Implementing a 3D histogram version of the Energy-Test in ROOT

April 11, 2016

E. O. Cohen[a], I. D. Reid[b], E. Piasetzky[a],
cohen.erez7@gmail.com, ivan.reid@brunel.ac.uk, eip@tauphy.tau.ac.il

[a] *School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel*
[b] *College of Engineering, Design and Physical Sciences, Brunel University London, Uxbridge UB8 3PH, UK*

Abstract

Comparing simulation and data histograms is of interest in nuclear and particle physics experiments; however, the leading three-dimensional histogram comparison tool available in ROOT, the 3D Kolmogorov-Smirnov test, exhibits shortcomings. Throughout the following, we present and discuss the implementation of an alternative comparison test for three-dimensional histograms, based on the Energy-Test by Aslan and Zech.
The software package can be found at `http://www-nuclear.tau.ac.il/~ecohen/`.

Keywords: Two-Sample Test, Goodness of Fit, Energy-Test, ROOT

1. Introduction

Goodness of Fit (GoF) comparisons are a recurrent task when analyzing nuclear physics and high-energy experiments. Particularly common are GoF comparisons between histograms of data and Monte Carlo (MC) simulation. Such comparisons typically serve to determine whether the data and an MC sample are consistent with being generated from the same parent distribution. Often multiple MC sets with different parameters are generated, and GoF comparisons are needed to determine which best describes the data (The null-hypothesis (distributions are the same) is well defined, and it is important to obtain appropriate GoF methods to check its validity).

One-dimensional comparison methods are well known in the literature. Some are designed for histogrammed data comparison (e.g, the χ^2 test), while others are intended for discrete data application (e.g, the Kolmogorov-Smirnov (KS) test), though also applicable to histogrammed data provided that the binning effects are considered.

GoF using the KS test (and other existing cumulative tests) is problematic for comparing multidimensional data, as it relies on the ordering of the data to obtain the Cumulative Distribution Function (CDF) and because of the large number of distinct ways of ordering the data in space ($2^d - 1$ in d-dimensional space). Another disadvantage of multidimensional GoF tests is the lack of metric invariance, which leads to an undesirable high sensitivity of the comparison on a scale factor - or the number of bins in the histogrammed case.

1.1. *Histograms comparisons in* ROOT

ROOT is the most widely used data analysis tool in high-energy physics experiments [1]. The major existing method for comparing 3-dimensional (3D) histograms in ROOT is the **Kolmogorov-Smirnov Test** (the KS test). ROOT also implements a 3D version of the χ^2 test, though due to exceptionally inferior performance in previous 2D investigations [4, 5], it was not considered in this work. The 3D extension of the KS test is complicated by the problem of ordering the data to build the CDF. In addressing this, ROOT computes six CDFs for each histogram, accumulating the binned data raster-wise, in all distinct possible patterns, so that the comparison yields six maximum differences to which the Kolmogorov function is applied to the averages, returning the null hypothesis probability (i.e., that the two histograms represent selections from the same distribution). However, at finer histogram binning, the order in which the binned data are accumulated approaches the order of the discrete data in the slowest varying dimension [4]. Consequently, the CDFs generated by the ROOT 3D-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 (see e.g. [5] for the 2D case).

1.2. *An alternative* 3D test

The Energy Test (ETest), first proposed by Aslan and Zech [2, 3], can serve as a powerful and robust tool for multidimensional data comparison. Although this test was originally designed for discrete data, applying it to histogrammed or clustered data may expedite calculations [3].

The ETest is a two-sample test, in which the null hypothesis to be examined is that both samples originate from the same distribution. The ETest can also be considered as a standard GoF test, if there is an MC sample large compared to a data sample. In this case, the null hypothesis is that the data follow the parent distribution of the MC sample. The difference between these two cases is important for obtaining the distribution of the ETest statistic. For model-dependent calculations, a large number of MC samples can be generated and compared with the data to accumulate a distribution of the Energy-Test statistic; however, in the case of two-samples originating from real experiments, this might not be possible. The only solution in this case is to perform the test multiple times using bootstrap samples of the data.

Reid et al. [4] have implemented a version of the ETest for 2-dimensional histogrammed data within the ROOT framework, provided some evaluations of its performance [4], and presented some of its advantages over χ^2-2D and KS-2D ROOT implementations. A revisit of the 2D histogrammed implementation of the ETest was introduced in 2012 to a wider audience, together with comparisons to available 2D tests (χ^2 and KS) [5].

In this work we follow [5] and introduce a 3D histogrammed implementation of the ETest, as well as demonstrate some of its performances.

2. The Energy-Test

Consider a sample of Data (D) and MC points in a d-dimensional space, consisting of $n_{\rm D}$ and $n_{\rm MC}$ charges, $\{\mathbf{x}_i^{\rm D}\}$ and $\{\mathbf{x}_j^{\rm MC}\}$, respectively. The hypothesis that they arise from the same parent distribution is to be examined.

If D (MC) represents a system of positive (negative) point charges $1/n_{\rm D}$ ($-1/n_{\rm MC}$), then, in the limit of $n_{\rm D} \to \infty$ and $n_{\rm MC} \to \infty$, the total electrostatic energy (for a $1/r$ potential) of the two samples will reach a minimum when both samples have the same distribution. The ETest generalizes this concept.

2.1. The test statistic

The ETest statistic consists of three terms, corresponding to the self-energies of the two samples, D and MC, and the interaction energy between the two samples, $\Phi = \Phi_{\rm D} + \Phi_{\rm MC} + \Phi_{\rm DMC}$, where

$$\begin{cases} \Phi_{\rm D} = \frac{1}{n_{\rm D}^2} \sum\limits_{i=2}^{n_{\rm D}} \sum\limits_{j=1}^{i-1} \psi(|\mathbf{x}_i^{\rm D} - \mathbf{x}_j^{\rm D}|) \\ \\ \Phi_{\rm MC} = \frac{1}{n_{\rm MC}^2} \sum\limits_{i=2}^{n_{\rm MC}} \sum\limits_{j=1}^{i-1} \psi(|\mathbf{x}_i^{\rm MC} - \mathbf{x}_j^{\rm MC}|) \\ \\ \Phi_{\rm DMC} = -\frac{1}{n_{\rm D} n_{\rm MC}} \sum\limits_{i=1}^{n_{\rm D}} \sum\limits_{j=1}^{n_{\rm MC}} \psi(|\mathbf{x}_i^{\rm D} - \mathbf{x}_j^{\rm MC}|) \end{cases}$$

and ψ is a continuous, monotonically-decreasing function of the Euclidean distance r between the charges. Following [5], we choose to use $\psi = -\ln(r+\epsilon)$, rather than the electrostatic potential $1/r$, since it renders a scale-invariant function for the test, and offers better rejection powers against alternatives to the null-hypothesis[1]. The value of the cutoff parameter ϵ 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.

2.2. *Implementation of a* 3D histogrammed version of the ETest in ROOT

The ETest was implemented as a compiled ROOT macro for equally-binned ($N \times N \times N$) histograms. Aslan and Zech [3] suggest that the ranges of the data can be normalized, to equalize the relative scales of the x, y, and z-coordinates. We found that for our specific application a similar normalization is not necessary. Underflow and overflow bins (with indices 0 and $N+1$,

[1] if all axes are scaled identically.

respectively, in ROOT notation) can be included with nominal widths of $1/N$ below or above the histogram limits, selected by a user input parameter.

The choice of the number of bins chosen can be based on statistical methods proposed in the literature. The authors found the Freedman-Diaconis rule to work well in practice [6]. In this approach the bin size is chosen by

$$\text{bin size} = 2\, n(x)^{-1/3}\, IRQ(x),$$

where $n(x)$ is the number of observations in the sample x, and $IRQ(x)$ is the interquartile distance[2]. For the example of 135,000 points uniformly distributed in a unit cube, this results in $N \sim 50$ bins in each direction.

Histograms neglect intrabin positional information as all points within a given bin are assigned a single position, i.e., the bin centre. Unlike the discrete case, the self-energy between points in the same bin must be taken into account. This means that the $r = 0$ case must be treated individually, i.e., when bin (i_1, i_2, i_3) is being compared to bin (i_1, i_2, i_3)

We assume the original points are randomly distributed within the bin limits, and take the average distance between pairs of random points in a unit cube to calculate an effective cutoff ϵ. This value is $\langle r \rangle = 0.66170...$ [3] [7], so we use $\epsilon = \langle r \rangle / N$ as the cutoff distance. See below the sensitivity study to the cutoff parameter and the number of points in each bin.

We also modified the calculation of the self-energy of k points within a given bin by the weight $k^2/2$ rather than the rigorous $k(k-1)/2$, to ensure that comparisons between identical histograms return exactly zero analytically.

To summarize, the implementation of the three terms in the energy sum when comparing two $N \times N \times N$ ROOT histograms, hD representing the data and hMC representing the Monte-Carlo expectation, with total content n_{D} and n_{MC}, respectively, is given by:

[2] The interquartile distance, sometimes also referred to as the midspread, is the difference between the upper and lower quartiles.

[3] $\langle r \rangle = \frac{1}{105}\left(4 + 17\sqrt{2} - 6\sqrt{3} + 21 \sinh^{-1} 1 + 42 \ln(2 + \sqrt{3}) - 7\pi\right)$

$$
\begin{cases}
\Phi_{\mathrm{D}} = \frac{1}{n_{\mathrm{D}}^2} \sum_{d_1=0}^{N+1} \sum_{d_2=0}^{N+1} \sum_{d_3=0}^{N+1} D_{d_1,d_2,d_3} \left(\begin{array}{l} \sum_{d_1'=0}^{d_1-1} \sum_{d_2'=0}^{N+1} \sum_{d_3'=0}^{N+1} D_{d_1',d_2',d_3'} \psi_{d_1,d_2,d_3}^{d_1',d_2',d_3'} \\ + \sum_{d_2'=0}^{d_2-1} \sum_{d_3'=0}^{N+1} D_{d_1,d_2',d_3'} \psi_{d_1,d_2,d_3}^{d_1,d_2',d_3'} \\ + \sum_{d_3'=0}^{d_3-1} D_{d_1,d_2,d_3'} \psi_{d_1,d_2,d_3}^{d_1,d_2,d_3'} \\ + 0.5 D_{d_1,d_2,d_3} \mathcal{D}_0 \end{array} \right), \\
\\
\Phi_{\mathrm{MC}} = \frac{1}{n_{\mathrm{MC}}^2} \sum_{m_1=0}^{N+1} \sum_{m_2=0}^{N+1} \sum_{m_3=0}^{N+1} MC_{m_1,m_2,m_3} \left(\begin{array}{l} \sum_{m_1'=0}^{m_1-1} \sum_{m_2'=0}^{N+1} \sum_{m_3'=0}^{N+1} MC_{m_1',m_2',m_3'} \psi_{m_1,m_2,m_3}^{m_1',m_2',m_3'} \\ + \sum_{m_2'=0}^{m_2-1} \sum_{m_3'=0}^{N+1} MC_{m_1,m_2',m_3'} \psi_{m_1,m_2,m_3}^{m_1,m_2',m_3'} \\ + \sum_{m_3'=0}^{m_3-1} MC_{m_1,m_2,m_3'} \psi_{m_1,m_2,m_3}^{m_1,m_2,m_3'} \\ + 0.5 MC_{m_1,m_2,m_3} \mathcal{D}_0 \end{array} \right), \\
\\
\Phi_{\mathrm{DMC}} = -\frac{1}{n_{\mathrm{D}} n_{\mathrm{MC}}} \sum_{d_1=0}^{N+1} \sum_{d_2=0}^{N+1} \sum_{d_3=0}^{N+1} D_{d_1,d_2,d_3} \sum_{m_1=0}^{N+1} \sum_{m_2=0}^{N+1} \sum_{m_3=0}^{N+1} MC_{m_1,m_2,m_3} \psi_{d_1,d_2,d_3}^{m_1,m_2,m_3},
\end{cases}
$$

where

$$
\psi_{i_1,i_2,i_3}^{j_1,j_2,j_3} = \begin{cases} \mathcal{D}_0 = -\ln(\langle r \rangle / N), & \text{if } i_1 = j_1, i_2 = j_2, i_3 = j_3, \\ \\ -\frac{1}{2} \ln \left[\frac{(i_1-j_1)^2 + (i_2-j_2)^2 + (i_3-j_3)^2}{N^2} \right], & \text{otherwise}, \end{cases}
$$

and D_{d_1,d_2,d_3}, MC_{m_1,m_2,m_3} are the contents of individual bins within the histograms.

2.3. Computation speed

The computation time complexity of the test statistic is $\mathcal{O}(n^2)$, and in terms of histogram dimensions $\mathcal{O}(N^6)$. In order to minimize computation time, time-consuming operations were eliminated by the following:

1. Allocating local arrays holding the histogram data to enable pointer indexing rather than using the time-consuming **GetCellContents()** method when retrieving bin counts.

2. Constructing a local array to hold the potential function $\psi_{i_1,i_2,i_3}^{j_1,j_2,j_3}$.

3. Skipping computations involving empty bins.

Table 1 shows the time expenditure for comparisons between histogram pairs filled with 10^6 randomly uniformly distributed points with various binning. The comparison of data samples with distribution of equally spaced points is meant for testing, and not to describe a real application. Despite attempts to reduce calculation time, the time expenditure for fine binning ($N \geq 50$) is very large, and time-reduction programming should be further studied to address this issue. We also note that ROOT experiences frequent memory crashes for 3-dimensional arrays with large sizes ($N > 60$), due to the fixed (and finite) memory size allocated on the stack. To address this, allocated variables were put in the heap so as to manually emulate 3D arrays. All calculations reported in this work were performed on a 3 GHz Intel Core i7 processor (8 GB 1600 MHz DDR3 memory) using ROOT version 5.34/21.

Table 1: Comparison time for 10^6 points histograms of various binning with the ROOT 3D-KS test and the ETest.

Histograms Size	ROOT 3D-KS	ETest
$10 \times 10 \times 10$	< 10 ms	< 10 ms
$30 \times 30 \times 30$	< 10 ms	5.3 s
$50 \times 50 \times 50$	30 ms	150 s
$100 \times 100 \times 100$	320 ms	$2{\times}10^4$ s

2.4. Testing resolving power

The ability of a test to discriminate against non-conforming data, usually referred to as the *power* of the test, serves as a measure for the test capability to reject incompatible data sets based on selected criterion. Determining the power is possible only if a confidence level for accepting the test result is established. A traditional criterion is a confidence level of 95% $\mathrm{CL}_{95\%}$.

In order to test our implementation of the 3D ETest, two reference sets were generated: (a) A unit cube filled with a constant distribution (no statistical fluctuations) of 37 points in each one of a $30 \times 30 \times 30$ bins, and (b) a continually re-generated sample of 1,000,000 points randomly and uniformly distributed in the unit cube. 10,000 tests were performed against these references using samples of 1,000,000 random points. The first sample served as a reference for a one-sample GoF test that can determine the consistency with the assumption of a constant distribution, and the second for a two-sample comparison test to determine if both resulted from the same parent distribution.

Fig. 1 shows the resulting test statistic distributions. The values for $\mathrm{CL}_{95}\%$ are 2.2×10^{-6} for a constant parent and 4.1×10^{-6} for comparison between uniform random distributions.

2.5. Gaussian contamination

The test for sensitivity to contamination was conducted by the following [5]. The comparisons described above in Section 2.4 were repeated 1,800 times with 1,000,000 points, but where $n = 0 - 20\%$ of the points from each sample were replaced by a trivariate $\mathcal{N}(\mu = 0.5, \sigma = 0.1)$ Gaussian distribution. The ETest discrimination power was determined as the fraction of comparison below the corresponding $\mathrm{CL}_{95\%}$. Results are presented in Fig. 2 and Table 2. As expected, for 0% contamination the result is consistent with the choice of $\mathrm{CL}_{95}\%$, which clearly rejects distributions with $n > 1\%$ contamination. The ETest exhibits superior performance than the 3D KS test.

Table 2: Discrimination power of the ETest and the ROOT 3D-KS ($30 \times 30 \times 30$ binning), as a function of the contamination. See text for details.

Gaussian Contamination	ETest power	ROOT 3D-KS power
0%	0.044	0.0
0.01%	0.051	0.0
0.1%	0.129	0.0
0.7%	1.0	0.0
1%	1.0	0.0
1.3%	1.0	0.0
3%	1.0	0.010
5%	1.0	0.260
10%	1.0	0.942
15%	1.0	0.999
20%	1.0	1.0

2.6. Binning effects

To study the effects of the different number of bins on the ETest resolving power, a set of 1,000,000 points uniformly distributed inside the unit cube was compared to 3,000 similar sets, each contaminated by a fixed fraction of $n = 0.1\%$ Gaussian distributed $\mathcal{N}(\mu = 0.5, \sigma = 0.1)$ points. The discrimination power for different binning (for $N = 10^3, 20^3, 30^3, 40^3$ and 50^3) is reported in Table 3. As expected, the discrimination power is improved with finer binning, though not drastically.

Table 3: ETest 95% confidence level for comparison between two sets of 1,000,000 uniform random distributed points and contamination of $n = 0.1\%$ as a function of the number of bins.

Histogram binning	ETest $\mathrm{CL}_{95\%}$	ETest power
$10 \times 10 \times 10$	3.35×10^{-6}	0.11
$20 \times 20 \times 20$	4.05×10^{-6}	0.11
$30 \times 30 \times 30$	4.10×10^{-6}	0.13
$40 \times 40 \times 40$	4.65×10^{-6}	0.12
$50 \times 50 \times 50$	4.85×10^{-6}	0.14

2.7. Cutoff parameter impact

To study the effects of different cutoff parameters values on the ETest results, the comparisons described in section 2.4 were repeated 3,000 times using cutoff parameters $\langle r \rangle$ in the range $0.1 - 1.0$. The Gaussian contamination was fixed at $n = 0.1\%$ Gaussian distribution $\mathcal{N}(\mu = 0.5, \sigma = 0.1)$.

Figure 3 shows results from this study. As expected, the choice of the cutoff parameter is not critical if its order of magnitude equals the mean intra-points distance in the densest distributions region.

2.8. 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 135,000 trivariate $\mathcal{N}(\mu = 0.5, \sigma = 0.1)$

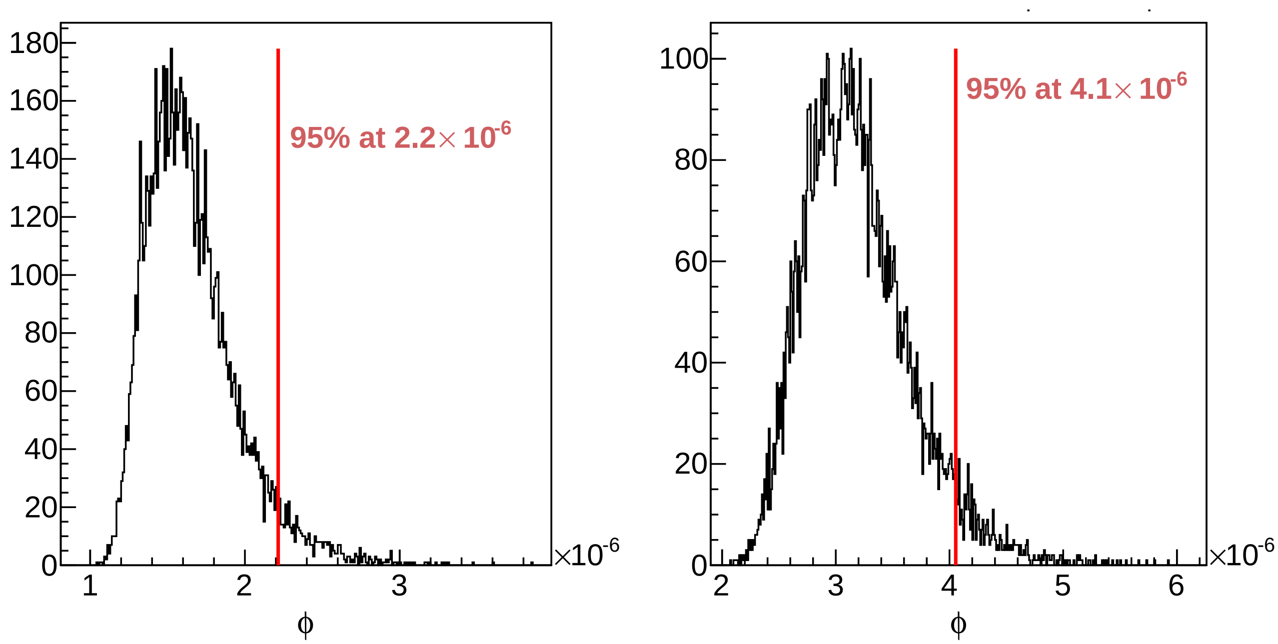

Figure 1: Distribution of the 3D ETest statistic. Compared are 10,000 sets of 1,000,000 randomly distributed points in the unit cube to a constant distribution and to a second uniform distribution, with $30 \times 30 \times 30$ bins.

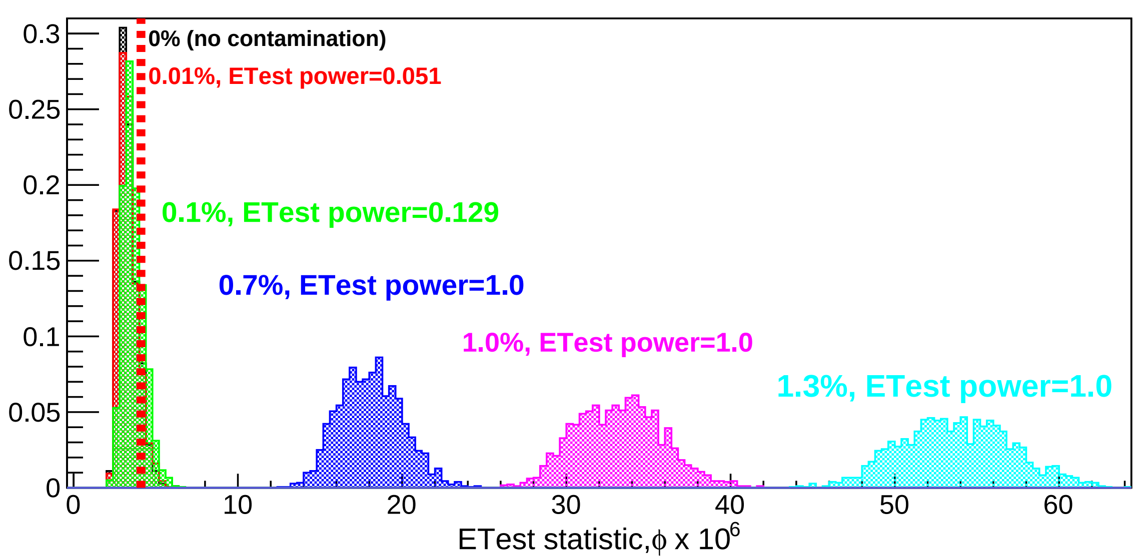

Figure 2: Same as Fig. 1 right, with one sample contaminated by $n = 0, 0.01, 0.1, 0.7, 1,$ and 1.3% trivariate $\mathcal{N}(\mu = 0.5, \sigma = 0.1)$ Gaussian distribution. The red doted line indicates the $\mathrm{CL}_{95\%}$.

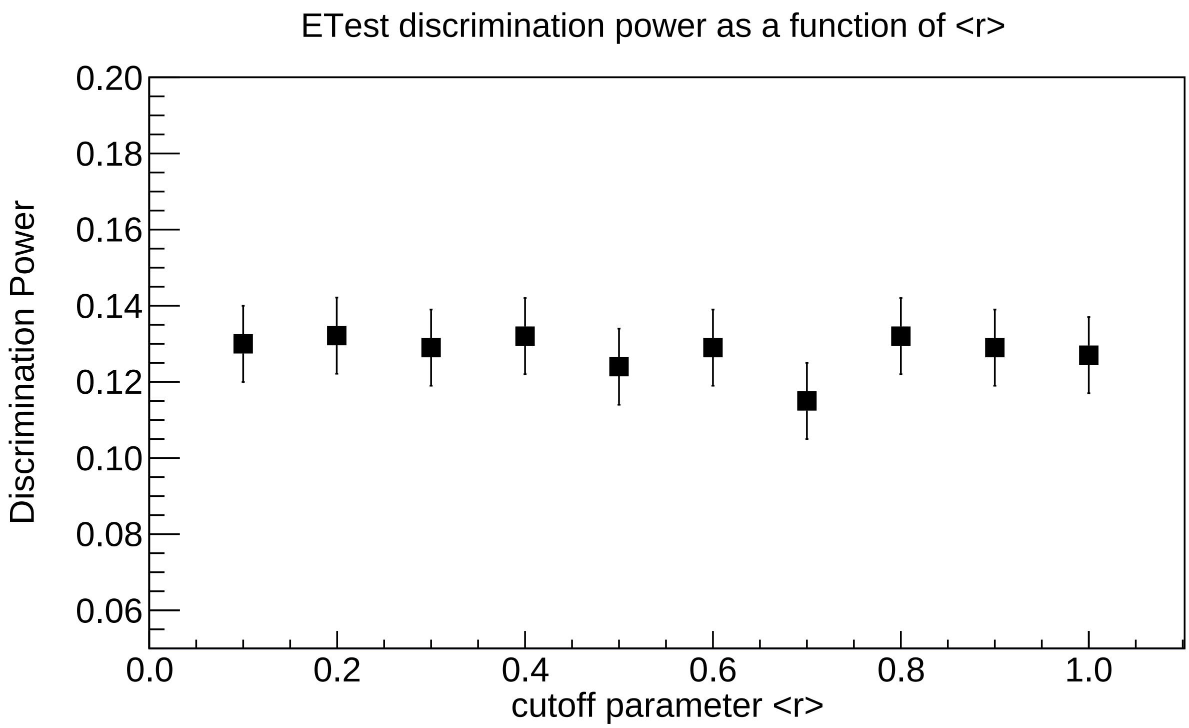

Figure 3: ETest discrimination power for different cutoff parameters $\langle r \rangle$ in the range $0.1 - 1.0$, with $30 \times 30 \times 30$ bins. Compared are sets of 1,000,000 uniformly distributed points inside the unit cube and contamination of 0.1% against a uniform reference.

distributed points and $30 \times 30 \times 30$ bins. The second distribution was shifted away $(0.5, 0.5, 0.5)$ by several values (δx). For the histogrammed Energy-Test, CL_{95} was taken from the test metric distribution obtained from 10,000 pair-wise comparisons at $\delta x = 0$, which yielded a value of 1.95×10^{-5} (Figure 4); The selection criteron for the ROOT 3D-KS tests was a 5% acceptance level. The calculated powers for the tests are given in Table 4. The histogrammed ETest provides significantly better rejection than the ROOT 3D-KS test, approaching full rejection at $\delta x = 0.002$ (about 6% of bin size), compared to $\delta x = 0.2$ for the 3D-KS test.

Table 4: Discrimination power of the ETest and the ROOT 3D-KS test for various δx displacements between trivariate $N(\mu = 0.5, \sigma = 0.1)$.

δx	ETest power	ROOT 3D-KS power
0.0001	0.150	0.0
0.0005	0.337	0.0
0.0007	0.477	0.0
0.001	0.910	0.0
0.002	0.999	0.0
0.003	1.0	0.0
0.004	1.0	0.002
0.1	1.0	0.350
0.15	1.0	0.790
0.2	1.0	1.0

3. Conclusions

A new implementation of the Energy Test of Aslan and Zech, for performing GoF comparisons between three-dimensional histograms, was introduced and investigated. The software package can be found at `http://www-nuclear.tau.ac.il/~ecohen/`.

Concluding this investigation, we show that the histogrammed ETest is superior to the only available ROOT **Kolmogorov-Smirnov Test**, for comparing synthetic data sets.

The main reason for this seems to be the fact that the histogrammed ETest is a global test that compares each pair of bins in the histograms, while the ROOT 3D-KS is sensitive to neighborhood variations, dependent on the way in which the CDFs are built.

The disadvantage of the histogrammed ETest is that its calculations are time consuming, especially with fine binnings. For moderately-sized histograms the penalty is slight, particularly if the time taken to construct the histograms is also considered.

An upgraded version of the 3D ETest, which also includes an un-binned test option, is planned for implementation in ROOT in the near future.

4. Acknowledgments

This work was supported by the United States-Israel Binational Science Foundation, as well as the Science and Technology Facilities Council, UK. Additionally, Erez O. Cohen would like to acknowledge the support

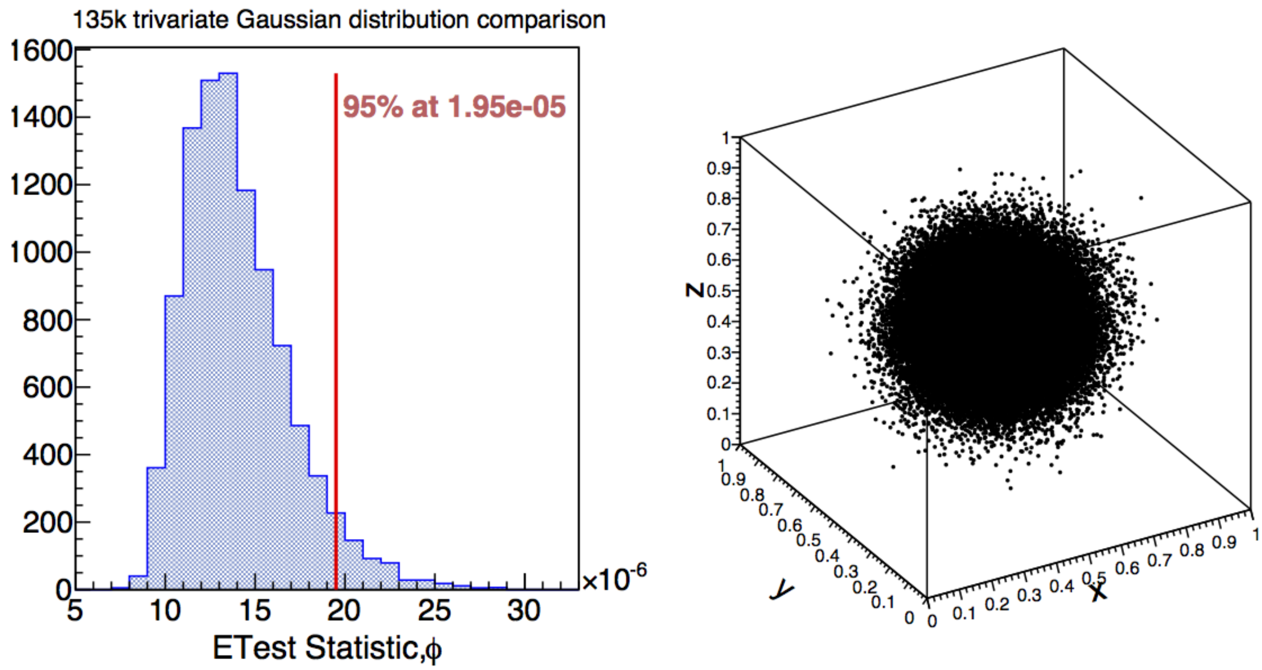

Figure 4: The distribution of results of the histogrammed ETest, comparing 10,000 pairs of histograms, each consisting of 135,000 points drawn from a trivariate $N(\mu = 0.5, \sigma = 0.1)$ distribution (shown on the right) in the unit cube.

of the Azrieli Foundation.

References

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

[2] B. Aslan, G. Zech, Nucl. Instr. Meth. A 537 (2005) 626.

[3] B. Aslan, G. Zech, J. Statist. Comput. Simul. 75 (2004) 109.

[4] I. D. Reid, R. H. C. Lopes and P. R. Hobson, CERN-CMS-NOTE-2008-023.

[5] I. D. Reid, R. H. C. Lopes and P. R. Hobson, J. Phys. Conf. Ser. **368**, 012046 (2012).

[6] Freedman, D., Diaconis, P.,"On the histogram as a density estimator: L2 theory", `http://dx.doi.org/10.1007/BF01025868`.

[7] Weisstein, Eric W. "Cube Line Picking." From MathWorld– `http://mathworld.wolfram.com/CubeLinePicking.html`.

[8] Philip, J. "probability distributions of distributions between two random points in a box", `https://people.kth.se/~johanph/habc.pdf`.