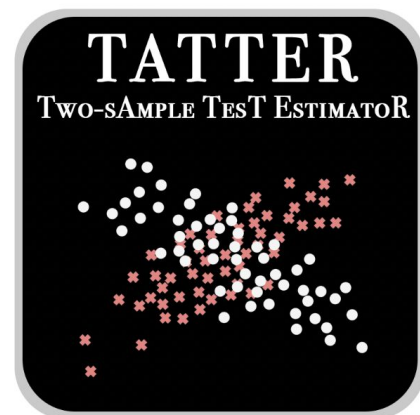


# TATTER: A hypothesis testing tool for multi-dimensional data



Farahi and Chen, 2021

Annie Moore

Paper: <https://www.sciencedirect.com/science/article/pii/S2213133720300998#appA>

Github: <https://github.com/afarahi/tatter/tree/e8dc8a33d16dbd5855568302d062f5641f319ade>

# Introduction

- Many different data sets exist from a variety of surveys
  - Large-scale structure, CMB
  - Need a test to determine if data samples from different surveys were generated from the same distribution
  - Done using two-sample hypothesis testing
  - Will quantify if combining two samples is statistically justified
- Can also be used to compare simulations to empirical data

# Two-sample hypothesis testing

- Performed on data from two independent samples
  - Testing to see if the difference in the populations is statistically significant
- Null hypothesis ( $H_0$ ):  $p(x) = q(y)$

Alternative hypothesis ( $H_A$ ):  $p(x) \neq q(y)$

- Will compute a null distribution assuming the null hypothesis is true
  - If observed value falls in the tail of the null distribution, then the null hypothesis is rejected
- Many different types exist
  - Use depends on the problem being solved

# Previous methods

- Kolmogorov-Smirnov (K-S) test
  - Compares cumulative distribution function of two sets of data
  - Popular choice for this two sample test
  - Non-parametric
  - Does not support multidimensional data distributions
- Kullback-Leibler (K-L) test
  - Quantifies the discrepancy between two samples
    - Measures how much information is lost when distribution  $q(y)$  is used to approximate  $p(x)$
  - Is not symmetric
  - Can only use for certain types of pdfs

# Introduction to maximum mean discrepancy test (MMD)

- Two-sample estimator that is based on distance
  - Developed to compute the distance between two multi-dimensional distributions
  - Relies on embedding sample into a high dimensional “feature space” through a kernel function (Gretton+, 2012a)
- Test statistic will determine how much test results differ from the null hypothesis
  - Value of zero indicated that two distributions are the same

# Design of MMD

- Measure of the Maximum Mean Discrepancy in this work is

$$\widehat{\text{MMD}}_u^2[k, x, y] = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n k(x_i, y_j)$$

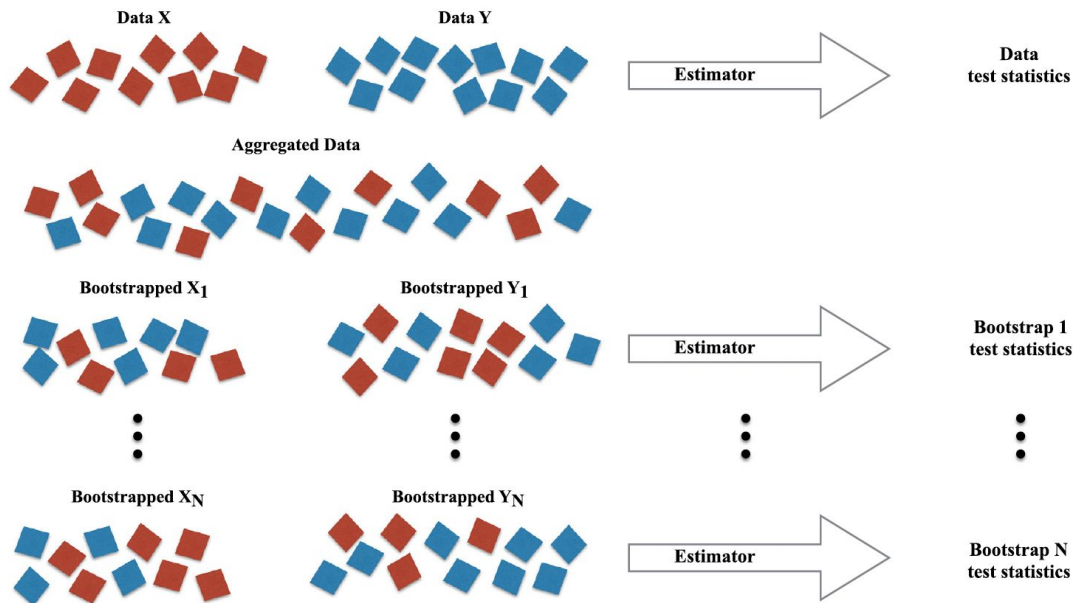
- $k(x_i, y_j)$  is the kernel
- $x_i$  and  $y_j$  are samples from distributions
- $m$  and  $n$  are size of sample data
  - Gaussian radial basis function

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

- Allows algorithm to capture complex, non-linear interactions

# Perform hypothesis testing for MMD

- Compute test statistic on previous slide using two lists drawn from probability distributions  $p(x)$  and  $q(y)$
- To compute the null distribution
  - Data sets are combined
  - Randomly draw  $m$  points for  $x_{\text{test}}$  and  $n$  points for  $y_{\text{test}}$
  - Uses bootstrap algorithm for resampling



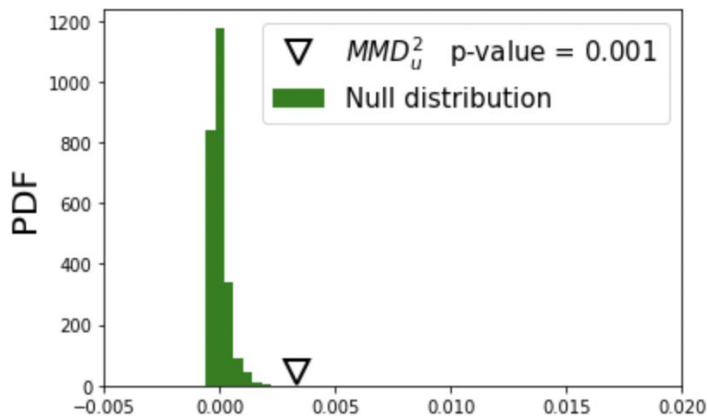
Farahi and Chen, 2021

# Perform hypothesis testing for MMD cont.

- Compute  $\text{MMD}_{\text{null}}^2$
- Calculate the p-value using

$$1 - p = \Pr(\text{MMD}_{\text{null}}^2[k, x, y] < \text{MMD}_{\text{data}}^2[k, x, y]).$$

- P-value < 0.05 indicates that null hypothesis can be rejected



---

**Algorithm 1** Our simulation-based bootstrap algorithm to estimate the null distribution and  $p$ -value

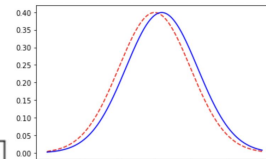
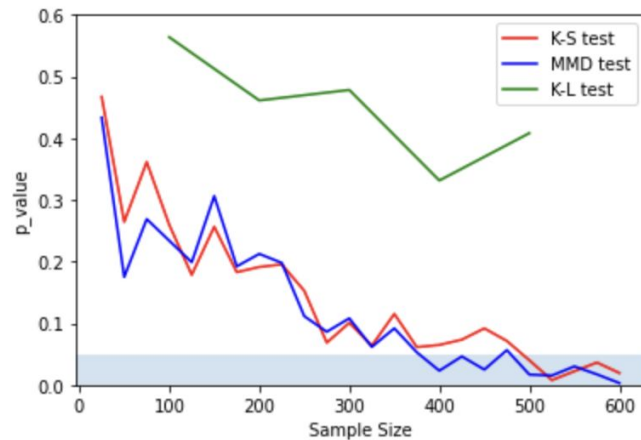
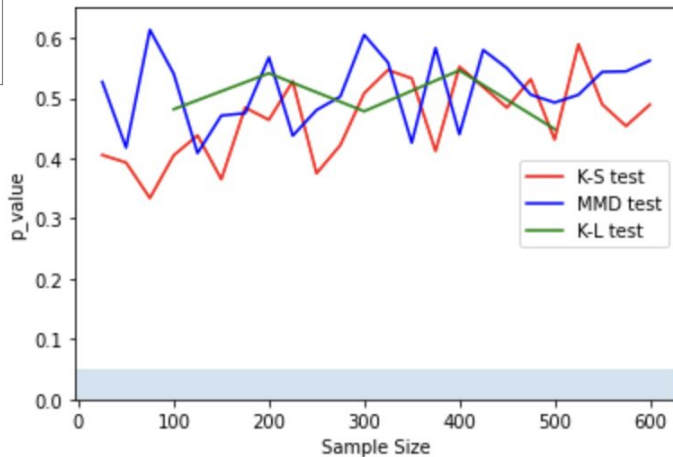
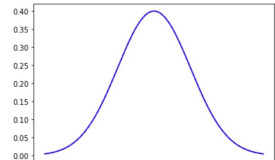
---

- 1: **Input:**  $x, y, N_b, k(\cdot, \cdot)$ : observed samples, the number of bootstraps, and the kernel function.
  - 2: **Output:**  $\widehat{\text{MMD}}^2$ , Null,  $p$ -value: an estimation of the  $\text{MMD}^2$  for the observed samples, drawn from the null distribution, and  $p$ -value.
  - 3: initialize the hyper parameters. ( $\gamma$ , see Eq. (4))
  - 4:  $\widehat{\text{MMD}}^2 = \text{MMD}^2(x, y, k)$ : compute  $\text{MMD}^2(x, y, k)$  with Eq. (2)
  - 5:  $Z \leftarrow$  aggregate observed samples.
  - 6:
  - 7: **for**  $i$  in  $\{1, \dots, N_b\}$  **do**
  - 8:    $x_{\text{boot}} \leftarrow$  randomly draw  $m$  data points from  $Z$  (with replacements)
  - 9:    $y_{\text{boot}} \leftarrow$  randomly draw  $n$  data points from  $Z$  (with replacements)
  - 10:   Null[ $i$ ]  $\leftarrow \text{MMD}^2(x_{\text{boot}}, y_{\text{boot}}, k)$
  - 11: **end for**
  - 12:
  - 13:  $p\text{-value} = \text{count}(\text{Null} > \widehat{\text{MMD}}^2) / N_b$
-



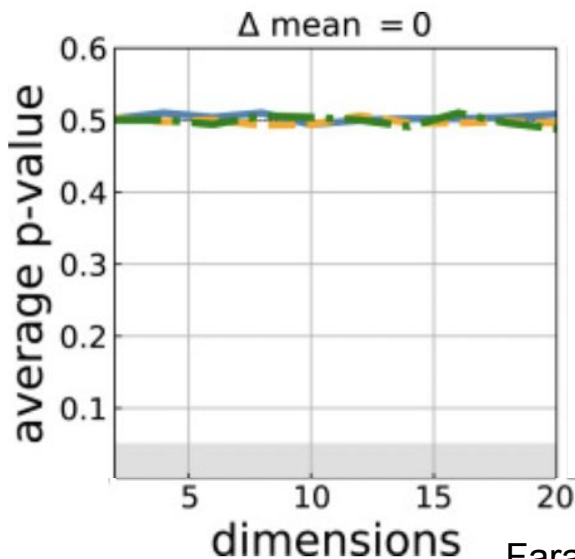
# Comparison with other test statistics

- Compare MMD, K-S, and K-L tests on two gaussian distributions
  - Once where means are the same and one where they differ by 0.2
  - Sample size is equal to 1
  - All correctly don't reject the null when the means are the same
  - MMD performance is comparable to K-S test

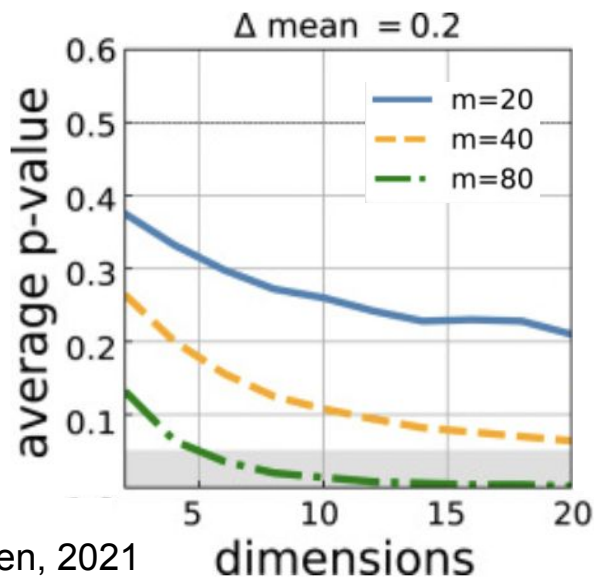


# Multi-dimensional data

- MMD is designed to work for multi-dimensional data
  - Short-coming of K-S test

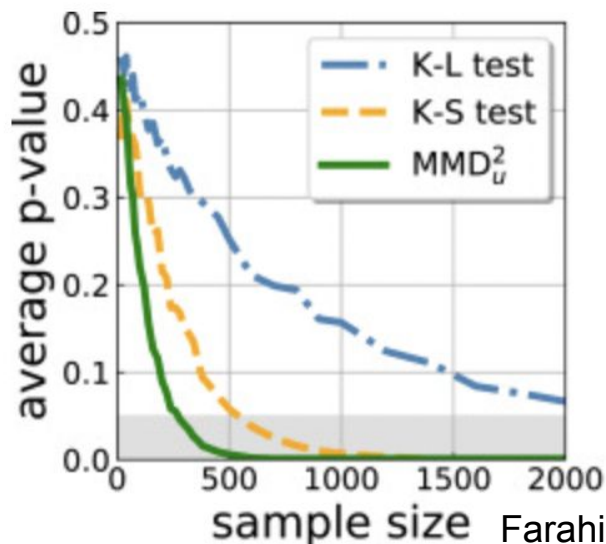


Farahi and Chen, 2021



# Different Distributions

- MMD is able to distinguish between data sets from different distributions
  - Able to do this without knowing the distributions beforehand



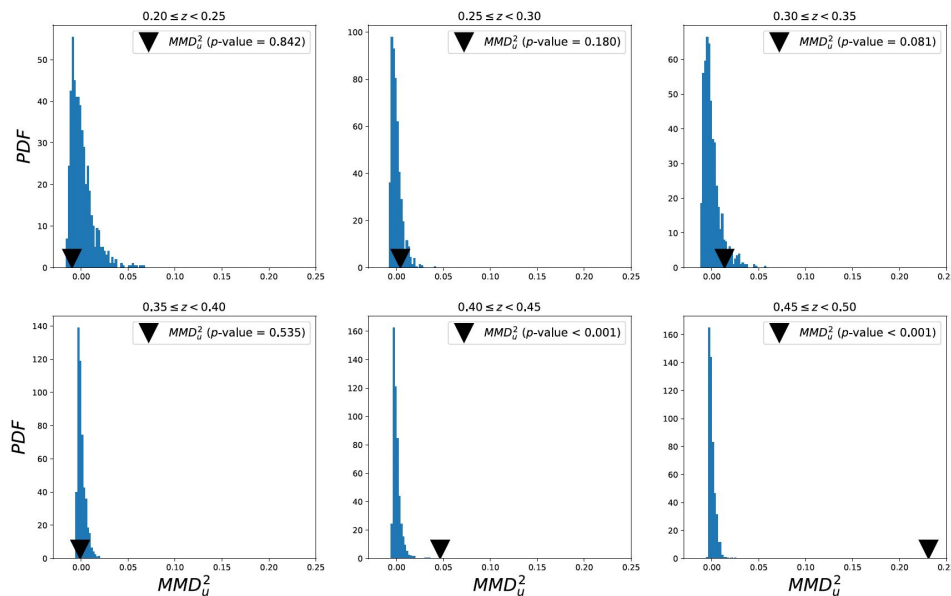
Farahi and Chen, 2021

# Application to astronomy

- Can be used to assess homogeneity of survey products and learn about systematics
- When looking at samples of galaxy clusters from two surveys
  - If galaxies are chosen in the same way and use the same cluster finding algorithm then distributions should be similar
  - Differences point to unaccounted for systematics
  - Will not be able to combine data samples

# Application to astronomy cont.

- Looked at optical richness and redshift from SDSS and DES data
  - Performed MMD test to compare similarity
  - Looks similar at low redshift but differs at high redshift indicating that there are systematics affecting the data



# Summary

- Maximum mean distance test is a new two-sample hypothesis test
  - Accuracy is comparable to other popular tests
  - Can be extended to multi-dimensional data
- Performance is comparable to or better than other two-sample tests used in astronomy (K-S and K-L tests)
- Can be used to determine if astronomical data sets from different surveys can be combined
  - Also as a test for potential systematics