Computational Statistics and Data Analysis 105 (2017) 53–58

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A fast algorithm for two-dimensional Kolmogorov–Smirnov two sample tests

Yuanhui Xiao

Department of Mathematics and Statistics, Mississippi State University, Mississippi State, MS 39762, United States

ARTICLE INFO

Article history:
Received 31 December 2015
Received in revised form 14 July 2016
Accepted 19 July 2016
Available online 30 July 2016

Keywords:
Kolmogorov–Smirnov test
Brute force algorithm

ABSTRACT

By using the brute force algorithm, the application of the two-dimensional two-sample Kolmogorov–Smirnov test can be prohibitively computationally expensive. Thus a fast algorithm for computing the two-sample Kolmogorov–Smirnov test statistic is proposed to alleviate this problem. The newly proposed algorithm is $O(n)$ times more efficient than the brute force algorithm, where n is the sum of the two sample sizes. The proposed algorithm is parallel and can be generalized to higher dimensional spaces.

1. A fast algorithm for one-dimensional Kolmogorov–Smirnov test

Given two continuous probability distribution functions F^1 and F^2 in one-dimensional space, consider the hypothesis test problem

$$H_0 : F^1 = F^2 \quad vs. \quad H_a : F^1 \neq F^2 \tag{1}$$

based on the samples $\{X_i^1\}_{i=1}^{n_1}$ and $\{X_j^2\}_{j=1}^{n_2}$ from the respective distributions. The classical Kolmogorov–Smirnov test uses the maximum difference of the empirical distribution functions (or cumulative frequency functions) at the observed values. Specifically, let $F_{n_k}^k$ $(k = 1, 2)$ be the empirical distribution function based on the sample $\{X_t^k\}_{t=1}^{n_k}$, that is,

$$F_{n_k}^k(x) = \frac{\#\{t : X_t^k \leq x,\ 1 \leq t \leq n_k\}}{n_k}, \quad \infty < x < \infty, \tag{2}$$

where # means "the number of", then the Kolmogorov–Smirnov test statistic D_{KS} is computed as (up to a multiple)

$$D_{KS} = \max\{\max_{1\leq i\leq n_1} |F_{n_1}^1(X_i^1) - F_{n_2}^2(X_i^1)|,\ \max_{1\leq j\leq n_2} |F_{n_1}^1(X_j^2) - F_{n_2}^2(X_j^2)|\}. \tag{3}$$

The value of D_{KS} is often computed by a brute force algorithm, which simply counts the number of sample values that are less than X_i^1 or X_j^2 for each $i = 1, 2, \ldots, n_1$ and $j = 1, 2, \ldots, n_2$. The number of comparisons needed by the brute force algorithm is $O(n^2)$, where $n = n_1 + n_2$.

However, there exists a faster algorithm. Let L be the least common multiple of n_1 and n_2, $d_1 = L/n_1$, $d_2 = L/n_2$, and let

$$\{X_{(t)}^0 : 1 \leq t \leq n\} = \{X_{(1)}^0 \leq X_{(2)}^0 \leq \cdots \leq X_{(n)}^0\} \tag{4}$$

E-mail address: xiao_yuanhui@hotmail.edu.

http://dx.doi.org/10.1016/j.csda.2016.07.014
0167-9473/

be the pooled sample arranged ascendingly. (Throughout this paper we assume all the observed values have no ties when necessary.) Define

$$h_t = L \times [F^1_{n_1}(X^0_{(t)}) - F^2_{n_2}(X^0_{(t)})], \quad 0 \le t \le n. \tag{5}$$

The value of h_0 is set to be 0. The reader can easily verify the following recurrence:

$$h_t = \begin{cases} h_{t-1} + d_1 & \text{if } X^0_{(t)} = X^1_i \text{ for some } i, \\ h_{t-1} - d_2 & \text{if } X^0_{(t)} = X^2_j \text{ for some } j. \end{cases} \tag{6}$$

See Burr (1963), Hájek and Šidàk (1967) and Xiao et al. (2007). The value of the Kolmogorov–Smirnov test statistic is the maximum value of $|h_t|/L$ over $1 \le t \le n$:

$$D_{KS} = \max_{0 \le t \le n} |h_t|/L. \tag{7}$$

If the quick sort method is used, this algorithm only needs $O(n \log_2 n)$ comparisons (Hoare, 1961), which is $O(n)$ times more efficient than the brute force algorithm. In addition, the use of L even speeds up the algorithm since all the intermediate results are integers.

2. Generalization to two-dimensional spaces

The generalization of the Kolmogorov–Smirnov test to high dimensional probability distributions is a challenge. To generalize the Kolmogorov–Smirnov test to two-dimensional space, Peacock (1983) proposed a procedure which makes the use of four (rather than just one) pairs of cumulative frequency functions. Denote the two given samples in a plane by $\{(X^k_i, Y^k_i)\}^{n_k}_{i=1}$, $k = 1, 2$, respectively, the four pairs of cumulative frequency functions used by Peacock's test are given by

$$F^k_{++}(x, y) = \#\{i : X^k_i > x,\ Y^k_i > y,\ 1 \le i \le n_k\}/n_k, \tag{8}$$

$$F^k_{+-}(x, y) = \#\{i : X^k_i > x,\ Y^k_i \le y,\ 1 \le i \le n_k\}/n_k, \tag{9}$$

$$F^k_{-+}(x, y) = \#\{i : X^k_i \le x,\ Y^k_i > y,\ 1 \le i \le n_k\}/n_k, \tag{10}$$

and

$$F^k_{--}(x, y) = \#\{i : X^k_i \le x,\ Y^k_i \le y,\ 1 \le i \le n_k\}/n_k, \tag{11}$$

where $\infty < x, y < \infty$ and $k = 1, 2$. Let $\{X^0_t : t = 1, 2, \dots, n\}$ be the pooled data set consisting of the values of the X-components of the given samples and $\{Y^0_t : t = 1, 2, \dots, n\}$ the pooled data set consisting of the values of the Y-components of the given samples. Define

$$D_{++} \overset{def}{=} \max_{1 \le s \le n,\ 1 \le t \le n} |F^1_{++}(X^0_s, Y^0_t) - F^2_{++}(X^0_s, Y^0_t)|, \tag{12}$$

$$D_{+-} \overset{def}{=} \max_{1 \le s \le n,\ 1 \le t \le n} |F^1_{+-}(X^0_s, Y^0_t) - F^2_{+-}(X^0_s, Y^0_t)|, \tag{13}$$

$$D_{-+} \overset{def}{=} \max_{1 \le s \le n,\ 1 \le t \le n} |F^1_{-+}(X^0_s, Y^0_t) - F^2_{-+}(X^0_s, Y^0_t)|, \tag{14}$$

and

$$D_{--} \overset{def}{=} \max_{1 \le s \le n,\ 1 \le t \le n} |F^1_{--}(X^0_s, Y^0_t) - F^2_{--}(X^0_s, Y^0_t)|. \tag{15}$$

Peacock's test is then defined as

$$D_{2DKS} = \max\{D_{++},\ D_{+-},\ D_{-+},\ D_{--}\}. \tag{16}$$

The test is often performed by a brute force algorithm and its application is very expensive in terms of computing time unless the sample sizes n_1 and n_2 are very small. Indeed, to compute the value of D_{--}, we need to compute the value of the difference of the cumulative frequency functions F^1_{--} and F^2_{--} at all the n^2 pairs (X_s, Y_t), X_s and Y_t being coordinates of any pairs in the given samples. It will need $O(n)$ comparisons to compute the value of the difference of the cumulative frequency functions F^1_{--} and F^2_{--} at a single point. Thus, it will take $O(n^3)$ comparisons to compute the value of D_{--}. Similar conclusions can be made for D_{++}, D_{+-}, D_{-+}.

To alleviate the problem, Fasano and Franceschini (1987, F&F, for short) revised Peacock's test by comparing the cumulative frequency functions at the observed sample points only, so the number of comparisons needed is only $O(n^2)$. The F&F test is widely used in practice. But it is a variant of Peacock's test, a different approach in essence.

In fact, there exists a fast algorithm for evaluating the value of Peacock's test statistic. Denote by $\{(X'_{(t)}, Y'_t) : 1 \le t \le n\}$ the pooled sample sorted ascendingly by the values of the X-components of the data points, and by $\{(X'_t, Y'_{(t)}) : 1 \le t \le n\}$