A small-sample multivariate kernel machine test for microbiome association studies

Xiang Zhan[1,†], Xingwei Tong[2, †], Ni Zhao[3], Arnab Maity[4], Michael C. Wu[1, *] Jun Chen[5, *]

[1]Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109
[2]School of Mathematical Sciences, Beijing Normal University, Beijing, China
[3]Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21202
[4]Department of Statistics, North Carolina State University, Raleigh, NC 27695
[5]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN 55905

†: Joint first author

*Address for Correspondence:

Michael C. Wu
Public Health Sciences Division
Fred Hutchinson Cancer Research Center
Email: mcwu@fhcrc.org

AND

Jun Chen
Division of Biomedical Statistics and Informatics
Mayo Clinic
Email: Chen.Jun2@mayo.edu

Abstract

High-throughput sequencing technologies have enabled large-scale studies of the role of the human microbiome in health conditions and diseases. Microbial community level association test, as a critical step to establish the connection between overall microbiome composition and an outcome of interest, has now been routinely performed in many studies. However, current microbiome association tests all focus on a single outcome. It has become increasingly common for a microbiome study to collect multiple, possibly related, outcomes to maximize the power of discovery. As these outcomes may share common mechanisms, jointly analyzing these outcomes can amplify the association signal and improve statistical power to detect potential associations. We propose the multivariate microbiome regression-based kernel association test (MMiRKAT) for testing association between multiple continuous outcomes and overall microbiome composition, where the kernel used in MMiRKAT is based on Bray-Curtis or UniFrac distance. MMiRKAT directly regresses all outcomes on the microbiome profiles via a semi-parametric kernel machine regression framework, which allows for covariate adjustment and evaluates the association via a variance-component score test. Since most of the current microbiome studies have small sample sizes, a novel small-sample correction procedure is implemented in MMiRKAT to correct for the conservativeness of the association test when the sample size is small or moderate. The proposed method is assessed via simulation studies and an application to a real data set examining the association between host gene expression and mucosal microbiome composition. We demonstrate that MMiRKAT is more powerful than large-sample based multivariate kernel association test, while controlling the type I error. A free implementation of MMiRKAT in R language is available at `http://research.fhcrc.org/wu/en.html`.

Key Words: Bray-Curtis; Kernel association test; Multivariate outcomes; Small sample; UniFrac.

Conflict of Interest Statement: The authors declear no conflict of interest.

1 Introduction

The human body is inhabited by a huge and complex microbial community called the microbiome. The number of microbes that live inside and on us is estimated to be ten times the number of our somatic and germ cells. The collective genomes of our microbiome contains two orders of magnitude more genes than the genes in the human genome and contributes to our normal physiology and disease predisposition [Turnbaugh et al., 2007]. Next generation sequencing (NGS) technology enables researchers to study the human microbiome using direct DNA sequencing techniques such as 16S ribosomal DNA-targeted [Lasken, 2012] and whole genome shotgun (WGS) sequencing [Turnbaugh et al., 2009]. As a result, there has been a surge of studies interrogating the relationship between the human microbiome and a wide range of diseases and phenotypes. Through these microbiome studies, many health conditions and diseases have been linked to the disorder of the human microbiome, such as obesity [Turnbaugh et al., 2009], inflammatory bowel disease [Morgan et al., 2015] and diabetes [Qin et al., 2012], providing new insights into the etiology of human diseases.

Within the context of microbiome association studies, a popular strategy for evaluating the association between the overall microbiome composition and an outcome of interest is using ecological distances. These ecological distances, also termed as β-diversities, summarize the between-sample variability and distance-based approaches circumvent the difficulty in direct modeling of the complex microbiome sequencing data, which are usually skewed, over dispersed and zero-inflated high-dimensional count data [Li, 2015]. For a typical 16S ribosomal DNA-targeted study, the 16S sequencing tags are first clustered on the basis of their sequence similarity to form Operational Taxonomic Units (OTUs), which are considered to be surrogates of biological taxa. These OTUs are related by a phylogenetic tree, which provides important prior information on the phylogenetic relationships among biological taxa. In a typical microbiome composition data, each OTU variable takes a nonnegative integer

value representing the count of the taxon detected in a sample. One can then calculate a distance matrix among the samples based on these OTU counts, with or without taking into account the phylogenetic tree information. The UniFrac distance and the Bray-Curtis dissimilarity are widely used in this pipeline [Li, 2015]. To further assess the association between microbiome profiles and outcome variables of interest, the variability of the microbiome, which is summarized in distance measures, is compared to the variability of the outcome variables [McArdle and Anderson, 2001]. A high correspondence usually suggests existence of association.

Despite their popularity, the distance-based approaches often suffer the following limitations. First, most distance-based tests need permutations to establish significance, which can be computationally expensive. Furthermore, the distance-based analysis framework does not easily accommodate adjustment of covariates and confounders that may affect both outcomes and OTUs. Moreover, it is challenging to handle multivariate outcomes in distance-based approaches. An alternative to the distance-based approach is kernel machine methods, where the complex microbiome effects are specified through a kernel function in a semi-parametric kernel machine regression (KMR) framework, which has been widely used in genetic association studies [Wu et al., 2010, 2011b, Zhan et al., 2016]. Recently, the approach was extended and tailored to microbiome data, via development of the microbiome regression-based kernel association test (MiRKAT) [Zhao et al., 2015]. In MiRKAT, the outcome variable is regressed on OTU abundances and covariates via KMR, which can simultaneously model OTU effects nonparametrically and covariate effects parametrically. Moreover, p-value of the kernel-based test is calculated analytically via a variance-component score test. Hence, the aforementioned limitations of distance-based approaches are avoided in MiRKAT.

It has become increasingly popular for a microbiome study to collect multiple, possibly related outcomes to maximize the power of discovery [Wu et al., 2011a]. Unfortunately, MiRKAT was designed to test the association between a single outcome and overall mi-

crobiome composition, and it cannot directly handle multiple outcomes. The advantage of jointly analyzing multiple outcomes is that the association signal is amplified and easier to detect by pooling information of multiple outcomes together. The power gain of joint analysis of multiple phenotypes has been clearly demonstrated in many genetic association studies [Maity et al., 2012, O' Reilly et al., 2012, Wu and Pankow, 2016, Broadaway et al., 2016]. In the same spirit, we propose the MMiRKAT, which extends the MiRKAT framework to test the association between microbiome composition and multiple outcomes simultaneously in order to improve power.

In the framework of MMiRKAT, the association is tested by examining whether the similarity in the microbiome composition across samples resembles similarity in the multivariate outcomes, where the similarity in microbiome composition is captured by a kernel function. The p-value is calculated analytically as a variation of the variance component score test used in the multivariate phenotype association test in genetic studies [Maity et al., 2012]. Compared to genetic studies, most current microbiome studies have moderate sample sizes. The asymptotic kernel association tests derived for large-sample genetic data may be conservative for microbiome data, leading to loss of power to detect associations. Such a small-sample conservativeness problem has been well observed in univariate outcome kernel association tests [Lee et al., 2012, Chen et al., 2016], which is also expected for multivariate outcomes situations. We thus implement a novel small-sample adjustment in our MMiRKAT to correct for the potential small-sample conservativeness issue.

The rest of the paper is organized as follows. We first propose a kernel-based model which describes the association between multiple outcomes and microbiome composition. Then, we introduce MMiRKAT within this kernel-based modeling framework and incorporate a correction procedure in MMiRKAT to improve its small sample behavior. Next, we use both simulation studies and a data example from a host transcriptome-microbiome association study to illustrate and evaluate the MMiRKAT. The paper concludes with discussion.

2 Methods

2.1 A kernel model for association analysis

Suppose we observe p continuous outcome variables $\mathbf{Y}_i = (y_{i1}, \ldots, y_{ip})'$ such as the expressions of p genes, q covariates $\mathbf{X}_i = (x_{i1}, \ldots, x_{iq})'$ such as age and gender, and m biological taxa (or OTUs) $\mathbf{Z}_i = (z_{i1}, \ldots, z_{im})'$ for each individual $i = 1, \ldots, n$. We relate our outcome variables ($\mathbf{Y}$'s) to OTUs ($\mathbf{Z}$'s) and covariates ($\mathbf{X}$'s) using the following model:

$$y_{il} = x'_{it}\beta_{tl} + h_l(\mathbf{Z}_i) + \epsilon_{il}, \;\; i = 1, \ldots, n, \;\; l = 1, \ldots, p, \;\; t = 1, \ldots, q, \tag{1}$$

where $h_l(\cdot) : R^m \mapsto R$ is a real function, $(\epsilon_{i1}, \ldots, \epsilon_{ip})'$ are distributed as $N_p(0, \mathbf{\Sigma})$, $\mathbf{\Sigma}_{ll'}$ is the covariance between the lth outcome and l'th outcome, and β_{tl} are the coefficients for effect of the covariate $\mathbf{X}_t$ on the lth outcome variable.

The relationship between microbiome profiles and outcome variables is modeled nonparametrically and is fully described by functions $h_1(\cdot), \ldots, h_p(\cdot)$, where $h_l(\cdot)$ characterizes the effect of microbiome composition on the lth outcome. The objective of this paper is to test whether microbiome profiles have any effect on the outcome variables after accounting for the effect of covariates. In other words, we are interested in testing the null hypothesis $H_0 : h_1(\cdot) = \cdots = h_p(\cdot) = 0$. In this paper, we specify each function $h_l(\cdot)$ using a common kernel function $k(\cdot, \cdot)$, that is, $h_l(\cdot)$ is assumed to lie in reproducing kernel Hilbert spaces (RKHS) spanned by the positive definite kernel function $k(\cdot, \cdot)$. According to Mercer's theorem [Cristianini and Shawe-Taylor, 2000], under some regularity conditions, a positive definite kernel function $k(\cdot, \cdot)$ implicity specifies a unique Hilbert space $\mathcal{H}$. Moreover, any function $h(x) \in \mathcal{H}$ can be expressed as $h(x) = \sum_{i=1}^{L} a_i k(x, x_i)$, for some coefficients a_i and observations x_i. This is called dual representation, which states that, $h_l(\cdot)$'s can be fully determined by the kernel function $k(\cdot, \cdot)$. Although functions $h_1(\cdot), \ldots, h_p(\cdot)$ are assumed in

the same space spanned by a common kernel $k(\cdot,\cdot)$, still they can be very different since the corresponding dual representation coefficients can be different.

Intuitively, $k(\mathbf{Z}_i, \mathbf{Z}_j)$ measures the similarity between two microbiome profiles $\mathbf{Z}_i$ and $\mathbf{Z}_j$, $i, j = 1, \ldots, n$. Hence, a natural approach of selecting an appropriate kernel is to utilize the relationship between kernels and distance or dissimilarity measures [Gower, 1966, Zhan et al., 2015b]. Let $\mathbf{K}$ be the kernel matrix corresponding to kernel function $k(\cdot,\cdot)$, that is $\mathbf{K}_{ij} = k(\mathbf{Z}_i, \mathbf{Z}_j)$. Let $\mathbf{D}$ be a distance matrix, with $\mathbf{D}_{ij}$ being the distance between $\mathbf{Z}_i$ and $\mathbf{Z}_j$. Based on Gower [1966], we can construct the kernel matrix from the distance matrix as

$$\mathbf{K} = -\frac{1}{2}(\mathbf{I}_n - \frac{\underline{1}\underline{1}'}{n})\mathbf{D}^2(\mathbf{I}_n - \frac{\underline{1}\underline{1}'}{n}), \tag{2}$$

where n is the sample size, $\mathbf{I}_n$ is the nth order identity matrix, $\underline{1}$ is a n-dimensional vector of ones, and $\mathbf{D}^2$ is the element-wise matrix square.

For microbiome composition data, there are a number of different distances or dissimilarities that can be used to construct the $\mathbf{D}$ matrix in (2). The taxa are related by a phylogenetic tree, which provides important prior information on relationships among the taxa. A distance metric that exploits the degree of divergence between different sequences can lead to more meaningful results than those ignore the phylogenetic tree information. One such distance metric is the UniFrac distance family including unweighted UniFrac distance d^U [Lozupone and Knight, 2005], weighted UniFrac distance d^W [Lozupone et al., 2007] and generalized UniFrac d^θ, where $\theta \in [0, 1]$ [Chen et al., 2012]. The UniFrac distance measures the phylogenetic distance between two microbial communities as the shared fraction of the branch length of the phylogenetic tree. The unweighted version is constructed based on he presence/absence information of OTUs, while the weighted version incorporates the relative taxa abundances, and the generalized UniFrac further attenuates the weight of branches with large proportions. Another widely used distance metric for microbiome samples is the

Bray-Curtis dissimilarity, which quantifies dissimilarity between two microbial communities on the basis of OTU counts without respect to the phylogenetic tree. It can be useful when the phylogenetic tree information is unavailable or unreliable. Based on these distance or dissimilarity metrics, we will construct kernels through (2) to build our association test.

2.2 Multivariate microbiome regression-based kernel association test

2.2.1 Kernel machine regression-based testing framework

To test the association between microbiome composition and multiple outcomes, we first rewrite model (1) in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \boldsymbol{\epsilon}, \tag{3}$$

where $\mathbf{Y} = \{y_{il}\}_{n\times p}, \mathbf{X} = \{x_{it}\}_{n\times q}, \boldsymbol{\beta} = \{\beta_{tl}\}_{q\times p}, \mathbf{h} = \{h_l(Z_i)\}_{n\times p}$, and $\boldsymbol{\epsilon}$ is a $n \times p$ error matrix with each row independent and identically distributed as p-dimensional normal with mean zero and covariance matrix $\boldsymbol{\Sigma}$. To test the microbiome effect on outcomes after accounting for the covariates effect, one needs to test $H_0 : \mathbf{h} = 0$ in model (3). In most kernel-based association tests, this is done by treating $\mathbf{h}$ as a random effect in an equivalent linear mixed model [Liu et al., 2007, 2008, Wu et al., 2010, 2011b, Maity et al., 2012, Zhan et al., 2015a, 2016]. Then $H_0 : \mathbf{h} = 0$ is tested as whether the corresponding variance component is zero in that linear mixed models via a score test. In particular, by stacking the $n \times p$ outcome matrix as a np-vector, the multivariate kernel machine (MVKM) statistic was proposed [Maity et al., 2012] as

$$T_{MVKM} = (\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta}^*)'\mathbf{V}_0^{-1}\mathbf{K}^*\mathbf{V}_0^{-1}(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta}^*), \tag{4}$$

where $\mathbf{Y}^* = (y_{11}, \ldots, y_{n1}, \ldots, y_{1p}, \ldots, y_{np})'$ is a np-vector which is the vectorization of matrix $\mathbf{Y}$ in (3), $\mathbf{X}^* = diag(\mathbf{X}_1, \ldots, \mathbf{X}_p)$, where $\mathbf{X}_1 = \cdots = \mathbf{X}_p = \mathbf{X}$ and $\mathbf{X}$ is the matrix in (3), $\boldsymbol{\beta}^* = (\beta_{11}, \ldots, \beta_{q1}, \ldots, \beta_{1p}, \ldots, \beta_{qp})'$ is a qp-vector stacked based on the $\boldsymbol{\beta}$ matrix in (3), $\mathbf{K}^* = diag(\mathbf{K}_1, \ldots \mathbf{K}_p)$, where $\mathbf{K}_l$ $(l = 1, \ldots, p)$ is a $n \times n$ kernel matrix with $\mathbf{K}_l(i, j) = k(\mathbf{Z}_i, \mathbf{Z}_j)$, and $\mathbf{V}_0 = \hat{\boldsymbol{\Sigma}} \otimes \mathbf{I}_n$, where $\hat{\boldsymbol{\Sigma}}$ is the estimated variance matrix under the null $H_0 : \mathbf{h} = 0$ in model (3), $\otimes$ is kronecker product, $\mathbf{I}_n$ is the nth order identity matrix. More details can be found in Maity et al. [2012].

The MVKM statistic (4) asymptotically distributed as mixture of χ^2 variables, which is further approximated by Davies's exact method [Davies, 1980, Duchesne and De Micheaux, 2010]. The p-values calculated in such a way works sufficiently well for large sample size [Wu et al., 2011b, Wu and Pankow, 2016]. However, when the sample size is small or modest (for example, less than 1000), current kernel-based association tests developed for large sample size can be very conservative, leading to potential power loss in detecting meaningful associations, especially for binary outcomes [Lee et al., 2012] and microbiome association studies [Chen et al., 2016]. Given small-sample conservatism, a feasible approach is the permutation test. However, the permutation test can be computationally expensive and does not allow for easy covariates adjustment. Instead, we develop a new small-sample correction procedure to analytically calculate the test p-value.

2.2.2 Small-sample correction

A small-sample adjustment of kernel-based association test was proposed for a single outcome [Chen et al., 2016]. The univariate small-sample adjustment accounts for estimation error of sample variance $\hat{\sigma}^2$ when calculating the p-value. However, it is not technically straightforward to account for the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ at the multivariate scenario. To facilitate small-sample p-value calculation, we propose a variant of MVKM statistic as

$$T = \frac{tr\left\{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{K}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right\}}{tr\left\{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right\}}, \tag{5}$$

where $tr(\cdot)$ calculates the trace of a matrix, $\hat{\boldsymbol{\beta}}$ is the ordinary least squares estimate of $\boldsymbol{\beta}$ under the null model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ in (3), and $\mathbf{K}$ is the kernel matrix on OTUs with $\mathbf{K}_{ij} = k(\mathbf{Z}_i, \mathbf{Z}_j), i, j = 1, \ldots, n$. When the dimension of the outcome is one, the test statistic (5) reduces to $(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{K}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/\hat{\sigma}^2$, which is the sequence kernel association test (SKAT) statistic [Wu et al., 2011b]. Moreover, the form of the test statistic (5) is similar to many statistics widely used in a range of fields, including the distance-based pseudo-F statistic [McArdle and Anderson, 2001]. Based on the new statistic (5), we derived a new small-sample correction to calculate its p-value. The detailed mathematical derivation is included in the Appendix.

It has been shown that the power of MVKM test increase as the correlation among outcomes increases [Maity et al., 2012]. This is mainly because correlation among outcomes can be captured by $\mathbf{V}_0^{-1}$ in the MVKM statistic (4). The corresponding un-stacked covariance matrix $\hat{\boldsymbol{\Sigma}}$, however, is missing in (5) for the sake of facilitation of the small-sample correction. To take the correlation among outcomes into consideration, we propose to use the de-correlated (or de-covarianced) outcomes $\mathbf{Y}^{de} = \mathbf{Y}\hat{\boldsymbol{\Sigma}}^{-1/2}$. We termed the statistic (5) calculated from de-correlated outcomes the MMiRKAT test statistic. The corresponding M-MiRKAT test calculates its p-value by incorporating the small-sample correction procedure fully described in the Appendix.

Finally, compared to the MVKM test, the proposed MMiRKAT has two advantages. First, instead of stacking the $n \times p$ $\mathbf{Y}$-matrix into a np-dimensional $\mathbf{Y}^*$ vector, MMiRKAT approach is computational faster than the MVKM test. In particular, the computational cost of MVKM is $O((np)^3)$ depending on the inverse of the $np \times np$ matrix $\mathbf{V}_0$, while computational cost of MMiRKAT is about $\max(O(n^3), O(p^3))$ depending on the eigendecomposition of a

$n \times n$ kernel matrix and a $p \times p$ outcome covariance matrix (detailed in the Appendix). Second, by replacing the inverse of sample correlation matrix term in MVKM by the trace of that matrix in the denominator in (5), the MMiRKAT test is shown in simulation studies to have better small-sample behavior.

3 Results

3.1 Simulation studies

We conducted extensive simulation studies to evaluate the performance of MMiRKAT. For comparison, we also included the large-sample based sequence kernel association test of multiple continuous phenotypes (denoted as MSKAT) [Wu and Pankow, 2016]. The MSKAT test was shown to have similar statistical power with MVKM under most scenarios, but computationally much more efficient than MVKM. The MSKAT approach was originally proposed as a genetic association test and could be easily adapted for microbiome association analysis. Let s_{jk} denote the score test statistic for the association between the kth outcome and jth OTU, and $\mathbf{S}_j = (s_{j1}, \ldots, s_{jp})'$ be the score vector of jth OTU and all outcomes. The MSKAT statistic has been proposed as MSKAT= $\sum_{j=1}^{M} \mathbf{S}_j' \hat{\mathbf{\Sigma}}^{-1} \mathbf{S}_j$ [Wu and Pankow, 2016], where M is the total number of OTUs and $\hat{\mathbf{\Sigma}}$ is the sample outcome covariance matrix. Besides, MMiRKAT and MSKAT, we also included the Bonferroni corrected minimum p-value approach as in Wu and Pankow [2016]. Here the p-values are calculated based on applying MiRKAT [Zhao et al., 2015] to each individual outcome. We denoted the minimum p-value approach as minMiRKAT in the rest of this paper. For clarification, the univariate small-sample adjustment [Chen et al., 2016] was implemented in each individual tests in minMiRKAT.

For all simulations, we first generated OTUs of each individual from a Dirichlet-multinomial distribution, which has distribution parameters estimated based on a real upper-respiratory-

tract microbiome data set [Charlson et al., 2010], to accommodate the over-dispersion of OTU counts. We simulated in total 856 OTUs using the estimated Dirichlet-multinomial distribution and assumed that each sample had 1000 OTU counts in total. We partitioned the 856 OTUs into 20 clusters (lineages) by performing PAM (Partition Around medoids) algorithm based on the OTU distance matrix and assumed that the outcome variables depend on the abundance of a relatively abundant OTU cluster, which consists of 43 OTUs and accounts for 19.4% of the total OTUs in the real upper-respiratory-tract microbiome data. We denoted this selected functional OTU cluster by $\mathcal{A}$ hereafter.

After the OTU variables were generated, we simulated our outcome variables using the following model:

$$y_{il} = \mathbf{X}_i^{'}\beta + f \cdot scale(\sum_{j \in \mathcal{A}} \mathbf{Z}_{ij}) + \epsilon_{il}, \;\; i = 1, \ldots, n, \;\; l = 1, \ldots, p. \tag{6}$$

The covariates $\mathbf{X}_i = (X_{i1}, X_{i2})$, with X_{i1} being a binary variable and X_{i2} being a standard normal random variable. We set the true value of $\beta = (0.5, 0.5)$. The effect size $f = 0, 0.1, 0.2, 0.3, 0.4, 0.5$, and the function $scale(\cdot)$ standardized the total OTU abundance $\mathbf{Z}$'s in the associated cluster to have mean zero and standard deviation 1. The error term $(\epsilon_{i1}, \ldots, \epsilon_{ip}) \sim N_p(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ has the compound symmetry structure with $\rho = 0.2, 0.5, 0.8$ representing low, moderate and high correlation among outcomes respectively. We used notations $\Sigma_{0.2}$, $\Sigma_{0.5}$ and $\Sigma_{0.8}$ to denote the corresponding covariance matrices respectively. The sample size $n = 200$ or 1000, and the dimension of outcomes $p = 10$. Under the null models, all $p = 10$ outcomes were simulated from $y_{il} = \mathbf{X}_i^{'}\beta + \epsilon_{il}$. Under the alternative model, however, it is possible that only a subset outcomes are related to OTUs in reality. To study the effect of number of relevant outcomes on the power of our test, we simulated the first p^* of p outcomes associated with OTUs according to (6), where $p^* = 2, 8$ representing sparse and dense outcome effect respectively. The rest $(p - p^*)$ outcomes were generated as

the Gaussian noises $\epsilon_{il}, i = 1, \ldots n,\ \ l = (p^* + 1), \ldots, p$, which did not depend on the OTUs.

After the data were simulated, we applied the MMiRKAT, minMiRKAT and MSKAT to test the association between OTUs and multiple outcomes. For each test, weighted UniFrac kernel K_w, unweighted UniFrac kernel K_u, generalized UniFrac kernels K_θ, where $\theta \in [0, 1]$ and Bray-Curtis kernel K_{BC} were used in simulation. For the generalized UniFrac distance-based kernels, we used the kernel $K_{0.5}$ as suggested [Li, 2015]. We studied the type I error rate of all tests by generating $B_0 = 10000$ data sets under the null model $f = 0$, and the test power by generating $B_1 = 1000$ data sets under alternative models $f = 0.1, 0.2, 0.3, 0.4, 0.5$ respectively. The empirical type I error rate and empirical power are calculated as the proportion of data sets with a p-value smaller than the nominal significance level $\alpha = 0.05$. Besides the simulations described here, some additional simulations under more specific scenarios were also conducted and reported in the Supplementary Materials.

3.1.1 Empirical type I error

The empirical type I error rates under $n = 200$ are reported in Table i. Except for K_w-based test, empirical type I error of the MSKAT is between 0.03 and 0.035, which is consistently lower than the nominal significance level 0.05. The performance of MSKAT under this relative small sample size ($n = 200$) is quite conservative. On the other hand, both the MMiRKAT and the minMiRKAT have correct type I error rates when the correlations between outcomes are low ($\mathbf{\Sigma}_{0.2}$) or moderate ($\mathbf{\Sigma}_{0.5}$). The type I error of minMiRKAT for K_{BC} under $\Sigma_{0.2}$ is even a little inflated (0.0557), which is probably due to Monte Carlo errors. However, when the correlation between outcomes is high ($\mathbf{\Sigma}_{0.8}$), minMiRKAT is very conservative with type I error around 0.03 for all kernels. The conservativeness is due to the fact that the effective number of tests is much smaller under high correlation and thus Bonferroni correction over corrects it. minMiRKAT can become even worse if we increase the dimension of outcomes (data not shown). When we increased the sample size to $n = 1000$,

the corresponding type I errors are reported in Table ii. With large samples, MSKAT is no longer conservative. This confirms that the conservativeness is caused by small sample size, and hence it is necessary to implement the small-sample correction in MMiRKAT for microbiome association analysis. With $n = 1000$ samples, minMiRKAT is still conservative when outcomes are highly correlated. Therefore, MMiRKAT is the only test that can always protect the nominal type I error rates across all simulation scenarios.

3.1.2 Empirical power

The empirical powers of the MMiRKAT, minMiRKAT and MSKAT are presented in Figure 1 and Figure 2 for $n = 200$ and $n = 1000$ respectively. For ease of presenting, all tests are based on the weighted UniFrac kernel. Power comparison under other kernels have a similar pattern among MMiRKAT, minMiRKAT and MSKAT, and hence are not reported.

Based on Figure 1, it can be seen that MMiRKAT is consistently more powerful than MSKAT, which is expected based on the small-sample conservativeness of MSKAT observed in type I error simulations. Interestingly, minMiRKAT is more powerful when the outcome correlation is low ($\rho = 0.2$) and the effect size is large. Under such a scenario, it is less beneficial to collectively analyze all outcomes together as in MMiRKAT and MSKAT, and minMiRKAT is powerful in capturing the effect of the strongest associated outcome and is robust to other noise outcomes.

As the correlation among outcomes increases, MMiRKAT becomes more powerful than minMiRKAT. The power gain is even more dramatic when the outcomes are highly correlated ($\rho = 0.8$). Under high correlation, the amplification effect achieved by collectively analyzing all outcomes dominates the negative effect caused by adding noise outcomes in the joint analysis. A similar phenomenon has also been observed in many other multivariate association tests [Maity et al., 2012, Wu and Pankow, 2016]. In contrast, minMiRKAT suffers from power loss with increasing outcome correlations, which is consistent with the type I er-

ror study. The decay of power in minMiRKAT as increasing outcome correlation is expected to be more serious if the number of outcomes p is much larger, due to the conservativeness of Bonferroni correction. Existing methods such as TATES [van der Sluis et al., 2013] can be used to address the conservativeness of minMiRKAT under such a high dimensional and high correlation scenario. More detailed comparison is out of the scope of the current paper.

The powers of three tests under $n = 1000$ are presented in Figure 2, where similar patterns have been observed as in Figure 1. MSKAT has improved power at a larger sample size, while is still less powerful than MMiRKAT. Due to the differences (e.g. the number of outcomes and different kernels being used) between our simulations and the original MSKAT simulation [Wu and Pankow, 2016], an even larger sample size (than 1000) is needed for MSKAT to detect the microbiome association considered in this simulation.

To summarize, both MMiRKAT and MSKAT can benefit from the correlation among outcomes, which can sometimes improve the statistical power to a large extent. Under all scenarios, MMiRKAT is more powerful than MSKAT especially when the sample size is small or moderate. On the other hand, minMiRKAT focuses on the most significantly associated outcome and can be most powerful only if the outcome association signal is sparse, strong, and the correlation among outcomes is low. Under any other scenarios (e.g. dense association signals, high correlation among outcomes), the proposed MMiRKAT can be much more powerful than minMiRKAT. Finally, MMiRKAT and MSKAT is computationally faster than minMiRKAT since the minMiRKAT need to eigendecompose a $n \times n$ matrix p times while the others only need it once.

3.2 Application to a host transcriptome and microbiome association study

To illustrate the potential usefulness of MMiRKAT, we apply it to a real data set from a study of pouchitis [Morgan et al., 2015]. It is well known that both host genetics and the microbiome influence the development of pouchitis. However, how they interact is less

well-understood. To gain insight into the host-microbe interactions in the epithelial mucosa, paired host transcriptome and microbial metagenome data were collected from the large J-pouch cohort, which consists of 265 patients [Morgan et al., 2015]. Most patients were biopsied both in the pouch and in the pre-pouch ileum (PPI). The data set consists of host gene expressions and microbiome OTU counts, obtained by microarray and 16S rRNA analysis respectively. After quality control, 255 samples (196 PPI samples and 59 pouch samples) are available. Besides host transcriptome and microbial data, some clinical metadata such as antibiotic use (yes/no), inflammation score (0-13), and disease outcome (familial adenomatous polyposis/FAP and non-FAP) are also available. This dataset is publicly available [Morgan et al., 2015].

The original analysis [Morgan et al., 2015] for testing the association between transcriptome and microbiome is based on multivariate analysis with linear modeling (MaAsLin). In particular, a multivariate linear model: gene $\sim$ OTU + antibiotic + inflammation score + outcomeFAP/non-FAP was used and the OTU regression coefficient was tested whether being zero. Only 196 PPI samples were used and hence tissue location (PPI/pocuh) was not adjusted in the linear model. In MaAsLin, one transcript and one OTU was tested each time. Since 33297 host transcripts and 7000 OTUs had been measured in the data, principal component analysis (PCA) based dimension reduction was performed in both transcripts and OTUs to reduce multiple testing burden of MaAsLin. In particular, 9 gene PC features (gPC) and 9 clade PC features were selected in order to explain 50% of total variance in host transcripts and OTUs respectively. Finally, MaAsLin was able to claim significance at a false discovery rate of 0.25 [Morgan et al., 2015].

Alternative to the individual analysis as conducted in MaAsLin, we performed the joint analysis MMiRKAT in this section. Both the host gene expression and OTUs were analyzed collectively. In particular, a Bray-Curtis kernel was constructed based on all the 7000 OTU counts, and the 9 gPCs were tested for association with all OTUs simultaneously. MMiRKAT

requires $n > p$, hence we use the 9 gPCs as outcomes rather than all 33297 host transcripts. We incorporate the additional 59 pouch samples by adding a new covariate of tissue location (1=PPI, 0=pouch). The working model for MMiRKAT is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \boldsymbol{\epsilon},$$

where $\mathbf{Y} = \{y_{il}\}_{255\times 9}$, $\mathbf{X} = \{x_{it}\}_{255\times 4}$, $\boldsymbol{\beta} = \{\beta_{tl}\}_{4\times 9}$, each column of $\mathbf{h}$ is distributed as $N_n(0, \tau K_{BC})$, and each row of $\boldsymbol{\epsilon}$ is distributed as $N_p(0, \boldsymbol{\Sigma})$ with unknown covariance matrix $\boldsymbol{\Sigma}$. Four covariates are antibiotic use, inflammation score, outcomeFAP/non-FAP and tissue location. Besides MMiRKAT, minMiRKAT and MSKAT were also implemented. The p-values are 1.3×10^{-5}, 8.3×10^{-4} and 3.6×10^{-3} respectively. MMiRKAT has the smallest p-value of the three, which indicates that it is most powerful in detecting potential association signal between host transcriptome and mucosal microbiome while adjusting for effects of covariates. A more detailed individual testing results are available in Table S5 of the online Supplementary Materials. Compared with the detected significance at the FDR level 0.25 of MaAsLin in the original study [Morgan et al., 2015], the association testing results of MMiRKAT, minMiRKAT and MSKAT are much more significant. This is probably because all MMiRKAT, minMiRKAT and MSKAT test the association between outcomes and the whole microbiome community rather than each OTU individually. This is concordant with recent findings that individual association between microbiome and metadata is often small and requires very large sample sizes to detect [Falony et al., 2016, Zhernakova et al., 2016]. Given the relatively small sample size ($n = 255$) in this host transcriptome and microbiome data, it is not surprising that microbiome community level analyses (MMiRKAT, minMiRKAT and MSKAT) are more powerful than individual analysis (MaAsLin).

To further demonstrate potential usefulness of the proposed method, we restricted our association analysis to a smaller gene set rather than the PCs of all 33297 host transcripts. It

has been found that the interleukin-12 pathway contains genes which have large magnitude of loadings on those microbially-associated gPCs [Morgan et al., 2015]. Inspired by this observation, we were interested in testing the association between the interleukin-12 pathway gene expression and OTUs. Twenty-one interleukin-12 pathway genes were found in this host transcriptome and microbiome data set. We used the expression values of these 21 genes as outcomes to test association with overall microbiome composition. The correlation among these gene expressions are reported in Figure S6 of the Supplementary Materials, where one can see that there are some high correlations among the gene expressions. The p-values of MMiRKAT, minMiRKAT and MSKAT are 0.011, 0.230, 0.028 respectively. MMiRKAT and MSKAT are much more significant than minMiRKAT. This is because both MMiRKAT and MSKAT can take advantage of the high correlations among the outcomes to boost the power of detecting associations. A more detailed individual gene testing results are reported in Table S6, where most individual association signals are relative weak, which explains the insignificant testing result of minMiRKAT.

Our analysis can not only provide formal testing for overall association, but also gain new biological insights on related pathways. Lastly, we emphasize that the proposed MMiRKAT is a global test and is conducted on the microbiome community level. It is the first step to associate the overall microbiome composition with multiple outcomes without multiple testing. More specific tests are the next step to improve biological interpretability, such as which group of taxa are more associated with outcomes. Such goals can be accomplished by many variable selection methods.

4 Discussion

In this paper, we propose the MMiRKAT to test for the association between microbial community composition and multiple outcomes of interest. Similar to MiRKAT [Zhao et al., 2015], MMiRKAT is able to control for potential confounding effects of covariates with-

in a principled regression framework by modeling the covariate effect parametrically and microbiome effect nonparametrically. Beyond that, MMiRKAT enjoys the following three additional nice features: 1) MMiRKAT can handle multiple outcomes simultaneously, which can improve the power to detect association due to amplification of the signals when outcomes are correlated. 2) MMiRKAT is computational efficient since it calculates the p-value analytically and works on relatively smaller matrix compared to the stacked MVKM statistic. 3) MMiRKAT has good performance even when the sample size is small or moderate. As a comparison, MVKM [Maity et al., 2012] only enjoys 1) and MSKAT [Wu and Pankow, 2016] only enjoys 1) and 2). Both simulation studies and real data analysis were conducted to illustrate and evaluate the MMiRKAT approach. Through those numerical studies, it has been shown that MMiRKAT can control the type I error and is overall more powerful than other existing methods in detecting potential association signals. Therefore, MMiRKAT provides a statistically powerful and computationally fast way to test associations between microbiome community composition and multiple outcomes of interest.

The current MMiRKAT approach requires sample size n to be greater than dimension of outcomes p, since $\hat{\mathbf{\Sigma}}^{-1/2}$ is calculated in the de-correlation procedure. The same condition is also required on other multivariate tests such as MVKM and MSKAT. To fix this issue in high-dimensional setting where $p > n$, we propose to perform principal component analysis (PCA) on the outcomes, just as illustrated in the host transcriptome and microbiome data. Then, some top PCs are taken as new outcome variables, and association is tested between those PCs and microbiome compositions. Both phylogeny-based UniFrac kernels and non-phylogeny-based Bray-Curtis kernel have been used in MMiRKAT. Different kernels represent different views of the microbial community. Clearly, one kernel is the best when it captures the relationship between the outcome and the microbiome composition. Therefore, each kernel has the best performing scenario depending on the underlying biological models [Chen et al., 2012]. It is helpful to consider several representative kernels in the association

tests in order not to miss important associations. A robust approach is to conduct an omnibus test which combines multiple candidate kernels. Due to page limit, we do not pursue such an omnibus test in the paper. Interested readers are referred to Wu et al. [2013] and Zhao et al. [2015] for further references.

Despite the fact that the MMiRKAT method presented in this paper aimed at testing the associations between multiple outcomes and microbiome composition, one can apply similar techniques to other types of "omics" data, such as the SNPs, gene expression, methylation, proteomics and metabolomics. MMiRKAT can be easily adapted to testing the association between multiple outcomes and other data types. One only need to replace the microbiome kernels with other suitable kernels which can accommodate the important features of that data type and apply the same testing methodology in MMiRKAT presented in this paper. It can be a potential useful association analysis tool in proteomics and metabolomics association studies, where sample size is usually relatively small.

In this paper, we focus on multiple continuous outcome variables. Some kernel-based association test for binary outcomes are also available in literature [Liu et al., 2008, Wu et al., 2010]. In the spirit of those works, the current continuous outcomes based MMiRKAT can be easily extended to the case of binary outcomes. With the development of techniques, there is a growing interest in applying microbiome studies to complex clinical and population-based studies. One issue in those complicated studies is the accommodation of more sophisticated outcomes (such as longitudinal outcomes and other highly structured outcomes). We will explore those extensions in future work.

Acknowledgments

We want to thank the editor and two reviewers for their constructive comments that have greatly improved the paper, and Dr. Baolin Wu for providing the MSKAT code.

References

K Alaine Broadaway, David J Cutler, Richard Duncan, Jacob L Moore, Erin B Ware, Min A Jhun, Lawrence F Bielak, Wei Zhao, Jennifer A Smith, Patricia A Peyser, et al. A statistical approach for testing cross-phenotype effects of rare variants. *The American Journal of Human Genetics*, 98(3):525–540, 2016.

Emily S Charlson, Jun Chen, Rebecca Custers-Allen, Kyle Bittinger, Hongzhe Li, Rohini Sinha, Jennifer Hwang, Frederic D Bushman, and Ronald G Collman. Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PloS One*, 5(12):e15216, 2010.

Jun Chen, Kyle Bittinger, Emily S Charlson, Christian Hoffmann, James Lewis, Gary D Wu, Ronald G Collman, Frederic D Bushman, and Hongzhe Li. Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics*, 28(16):2106–2113, 2012.

Jun Chen, Wenan Chen, Ni Zhao, Michael C Wu, and Daniel J Schaid. Small sample kernel association tests for human genetic and microbiome association studies. *Genetic Epidemiology*, 40(1):5–19, 2016.

Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

Robert B Davies. Algorithm as 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3): 323–333, 1980.

Pierre Duchesne and Pierre Lafaye De Micheaux. Computing the distribution of quadratic forms: Further comparisons between the liu–tang–zhang approximation and exact methods. *Computational Statistics & Data Analysis*, 54(4):858–862, 2010.

Gwen Falony, Marie Joossens, Sara Vieira-Silva, Jun Wang, Youssef Darzi, Karoline Faust, Alexander Kurilshikov, Marc Jan Bonder, Mireia Valles-Colomer, Doris Vandeputte, et al. Population-level analysis of gut microbiome variation. *Science*, 352(6285):560–564, 2016.

John C Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.

Roger S Lasken. Genomic sequencing of uncultured microorganisms from single cells. *Nature Reviews Microbiology*, 10(9):631–640, 2012.

Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, ESP Lung Project Team, David C Christiani, Mark M Wurfel, Xihong Lin, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, 2012.

H Li. Microbiome, metagenomics and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

Dawei Liu, Xihong Lin, and Debashis Ghosh. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088, 2007.

Dawei Liu, Debashis Ghosh, and Xihong Lin. 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(1):292, 2008.

Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, 2005.

Catherine A Lozupone, Micah Hamady, Scott T Kelley, and Rob Knight. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, 73(5):1576–1585, 2007.

Arnab Maity, Patrick F Sullivan, and Jun-ing Tzeng. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic Epidemiology*, 36(7):686–695, 2012.

Brian H McArdle and Marti J Anderson. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82(1):290–297, 2001.

Xochitl C Morgan, Boyko Kabakchiev, Levi Waldron, Andrea D Tyler, Timothy L Tickle, Raquel Milgrom, Joanne M Stempak, Dirk Gevers, Ramnik J Xavier, Mark S Silverberg, et al. Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biology*, 16(1):67, 2015.

Paul F O' Reilly, Clive J Hoggart, Yotsawat Pomyen, Federico CF Calboli, Paul Elliott, Marjo-Riitta Jarvelin, and Lachlan JM Coin. Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PloS One*, 7(5):e34861, 2012.

Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012.

Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, 449(7164):804, 2007.

Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, 2009.

Sophie van der Sluis, Danielle Posthuma, and Conor V Dolan. Tates: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genetics*, 9(1): e1003235, 2013.

B Wu and JS Pankow. Sequence kernel association test of multiple continuous phenotypes. *Genetic Epidemiology*, 40(2):91–100, 2016.

Gary D Wu, Jun Chen, Christian Hoffmann, Kyle Bittinger, Ying-Yu Chen, Sue A Keilbaugh, Meenakshi Bewtra, Dan Knights, William A Walters, Rob Knight, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011a.

Michael C Wu, Peter Kraft, Michael P Epstein, Deanne M Taylor, Stephen J Chanock, David J Hunter, and Xihong Lin. Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942, 2010.

Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011b.

Michael C Wu, Arnab Maity, Seunggeun Lee, Elizabeth M Simmons, Quaker E Harmon, Xinyi Lin, Stephanie M Engel, Jeffrey J Molldrem, and Paul M Armistead. Kernel machine snp-set testing under multiple candidate kernels. *Genetic Epidemiology*, 37(3):267–275, 2013.

Xiang Zhan, Michael P Epstein, and Debashis Ghosh. An adaptive genetic association test using double kernel machines. *Statistics in Biosciences*, 7(2):262–281, 2015a.

Xiang Zhan, Andrew D Patterson, and Debashis Ghosh. Kernel approaches for differential expression analysis of mass spectrometry-based metabolomics data. *BMC Bioinformatics*, 16:77, 2015b.

Xiang Zhan, Santhosh Girirajan, Ni Zhao, Michael C Wu, and Debashis Ghosh. A novel copy number variants kernel association test with application to autism spectrum disorders studies. *Bioinformatics, In press, DOI: 10.1093/bioinformatics/btw500*, 2016.

Ni Zhao, Jun Chen, Ian M Carroll, Tamar Ringel-Kulka, Michael P Epstein, Hua Zhou, Jin J Zhou, Yehuda Ringel, Hongzhe Li, and Michael C Wu. Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test. *The American Journal of Human Genetics*, 96(5):797–807, 2015.

Alexandra Zhernakova, Alexander Kurilshikov, Marc Jan Bonder, Ettje F Tigchelaar, Melanie Schirmer, Tommi Vatanen, Zlatan Mujagic, Arnau Vich Vila, Gwen Falony, Sara Vieira-Silva, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, 352(6285):565–569, 2016.

Appendix

Derivation of small-sample correction

Without loss of generality, we omit the superscript of the de-correlated outcomes $\mathbf{Y}^{de}$ for ease of presenting and use $\mathbf{Y}$ instead. The same kernel machine regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \boldsymbol{\epsilon}$$

and the same statistic

$$T = \frac{tr\left((\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{K}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right)}{tr\left((\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right)}$$

are assumed in this appendix section. The intuition of T is from the univariate small-sample correction [Chen et al., 2016], where mathematical derivation has been developed to account for the variability of the variance estimator in the denominator, in order to derive a distribution that is closer to the true finite-sample distribution. In the multivariate case, the exact form of T facilitates a multivariate small-sample adjustment described below.

Existing tests [Maity et al., 2012, Wu and Pankow, 2016] often calculate their p-values based on asymptotic distribution of T. However, it has been observed that the p-values calculated in this way is often over-protected when the sample size is small or moderate [Chen et al., 2016]. To overcome the potential small-sample conservatism, we propose the following procedure to calculate the p-value. Let $\mathbf{P} := \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ be the projection matrix of the null model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{I}_n$ is the nth order identity matrix. Then the test statistic can be calculated as

$$T = \frac{tr\left((\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{K}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right)}{tr\left((\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right)} = \frac{tr\left(\boldsymbol{\epsilon}'\mathbf{P}\mathbf{K}\mathbf{P}\boldsymbol{\epsilon}\right)}{tr\left(\boldsymbol{\epsilon}'\mathbf{P}\boldsymbol{\epsilon}\right)},$$

and we have that

$$
\begin{aligned}
Pr\left(\frac{tr\left((\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{K}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})\right)}{tr\left((\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})\right)} > t_0\right) &= Pr\left(\frac{tr\left(\boldsymbol{\epsilon}'\mathbf{PKP}\boldsymbol{\epsilon}\right)}{tr\left(\boldsymbol{\epsilon}'\mathbf{P}\boldsymbol{\epsilon}\right)} > t_0\right) \\
&= Pr\left(tr\left(\boldsymbol{\epsilon}'\mathbf{P}[\mathbf{K}-t_0\mathbf{I}_n]\mathbf{P}\boldsymbol{\epsilon}\right) > 0\right) \\
&= Pr\left(\sum_{i=1}^{n}\lambda_i tr(\epsilon_i\epsilon_i') > 0\right) \\
&= Pr\left(\sum_{i=1}^{n}\lambda_i\epsilon_i'\epsilon_i > 0\right),
\end{aligned} \tag{7}
$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the matrix $\mathbf{P}[\mathbf{K}-t_0\mathbf{I}_n]\mathbf{P}$. Let $\epsilon_i := \boldsymbol{\Sigma}^{1/2}\eta_i$ where $\boldsymbol{\Sigma} = Cov(\epsilon_i)$. Then $\eta_i \sim N(0, \mathbf{I}_p)$ and thus

$$
Pr\left(\sum_{i=1}^{n}\lambda_i\epsilon_i'\epsilon_i > 0\right) = Pr\left(\sum_{i=1}^{n}\lambda_i\eta_i'\Sigma\eta_i > 0\right) = Pr\left(\sum_{i=1}^{n}\lambda_i\sum_{j=1}^{p}\mu_j\eta_{ij}^2 > 0\right), \tag{8}
$$

where $\mu_i, \ldots, \mu_p$ are the eigenvalues of $\boldsymbol{\Sigma}$. Since η_{ij}^2 follows χ^2 distribution with 1 degree of freedom, then combining (7) and (8), we have

$$
Pr\left(\frac{tr\left((\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{K}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})\right)}{tr\left((\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})\right)} > t_0\right) = Pr\left(\sum_{i=1}^{n}\lambda_i\sum_{j=1}^{p}\mu_j\chi_{ij}^2(1) > 0\right), \tag{9}
$$

where $\chi_{ij}^2(1), i = 1, \ldots, n, j = 1, \ldots, p$ are i.i.d. χ^2 random variables with 1 degree of freedom. The last probability on the right hand side of equation (9) can be calculated using the Davies's exact method [Davies, 1980, Duchesne and De Micheaux, 2010]. The procedure requires the p eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$. In practice, we use the eigenvalues of the sample covariance matrix of the residuals. Extensive simulation studies have been conducted in this paper to evaluate the proposed small-sample correction procedure. Based on the those numerical studies, our small-sample correction procedure works very well as long as the dimension of the outcomes is smaller than the sample size.

Table i: Empirical type I error of MMiRKAT, minMiRKAT and MSKAT when $n = 200$.

Method	Σ	K_w	K_u	$K_{0.5}$	K_{BC}
MMiRKAT	$\Sigma_{0.2}$	0.0464	0.0489	0.0456	0.0434
	$\Sigma_{0.5}$	0.0446	0.0461	0.0460	0.0431
	$\Sigma_{0.8}$	0.0428	0.0457	0.0474	0.0447
minMiRKAT	$\Sigma_{0.2}$	0.0516	0.0512	0.0519	0.0557
	$\Sigma_{0.5}$	0.0392	0.0527	0.0448	0.0445
	$\Sigma_{0.8}$	0.0270	0.0280	0.0336	0.0277
MSKAT	$\Sigma_{0.2}$	0.0422	0.0338	0.0322	0.0355
	$\Sigma_{0.5}$	0.0435	0.0327	0.0318	0.0338
	$\Sigma_{0.8}$	0.0403	0.0314	0.0337	0.0344

Table ii: Empirical type I error of MMiRKAT, minMiRKAT and MSKAT when $n = 1000$.

Method	Σ	K_w	K_u	$K_{0.5}$	K_{BC}
MMiRKAT	$\Sigma_{0.2}$	0.0503	0.0485	0.0503	0.0496
	$\Sigma_{0.5}$	0.0451	0.0511	0.0502	0.0550
	$\Sigma_{0.8}$	0.0503	0.0485	0.0444	0.0499
minMiRKAT	$\Sigma_{0.2}$	0.0510	0.0485	0.0468	0.0500
	$\Sigma_{0.5}$	0.0375	0.0387	0.0399	0.0416
	$\Sigma_{0.8}$	0.0247	0.0283	0.0275	0.0281
MSKAT	$\Sigma_{0.2}$	0.0497	0.0438	0.0448	0.0495
	$\Sigma_{0.5}$	0.0478	0.0467	0.0484	0.0493
	$\Sigma_{0.8}$	0.0496	0.0448	0.0413	0.0471

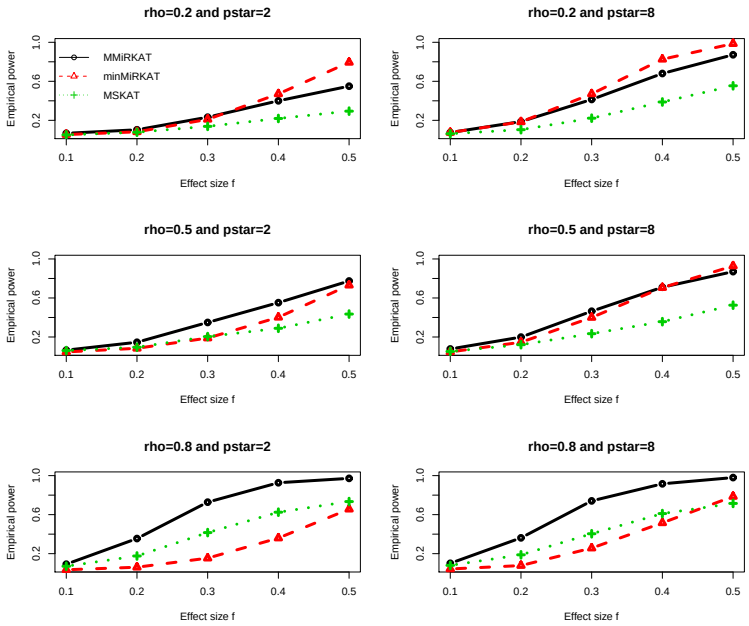

Figure 1: Empirical powers of MMiRKAT, minMiRKAT and MSKAT under $n = 200$. $\circ$, $\triangle$ and $+$ represent MMiRKAT, minMiRKAT and MSKAT respectively.

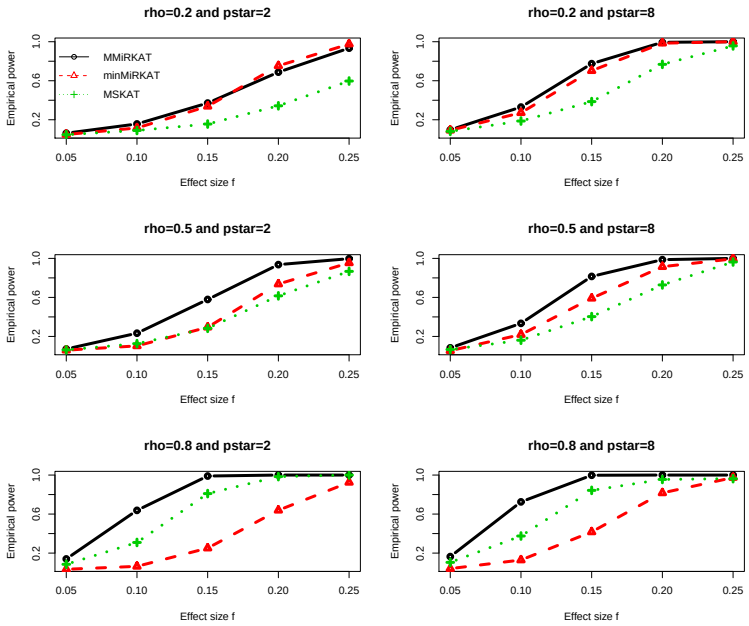

Figure 2: Empirical powers of MMiRKAT, minMiRKAT and MSKAT under $n = 1000$. $\circ$, $\triangle$ and $+$ represent MMiRKAT, minMiRKAT and MSKAT respectively.