reassembled protein pairs, originated from (a), for our subsequent studies.

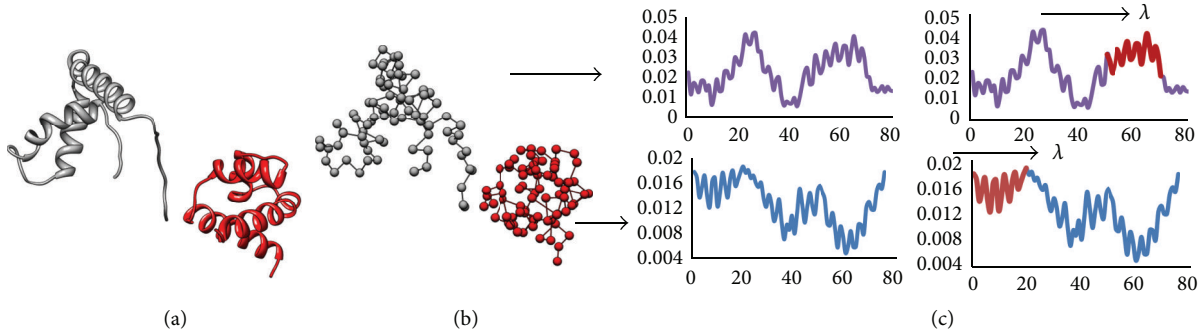


Figure 2: (a) The protein pair containing Sox-2 (gray) and POUS (red) in 1GT0. (b) The refined structure of the protein pair, where nodes represent residues. In each mode, a refined protein structure in the pair corresponds to a motion magnitude curve. (c) The searching process for the most cooperative segments of length λ in the protein pair in a specific mode.

composed of a POU factor Oct-1 (very similar to Oct-4), an HMG factor Sox-2 and an enhancer element. Figure 1(a) displays the 3D structure of complex 1GT0 and the diagram is produced using UCSF Chimera [19]. In 1GT0, the bounded DNA piece is a fibroblast growth factor 4 enhancer (FGF4) [20]; in 1O4X, the homeobox B1 (Hoxb1) enhancer is bounded by the two TFs [21].

Furthermore, each Oct protein contains two subunits (POUS and POUHD) that are connected by a flexible linker and control DNAs in a bipartite manner [21]. Based on the special structure of Oct proteins, we regard Oct-1 and Sox-2 in each complex as the two protein pairs for further investigation, namely, POUHD and Sox-2 as pair 1 and POUS and Sox-2 as pair 2, both of which are shown in Figure 1(b).

2.2.2. Analysis of Correlative Motions. After generating the motions of the two POU/HMG/DNA ternary complexes using NMA, we observe how the two protein pairs behave at the enhancer binding sites in these most collective or cooperative motions.

For each protein pair in each ternary complex, we analyze the first 10 obtained essential modes. In each mode, we firstly refine an observed pair at the residue level from a view of motion magnitude. This can be achieved by calculating the motion magnitudes for all the atoms in each protein and subsequently computing the motion magnitude of each residue in this protein by averaging the motion magnitudes of all component atoms (see (5)):

$$MR_i = \frac{1}{N}\sum_{j=1}^{N_i} MA_{ij} = \frac{1}{N}\sum_{j=1}^{N_i}\sqrt{\left(x_{ij}-x_{ij}^0\right)^2+\left(y_{ij}-y_{ij}^0\right)^2+\left(z_{ij}-z_{ij}^0\right)^2}. \quad (5)$$

Here atoms $j = 1 \sim N_i$ comprise the residue i; $(x_{ij}^0, y_{ij}^0, x_{ij}^0)$ and (x_{ij}, y_{ij}, x_{ij}) represent the positions of atom j in its equilibrium position and in a specific mode, respectively. Therefore, for each mode, we will obtain a motion magnitude curve for each protein in an observed pair, and each curve point corresponds to a residue along the protein sequence (Figures 2(a) and 2(b)).

Next, in each protein pair we observe the potential protein-protein cooperativity in these motions based on the correlations of motion magnitude functions. An effective method to measure the dependence between two quantities is the Pearson product-moment correlation coefficient [22–24] also is usually called the correlation coefficient. This coefficient is calculated based on the expected values (μ) and standard deviations (σ) of the two variables (**X** and **Y**), as shown in (6):

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X \sigma_Y}. \tag{6}$$

We adopt this correlation coefficient in our studies. However, since each protein has a different length, we investigate the most cooperative/correlated segments among each protein pair in each mode. We introduce a segment length parameter λ here. For an observed pair of proteins that have different lengths of x and y, with a specific $\lambda\,(\lambda \le x < y)$ defined in a mode, we shift one motion magnitude function along the other to find the λ-length-segments which share the largest absolute correlation value (Figure 2(c)). We could further describe the process as follows:

$$\max_{ij} C(i,j) = \left|\operatorname{corr}\left(\mathrm{Seg}_i, \mathrm{Seg}_j\right)\right|$$
$$\text{So that} \begin{cases} \mathrm{Seg}_i \in F_1 \\ \mathrm{Seg}_j \in F_2 \\ \left\|\mathrm{Seg}_i\right\| = \left\|\mathrm{Seg}_j\right\| = \lambda. \end{cases} \tag{7}$$

Here F_1 and F_2 represent motion magnitude functions of the two proteins in an observed pair, in a specific essential mode; Seg_i and Seg_j denote λ-length-segments of F_1 and F_2, respectively.

In each modes with a list of λ values defined for each protein pair, we obtain a series of most cooperative segment pairs having correlation values c_{mn}, where m denotes different λ values and n (1~10) represents different modes. Here we replace λ by $p = \lambda/x$ for easier illustration and x is the shorter length in the observed pair. Since larger absolute value of correlation demonstrates more correlated segments (positively or negatively), we investigate how $|c_{mn}|$ distribute for the two protein pairs in each complex. For each p in an observed pair, the median value (8) is extracted and explored. Furthermore, the performances (based on $\tilde{c}_m$) of the two pairs in each complex are compared:

$$\tilde{c}_m = \mathrm{MED}_{n=1}^{10}\left\{\left|c_{mn}\right|\right\} = \mathrm{MED}\left\{\left|c_{m1}\right|, \left|c_{m2}\right|, \ldots, \left|c_{m10}\right|\right\}, \quad m = 1, \ldots, 6. \tag{8}$$

Now, we use medians in (8) as a filter and investigate how those $|c_{mn}|$ larger than $\tilde{c}_m$ (supposed to be significant) distribute. For each protein in an observed pair, we can obtain a logic matrix L that reflects this process:

$$L = [l_{mn}] = \left([|c_{mn}|] > [\tilde{c}_m]?\right) = \left([C_m] > [\tilde{c}_m]?\right) = \left(\begin{pmatrix} C_1 \\ C_2 \\ \vdots \\ C_6 \end{pmatrix} > \begin{pmatrix} \tilde{c}_1 \\ \tilde{c}_2 \\ \vdots \\ \tilde{c}_6 \end{pmatrix}?\right)$$
$$\text{So that} \begin{cases} C_m = \left(\left|c_{m1}\right|, \left|c_{m2}\right|, \ldots, \left|c_{m10}\right|\right) \\ m = 1, \ldots, 6 \\ n = 1, \ldots, 10. \end{cases} \tag{9}$$

Here $[x]$ gives the matrix that is composed of x.

We subsequently examine the relationship between the two protein pairs in each complex based on these logic matrices. The idea is to explore that in a single essential mode whether only one significantly correlated segment pair (either in protein pair 1 or pair 2) is involved or both pairs are involved. To balance the segment lengths (p) used by the two pairs, we take into consideration all the length pairs (p^1, p^2) between the two pairs, as presented in Figure 3. Here we use superscripts to distinguish pairs 1 and 2.

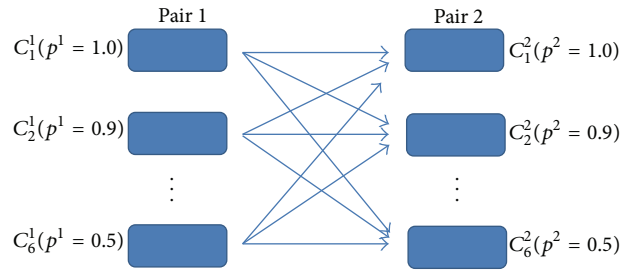


Figure 3: This diagram shows all the possibilities of the length pair (p^1, p^2). Specifically, for an observed complex, the significantly correlated segment pair (with a length parameter of p^1) in protein pair 1 and that (with a length parameter of p^2) in protein pair 2 are comparably investigated, with p^1 and p^2 taking all possible values (m = 1 ~6) as previously stated.

To fulfill the aforementioned operation, we conduct several iterations for all the p^i values and combine the results of these iterations. We now take rows in L^1 (denoting a specific p^1 value) and show how the whole procedure is accomplished. In each iteration, we firstly expand the involved row (identified with a subscript m) into a matrix in (10) and then carry out statistics on the cases where three situations occur: (a) S_1 (index s_1)—only the significantly correlated segment pair in pair 1 is detected in a single essential mode with a length pair (p^1, p^2), (b) S_1 (index s_2)—only the significantly correlated segment pair in pair 2 is detected, and (c) D (index d)—both pairs 1 and 2 are detected. The statistical analysis is based on logic operations, as shown in (11), which combines all the iterations to derive the final indexes for the two pairs:

$$L^{1,m} = \frac{\left(\underbrace{L_m^1, L_m^1, \ldots, L_m^1}_{6}\right)^T}{}, \tag{10}$$

$$s_1 = \sum_{m=1}^{6} \mathrm{sum}\left(L^{1,m} \cdot \times \left(\neg L^2\right)\right),$$

$$s_2 = \sum_{m=1}^{6} \mathrm{sum}\left(\left(\neg L^{1,m}\right) \cdot \times L^2\right),$$

$$d = \sum_{m=1}^{6} \mathrm{sum}\left(L^{1,m} \cdot \times L^2\right). \tag{11}$$

Here "$\cdot\times$" means a batch of multiplications of the corresponding elements in two matrices, and sum(X) counts the number of ones (a logic "*true*" value) in a logic matrix X. Indexes s_1, s_2, and d separately show three cooperative modes (corresponding to the aforementioned three cases) between the two protein pairs in a POU/HMG/DNA complex. We exhibit some representative cooperative modes in Section 3, where we also list the above-mentioned indexes for the two complexes. Furthermore, to take the signs of the correlations c_{mn} into consideration, we introduce another logic matrix Z that describes the signs of c_{mn}, as stated in (12). Through combining logic operations of L^i and Z^i (13), we can divide the situations (S_1, S_2, and D) into subtypes (positive and negative), and all these subtypes are analyzed in Section 3:

$$Z = [z_{mn}] = ([c_{mn}] > [0]?) \quad \text{So that} \begin{cases} m = 1, \ldots, 6 \\ n = 1, \ldots, 10, \end{cases} \tag{12}$$

$$Z^{1,m} = \underbrace{\left(Z_m^1, Z_m^1, \ldots, Z_m^1\right)^T}_{6},$$

$$s_{1,\mathrm{positive}} = \sum_{m=1}^{6} \mathrm{sum}\left(L^{1,m} \cdot \times \left(\neg L^2\right) \cdot \times Z^{1,m}\right)$$

$$s_{1,\mathrm{negative}} = \sum_{m=1}^{6} \mathrm{sum}\left(L^{1,m} \cdot \times \left(\neg L^2\right) \cdot \times \left(\neg Z^{1,m}\right)\right),$$

$$s_{2,\mathrm{positive}} = \sum_{m=1}^{6} \mathrm{sum}\left(\left(\neg L^{1,m}\right) \cdot \times L^2 \cdot \times Z^2\right),$$

$$s_{2,\mathrm{negative}} = \sum_{m=1}^{6} \mathrm{sum}\left(\left(\neg L^{1,m}\right) \cdot \times L^2 \cdot \times \left(\neg Z^2\right)\right),$$

$$d_{\mathrm{positive}} = \sum_{m=1}^{6} \mathrm{sum}\left(L^{1,m} \cdot \times L^2 \cdot \times \left[Z^{1,m} \cdot \times Z^2 + \left(\neg Z^{1,m}\right)\right] \cdot \times \left(\neg Z^2\right)\right),$$

$$d_{\mathrm{negative}} = \sum_{m=1}^{6} \mathrm{sum}\left(L^{1,m} \cdot \times L^2 \cdot \times \left[Z^{1,m} \cdot \times \left(\neg Z^2\right) + \left(\neg Z^{1,m}\right) \cdot \times Z^2\right]\right). \tag{13}$$

To compare the scenarios where different filters are applied, we, respectively, apply the first tertile, the first quartile, and the mean value as filters to investigate the corresponding results. The mean filter can be described as (14), and the quantile filter as (15), where Pr represents probability. Specifically, the tertile and quartile filters correspond to situations where $p = 1/3$ and $p = 1/4$, respectively. A series of operations are then carried out based on these filters, to reveal how the observed complexes behave in these situations:

$$\bar{c}_m = \frac{\sum_{n=1}^{10} \{|c_{mn}|\}}{10}, \quad m = 1, \ldots, 6, \tag{14}$$

$$\tilde{c}_m(p) = \inf\{|c_{mn}| \mid F(|c_{mn}|) \geq p\},$$

$$F(|c_{mn}|) = \Pr(C_m \leq |c_{mn}|), \quad m = 1, \ldots, 6. \tag{15}$$

Finally, to gain a deep insight into the motions of these two complexes, we have also observed the rotation angles of the corresponding protein chains. In the above discussion, we regard residues as basic units in protein sequences, and here we consider the links between each consecutive two residues (Figure 2(b)). The angles between each pair of corresponding links in the original structure and in deformed structures (modes) are studied. We obtain a rotation angle function for each protein of a protein pair in each essential mode. Afterwards, we conduct a similar analysis as aforementioned on these rotation angle functions as a supplementary study. Principal component analysis (PCA) is implemented to reduce the effect of noisy rotation angles. We also investigate the suitability of the Fourier transform for data analysis.

3. Results and Discussion

3.1. Motion Magnitude Functions. For each protein in an observed pair of a ternary complex, we calculate the motion magnitude functions (5) for the first 10 essential modes. Figure 4 shows the motion magnitude curves for the two observed protein pairs in 1GT0 for the first essential mode.

After defining a list of p values, we calculate the most cooperative/correlated segment pairs among each protein pair in a complex for the 10 essential modes, using the mechanism discussed in Section 2.2.2. Since small p values correspond to shorter segment matching, whose results may be trivial due to the high correlation possibilities, we use a set of p values starting at 0.5 to 1.0 at a step of 0.1. Table 1 shows the results of correlations c_{mn} for the most cooperative segment pairs among protein pair 1 of 1GT0.

The larger the absolute value of correlation is, the more the two compared segments correlate with each other, either positively or negatively. Now we examine how the absolute correlation values $|c_{mn}|$ distribute, for the two protein pairs in each complex. The values are presented in Figure 5, where parts (a) and (b), respectively, show the values for the two pairs in 1GT0, and parts (d) and (e) show those for 1O4X. We can see from these diagrams that $|c_{mn}|$ becomes larger when p gets smaller, and this can also be detected from the median value $\tilde{c}_m$ shown with a pink circle in each box (denoting a specific p). To give a comparison between the performances of the two pairs in each complex, we extract the above-mentioned median values $\tilde{c}_m$ for each pair and present them in parts (c) (1GT0) and (f) (1O4X). In diagrams (c) and (f), especially (f), pair 2 presents a higher $\tilde{c}_m$ than pair 1, which

Table 1: Motion correlations between POUHD and Sox-2 in protein pair 1 of 1GT0.

p	Mode									
	Mode 7	Mode 8	Mode 9	Mode 10	Mode 11	Mode 12	Mode 13	Mode 14	Mode 15	Mode 16
1	−0.696	0.606	−0.477	0.324	−0.265	0.383	0.202	−0.326	−0.382	−0.520
0.9	−0.738	0.711	0.697	0.340	−0.419	0.772	0.454	−0.429	−0.420	0.739
0.8	−0.853	0.797	0.819	0.342	0.651	0.838	−0.607	−0.463	0.569	0.730
0.7	−0.859	0.836	0.820	−0.445	0.639	0.820	−0.621	−0.562	−0.716	0.769
0.6	−0.856	0.851	0.850	0.537	0.814	0.849	−0.699	−0.684	0.762	0.806
0.5	−0.865	−0.862	0.856	−0.757	0.858	−0.901	−0.733	−0.761	0.806	0.819

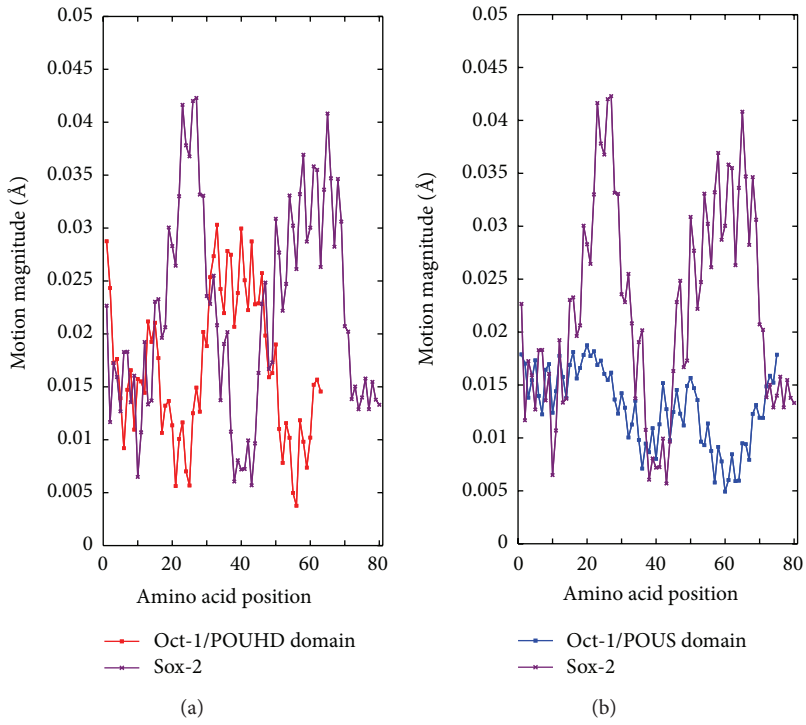


Figure 4: (a) The motion magnitude curves in mode 7 for proteins POUHD (red) and Sox-2 (purple) in pair 1 of 1GT0. (b) The motion magnitude curves in mode 7 for proteins POUS (blue) and Sox-2 (purple) in pair 2 of 1GT0.

to some extent implies that pair 2 may behave as a leading role in the Oct/Sox interactions.

Next, we use the above-mentioned medians as a filter and investigate how those $|c_{mn}|$ larger than $\tilde{c}_m$ (supposed to be significant) distribute. For each protein in an observed pair, we calculate its logic matrices L and Z (Section 2.2.2), which correspond to (9) and (12), respectively. We subsequently study the logic matrices of the two protein pairs (L^1 and Z^1, L^2 and Z^2) in each complex, after which we propose several cooperative modes between the two pairs and conduct statistical analysis according to (10) and (11). In detail, these modes include (a) mode S_1 (index s_1)—only the significantly correlated segment pair in pair 1 is detected in a single essential mode with a length pair (p^1, p^2), (b) mode S_1 (index s_2)—only the significantly correlated segment pair in pair 2 is detected, and (c) mode D (index d)—both pairs 1 and 2 are detected. To visually show the cooperative modes S and D, we select parts of the results of 1GT0 for $p^1 = p^2 = 0.5$ as a display in Figure 6, in which S_1, S_2, and D modes are, respectively, presented with the significantly correlated segment pairs colored.

Mode S denotes that only one protein pair, either pair 1 (S_1) or pair 2 (S_2), is significantly involved in a specific collective motion. This indicates that only one subunit, either POUHD or POUS, is significantly involved in the cooperativity with Sox-2 in an essential mode. Mode D implies that both subunits are involved in the interactions with Sox-2. Detailed statistical results are reported in Table 2. In this table, cooperative mode D occurs more frequently than modes S_1 and S_2 in the two complexes, which have the tuples of (82, 82, 98) and (85, 85, 95) for the indexes (s_1, s_2, d), respectively. This implies that, compared with mode S_1 or S_2, both subunits of Oct-1 frequently participate in the interactions with Sox-2 at the same time, as mode D.

Furthermore, we divide the modes S and D into subtypes, positive subtype and negative subtype, and their statistics are evaluated using (13) and listed in Table 2. In modes S_1 and S_2, the positive subtype ($s_{1,\text{positive}}$ and $s_{2,\text{positive}}$) shows

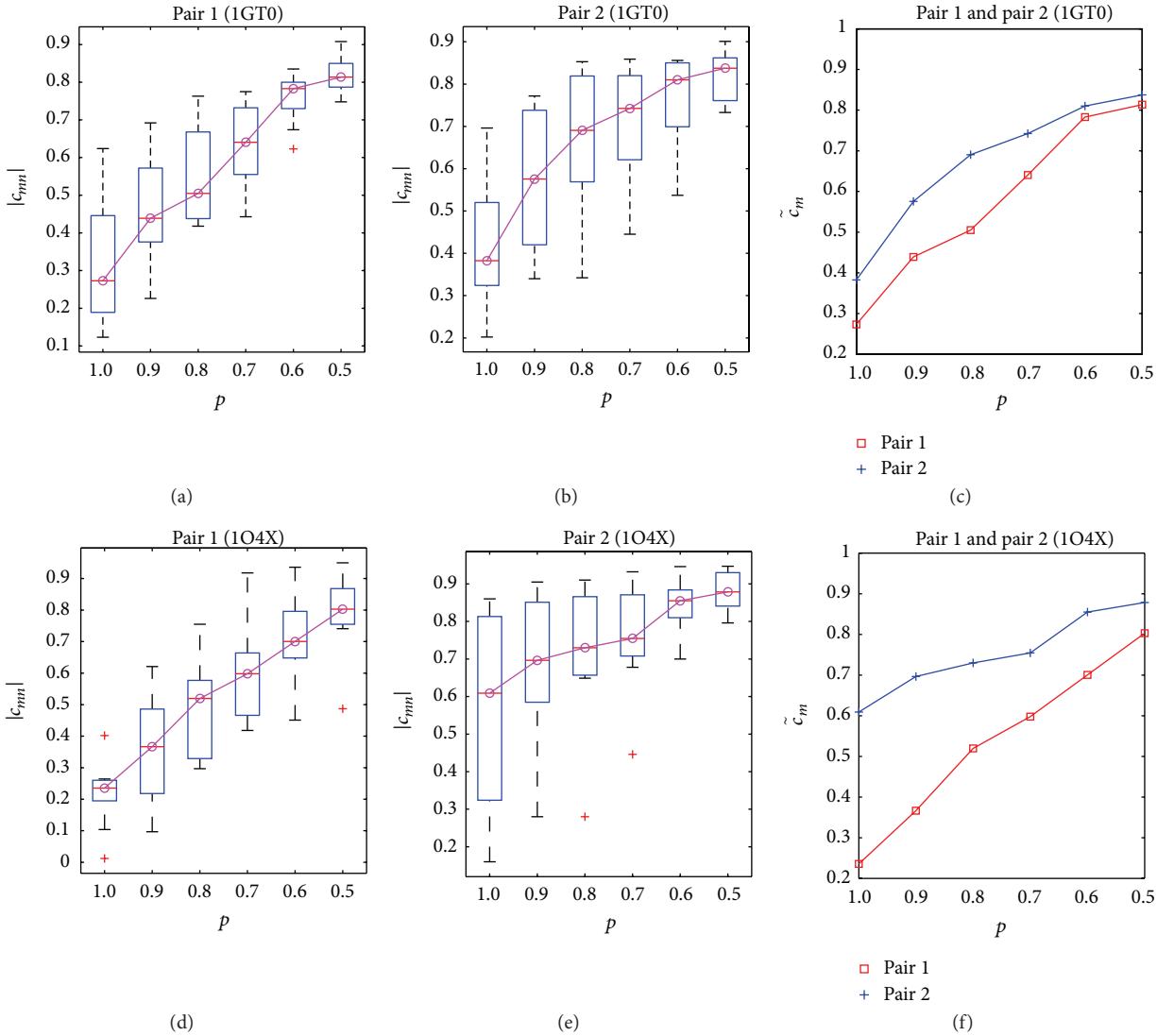


FIGURE 5: (a) and (b) show the distributions of absolute correlation values $|c_{mn}|$ for the two protein pairs in complex 1GT0, respectively. In (c), the median absolute correlation value $\tilde{c}_m$ for each p is extracted for the two pairs from (a) and (b). Similarly, (d), (e), and (f) are the plots for complex 1O4X.

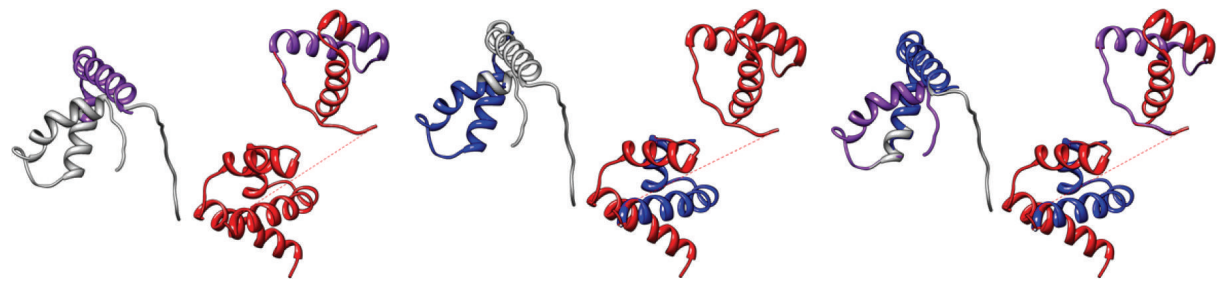


FIGURE 6: Parts of the results of cooperative modes S and D in complex 1GT0, with TFs Oct-1 and Sox-2 colored red and gray, for $p^1 = p^2 = 0.5$ in the first 10 essential modes. (a) displays mode S_1 in normal mode 13, with the significantly correlated segments in protein pair 1 colored purple; (b) displays mode S_2 in normal mode 12 and the correlated segments are colored blue in protein pair 2; (c) presents mode D in normal mode 16 with the correlated segment pairs colored purple and blue, respectively, in both protein pairs.

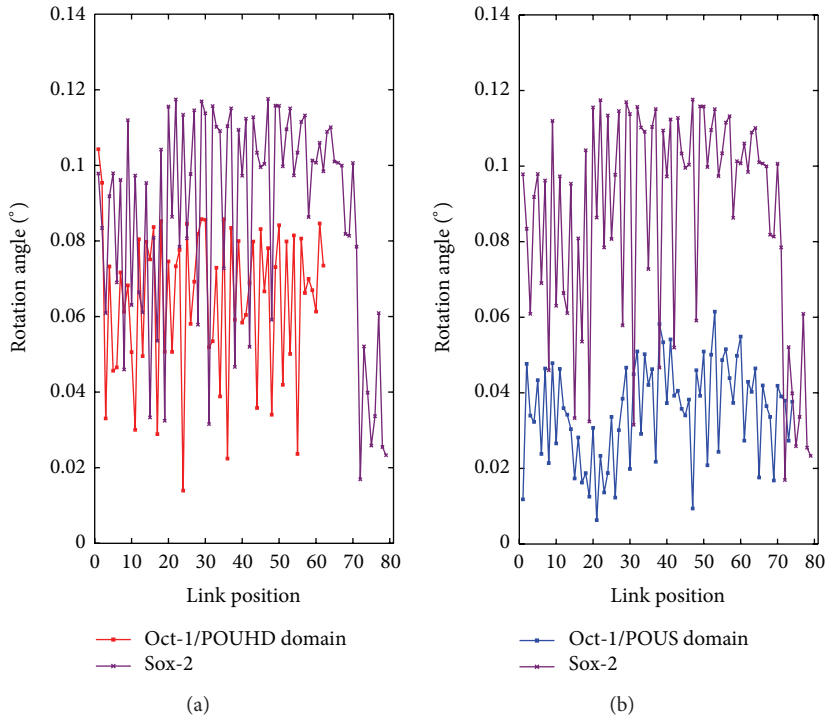


Figure 7: (a) displays the rotation angle curves for proteins POUHD (red) and Sox-2 (purple) in pair 1 of 1GT0 in mode 7; (b) shows the rotation angle curves for proteins POUS (blue) and Sox-2 (purple) in pair 2 of 1GT0 in mode 7.

a positive sign of c_{mn} for the significantly correlated segment pair in protein pair 1 or 2, and the negative one ($s_{1,\text{negative}}$ and $s_{2,\text{negative}}$) indicates a negative sign. In mode D, the positive subtype (d_{positive}) denotes a scenario where both significantly correlated segment pairs in the two protein pairs share the same sign of c_{mn} (+/+ or −/−), and the negative one (d_{negative}) represents two different signs (+/− or −/+). From Table 2 we notice that, for mode S_2 in both complexes 1GT0 and 1O4X, the positive subtype has a lead; for modes S_1 and D, the negative subtype is in the lead for 1GT0 while the positive one is in a dominant position for 1O4X.

We have also applied the first tertile, the first quartile and the mean value as filters and similarly conducted the statistical analysis as illustrated above. Tables 3, 4, and 5 present the results for these three scenarios, respectively. As shown in these tables, the gap between the occurrence frequencies of mode D and mode S_1 (or S_2) becomes larger, and the dominant occurrences of mode D are demonstrated. Besides, for modes S_1 and D, complexes 1GT0 and 1O4X have the opposite subtype distributions, while for mode S_2, they present a similar distribution. Overall, these additional results are consistent with the previous one (the median filter).

3.2. Rotation Angle Functions. Subsequently, we calculate the rotation angle functions for each protein in each complex in the first 10 essential normal modes (described in Section 2.2.2). Figure 7 shows the rotation angle curves of proteins in the two protein pairs of 1GT0 in the first essential mode.

Table 2: Statistics on the occurrences of the cooperative modes S and D, and their subtypes, in the 10 essential modes for all the pairs (p^1, p^2), using the median filter for motion correlations.

1GT0								
	s_1			s_2			d	
82	$s_{1,\text{positive}}$	$s_{1,\text{negative}}$	82	$s_{2,\text{positive}}$	$s_{2,\text{negative}}$	98	d_{positive}	d_{negative}
	36	46		49	33		35	63
1O4X								
	s_1			s_2			d	
85	$s_{1,\text{positive}}$	$s_{1,\text{negative}}$	85	$s_{2,\text{positive}}$	$s_{2,\text{negative}}$	95	d_{positive}	d_{negative}
	62	23		75	10		68	27

Table 3: Statistics on the occurrences of the cooperative modes S and D, and their subtypes, in the 10 essential modes for all the pairs (p^1, p^2), using the first tertile as a filter for motion correlations.

1GT0								
	s_1			s_2			d	
71	$s_{1,\text{positive}}$	$s_{1,\text{negative}}$	71	$s_{2,\text{positive}}$	$s_{2,\text{negative}}$	145	d_{positive}	d_{negative}
	28	43		47	24		70	75
1O4X								
	s_1			s_2			d	
77	$s_{1,\text{positive}}$	$s_{1,\text{negative}}$	71	$s_{2,\text{positive}}$	$s_{2,\text{negative}}$	139	d_{positive}	d_{negative}
	57	20		62	9		98	41

Since the rotation angle functions contain a lot of noise, we apply the principal component analysis (PCA) to the 10

Table 4: Statistics on the occurrences of the cooperative modes S and D, and their subtypes, in the 10 essential modes for all the pairs (p^1, p^2), using the first quartile as a filter for motion correlations.

1GT0								
	s_1			s_2			d	
62	$s_{1,\text{positive}}$	$s_{1,\text{negative}}$	62	$s_{2,\text{positive}}$	$s_{2,\text{negative}}$	190	d_{positive}	d_{negative}
	26	36		33	29		85	105
1O4X								
	s_1			s_2			d	
70	$s_{1,\text{positive}}$	$s_{1,\text{negative}}$	64	$s_{2,\text{positive}}$	$s_{2,\text{negative}}$	182	d_{positive}	d_{negative}
	47	23		52	12		119	63

Table 5: Statistics on the occurrences of the cooperative modes S and D, and their subtypes, in the 10 essential modes for all the pairs (p^1, p^2), using the mean value as a filter for motion correlations.

1GT0								
	s_1			s_2			d	
66	$s_{1,\text{positive}}$	$s_{1,\text{negative}}$	78	$s_{2,\text{positive}}$	$s_{2,\text{negative}}$	108	d_{positive}	d_{negative}
	33	33		43	35		44	64
1O4X								
	s_1			s_2			d	
73	$s_{1,\text{positive}}$	$s_{1,\text{negative}}$	85	$s_{2,\text{positive}}$	$s_{2,\text{negative}}$	107	d_{positive}	d_{negative}
	52	21		72	13		77	30

Table 6: Correlations between PC curves of rotation angle functions for the two protein pairs in 1GT0 and 1O4X.

p	1GT0		1O4X	
	Pair 1	Pair 2	Pair 1	Pair 2
1.0	0.682	−0.278	−0.875	−0.310
0.9	0.836	−0.339	−0.884	−0.764
0.8	0.844	−0.354	−0.884	−0.821
0.7	0.854	−0.328	−0.884	−0.829
0.6	0.863	−0.498	−0.890	−0.856
0.5	0.876	0.703	−0.915	−0.859

Table 7: Statistics on the occurrences of the cooperative modes S and D, and their subtypes, in the 10 essential modes for all the pairs (p^1, p^2), using the first quartile as a filter for the correlations of rotation angle functions.

1GT0								
	s_1			s_2			d	
108	$s_{1,\text{positive}}$	$s_{1,\text{negative}}$	108	$s_{2,\text{positive}}$	$s_{2,\text{negative}}$	144	d_{positive}	d_{negative}
	108	0		108	0		144	0
1O4X								
	s_1			s_2			d	
48	$s_{1,\text{positive}}$	$s_{1,\text{negative}}$	48	$s_{2,\text{positive}}$	$s_{2,\text{negative}}$	204	d_{positive}	d_{negative}
	48	0		48	0		204	0

rotation angle curves of each protein in the two complexes to obtain the first principal component (PC), leading the rotation angle curves ($n = 1$~10) of each protein to a single condensed PC curve. We similarly carry out the correlation analysis of the PC curves in each pair. Table 6 shows the statistical results for the two complexes. Intuitively, 1GTO presents the distinct cooperative mode of S_1, where pair 1 shows more significantly correlated segment pairs (with a positive subtype), while mode D is the dominant one in 1O4X, where many significantly correlated segment pairs occur in both pairs (with a positive subtype).

Now we apply the Fourier transform to analyze these noisy rotation angle values. Simply, the magnitudes of the transformed signals are regarded as our new data. The segmentation and correlation calculation are implemented, after which the statistical analysis is carried out. As an example, we use the first quartile as a filter for the correlations of rotation angle functions. The results are listed in Table 7, where we can see that the negative subtype of each cooperative mode is concealed after the transform. This implies that the Fourier transform may not be a suitable tool for handling these rotation angle values. More efficient strategies should be explored in the future to deal with these data.

4. Conclusions

In this paper, we performed NMA to study the collective motions of two TFs, Oct-1 and Sox-2, at their enhancer binding sites, aiming to gain an insight into the cooperative manner of these two TFs through the dynamics of their enhancer-bounded complexes. Based on the special structure of Oct proteins, we treated an Oct/Sox group as two protein pairs and comparably investigated how these two pairs behave in the collective motions. A segmentation idea was introduced to explore the most correlated segments in each protein pair, according to the correlations of motion magnitude curves (or their segments). A median analysis on these correlations was conducted, which shows the leading role of subunit POUS (pair 2). Furthermore, based on statistics of the correlated segment pairs having a correlation value above the corresponding median, we proposed several motion cooperative modes (S_1, S_2, and D) and their subtypes (positive or negative). The first tertile, the first quartile, and the mean value provide consistent results. Moreover, the supplementary study on the rotation angle functions presents a consensus about these modes. These proposed modes provide a clue that when binding to different regulatory DNA regions or involved in different collective motions, Oct-1 has a synergistic relationship with Sox-2 either with one of the components, POUS or POUHD, or both of them, POUS and POUHD at the same time.

Cooperativity, in protein-DNA [25] and protein-protein [26] interactions, is an important feature in biomolecular interactions. In our work, we carried out a series of studies on the cooperative manner of Oct and Sox at their enhancer binding sites, which are important elements in the transcriptional regulation of embryonic stem cells. This work reveals how the two proteins work together physically and structurally at two specific DNA biding sites. The method developed here can be useful for the analysis of molecular interactions in other protein-protein and protein-DNA complexes.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work is supported by the City University of Hong Kong (Project 7002843).

References

[1] L. A. Boyer, I. L. Tong, M. F. Cole et al., "Core transcriptional regulatory circuitry in human embryonic stem cells," *Cell*, vol. 122, no. 6, pp. 947–956, 2005.

[2] M. Pesce and H. R. Schöler, "Oct-4: gatekeeper in the beginnings of mammalian development," *Stem Cells*, vol. 19, no. 4, pp. 271–278, 2001.

[3] V. Botquin, H. Hess, G. Fuhrmann et al., "New POU dimer configuration mediates antagonistic control of an osteopontin preimplantation enhancer by Oct-4 and Sox-2," *Genes and Development*, vol. 12, no. 13, pp. 2073–2090, 1998.

[4] M. Nishimoto, A. Fukushima, A. Okuda, and M. Muramatsu, "The gene for the embryonic stem cell coactivator UTF1 carries a regulatory element which selectively interacts with a complex composed of Oct-3/4 and Sox-2," *Molecular and Cellular Biology*, vol. 19, no. 8, pp. 5453–5465, 1999.

[5] L. Dailey and C. Basilico, "Coevolution of HMG domains and homeodomains and the generation of transcriptional regulation by Sox/POU complexes," *Journal of Cellular Physiology*, vol. 186, pp. 315–328, 2001.

[6] M. J. Borok, D. A. Tran, M. C. W. Ho, and R. A. Drewell, "Dissecting the regulatory switches of development: lessons from enhancer evolution in Drosophila," *Development*, vol. 137, no. 1, pp. 5–13, 2010.

[7] A. Warshel, "Molecular dynamics simulations of biological reactions," *Accounts of Chemical Research*, vol. 35, pp. 385–395, 2002.

[8] G. G. Dodson, D. P. Lane, and C. S. Verma, "Molecular simulations of protein dynamics: new windows on mechanisms in biology," *EMBO Reports*, vol. 9, no. 2, pp. 144–150, 2008.

[9] I. Bahar, T. R. Lezon, A. Bakan, and I. H. Shrivastava, "Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins," *Chemical Reviews*, vol. 110, no. 3, pp. 1463–1497, 2010.

[10] I. Bahar and A. J. Rader, "Coarse-grained normal mode analysis in structural biology," *Current Opinion in Structural Biology*, vol. 15, no. 5, pp. 586–592, 2005.

[11] B. Erman, "The Gaussian network model: precise prediction of residue fluctuations and application to binding problems," *Biophysical Journal*, vol. 91, no. 10, pp. 3589–3599, 2006.

[12] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model," *Biophysical Journal*, vol. 80, no. 1, pp. 505–515, 2001.

[13] S. Hayward, A. Kitao, and H. J. Berendsen, "Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme," *Proteins*, vol. 27, pp. 425–437, 1997.

[14] S. Hayward, A. Kitao, and N. Go, "Harmonic and anharmonic aspects in the dynamics of BPTI: a normal mode analysis and principal component analysis," *Protein Science*, vol. 3, no. 6, pp. 936–943, 1994.

[15] D. G. Teotico, M. L. Frazier, F. Ding, N. V. Dokholyan, B. R. S. Temple, and M. R. Redinbo, "Active nuclear receptors exhibit highly correlated AF-2 domain motions," *PLoS Computational Biology*, vol. 4, no. 7, Article ID e1000111, 2008.

[16] E. Lindahl, C. Azuara, P. Koehl, and M. Delarue, "NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis," *Nucleic Acids Research*, vol. 34, pp. W52–W56, 2006.

[17] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, Philadelphia, Pa, USA, 1998.

[18] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[19] E. F. Pettersen, T. D. Goddard, C. C. Huang et al., "UCSF Chimera—a visualization system for exploratory research and analysis," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1605–1612, 2004.

[20] A. Reményi, K. Lins, L. J. Nissen, R. Reinbold, H. R. Schöler, and M. Wilmanns, "Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers," *Genes and Development*, vol. 17, no. 16, pp. 2048–2059, 2003.

[21] D. C. Williams Jr., M. Cai, and G. M. Clore, "Molecular basis for synergistic transcriptional activation by Oct1 and Sox2 revealed from the solution structure of the 42-kDa Oct1.Sox2.Hoxb1-DNA ternary transcription factor complex," *Journal of Biological Chemistry*, vol. 279, no. 2, pp. 1449–1457, 2004.

[22] B. G. Häne, K. Jäger, and H. G. Drexler, "The Pearson product-moment correlation coefficient is better suited for identification of DNA fingerprint profiles than band matching algorithms," *Electrophoresis*, vol. 14, no. 10, pp. 967–972, 1993.

[23] T. R. Derrick, B. T. Bates, and J. S. Dufek, "Evaluation of time-series data sets using the Pearson product-moment correlation coefficient," *Medicine and Science in Sports and Exercise*, vol. 26, no. 7, pp. 919–928, 1994.

[24] A. K. Gayen, "The frequency distribution of the product-moment correlation coefficient in random samples of any size drawn from non-normal universes," *Biometrika*, vol. 38, no. 1-2, pp. 219–247, 1951.

[25] D. D. Wang and H. Yan, "The relationship between periodic dinucleotides and the nucleosomal DNA deformation revealed by normal mode analysis," *Physical Biology*, vol. 8, no. 6, Article ID 066004, 2011.

[26] Y.-N. Zhang, X.-Y. Pan, Y. Huang, and H.-B. Shen, "Adaptive compressive learning for prediction of protein-protein interactions from primary sequence," *Journal of Theoretical Biology*, vol. 283, no. 1, pp. 44–52, 2011.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 436493, 8 pages
http://dx.doi.org/10.1155/2013/436493

Research Article

Variable Selection in ROC Regression

Binhuan Wang

New York University School of Medicine, New York, NY 10016, USA

Correspondence should be addressed to Binhuan Wang; binhuan.wang@nyumc.org

Received 6 August 2013; Accepted 18 September 2013

Academic Editor: Gengsheng Qin

Regression models are introduced into the *receiver operating characteristic* (ROC) analysis to accommodate effects of covariates, such as genes. If many covariates are available, the variable selection issue arises. The traditional induced methodology separately models outcomes of diseased and nondiseased groups; thus, separate application of variable selections to two models will bring barriers in interpretation, due to differences in selected models. Furthermore, in the ROC regression, the accuracy of area under the curve (AUC) should be the focus instead of aiming at the consistency of model selection or the good prediction performance. In this paper, we obtain one single objective function with the group SCAD to select grouped variables, which adapts to popular criteria of model selection, and propose a two-stage framework to apply the focused information criterion (FIC). Some asymptotic properties of the proposed methods are derived. Simulation studies show that the grouped variable selection is superior to separate model selections. Furthermore, the FIC improves the accuracy of the estimated AUC compared with other criteria.

1. Introduction

In modern medical diagnosis or genetic studies, the *receiver operating characteristic* (ROC) curve is a popular tool to evaluate the discrimination performance of a certain biomarker on a disease status or a phenotype. For example, in a continuous-scale test, the diagnosis of a disease is dependent upon whether a test result is above or below a specified cutoff value. Also, genome-wide association studies in human populations aim at creating genomic profiles which combine the effects of many associated genetic variants to predict the disease risk of a new subject with high discriminative accuracy [1]. For a given cutoff value of a biomarker or a combination of biomarkers, the sensitivity and the specificity are employed to quantitatively evaluate the discriminative performance. By varying cutoff values throughout the entire real line, the resulting plot of sensitivity against 1-specificity is a ROC curve. The area under the ROC curve (AUC) is an important one-number summary index of the overall discriminative accuracy of a ROC curve, by taking the influence of all cutoff values into account. Let Y_D be the response of a diseased subject, and let $Y_{\overline{D}}$ be the response of a nondiseased subject; then, the AUC can be expressed as $P(Y_D > Y_{\overline{D}})$ [2]. Pepe [3] and Zhou et al. [4] provided broad reviews on many statistical methods for the evaluation of diagnostic tests.

Traditional ROC analyses do not consider the effect of characteristics of study subjects or operating conditions of the test, so test results may be affected in the way of influencing distributions of test measurements for diseased and/or nondiseased subjects. Additionally, although the number of genes is large, there may be only a small number of them associated with the disease risk or phenotype. Therefore, regression models are introduced into the ROC analysis. Chapter Six in Pepe [3] offered a wonderful introduction to the adjustment for covariates in ROC curves. As reviewed in Rodríguez-Álvarez et al. [5], there are two main methodologies of regression analyses in ROC: (1) "induced" methodology, which firstly models outcomes of diseased and nondiseased subjects separately and then uses these outcomes to induce ROC and AUC and (2) "direct" methodology, which directly models the AUC on all covariates. In this paper, we focus on the induced methodology, to which current model selection techniques may be extended.

If there are many covariates, the variable selection issue arises in terms of the consideration of model interpretation and estimability. There are two main groups of variable selection procedures. One is the best-subset selection associated with criteria such as cross-validation (CV, [6]), generalized cross-validation (GCV, [7]), AIC [8], and BIC [9]. The other is based on regularization methods such as LASSO [10],

SCAD [11], and adaptive LASSO [12], with tuning parameters selected by the same criteria such as CV and BIC. Procedures in the second group have recently become popular because they are stable [13] and applicable for high-dimensional data [14].

So far, not much attention has been drawn on the topic of variable selection in the ROC regression. Two possible reasons may account for this situation. Firstly, if we model outcomes of diseased and nondiseased subjects separately, selected submodels may be different. The difference will result in difficulties in interpretation, because it is natural to expect that the same set of variables contributes to discriminating diseased and nondiseased subjects. Secondly, most current criteria for variable selection procedures focus on the prediction performance or variable selection consistency. However, in the ROC regression, instead of prediction or model selection, our focus is the precision of an estimated AUC, which means that most popular criteria may not be appropriate. Claeskens and Hjort [15] argued that these "one-fit-all" model selection criteria aim at selecting a single model with good overall properties. Alternatively, they developed the focused information criterion (FIC), which focuses on a parameter singled out for interests. The insight behind this criterion is that a model that gives good precision for one estimand may be worse when used in inference for another estimand. Wang and Fang [16] successfully applied the FIC to variable selection in linear models and demonstrated that the FIC exactly improved the estimation performance of singled-out parameters. This "individualized" criterion exactly fits the ROC regression.

The remaining parts of this paper are organized as follows. In Section 2, we rewrite the ROC regression into a grouped variable selection form so that current criteria can be applied. Then, a general two-stage framework with a BIC selector for the group SCAD under the local model assumption is proposed in Section 3. Simulation studies and a real data analysis are given in Sections 4 and 5. A brief discussion is provided in Section 6. All proofs are presented in the Supplement; see Supplementary Materials available online at http://dx.doi.org/10.1155/2013/436493.

2. ROC Regression

In this section, we rewrite the penalized ROC regression with induced methodology into a problem of the grouped variable selection by SCAD. Initially, we require that all covariates be centered at 0 for the consideration of comparability. Also, for notation simplicity, response variables are centered. If not, we can center responses to finish the model selection and then add centers back to evaluate the AUC. By following notations of the local model, which generalizes the commonly used sparsity assumption, homoscedastic regression models for diseased and nondiseased subjects are assumed as follows:

$$\begin{aligned} y_D &= z^T\theta_D + \sigma_D\varepsilon_D = x^T\beta_{D0} + u^T\gamma_D + \sigma_D\varepsilon_D, \\ y_{\overline{D}} &= z^T\theta_{\overline{D}} + \sigma_{\overline{D}}\varepsilon_{\overline{D}} = x^T\beta_{\overline{D}0} + u^T\gamma_{\overline{D}} + \sigma_{\overline{D}}\varepsilon_{\overline{D}}, \end{aligned} \tag{1}$$

where x includes p variables added always, u includes q variables which may or may not be added, $z = (x^T, u^T)^T$, β_{D0} and $\beta_{\overline{D}0}$ are p dimensional vectors, $\gamma_D = \delta_D/\sqrt{n_D}$ and $\gamma_{\overline{D}} = \delta_{\overline{D}}/\sqrt{n_{\overline{D}}}$ are q dimensional vectors with n_D and $n_{\overline{D}}$ as sample sizes for diseased and nondiseased groups, respectively, $\theta_D = (\beta_{D0}^T, \gamma_D^T)^T = (\theta_{D1}, \ldots, \theta_{Dd})^T$ and $\theta_{\overline{D}} = (\theta_{\overline{D}1}, \ldots, \theta_{\overline{D}d})^T$ are $d \triangleq p + q$ dimensional vectors, and ε_D and $\varepsilon_{\overline{D}}$ independently follow $\mathcal{N}(0, 1)$. Especially, if $\delta_D = \delta_{\overline{D}} = \mathbf{0}_q$, a sparse model is given. Then, the AUC given z can be written as

$$\mathrm{AUC}_z = \Pr\left(y_D \geq y_{\overline{D}} \mid z\right) = \Phi\left(\frac{z^T\left(\theta_D - \theta_{\overline{D}}\right)}{\sqrt{\sigma_{\overline{D}}^2 + \sigma_D^2}}\right), \tag{2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. Clearly, the narrow model is $\mathcal{S}_0 = \{1, \ldots, p\}$, including all constant effects β_{D0} and $\beta_{\overline{D}0}$. More details of the local model assumption are provided in the following section.

Assume that observed i.i.d that the samples are $\{(y_{Di}, z_{Di})\}$, $i = 1, \ldots, n_D$, and $\{(y_{\overline{D}j}, z_{\overline{D}j})\}$, $j = 1, \ldots, n_{\overline{D}}$. Instead of selecting separate models, we consider the following single objective function with a group penalty, given a tuning parameter λ:

$$\begin{aligned} Q_\lambda\left(\theta_D, \theta_{\overline{D}}\right) = {} & \frac{1}{2n\sigma_D^2}\sum_{i=1}^{n_D}\left(y_{Di} - z_{Di}^T\theta_D\right)^2 \\ & + \frac{1}{2n\sigma_{\overline{D}}^2}\sum_{j=1}^{n_{\overline{D}}}\left(y_{\overline{D}j} - z_{\overline{D}j}^T\theta_{\overline{D}}\right)^2 \\ & + \lambda\sum_{s=1}^{d} p_\lambda\left(\|\boldsymbol{\theta}_s\|\right), \end{aligned} \tag{3}$$

where $\boldsymbol{\theta}_s = (\theta_{\overline{D}s}, \theta_{Ds})^T$, a 2-dimensional vector, with the sth component θ_{Ds} of θ_D and the sth component $\theta_{\overline{D}s}$ of $\theta_{\overline{D}}$, and $\partial p_\lambda(w)/\partial w = \lambda I(w \leq \lambda) + \max(0, a\lambda - w)I(w > \lambda)/(a-1)$ with $a = 3.7$. More generally, instead of the L_2 norm for $\boldsymbol{\theta}_s$, we can define $\|\boldsymbol{\theta}_s\|_{K_s} = \sqrt{\boldsymbol{\theta}_s^T K_s \boldsymbol{\theta}_s}$ with a positive definite 2×2 matrix W_s. Then, given λ, the minimizer of (3) can be obtained as an estimate of $(\theta_D^T, \theta_{\overline{D}}^T)^T$. The motivation of considering such a penalty on $\boldsymbol{\theta}_s$ jointly rather than separately is that the inclusion or exclusion of the effect of a certain variable should be simultaneous for both diseased and nondiseased groups. It may not be appropriate to include either θ_{Dk} or $\theta_{\overline{D}k}$ in the model only, which will bring troubles in interpretation of the resulting model. This is exactly the motivation of the group LASSO method by Yuan and Lin [17] to handle categorical variables, and the group SCAD by Wang et al. [18] to address spline bases.

Note that there are two separate summations of residual squares in (3). In order to comply with the framework of selecting grouped variables, a modified version of the objective function (3) is required. Let $\otimes$ be the Kronecker product operator. Define $\boldsymbol{\theta} = \theta_{\overline{D}} \otimes (1, 0)^T + \theta_D \otimes (0, 1)^T$,

$\mathbf{z}_{\overline{D}j} = z_{\overline{D}j} \otimes (1,0)^T$, $j = 1,\ldots,n_{\overline{D}}$, and $\mathbf{z}_{Di} = z_{Di} \otimes (0,1)^T$, $i = 1,\ldots,n_D$. In matrix form, we have

$$Y = \left(y_{\overline{D}1},\ldots,y_{\overline{D}n_{\overline{D}}},y_{D1},\ldots,y_{Dn_D}\right)^T,$$
$$Z = \left(\mathbf{z}_{\overline{D}1},\ldots,\mathbf{z}_{\overline{D}n_{\overline{D}}},\mathbf{z}_{D1},\ldots,\mathbf{z}_{Dn_D}\right)^T, \quad (4)$$

where Y is an $n \triangleq n_{\overline{D}} + n_D$ dimensional vector with components y_i, $i = 1,\ldots,n$, and Z is an $n \times 2d$ dimensional matrix. Clearly, there are d grouped variables, and Z can be split into d submatrices $Z = (Z_l,\ldots,Z_d)$, each of which includes two consecutive columns of Z in turn. Similarly, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T,\ldots,\boldsymbol{\theta}_d^T)^T$ with $\boldsymbol{\theta}_m = (\theta_{\overline{D}m},\theta_{Dm})^T$, $m = 1,\ldots,d$. Additionally, due to different variances of healthy and diseased subjects, weighted least squares should be applied. Let W be a diagonal matrix, with each diagonal entry

$$W_{ii} = \begin{cases} \sigma_{\overline{D}}^{-2} & \text{if } i = 1,\ldots,n_{\overline{D}}, \\ \sigma_D^{-2} & \text{if } i = n_{\overline{D}}+1,\ldots,n. \end{cases} \quad (5)$$

Then, the objective function (3) is written as

$$Q_\lambda(\boldsymbol{\theta}) = \frac{1}{2n}\left\|Y - \sum_{m=1}^{d} Z_m\boldsymbol{\theta}_m\right\|_W^2 + \lambda\sum_{m=1}^{d} p_\lambda\left(\|\boldsymbol{\theta}_m\|\right). \quad (6)$$

Furthermore, in order to facilitate computation with current R packages, we would define transformed observations by weighting. Simply, put $\widetilde{Y} = W^{1/2}Y$ and $\widetilde{Z}_m = W^{1/2}Z_m$. Therefore,

$$Q_\lambda(\boldsymbol{\theta}) = \frac{1}{2n}\left\|\widetilde{Y} - \sum_{m=1}^{d} \widetilde{Z}_m\boldsymbol{\theta}_m\right\|^2 + \lambda\sum_{m=1}^{d} p_\lambda\left(\|\boldsymbol{\theta}_m\|\right). \quad (7)$$

Finally, the penalized ROC regression (3) has been written into a group SCAD-type problem (7). Then, current model selection criteria, like CV, GCV, AIC, and BIC, can be applied to select a final model. For this specific ROC regression problem, where AUC is the focus, these criteria may not be appropriate. Therefore, as argued by Claeskens and Hjort [15], the FIC can play a role here.

Under the local model assumption, a novel procedure of applying the FIC to the grouped variable selection is developed, which is motivated by Wang and Fang [16]. Briefly speaking, the procedure consists of two steps. Firstly, a narrow model, containing variables added always, is identified through the objective function (7). Secondly, the FIC is applied to select a subgroup of remaining variables. As a consequence, the final model is the combination of variables selected in both two steps. Details are provided in the following section. In terms of FIC, naturally, the focus parameter is the AUC at a given z_0; that is, $\mu(\boldsymbol{\theta}) = \Phi((z_0^T \otimes (-1,1))\boldsymbol{\theta}/\sqrt{\sigma_{\overline{D}}^2 + \sigma_D^2})$ with $\partial\mu/\partial\boldsymbol{\theta} = \phi((z_0^T \otimes (-1,1))\boldsymbol{\theta}/\sqrt{\sigma_{\overline{D}}^2 + \sigma_D^2})(z_0 \otimes (-1,1)^T/\sqrt{\sigma_{\overline{D}}^2 + \sigma_D^2})$.

Later, in simulation studies, the separate variable selection for diseased and nondiseased models will also be utilized to make a comparison. We expect, the group selection is superior to the separate selection.

3. A BIC Selector for Group SCAD under the Local Model Assumption

This section follows notations used in the two fundamental papers of the FIC: Hjort and Claeskens [19] and Claeskens and Hjort [15]. Furthermore, we allow grouped variables, each of which stands for a factor, such as a series of dummy variables coded from a multilevel categorical variable. The starting assumption of the FIC is that some variables are added to the regression model always and the others may or may not be added; that is,

$$y_i = x_i^T\beta_0 + u_i^T\gamma + \varepsilon_i, \quad i = 1,\ldots,n, \quad (8)$$

where x_i includes p variables which are added always, u_i includes q variables which may or may not be added, and $\varepsilon_i \sim \mathcal{N}(0,\sigma_\varepsilon^2)$. Without loss of generality, both x_i and y_i are standardized to remove the intercept term. Furthermore, we assume that x_i actually consists of K factors, that is, $x_i = (x_{i1}^T,\ldots,x_{iK}^T)^T$, and the corresponding $\beta_0 = (\beta_{01}^T,\ldots,\beta_{0K}^T)^T$, with dimensions p_k for each x_{ik} and β_{0k}, $k = 1,\ldots,K$, such that $\sum_{k=1}^K p_k = p$. Similarly, u_i consists of L factors, that is, $u_i = (u_{i1}^T,\ldots,u_{iL}^T)^T$, and the corresponding $\gamma = (\gamma_1^T,\ldots,\gamma_L^T)^T$, with dimensions q_l for each u_{il} and γ_l, $l = 1,\ldots,L$, such that $\sum_{l=1}^L q_l = q$. Let $z_i^T = (x_i^T,u_i^T) = (z_{i1}^T,\ldots,z_{iM}^T)$, with $d \triangleq p+q$ dimensions, and each z_{im} has d_m dimensions, $m = 1,\ldots,M$, such that $M \triangleq K+L$ and $\sum_{m=1}^M d_m = d$. Let $Y = (y_1,\ldots,y_n)^T$, $X = (x_1,\ldots,x_n)^T$, $U = (u_1,\ldots,u_n)^T$, and $Z = (z_1,\ldots,z_n)^T$. For simplicity, assume that the residual variance σ_ε^2 is estimated based on the full model and is not considered as a parameter.

In the literature of the variable selection, in order to show the selection consistency of a variable selection procedure, usually, the true model is assumed to be sparse. Thus, the sparsity assumption plays a critical role in the current model selection literature. Many procedures have been shown to be selection consistent under this sparsity assumption [20]. For example, the SCAD with tuning parameter selected via BIC has been shown to be selection consistent by Wang et al. [21, 22], and Zhang et al. [23].

However, it is questionable or too strict to assume that the true model is sparse. It is more reasonable and flexible to consider the local model (8) with $\theta_{\text{true}}^T = (\beta_0^T,\gamma^T)$ and $\gamma = \gamma_0 + \delta/\sqrt{n}$ as a true model, where $\gamma_0 = \mathbf{0}_q$ for the purpose of variable selection, under which the FIC is developed. This model is close to the sparse model, but it is different from it by $\gamma - \mathbf{0}_q = \delta/\sqrt{n}$. The sparsity assumption, with notations in this paper, is equivalent to assume that $\delta = \mathbf{0}_q$ and $\theta_{\text{true}}^T = (\beta_0^T,\mathbf{0}_q^T)$. Therefore, the local model assumption used here is a natural extension of the sparsity assumption. All "consistency" results obtained in this paper still apply to sparse models with grouped variables.

The FIC centers at the inference on a certain estimand or focus, denoted by $\mu_{\text{true}} = \mu(\theta_{\text{true}})$. It is well known that using a bigger model would typically mean smaller bias but bigger variance. Therefore, the FIC tries to balance the bias and the variance of estimating a certain parameter estimand. To be specific, like what any existing criterion does, among

a possible model range, the FIC starts with a narrow model that includes only variables in x_i and searches over submodels including some factors in u_i. The whole process leads to totally 2^L submodels, one for each subset of $\{1, \ldots, L\}$.

In this framework, various estimators of the focus parameter range from $\widehat{\mu}_{\text{full}} = \mu(\widehat{\beta}_{\text{full}}, \widehat{\gamma}_{\text{full}})$ to $\widehat{\mu}_{\text{narr}} = \mu(\widehat{\beta}_{\text{narr}}, \mathbf{0}_q)$. In general, the FIC attempts to select a subset $\widehat{\mathcal{S}}$ associated with the smallest mean squared error (MSE) of $\widehat{\mu}_{\mathcal{S}} = \mu(\widehat{\beta}_{\mathcal{S}}, \widehat{\gamma}_{\mathcal{S}}, \gamma_{0,\mathcal{S}^c})$, where $\mathcal{S}^c$ is the complement of $\mathcal{S}$ and the subscript $\mathcal{S}$ means a subset of corresponding vectors indexed by $\mathcal{S}$.

3.1. Stage 1: Consistent Selection of the Narrow Model. Once assuming the true model (8) with $\theta_{\text{true}}^T = (\beta_0^T, \gamma^T)$ and $\gamma = \delta/\sqrt{n}$ as well as grouped variables, here arises the first important question regarding whether we can select the narrow model $\mathcal{S}_0 = \{1, \ldots, K\}$ consistently. A similar question has been addressed by Wang and Fang [16], where they considered nongrouped variables. In the following, we show that the group SCAD with a tuning parameter selected via BIC can consistently select the narrow model.

Wang et al. [18] extended the SCAD, proposed by Fan and Li [11], to grouped variables and established its oracle property, following an elegant idea of the group LASSO [17]. The group SCAD generates an estimate via following penalized least squares:

$$\widehat{\theta}_\lambda = \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|Y - Z\theta\|^2 + \sum_{m=1}^{M} p_\lambda \left(\|\theta_m\| \right) \right\}, \quad (9)$$

where $\theta = (\theta_1^T, \ldots, \theta_M^T)^T$ with d_m-dimensional θ_m, and $p_\lambda(\cdot)$ is defined in the previous section. Let $\widehat{\mathcal{S}}_\lambda = \{m : \widehat{\theta}_{\lambda m} \neq \mathbf{0}_{d_m}\}$ be the selected narrow model for a given λ. With similar arguments in the previous section, the L_2 norm used in the penalty can be replaced by any metric with the form $\|\theta_m\|_{K_m} \triangleq (\theta_m^T K_m \theta_m)^{1/2}$ such that K_m is a symmetric $d_m \times d_m$ positive definite matrix.

Under the local model assumption with no grouped variables, Wang and Fang [16] showed that, with a tuning parameter λ selected via BIC, the SCAD is selection consistent; that is, with probability tending to one, the narrow model can be identified. Similarly, a BIC selector can be defined based on the group SCAD as follows:

$$\widehat{\lambda}_{\text{B}} = \underset{\lambda}{\operatorname{argmin}} \left\{ \log\left(\widehat{\sigma}_\lambda^2\right) + \frac{\text{df}_\lambda \log(n)}{n} \right\}, \quad (10)$$

where $\widehat{\sigma}_\lambda^2 = \|Y - Z\widehat{\theta}_\lambda\|^2/n$ and $\text{df}_\lambda = \sum_{m \in \widehat{\mathcal{S}}_\lambda} d_m$. We expect that the group SCAD is still selection consistent in the sense that $\Pr(\widehat{\mathcal{S}}_{\widehat{\lambda}_{\text{B}}} = \mathcal{S}_0) \to 1$ as $n \to \infty$, provided that $\mathcal{S}_0$ is the narrow model.

Formally, within the framework of FIC, assuming that the local model (8) is the true model and that $\mathcal{S}_0$ is the narrow model, we show the following theorem. Proofs can be found in the Supplement.

Theorem 1. *Under some mild conditions (see the Supplement for details), one has that*

$$\Pr\left(\widehat{\mathcal{S}}_{\widehat{\lambda}_{\text{B}}} = \mathcal{S}_0\right) \longrightarrow 1, \quad \text{as } n \longrightarrow \infty, \quad (11)$$

provided that model (8) *with* $\theta_{true} = (\beta_0^T, \gamma^T)^T$ *and* $\gamma = \delta/\sqrt{n}$ *is the true model.*

Remark 2. If we assume that $\delta = \mathbf{0}_q$, that is, the model is sparse, then Theorem 1 provides a BIC selector for the tuning parameter in the group SCAD, which can consistently identify nonzero effects. In other words, we extend the BIC selector for the SCAD proposed by Wang et al. [21] to the situation with the group SCAD.

Theorem 1 also implies both advantages and disadvantages of the BIC, which have been discussed by Wang and Fang [16]. Briefly speaking, the BIC sacrifices prediction consistency [24] in the sense of filtering all of the variables whose effect sizes are of order $O(1/\sqrt{n})$ to achieve the model selection consistency. The previous theorem provides a data-driven method to consistently specify a narrow model, which is critical before applying FIC. In the following subsection, we suggest a two-stage framework to apply the FIC based upon a narrow model selected via the BIC, in order to recover part of the variables filtered by the BIC.

3.2. Stage 2: FIC. In Stage 1, a narrow model, $\widehat{\mathcal{S}}_0 = \{1, \ldots, \widehat{K}\}$, has been identified via the group SCAD with a tuning parameter selected via BIC. In Stage 2, any subset of $\widehat{\mathcal{S}}_0^c = \{\widehat{K}+1, \ldots, M = \widehat{K} + \widehat{L}\}$ can be added to $\widehat{\mathcal{S}}_0$. A direct application of the FIC proposed by Claeskens and Hjort [15] is not plausible even for moderate size of $\widehat{L}$, because there are $2^{\widehat{L}}$ subsets of $\widehat{\mathcal{S}}_0^c$. Furthermore, the best-subset selection is unstable [13]. Therefore, similar to Wang and Fang [16], without double minimizations through both subsets and tuning parameters proposed by Claeskens [25], we suggest limiting the search domain to those subsets on the solution path from any group regularization procedure such as group LASSO or group SCAD.

With a selected narrow model $\widehat{\mathcal{S}}_0 = \{1, \ldots, \widehat{K}\}$, let $\widetilde{x}_i = (z_{i1}^T, \ldots, z_{i\widehat{K}}^T)^T$, $\widetilde{u}_i = (z_{i,\widehat{K}+1}^T, \ldots, z_{iM}^T)^T$, $\widetilde{\beta} = (\theta_1^T, \ldots, \theta_{\widehat{K}}^T)^T$, and $\widetilde{\gamma} = (\theta_{\widehat{K}+1}^T, \ldots, \theta_M^T)^T$. Then, a solution path is generated from the following group LASSO procedure (or group SCAD):

$$\left(\widehat{\beta}_\tau, \widehat{\gamma}_\tau\right) = \underset{\widetilde{\beta}, \widetilde{\gamma}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \frac{\left(y_i - \widetilde{x}_i^T \widetilde{\beta} - \widetilde{u}_i^T \widetilde{\gamma}\right)^2}{2n} + \tau \sum_{l=1}^{\widehat{L}} \|\widetilde{\gamma}_l\| \right\}, \quad (12)$$

where the tuning parameter τ controls the grouped variables included in the subset $\widehat{\mathcal{A}}_\tau = \{l : \widehat{\gamma}_{\tau l} \neq \mathbf{0}_{q_l}\}$. As the tuning parameter τ varies from some large value to 0, $\widehat{\mathcal{A}}_\tau$ increases from an empty set to a "full" set $\{1, \ldots, \widehat{L}\}$. Then, we utilize the FIC to guide the selection of τ in (12) over the resulting $\widehat{\mathcal{A}}_\tau$'s, which consist of a search domain.

Now, Stage 2 of the FIC for a certain focus $\mu_{\text{true}} = \mu(\theta_{\text{true}})$ is summarized as follows. For a given τ, a subset $\widehat{\mathcal{A}}_\tau$ is provided by indices of nonzero factors from (12). Then, based on

the submodel $\mathcal{S} = \widehat{\mathcal{S}}_0 \cup \widehat{\mathcal{A}}_\tau$, the FIC_τ is evaluated according to a formula developed in Claeskens and Hjort [15, formula (3.3)], which is essentially a parametric estimate of the MSE of μ_{true} on a model $\mathcal{S}$. Consequently, τ is selected as

$$\widehat{\tau}_{\mathrm{F}} = \underset{\tau}{\operatorname{argmin}}\, \mathrm{FIC}_\tau, \tag{13}$$

and the final submodel is selected as $\widehat{\mathcal{S}}_{\mathrm{F}} = \widehat{\mathcal{S}}_0 \cup \widehat{\mathcal{A}}_{\widehat{\tau}_{\mathrm{F}}}$.

4. Simulation

Simulated data are generated under models (1) with 0 as intercepts. Moderate sample sizes are set to $n_{\overline{D}} = 50$ and $n_D = 50$, compared with 8 and 20 as numbers of covariates. Three scenarios of parameters are considered in the following:

(1) $\sigma_{\overline{D}} = \sigma_D = 2$, $\beta_D = (1.5, 2, 3)$, $\gamma_{Dj} = (3 - 0.5(j-1))/\sqrt{n_D}$, $j = 1, \ldots, 5$, $\theta_D = (\beta_D^T, \gamma_D^T)^T$, $\beta_{\overline{D}} = (0.5, 1, 2)$, $\gamma_{\overline{D}j} = (1 - 0.2(j-1))/\sqrt{n_{\overline{D}}}$, $j = 1, \ldots, 5$, $\theta_{\overline{D}} = (\beta_{\overline{D}}^T, \gamma_{\overline{D}}^T)^T$, $p = 3$, $q = 5$, $d = 8$;

(2) $\sigma_{\overline{D}} = \sigma_D = 2$, $\beta_D = (1.5, 2, 3)$, $\gamma_{Dj} = (2 - 0.05(j-1))/\sqrt{n_D}$, $j = 1, \ldots, 17$, $\theta_D = (\beta_D^T, \gamma_D^T)^T$, $\beta_{\overline{D}} = (0.5, 1, 2)$, $\gamma_{\overline{D}j} = (1 - 0.05(j-1))/\sqrt{n_{\overline{D}}}$, $j = 1, \ldots, 17$, $\theta_{\overline{D}} = (\beta_{\overline{D}}^T, \gamma_{\overline{D}}^T)^T$, $p = 3$, $q = 17$, $d = 20$;

(3) $\sigma_{\overline{D}} = \sigma_D = 1$, $\theta_{Dj} = 3/2j$, $\theta_{\overline{D}j} = 2/2j$, $j = 1, \ldots, 8$, $d = 8$.

Clearly, the narrow model of the first two settings is $\{1, 2, 3\}$, whereas, for the third one, no clear boundary is specified between big effects and small effects.

Corresponding to each setting, test datasets z_{01}, z_{02}, and z_{03} are selected to generate AUC around 0.6, 0.8, and 0.95 to accommodate low-, moderate-, and high-accuracy cases, respectively. Consider the following:

(1) $z_{01} = (0.2, \ldots, 0.2)^T$, AUC = 0.611; $z_{02} = (0.7, \ldots, 0.7)^T$, AUC = 0.838; $z_{03} = (1.2, \ldots, 1.2)^T$, AUC = 0.955;

(2) $z_{01} = (0.15, \ldots, 0.15)^T$, AUC = 0.613; $z_{02} = (0.45, \ldots, 0.45)^T$, AUC = 0.805; $z_{03} = (0.9, \ldots, 0.9)^T$, AUC = 0.957;

(3) $z_{01} = (0.3, \ldots, 0.3)^T$, AUC = 0.613; $z_{02} = (0.9, \ldots, 0.9)^T$, AUC = 0.806; $z_{03} = (1.8, \ldots, 1.8)^T$, AUC = 0.958.

Besides the proposed two-stage framework (FIC) with group SCAD, for comparison purpose, four popular variable selection criteria, including 5-fold CV, GCV, AIC, and BIC, are also employed. Additionally, the SCAD penalty is applied to diseased and healthy groups separately to show the gain of applying the group SCAD.

Two popular measurements, $\mathrm{MSE} = E(\mu(\widehat{\theta}_{\widehat{\mathcal{S}}_{\mathrm{F}}}) - \mu(\theta_{\text{true}}))^2$ and the mean absolute error (MAE), defined by $E|\mu(\widehat{\theta}_{\widehat{\mathcal{S}}_{\mathrm{F}}}) - \mu(\theta_{\text{true}})|$, are utilized to evaluate the prediction performance of selected models based on different criteria, where $\widehat{\theta}_{\widehat{\mathcal{S}}_{\mathrm{F}}}$ is an estimate of θ based on the final model $\widehat{\mathcal{S}}_{\mathrm{F}}$ selected by a certain selection criterion. Due to the limited range of AUC and skewed distributions of estimates of AUC especially at boundaries, the MAE is supposed to be more appropriate.

Table 1: Model selection performance for group SCAD.

Setting	Method	F-measure (%)
1	CV	71.3
	GCV	72.8
	AIC	70.9
	BIC	77.4
2	CV	66.0
	GCV	66.0
	AIC	66.2
	BIC	67.8

In this paper, a composite measurement, the F-measure, is employed to evaluate the performance of selecting the narrow model among various methods, including commonly used proportions of selecting underfitting, correct, and overfitting models separately. As noted by Lim and Yu [26], a high F-measure means that both false-positive and false-negative rates are low. Define Precision = true positivity, Recall = true discovery and then, F-measure $\triangleq$ (2 · Precision · Recall)/(Precision+Recall). All results are summarized based on 500 repetitions according to simulation settings in Tables 1, 2, and 3.

Table 1 indicates that the BIC has the best performance to identify the narrow model, compared with others. Also, if there are more weak signals, like Setting 2, the performance is not as good as that of Setting 1. This is reasonable, because, with increasing number of variables given the sample size, it is more challenging to filter weak signals, even under the sparsity assumption. From Table 2, we can see that, in all three settings, these five methods perform well. Specifically, for moderate and large AUC cases, the FIC performs slightly better, providing smaller MAE. Additionally, in these cases, the FIC improves the BIC substantially, which once again indicates that the BIC would filter weak signals.

In order to show how we can benefit from applying the grouped variable selection, separate model selections for diseased and healthy subjects are also considered, and results are summarized in Table 3. By comparing Tables 2 and 3, in most cases, the group penalty provides smaller MSE and MAE for every criterion. Due to limited range of the AUC, all MSE and MAE values in Tables 2 and 3 are small, but the group selection can improve separate selections by as high as 25%. It is not surprising to see that, in high AUC situations, differences are small, and separate selections with BIC are better. Possible reasons are the following: (1) there is no much room for an estimated AUC to vary when it is close to 1; (2) separate selections with BIC offer a larger flexibility to obtain a sparse model.

5. Real Data Analysis

In this section, we demonstrate the proposed procedure by the audiology data reported by Stover et al. [27], which has

Table 2: Prediction of AUC at z_{0k} with group SCAD. Size means the number of selected factors, where each factor contains two variables.

Setting	Methods	z_{01}			z_{02}			z_{03}		
		MSE	MAE	Size	MSE	MAE	Size	MSE	MAE	Size
1	CV	0.00345	0.0467	4.72	0.00295	0.0437	4.72	0.00114	0.0255	4.72
	GCV	0.00345	0.0467	4.49	0.00294	0.0432	4.49	0.00114	0.0251	4.49
	AIC	0.00349	0.0468	4.90	0.00291	0.0428	4.90	0.00110	0.0246	4.90
	BIC	0.00335	0.0461	3.62	0.00317	0.0450	3.62	0.00147	0.0278	3.62
	FIC	0.00339	0.0464	4.23	0.00283	0.0426	4.23	0.00108	0.0247	4.23
2	CV	0.00328	0.0461	8.31	0.00279	0.0428	8.31	0.00073	0.0209	8.31
	GCV	0.00339	0.0470	9.43	0.00281	0.0433	9.43	0.00066	0.0206	9.43
	AIC	0.00344	0.0472	12.05	0.00285	0.0434	12.05	0.00064	0.0204	12.05
	BIC	0.00324	0.0458	6.14	0.00328	0.0462	6.14	0.00129	0.0259	6.14
	FIC	0.00327	0.0459	7.97	0.00290	0.0440	7.97	0.00081	0.0224	7.97
3	CV	0.00369	0.0483	6.67	0.00439	0.0535	6.67	0.00199	0.0317	6.67
	GCV	0.00367	0.0483	6.14	0.00436	0.0533	6.14	0.00197	0.0316	6.14
	AIC	0.00369	0.0484	6.36	0.00441	0.0534	6.36	0.00201	0.0318	6.36
	BIC	0.00367	0.0482	5.14	0.00473	0.0549	5.14	0.00247	0.0345	5.14
	FIC	0.00368	0.0483	5.46	0.00451	0.0532	5.49	0.00219	0.0324	5.49

Table 3: Prediction of AUC at z_{0k} with models on diseased and healthy groups separately. Size means the sum of numbers of selected variables in diseased and non-diseased groups.

Setting	Methods	Size	z_{01}		z_{02}		z_{03}	
			MSE	MAE	MSE	MAE	MSE	MAE
1	CV	8.78	0.00384	0.0490	0.00303	0.0439	0.00107	0.0247
	GCV	8.23	0.00383	0.0488	0.00298	0.0432	0.00103	0.0239
	AIC	8.18	0.00383	0.0488	0.00397	0.0432	0.00102	0.0239
	BIC	6.96	0.00383	0.0490	0.00313	0.0447	0.00114	0.0254
2	CV	14.47	0.00483	0.0566	0.00351	0.0481	0.00079	0.0218
	GCV	16.31	0.00553	0.0609	0.00390	0.0515	0.00069	0.0218
	AIC	15.46	0.00545	0.0606	0.00388	0.0516	0.00069	0.0218
	BIC	12.67	0.00514	0.0590	0.00384	0.0502	0.00092	0.0225
3	CV	12.29	0.00405	0.0494	0.00481	0.0558	0.00225	0.0332
	GCV	10.55	0.00403	0.0497	0.00461	0.0541	0.00208	0.0320
	AIC	10.55	0.00402	0.0497	0.00461	0.0541	0.00209	0.0320
	BIC	9.53	0.00403	0.0499	0.00468	0.0550	0.00206	0.0325

been analyzed by Pepe [3, 28]. The dataset contains results of distortion product otoacoustic emissions (DPOAE) test used to diagnose the hearing impairment. There are 208 subjects who were examined at different combinations of three frequencies (f) and three intensities (L) of the DPOAE device. An audiometric threshold can be obtained for each combination. At a particular frequency, if the audiometric threshold is greater than 20 dB HL, an ear was classified as hearing impaired. In the original dataset, there are multiple records for each subject. In this study, we randomly select one record for each subject, and among 208 subjects there are 55 subjects with hearing impairment. The test result is the negative signal-to-noise ratio, −SNR. The covariates used in Dodd and Pepe [29] are z_f = frequency Hz/100, z_L = intensity dB/10, and z_D = (hearing threshold − 20) dB/10. In order to encourage the model selection, we incorporate two-way interaction terms. Quadratic terms are not included due to the high correlation between each variable and its quadratic term. Therefore, z is the centered $(z_f, z_L, z_D, z_f z_L, z_f z_D, z_L z_D)^T$ for each element.

Former studies on this dataset showed that −SNR provided quite high discriminative performance and that z_f had a small effect. In order to avoid specifying inappropriate covariates, we randomly select three centered observations from the whole dataset as focused subjects.

Table 4 shows AUC values of models selected by each method as well as corresponding model sizes. CV, AIC, and GCV tend to select a full model. On the contrary, BIC tends to select a sparse model, only containing z_D. The full model may not provide the largest AUC, because a large model will bring instability and ruin the AUC. As indicated in the table, for the second test point, both BIC and FIC provide a higher AUC than the full model. But a single variable selected by the BIC seems to be too strict. By focusing on the precision of

Table 4: Estimated AUC at three test points. Size means the number of selected factors.

Methods	Test point 1		Test point 2		Test point 3	
	AUC	Size	AUC	Size	AUC	Size
CV	0.971	6	0.916	6	0.982	6
AIC	0.971	6	0.916	6	0.982	6
GCV	0.971	6	0.916	6	0.982	6
BIC	0.949	1	0.957	1	0.944	1
FIC	0.963	3	0.957	2	0.944	1

estimated focus parameter, the FIC provides a customized way to fill the gap: for the first test point, three main effects are selected; for the second one, z_L and z_D are selected; for the third one, only z_D is selected. Based on the precision of estimated AUC, the FIC performs as a compromise, selecting models to generate AUC values in the middle.

6. Discussion

In this paper, we rewrite the model selection problem of the ROC regression into a grouped factor selection form with induced methodology. Also, we develop a two-stage framework to apply the FIC to select a final model with group SCAD under the local model assumption. Specifically, if the true model is sparse, our framework naturally accommodates current model selection criteria. Furthermore, the BIC selector is proved to be model selection consistent if either a sparse or a local model is assumed, in the sense of selecting a sparse model or a narrow model.

Most current model selection criteria aim at the prediction performance or model selection consistency; thus, in the ROC regression where the AUC is a focus parameter, they may not be appropriate. This observation motivates an application of FIC, which is shown to perform well through simulation studies. Therefore, our method has a potential application in genetic studies, where the number of gene arrays is always large, compared with the sample size.

For the direct methodology, the literature based on generalized estimating equations is prosperous, which is motivated by the range $[0, 1]$ of the AUC, similar to the probability of a binary random variable. Our future work will extend the framework developed here to generalized estimating equations and apply it to the ROC regression with the direct methodology.

As discussed by one referee, it is possible that some coefficients are the same for both Y_D and $Y_{\overline{D}}$. As in (1), modeling them separately will increase the degree of freedom in (3), especially when a large number of genes are covariates. If the shrinkage of a coefficient, which is known a priori to be the same in both diseased and healthy groups, is not necessary, then it is natural for the FIC to include it in the narrow model with a single coefficient. By using the proposed objective function, a fused LASSO type of penalty may be applied to obtain such kind of structure, in addition to the group LASSO/SCAD. Friedman et al. [30] provided a note on the group LASSO and the sparse group LASSO, which could shed light on the question here. It will be also an interesting topic in the future.

Conflict of Interests

There is no conflict of interests regarding the publication of this article.

Acknowledgments

The authors would like to thank Dr. Yixin Fang for his invaluable suggestions and generous support which make this paper publishable. They also thank the editor, the associate editor, and the referees for their valuable comments which led to substantial improvements of this paper.

References

[1] N. R. Wray, J. Yang, M. E. Goddard, and P. M. Visscher, "The genetic interpretation of area under the ROC curve in genomic profiling," *PLoS Genetics*, vol. 6, no. 2, Article ID e1000864, 2010.

[2] D. Bamber, "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology*, vol. 12, no. 4, pp. 387–415, 1975.

[3] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, New York, NY, USA, 2003.

[4] X. H. Zhou, N. A. Obuchowski, and D. M. McClish, *Statistical Methods in Diagnostic Medicine*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2011.

[5] M. X. Rodríguez-Álvarez, P. G. Tahoces, C. Cadarso-Suárez, and M. J. Lado, "Comparative study of ROC regression techniques-applications for the computer-aided diagnostic system in breast cancer detection," *Computational Statistics and Data Analysis*, vol. 55, no. 1, pp. 888–902, 2011.

[6] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society B*, vol. 36, no. 2, pp. 111–147, 1974.

[7] P. Craven and G. Wahba, "Smoothing noisy data with spline functions—estimating the correct degree of smoothing by the method of generalized cross-validation," *Numerische Mathematik*, vol. 31, no. 4, pp. 377–403, 1979.

[8] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proceedings of the 2nd International Symposium Information Theory*, B. N. Petrov and F. Csaki, Eds., pp. 267–281, Akademia Kiado, Budapest, Hungary, 1973.

[9] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society B*, vol. 58, no. 1, pp. 267–288, 1996.

[11] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[12] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[13] L. Breiman, "Heuristics of instability and stabilization in model selection," *Annals of Statistics*, vol. 24, no. 6, pp. 2350–2383, 1996.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, NY, USA, 2009.

[15] G. Claeskens and N. L. Hjort, "The focused information criterion," *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 900–916, 2003.

[16] B. Wang and Y. Fang, "On the focused information criterion for variable selection," submitted.

[17] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society B*, vol. 68, no. 1, pp. 49–67, 2006.

[18] L. Wang, G. Chen, and H. Li, "Group SCAD regression analysis for microarray time course gene expression data," *Bioinformatics*, vol. 23, no. 12, pp. 1486–1494, 2007.

[19] N. L. Hjort and G. Claeskens, "Frequentist model average estimators," *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 879–899, 2003.

[20] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, 2011.

[21] H. Wang, R. Li, and C.-L. Tsai, "Tuning parameter selectors for the smoothly clipped absolute deviation method," *Biometrika*, vol. 94, no. 3, pp. 553–568, 2007.

[22] H. Wang, B. Li, and C. Leng, "Shrinkage tuning parameter selection with a diverging number of parameters," *Journal of the Royal Statistical Society B*, vol. 71, no. 3, pp. 671–683, 2009.

[23] Y. Zhang, R. Li, and C.-L. Tsai, "Regularization parameter selections via generalized information criterion," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 312–323, 2010.

[24] Y. Yang, "Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation," *Biometrika*, vol. 92, no. 4, pp. 937–950, 2005.

[25] G. Claeskens, "Focused estimation and model averaging with penalization methods: an overview," *Statistica Neerlandica*, vol. 66, no. 3, pp. 272–287, 2012.

[26] C. Lim and B. Yu, "Estimation Stability with Cross Validation (ESCV)," http://arxiv.org/abs/1303.3128.

[27] L. Stover, M. P. Gorga, S. T. Neely, and D. Montoya, "Toward optimizing the clinical utility of distortion product otoacoustic emission measurements," *Journal of the Acoustical Society of America*, vol. 100, no. 2, part 1, pp. 956–967, 1996.

[28] M. S. Pepe, "Three approaches to regression analysis of receiver operating characteristic curves for continuous test results," *Biometrics*, vol. 54, no. 1, pp. 124–135, 1998.

[29] L. E. Dodd and M. S. Pepe, "Semiparametric regression for the area under the receiver operating characteristic curve," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 409–417, 2003.

[30] J. Friedman, T. Hastie, and R. Tibshirani, *A Note on the Group Lasso and a Sparse Group Lasso*, 2010.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 843563, 12 pages
http://dx.doi.org/10.1155/2013/843563

Research Article

Robust Joint Analysis with Data Fusion in Two-Stage Quantitative Trait Genome-Wide Association Studies

Dong-Dong Pan,[1] Wen-Jun Xiong,[2] Ji-Yuan Zhou,[3] Ying Pan,[4] Guo-Li Zhou,[5] and Wing-Kam Fung[6]

[1] *Department of Statistics, Yunnan University, Kunming 650091, China*
[2] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*
[3] *Department of Biostatistics, School of Public Health and Tropical Medicine, Southern Medical University, Guangzhou 510515, China*
[4] *Department of Biology, Nanjing University, Nanjing 210093, China*
[5] *College of Mathematics and Statistics, Chongqing University, Chongqing 40044, China*
[6] *Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong*

Correspondence should be addressed to Dong-Dong Pan; ddpan@ynu.edu.cn and Wing-Kam Fung; wingfung@hku.hk

Received 6 July 2013; Accepted 29 July 2013

Academic Editor: Qizhai Li

Genome-wide association studies (GWASs) in identifying the disease-associated genetic variants have been proved to be a great pioneering work. Two-stage design and analysis are often adopted in GWASs. Considering the genetic model uncertainty, many robust procedures have been proposed and applied in GWASs. However, the existing approaches mostly focused on binary traits, and few work has been done on continuous (quantitative) traits, since the statistical significance of these robust tests is difficult to calculate. In this paper, we develop a powerful F-statistic-based robust joint analysis method for quantitative traits using the combined raw data from both stages in the framework of two-staged GWASs. Explicit expressions are obtained to calculate the statistical significance and power. We show using simulations that the proposed method is substantially more robust than the F-test based on the additive model when the underlying genetic model is unknown. An example for rheumatic arthritis (RA) is used for illustration.

1. Introduction

Genome-wide association studies (GWASs) have identified a large number of genomic regions (especially single-nucleotide polymorphisms (SNPs)) with a wide variety of complex traits/diseases. In a GWAS, two most common types of data, qualitative (or binary) and quantitative (or continuous) traits, are analyzed and two contentious points are often faced; one is how to construct the test statistic considering the genetic model uncertainty and the other is how to evaluate the statistical significance for controlling the false positive rates efficiently (e.g., [1, 2]). Considering these issues, a lot of work has been done on the binary trait in the past 10 years (e.g., [3–7]). Computer algorithms have also been developed to calculated the significance level of robust tests in GWASs, taking into account the genetic model uncertainty [8]. However, few work has been done on continuous traits, only recently So and Sham [9] proposed a MAX3 based on score test statistics, and Li et al. [10] gave a MAX3 based on F-test statistics. Note that these tests just focus on single-marker analysis in one-stage analysis.

Although the costs of whole-genome genotyping are decreasing with the high-throughput biological technology, the total costs for a GWAS are still very expensive due to the thousands of sampling units and huge amounts of single-nucleotide polymorphisms. In order to save the costs, the two-stage design and the corresponding statistical analysis where all the SNPs are genotyped in Stage 1 on a portion of the samples and the promising SNPs with small P-values (e.g., <0.001) based on some efficient tests are further screened on the remaining subjects, are often adopted in practice (e.g., [11–15]).

In genetic association studies, especially GWASs, genetic markers are routinely tested under the assumption of additive

effects. Although convenient to use, those tests are optimal only when the true underlying genetic model is additive so that they are not robust against the genetic model misspecification. To our best knowledge, few work has been done on the two-stage joint analysis for quantitative trait GWASs allowing for genetic model uncertainty. Here, we attempt to develop a joint analysis method with data fusion in the two-stage design using F-statistic, since F-test is commonly employed from the linear regression model for quantitative trait, and Li et al. [10] show that MAX3 based on F-statistics is more powerful than So and Sham's method by extensively numerical simulation.

The content of this paper is organized as follows. In Section 2, we give some notations and the proposed robust joint test statistics. Further, we derive the asymptotic distribution of the test statistics under the null and the alternative hypotheses. In Section 3, we show that the proposed joint analysis method is substantially more robust than the additive-model-based F-test from the numerical results of power comparison when the real genetic model is unknown. After that, an illustrative example for rheumatic arthritis (RA) is presented. Finally, we give some discussion of this paper in Section 4.

2. Methods

2.1. Notations. Assume that n individuals are randomly selected to be genotyped in a two-staged GWAS for a certain quantitative trait and that π is the sampling proportion in Stage 1. Let $n_1 = n\pi$ and $n_2 = n(1-\pi)$ be the sample sizes for Stages 1 and 2, respectively. Consider a biallelic marker with two alleles G and g. Without loss of generality, we assume that G is the minor or high-risk allele. We suppose that the total m SNPs are genotyped on the samples of Stage 1, and SNPs with P-values less than γ in Stage 1 will be further genotyped and tested in Stage 2. Let the significance level be α, and then the genome-wide significance level per SNP is α/m with the Bonferroni adjustments. Let $\mathbf{Y}_1 = (y_1, y_2, \ldots, y_{n_1})'$ and $\mathbf{Y}_2 = (y_{n_1+1}, y_{n_1+2}, \ldots, y_n)'$ be the observed quantitative outcome vectors for Stage 1 and Stage 2, respectively. Without loss of generality, we assume that the first n_{10} individuals in Stage 1 have the genotype gg, the second n_{11} individuals in Stage 1 have the genotype Gg, and the last n_{12} subjects in Stage 1 possess the genotype GG. Similarly, the first n_{20} subjects in Stage 2 have the genotype gg, the second n_{21} individuals in Stage 2 have the genotype Gg, and the last n_{22} subjects in Stage 2 possess the genotype GG. Let $\mathbf{0}_k = (0, 0, \ldots, 0)'_{k\times 1}$ and $\mathbf{1}_k = (1, 1, \ldots, 1)'_{k\times 1}$, and let $\mathbf{O}_{k\times j}$ be the $k \times j$ matrix with all its entries being zero and $\mathbf{I}_n$ be the $n \times n$ identity matrix.

2.2. F-Statistic-Based Robust Joint Analysis. We firstly briefly introduce F-statistic-based MAX3 by Li et al. [10] just using the data from Stage 1. Consider the following linear regression model:

$$y_i = \beta_0 + g_i\beta_1 + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \sigma^2\right), \ i = 1, 2, \ldots, n_1, \quad (1)$$

where β_0 is the nuisance parameter for the intercept, β_1 is the parameter of interest for genetic effect, and g_i is the genotype value, which takes 0, 1, or 2 corresponding to the count of G at a marker locus for the ith subject, $i = 1, 2, \ldots, n_1$. The hypotheses of interest are

$$H_0 : \beta_1 = 0 \longleftrightarrow H_1 : \beta_1 \neq 0. \quad (2)$$

The variable g_i in the previously stated equation is coded differently for the three common genetic models. Let $\mathbf{X}_{1R} = (\mathbf{1}_{n_1}, \mathbf{G}_{1R})$, $\mathbf{X}_{1A} = (\mathbf{1}_{n_1}, \mathbf{G}_{1A})$, and $\mathbf{X}_{1D} = (\mathbf{1}_{n_1}, \mathbf{G}_{1D})$ be the design matrices under three commonly used genetic models, where $\mathbf{G}_{1R} = (\mathbf{0}'_{n_{10}+n_{11}}, \mathbf{1}'_{n_{12}})'$ corresponds to the recessive model, $\mathbf{G}_{1A} = (g_1, g_2, \ldots, g_{n_1})'$ corresponds to the additive model, and $\mathbf{G}_{1D} = (\mathbf{0}'_{n_{10}}, \mathbf{1}'_{n_{11}+n_{12}})'$ is for the dominant model. Denote $\mathbf{X}_1 = (\mathbf{1}_{n_1}, \mathbf{x}_{11}, \mathbf{x}_{12})$, where $\mathbf{x}_{11} = (\mathbf{0}'_{n_{10}}, \mathbf{1}'_{n_{11}}, \mathbf{0}'_{n_{12}})'$ and $\mathbf{x}_{12} = (\mathbf{0}'_{n_{10}}, \mathbf{0}'_{n_{11}}, \mathbf{1}'_{n_{12}})'$. The modified F-test statistics under the recessive, additive, and dominant models for Stage 1 are given by

$$\begin{aligned}
F_1^R &= \frac{\mathbf{Y}_1'\left[\mathbf{X}_{1R}\left(\mathbf{X}_{1R}'\mathbf{X}_{1R}\right)^{-1}\mathbf{X}_{1R}' - \mathbf{1}_{n_1}\left(\mathbf{1}_{n_1}'\mathbf{1}_{n_1}\right)^{-1}\mathbf{1}_{n_1}'\right]\mathbf{Y}_1}{\mathbf{Y}_1'\left[\mathbf{I}_{n_1} - \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right]\mathbf{Y}_1/\left(n_1-3\right)} \\
&= \frac{\left(Z_1^R\right)^2}{\mathrm{RSS}_1/\left(n_1-3\right)}, \\
F_1^A &= \frac{\mathbf{Y}_1'\left[\mathbf{X}_{1A}\left(\mathbf{X}_{1A}'\mathbf{X}_{1A}\right)^{-1}\mathbf{X}_{1A}' - \mathbf{1}_{n_1}\left(\mathbf{1}_{n_1}'\mathbf{1}_{n_1}\right)^{-1}\mathbf{1}_{n_1}'\right]\mathbf{Y}_1}{\mathbf{Y}_1'\left[\mathbf{I}_{n_1} - \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right]\mathbf{Y}_1/\left(n_1-3\right)} \\
&= \frac{\left(Z_1^A\right)^2}{\mathrm{RSS}_1/\left(n_1-3\right)}, \\
F_1^D &= \frac{\mathbf{Y}_1'\left[\mathbf{X}_{1D}\left(\mathbf{X}_{1D}'\mathbf{X}_{1D}\right)^{-1}\mathbf{X}_{1D}' - \mathbf{1}_{n_1}\left(\mathbf{1}_{n_1}'\mathbf{1}_{n_1}\right)^{-1}\mathbf{1}_{n_1}'\right]\mathbf{Y}_1}{\mathbf{Y}_1'\left[\mathbf{I}_{n_1} - \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right]\mathbf{Y}_1/\left(n_1-3\right)} \\
&= \frac{\left(Z_1^D\right)^2}{\mathrm{RSS}_1/\left(n_1-3\right)},
\end{aligned} \quad (3)$$

where

$$\begin{aligned}
Z_1^R &= \sqrt{\frac{\left(n_{10}+n_{11}\right)n_{12}}{n_1}}\left(\overline{y}_{n_{10}+n_{11}} - \overline{y}_{n_{12}}\right), \\
Z_1^D &= \sqrt{\frac{n_{10}\left(n_{11}+n_{12}\right)}{n_1}}\left(\overline{y}_{n_{10}} - \overline{y}_{n_{11}+n_{12}}\right), \\
Z_1^A &= \left(n_{10}\left(n_{11}+2n_{12}\right)\overline{y}_{n_{10}} - n_{11}\left(n_{10}-n_{12}\right)\overline{y}_{n_{11}}\right. \\
&\quad \left. - n_{12}\left(2n_{10}+n_{11}\right)\overline{y}_{n_{12}}\right) \\
&\quad \times \left(\sqrt{n_1\left[n_{10}\left(n_{11}+4n_{12}\right)+n_{11}n_{12}\right]}\right)^{-1}, \\
\overline{y}_{n_{10}} &= \frac{1}{n_{10}}\sum_{j=1}^{n_{10}} y_j,
\end{aligned}$$

$$\overline{y}_{n_{11}} = \frac{1}{n_{11}} \sum_{j=n_{10}+1}^{n_{10}+n_{11}} y_j,$$

$$\overline{y}_{n_{12}} = \frac{1}{n_{12}} \sum_{j=n_{10}+n_{11}+1}^{n_1} y_j,$$

$$\overline{y}_{n_{10}+n_{11}} = \frac{1}{n_{10}+n_{11}} \sum_{j=1}^{n_{10}+n_{11}} y_j,$$

$$\overline{y}_{n_{11}+n_{12}} = \frac{1}{n_{11}+n_{12}} \sum_{j=n_{10}+1}^{n_1} y_j. \tag{4}$$

The robust test statistic in Stage 1 is

$$F_1^{\mathrm{MAX}} = \max\left\{F_1^R, F_1^A, F_1^D\right\}. \tag{5}$$

We now give the proposed robust joint analysis. In the framework of two-stage design GWAS of quantitative traits, the SNPs with P-values less than γ will be genotyped on the remaining n_2 subjects in Stage 2. Following the previous notation for Stage 1, corresponding to the recessive, additive, and dominant models, the genotype data in Stage 2 are denoted by $\mathbf{G}_{2R} = (\mathbf{0}'_{n_{20}+n_{21}}, \mathbf{1}'_{n_{22}})'$, $\mathbf{G}_{2A} = (g_{n_1+1}, g_{n_1+2}, \ldots, g_n)'$, and $\mathbf{G}_{2D} = (\mathbf{0}'_{n_{20}}, \mathbf{1}'_{n_{21}+n_{22}})'$, respectively, and the design matrices are $\mathbf{X}_{2R} = (\mathbf{1}_{n_2}, \mathbf{G}_{2R})$, $\mathbf{X}_{2A} = (\mathbf{1}_{n_2}, \mathbf{G}_{2A})$, and $\mathbf{X}_{2D} = (\mathbf{1}_{n_2}, \mathbf{G}_{2D})$, respectively. Denote $\mathbf{X}_2 = (\mathbf{1}_{n_2}, \mathbf{x}_{21}, \mathbf{x}_{22})$, where $\mathbf{x}_{21} = (\mathbf{0}'_{n_{20}}, \mathbf{1}'_{n_{21}}, \mathbf{0}'_{n_{22}})'$ and $\mathbf{x}_{22} = (\mathbf{0}'_{n_{20}}, \mathbf{0}'_{n_{21}}, \mathbf{1}'_{n_{22}})'$. Then, we can obtain three modified F-test statistics under the recessive, additive, and dominant models for Stage 2 similarly, and denote them by F_2^R, F_2^A, and F_2^D. Let $\mathbf{Y} = (\mathbf{Y}'_1, \mathbf{Y}'_2)'$, $\mathbf{G}_R = (\mathbf{G}'_{1R}, \mathbf{G}'_{2R})'$, $\mathbf{G}_A = (\mathbf{G}'_{1A}, \mathbf{G}'_{2A})'$, and $\mathbf{G}_D = (\mathbf{G}'_{1D}, \mathbf{G}'_{2D})'$. Denote $N_0 = n_{10} + n_{20}$, $N_1 = n_{11} + n_{21}$, and $N_2 = n_{12} + n_{22}$ for the combined sample sizes from two stages, corresponding to three genotypes. Then the proposed F-test statistics under three genetic models on the basis of the combined data are as follows:

$$\begin{aligned}
F_J^R &= \frac{\mathbf{Y}'\left[\mathbf{X}_R\left(\mathbf{X}'_R\mathbf{X}_R\right)^{-1}\mathbf{X}'_R - \mathbf{1}_n\left(\mathbf{1}'_n\mathbf{1}_n\right)^{-1}\mathbf{1}'_n\right]\mathbf{Y}}{\mathbf{Y}'\left[\mathbf{I}_n - \mathbf{W}\left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\right]\mathbf{Y}/(n-6)} \\
&= \frac{\left(Z_J^R\right)^2}{\mathrm{RSS}_J/(n-6)}, \\
F_J^A &= \frac{\mathbf{Y}'\left[\mathbf{X}_A\left(\mathbf{X}'_A\mathbf{X}_A\right)^{-1}\mathbf{X}'_A - \mathbf{1}_n\left(\mathbf{1}'_n\mathbf{1}_n\right)^{-1}\mathbf{1}'_n\right]\mathbf{Y}}{\mathbf{Y}'\left[\mathbf{I}_n - \mathbf{W}\left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\right]\mathbf{Y}/(n-6)} \\
&= \frac{\left(Z_J^A\right)^2}{\mathrm{RSS}_J/(n-6)}, \\
F_J^D &= \frac{\mathbf{Y}'\left[\mathbf{X}_D\left(\mathbf{X}'_D\mathbf{X}_D\right)^{-1}\mathbf{X}'_D - \mathbf{1}_n\left(\mathbf{1}'_n\mathbf{1}_n\right)^{-1}\mathbf{1}'_n\right]\mathbf{Y}}{\mathbf{Y}'\left[\mathbf{I}_n - \mathbf{W}\left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\right]\mathbf{Y}/(n-6)} \\
&= \frac{\left(Z_J^D\right)^2}{\mathrm{RSS}_J/(n-6)},
\end{aligned} \tag{6}$$

where $\mathbf{X}_R = (\mathbf{1}_n, \mathbf{G}_R)$, $\mathbf{X}_A = (\mathbf{1}_n, \mathbf{G}_A)$, $\mathbf{X}_D = (\mathbf{1}_n, \mathbf{G}_D)$, and $\mathbf{W} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{o}_{n_1\times 3} \\ \mathbf{o}_{n_2\times 3} & \mathbf{X}_2 \end{pmatrix}$,

$$\begin{aligned}
Z_J^R &= \sqrt{\frac{(N_0+N_1)N_2}{n}}\left(\overline{y}_{01} - \overline{y}_2\right), \\
Z_J^D &= \sqrt{\frac{N_0(N_1+N_2)}{n}}\left(\overline{y}_0 - \overline{y}_{12}\right), \\
Z_J^A &= \left(N_0(N_1+2N_2)\overline{y}_0 - N_1(N_0-N_2)\overline{y}_1\right. \\
&\quad \left. -N_2(2N_0+N_1)\overline{y}_2\right) \\
&\quad \times\left(\sqrt{n\left[N_0(N_1+4N_2)+N_1N_2\right]}\right)^{-1}, \\
\overline{y}_0 &= \frac{1}{N_0}\left(\sum_{j=1}^{n_{10}} y_j + \sum_{j=n_1+1}^{n_1+n_{20}} y_j\right), \\
\overline{y}_1 &= \frac{1}{N_1}\left(\sum_{j=n_{10}+1}^{n_{10}+n_{11}} y_j + \sum_{j=n_1+n_{20}+1}^{n_1+n_{20}+n_{21}} y_j\right), \\
\overline{y}_2 &= \frac{1}{N_2}\left(\sum_{j=n_{10}+n_{11}+1}^{n_1} y_j + \sum_{j=n_1+n_{20}+n_{21}+1}^{n} y_j\right), \\
\overline{y}_{01} &= \frac{N_0\overline{y}_0 + N_1\overline{y}_1}{N_0+N_1}, \\
\overline{y}_{12} &= \frac{N_1\overline{y}_1 + N_2\overline{y}_2}{N_1+N_2}.
\end{aligned} \tag{7}$$

Furthermore, we propose the joint testing statistic as

$$F_J^{\mathrm{MAX}} = \max\left\{F_J^R, F_J^A, F_J^D\right\}. \tag{8}$$

In order to calculate the power of the proposed joint analysis, we have to get the thresholds, which is determined by the significance level. Denote the threshold for choosing the promising SNPs in Stage 1 by u_1, which is the solution of

$$\Pr_{H_0}\left(F_1^{\mathrm{MAX}} > u_1\right) = \gamma. \tag{9}$$

Since the genome-wide significance level is α/m, in order to control the false positive rate, we have

$$\Pr_{H_0}\left(F_1^{\mathrm{MAX}} > u_1, F_J^{\mathrm{MAX}} > u_J\right) = \alpha/m, \tag{10}$$

where u_J is the cut-off point for the joint statistic. Once we have u_1 and u_J, the power is calculated by

$$\Pr_{H_1}\left(F_1^{\mathrm{MAX}} > u_1, F_J^{\mathrm{MAX}} > u_J\right). \tag{11}$$

We now give the detail to calculate the cut-off point and power above. The left side of (10) can be further expressed as

$$\begin{aligned}
&\Pr_{H_0}\left(F_1^{\mathrm{MAX}} > u_1, F_J^{\mathrm{MAX}} > u_J\right) \\
&\quad = 1 - \Pr_{H_0}\left(F_1^{\mathrm{MAX}} \le u_1\right) - \Pr_{H_0}\left(F_J^{\mathrm{MAX}} \le u_J\right) \\
&\qquad + \Pr_{H_0}\left(F_1^{\mathrm{MAX}} \le u_1, F_J^{\mathrm{MAX}} \le u_J\right).
\end{aligned} \tag{12}$$

For controlling the type I error rate and calculating the power, we need to know the distribution or the asymptotic distribution of $(Z_1^R, Z_1^A, Z_1^D, Z_J^R, Z_J^A, Z_J^D, \mathrm{RSS}_1, \mathrm{RRS}_J)'$ under both H_0 and H_1.

Note that whether H_0 or H_1 holds, RSS_1 and RSS_J and $(Z_1^R, Z_1^A, Z_1^D, Z_J^R, Z_J^A, Z_J^D)'$ are mutually independent (the proof is given in Appendix A). Denote the correlation matrix of $(Z_1^R, Z_1^A, Z_1^D)'$ by $\mathbf{V}_1 = (v_{kl})_{3\times 3}$, whose entries are $v_{11} = v_{22} = v_{33} = 1$, $v_{12} = v_{21} = \sqrt{n_{12}}\,(2n_{10} + n_{11})/\sqrt{(n_{10}+n_{11})[n_{10}(n_{11}+4n_{12})+n_{11}n_{12}]}$, $v_{13} = v_{31} = \sqrt{n_{10}n_{12}}/\sqrt{(n_{10}+n_{11})(n_{11}+n_{12})}$, and $v_{23} = v_{32} = \sqrt{n_{10}}\,(2n_{12} + n_{11})/\sqrt{(n_{11}+n_{12})[n_{10}(n_{11}+4n_{12})+n_{11}n_{12}]}$, respectively. Similarly, let $\mathbf{V}_J = (v_{kl}^*)_{3\times 3}$ be the correlation matrix of $(Z_J^R, Z_J^A, Z_J^D)'$ with $v_{11}^* = v_{22}^* = v_{33}^* = 1$, $v_{12}^* = v_{21}^* = \sqrt{N_2}(2N_0 + N_1)/\sqrt{(N_0+N_1)[N_0(N_1+4N_2)+N_1N_2]}$, $v_{13}^* = v_{31}^* = \sqrt{N_0N_2}/\sqrt{(N_0+N_1)(N_1+N_2)}$, and $v_{23}^* = v_{32}^* = \sqrt{N_0}(2N_2+ N_1)/\sqrt{(N_1+N_2)[N_0(N_1+4N_2)+N_1N_2]}$. Then, we can derive that $\mathrm{RRS}_1/\sigma^2 \sim \chi^2_{n_1-3}$, $\mathrm{RSS}_J/\sigma^2 \sim \chi^2_{n-6}$, and

$$\left(Z_1^R, Z_1^A, Z_1^D, Z_J^R, Z_J^A, Z_J^D\right)'\Big|_{H_0} \sim N_6\left(\mathbf{0}_6, \sigma^2\begin{pmatrix}\mathbf{V}_1 & \boldsymbol{\rho}\\ \boldsymbol{\rho}' & \mathbf{V}_J\end{pmatrix}\right), \tag{13}$$

where $\boldsymbol{\rho} = (\rho_{kl})_{3\times 3}$ is the correlation matrix between $(Z_1^R, Z_1^A, Z_1^D)'$ and $(Z_J^R, Z_J^A, Z_J^D)'$, with

$$\begin{aligned}
\rho_{11} &= \mathrm{Corr}\left(Z_1^R, Z_J^R\right) = \sqrt{\frac{n\left(n_{10}+n_{11}\right)n_{12}}{n_1\left(N_0+N_1\right)N_2}},\\
\rho_{12} &= \mathrm{Corr}\left(Z_1^R, Z_J^A\right)\\
&= \frac{\sqrt{nn_{12}}\left(2n_{10}+n_{11}\right)}{\sqrt{n_1\left(n_{10}+n_{11}\right)\left[N_0\left(N_1+4N_2\right)+N_1N_2\right]}},\\
\rho_{13} &= \mathrm{Corr}\left(Z_1^R, Z_J^D\right) = \frac{\sqrt{nn_{12}}n_{10}}{\sqrt{n_1\left(n_{10}+n_{11}\right)N_0\left(N_1+N_2\right)}},\\
\rho_{21} &= \mathrm{Corr}\left(Z_1^A, Z_J^R\right)\\
&= \frac{\sqrt{n}n_{12}\left(2n_{10}+n_{11}\right)}{\sqrt{n_1\left[n_{10}\left(n_{11}+4n_{12}\right)+n_{11}n_{12}\right]\left(N_0+N_1\right)N_2}},\\
\rho_{22} &= \mathrm{Corr}\left(Z_1^A, Z_J^A\right) = \sqrt{\frac{n\left[n_{10}\left(n_{11}+4n_{12}\right)+n_{11}n_{12}\right]}{n_1\left[N_0\left(N_1+4N_2\right)+N_1N_2\right]}},\\
\rho_{23} &= \mathrm{Corr}\left(Z_1^A, Z_J^D\right)\\
&= \frac{\sqrt{n}n_{10}\left(n_{11}+2n_{12}\right)}{\sqrt{n_1\left[n_{10}\left(n_{11}+4n_{12}\right)+n_{11}n_{12}\right]N_0\left(N_1+N_2\right)}},\\
\rho_{31} &= \mathrm{Corr}\left(Z_1^D, Z_J^R\right) = \frac{\sqrt{nn_{10}}n_{12}}{\sqrt{n_1\left(n_{11}+n_{12}\right)\left(N_0+N_1\right)N_2}},\\
\rho_{32} &= \mathrm{Corr}\left(Z_1^D, Z_J^A\right)\\
&= \frac{\sqrt{nn_{10}}\left(n_{11}+2n_{12}\right)}{\sqrt{n_1\left(n_{11}+n_{12}\right)\left[N_0\left(N_1+4N_2\right)+N_1N_2\right]}},\\
\rho_{33} &= \mathrm{Corr}\left(Z_1^D, Z_J^D\right) = \sqrt{\frac{nn_{10}\left(n_{11}+n_{12}\right)}{n_1N_0\left(N_1+N_2\right)}}.
\end{aligned} \tag{14}$$

Under H_1, for a given odds ratio $\mathrm{OR} = \exp(\beta_1)$ for subjects with two copies of risk allele corresponding to recessive model or one copy of risk allele corresponding to additive or dominant models, we have the following:

(i) when the true genetic model is recessive,

$$\left(Z_1^R, Z_1^A, Z_1^D, Z_J^R, Z_J^A, Z_J^D\right)'\Big|_{H_1} \sim N_6\left(\boldsymbol{\mu}^R, \sigma^2\begin{pmatrix}\mathbf{V}_1 & \boldsymbol{\rho}\\ \boldsymbol{\rho}' & \mathbf{V}_J\end{pmatrix}\right), \tag{15}$$

where $\boldsymbol{\mu}^R = (\mu_1^{RR}, \mu_1^{RA}, \mu_1^{RD}, \mu_J^{RR}, \mu_J^{RA}, \mu_J^{RD})'$ with

$$\begin{aligned}
\mu_1^{RR} &= -\sqrt{\frac{\left(n_{10}+n_{11}\right)n_{12}}{n_1}}\beta_1,\\
\mu_1^{RA} &= \frac{-n_{12}\left(2n_{10}+n_{11}\right)\beta_1}{\sqrt{n_1\left[n_{10}\left(n_{11}+4n_{12}\right)+n_{11}n_{12}\right]}},\\
\mu_1^{RD} &= \frac{-\sqrt{n_{10}}n_{12}\beta_1}{\sqrt{n_1\left(n_{11}+n_{12}\right)}},\\
\mu_J^{RR} &= -\sqrt{\frac{\left(N_0+N_1\right)N_2}{n}}\beta_1,\\
\mu_J^{RA} &= \frac{-N_2\left(2N_0+N_1\right)\beta_1}{\sqrt{n\left[N_0\left(N_1+4N_2\right)+N_1N_2\right]}},\\
\mu_J^{RD} &= \frac{-\sqrt{N_0}N_2\beta_1}{\sqrt{n\left(N_1+N_2\right)}},
\end{aligned} \tag{16}$$

(ii) when the true genetic model is additive,

$$\left(Z_1^R, Z_1^A, Z_1^D, Z_J^R, Z_J^A, Z_J^D\right)'\Big|_{H_1} \sim N_6\left(\boldsymbol{\mu}^A, \sigma^2\begin{pmatrix}\mathbf{V}_1 & \boldsymbol{\rho}\\ \boldsymbol{\rho}' & \mathbf{V}_J\end{pmatrix}\right), \tag{17}$$

where $\boldsymbol{\mu}^A = (\mu_1^{AR}, \mu_1^{AA}, \mu_1^{AD}, \mu_J^{AR}, \mu_J^{AA}, \mu_J^{AD})'$ with

$$\begin{aligned}
\mu_1^{AR} &= \frac{-\sqrt{n_{12}}\left(n_{11}+2n_{10}\right)\beta_1}{\sqrt{n_1\left(n_{10}+n_{11}\right)}},\\
\mu_1^{AA} &= -\sqrt{\frac{n_{10}\left(n_{11}+4n_{12}\right)+n_{11}n_{12}}{n_1}}\beta_1,\\
\mu_1^{AD} &= \frac{-\sqrt{n_{10}}\left(n_{11}+2n_{12}\right)\beta_1}{\sqrt{n_1\left(n_{11}+n_{12}\right)}},\\
\mu_J^{AR} &= \frac{-\sqrt{N_2}\left(N_1+2N_0\right)\beta_1}{\sqrt{n\left(N_0+N_1\right)}},
\end{aligned}$$

$$\mu_J^{AA} = -\sqrt{\frac{N_0\left(N_1+4N_2\right)+N_1N_2}{n}}\beta_1,$$
$$\mu_J^{AD} = \frac{-\sqrt{N_0}\left(N_1+2N_2\right)\beta_1}{\sqrt{n\left(N_1+N_2\right)}},$$
(18)

(iii) when the true genetic model is dominant,

$$\left(Z_1^R, Z_1^A, Z_1^D, Z_J^R, Z_J^A, Z_J^D\right)'\Big|_{H_1} \sim N_6\left(\boldsymbol{\mu}^D, \sigma^2\begin{pmatrix}\mathbf{V}_1 & \boldsymbol{\rho}\\ \boldsymbol{\rho}' & \mathbf{V}_J\end{pmatrix}\right),$$
(19)

where $\boldsymbol{\mu}^D = (\mu_1^{DR}, \mu_1^{DA}, \mu_1^{DD}, \mu_J^{DR}, \mu_J^{DA}, \mu_J^{DD})'$ with

$$\mu_1^{DR} = \frac{-\sqrt{n_{12}}n_{10}\beta_1}{\sqrt{n_1\left(n_{10}+n_{11}\right)}},$$
$$\mu_1^{DA} = \frac{-n_{10}\left(n_{11}+2n_{12}\right)\beta_1}{\sqrt{n_1\left[n_{10}\left(n_{11}+4n_{12}\right)+n_{11}n_{12}\right]}},$$
$$\mu_1^{DD} = -\sqrt{\frac{n_{10}\left(n_{11}+n_{12}\right)}{n_1}}\beta_1,$$
$$\mu_J^{DR} = \frac{-\sqrt{N_2}N_0\beta_1}{\sqrt{n\left(N_0+N_1\right)}},$$
$$\mu_J^{DA} = \frac{-N_0\left(N_1+2N_2\right)\beta_1}{\sqrt{n\left[N_0\left(N_1+4N_2\right)+N_1N_2\right]}},$$
$$\mu_J^{DD} = -\sqrt{\frac{N_0\left(N_1+N_2\right)}{n}}\beta_1.$$
(20)

We develop a method for simplifying the calculations of $\Pr_{H_0}(F_1^{\mathrm{MAX}} \le u_1)$ and $\Pr_{H_0}(F_J^{\mathrm{MAX}} \le u_J)$ and $\Pr_{H_0}(F_1^{\mathrm{MAX}} \le u_1, F_J^{\mathrm{MAX}} \le u_J)$. The details are included in Appendix B, and the calculations of $\Pr_{H_1}(F_1^{\mathrm{MAX}} \le u_1)$ and $\Pr_{H_1}(F_J^{\mathrm{MAX}} \le u_J)$ and $\Pr_{H_1}(F_1^{\mathrm{MAX}} \le u_1, F_J^{\mathrm{MAX}} \le u_J)$ are essentially similar.

3. Results

3.1. Power Comparison. We conduct simulation studies to evaluate the performance of the proposed method under three commonly used genetic models (recessive, additive, and dominant models). We mainly compare the power of two approaches; one is the proposed method in this paper, and the other is the joint analysis based on the F-test statistics F_1^A and F_J^A. For convenience, we refer to the proposed method as MAXFJ and AFJ for the other one. We choose the sample size $n = 2000$, and $m = 5 \times 10^5$. The proportion of subjects genotyped in Stage 1 has three levels $\pi = 0.3, 0.4, 0.5$. We set the genome-wide significance level as $\alpha = 0.05$ and that the significance level per SNP as $\alpha/m = 1 \times 10^{-7}$. In Stage 1, the P-value threshold for SNPs selected for followup is set to be 1×10^{-4} and 2×10^{-4}. We assume that the Hardy-Weinberg equilibrium holds in the general sample population, and then there are on average $n \times (1 - \mathrm{MAF})^2$, $2n \times \mathrm{MAF} \times (1 - \mathrm{MAF})$, and $n \times \mathrm{MAF}^2$ individuals with genotype gg, Gg, and GG, respectively, where the minor allele frequency is set to be 0.15, 0.30 and 0.45. To make the power comparison more distinctly, we specify different genetic effect parameters β_1 under three genetic models as follows: $\beta_1 = 0.5$ for the recessive model, $\beta_1 = 0.3$ for the additive model, and $\beta_1 = 0.4$ for the dominant model.

Table 1: Power comparison ($n = 2000$, $\gamma = 1 \times 10^{-4}$, $\alpha = 0.05$, and $m = 5 \times 10^5$).

π	MAF	REC		ADD		DOM	
		AFJ	MAXFJ	AFJ	MAXFJ	AFJ	MAXFJ
0.30	0.15	$7.5e-5$	0.005	0.426	0.365	0.610	0.618
	0.30	0.052	0.285	0.811	0.759	0.698	0.784
	0.45	0.487	0.785	0.893	0.854	0.449	0.647
0.40	0.15	$1.1e-4$	0.009	0.651	0.589	0.826	0.837
	0.30	0.086	0.470	0.945	0.922	0.887	0.938
	0.45	0.711	0.938	0.979	0.968	0.677	0.859
0.50	0.15	$1.0e-4$	0.010	0.802	0.751	0.933	0.941
	0.30	0.121	0.639	0.987	0.980	0.965	0.986
	0.45	0.856	0.987	0.997	0.995	0.826	0.953

Table 2: Power comparison ($n = 2000$, $\gamma = 2 \times 10^{-4}$, $\alpha = 0.05$, and $m = 5 \times 10^5$).

π	MAF	REC		ADD		DOM	
		AFJ	MAXFJ	AFJ	MAXFJ	AFJ	MAXFJ
0.30	0.15	$1.3e-4$	0.006	0.489	0.426	0.676	0.681
	0.30	0.066	0.340	0.852	0.806	0.754	0.828
	0.45	0.556	0.833	0.922	0.891	0.516	0.706
0.40	0.15	$1.2e-4$	0.011	0.709	0.651	0.866	0.876
	0.30	0.101	0.529	0.961	0.943	0.916	0.956
	0.45	0.765	0.957	0.987	0.979	0.732	0.892
0.50	0.15	$1.7e-4$	0.012	0.838	0.793	0.951	0.958
	0.30	0.133	0.683	0.992	0.987	0.975	0.991
	0.45	0.888	0.992	0.998	0.997	0.860	0.967

The power results are displayed in Tables 1 and 2 for $\gamma = 1 \times 10^{-4}$ and $\gamma = 2 \times 10^{-4}$, respectively. They indicate that MAXFJ is more efficiency robust than AFJ across various inheritance models. As expected, AFJ is more powerful than MAXFJ under the additive model. However, MAFJ performs much more powerful than AFJ when the true genetic model is recessive. For instance, in Table 2, with $\pi = 0.4$ and MAF = 0.3, the powers of AFJ and MAXFJ are 0.101 and 0.529, respectively. In summary, MAXFJ is substantially more powerful than AFJ in two-staged GWAS of quantitative traits, when the model for AFJ is misspecified.

3.2. An Illustration Example: Rheumatoid Arthritis. Rheumatoid arthritis (RA) is an autoimmune disease (resulting in a chronically systemic inflammatory disorder) which mainly attacks synovial joints. About 1% of the common adult population worldwide is affected by RA [16]. It has been

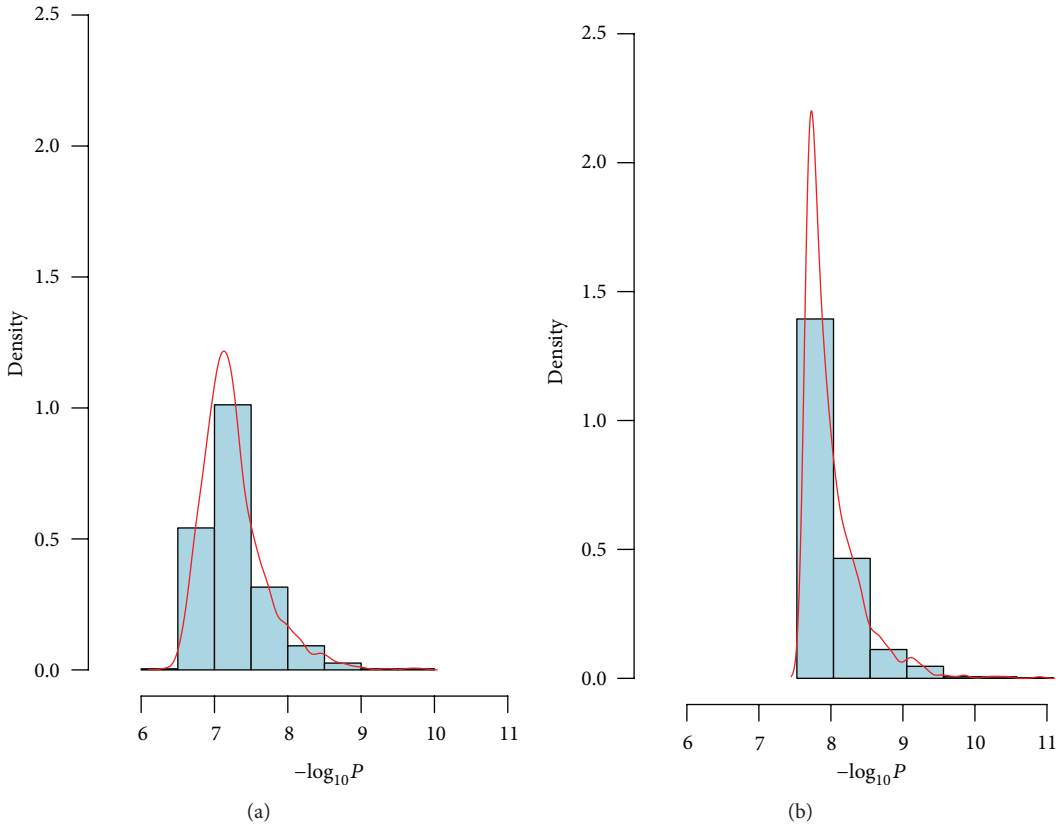


FIGURE 1: The histogram and density of $-\log_{10}P$ when $\pi = 0.3$ (the left subgraph corresponds to MAXFJ while the right one for AFJ).

pointed out that the genetic variants might play a major role in RA susceptibility [17]. Genetic Analysis Workshop 16 (GAW16) based on the North American Rheumatoid Arthritis Consortium (NARAC) is a GWAS testing association with RA using about 5×10^5 SNPs [18–20]. It included 868 individuals who were RA positive (cases) and also had continuous trait anticyclic citrullinated peptide (anti-CCP) measures and 1194 controls sampled from the New York Cancer Project (NYCP) without RA which had no anti-CCP measures. Huizinga et al. [21] pointed out that a greater anti-CCP would be linked to better prediction of increased risk developing RA. Chen et al. [22] showed that SNP rs2476601 located in PTPN22 had the most significant association with RA. Here, we only focus on SNP rs2476601 and apply two joint analysis methods (AFJ and MAXFJ) to evaluate its statistical significance. The minimum of anti-CCP among 868 cases was affected to each control, and a log transformation of anti-CCP was applied in the analysis. Then, we considered $\pi = 0.3, 0.4, 0.5$ three simulation circumstances. For $\pi = 0.3$, thirty percent of individuals were randomly sampled from all cases and controls and were used as the data from Stage 1, and the rest of individuals were treated as the data of Stage 2. The P-values of AFJ and MAXFJ were calculated, respectively. We repeated the above procedure 1,000 times and saved the corresponding P-values. A base-10 logarithm transformation and an opposite transformation were successively applied to these P-values, and the histogram and density of these transformed data were obtained (Figure 1). Similarly, we conducted the simulation and calculation for $\pi = 0.4$ and 0.5, and the corresponding histogram and density were presented in Figures 2 and 3. Examination of Figures 1–3 showed that the P-values of MAXFJ are more stable than those of AFJ and the estimated density curves of MAXFJ are more closer to the symmetrical normal distribution while the estimated density curves of AFJ are rather skewed, which indicated that MAXFJ possesses more robust performance when the real genetic models are unknown.

4. Discussion

We have developed a feasible two-stage design and the corresponding robust joint analysis approach for quantitative trait GWASs. The method is based on the F-statistics over three different genetic models. The denominator of the used F-statistic, which is constructed without assuming any genetic model, is different from the commonly used one. This adoption can reduce the computation intensity. Taking advantage of an ingenious design matrix, we successfully construct the common denominator of three F-test statistics for the joint analysis with combined raw data from both stages. The statistical significance (P-value) for the proposed joint analysis method can be calculated with the derived analytic expressions on the basis of the asymptotic distributions, which greatly reduce the complexity and computational intensity compared with the resampling-type permutation and bootstrap procedures. Our numerical results demonstrate

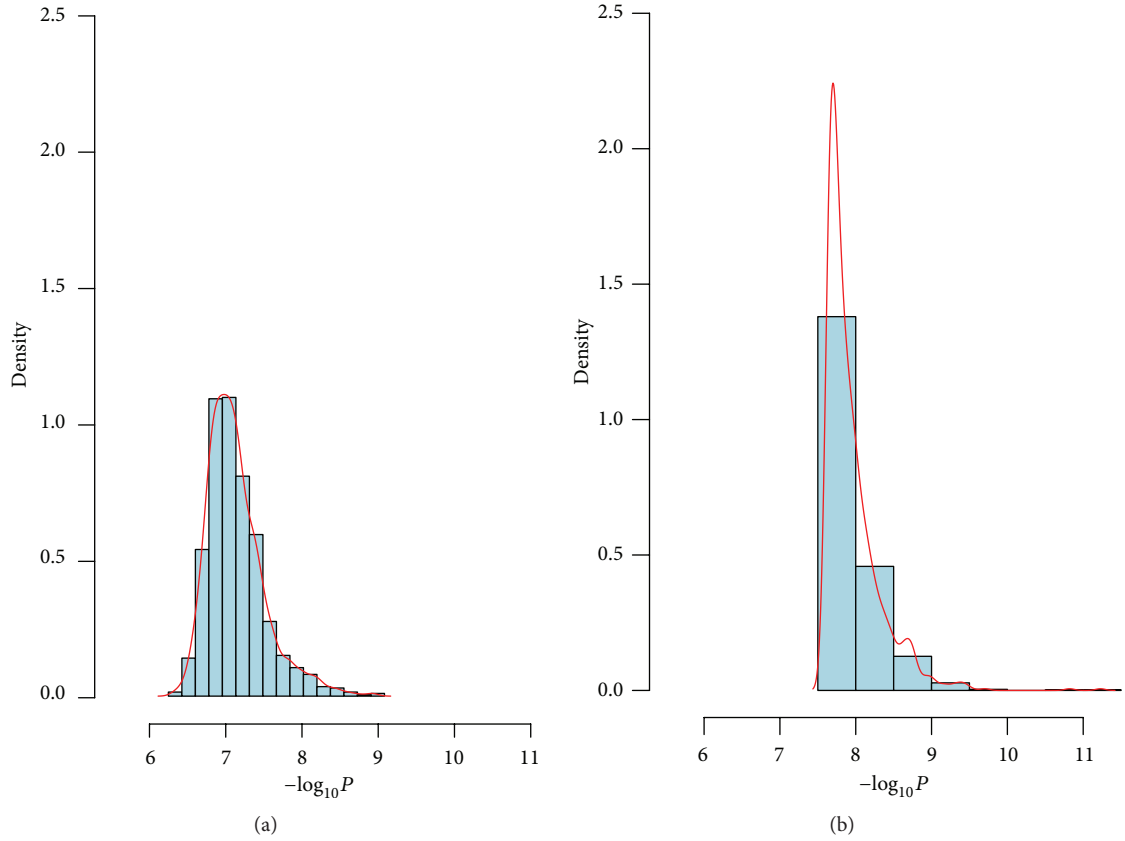


FIGURE 2: The histogram and density of $-\log_{10}P$ when $\pi = 0.4$ (the left subgraph corresponds to MAXFJ while the right one for AFJ).

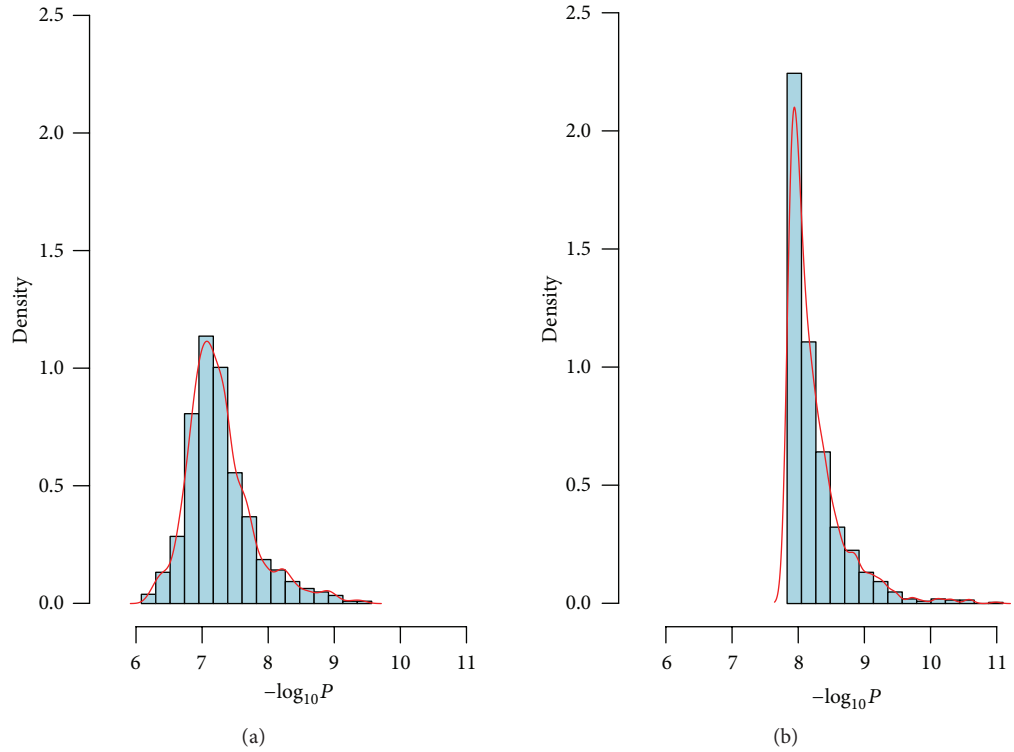


FIGURE 3: The histogram and density of $-\log_{10}P$ when $\pi = 0.5$ (the left subgraph corresponds to MAXFJ while the right one for AFJ).

that this novel approach has the greater efficiency robustness for genetic model uncertainty than the F-statistic-based joint analysis which assumes the additive genetic model.

In this work, we did not investigate the power of joint analysis based on other existing robust association methods for quantitative traits such as So and Sham's method. We find that it is very difficult to extend So and Sham's method (score test-based MAX3) to two-staged GWASs with quantitative outcomes, since it is almost impossible to derive the joint distribution of score tests from two stages.

For simplicity, here we do not take into account the effects of covariates in the considered two-stage design. However, in real application, the proposed method can be easily applied to the situation including one or more covariates as shown by the original MAXF by Li et al. [10]. It is important to stress that we combine the raw data from two stages to construct the joint statistic, unlike the joint analysis for binary traits using the weighted sum of two statistics in Stages 1 and 2 [12]. Furthermore, one basic assumption in this paper is that the effect sizes of genetic variants between two stages are identical (i.e., no heterogeneity exists), which is the natural and reasonable precondition for the data fusion strategy. In addition, the population-based genetic association studies may be affected by the population stratification, and this needs future research to examine it.

Appendix

A. The Derivation of the Asymptotic Properties of F_J^R, F_J^A, F_J^D

Consider the linear model for the combined raw data from Stage 1 and Stage 2 as follows:

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\vartheta} + \boldsymbol{\varepsilon}, \tag{A.1}$$

where $\boldsymbol{\vartheta} = (\beta_0, \zeta_1, \zeta_2, \beta_0^*, \zeta_1^*, \zeta_2^*)'$ and $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$.

Denote

$$\mathbf{C}_R = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{C}_D = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{C}_A = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 2 & -1 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{C}_0 = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}. \tag{A.2}$$

Based on the design matrix

$$\mathbf{W} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{O}_{n_1\times 3} \\ \mathbf{O}_{n_2\times 3} & \mathbf{X}_2 \end{pmatrix} \tag{A.3}$$

for the expanded full model above, we can get the ordinary least square estimator of $\boldsymbol{\vartheta}$ by $\widehat{\boldsymbol{\vartheta}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}$, and the residual sum of square is given by $\mathrm{RSS}_J = \mathbf{Y}'[\mathbf{I}_n - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}']\mathbf{Y}$. Furthermore, we denote the residual sum of squares under the following constraints: $\mathbf{C}_R\boldsymbol{\vartheta} = \mathbf{0}_4$, $\mathbf{C}_D\boldsymbol{\vartheta} = \mathbf{0}_4$, $\mathbf{C}_A\boldsymbol{\vartheta} = \mathbf{0}_4$, and $\mathbf{C}_0\boldsymbol{\vartheta} = \mathbf{0}_5$, by RSS_R, RSS_D, RSS_A, and RSS_0, respectively. After some algebras, we can obtain

$$F_J^R = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_R}{\mathrm{RSS}_J/(n-6)} = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_J) - (\mathrm{RSS}_R - \mathrm{RSS}_J)}{\mathrm{RSS}_J/(n-6)},$$

$$F_J^A = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_A}{\mathrm{RSS}_J/(n-6)} = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_J) - (\mathrm{RSS}_A - \mathrm{RSS}_J)}{\mathrm{RSS}_J/(n-6)},$$

$$F_J^D = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_D}{\mathrm{RSS}_J/(n-6)} = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_J) - (\mathrm{RSS}_D - \mathrm{RSS}_J)}{\mathrm{RSS}_J/(n-6)}. \tag{A.4}$$

On the one hand, according to

$$\mathbf{W}'\mathbf{W} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{O}_{3\times 3} \\ \mathbf{O}_{3\times 3} & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}, \tag{A.5}$$

it follows that

$$\left(\mathbf{W}'\mathbf{W}\right)^{-1} = \begin{pmatrix} \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1} & \mathbf{O}_{3\times 3} \\ \mathbf{O}_{3\times 3} & \left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1} \end{pmatrix},$$

$$\mathbf{W}\left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}' = \begin{pmatrix} \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1' & \mathbf{O}_{n_1\times n_2} \\ \mathbf{O}_{n_2\times n_1} & \mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2' \end{pmatrix}. \tag{A.6}$$

So, we can get that

$$\mathbf{Y}'\left[\mathbf{I}_n - \mathbf{W}\left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\right]\mathbf{Y} = \mathbf{Y}_1'\left[\mathbf{I}_{n_1} - \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right]\mathbf{Y}_1 + \mathbf{Y}_2'\left[\mathbf{I}_{n_2} - \mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\right]\mathbf{Y}_2. \tag{A.7}$$

That is, $\mathrm{RSS}_J = \mathrm{RSS}_1 + \mathrm{RSS}_2$. Furthermore, $\mathrm{RSS}_J/(n-6)$ is also the unbiased estimator of the variance of the residual σ^2 based on the independence and unbiasedness of RSS_1 and RSS_2.

On the other hand, $\mathrm{RSS}_0 - \mathrm{RSS}_J$ and $\mathrm{RSS}_R - \mathrm{RSS}_J$ are both independent of RSS_J, so $(Z_J^R)^2 = \mathrm{RSS}_0 - \mathrm{RSS}_R = (\mathrm{RSS}_0 - \mathrm{RSS}_J) - (\mathrm{RSS}_R - \mathrm{RSS}_J)$ is also independent of RSS_J, and $(\mathrm{RSS}_0 - \mathrm{RSS}_R)/\sigma^2 \sim \chi_1^2$, $\mathrm{RSS}_J/\sigma^2 \sim \chi_{n-6}^2$. Consequently, we have $F_J^R \sim F_{1,n-6}$. Similarly, we can get $F_J^A \sim F_{1,n-6}$ and $F_J^D \sim F_{1,n-6}$.

B. The Detailed Calculation of $\Pr_{H_0}(F_1^{\mathrm{MAX}} \le u_1)$ and $\Pr_{H_0}(F_J^{\mathrm{MAX}} \le u_J)$ and $\Pr_{H_0}(F_1^{\mathrm{MAX}} \le u_1, F_J^{\mathrm{MAX}} \le u_J)$

Denote $w_0 = \sqrt{(n_{10}+n_{11})n_{12}/(n_{10}(n_{11}+4n_{12})+n_{11}n_{12})}$ and $w_1 = \sqrt{n_{10}(n_{11}+n_{12})/(n_{10}(n_{11}+4n_{12})+n_{11}n_{12})}$. For a given $c > 0$,

$$\Pr_{H_0}\left(\left|\frac{Z_1^R}{\sigma}\right| \le c, \left|\frac{Z_1^A}{\sigma}\right| \le c, \left|\frac{Z_1^D}{\sigma}\right| \le c\right) = \oint_{\Omega_0} f(z_0, z_1; \boldsymbol{\Sigma}_0)\, dz_0 dz_1, \tag{B.1}$$

where $\Omega_0 = \{(z_0, z_1) : |z_0| \le c, |w_0 z_0 + w_1 z_1| \le c, |z_1| \le c\}$ and $f(z_0, z_1; \boldsymbol{\Sigma}_0)$ is the bivariate normal density function for $(Z_1^R/\sigma, Z_1^D/\sigma)'$ with mean $\mathbf{0}_2$ and variance-covariance matrix $\boldsymbol{\Sigma}_0 = \begin{pmatrix} 1 & \nu_{13} \\ \nu_{13} & 1 \end{pmatrix}$. Taking advantage of the symmetry of bivariate normal distribution, the above twofold integration can be only calculated at the right half space of Ω_0 and then multiplied by 2, which is

$$\begin{aligned}&\oint_{\Omega_0} f(z_0, z_1; \boldsymbol{\Sigma}_0)\, dz_0 dz_1 \\ &= 2\left[\int_0^{c(1-w_1)/w_0} dz_0 \int_{-c}^{c} f(z_0, z_1; \boldsymbol{\Sigma}_0)\, dz_1 \right. \\ &\quad \left. + \int_{c(1-w_1)/w_0}^{c} dz_0 \int_{-c}^{(c-w_0 z_0)/w_1} f(z_0, z_1; \boldsymbol{\Sigma}_0)\, dz_1\right].\end{aligned} \tag{B.2}$$

Based on the property of conditional distributions of the multivariate normal distribution, we have

$$f(z_0, z_1; \boldsymbol{\Sigma}_0) = \phi(z_0)\, f(z_1 \mid z_0; \nu_{13}), \tag{B.3}$$

where $\phi(z_0)$ is the probability density function of $N(0,1)$ and $f(z_1 \mid z_0; \nu_{13})$ is the density function of the conditional normal distribution $N(\nu_{13} z_0, 1-\nu_{13}^2)$. That is,

$$f(z_1 \mid z_0; \nu_{13}) = \frac{1}{\sqrt{1-\nu_{13}^2}} \phi\left(\frac{z_1 - \nu_{13} z_0}{\sqrt{1-\nu_{13}^2}}\right). \tag{B.4}$$

Then, it follows that

$$\begin{aligned}&\int_0^{c(1-w_1)/w_0} dz_0 \int_{-c}^{c} f(z_0, z_1; \boldsymbol{\Sigma}_0)\, dz_1 \\ &\quad = \int_0^{c(1-w_1)/w_0} \left[\Phi\left(\frac{c - \nu_{13} z_0}{\sqrt{1-\nu_{13}^2}}\right) - \Phi\left(\frac{-c - \nu_{13} z_0}{\sqrt{1-\nu_{13}^2}}\right)\right] \\ &\qquad \times \phi(z_0)\, dz_0, \\ &\int_{c(1-w_1)/w_0}^{c} dz_0 \int_{-c}^{(c-w_0 z_0)/w_1} f(z_0, z_1; \boldsymbol{\Sigma}_0)\, dz_1 \\ &\quad = \int_{c(1-w_1)/w_0}^{c} \left[\Phi\left(\frac{(c - w_0 z_0)/w_1 - \nu_{13} z_0}{\sqrt{1-\nu_{13}^2}}\right)\right. \\ &\qquad \left. - \Phi\left(\frac{-c - \nu_{13} z_0}{\sqrt{1-\nu_{13}^2}}\right)\right] \phi(z_0)\, dz_0.\end{aligned} \tag{B.5}$$

Thus,

$$\begin{aligned}&\oint_{\Omega_0} f(z_0, z_1; \boldsymbol{\Sigma}_0)\, dz_0 dz_1 \\ &= 2\left[\int_0^{c(1-w_1)/w_0} \Phi\left(\frac{c - \nu_{13} z_0}{\sqrt{1-\nu_{13}^2}}\right) \phi(z_0)\, dz_0 \right. \\ &\quad + \int_{c(1-w_1)/w_0}^{c} \Phi\left(\frac{(c - w_0 z_0)/w_2 - \nu_{13} z_0}{\sqrt{1-\nu_{13}^2}}\right) \\ &\qquad \times \phi(z_0)\, dz_0 \\ &\quad \left. - \int_0^{c} \Phi\left(\frac{-c - \nu_{13} z_0}{\sqrt{1-\nu_{13}^2}}\right) \phi(z_0)\, dz_0\right].\end{aligned} \tag{B.6}$$

Denote $w_0^* = \sqrt{(N_0+N_1)N_2/(N_0(N_1+4N_2)+N_1N_2)}$ and $w_1^* = \sqrt{N_0(N_1+N_2)/(N_0(N_1+4N_2)+N_1N_2)}$. For any given $c_1, c_2 > 0$,

$$\begin{aligned}&\Pr_{H_0}\left(\left|\frac{Z_1^R}{\sigma}\right| \le c_1, \left|\frac{Z_1^A}{\sigma}\right| \le c_1, \left|\frac{Z_1^D}{\sigma}\right| \le c_1, \right. \\ &\qquad \left. \left|\frac{Z_J^R}{\sigma}\right| \le c_2, \left|\frac{Z_J^A}{\sigma}\right| \le c_2, \left|\frac{Z_J^D}{\sigma}\right| \le c_2\right) \\ &\quad = \oint_{\Omega_1} f(z_0, z_1, z_2, z_3; \boldsymbol{\Sigma}_1)\, dz_0 dz_1 dz_2 dz_3,\end{aligned} \tag{B.7}$$

where

$$\Omega_1 = \{(z_0, z_1, z_2, z_3) : |z_0| \le c_1, |w_0 z_0 + w_1 z_1| \le c_1, \\ |z_1| \le c_1, |z_2| \le c_2, |w_0^* z_2 + w_1^* z_3| \le c_2, |z_3| \le c_2\} \tag{B.8}$$

and $f(z_0, z_1, z_2, z_3; \mathbf{\Sigma}_1)$ is the multivariate normal density function for $(Z_1^R/\sigma, Z_1^D/\sigma, Z_J^R/\sigma, Z_J^D/\sigma)'$ with mean $\mathbf{0}_4$ and variance-covariance matrix

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 1 & \nu_{13} & \rho_{11} & \rho_{13} \\ \nu_{13} & 1 & \rho_{31} & \rho_{33} \\ \rho_{11} & \rho_{31} & 1 & \nu_{13}^* \\ \rho_{13} & \rho_{33} & \nu_{13}^* & 1 \end{pmatrix}. \tag{B.9}$$

Note that $\oint_{\Omega_1} f(z_0, z_1, z_2, z_3; \mathbf{\Sigma}_1) dz_0 dz_1 dz_2 dz_3$ is the sum of nine integrations as follows:

$$L_1 = \int_{-c_1(1-w_1)/w_0}^{c_1(1-w_1)/w_0} dz_0 \int_{-c_1}^{c_1} dz_1 \int_{-c_2(1-w_1^*)/w_0^*}^{c_2(1-w_1^*)/w_0^*} dz_2 \int_{-c_2}^{c_2} f(z_0, z_1, z_2, z_3; \mathbf{\Sigma}_1)\, dz_3,$$

$$L_2 = \int_{-c_1(1-w_1)/w_0}^{c_1(1-w_1)/w_0} dz_0 \int_{-c_1}^{c_1} dz_1 \int_{-c_2}^{-c_2(1-w_1^*)/w_0^*} dz_2 \int_{-(c_2+w_0^* z_2)/w_1^*}^{c_2} f(z_0, z_1, z_2, z_3; \mathbf{\Sigma}_1)\, dz_3,$$

$$L_3 = \int_{-c_1(1-w_1)/w_0}^{c_1(1-w_1)/w_0} dz_0 \int_{-c_1}^{c_1} dz_1 \int_{c_2(1-w_1^*)/w_0^*}^{c_2} dz_2 \int_{-c_2}^{(c_2-w_0^* z_2)/w_1^*} f(z_0, z_1, z_2, z_3; \mathbf{\Sigma}_1)\, dz_3,$$

$$L_4 = \int_{c_1(1-w_1)/w_0}^{c_1} dz_0 \int_{-c_1}^{(c_1-w_0 z_0)/w_1} dz_1 \int_{-c_2(1-w_1^*)/w_0^*}^{c_2(1-w_1^*)/w_0^*} dz_2 \int_{-c_2}^{c_2} f(z_0, z_1, z_2, z_3; \mathbf{\Sigma}_1)\, dz_3,$$

$$L_5 = \int_{c_1(1-w_1)/w_0}^{c_1} dz_0 \int_{-c_1}^{(c_1-w_0 z_0)/w_1} dz_1 \int_{-c_2}^{-c_2(1-w_1^*)/w_0^*} dz_2 \int_{-(c_2+w_0^* z_2)/w_1^*}^{c_2} f(z_0, z_1, z_2, z_3; \mathbf{\Sigma}_1)\, dz_3,$$

$$L_6 = \int_{c_1(1-w_1)/w_0}^{c_1} dz_0 \int_{-c_1}^{(c_1-w_0 z_0)/w_1} dz_1 \int_{c_2(1-w_1^*)/w_0^*}^{c_2} dz_2 \int_{-c_2}^{(c_2-w_0^* z_2)/w_1^*} f(z_0, z_1, z_2, z_3; \mathbf{\Sigma}_1)\, dz_3,$$

$$L_7 = \int_{-c_1}^{-c_1(1-w_1)/w_0} dz_0 \int_{-(c_1+w_0 z_0)/w_1}^{c_1} dz_1 \int_{-c_2(1-w_1^*)/w_0^*}^{c_2(1-w_1^*)/w_0^*} dz_2 \int_{-c_2}^{c_2} f(z_0, z_1, z_2, z_3; \mathbf{\Sigma}_1)\, dz_3,$$

$$L_8 = \int_{-c_1}^{-c_1(1-w_1)/w_0} dz_0 \int_{-(c_1+w_0 z_0)/w_1}^{c_1} dz_1 \int_{-c_2}^{-c_2(1-w_1^*)/w_0^*} dz_2 \int_{-(c_2+w_0^* z_2)/w_1^*}^{c_2} f(z_0, z_1, z_2, z_3; \mathbf{\Sigma}_1)\, dz_3,$$

$$L_9 = \int_{-c_1}^{-c_1(1-w_1)/w_0} dz_0 \int_{-(c_1+w_0 z_0)/w_1}^{c_1} dz_1 \int_{c_2(1-w_1^*)/w_0^*}^{c_2} dz_2 \int_{-c_2}^{(c_2-w_0^* z_2)/w_1^*} f(z_0, z_1, z_2, z_3; \mathbf{\Sigma}_1)\, dz_3. \tag{B.10}$$

Moreover, we can obtain that $L_2 = L_3$, $L_4 = L_7$, $L_5 = L_9$, and $L_6 = L_8$ based on the symmetry of the integration domain for (z_0, z_1) and (z_2, z_3), respectively.

We have

$$\begin{aligned} &f(z_0, z_1, z_2, z_3; \mathbf{\Sigma}_1) \\ &\quad = \phi(z_0) f(z_1 \mid z_0; \nu_{13}) \\ &\qquad \times f(z_2 \mid z_0, z_1; \nu_{13}, \rho_{11}, \rho_{31}) \\ &\qquad \times f(z_3 \mid z_0, z_1, z_2; \nu_{13}, \nu_{13}^*, \rho_{11}, \rho_{13}, \rho_{31}, \rho_{33}), \end{aligned} \tag{B.11}$$

where $f(z_2 \mid z_0, z_1; \nu_{13}, \rho_{11}, \rho_{31})$ is the conditional normal density function as

$$\begin{aligned} &f(z_2 \mid z_0, z_1; \nu_{13}, \rho_{11}, \rho_{31}) \\ &\quad = \frac{1}{\sqrt{1 - \left(\left(\rho_{11}^2 - 2\rho_{11}\rho_{31}\nu_{13} + \rho_{31}^2\right) / \left(1 - \nu_{13}^2\right)\right)}} \\ &\qquad \times \phi\Bigg[\Big(\Big(z_2 - \big[z_0 (\rho_{11} - \nu_{13}\rho_{31}) \\ &\qquad\qquad + z_1 (\rho_{31} - \rho_{11}\nu_{13}) \big] \Big) / \left(1 - \nu_{13}^2\right) \Big) \\ &\qquad \times \Big(\sqrt{1 - \left(\left(\rho_{11}^2 + 2\rho_{11}\rho_{31}\nu_{13} + \rho_{31}^2\right) / \left(1 - \nu_{13}^2\right)\right)} \Big)^{-1} \Bigg], \end{aligned} \tag{B.12}$$

and $f(z_3 \mid z_0, z_1, z_2; \nu_{13}, \nu_{13}^*, \rho_{11}, \rho_{13}, \rho_{31}, \rho_{33})$ is the conditional normal density function given by

$$f(z_3 \mid z_0, z_1, z_2; \nu_{13}, \nu_{13}^*, \rho_{11}, \rho_{13}, \rho_{31}, \rho_{33}) = \frac{1}{\sqrt{1-(\rho_{13}, \rho_{33}, \nu_{13}^*)\,\boldsymbol{\Gamma}^{-1}(\rho_{13}, \rho_{33}, \nu_{13}^*)'}} \times \phi\left(\frac{z_3 - (\rho_{13}, \rho_{33}, \nu_{13}^*)\,\boldsymbol{\Gamma}^{-1}(z_0, z_1, z_2)'}{\sqrt{1-(\rho_{13}, \rho_{33}, \nu_{13}^*)\,\boldsymbol{\Gamma}^{-1}(\rho_{13}, \rho_{33}, \nu_{13}^*)'}}\right) \tag{B.13}$$

with $\boldsymbol{\Gamma}$ the submatrix of $\boldsymbol{\Sigma}_1$ formed by first three rows and three columns.
Denote $(\sigma^*)^2 = (\rho_{13}, \rho_{33}, \nu_{13}^*)\boldsymbol{\Gamma}^{-1}(\rho_{13}, \rho_{33}, \nu_{13}^*)'$. Then, we have

$$\begin{aligned}
L_1 &= \int_{-c_1(1-w_1)/w_0}^{c_1(1-w_1)/w_0} \phi(z_0)\,dz_0 \int_{-c_1}^{c_1} \frac{1}{\sqrt{1-\nu_{13}^2}}\phi\left(\frac{z_1-\nu_{13}z_0}{\sqrt{1-\nu_{13}^2}}\right)dz_1 \int_{-c_2(1-w_1^*)/w_0^*}^{c_2(1-w_1^*)/w_0^*} \left[\Phi\left(\left(c_2-(\rho_{13}, \rho_{33}, \nu_{13}^*)\,\boldsymbol{\Gamma}^{-1}(z_0, z_1, z_2)'\right)\left(\sqrt{1-(\sigma^*)^2}\right)^{-1}\right)\right. \\
&\quad \left. -\Phi\left(\left(-c_2-(\rho_{13}, \rho_{33}, \nu_{13}^*)\,\boldsymbol{\Gamma}^{-1}(z_0, z_1, z_2)'\right)\left(\sqrt{1-(\sigma^*)^2}\right)^{-1}\right)\right]dz_2, \\
L_2 &= \int_{-c_1(1-w_1)/w_0}^{c_1(1-w_1)/w_0} \phi(z_0)\,dz_0 \int_{-c_1}^{c_1} \frac{1}{\sqrt{1-\nu_{13}^2}}\phi\left(\frac{z_1-\nu_{13}z_0}{\sqrt{1-\nu_{13}^2}}\right)dz_1 \int_{-c_2}^{-c_2(1-w_1^*)/w_0^*} \left[\Phi\left(\left(c_2-(\rho_{13}, \rho_{33}, \nu_{13}^*)\,\boldsymbol{\Gamma}^{-1}(z_0, z_1, z_2)'\right)\left(\sqrt{1-(\sigma^*)^2}\right)^{-1}\right)\right. \\
&\quad \left. -\Phi\left(\left(-\frac{c_2+w_0^*z_2}{w_1^*}-(\rho_{13}, \rho_{33}, \nu_{13}^*)\,\boldsymbol{\Gamma}^{-1}(z_0, z_1, z_2)'\right)\left(\sqrt{1-(\sigma^*)^2}\right)^{-1}\right)\right]dz_2, \\
L_4 &= \int_{c_1(1-w_1)/w_0}^{c_1} \phi(z_0)\,dz_0 \int_{-c_1}^{(c_1-w_0z_0)/w_1} \frac{1}{\sqrt{1-\nu_{13}^2}}\phi\left(\frac{z_1-\nu_{13}z_0}{\sqrt{1-\nu_{13}^2}}\right)dz_1 \int_{-c_2(1-w_1^*)/w_0^*}^{c_2(1-w_1^*)/w_0^*} \left[\Phi\left(\left(c_2-(\rho_{13}, \rho_{33}, \nu_{13}^*)\,\boldsymbol{\Gamma}^{-1}(z_0, z_1, z_2)'\right)\left(\sqrt{1-(\sigma^*)^2}\right)^{-1}\right)\right. \\
&\quad \left. -\Phi\left(\left(-c_2-(\rho_{13}, \rho_{33}, \nu_{13}^*)\,\boldsymbol{\Gamma}^{-1}(z_0, z_1, z_2)'\right)\left(\sqrt{1-(\sigma^*)^2}\right)^{-1}\right)\right]dz_2, \\
L_5 &= \int_{c_1(1-w_1)/w_0}^{c_1} \phi(z_0)\,dz_0 \int_{-c_1}^{(c_1-w_0z_0)/w_1} \frac{1}{\sqrt{1-\nu_{13}^2}}\phi\left(\frac{z_1-\nu_{13}z_0}{\sqrt{1-\nu_{13}^2}}\right)dz_1 \int_{-c_2}^{-c_2(1-w_1^*)/w_0^*} \left[\Phi\left(\left(c_2-(\rho_{13}, \rho_{33}, \nu_{13}^*)\,\boldsymbol{\Gamma}^{-1}(z_0, z_1, z_2)'\right)\left(\sqrt{1-(\sigma^*)^2}\right)^{-1}\right)\right. \\
&\quad \left. -\Phi\left(\left(-\frac{c_2+w_0^*z_2}{w_1^*}-(\rho_{13}, \rho_{33}, \nu_{13}^*)\,\boldsymbol{\Gamma}^{-1}(z_0, z_1, z_2)'\right)\left(\sqrt{1-(\sigma^*)^2}\right)^{-1}\right)\right]dz_2, \\
L_6 &= \int_{c_1(1-w_1)/w_0}^{c_1} \phi(z_0)\,dz_0 \int_{-c_1}^{(c_1-w_0z_0)/w_1} \frac{1}{\sqrt{1-\nu_{13}^2}}\phi\left(\frac{z_1-\nu_{13}z_0}{\sqrt{1-\nu_{13}^2}}\right)dz_1 \int_{c_2(1-w_1^*)/w_0^*}^{c_2} \left[\Phi\left(\left(\frac{c_2-w_0^*z_2}{w_1^*}-(\rho_{13}, \rho_{33}, \nu_{13}^*)\,\boldsymbol{\Gamma}^{-1}(z_0, z_1, z_2)'\right)\left(\sqrt{1-(\sigma^*)^2}\right)^{-1}\right)\right. \\
&\quad \left. -\Phi\left(\left(-c_2-(\rho_{13}, \rho_{33}, \nu_{13}^*)\,\boldsymbol{\Gamma}^{-1}(z_0, z_1, z_2)'\right)\left(\sqrt{1-(\sigma^*)^2}\right)^{-1}\right)\right]dz_2.
\end{aligned} \tag{B.14}$$

Finally,

$$\oint_{\Omega_1} f(z_0, z_1, z_2, z_3; \boldsymbol{\Sigma}_1)\,dz_0dz_1dz_2dz_3 = L_1 + 2(L_2 + L_4 + L_5 + L_6). \tag{B.15}$$

Acknowledgments

The work was partially supported by the National Natural Science Foundation of China (no. 11225103; 11161054; 11171293; 81072386); the Key Fund of Yunnan Province (no. 2010CC003); The Youth Program of Applied Basic Research

Programs of Yunnan Province (2013FD001); the Foundation of Yunnan University (no. 2012CG018).

References

[1] B. Freidlin, G. Zheng, Z. Li, and J. L. Gastwirth, "Trend tests for case-control studies of genetic markers: power, sample size and robustness," *Human Heredity*, vol. 53, no. 3, pp. 146–152, 2002.

[2] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.

[3] K. Song and R. C. Elston, "A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies," *Statistics in Medicine*, vol. 25, no. 1, pp. 105–126, 2006.

[4] G. Zheng and J. L. Gastwirth, "On estimation of the variance in Cochran-Armitage trend tests for genetic association using case-control studies," *Statistics in Medicine*, vol. 25, no. 18, pp. 3150–3159, 2006.

[5] Q. Li, K. Yu, Z. Li, and G. Zheng, "MAX-rank: a simple and robust genome-wide scan for case-control association studies," *Human Genetics*, vol. 123, no. 6, pp. 617–623, 2008.

[6] Q. Li, G. Zheng, X. Liang, and K. Yu, "Robust tests for single-marker analysis in case-control genetic association studies," *Annals of Human Genetics*, vol. 73, no. 2, pp. 245–252, 2009.

[7] Y. Zang and W. K. Fung, "Robust tests for matched case-control genetic association studies," *BMC Genetics*, vol. 11, article 91, 2010.

[8] Y. Zang, W. K. Fung, and G. Zheng, "Simple algorithms to calculate asymptotic null distributions of robust tests in case-control genetic association studies in R," *Journal of Statistical Software*, vol. 33, no. 8, pp. 1–24, 2010.

[9] H. C. So and P. C. Sham, "Robust association tests under different genetic models, allowing for binary or quantitative traits and covariates," *Behavior Genetics*, vol. 41, no. 5, pp. 768–775, 2011.

[10] Q. Li, W. Xiong, J. B. Chen et al., "A robust test for quantitative trait analysis with model uncertainty in genetic association studies," *Statistics and Its Interface*. In press.

[11] J. M. Satagopan, E. S. Venkatraman, and C. B. Begg, "Two-stage designs for gene-disease association studies with sample size constraints," *Biometrics*, vol. 60, no. 3, pp. 589–597, 2004.

[12] A. D. Skol, L. J. Scott, G. R. Abecasis, and M. Boehnke, "Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies," *Nature Genetics*, vol. 38, no. 2, pp. 209–213, 2006.

[13] K. Yu, N. Chatterjee, W. Wheeler et al., "Flexible design for following up positive findings," *American Journal of Human Genetics*, vol. 81, no. 3, pp. 540–551, 2007.

[14] R. Sladek, G. Rocheleau, J. Rung et al., "A genome-wide association study identifies novel risk loci for type 2 diabetes," *Nature*, vol. 445, no. 7130, pp. 881–885, 2007.

[15] D. Pan, Q. Li, N. Jiang, A. Liu, and K. Yu, "Robust joint analysis allowing for model uncertainty in two-stage genetic association studies," *BMC Bioinformatics*, vol. 12, article 9, 2011.

[16] A. J. Silman and J. E. Pearson, "Epidemiology and genetics of rheumatoid arthritis," *Arthritis Research & Therapy*, vol. 4, supplement 3, pp. S265–S272, 2002.

[17] A. J. MacGregor, H. Snieder, A. S. Rigby et al., "Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins," *Arthritis & Rheumatism*, vol. 43, no. 1, pp. 30–37, 2000.

[18] C. I. Amos, W. V. Chen, M. F. Seldin et al., "Data for Genetic Analysis Workshop 16 Problem 1, association analysis of rheumatoid arthritis data," *BMC Proceedings*, vol. 3, supplement 7, article S2, 2009.

[19] F. Xia, J. Y. Zhou, and W. K. Fung, "A powerful approach for association analysis incorporating imprinting effects," *Bioinformatics*, vol. 27, no. 18, pp. 2571–2577, 2011.

[20] G. Zheng, C. O. Wu, M. Kwak, W. Jiang, J. Joo, and J. A. C. Lima, "Joint analysis of binary and quantitative traits with data sharing and outcome-dependent sampling," *Genetic Epidemiology*, vol. 36, no. 3, pp. 263–273, 2012.

[21] T. W. J. Huizinga, C. I. Amos, A. H. M. van der Helm-Van Mil et al., "Refining the complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins," *Arthritis & Rheumatism*, vol. 52, no. 11, pp. 3433–3438, 2005.

[22] L. Chen, M. Zhong, W. V. Chen, C. I. Amos, and R. Fan, "A genome-wide association scan for rheumatoid arthritis data by Hotellings T^2 tests," *BMC Proceedings*, vol. 3, supplement 7, article S6, 2009.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 732989, 9 pages
http://dx.doi.org/10.1155/2013/732989

Research Article

A Statistical Method for Synthesizing Meta-Analyses

Liansheng Larry Tang,[1] Michael Caudy,[2] and Faye Taxman[2]

[1] *Department of Statistics, George Mason University, Fairfax, VA 22030, USA*
[2] *Department of Criminology, Law and Society, George Mason University, Fairfax, VA 22030, USA*

Correspondence should be addressed to Liansheng Larry Tang; tang7814@yahoo.com

Received 16 May 2013; Accepted 28 August 2013

Academic Editor: Gengsheng Qin

Multiple meta-analyses may use similar search criteria and focus on the same topic of interest, but they may yield different or sometimes discordant results. The lack of statistical methods for synthesizing these findings makes it challenging to properly interpret the results from multiple meta-analyses, especially when their results are conflicting. In this paper, we first introduce a method to synthesize the meta-analytic results when multiple meta-analyses use the same type of summary effect estimates. When meta-analyses use different types of effect sizes, the meta-analysis results cannot be directly combined. We propose a two-step frequentist procedure to first convert the effect size estimates to the same metric and then summarize them with a weighted mean estimate. Our proposed method offers several advantages over existing methods by Hemming et al. (2012). First, different types of summary effect sizes are considered. Second, our method provides the same overall effect size as conducting a meta-analysis on all individual studies from multiple meta-analyses. We illustrate the application of the proposed methods in two examples and discuss their implications for the field of meta-analysis.

1. Introduction

The last two decades have seen an exponential growth in the popularity of meta-analyses across scientific disciplines including medical research [1] and diagnostic medicine [2]. The purpose of meta-analysis is to assess the consistency and robustness of findings across populations, settings, and contextual factors in order to help ensure that a practice is likely to produce similar results when it is implemented. A single study cannot determine, with certainty, that an intervention works or does not work. Instead, studies that are combined together, across different settings, and conducted over time can establish a pattern of consistent findings that may be useful to justify new or refined practice. Several studies combined can establish both significance and repeatability of results [3]. Common statistical methods to combine studies in meta-analyses are based on a fixed-effects model which assumes a homogeneous treatment effect among studies or a more general random-effects model which allows heterogeneity among studies [4].

Multiple meta-analyses are sometimes conducted to investigate the effect of the same topic or intervention. The synthesis of these meta-analyses can be used to highlight what is known on a particular topic or intervention or what can contribute to more complete understanding of the extant empirical evidence [5]. However, the summarizing of meta-analyses can be very challenging because existing meta-analyses may be conflicting, may be reported differently or incompletely across studies, or may be more or less valid based on the quality of the research synthesis methods that were employed in conducting the reviews. This raises challenges for interpreting and drawing conclusions about what the results of these studies mean and how they should be used to inform future research, theory development, and practice. Systematic reviews are further complicated when a diverse array of health-relevant outcomes are assessed. This is essentially the same issue researchers and practitioners face when attempting to summarize primary studies, the same issue that makes meta-analysis appealing.

Several nonstatistical approaches to summarizing meta-analyses of the same topic or intervention are currently available. The Cochrane Collaboration has developed a set of recommended procedures for conducting overviews of

reviews when multiple meta-analyses exist regarding different treatments for the same clinical condition [6–8]. Nonstatistical approaches to summarizing multiple reviews, which may be used for meta-analyses, include vote counting and using decision algorithms to identify the review(s) that are most salient [7]. However, narrative reviews can be far too subjective to reflect the knowledge that has been gained through research [9]. An ideal statistical approach to summarizing meta-analyses is to conduct a new meta-analysis of all primary studies included in multiple meta-analyses. That is, individual studies in all the related meta-analyses are identified, and their effect sizes are combined to calculate a summary statistic using meta-analytic techniques such as the random-effects model in DerSimonian and Laird [4]. However, while this approach is ideal and appealing when there are only two or three meta-analyses to combine, doing so is not as efficient as directly summarizing effect sizes reported in existing meta-analyses on similar topics. Recently, Hemming et al. [10] provided a Bayesian method to summarize multiple reviews. Their method assumes that estimated summary effect sizes follow a random-effects model based on exchangeability assumption. Differing types of effect sizes are not considered in their method. In addition, the authors did not study the fixed-effects models which may lead to nice properties for the combined overall effect size.

In this paper we start with describing methods to meta-analyze effect size estimates from individual studies extracted from several existing meta-analyses. However, the methods require a substantial amount of time and resources. Thus, we also introduce a more efficient statistical method to directly summarize information from existing meta-analyses without going back to individual studies. The paper is organized as follows. In Section 2 we introduce the notations and our method for synthesizing meta-analyses with the same type of summary statistics. We also describe two-step methods for combining different types of summary statistics commonly used in meta-analyses. We start with the fixed-effects model and show that summarizing the reviews yields the same overall effect size as conducting a meta-analysis on all possible individual studies. We then describe the method for the random-effects model. Section 3 applies the proposed method to two examples. Section 4 discusses some potential applications of this method.

2. Methods

The use of quantitative research synthesis techniques such as meta-analysis overcomes many of the limitations of traditional narrative literature reviews. Meta-analysis provides an objective and quantitative approach to research synthesis by taking into account factors such as sample size, magnitude, and direction of relationships and the methodological quality of the various studies analyzed [11–13]. Meta-analyses are not limited by a reliance on traditional indicators of statistical significance but instead rely upon effect sizes to give a picture of the size and scope of the impact of an intervention [14]. However, the more meta-analyses that are conducted on a given topic, the more difficult it can be to analyze and determine the cumulative findings from a body of literature.

Calculating the summary effect sizes across existing meta-analyses is a potentially valuable way to synthesize the knowledge base on a given topic. This process should be guided by the same rule that guides traditional meta-analysis which requires that studies being combined should be assessing a common clinical research question. Suppose several meta-analyses consider a common clinical question, and use similar search criteria to include and exclude individual studies. Summarizing these meta-analyses can provide better understanding of the clinical question than any single meta-analysis.

When multiple meta-analyses address the same topic, we propose that researchers should explore methods to synthesize the summary effect sizes from the existing meta-analyses. This proposed approach only requires summary effect sizes and their variances which are reported in the existing meta-analyses. The calculated overall effect size for the combined meta-analyses is essentially a weighted average of the summary effect sizes with the weights being the inverse of the variances; this approach is analogous to the techniques used to conduct a meta-analysis of primary studies.

Popular effect size estimates used in meta-analyses include standardized mean differences, odds ratios, and correlation coefficients [14]. In this section, we first consider the simplest case of combining multiple meta-analyses when all analyses use the same type of effect size estimates in a fixed-effects model. We then extend the method to the more commonly used random-effects model. Finally, we present a method for combining meta-analyses with different types of summary statistics in the fixed-effects or random-effects model.

2.1. Combining Effect Sizes from the Same Measure

2.1.1. Synthesizing All Individual Studies. Suppose J meta-analyses using similar search criteria on a common topic are identified and need to be combined to obtain an overall summary statistic. We use θ_{ij} and $\widehat{\theta}_{ij}$ to denote a true effect parameter and its estimate, respectively, for study i, $i = 1, \ldots, k_j$, in meta-analysis j. The estimate $\widehat{\theta}_{ij}$ is commonly assumed to follow a normal distribution as $\widehat{\theta}_{ij} \sim N(\theta_{ij}, \sigma_{ij}^2)$ where σ_{ij}^2 is the variance of $\widehat{\theta}_{ij}$. The variance estimator, $\widehat{\sigma}_{ij}^2$, obtained from the study tends to be smaller for a larger sample size. If all $\widehat{\theta}_{ij}$'s and $\widehat{\sigma}_{ij}^2$'s are available, researchers can conduct a meta-analysis using existing statistical methods based on a random-effects model (REM) or a fixed-effects model (FEM). As a popular model in meta-analysis, the REM assumes heterogeneity among θ_{ij} and is given as follows:

$$\begin{aligned} \widehat{\theta}_{ij} &= \theta_j + \epsilon_{ij}, \\ \theta_j &= \theta + u_j, \end{aligned} \tag{1}$$

where $\epsilon_{ij} \sim N(0, \sigma_{ij}^2)$ and $u_j \sim N(0, \tau^2)$. Here σ_{ij}^2 is estimated using the sample variance $\widehat{\sigma}_{ij}^2$ from study i in meta-analysis j. Conventionally, the true variance σ_{ij}^2 is assumed to be known and equal to $\widehat{\sigma}_{ij}^2$, and to simplify the notation we will use σ_{ij}^2

instead of $\widehat{\sigma}_{ij}^2$ throughout the paper. The random effect, u_j, represents heterogeneous θ_j among all studies, and its variance can be estimated using a moment estimator $\widehat{\tau}^2$ by DerSimonian and Laird [4]:

$$\widehat{\tau}^2 = \frac{Q - \left(\sum_{j=1}^{J} k_j - 1\right)}{\sum_{i,j} w_{ij} - \sum_{i,j} w_{ij}^2 / \sum_{i,j} w_{ij}}, \tag{2}$$

where

$$Q = \sum_{i,j} w_{ij} \widehat{\theta}_{ij}^2 - \frac{\left(\sum_{i,j} w_{ij} \widehat{\theta}_{ij}\right)^2}{\sum_{i,j} w_{ij}}, \tag{3}$$

and the weight w_{ij} is given by $1/\sigma_{ij}^2$ indicating a larger weight for a larger sample size. The moment estimate, $\widehat{\tau}_j^2$, should always be nonnegative. If a negative value of $\widehat{\tau}_j$ is obtained, $\widehat{\tau}$ is set to 0 according to DerSimonian and Laird [4]. Using $w_{ij}^* = 1/(\sigma_{ij}^2 + \widehat{\tau}^2)$ as a weight for study i in meta-analysis j, an estimate for the summary effect θ is given by a weighted average of effect estimates:

$$\widehat{\theta} = \frac{\sum_{j=1}^{J} \sum_{i=1}^{k_j} w_{ij}^* \widehat{\theta}_{ij}}{\sum_{j=1}^{J} \sum_{i=1}^{k_j} w_{ij}^*}. \tag{4}$$

In the FEM, the random-effect component in (1) is 0 by assuming true effect parameters are homogenous across studies, and the REM in (1) becomes

$$\widehat{\theta}_{ij} = \theta + \epsilon_{ij}. \tag{5}$$

The estimate for θ can be obtained using (4) by setting $\widehat{\tau} = 0$.

2.1.2. Synthesizing Meta-Analyses. Obtaining $\widehat{\theta}$ with results from all individual studies as in (4) is an ideal way to synthesize multiple meta-analyses. In the traditional meta-analysis, hundreds of studies are often available on a similar topic, and when authors of some individual studies are not contactable, criteria can be set accordingly to exclude these studies. However, extracting individual study data requires tremendous time and resources. Alternatively, we propose to directly summarize meta-analyses. We start with showing the equivalence of synthesizing meta-analyses and synthesizing all individual studies under the simple FEM which assumes the same effect parameter, θ, across studies. The effect parameter θ_j in meta-analysis j is estimated using

$$\widehat{\theta}_j = \frac{\sum_{i=1}^{k_j} w_{ij} \widehat{\theta}_{ij}}{\sum_{i=1}^{k} w_{ij}}, \tag{6}$$

where $w_{ij} = 1/\sigma_{ij}^2$. The variance estimate of $\widehat{\theta}_j$ is given by $\text{var}(\widehat{\theta}_j) = 1/\sum_{i=1}^{k} w_{ij}$. Using $w_j = 1/\text{var}(\widehat{\theta}_j)$ as the weight for meta-analysis j, we can obtain a weighted average of J meta-analysis results as follows:

$$\widehat{\theta}_f = \frac{\sum_{j=1}^{J} w_j \widehat{\theta}_j}{\sum_{j=1}^{J} w_j}. \tag{7}$$

After some math, we see that the expression above is the same as (4) with $\widehat{\tau} = 0$. The result implies that conducting a meta-analysis by pooling summary effect sizes of the meta-analyses of interest is equivalent to a meta-analysis of combining all individual studies from these meta-analyses. The variance of the weighted average is given by $\text{var}(\widehat{\theta}_f) = 1/(\sum_{j=1}^{J} w_j)$.

The assumption of the same effect parameter across the studies may not be realistic considering studies' heterogeneous characteristics. This assumption is relaxed in the REM as shown in (1). Using the data from meta-analysis j, the parameter τ accounting for the random component is estimated by the following moment estimator:

$$\widehat{\tau}_j^2 = \max\left\{0, \frac{Q_j - \left(k_j - 1\right)}{\sum_i w_{ij} - \left(\sum_i w_{ij}^2 / \sum_i w_{ij}\right)}\right\}, \tag{8}$$

where

$$Q_j = \sum_i w_{ij} \widehat{\theta}_{ij}^2 - \frac{\left(\sum_i w_{ij} \widehat{\theta}_{ij}^2\right)}{\sum_i w_{ij}}. \tag{9}$$

The moment estimator $\widehat{\tau}_j^2$ has the same expression as the one given in (8) except that it is now obtained using only data from meta-analysis j. Meta-analysis j then uses $\widetilde{w}_{ij}^* = 1/(\sigma_{ij}^2 + \widehat{\tau}_j^2)$ as a weight for study i to obtain an estimate for the summary effect θ_j by

$$\widehat{\widetilde{\theta}}_j^* = \frac{\sum_{i=1}^{k_j} \widetilde{w}_{ij}^* \widehat{\theta}_{ij}}{\sum_{i=1}^{k_j} \widetilde{w}_{ij}^*}. \tag{10}$$

To synthesize J meta-analyses, a possible way is to use $\widetilde{w}_j^* = 1/\sum_{i=1}^{k_j} \widetilde{w}_{ij}^*$ as the weight for meta-analysis j and calculate a weighted average of $\widehat{\theta}_j$'s as follows:

$$\widehat{\theta}_r = \frac{\sum_{j=1}^{J} \widetilde{w}_j^* \widehat{\widetilde{\theta}}_j^*}{\sum_{j=1}^{J} \widetilde{w}_j^*}. \tag{11}$$

Comparison between the estimate above and the one in (4) shows some difference. We can write (4) as the weighted average of meta-analyses summary statistics in the following expression:

$$\widehat{\theta} = \frac{\sum_{j=1}^{J} w_j^* \widehat{\theta}_j^*}{\sum_{j=1}^{J} w_j^*}, \tag{12}$$

where

$$\widehat{\theta}_j^* = \frac{\sum_{i=1}^{k_j} w_{ij}^* \widehat{\theta}_{ij}}{\sum_{i=1}^{k_j} w_{ij}^*}, \tag{13}$$

and $w_j^* = 1/\sum_{i=1}^{k_j} w_{ij}^*$. A close look at $\widehat{\widetilde{\theta}}_j^*$ and $\widehat{\theta}_j^*$ reveals a slight difference. This is due to different weight estimates, $\widetilde{w}_{ij}^*$ and w_{ij}^*, where the former uses studies in meta-analysis

j to estimate τ and the latter pools studies from all meta-analyses to estimate τ. In practice, a well-planned meta-analysis includes a large number of studies, and $\widehat{\tau}_j^2$'s and $\widehat{\tau}^2$ should all be close to the true parameter τ. Subsequently, the difference between $\widehat{\theta}_r$ and $\widehat{\theta}$ can be possibly ignored. Thus, although the weights used in (11) may not be as accurate as the ones obtained by pooling all individual studies, $\widehat{\theta}_r$ is still a statistically sound estimate for θ.

2.2. Combining Effect Sizes from Different Types of Statistics. The summary statistics in meta-analyses may be of different types. When individual studies in a meta-analysis have continuous outcomes, the statistic may appear in the form of the sample mean difference. Sometimes, other studies address the same topic, but the outcome is dichotomized, and the odds ratio is commonly used to summarize the difference between two groups. The Pearson correlation coefficient r is frequently used to evaluate the correlation between two continuous variables. When one variable is a continuous outcome and the other variable is dichotomized to indicate the group status, r can be used to compare outcomes between two groups. When meta-analyses addressing the same topic use different types of statistics, one should convert these statistics to the same type of summary statistics before synthesizing them.

2.2.1. Synthesizing All Individual Studies. We now discuss how to combine meta-analyses with these types of summary statistics using a two-step procedure. The first step is to convert various types of statistics to the same type. In this paper, we discuss how to convert the log-transformed odds ratio (OR) and the correlation coefficient to the same scale of the sample mean difference. The log-transformed OR can be converted to the sample mean difference using their linear relationship as will be discussed. However, a Taylor expansion is needed to convert the correlation coefficient to the sample mean difference via linear relationship. We focus on the linear relations because the weighted average in a meta-analysis is a linear combination of effect estimates, and preserving the linear nature in the conversion will facilitate the future calculation for synthesizing the meta-analyses.

Suppose meta-analysis j, $j = 1, \ldots, J$, is conducted using the standardized sample mean difference $\widehat{\theta}_{j,m}$ with its variance of $\sigma_{j,m}^2$ and meta-analysis j', $j' = 1, \ldots, J'$, has the estimated log-transformed odds ratio $\widehat{\theta}_{j',\mathrm{OR}}$ as the summary effect and its sample variance of $\sigma_{j',\mathrm{OR}}^2$.

Chinn gives the detailed discussion on how to convert between the standardized mean difference and the estimated odds ratios from individual studies [15]. Since the log-transformed odds ratio, $\widehat{\theta}_{ij',\mathrm{OR}}$, in study i of meta-analysis j follows approximately a logistic distribution which differs from a standard normal distribution mainly in the tail area, dividing $\widehat{\theta}_{ij',\mathrm{OR}}$ by $\pi/\sqrt{3}$, or 1.81, converts $\widehat{\theta}_{ij',\mathrm{OR}}$ to an approximate standardized sample mean difference.

In meta-analysis j'', $j'' = 1, \ldots, J''$, with correlation coefficients, the coefficient $\widehat{\theta}_{ij'',r}$ from study i is converted to Fisher's z-score $\widehat{\theta}_{ij'',z}$ before the further calculation [14]. The conversion is given by

$$\widehat{\theta}_{ij'',z} = \frac{1}{2}\ln\left(\frac{1+\widehat{\theta}_{ij'',r}}{1-\widehat{\theta}_{ij'',r}}\right). \tag{14}$$

Since a correlation coefficient $\widehat{\theta}_r$ can be converted to a standardized mean difference $\widehat{\theta}_m$ using $\widehat{\theta}_m = 2\widehat{\theta}_r/\sqrt{1-\widehat{\theta}_r^2}$, we can convert Fisher's z-score $\widehat{\theta}_z$ to $\widehat{\theta}_m$ using $\widehat{\theta}_m = e^{\widehat{\theta}_z} - e^{-\widehat{\theta}_z}$. The relationship above between $\widehat{\theta}_z$ and $\widehat{\theta}_m$ is nonlinear. The second order Taylor expansion of

$$\widehat{\theta}_m = 1 + \widehat{\theta}_z + \frac{\widehat{\theta}_z^2}{2} - \left(1 - \widehat{\theta}_z + \frac{\widehat{\theta}_z^2}{2}\right) = 2\widehat{\theta}_z \tag{15}$$

gives the approximate linear relationship between $\widehat{\theta}_z$ and $\widehat{\theta}_m$.

After the effect estimates are converted to the same scale of $\widehat{\theta}_{j,m}$, they can be used to estimate the overall effect parameter θ in the following REM model:

$$\begin{aligned} &\widehat{\theta}_{ij,m} = \theta_{ij} + \epsilon_{ij}, \qquad \frac{\widehat{\theta}_{ij',\mathrm{OR}}}{1.81} = \theta_{ij'} + \epsilon_{ij'}, \\ &2\widehat{\theta}_{ij'',z} = \theta_{ij''} + \epsilon_{ij''}, \qquad \theta_{ij} = \theta + u_{ij}, \\ &\theta_{ij'} = \theta + u_{ij'}, \qquad \theta_{ij''} = \theta + u_{ij''}, \end{aligned} \tag{16}$$

where $\epsilon_{ij} \sim N(0, \sigma_{ij}^2)$, $\epsilon_{ij'} \sim N(0, \sigma_{ij'}^2)$, $\epsilon_{ij''} \sim N(0, \sigma_{ij''}^2)$, and $u_{ij}, u_{ij'}, u_{ij''} \sim N(0, \tau^2)$.

In the second step, $\widehat{\theta}_{ij,m}$, $\widehat{\theta}_{ij,\mathrm{OR}}$, and $\widehat{\theta}_{ij,z}$ are combined to obtain the overall summary effect estimate $\widehat{\theta}$. Let $w_{ij'} = 1/\sigma_{ij}^2$, $w_{ij'}/\sigma_{ij'}^2$, and $w_{ij''} = 1/\sigma_{ij''}^2$. Again, the parameter τ^2 can be estimated by a moment estimator $\widehat{\tau}^2$ similar to (8) using all effect estimates and their variances:

$$\widehat{\tau}^2 = \max\left\{0, \frac{Q - \left(\sum_{j=1}^{J} k_j + \sum_{j'=1}^{J} k_{j'} + \sum_{j''=1}^{J} k_{j''} - 1\right)}{w - w^2/w}\right\}, \tag{17}$$

where

$$\begin{aligned} w &= \sum_{i,j} w_{ij} + \sum_{i,j'} w_{ij'} + \sum_{i,j''} w_{ij''}, \\ w^2 &= \sum_{i,j} w_{ij}^2 + \sum_{i,j'} w_{ij'}^2 + \sum_{i,j''} w_{ij''}^2, \\ Q &= \sum_{i,j} w_{ij}\widehat{\theta}_{ij,m}^2 + \sum_{i,j'} w_{ij'}\left(\frac{\widehat{\theta}_{ij',\mathrm{OR}}}{1.81}\right)^2 + \sum_{i,j''} w_{ij''}\left(2\widehat{\theta}_{ij'',r}\right)^2 \\ &\quad - \frac{1}{w}\left(\sum_{i,j} w_{ij}\widehat{\theta}_{ij,m} + \sum_{i,j'} w_{ij'}\frac{\widehat{\theta}_{ij',\mathrm{OR}}}{1.81} + \sum_{i,j''} 2w_{ij''}\widehat{\theta}_{ij'',r}\right)^2. \end{aligned} \tag{18}$$

The resulting estimate for the summary effect θ using the REM is given by

$$\widehat{\theta} = \frac{\sum_{i,j} w^*_{ij}\widehat{\theta}_{ij,m} + \sum_{i,j'} w^*_{ij'}\widehat{\theta}_{ij',\mathrm{OR}}/1.81 + \sum_{i,j''} 2w^*_{ij''}\widehat{\theta}_{ij'',r}}{\sum_{i,j} w^*_{ij} + \sum_{i,j'} w^*_{ij'} + \sum_{i,j''} w^*_{ij''}}, \tag{19}$$

where $w^*_{ij'} = 1/(\sigma^2_{ij} + \widehat{\tau}^2)$, $w^*_{ij'} = 1/(\sigma^2_{ij'} + \widehat{\tau}^2)$, and $w^*_{ij''} = 1/(\sigma^2_{ij''} + \widehat{\tau}^2)$. The variance of $\widehat{\theta}$ is given by $\mathrm{var}(\widehat{\theta}) = 1/(\sum_{i,j} w^*_{ij} + \sum_{i,j'} w^*_{ij'} + \sum_{i,j''} w^*_{ij''})$.

2.2.2. Synthesizing Meta-Analyses. Combining meta-analyses with different types of statistics is similar to synthesizing meta-analyses using the same type of statistics except for the appropriate conversion of summary statistics. We let $\widehat{\tau}_j$ be an estimate for τ in the meta-analysis j using the standardized sample mean difference, $\widehat{\tau}_{j'}$ in the meta-analysis j' using the log-transformed odds ratio, and $\widehat{\tau}_{j''}$ in the meta-analysis j'' using the correlation coefficient. These estimates take the similar form as (17) using data from the corresponding meta-analysis. The estimate for the summary effect in each meta-analysis is derived by using these estimates for τ and the variances of the effect estimates in the weights and is given by

$$\begin{aligned}
\widehat{\theta}_{j,m} &= \frac{\sum_{i=1}^{k_j} \widetilde{w}^*_{ij}\widehat{\theta}_{ij,m}}{\sum_{i=1}^{k_j} \widetilde{w}^*_{ij}}, \\
\widehat{\theta}_{j',\mathrm{OR}} &= \frac{\sum_{i=1}^{k_{j'}} \widetilde{w}^*_{ij'}\widehat{\theta}_{ij',\mathrm{OR}}}{\sum_{i=1}^{k_{j'}} \widetilde{w}^*_{ij'}}, \\
\widehat{\theta}_{j'',z} &= \frac{\sum_{i=1}^{k_{j''}} \widetilde{w}^*_{ij''}\widehat{\theta}_{ij'',z}}{\sum_{i=1}^{k_{j''}} \widetilde{w}^*_{ij''}},
\end{aligned} \tag{20}$$

where $\widetilde{w}^*_{ij} = 1/(\sigma^2_{ij} + \widehat{\tau}^2_j)$, $\widetilde{w}^*_{ij'} = 1/(\sigma^2_{ij'} + \widehat{\tau}^2_{j'})$, and $\widetilde{w}^*_{ij''} = 1/(\sigma^2_{ij''} + \widehat{\tau}^2_{j''})$. To directly synthesize these meta-analysis results, we use a weighted average of converted effect estimates as follows:

$$\widehat{\theta} = \frac{\sum_{j=1}^{J} w^*_j\widehat{\theta}_{j,m} + \sum_{j'=1}^{j'} w^*_{j'}\widehat{\theta}_{j',\mathrm{OR}}/1.81 + \sum_{j''=1}^{J''} 2w^*_{j''}\widehat{\theta}_{j,r}}{\sum_{j=1}^{J} w^*_j + \sum_{j'=1}^{J'} w^*_{j'} + \sum_{j''=1}^{J''} w^*_{j'}}, \tag{21}$$

where $\widetilde{w}^*_j = \sum_{j=1}^{J} \widetilde{w}^*_{ij}$, $\widetilde{w}^*_{j'} = \sum_{j'=1}^{J'} \widetilde{w}^*_{ij'}$, and $\widetilde{w}^*_{j''} = \sum_{j''=1}^{J''} \widetilde{w}^*_{ij'}$. The variance of $\widehat{\theta}$ is given by $\mathrm{var}(\widehat{\theta}) = 1/(\sum_{j=1}^{J} w^*_j + \sum_{j'=1}^{J'} w^*_{j'} + \sum_{j''=1}^{J''} w^*_{j'})$. The weights $\widetilde{w}^*_j$, $\widetilde{w}^*_{j'}$, and $\widetilde{w}^*_{j''}$ are usually calculated from the variances, P values, or the confidence intervals found in original meta-analyses being combined. Articles which only report effect estimates do not have sufficient information on the true effect and should be excluded in the synthesis. Following similar lines in previous sections, we can show that the estimated overall effect is approximately the same as the one by combining individual studies.

The effect estimates used in the expression (21) are already converted to the same type. The estimates reported in original meta-analyses may not be in the same form. For example, a meta-analysis using the correlation coefficient is likely to report the correlation coefficient instead of Fisher's z-score. Suppose $\widehat{\mathrm{OR}}_{j'}$ is the estimated odds ratio in meta-analysis j' and $r_{j''}$ is the estimated correlation coefficient in meta-analysis j''. In terms of these indices and their variances originally reported in the meta-analyses, the overall effect estimate can be written as the following expression:

$$\begin{aligned}
\widehat{\theta} = \Bigg(& \sum_{j=1}^{J} w_{j,m}\widehat{\theta}_{j,m} + 1.81 \sum_{j'=1}^{J'} w_{j',\mathrm{OR}} \ln\left(\widehat{\mathrm{OR}}_{j'}\right) \\
& + \frac{1}{2} \sum_{j''=1}^{J''} w_{j'',r} \ln\left(\frac{1 + r_{j''}}{1 - r_{j''}}\right)\Bigg) \\
& \times \left(\sum_{j=1}^{J} w_{j,m} + 1.81^2 \sum_{j'=1}^{J'} w_{j',\mathrm{OR}} + \frac{1}{4} \sum_{j''=1}^{J''} w_{j'',r}\right)^{-1},
\end{aligned} \tag{22}$$

where $w_{j,m} = 1/\mathrm{var}(\widehat{\theta}_{j,m})$, $w_{j',\mathrm{OR}} = 1/\mathrm{var}(\ln(\widehat{\mathrm{OR}}_{j'}))$, and $w_{j'',r} = 1/\{\mathrm{var}(r_{j''})/(1 - r_{j''^2})^2\}$. And its variance is

$$\mathrm{var}\left(\widehat{\theta}\right) = \frac{1}{\sum_{j=1}^{J} w_{j,m} + 1.81^2 \sum_{j'=1}^{J'} w_{j',\mathrm{OR}} + 1/4 \sum_{j''=1}^{J''} w_{j'',r}}. \tag{23}$$

The proposed estimate in (22) is a summary standardized sample mean difference. Its expression represents two steps. The first step is to convert the odds ratios and correlation coefficients to the same scale of the standardized sample mean difference, and the second step is to calculate the overall mean difference from all meta-analyses.

3. Examples

We illustrate the proposed methods in two examples. The first example illustrates that combining two meta-analyses using our methods produces the same overall effect size as conducting a meta-analysis from all includable primary studies based on either the fixed-effects model or the random-effects model. The second example illustrates the process of combining multiple meta-analyses of the same clinical question which have concordant findings. The calculation is conducted using the R package "metafor."

The first example is based on a meta-analysis by Swift and Callahan [16], which compared treatment outcomes between appropriately and inappropriately matched groups using 26 individual studies [16]. The treatments in these studies include pharmacotherapy, cognitive behavioral therapy, and group therapy. The treatment outcomes range from substance use days to weight loss. All the studies used the Pearson correlation coefficient as the effect estimate, and the estimates

Table 1: Effect estimates with lower limits (LL) and upper limits (UL) of 95% confidence interval (CI).

Study number	r	95% CI		var(r)
		LL	UL	
1	0.25	0.04	0.44	0.0104
2	0.21	−0.02	0.43	0.0132
3	0.18	0.00	0.35	0.0080
4	0.05	−0.15	0.24	0.0099
5	0.17	0.00	0.33	0.0071
6	0.33	0.09	0.53	0.0126
7	0.26	0.02	0.47	0.0132
8	0.51	0.24	0.71	0.0144
9	0.25	0.03	0.45	0.0115
10	−0.26	−0.45	−0.04	0.0109
11	0.23	−0.04	0.48	0.0176
12	0.04	−0.24	0.32	0.0204
13	0.50	0.25	0.69	0.0126
14	0.11	−0.28	0.47	0.0366
15	0.15	−0.04	0.33	0.0089
16	0.10	0.00	0.21	0.0029
17	0.04	−0.19	0.27	0.0138
18	0.04	−0.12	0.21	0.0071
19	0.12	−0.03	0.27	0.0059
20	0.55	0.15	0.80	0.0275
21	−0.07	−0.31	0.19	0.0163
22	0.18	−0.26	0.56	0.0438
23	0.23	0.00	0.44	0.0126
24	0.11	−0.14	0.34	0.0150
25	0.21	−0.14	0.52	0.0283
26	−0.04	−0.15	0.07	0.0031

and the confidence intervals are presented in Table 1. The summary correlation coefficient estimate using all 26 studies is $r = 0.1334$ with 95% confidence interval $(0.0951, 0.1717)$ based on a fixed-effects model. We conduct two separate meta-analyses based on studies 1–12 and studies 13–26 as cited in the reference section of the review in Swift and Callahan [16]. The summary effect sizes for these two meta-analyses are calculated to be $\widehat{\theta}_1 = 0.1774$ with its sample variance being $\text{var}(\widehat{\theta}_1) = 0.00095$ and $\widehat{\theta}_2 = 0.1041$ with its sample variance being $\text{var}(\widehat{\theta}_2) = 0.00064$. It is of interest to investigate whether combining these two effect sizes gives a similar result as conducting the meta-analysis using all 26 individual studies. We use the inverse of the sample variances as the weights and calculate the weighted average of the summary effect sizes. The resulting overall effect size is

$$\widehat{\theta}_f = \frac{\left(\sum_{j=1}^{2} \widehat{\theta}_j / \text{var}\left(\widehat{\theta}_j\right)\right)}{\left(\sum_{j=1}^{2} 1/\text{var}\left(\widehat{\theta}_j\right)\right)} = 0.1334, \tag{24}$$

with its variance being $\text{var}(\widehat{\theta}_f) = 0.00038$. The 95% confidence interval is also $(0.0951, 0.1717)$. The results are identical to those found using all individual effect sizes indicating the potential utility of this procedure.

To further illustrate the technique we repeat the analyses with a random-effects model (REM). We again use the 26 studies from Swift and Callahan [16] and illustrate that the proposed summary effect size obtained under REM may differ slightly from the one achieved from combining all individual studies. Under the REM, the summary correlation coefficient using all 26 studies is $\widehat{\theta} = 0.1597$ with 95% confidence interval $(0.0941, 0.2253)$. We repeat the process described above based on studies 1–12 and studies 13–26, respectively. Using the random-effects model, the summary correlation coefficients for these two meta-analyses are $\widehat{\widehat{\theta}}_1^* = 0.1827$ with its variance being $\text{var}(\widehat{\widehat{\theta}}_1^*) = 0.0027$ and $\widehat{\widehat{\theta}}_2^* = 0.1370$ with its variance being $\text{var}(\widehat{\widehat{\theta}}_2^*) = 0.0018$. Combining these two effect sizes produces an overall weighted mean correlated coefficient $\widehat{\theta}_r = 0.1553$, with its variance being $\text{var}(\widehat{\theta}_r) = 0.00108$. This variance is smaller than those of separate meta-analyses, indicating more accuracy in estimating the overall correlation coefficient by including more individual studies. The 95% confidence interval is $(0.0909, 0.2197)$. Although the estimates of the between-studies variance differ slightly across the two meta-analyses, combining the two pseudoanalyses produces a similar finding to that of combining all of the includable individual studies.

To illustrate the method for synthesizing different types of effect estimates, we create a synthetic dataset using the studies from Swift and Callahan [16]. We convert the correlation coefficients to standardized sample mean differences for studies 1–9 and to log-transformed odds ratios for studies 10–17. The estimates are kept unchanged for studies 18–26. The converted estimates and the corresponding confidence intervals are listed in Table 2. The variances of the estimates are obtained using $(\text{UL} - \text{LL})^2/(2 \times 1.96)^2$. We use the fixed-effects model to conduct meta-analyses on these three sets of studies. The summary effect estimates are $\widehat{\theta}_{1,m} = 0.4355$ with $\text{var}(\widehat{\theta}_{1,m}) = 0.0059$, $\widehat{\theta}_{1,\text{OR}} = 0.3304$ with $\text{var}(\widehat{\theta}_{1,\text{OR}}) = 0.0183$, and $\widehat{\theta}_{1,r} = 0.0711$ with $\text{var}(\widehat{\theta}_{1,r}) = 0.0011$. The overall effect estimate in mean difference is calculated using (22) to be $\widehat{\theta} = 0.2973$ with variance $\text{var}(\widehat{\theta}) = 0.0017$. We see that this result is quite close to the summary effect size calculated from combining individual studies at the beginning of the section which is 0.2692 with the variance 0.0016 after converting to the mean effect size. This indicates that the proposed procedure for directly combining results from meta-analyses produces similar results as the ideal but much more tedious procedure which instead combines individual studies subtracted from meta-analyses.

The anterior cruciate ligament (ACL) is a commonly reconstructed ligament of the knee. The meta-analyses conducted by Biau et al. [17] and M. C. Forster and I. W. Forster [18] indicate that the graft failure using hamstring autografts is not significantly different compared with bone-patellar tendon-bone autografts. We sought to apply our approach to synthesize these two meta-analyses on the graft choices in the ACL reconstruction. The summary odds ratios of graft fracture in Biau et al. [17] are $\widehat{\text{OR}}_1 = 1.33$ with the 95% confidence interval $(0.73, 2.44)$ and $\widehat{\text{OR}}_2 = 1.09$ with

Table 2: Effect estimates of different types with lower limits (LL) and upper limits (UL) of 95% confidence interval (CI). Summary effect estimates are calculated based on a fixed-effects model for three sets of studies.

Study no.	$\widehat{\theta}_m$	95% CI		Study no.	$\widehat{\theta}_{\mathrm{OR}}$	95% CI		Study no.	$\widehat{\theta}_r$	95% CI	
		LL	UL			LL	UL			LL	UL
1	0.52	0.08	0.98	10	−0.97	−1.82	−0.14	18	0.04	−0.12	0.21
2	0.43	−0.04	0.95	11	0.86	−0.14	1.98	19	0.12	−0.03	0.27
3	0.37	0.00	0.75	12	0.14	−0.89	1.22	20	0.55	0.15	0.80
4	0.10	−0.30	0.49	13	2.09	0.93	3.45	21	−0.07	−0.31	0.19
5	0.35	0.00	0.70	14	0.40	−1.06	1.93	22	0.18	−0.26	0.56
6	0.70	0.18	1.25	15	0.55	−0.14	1.27	23	0.23	0.00	0.44
7	0.54	0.04	1.06	16	0.36	0.00	0.78	24	0.11	−0.14	0.34
8	1.19	0.49	2.02	17	0.14	−0.70	1.02	25	0.21	−0.14	0.52
9	0.52	0.06	1.01					26	−0.04	−0.15	0.07

the 95% confidence interval $(0.40, 2.96)$ in M. C. Forster and I. W. Forster [18]. We first convert the odds ratios and their confidence limits to the natural log scale since meta-analyses using odds ratios are commonly conducted using the weighted average of the log-transformed odds ratios. After the conversion, we obtain the log-transformed odds ratios $\widehat{\theta}_{1,\mathrm{OR}} = 0.2852$ with the variance 0.9048 and $\widehat{\theta}_{2,\mathrm{OR}} = 0.0862$ with the variance 0.3607. The resulting overall estimate of log odds ratio is

$$\widehat{\theta} = \sum_{j=1}^{2} \left(\frac{\widehat{\theta}_{j,\mathrm{OR}} / \mathrm{var}\left(\widehat{\theta}_{j,\mathrm{OR}}\right)}{\sum_{j=1}^{2} \left(1/\mathrm{var}\left(\widehat{\theta}_{j,\mathrm{OR}}\right)\right)} \right) = 0.2321, \tag{25}$$

with the variance 0.0695. This variance is smaller than either of the meta-analyses due to the inclusion of more studies. The 95% confidence interval of $\widehat{\theta}$ is $(-0.2846, 0.2321)$. Converting back to the original scale, the overall estimated odds ratio based on the two meta-analyses is 1.2613 with 95% confidence interval $(0.7523, 1.2613)$. Although this confidence interval does not show significant difference between two graft choices, it is narrower than those in Biau et al. [17] and M. C. Forster and I. W. Forster [18]. This is expected since the overall estimated odds ratio calculated across meta-analyses provides a more accurate indicator than those calculated from separate meta-analyses.

4. Discussion

Despite the advantages of meta-analyses over traditional narrative review methods, these techniques are not without limitations. As is the case with primary studies, multiple meta-analyses often produce conflicting findings even when they examine the same body of literature [7] given the numerous methodological decisions that are made during the review process. These decisions affect the search strategy, coding of studies, moderators examined, inclusion/exclusion criteria, and the reporting of effect sizes, and each has the potential to impact the review findings. Despite the transparency of the method, there is always a certain degree of subjectivity in research synthesis [14]. We have presented methods which can synthesize meta-analysis findings with minimal subjectivity.

Attempting to make statements about the overall findings of a body of systematic review literature raises a number of challenges. Part of the challenge is the lack of consistency in how authors report the findings of meta-analyses [19, 20]. Since no standardized reporting procedures have been adopted, there is a great deal of variability in how meta-analytic findings are reported both within and across academic disciplines. Additionally, many extant meta-analyses do not completely report effect sizes; confidence intervals, variance measures, and precise P values are often omitted. This lack of consistency and quality assurance is a major limitation of this body of research and potentially limits the transportability of meta-analysis findings [21]. Like primary studies, meta-analyses on similar constructs may differ from one another in many ways. To address this issue and make meta-analyses more compatible and therefore more appropriate for synthesis and better equipped to inform practice, it is apparent that the research community needs to continue to work towards adopting a set of standardized best practices for conducting and reporting meta-analyses in behavioral health (see, e.g., [22]).

While the examples provided here illustrate that it is possible to calculate summary effect sizes across multiple meta-analyses of the same intervention, our testing of these techniques is still ongoing. A potential barrier to this approach is the lack of independence that results when multiple meta-analyses include many of the same primary studies. Treating these different outcomes as independent can produce incorrect estimates of the variance for the summary effect, but it does not invalidate the calculation of the effect size itself. This can be seen from (4) for combining individual studies. For example, suppose J meta-analyses use the exactly same set of studies. Equation (4) becomes

$$\widehat{\theta} = \frac{J \sum_{i=1}^{k_1} w_{i1}^{*} \widehat{\theta}_{i1}}{J \sum_{i=1}^{k_1} w_{i1}^{*}}, \tag{26}$$

which gives the same estimate as the summary effect estimate in either of the meta-analyses. Ideally, the resulting variance

of $\widehat{\theta}$ should be the same as the one in either of the meta-analyses. But instead, we have

$$\operatorname{var}\left(\widehat{\theta}\right) = \frac{1}{J\sum_{i=1}^{k_1} w_{i1}^*}, \tag{27}$$

which tends to be considerably smaller than the appropriate value of $1/\sum_{i=1}^{k_1} w_{i1}^*$. This limitation makes the proposed method more appropriate for meta-analyses with no or a limited number of overlapping studies. A possible inflation factor of J can be multiplied with the variance in this case to correct for the biased variance estimator. More generally, with p overlapping studies in all meta-analyses, the unbiased variance of the summary estimator can be written as

$$\begin{aligned}\operatorname{var}\left(\widehat{\theta}\right) &= \frac{1}{\sum_{j=1}^{J}\sum_{i=1}^{k} w_{ij}} \\ &\quad\times \frac{J^2\sum_{i=1}^{p} w_{i1} + \sum_{j=1}^{J}\sum_{i=p+1}^{k} w_{ij}}{J\sum_{i=1}^{p} w_{i1} + \sum_{j=1}^{J}\sum_{i=p+1}^{k} w_{ij}} \\ &= \frac{1}{\sum_{j=1}^{J}\sum_{i=1}^{k} w_{ij}} \times \mathrm{IF},\end{aligned} \tag{28}$$

where the inflation factor IF corrects for the underestimated variance and we have IF $\geq$ 1. If w_{ij}'s can be retrieved from the overlapped studies, the factor can be precisely calculated. Otherwise, the investigator may have to guess an appropriate value for the inflation factor.

While further research is needed to refine the technique proposed in this paper, the findings of these initial validations suggest the utility of this method for combining mean effect sizes from multiple meta-analyses of the same intervention or treatment. The methodology detailed in this paper has several possible applications. The technique has its most utility when several meta-analyses have been conducted on the same treatment and have produced varying results. Even when multiple meta-analyses report similar findings regarding the magnitude and direction of effect sizes, the proposed technique can be used to summarize across the extant meta-analytic evidence base. The technique also offers a solution when meta-analyses have been conducted to update prior research syntheses; rather than ignoring findings predating the inclusion criteria of the updated study, this technique can be used to summarize the full range of available evidence. These and other potential applications make this a particularly appealing technique for researchers and practitioners alike who are faced with the challenge of summarizing the rapidly expanding body of meta-analytic research in medicine and many other scientific disciplines.

Acknowledgments

The authors would like to thank the associate editor and two referees for their constructive comments.

References

[1] A. J. Sutton and J. P. T. Higgins, "Recent developments in meta-analysis," *Statistics in Medicine*, vol. 27, no. 5, pp. 625–650, 2008.

[2] K. Zou, A. Liu, A. Bandos, L. Ohno-Machado, and H. Rockette, *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis*, CRC Press, New York, NY, USA, 2010.

[3] T. D. Cook and L. C. Leviton, "Reviewing the literature: a comparison of traditional methods with meta-analysis," *Journal of Personality*, vol. 48, pp. 449–472, 1980.

[4] R. DerSimonian and N. Laird, "Meta-analysis in clinical trials," *Controlled Clinical Trials*, vol. 7, no. 3, pp. 177–188, 1986.

[5] F. Taxman and S. Belenko, *Implementing Evidence-Based Practices in Community Corrections and Addiction Treatment*, Springer Series on Evidence-Based Crime Policy, Springer, New York, NY, USA, 2011.

[6] L. A. Becker and O. D. Oxman, "Overviews of reviews," in *Cochrane Handbook for Systematic Reviews of Interventions*, J. P. T. Higgins and S. Green, Eds., Cochrane, 2011.

[7] A. R. Jadad, D. J. Cook, and G. P. Browman, "A guide to interpreting discordant systematic reviews," *Canadian Medical Association Journal* , vol. 156, no. 10, pp. 1411–1416, 1997.

[8] V. Smith, D. Devane, C. M. Begley, and M. Clarke, "Methodology in conducting a systematic review of systematic reviews of healthcare interventions," *BMC Medical Research Methodology*, vol. 11, article 15, 2011.

[9] M. Eccles, "An overview of reviews evaluating the effects of financial incentives in changing healthcare professional behaviours and patient outcomes (protocol)," *Cochrane Database of Systematic Reviews*, no. 7, Article ID CD008608, 2010.

[10] K. Hemming, R. J. Bowater, and R. J. Lilford, "Pooling systematic reviews of systematic reviews: a Bayesian panoramic meta-analysis," *Statistics in Medicine*, vol. 31, no. 3, pp. 201–216, 2012.

[11] X.-H. Zhou, N. Hu, G. Hu, and M. Root, "Synthesis analysis of regression models with a continuous outcome," *Statistics in Medicine*, vol. 28, no. 11, pp. 1620–1635, 2009.

[12] M. Egger and G. Davey Smith, "Meta-analysis: bias in location and selection of studies," *British Medical Journal*, vol. 316, no. 7124, pp. 61–66, 1998.

[13] M. Egger, G. D. Smith, and A. N. Phillips, "Meta-analysis: principles and procedures," *British Medical Journal*, vol. 315, no. 7121, pp. 1533–1537, 1997.

[14] M. Borenstein, L. Hedges, J. Higgins, and H. Rothstein, *Introduction To Meta-Analysis*, John Wiley & Sons, New York, NY, USA, 2009.

[15] S. Chinn, "A simple method for converting an odds ratio to effect size for use in meta-analysis," *Statistics in Medicine*, vol. 19, no. 22, pp. 3127–3131, 2000.

[16] J. K. Swift and J. L. Callahan, "The impact of client treatment preferences on outcome: a meta-analysis," *Journal of Clinical Psychology*, vol. 65, no. 4, pp. 368–381, 2009.

[17] D. J. Biau, C. Tournoux, S. Katsahian, P. J. Schranz, and R. S. Nizard, "Bone-patellar tendon-bone autografts versus hamstring autografts for reconstruction of anterior cruciate ligament: meta-analysis," *British Medical Journal*, vol. 332, no. 7548, pp. 995–998, 2006.

[18] M. C. Forster and I. W. Forster, "Patellar tendon or four-strand hamstring? A systematic review of autografts for anterior cruciate ligament reconstruction," *The Knee*, vol. 12, no. 3, pp. 225–230, 2005.

[19] D. Moher, A. Jones, and L. Lepage, "Use of the consort statement and quality of reports of randomized trials: a comparative beforeand-after evaluation," *The Journal of the American Medical Association*, vol. 285, pp. 1992–1995, 2001.

[20] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," *International Journal of Surgery*, vol. 8, no. 5, pp. 336–341, 2010.

[21] PLoS Medicine Editors, "Many reviews are systematic but some are more transparent and completely reported than others," *PLoS Medicine*, vol. 4, p. e147, 2007.

[22] J. P. T. Higgins and S. Green, Eds., *Overviews of Reviews*, Cochrane, 2011.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 340678, 8 pages
http://dx.doi.org/10.1155/2013/340678

Research Article

SNP Selection in Genome-Wide Association Studies via Penalized Support Vector Machine with MAX Test

Jinseog Kim,[1] Insuk Sohn,[2] Dennis (Dong Hwan) Kim,[3] and Sin-Ho Jung[2,4]

[1] *Department of Statistics and Information Science, Dongguk University, Gyeongju 780-714, Republic of Korea*
[2] *Samsung Cancer Research Institute, Samsung Medical Center, Seoul 137-710, Republic of Korea*
[3] *Department of Medical Oncology and Hematology, Princess Margaret Hospital, University of Toronto, Toronto, ON, Canada M5G 2M9*
[4] *Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, USA*

Correspondence should be addressed to Sin-Ho Jung; sinho.jung@duke.edu

Received 22 May 2013; Revised 14 August 2013; Accepted 22 August 2013

Academic Editor: Wenqing He

One of main objectives of a genome-wide association study (GWAS) is to develop a prediction model for a binary clinical outcome using single-nucleotide polymorphisms (SNPs) which can be used for diagnostic and prognostic purposes and for better understanding of the relationship between the disease and SNPs. Penalized support vector machine (SVM) methods have been widely used toward this end. However, since investigators often ignore the genetic models of SNPs, a final model results in a loss of efficiency in prediction of the clinical outcome. In order to overcome this problem, we propose a two-stage method such that the the genetic models of each SNP are identified using the MAX test and then a prediction model is fitted using a penalized SVM method. We apply the proposed method to various penalized SVMs and compare the performance of SVMs using various penalty functions. The results from simulations and real GWAS data analysis show that the proposed method performs better than the prediction methods ignoring the genetic models in terms of prediction power and selectivity.

1. Introduction

We consider a genome-wide association study (GWAS) on a complex disease. One of the popular study objectives of such study is to predict a binary clinical outcome, such as benign versus malignant and response versus no response with respect to a specific regimen, based on single-nucleotide polymorphisms (SNPs) data. A fitted prediction model will be used to predict the diagnostic or prognostic outcomes of future patients. Recently, penalization approaches incorporating logistic model or support vector machines have been actively proposed to fit prediction models with binary outcomes. These are well known to achieve both predictive accuracy and variable selection simultaneously.

By introducing shrinkage priors of the normal exponential-gamma (NEG) distribution family, Hoggart et al. [1] suggested a stochastic search method for penalized logistic regression models with SNPs. Ayers and Cordell [2] showed that the NEG priors have better performance than other competing penalized methods using simulations, while it is very computing intensive to produce the results. Wu et al. [3] considered lasso-penalized logistic regression [4] with a large number of SNPs and proposed a cyclic coordinate descent algorithm [5] to implement the computation. Kooperberg et al. [6] removed SNPs that had a Hardy-Weinberg P value smaller than 10^{-5} and applied logistic regression models with lasso and Elastic net [7] penalties using a set of SNPs preselected by a cross-validation procedure. On the other hand, Wei et al. [8] proposed selecting SNPs using EigenStrat algorithm [9] and applying the SVM and logistic regression as predictive models. Abraham et al. [10] showed that the two penalized methods, l_1 and Elastic-net SVM, were robust in case/control predictive performance based on simulation studies and real data analyses. These simultaneous analysis methods ignored the genetic models of SNPs [6] or assumed the additive model for all SNPs [6, 8, 10].

The statistical tests such as the Pearson's chi-squared test or the Cochran-Armitage trend test (CATT) are frequently used to test if an SNP is associated with a binary outcome by assuming a specific genetic model. Oftentimes, however, the true genetic model is unknown. We can improve the testing power if we know the true genetic model of an SNP [11]. Toward this end, the test based on the maximum over the three CATT statistics (MAX test) has been presented by several authors [12, 13]. Kim et al. [14] recently proposed a prediction method for time-to-event traits using SNPs and showed that a prediction model based on the best fitting genetic models of SNPs can improve the prediction efficiency. We extend their approach to the prediction of binary outcomes using SVMs.

In this paper, we propose a prediction method combining the MAX test and penalized SVM to predict binary outcome using SNPs. The proposed method consists of two phase procedures: (i) to select candidate prognostic SNPs and identify their genetic models using MAX test, and (ii) to fit a prediction model using the penalized SVM with appropriate scores for the selected SNPs based on their genetic types. We compare the performance of the proposed method using a different penalized SVM method through simulations and a real GWAS data analysis. Each SVM method is combined with MAX test or the general practice ignoring the genetic types of the SNPs.

To facilitate and enable MAX test, we provide the R package called `SNPselect` in http://datamining.dongguk.ac.kr/Rlib/SNPselect which uses the penalized SVM R package [15] to implement SVM with SCAD, l_1, and Elastic Net penalties.

2. Methods

2.1. Penalized Support Vector Machine. Suppose that there are n subjects. For the subject i $(= 1, \ldots, n)$, we have an input vector $x_i \in R^p$ and a class label $y_i \in \{-1, 1\}$. The SVM [16, 17] is to find the optimal hyperplane which separates data points into two classes with the largest margin.

Wahba et al. [18] and Hastie et al. [19] found that the optimization problem of the SVM can be represented as a penalized optimization problem:

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} \left[1 - y_i \left(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\right)\right]_+ + p_\lambda \left(\boldsymbol{\beta}\right), \tag{1}$$

where $[1 - yf]_+ = \max(1 - yf, 0)$ is called the hinge loss and p_λ is a penality function with regularization parameter λ. The SVM using an l_2-norm, $p_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$, as a penalty function is called the standard SVM or l_2-SVM.

The l_2-SVM has been successfully applied to classification with high-dimensional data such as gene microarrays and SNPs, but it does not select the variables affecting the response class label. For feature selection with l_2-SVM, Guyon et al. [20] proposed the SVM-REF procedure which combines the recursive feature elimination (RFE) with the l_2-SVM. This procedure consists of a two-step procedure using an external gene selection method.

In order to achieve classification accuracy and feature selection simultaneously, variants of SVM have been proposed by replacing the penalty function in (1) with other types of penalty functions, for example, SVM with 1-norm [21, 22], adaptive lasso [23], or smoothly clipped and absolute deviation (SCAD) [24, 25] penalties. The SVM with 1-norm (or l_1-SVM) adapts the lasso (or l_1) penalty, $p_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$, originally proposed by Tibshirani [4] as a practical alternative to l_2 penalty. Due to the l_1 penalty, the l_1-SVM automatically selects variables by shrinking the small coefficients of the hyperplane to exactly zero.

One of major drawbacks of the l_1 penalty is that it tends to select only one variable when there are many correlated input variables in data. To overcome this limitation of LASSO, Zou and Hastie [7] proposed the Elastic Net penalty by combineing l_1 and l_2 penalties:

$$p_\lambda \left(\boldsymbol{\beta}\right) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2. \tag{2}$$

The Elastic Net penalty provides variable selection owing to l_1 penalty, while finding highly correlated variables, called grouping effect. Wang et al. [26] applied the Elastic Net penalty to SVM classification problems.

Fan and Li [24] proposed the smoothly clipped absolute deviation (SCAD) penalty given as

$$p_\lambda \left(\boldsymbol{\beta}\right) = \sum_{j=1}^{p} p_\lambda \left(\beta_j; a\right), \tag{3}$$

where

$$p_\lambda \left(\beta; a\right) = \begin{cases} \lambda \left|\beta\right| & \text{if } \left|\beta\right| < \lambda \\ -\dfrac{\left|\beta\right|^2 - 2a\lambda \left|\beta\right| + \lambda^2}{2\left(a - 1\right)} & \text{if } \lambda \le \left|\beta\right| \le a\lambda \\ \dfrac{\left(a + 1\right)\lambda^2}{2} & \text{if } \left|\beta\right| \ge a\lambda. \end{cases} \tag{4}$$

Here, a (>2) and λ (>0) are tuning parameters. Fan and Li [24] showed that the prediction with SCAD penalty is not sensitive to the tuning parameter a and recommended to use $a = 3.7$.

The SCAD yields the same behavior as l_1 for small coefficients β_j, $j = 1, \ldots, p$, but assigns a constant penalty for large coefficients. This property reduces the estimation bias. Fan and Li [24] demonstrate more desirable theoretical properties of the SCAD penalty compared to the l_1 penalty. Later, Zhang et al. [25] proposed the SVM with the SCAD penalty for feature selection.

2.2. Genetic Models for SNPs. Let AA, AB, and BB be three possible genotypes where B is a risk allele for a given SNP. We denote the number of B alleles in a genotype by k; that is, $k = 0, 1$, or 2 if the genotype is AA, AB, or BB, respectively. For a given SNP, the data from n patients are summarized in Table 1.

Let p_k denote the response probability given a genotype $k = 0, 1, 2$. If B is the response allele, the response probability increases as the number of B alleles in the SNP increases; that is, $p_0 \le p_1 \le p_2$. In this paper, we will consider three popular

Table 1: Genotype frequencies.

	AA	AB	BB	Total
Response	r_0	r_1	r_2	r
No response	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

genetic models satisfying this assumption:

(i) *recessive* model: $p_0 = p_1 < p_2$;

(ii) *dominant* model: $p_0 < p_1 = p_2$;

(iii) *additive* model: $p_0 < p_1 = (p_0 + p_2)/2$.

2.3. Trend Test and MAX Test. For testing association between an SNP and a clinical outcome in case-control studies, the statistical tests such as the Pearson's chi-squared test or CATT are frequently used when the true genetic model is known. In this case, the CATT is usually more powerful than Pearson's chi-squared test when $p_0 \le p_1 \le p_2$ [12]. For a single SNP, borrowing the notations of Table 1, the CATT statistic can be written as

$$T_c = \frac{n^{1/2} \sum_{k=0}^{2} c_k \left(sr_k - rs_k\right)}{\sqrt{rs\left\{n \sum_{k=0}^{2} c_k^2 n_k - \left(\sum_{k=0}^{2} x_k n_k\right)^2\right\}}}, \tag{5}$$

where (c_0, c_1, and c_2) is a set of scores assigned to genotypes (AA, AB, and BB) with respect to a specific genotype. The trend test is invariant under a linear transformation with $c_0 \le c_1 \le c_2$, so that the typical choice of these scores is $c_0 = 0$ and $c_2 = 1$, but c_1 can take a different value according to a specific genetic model. From the results of Sasieni [27] and Zheng et al. [12, 28], the optimal choices of c_1 are 0, 1/2 and 1 for the recessive, additive, and dominant models, respectively. Let p_k denote the response probability for genotype group $k = 0, 1, 2$. Under the null hypothesis of no association, $H_0 : p_0 = p_1 = p_2$, T_c approximately follows $N(0, 1)$ for large n.

When the true genetic model is unknown, the test based on multiple CATTs for different genetic models can lead to substantial reduction in statistical power [11] or inflated type I error rate. To address this issue, the test based on the maximum over the three CATT statistics (MAX test) has been proposed by several authors [12, 13]. Let T_R, T_A, and T_D denote the CATT statistics using the scores for recessive, additive, and dominant models, respectively. Based on the three CATT statistics, the MAX test statistic is defined as

$$T_{\max} = \max\left(\left|T_R\right|, \left|T_A\right|, \left|T_D\right|\right). \tag{6}$$

The MAX test has robust properties [29] and is more powerful than the Pearson's chi-squared test [12] when the underlying genetic model is unknown.

Even though one can easily calculate the MAX test statistic from (5) and (6), it is not simple to compute its P value. One approach of obtaining the P value is based on a Monte-Carlo simulation. Under H_0, Zheng et al. [12] showed that (T_R, T_D, T_A) is asymptotically normal with covariances

$$\begin{aligned}
\operatorname{cov}\left(T_R, T_A\right) &= \frac{f_2\left(f_1 + 2f_0\right)}{\sqrt{f_2\left(1 - f_2\right)}\sqrt{f_0\left(f_1 + 2f_2\right) + f_2\left(f_1 + 2f_0\right)}}, \\
\operatorname{cov}\left(T_R, T_D\right) &= \frac{f_0 f_2}{\sqrt{f_0\left(1 - f_0\right)}\sqrt{f_2\left(1 - f_2\right)}}, \\
\operatorname{cov}\left(T_A, T_D\right) &= \frac{f_0\left(f_1 + 2f_2\right)}{\sqrt{f_0\left(1 - f_0\right)}\sqrt{f_0\left(f_1 + 2f_2\right) + f_2\left(f_1 + 2f_0\right)}},
\end{aligned} \tag{7}$$

where f_k denotes the relative frequency of genotype $k = 0, 1, 2$. Thus we can approximate the P value of MAX test based on Monte-Carlo samples from multivariate normal distribution with estimated variance-covariance matrix $\widehat{\Sigma}$ which is obtained by replacing f_k in the above covariances with $\widehat{f}_k = r_k/n_k$ for $k = 0, 1, 2$ ($f_0 + f_1 + f_2 = 1$).

There have been some studies on variants of MAX test for binary clinical outcomes. Zheng et al. [12] developed a robust ranking method, called MAX-rank test. Conneely et al. [30] proposed an efficient P value computation method that is shown to be more accurate than that using permutations by adjusting for correlated test statistics. Li et al. [31] proposed the P-rank test approximating the P value for the MAX test with or without covariate adjustment. Li et al. [32] compared the performance of the MAX-rank and P-rank tests. For more detailed discussions on MAX test, see [11] or [32].

2.4. Classification via SVM with MAX Test. For patient $i = 1, \ldots, n$, let y_i denote the binary clinical outcome taking 1 if responded or -1 if not responded and $(k_{i1}, \ldots, k_{im})$ the encoded data on m SNPs, that is, $k_{ij} = 0, 1, 2$, the number of the risk allele for SNP j $(= 1, \ldots, m)$. To build a classification model with this data set, we propose a method combining a penalized SVM and the MAX test. Our method consists of two-phase procedures: (i) prescreening SNPs and identifying the genetic models for the selected SNPs using the MAX test and (ii) applying the penalized SVM to fit a classification model. Our method can be summarized as follows.

(1) Read in the clinical outcomes $(y_1, \ldots, y_n)$ and SNP data $\{(k_{i1}, \ldots, k_{im}),\ i = 1, \ldots, n\}$.

(2) For SNP j $(= 1, \ldots, m)$,

(a) using the original data, calculate test statistics $(T_{j,R}, T_{j,A}, T_{j,D})$ and their two-sided P values $(p_{j,R}, p_{j,A}, p_{j,D})$ and MAX test statistic $T_{j,\max} = \max(|T_{j,R}|, |T_{j,A}|, |T_{j,D}|)$.

(b) compute the approximate P value of MAX test by Monte-Carlo simulation:

(i) estimate the variance-covarince matrix $\widehat{\Sigma}_j$;

(ii) generate $(t_{j,R}^{(b)}, t_{j,A}^{(b)}, t_{j,D}^{(b)})$ from $N(0, \widehat{\Sigma}_j)$ for $b = 1, \ldots, B$ (=100,000, say);

(iii) approximate the P value for MAX test by

$$p_j = B^{-1}\sum_{b=1}^{B} I\left(t_{j,\max}^{(b)} \geq T_{j,\max}\right), \quad (8)$$

$$\text{where } t_{j,\max}^{(b)} = \max(|t_{j,R}^{(b)}|, |t_{j,A}^{(b)}|, |t_{j,D}^{(b)}|).$$

(3) SNP screening: select J ($\ll m$) SNPs with $p_j < \alpha$ for a prespecified α value, such as 0.01.

(4) For SNP j, identify the genetic model by the smallest P value among $p_{j,R}$, $p_{j,A}$, and $p_{j,D}$.

(5) Assign covariate values $(z_{i1}, \ldots, z_{iJ})$ using the score corresponding to the identified genetic model.

(6) Standardize the covariates; that is,

$$z'_{ij} = \frac{z_{ij} - \overline{z}_j}{s_j}, \quad (9)$$

where $\overline{z}_j = n^{-1}\sum_{i=1}^{n} z_{ij}$ and $s_j^2 = n^{-1}\sum_{i=1}^{n}(z_{ij} - \overline{z}_j)^2$.

(7) Apply the penalized SVM to the response data $(y_1, \ldots, y_n)$ and the standardized covariates $\{(z'_{i1}, \ldots, z'_{iJ}),\ i = 1, \ldots, n\}$.

3. Results

3.1. Simulation Studies. At first, we generate IID $N(0,1)$ random variables $\epsilon_{i1}, \ldots, \epsilon_{im}$ and, for $\rho \in (0,1)$, set

$$x_{ij} = \begin{cases} \epsilon_{ij}, & j = 1 \\ \rho x_{i,j-1} + \sqrt{1-\rho^2}\,\epsilon_{ij}, & j = 2, \ldots, m. \end{cases} \quad (10)$$

Note that $x_{i1}, \ldots, x_{im}$ have an AR(1) correlation structure with autocorrelation coefficient ρ as in [14]. Correlated SNP data are generated by

$$z_{ij} = \begin{cases} 0, & x_{ij} < u_{f_0} \\ 1, & u_{f_0} \leq x_{ij} < u_{(f_0+f_1)} \\ 2, & \text{otherwise}, \end{cases} \quad (11)$$

where u_q denotes the qth quantile of the standard normal distribution. The binary clinical outcome of patient i is generated using response probability p_i which is related to the covariates by

$$\text{logit}\left(p_i\right) = \sum_{j=1}^{m} \beta_j z_{ij}. \quad (12)$$

To consider the cases of uncorrelated or moderately correlated SNPs in our experiment, we set $\rho = 0$ or 0.3. We generate $m = 1000$ encoded SNPs with $(f_0, f_1) = (1/4, 1/2)$ for $j = 1, \ldots, 6$ and $(f_0, f_1) = (1/3, 1/3)$ for $j = 7, \ldots, 1000$. SNPs 1 and 2 have recessive models; SNPs 3 and 4 have dominant models, and SNPs 5 and 6 have additive models, the regression coefficients for these six prognostic SNPs are set at $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0.8$. According to the above data generation scheme, we have generated simulation data sets of size 200, and each data set is partitioned into 2/3 training set and 1/3 test set. For a classification model fitting, the SVM with one of the three penalty functions, SCAD (SCAD-SVM), l_1 (l_1-SVM), and Elastic Net (Enet-SVM), is applied to the SNPs selected using $\alpha = 0.01$. To choose a final classification model, we use 5-fold cross-validation for selecting the tuning parameters. One of the standard practice in the classification model fitting using SNP data will be assuming an equal genetic model for all SNPs. In order to evaluate the performance of the model fitting methods combined with the MAX test, we also have fitted a classification model by assuming one genetic model for all SNPs.

For each model fitting method, we calculate three performance measures such as the number of the selected SNPs, the number of the selected prognostic SNPs by the penalized SVM, and the misclassification error. Here, the selected SNPs are selected by penalized SVM among SNPs after a prescreening step, and the selected prognostic SNPs are the prognostic ones included in the selected SNPs. The misclassification errors are estimated using test data set; that is,

$$\frac{1}{n}\sum_{i=1}^{n} I\left(y_i \neq \text{sign}\,\widehat{f}\left(\mathbf{z}_i\right)\right), \quad (13)$$

where $I(x)$ is an indicator function, $\widehat{f}(\mathbf{z}) = \widehat{\beta}_1 z'_1 + \cdots + \widehat{\beta}_J z'_J$ denote the predicted response score predicted for the test set, and z'_js are standardized covariates in the test set using the means and standard errors calculated from the training set. In order to assess the variability of the experiments, we replicate the whole process 100 times. Table 2 summarizes the three averaged performance measures from our simulations.

When comparing the number of selected SNPs in Table 2, we observe that Enet-SVM tends to select more SNPs but SCAD-SVM selects lower SNPs except for the case of $\rho = 0.3$ and dominant model. In view of different genetic models, the proposed method selects more SNPs when applying l_1-SVM or Enet-SVM. However, the combination of proposed method and SCAD-SVM selects much less SNPs than other combinations. Comparing the numbers of prognostic SNPs, Enet-SVM or l_1-SVM performs better than SCAD-SVM and assuming the proposed method or additive model has good selectivities of the true prognostic SNPs. In results with correlated SNPs ($\rho = 0.3$), Enet-SVM and l_1-SVM with the proposed method result in better selectivities for true prognostic SNPs than those with the additive model. However, the proposed methods can be the worst when SCAD-SVM is used for uncorrelated SNP data. We also compare the misclassification errors. Even if there are a little differences between Enet-SVM and l_1-SVM, Enet-SVM performs better than other penalized methods. SCAD-SVM produces the worst misclassification errors for all cases. We also find that the proposed method has the lowest misclassification errors whatever the penalized SVM method used except the case of applying SCAD-SVM for $\rho = 0.3$. Based on the discussions on the simulation results so far,

Table 2: The result of simulations with 100 replications: selected SNPs and prognostic SNPs indicate the averaged numbers of the selected SNPs and the selected prognostic SNPs, respectively, in the fitted models; standard error is reported in the parentheses.

ρ	Genetic model	Selected SNPs			Prognostic SNPs			Misclassification error		
		l_1	Enet	SCAD	l_1	Enet	SCAD	l_1	Enet	SCAD
0	Proposed	43.10	48.66	20.31	5.11	5.46	3.31	0.1766	0.1567	0.2736
		(0.54)	(0.70)	(0.65)	(0.09)	(0.07)	(0.14)	(0.0048)	(0.0054)	(0.0062)
	Recessive	40.50	41.38	25.28	4.62	4.74	3.66	0.2408	0.2518	0.2912
		(0.56)	(1.11)	(0.66)	(0.10)	(0.10)	(0.14)	(0.0054)	(0.0065)	(0.0047)
	Additive	42.71	45.58	24.12	5.23	5.35	4.07	0.2161	0.2118	0.3272
		(0.49)	(0.87)	(0.91)	(0.08)	(0.08)	(0.18)	(0.0048)	(0.0064)	(0.0076)
	Dominant	41.35	43.46	23.99	4.70	4.86	3.38	0.2457	0.2347	0.2995
		(0.58)	(0.98)	(0.70)	(0.10)	(0.09)	(0.12)	(0.0056)	(0.0063)	(0.0042)
0.3	Proposed	42.92	45.68	19.56	5.12	5.20	3.40	0.1690	0.1541	0.2833
		(0.50)	(0.80)	(0.48)	(0.08)	(0.08)	(0.10)	(0.0049)	(0.0047)	(0.0060)
	Recessive	39.49	41.09	27.23	4.34	4.47	3.03	0.2383	0.2368	0.2741
		(0.58)	(0.87)	(0.59)	(0.10)	(0.11)	(0.03)	(0.0057)	(0.0057)	(0.0019)
	Additive	42.07	43.90	21.89	5.06	5.04	3.74	0.2126	0.2074	0.3338
		(0.52)	(0.96)	(0.67)	(0.08)	(0.08)	(0.08)	(0.0052)	(0.0057)	(0.0056)
	Dominant	39.97	38.56	24.97	4.56	4.29	2.04	0.2502	0.2338	0.2607
		(0.62)	(1.03)	(0.53)	(0.10)	(0.11)	(0.03)	(0.0065)	(0.0059)	(0.0039)

the proposed method combined with Enet-SVM or l_1-SVM could improve the selectivity for true prognostic SNPs and the ablility of precdiction than other methods using a prefixed genetic model.

3.2. Real Data Analysis Example. Kim et al. [33] performed a GWAS using Affymetic Genome-wide Human SNP Arrays 6.0 (San Diego, CA, USA) on 190 patients with chronic myelogenous leukemia (CML). After excluding the SNPs with one missing case and those with the same genotype for all 190 patients, we use 330,353 autosomal SNPs in the further data analysis. The clinical endpoint is the achievement of major molecular response by 18 months to an induction chemotherapy. BCR/ABL transcript levels were measured to determine molecular response to imatinib therapy as described before by Kim et al. [34] and presented using the international scale. Major molecular response (MMR) was defined as <0.1% of the BCR/ABL fusion gene transcript level on an international scale by quantitative PCR. Among the 190 patients, 115 responded.

We randomly partition the CML data into 126 training samples and 64 test samples and then calculate the predictive performance measures for the methods over 100 random partitions. Table 3 summarizes the number of selected SNPs and the mean misclassification errors with their standard errors in parentheses over 100 random partitions. Similar to the simulation results, l_1-SVM and Enet-SVM using the MAX test slightly increase the number of selections, but produce lower misclassification errorr. Among the three penalized methods, Enet-SVM selects the largest number of SNPs but has the lowest misclassification error regardless of the use of the MAX test. However, SCAD-SVM selects the lowest SNPs, while it has poor prediction performances for any assumption for genetic models, which is the same observation in the simulation results.

Table 4 shows the list of 51 SNPs selected commonly by three penalized methods from 126 training samples of one of 100 random partitions. TGFBR1 gene (rs420549, located in 3′UTR region) among 51 SNPs, transforming growth factor beta receptor 1, interacts with TGF beta 1 [35, 36] and TGF beta receptor 2 [37, 38] and is located in 9q22. TGF beta is playing an important role of maintaining the growth and differentiation balance of hematopoietic cells [39, 40] and is known to have bidirectional properties of tumor suppressing and promoting function [41]. TGF-β-FOXO signaling pathway is involved in the maintenance of leukemia-initiating cells in CML, contributing to intrinsic resistance of CML LSCs to tyrosine kinase inhibitor [42, 43]. Accordingly, intrinsic trait of receptor affinity on TGF-β might contribute to different sensitivities to TGF-β; thus, it is potentially explainable that the response to imatinib therapy is dependent on the TGFBR1 genotype.

4. Conclusions

Although the penalized methods have been considered as successful ones for prediction in GWAS, they are still subject to high misclassification error by ignoring the genetic models of prognostic SNPs. In this paper, we proposed a two-phase procedure: (i) carrying out the MAX test for screening out noncandidate SNPs and identifying the genetic models of the selected SNPs at the first stage and then (ii) applying a penalized SVM to the selected SNPs for fitting a classification model at the second stage. We have compared the performances of the proposed method with the conventional methods ignoring the genetic type of prognostic

Table 3: The results of CML data: number of selected SNPs and misclassification error are calculated on average over 100 random partitions; standard error is reported in the parentheses.

Genetic model	Average number of selected SNPs			Misclassification error		
	l_1-SVM	Enet-SVM	SCAD-SVM	l_1-SVM	Enet-SVM	SCAD-SVM
Proposed	70.38	99.80	55.90	0.0737	0.0590	0.1098
	(1.29)	(4.10)	(0.52)	(0.0036)	(0.0062)	(0.0013)
Recessive	55.24	120.46	27.82	0.1184	0.0562	0.2003
	(1.19)	(4.73)	(2.33)	(0.0048)	(0.0051)	(0.0044)
Additive	66.32	120.76	43.50	0.1063	0.0667	0.1530
	(1.12)	(5.00)	(0.27)	(0.0051)	(0.0061)	(0.0026)
Dominant	51.90	91.92	50.90	0.1013	0.0702	0.1663
	(0.89)	(4.81)	(1.30)	(0.0062)	(0.0069)	(0.0044)

Table 4: List of SNPs selected commonly by three penalized methods.

RS ID	Genetic model	P value	RS ID	Genetic model	P value	RS ID	Genetic model	P value
rs3750551	D	0.000510	rs9289221	R	0.000160	rs6621316	A	0.000890
rs3886721	A	0.000040	rs16972014	A	0.000170	rs9890262	R	0.000210
rs2938451	A	0.000000	rs3013492	R	0.000760	rs6779769	A	0.000510
rs6429646	R	0.000050	rs7095688	A	0.000920	rs9502826	D	0.000690
rs6426870	R	0.000230	rs1439691	R	0.000100	rs9896683	R	0.000850
rs4784924	R	0.000100	rs7123207	R	0.000490	rs12907966	D	0.000220
rs8075266	R	0.000190	rs16830058	A	0.000830	rs5979009	D	0.000150
rs4851920	R	0.000130	rs10484180	R	0.000930	rs17157980	D	0.000730
rs9809817	R	0.000190	rs1952096	A	0.000250	rs2865510	R	0.000160
rs342735	A	0.000180	rs2842068	D	0.000600	rs12457620	D	0.000810
rs17066311	D	0.000790	rs420549	D	0.000440	rs4510937	R	0.000390
rs6627852	A	0.000470	rs16822723	A	0.000590	rs8073928	R	0.000510
rs11841074	D	0.000130	rs2492664	A	0.000270	rs10409991	R	0.000290
rs9447907	R	0.000650	rs2029866	R	0.000730	rs1871332	A	0.000150
rs16873423	D	0.000360	rs764515	A	0.000030	rs1264547	D	0.000670
rs315025	A	0.000390	rs11197596	A	0.000240	rs2016016	A	0.000360
rs2355615	A	0.000130	rs9344734	D	0.000690	rs6605081	R	0.000150

SNPs through simulations and real data example. In the simulations, we observed that Enet-SVM and l_1-SVM select more SNPs but have higher selectivities for true prognostic SNPs and lower misclassification errors among the three penalized SVM methods. Combining the proposed method which selects candidate SNPs and estimates their genetic models, we observed that the penalized SVMs except for SCAD-SVM could improve the performances in terms of the selection of the true prognostic SNPs and misclassification errors. Furthermore, the differences of misclassification errors among the three methods with the proposed method become much smaller. Hence, whichever a penalized SVM for model fitting we use, combining it with the MAX test to identify the genetic models of candidate prognostic SNPs could help to improve its performances. We made similar observations from a real data example. Even so, the selection of candidate SNPs could vary according to the choice of a prespecified α; thus, the prescreening by the MAX test could not select a part of true prognostic SNPs. We will consider this point in future work.

Authors' Contribution

Jinseog Kim and Insuk Sohn contributed equally to this work.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (no. 2010-0023302).

References

[1] C. J. Hoggart, J. C. Whittaker, M. De Iorio, and D. J. Balding, "Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies," *PLoS Genetics*, vol. 4, no. 7, Article ID e1000130, 2008.

[2] K. L. Ayers and H. J. Cordell, "SNP Selection in genome-wide and candidate gene studies via penalized logistic regression," *Genetic Epidemiology*, vol. 34, no. 8, pp. 879–891, 2010.

[3] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, vol. 25, no. 6, pp. 714–721, 2009.

[4] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society B*, vol. 73, no. 3, pp. 273–282, 2011.

[5] J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani, "Pathwise coordinate optimization," *Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.

[6] C. Kooperberg, M. LeBlanc, and V. Obenchain, "Risk prediction using genome-wide association studies," *Genetic Epidemiology*, vol. 34, no. 7, pp. 643–652, 2010.

[7] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society B*, vol. 67, no. 2, pp. 301–320, 2005.

[8] Z. Wei, K. Wang, H.-Q. Qu et al., "From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes," *PLoS Genetics*, vol. 5, no. 10, Article ID e1000678, 2009.

[9] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[10] G. Abraham, A. Kowalczyk, J. Zobel, and M. Inouye, "Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease," *Genetic Epidemiology*, vol. 37, no. 2, pp. 184–195, 2013.

[11] B. Freidlin, G. Zheng, Z. Li, and J. L. Gastwirth, "Trend tests for case-control studies of genetic markers: power, sample size and robustness," *Human Heredity*, vol. 53, no. 3, pp. 146–152, 2002.

[12] G. Zheng, B. Freidlin, and J. L. Gastwirth, "Comparison of robust tests for genetic association using case-control studies," in *IMS Lecture Notes-Monograph Series 2nd Lehmann Symposium—Optimality*, vol. 49, pp. 253–265, 2006.

[13] R. Sladek, G. Rocheleau, J. Rung et al., "A genome-wide association study identifies novel risk loci for type 2 diabetes," *Nature*, vol. 445, no. 7130, pp. 881–885, 2007.

[14] J. Kim, I. Sohn, D. Son, D. H. Kim, T. Ahn, and S. Jung, "Prediction of a time-to-event trait using genome wide SNP data," *BMC Bioinformatics*, vol. 14, p. 58, 2013.

[15] N. Becker, W. Werft, G. Toedt, P. Lichter, and A. Benner, "PenalizedSVM: a R-package for feature selection SVM classification," *Bioinformatics*, vol. 25, no. 13, pp. 1711–1712, 2009.

[16] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1996.

[17] B. Scholkopf and A. Smola, *Learning With Kernels—Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press, 2002.

[18] G. Wahba, Y. Lin, and H. Zhang, "Gacv for support vector machines," in *Advances in Large Margin Classiers*, A. J. Smola, P. L. Bartlett, B. Scholkopf, and D. Schuurmans, Eds., pp. 297–211, 2000.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference and Prediction*, Springer, New York, NY, USA, 2001.

[20] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[21] P. Bradley and O. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Morgan Kaufmann (ICML '98)*, 1998.

[22] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm support vector machines," in *Neural Information Processing Systems 16. Massachusettes*, MIT Press, 2003.

[23] H. Zou, "An improved 1-norm SVM for simultaneous classification and variable selection," in *Proceedings of the Eleventh International Conference on Articial Intelligence and Statistics*, vol. 2, pp. 675–681, 2007.

[24] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[25] H. H. Zhang, J. Ahn, X. Lin, and C. Park, "Gene selection using support vector machines with non-convex penalty," *Bioinformatics*, vol. 22, no. 1, pp. 88–95, 2006.

[26] L. Wang, J. Zhu, and H. Zou, "The doubly regularized support vector machine," *Statistica Sinica*, vol. 16, no. 2, pp. 589–615, 2006.

[27] P. D. Sasieni, "From genotypes to genes: doubling the sample size," *Biometrics*, vol. 53, no. 4, pp. 1253–1261, 1997.

[28] G. Zheng, B. Freidlin, Z. Li, and J. L. Gastwirth, "Choice of scores in trend tests for case-control studies of candidate-gene associations," *Biometrical Journal*, vol. 45, no. 3, pp. 335–348, 2003.

[29] J. L. Gastwirth, "The use of maximin efficiency robust tests in combining contingency tables and survival analysis," *Journal of the American Statistical Association*, vol. 80, no. 390, pp. 380–384, 1985.

[30] K. N. Conneely and M. Boehnke, "So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests," *American Journal of Human Genetics*, vol. 81, no. 6, pp. 1158–1168, 2007.

[31] Q. Li, K. Yu, Z. Li, and G. Zheng, "MAX-rank: a simple and robust genome-wide scan for case-control association studies," *Human Genetics*, vol. 123, no. 6, pp. 617–623, 2008.

[32] Q. Li, G. Zheng, Z. Li, and K. Yu, "Efficient approximation of P-value of the maximum of correlated tests, with applications to genome-wide association studies," *Annals of Human Genetics*, vol. 72, no. 3, pp. 397–406, 2008.

[33] D. Kim et al., "Genome-wide genotype-based prognostic stratification of treatment outcomes following Imatinib therapy in chronic myeloid leukemia in chronic phase," In submission, 2013.

[34] D. H. Kim, J. H. Kong, J. Y. Byeun et al., "The IFNG (IFN-γ) genotype predicts cytogenetic and molecular response to imatinib therapy in chronic myeloid leukemia," *Clinical Cancer Research*, vol. 16, no. 21, pp. 5339–5350, 2010.

[35] R. Ebner, R.-H. Chen, S. Lawler, T. Zioncheck, and R. Derynck, "Determination of type I receptor specificity by the type II receptors for TGF-β or activin," *Science*, vol. 262, no. 5135, pp. 900–902, 1993.

[36] S. P. Oh, T. Seki, K. A. Goss et al., "Activin receptor-like kinase 1 modulates transforming growth factor-β1 signaling in the regulation of angiogenesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 6, pp. 2626–2631, 2000.

[37] B. Razani, X. L. Zhang, M. Bitzer et al., "Caveolin-1 regulates transforming growth factor (TGF)-β/SMAD signaling through an interaction with the TGF-β type I receptor," *The Journal of Biological Chemistry*, vol. 276, no. 9, pp. 6727–6738, 2001.

[38] M. Kawabata, A. Chytil, and H. L. Moses, "Cloning of a novel type II serine/threonine kinase receptor through interaction with the type I transforming growth factor-β receptor," *The Journal of Biological Chemistry*, vol. 270, no. 10, pp. 5625–5630, 1995.

[39] S.-J. Kim and J. Lettirio, "Transforming growth factor-β signaling in normal and malignant hematopoiesis," *Leukemia*, vol. 17, no. 9, pp. 1731–1737, 2003.

[40] N. O. Fortunel, J. A. Hatzfeld, M.-N. Monier, and A. Hatzfeld, "Control of hematopoietic stem/progenitor cell fate by transforming growth factor-β," *Oncology Research*, vol. 13, no. 6–10, pp. 445–453, 2002.

[41] B. Bierie and H. L. Moses, "Tumour microenvironment—TGFB: the molecular Jekyll and Hyde of cancer," *Nature Reviews Cancer*, vol. 6, no. 7, pp. 506–520, 2006.

[42] K. Naka, T. Hoshii, T. Muraguchi et al., "TGF-B-FOXO signalling maintains leukaemia-initiating cells in chronic myeloid leukaemia," *Nature*, vol. 463, no. 7281, pp. 676–680, 2010.

[43] K. Miyazono, "Tumour promoting functions of TGF-B in CML-initiating cells," *The Journal of Biochemistry*, vol. 152, no. 5, pp. 383–385, 2012.