Multivariate Distribution Equality Hypothesis Test

Kang Gong
Supervisor: Prof. KaWai Tsang
CUHKSZ, Nov 20, 2017

1. Introduction

Suppose that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are independent random samples of $\mathbb{R}^d$-valued random vectors, $d > 1$, with respective distributions F_1 and F_2. The underlying problem is to test:

$$H_0 : F_1 = F_2$$

Previously, Székely and Rizzo (2004) [1] proposed a nonparametric test based on Euclidean distance and and resampling without replacement . As we will show in **Implement** section, the method they proposed is extremely time-consuming when n and m is large. Worse yet, the method is not sensitive with respect to the difference between two distribution.

Here, we propose two new nonparametric test: one is based on methods of clustering(e.g. k-means and Hierarchical tree) and another is based on regression analysis.

2. Methods

2.1 Clustering Method proposed by Gong Kang, Pan Lishuo, Cheng Yuxiao

Here we separate the pooled sample space $\mathcal{P} = \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ into k parts $(\mathcal{P}_1, \ldots, \mathcal{P}_k)$. For $Z_i \in \mathcal{P}$, let $p_{j(i)}$ be the conditional probability $Pr(Z_i \in \mathcal{P}_j | Z_i \sim F_i)$, $i = 1, 2; j = 1, \ldots, k$. Under $H_0 : F_1 = F_2$, we have:

$$p_{j(1)} = p_{j(2)}, \forall j = 1, \ldots, k.$$

Thus, the test procedure is firstly to clustering the pooled sample space based on some clustering methods. Let n_{ij} denote the number of the pooled sample points $Z_i \sim F_i$ clustered in $\mathcal{P}_j$. Under H_0,

$$e_{ij} = E_0[n_{ij}] = \frac{\sum_{j=1}^{k} n_{ij} \sum_{i=1}^{2} n_{ij}}{n+m}.$$

Then, we can calculate the Pearson Goodness of Fit test statistics

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{k-1},$$

and reject H_0 when $X^2 > \chi^2_{k-1,\alpha}$.

2.2 Logistic Regression Method proposed by Prof. Tsang

Here we propose two logistic regression method, one is straightforward and the other is modified with resampling.

Simple version

We label each Z_i with 0 if $Z_i \sim F_1$ or with 1 if $Z_i \sim F_2$. Let's say the train ratio is $\gamma \in (0, 1)$, then we random select $100\gamma\%$ of X_i and Y_j as the train set to fit a logistic regression. The remaining $100(1-\gamma)\%$ X_i and Y_j are denoted as the test sets $\mathcal{X}_{\text{test}}$ and $\mathcal{Y}_{\text{test}}$ respectively.

Then, we use the fitted model to predict $q_i^x = Pr(X_i \sim F_1|X_i), X_i \in \mathcal{X}_{\text{test}}$ and $q_j^y = Pr(Y_j \sim F_1|Y_j), Y_j \in \mathcal{Y}_{\text{test}}$.

Under $H_0 : F_1 = F_2$, q_i^x and q_j^y would follow the same distribution, and since they are one-dimensional random variable, we can use the existing hypothesis test, e.g. KS test.

Robust version

We denote $\mathcal{X}^{(0)}$ and $\mathcal{Y}^{(0)}$ as the original observed the $\{X_1, \dots X_m\}$ and $\{Y_1, \dots Y_n\}$ respectively. Each element in $\mathcal{X}^{(0)}$ or $\mathcal{Y}^{(0)}$ is labeled with 0 or 1 respectively. Then, we fit the logistic regression model based on $\mathcal{X}^{(0)}$ and $\mathcal{Y}^{(0)}$, and estimate $q_i^x = Pr(X_i \sim F_1|X_i), X_i \in \mathcal{X}^{(0)}, i = 1, \dots, m$; $q_j^y = Pr(Y_j \sim F_1|Y_j), Y_j \in \mathcal{Y}^{(0)}, j = 1, \dots, n$. Based on these estimated probabilities, we can use some one-dimensional test for distribution equality, e.g. KS test, to get the original p-value, $p^{(0)}$.

To simulate the distribution of the p-value, for $b^* = 1, \dots, b$, we select n elements from the pooled sample $\mathcal{P}$ without replacement to form $\mathcal{X}^{(b^*)}$; and $\mathcal{Y}^{(b^*)} = \{Z_i : Z_i \in \mathcal{P}, Z_i \notin \mathcal{X}^{(b^*)}\}$. Then, we fit the logistic regression model based on $\mathcal{X}^{(b^*)}$ and $\mathcal{Y}^{(b^*)}$, and estimate $q_i^x = Pr(X_i \sim F_1|X_i), X_i \in \mathcal{X}^{(b^*)}, i = 1, \dots, m$; $q_j^y = Pr(Y_j \sim F_1|Y_j), Y_j \in \mathcal{Y}^{(b^*)}, j = 1, \dots, n$. Therefore, we can calculate the p-value $p^{(b^*)}$ based on q_i^x, q_j^y and some one-dimensional test for distribution equality.

Therefore, we can reject H_0 if $p^{(0)} \notin (p_{\alpha/2}, p_{1-\alpha/2})$, where $p_{\alpha/2}$ and $p_{1-\alpha/2}$ are the $(100\alpha/2)^{\text{th}}$ and $(100(1-\alpha/2))^{\text{th}}$ percentile of $\{p^{(1)}, \dots, p^{(b)}\}$.

3. Implement

Here we simulate some data from different two multivariate norm distribution and estimate the reject ratio of these 3 methods, compared with Székely and Rizzo's method (2004) [1]. The Clustering method we choose is k-means. We set testing time is 1000.

	$\mu_1 = (0, 0.2, 0.4, 0.6, 0.8)$, $\mu_2 = (0.1, 0.3, 0.5, 0.7, 0.9)$, same Σ, $k=5$, $\gamma = 0.7$			
	Rizzo's	Clustering	Simple Logistic	Robust Logistic
n=20, m=20	0.051	0.032	0.801	1
n=20, m=30	0.058	0.033	0.865	1
n=50, m=100	0.068	0.061	1	1

	$\mu_1 = (0, 0.2, 0.4, 0.6, 0.8)$, $\mu_2 = (1, -0.5, 0, 0.5, -1)$, same Σ, $k=5$, $\gamma = 0.7$			
	Rizzo's	Clustering	Simple Logistic	Robust Logistic
n=20, m=20	0.999	0.961	1	1
n=30, m=50	1	0.993	1	1
n=50, m=100	1	1	1	1

4. Reference

[1] Gábor J Székely and Maria L Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10), 2004.