A Proposed High Dimensional Kolmogorov-Smirnov Distance

Alex Hagen[1], Jan Strube[1], Isabel Haide[2], James Kahn[2], Shane Jackson[1], and Connor Hainje[1]

[1]Pacific Northwest National Laboratory, Richland, WA, USA
[2]Karlsruhe Institute of Technology, Karlsruhe, DE

Abstract

We present a d-dimensional test statistic inspired directly by the Kolmogorov–Smirnov (KS) test statistic and Press' extension of the KS test to two dimensions. We call this the ddKS statistic. To preclude the high computational cost associated with working in higher dimensions, we present an implementation using tensor primitives. This allows parallel computation on CPU or GPU. We explore the behavior of the test statistic in comparing two three-dimensional samples, and use a standard statistical method - the permutation method - to explore its significance. We show that, while the Kullback–Leibler divergence is a good choice for general distribution comparison, ddKS has properties that make it more desirable for surrogate model training and validation than Kullback–Leibler divergence.

Diverse fields of research, especially the physical sciences and studies including surrogate modeling and normalizing flows, require the comparison of high-dimensional distributions. In many cases, analysts of these distributions revert to using one dimensional comparison tools, such as the Kolmogorov–Smirnov test, simple formulation of the Earth Mover's Distance, or even the mean integrated squared error between histograms. These one-dimensional comparisons suffer from the inability to identify differences in the covariances encoded by each distribution. Therefore, efficient (both statistically and computationally) comparisons between high-dimensional distributions are essential to advancing the use of machine learning and statistical analysis in the physical sciences. We present one such technique.

One of the most used, and arguably most powerful, two-sample tests is that proposed by Kolmogorov and tabulated by Smirnov, the so-called Kolmogorov–Smirnov (KS) test (1; 2). This test is widely used to compare two samples to determine whether they come from the same one-dimensional distribution. It does so by creating a test statistic, which is the maximum difference between the cumulative density function of two samples. The significance can then be easily calculated from that test statistic, regardless of whether either sample came from a well-defined distribution or not. This test has several important properties for machine learning, including its fast computation and ability to test non-parametric samples. However, it comes with one major drawback: that it is only applicable in one dimension, whereas many problems in data science cannot be compressed to one dimension without loss of information.

Several authors, mostly notably Press *et al.* (3), have defined similar tests in higher dimensions. The difficulty in extension to many dimensions is the ambiguity in the cumulative density function for a many dimensional distribution. Press *et al.* define a test statistic in two dimensions using the class membership in each of the four quadrants surrounding each test point. The maximum of the differences between membership vectors becomes a test statistic very similar to that in the KS test, however the same properties for calculating the significance are not retained. Press' test is of polynomial time complexity, and perhaps due to this, has not seen wide acceptance (or possibly awareness) in the data science community.

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

We present a d-dimensional KS test, inspired directly by Press' methodology of extension, but we present two novel changes. Firstly, we present a tensor-based calculation of the statistic, which reduces the computational complexity for small sample sizes. Secondly, we use the permutation method to define significance for our examples, and illustrate how it could be used for practicing data scientists with our test statistic.

1 The d-dimensional Kolmogorov–Smirnov test statistic

We begin with two samples (a predicted, X_p, and true, X_t, sample) of N points, each point having d dimensions. We seek to compare the cumulative density functions (CDF) between these samples. The construction of a CDF is ambiguous in more than one dimension, however an often used surrogate is the membership in hyperspace regions partitioned at a given test point. In one dimension, this is equivalent to choosing a test point, and measuring two numbers: counting membership greater than and less than the given test point. This concept generalizes to many dimensions by using each test point as an origin to delineate regions in hyperspace and forming a 2^d membership vector whose components correspond to the number of points in each region. The ddKS distance using points from X_p to partition between the two samples is designated D_p. Let $x_i \in X_p$ and $V_j^p(x_i), V_j^t(x_i)$ be the jth component of the membership vectors for the predicted and true samples generated by the partition of space due to the point x_i. The ddKS divergence between X_p and X_t using X_p is the largest element of the difference of the two membership vectors:

$$D_p = \max_{i,j} |V_j^t(x_i) - V_j^p(x_i)| \tag{1}$$

Where the subscript p indicates use of the predicted dataset as the partitioning points; a subscript t indicates use of the true dataset as the partitioning points. The ddKS test statistic, D, is then the average of D_p and D_t.

The most straightforward implementation of ddKS can be constructed using loop-based logic. For each point in each sample, the membership of the hyperspace regions using that point as a partition is counted. As described above, we evaluate the region membership of both, using predicted and true samples as partitioning points, and average the maximum differences of their membership vectors.

The naive loop-based implementation has high computational cost, on the order of $\mathcal{O}\left(2^d N^2\right)$ for all N. Because of this, we have developed a tensor-based implementation which utilizes `pytorch`'s (4) implicit parallelization on CPU or GPU. The computational complexity of the tensor form of ddKS is $\mathcal{O}\left(2^d\right)$ for small N, with memory complexity $\mathcal{O}\left(N^2\right)$. We implemented this tensor-based method with `pytorch`, enabling use of modern GPU hardware.

1.1 Monte Carlo

We illustrate the utility of ddKS on two non-parametric distributions. In many physics applications, the correlation between two or more variables is of interest, more than simply the distribution of either variable independently. This is true in coincidence spectroscopy, particle transport, and many other applications. We create two pathological distributions: one to illustrate the problem with using one dimensional test statistics, the other to demonstrate comparison of signals in varied background - an oft-encountered detection physics problem.

As an illustration of ddKS's utility on problems in modern physical sciences, we present a non-parametric, toy problem germane to modern high energy physics. We simulate a Cherenkov ring detector made of quartz of height $20\mathrm{cm}$. The bottom surface of the quartz medium is an ideal detection plane, and detections are recorded when an optical photon passes the surface with perfect spatial and timing resolution. We simulate a Cherenkov cone with the charged particle traveling directly down the z axis, the center of the detection plane. We also simulate a uniform background: optical photons are emitted isotropically from the top plane of the cylinder, uniformly in time. As shown in Figure 1, the one-dimensional histograms of the background versus the Cherenkov signal are identical (shown by the silver a copper hatched regions overlapping). Clearly, the one-dimensional KS test would not reject the null hypothesis, and this signal would not be distinguishable from the background. However, when looking at the distributions in three dimensions, as shown on the three-dimensional scatter plot, there is a clear difference in the distributions of the Cherenkov cone and background.

We also generate another dataset, where photons are generated and detected as before, however this time in the presence of a volumetric radiological contamination. Background photons are emitted

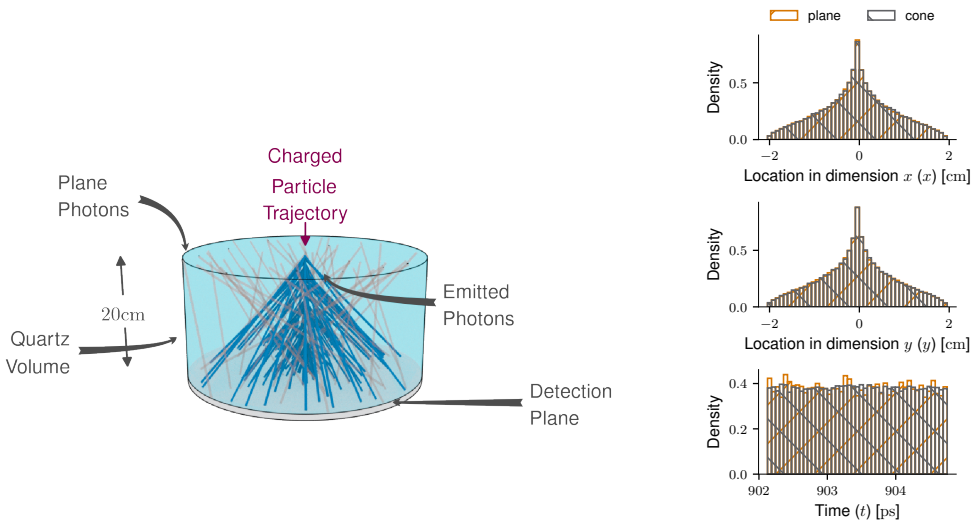


Figure 1: Dataset constructed mimicking photon emission during Cherenkov process. Histograms of detection position and time (silver and copper hatched regions) overlap almost exactly.

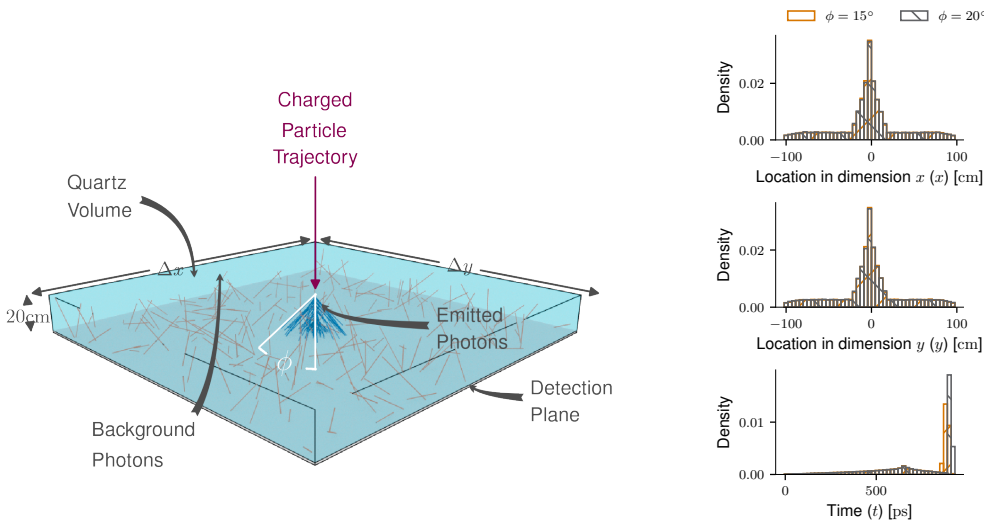
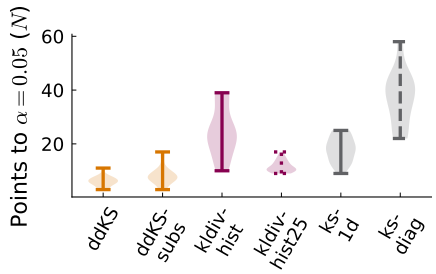


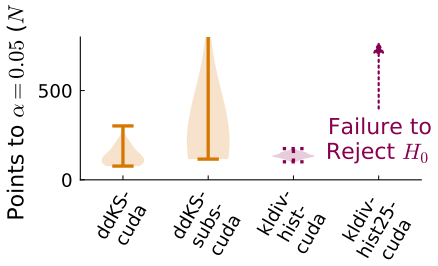
Figure 2: Dataset constructed mimicking photon emission during Cherenkov process with a volumetric background. Histograms of detection position and time (silver and copper hatched regions) overlap closely.

isotropically, uniformly distributed throughout a large quartz volume. A comparison between two different charged particles (with properties such that $\varphi = 15°$ and $20°$) in the presence of the same background is then attempted. As shown in Figure 2, the one dimensional histograms of the two different cones overlap closely. In the presence of the broad background, a very fine binned histogram (and therefore many samples) is necessary to disambiguate the two distributions.

The examples we present have been chosen specifically to illustrate issues with single dimensional distribution comparisons in modern particle physics; however they illustrate a fact that is important for both simulation validation and generative modeling. Correctly modeling and accounting for the correlations between variables is essential in real world physics. In generative modeling, it is even more important: a generative model could learn an "easier" distribution and one-dimensional test



(a) P-Value versus number of points for KL, 1d KS, ddKS tests on the comparison between a Cherenkov cone and a volume source. Each permutation test was performed using 100 permutations, and trials were repeated 25 times.



(b) P-Value versus number of points for KL, 1d KS, ddKS tests on the comparison between two Cherenkov cones with θ of 15° and 20° in a wide background of volumetric photon emissions. Each permutation test was performed using 100 permutations, and trials were repeated 25 times.

Figure 3: Efficiency metrics for different statistical tests.

statistics would indicate the generative model had larger skill than it truly had. ddKS provides a better interface for doing this than one-dimensional tests.

2 Comparison to other test statistics

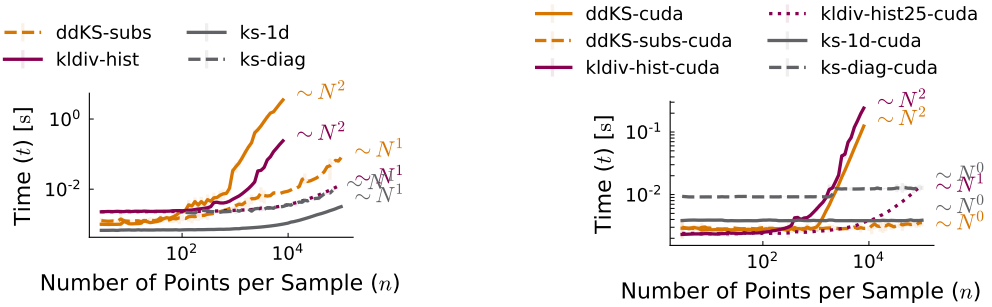
Because of the permutation test, we can use any distance or divergence as a test statistic. To show ddKS's utility for physical sciences, we compare it to two other test statistics: the oft-used Kullback–Leibler (KL) divergence, and against a combination of one-dimensional KS tests. We calculate the KS test statistic on each dimension individually, summing those to create a pseudo-multi-dimensional test statistic. We indicate this as `ks-1d` on figures. We also calculate the diagonal distance of each point in each pairwise dimension using the l_2 norm, subsequently summing each dimension's KS test statistic as above. We indicate this as `ks-diag` on figures. We calculate the KL divergence between an estimated probability density function of the two distributions. We use a histogram using Scott's (5) rule, sizing the number of bins per dimension proportional to $N^{\frac{d}{d+2}}$, to estimate probability density. We indicate this as `kldiv-hist` on figures. Finally, we calculate a lower resolution probability density using only 3 bins in each dimension, subsequently calculating the KL divergence as above. We indicate this as `kldiv-hist25` on figures. The label `-cuda` indicates that computations were performed on a CUDA-enabled GPU, which decreases wall-time for computation without changing the value or significance of the test statistic.

Shown in Figure 3a, we can see the utility of ddKS compared to other metrics. For the dataset with cylindrical geometry, ddKS variants reject the null hypothesis to $\alpha = 0.05$ by 5 points per sample, whereas KL divergence variants reject the same null hypothesis by 15-20 points per sample. One-dimensional test statistics require > 20 points per sample to reject the test statistic.

Figure 3b shows the statistical efficiency of all metrics for the background included dataset. ddKS again performs well, rejecting the null hypothesis at around 100 points per sample, with KL divergence rejecting the null hypothesis by 125 points per sample. Both accelerated computation methods (ddKS with subsampling and KL divergence ($\sim$25 bins) have a difficult time consistently rejecting the null hypothesis.

2.1 Time Complexity

Compared to one-dimensional test statistics, multidimensional test statistics incur a high computational cost. Illustrated on Figure 4, both KL divergence and ddKS are $\mathcal{O}\left(N^2\right)$ at high N. However, the implementation using tensor primitives and the implicit parallelization of modern matrix math libraries (`pytorch` in this case) make all test statistics $\mathcal{O}\left(1\right)$ for low N, making them possible to calculate continuously in a machine learning context. Loop-based implementations (not shown on



(a) CPU (b) GPU

Figure 4: Time to evaluate versus number of points for metrics considered. Time is recorded for permutation tests using 100 permutations, therefore the number to evaluate the test statistic once is $\leq 100\times$ that recorded on this chart. Estimated time complexities as $N \to \infty$ are printed to the right side of each line.

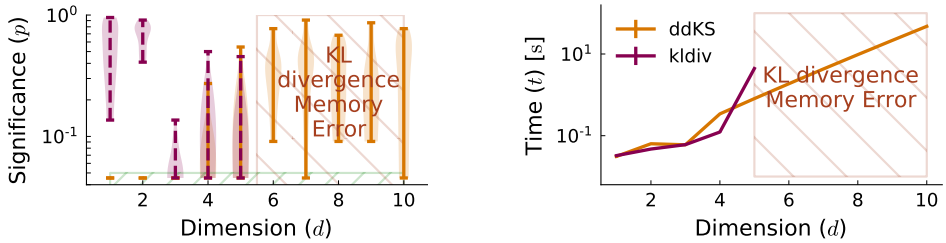


Figure 5: Time complexity and significance with increasing dimensions for ddKS and KL divergence using Scott's rules for constructing the histogram. Each permutation test was performed with 20 permutations, and each trial was performed 10 times. The green region indicates regions where H_0 could be rejected to $\alpha = 0.05$.

Figure 4) exhibit the same time complexity for all N, so while sometimes accelerated, are much slower at small N.

2.2 Higher Dimensions

One of the main benefits of ddKS is its direct application to higher dimensions. Because the Cherenkov and Background Included datsets included effects that were inherently 3-dimensional, they were inappropriate to illustrate effects in any other number of dimensions. An abstract test set for higher (and lower) dimensions was constructed. Both samples were filled with 50% background from a uniform distribution of dimension d from -100 to 100. Then, a hypersphere of dimension d was constructed, the radius 50 and 45 in each respective distribution. The results of evaluating ddKS and KL divergence on this dataset are shown in Figure 5.

ddKS is able to reject the null hypothesis for every trial up to dimension 3 at $\alpha = 0.05$, and is sometimes able to reject the null hypothesis in dimension $4 \leq d \leq 8$. Above dimension $d \geq 9$, ddKS is no longer able to reject the null hypothesis. In general, KL divergence is not as consistent as ddKS in higher dimensions. KL divergence is able to reject the null hypothesis for $\alpha = 0.05$ sometimes in dimensions $d \in [1, 3, 4, 5]$. Using Scott's rules for histogram construction lead to huge memory requirements for PDF estimation in the KL divergence calculation. Above dimension $d \geq 5$, KL divergence requires $\geq$ 32GB of memory, including requesting 374PiB of memory for dimension $d = 10$.

3 Conclusions

In conclusion, we find ddKS to be a generally useful test statistic for high dimensional physical science problems. It compares favorably against other possible test statistics, both in statistical and computational efficiency. Particularly, it is more consistent, and doesn't require the high effort of constructing an appropriate estimate of the PDF that KL Divergence requires. ddKS is also a metric, and those properties make it more desirable for use as a loss function in surrogate modeling problems. In general data science applications, ddKS could place statistical significance on predictions from other machine learning techniques with high dimensional latent spaces.

Broader Impact

We see applications of ddKS in a wide variety of fields. While our specific motivation - to validate the correct behavior of a surrogate model in high energy physics - is a natural application, high dimensional distribution distances are also important in generative modeling. As ddKS is closer to basic statistics than many applied machine learning algorithms. We do not foresee any ethical implications of its publication.

References

[1] A. Kolmogorov, Sulla determinazione empirica di una lgge di distribuzione, Inst. Ital. Attuari, Giorn. 4 (1933) 83–91.

[2] N. Smirnov, Table for estimating the goodness of fit of empirical distributions, Ann. Math. Statist. 19 (2) (1948) 279–281. doi:10.1214/aoms/1177730256.
URL https://doi.org/10.1214/aoms/1177730256

[3] W. H. Press, S. A. Teukolsky, Kolmogorov-Smirnov Test for Two-Dimensional Data, Citation: Computers in Physics 2 (1988) 74. doi:10.1063/1.4822753.
URL https://doi.org/10.1063/1.4822753

[4] A. e. A. Paszke, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.
URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[5] D. W. Scott, S. R. Sain, Multi-dimensional Density Estimation, Handbook of Statistics vol 23 Data Mining and Computational Statistics (August 2004) (2004) 1–39. doi:10.1016/S0169-7161(04)24009-3.