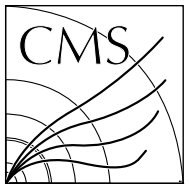
Available on CMS information server

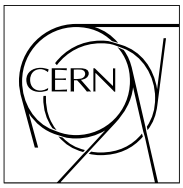
CMS NOTE-2008/023



The Compact Muon Solenoid Experiment

CMS Note

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland



14 February 2008

Comparison of Two-Dimensional Binned Data Distributions Using the Energy Test

I.D. Reid, R.H.C. Lopes and P.R. Hobson
School of Engineering and Design,
Brunel University, Uxbridge UB8 3PH, UK

Abstract

For the purposes of monitoring HEP 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 experiments now starting up, there is a need for automation of this task since the volume of comparisons would overwhelm human operators. However, the two-dimensional histogram comparison tools currently available in ROOT have noticeable shortcomings. We present a new comparison test for 2D histograms, based on the Energy Test of Aslan and Zech, which provides more decisive discrimination between histograms of data coming from different distributions.

1 Introduction

A traditional task when monitoring HEP experiments has been the comparison between regularly acquired histograms of data and reference histograms which represent the baseline state of the equipment. Histograms are typically used rather than the raw data points because of the compactness they afford, both in data storage 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 larger experiments now starting up, such as CMS, there is a need for automation in the comparison task since the number of histograms would overwhelm human operators.

These types of comparison are called goodness-of-fit (GoF) tests, and can be subdivided into two broad types: the determination of whether a given data sample is consistent with being generated from some specified distribution is sometimes called a one-sample GoF test, while a two-sample GoF test considers the hypothesis that two data samples are derived from the same 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 are well known, one of the more widespread being the Kolmogorov-Smirnov test [1]. This 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 a CDF 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 Currently available 2D tests for histogrammed data

The most-widely used data-handling and analysis package in HEP today is undoubtedly ROOT [3], which provides two methods for comparing histograms, the Chi2Test (χ^2) [4, 5] and the KolmogorovTest (KS) [6]. Details of the ROOT χ^2 test may be found in Appendix A. The ROOT KS test operates as described above, by finding the maximum difference $D_{\max}$ between the CDFs for the two histograms. The Kolmogorov distribution function [7] is applied to the normalised maximum distance $D_{\max}\{(n_A * n_B)/(n_A + n_B)\}^{1/2}$, where n_A, n_B are the sums of the histogram contents, to return the probability P of the null hypothesis (i.e., that the two histograms represent selections from the same distribution). The returned value is calculated such that it will be uniformly distributed between zero and one for compatible histograms, provided the data are not binned (or the number of bins is very large compared with the number of events) [6]. In practice, binning 1D data into histograms skews the distribution of P [6] and 2D histograms appear also to distort the distribution – as will be seen later – so that selecting an acceptance criterion of, say, 5%, will in fact reject fewer than 5% of compatible histograms.

In an attempt to deal with the 2-dimensional ordering problem, the ROOT 2D-KS test 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., $\sum_x \sum_y$ and $\sum_y \sum_x$). Thus two values of $D_{\max}$ are calculated, and the Kolmogorov function is evaluated for their average, normalised as above, to return the value of P. See Appendix B for details of the CDF calculations.

To illustrate the ROOT tests in two dimensions, two sets of 100 000 (x, y) synthetic data points were generated using ROOT [1)]. Each set had a normal N(μ,σ^2)=N(0,1) distribution in x and a Landau distribution [8] Landau(*mpv*,σ) in y, where the most-probable value *mpv* was set to 2.0 and σ was 0.50 and 0.62 for the two data sets, respectively. The data points are plotted as 50x50 histograms over $(-3 \leq x \leq 3,\ 0 \leq y \leq 30)$ in Figure 1. Results of the maximum distances $D_{\max}$ and probability values P for ROOT 2D-KS comparisons of the two data sets at different binnings are given in Table 1.

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 (see Appendix B). Consequently the CDFs generated by the ROOT 2D-KS test approach those of the discrete data ordered in one dimension along each coordinate

[1)] Unless otherwise noted, results in this work were obtained with programmes and libraries distributed with CMSSW_1_2_3 [9] – ROOT V5.13/04e, Python 2.4.2, and GCC 3.2.3 – run under Scientific Linux CERN 3.0.8. on a 2.8 GHz Pentium D. Where ROOT 2D-χ^2 test calculations involved outliers, ROOT V5.17/04 was used to avoid a bug in the distributed version.

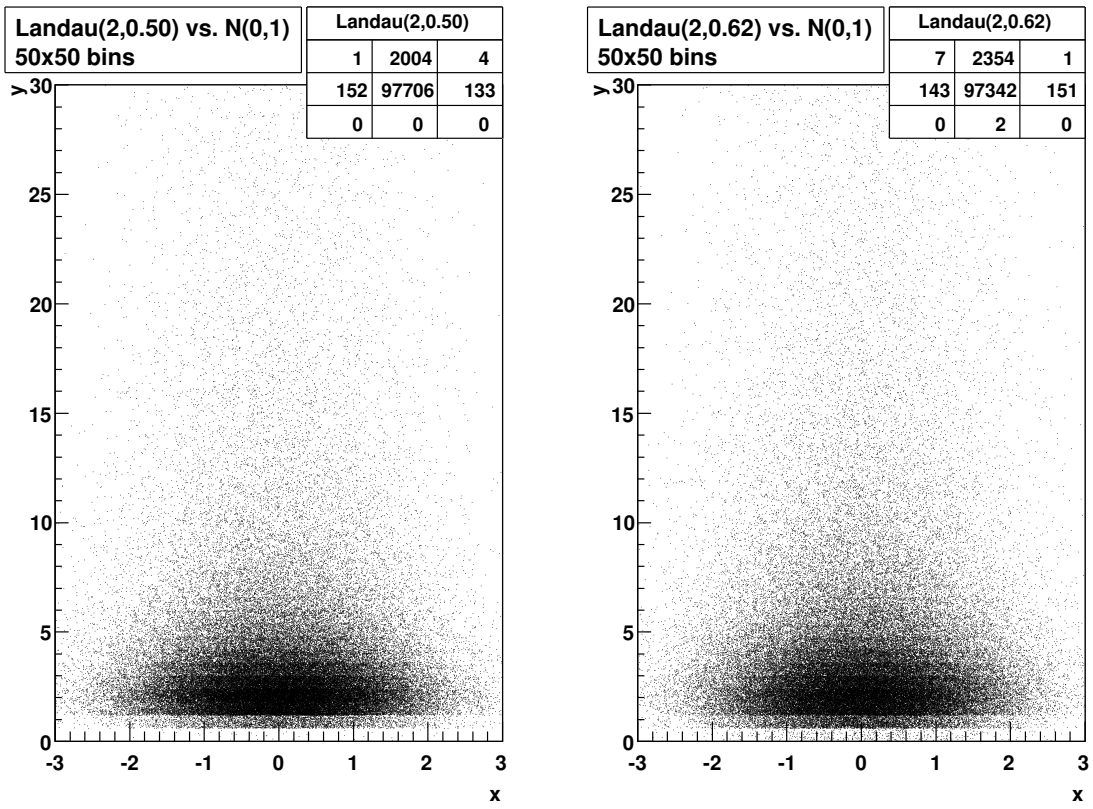


Figure 1: The two sets of 100 000-point synthetic data generated to illustrate ROOT's 2D comparison methods. Both sets have a normal distribution N(0,1) in x and a Landau distribution Landau(2,σ) in y, with σ being 0.50 (left) and 0.62 (right), respectively. The distributions are plotted here at a binning of 50x50. The statistic boxes show the number of points within the histogram limits and the distribution of outliers.

Histogram Size	$D_{\max}(y)$	$D_{\max}(x)$	2D-KS P	2D-χ^2 P
10x10	0.012082	0.039516	0.0	0.0
20x20	0.007833	0.041776	0.0	0.0
25x25	0.007885	0.043550	0.0	0.0
50x50	0.006493	0.043829	0.0	0.0
100x100	0.005404	0.043568	0.0	0.0
200x200	0.005051	0.043428	0.0	1.0
500x500	0.004773	0.043287	0.0	1.0
1000x1000	0.004802	0.043367	0.0	1.0
RPy 1D KS	0.004640	0.044510		

Table 1: The ROOT 2D comparison tests applied at different binnings to the two 100 000-sample data sets shown in Figure 1. Shown are the maximum differences $D_{\max}$ between the CDFs obtained from the two different orderings of the histogram bins, using a customised ROOT 2D-KS method. Discrete 1D KS test results, from the statistics package RPy [10] applied to the y and x data separately, are included for comparison. Also given are the probabilities P returned by the 2D-KS and 2D-χ^2 tests. Outliers were ignored in all the histogram comparisons.

separately. Table 1 illustrates this by showing how the individual $D_{\max}$ differences approach the 1D KS differences computed with the statistics package RPy [10]. 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 2.

ROOT's documentation [6] suggests that the KS test gives better results than its χ^2 test, especially at low occupancy. In practice, we have found the χ^2 test to be sensitive to binning choices and counter-intuitive when applied to 2D histograms. For example, Table 1 also gives the results of comparing the two synthetic data sets using the ROOT 2D-χ^2 test. While the ROOT 2D-KS test returns zero for P at all binnings, indicating that the two data sets are probably from different parent distributions, the 2D-χ^2 test gives zero P only for coarser binnings; at 200x200 and above it returns P=1.0, indicating compatability. This point is discussed in more detail in Appendix A.

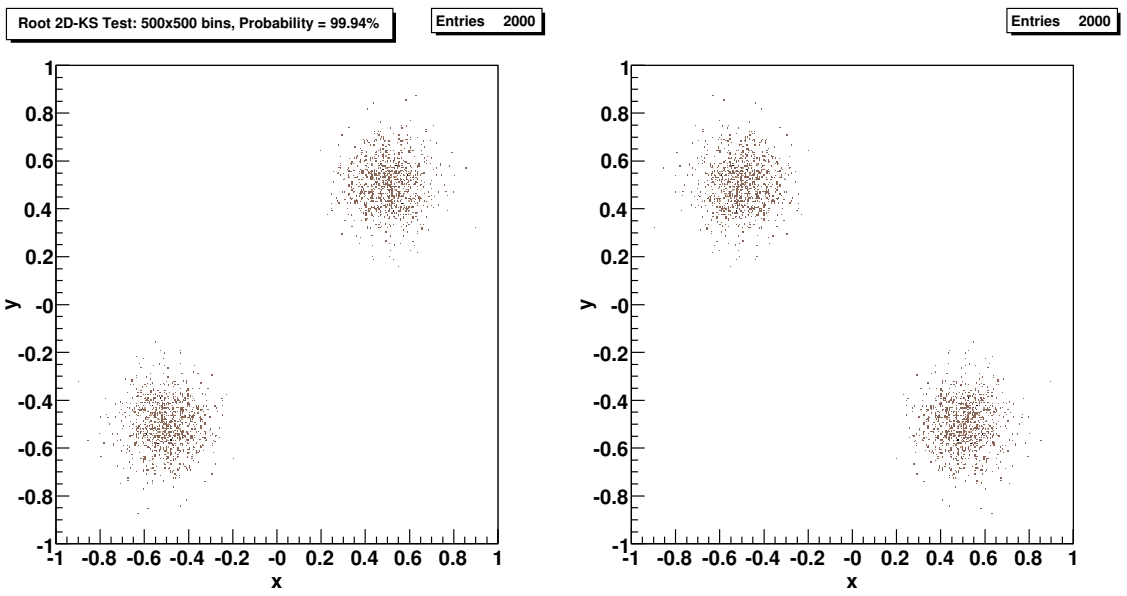


Figure 2: A ROOT 2D-KS comparison of two 2000-point histograms binned at 500x500. The test returns a high probability (P=99.94%) that the both sets of data come from the same distribution. This is because they each have the same projections onto the axes.

1.2 An alternative 2D test

We have recently demonstrated [11] efficient algorithms for a Kolmogorov-Smirnov test for discrete 2D data, as described by Peacock [2] and modified by Fasano and Franceschini [12]. Both methods are improved by using range-counting trees, and a linear speedup of the Peacock algorithms was obtained by parallel processing. Fasano and Franceschini's method is much faster than Peacock's, but no efficient parallel-processing method was found for it. Unfortunately, neither method is suitable for processing histogrammed data.

Another method for comparing distributions in more than one dimension is the Energy Test presented in recent years by Aslan and Zech [13, 14, 15]. 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 [13]. We present here an implementation of the Energy Test for histogrammed data within the ROOT framework, and provide some evaluations of its performance.

2 The Energy Test

Consider two samples of data in a d-dimensional domain, A: $\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_3,\ldots,\mathbf{X}_n$ and B: $\mathbf{Y}_1,\mathbf{Y}_2,\mathbf{Y}_3,\ldots,\mathbf{Y}_m$ whose compatability with the hypothesis that they arise from the same distribution is to be to tested. If A is taken as a system of positive charges, each $1/n$, and B as a system of negative charges $1/m$ (i.e., normalised so that the total charge over each system is one unit), 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 minimum if both charge samples have the same distribution. The energy test generalises this scenario.

2.1 The test statistic

The test statistic Φ_{nm} consists of three terms, corresponding to the self-energies of the samples A and B (Φ_A and Φ_B, respectively) and the interaction energy between the samples (Φ_{AB}):

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

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

where R is a continuous, monotonically-decreasing function of the Euclidian distance r between the charges. In most analyses Aslan and Zech chose $R(r) = -\ln r$ 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. In practice, to avoid singularities, one must use a cutoff such as $R(r) = -\ln(r + \epsilon)$, where the value of 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.

It was shown [14] 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 [13, 15] shows that when the samples have the same number of points, Φ_{nm} has a minimum when the points are pairwise coincident. Note that by inspection the calculation of Φ_{nm} is $\mathbf{O}(n^2)$.

2.2 Implementing a 2D histogram version of the energy test

A version of the energy test for ROOT 2D histograms was implemented first in Python, and then as a compiled ROOT macro for speed. Since the test is $\mathbf{O}(n^2)$ and the number of bins in a ROOT histogram is $(N{+}2)^2$ for NxN binning, calculation time rapidly increases as the binning is made finer. The first implementation reported here compares "square" (NxN) histograms, but it can easily be generalised to NxM histograms.

The implementation is straightforward, but slightly complicated by 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 care has to be taken when $r = 0$, 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). In this case 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 square as the effective cutoff ϵ. This value is $<r> = \frac{1}{15}(2 + \sqrt{2} + 5\sinh^{-1} 1)$ = 0.521 405 433… [16]. Distances for other bin combinations are calculated simply as the Euclidian distance between bin centres (with no need for an ϵ cutoff), justified by the proximity of this value to the average distance between random points in the two bins as determined by 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 (see Appendix C).

Aslan and Zech [14] 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 implemented here by taking the histogram limits to be zero and unity (i.e., the distance between adjacent rows or columns is set to $1/N$), on the grounds that a well-designed histogram will have limits chosen to adequately span anticipated data sets. In this implementation underflow and overflow bins (with indices 0 and N+1, respectively, in ROOT notation) are included and placed at $1/N$ below or above the histogram limits. A production version should make inclusion of these bins optional, as in the current ROOT 2D tests.

In equations, our implementation of the three terms in the energy sum when comparing two NxN ROOT histograms A and B with total contents n and m, respectively, is given by

$$\Phi_A \quad = \quad \frac{1}{n^2}\sum_{i=0}^{N+1}\sum_{j=0}^{N+1} A(i,j)\left(\sum_{k=0}^{i-1}\sum_{l=0}^{N+1} A(k,l)R(i,j,k,l) + \sum_{l=0}^{j-1} A(i,l)D(j,l) + 0.5A(i,j)D_0\right) \qquad (2)$$

$$\Phi_B \quad = \quad \frac{1}{m^2}\sum_{i=0}^{N+1}\sum_{j=0}^{N+1} B(i,j)\left(\sum_{k=0}^{i-1}\sum_{l=0}^{N+1} B(k,l)R(i,j,k,l) + \sum_{l=0}^{j-1} B(i,l)D(j,l) + 0.5B(i,j)D_0\right)$$

$$\Phi_{AB} \quad = \quad -\frac{1}{nm}\sum_{i=0}^{N+1}\sum_{j=0}^{N+1} A(i,j)\sum_{k=0}^{N+1}\sum_{l=0}^{N+1} B(k,l)R(i,j,k,l)$$

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.

As noted above, the number of computations rapidly increases with finer binning, so it is essential to reduce each calculation to the absolute minimum. This has been done by eliminating as much as possible all "expensive" operations in the calculations. Measures include:

- Allocating local arrays holding the histogram data to enable pointer indexing rather than the expensive method `GetCellContents()` when retrieving bin counts.
- Constructing a local array to hold the potential function $R(i,j,k,l)$ of distances between bin centres, as a two-dimensional array indexed by $(|i-k|,|j-l|)$. This avoids repeating the expensive *ln* and *sqrt* functions, although in practice the *sqrt* used in calculating the Euclidian distance r can be folded into the *ln* calculation as a factor of 0.5 in the accumulated sums.
- Skipping calculations involving empty bins.

In addition, to reduce potential numerical round-off errors due to the addition of numbers of greatly varying magnitude, running sums are accumulated in the outer loops and updated with interior sums from the inner loops.

3 Performance

First evaluations of the energy test involved reconstructions of simulated muon tracks from $\mathbf{Z}^0$-decays in the CMS Silicon Tracker. The data, obtained using the CMSSW software framework [9], are given as histograms in Figure 3, showing the relationship between the reduced χ^2 of the track fit and the track pseudorapidity η. The first histogram gives results for perfect detector alignment while the second histogram was obtained after introducing small displacements, representative of probable initial position errors [17], to the positions of individual detector modules; these data sets are referred to hereafter as *aligned* and *misaligned*, respectively. The data are binned at 20x20 resolution; the blockiness is due to ROOT's dithering each bin to fill its area proportionally to its contents. It is noticeable that χ^2 is generally higher in the second histogram around $\eta = \pm 1.5$, where tracks pass through the transition between cylindrical "barrel" detectors and circular "end caps" [9].

3.1 Discrete *vs* binned comparisons

A Kolmogorov-Smirnov comparison of the two discrete data sets using Fasano and Franceschini's range-counting tree method [11] took 3 minutes 11 seconds on our 2.8 GHz Pentium D, returning a KS distance $D_{\max}$ of 0.004251;

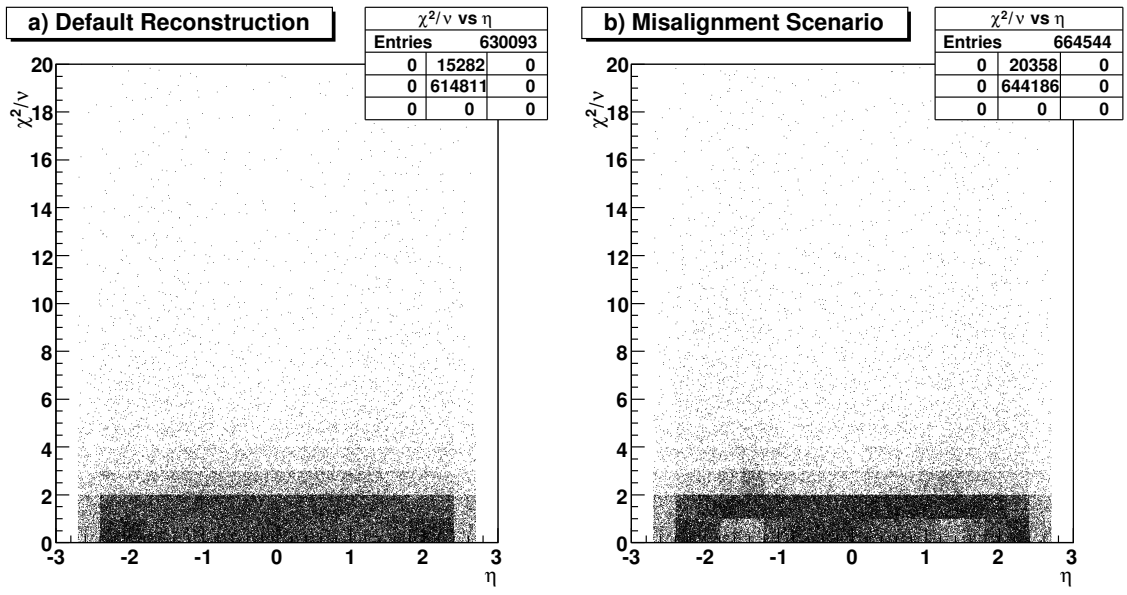


Figure 3: Data used for performance calculations. a) Reduced χ^2 of the fit vs pseudorapidity η for muon tracks reconstructed within the CMS Silicon Tracker with ideal geometry. b) Reconstruction of the same event data after the introduction of small perturbations to the positions of Tracker detectors, of the order expected when the CMS experiment first starts operation.

Binning	Φ_A	Φ_B	Φ_{AB}	Φ_{nm}	CPU Time (Real Time)	ROOT 2D-KS			ROOT 2D-χ^2	
						P	$D_{\max}(\eta)$	Time	P	Time
Discrete	0.721723	0.709668	-1.429906	0.001485	793 m 27 s (531 m 12 s)					
5x5	0.636990	0.622299	-1.259104	0.000184	<10 ms	0	0.005162	<10 ms	0	<10 ms
10x10	0.692381	0.677274	-1.369297	0.000358	<10 ms	0	0.010779	<10 ms	0	<10 ms
20x20	0.704278	0.690849	-1.394033	0.001094	<10 ms	0	0.012959	<10 ms	0	<10 ms
50x50	0.719045	0.706722	-1.424479	0.001288	0.02 s	0	0.006867	<10 ms	0	<10 ms
100x100	0.721143	0.709015	-1.428725	0.001433	0.29 s	0	0.004902	<10 ms	0	<10 ms
200x200	0.721699	0.709646	-1.429860	0.001484	3.67 s	0	0.003615	0.01 s	0	0.02 s
500x500	0.721881	0.709852	-1.430234	0.001499	131.33 s	0	0.002817	0.08 s	0	0.14 s

Table 2: Comparisons between the aligned and misaligned track data of Figure 3 using the discrete energy test and the histogrammed energy test on ROOT histograms binned at various levels. Probabilities P from the ROOT 2D-KS and 2D-χ^2 tests are also shown. The discrete calculations were performed in parallelised Fortran on a 2.2 GHz dual-core Athlon64 (SPECfp2000=1466), the ROOT calculations on a 2.8 GHz Pentium D (SPECfp2000=1664).

since this method runs as $\mathbf{O}(n\log n)$ and the similar Peacock method takes $\mathbf{O}(n^2\log n)$, no attempt was made at a years-long Peacock comparison. In contrast, a discrete energy test carried out on a comparable Athlon64 dual-core 2.2 GHz processor running 64-bit Scientific Linux CERN 4, using Intel Fortran with parallelisation options[2)], took 793 minutes 27 seconds CPU time, in 531 minutes 12 seconds of real time.

The C++ comparison code for histogrammed data was loaded into ROOT as a compiled and optimised library [3)] (loading as an interpreted macro produced runtimes ∼150 times slower) and comparisons were made between the two histogrammed data sets at different binning levels. The results are summarised and compared with the discrete result in Table 2. When run as a standalone C++ programme using the ROOT libraries, the histogrammed tests ran 5-10% faster with the use of the `-funroll-loops` compiler flag [4)] but gave identical results. It is seen that at binnings of 100x100 and above, the histogrammed comparisons gave results quite close to that of the discrete test.

However, the histogram comparisons ran much faster than the discrete test, by a large margin. Some of the speed increase is due to the smaller problem size (e.g., 102^4 = 1.08e8 for the 100x100 histogram case compared with 630093x664544 = 4.18e11 for the the discrete case), but for the 500x500 comparison the problem size is only a factor of 7 smaller (502^4 = 6.35e10) while the runtime is smaller by a factor of 360. This reduction can be mainly attributed to the lookup table for the inter-bin distance function $R(r)$. The discrete comparison needed to make some 8.4e11 evaluations of $ln((x_1 - x_2)^2 + (y_1 - y_2)^2 + \epsilon)$, a variation on the cutoff scheme which allows *sqrt* calculations to be eliminated, while the histogrammed comparison made just 125 750 similar calculations to build the 502x502 lookup table. A further reduction in time was afforded by the skipping of empty bins where possible.

Table 2 also shows results obtained by comparing the histograms using the ROOT 2D-KS and 2D-χ^2 tests. In all cases a zero result was returned, and the running times were almost instantaneous. Note that the ROOT 2D-KS debug option only provides one Kolmogorov-Smirnov distance $D_{\max}$, from the column-major accumulation, so the results may not be directly comparable to that obtained with Fasano and Franceschini's method.

3.2 Region of validity

To explore the limitations of binning and sample size, comparisons were made between samples drawn randomly without replacement from the two sets of track data, with sample sizes between 5 000 and 600 000 tracks, and binning ranging from 5x5 to 500x500. Twenty comparisons were made for each set of conditions and the average results and r.m.s. residuals are summarised in Table 3. For binnings of 50x50 and above and for sample sizes above 20 000 the results are consistently close to the 1.5e-3 obtained with the full data sets. Note that for the 600 000-point samples the spread in the results is small, because without replacement the samples comprise almost the full parent distributions so there is little variation between samples. This could have been avoided by sampling with replacement, but for consistency without-replacement sampling has been used throughout this study.

[2)] `ifort -O3 -parallel -ipo0`

[3)] `.L twoDenergy.C++O`

[4)] `g++ -O3 -funroll-loops 'root-config --cflags' 'root-config --libs' twoDenergy.cpp\ GetFullEnergy.cpp`

Binning	Sample Size							
	5 000	10 000	20 000	40 000	80 000	150 000	300 000	600 000
5x5	3.296e-4 ±2.86e-5	2.427e-4 ±4.30e-5	2.465e-4 ±6.89e-5	1.872e-4 ±1.29e-5	1.937e-4 ±3.23e-5	1.864e-4 ±8.35e-6	1.891e-4 ±2.27e-5	1.848e-4 ±6.04e-7
10x10	6.885e-4 ±1.04e-4	5.296e-4 ±2.09e-5	4.346e-4 ±8.10e-5	3.890e-4 ±1.35e-5	3.676e-4 ±3.35e-5	3.681e-4 ±3.96e-5	3.627e-4 ±9.22e-6	3.570e-4 ±1.03e-6
20x20	1.529e-3 ±3.39e-4	1.323e-3 ±2.33e-4	1.197e-3 ±5.70e-5	1.145e-3 ±2.97e-5	1.116e-3 ±4.50e-5	1.109e-3 ±8.25e-6	1.096e-3 ±4.11e-5	1.093e-3 ±2.38e-6
50x50	1.856e-3 ±1.81e-4	1.647e-3 ±2.78e-5	1.423e-3 ±8.99e-6	1.375e-3 ±6.38e-5	1.314e-3 ±3.76e-5	1.298e-3 ±1.51e-6	1.299e-3 ±3.20e-6	1.287e-3 ±3.76e-6
100x100	2.135e-3 ±1.75e-4	1.784e-3 ±1.61e-4	1.599e-3 ±1.39e-4	1.522e-3 ±2.75e-5	1.475e-3 ±6.13e-5	1.446e-3 ±1.36e-5	1.429e-3 ±1.37e-5	1.431e-3 ±4.15e-7
200x200	2.358e-3 ±4.04e-4	1.973e-3 ±2.34e-4	1.695e-3 ±2.98e-5	1.621e-3 ±1.90e-4	1.539e-3 ±2.90e-5	1.510e-3 ±3.97e-5	1.489e-3 ±2.13e-5	1.485e-3 ±5.44e-6
500x500	2.671e-3 ±1.51e-4	1.940e-3 ±8.82e-5	1.814e-3 ±6.74e-5	1.607e-3 ±1.52e-4	1.565e-3 ±2.74e-6	1.512e-3 ±5.93e-6	1.513e-3 ±2.00e-5	1.501e-3 ±1.14e-6

Table 3: The average energy metric and r.m.s. residuals obtained for a series of comparisons between samples drawn randomly without replacement from the two sets of track data, at varying histogram binning and sample size. Twenty pairs of samples were compared at each point.

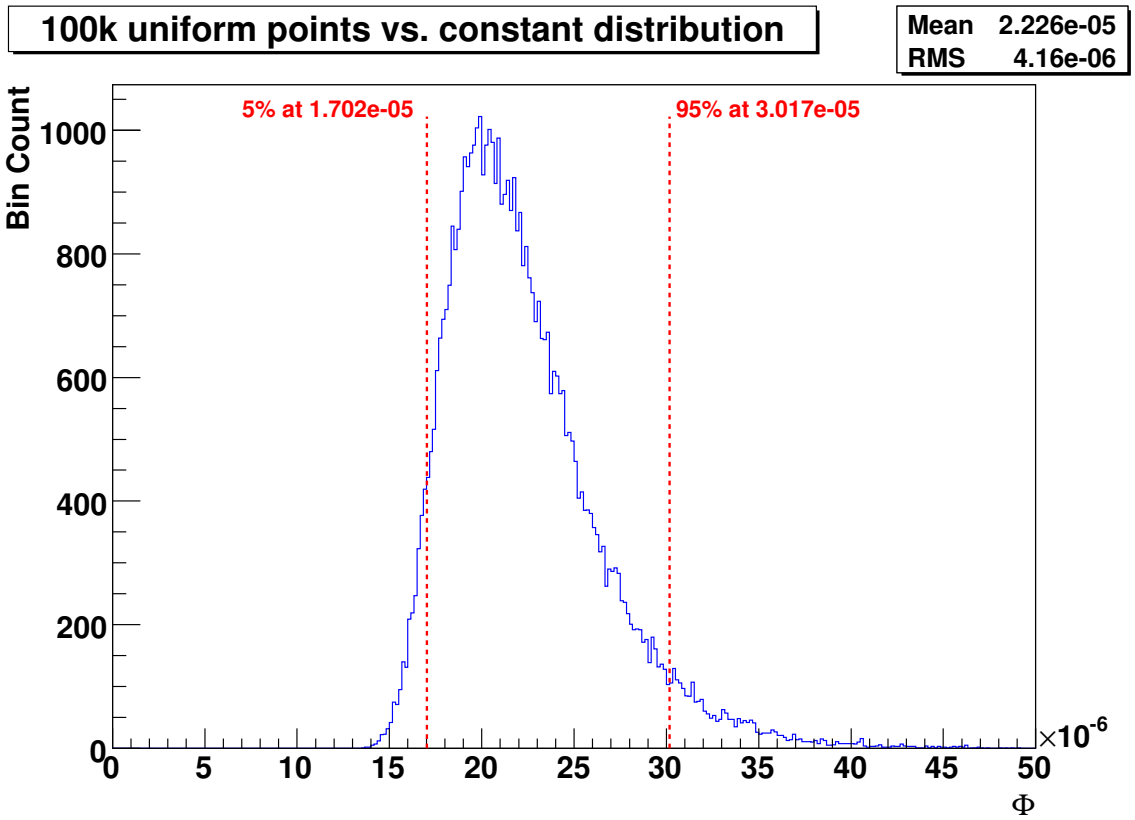


Figure 4: The distribution of results of the histogrammed energy test, comparing 50 000 sets of 100 000 randomly distributed points on the unit square to a constant distribution at 100x100 binning. The 95% confidence level is 3.017e-5.

3.3 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 – the value of a test beyond which only 5% of valid comparisons will lie.

As a reference for several tests a constant distribution (i.e., no statistical fluctuation) in a 100x100 histogram across the unit square was used. 50 000 tests were performed against this reference using samples of 100 000 points

randomly and uniformly distributed across the square. The resulting test statistic distribution is shown in Figure 4. Aslan and Zech [15, 18] found that the form of their test distribution is well described by a generalised extreme value distribution [19] but because they were unable to calculate the parameters of the distribution from first principles, and because of the high speed of modern computers, they recommended determining the distribution by Monte Carlo methods. We also found that a generalised extreme value distribution fits well to the data in Figure 4 (see Appendix D) but have used the experimental distribution rather than a fit to it to determine percentile values. The reference distribution gives the 95% confidence level at 3.017e-05, as shown in the Figure. The result distribution scales inversely with the total number of points per sample (Fig. 5), since common factors can be removed in Equation 2 so that $A(i,j)$=1, and the distribution of $B(i,j)$ approaches N(M,M)/M = N(1,1/M), where M is the average bin content, as M increases.

3.3.1 The Cook-Johnson distribution

The power of the histogrammed energy test to determine deviations from the constant distribution was tested using various levels of the Cook-Johnson distribution, one of the tests used by Aslan and Zech for their discrete energy test [14]. The Cook-Johnson distribution is the multivariate uniform distribution given by

$$(X_1, \ldots, X_d) = ((1 + \frac{E_1}{S})^{-a}, \ldots, (1 + \frac{E_d}{S})^{-a}) \tag{3}$$

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 [20]. 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 6 for examples of the 2D Cook-Johnson distribution).

Figure 7 shows the distribution of results from the histogrammed energy test and the ROOT 2D-KS and 2D-χ^2 tests for 1000 comparisons of 100 000 random points from a 2D Cook-Johnson distribution to a constant distribution, for 100x100 histograms and values of a ranging from 0.6 to 200.

The power of the energy test and the ROOT 2D-KS and 2D-χ^2 tests for comparing the various Cook-Johnson distributions against the constant reference are given in Table 4. The selection criteria are the 95% confidence level established in Section 3.3 with a uniform distribution for the energy test, and a 5% acceptance level for the ROOT 2D-KS and 2D-χ^2 tests, as shown in Figure 7. From the Table and the Figure 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 $a \leq 2$. It is noticeable that the Cook-Johnson distributions with $a \geq 10$ result in quite similar probability distributions in the ROOT 2D-KS comparisons, producing identical powers for these tests. Indeed these would all have similar powers whatever the acceptance criterion. The ROOT 2D-χ^2 test, on the other hand, shows an abrupt changeover between rejection, for $a \leq 2$, and acceptance, for $a \geq 5$.

Cook-Johnson parameter a	Energy Test power	ROOT 2D-KS power	ROOT 2D-χ^2 power
0.6	1.0	1.0	1.0
0.8	1.0	1.0	1.0
1	1.0	1.0	1.0
2	1.0	0.37	1.0
5	1.0	0.0	0.0
10	1.0	0.0	0.0
20	1.0	0.0	0.0
50	0.819	0.0	0.0
100	0.186	0.0	0.0
200	0.076	0.0	0.0

Table 4: The discrimination power of the histogrammed energy test and the ROOT 2D tests comparing 2D Cook-Johnson distributions to a constant reference, from the distributions and selection criteria shown in Figure 7.

3.3.2 Gaussian contamination

As a test of sensitivity to contamination, similar comparisons were made between a constant reference distribution and 1000 samples of a uniform distribution where n% (n = 0,1,. . .,5,10,15) of the 100 000 points in each sample were replaced by points from a rotationally-symmetric N(0,1) (Gaussian) distribution [5)]. The extent of the 100x100

[5)] That is, (r, ϕ) in polar coordinates, where r is from a N(0,1) distribution and ϕ uniformly distributed in $(0, 2\pi)$.

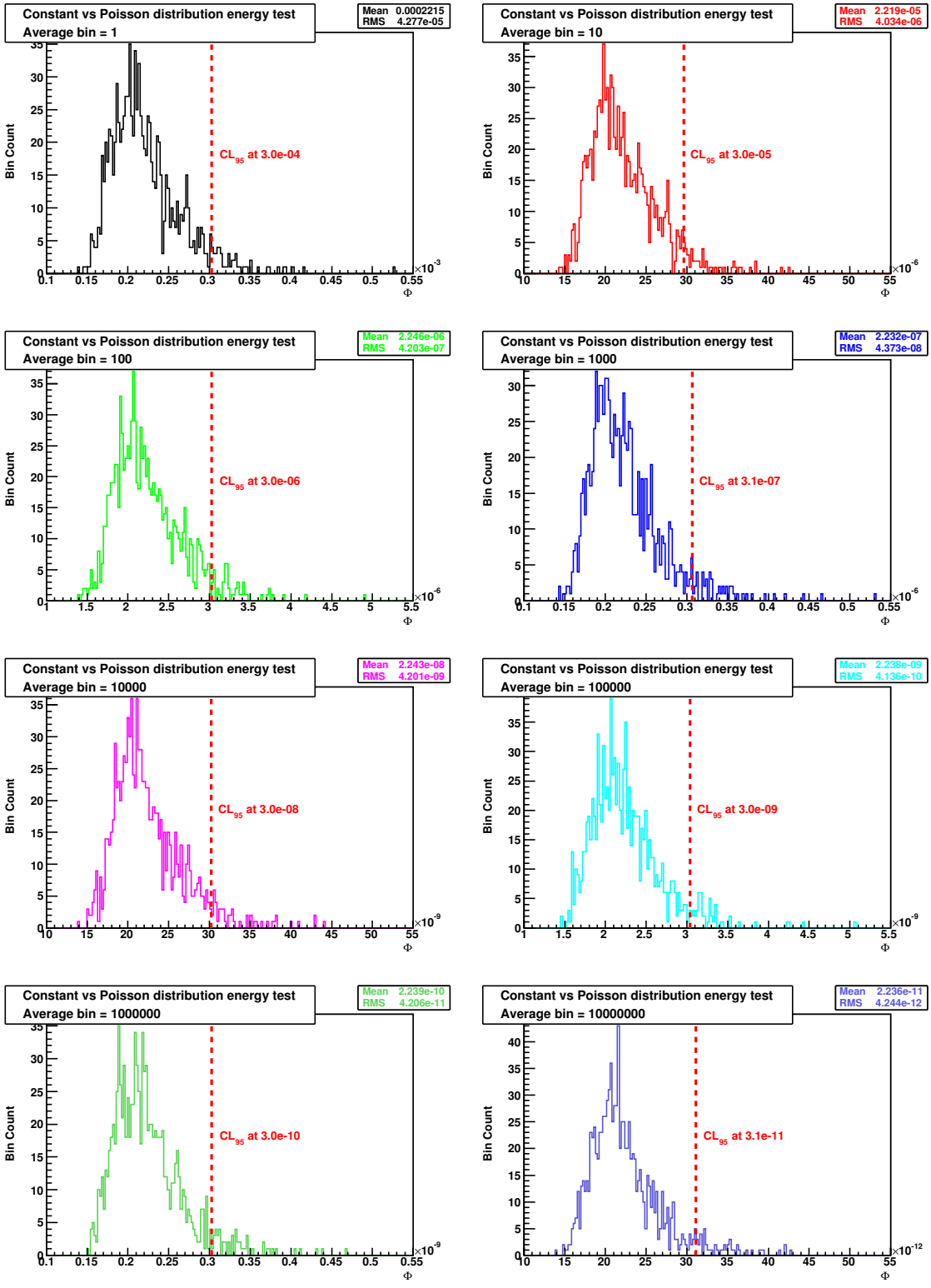


Figure 5: The distribution of results of the histogrammed energy test, comparing 1000 sets of 100x100 histograms with bins filled from a Poisson distribution, with averages from 1 to 1e7 per bin, to a constant distribution (the same value in each bin). The result distributions, including the 95% confidence level CL_{95}, scale inversely with the average bin value. The distribution with average = 10 corresponds to the 100 000-point comparisons of Figure 4.

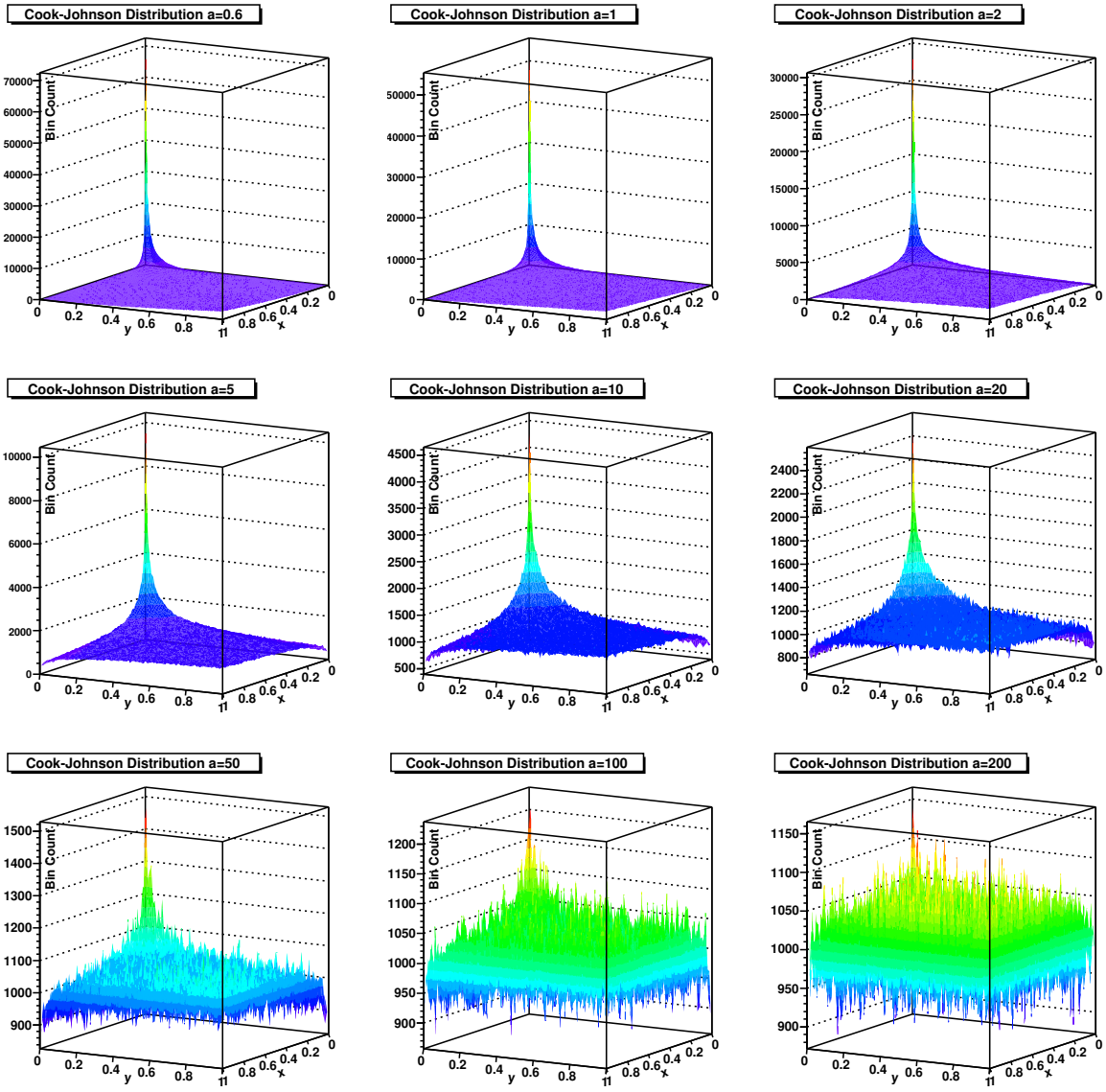


Figure 6: The two-dimensional Cook-Johnson distribution, for parameter a=0.6, 1, 2, 5, 10, 20, 50, 100, and 200, plotted as 100x100 histograms. Each histogram contains 1e7 points (i.e., an average of 1000 points per bin). Note the change in the z range of the distribution as a becomes smaller.

histograms was increased to [-3,3] in each dimension to ensure a very small proportion ($\sim$0.13%) of outliers from the tails of the Gaussian; because of the normalisation in the energy test, the same confidence level as in Section 3.3 is expected for 100 000-point uniform distributions. The histogrammed energy tests considered outliers by default, so the ROOT 2D-KS and 2D-χ^2 tests also included them.

Figure 8 shows the distributions of the results from all tests, including the selection criteria as above, and Table 5 gives the discrimination power of the two tests. As expected, the observed power of the energy test for 0% contamination is consistent with the selection of the 95% confidence level. The chosen confidence level almost completely rejects the distributions with 1% contamination and totally rejects distributions with higher contamination. However, the ROOT 2D-KS test only shows high discrimination power at 2% contamination and above, while the ROOT 2D-χ^2 test does not reject any contaminated distributions below an impurity level of 15%.

3.3.3 Displacement sensitivity

The sensitivity of the tests to a shift in the position of a histogrammed sample was investigated by comparing 1000 pairs of 100 000-point rotationally-symmetric N(0,1) distributions as defined in Section 3.3.2, in 100x100 histograms with a range of [-3,3] in each dimension, while the second distribution was shifted away from (0,0) in x-increments of 0.003 (1/20th of a bin width). Outliers were considered in all the comparisons.

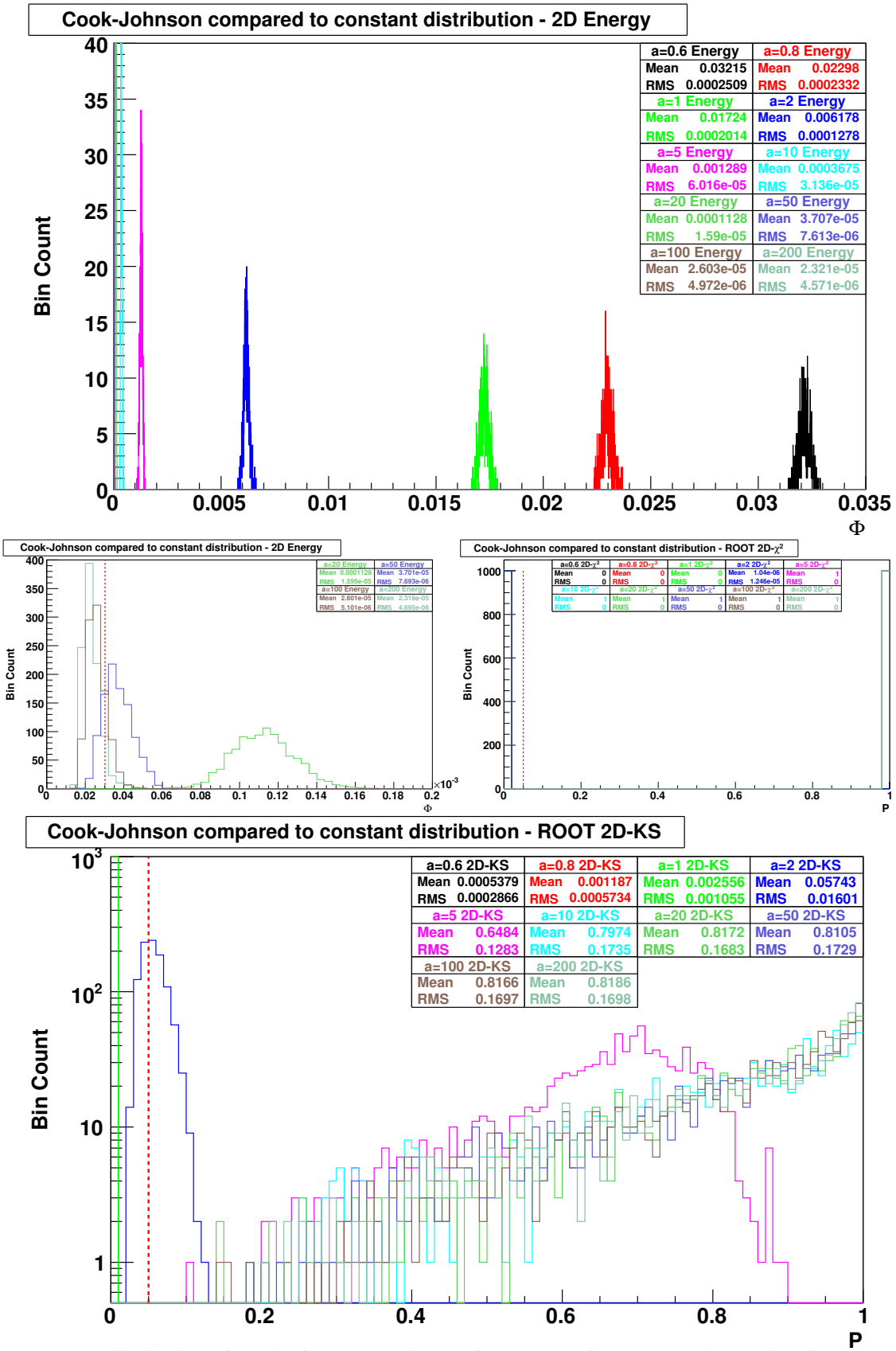


Figure 7: The distribution of results from comparisons of 1000 sets of 100 000 random points from 2D Cook-Johnson distributions (a=0.6, 0.8, 1, 2, 5, 10, 20, 50, 100, and 200) to a constant distribution, at 100x100 binning, for the histogrammed energy test (top, expanded middle left) and the ROOT 2D-χ^2 and 2D-KS tests tests (middle right and bottom). The vertical dashed lines give the 95% confidence level (at 3.017e-5) for the energy test and the 5% acceptance criteria for the KS and χ^2 tests.

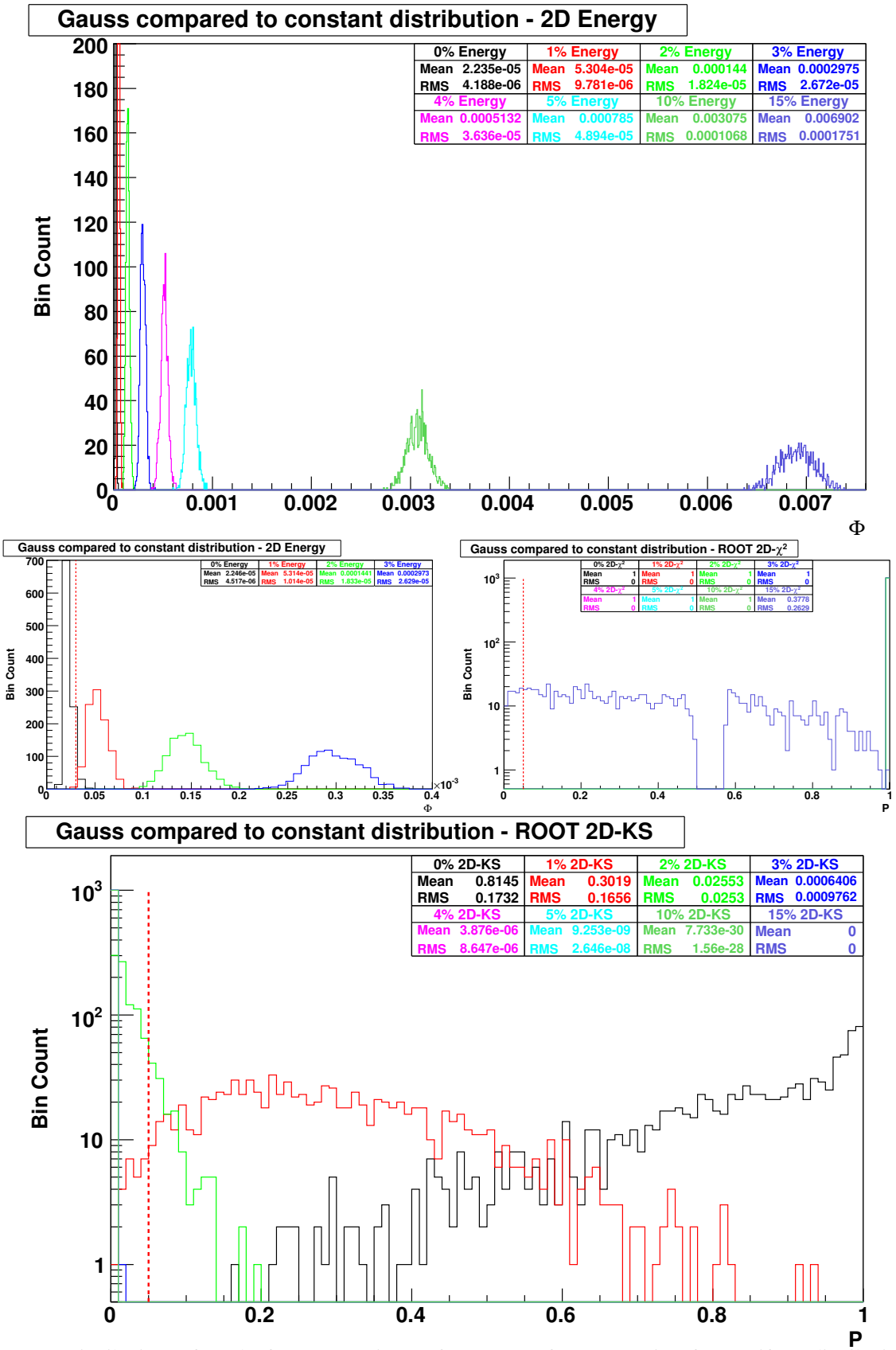


Figure 8: Distributions of results from comparisons of 1000 sets of 100 000 points from uniform distributions in -3<x, y<3 with contamination from a rotationally-symmetric N(0,1) distribution at levels of 0, 1, 2, 3, 4, 5, 10, and 15 percent to a constant distribution, at 100x100 binning, for the histogrammed energy test (top, expanded middle left) and the ROOT 2D–χ^2 and 2D-KS tests (middle right, bottom). The vertical dashed lines give the 95% confidence level (at 3.017e-5) for the energy test and the 5% acceptance criteria for the KS and χ^2 tests.

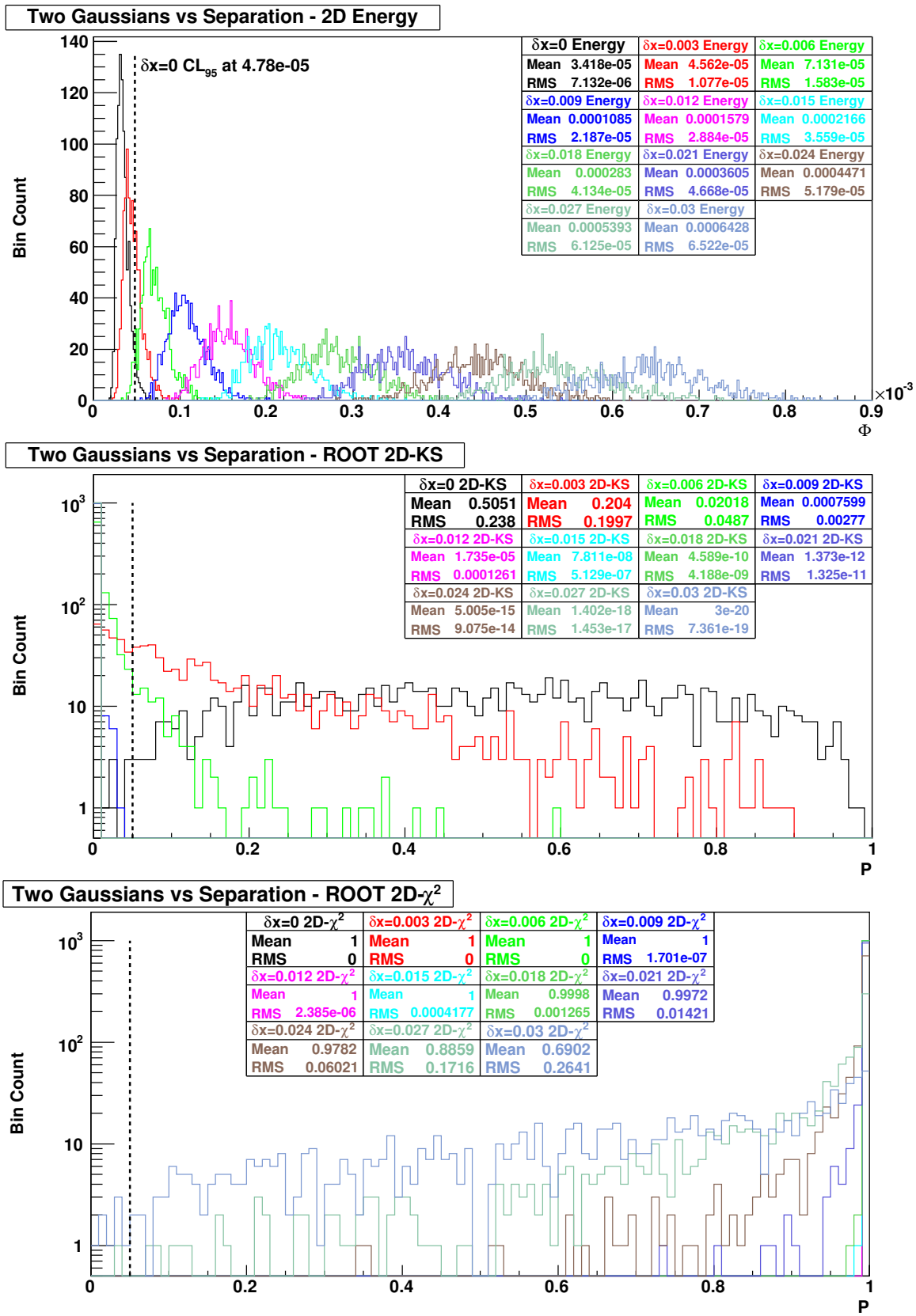


Figure 9: Results from comparisons of 1000 pairs of 100x100 histograms of 100 000 random points from rotationally symmetric N(0,1) distributions as a function of relative displacement δx: energy test (top); ROOT 2D-KS test (middle); and ROOT 2D-χ^2 test (bottom). Vertical dashed lines give the CL_{95} for δx=0 (4.78e-5) for the energy test and the 5% acceptance criteria for the ROOT tests.

Gaussian contamination	Energy Test power	ROOT 2D-KS power	ROOT 2D-χ^2 power
0%	0.049	0.0	0.0
1%	0.996	0.024	0.0
2%	1.0	0.867	0.0
3%	1.0	1.0	0.0
4%	1.0	1.0	0.0
5%	1.0	1.0	0.0
10%	1.0	1.0	0.0
15%	1.0	1.0	0.079

Table 5: 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 N(0,1) distribution contamination in a uniform distribution in -3$<x, y<$3 against a constant reference (see text), calculated from the result distributions and the selection criteria shown in Figure 8.

Centroid separation	Energy Test power	ROOT 2D-KS power	ROOT 2D-χ^2 power
0.000	0.05	0.01	0.0
0.003	0.364	0.246	0.0
0.006	0.960	0.902	0.0
0.009	1.0	1.0	0.0
0.012	1.0	1.0	0.0
0.015	1.0	1.0	0.0
0.018	1.0	1.0	0.0
0.021	1.0	1.0	0.0
0.024	1.0	1.0	0.0
0.027	1.0	1.0	0.001
0.030	1.0	1.0	0.008

Table 6: The discrimination power of the histogrammed energy test and the ROOT 2D-KS and 2D-χ^2 tests comparing increasingly separated N(0,1) distributions over -3$<x, y<$3 (see text), calculated from the distributions and selection criteria shown in Figure 9.

For the histogrammed energy test the confidence level was taken as the 95th percentile of the δx=0 comparison (i.e., comparison between samples from the same distribution); acceptance criteria for the ROOT 2D-KS and 2D-χ^2 tests were 5%. Distributions of the results are shown in Figure 9 and calculated powers in Table 6. The energy and ROOT 2D-KS tests show similar performance, both having high rejections at δx=0.006 (1/10th of a bin width) and above. The ROOT 2D-χ^2 test, however, shows essentially no rejection across the range of separations studied.

3.3.4 Shape sensitivity

To investigate the sensitivity of the tests to changes in the shapes of distributions, 1000 10^5-point rotationally-symmetric N(0,1) distributions as in Section 3.3.3 were each compared to another N(0,σ^2) distribution where σ took values from 0.95 to 1.05. Outliers were considered in both the energy tests and the ROOT tests.

The confidence level for the histogrammed energy test was set at the 95th percentile of the σ=1 comparison, a value of 4.96e-5. This level should match that used in Section 3.3.3 (4.78e-5) as the comparisons were made between samples from the N(0,1) distribution in each case; the discrepancy gives an indication of the repeatability of the statistic distribution. The acceptance criteria for the ROOT 2D-KS and 2D-χ^2 tests were 5%. The distributions of the results are shown in Figure 10 and the calculated powers in Table 7. Here the histogrammed energy test again performs slightly better than the ROOT 2D-KS test, providing high discrimination power for $|1-\sigma|\geq$0.02 while the KS test only shows high power at $|1-\sigma|\geq$0.03. The ROOT 2D-χ^2 test again showed poor rejection power as it returned the result P=1.0 for all comparisons. Further trials showed that this test would only start rejecting the N(0,σ^2) distributions when $|1-\sigma|$ exceeded approximately 0.15.

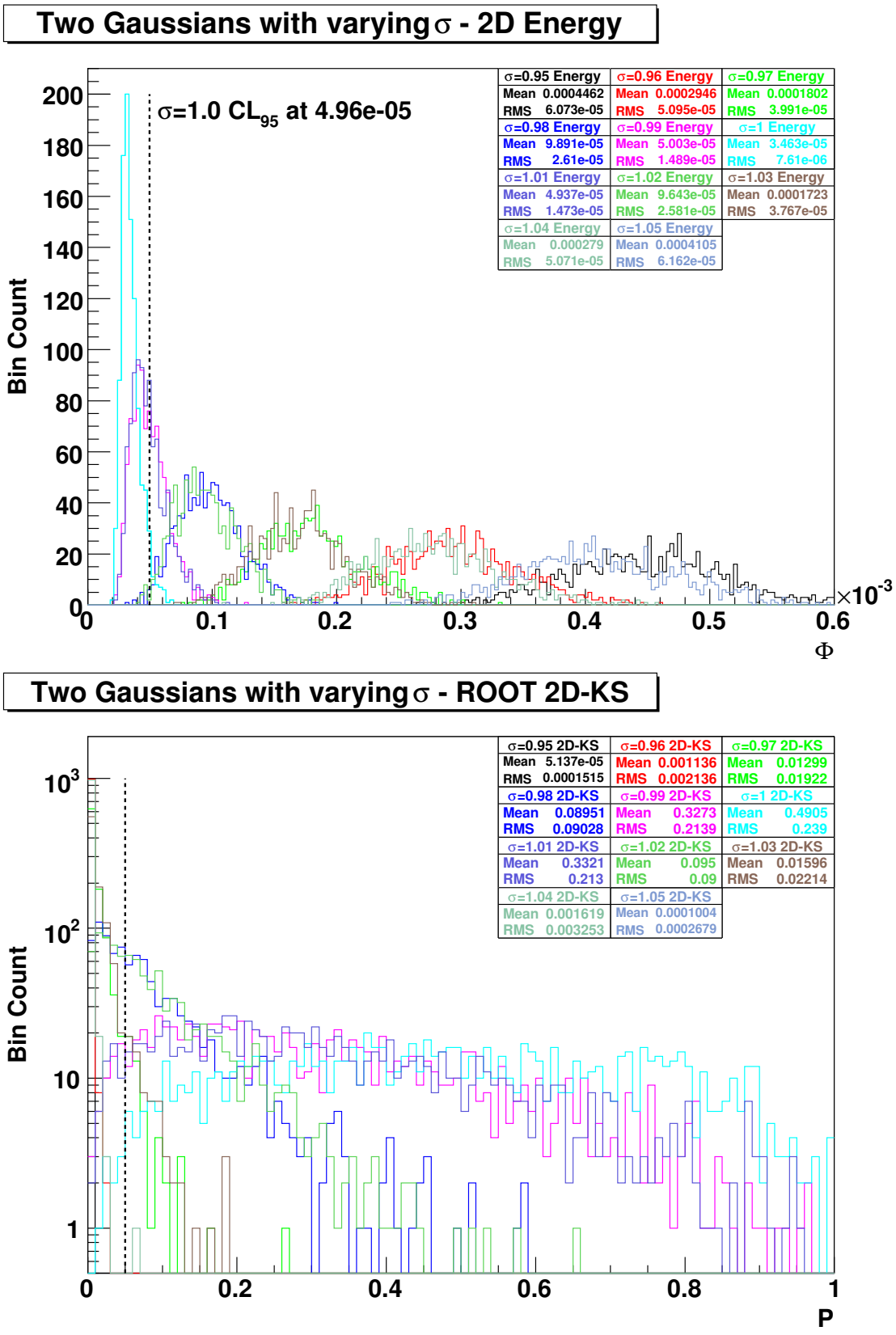


Figure 10: Distributions of results from comparisons of 1000 pairs of sets of 100 000 random points from N(0,1) and N(0,σ^2) distributions as a function of σ, at 100x100 binning, for the histogrammed energy test (top) and the ROOT 2D-KS test (bottom). The vertical dashed lines give the 95% confidence level derived from the σ=1.0 distribution (4.96e-5) for the energy test and the 5% acceptance criterion for the KS test. The distributions of results from the ROOT 2D-χ^2 test are not shown as it returned unity for all comparisons within this range of σ.

Distribution σ	Energy Test power	ROOT 2D-KS power	ROOT 2D-χ^2 power
0.95	1.0	1.0	0.0
0.96	1.0	1.0	0.0
0.97	1.0	0.95	0.0
0.98	0.988	0.425	0.0
0.99	0.456	0.05	0.0
1.0	0.05	0.009	0.0
1.01	0.418	0.048	0.0
1.02	0.987	0.391	0.0
1.03	1.0	0.932	0.0
1.04	1.0	0.999	0.0
1.05	1.0	1.0	0.0

Table 7: The discrimination power of the histogrammed energy test and the ROOT 2D-KS and 2D-χ^2 tests for comparisons of N(0,1) and N(0,σ^2) distributions (see text), calculated from the distributions and selection criteria shown in Figure 10.

4 Results with CMS Reconstructed Simulated Track Data

The simulated data introduced in Section 3 were further investigated to ascertain how the histogrammed energy test might perform in real data-monitoring situations.

4.1 Detection of contaminated data

Guided by the region of validity in the result matrix of Table 3, 80 000-sample 100x100 histograms were used to test the limit of detection of contamination in the data due to a change in detector alignment during the course of an experiment. The sample size ensures that each data point has at most one chance in eight of being selected in a given histogram, while the binning is a compromise between retaining fine structure and limiting calculation times. Reference samples consisted of selections drawn from the aligned data set while test samples were made up of $n\%$ drawn from the misaligned data with the balance drawn from the aligned data; all selections for a given comparison were made without replacement. Comparisons were made using the histogrammed energy test and the ROOT 2D-KS and 2D-χ^2 tests between the reference and test samples, repeated 10 000 times for each value of n.

The distributions of results obtained from the three tests are shown in Figure 11. The power of the tests is given in Table 8, using the 95th percentile of the non-contaminated sample tests (7.39e-5) for the detection criterion for the energy tests, and the 5% acceptance level for the KS and χ^2 tests. Here the 2D-KS test has slightly better performance than the histogrammed energy test for samples with 15% contamination, but both tests perform almost ideally at 20% contamination. The 2D-χ^2 test failed to reject any samples until contamination exceeded 30%.

4.2 Early detection of flawed data

Further experiments were performed to determine when a sample of flawed data was large enough for a comparison to clearly detect it as being different from the reference. The whole aligned data set was taken as the reference histogram, at 100x100 binning, and compared to 10 000 randomly-drawn histograms of varying sample size from

Contamination (%)	Energy Test power	ROOT 2D-KS power	ROOT 2D-χ^2 power
0	0.05	0.0102	0.0
5	0.07	0.0614	0.0
10	0.178	0.4038	0.0
15	0.5948	0.8992	0.0
20	0.9864	0.9981	0.0
25	1.0	1.0	0.0
30	1.0	1.0	0.0
35	1.0	1.0	0.0043

Table 8: The discrimination power of the histogrammed energy test and the ROOT 2D-KS and 2D-χ^2 tests for comparisons of aligned data histograms to histograms with $n\%$ contamination of misaligned data, calculated from the distributions and selection criteria shown in Figure 11.

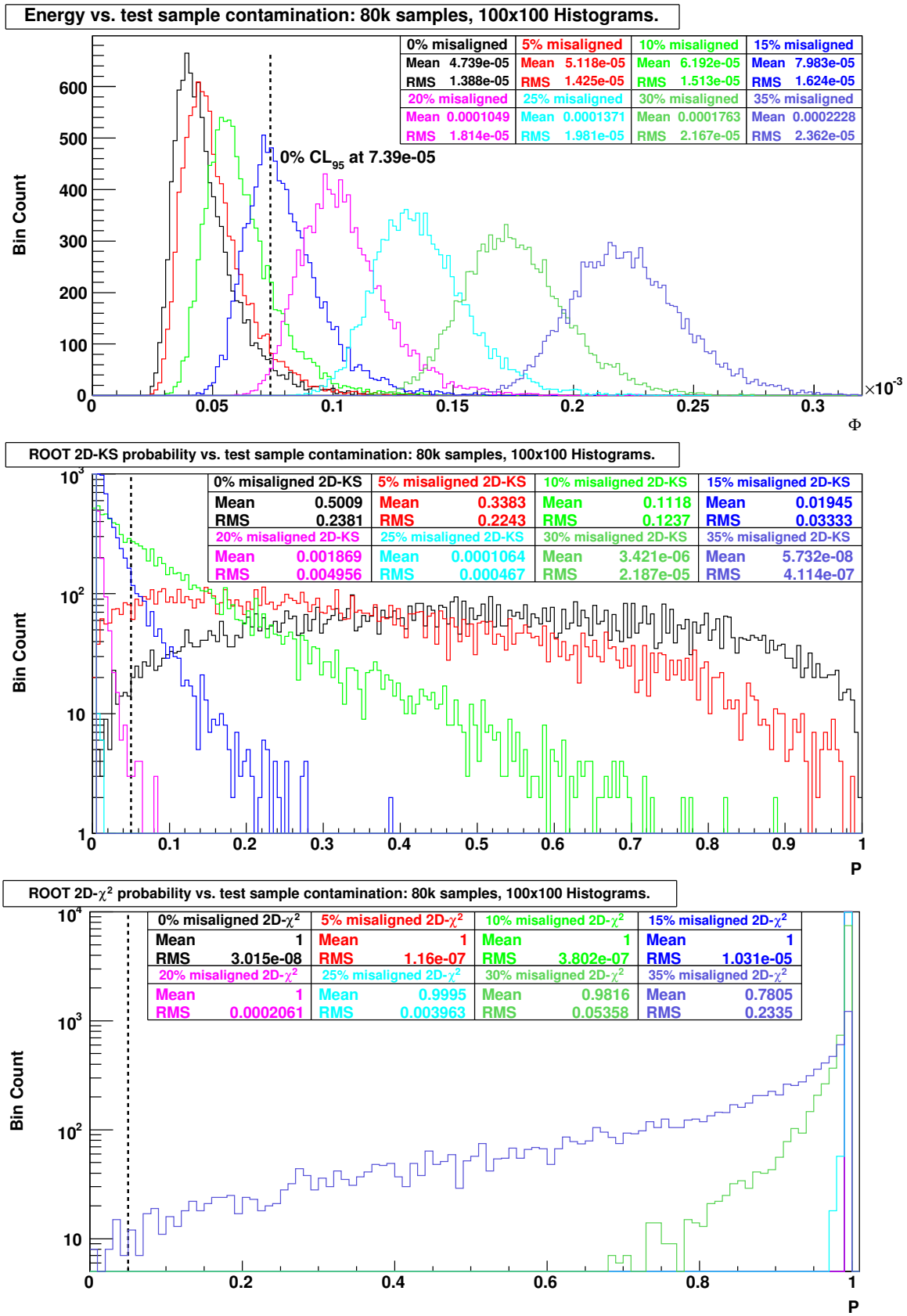


Figure 11: The distribution of histogrammed energy test (top) and ROOT 2D-KS (middle) and 2D-χ^2 test (bottom) results from comparisons between 10 000 pairs of 100x100 histograms of 80 000 points each selected randomly from the aligned data, where the second histogram is contaminated with $n\%$ of its points selected from the misaligned data. The vertical dashed lines give the 95% confidence level calculated from the uncontaminated distribution (7.39e-5) for the energy tests and the 5% acceptance criterion for the ROOT 2D tests.

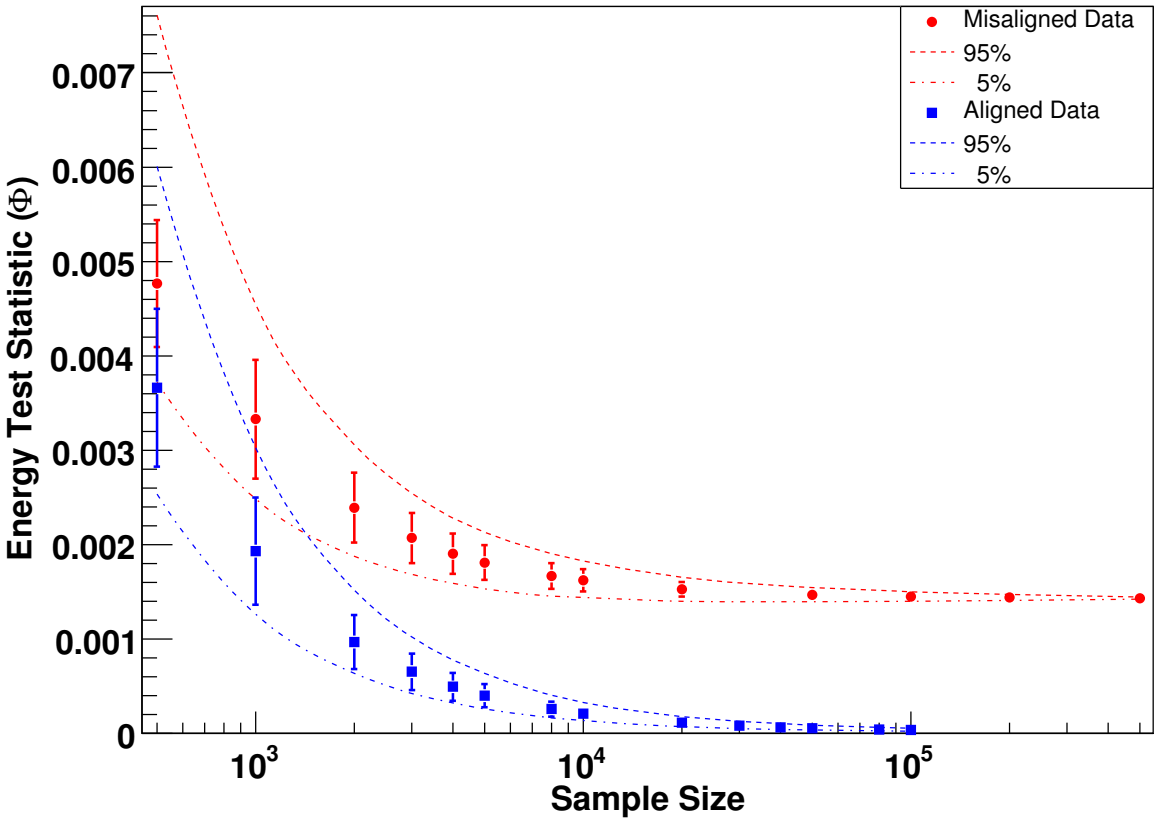


Figure 12: Variation in the histogrammed energy test results with sample size when compared to a large reference sample. The "misaligned" results compare 10 000 samples drawn from data obtained using misaligned detectors to the full data set from aligned detectors; the "aligned" results compare 10 000 samples drawn from one half of the aligned data set to all the other half. Shown are the average and r.m.s. deviation of the resultant distributions, as well as the 5th and 95th percentiles.

Sample Size	Energy Test aligned CL_{95}	ROOT 2D-KS aligned power	Energy Test misaligned power	ROOT 2D-KS misaligned power
500	0.006008	0.0129	0.2194	0.5509
1 000	0.003027	0.0103	0.6531	0.9501
2 000	0.001519	0.0124	0.9996	1.0
3 000	0.001021	0.0112	1.0	1.0
4 000	0.000780	0.0143	1.0	1.0
5 000	0.000636	0.0156	1.0	1.0
8 000	0.000407	0.0159	1.0	1.0
10 000	0.000328	0.0185	1.0	1.0

Table 9: The discrimination power of the histogrammed energy test and the ROOT 2D-KS test for comparisons of reference aligned data histograms to 10 000 histograms of smaller sample sizes, both aligned and misaligned (see text), calculated from the distributions and selection criteria shown in Figures 13 and 14.

the misaligned data set. As a control, similar experiments used half the aligned data set as reference, compared to histograms drawn randomly from the other half. All sampling was done without replacement.

Results from the three tests are shown in Figures 12–14 and summarised in Table 9. The 95th percentile confidence limits for the histogrammed energy test were determined from the comparisons of aligned data samples to aligned reference data, then used to determine the power of the tests with misaligned samples as shown in Table 9. For the KS and χ^2 tests, the 5% acceptance level determined the power. In this case, the ROOT 2D-KS test proved slightly superior to the histogrammed energy test, rejecting 95% of the 1000-sample histograms of misaligned data, where the energy test only rejected 65%. However, both tests performed well on 2000-sample histograms of misaligned data, rejecting 100% and 99.96% of the histograms, respectively. In contrast, the ROOT 2D-χ^2 test, whilst rejecting all the misaligned samples with a statistic of 0.0, also rejected aligned samples for samples with fewer than 50 000 points. The third column of Table 9 shows that at the 5% acceptance level the ROOT 2D-KS

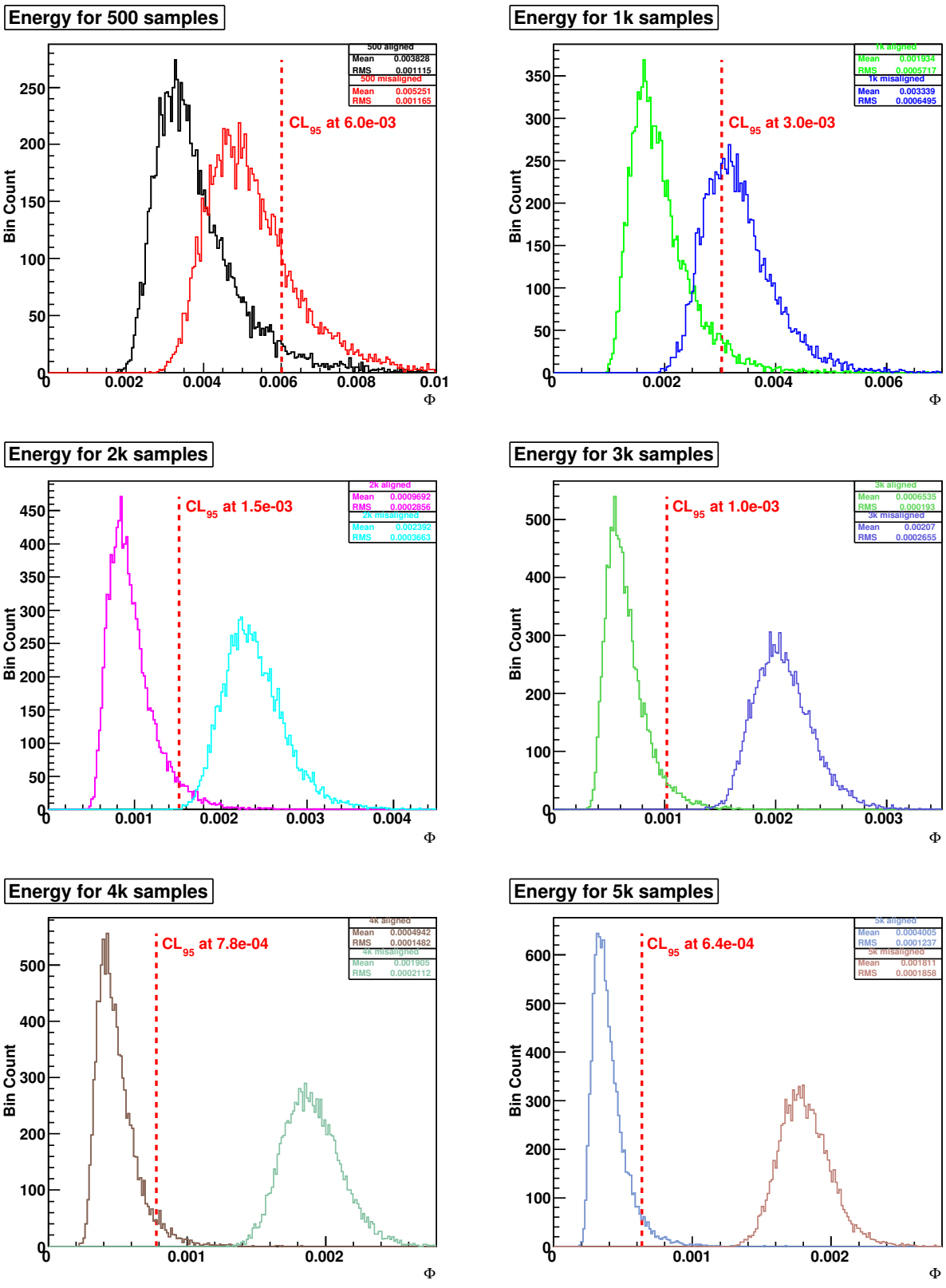


Figure 13: Distributions of the histogrammed energy test results with sample size when compared to a large reference sample. The "misaligned" results are from 10 000 comparisons of different-sized selections from tracks reconstructed using misaligned detectors to the full data set of tracks reconstructed with aligned detectors; the "aligned" results are from comparisons of half the aligned data set to 10 000 samples drawn from the other half. The dashed vertical lines show the 95th percentiles of the aligned distributions, used as confidence levels to determine the discrimination power of the energy test as a function of sample size (see Table 9).

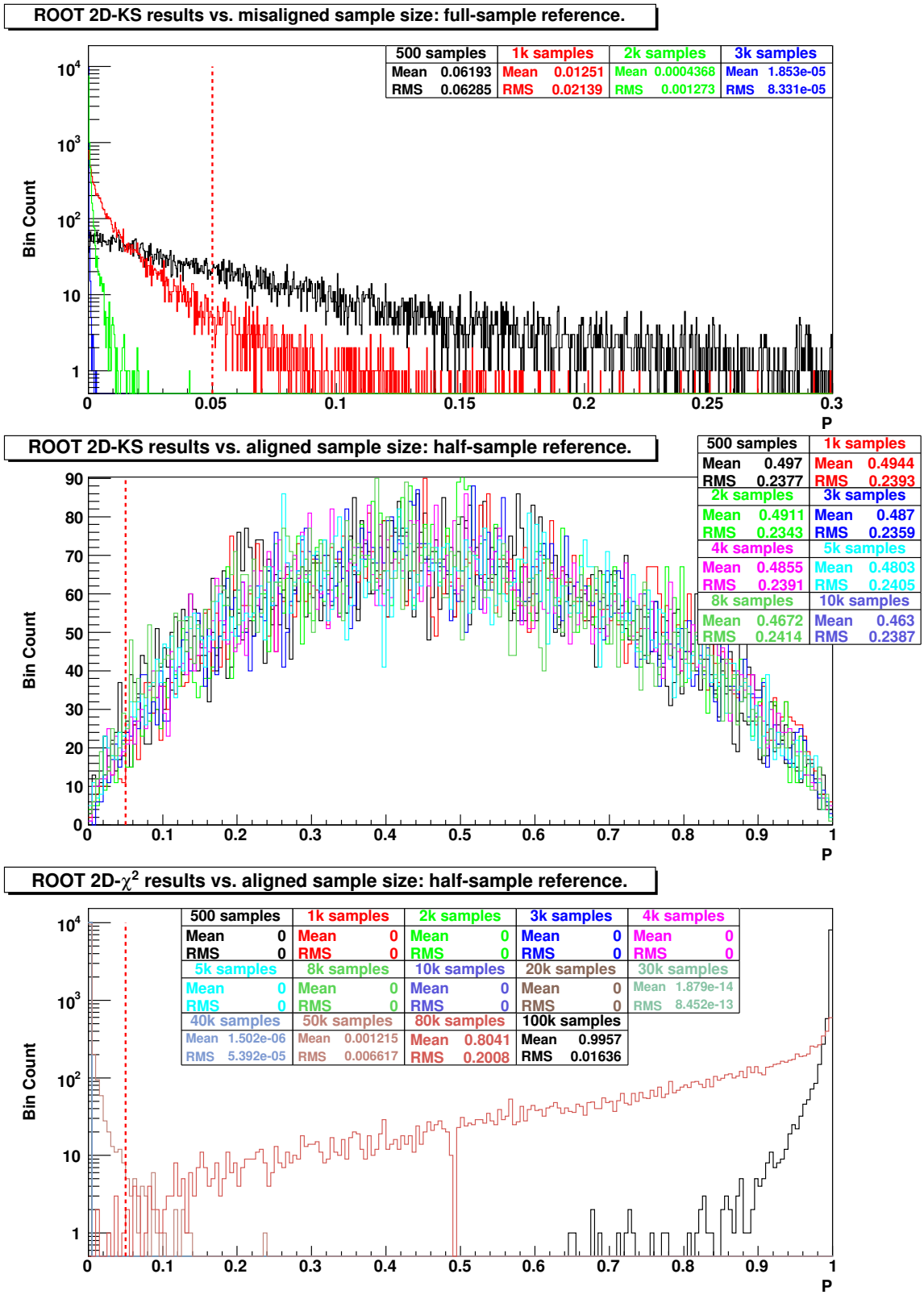


Figure 14: Distributions of the ROOT 2D-KS and 2D-χ^2 test results with sample size when compared to a large reference sample. The "misaligned" 2D-KS test results (top) are from 10 000 comparisons of different-sized selections from tracks reconstructed using misaligned detectors to the full data set of tracks reconstructed with aligned detectors. Results from samples larger than 3000 points are not shown, nor are the 2D-χ^2 test distributions, as these all returned a value of zero. The "aligned" results (lower plots) are from comparisons of half the aligned data set to 10 000 samples drawn from the other half. The dashed vertical lines show the 5% acceptance level.

test rejected only $\sim$1% of histograms of aligned data for small sample sizes, reinforcing the observation made in Section 1.1. The distortion of the probability distribution away from a uniform distribution when comparing identical 2D histograms with the ROOT 2D-KS test is also clearly illustrated in Figure 14.

5 Implementation details

Given that the histogrammed energy test has been shown to be a worthwhile addition to the methods for comparing 2D histograms, work can now proceed on providing a release version to the community. Modifications to the current implementation for general use are, for the most part, small and easily added:

- More extensive and robust error-handling. In particular, the tests used to check the suitability of the input histogrammes to the ROOT 2D-KS test can be adopted for the histogrammed energy test. *[Implemented]*
- Removal of the restriction that both axes of the histograms have the same binning. This will require the extension of the square line picking result [16] to rectangular regions.
- Allowing the inclusion of overflows and/or underflows to be specified by an option. *[Implemented]*
- Allowing the inclusion of the artificial offset (Appendix C) to be controlled by an options flag. Since the shift can be precisely calculated outside the loops which accumulate the energy sums, it can be accommodated with negligible penalty. *[Implemented]*
- Where the test is incorporated into a histogram class (as with the two ROOT tests) rather than used as a standalone test, further efficiencies can be gained by being able to store private copies of such data as the array of the inter-bin potential function $R(r)$. The self-potential sum Φ_A also needs only be calculated once, saving approximately one-quarter of the time for subsequent calculations when one histogram is used as a reference for many comparisons with other data. In this case the class must have a robust method of invalidating the local data whenever a change is made to the histogram data, and also when the test method is applied with changed options.

6 Conclusions

We have presented our investigations into a new test for comparing 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, but performed slighly below it on tests with reconstructed track data. It is more consistent than the ROOT 2D-KS, returning similar statistics across a range of histogram parameters, whereas the ROOT test can return quite different results depending on how the data are binned.

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.

While far-ranging, our investigation has been by no means exhaustive and we encourage members of the community to evaluate the new test on data sets of interest to them.

Acknowledgments

The authors would like to thank the CMS Tracker community and CMSSW developers for providing the tools used to produce the data sets. We also thank Israel Goitom for useful feedback and discussions, and the referees for many helpful suggestions. This work has been funded by the Science and Technology Facilities Council, UK.

Appendix A The ROOT 2D Chi2Test

One obvious finding of the comparisons carried out in this study is that the ROOT 2D-χ^2 method is unsuitable for the type of comparisons being considered. There are a number of reasons for this, the most important being the influence of regions of low density in the histograms. The sum which is accumulated to give the χ^2 statistic when comparing two histograms is [5]

$$\frac{1}{nm}\sum_{i=1}^{r}\frac{(mn_i - nm_i)^2}{n_i + m_i} \tag{4}$$

where the contents of the ith bin (of r) are given as n_i and m_i respectively for the two histograms, and n and m are the respective sums of the events in the histograms. The number of degrees of freedom ν is taken as one less than the number of histogram bins (i.e., r-1). ν is reduced by one for every bin which has zero content in *either* of the histograms being compared, while the summation term for that bin is ignored. The probability which is then returned is the function `TMath::Prob`(χ^2, ν) [21] which changes very rapidly between 1.0 and 0.0 in the vicinity of ν; e.g., for $\nu = 9999$ (100x100 histograms) the function is 0.95 at $\chi^2 = 9769.0$ ($\chi^2/\nu = 0.977$) and 0.05 at $\chi^2 = 10232.7$ ($\chi^2/\nu = 1.023$).

The reason for skipping bins with zero contents is obviously to avoid division by zero when both n_i and m_i are zero. However, when only *one* of the two bins is zero, the contents of the other one are still significant and should be included, so the algorithm has two competing effects in this case – the unnecessary reduction of ν and the reduction of the accumulating χ^2 sum. For (say) $m_i = 0$, the missing term in the accumulation is mn_i/n or, when $n \sim m$, approximately n_i. On the assumption that this is the appropriate term to include in the summation then, since $n_i \geq 1$, the true value of χ^2/ν will be changed to $(\chi^2 - p - \delta)/(\nu - p)$, where p is the number of skipped bins and $\delta \geq 0$.

It is a simple matter to include these terms in the summation, by changing the coding of the Chi2Test algorithm so that a logical OR (||) in the C++ code which decides when to skip a bin and decrement ν [22] is replaced by a logical AND (&&).

```
if (bin1 == 0 && bin2 == 0) {
   --ndf;  //no data means one degree of freedom less
} else {...}
```

When this is done, the results of the 2D-χ^2 test fall more into line with expectations. Table 10 shows the results of comparing the data sets given in Figure 2, with the addition of two identical points in each set to prevent ν being decremented to -1, at various binnings using the ROOT 2D-KS test, the ROOT 2D-χ^2 test, the modified ROOT 2D-χ^2 test, and the histogrammed energy test. The results of the ROOT 2D-KS test follow expectations; as the binning becomes finer and the two CDFs approach those for each coordinate separately, the reported probability increases to unity. For the ROOT 2D-χ^2 test, only the two bins containing the identical points are retained, all others being discarded because they are empty in one or other of the histograms. Because the remaining bins have identical contents, the accumulated sum is zero, and because all but two of the bins are discarded, ν is reduced to 1; therefore the probability function returns unity. For the modified 2D-χ^2 test, ν is only reduced to the number of bins which are non-zero in *either* histogram, less 1, while the accumulated sum becomes $\sum_i |n_i - m_i|$ (see above) or 4000 in this case. Only at finer binnings, where the probability that each bin contains at most one data point approaches unity, does ν begin to approach χ^2 and the modified test return a significant probability. The histogrammed energy test, on the other hand, returns an almost-constant result over the whole range of binning.

Table 11 compares the ROOT 2D-χ^2 results from Table 1 with those obtained from the modified test. The values of ν for the modified tests are significantly higher than the original test, and the accumulated sums even more so. Consequently the modified test returns a zero probability for all the comparisons, except at a binning of 1000x1000, where the increase in ν as the number of populated bins increases leads to a probability of 2.6% being returned.

In fairness, the ROOT documentation [4] and Gagunashvili [5] warn about using the χ^2 test for histograms with low occupation values. A recommendation attributed to Lewontin and Felsenstein [23] is that all expectations should be equal to or greater than unity for both histograms, although it is pointed out that this limit can be relaxed to 0.5 [5]. For the cases which have been considered, this criterion should have been met for the Cook-Johnson distributions of Section 3.3.1 and the Gaussian-contaminated uniform distributions of Section 3.3.2, so the performance of the ROOT 2D-χ^2 test in these comparisons cannot be attributed to mis-application (nor to the apparent shortcoming discussed above). However, it should be noted that when using a χ^2 test it is often recommended to use bins of equal probability rather than bins of equal size, as this tends to yield more power against general alternatives.

Histogram Size	ROOT 2D-KS P	ROOT 2D-χ^2 ν	ROOT 2D-χ^2 χ^2	ROOT 2D-χ^2 P	Modified 2D-χ^2 ν	Modified 2D-χ^2 χ^2	Modified 2D-χ^2 P	Energy Test
20x20	0.0	1	0.0	1.0	157	4000	0.0	0.7329
25x25	0.0	1	0.0	1.0	229	4000	0.0	0.7507
50x50	0.0	1	0.0	1.0	633	4000	0.0	0.7680
100x100	0.0231	1	0.0	1.0	1509	4000	0.0	0.7782
200x200	0.5602	1	0.0	1.0	2833	4000	0.0	0.7824
500x500	0.9977	1	0.0	1.0	3745	4000	0.0019	0.7835
1000x1000	1.0	1	0.0	1.0	3941	4000	0.2518	0.7841

Table 10: Comparisons of the two mirror-image data sets of Figure 2, with the addition points at (0.0,0.0) and (0.99,0.99) in each set, at different binning levels for the two ROOT 2D comparison tests, the modified ROOT χ^2 test, and the histogrammed energy test.

Histogram Size	ROOT 2D-χ^2 ν	ROOT 2D-χ^2 χ^2	ROOT 2D-χ^2 P	Modified 2D-χ^2 ν	Modified 2D-χ^2 χ^2	Modified 2D-χ^2 P
10x10	97	537.373	0.0	98	539.380	0.0
20x20	363	1596.58	0.0	382	1627.60	0.0
25x25	550	2253.85	0.0	1768	3271.09	0.0
50x50	1768	3271.09	0.0	2138	4045.38	0.0
100x100	4426	5226.93	0.0	6763	8920.67	0.0
200x200	9681	8422.96	1.0	17910	19828.7	0.0
500x500	22774	14954.6	1.0	53554	55243.4	0.0
1000x1000	28820	10549.8	1.0	104854	105746	0.025982

Table 11: Comparisons of the original and modified ROOT 2D-χ^2 test for the data shown in Figure 1.

Unfortunately ROOT does not implement adaptive binning, but we have already performed explorative studies on an adaptive approach to the energy test by using clustering of the discrete input points rather than histogramming.

A further check on the modified 2D-χ^2 test was made by obtaining the distributions of the probabilities from a large number of tests comparing two samples from the same 2D distribution. The expectation is that the result distribution should be uniform between zero and unity. Samples drawn from a uniform distribution on the unit square were placed into 100x100 histograms. 10 000 comparisons were made with both 2D-χ^2 methods for samples of 500, 1000, 2000, 5000, 10 000, 20 000, 40 000, 60 000, 80 000, and 100 000 points (average bin populations of 0.05 to 10). The result distributions are shown in Figure 15. As expected, the ROOT 2D-χ^2 test only produced a uniform distribution at the highest occupancies, while the modified 2D-χ^2 test gave a uniform spread at lower occupancies, from about 2 per cell. Below that, however, its distributions are noticeably peaked around 0.5, suggesting that the modification may not be the ultimate solution to the problem of low populations in 2D-χ^2 tests.

Another minor problem with the ROOT χ^2 test can be seen in the statistic distributions in Figures 8, 9, 14, and 15 where gaps appear in the probability distributions around 0.5. This was traced to the `TMath::Gamma`(ν/2,χ^2/2) function [24] used by `TMath::Prob`. `Gamma`(a,x) calls two separate implementations, a series expansion for $x < a+1$ and a continued fraction method otherwise. A shortcoming in the series expansion leads to it diverging from the true function value near the crossover point and thus a step change in the `Prob` function close to 0.5.

We understand that both these problems with Chi2Test will be addressed in future releases of ROOT [6)], including consideration of the proper term to include in the χ^2 sum when only one bin is zero.

Appendix B The Generation of CDFs by the ROOT 2D-KS Test

As discussed in Section 1.1, the ROOT 2D-KS test generates two CDFs for both of the histograms being compared by accumulating the histogrammed data bins rasterwise, in column- and row-major fashion respectively. This process is illustrated for various binnings of the first data set of Section 1.1 in Figure 16. As the binning becomes finer, the excursions in the CDFs become smaller and the curves approach the CDF for the discrete data, ordered in the appropriate dimension. This is the reason why the comparison of the two distributions in Figure 2 returns an incorrect probability, as the comparison essentially ignores the connectivity of the data, registering only its distribution in each dimension (see also Table 10).

[6)] `https://savannah.cern.ch/bugs/?32884` and L. Moneta, private communication.

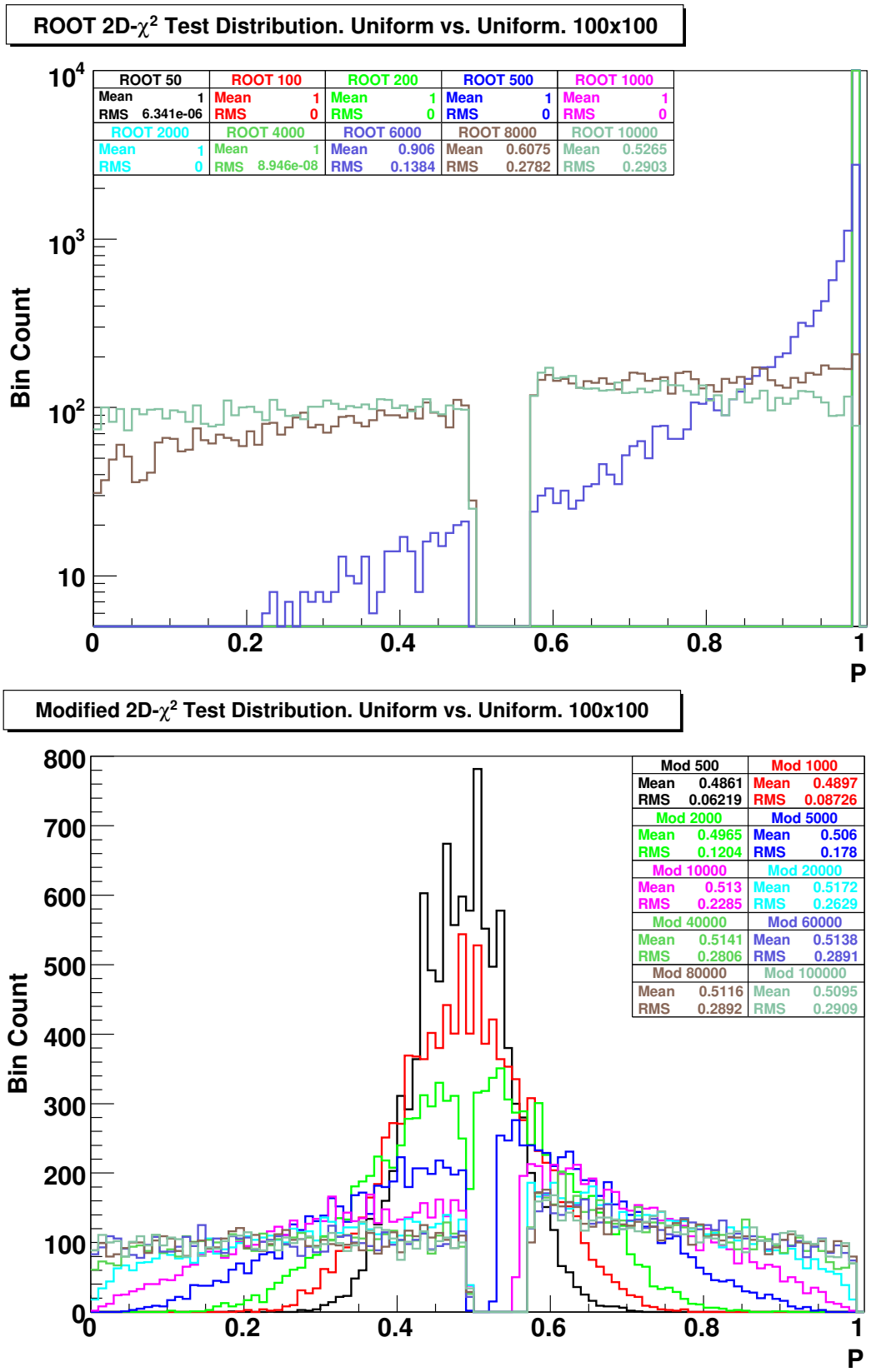


Figure 15: Results from comparisons of 10 000 pairs of 100x100 histograms of random points from uniform distributions in the unit square as a function of the size of the samples for the standard ROOT 2D-χ^2 test (left) and the modified test (right). Sample sizes were 500, 1000, 2000, 5000, 10 000, 20 000, 40 000, 60 000, 80 000, and 100 000 points.

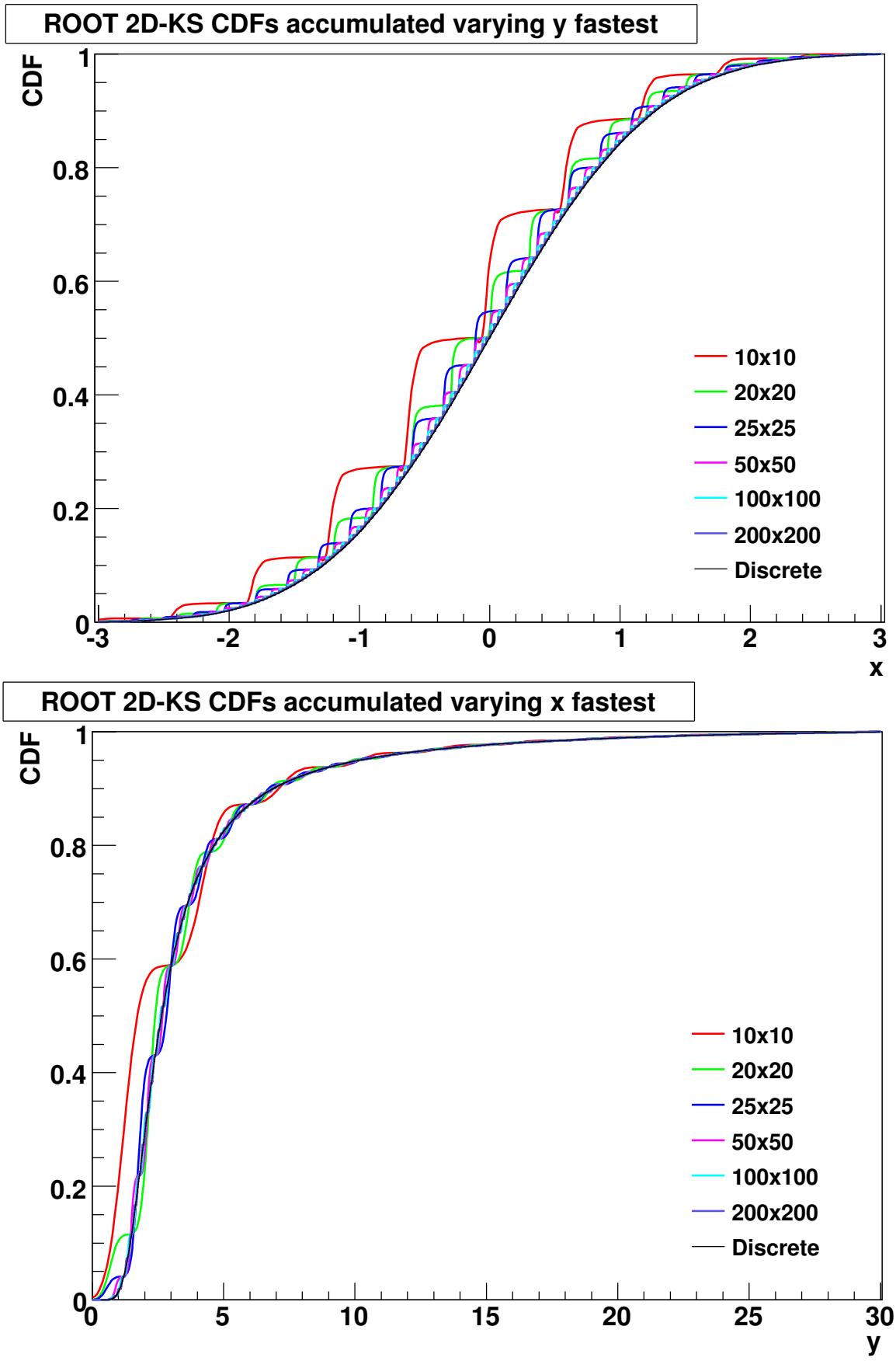


Figure 16: The CDFs calculated by ROOT's 2D-KS test for a 100 000-point data set from Figure 1, excluding outliers, at several binnings: top, column-major ordering; bottom, row-major ordering. The CDFs for the discrete data ordered along the two dimensions are also given.

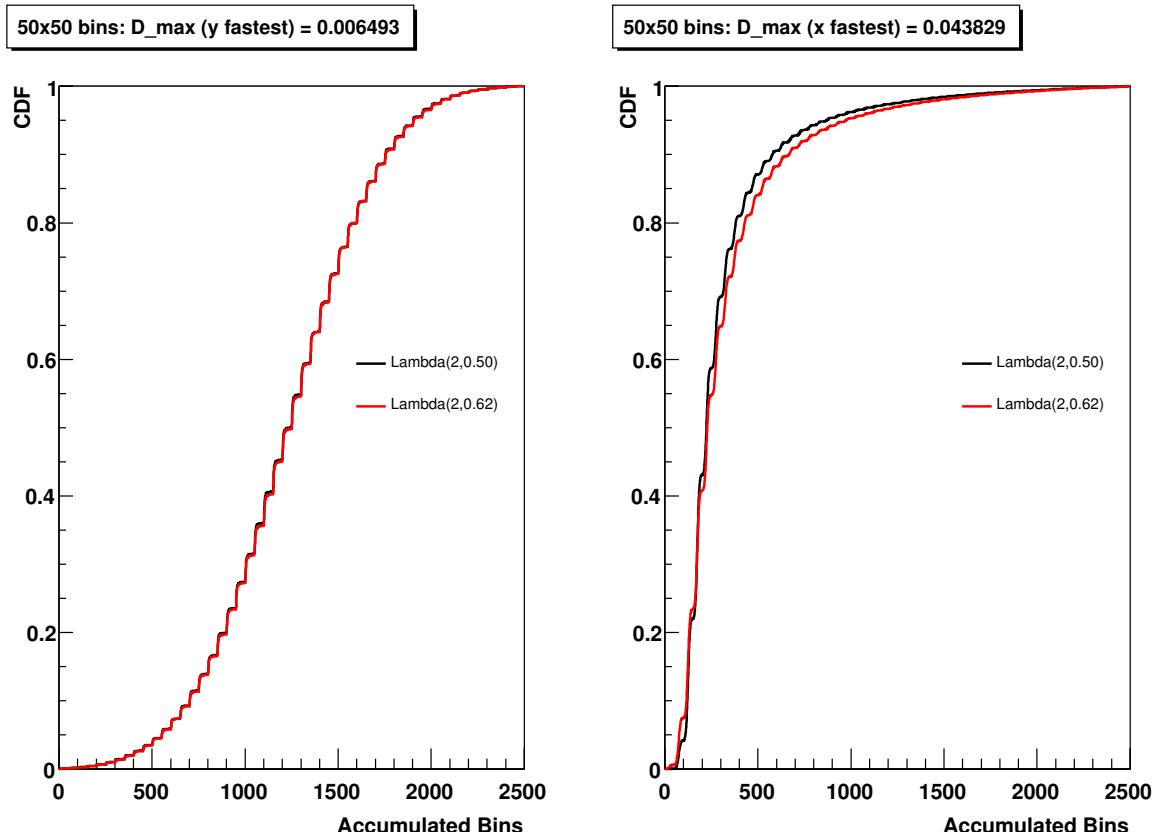


Figure 17: The two pairs of CDFs obtained by ROOT's 2D-KS test while comparing the two 100 000-point data sets, excluding outliers, binned at 50x50 (Figure 1): left, column-major ordering; right, row-major ordering. The maximum distances $D_{\max}$ between each pair are averaged and this distance used to calculate the probability P of the comparison. A customised version of the ROOT method was used in order to obtain data not normally reported.

Two Kolmogorov-Smirnov distances $D_{\max}$ are then calculated from the maximum distance between the appropriate CDFs from each histogram (see Figure 17). The average of these distances is then normalised by $\{(n_A * n_B)/(n_A + n_B)\}^{1/2}$, where n_A, n_B are the sums of the histogram contents, and the probability P is obtained by evaluating the Kolmogorov distribution function [7] at that value.

Appendix C Calculation of the Self-Energy Φ_A and Φ_B

As mentioned in Section 2.2, the implementation of the energy test used here makes a small variation from mathematical rigour in calculating the self-energies Φ_A and Φ_B in order that comparisons of identical histograms should return zero (within the limits of numerical calculation). The term for the interactions between the n_k points in the kth histogram bin are weighted as $n_k^2/2$ rather than the actual number of interactions $n_k(n_k - 1)/2$, in order to cancel with the corresponding term in Φ_{AB} which is weighted as $n_k m_k \rightarrow n_k^2$ when the histograms are identical. This means that $\Phi_{A,B}$ are increased by $-(n_k/2)ln(0.5214/N)$ for every bin so that the total energy sum is increased by -0.5$(1/n + 1/m)ln(0.5214/N)$ where n, m are the total number of points in each NxN histogram.

For example when n, m = 100 000 and N = 100, this amounts to 5.26e-5 which is similar to the 95% confidence levels established for comparisons of samples of that size from the same distributions (Sections 3.3.3 and 3.3.4), implying that the unshifted distributions will lie around zero and below. Since the shift in Φ_{nm} depends only on n, m, and N (which must be the same for both histograms), this should not present any great problem so long as n and m do not vary enough to significantly affect the factor $(1/n + 1/m)$ during any given set of comparisons.

The only comparison reported here which may be affected by changing offsets is the early-detection tests of Section 4.2, mainly because the size of the reference data set was different between the tests with misaligned samples (n = 630 093) and those with aligned samples (n = 315 046). In practice, however, the $1/m$ term dominates the offset (m =500. . .100 000) and in fact the smaller offset for the larger reference histogram slightly reduces the separation of the misaligned results from the aligned results. The results from Section 4.2 were recalculated without the artificial offset and are given in Table 12. The power of the histogrammed energy test without offset shows a very slight improvement in this case because of the separation effect noted above.

Sample Size	Aligned offset	Misaligned offset	With Offset		Without Offset	
			Aligned CL_{95}	Power	Aligned CL_{95}	Power
500	0.0052647	0.0052606	0.006008	0.2194	0.000744	0.2203
1 000	0.0026365	0.0026324	0.003027	0.6531	0.000390	0.6557
2 000	0.0013224	0.0013183	0.001519	0.9996	0.000196	0.9996
3 000	0.0008844	0.0008802	0.001021	1.0	0.000137	1.0
4 000	0.0006654	0.0006612	0.000780	1.0	0.000114	1.0
5 000	0.0005340	0.0005299	0.000636	1.0	0.000102	1.0

Table 12: The effect of removing the artificial offset from the histogrammed energy test. The results of Section 4.2 are compared to those obtained when the offset is removed.

Appendix D The Distribution of Energy Test Results

An important foundation-stone of statistics is the central limit theorem which states that, for a set of random independent and identically distributed variables $\{X_1,X_2,X_3,\ldots,X_n\}$, the distribution of the arithmetic mean $\sum_n X_i/n$ will be Gaussian as $n \to \infty$ whatever the distribution of X. A similar theorem states that the distribution of the maximum value of the set, $\max(\{X_1,X_2,X_3,\ldots,X_n\})$, will tend to the distribution known as the generalised extreme value (GEV) distribution [19, 25]. This distribution is given by

$$f(x) = \frac{1}{\sigma}\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}-1}\exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right\} \tag{5}$$

for $(1+\xi\frac{x-\mu}{\sigma}) > 0$, where μ is a location parameter, $\sigma > 0$ is a scale parameter, and ξ is a shape parameter.

Aslan and Zech found that the distributions of their energy test results closely followed such a form independently of the choice of the distance function $R(r)$, but did not find a means of generating the parameters from first principles [15, 18]. Because of the high speed of modern computers, it was recommended that distributions be determined empirically, with the possibility of reducing the number of samples needed by determining the GEV distribution parameters from a fit or from the first three moments of the distribution.

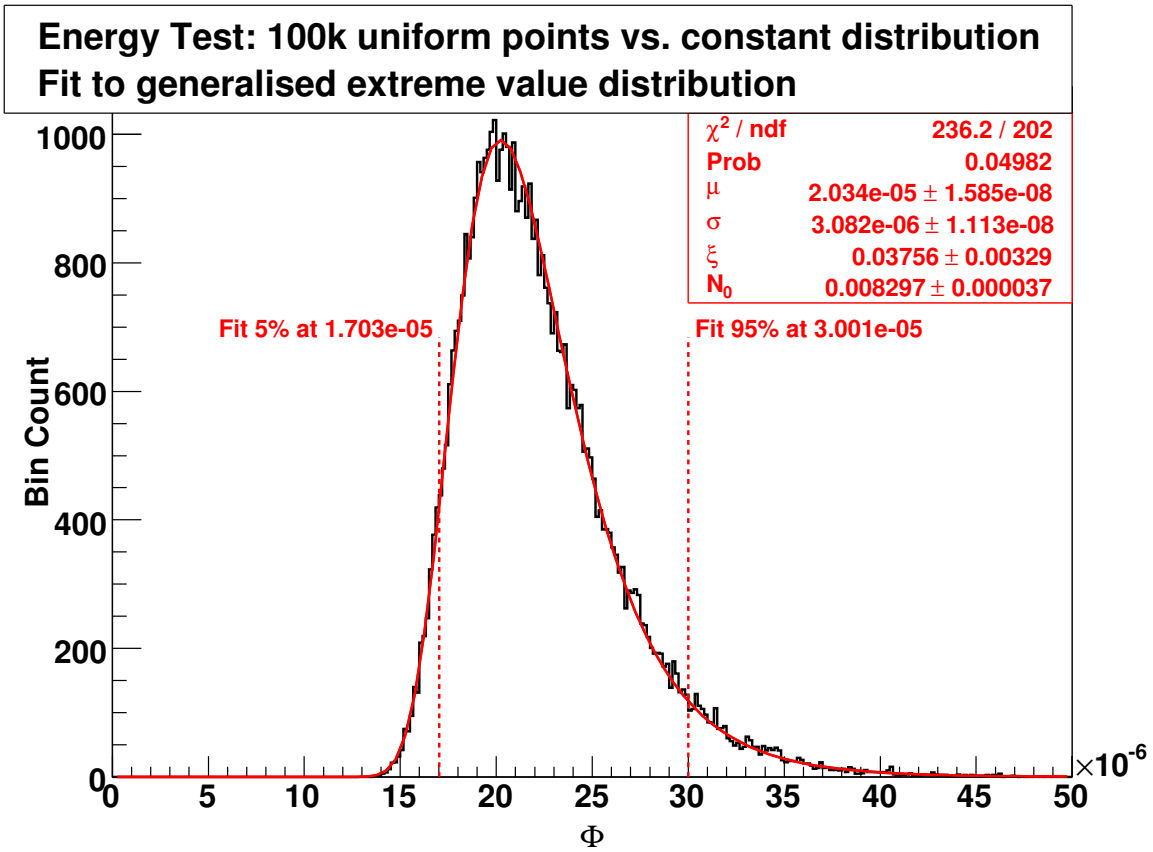


Figure 18: The distribution of results of the histogrammed energy test, comparing 50 000 sets of 100 000 randomly distributed points on the unit square to a constant distribution at 100x100 binning. The distribution is fitted to a GEV distribution (Equation 5) using ROOT. Also shown are the 5th and 95th percentiles as given by the fit.

To test whether the distribution of results from the histogrammed energy test also followed a GEV distribution, the reference distribution generated in Section 3.3 (Figure 4) was fitted to the GEV form using ROOT. The results are given in Figure 18 together with the 5% and 95% levels of the fitted distribution. A factor N_0 was included in the fit to normalise the integral of the distribution. This was known in advance from the number of samples and the histogram bin width to be 8.333e-3; the fitted value is 8.297$\pm$0.037e-5. It can be seen from the Figure that the GEV distribution does fit all portions of the histogram quite well, although the ROOT statistics are do not fully support this observation. It would appear from the value of 202 given for the degrees of freedom that the statistical analysis discounts bins with zero content. From the restriction given with Equation 5 it can be seen that the fit is valid for $\Phi > \mu - \sigma/\xi$, or $\Phi >$ - 6.17e-5. The Figure also shows the 5th and 95th percentiles calculated from the fit. These differ by just one bin-width from those derived directly from the histogram in Section 3.3.

From this it can be seen that, while the result distribution does appear to closely follow a GEV distribution, there is no immediate advantage to using this fact in deriving the levels of confidence in discriminatory tests.

References

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

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

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

[4] `http://root.cern.ch/root/html514/TH1.html#TH1:Chi2Test`

[5] N. Gagunashvili, *"χ^2 test for comparison of weighted and unweighted histograms"*, Statistical Problems in Particle Physics, Astrophysics and Cosmology, Proc. of PHYSTAT05, Oxford, UK, 12-15 September 2005 (London: Imperial College Press), 43–4 (2006).

N. Gagunashvili, *"Comparison of weighted and unweighted histograms"*, `arXiv:physics/0605123` (2006).

[6] `http://root.cern.ch/root/html514/TH1.html#TH1:KolmogorovTest`

[7] `http://root.cern.ch/root/html514/TMath.html#TMath:KolmogorovProb`

[8] `http://root.cern.ch/root/html514/TRandom.html#TRandom:Landau`

[9] CMS Collaboration, *CMS Physics Technical Design Report: Volume I, Detector Performance and Software*, ed. D. Acosta (2006).
`http://cmsdoc.cern.ch/cms/cpt/tdr/ptdr1_final_colour.pdf`

[10] `http://rpy.sourceforge.net`

[11] R.H.C. Lopes, P.R. Hobson and I.D. Reid, *"Computationally efficient algorithms for the two-dimensional Kolmogorov-Smirnov test"*, Int. Conf. on Computing in High Energy and Nuclear Physics, September 2–7 2007, Victoria BC; J. Phys.: Conf. Ser. **120**, 042019 (2008).
`http://bura.brunel.ac.uk/handle/2438/2571`

[12] G. Fasano and A. Franceschini, *"A multidimensional version of the Kolmogorov-Smirnov test"*, Mon. Not. R. Astron. Soc. **225**, 155–70 (1987).

[13] B. Aslan and G. Zech, *"A new class of binning-free, multivariate goodness-of-fit tests: the energy tests"*, `arXiv:hep-exp/0203010` (2003).

[14] G. Zech and B. Aslan, *"A new test for the multivariate two-sample problem based on the concept of minimum energy"*, `arXiv:math/0309164` (2003).

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

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

http://www.research.att.com/∼njas/sequences/A091505

[17] L. Barbone, N. De Filippis, O. Buchmueller, F.P. Schilling, T. Speer and P. Vanlaer, *"Impact of CMS silicon tracker misalignment on track and vertex reconstruction"*, Nucl. Instr. and Meth. A **566**, 45–9 (2006).

[18] B. Aslan and G. Zech, *"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 (2005).

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

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

[21] http://root.cern.ch/root/html514/TMath.html#TMath:Prob

[22] http://root.cern.ch/root/html514/src/TH1.cxx.html#ImYxrE

[23] R.C. Lewontin and J. Felsenstein, *"The robustness of homogeneity test in* $2 \times N$ *tables"*, Biometrics **21**, 19–33 (1965).

[24] http://root.cern.ch/root/html514/TMath.html#TMath:Gamma

[25] H.W. Park and H. Sohn, *"Parameter estimation of the generalized extreme value distribution for structural health monitoring"*, Probabilistic Eng. Mech. **21**, 366-376 (2006).