

A Multivariate Method for Comparing N-dimensional Distributions

James D. Loudin and Hannu E. Miettinen
Rice University, Houston, TX 77005, U.S.A.

We propose a new multivariate method for comparing two N -dimensional distributions. We first use kernel estimation to construct probability densities for the two data sets, and then define two discriminant functions, one appropriate for the null hypothesis and another appropriate for the actual data. Distributions of the two discriminant functions at random test points are then compared using the one-dimensional K-S test. The performance of the method is illustrated with Monte Carlo data.

1. INTRODUCTION

A comparison of two distributions is often a crucial part of data analysis. We may want to know whether a measured distribution is consistent with some hypothesis, or we want to compare “signal” and “background” distributions to see if they are different. In case of one-dimensional distributions the Kolmogorov-Smirnov (K-S) test [1] provides a tried-and-true method for such a comparison. The test uses the maximum distance d between the cumulative distribution functions of two histograms or probability densities as a measure of their similarity. The K-S test is non-parametric and independent of the shapes of the underlying distributions. However, the K-S test does not generalize naturally to higher dimensions, and there is no widely accepted test for comparing N -dimensional distributions.

In this note we propose a method which combines the information contained in N -dimensional distributions with the simplicity of the K-S test in one dimension. We first construct the relevant N -dimensional probability densities by using kernel estimation. We then define two discriminant functions, one for the null hypothesis and another for the actual data, and their distributions at randomly selected test points are compared using the standard K-S test. The method is mathematically uncomplicated and it appears to work well, based on test results in two dimensions.

2. THE METHOD

Consider two data sets A and B containing n_A and n_B data points (or “events”), respectively. Let each set be described by N variables x_i which are combined into a feature vector $\mathbf{x} = (x_1, \dots, x_N)$. We assume that the data have been binned in each dimension using some reasonable criteria so that N -dimensional scatter plots for both sets are available.

Our method, sketched in Figure 1, consists of three steps:

1. We first construct probability densities $f_A(\mathbf{x})$ and $f_B(\mathbf{x})$ using kernel estimation. In this study we have used the PDE method [2], whereby

the densities are estimated by adding up N -dimensional Gaussians placed at data points. The width of the Gaussian in each dimension, h_i , is proportional to the standard deviation σ_i of the i^{th} variable: $h_i = h \cdot \sigma_i$. Here h is a global smoothing parameter which sets the overall scale for the widths of the Gaussians. The value of h is optimized by minimizing the error $\int [d - f(\mathbf{x})]^2$, where d is the number of data points in a given bin and f is the corresponding density estimate. Figure 2 illustrates the density estimation and the optimization of h in the case when the feature space is two-dimensional.

2. We define a discriminant function D associated with the data as

$$D(\mathbf{x}) = \frac{f_A(\mathbf{x})}{f_A(\mathbf{x}) + f_B(\mathbf{x})} \quad (1)$$

Possible values of D are obviously in the range $0 \leq D \leq 1$. We similarly define another discriminant function D^* by replacing f_B with f_A which is obtained by generating n_B random data points distributed according to f_A and then constructing the density as described above. Thus D^* is a discriminant function associated with the null hypothesis that A and B come from the same underlying density. In order to reduce statistical fluctuations we actually form D^* a number of times and use their average $\langle D^* \rangle$ as the discriminant function appropriate for the null hypothesis. Distributions of D , D^* , and $\langle D^* \rangle$ are obtained by evaluating each function at randomly generated test points pulled from the density f_A . If A and B are similar, the distribution of D should peak near $D = 0.5$. The distribution of $\langle D^* \rangle$ is independent of B and should ideally be narrow and peaked at $\langle D^* \rangle = 0.5$.

3. We compute the cumulative distribution function F given by

$$F(x) = \int_0^x f_D(t) dt \quad (2)$$

where f_D is the probability density for D , obtained from the distribution of D . Similar cumulative distribution functions F^* and $\langle F^* \rangle$

are computed from the distributions of D^* and $\langle D^* \rangle$. The K-S distance d between F and $\langle F^* \rangle$ is a measure of similarity between sets A and B . The distribution of K-S distances d^* between F^* and $\langle F^* \rangle$ tells us what to expect for the null hypothesis. We can then compute the significance function $S(x)$, defined as

$$S(x) = \int_x^1 f_{d^*}(t) dt \quad (3)$$

where f_{d^*} is the density function for d^* . If we obtain a K-S distance d for the data, then $S(d)$ is the probability that the K-S distance would exceed d under the null hypothesis, i.e. the probability that purely random fluctuations could produce the observed value.

3. RESULTS

We generated the data sets A and B from two-dimensional Gaussian densities of equal widths in the two dimensions. We chose $n_A = 1000$ and $n_B = 50$, motivated by a “typical” analysis situation where we might have ~ 50 interesting data events to be compared with a larger control sample of e.g. Monte Carlo events. The relevant discriminant functions were constructed as described above, and they were evaluated at 1000 random test points to get the associated distributions. Figure 3 shows the distribution of $\langle D^* \rangle$, averaged over 500 distributions of D^* . The distribution is narrow and peaked at $\langle D^* \rangle \simeq 0.5$ as it should be.

We have studied the shapes of the D^* distributions as a function of the parameter h which determines the widths of the Gaussians used in kernel estimation. We find that if h is too small, there will be “valleys” in the density f_A^* due to the small sample size, and these valleys give rise to values of D^* near 1. If h is too large, all densities are too flat, and values of D^* much above 0.5 cannot occur. It is therefore important to optimize the value of h fairly carefully, otherwise the distribution of $\langle D^* \rangle$ will be asymmetric.

Figure 4 shows the cumulative distribution function $\langle F^* \rangle$ associated with $\langle D^* \rangle$. We also show F for a data set where the widths of the Gaussians for sets A and B are equal. The K-S distance d is indicated in the figure.

Figure 5a shows the (normalized) distribution of d^* for the null hypothesis. The mean value is $\langle d^* \rangle \simeq 0.10$, and there are no distances beyond $d^* \simeq 0.2$. The associated significance curve is shown in Figure 5b. We have verified that the shape of the significance

curve for the null hypothesis is nearly independent of the widths of the Gaussians used in kernel estimation.

In order to test the “resolving power” of the method we have applied it to data sets A and B (1000 and 50 events, respectively) pulled from two-dimensional Gaussian densities whose widths differ by 10%, 20%, and 50%. In each case the B set was generated 200 times. The distributions of the K-S distance d are shown in Figure 6.

We find that in the 10% case the mean value is $\langle d \rangle \simeq 0.12$, indicating that there is a $\sim 30\%$ probability that the *average* d would exceed this value for two identical densities. Thus we cannot distinguish between the two data sets. In the 20% case we find $\langle d \rangle \simeq 0.15$, and the null hypothesis probability for the average d has dropped to $\sim 10\%$. However, nearly half the time d is below 0.15, meaning that such values *could* arise from random fluctuations, and one quarter of the time d is below 0.10, meaning that such values are *likely* to arise from random fluctuations. In the 50% case it is easy to distinguish between A and B most of the time, but there is still a 15-20% chance that random fluctuations could explain the observed value of d .

4. CONCLUSIONS

We have outlined a method for comparing N -dimensional distributions which combines a multivariate approach with the standard K-S test. The method provides a precise way of quantifying the degree of similarity between two distributions. We have tested the method in two dimensions by comparing two Gaussian distributions of different widths, and find that the method performs well even when one of the data sets is relatively small.

Acknowledgments

We thank Sang-Joon Lee for discussions and for a critical reading of the manuscript.

References

- [1] F.J. Massey, J. Amer. Stat. Assoc. **46**, 68-78 (1951).
- [2] L. Holmström, S.R. Sain, and H.E. Miettinen, Comp. Phys. Communications **88** (1995) 195-210.

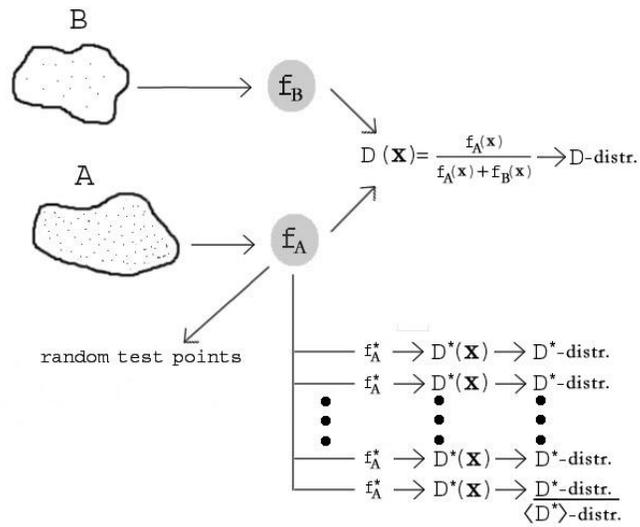


Figure 1: Sketch of the analysis method.

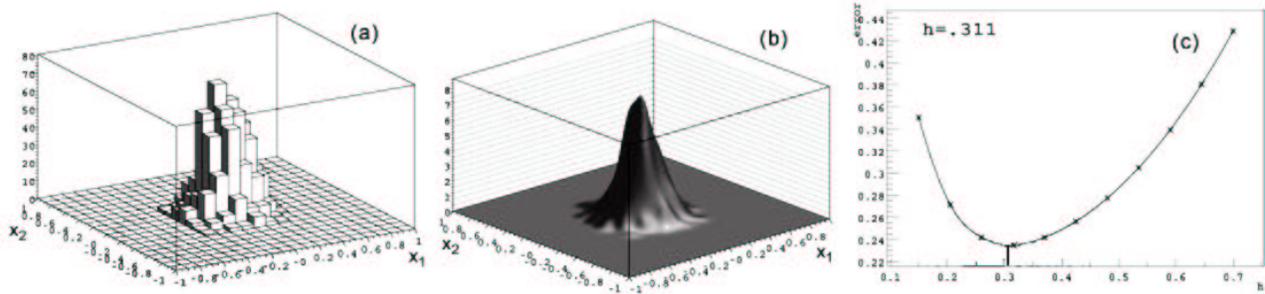


Figure 2: (a) Lego plot of a two-dimensional Gaussian. (b) The corresponding density estimate f . (c) Optimization curve for h .

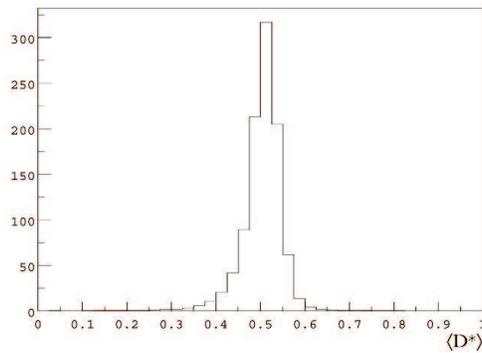


Figure 3: Distribution of $\langle D^* \rangle$, the discriminant function for the null hypothesis.

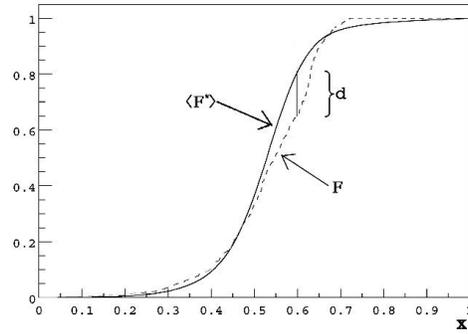


Figure 4: Cumulative distribution functions $\langle F^* \rangle$ for the null hypothesis (solid curve) and F for the data (dashed curve). The K-S distance d is also shown.

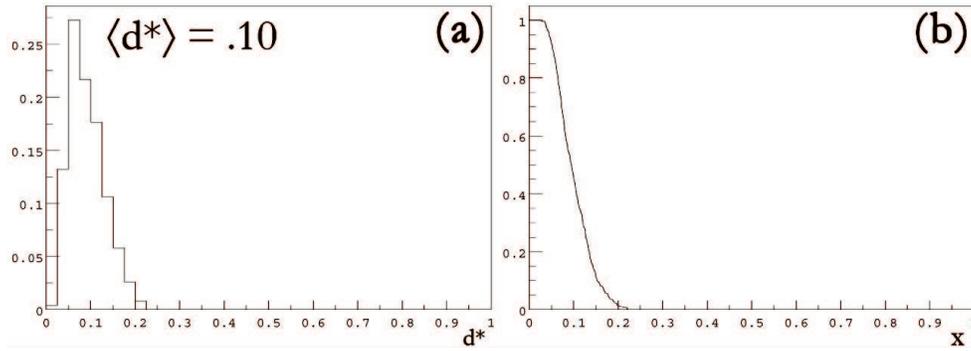


Figure 5: (a) Distribution of the K-S distance d^* for the null hypothesis. (b) The associated significance curve.

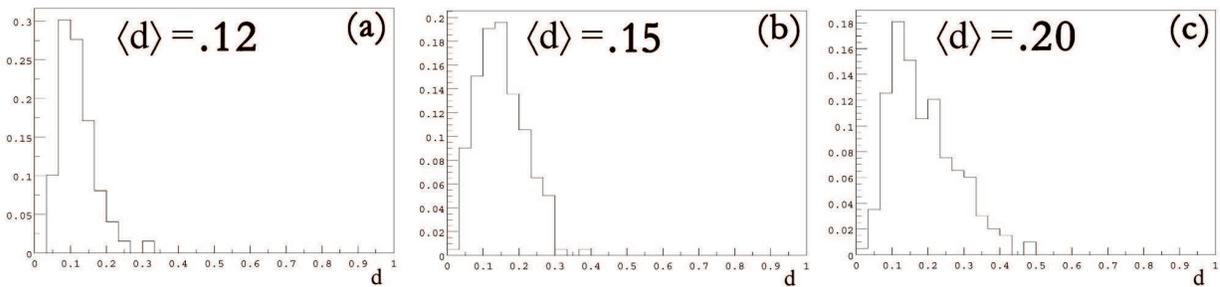


Figure 6: The distributions of the K-S distance d when the widths of the Gaussian densities for data sets A and B differ by (a) 10% (b) 20% (c) 50%.