An optimal kernel-based method for gene set association analysis

Tao He[1], Shaoyu Li[2], Ping-Shou Zhong[3a] and Yuehua Cui[3a]

[1]*Department of Mathematics, San Francisco State University, San Francisco, CA, 94132*

[2]*Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC, 28270*

[3]*Department of Statistics & Probability, Michigan State University, East Lansing, MI, 48824*

ABSTRACT

Single-variant based genome-wide association studies have successfully detected many genetic variants that are associated with many complex traits. However, their power is limited due to weak marginal signals and ignoring potential complex interactions among genetic variants. Set-based strategy was proposed to provide a remedy where multiple genetic variants in a given set (e.g., gene or pathway) are jointly evaluated, so that the systematic effect of the set is considered. Among many, the kernel-based testing (KBT) framework is one of the most popular and powerful methods in set-based association studies. Given a set of candidate kernels, method has been proposed to choose the one with the smallest p-value. Such a method, however, can yield inflated type I error, especially when the number of variants in a set is large. Alternatively one can get p-values by permutations which, however, could be very time consuming. In this work, we proposed an efficient testing procedure that can not only control type I error rate but also generate power close to the one obtained under the optimal kernel. Our method is built upon the KBT framework and is based on asymptotic results under a high-dimensional setting. Hence it can efficiently deal with the case where

[a]Corresponding author: {pszhong, cui}@stt.msu.edu

the number of variants in a set is much larger than the sample size. Both simulation and real data analysis demonstrate the advantages of the method compared with its counterparts.

Key words: Multiple kernels; Gene-set association; Pathway association; High dimension; Non-linear effect

1 Introduction

Driven by the advancements in microarray and next generation sequencing technologies, increasing number of genetic variants such as single nucleotide polymorphisms (SNPs), indels and copy number variation, are generated in a daily basis. Traditional genome-wide association studies (GWAS), aiming at associating single SNPs with complex traits, have been proven to be a powerful tool to unveil the genetic architecture of many complex traits. However, the power of traditional GWAS analyses by assessing the effect of SNPs one at a time, is limited due to weak marginal signals and the lack of considering potential interactions among genetic variants. Such limitation has been partially addressed by the recent wave of set-based association studies (e.g., Subramanian et al., 2005). The extension to a set-based analysis is a natural choice because genetic variants in a set (e.g., a gene or a pathway) tend to work coordinately to fulfill their joint task. On one hand, the subtle effects in multiple variants can be combined so that the joint signal of the set could be potentially boosted and be detected. On the other hand, the set-based strategy improves the power by capturing the complicated interactions among variants if any. There are a variety of biologically meaningful methods to create a SNP-set or gene-set, such as the annotated gene models (for SNP-set), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (Kanehisa, 2000), Reactome (Croft et al., 2011) and Gene Ontology (Ashburner et al., 2000).

The kernel-based testing (KBT) framework, which measures the similarity between ge-

netic variants through a kernel function then compares with the phenotype similarity, is one of the most popular and powerful methods in set-based association studies (Liu et al. 2007; Liu et al. 2008; Kwee et al. 2008). KBT is a very general framework and many other similarity based approaches (e.g., Reiss et al. 2010; Wessel and Schork 2006; Mukhopadhyay et al. 2010; Tzeng et al. 2009) are closely related to it. As observed in the literature (e.g., Wessel and Schork 2006; Wu et al. 2010; Lin et al. 2011), the power of KBT generally depends on the choice of the kernel function. Assuming that the relationship between a gene set and a disease phenotype can be described by a function $h(\cdot)$, if the true function $h(\cdot)$ comes from the function space generated by the specified kernel, then analysis based on the corresponding kernel will ideally achieve the optimal power. However, the underlying genetic function on a phenotypic response, hence the true function $h(\cdot)$, is generally unknown in practice. As a result, it is difficult to choose what kernel should be used. Given a set of candidate kernels in the KBT framework, a common practice is to choose the one leading to the smallest p-value. This, however, could inflate the type I error rate due to the choice of kernel selection. To overcome this, Wu et al. (2010) proposed a data dependent perturbation method. However, this strategy is over-conservative in a high-dimensional setting in which the number of variants could be much larger than the sample size. Moreover, it needs computationally intensive procedures to evaluate the statistical significance. The computation burden can further hamper its applicability to large scale genomic data.

In a gene-set association analysis, the number of variants (e.g., SNPs), denoted as p, is typically larger than the sample size, denoted as n, especially in a pathway-based analysis. Such a large p small n problem brings statistical challenges in developing a set-based testing procedure. Therefore, our interest is to find an efficient kernel testing procedure that can maintain nominal type I error rate while achieving high power in a high-dimensional setting (i.e., $p > n$), under the KBT framework. We mainly focus on a high-dimensional setting and assume a set of candidate kernels are given. We propose an effective and efficient

testing procedure when multiple candidate kernels are available. We introduce a new test statistic taking the maximum of the test statistics using the standardized kernels across the candidate set under a high-dimensional setting. We demonstrate the performance of the strategy through a real data application and extensive simulation studies under both continuous and discrete variable settings. The simulation studies show that the proposed approach can maintain the nominal type I error rate, and the maximum method enables the power to be close to the one obtained using the best candidate kernel function in a set, while the perturbation method proposed by Wu et al. (2010) suffers from power loss. Our method enriches the literature of kernel based association methods in genetic association studies, and has broad applications in other fields where the interest is to evaluate the joint (potentially nonlinear) effect of a set of variants with a response.

2 Statistical methods

2.1 The model setup

We assume that n independent subjects from a population are observed in a study design. For the ith subject, let Y_i be the quantitative measurement; $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^T$ be a vector of measurements for a gene set, which could be SNP genotypes in a SNP set, or gene expression profiles in a gene expression set; $\mathbf{W}_i = (W_{i1}, W_{i2}, \cdots, W_{iL})^T$ be a vector of L-dim covariates, where L is finite and $i = 1, 2, \cdots, n$. These covariates can be any clinical variables such as age, gender, and smoking status. In this work, we focus our attention on a p-dim SNP set or gene expression set, where p is assumed to be large and could be larger than the sample size n. For SNP genotype values, an additive model is assumed where X_{ij} is typically coded as 0, 1 and 2 corresponding to the number of minor alleles that subject i possesses at the jth specific locus. The gene expression values are measured as intensity in microarray studies or FPKM values in RNA-seq studies. In this work, we do not assume

any specific distribution assumption on $\mathbf{X}_i$. This makes our method more general in which it can deal with gene set based association analysis for both gene expression and SNP data. In the follows, we use gene set to denote a SNP set or gene expression set.

To model the relationship between a quantitative trait and a gene set, we consider the following semi-parametric regression model,

$$Y_i = \mu + \boldsymbol{\alpha}^T \mathbf{W}_i + h(\mathbf{X}_i) + \epsilon_i, \quad i = 1, 2, \cdots, n, \tag{1}$$

where $h(\cdot)$ is an unknown function, ϵ_i is a random subject-specific error term following a certain distribution (not necessarily normal) with $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$ and is independent of $(\mathbf{X}_i, \mathbf{W}_i)$. The identifiability of the h function is assured by the side condition $\mathrm{E}[h(\mathbf{X}_i)] = 0$. Our interest is to test association between a gene set and a continuous trait of interest, which can be done by testing the following hypotheses,

$$\mathrm{H}_0 : h(\cdot) = 0 \quad \text{vs} \quad \mathrm{H}_1 : h(\cdot) \neq 0. \tag{2}$$

2.2 Kernel function

Our method is built upon the KBT idea (Liu et al. 2007; Kwee et al. 2008), but with a different testing strategy. Before proceeding to the KBT statistic, we introduce some basics about the kernel function, which is widely used to measure the similarity between two subjects. Kernel function is commonly used to generate the functional space for the underlying true function $h(\cdot)$. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel function if it is symmetric and positive semi-definite (i.e., $K(x_1, x_2) = K(x_2, x_1)$ for any $x_1, x_2 \in \mathcal{X}$, and the $N \times N$ kernel matrix $\mathbf{K} = (K_{ij})_{i,j=1}^N$ is positive semi-definite, for any $x_1, x_2, \cdots, x_N \in \mathcal{X}$) with the (i, j)-th component $K_{ij} = K(x_i, x_j)$. In our context, $K(\mathbf{X}_i, \mathbf{X}_j)$ is a measure of similarity between the ith and the jth subject based on the SNP genotype or gene expression values.

For any positive definite kernel K^* with corresponding matrix $\mathbf{K}^*$, we can defined its centralized kernel

$$K(x_1, x_2) = K^*(x_1, x_2) - K_1^*(x_1) - K_1^*(x_2) + \mu_{K^*} \tag{3}$$

satisfying $\mathrm{E}\{K(X_1, X_2)\} = 0$, where $K_1^*(x_1) = \mathrm{E}\{K^*(x_1, X_2)\}$, and $\mu_{K^*} = \mathrm{E}\{K^*(X_1, X_2)\}$. Empirically, centralized kernel matrix $\mathbf{K}$ can be replaced by its estimator

$$\mathbf{K}_n = \mathbf{K}^* - (n-1)^{-1}[\mathbf{J}(\mathbf{K}^*)^0 + (\mathbf{K}^*)^0\mathbf{J}] + n^{-1}(n-1)^{-1}\mathbf{J}(\mathbf{K}^*)^0\mathbf{J},$$

where $\mathbf{J}$ is an $n \times n$ matrix with all the elements as 1, and $\mathbf{D}^0 = \mathbf{D} - \mathrm{diag}(\mathbf{D})$ is a zero-diagonal matrix defined for any square matrix $\mathbf{D}$ which sharing all non-diagonal elements with $\mathbf{D}$. For notation simplicity, hereafter we use K^*, K and $\mathbf{K}_n$ to represent the original kernel function, centralized kernel function and the empirical version of the centralized kernel matrix, respectively.

Some commonly used kernel functions include linear kernel $K^*(x_1, x_2) = x_1^T x_2$, polynomial kernel $K^*(x_1, x_2) = (x_1^T x_2 + c)^d$, Gaussian kernel $K^*(x_1, x_2) = \exp(-\|x_1 - x_2\|^2/\rho)$ where $c, \rho > 0, d \in \mathbb{N}$ are tuning parameters, and IBS kernel defined as $K^*(x_1, x_2) = (2p)^{-1}\sum_{j=1}^n \mathrm{IBS}(x_{1j}, x_{2j}) = (2p)^{-1}\sum_{j=1}^n (2 - |x_{1j} - x_{2j}|)$. The IBS kernel is for discrete genotype data only. For a review of genomic similarity and more kernel functions, please refer to Schaid (2010a, 2010b).

Throughout this work, we focus on centralized kernel in the testing since the asymptotic distribution of the test statistic using non-centralized kernel is largely determined by the centralized kernel except a location shift. More benefits of using centralized kernel can be found in Lindsay et al. (2008, 2014). Furthermore, we can define the standardized kernel

$$\mathcal{K}(x_1, x_2) = K(x_1, x_2)/\mathrm{E}\{K(X, X)\}$$

from which it is easy to verify that $\mathrm{E}\{\mathcal{K}(\mathbf{X}, \mathbf{X})\} = 1$. Next let us briefly look at the eigen-decomposition of a kernel function, which is an important way to characterize a kernel

function. Assume $K(\cdot,\cdot)$ is a kernel function defined on $\mathcal{X}\times\mathcal{X}$. Then the spectral decomposition theorems (Lemma 1 of Chapter 2, Steinwart and Scovel, 2012) implies that the standardized kernel $\mathcal{K}(\cdot,\cdot)$ enjoys the following representation

$$\mathcal{K}(x_1,x_2)=\sum_{m=1}^{S}\lambda_{\mathcal{K},m}\psi_m(x_1)\psi_m(x_2),\quad \forall x_1,x_2\in\mathcal{X},$$

where the eigenfunctions $\{\psi_m(\cdot)\}_{m=1}^{S}$ form a complete orthonormal system (i.e., $\mathrm{E}\{\psi_m^2(X)\}=1$ for any m, $\mathrm{E}\{\psi_m(X)\psi_{m'}(X)\}=0$ for $m\neq m'$), and $\lambda_{\mathcal{K},1}\geq\lambda_{\mathcal{K},2}\geq\cdots\geq\lambda_{\mathcal{K},S}>0$ are the non-zero eigenvalues satisfying $\sum_{m=1}^{S}\lambda_{\mathcal{K},m}=1$. The standardization is required because $\mathrm{E}\{K(\mathbf{X},\mathbf{X})\}$ could diverge in the high-dimensional case, and it ensures $\mathrm{E}\{\mathcal{K}(\mathbf{X},\mathbf{X})\}<\infty$ so that the eigen-decomposition can be properly defined. By denoting $\lambda_m=\mathrm{E}\{K(\mathbf{X},\mathbf{X})\}\lambda_{\mathcal{K},m}$, we can get the pseudo eigen-decomposition of kernel function $K(\cdot,\cdot)$

$$K(x_1,x_2)=\sum_{m=1}^{S}\lambda_m\psi_m(x_1)\psi_m(x_2),\quad \forall x_1,x_2\in\mathcal{X}.$$

It should be noticed that the eigen-decomposition not only depends on the expression of the kernel, but also implicitly depends on the space $\mathcal{X}$ (e.g., dimension p).

A functional space $\mathcal{H}_K$, namely a reproducing kernel Hilbert space (RKHS), can be generated by any positive semi-definite kernel function $K(\cdot,\cdot)$. The form of the functions that reside in $\mathcal{H}_K$ is characterized by the kernel function K. Here we assume that the $h(\cdot)$ function in model (1) is a member of the RKHS $\mathcal{H}_K$. Therefore, by specifying the kernel function, we assume that $h(\cdot)$ function has some structure defined by $\mathcal{H}_K$. For example, linear kernel indicates that the overall genetic effect is a linear combination of the individual effects in the set, i.e., $h(\mathbf{X}_i)=\beta^T\mathbf{X}_i$; polynomial kernel with $(c,d)=(1,2)$ implies a quadratic model $h(\mathbf{X}_i)=\beta^T\mathbf{X}_i+\mathbf{X}_i^T\Lambda\mathbf{X}_i$, where interactions are modeled in addition to the linear effects, β and Λ are coefficient vector and matrix, respectively. Because different kernel functions are associated with different functional spaces, the kernel based approach is very flexible for modeling different types of functions as well as complicated (potentially nonlinear) interactions among variants. On the other hand, challenges arises given that the true function

is generally unknown in practice. It is expected that the power of KBT is limited if a kernel function is misspecified. In the following sections, we start with the hypothesis testing problem using a single kernel function, followed by the ones using multiple kernel functions through which the power can be greatly boosted.

2.3 Hypothesis test based on a single kernel

We consider the following kernel-based U-statistic (KU)

$$T_n = \frac{1}{n(n-1)} \sum_{i \neq j} K(\mathbf{X}_i, \mathbf{X}_j)(Y_i - \hat{Y}_i)(Y_j - \hat{Y}_j)/\hat{\sigma}^2, \tag{4}$$

where $\hat{Y}_i$ and $\hat{\sigma}^2$ are the sample estimates under the null model $Y_i = \mu + \boldsymbol{\alpha}^T W_i + \epsilon_i$. Specifically, let $\tilde{\mathbf{W}}_{n\times(L+1)} = [\mathbf{1}_n, \mathbf{W}_{n\times L}]$, $\mathbf{A} = \tilde{\mathbf{W}}(\tilde{\mathbf{W}}^T\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}^T$, then $\hat{\mathbf{Y}} = \mathbf{A}\mathbf{Y}$ and $\hat{\sigma}^2 = \mathbf{Y}^T(\mathbf{I}-\mathbf{A})\mathbf{Y}/(n-L-1)$. Define $V_k = \sum_{m=1}^{\infty} \lambda_m^k$ for any positive integer k. Then the asymptotic normality of the test statistic T_n under the null hypothesis is stated in the following theorem.

Theorem 1 *Assume the density function of error ϵ is symmetric around 0 with $E(\epsilon_i^4) = \tau_4 < \infty$. Then, (i) under the null hypothesis of no genetic effect (i.e., $h(\cdot) = 0$),*

$$\sigma_{T_n}^{-1} n T_n \xrightarrow{d} N(0,1),$$

if

$$V_4/V_2^2 \to 0 \quad as \quad p(n) \to \infty, \tag{5}$$

where $\sigma_{T_n}^2$ is the variance of nT_n and can be estimated by the following estimator

$$\hat{\sigma}_{T_n}^2 = \frac{1}{n^2}\Big\{(2 - \frac{12}{n^2} + \frac{6\hat{\Delta}}{n})tr(\mathbf{B}^2) - (\frac{2}{n} + \frac{\hat{\Delta}}{n})tr^2(\mathbf{B}) + \hat{\Delta}tr(\mathbf{B}\circ\mathbf{B})\Big\},$$

where $\mathbf{B} = \mathbf{H}\mathbf{K}_n^0\mathbf{H}$; $\mathbf{H} = \mathbf{I} - \mathbf{A}$; $\circ$ denotes the Hadamard product (elementwise product); and $\hat{\Delta} = n^{-1}\sum_{i=1}^n [(Y_i - \hat{Y}_i)/\hat{\sigma}]^4 - 3$. (ii) Assume that $E\{\psi_m^4(\mathbf{X})\} < \infty$ for all integers m. Under

the local alternative $H_{1n} : h(\mathbf{x}) = d_n(\mathbf{x})$*, where* d_n *satisfies two conditions:* $n\delta_K = O(\sqrt{V_2})$ *with* $\delta_K = E\{K(\mathbf{X}_1, \mathbf{X}_2)d_n(\mathbf{X}_1)d_n(\mathbf{X}_2)\}$ *and* $n^2E\{d_n^8(\mathbf{X})\} = o(V_2^2/V_1^4)$*, we have*

$$\sigma_{T_n}^{-1} nT_n - \Psi(d_n) \xrightarrow{d} N(0,1)$$

if (5) holds, where $\Psi(d_n) = n\delta_K/(\sigma^2\sigma_{T_n})$ *is the location shift.*

A sketch of proof to Theorem 1 is relegated to Appendix. Given the asymptotic normality, we can then obtain the p-value for testing $H_0 : h(.) = 0$, i.e.,

$$p\text{-value} = 1 - \Phi(\sigma_{T_n}^{-1} nT_n), \tag{6}$$

where $\Phi(\cdot)$ is the cumulative density function for a standard normal distribution. As we can see from Theorem 1, the asymptotic normality holds if the condition $V_4/V_2^2 \to 0$ holds. It was mentioned earlier that this ratio depends on the kernel function, the dimension p of the space where the kernel is defined, and the probability measure on $\mathcal{X}$. To highlight the effect of dimension, define $\pi_p = V_4/V_2^2$. In the following, we take a further look at the conditions for some commonly used kernel functions.

Example 1 *Consider the linear kernel* $K^*(x_1, x_2) = x_1^T x_2$*, and assume a multivariate random variable* $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})$ *with covariance matrix* $\mathbf{\Sigma}$*,* $i = 1, \cdots, n$*. Then* $\pi_p = tr(\mathbf{\Sigma}^4)/tr^2(\mathbf{\Sigma}^2)$*.*

Example 2 *Consider the quadratic kernel* $K^*(x_1, x_2) = (x_1^T x_2 + 1)^2$*, which is a special polynomial kernel. Denote* $\mathbf{X}_i^B = (X_{i1}^2, \cdots, X_{ip}^2, \sqrt{2}X_{i1}X_{i2}, \cdots, \sqrt{2}X_{i(p-1)}X_{ip}, \sqrt{2}X_{i1}, \cdots, \sqrt{2}X_{ip})^T$ *as a* J*-dim random vector with covariance matrix* $\mathbf{\Sigma}_B$*, where* $J = (p^2 + 3p)/2$*. Then* $\pi_p = tr(\mathbf{\Sigma}_B^4)/tr^2(\mathbf{\Sigma}_B^2)$*.*

Example 3 *Consider the IBS kernel*

$$K^*(x_1, x_2) = (2p)^{-1} \sum_{m=1}^{p} IBS(x_{1m}, x_{2m}) = (2p)^{-1} \sum_{m=1} (2 - |x_{1m} - x_{2m}|).$$

Denote $\mathbf{X}_i^I = (X_{i1}, \cdots, X_{ip}, 1_{\{X_{i1}=1\}}, \cdots, 1_{\{X_{ip}=1\}})^T$ *as a* $2p$*-dim random vector with covariance matrix* $\mathbf{\Sigma}_I$*. Then* $\pi_p = tr(\mathbf{\Sigma}_I^4)/tr^2(\mathbf{\Sigma}_I^2)$*.*

The proofs to Example 1-3 are relegated to Appendix. From the above examples we can see that under the three widely-used kernels, condition (5) is equivalent to a condition on the covariance matrix of a random vector whose length depends on p. Besides, it is a weak condition that brings little constraint to the growth rate of p relative to n. Moreover, if all the eigenvalues of the covariance matrix Σ is bounded, then it is not difficult to see that π_p is of orders p^{-1}, p^{-2} and p^{-1} respectively for the linear, quadratic and IBS kernel defined earlier, and $\pi_p \to 0$ as $p \to 0$. For more discussion on the condition $\mathrm{tr}(\mathbf{\Sigma}^4)/\mathrm{tr}^2(\mathbf{\Sigma}^2) \to 0$, please refer to Chen et al. (2010).

Although the explicit condition for the covariance matrix of many kernel functions is typically unknown, there do exist consistent estimators for V_2 and V_4 that can provide us the empirical version of π_p. Specifically, $\hat{V}_2 = (P_n^2)^{-1}\mathrm{tr}\{(\mathbf{K}_n^0)^2\}$, $\hat{V}_4 = (P_n^4)^{-1}\mathrm{tr}\{(\mathbf{K}_n^0)^4\}$, $\hat{\pi}_p = \hat{V}_4/\hat{V}_2^2$, and P_n^k is the number of k-permutations of n.

2.4 Hypothesis test under multiple candidate kernels

In the previous section, we proposed a test statistic based on a single candidate kernel, and we showed its asymptotic normality under a high-dimensional setting. Since the optimal kernel is generally unknown in practice, we consider a set of M (finite) candidate kernel functions $K_1(\cdot,\cdot), K_2(\cdot,\cdot), \cdots, K_M(\cdot,\cdot)$ with kernel matrix $\mathbf{K}_{n,1}, \mathbf{K}_{n,2}, \cdots, \mathbf{K}_{n,M}$. Two testing methods are proposed under this setting. In the first one, a new kernel function is generated by taking the simple average of the normalized candidate kernels and then apply it to the single kernel based testing procedure. The second method uses a maximum test statistic and the well-developed results on multivariate normal distribution. Both methods are computationally efficient and easy to implement in practice.

2.4.1 Test based on kernel average

Without any prior knowledge of the nonparametric function $h(\cdot)$ in (1), taking the simple average among a set of normalized kernels is a natural choice, where the normalization is necessary for equal-metric consideration. In particular, denote the standardized kernels with their empirical matrix forms as

$$\mathcal{K}_m(\cdot,\cdot) = \frac{K_m(\cdot,\cdot)}{\mathrm{E}\{K_m(\mathbf{X},\mathbf{X})\}}, \quad \mathbb{K}_{n,m} = \frac{n\mathbf{K}_{n,m}}{\mathrm{tr}(\mathbf{K}_{n,m})}, \quad m = 1, 2, \cdots, M$$

and the simple average kernel with its matrix form as

$$\tilde{K}(\cdot,\cdot) = \frac{1}{M}\sum_{m=1}^{M}\mathcal{K}_m(\cdot,\cdot), \quad \tilde{\mathbf{K}}_n = \frac{1}{M}\sum_{m=1}^{M}\mathbb{K}_{n,m}.$$

Intuitively, the performance of the test using $\tilde{K}$ is most likely a compromise between the best and the worst ones. Its power will not be close to the optimal one among a candidate set, but it is a conservative option to improve the power over the weakest choice in the set given the fact that the truth is unknown in practice. We call this test as the simple average test.

2.4.2 Maximum test among a candidate set

An alternative strategy to the average kernel testing is to perform the test for individual kernels, then taking the maximum as the test statistic. Taking the maximum test statistic is the same as taking the minimum p-value which has been proposed in literature. However, the minimum p-value method often requires computationally expensive techniques such as permutation or perturbation to evaluate the null distribution. Here we focus on the maximum test statistic among all the candidate kernels and take advantage of the derived asymptotic normality under the high dimensional assumption. Let $nT_{n,m}$ and $\sigma^2_{T_{n,m}}$ be respectively the test statistic and the corresponding variance using the mth kernel function, and denote $Q_m = \sigma^{-1}_{T_{n,m}} nT_{n,m}$, $m = 1, \cdots, M$. As we can see from (6), the p-value is fully determined

by Q_m, hence maximizing Q_m is equivalent to minimizing a nonlinear function of p-values. We focus on the following maximum statistic

$$Q_{max} = \max_{1 \le m \le M} Q_m.$$

Let $\rho_{kl,n} = \mathrm{cov}(Q_k, Q_l)$ and $\rho_{kl,n} \to \rho^0_{kl}$ as $n \to \infty$, $k, l = 1, \cdots, M$. The following theorem states the the asymptotic distribution of the maximum statistic Q_{max}.

Theorem 2 *Assume condition (5) in Theorem 1 is satisfied for each candidate kernel K_m, then*

$$Q_{max} \xrightarrow{d} \max_{1 \le m \le M} Z_m,$$

where $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_M)^T$ follows a multivariate normal distribution with mean $\mathbf{0}_M$ and covariance matrix $\mathbf{\Omega}^0 = (\rho^0_{kl})$. Moreover, under the local alternative H_{1n}, the power of the maximum test achieves what the optimal one does among a candidate set, when the location shift of the optimal kernel, specified in Theorem 1, is large enough.

The proof of Theorem 2 is relegated to Appendix. Based on Theorem 2, the p-value of the maximum test can be calculated as

$$\mathrm{P}(Q_{max} > q_{max}) = 1 - \mathrm{P}(Q_{max} \le q_{max}) = [1 - P(\mathbf{Z} \le q_{max}\mathbf{1}_M)]\,\{1 + o(1)\},$$

where the leading order term can be efficiently and accurately calculated in many popular platforms (e.g., *mvnorm* package in R). Although the true covariance matrix $\mathbf{\Omega}^0$ is unknown, it can be approximately substituted by its consistent estimator $\hat{\mathbf{\Omega}}_n = (\hat{\rho}_{kl,n})$, where

$$\hat{\rho}_{kl,n} = \frac{1}{n^2}\Big\{(2 - \frac{12}{n^2} + \frac{6\hat{\Delta}}{n})\mathrm{tr}(\tilde{\mathbf{B}}_k\tilde{\mathbf{B}}_l) - (\frac{2}{n} + \frac{\hat{\Delta}}{n})\mathrm{tr}(\tilde{\mathbf{B}}_k)\mathrm{tr}(\tilde{\mathbf{B}}_l) + \hat{\Delta}\mathrm{tr}(\tilde{\mathbf{B}}_k \circ \tilde{\mathbf{B}}_l)\Big\},$$

where $\tilde{\mathbf{B}}_m = \mathbf{H}\mathbf{K}^0_{n,m}\mathbf{H}/\hat{\sigma}_{T_{n,m}}$, $i = 1, \cdots, M$. This maximum test strategy enjoys several merits. First, the nature of maximum strategy enables the best power among a set of candidate kernels. Second, the asymptotic normality results obtained under the high-dimensional

asymptotics greatly reduce our computational burden, and protects the size from being inflated or over-conservative. Although the maximum method is designed for the high-dimensional case, we found in the extensive simulation studies that the method is also applicable when the dimension p is low. Specifically, type I error rate was well-protected and only slightly conservative when p is very low (e.g., $p = 10$). Under a low-dimensional case, the distribution of Q_{max} can be approximately viewed as the maximum among M correlated chi-square random variables. Although its asymptotic behavior is beyond the scope of this paper, from our simulation studies, we found that as p grows ($p \geq 20$), the empirical type I error is very close to the nominated level and the shape of the distributions $Q_m(m = 1, ..., M)$ gets closer to a normal distribution. As the number of variants in a gene set (e.g., in a pathway) is typically large, the proposed test is generally safe to apply in practice. We call this test as the maximum test.

3 Simulation studies

Extensive simulation studies were conducted to evaluate the type I error rate and the empirical power of the proposed methods. A continuous trait was simulated from the following model,

$$Y_i = 0.03W_{i1} + 0.5W_{i2} + h(\mathbf{X}_i) + \epsilon_i, \quad i = 1, \cdots, n,$$

where ϵ_i are independent and identically distributed random errors generated from $N(0, 1)$ distribution, $W_{i1} \sim N(2, 1)$ and $W_{i2} \sim Ber(0.6)$ are independent covariates, and $\mathbf{X}_i$ is a p-dim discrete or continuous vector representing genotypes or gene expression profiles. To evaluate the type I error, we generated data sets under the null hypothesis of no association (i.e., $h(\cdot) = 0$), and recorded the proportion of (incorrectly) rejecting the null hypothesis. To assess the power, we generated data sets by specifying the h function, and recorded the proportion of (correctly) rejecting the null hypothesis. We conducted 1000 simulation

Table 1: Empirical type I error rates of different tests under the continuous variant setting

p	n	Gau	Linear	Poly	SimplyAv	Pertb	Max
50	500	0.061	0.056	0.046	0.056	0.020	0.054
	1000	0.055	0.058	0.047	0.057	0.025	0.056
	2000	0.052	0.048	0.050	0.050	0.036	0.049
100	500	0.055	0.051	0.063	0.051	0.012	0.058
	1000	0.055	0.055	0.056	0.053	0.015	0.058
	2000	0.052	0.051	0.046	0.051	0.021	0.047

replications in each case and set the significance level as 0.05. In the following, we assessed the performance of the proposed methods under the continuous and discrete variant settings separately.

3.1 Continuous variants

Under the continuous variant setting, we simulated $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})$ from a multivariate normal distribution with mean $\mathbf{0}_p$ and covariance matrix $\Gamma = (0.6^{|j-k|})$, where $p = 50, 100$ and $i = 1, \cdots, n$. The sample was assumed to be $n = 500, 1000, 2000$. The candidate set consists of three commonly used kernels, including linear kernel, polynomial kernel ($c = 1, d = 2$) and Gaussian kernel $K^*(x_1, x_2) = \exp(-\|x_1 - x_2\|^2/p)$. In addition to a single kernel based test, the kernel average method (denoted as SimplyAv), the perturbation method (denoted as Pertb) (Wu et al. 2010) and the maximum method (denoted as Max) were also applied. Table 1 reports the type I error rates of tests with varying sample size. We can see that the type I error was not well-protected using the perturbation method, and others are reasonably controlled (close to the nominal level 0.05). This finding implies that the perturbation method is relatively conservative under the high-dimensional setup, while the other method works reasonably well.

To evaluate the testing power, we considered four different scenarios. Under each scenario,

the $h(\cdot)$ function was set differently as follows:

$$
\begin{aligned}
A: \quad & h(x) = \tau_p\{0.4x_1x_3\}, \\
B: \quad & h(x) = \tau_p\{0.1x_1 + 0.1x_3 + 0.4x_1x_3\} \\
C: \quad & h(x) = \tau_p\{0.1(x_1 - x_3) + 0.8\cos(x_3)\exp(-x_3^2/5)\}, \\
D: \quad & h(x) = \sum_{k=1}^{S_p}\left\{(-0.01)^k x_k + 2\exp(-x_k^2/100)H_2(x_k/100)\right\} + 0.01\{x_1x_3 + \cos x_3^2\},
\end{aligned}
$$

where $H_k(\cdot)$ is the kth order Hermite polynomial, τ_p and S_p are two constants that were set differently for each p to adjust for the overall effect. Specifically, $(\tau_p, S_p) = (0.8, 8)$ when $p = 50$, and $(\tau_p, S_p) = (1, 30)$ when $p = 100$. For each scenario, 1000 simulation replicates were generated to estimate the empirical power. Figure 1 and Figure 2 show the empirical power under different scenarios for $p = 50$ and $p = 100$ respectively. We can see that different kernels have different powers, depending on the underlying trait architecture. Simple average kernel gives intermediate power among the candidate kernels, and the power of maximum test under each scenario was generally close to the optimal kernel. For example, under scenario A the polynomial kernel was the optimal kernel in terms of best power. Among the three competitive ones (i.e., SimplyAv, Pertb and Max), the maximum test gives power more close to the best one and performs the best among the three. The same pattern can be seen under other three scenarios for both $p = 50$ and $p = 100$ settings. It is also worth mentioning that the perturbation method suffers tremendously from power loss when the sample size is small (see cases with $n = 500$ and $n = 1000$). This implies that the Pertb method was not suitable for large p. The simulation study demonstrates that the maximum strategy is a good solution in practice to maintain proper power over the weak choices of kernels under the high-dimensional setting.

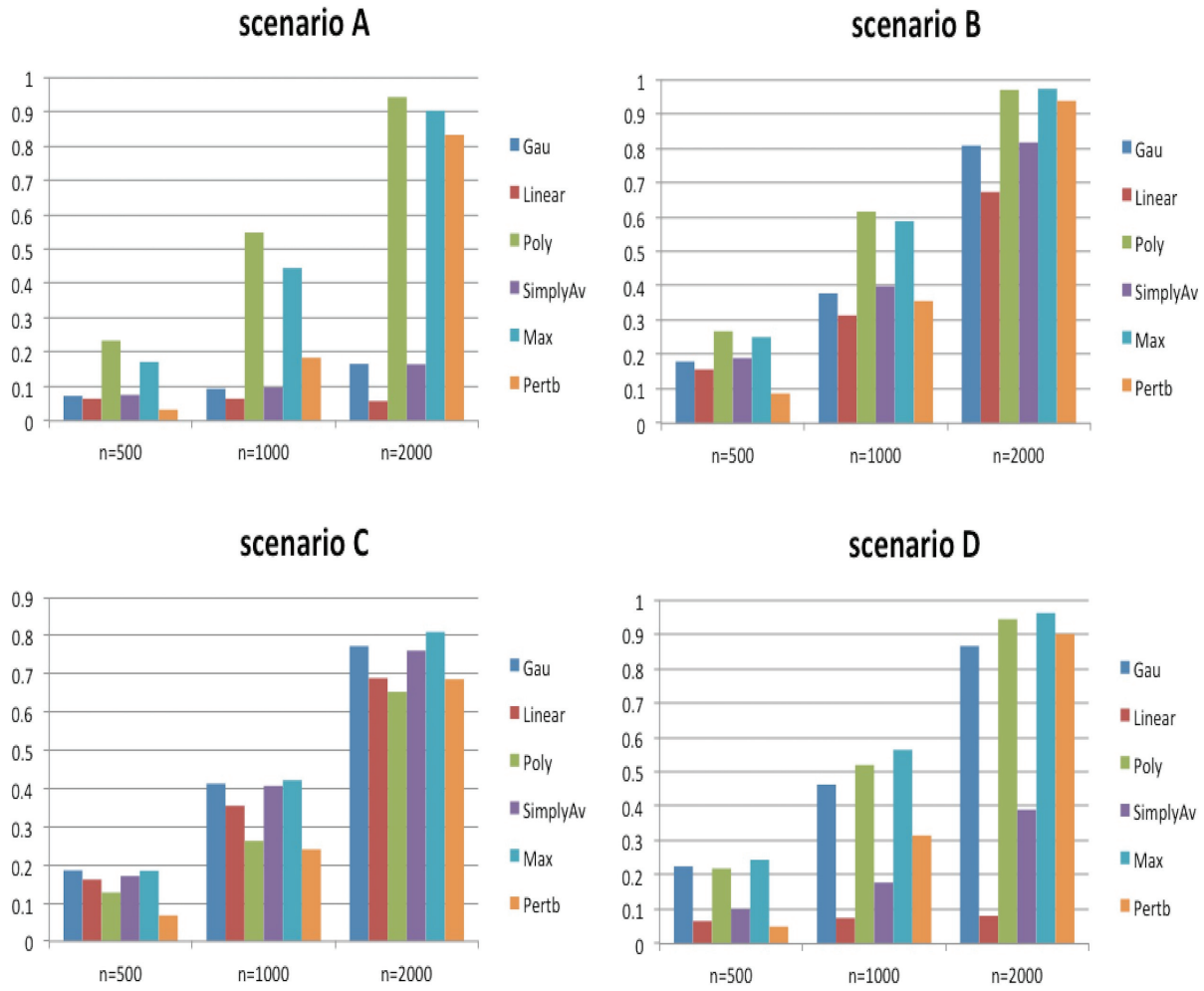

Figure 1: Empirical testing power of different tests under different scenarios and sample sizes with the continuous variant setting when $p = 50$.

3.2 Discrete variants

For the discrete variant setting, we generated genotypes based on 378 HAPMAP SNPs located within the KEGG thyroid cancer pathway using the HAPGEN software (Marchini et al. 2007). This pathway was detected as a significant pathway associated with birth weight in our real data analysis given in Section 4. We simulated the quantitative trait for $n = 1000, 2000$, under three scenarios E, F, and G. Under scenario E, we let the $h(\cdot)$ function take the form of

$$h(x) = 0.2(x_1 - x_4) + \cos(x_4)\exp(-x_4^2/5)$$

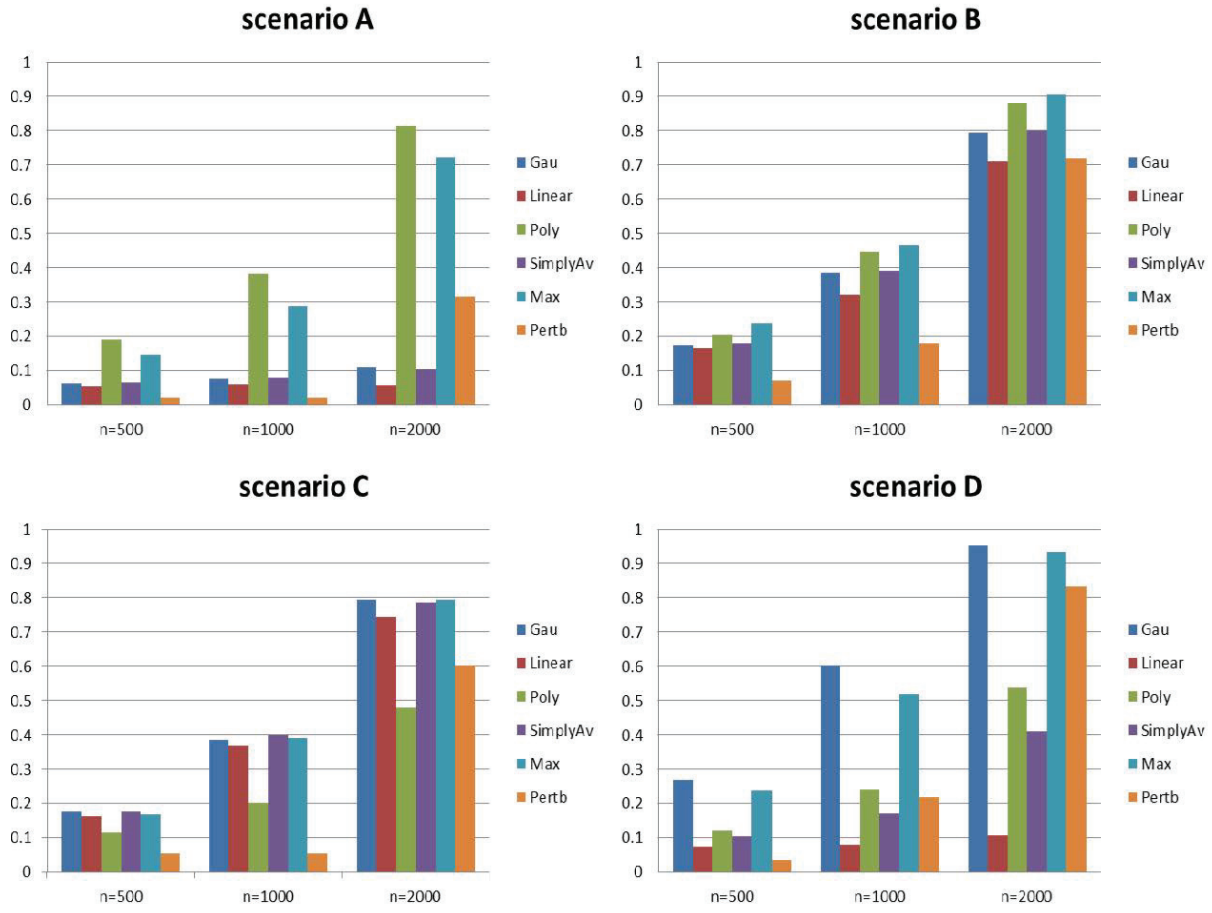

Figure 2: Empirical testing power of different tests under different scenarios and sample sizes with the continuous variant setting when $p = 100$.

where the fourth SNP has a nonlinear effect on the response in addition to the main effects of SNP 1 and 4 with different effect directions.

To mimic the situation where a large number of SNPs contributes to the trait variation, we considered the following model,

$$h(x) = a_M \sum_{k \in S_M} \beta_k x_k + a_I \sum_{(k,k') \in S_I} \alpha_{kk'} x_k x_{k'},$$

where S_M is a pre-defined set of 30 SNPs with main effects, S_I consists of 60 SNP-pairs representing 60 simple interactions. Both $\{\beta_k, k \in S_M\}$ and $\{\alpha_{kk'}, (k,k') \in S_I\}$ were independently generated from $Unif(0,\ 0.02)$, and were fixed once generated for all simula-

tion replicates. We set the coefficients $(a_M, a_I) = (0.01, 1.5)$ under scenario F, indicating the combination of weak main effects and relatively strong interaction effects. We let $(a_M, a_I) = (3.5, 0)$ under scenario G, which implies a pure main-effect model.

In addition to linear and polynomial kernels, we added the IBS kernel to the candidate set, since it is commonly used to measure SNP similarity between two subjects in genetic association studies. Similar to the previous section, the SimplyAv, Pertb and Max methods were applied. Table 2 displays the type I error rates of different tests under different sample sizes. We can see that all the tests maintained reasonable type I error rate except the Pertb method which is a little conservative. Again, the reason might be due to the high dimensionality of which the Pertb method cannot handle very well.

Table 2: Empirical type I error rates of different tests under the discrete variant setting.

n	IBS	Linear	Poly	SimplyAv	Pertb	Max
1000	0.052	0.050	0.047	0.050	0.037	0.054
2000	0.053	0.045	0.046	0.043	0.038	0.050

The power simulation results are shown in Table 3, where the best and second best powers among all the tests are shown with the underline and bold font, respectively. Again, we observed the power difference of applying different kernels. Among the different methods, the perturbation method has the smallest power which might be due to the issue of high-dimensionality. The maximum test always achieves the power as close as the best power indicating the robustness of the testing procedure by taking the maximum among the three individual ones. We also noticed the power improvement as the sample size increases.

In summary, the simulation results indicate that it is generally safe to apply the maximum test strategy given a set of candidate kernels. The maximum test can control the type I error reasonably well, while it also maintains relatively high power. Without knowing the underlying truth, the maximum test procedure is safely recommended in practice under a

Table 3: Empirical power of testing with single kernel and multiple kernels under the discrete variants setting*

n	Scenario	IBS	Linear	Poly	SimplyAv	Pertb	Max
1000	E	<u>0.526</u>	0.457	0.429	0.480	0.388	**0.488**
	F	0.397	0.412	<u>0.475</u>	0.428	0.383	**0.452**
	G	0.390	0.423	<u>0.444</u>	0.422	0.356	**0.431**
2000	E	<u>0.967</u>	0.932	0.913	0.954	0.927	**0.961**
	F	0.738	0.753	<u>0.813</u>	0.781	0.745	**0.796**
	G	0.769	0.790	<u>0.808</u>	0.798	0.748	**0.799**

* The best power across all the tests is underlined, and the second best is shown as bold font.

high-dimensional setup.

4 Application to real data

We illustrated our methods via the analysis of a Thai baby birth weight data set to investigate significant pathways that are associated with birth weight. As part of Hyperglycemia and Adverse Pregnacy Outcome (HAPO) study, this data collect genotype and phenotype information for 1209 Thai infants and their mothers. For more details about the HAPO study, please refer to `http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000096.v4.p1&phv=163690&phd=2831&pha=&pht=2446&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1`. We removed infants with large proportion of missing SNPs ($> 10\%$), and SNPs with minor allele frequency (MAF) less than 0.05 or showing deviation from Hardy-Weinberg equilibrium (p-value< 0.001). The final data set contains 970,342 SNPs in 1189 infants (580 males, 509 females). The pathways were defined by Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). SNPs that are within 5kb up- and down-stream of a gene were firstly assigned to the corresponding gene based on Human Genome Build v38, and then grouped into 186 pathways based on the KEGG pathway information retrieved from the Molecular Signature Database (MSigDB)

(Subramanian et al., 2005). The size of the pathways ranges from 167 to 9,912 (SNPs), where $> 86\%$ of the pathways are of dimension higher than 500.

We tested the association of each pathway with birth weight, adapting gender (1=male, 2=female) and baby's gestational age at delivery (in weeks) as two covariates. Since we had little knowledge about the underlying true functional mechanism, we applied three different kernels in the test, including IBS kernel, linear kernel and polynomial kernel ($c = 1, d = 2$). We applied simple average kernel test, the perturbation method (Wu et al. 2010) and the maximum statistic method. The false discovery rate was controlled using q-value significance levels (0.05 and 0.1) (Storey and Tibshirani, 2003). Table 4 summarizes the significant KEGG pathway index using different methods. The corresponding p-values and information of the significant pathways are reported in Table 5.

Table 4: Significant KEGG pathway index using different methods.

q-level	IBS	Linear	Poly	SimAv	Perturb	Max
0.10	{36,44,101,169}	{36,80,123,169}	{36,48,80,169}	{36,80,169}	NA	{36,44,80,101,169}
0.05	{101,169}	{36,80,169}	{36,169}	{169}	NA	{36,101,169}

The result shows that the perturbation method (Wu et al. 2010) fails to detect any signal, which is probably due to the over-conservative behavior under the high-dimensional setting. Among the seven distinct pathways detected by the three kernels at q-level 0.1, the maximum test was able to capture five of them, while individual kernel and simple average kernel identified four and three of them, respectively. At q-level 0.05, the observations were quite similar. One important observation is that the p-value of the maximum test is generally close to the smallest p-value among the three kernels, which implies that the maximum test tends to improve the power over the weak choice of kernels. Simply taking average did not achieve the power as the maximum test did.

Table 5: List of significant KEGG pathways and the p-values using the corresponding kernel functions.

idx	# of SNPs	Name*	IBS	Linear	Poly	SimAv	Max
169	485	KTC	1.93E-05	1.42E-07	1.32E-08	1.95E-07	1.32E-08
101	914	KP	1.71E-04	2.27E-02	2.38E-02	5.50E-03	2.84E-04
36	785	KGBCS	3.13E-03	8.14E-04	5.25E-04	9.76E-04	8.54E-04
80	914	KPSP	1.16E-02	1.06E-03	1.53E-03	2.20E-03	1.77E-03
44	1052	KAAM	1.38E-03	8.06E-02	8.28E-02	2.68E-02	2.32E-03
48	419	KGBLANS	3.55E-02	5.18E-03	3.19E-03	7.56E-03	5.41E-03
123	555	KNLRSP	1.32E-02	3.78E-03	7.43E-03	6.02E-03	5.99E-03

*KTC: KEGG thyroid cancer; KP: KEGG peroxisome; KGBCS: KEGG glycosaminoglycan biosynthesis chondroitin sulfate; KPSP: KEGG ppar signaling pathway; KAAM: KEGG arachidonic acid metabolism; KGBLANS: KEGG glycosphingolipid biosynthesis lacto and neolacto series; KNLRSP: KEGG nod like receptor signaling pathway.

5 Discussion

In this work, we developed testing procedures to test relationship between multiple variants in a gene set and a quantitative trait, while adjusting for other covariates' effects. We considered a general setting where the variants work coordinately in a (non)linear way, and the dimension of the variants p is high in the sense that p can go to infinity as sample size n goes to infinity. We first proposed a test statistic based on a single kernel function, and derived its asymptotic distribution under the null hypothesis. Based on this, we proposed a practical and efficient testing strategy when multiple candidate kernels are available. We demonstrated, via extensive simulation studies and real data analysis, that under a high-dimensional setting the maximum method can reasonably control the false positive rate while they can also improve the power over a set of weaker choices of kernels. In particular, the maximum method performs as good as the optimal one for a given set of candidate kernels, hence should be recommended in practice. Compared to the perturbation method (Wu et al., 2013), the maximum method outperformed it uniformly in various simulation settings.

Our methods enjoy several advantages as described below. The first advantage lies on the ability to accommodate high-dimensional variants and to maintain reasonable type I

error rate, even if the utilized kernel functions do not reflect the underlying relationship between the variants and the trait. Another advantage is the flexibility, which is revealed in two aspects. On one hand, we consider a general model which can potentially capture any complex interaction mechanism and is different from many models restricted to linear relation and/or linear interactions. On the other hand, when there are a range of kernels that can be selected to form the candidate set, the proposed maximum kernel testing strategy is shown to maintain improved power over the poor choices of kernels in the set, without the prior knowledge of the underlying genetic function.

Thirdly, our method is easy to implement and is free of computational burden, by applying the asymptotic result of the test statistic. This can greatly facility the applications in pathway (or gene-set) association studies where the variants (SNPs or gene expression profiles) are typically in high dimensions. The unique feature of our method under a high-dimensional setup distinguishes itself from many existing ones. Given the typical norm of high-dimensionality in gene set association studies, our method should be a good choice to implement. Our method relies on the asymptotic results where the dimension p is relatively large. Although large p is very typical in gene set association studies, in case of low dimension our method still performs well and can be an alternative to the perturbation method by Wu et al. (2010).

In our proposed methods, we only consider continuous responses. Extension to a binary response is natural and will be considered in our future investigation. Besides, our current methods were developed without prior knowledge. However, the kernel function actually allows for the inclusion of known information, such as the minor allele frequencies or association signals from an independent study. For example, weighted linear, quadratic, or IBS kernels can be constructed by assigning weights to variables individually. Thus, extension to weighted kernel is another direction that needs further investigation. With the next-generation sequencing data, identifying rare variants under the kernel machine framework

has been a standard means in rare variants detection (Wu et al. 2011). Our method can also be applied to sequencing data under the KBT framework to improve power by integrating multiple kernel functions. This will also be investigated in our future work.

Acknowledgments

This work was partially supported by grants from NSF (DMS-1209112 and DMS-1309156). Funding support for the GWA mapping: Maternal Metabolism-Birth Weight Interactions study was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01HG004415). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap through dbGaP accession number phs000096.v4.p1. Code for implementing the method was written in R, and is available at https://github.com/hetao12/multi-kernel-test.

Appendix

Sketch Proof of Theorem 1: Under the null hypothesis, the leading order of the test statistic can be written as the sum of U-statistics of different orders. The asymptotic distribution can be then studied using U-Statistic theory (Lee, 1990). Under the local alternative, the test statistics can be decomposed into two parts, where the first part corresponds to the null distribution, and the leading order of second part converges to the location shift. See detailed proof in (He et. al 2018).

Proof of Example 1: By the definition of centralized kernel in (3), we can obtain the centralized linear kernel as $K(x_1, x_2) = (x_1 - \boldsymbol{\mu})^T(x_2 - \boldsymbol{\mu})$, where $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)^T$ is the mean of random vectors $\mathbf{X}_i$, $i = 1, \cdots, n$. Assuming the covariance matrix has decomposition $\boldsymbol{\Sigma} = \boldsymbol{Q^T \Lambda Q}$ with $\boldsymbol{\Lambda}$ being the diagonal matrix. Let $\tilde{\mathbf{X}}_i = \boldsymbol{\Lambda^{-1/2} Q^T}(\mathbf{X}_i - \boldsymbol{\mu})$, where it is obvious to see $\mathrm{E}(\tilde{\mathbf{X}}_i) = \mathbf{0}$ and $\mathrm{Var}(\tilde{\mathbf{X}}_i) = \mathbf{I}$, for $i = 1, \cdots, n$. Noting that the

centralized kernel can be written as

$$K(x_1, x_2) = \tilde{\mathbf{X}}_1^T \mathbf{\Lambda} \tilde{\mathbf{X}}_2 = \sum_{m=1}^{p} \Lambda_{mm} \tilde{\mathbf{X}}_{1m} \tilde{\mathbf{X}}_{2m}.$$

We can obtain our claim by letting $\lambda_m = \Lambda_{mm}$ and $\phi_m(x_1) = \tilde{\mathbf{X}}_{1m}$, $m = 1, \cdots, p$.

Proof of Example 2: We first derive the closed form of the centralized kernel function for the quadratic kernel $K^*(x_1, x_2) = (x_1^T x_2 + 1)^2$. Decompose the kernel K^* into the sum of three parts

$$K^*(x_1, x_2) = (x_1^T x_2)^2 + 2x_1^T x_2 + 1. \tag{A.1}$$

In the following we study each part separately, because the centralized function of the K^* is essentially the sum of individual centralized functions. For the constant 1, the corresponding centralized version is 0. Since we have studied the centralized version of inner product $x_1^T x_2$ in Example 1, it remains to investigate the first term $(x_1^T x_2)^2$. It is easy to show that

$$\begin{aligned} &\mathrm{E}(x_1^T \mathbf{X}_2)^2 = x_1^T \mathbf{R} x_1, \\ &\mathrm{E}(\mathbf{X}_1^T \mathbf{X}_2)^2 = \mathrm{E}\{\mathbf{X}_1^T \mathbf{R} \mathbf{X}_1\} = \mathrm{tr}(\mathbf{R}\mathbf{\Sigma}_0) + \boldsymbol{\mu}^T \boldsymbol{R} \boldsymbol{\mu}, \end{aligned}$$

where $\mathbf{R} = (R_{ij}) = \mathbf{\Sigma}_0 + \boldsymbol{\mu}\boldsymbol{\mu}^T$ is a constant matrix, and $\boldsymbol{\mu}$, $\mathbf{\Sigma}_0$ are the mean and covariance matrix of $\mathbf{X}_i$ respectively. Thus the centralized version of $(x_1^T x_2)^2$ is

$$\begin{aligned} & (x_1^T x_2)^2 - x_1^T \mathbf{R} x_1 - x_2^T \mathbf{R} x_2 + \mathrm{tr}(\mathbf{R}\mathbf{\Sigma}_0) + \boldsymbol{\mu}^T \boldsymbol{R} \boldsymbol{\mu} \\ = & \sum_{i,j=1}^{p} (x_{1i} x_{1j} - R_{ij})(x_{2i} x_{2j} - R_{ij}) \\ = & \sum_{i=1}^{p} (x_{1i}^2 - R_{ii})(x_{2i}^2 - R_{ii}) + \sum_{i<j} (\sqrt{2} x_{1i} x_{1j} - \sqrt{2} R_{ij})(\sqrt{2} x_{2i} x_{2j} - \sqrt{2} R_{ij}). \end{aligned}$$

Combing the centralized expansions for the three terms in (A.1), we can rewrite

$$\begin{aligned} K(x_1, x_2) = & \sum_{i=1}^{p} (x_{1i}^2 - R_{ii})(x_{2i}^2 - R_{ii}) + \sum_{i<j} (\sqrt{2} x_{1i} x_{1j} - \sqrt{2} R_{ij})(\sqrt{2} x_{2i} x_{2j} - \sqrt{2} R_{ij}) \\ & + \sum_{i=1}^{p} (\sqrt{2} x_{1i} - \sqrt{2} \mu_i)(\sqrt{2} x_{2i} - \sqrt{2} \mu_i). \end{aligned}$$

Assuming random vector $\mathbf{X}_i^B = (X_{i1}^2, \cdots, X_{ip}^2, \sqrt{2}X_{i1}X_{i2}, \cdots, \sqrt{2}X_{i(p-1)}X_{ip}, \sqrt{2}X_{i1}, \cdots, \sqrt{2}X_{ip})$ follow some distribution with covariance matrix $\mathbf{\Sigma} = \boldsymbol{Q}^{\boldsymbol{T}}\boldsymbol{\Lambda}\boldsymbol{Q}$, then we can achieve our conclusion, i.e., $\pi_p = \mathrm{tr}(\mathbf{\Sigma}_B^4)/\mathrm{tr}^2(\mathbf{\Sigma}_B^2)$, by performing the similar orthogonal transformations we proposed in the proof of Example 1.

Proof of Example 3: For the IBS kernel taking the form of

$$K^*(x_1, x_2) = \frac{1}{2p}\sum_{m=1}(2 - |x_{1m} - x_{2m}|),$$

it is defied based on the total number of alleles shared identical by state (IBS) by two subjects at the SNPs within a SNP set. Noticing $X_{im} \in \{0, 1, 2\}(1 \le i \le n, 1 \le m \le p)$, it is not difficult to verify that K^* has an alternative form of

$$K^*(x_1, x_2) = \frac{1}{2p}\sum_{m=1}^{p}\frac{1}{2}(x_{1m} - 2)(x_{2m} - 2) + \frac{1}{2}x_{1m}x_{2m} + 1_{\{x_{1m}=1\}}1_{\{x_{2m}=1\}},$$

hence the centralized kernel has the following expansion

$$K(x_1, x_2) = \frac{1}{2p}\sum_{m=1}^{p}(x_{1m} - 2q_m)(x_{2m} - 2q_m) + \left[1_{\{x_{1m}=1\}} - \theta_m\right]\left[1_{\{x_{2m}=1\}} - \theta_m\right],$$

where q_m is the minor allele frequency of the mth SNP, and $\theta_m = \mathrm{P}(x_{im} = 1) = 2q_m(1 - q_m)$. Using the similar arguments as the proof of Example 1, we can obtain the result.

Proof of Theorem 2: Assume condition (5) is satisfied for each candidate kernel K_m, then

$$Q_m \xrightarrow{d} Z_m,\ m = 1, \cdots, M.$$

By using Cramèr-Wold device, $(Q_1, \cdots, Q_M)^T \xrightarrow{d} \mathbf{Z}$ where $\mathbf{Z}$ follows a standard multivariate normal distribution. Then the first conclusion can be immediately obtained through the continuous mapping theorem. Since by Theorem 1 the power of single kernel test Q_m depends on the location shift Ψ_m, we thus denote the m^*th kernel that has largest location shift Ψ_{m^*} (i.e., $m^* = \arg\max_{1\le m\le M}\Psi_m$) as the optimal kernel in the candidate set. Let $Z_{max,1-\alpha}^M$ be

the critical value of the maximum test at level α, then the power of the maximum test is

$$\begin{aligned} P(Q_{max} \geq Z^M_{max,1-\alpha}) &= 1 - P(Q_{max} < Z^M_{max,1-\alpha}) = 1 - P(Q_1 < Z^M_{max,1-\alpha}, ..., Q_M < Z^M_{max,1-\alpha}) \\ &\geq 1 - P(Q_m < Z^M_{max,1-\alpha}) = P(Q_m \geq Z^M_{max,1-\alpha}) \\ &= P(Q_m - \Psi_m \geq Z^M_{max,1-\alpha} - \Psi_m) = 1 - \Phi(Z^M_{max,1-\alpha} - \Psi_m). \end{aligned}$$

Therefore, the power of $Q_{max} \geq \max_{1\leq m\leq M} 1-\Phi(Z^M_{max,1-\alpha}-\Psi_m) = 1-\Phi(Z^M_{max,1-\alpha}-\Psi_{m^*}) = 1 - \Phi(Z_{1-\alpha} - \Psi_{m^*}) + \Phi(Z_{1-\alpha} - \Psi_{m^*}) - \Phi(Z^M_{max,1-\alpha} - \Psi_{m^*})$, where we can reach the second conclusion because $\Phi(Z_{1-\alpha} - \Psi_{m^*}) - \Phi(Z^M_{max,1-\alpha} - \Psi_{m^*}) = o(1)$ when Ψ_{m^*} is large enough.

References

[1] Ashburner, M., et al. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, **25**, 25-29.

[2] Chen, S., Zhang, L. and Zhong, P. (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association* **105**, 810-819.

[3] Croft, D., O'Kelly, G., Wu, G., et al. (2010). Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, gkq1018.

[4] He, T., Zhong, P., Cui, Y., and Mandraka, V. (2018) Tests for high-dimensional nonparametric functions in RKHS with kernel selection and regularization. Manuscript submitted for publication.

[5] Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**, 27-30.

[6] Kumar, A. (1973). Expectation of product of quadratic forms. *Sankhy: The Indian Journal of Statistics, Series B*, **35**, 359-362.

[7] Kwee, L, Liu, D., Lin, X., Ghosh, D., and Epstein, M. (2008) A powerful and flexible multilocus association test for quantitative traits. *American Journal of Human Genetics*, **82**, 386-397.

[8] Lee, A. (1990). *U-statistics: Theory and Practice*. New York, Basel: Marcel Dekker, Inc.

[9] Li, S. and Cui, Y. (2012). Gene-centric gene-gene interaction: a model-based kernel machine method. *Annals of Applied Statistics*, **6**, 1134-1161.

[10] Lin, X., Cai, T., Wu, M., Zhou, Q., Liu, G., Christiani, D. and Lin, X. (2011). Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genetic Epidemiology*, **35** , 82-93

[11] Lindsay, B., Markatou, M., Ray, S., Yang, K. and Chen, S.(2008). Quadratic distances on probabilities: a unified foundation. *The Annals of Statistics*, **36**, 983-1006.

[12] Lindsay, B., Markatou, M., and Ray, S. (2014). Kernels, Degrees of Freedom, and Power Properties of Quadratic Distance Goodness-of-Fit Tests. *Journal of the American Statistical Association*, **109**, 395-410.

[13] Liu, D., Ghosh, D. and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*,**9**, 292.

[14] Liu, D., Lin, X. and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, **63**, 1079-1088.

[15] Mukhopadhyay, I., Feingold, E., Weeks, D. and Thalamuthu, A. (2010). Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genetic Epidemiology*, **34**, 213-221.

[16] Reiss, P., Stevens, M., Shehzad, Z., Petkova, E and Milham, M. (2010). On distance-based permutation tests for between-group comparisons. *Biometrics*, **66**, 636-643.

[17] Schaid, D. (2010) Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations.*Human heredity*, **70**, 109-131.

[18] Schaid, D. (2010) Genomic Similarity and Kernel Methods II: Methods for Genomic Information. *Human heredity*, **70**, 132-140.

[19] Steinwart, I. and Scovel, C. (2012) Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, **35**, 363-417.

[20] Storey, J. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**, 9440-9445.

[21] Subramanian, A. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545-15550.

[22] Tzeng, J., Zhang, D., Chang, S., Thomas, D. and Davidian, M. (2009). Gene-trait similarity regression for multimarker-based association analysis. *Biometrics*, **65**, 822-832.

[23] Wang, K., Li, M. and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, **81**, 1278-1283.

[24] Wang, K., Li, M. and Hakonarson, H. (2010). Analyzing biological pathway in genome-wide association studies. *Nature Reviews Genetics*, **11**, 843-854.

[25] Wessel, J. and Schork, N. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. *The American Journal of Human Genetics*, **79**, 792-806.

[26] Wu, M. et al. (2010) Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, **86**, 929-942.

[27] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011) Rare Variant Association Testing for Sequencing Data Using the Sequence Kernel Association Test (SKAT). *American Journal of Human Genetics*, **89**, 82-93.

[28] Zuk, O., Hechter, E., Sunyaev, S. and Lander, E. (2012). They mystery of missing heritability: Genetic interactions creat phantom heritability. *Proceedings of the National Academy of Sciences*, **109**, 1193-1198.