Manual of Testing for Equal Distributions Based on E-statistics

REN YAN

September 30, 2017

1 Statement of problem

Let X and Y be random vectors in $\mathbb{R}^d$ with distributions F_1 and F_2 respectively, $d \geq 1$. Let $\mathcal{A}_1 = \{X_1, \ldots, X_{n_1}\}$, $\mathcal{A}_2 = \{Y_1, \ldots, Y_{n_2}\}$ be finite sample sets containing random samples X_i, $Y_j \in \mathbb{R}^d$ of X and Y, $i = 1, 2, \ldots, n_1$, $j = 1, 2, \ldots, n_2$. Need to test $H_0 : F_1 = F_2$ against $H_1 : F_1 \neq F_2$ at significance level α.

2 Test statistic

The test statistic, namely E-statistic, is given by

$$\begin{aligned}\varepsilon_{n_1,n_2} = \frac{n_1 n_2}{n_1 + n_2} \Bigg(& \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{m=1}^{n_2} \|X_i - Y_m\| - \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \|X_i - X_j\| \\ & - \frac{1}{n_2^2} \sum_{l=1}^{n_2} \sum_{m=1}^{n_2} \|Y_l - Y_m\| \Bigg)\end{aligned} \tag{1}$$

where $\|\mathbf{a}\|$ denotes the Euclidean norm of some vector $\mathbf{a}$. The proposed test statistic is based on the V-statistic

$$V_{n_1,n_2} = \frac{1}{n_1^2 n_2^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \sum_{l=1}^{n_2} \sum_{m=1}^{n_2} h(x_i, x_j; y_l, y_m)$$

associated with the symmetric kernel function

$$h(x_i, x_j; y_l, y_m) = \|x_i - y_l\| + \|x_j - y_m\| - \|x_i - x_j\| - \|y_l - y_m\|$$

Under H_0, we have $E[h(x_i, x_j; y_l, y_m)] = 0$. And since $g(x, y) = E[h(x, X_1; y, Y_1)] = 0$ for almost all (x, y), V_{n_1,n_2} is a degenerate kernel V-statistic. In this case, it can be proved that with necessary moment conditions (on h), degenerate kernel V-statistic V_n satisfies

$$nV_n \xrightarrow{\mathrm{D}} \sum_{i=1}^{\infty} \lambda_i Z_i^2$$

where Z_i^2s are independent $\chi^2(1)$ random variables and λ_is are constants dependent on distribution F. Since computation shows that

$$\varepsilon_{n_1,n_2} = \frac{n_1 n_2}{n_1 + n_2} V_{n_1,n_2}$$

the test statistic is an asymptotic weighted sum of χ^2 variables. H_0 should be rejected if ε_{n_1,n_2} is too large.

3 Test procedure

Concerning the data of us, the total sample size $n = n_1 + n_2$ is large. As a result, the approximate permutation test, instead of the exact permutation test would be more feasible. Procedure:

1. Acquire 2 sample sets $\mathcal{A}_1$, $\mathcal{A}_2$ with cardinalities n_1, n_2 respectively.
2. Acquire pooled samples $\{W_1, \dots, W_n\} = \mathcal{A}_1 \cup \mathcal{A}_2$, $n = n_1 + n_2$.
3. Calculate the observed value of test statistic ε_n^{obs} based on $\mathcal{A}_1$, $\mathcal{A}_2$.
4. Define $m_j := \sum_{i=1}^{j} n_i$, $j = 1, 2$, $m_0 = 0$. Determine positive integer B such that $(B+1)\alpha$ is an integer.
5. Monte Carlo sampling (without replacement): for $b = 1, 2, \dots, B$, do:

(a). Acquire $\{W_1^{(b)}, \dots, W_n^{(b)}\}$, a random permutation of $\{W_1, \dots, W_n\}$.

(b). Let $\mathcal{A}_i^{(b)} = \{W_{m_{i-1}+1}^{(b)}, \dots, W_{m_i}^{(b)}\}$, $i = 1, 2$.

(c). Calculate $\varepsilon_n^{(b)}$ based on $\mathcal{A}_1^{(b)}$, $\mathcal{A}_2^{(b)}$.

6. Define *edf* of ε_n: $F_n(t) = P_n(\varepsilon_n \leqslant t) := \frac{1}{B} \sum_{b=1}^{B} I_{\{\varepsilon_n^{(b)} \leqslant t\}}$. Reject H_0 if ε_n^{obs} exceeds $100(1-\alpha)$ percent of the replicates $\varepsilon_n^{(b)}$. OR
7. Estimate p-value: $\hat{p} = \frac{1}{B} \sum_{b=1}^{B} I_{\{\varepsilon_n^{(b)} \geqslant \varepsilon_n^{obs}\}}$. Reject H_0 if $\hat{p} \leqslant \alpha$.

Note that Gandy, A. (2009) *Sequential Implementation of Monte Carlo Tests with Uniformly Bounded Resampling Risk* provides a good alternative in estimating p-value and bounding power loss of Monte Carlo tests compared with corresponding theoretical tests, which may be used for reference.