A NEW TEST FOR THE MULTIVARIATE TWO-SAMPLE PROBLEM BASED ON THE CONCEPT OF MINIMUM ENERGY

G. Zech* and B. Aslan†
University of Siegen, Germany

August 28, 2018

Abstract

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 was shown to be 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

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

against the general alternative

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

where the distribution functions F and G are unknown.

Testing whether two samples are consistent with a single unknown distribution is a task that occurs in many areas of research. A natural and simple approach is to compare the first two moments of the sample which measure location and scale. Many tests of this type can be found in the literature [Duran (1976), Conover *et al.* (1981), Buening (1991)] but distributions may differ in a more subtle way. Other tests require binning of data like the power-divergence statistic test [Read and Cressie (1988)] 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* [Scott (1992)], hence these tests 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 extension of the Wald-Wolfowitz run test [Wald and Wolfowitz (1940)]

*Corresponding author. E-mail: zech@physik.uni-siegen.de
†E-mail: aslan@physik.uni-siegen.de

and the nearest neighbor test [Henze (1988)] avoid these caveats but it is not obvious that they are sensitive to all kind of differences in the parent distributions from which the samples are drawn.

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 [Bowman and Foster (1993)] but whereas this test is based on probability density estimation and local comparison, the energy test explores long range correlations.

In Section 2 we define the test statistic Φ_{nm}. Even though the energy test has been designed for multivariate applications, we apply it in Section 3 to univariate samples because there a unbiased comparison to several well established univariate tests is easily possible. A selection of examples and tests investigated by [Buening (1999)] are considered. These are the Kolmogorov-Smirnov, Cramèr-von Mises, Wilcox [Wilcox (1997)] and Lepage [Lepage (1971)] tests. We have added the χ^2 test with equal probability bins.

In Section 4 we study the power of the energy test in two and four dimensions and compare it to the Friedman-Rafsky [Friedman and Rafsky (1979)] and the nearest neighbor tests.

We conclude in Section 5 with a short summary.

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 the Appendix we show that also in this case, large values of energy indicate significant deviations between the parent populations of the two samples.

2.1 The test statistic

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

$$\begin{aligned}
\Phi_{nm} &= \Phi_A + \Phi_B + \Phi_{AB} \\
\Phi_A &= \frac{1}{n^2} \sum_{i<j}^{n} R\left(|\mathbf{x}_i - \mathbf{x}_j|\right), \\
\Phi_B &= \frac{1}{m^2} \sum_{i<j}^{m} R\left(|\mathbf{y}_i - \mathbf{y}_j|\right), \\
\Phi_{AB} &= -\frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} R\left(|\mathbf{x}_i - \mathbf{y}_j|\right),
\end{aligned}$$

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. For the present analysis, we select $R(r) = -\ln r$ instead of the electrostatic potential $1/r$.

With this choice 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 [Efron and Tibshirani (1993)] 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.

2.2 Normalization of the distance

The Euclidean distances between two observations $\mathbf{z}_i$ and $\mathbf{z}'_j$ in $\mathbb{R}^d$ is

$$\left|\mathbf{z}_i - \mathbf{z}'_j\right| = \sqrt{\sum_{k=1}^{d}(z_{ik} - z'_{jk})^2}$$

with projections z_{ik} and z'_{jk}, $k = 1, \ldots, d$ of the vectors $\mathbf{z}_i$ and $\mathbf{z}'_j$.

Since the relative scale of the different variates usually is arbitrary we propose to normalize the projections by the following transformation

$$\overset{*}{z}_{ik} = \frac{z_{ik} - \mu_k}{\sigma_k} \qquad \begin{array}{l} i = 1, \ldots, n \\ k = 1, \ldots, d \end{array}$$

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 that a single projection dominates the value of the energy and that other projections contribute only marginally to it.

We did not apply this transformation in the present study, because this might have biased the comparison with other tests and because the different variates had similar variances.

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.

For the null hypothesis we determine the distribution of Φ_{nm} with the permutation technique, as mentioned above. We followed [Efron and Tibshirani (1993)] and generated 1000 randomly selected two 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 α, we obtain the interval $[0.036, 0.063]$, see Table 1.

3.1 One dimensional case

Even though the energy test has been designed for multivariate applications, we investigate its power in one dimension because there a comparison with several well established

Table 1: Confidence intervals as a function of the number of permutations for nominal $\alpha = 0.05$.

# of permutations	$CL(95\%)$ for α
100	$[0.006, 0.095]$
300	$[0.025, 0.075]$
500	$[0.031, 0.068]$
1000	$[0.036, 0.063]$

tests is possible . To avoid a personal bias we drew the two samples from the probability distributions, which have also been investigated by [Buening (2001)]:

$$
\begin{aligned}
f_1(x) &= \begin{cases} 1 & -\sqrt{3} \le x \le \sqrt{3} \\ 0 & \text{otherwise} \end{cases} \\
f_2(x) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\
f_3(x) &= \frac{1}{2} e^{-|x|} \\
f_4(x) &= \frac{1}{\pi} \frac{1}{1+x^2} \quad , \quad \text{Cauchy} \\
f_5(x) &= e^{-(x+1)} \quad , \quad x \ge -1 \\
f_6(x) &= \chi_3^2 \qquad \begin{array}{c} \chi^2 \text{ with 3 degrees of freedom,} \\ \text{transformed to mean 0, variance 1} \end{array} \\
f_7(x) &= \frac{1}{2} N(1.5, 1) + \frac{1}{2} N(-1.5, 1) \\
f_8(x) &= 0.8 N(0, 1) + 0.2 N(0, 4^2) \\
f_9(x) &= \frac{1}{2} N(1, 2^2) + \frac{1}{2} N(-1, 1)
\end{aligned}
$$

This set f_1 to f_9 of probability distributions covers a variety of cases of short tailed up to very long tailed probability distributions as well as skewed ones.

To evaluate the power of the tests we generated 1000 pairs of samples for small $n = m = 25$, moderate $n = 50$, $m = 40$ and "large" $n = 100$, $m = 50$, for seven different scenarios. We have transformed the variates $Y_i^* = \theta + \tau Y_j$, $j = 1, \ldots, m$ of the second sample, corresponding to the alternative distribution, with different location parameters θ and scale parameters τ. Powers were simulated in all cases by counting the number of times a test resulted in a rejection divided by 1000. All tests have a nominal significance level of 0.05.

Table 2 shows the estimated power for small sample sizes, $n = 25$, $m = 25$, of the selected tests. These are the Kolmogorov-Smirnov (KS), Cramèr-von Mises (CvM), Wilcox (W), Lepage (L). We have added the χ^2 test with 5 equal probability bins. Tables 3 and 4 present the results for $n = 50$, $m = 40$ and $n = 100$, $m = 50$, respectively. For the large sample the number of χ^2 bins was increased to 10.

It is apparent that none of the considered tests performs better than all other tests for all alternatives. 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 KS tests and comparable to that of the CvM test. For long tailed distributions, e.g. for combinations (f_8, f_4), the energy test is the most powerful test. This is not unexpected since $R(x) = -\ln(x)$ is long range. Lepage and Wilcox tests are powerful tests for all combinations and sample sizes considered, however,

Table 2: Power of the selected tests for n=m=25, $\alpha = 0.05$, $x \to \theta + \tau x$

P_1	P_2	θ, τ	KS	CvM	W	L	$\Phi_{25,25}$	χ^2
$f_1(x)$	$f_7(x)$	$0.4; 1.4$	0.12	0.18	0.48	0.38	0.24	0.11
		$0.6, 1.6$	0.37	0.41	0.87	0.69	0.54	0.17
		$0.6; 0.8$	0.40	0.55	0.66	0.50	0.45	0.52
		$0.5; 0.5$	0.70	0.70	0.93	0.86	0.85	0.85
$f_7(x)$	$f_2(x)$	$0.4; 1.4$	0.08	0.13	0.22	0.14	0.13	0.08
		$0.6, 1.6$	0.20	0.29	0.46	0.34	0.31	0.14
		$0.6; 0.8$	0.34	0.46	0.57	0.51	0.44	0.45
		$0.5; 0.5$	0.72	0.69	0.93	0.93	0.89	0.88
$f_2(x)$	$f_3(x)$	$0.4; 1.4$	0.17	0.23	0.22	0.19	0.19	0.14
		$0.6, 1.6$	0.33	0.44	0.42	0.37	0.38	0.24
		$0.6; 0.8$	0.64	0.70	0.67	0.66	0.67	0.60
		$0.5; 0.5$	0.74	0.77	0.84	0.91	0.89	0.84
$f_2(x)$	$f_9(x)$	$0.4; 1.4$	0.06	0.09	0.19	0.13	0.12	0.07
		$0.6, 1.6$	0.14	0.22	0.35	0.29	0.21	0.11
		$0.6; 0.8$	0.46	0.59	0.65	0.59	0.45	0.54
		$0.5; 0.5$	0.71	0.72	0.89	0.88	0.82	0.85
$f_6(x)$	$f_5(x)$	$0.4; 1.4$	0.10	0.16	0.15	0.12	0.12	0.12
		$0.6, 1.6$	0.16	0.25	0.24	0.22	0.16	0.20
		$0.6; 0.8$	0.95	0.94	1.00	0.97	0.98	0.97
		$0.5; 0.5$	0.99	0.98	1.00	1.00	1.00	1.00
$f_3(x)$	$f_8(x)$	$0.4; 1.4$	0.26	0.34	0.30	0.25	0.28	0.21
		$0.6, 1.6$	0.55	0.64	0.56	0.53	0.50	0.43
		$0.6; 0.8$	0.85	0.89	0.86	0.84	0.85	0.79
		$0.5; 0.5$	0.90	0.93	0.94	0.97	0.95	0.91
$f_8(x)$	$f_4(x)$	$0.4; 1.4$	0.34	0.42	0.45	0.52	0.54	0.32
		$0.6, 1.6$	0.60	0.67	0.68	0.77	0.80	0.51
		$0.6; 0.8$	0.80	0.85	0.78	0.72	0.82	0.70
		$0.5; 0.5$	0.81	0.84	0.81	0.71	0.84	0.72

the Lepage test is based on the first two moments of the null distribution and therefore specifically adapted to the type of study presented here.

3.2 Multi-dimensional case

For the general multivariate two-sample problem, only a few binning-free tests have been proposed. The Friedman-Rafsky test and the nearest neighbor test like the energy test are based on the distance between the observations. The Bowman-Foster goodness-of-fit test uses the probability density estimation (PDE) to deduce a p.d.f. from a sample. If this technique is applied to both samples, it obviously can be used as a two sample test. With a Gaussian Kernel it is almost identical to the energy test with a Gaussian distance function. We prefer the logarithmic function.

3.2.1 Friedman-Rafsky test

The Friedman-Rafsky test can be seen as a generalization of the univariate run test. 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

Table 3: Power of the selected tests for n=50, m=40, $\alpha = 0.05$, $x \to \theta + \tau x$

P_1	P_2	$\theta;\tau$	KS	CvM	W	L	$\Phi_{50,40}$	χ^2
$f_1(x)$	$f_7(x)$	$0.3;1.3$	0.22	0.18	0.67	0.41	0.25	0.14
		$0.4;0.8$	0.49	0.47	0.67	0.53	0.46	0.62
$f_7(x)$	$f_2(x)$	$0.3;1.3$	0.15	0.17	0.34	0.16	0.20	0.10
		$0.4;0.8$	0.62	0.56	0.66	0.70	0.58	0.58
$f_2(x)$	$f_3(x)$	$0.3;1.3$	0.28	0.29	0.25	0.22	0.26	0.14
		$0.4;0.8$	0.67	0.66	0.65	0.70	0.68	0.61
$f_2(x)$	$f_9(x)$	$0.3;1.3$	0.07	0.07	0.27	0.12	0.14	0.07
		$0.4;0.8$	0.51	0.50	0.66	0.58	0.46	0.61
$f_6(x)$	$f_5(x)$	$0.3;1.3$	0.13	0.14	0.18	0.12	0.11	0.14
		$0.4;0.8$	1.00	0.97	1.00	0.99	1.00	1.00
$f_3(x)$	$f_8(x)$	$0.3;1.3$	0.38	0.39	0.32	0.29	0.31	0.25
		$0.4;0.8$	0.84	0.84	0.80	0.78	0.86	0.74
$f_8(x)$	$f_4(x)$	$0.3;1.3$	0.50	0.52	0.59	0.66	0.64	0.39
		$0.4;0.8$	0.76	0.77	0.70	0.63	0.79	0.59

Table 4: Power of the selected tests for n=100, m=50, $\alpha = 0.05$, $x \to \theta + \tau x$

P_1	P_2	$\theta;\tau$	KS	CvM	W	L	$\Phi_{100,50}$	χ^2
$f_1(x)$	$f_7(x)$	$0.3;1.3$	0.32	0.33	0.97	0.62	0.44	0.28
		$0.4;0.8$	0.68	0.73	0.85	0.76	0.76	0.66
$f_7(x)$	$f_2(x)$	$0.3;1.3$	0.21	0.27	0.67	0.28	0.33	0.26
		$0.4;0.8$	0.82	0.79	0.82	0.90	0.82	0.64
$f_2(x)$	$f_3(x)$	$0.3;1.3$	0.31	0.37	0.41	0.29	0.34	0.17
		$0.4;0.8$	0.85	0.86	0.82	0.90	0.89	0.72
$f_2(x)$	$f_9(x)$	$0.3;1.3$	0.08	0.10	0.51	0.18	0.21	0.17
		$0.4;0.8$	0.65	0.66	0.79	0.77	0.67	0.63
$f_6(x)$	$f_5(x)$	$0.3;1.3$	0.13	0.19	0.25	0.18	0.18	0.22
		$0.4;0.8$	1.00	1.00	1.00	1.00	1.00	1.00
$f_3(x)$	$f_8(x)$	$0.3;1.3$	0.46	0.52	0.44	0.44	0.47	0.25
		$0.4;0.8$	0.95	0.97	0.92	0.94	1.00	0.86
$f_8(x)$	$f_4(x)$	$0.3;1.3$	0.61	0.68	0.81	0.84	0.86	0.63
		$0.4;0.8$	0.90	0.93	0.87	0.86	0.92	0.74

Table 5: Two dimensional distributions used to generate the samples.

case	P^X	P^Y
1	$N(\mathbf{0},\mathbf{I})$	$C(\mathbf{0},\mathbf{I})$
2	$N(\mathbf{0},\mathbf{I})$	$N_{\log}(0,\mathbf{I})$
3	$N(\mathbf{0},\mathbf{I})$	$N\left(\mathbf{0},\begin{matrix}1 & 0.6\\ 0.6 & 1\end{matrix}\right)$
4	$N(\mathbf{0},\mathbf{I})$	$N\left(\mathbf{0},\begin{matrix}1 & 0.9\\ 0.9 & 1\end{matrix}\right)$
5	$N(\mathbf{0},\mathbf{I})$	Student's t_2
6	$N(\mathbf{0},\mathbf{I})$	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	$U(\mathbf{0},\mathbf{1})$	$80\%U(\mathbf{0},\mathbf{1})+20\%N\left(\mathbf{0.5},0.05^2\mathbf{I}\right)$
14	$U(\mathbf{0},\mathbf{1})$	$50\%U(\mathbf{0},\mathbf{1})+50\%N\left(\mathbf{0.5},0.2^2\mathbf{I}\right)$

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 is formed. The test statistic R_{nm} equals the number of connections between observations from different samples.

Obviously, in one dimension the test reduces to the run test. Small values of R_{nm} lead to a rejection of H_0. The statistic R_{nm} is asymptotically distribution-free under the null hypothesis [Henze and Penrose (1998)].

3.2.2 The nearest neighbor test

The nearest neighbor test statistic N_{nm} is the sum of the number of vectors $\mathbf{Z}_i$ of the pooled sample $(\mathbf{Z}_1,\ldots,\mathbf{Z}_{n+m})$ where the nearest neighbor of $\mathbf{Z}_i$, denoted by $N(\mathbf{Z}_i)$, is of the same type as $\mathbf{Z}_i$:

$$N_{nm}=\sum_{i=1}^{n+m} I\left(\mathbf{Z}_i \text{ and } N(\mathbf{Z}_i) \text{ belong to the same sample}\right)$$

Here I is the indicator function. $N(\mathbf{Z}_i)$ can be determined by a fixed but otherwise arbitrary norm on $\mathbb{R}^d$. We selected the Euclidean norm. In [Henze (1988)] it is shown that the limiting distribution of N_{nm} is normal in the limit $\min(n,m)\to\infty$ and $n/(n+m)\to\tau$ with $0<\tau<1$. Large values of N_{nm} lead to a rejection of H_0.

3.2.3 Comparison of the 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 considered problems in dimensions $d=2$ and 4. In Table 5 and Table 6 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

Table 6: 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	$N(\mathbf{0},\mathbf{I})$	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	$U(\mathbf{0},\mathbf{1})$	$50\%U(\mathbf{0},\mathbf{1})+50\%N(\mathbf{0.5},0.2^2\mathbf{I})$

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 [Devroye (1986] 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 [Ahrens and Dieter (1977)]. The distribution denoted by $N_{\log}$ is obtained by the variable transformation $x \to \ln |x|$ applied to each coordinate of a multidimensional normal distribution and is not to be mixed up with the log-normal distribution. It is extremely asymmetric. Some of the considered probability densities have also been used in a power study in [Bahr (1996)].

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. Table 7 and Table 8 illustrate the power of the three considered tests calculated from 1000 replications.

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 this unit cube. The multivariate student distribution represents very mild departures from normality, but nevertheless the rejection rate of the energy test is high.

Table 7: Power at significance level α =0.05, calculated from 1000 repetitions, $n = m = 30$, $n = m = 50$ and $n = m = 100$, $d = 2$

case	$R_{30,30}$	$N_{30,30}$	$\Phi_{30,30}$	$R_{50,50}$	$N_{50,50}$	$\Phi_{50,50}$
1	0.25	0.23	0.57	0.44	0.41	0.86
2	0.33	0.30	0.58	0.53	0.50	0.89
3	0.14	0.12	0.13	0.17	0.20	0.21
4	0.63	0.57	0.53	0.87	0.83	0.87
5	0.14	0.14	0.32	0.18	0.20	0.49
6	0.07	0.07	0.11	0.08	0.08	0.13
7	0.04	0.07	0.05	0.04	0.07	0.05
8	0.05	0.06	0.08	0.05	0.08	0.06
9	0.08	0.08	0.10	0.08	0.10	0.14
10	0.13	0.12	0.15	0.18	0.18	0.23
11	0.15	0.14	0.18	0.23	0.22	0.30
12	0.20	0.20	0.25	0.33	0.31	0.45
13	0.11	0.10	0.14	0.16	0.15	0.33
14	0.09	0.09	0.14	0.12	0.11	0.22

case	$R_{100,100}$	$N_{100,100}$	$\Phi_{100,100}$
1	0.70	0.60	1.00
2	0.82	0.74	1.00
3	0.31	0.28	0.47
4	0.99	0.97	1.00
5	0.34	0.29	0.86
6	0.10	0.11	0.24
7	0.04	0.05	0.10
8	0.05	0.05	0.09
9	0.10	0.11	0.24
10	0.25	0.23	0.52
11	0.32	0.29	0.66
12	0.56	0.48	0.90
13	0.23	0.19	0.78
14	0.16	0.16	0.56

Table 8: Power at significance level α =0.05, calculated from 1000 repetitions, $n = m = 30$, $n = m = 50$ and $n = m = 100$, $d = 4$

case	$R_{30,30}$	$N_{30,30}$	$\Phi_{30,30}$	$R_{50,50}$	$N_{50,50}$	$\Phi_{50,50}$
1	0.15	0.19	0.68	0.22	0.39	0.93
2	0.46	0.51	0.90	0.68	0.78	1.00
3	0.12	0.13	0.23	0.14	0.17	0.47
4	0.08	0.17	0.13	0.09	0.26	0.18
5	0.16	0.21	0.73	0.22	0.35	0.95
6	0.06	0.07	0.17	0.08	0.10	0.31
7	0.05	0.06	0.06	0.06	0.06	0.07
8	0.06	0.07	0.08	0.06	0.09	0.07
9	0.10	0.12	0.18	0.13	0.20	0.30
10	0.16	0.26	0.30	0.27	0.42	0.62
11	0.23	0.37	0.45	0.39	0.58	0.77
12	0.35	0.51	0.65	0.58	0.76	0.93
13	0.15	0.16	0.27	0.20	0.20	0.62
14	0.11	0.13	0.17	0.14	0.18	0.31

case	$R_{100,100}$	$N_{100,100}$	$\Phi_{100,100}$
1	0.47	0.73	1.00
2	0.93	0.98	1.00
3	0.28	0.25	0.94
4	0.15	0.46	0.49
5	0.45	0.63	1.00
6	0.12	0.16	0.62
7	0.06	0.08	0.10
8	0.05	0.09	0.12
9	0.19	0.29	0.60
10	0.49	0.66	0.97
11	0.69	0.84	0.99
12	0.88	0.96	1.00
13	0.31	0.29	0.99
14	0.22	0.23	0.68

4 SUMMARY

We have introduced the statistic *energy* which is a simple function of the distances of observations in the sample space. It can be used as a powerful measure of the compatibility of two samples. It is easy to compute, efficient and applicable in arbitrary dimensions of the sample space. The comparison to the Friedman-Rafsky and the nearest neighbor tests demonstrates its high performance in the multi-dimensional case.

5 APPENDIX

We define the energy of the difference of two probability density functions by

$$\phi = \frac{1}{2} \int \int \left[f(\mathbf{x}) - f_0(\mathbf{x}) \right] \left[f(\mathbf{x}') - f_0(\mathbf{x}') \right] R(\mathbf{x}, \mathbf{x}') d\mathbf{x} d\mathbf{x}'.$$

Here and in what follows, an unspecified integral denotes integration over $\mathbb{R}^d$. Substituting $g(\mathbf{x}) = f(\mathbf{x}) - f_0(\mathbf{x})$ we obtain

$$\phi = \frac{1}{2} \int \int g(\mathbf{x}) g(\mathbf{x}') R(\mathbf{x}, \mathbf{x}') d\mathbf{x} d\mathbf{x}'. \tag{1}$$

We replace in the Eq.(1) the distance function $R(\mathbf{x}, \mathbf{x}') = R(|\mathbf{x} - \mathbf{x}'|)$ by its Fourier integral

$$R(|\mathbf{x} - \mathbf{x}'|) = \int F(\mathbf{k}) e^{i\mathbf{k} \cdot (\mathbf{x} - \mathbf{x}')} d\mathbf{k}$$

and obtain

$$\begin{aligned} \phi &= \frac{1}{2} \int \int \int g(\mathbf{x}) g(\mathbf{x}') F(\mathbf{k}) e^{i\mathbf{k} \cdot (\mathbf{x} - \mathbf{x}')} d\mathbf{x} d\mathbf{x}' d\mathbf{k} \\ &= \frac{1}{2} \int \left| G(\mathbf{k}) \right|^2 F(\mathbf{k}) d\mathbf{k}. \end{aligned} \tag{2}$$

where $G(\mathbf{k})$ is the Fourier transform of $g(\mathbf{x})$.

For the function $R(|\mathbf{r}|) = R(r) = \frac{1}{r^\kappa}$ with $d > \kappa$, where d is the dimension of $\mathbf{r}$, the Fourier transformation $F(\mathbf{k})$ is [Gel'fand and Shelov (1964)]:

$$F(k) = 2^{d-\kappa} \pi^{d/2} \frac{\Gamma\left(\frac{d-\kappa}{2}\right)}{\Gamma\left(\frac{\kappa}{2}\right)} k^{\kappa - d} > 0.$$

with $k = |\mathbf{k}|$.

From Eq.(2) follows that ϕ is positive. The minimum $\phi_{\min} = 0$ is obtained for $g_{\min}(\mathbf{x}) \equiv \mathbf{0}$ or $f(\mathbf{x}) \equiv f_0(\mathbf{x})$. The result $g_{\min}(\mathbf{x}) \equiv \mathbf{0}$ holds also for the logarithmic distance function $R(r) = -\ln r$ which can be considered as the $\kappa = 0$ limit of the power law distance function:

$$-\ln r = \lim_{n \to \infty} n \left(\left(\frac{1}{r} \right)^{1/n} - 1 \right),$$

The additional constant in the distance function does not contribute to the integral (1).

Expanding (1) we get a sum of three expectation values of $R(\mathbf{x}, \mathbf{x}')$

$$\phi = \frac{1}{2} \int \int \left[f(\mathbf{x}) f(\mathbf{x}') - 2 f_0(\mathbf{x}) f(\mathbf{x}') + f_0(\mathbf{x}) f_0(\mathbf{x}') \right] R(\mathbf{x}, \mathbf{x}') d\mathbf{x} d\mathbf{x}'$$

which can be estimated from two samples drawn from f and f_0, respectively, as defined in Section 2:

$$\phi = \lim_{n,m\to\infty} \left[\frac{1}{n(n-1)} \sum_{i<j}^{n} R\left(|\mathbf{x}_i - \mathbf{x}_j|\right) + \right.$$
$$- \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} R\left(|\mathbf{x}_i - \mathbf{y}_j|\right) +$$
$$\left. + \frac{1}{m(m-1)} \sum_{i<j}^{m} R\left(|\mathbf{y}_i - \mathbf{y}_j|\right) \right].$$

The quantity in the brackets is up to a minor difference in the denominators equal to our test quantity ϕ_{nm}. In the limit $n, m \to \infty$ the statistic ϕ_{nm} is minimum if the two samples are from the same distribution.

References

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

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

[Bowman and Foster (1993)] Bowman, A. and Foster, P. (1993). Adaptive smoothing and density-based tests of multivariate normality. *J. Amer. Statist. Assoc.*, **88**, 529-537.

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

[Buening (2001)] Buening, H. (2001). Kolmogorov-Smirnov- and Cramèr-von Mises type two-sample tests with various weight functions. *Communications in Statistics-Simulation and Computation*, **30**, 847-865.

[Buening (1999)] Buening, H. (1999). Power comparison of several two-sample tests for general alternatives. *Allgemeines Statistisches Archiv*, **83**, 190-210.

[Conover *et al.* (1981)] Conover, W. J., Johnson, M. E. and Johnson, M. M. (1981). A comparative study of tests of homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, **23**, 351-361.

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

[Duran (1976)] Duran, B. S. (1976). A survey of nonparametric tests for scale. *Communications in statistics- Theory and Methods*, **5**, 1287-1312.

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

[Friedman and Rafsky (1979)] Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wolfowitz and Smirnov two-sample tests. *Annals of Statistics,* **7**, 697-717.

[Gel'fand and Shelov (1964)] I. M. Gel'fand and G. E. Shelov (1964). *Generalized Functions, Vol.1: Properties and Operations.* Academic Press, New York.

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

[Henze and Penrose (1998)] Henze, N. and Penrose, M. D. (1998). On the multivariate runs test. *Annals of Statistics,* **27**, 290-298.

[Lepage (1971)] Lepage, Y. A. (1971). A combination of Wilcoxon's and Ansari-Bradley's statistics. *Biometrica*, **58**, 213-217.

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

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

[Wald and Wolfowitz (1940)] Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Ann. Math. Statist.,* **11***,* 147-162.

[Wilcox (1997)] Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing.* Academic Press, San Diego.