MiRKAT Package

Haotian Zheng, Xiang Zhan, Anna Plantinga, Michael Wu, Ni Zhao

July 25, 2017

1 Overview

MiRKAT package (v1.0) has functions to test the association between a microbiome community and phenotype of different types, such as univariate continuous or binary phenotypes, survival outcomes, multivariate and structured phenotypes. For all these effect, the microbiome community effect was modeled nonparametrically through a kernel function, which can incorporate the phylogenetic tree information.

2 Changes from v0.02

Three additional functions are included in the package than v0.02. They were known as MiRKAT-S (for survival outcome), MMiRKAT (multivariate outcome), and KRV (for structured outcome).

3 Required functions

The packages "survival", "PearsonDS", "GUniFrac" and "MASS" are required for MiRKAT. All these four required packages are available on CRAN (https://cran.r-project.org/web/packages/).

4 MiRKAT: association testing between microbiome composition and a continuous or binary outcome

4.1 Example Dataset

We use the throat microbiome data (Charlson et al 2010) from package GUniFrac to demonstrate our methods. The throat data contains 60 subjects with 28 smokers and 32 nonsmokers. Microbiome data were collected from right and left nasopharynx and oropharynx regions to form an OTU table with 856 OTUs. We want to evaluate whether smoking can affect the microbiome composition in

the upper respiratory tract, taking into consideration additional covariates including gender and antibiotic use within 3 months. We also use simulated data based on the throat dataset to demonstrate the usage of MiRKATS, MMiRKAT and KRV.

```
> library(MiRKAT, quietly=TRUE)
> library(GUniFrac, quietly=TRUE)
> data(throat.tree)
> data(throat.otu.tab)
> data(throat.meta)
> attach(throat.meta)
```

4.2 Prepare the data

```
> set.seed(123)
> Male = (Sex == "Male")**2
> Smoker =(SmokingStatus == "Smoker") **2
> anti =  (AntibioticUsePast3Months_TimeFromAntibioticUsage != "None")^2
> cova = cbind(Male,  anti)
> detach(throat.meta)
```

4.3 Create the UniFrac Distances

Many options exist for distance or dissimilarity metrics. Below, we calculate four distances: the unweighted and weighted UniFrac distances, the generalized UniFrac distance with $\alpha = 0.5$, and the Bray-Curtis dissimilarity.

```
> otu.tab.rff <- Rarefy(throat.otu.tab)$otu.tab.rff
> unifracs <- GUniFrac(otu.tab.rff, throat.tree, alpha=c(0, 0.5, 1))$unifracs
> D.weighted = unifracs[,,"d_1"]
> D.unweighted = unifracs[,,"d_UW"]
> D.generalized = unifracs[,,"d_0.5"]
> D.BC= as.matrix(vegdist(otu.tab.rff , method="bray"))
>
```

4.4 Convert Distances to kernel matrices

The D2K function in MiRKAT converts distance matrices to kernel matrices via

$$K = -\frac{1}{2}\left(I - \frac{11'}{n}\right) D^2 \left(I - \frac{11'}{n}\right).$$

Here, I is the identity matrix and 1 is an n-vector of ones. To ensure that K is positive semi-definite, we replace negative eigenvalues with zero. That is, we perform an eigenvalue decomposition $K = U\Lambda U$, where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_n)$, and then reconstruct the kernel matrix using the nonnegative eigenvalues $\Lambda^* = \text{diag}(\max(\lambda_1, 0), ..., \max(\lambda_n, 0))$ so that $K = U\Lambda^* U$.

```
> K.weighted = D2K(D.weighted)
> K.unweighted = D2K(D.unweighted)
> K.generalized = D2K(D.generalized)
> K.BC = D2K(D.BC)
```

4.5 Testing using a single kernel

```
> MiRKAT(y = Smoker, Ks = K.weighted, X = cbind(Male, anti),
+        out_type = "D", method = "davies")

[1] 0.004654662
```

"Method" indicates which method to use to compute kernel specific p-value. "davies" represents an exact method that computes the p-value by inverting the characteristic function of the mixture chi square distribution. We adopt an exact variance component tests because most of the studies concerning microbiome compositions have modest sample size. "permutation" represents a residual permutation approach. "moment" represents an approximation method that matches the first two moments. When out_type = "C" (continuous outcome y), the "moment" method is the Satterthwaite approximation. When out_type = "D" (dichotomous outcome), the "moment" method is the small sample adjustment in Lee et al (2012). When sample size is modest($n < 100$ for continuous or $n < 200$ for dichotomous outcome), the "moment" method can be inflated at very small size (such as $\alpha = 0.001$), although the type I error at $\alpha = 0.05$ is usually sustained. Therefore, we suggest using "davies" or permutation approach for such situations.

Please note that the "method" only concerns with the way that a kernel specific p-value is produced.

4.6 Properties of different kernels

How to choose an appropriate distance matrix and kernel for testing is a difficult question. However, it is important, since the distance matrix used to generate the kernel strongly affects the power of our tests. In particular, MiRKAT has highest power when the form of association between the microbiota and the outcome assumed by the kernel matches the true form of association. Poor choice of kernel will lead to reduced power, although the type I error will be preserved.

In the case of the UniFrac families and the Bray-Curtis dissimilarity, the factors at play are (1) the abundance of the associated taxa and (2) whether closely related taxa (phylogenetically) tend to be related or not related to the outcome as a group. For example, the following are some of the distance metrics that have been used for studies of the microbiome:

Distance	Phylogeny?	Abundance?	Other notes	Reference
Unweighted UniFrac	Yes	No		1
Weighted UniFrac	Yes	Yes		2
Generalized UniFrac	Yes	(Yes)	Parameter alpha defines extent to which abundance is taken into account	3
Jaccard	No	No	1 - (taxa in both)/(taxa in either); typically presence/absence, but can be extended to an abundance-weighted version	4,5
Bray-Curtis	No	Yes	Similar to Jaccard, but uses counts	6

In the table above, "Yes" indicates the distance or dissimilarity metric has the feature; "(Yes)" indicates that the feature is present either in some variations of the metric or is present to some extent; and "No" indicates that the feature is not present.

Depending on which of these characteristics are expected to be present in a particular study (based on prior knowledge or intuition), an appropriate distance or dissimilarity can be selected. If the study is exploratory and strong protection of type 1 error is not needed, several distance metrics can be explored. Depending on which one(s) are highly significant, some information can be gained about the nature of any association between the microbiota and the outcome.

4.7 Testing using multiple kernels

We provide an omnibus test that takes into account of multiple kernels simultaneously. The method is robust that it has substantial power gain compared to when an improper kernel is used, and has little power loss compared to when the best kernel is used.

```
> Ks = list(K.weighted, K.unweighted, K.BC)
> MiRKAT(y = Smoker, Ks = Ks, X = cbind(Male, anti),out_type = "D" ,
+        nperm = 9999, method = "davies")

$indivP
[1] 0.004654662 0.014189982 0.001973473

$omnibus_p
[1] 0.00595
```

This function outputs p-values for association using each single kernel and an omnibus p-value considering all kernels. The omnibus p-value is obtained

through residual permutation where the minimum p-values from each of the individual tests are used as test statistics. Still, "method" option only indicates the method that is used for generating individual kernel p-value.

5 MiRKAT-S: association testing between microbiome community and survival outcome

5.1 Simulate time to event data

We still use the throat data set to demonstrate the use of MiRKATS. Data loading and preparation are the same as in the previous section. Because the original dataset has a binary phenotype (smoking) rather than a measure of censored time to event outcomes, we consider smoking status and gender as covariates and generate null outcome data from the Exponential distribution. Specifically, we generate survival times as $S \sim \text{Exponential}(1 + I(\text{smoke}) + I(\text{male}))$, and censoring times as $C \sim \text{Exponential}(0.75)$. Then the outcome measure consists of $T = \min(S, C)$ and $\Delta = I(S \leq C)$. This simulation procedure results in approximately 33% censoring.

```
> # Simulate outcomes
> # Here, outcome is associated with covariates but unassociated with microbiota
> # 33% censoring
> SurvTime <- rexp(60, (1 + Smoker + Male))
> CensTime <- rexp(60, 0.75)
> Delta <- as.numeric( SurvTime <= CensTime )
> ObsTime <- pmin(SurvTime, CensTime)
```

5.2 Test using a single kernel

Here, we run MiRKAT-S to test the association between the microbiota and simulated survival times, adjusting for gender and smoking status.

```
> # use kernel matrix with distance=FALSE
> MiRKATS(kd = K.generalized, distance = FALSE, obstime = ObsTime,
+         delta = Delta, covar = cbind(Male, Smoker))

[1] 0.3832066

> # equivalently, use distance matrix with distance=TRUE
> MiRKATS(kd = D.generalized, distance = TRUE, obstime = ObsTime,
+         delta = Delta, covar = cbind(Male, Smoker))

[1] 0.3832066

> # Permutation version of the test
> MiRKATS(kd = K.generalized, distance = FALSE, obstime = ObsTime,
+         delta = Delta, covar = cbind(Male, Smoker), perm=TRUE, nperm=1000)
```

```
[1] 0.459
```

The argument "distance" indicates whether "kd" is a distance matrix (TRUE) or kernel matrix (FALSE). "delta" give censoring status for each individual. When delta = 1, the corresponding "obstime" is the survival time, and when delta = 0, the corresponding observation is censored. The output is the p-value for the test using Davies' exact method, which computes the p-value based on a mixture of chi-square distributions. We use a small-sample correction to account for the modest sample sizes and sparse OTU count matrices that often result from studies of the microbiome. "perm = T" indicates that a permutation p-value is calculated and "perm = F" indicates that the p-value is obtained via the davies method. Overall, permutation is recommended when the sample size is small, as the davies method may be slightly anti-conservative with very small sample sizes. MiRKAT-S will generate a warning when permutation is not used for sample sizes $n \leq 50$. "nperm" indicates the number of permutations to perform to generate the p-value (default = 1000).

6 MMiRKAT: association testing between microbiome composition and a multivariate continuous outcome

Compared to KRV approach, which will be demonstrated in the next section, MMiRKAT is designed to test the association between microbiome with a low-dimensional multivariate continuous outcome. For structural or high dimensional outcome, KRV is recommended.

We still use the throat microbiome data to demonstrate the use of MMiRKAT. We generate multivariate outcomes $Y \sim N_{3n}(0, I_{3n})$.

6.1 Simulate multivariate outcome

```
> set.seed(123)
> n = nrow(throat.otu.tab)
> Y = matrix(rnorm(n * 3, 0, 1),n ,3)
```

6.2 MMiRKAT testing

```
> MMiRKAT(Y = Y, K = K.weighted, X = cbind(Male, anti))
```

```
[1] 0.1840602
```

MMiRKAT uses a small-sample correction procedure for p-value calculation.

7 KRV: association testing between microbiome composition and structured outcome

KRV (kernel RV coefficient) is designed to evaluate the association between microbiome composition and a structured, potentially high-dimensional phenotype, such as gene expression of a set of genes which are functionally related. The KRV statistic can capture nonlinear correlations and complex relationships among the individual data types and between the complex multivariate phenotype and microbiome composition through measuring general dependency. Two kernels are involved in KRV statistics, one kernel for microbiome composition, which can be obtained by transforming the distance metrics just as the previous sections, and one kernel capturing similarities in the phenotypes. As an alternative, KRV can also take as input a kernel matrix for microbiome composition, and a multivariate phenotype y, and a set of additional covariates X. In this setting, a multivariate regression was first carried out, the residuals of which were subsequently used to construct a kernel matrix for phenotypes.

7.1 Create the kernel matrix of simulated data

```
> library(MASS)
> set.seed(123)
> rho = 0.2
> Va = matrix(rep(rho, (2*n)^2), 2*n, 2*n)+diag(1-rho, 2*n)
> G = mvrnorm(n, rep(0, 2*n), Va)
```

7.2 KRV test

```
> KRV(kernel.otu = K.weighted, kernel.y = G %*% t(G))

          [,1]
[1,] 0.1459494

> KRV(kernel.otu = K.weighted, y = G, X = cova, kernel.y = "linear")

          [,1]
[1,] 0.1442477
```

In the first example, KRV tests the association between microbiome and y through two kernels: kernel.otu and kernel.y. No additional covariates is considered. In the second example, KRV tests the association between microbiome and phenotype y, adjusting for X. Input for "kernel.y" can be selected from "Gaussian" and "linear", or a n by n numeric matrix.

8 Reference

Zhao, N., Chen, J.,Carroll, I. M., Ringel-Kulka, T., Epstein, M.P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H. and Wu, M.C. (2015)). Microbiome Regression-

based Kernel Association Test (MiRKAT). American Journal of Human Genetics, 2015 May 7;96(5):797-807.

Plantinga, A., Zhan, X., Zhao, N., Chen, J., Jenq, R., and Wu, M.C. MiRKAT-S: a distance-based test of association between microbiome composition and survival times. Microbiome, 2017, 5:17. DOI: 10.1186/s40168-017-0239-9

Zhan, X., Tong, X., Zhao, N., Maity,A., Wu, M.C., and Chen,J. A small-sample multivariate kernel machine test for microbiome association studies. Genetic Epidemiology, 41(3), 210-220. DOI: 10.1002/gepi.22030

Zhan, X., Plantinga, A., Zhao, N., and Wu, M.C. A Fast Small-Sample Kernel Independence Test for Microbiome Community-Level Association Analysis. Biometrics, 2017 Mar 10. doi: 10.1111/biom.12684.

Davies R.B. (1980) Algorithm AS 155: The Distribution of a Linear Combination of chi-2 Random Variables, *Journal of the Royal Statistical Society. Series C*, 29, 323-333.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. Biom. Bull. 2, 110-114.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA; NHLBI GO Exome Sequencing Project-ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X. (2012) Optimal unified approach for rare variant association testing with application to small sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, 91, 224-237.

Chen, J., Chen, W., Zhao, N., Wu, M. C. and Schaid, D. J. (2016), Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. Genet. Epidemiol., 40: 5-19. doi: 10.1002/gepi.21934

Zhou., J. J. and Zhou, H.(2015) Powerful Exact Variance Component Tests for the Small Sample Next Generation Sequencing Studies (eVCTest), in submission.

Efron, B. (1977) "The efficiency of Cox's likelihood function for censored data." Journal of the American statistical Association 72(359):557-565.

Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, et al. "Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers." PLoS ONE 5(12): e15216.