A Multivariate Two-Sample Test Based on the Concept of Minimum Energy

G. Zech and B. Aslan
University of Siegen, 57072 Siegen, Germany

We introduce a new statistical quantity the *energy* to test whether two samples originate from the same distributions. The energy is a simple logarithmic function of the distances of the observations in the variate space. The distribution of the test statistic is determined by a resampling method. The power of the energy test in one dimension was studied for a variety of different test samples and compared to several nonparametric tests. In two and four dimensions a comparison was performed with the Friedman-Rafsky and nearest neighbor tests. The two-sample energy test is especially powerful in multidimensional applications.

1. INTRODUCTION

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ be two samples of independent random vectors with distributions F and G, respectively. The classical two-sample problem then consists of testing the hypothesis

$$H_0 : F(\mathbf{x}) = G(\mathbf{x}), \text{ for every } \mathbf{x} \in \mathbb{R}^d,$$

against the general alternative

$$H_1 : F(\mathbf{x}) \neq G(\mathbf{x}), \text{ for at least one } \mathbf{x} \in \mathbb{R}^d,$$

where the distribution functions F and G are unknown.

Testing whether two samples, for example, two data sets taken at different times, are consistent with a single unknown distribution is a task that occurs in many areas of research. Clearly tests based on moments [1–3] are not sensitive to all alternatives H_1. Other tests require binning of data like the power-divergence statistic test [4] and tests of the χ^2 type. However, a high dimensional space is essentially empty, as is expressed in the literature by the term *curse of dimensionality* [5], hence tests based on binning are rather inefficient unless the sample sizes are large. Binning-free tests based on rank statistics are restricted to univariate distributions, and, when applied to the marginal distributions, they neglect correlations. The Friedman-Rafsky test [6] and the nearest neighbor test [7] avoid these caveats.

The *Friedman-Rafsky test* can be seen as a generalization of the univariate Wald-Wolfowitz run test [8]. The problem in generalizing the run test to more than one dimension is that there is no unique sorting scheme for the observations. The minimum spanning tree can be used for this purpose. It is a graph which connects all observations in such a way that the total Euclidean length of the connections is minimum. Closed cycles are inhibited. The minimum spanning tree of the pooled sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ is formed. The test statistic R_{nm} equals the number of connections between observations from different samples. Small values of R_{nm} lead to a rejection of H_0. The statistic R_{nm} is asymptotically distribution-free under the null hypothesis [9].

The *nearest neighbor test* statistic N_{nm} is the sum of the number of observations of the pooled sample where the nearest neighbor is of the same type. In [7] it is shown that the limiting distribution of N_{nm} is normal in the limit $\min(n, m) \to \infty$. Large values of N_{nm} lead to rejection of H_0.

In this paper we propose a new test for the two-sample problem - the *energy test* - which shows high performance independent of the dimension of the variate space and which is easy to implement. Our test is related to Bowman-Foster test [10] but whereas this test is based on probability density estimation and local comparison, the energy test explores long range correlations.

2. THE TWO-SAMPLE *ENERGY* TEST

The basic idea behind using the quantity *energy* to test the compatibility of two samples is simple. We consider the sample $A : \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ as a system of positive charges of charge $1/n$ each, and the second sample $B : \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ as a system of negative charges of charge $-1/m$. The charges are normalized such that each sample contains a total charge of one unit. From electrostatics we know that in the limit of where n, m tend to infinity, the total potential energy of the combined samples computed for a potential following a one-over-distance law will be minimum if both charge samples have the same distribution. The energy test generalizes these conditions. For the two-sample test we use a logarithmic potential in $\mathbb{R}^d$. In Ref. [11] we show that also in this case, large values of energy indicate significant deviations between the parent populations of the two samples. The proof relies on the fact that the Fourier transform of the kernel function $1/r^\kappa$ is positive definite. The logarithmic function is equivalent to the inverse power function in

the limit where the exponent tends to zero.

The test statistic Φ_{nm} consists of three terms, which correspond to the energies of samples A, B and the interaction energy of the two samples

$$\Phi_{nm} = \frac{1}{n^2}\sum_{i<j}^{n} R\left(|\mathbf{x}_i - \mathbf{x}_j|\right) + \\ + \frac{1}{m^2}\sum_{i<j}^{m} R\left(|\mathbf{y}_i - \mathbf{y}_j|\right) + \\ - \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} R\left(|\mathbf{x}_i - \mathbf{y}_j|\right)$$

where $R(r)$ is a continuous, monotonic decreasing function of the Euclidean distance r between the charges. The choice of R may be adjusted to a specific statistical problem. With the choice $R(r) = -\ln r$ the test is scale invariant and offers a good rejection power against many alternatives to the null hypothesis.

To compute the power of the new two-sample *energy* test we use the permutation method [12] to evaluate the distribution of Φ_{nm} under H_0. We merge the $N = m + n$ observations of both samples and draw from the combined sample a subsample of size n without replacement. The remaining m observations represent a second sample. The probability distribution under H_0 of Φ_{nm} is evaluated by determining the values of Φ_{nm} of all $\binom{N}{m} = \frac{N!}{n!m!}$ possible permutations. For large N this procedure can become computationally too laborious. Then the probability distribution is estimated from a random sample of all possible permutations.

We propose to normalize the vectors $\mathbf{z}_i$, $i = 1, 2, \ldots, N$ of the pooled sample to unit variance in all projections, $z^*_{ik} = (z_{ik} - \mu_k)/\sigma_k$, where μ_k, σ_k are mean value and standard deviation of the projection $z_{1k}, \ldots, z_{Nk}$ of the coordinates of the observations of the pooled sample. In this way we avoid situations in which a single projection dominates the value of the distance and consequently of the energy and that other projections contribute only marginally to it. We have not studied the effect of this scaling procedure which probably is sensible for all multidimensional goodness-of fit tests. In the following power comparison of our method with the competing methods, the different projections were not normalized.

3. POWER COMPARISONS

The performance of various tests were assessed for finite sample sizes by Monte Carlo simulations in $d = 1$, 2 and 4 dimensions. Also the critical values of all considered tests were calculated by Monte Carlo simulation. We chose a 5% significance level.

Table I Four dimensional distributions used to generate the samples.

case	P^X	P^Y
1	$N(\mathbf{0},\mathbf{I})$	$C(\mathbf{0},\mathbf{I})$
2	$N(\mathbf{0},\mathbf{I})$	$N_{\log}(\mathbf{0},\mathbf{I})$
3	$N(\mathbf{0},\mathbf{I})$	$80\%N(\mathbf{0},\mathbf{I})+20\%N\left(\mathbf{0},0.2^2\mathbf{I}\right)$
4	$N(\mathbf{0},\mathbf{I})$	$50\%N(\mathbf{0},\mathbf{I})+50\%N\left(\mathbf{0},\begin{pmatrix} 1 & 0.4 & 0.5 & 0.7 \\ 0.4 & 1 & 0.6 & 0.8 \\ 0.5 & 0.6 & 1 & 0.9 \\ 0.7 & 0.8 & 0.9 & 1 \end{pmatrix}\right)$
5	$N(\mathbf{0},\mathbf{I})$	Student's t_2
6	$N(\mathbf{0},\mathbf{I})$	Student's t_4
7	$U(\mathbf{0},\mathbf{1})$	$CJ(10)$
8	$U(\mathbf{0},\mathbf{1})$	$CJ(5)$
9	$U(\mathbf{0},\mathbf{1})$	$CJ(2)$
10	$U(\mathbf{0},\mathbf{1})$	$CJ(1)$
11	$U(\mathbf{0},\mathbf{1})$	$CJ(0.8)$
12	$U(\mathbf{0},\mathbf{1})$	$CJ(0.6)$
13	$U(\mathbf{0},\mathbf{1})$	$80\%U(\mathbf{0},\mathbf{1}) + 20\%N(\mathbf{0.5},0.05^2\mathbf{I})$
14	$U(\mathbf{0},\mathbf{1})$	$50\%U(\mathbf{0},\mathbf{1}) + 50\%N(\mathbf{0.5},0.2^2\mathbf{I})$

For the null hypothesis we determine the distribution of Φ_{nm} with the permutation technique, as mentioned above. We followed [12] and generated 1000 randomly selected two-sample subsets in each case and determined the critical values ϕ_c of ϕ_{nm}. For the specific case $n = m = 50$ and samples drawn from a uniform distribution we studied the statistical fluctuations. Transforming the confidence interval of ϕ_c into limits for nominal level $\alpha = 0.05$, we obtain the interval $[0.036, 0.063]$.

Even though the energy test has been designed for multivariate applications, we also investigated its power in one dimension because there a comparison with several well established tests is possible. To avoid a personal bias we drew the two samples from the same probability distributions which have been investigated by [13]. We compared the energy test to the Kolmogorov-Smirnov, Cramèr-von Mises, Wilcox, Lepage and χ^2 test. Details of the comparison are given in Ref. [11]. The results indicate that the power of the energy test in most of the cases is larger than that of the well known χ^2 and Kolmogorov-Smirnov test and comparable to that of the Cramèr-von Mises test.

In the multivariate case we compared the energy test with the Friedman-Rafsky and the nearest neighbor tests.

In order to investigate how the performance of the tests using Φ_{nm}, R_{nm} and N_{nm} changes with the dimension, we have investigated the power in dimensions $d = 2$ and 4. Since the results in both cases are similar, we present in this short writeup only those

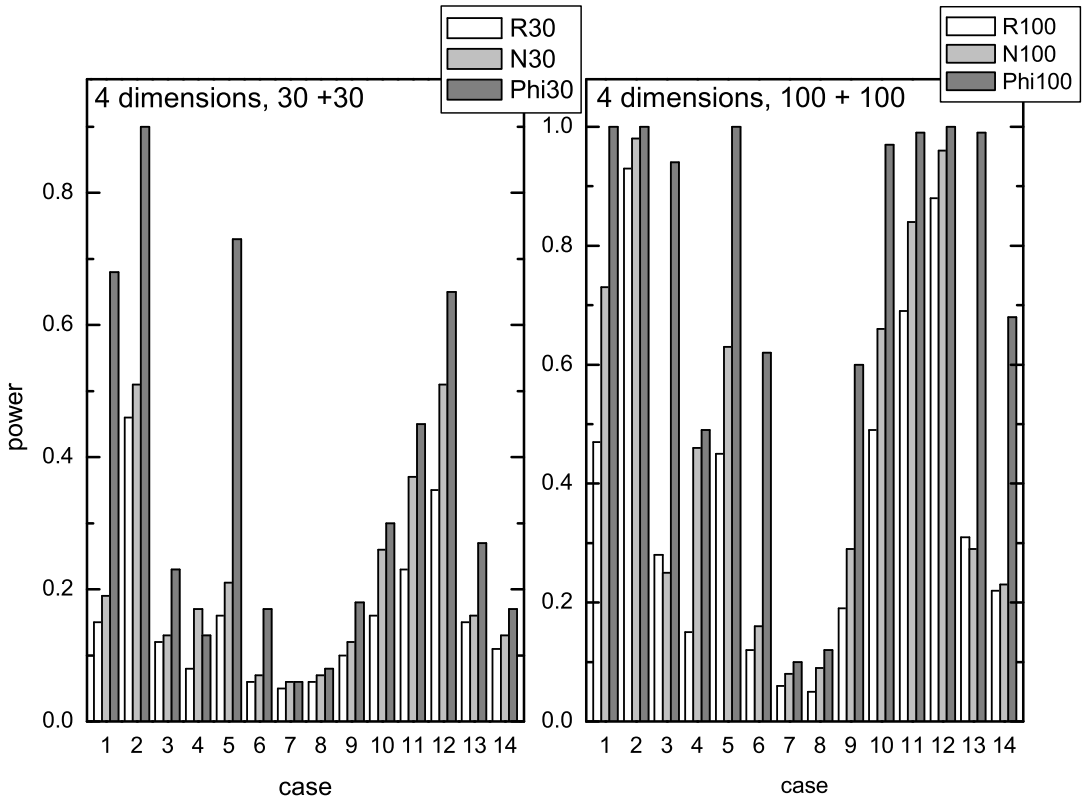

Figure 1: Rejection power of three two-sample tests for different alternatives. The sample sizes are 30 + 30 (left hand) and 100 +100 (right hand). R, N, Phi denote the Friedman-Rafsky test, the nearest neighbor test, and the energy test.

of the four dimensional case. In Table I we summarize the alternative probability distributions P^X and P^Y from which we drew the two samples. The first sample was drawn either from $N(\mathbf{0}, \mathbf{I})$ or from $U(\mathbf{0}, \mathbf{1})$ where $N(\mu, \mathbf{V})$ is a multivariate normal probability distribution with the indicated mean vector μ and covariance matrix $\mathbf{V}$ and $U(\mathbf{0}, \mathbf{1})$ is the multivariate uniform probability distribution in the unit cube. The parent distributions of the second sample were the Cauchy distribution C, the $N_{\log}$ distribution (explained below), correlated normal distributions, the Student's distributions, t_2 and t_4, and Cook-Johnson $CJ(a)$ distributions [14] with correlation parameter $a > 0$. $CJ(a)$ converges for $a \to \infty$ to the independent multivariate uniform distribution and $a \to 0$ corresponds to the totally correlated case $X_{i1} = X_{i2} = ... = X_{id}, i = 1, ..., n$. We generated the random vectors from $CJ(a)$ via the standard gamma distribution with shape parameter a, following the prescription proposed by [15]. The distribution denoted by $N_{\log}$ is obtained by the variable transformation $x \to x' = \ln|x|$ applied to each coordinate of a multidimensional normal distribution and is not to be confused with the log-normal distribution. It is extremely asymmetric. Some of the considered probability densities have also been used in a power study in [16].

The various combinations emphasize different types of deviations between the populations. These include location and scale shifts, differences in skewness and kurtosis as well as differences in the correlation of the variates.

The test statistics Φ_{nm}, R_{nm} and N_{nm} were evaluated.

The power was again computed for 5% significance level and samples of equal size $n = m = 30$, 50, and 100 (small, moderate and large) in two and four dimensions. In Figure 1 we show some of the results. More details can be found in [11].

The Friedman-Rafsky and the nearest neighbor tests show very similar rejection power. For all three sample sizes and dimensions the energy test performed better than the other two tests in almost all considered alternatives. This is astonishing because the logarithmic distance function is long range and the probability distributions in the cases 11 and 12 have a sharp

peak in one corner of a d dimensional unit cube and in case 13 a sharp peak in the middle of the unit cube. The multivariate student distribution represents very mild departures from normality, but nevertheless the rejection rate of the energy test is high.

Acknowledgments

We would like to thank the organizers of the conference for this beautiful conference. We acknowledge the financial support by the German Bundesministerium für Bildung und Forschung, Frderkennzeichen 05 HB1PSA/4.

References

[1] Duran, B. S., "A survey of nonparametric tests for scale", Communications in Statistics - Theory and Methods 5 (1976) 1287.

[2] Conover, W. J., Johnson, M. E., and Johnson, M. M., "A comparative study of tests of homogeneity of variances, with applications to the outer continental shelf bidding data", Technometrics 23 (1981) 351.

[3] Buening, H., "Robuste und adaptive Tests", Berlin: De Gruyter (1991).

[4] Read, T. R. C., and Cressie, N. A. C., "Goodness-of-fit statistics for discrete multivariate data", New York: Springer-Verlag (1988).

[5] Scott, D. W., "Multivariate Density Estimation: Theory, Practice and Visualisation", Wiley, New York (1992).

[6] Friedman, J. H., and Rafsky, L. C., "Multivariate generalizations of the Wolfowitz and Smirnov two-sample tests", Annals of Statistics 7 (1979) 697.

[7] Henze, N., "A multivariate two-sample test based on the number of nearest neighbor type coincidences", Annals of Statistics 16 (1988) 772.

[8] Wald, A., and Wolfowitz, J., "On a test whether two samples are from the same population", Ann. Math. Statist. 11 (1940) 147.

[9] Henze, N., and Penrose, M. D., "On the multivariate runs test", Annals of Statistics 27 (1998) 290.

[10] Bowman, A., and Foster, P., "Adaptive smoothing and density-based tests of multivariate normality", J. Amer. Statist. Assoc. 88 (1993) 529.

[11] Aslan, B. and Zech, G., "A new test for the multivariate two-sample problem based on the concept of minimum energy", available on http://arxiv.org/abs/math.PR/0309164 (2003).

[12] Efron, B., and Tibshirani, R., "An Introduction to the Bootstrap", New York: Chapman and Hall (1993).

[13] Buening, H., "Kolmogorov-Smirnov- and Cramèr-von Mises type two-sample tests with various weight functions", Communications in Statistics-Simulation and Computation 30 (2001) 847.

[14] Devroye, L., "Non-Uniform Random Variate Generation", New York: Springer-Verlag (1986).

[15] Ahrens, J. H., and Dieter, U., "Pseudo-Random Numbers", New York: Wiley (1977).

[16] Bahr, R., "A new test for the multi-dimensional two-sample problem under general alternative", Ph.D. Thesis, University of Hannover (1996) (in German).