Published in final edited form as:

Biometrics. 2018 September ; 74(3): 1120–1124. doi:10.1111/biom.12823.

On the fast small-sample kernel independence test for microbiome community-level association analysis

Bin Guo and Baolin Wu*

Division of Biostatistics, School of Public Health, University of Minnesota

Summary

Zhan et al. (2017) presented a kernel RV coefficient (KRV) test to evaluate the overall association between host gene expression and microbiome composition, and showed its competitive performance compared to existing methods. In this paper, we clarify the close relation of KRV to the existing generalized RV (GRV) coefficient, and show that KRV and GRV have very similar performance. Although the KRV test could control the type I error rate well at 1% and 5% levels, we show that it could largely under-estimate p-values at small significance levels leading to significantly inflated type I errors. As a partial remedy, we propose an alternative p-value calculation, which is efficient and more accurate than KRV p-value at small significance levels. We recommend that small KRV test p-values should always be accompanied and verified by the permutation p-value in practice. In addition we analytically show that KRV can be written as a form of correlation coefficient, which can dramatically expedite its computation and make permutation p-value calculation more efficient.

Keywords

Gamma distribution; Kurtosis; Microbiome; RV coefficient; Skewness

1 Introduction

Zhan et al. (2017) proposed a kernel RV (KRV) coefficient to test the overall association between host gene expression and microbiome composition. KRV is closely related to the generalized RV (GRV) coefficient of Minas et al. (2013), and both are broadly applicable to test the association between two sets of multivariate outcomes, X and Y , measured on a same cohort of n subjects. Both are essentially the correlations between two similarity matrices: KRV uses kernel matrix and GRV uses Gower matrix (Gower, 1966) respectively. The similarity matrix is typically computed from some distance matrix. Specifically given two $n \times n$ distance matrices D_X and D_Y computed from X and Y respectively, we compute the Gower matrices

* baolin@umn.edu.

5 Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 3.2, and an R package implementing the proposed method are available with this paper at the Biometrics website on Wiley Online Library.

$$G_x = -\frac{1}{2}HD_x^2H, \quad G_y = -\frac{1}{2}HD_y^2H, \quad H = I_n - J_n/n,$$

where I_n is the n -th order identity matrix, J_n is a $n \times n$ matrix of ones, D_x^2 and D_y^2 are the element-wise matrix squares. GRV can be computed as

$$Q_g = \frac{\text{tr}(G_x G_y)}{\sqrt{\text{tr}(G_x G_x)} \sqrt{\text{tr}(G_y G_y)}},$$

where $\text{tr}()$ denotes the matrix trace. To compute KRV, we first transform the Gower matrix G into a kernel matrix K following the approach of Zhao et al. (2015). Specifically given the eigen decomposition $G = U \Lambda U^T$ with $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ being the eigen values, we calculate $K = U \tilde{\Lambda} U^T$, where $\tilde{\Lambda} = \text{diag}(|\lambda_1|, \dots, |\lambda_n|)$. Let K_x and K_y be the transformed kernel matrices from G_x and G_y respectively. The KRV is then computed as

$$Q_k = \frac{\text{tr}(\tilde{K}_x \tilde{K}_y)}{\sqrt{\text{tr}(\tilde{K}_x \tilde{K}_x)} \sqrt{\text{tr}(\tilde{K}_y \tilde{K}_y)}},$$

where $\tilde{K}_x = HK_x H$ and $\tilde{K}_y = HK_y H$ are the centered kernel matrices. Here the kernel matrix is essentially the Gower matrix with a positive semi-definiteness correction. In this spirit, the KRV is equivalent to the GRV. So when the kernel matrix is directly available (not computed from some distance matrix), we can substitute the centered kernel matrix as Gower matrix to calculate GRV (in section 3, we will see one such example). For most commonly used and well-defined distances, the Gower matrix is typically positive-definite, and GRV and KRV are identical. In our numerical studies, we have observed that GRV and KRV give very similar results.

2 Computation

We first show that the RV-type statistic can be written as a sample correlation. Given two symmetric and centered $n \times n$ matrices $A = (a_{ij})$, $B = (b_{ij})$, we can easily check that $\text{tr}(AB) = \sum_{i,j} a_{ij} b_{ij}$, $\text{tr}(AA) = \sum_{i,j} a_{ij}^2$, $\text{tr}(BB) = \sum_{i,j} b_{ij}^2$. Hence

$$Q = \frac{\text{tr}(AB)}{\sqrt{\text{tr}(AA)} \sqrt{\text{tr}(BB)}} = \frac{\sum_{i,j} a_{ij} b_{ij}}{\sqrt{\sum_{i,j} a_{ij}^2} \sqrt{\sum_{i,j} b_{ij}^2}}.$$

Therefore $Q = \text{cor}[\text{vec}(A), \text{vec}(B)]$, where $\text{vec}()$ is the vector operator that stacks the columns of a matrix into a vector.

When both A and B are positive semi-definite kernel matrices, denote the Cholesky decompositions $A/\sqrt{\text{tr}(AA)} = CC^T$, $B/\sqrt{\text{tr}(BB)} = DD^T$, where C and D are $n \times n$ lower-triangular matrices. Denote $Z = D^T C = (z_{ij})$, which can be treated as a matrix of Score type statistics testing the association between two sets of “working outcomes”: the n columns of C and D . We then have $Q = \text{tr}[(D^T C)^T (D^T C)] = \sum_{i,j} z_{ij}^2 \geq 0$, which is closely related to the multi-trait variant set association test statistic (Wu and Pankow, 2016).

2.1 Permutation p-value calculation

The distributions of GRV and KRV are generally unknown, and permutations have been typically used to obtain their p-values (Minas et al., 2013; Zhan et al., 2017). Specifically their null distributions are obtained by randomly shuffling rows and columns of one kernel matrix simultaneously. Note that randomly shuffling the columns and rows of B simultaneously does not change $\text{tr}(BB)$. Therefore computationally we just need to calculate $\text{vec}(A)^T \text{vec}(B)$ under permutation to compute the p-value of Q , which can be done very efficiently.

2.2 Permutation distribution approximation

To further improve the computation efficiency, an approximate Gamma distribution is commonly used to compute the p-value analytically for the RV-type statistics. The approximate Gamma distribution is estimated by matching the first three moments with the permuted RV-type statistics.

Denote the observed RV statistic as Q , and its mean, standard error, and skewness under permutation as (μ, σ, γ) , where $\mu = E(Q)$, $\sigma = \sqrt{E[(Q - \mu)^2]}$, and $\gamma = E[(Q - \mu)^3]/\sigma^3$. Here the expectation is with respect to the permutation distribution of Q . We approximate the permutation null distribution of $\text{sign}(\gamma)(Q - \lambda)$ with a $\text{Gamma}(k, \theta)$ distribution, where k is the shape, θ is the scale and

$$k = 4/\gamma^2, \quad \theta = \sigma / \gamma / 2, \quad \lambda = \mu - 2\sigma/\gamma.$$

By matching the first three moments, the approximated Gamma distribution has been shown to control type I errors well at the 0.05 significance level.

For small p-values, the accuracy of approximation critically depends on the tail of Gamma distribution. In our numerical studies, we have found that the $\text{Gamma}(k, \theta)$ approximation could largely under-estimate small p-values (leading to significantly inflated type I errors at small significance levels). For example, when analyzing the association of host gene expression and microbiome profiles in 196 pre-pouch ileum (PPI) samples using KRV test (see Section 3), the Gamma approximation gives a p-value of 6.5×10^{-5} , while the actual permutation p-value is 1.5×10^{-4} (based on 10^7 permutations).

To improve the performance of Gamma approximation, we propose to match kurtosis instead of skewness as follows. Denote $\nu = E[(Q - \mu)^4]/\sigma^4$ as the kurtosis under

permutation, then we approximate the permutation null distribution of $\text{sign}(\nu - 3)(Q - \tilde{\lambda})$ with a Gamma($\tilde{k}, \tilde{\theta}$) distribution, where

$$\tilde{k} = 6 / |\nu - 3|, \quad \tilde{\theta} = \sigma \sqrt{|\nu - 3| / 6}, \quad \tilde{\lambda} = \mu - \sigma \sqrt{6 |\nu - 3|} / (\nu - 3).$$

We can easily verify that it leads to the same mean, variance and kurtosis as permuted RV statistics. This new Gamma approximation gives a p-value of 1.3×10^{-4} for the 196 PPI samples.

3 Numerical studies

3.1 Application to PPI study

For illustration, we reanalyze the association of host gene expression and microbiome profiles in 196 pre-pouch ileum (PPI) samples (Morgan et al., 2015). Specifically 7000 OTU counts measured by 16S rRNA and around 20,000 microarray gene expressions were analyzed following Zhan et al. (2017).

For the host transcript, we calculate a Gaussian kernel, $K_{ij} = \exp(-\|Y_i - Y_j\|^2 / \sigma^2)$, where Y_i denotes the centered expression values for the i -th sample, and σ^2 is taken as the median of all pairwise distances. For the OTU data, we computed the Bray-Curtis distance kernel (denoted as K_{BC}) directly from the counts, and the following five UniFrac-type distance kernels based on a trained phylogenetic tree (Li, 2015; Zhan et al., 2017): (1) weighted UniFrac distance kernel (denoted as K_w); (2) unweighted UniFrac distance kernel (denoted as K_u); (3) variance adjusted weighted UniFrac distance kernel (denoted as K_v); (4) generalized UniFrac distance kernel K_θ with $\theta = 0$ (denoted as K_0); (5) generalized UniFrac distance kernel K_θ with $\theta = 0.5$ (denoted as $K_{0.5}$). We apply the KRV, GRV, and the proposed method (denoted as “RRV” for revised RV test). As a benchmark, we also included the permutation approach based on 10^7 random permutations (denoted as “Perm”), which can be treated as the gold standard. Table 1 shows the results. Overall, we can see that KRV and GRV have very similar performance. KRV generally under-estimates the significance p-values compared to the permutation approach. The proposed RRV mitigates the inflation of KRV, and is closer to the permutation p-values.

3.2 Simulation study

For illustration, we fix the six microbiome kernels we computed previously for the 196 PPI samples, and randomly simulate the expressions of 100 genes for these 196 samples from a multivariate normal distribution with zero means and compound symmetry covariance matrix with unit variance and pairwise covariance being 0.25. We compute the Gaussian kernel for these gene expression profiles as previously. We consider three tests: KRV, the proposed RRV based on matching kurtosis, and the permutation test based on randomly shuffling the columns and rows of the expression Gaussian kernel. Permutation p-values (denoted “Perm”) are obtained adaptively as follows. We first used 10^5 random permutations to estimate p-values. For those p-values less than 10^{-3} , we then increased the permutations to 10^6 to obtain more accurate estimates. We conducted 10^7 simulations to estimate the type

I errors at significance level $\alpha = 0.05, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$. The results are summarized at Table 2. Overall we can see that the permutation approach very accurately controls the type I errors. The KRV test has increased inflations at more stringent significance levels. The proposed RRV has reduced inflations compared to KRV, though it still has increased inflations at more stringent significance levels.

Figure 1 shows the QQ-plots for the KRV and RRV p-values based on the K_W and K_0 kernels. We can clearly see the trend of increasing inflations at more stringent significance levels.

We also investigated the performance of different methods under linear kernel and with different dependence structures. Similar results are observed consistently across all simulation scenarios. The complete results (QQ plots and type I errors at various α levels) are available at the supplementary materials.

4 Discussion

Another closely related and widely used approach in ecology is the Mantel test (Mantel, 1967). It essentially calculates the correlation between two distance matrices, in contrast to the Gower/Kernel (roughly squared distance) matrices used in the RV statistics. In the special case of using linear kernel based on a single variable, the RV type statistic reduces to the Pearson correlation coefficient, while the Mantel test does not. In our numerical studies, the Mantel test performs worse than the RV statistics (see supplementary materials for results), which are consistent with the simulation studies of Minas et al. (2013).

When assessing the significance of RV-type statistics, we need to be cautious when using the existing Gamma approximation based p-values. In light of their simple structure as a correlation and the common practice of comparing different methods based on the computed p-values, we recommend that small Gamma approximation p-values should always be verified by the actual permutation p-values.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by NIH grant GM083345 and CA134848. We would like to thank the reviewers for their valuable comments, which have greatly improved the presentation of the paper. We acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported within this paper.

References

- Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*. 1966; 53:325–338.
- Li H. Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis. *Annual Review of Statistics and Its Application*. 2015; 2:73–94.
- Mantel N. The detection of disease clustering and a generalized regression approach. *Cancer Research*. 1967; 27:209–220. [PubMed: 6018555]

- Minas C, Curry E, Montana G. A distance-based test of association between paired heterogeneous genomic data. *Bioinformatics*. 2013; 29:2555–2563. [PubMed: 23918252]
- Morgan XC, Kabakchiev B, Waldron L, Tyler AD, Tickle TL, Milgrom R, Stempak JM, Gevers D, Xavier RJ, Silverberg MS, Huttenhower C. Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biology*. 2015; 16:67. [PubMed: 25887922]
- Wu B, Pankow JS. Sequence Kernel Association Test of Multiple Continuous Phenotypes. *Genetic Epidemiology*. 2016; 40:91–100. [PubMed: 26782911]
- Zhan X, Plantinga A, Zhao N, Wu MC. A fast smallsample kernel independence test for microbiome community-level association analysis. *Biometrics*. 2017; doi: 10.1111/biom.12684
- Zhao N, Chen J, Carroll I, Ringel-Kulka T, Epstein M, Zhou H, Zhou J, Ringel Y, Li H, Wu M. Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *The American Journal of Human Genetics*. 2015; 96:797–807. [PubMed: 25957468]

Figure 1.

QQ plots: the dashed line is the diagonal line. We have plotted the p-values on the log10 scale for better visualization.

P-values of testing around 20,000 gene expressions and microbiome association in 196 PPI samples. Permutation p-values (denoted “Perm”) are obtained based on 10^7 Monte Carlo permutations, and can be treated as the gold standard. The proposed method is denoted “RRV” (for revised RV test).

Table 1

Test	K_w	K_u	$K_{0.5}$	K_{BC}	K_v	K_0
Perm	1.23×10^{-2}	1.46×10^{-4}	5.12×10^{-3}	7.82×10^{-3}	1.96×10^{-2}	4.58×10^{-4}
KRV	1.20×10^{-2}	6.45×10^{-5}	4.64×10^{-3}	7.44×10^{-3}	1.97×10^{-2}	2.57×10^{-4}
GRV	1.22×10^{-2}	6.45×10^{-5}	4.43×10^{-3}	8.09×10^{-3}	2.03×10^{-2}	2.57×10^{-4}
RRV	1.34×10^{-2}	1.26×10^{-4}	5.72×10^{-3}	8.46×10^{-3}	2.20×10^{-3}	4.75×10^{-4}

Table 2

Ratio of empirical type I errors divided by the significance level α estimated over 10^7 simulations. Listed within parenthesis are the standard errors divided by α . Permutation p-values (denoted “Perm”) are obtained adaptively as follows. We first used 10^5 Monte Carlo simulations to estimate p-values. For those p-values less than 10^{-3} , we then used 10^6 Monte Carlo permutations to obtain more accurate estimates.

K_w										K_u				
α	0.05	10^{-2}	10^{-3}	10^{-4}	10^{-5}	0.05	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
KRV	0.94 (0.001)	1.04 (0.003)	1.36 (0.01)	1.86 (0.04)	2.92 (0.17)	0.92 (0.001)	1.05 (0.003)	1.49 (0.01)	2.41 (0.05)	4.12 (0.20)				
RRV	0.93 (0.001)	0.93 (0.003)	1.07 (0.01)	1.25 (0.04)	1.68 (0.13)	0.90 (0.001)	0.91 (0.003)	1.06 (0.01)	1.38 (0.04)	1.66 (0.13)				
Perm	1.00 (0.001)	1.00 (0.003)	1.02 (0.01)	0.99 (0.03)	0.99 (0.10)	1.00 (0.001)	1.00 (0.003)	1.00 (0.01)	1.02 (0.03)	0.99 (0.10)				
$K_{0.5}$										K_{BC}				
α	0.05	10^{-2}	10^{-3}	10^{-4}	10^{-5}	0.05	10^{-2}	10^{-3}	10^{-4}	10^{-5}				
KRV	0.93 (0.001)	1.05 (0.003)	1.44 (0.01)	2.15 (0.05)	3.50 (0.19)	0.95 (0.001)	1.05 (0.003)	1.39 (0.01)	2.01 (0.04)	3.10 (0.18)				
RRV	0.92 (0.001)	0.92 (0.003)	1.06 (0.01)	1.34 (0.04)	1.86 (0.14)	0.93 (0.001)	0.92 (0.003)	1.02 (0.01)	1.23 (0.04)	1.52 (0.12)				
Perm	1.00 (0.001)	1.00 (0.003)	1.01 (0.01)	1.01 (0.03)	0.98 (0.10)	1.00 (0.001)	1.00 (0.003)	1.00 (0.01)	1.02 (0.03)	0.98 (0.10)				
K_v										K_0				
α	0.05	10^{-2}	10^{-3}	10^{-4}	10^{-5}	0.05	10^{-2}	10^{-3}	10^{-4}	10^{-5}				
KRV	0.92 (0.001)	1.03 (0.003)	1.44 (0.01)	2.23 (0.05)	3.62 (0.19)	0.92 (0.001)	1.06 (0.003)	1.60 (0.01)	2.80 (0.05)	5.10 (0.23)				
RRV	0.90 (0.001)	0.91 (0.003)	1.06 (0.01)	1.33 (0.04)	1.83 (0.14)	0.89 (0.001)	0.90 (0.003)	1.06 (0.01)	1.46 (0.04)	2.09 (0.14)				
Perm	1.00 (0.001)	1.00 (0.003)	0.99 (0.01)	0.97 (0.03)	1.01 (0.10)	1.00 (0.001)	1.00 (0.003)	1.00 (0.01)	1.01 (0.03)	1.00 (0.10)				