Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Juniper Publishers
key to the Researchers

Review Article
Volume 5 Issue 4 - February 2018
DOI: 10.19080/BBOAJ.2018.04.555668

Biostat Biometrics Open Acc J

Kernel-Based Hypothesis Testing

Wonkyung J[1] and Ilmun K[2]*

[1]*Department of Statistics, The University of North Carolina at Chapel hill, USA*

[2]*Department of Statistics & Data Science, Carnegie Mellon University, USA*

Submission: January 30, 2018; **Published:** February 28, 2018

***Corresponding author:** Ilmun Kim, School of Education, The University of North Carolina at Chapel hill, USA; Email: ilmunk@andrew.cmu.edu

Abstract

With the advent of big and complex data, there have been recent developments in statistics and related fields that overcome limitations of classical parametric inference. For instance, Gretton et al. [1] introduce kernel maximum mean discrepancy (MMD) and propose two sample testing procedures based on MMD. Unlike the classical t-test, which is only sensitive to mean differences, MMD test can detect an arbitrary difference between two multivariate distributions without imposing parametric assumptions. Such tests have application in a number of complex high-dimensional real world problems. In this article, we review some aspects of MMD and propose kernel distance components (kernel DISCO) as an extension of MMD to the k-sample problems.

Keywords : Distance components; Chemometrics; Microarray data; Mean discrepancy; K-sample problems

Abbreviations : MMD: Maximum Mean Discrepancy; kernel DISCO: kernel Distance Components; RKHS: Reproducing Kernel Hilbert Space

Introduction

The object of two-sample tests (homogeneity tests) is to determine whether the underlying distributions behind two sets of data are equivalent. More precisely, given two independent samples $X_1,...,X_m \sim F_x$ and $Y_1,...,Y_n \sim G_y$, we are interested in testing $H_0 : F_x = G_y$ against $H_1 : F_x \neq G_y$. Equation Section (Next) (1)

The two-sample test has a myriad of application in diverse areas. In bioinformatics, it is of interest to compare high-dimensional low sample size data from medical imaging techniques (like computed tomography or X-ray radiography), chemometrics and microarray data (proteomics and transcriptomics) [2]. The test can also be applied to public health and social science studies when working with multivariate real world data with complex dependence structure and the heavy-tailed or skewed population distributions [3]. One of the most classic methods for testing homogeneity is two-sample t-test, which compares the means of two populations. In recent years, Gretton et al. [4] propose a nonparametric two sample t-test called kernel Maximum Mean Discrepancy (MMD), which compares the means of two distributions in a Reproducing Kernel Hilbert Space (RKHS). Contrary to the t-test, which is only sensitive to mean differences, the test based on MMD can be sensitive to an arbitrary difference between two multivariate distributions. The distinguishing features of MMD are summarized as follows:

Ability to detect any difference between two multivariate distributions

Applicability to complex and high-dimensional data

Flexibility of the approach depending on the choice of kernel

Computational efficiency without involving density estimates

Despite its attractive properties, MMD has not been fully introduced to researchers outside of theoretical statistics and machine learning communities. The purpose of this article is to review the selected aspects of MMD and describe its testing procedure. We also briefly discuss an extension of MMD to the multi-sample problems where the interest is in testing

$H_0 = F_1 = F_2 \ldots = F_K$ against H_1: at least one of F_i is different.

Maximum mean discrepancy

Let and be random vectors from and defined on a domain x. It is well-known that X and Y have the same distribution if and only if $\mathbb{E}[f(x)] = \mathbb{E}[f(y)]$ for all continuous bounded functions f The key insight of MMD is that it is possible to reduce the class of functions $f\cdot$ while maintaining the ability to distinguish between distributions F_X and G_Y. The definition of MMD is as follows:

Definition (Maximum mean discrepancy): Let $\mathcal{F}$ be a class of functions $f : x \to \mathbb{R}$. Maximum mean discrepancy is defined as

$$\mathrm{MMD}(\mathcal{F}, X, Y) \sup_{f \in \mathcal{F}} \left(\mathrm{E}[f(x)] - \mathrm{E}[f(y)] \right). \qquad (2)$$

Suppose the function class $\mathcal{F}$ is a unit ball in a universal RKHS. Then MMD has the characteristic property.

Theorem (Characteristic property of MMD): Let $\mathcal{F}$ be a unit ball in a universal RKHS $\mathcal{H}$, defined on the compact metric space x, with associated kernel k (•, •). Then $\mathrm{MMD}(\mathcal{F}, X, Y)=0$ if and only if $F_X = G_Y$. Let X, $X' \overset{i.i.d}{\sim} F_X$ and $Y, Y' \overset{i.i.d}{\sim} G_Y$. From a practical point of view, it is not trivial how to estimate (2). However, if the unit ball is used as a function class, MMD has a compact representation in terms of the expected values of pair wise kernels:

$$\mathrm{MMD}^2(F, p, q) = \mathrm{E}\left[k(X, X')\right] - 2\mathrm{E}\left[k(X, Y)\right] + \mathrm{E}\left[k(Y, Y')\right]$$

Which results in a direct empirical estimate based on a U-statistic:

$$U_{m,n} = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(x_i, x_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(y_i, y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i, y_j).$$

An important issue of the MMD-based tests is in the choice of kernel. As an approach, Gretton et al. [1] suggest a heuristic way to select a radial basis function kernel with a kernel bandwidth to be the median distance between points in the aggregate sample. Another way to choose the bandwidth is to maximize the test statistic. Gretton et al. [4] and Sutherland et al. [5] introduce ways to select the kernel that maximizes asymptotic power of the test. However, the last two approaches rely on sample splitting and thus less efficient than the other methods in small sample size. Despite all of the efforts, the optimal choice of kernel still remains as an open question. There are many ways to implement the two-sample test based on MMD. Here, we focus on the permutation test which controls an exact type I error under finite sample size. The testing procedure is summarized as follows:

1. Generate random permutations $\pi_1, \pi_2, \ldots, \pi_{m+n}$ among $\{1, \ldots, m+n\}$.
2. Let $(Z_1, \ldots, Z_{m+n}) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$. Calculate $U_{m,n}$ based on $\left(Z_{\pi_1}, \ldots, Z_{\pi_m}\right)$ and $\left(Z_{\pi_{m+1}}, \ldots, Z_{\pi_{m+n}}\right)$.
3. Repeat the previous steps B times to obtain $U_{m,n}^{(1)}, \ldots, U_{m,n}^{(B)}$.
4. Let $U_{m,n}$ be the statistic calculated based on the original samples. Then evaluate the p-value as
$$p = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} I\left(U_{m,n}^{(b)} > U_{m,n}\right)\right)$$
5. Reject the null hypothesis if $p < \alpha$ where α is a pre-fixed significant level.

Kernel DISCO: Extension to multi-sample problem

Beyond the two-sample hypothesis, we can consider the K-sample hypothesis $H_0 : F_1 = \cdots = F_K, K \geq 2$ versus $H_1 : F_i \neq F_j$ for some $i \neq j$. Analogous to the ANOVA decomposition of variance, Rizzo & Székely [6] propose distance components (DISCO). DISCO is a nonparametric extension of ANOVA based on the partition of the total dispersion into between and within components. In the next section, we briefly describe the definition of it.

DISCO

Let $A = \{a_1, \ldots, a_{n_1}\}$, $B = \{b_1, \ldots, b_{n_2}\}$ be two samples and define

$$g_\alpha(A, B) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \| a_i - b_j \|^\alpha,$$

for $0 < \alpha \leq 2$. Let $A_1, \ldots, A_K$ be samples of sizes $n_1, \ldots, n_K$ and $N = \sum_{j=1}^{K} n_j$. Based on these notations, the within-sample dispersion statistic is defined by

$$W_\alpha = \sum_{j=1}^{K} \frac{n_j}{2} g_\alpha(A_j, A_j). \quad (3)$$

Similarly, the total dispersion of the observed response is defined by

$$T_\alpha = \frac{N}{2} g_\alpha(A, A), \quad (4)$$

Where, A is the pooled sample. Lastly, the between-sample energy statistic is given as

$$S_\alpha = \sum_{1\leq j\leq k\leq K} \left(\frac{n_j + n_k}{2}\right)\left[\frac{n_j n_k}{n_j + n_k}\varepsilon_{n_j, n_k}^{(\alpha)}(A_j, A_k)\right] \quad (5)$$

$$= \sum_{1\leq j\leq k\leq K}\left[\frac{n_j n_k}{2N}(2g_\alpha(A_j, A_k) - g_\alpha(A_j, A_j) - g_\alpha(A_k, A_k))\right]. \quad (6)$$

Rizzo & Székely [6] show that $T_\alpha = S_\alpha + W_\alpha$, where both S_α and W_α are nonnegative. For every $0 < \alpha < 2$, S_α determines a consistent test of the multi-sample hypothesis of equal distributions. For $\alpha = 2$, DISCO decomposition is equivalent to classical ANOVA decomposition as $T_2 = S_2 + W_2$. However, the ANOVA test statistic measures differences in means, not distributions. In this sense, DISCO can be considered as a nonparametric generalization of classical ANOVA.

Analogous to the ANOVA test, the K-sample DISCO test is carried out by using the ratio statistic

$$F_\alpha = \frac{S_\alpha/(K-1)}{W_\alpha/(N-K)},$$

and reject the null for a large value of F_α. The p-value can be evaluated based on the permutation procedure.

Kernel DISCO

Sejdinovic et al. [7] provide a framework that explains the equivalence of MMD and the energy distance when a special type of kernel, termed distance kernel, is employed. Motivated by this observation, we propose kernel distance components (Kernel DISCO) for testing the K-sample hypothesis. Let (z, ρ) be a semimetric space of negative type. Define W_ρ, T_ρ and S_ρ by replacing $g_\alpha(A, B)$ in (3), (4) and (5) with $g_\rho(A, B) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\rho(a_i, b_j)$.

Followed by Sejdinovic et al. [7], there exists a kernel such that

How to cite this article: Wonkyung J, Ilmun K. Kernel-Based Hypothesis Testing. Biostat Biometrics Open Acc J. 2018; 5(4): 555668. DOI: 10.19080/BBOAJ.2018.05.555668

$\rho(z,z') = k(z,z) + k(z',z') - 2k(z,z')$. When we further consider k to be characteristic, which means that MMD_k $MMD_k\left(F_i, F_j\right) = 0$ if and only if (F_i, F_j), then the following statements hold.

Theorem: Let ρ be the semimetric associated with a characteristic kernel. For all p-dimensional samples $A_1, \ldots, A_K$, where $K \geq 2$, we have

1. $S_\rho(A_1, \ldots, A_K) \geq 0$.
2. $S_\rho(A_1, \ldots, A_K) = 0$ if and only if $A_1 = \cdots = A_K$.

Proof: Let $A_j = \{a_1, \ldots, a_{n_j}\}$ and $A_k = \{b_1, \ldots, b_{n_k}\}$. Define i.i.d. random variables X and X' uniformly distributed on A_j, and define i.i.d. random variables Y and Y' uniformly distributed on A_K. Then $\text{E}\rho(X,Y) = g_\rho(A_j, A_k)$, $\text{E}\rho(X,X') = g_\rho(A_j, A_j)$, $\text{E}\rho(Y,Y') = g_\rho(A_k, A_k)$ and $\frac{n_1 n_2}{n_1 + n_2}\varepsilon_\rho(X,Y) = \frac{2n_1 n_2}{n_1 + n_2}\text{MMD}^2(p,q) = 2S_\rho$.

This implies that $S_\rho(A_j, A_k) \geq 0$ and the equality becomes zero if and only if X and Y have the same distribution since the kernel is characteristic. The result for $K \geq 2A$ follows by induction. The next theorem is the kernel DISCO decomposition of total dispersion into between-sample, which is a weighted combination of pair wise MMD statistics and within-sample components.

Theorem: For all integers $K \geq 2$, the total dispersion T_ρ of K sample can be decomposed into

$$T_\rho(A_1, \ldots, A_K) = S_\rho(A_1, \ldots, A_K) + W_\rho(A_1, \ldots, A_K),$$

Where, $S_\rho \geq 0$ and $W_\rho \geq 0$ are the between-sample and within-sample measures of dispersion, respectively.

Proof: The proof is directly followed by Theorem 2 of Rizzo & Székely [6] except g_{jk} is replaced by $g_\rho(A_j, A_K)$.

In order to carry out the test, we define the ratio between S_ρ and W_ρ as

$$F_\rho = \frac{S_\rho(K-1)}{W_\rho / (N-K)},$$

and reject the null for a large value of F_ρ. The test can be carried out by using the permutation procedure. Similar to MMD, kernel DISCO statistic also relies on the choice of kernel. We recommend to use a radial basis kernel with the median heuristic but more theoretical and empirical studies should be followed up in the future.

Conclusion

Modern scientific studies in different fields including public health and social science have a common interest in comparing non-normal high-dimensional data. Many of classical methods for comparing distributions often fail when the dimension of the data exceeds the sample size. In addition, some of the methods heavily rely on parametric assumptions that are hardly true in practice. The MMD test, combined with the permutation procedure, addresses these issues by being fully nonparametric and applicable to an arbitrary dimension. In this article, we reviewed some of the properties of MMD and introduced kernel DISCO as an extension of MMD to the k-sample problems. We hope that this article encourages practitioners to consider the newly developed statistical methods and take them into consideration for their applications in the future.

References

1. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A (2012) A kernel two-sample test. Journal of Machine Learning Research 13: 723-773.
2. Marozzi M (2015) Multivariate multi distance tests for high-dimensional low sample size case control studies. Statistics in medicine 34(9): 1511-1526.
3. Linebach JA, Tesch BP, Kovacsiss LM (2014) Nonparametric statistics for applied research. Springer, USA.
4. Gretton A, Sejdinovic D, Strathmann H, Balakrishnan S, Pontil M, et al. (2012) Optimal kernel choice for large-scale two-sample tests. In Advances in neural information processing systems 1: 1205-1213.
5. Sutherland DJ, Tung HY, Strathman H, De S, Ramdas A, et al. (2016) Generative models and model criticism via optimized maximum mean discrepancy.
6. Rizzo ML, Székely GJ (2010) Disco analysis: A nonparametric extension of analysis of variance. The Annals of Applied Statistics 4(2): 1034-1055.
7. Sejdinovic D, Sriperumbudur B, Gretton A, Fukumizu K (2013) Equivalence of distance based and RKHS-based statistics in hypothesis testing. The Annals of Statistics 41(5): 2263-2291.

DOI: 10.19080/BBOAJ.2018.05.555668

How to cite this article: Wonkyung J, Ilmun K. Kernel-Based Hypothesis Testing. Biostat Biometrics Open Acc J. 2018; 5(4): 555668. DOI: 10.19080/BBOAJ.2018.05.555668