Introduction to TATTER: code built to perform two-sample hypothesis testing

Annie Moore

December 14, 2022

Abstract

Different surveys generate data sets of the same observable. It is important to know if it is statistically favorable to combine data sets from different surveys. One way this is done is to use a two-sample hypothesis test, the form of which is chosen to fit the problem at hand. In the case of astronomical data, the test needs to be able to analyze general, multi-dimensional data. Popular tests such as the Kolmogorov-Simirnov test and the Kullback-Leibler tests have limitations that make it difficult to use them to study astronomical data. The most successful test so far that improves upon these limitations is the Maximum Mean Discrepancy (MMD) test. It is non-parametric and can be extended to analyze multi-dimensional data. In this paper, he results of these three tests are compared when analyzing one-dimensional data, multi-dimensional data, and data generated from different distributions. It was found that the MMD test performed as well or better than the others. The MMD test is then used to study the compatibility of two astronomical data sets as an example of how it can be used in future analyses.

1. Introduction

Recently, astronomers have been able to collect massive amounts of data. As a result the number of public data sets has greatly increased. These data sets cover a large range of astronomical data from large scale structure to measurements of single objects in multiple wavelengths.

With this increase in data, one of the main questions in astronomy today is if two data sets from independent surveys are the same (or similar). If they are, then the two data sets can be combined and used to learn more about our universe. In theory, two data sets that were created using the same process (same object/observable was measured, same selection criteria was used to select observable, etc.) can always be combined. However, observational and theoretical systematic errors are often present that can cause these measurements to differ. If not corrected, these errors can be large enough that data sets describing the same observable cannot be combined.

Another way to think about this is that systematics will cause the data sets to appear as if they were drawn from different probability distributions. Because of this, it is important to have a test that can be used to compare two sets of data and determine if they could have been generated from the same distribution.

This can be accomplished using the two-sample hypothesis test. It defines the null hypothesis as two data sets are equal $p(x) = q(y)$. The alternative hypothesis is then that the two data sets are not equal $p(x) \neq q(y)$. To perform the test, the test statistic (measures the agreement between data and the null hypothesis) is computed many times assuming that the null hypothesis is true. This gives the test statistic distribution. The test statistic is computed for the two data sets and the result is compared with the test statistic distribution. If the test statistic computed from the observed data falls within the tails of the test distribution below some cutoff point, then that indicates that the null hypothesis is rejected (maybe include image). The cutoff point is chosen beforehand.

There are two tests that are popular in astronomy. The first is the Kolmogorov-Simirnov test (K-S test). The K-S test is used to compare the cumulative distribution function of two data sets or compare simulations with empirical data. This is is popular in astronomy due to the fact that it is non-parametric, so the distributions do not need to be known beforehand, and it is simple. The test statistic is computed in the following way

$$D_{KS} = \sup|\mathrm{CDF}(p(x)) - \mathrm{CDF}(q(y))|. \tag{1}$$

It has been used in ... The major drawback to this test is that it is not easy to extend it to multi-dimensional data. This has been attempted, but attempts are less accurate and are not publicly available. A proposed

improvement to this test is the Kullback-Leibler (K-L) divergence test. This quantifies the information lost when one distribution is used to approximate another (discrepancy between two distributions). This test is asymmetric, meaning that using some distribution q(y) to approximate another distribution p(x) will yield a different result than using p(X) to approximate q(y). The test statistic is computed through

$$D_{KL}(p|q) = \int p(x) \ln[\frac{p(x)}{q(x)}]dx \tag{2}$$

This can be easily applied to multi-dimensional data. However, it does require an explicit form of the probability density functions for finite samples. Studies of astronomical data would prefer more general tests.
The Maximum Mean Discrepancy (MMD) test is proposed as a solution to the above problems. It is a non-parametric test that is easily extended to multi-dimensional data. The development and use of this test will be discussed in the remainder of this paper.
The rest of the paper will be organized in the following way. In section 2 I will introduce the background material for the Maximum Mean Discrepancy test. Section 3 I will use the MMD, K-S, and K-L tests to compare pairs of distributions and discuss the results. The results in this section were computed using TATTER[1], software designed to implement these three tests. In section 4 I will demonstrate potential application to astronomical data. I will conclude in section 5.

2. Methods

The Maximum Mean Discrepancy (MMD) test is a measurement of the distance between two probability distributions. A test statistic was developed so that it returns a unique value only when two distributions $p(x), q(y)$ are equal. In the case of a distance metric, the test statistic will go to zero when the two are equal.
This test was first introduced in (1) as a way to compare two multi-dimensional probability distributions. This specific test relies on a kernel based estimator. The function of the kernel is to embed the sample into a high dimensional feature space called a reproducing kernel Hilbert space (RKHS). The dimensionality of that space is determined by the form of the kernel function.
To better understand the use of the kernel function, it is important to introduce the concept of feature means. Given some random variables X, a feature map ϕ maps X to another space denoted by $\mathcal{F}$. Following Gretton et al, it is possible to use kernel functions to compute the inner product in $\mathcal{F}$ by

$$X, Y \text{such that} k(X, Y) = \langle \phi(X), \phi(Y) \rangle_{\mathcal{F}}. \tag{3}$$

To find the required feature means, we need another feature map that takes $\phi(X)$ and maps it to the means of every coordinate in $\phi(X)$. This could take the form

$$\mu_p(\phi(X)) = [E(\phi(X_1)), ..., E(\phi(X_m))]^T. \tag{4}$$

The inner product of these feature means can then be written as a kernel function

$$\langle \mu_p(\phi(X)), \mu_q(\phi(Y)) \rangle_{\mathcal{F}} = E_{p,q}[\langle \phi(X), \phi(Y) \rangle_{\mathcal{F}}] \tag{5}$$

Using that the Maximum Mean Discrepancy is the distance between feature means

$$MMD^2(p, q) = ||\mu_p - \mu_q||^2 \tag{6}$$

where $||x|| = \sqrt{\langle x, x \rangle}$, we can write this as a function of the kernels

$$\begin{aligned} MMD^2(k, p, q) &= \langle \mu_p, \mu_p \rangle - 2\langle \mu_p, \mu_q \rangle + \langle \mu_q, \mu_q \rangle & (7) \\ &= E_p[k(X, X)] - 2E_{p,q}[k(X, Y)] + E_q[k(Y, Y)]. & (8) \end{aligned}$$

this estimate will be biased if instead we choose to study the empirical expectations computed from samples instead of the population expectations. This bias is currently unavoidable as we don't know the probability distribution beforehand. For finite sample size, the above can be approximated as

$$MMD_u^2(k, p, q) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1i}^{n} k(x_i, y_j) \tag{9}$$

[1] https://github.com/afarahi/tatter/tree/e8dc8a33d16dbd5855568302d062f5641f319ade

where k(;) is the kernel function, m and n are the sample sizes, and x_i and y_j are individual samples of the distribution. This can only be used as a metric when the Hilbert spaces are universal. It was shown in ... that the Gaussian and Laplace kernel functions satisfy these requirements.
Here the kernel function was chosen to be a Gaussian radial basis function of the form

$$k(x, y) = \exp(-\gamma||x - y||^2). \tag{10}$$

This has had great success in the machine learning community for a broad range of applications. It projects the data into an infinite dimensional feature space, allowing it to capture information from higher order moments in the distribution. The algorithm is able to capture complex, non-linear interactions between different observables. This can be seen by looking at the feature map associated with this kernel function

$$\phi(x) = \exp(-\gamma x^2)[1, \sqrt{\frac{2\gamma}{1!}}, \sqrt{\frac{4\gamma^2}{2!}}, \sqrt{\frac{8\gamma^3}{3!}}, ...] \tag{11}$$

. The kernel function also includes the hyper parameter γ which can be interpreted as the weight of each moment in the computation. The specific value of γ was not explored in this paper (maybe in another one) and was just set to .1 throughout. This will just decrease the importance of higher order moments. This is ok for the majority of the examples in this paper as they follow a Gaussian distribution. As this code is used to examine distributions where higher order moments are important, this will need to be adjusted.
To perform the MMD test the test statistic is computed for the observed data. It is then assumed that the two data sets were drawn from the same distribution (p(x) = q(y)). The two sets are aggregated and the bootstrap algorithm is employed to generate two new data sets x_{test1} and y_{test1}. This is done N times and the test statistic is computed each time. This forms the null distribution. We can now determine the probabily of the null distribution exceeding that of the observations, the p_{value}. This is defined as

$$p_{value} = \text{count}(\text{Null} > \text{MMD}^2)/N \tag{12}$$

If this value is less than 0.05 then the null hypothesis is rejected.

3. Results

Here to describe the different results in the paper. The first is two test the accuracy of the MMD test as compared to the KS and KL tests. For this one important adjustment was made when computing the KL test statistic. The MMD and KS test statistics are symmetric while the KL test is not. In order to accurately compare the three different distributions, the KL test statistic needed to be made symmetric. To do this the average is taken of the two possible values of the KL test (using p to analyze q and using q to analyze p).

$$D_{KL,symm} = \frac{D_{KL}(p|q) + D_{KL}(q|p)}{2} \tag{13}$$

The first test compares two samples drawn from the same distribution, a Gaussian with $\mu = 0$ and $\sigma = 1$.. Because we know that the distributions are the same, we expect that the p-value computed will always be above 0.05. That is what we see in fig. 1a.

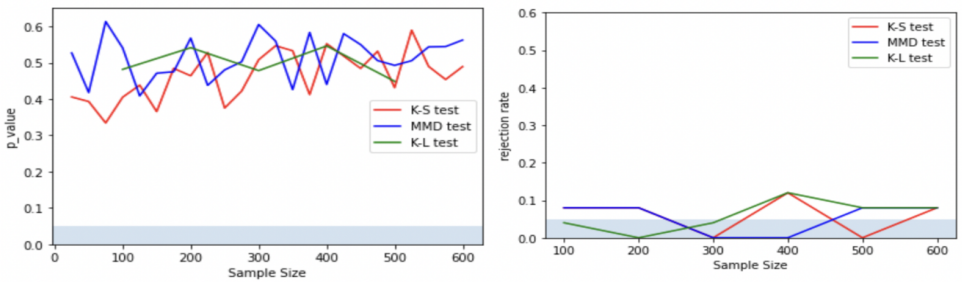

Figure 1: Results of the hypothesis test for two Gaussian distributions with the same mean. The left shows the average p_{value} from 25 iterations as a function of sample size. The right shows the rejection rate of those 25 iterations as a function of sample size.

All three tests determined that the two data sets were generated from the same distribution regardless of the sample size. Fig. 1b shows the rejection rate of the three tests. Each test rarely returned a p_{value} that would reject the null hypothesis.
Fig. 2 shows the results of the three tests when the data sets were generated from two different distributions ($\mu = 0.0and0.2$, $\sigma = 1.$).

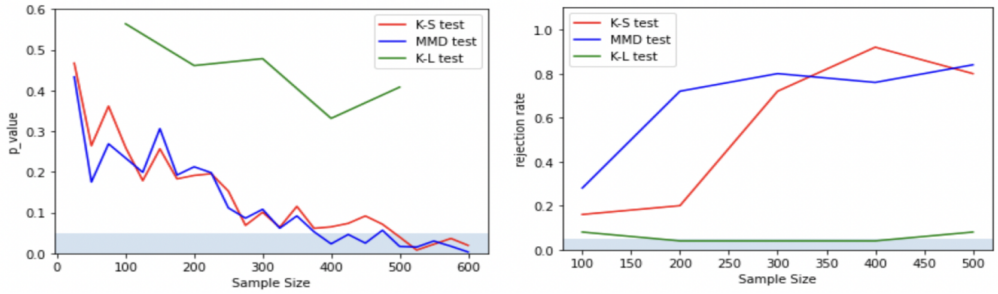

Figure 2: Results of the hypothesis test for two Gaussian distributions with different distributions. On the left we see that the MMD and K-S tests consistently return a $p_{value} < 0.05$ correctly rejecting the null hypothesis. On the right we see the rejection rate as a function of sample size. The MMD and K-S tests approach a 100% rejection rate as the sample size increases.

Fig. 2a shows that the MMD and K-S tests have similar performances and are able to correctly reject the null hypothesis at a large enough sample size. The K-L test is not as accurate, which is to be expected given what is known about the test. This can also be seen in the rejection rates plotted in Fig. 2b The improvement with the increase in sample size is expected as well. As the sample size increases, the probability distribution used to generate it becomes known (if we had an infinitely large sample the distribution would be known). The better the distribution is known the more likely it is that the test will correctly accept or reject the null hypothesis.
The second test will extend the used of the MMD test to multi-dimensional data sets. The test was performed for data sets drawn from multi-dimensional Gaussian distributions with different means ($\mu = 0.and0.2$). The inputs used to generate the distributions were a d-dimensional array equal to the value of the mean and the identity matrix as the covariance. This was done for three different sample sizes (20, 40, 80). The results are shown in Fig. 3.

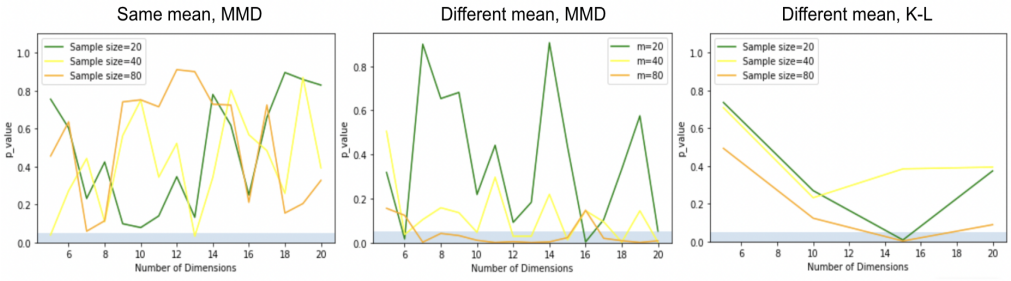

Figure 3: Left: The average p_{value} as a function of dimensionality computed using the MMD test for different sample sizes. The two multi-variate Gaussians used to produce the samples have the same mean and covariance. The null hypothesis is always accepted. Middle: The two Gaussians have different means. The results of the MMD test show that the null is correctly rejected for higher dimensions when the sample size is larger. Right: The K-L test is used to compare the two multi-dimensional Gaussians with different means. It is only able to reject the null hypothesis at high dimensions.

When the two distributions are the same (as shown in fig. 3a), the test does not reject the null hypothesis. We the distributions are different (fig. 3b) we find what we expect in that the MMD test is able to correctly reject the null hypothesis for multidimensional data for larger sample sizes. Fig. 3c shows the the results of comparing the different distributions using the K-L test. Similar to the 1-D case, the K-L test is not as accurate as the MMD test.

This test should also be able to differentiate between sets generated from probability distributions of different shapes. The results of this are shown in fig. 4.

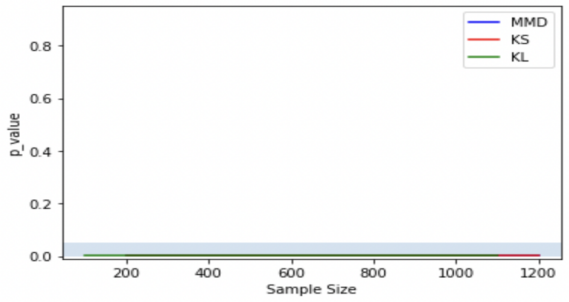

Figure 4: Average p_{value} as a function of sample size when comparing a Gaussian and lognormal distribution. All three tests always reject the null hypothesis.

One data set was generated from a Gaussian distribution and the other was generated from a log normal distribution. Both have the same mean and variance. The null hypothesis is correctly rejected for all three tests.

4. Astronomical Applications

Now for an example of how this test can be useful when studying astronomical data. This comparison can be done to study features of multiple surveys or to compare simulations. It will allow us to address the homogeneity of the surveys and the two-sample hypothesis test can quantify if combining the two data sets is statistically justified. If they are not, that could point to required work on the systematics of the survey.
An example of this application is shown for optical richness of red luminous galaxies at different redshifts from the Dark Energy Survey. An initial test compares two data sets from the same survey. The results should indicate that both data sets are the same. The results of this are shown in fig. 5.

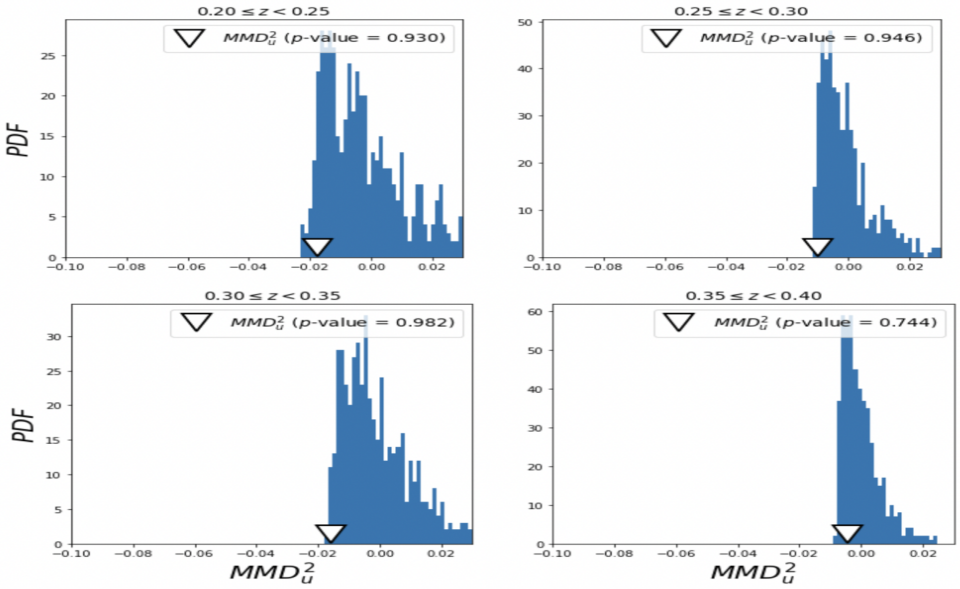

Figure 5: Results of the MMD test comparing the optical richness from two DES catalogs at different redshifts. All support the claim that the data sets are the same.

We plot the null distribution in blue and the value of the test statistic computed using data with the del. We see that the value for MMD is close to zero at all redshifts, indicating that the samples are indeed from the same distribution. An extension to this test is shown in fig. 7 of (2). This was done using data from DES and

SDSS optical richness. They found that the data agreed at low redshift but differed at high redshift suggesting that some systematics remained that were affecting the measurement. The difference is large enough that the two data sets cannot be combined.

5. Conclusion

With the current and expected astronomical data that is available, it will be important to know how they all fit together in order to learn more about our universe. For this we will need to find if two independent data sets describing the same observable are the same. This can be done using various two-sample hypothesis tests. In this paper, the Maximum Mean Discrepancy (MMD) test has been introduced. It is a distance based metric that is used to quantify the difference between two independent samples. The distribution of the data does not need to be known beforehand (non-parametric) and it can be used to analyze multi-dimensional data. This is an improvement over other popular tests in astronomy such as the K-S and K-L tests.
This test can be used to test the homogeneity of two samples to quantify the statistical significance of combining them. Two samples that describe identical observables should be able to be combined. Any other finding would indicate that unmodeled systematics are present in the measurement. This can help surveys learn about different systematics and the effects that they can have on their data.

References

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. 25, pp. 723–773, 2012.

A. Farahi and Y. Chen, "Tatter: A hypothesis testing tool for multi-dimensional data," *Astronomy and Computing*, vol. 34, p. 100445, 2021.