A Multivariate Two-Sample Test using the Jaccard Distance

Stuart B. Heinrich

717 Market St, San Francisco, CA

Abstract

A common need in statistics is to assess whether two samples come from the same underlying population distribution. Existing two-sample tests often make limiting a priori assumptions, or cannot be easily generalized to multivariate data. We derive a new multivariate two-sample test that makes no a priori assumptions, has higher statistical power than previous tests, has better runtime performance, has an easily understood geometrical interpretation, and is simple to implement.

Keywords: two-sample test, multivariate, distribution equality, Jaccard distance

1. Introduction

A two-sample test is a statistical test that attempts to determine if two samples come from the same population. One of the simplest and most commonly used two-sample tests is the Student's *t*-test [25], which makes the *a priori* assumptions that the samples are normally distributed and homoscedastic (i.e., that they have equal variance), thereby reducing the problem to a question of whether the population means of the two samples are equal. This test can be easily generalized to multivariate data to test for equality of the mean vectors, known as Hotelling's *t*-squared test.

Whereas the assumption of normality can *sometimes* be a reasonable assumption based on the problem domain, homoscedasticity can rarely be assumed, and hence must usually be validated by some other test. From a pedantic standpoint, homoscedasticity is only *truly* satisfied when the null-hypothesis is true.

Nonetheless, the two-phase approach of first testing for (approximately) equal variance and then testing for equal mean is often sufficient for making simple binary judgements of equality between two distributions. However, it is clearly suboptimal in comparison to a test that considers all information simultaneously in order to make a single decision. Moreover, if a real-valued metric for overall similarity is needed for some purpose other than binary hypothesis testing, such a two-phase approach is inapplicable.

The Behrens-Fisher problem (BFP) [1, 9], which also assumes normality, and can also be generalized to multivariate data [15, 2], is sometimes thought to overcome the limiting assumption of homoscedasticity. In fact, the assumption of homoscedasticity is merely replaced with the even more unreasonable *a priori* assumption of heteroskedasticity. As pointed out by Sawilowsky [20], there are arguably no practical circumstances under which this assumption is reasonable, because if the samples do come from the same population, they should have equal variance.

Clearly, it would be preferable to have a test that does not make any *a priori* assumptions about the sameness of the population variances. There are several well-known two-sample tests in this category, such as the non-parametric two-sample Kolmogorov-Smirnov (KS) test [17, 24], the two-sample Anderson-Darling test [19], Wilcoxon's signed-rank test [28] and the Mann-Whitney U-test [18]. However, unlike the *t*-test and the BFP, these statistics require the random variables in each sample to be ordinal, and hence they do not naturally generalize to multivariate data (although there have been some marginally successful attempts [8, 16]).

Indeed, there are not many two-sample tests that are designed for multivariate data. Perhaps the first such test was made by attempting to generalize the classic Smirnov test [4, 10], followed by several variants of the non-parametric nearest neighbor (NN) test [21, 12, 11]. Zech and Aslan [29] proposed a method based on minimum energy [29], and more recently the non-parametric $\mathcal{E}$-test was proposed [26].

In this manuscript, we derive a new test statistic that does not make any *a priori* assumptions about the variances, and greatly outperforms these aforementioned existing multivariate two-sample tests. It is simply generalized to multivariate distributions, and can also be generalized for non-parametric data, making the assumption of normality optional instead of mandatory. It has better receiver operating characteristics (ROC) and higher statistical power than previous multivariate two-sample tests, making it more reliable for small sample sizes. Under the assumption of normality, it has linear time complexity (which is superior to previous algorithms). Finally, it has an easily understood geometric interpretation.

2. Approach

Given a sample of independent and identically distributed (IID) random variables $S_A = \{\mathbf{a}_1, \ldots, \mathbf{a}_{n_1}\}$ from distribution A and a sample of IID random variables $S_B = \{\mathbf{b}_1, \ldots, \mathbf{b}_{n_2}\}$ from distribution B, where $\mathbf{a}_i \in \mathbb{R}^d$ and $\mathbf{b}_i \in \mathbb{R}^d$, our objective is to assess the statistical significance of the difference between the

distributions A and B. Thus, the null hypothesis is $H_0 : A = B$ and the alternative is $H_a : A \neq B$.

Any probability density function (PDF) denoted by $f(x)$ may be equivalently interpreted as the set of all points under the probability density curve,

$$\{(x, y) \in \mathbb{R}^2 | f(x) > y\}. \tag{1}$$

Given two sets S_1 and S_2, the Jaccard [14] distance $J(S_1, S_2) \in [0, 1]$ is given by

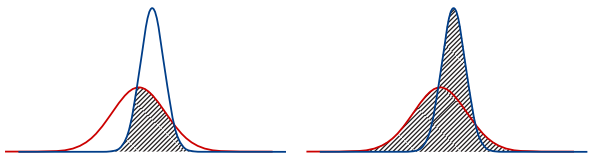
$$J(S_1, S_2) = 1 - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}, \tag{2}$$

or in plain terms, it is 1 minus the ratio of the overlap (intersection) area to the total (union) area.

The basic approach of our method is to estimate the two sample distributions, interpret these distributions as sets, and then measure similarity between the sets using the Jaccard distance [14].

When PDFs are represented as sets according to (1), the area of the intersection is given by integrating over the minimum between the two PDFs, and the area of the union is given by integrating over the maximum (Fig. 1); thus, given two PDFs P_A and P_B, the Jaccard distance between the two distributions is given by

$$1 - \frac{\int_{-\infty}^{\infty} \min\{P_A(x), P_B(x)\}\, dx}{\int_{-\infty}^{\infty} \max\{P_A(x), P_B(x)\}\, dx}. \tag{3}$$



(a) (b)

Figure 1. Example of the intersection (left) and union (right) between the set-representation of two normal probability density functions.

It is straightforward to generalize (1) for multivariate distributions. Given a multivariate PDF $f(\mathbf{x})$ where $\mathbf{x} \in \mathbb{R}^d$, the equivalent set representation is

$$\{(\mathbf{x}, y) \in \mathbb{R}^{d+1} | f(\mathbf{x}) > y\}, \tag{4}$$

and hence (3) is generalized by simply integrating over all dimensions,

$$1 - \frac{\int\int\cdots\int_{\mathbb{R}^d} \min\{P_A(\mathbf{x}), P_B(\mathbf{x})\}\, d\mathbf{x}}{\int\int\cdots\int_{\mathbb{R}^d} \max\{P_A(\mathbf{x}), P_B(\mathbf{x})\}\, d\mathbf{x}}. \tag{5}$$

In some cases, this infinite integral can be computed exactly; for example, when P_A and P_B are univariate normal distributions, (3) can be computed from the normal CDF after finding the roots of $P_A(x) - P_B(x)$.

When P_A and P_B are not both normal, or when $d > 1$, computation of an exact solution is no longer practical. However, a good approximation can be computed using numerical integration. An approximation to the integral of any function $f(x)$ is given by

$$\int f(x)dx \approx \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)}{p(x_i)}, \quad n \gg 1, \tag{6}$$

where $x_i \sim p$, for some arbitrary sampling distribution p [23, p.289].

This approximation becomes increasingly accurate as $n \to \infty$, and converges faster when p is similar to f. A logical choice for p is to use the pooled sample distribution, because this tightly envelopes both the intersection and union functions to be integrated. Furthermore, it makes sense to choose a number of Monte Carlo samples proportional to the number of pooled samples so that the integration accuracy does not lag behind the information content of the samples.

It would be computationally wasteful to attempt to derive a parametric model for p based on the pooled samples, only to draw this many samples from it again. Thus, we propose to ignore p and simply use the pooled sample points as the Monte Carlo integration points; that is, if we denote the pooled sample by $S_C = S_A \cup S_B = \{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$, then (5) is approximated by

$$1 - \frac{\frac{1}{n}\sum_{i=1}^{n} \min\{P_A(\mathbf{c}_i), P_B(\mathbf{c}_i)\}}{\frac{1}{n}\sum_{i=1}^{n} \max\{P_A(\mathbf{c}_i), P_B(\mathbf{c}_i)\}}. \tag{7}$$

After cancelling out $1/n$, we obtain a statistic of

$$J(S_A, S_B) = 1 - \frac{\sum_{i=1}^{n} \min\{P_A(\mathbf{c}_i), P_B(\mathbf{c}_i)\}}{\sum_{i=1}^{n} \max\{P_A(\mathbf{c}_i), P_B(\mathbf{c}_i)\}}. \tag{8}$$

If one assumes multivariate normal populations, then

$$P_A(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mu_A)^{\mathsf{T}}\Sigma_A^{-1}(\mathbf{x}-\mu_A)} / \left(\sqrt{(2\pi)^d |\Sigma_A|}\right), \tag{9}$$

where μ_A and Σ_A are the sample mean and covariance of S_A (and similarly for P_B). The advantage of using a simple multivariate model is that these distributions can be estimated and sampled very efficiently. Moreover, because the multivariate normal has very few degrees of freedom, it will have high statistical power and therefore perform very well on small sample sizes (i.e., perhaps less than 10).

When the assumption of normality is too restrictive, (8) may be easily generalized into a non-parametric statistic by using a multivariate kernel density estimate (KDE) for P_A, P_B; that is,

$$P_A(\mathbf{x}) = \frac{1}{n_1 \sqrt{|\mathbf{H}|}} \sum_{i=1}^{n_1} K\left(\frac{\mathbf{x} - \mathbf{a}_i}{\sqrt{|\mathbf{H}|}}\right), \tag{10}$$

where $K(.)$ is the kernel, typically the standard multivariate normal kernel,

$$K(\mathbf{x}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{x}\right), \tag{11}$$

and $\mathbf{H}$ is the bandwidth matrix, which can be chosen in a data-dependent fashion using various automatic methods [5, 27, 6]. The KDE for P_B is defined similarly.

2.1. Permutation Distribution

The significance of the test statistic (8) may be assessed by calculating the probability (p-value) of obtaining a statistic at least as extreme as the one observed under the null hypothesis. The null hypothesis may then be rejected if the p-value is less than the desired significance level of $\alpha = 0.05$, for example.

A simple and general method of obtaining a p-value that can be used with any test statistic is to use the permutation distribution (see Efron [7] and Higgins [13, p.31]). Under the null hypothesis, the populations $A = B = C$ and hence S_C is a set of IID random variables; thus, one should be able to randomly rearrange the samples without changing the expected value of the test statistic. Specifically, for the bth permutation sample, one can randomize the order of S_C, reassign the elements to S_A^b and S_B^b such that $\#S_A^b = \#S_A$ and $\#S_B^b = \#S_B$, and then compute the test statistic T^b between S_A^b and S_B^b.

Because larger values of the test statistic indicate divergence from the null hypothesis, the permutation estimate of the p-value is then

$$p = P(T^b \geq T) = \lim_{B\to\infty} \frac{1}{B} \sum_{b=1}^{B} I(T^b \geq T), \tag{12}$$

where $I(.) \in \{0, 1\}$ is the indicator function, and B is the number of permutation samples.

In this paper, we always use this approach (with $B = 1000$) for converting test statistics into p-values.

2.2. Time Complexity

Using the multivariate normal model, computation of the means and covariances in (9) is linear in the sample size. Inversion of the covariance matrix may be pre-computed and takes $O(d^3)$ time. At this point, evaluation of (8) is again linear in the sample size. Thus, the total time complexity is just $O(d^3 + n)$. For most practical problems $d \leq 3$, and can be regarded as a constant.

The non-parametric version of the J-statistic requires evaluating the multivariate KDE at each point in the samples. The multivariate KDE can be most efficiently evaluated by solving the fixed-radius near neighbor problem [3] in order to find all sample points within the support region of the bandwidth matrix. Using a kd-tree structure this can be done in amortized $O(\log n)$ time, leading to a total amortized runtime of $O(n \log n)$.

The nearest neighbor test requires finding the kth nearest neighbor for each of the n sample points, which can be computed in $O(\log n)$ amortized time after building a kd-tree[22]. Thus, the overall amortized runtime is $O(n \log n)$, which is the same as the time needed to build the kd-tree. It may be easily verified that the time complexity of the $\mathcal{E}$-statistic [26] is $O(n^2)$.

3. Experiments & Results

3.1. Runtime Performance

Runtime performance of each method was measured on an Intel Core i7-2600 CPU (2.4 GHz) machine. From the linearly scaled plot (Fig. 2, top) it is evident that the J-test has by far the best performance on large sample sizes, with a runtime of 0.014 seconds for 70,000 samples. The NN-test was the next fastest at 4.67 seconds. In contrast, the $\mathcal{E}$-test and non-parametric J-test were comparatively very slow, taking 189 seconds and 75 seconds, respectively.

From the log-log plot of performance (Fig. 2, bottom), we see that, despite having the highest time complexity, the $\mathcal{E}$-test actually has the best performance for extremely small sample sizes (i.e., less than 200). This is because the algorithm is so simple that there is almost no memory or runtime overhead. However, the algorithms are all so fast with these small number of samples that the advantage is inconsequential.

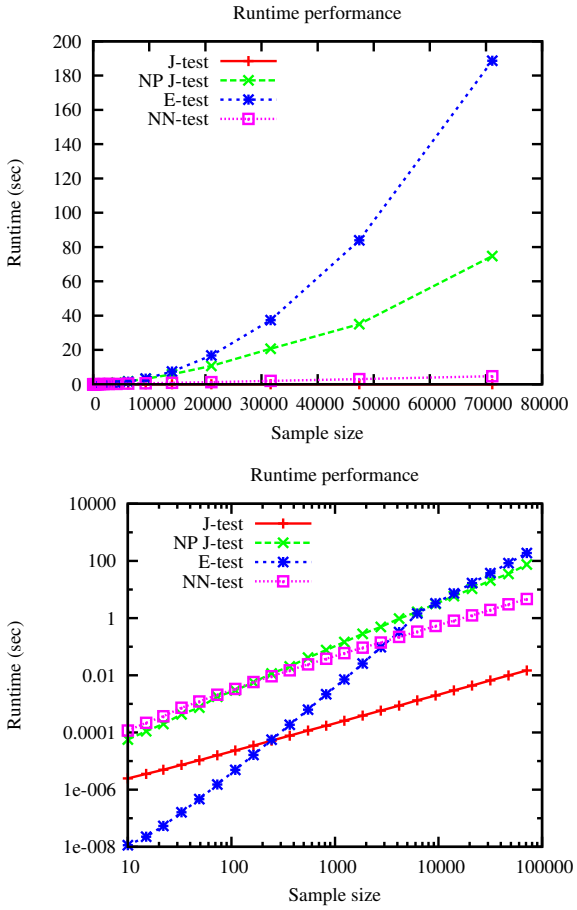


Figure 2. Runtime performance as a function of sample size on 3-dimensional multivariate data. Top: linear scale. Bottom: log-log scale.

3.2. Statistical Performance

We begin with a graphical analysis showing the raw statistic response and p-values in comparison to other univariate two-sample statistics (Fig. 3). We use the standard normal distribution as a reference for population A and then vary the mean and standard deviation for population B. We compare the J-test (assuming normality), the non-parametric (NP) J-test (using the KDE), $\mathcal{E}$-test, the NN-test using 3rd nearest neighbors (as recommended in Szekely and Rizzo [26]) and the KS-test.

In the first row, we show the raw statistics using a small sample size of $n_1 = n_2 = 10$. In the second row, we increase the sample sizes to $n_1 = n_2 = 100$. The third and fourth rows show the corresponding p-values corresponding to the first and second rows, respectively.

As expected, each statistic has a minimum value when $\mu_2 = 0$, $\sigma_2 = 1$ because this is when the two populations are actually equal. However, the profile of the statistic as the populations diverge is very different. The J-test increases most rapidly as the standard deviations diverge, whereas, for example, the KS-test is less sensitive. The p-values calculated from each statistic with this number of samples look approximately equivalent, being essentially a delta function that properly rejects the null hypothesis when the distributions are unequal, although the NN-statistic has a notably less-well defined peak.

The difference between test statistics is more apparent for smaller sample sizes (top row), where it can be seen that the statistics have different levels of 'background noise.' This background noise is effectively a visual indicator of statistical power, because it shows how sensitive they are to random chance. The parametric J-test has the least background noise, followed closely by the non-parametric J-test, then the $\mathcal{E}$-test, NN-test, and KS-test.

Formally, the statistical power of a test is the probability of rejecting the null-hypothesis at a particular statistical significance level for a particular magnitude of the difference between populations. We make the differences in statistical power more explicit by plotting power as a function of the difference in standard deviation for each statistic, based on 1000 replications, using a sample size of $n_1 = n_2 = 10$ and a significance level of $\alpha = 0.05$. We fix $\sigma_1 = 1$ and let $\sigma_2 = 10, 8, 6.4 \ldots 1.342$ in multiplicative increments of 0.8 (see Fig. 4).

Our results confirm what was visually apparent from Fig. 3: that the J-test (either parametrically or non-parametrically) has the highest statistical power for detecting a change in standard deviation; this is followed by the $\mathcal{E}$-test and NN-test which have approximately equal statistical power, and finally the KS-test, which has significantly lower statistical power.

All other things beings equal, a statistic with higher statistical power is superior because it can be used to detect a change from a smaller sample size. Thus, for a given statistical power of 0.8, for example, we can calculate the minimum number of samples necessary to detect a change in the standard deviation by iteratively increasing the sample size and re-computing statistical power until the desired threshold is exceeded. Using this approach, we have calculated the minimum sample size to detect a change in standard deviation using each test in Fig. 6.

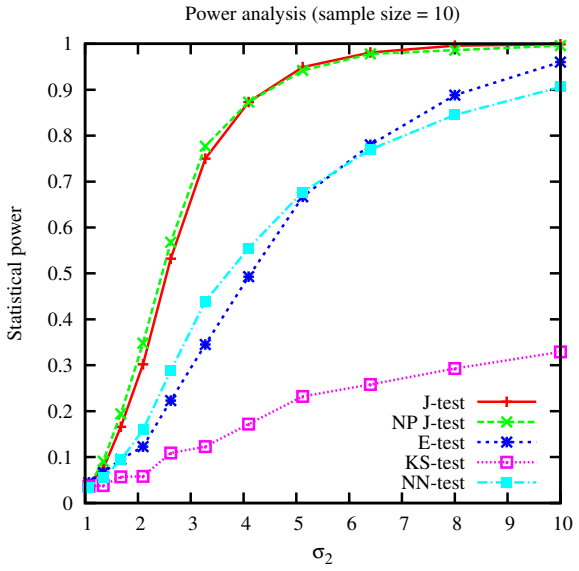


Figure 4. Statistical power for detecting a change from the standard normal distribution at a significance level of $\alpha = 0.05$, estimated from 1000 replications, using a sample size of $n_1 = n_2 = 10$.

As the sample standard deviations converge (that is, as σ_2 approaches $\sigma_1 = 1$ from the right), the number of samples required to detect a change grows faster than exponentially. Due to computational limitations, we were not able to acquire enough samples to detect the most subtle difference of $\sigma_2 = 1.34$ using the NN-test, so this data point is omitted from the plot.

For relatively large differences in the standard deviation, for example when $\sigma_2 = 10$, most tests require 6-9 samples, except for the NN-test which requires almost 20 samples. However, for more subtle differences in standard deviation, the change in the number of samples is more significant. For example, when $\sigma_2 = 1.34$, in order to achieve power ≥ 0.8 with a significance level of $\alpha = 0.05$, the J-test requires a sample size of about 100, the non-parametric J-test requires a sample size of about 150, the $\mathcal{E}$-test requires a sample size of about 240, and the KS-test requires a sample size of about 400.

Finally, we examine statistical power in higher dimensional problems, where the KS-test is no longer applicable. Specifically, we use $d = 3$ with a standard trivariate normal reference distribution. We compare the statistical power of the J-test, the $\mathcal{E}$-test and the NN-test for their ability to detect a change in the mean-vector, the overall scale of the covariance matrix, and the rotation of the covariance matrix about the z-axis. For this latter test, we use an anisotropic reference distribution having covariance $\Sigma_1 = \mathrm{diag}(10^2, 1, 1)$. Statistical power is assessed based on 1000 replications and the null-hypothesis is rejected at a significance level of $\alpha = 0.05$.

We observe that the $\mathcal{E}$-test has a minor advantage in statistical power for detecting changes in the mean vector (Fig. 5, left). For detecting changes in the overall scale of the covariance matrix, the J-test is significantly more powerful, and the $\mathcal{E}$-test and NN-test are about the same (Fig. 5, middle). Finally, we observe that the $\mathcal{E}$-test fails to detect changes in the rotation of the covariance matrix regardless of sample size (Fig. 5, right);

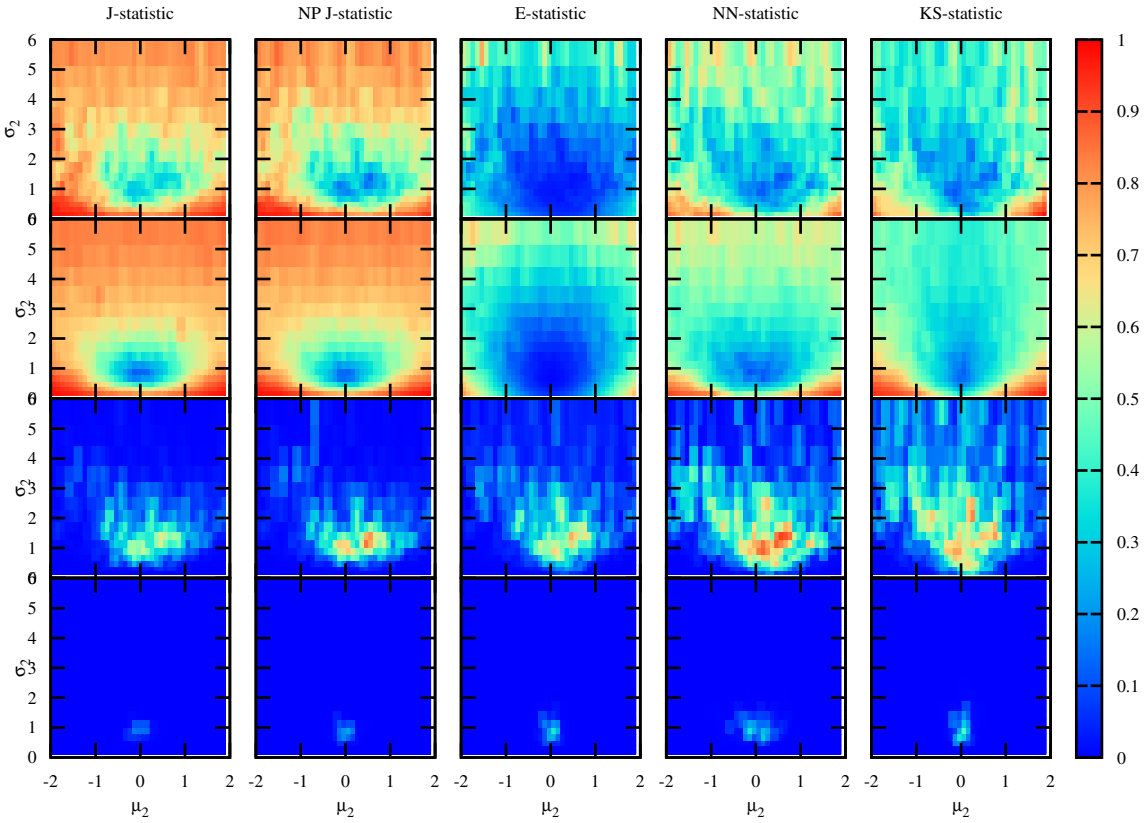


Figure 3. Comparison of test statistics for detecting a change from the standard normal distribution. First row: raw test statistic with sample size of $n_1 = n_2 = 10$. Second row: raw test statistic with sample size of $n_1 = n_2 = 100$. Third row: p-values corresponding to test statistic from first row. Fourth row: p-values corresponding to test statistic from second row.

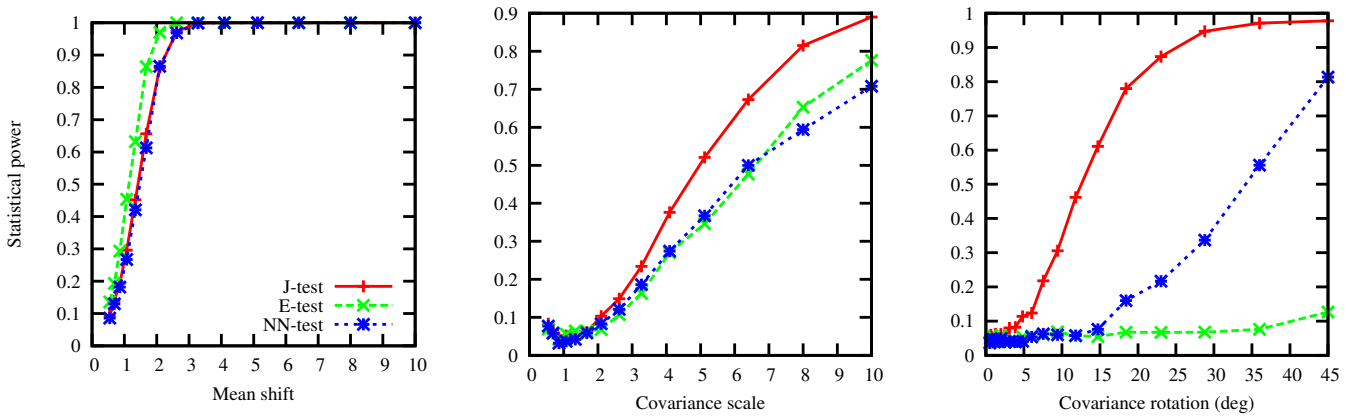


Figure 5. Statistical power for detecting changes in a 3-dimensional multivariate normal distribution at a significance level of $\alpha = 0.05$, estimated from 1000 replications, using a sample size of $n_1 = n_2 = 10$; (a) from shifting the mean-vector; (b) from scaling the covariance matrix; (c) from rotating the covariance matrix, in comparison to a reference distribution with covariance $\Sigma_1 = \mathrm{diag}(10^2, 1, 1)$.

the NN-test can detect changes, but is not nearly as powerful as the J-test, which continues to perform quite well.

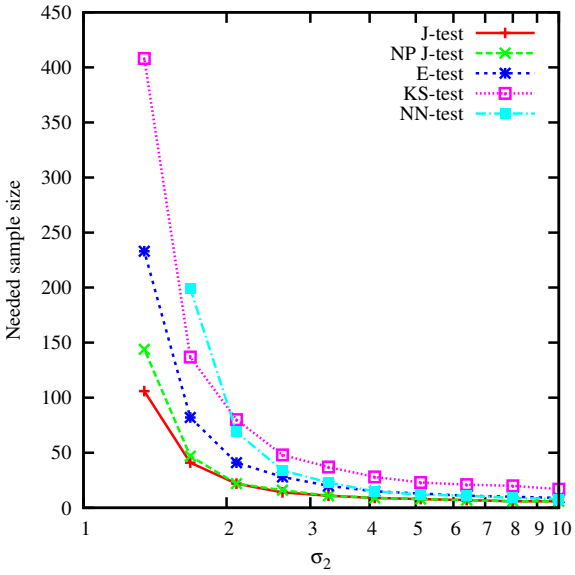


Figure 6. Minimum sample size ($n_1 = n_2$) in order to obtain statistical power ≥ 0.8 for detecting a change at a significance level of $\alpha = 0.05$.

Finally we compare the receiver operating characteristics (ROC) of the various two-sample tests on both univariate (Fig. 7) and multivariate (Fig. 8) data. In the univariate case, we generate 1000 unique normal distributions having mean uniformly distributed in $(-1, 1)$ and standard deviation uniformly distributed in $(1, 100)$. We perform 100 trials, and on trial we randomly select two equal distributions (25% of the time) or two unequal distributions (75% of the time). We generate 50 random samples from the two selected distributions and then classify the distributions as either equal or unequal based on the computed p-value for each test statistic, as we vary the significance level from $\alpha = 0.01, 0.02, 0.05, 0.1 \ldots 0.9$.

The multivariate test is performed identically, except that the distributions are trivariate normal. The random trivariate normal distributions are constructed starting from a diagonal covariance matrix Σ with variances randomly chosen in $(1, 100)$. We then construct a random rotation $\mathbf{R}$ from a random axis and angle, and rotate the covariance matrix to obtain $\Sigma' = \mathbf{R}^{-1}\Sigma\mathbf{R}$. The mean vector is randomly chosen in the range $(-1, 1)$.

In the univariate case it can be seen that the J-test performs the best, obtaining the highest true positive rate (TPR) as a function of false positive rate (FPR), while the KS-test performs the worst.

In the multivariate case the KS-test is dropped because it does not work for multivariate data. Here we see that the J-test again performs the best, this time by an even more significant margin, and the NN-test performs the worst (when operating at less than 30% false positive rate).

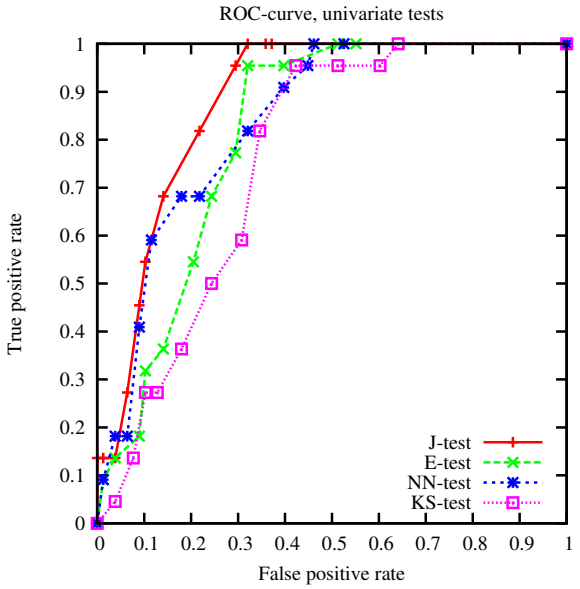


Figure 7. Receiver operating characteristics (ROC) of the test statistics for detecting a difference between normal distributions as a function of the significance level.

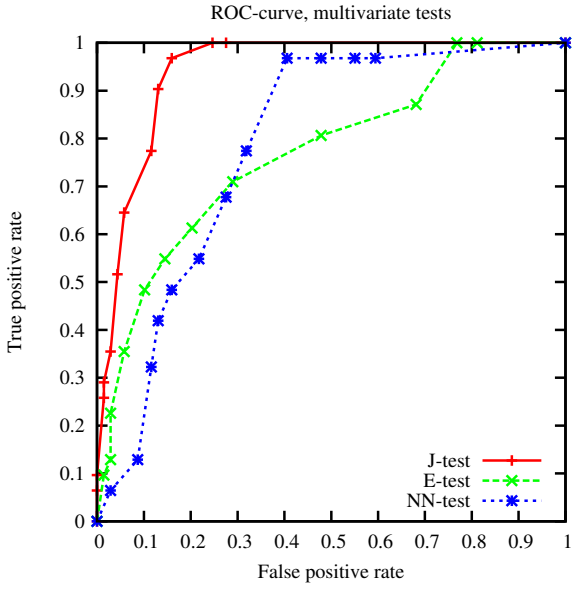


Figure 8. Receiver operating characteristics (ROC) of the test statistics for detecting a difference between trivariate normal distributions as a function of the significance level.

4. Conclusions

We have shown that previous two-sample tests are often lacking, being either limited to univariate data, or inefficient to calculate on high-dimensional data, or incapable of detecting all possible types of changes in a normal distribution such as rotation of the covariance matrix.

In order to meet these needs, we have presented the J-statistic, a multivariate two-sample test that can be used parametrically or non-parametrically. Computation of the statistic

is simple to implement, and can be done in linear time under the assumption of normality, which is faster than any existing multivariate two-sample tests. We have shown that the J-statistic has better receiver operating characteristics (ROC), and higher statistical power then previous univariate or multivariate statistics, particularly when it comes to detecting changes in the standard deviation or 'rotation' of the covariance matrix (which cannot be detected by the previous $\mathcal{E}$-statistic).

References

[1] B. V. Behrens. Ein beitrag zur fehlerberechnung bei wenige beobachtungen. *Landwirtschaftliches Jahresbuch*, 68:807–837, 1929.

[2] Alexandre Belloni. On the behrens-fisher problem: A globally convergent algorithm and a finite-sample study of the wald, lr and lm tests. *Annals of Statistics*, 36(5):2377–2408, 2008.

[3] Jon Louis Bentley. A survey of techniques for fixed-radius near neighbor searching. Technical report, Stanford Linear Accelerator Center, 1975. Technical Report SLAC-186 and STAN-CS-75-513.

[4] P.J. Bickel. A distribution free version of the smirnov two sample test in the p-variate case. *Annals of Mathematical Statistics*, 40(1):1–23, 1969.

[5] R. Cao, A. Cueveas, and W.G. Manteiga. A comparative study of several smoothing methods in density estimation. *Computational Statistics and Data Analysis*, 17:153–176, 1994.

[6] T. Duong and M.L. Hazelton. Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, 15:17–30, 2003.

[7] Bradley Efron. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

[8] G. Fasano and A. Franceschini. A multidimensional version of the kolmogorov-smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225:155–170, 1987.

[9] R.A. Fisher. The fiducial argument in statistical inference. *The Annals of Eugenics*, 11:141–172, 1935.

[10] Jerome H. Friedman and Lawrence C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717, 1979.

[11] Peter Hall and Nader Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002.

[12] N. Henze. A multivariate two-sample test based on the number of nearest neighbor coincidences. *Annals of Statistics*, 16:772–783, 1988.

[13] James J. Higgins. *Introduction to Modern Nonparametric Statistics*. Brooks/Cole, Pacific Grove, CA, 2004.

[14] Paul Jaccard. Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[15] Richard A. Johnson and Samaradasa Weerahandi. A bayesian solution to the multivariate behrens-fisher problem. *Journal of the American Statistical Association*, 83(401):145–149, 1998.

[16] Ana Justel, Daniel Pena, and Ruben Zamar. A multivariate kolmogorov-smirnov test of goodness of fit. *Statistics Probability Letters*, 35(3):251–259, 1997.

[17] A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933.

[18] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[19] A. N. Pettitt. A two-sample anderson-darling rank statistic. *Biometrika*, 63(1):161–168, 1976.

[20] Shlomo S. Sawilowsky. Fermat, schubert, einstein, and behrens-fisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$. *Journal of Modern Applied Statistical Methods*, 1(2):461–472, 2002.

[21] M. F Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81:799–806, 1986.

[22] G. Shaknarovich, T. Darrell, and P. Indyk. *Nearest-Neighbor Methods in Learning and Vision*. MIT Press, Cambridge, Massachusetts, 2005.

[23] Peter Shirley, Michael Ashikhmin, Michael Gleicher, Stephen Marschner, Erik Reinhard, Kelvin Sung, William Thompson, and Peter Willemsen. *Fundamentals of Computer Graphics, Second Ed.* A. K. Peters, Ltd., Natick, MA, USA, 2005. ISBN 1568812698.

[24] N. Smirnov. Sur la distribustion de w^2. *Comptes Rendus de l'Academie des Sciences*, 202:449–452, 1936.

[25] Student. The probable error of a mean. *Biometrika*, 6:1–15, 1908.

[26] Gabor J. Szekely and Maria L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, Nov 2004. `http://interstat.statjournals.net/YEAR/2004/abstracts/0411005.ph`

[27] M.P. Wand and M.C. Jones. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9:97–177, 1994.

[28] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[29] G. Zech and B. Aslan. A multivariate two-sample test based on the concept of minimum energy. In *In PHYSTAT*, 2003.