School of Engineering and Design

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, 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 [1] have been noted in the past to exhibit shortcomings [2]. We discuss a newer comparison test for 2D histograms, based on the Energy Test of Aslan and Zech [3], which provides more decisive discrimination between histograms of data coming from different distributions, and compare it with a recent ROOT release.*

Introduction. Methods for comparing one-dimensional data are well known, one of the more widely used being the Kolmogorov-Smirnov (KS) test [4] which compares cumulative distribution functions (CDF) for two sets of data, taking as a statistic D_{max}, 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 a CDF in a d-dimensional space [5]. Multidimensional goodness-of-fit 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.

The widely-used data-handling and analysis package ROOT [3] provides two methods for comparing histograms, a KS test and a Chi2Test (χ^2). In an attempt to deal with the 2-dimensional ordering problem, the ROOT 2D-KS test for histogrammed data generates two pairs of CDFs by accumulating the binned data in the histograms being compared rasterwise, in column- and row-major fashion respectively (i.e., $\Sigma_x\Sigma_y$ and $\Sigma_y\Sigma_x$). Thus two values of D_{max} are calculated, and the Kolmogorov function is evaluated for their average to return the probability P of the null hypothesis (i.e. that the two histograms represent selections from the same distribution). This separation of coordinates makes it possible to obtain a very high value of P for significantly different distributions so long as their projections in each dimension are similar. An extreme example of this is shown in Figure 1.

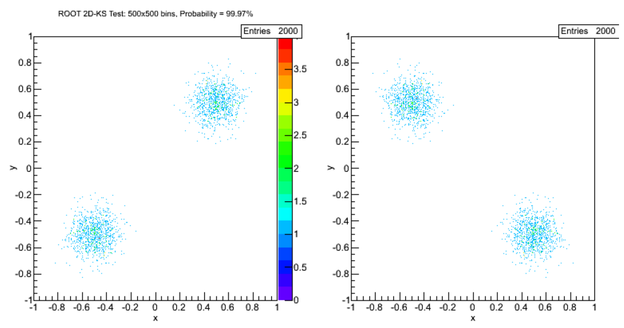

Figure 1. A ROOT 2D-KS comparison of two 2000-point histograms binned at 500x500. The test returns a high probability (99.8%) that the two histograms come from the same distribution because they each have the same projection onto the axes and hence the same cumulative distribution functions along both axes.

The Energy Test. Introduced by Aslan and Zech [3], the *Energy Test* is based on considering two data sets to be compared, $A_{1..n}$ $B_{1..m}$, to test if they arise from the same distribution, as sets of +1/n and –1/m charges respectively. In the limit $n,m\to\infty$ the total potential energy of the system will be a minimum if both sets of charges have the same distribution. The test statistic is thus $\Phi_{nm} = \Phi_A + \Phi_B + \Phi_{AB}$ where

$$\Phi_A = \frac{1}{n^2}\sum_{1=2}^{n}\sum_{j=1}^{i-1} R\left(\left|x_i - x_j\right|\right),\quad \Phi_B = \frac{1}{m^2}\sum_{1=2}^{m}\sum_{j=1}^{i-1} R\left(\left|y_i - y_j\right|\right) \text{ and } \Phi_{AB} = -\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} R\left(\left|x_i - y_j\right|\right)$$

and R is a continuous, monotonically decreasing function of the distance r between charges, e.g. $R(r) = -\ln r$ or, in practice, $R(r) = -\ln(r+\varepsilon)$. The test statistic is positive and has a minimum when the two samples are from the same distribution.

We have implemented this test to compare ROOT 2D NxN histograms by taking r as the distance between bin centres, where the axes have been normalised to [0,1], i.e. each bin has width 1/N. Where r=0 we use as an effective cutoff ε the scaled average distance between pairs of random points in a unit square, $\langle r\rangle = (2 + \sqrt{2} + 5\sinh^{-1}1)/15$ or 0.521405433[1] [6]. We calculate the three terms in the energy sum when comparing two histograms A and B with total contents n and m, respectively, as

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

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.

1) For a NxM histogram one would use instead the average distance between random points in a 1/N x1/M rectangle [7]; taking a=1/N, b=1/M and $\rho=\sqrt{(a^2+b^2)}$

$$\langle r\rangle = \frac{1}{3}\rho + \frac{a^2}{6b}\ln\left(\frac{b+\rho}{a}\right) + \frac{b^2}{6a}\ln\left(\frac{a+\rho}{b}\right) + \frac{a^5+b^5-\rho^5}{15a^2b^2}$$

Comparisons. We established a confidence level at the 95th percentile CL_{95} for comparisons against a constant distribution by testing 100x100 histograms of 100,000 points, uniformly and randomly distributed across them, against a histogram of the same size with 10 points per bin. 50,000 comparisons were made and CL_{95} established from the value below which 95% of the results fell (Figure 2). This value was then used as the basis for evaluating the test's *power* (the fraction of non-compatible data rejected by the test for a given criterion) in comparisons of various distributions, and evaluations were made against results from the ROOT 2D-KS and 2D-χ^2 tests.

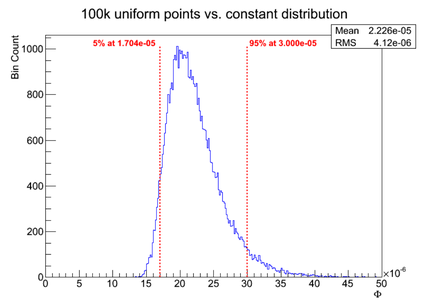

Figure 2. Distribution of the test statistic Φ for 50k comparisons of 100k random points in 100x100 histograms to a flat distribution, establishing CL_{95}=3.0E-5.

Cook-Johnson distribution. One of the the distributions used to test the discrete energy test [3] is the multivariate uniform Cook-Johnson (CJ) distribution [8], shown on the unit square in Figure 3 for varying values of its parameter a. Table 1 shows the rejection power of the three tests against the hypothesis that 100k-point 100x100 histograms from these distributions are drawn from a flat parent distribution. 1,000 histograms were tested at each value of a; rejection criteria were $\Phi > CL_{95}$ for the Energy Test, and probability $P<0.05$ for the ROOT tests. The Energy Test clearly has a higher power for rejecting these non-constant distributions.

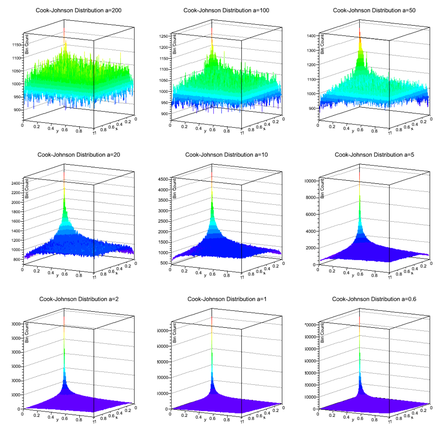

Figure 3. The Cook-Johnson distribution on the unit square for selected values of the parameter a. Each 100x100 histogram contains ten million points (av. 1,000/bin).

CJ parameter a	EnergyTest power	2D-KS power	2D-χ^2 power
200	0.092	0.0	0.0
100	0.190	0.0	0.0
50	0.806	0.0	0.0
20	1.0	0.0	0.0
10	1.0	0.0	0.0
5.0	1.0	0.0	0.0
2.0	1.0	0.379	1.0
1.0	1.0	1.0	1.0
0.6	1.0	1.0	1.0

Table 1. The rejection power for the 2D histogram comparison tests, the Energy Test and ROOT's 2D-KS and 2D-χ^2 tests, when comparing Cook-Johnson distributions against a flat distribution. Comparisons were made for 1,000 100x100 histograms, each with 100k random points from CJ distributions of varying parameter a. The power is the fraction of histograms rejected by the criteria given in the main text.

Gaussian Contamination. We have also compared the sensitivity of the tests to contamination of a uniform distribution. Again 1,000 100x100 histograms of 100k points each were compared to a flat histogram. The distributions were randomly and uniformly distributed across a range of [-3, 3] in each dimension, with an increasing proportion of points drawn from an independent bivariate N(0,1) distribution. Table 2 gives the rejection powers observed for the three tests, using the same criteria as above. The Energy Test gave ~5% rejection for zero contamination, as expected, and full rejection above 1%. The ROOT 2D-KS test provided almost full rejection at 3% contamination whereas the 2D-χ^2 test did not give any rejection at all up to 8% contamination.

Timing. The Energy Test is somewhat slower than the ROOT tests, since its complexity is $O((NxM)^2)$. This becomes noticeable beyond about 100x100 binning – see Table 3.

Conclusion. We have presented a 2D histogram comparison test based on the Energy Test. It has been shown to have a higher rejection power for histograms drawn from dissimilar populations than the existing Kolmogorov-Smirnov and chi-squared tests. However, this is at the expense of increased run-time for very fine histogram binning.

Contamination Level	EnergyTest power	2D-KS power	2D-χ^2 power
0%	0.041	0.0	0.0
1%	0.717	0.003	0.0
2%	1.0	0.194	0.0
3%	1.0	0.964	0.0
4%	1.0	1.0	0.0
5%	1.0	1.0	0.0
6%	1.0	1.0	0.0
7%	1.0	1.0	0.0
8%	1.0	1.0	0.0

Table 2. The rejection power for the 2D histogram comparison tests, the Energy Test and ROOT's 2D-KS and 2D-χ^2 tests, when comparing uniform distributions with contamination from an independent bivariate Gaussian [N(0,1)] distribution against a flat distribution. Comparisons were made for 1,000 100x100 histograms, each with 100k points, at each level of contamination.

Histo. Size	ROOT 2D-KS	ROOT 2D-χ^2	Energy Test
25x25	<10 ms	<10 ms	<10 ms
50x50	<10 ms	<10 ms	10 ms
100x100	<10 ms	<10 ms	160 ms
250x250	<10 ms	10 ms	6.1 s
500x500	30 ms	30 ms	96.3 s

Table 3. Times for the three tests at various histogram binnings, comparing 1e6 points of random uniform and constant distributions. (2.67 GHz X5550 CPU)

References:

[1] http://root.cern.ch -- Version 5.30/00 was used in this study.
[2] *Comparison of two-dimensional binned data distributions using the Energy Test*, ID Reid, RHC Lopes and PR Hobson, http://bura.brunel.ac.uk/handle/2438/2763 (2008).
[3] *A new class of binning free, multivariate goodness-of-fit tests: the energy tests*, B Aslan and G Zech, http://arxiv.org/abs/hep-ex/0203010 (2003).
[4] *Handbook of Methods of Applied Statistics, Vol. 1*, IM Chakravati, RG Laha and J Roy, (New York: John Wiley and Sons), 392-4 (1967).
[5] *Two-dimensional goodness-of-fit testing in astronomy*, JA Peacock, Mon. Not. R. Astronom. Soc. **202**, 615-27 (1983).
6] *Square Line Picking*, EW Weisstein, from MathWorld – A Wolfram Web Resource. http://mathworld.wolfram.com/SquareLinePicking.html
[7] *Expected distances between two uniformly distributed random points in rectangles and rectangular parallelepipeds*, B Gaboune, G Laporte and F Soumis, J. Opl Res. Soc. 44, 513 (1993).
[8] *Non-Uniform Random Variate Generation*, L Devroye, (New York: Springer-Verlag), (1986). http://cg.scs.carleton.ca/~luc/rnbookindex.html

Acknowledgement: This work was funded by the Science and Technology Facilities Council, UK.

Brunel UNIVERSITY LONDON