Non-parametric comparison of histogrammed two-dimensional data distributions using the Energy Test

2012 J. Phys.: Conf. Ser. 368 012046

(http://iopscience.iop.org/1742-6596/368/1/012046)

Non-parametric comparison of histogrammed two-dimensional data distributions using the Energy Test

Ivan D Reid, Raul H C Lopes and Peter R Hobson
School of Engineering and Design, Brunel University, Uxbridge UB8 3PH, UK
E-mail: ivan.reid@brunel.ac.uk

Abstract. When monitoring complex experiments, comparison is often made between regularly acquired histograms of data and reference histograms which represent the ideal state of the equipment. With the larger HEP experiments now ramping up, there is a need for automation of this task since the volume of comparisons could overwhelm human operators. However, the two-dimensional histogram comparison tools available in ROOT have been noted in the past to exhibit shortcomings. We discuss a newer comparison test for two-dimensional histograms, based on the Energy Test of Aslan and Zech, which provides more conclusive discrimination between histograms of data coming from different distributions than methods provided in a recent ROOT release.

1. Introduction

A traditional task when monitoring HEP experiments has been the comparison between regularly acquired histograms of data, representing some aspect of the machine performance, and reference histograms which reflect the baseline or ideal state of the equipment. Histograms are typically used rather than the raw data points because of the compactness they provide, both in data storage space and in visual presentation. When a discrepancy is seen, it is flagged and the problem passed to an appropriate expert to decide what action is to be taken. With the advent of larger experiments, such as at the LHC, there is the potential for automated comparisons to ease the load on control-room shift-workers faced with a wide array of such histograms.

These types of comparison are called goodness-of-fit (GoF) tests, and can be subdivided into two broad types: the one-sample GoF test considers whether a given data sample is consistent with being generated from some specified distribution, while the two-sample GoF test examines the hypothesis that two data samples are derived from the same parent distribution. In general, similar methods can be applied to both types of tests. However, the problems are ill-posed – only the null hypothesis (that the distributions are the same) is well defined, the alternative hypothesis (that the distributions do not match) is not fully specified. It is important, therefore, to determine the most appropriate GoF method for any given problem.

Methods for comparing one-dimensional data distributions are well known, one of the more widespread being the Kolmogorov-Smirnov (KS) test [1] which compares cumulative distribution functions (CDF) for the two sets of data and takes as a statistic the maximum difference between them. Although this test is intended to be applied to discrete data, it is feasible to apply it to histogrammed data as well, provided that the effects of the binning on the test are taken into account. Applying this test in more than one dimension is problematic since it relies on an ordering of the data to obtain the CDFs, but there are 2^d-1 distinct ways of defining an ordering in a d-dimensional space [2]. Multidimensional GoF tests are also ill-posed in that they lack metric invariance. That is, the choice of scale factor or, in the case of histogrammed data, the number of bins can greatly affect the comparison result.

1.1. Histogram comparisons in ROOT
The widely-used data-handling and analysis package ROOT [3] provides two methods for comparing histograms, `Chi2Test` (χ^2) and `KolmogorovTest` (a KS test). Details of these tests may be found elsewhere [3, 4, 5] but in brief the χ^2 test compares histograms on a bin-by-bin basis while the KS test compares neighbourhoods, using the CDFs as described above. Extension of the χ^2 test to two dimensions (2D) is relatively straightforward, but the 2D-KS test is complicated by the ordering problem. ROOT addresses this by computing two CDFs for each histogram, accumulating the binned data rasterwise, in column- and row-major patterns, so that the comparison yields two maximum differences and the Kolmogorov function is applied to their average to return the probability P of the null hypothesis (i.e., that the two histograms represent selections from the same distribution). However, as 2D histograms are more finely binned, the order in which the binned data are accumulated approaches the order of the discrete data in the most-slowly varying dimension [5]. Consequently the CDFs generated by the ROOT 2D-KS test approach those of the discrete data ordered in one dimension along each coordinate separately. In extreme cases this can lead to false positives as histograms with similar projections onto the axes are compared (figure 1).

1.2. An alternative 2D test
Another method for comparing distributions in more than one dimension is the Energy Test presented in recent years by Aslan and Zech [6, 7, 8]. While this is again originally designed for discrete data, the authors postulated that speed gains may be obtained by applying it to histogrammed or clustered data sets [6]. We have previously shown a version of the Energy Test for histogrammed data within the ROOT framework, and provided some evaluations of

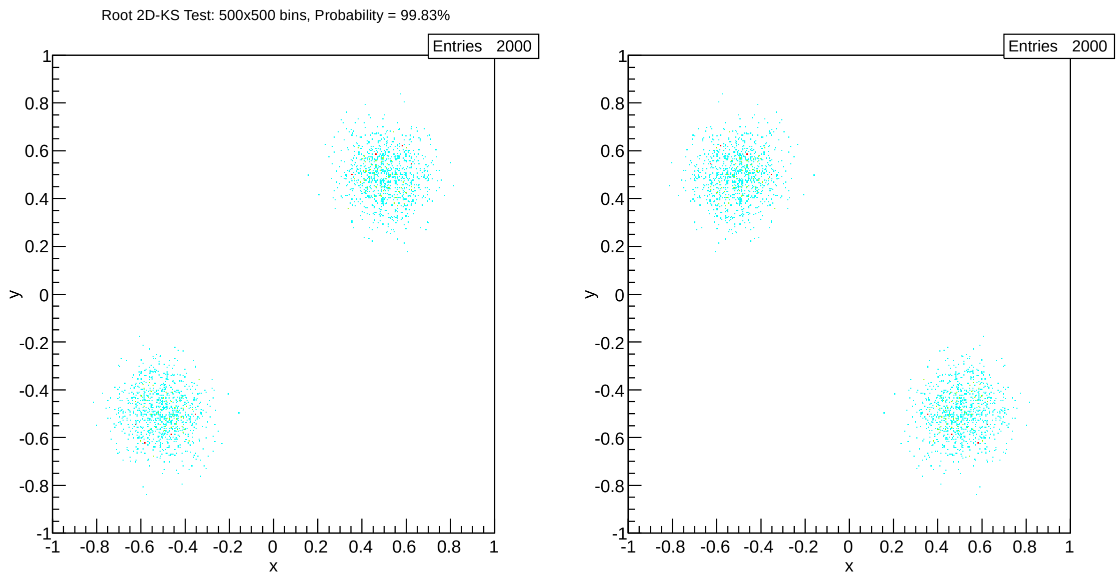


Figure 1. A ROOT 2D-KS comparison of two 2 000-point histograms binned at 500×500. The test returns a high probability (P=99.83%) that the both sets of data come from the same distribution. This is because they each have the same projections onto the axes.

its performance [5]. Some shortcomings in the ROOT implementation of the 2D-χ^2 Test were found, which have since been improved, and we revisit here some comparisons with the latest version of ROOT, in order to introduce the method to a wider audience.

2. The Energy Test

Consider two samples of data points in a d-dimensional domain, A of n points $\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\ldots,\mathbf{x}_n$ and B of m points $\mathbf{y}_1,\mathbf{y}_2,\mathbf{y}_3,\ldots,\mathbf{y}_m$, whose compatibility with the hypothesis that they arise from the same distribution is to be tested. If A is considered a system of positive charges, each $1/n$, and B a system of negative charges, each -$1/m$ (i.e., each system is normalised to unity charge), then from electrostatics in the limit of $n \to \infty$, $m \to \infty$ the total potential energy of the combined samples, computed for a $1/r$ potential, will be a minimum when both charge samples have the same distribution. This concept is generalised in the Energy Test.

2.1. The test statistic

The test statistic Φ_{nm} for the Energy Test consists of three terms, corresponding to the self-energies (from the repulsive forces between like charges) of the samples A and B (Φ_A and Φ_B, respectively) and the interaction energy (from the attraction between opposite charges) between the samples (Φ_{AB}) – a simple example is shown in figure 2:

$$\Phi_{nm} = \Phi_A + \Phi_B + \Phi_{AB} \tag{1}$$

$$\Phi_A = \frac{1}{n^2}\sum_{i=2}^{n}\sum_{j=1}^{i-1} R(|\mathbf{x}_i - \mathbf{x}_j|), \Phi_B = \frac{1}{m^2}\sum_{i=2}^{m}\sum_{j=1}^{i-1} R(|\mathbf{y}_i - \mathbf{y}_j|), \Phi_{AB} = -\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} R(|\mathbf{x}_i - \mathbf{y}_j|)$$

where R is a continuous, monotonically-decreasing function of the Euclidean distance r between the charges. Usually $R(r) = -\ln(r + \epsilon)$ is chosen rather than the electrostatic potential $1/r$ because it renders the test scale-invarant (although this is strictly true only if the same scale is applied in all dimensions) and offers a good rejection power against many alternatives to the null hypothesis. The value of the cutoff parameter ϵ, used to avoid singularities, is not critical so long as it is of the order of the mean distance between points at the densest region of the sample distributions.

It has been shown [7] that the test statistic is positive and has a minimum when the two samples are from the same distribution, in the limit of $n \to \infty$, $m \to \infty$, while another argument [6, 8] shows that when the samples have the same number of points, Φ_{nm} has a minimum when the points are pairwise coincident.

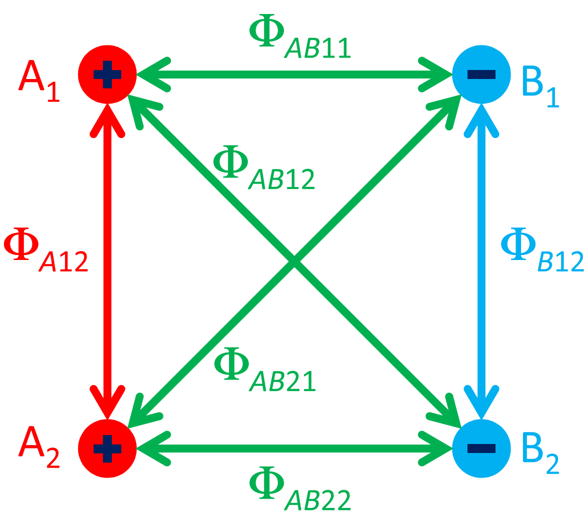


Figure 2. A simple example of the terms involved in the Energy Test. In this two-dimensional case the distributions A and B each have two points. The red arrow illustrates the (repulsive) self-energy between the positive particles in A, and the blue arrow shows that of the negative particles in B. The green arrows illustrate the (attractive) interaction energy between the oppositely-charged particles.

2.2. Implementing the Energy Test for 2D histograms

The test was implemented as a compiled ROOT macro, for equally-binned ($N \times N$) histograms in the first instance. Aslan and Zech [7] suggest that the ranges of the data can be normalised, to equalise the relative scales of the x- and y-coordinates. A similar normalisation is realised here by taking the histogram limits to be zero and unity (i.e., the distance between adjacent rows or columns is set to $1/N$). Underflow and overflow bins (with indices 0 and N+1, respectively, in ROOT notation) are included with nominal widths of $1/N$ below or above the histogram limits; arguments to the routine determine whether or not they are included in the comparison.

There is a slight complication from the fact that histograms do not preserve positional information about the points within a given bin so they must all be assigned a single position, for example the bin centre. This means that the case where $r = 0$ must be treated specially – i.e., when bin (i,j) is being compared to bin (i,j), either when computing Φ_{AB} (different histograms) or when calculating Φ_A and Φ_B (same histogram; unlike the discrete case, the self-energy between points in the same bin must be taken into account). We assume the original points are randomly distributed within the (square) bin limits and take the average distance between pairs of random points in a unit square to calculate an effective cutoff ϵ.

This value is $<r> = \frac{1}{15}(2 + \sqrt{2} + 5\sinh^{-1} 1) = 0.521\,405\,433\ldots$ [9] so we use $\epsilon = <r>/N$ as the distance. For an $N \times M$ histogram one would use instead the expression for the average distance between uniform random points in a $1/N \times 1/M$ rectangle [10],

$$\epsilon = \frac{1}{3}\rho + \frac{a^2}{6b}\ln\frac{b+\rho}{a} + \frac{b^2}{6a}\ln\frac{a+\rho}{b} + \frac{a^5+b^5-\rho^5}{15a^2b^2} \tag{2}$$

where $a = 1/N$, $b = 1/M$ and $\rho = \sqrt{a^2+b^2}$. Distances for other bin combinations are calculated simply as the Euclidean distance between bin centres, justified by the proximity of this value to the average distance between random points in the two bins from Monte Carlo simulations.

A minor modification to the calculation of the self-energy of the k points within a given bin is to weight by $k^2/2$ rather than the rigorous $k(k-1)/2$, as this ensures that comparisons between identical histograms return exactly zero analytically. An added benefit is that any scaling factors applied across individual histograms will be cancelled out rather than producing an offset that is dependent on the total histogram content [5]. If desired the rigorous behaviour can be obtained by the use of a parameter flag.

To summarise, the implementation of the three terms in the energy sum when comparing two $N \times N$ ROOT histograms A and B with total contents n and m, respectively, is given by

$$\begin{aligned}
\Phi_A &= \frac{1}{n^2}\sum_{i=0}^{N+1}\sum_{j=0}^{N+1} A(i,j)\left(\sum_{k=0}^{i-1}\sum_{l=0}^{N+1} A(k,l)R(i,j,k,l) + \sum_{l=0}^{j-1} A(i,l)D(j,l) + 0.5A(i,j)D_0\right) \quad (3)\\
\Phi_B &= \frac{1}{m^2}\sum_{i=0}^{N+1}\sum_{j=0}^{N+1} B(i,j)\left(\sum_{k=0}^{i-1}\sum_{l=0}^{N+1} B(k,l)R(i,j,k,l) + \sum_{l=0}^{j-1} B(i,l)D(j,l) + 0.5B(i,j)D_0\right)\\
\Phi_{AB} &= -\frac{1}{nm}\sum_{i=0}^{N+1}\sum_{j=0}^{N+1} A(i,j)\sum_{k=0}^{N+1}\sum_{l=0}^{N+1} B(k,l)R(i,j,k,l)
\end{aligned}$$

where $D_0 = -\ln(<r>/N)$, $R(i,j,k,l) = D_0$ when (i=k, j=l) or $-\frac{1}{2}\ln(((i-k)^2+(j-l)^2)/N^2)$ otherwise, $D(j,l) = R(i,j,i,l) = -\ln(|j-l|/N)$, and $A(i,j)$, $B(i,j)$ are the contents of individual bins within the histograms.

2.3. Speed

The calculation, and thus the time complexity, of the test statistic Φ_{nm} is, by inspection, $\mathbf{O}(n^2)$ so that in terms of histogram dimensions the cost is $\mathbf{O}(N^4)$. Table 1 shows the times taken for

Table 1. Times taken to compare 10^6-point histograms of various binnings with the ROOT 2D-KS and 2D-χ^2 tests and the Energy Test.

Size	ROOT 2D-KS	ROOT 2D-χ^2	Energy Test
25 × 25	<10 ms	<10 ms	<10 ms
50 × 50	<10 ms	<10 ms	10 ms
100 × 100	<10 ms	<10 ms	160 ms
250 × 250	<10 ms	10 ms	6.1 s
500 × 500	30 ms	30 ms	96.3 s

comparison of several pairs of histograms, filled with 10^6 points in a random uniform distribution and as a constant bin content, respectively. Despite all efforts to reduce calculations as much as possible, the increase in CPU time for the Energy Test beyond a binning of about 100×100 is evident. As with all the results reported in this work, the calculations were performed on a 2.67 GHz Xeon X5550 computer using ROOT Version 5.30/00. Outlier bins were always excluded from the comparisons.

3. Performance

3.1. Testing the power

The *power* of a comparison test is its ability to discriminate against non-conforming data, i.e., the fraction of non-compatible data which is rejected based on a selection criterion. In order to determine the power, the confidence level for accepting a test result must first be established. A common criterion is the 95th percentile CL_{95} – the value of a test beyond which only 5% of valid comparisons will lie.

Two references were developed for several tests; a constant distribution (i.e., no statistical fluctuation) of 10 points in each bin of a 100×100 histogram across the unit square, and a continually re-generated sample of 100 000 points randomly and uniformly distributed across the square. 50 000 tests were performed against these references using further samples of 100 000 random points. These are thus, respectively, references for one-sample GoF tests, to determine

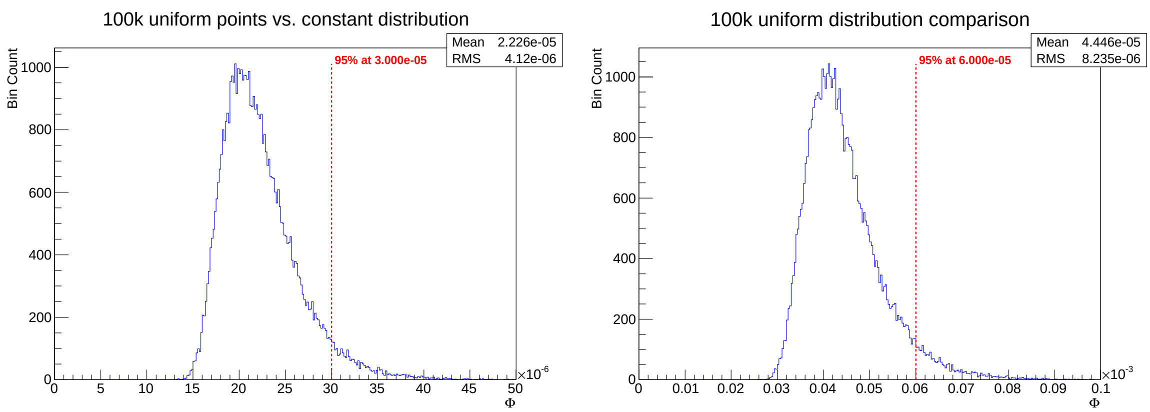


Figure 3. The distribution of results of the histogrammed energy test, comparing 50 000 sets of 100 000 randomly distributed points on the unit square to a constant distribution and to a second uniform distribution at 100×100 binning.

if samples are consistent with arising from a constant distribution, and two-sample GoF tests to test if samples come from the same parent distribution as the random distribution.

The resulting test statistic distributions are shown in figure 3. Aslan and Zech [8, 11] found that their test distribution is well described by a generalised extreme value (GEV) distribution [12] but were unable to calculate the parameters of the distribution from first principles so they recommended determining the distribution by Monte Carlo methods. We also have found that GEV distributions fit well to data such as in figure 3 [5] but have used the experimental distributions rather than fits to determine percentile values. These give values for CL_{95} of 3×10^{-5} for a constant parent and 6×10^{-5} for comparisons against a uniform random distribution, as shown in the figure.

3.1.1. The Cook-Johnson distribution The power of the histogrammed energy test to determine deviations from the constant and random distributions was tested using various levels of the Cook-Johnson distribution, one of the tests used by Aslan and Zech for their discrete energy test [7]. The Cook-Johnson distribution is the multivariate uniform distribution given by

$$(X_1, \ldots, X_d) = ((1 + \frac{E_1}{S})^{-a}, \ldots, (1 + \frac{E_d}{S})^{-a}) \tag{4}$$

where $E_1, \ldots, E_d$ are independent and identically distributed exponential random variables, S is an independent gamma(a) random variable and $a > 0$ is a parameter [13]. For $a \to \infty$ this approaches a uniform distribution within the d-dimensional hypercube; as $a \to 0$ the distribution becomes correlated, $X_1 = \ldots = X_d$ (see figure 4 for examples in two dimensions).

The power of the energy test and the ROOT 2D-KS and 2D-χ^2 tests for comparing the various Cook-Johnson distributions on the unit square against the constant reference and uniform distributions are given in table 2. The selection criteria are the 95% confidence levels established in section 3.1 for the energy test, and a 5% acceptance level for the probability P returned by the ROOT 2D-KS and 2D-χ^2 tests. From the table it is evident that the histogrammed energy test has a much higher power than the ROOT 2D tests, rejecting Cook-Johnson distributions up to $a = 50$, whereas the ROOT tests only reject distributions with much lower values of a. Both of the ROOT tests perform rather better at rejecting the hypothesis that the Cook-Johnson

Table 2. The discrimination power of the histogrammed energy test and the ROOT 2D tests comparing 2D Cook-Johnson distributions to a constant reference distribution and a uniform distribution.

	Constant reference			Uniform reference		
Cook-Johnson parameter a	Energy Test power	2D-KS power	2D-χ^2 power	Energy Test power	2D-KS power	2D-χ^2 power
200	0.092	0.0	0.0	0.076	0.012	0.035
100	0.19	0.0	0.0	0.114	0.012	0.051
50	0.806	0.0	0.0	0.412	0.013	0.05
20	1.0	0.0	0.0	1.0	0.023	0.172
10	1.0	0.0	0.0	1.0	0.018	0.87
5	1.0	0.0	0.0	1.0	0.01	1.0
2	1.0	0.379	1.0	1.0	0.491	1.0
1	1.0	1.0	1.0	1.0	1.0	1.0
0.6	1.0	1.0	1.0	1.0	1.0	1.0

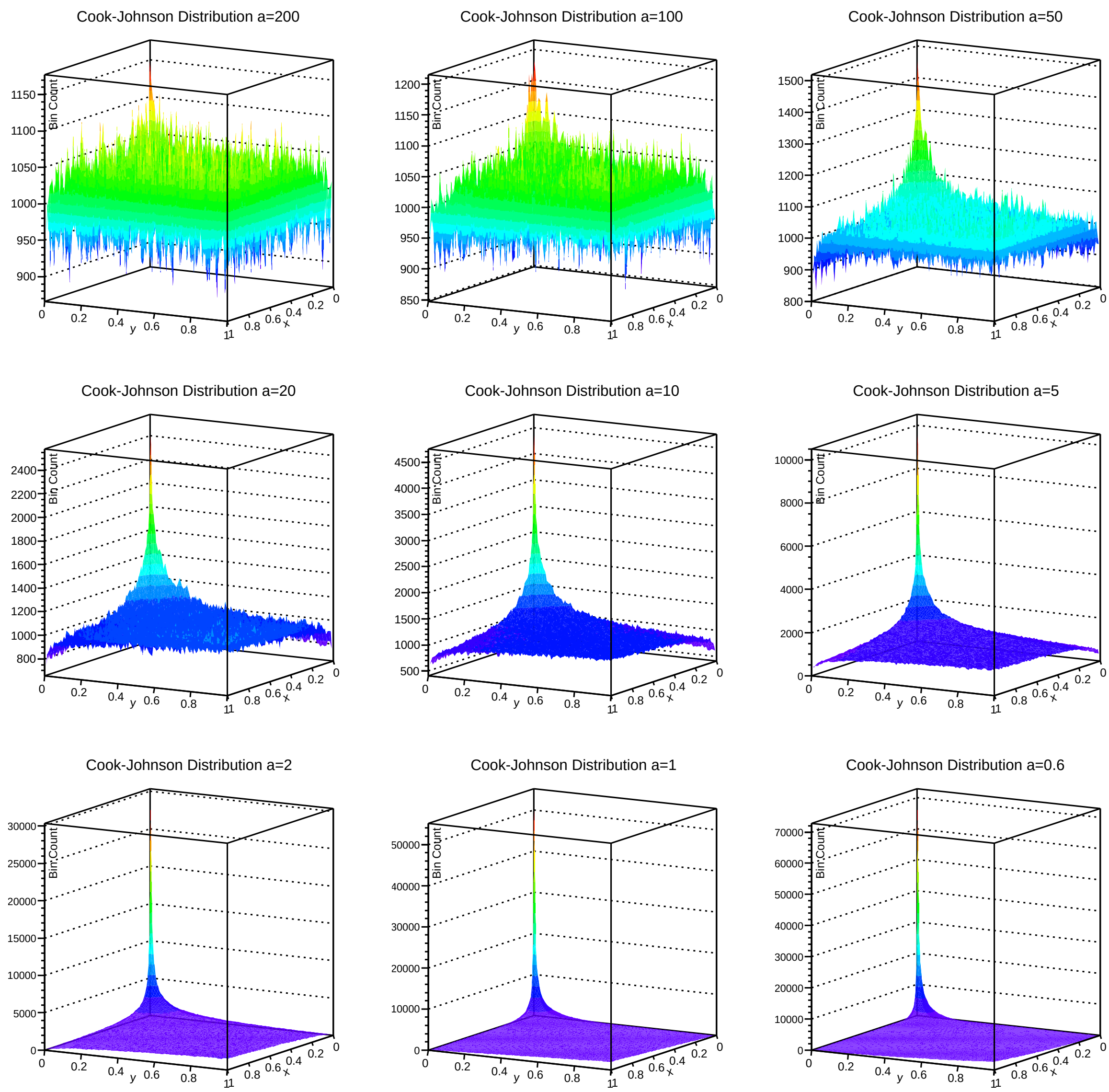


Figure 4. The two-dimensional Cook-Johnson distribution, for parameter a=200, 100, 50, 20, 10, 5, 2, 1 and 0.6, plotted as 100×100 histograms on the unit square. Each histogram contains 10^7 points (i.e., an average of 1 000 points per bin). Note the change in the z range of the distribution as a becomes smaller.

distributions are from a randomly distributed population than in rejecting them as from a constant distribution.

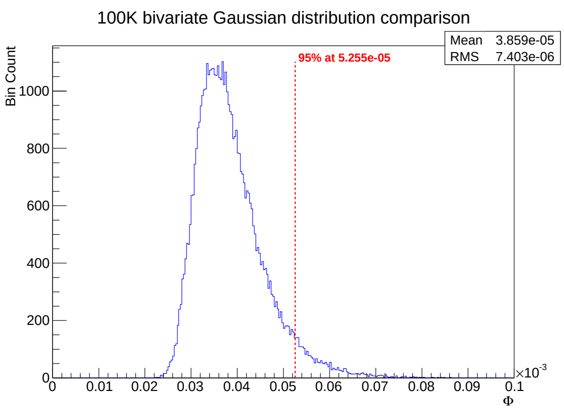
3.1.2. Gaussian contamination As a test of sensitivity to contamination, similar comparisons were made between the constant and uniform reference distributions and 1 000 samples of a uniform distribution where n% ($n = 0,1,\ldots,8$) of the 100 000 points in each sample were replaced by points from a bivariate N(0,1) (Gaussian) distribution (see figure 5). The 100×100 histograms' limits were [-3,3] in each dimension with the bivariate distribution truncated at these limits; because of the normalisation in the energy test, the same CL_{95} values as in section 3.1 are expected for 100 000-point uniform distributions. Acceptance criteria for the ROOT tests were the same as in section 3.1.1.

Table 3 gives the discrimination power of the three tests. As expected, the observed power of the energy test for 0% contamination is consistent with the choices of CL_{95}, which strongly reject distributions with 2% contamination or higher. However, the ROOT 2D-KS test only shows high discrimination power at 3% contamination and above for both hypotheses, while the ROOT 2D-χ^2 test does not reject any distributions in the case of a constant reference but performs slightly better at considering the contaminated samples not to match a uniform distribution.

Table 3. The discrimination power of the histogrammed energy test and the ROOT 2D-KS and 2D-χ^2 tests for comparisons of increasing levels of a bivariate N(0,1) distribution contamination in a uniform distribution in -3<x, y<3 against a constant and a uniformly distributed reference.

	Constant reference			Uniform reference		
Gaussian contamination	Energy Test power	2D-KS power	2D-χ^2 power	Energy Test power	2D-KS power	2D-χ^2 power
0%	0.044	0.0	0.0	0.045	0.005	0.048
1%	0.731	0.003	0.0	0.351	0.083	0.054
2%	1.0	0.194	0.0	0.982	0.560	0.075
3%	1.0	0.964	0.0	1.0	0.983	0.114
4%	1.0	1.0	0.0	1.0	1.0	0.238
5%	1.0	1.0	0.0	1.0	1.0	0.454
6%	1.0	1.0	0.0	1.0	1.0	0.690
7%	1.0	1.0	0.0	1.0	1.0	0.917
8%	1.0	1.0	0.0	1.0	1.0	0.983

3.1.3. Displacement sensitivity The sensitivity of the tests to a shift in the position of a histogrammed sample was investigated by comparing 1 000 pairs of 100 000-point bivariate N(0,1) distributions, in 100×100 histograms with a range of [-3,3] in each dimension, while the second distribution was moved away from (0,0) in x-increments of 0.01 (1/6th of a bin width). All



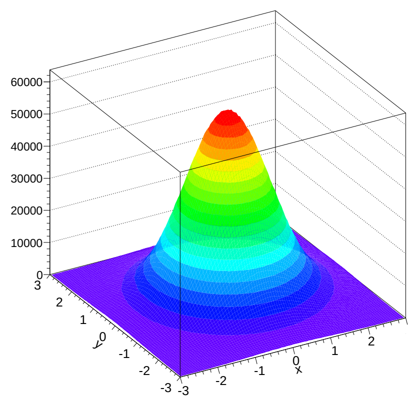


Figure 5. The distribution of results of the histogrammed energy test, comparing 50 000 pairs of histograms, each consisting of 100 000 points drawn from a bivariate N(0,1) distribution (shown on the right) in the region -3<x, y<3.

distributions were truncated at the histogram limits.

For the histogrammed energy test CL_{95} was taken from the test metric distribution obtained from 50 000 pair-wise comparisons at δx=0 which yielded a value of 5.26×10^{-5} (figure 5); acceptance criteria for the ROOT 2D-KS and 2D-χ^2 tests were again P >5%.

The calculated powers for the tests are given in table 4. The histogrammed energy test provides slightly better rejection than the ROOT 2D-KS test, approaching full rejection at δx=0.02 (1/3rd of a bin width) compared to 0.03 for the 2D-KS test. The ROOT 2D-χ^2 test, however, does not provide high rejection until the separation approaches two bin-widths (0.12).

Table 4. The discrimination power of the histogrammed energy test and the ROOT 2D-KS and 2D-χ^2 tests for comparisons of increasing displacement δx between bivariate N(0,1) distributions in 100×100 histograms covering $-3<x, y<3$.

δx	Energy Test power	2D-KS power	2D-χ^2 power	δx	Energy Test power	2D-KS power	2D-χ^2 power
0	0.045	0.009	0.026	0.07	1.0	1.0	0.490
0.01	0.380	0.135	0.021	0.08	1.0	1.0	0.713
0.02	0.965	0.723	0.036	0.09	1.0	1.0	0.884
0.03	1.0	0.990	0.057	0.10	1.0	1.0	0.958
0.04	1.0	1.0	0.100	0.11	1.0	1.0	0.992
0.05	1.0	1.0	0.169	0.12	1.0	1.0	1.0
0.06	1.0	1.0	0.305				

4. Conclusions

We have presented our investigations into a new test for performing goodness-of-fit comparisons between two-dimensional histograms, based upon the Energy Test of Aslan and Zech.

Compared with the two existing ROOT tests for 2D histograms, the histogrammed energy test proves far superior to the ROOT `Chi2Test` and outperformed the ROOT `KolmogorovTest` in our comparisons of synthetic data sets.

The main reason for this ranking in performance seems to be that the histogrammed energy test is a global test, with comparisons between every pair of bins in the histograms entering into the result, while the ROOT 2D-KS is a regional test more influenced by neighborhood variations as the CDFs are built up. The ROOT χ^2 test for its part is strictly a localised test with each bin in the histogram only being compared to its counterpart.

The disadvantage of the histogrammed energy test is that it takes longer to perform, especially at the highest binnings, but for moderately-sized histograms the penalty is slight, particularly when the time taken to construct the histograms is also considered.

Acknowledgments

This work has been funded by the Science and Technology Facilities Council, UK.

References

[1] Chakravati I M, Laha R G and Roy J 1967 *Handbook of Methods of Applied Statistics, Volume I* (New York: John Wiley and Sons), 392–4

[2] Peacock J A 1983 Two-dimensional goodness-of-fit testing in astronomy *Mon. Not. R. Astron. Soc.* **202** 615–27

[3] http://root.cern.ch

[4] Lopes R H C, Hobson P R and Reid I D 2008 Computationally efficient algorithms for the two-dimensional Kolmogorov-Smirnov test, *Int. Conf. on Computing in High Energy and Nuclear Physics* (2–7 September 2007, Victoria BC) *J. Phys.: Conf. Ser.* **119** 042019 http://iopscience.iop.org/1742-6596/119/4/042019/pdf/1742-6596_119_4_042019.pdf

[5] Reid I D, Lopes R H C and Hobson P R 2008 Comparison of two-dimensional binned data distributions using the energy test *CMS NOTE-2008/023* (Geneva: CERN) http://cdsweb.cern.ch/record/1127567/files/NOTE2008_023.pdf

[6] Aslan B and Zech G 2003 A new class of binning-free, multivariate goodness-of-fit tests: the energy tests *Preprint* arXiv:hep-exp/0203010v5

[7] Zech G and Aslan B 2003 A new test for the multivariate two-sample problem based on the concept of minimum energy *Preprint* arXiv:math/0309164v1

[8] Aslan B 2004 *The concept of energy in nonparametric statistics: Goodness-of-Fit problems and deconvolution* Thesis, Universität Siegen http://deposit.ddb.de/cgi-bin/dokserv?idn=972172122

[9] Weisstein E W *Square Line Picking*, From MathWorld – A Wolfram Web Resource. http://mathworld.wolfram.com/SquareLinePicking.html

[10] Gaboune B, Laporte G and Soumis F 1993 Expected distances between two uniformly distributed random points in rectangles and rectangular parallelepipeds *J. Opl Res. Soc.* **44** 513

[11] Aslan B and Zech G 2005 Statistical energy as a tool for binning-free, multivariate goodness-of-fit tests, two-sample comparison and unfolding *Nucl. Instr. and Meth.* A **537** 626–36

[12] Kotz S and Nadarajah S 2000 *Extreme Value Distributions: Theory and Applications* (Singapore: World Scientific)

[13] Devroye L 1986 *Non-Uniform Random Variate Generation* (New York: Springer-Verlag) http://cg.scs.carleton.ca/∼luc/rnbookindex.html