Astronomy and Computing 34 (2021) 100445

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

TATTER: A hypothesis testing tool for multi-dimensional data[☆]

A. Farahi[a,*], Y. Chen[b]

[a] The Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI 48109, United States
[b] Department of Statistics, University of Michigan, Ann Arbor, MI 48109, United States

ARTICLE INFO

Article history:
Received 9 June 2020
Accepted 23 December 2020
Available online 5 January 2021

Keywords:
Methods
Data analysis — methods
Statistical

ABSTRACT

The two-sample hypothesis test quantifies whether distributions p and q are different, given the corresponding finite samples drawn from each. This problem appears in a legion of applications in astronomy, ranging from data mining to data analysis and inference. For decades, the Kolmogorov–Smirnov test has been astronomers' first choice to answer this question, but it has a major drawback, a generalization to multi-dimensional data sets is not straightforward. To fill this gap, we present a nonparametric estimator for comparing given multi-dimensional distributions drawn from them. This method employs a kernel function to construct an unbiased estimator of the Maximum Mean Discrepancy (MMD) distance between the two distributions that generated the observed data. We perform controlled numerical experiments in Gaussian, non-Gaussian, and multi-dimensional finite sample settings and test the performance of MMD estimator in each experiment. We then discuss some of the applications of this method in astronomy data analysis.

1. Introduction

Astronomical data are ubiquitous, multi-dimensional, and time-dependent with varying quality. Modern observatories collect massive amounts of data through large-scale surveys, medium and small size programs, and single targeted observations. On-earth and space-based observatories scan the sky over the entire electromagnetic spectrum, from X-ray wavelengths to radio wavelengths. And now gravitational wave observatories, with the recent discovery of gravitational waves from binary black hole and binary neutron star systems (Abbott et al., 2016, 2017), ushered a new dimension in exploring astronomical systems. The scale and complexity of these multi-wavelength, multi-messenger and time-dependent astronomical data from current and near-future surveys exceed the capacities of traditional data analysis tools and computational methods. Novel inference algorithms, computational frameworks, and validation techniques are inevitable necessities to enable new scientific discoveries.

A fundamental question that many inference algorithms and population studies are concerned about is that if two data samples are generated from the same distribution, i.e. *the two-sample hypothesis testing*. Suppose we have samples $\{x_i | x \in \mathcal{X}, i \in [1, \ldots, m]\}$ and $\{y_i | y \in \mathcal{X}, i \in [1, \ldots, n]\}$ drawn independently from distributions p and q respectively. The two-sample hypothesis testing asks if $p = q$. This seemingly simple but rather sophisticated and profound question appears in a wide range of astronomy applications, such as model comparison, model selection, hypothesis testing, goodness-of-fit evaluation, and inference, to name a few. Despite its practical importance, the astronomical literature has not seen a major leap in this direction. Some authors explored novel approaches (Mondal et al., 2008; Freeman et al., 2017; Modak and Bandyopadhyay, 2019); yet such studies are scarce.

Astronomers' first choice of the two-sample hypothesis test is the Kolmogorov–Smirnov (K–S) test. The K–S test compares the cumulative distribution function (CDF) of two sets of data or one probability distribution function and one set of data. While this simple, non-parametric test is popular among astronomers, it has a major drawback that impedes its application to a large class of problems. In its original form, this test can only handle one-dimensional distributions for which the CDF is uniquely defined. Despite a few attempts to extend the K–S test beyond one-dimensional distributions (Peacock, 1983; Fasano and Franceschini, 1987; Gosset, 1987; Lopes et al., 2008; Harrison et al., 2015), the current popular implementations of software packages do not support multi-dimensional distributions. Moreover, its multi-dimensional implementations often lack theoretical guarantees or are limited in scope and application. Other variants of the K–S test are the Cramer–von Mises test (Darling, 1957) and Anderson–Darling (A–D) test that inherit the same limitations of the K–S test. To this end, statisticians proposed various approaches to perform two-sample hypothesis testing in a multi-dimensional setting, e.g., Weiss (1960), Justel et al. (1997), Baringhaus and Franz (2004), Gretton et al. (2012a).

[☆] This code is registered at the ASCL with the code entry ascl:2006.007.
[*] Corresponding author.
E-mail address: aryaf@umich.edu (A. Farahi).

https://doi.org/10.1016/j.ascom.2020.100445

An alternative approach to quantifying the discrepancy between two samples is to use the Kullback–Leibler (K–L) divergence (Kullback and Leibler, 1951). The K–L divergence is an entropy-based method which is rapidly gaining popularity among the astronomy community (*e.g.,* Seehars et al., 2014; Ben-David et al., 2015; Zhao et al., 2017; Nicola et al., 2019). However, the K–L divergence has directionality and requires a base probability, i.e. is an asymmetric metric. As a result, this measure is not well-suited to the problem of two-sample hypothesis testing; nevertheless, in certain problems, the asymmetry goes away, or it can be symmetrized. Variations of this measure are employed in the literature to perform hypothesis testing (*e.g.,* Ben-David et al., 2015; De Simone and Jacques, 2019; Nicola et al., 2019). In contrast to the K–S test, the K–L divergence can be readily applied to multi-dimensional distributions. In a finite sample setting, one limitation of this method is that the K–L divergence evaluation relies on intermediate steps to estimate the probability density of data because it requires an explicit form for the probability density functions. Although assuming a specific probability density function simplifies the computation, it implies a restriction on the range of problems to which the method can be applied to, and does not satisfy our goal of developing a general methodology. Furthermore, computational complexity is another major drawback of this approach. Complex space partitioning or bias correction strategies complicate the computation of the K–L divergence (see Wang et al., 2006, for more detailed discussions).

In this work, we introduce a distance-based two-sample test estimator proposed by Gretton et al. (2012a). Distance metrics are often a popular method to quantify the discrepancy between two probability distributions. And their applications are not limited to the two-sample hypothesis test. For example, a distance metric that handles multi-dimensional and population-level data can be utilized in approximate inference algorithms (Herbel et al., 2017). A diverse set of likelihood-free methods, such as the Approximate Bayesian Computation (ABC, Weyant et al., 2013; Akeret et al., 2015; Ishida et al., 2015; Jennings and Madigan, 2017), rely on comparing derived summary statistics instead of comparing two data sets at the population level directly, an extension of these algorithms that can compare two populations directly improves their constraining power and broadens their applications.

The maximum mean discrepancy (MMD), a family of distance metrics, has been developed to compute the distance between two multi-dimensional distributions. Gretton et al. (2012a) proposed a non-parametric and kernel-based estimator of MMD given data drawn from two distributions. This estimator relies on embedding the sample into a high dimensional "feature space", known as a reproducing kernel Hilbert space (RKHS). This technique has been successfully applied to a diverse set of problems in machine learning literature (*e.g.,* Mitrovic et al., 2016; Muandet et al., 2016; Ramaswamy et al., 2016; Li et al., 2017). Gretton et al. (2012a)'s estimator does not rely on intermediate density estimation or data binning to estimate the distance between two probability distributions. Most importantly, this estimator comes with theoretical guarantees. It is unbiased, is computationally tractable, and has a fast convergence rate (Sriperumbudur et al., 2012). These features make this novel distance estimator an enticing tool for a broad range of astronomy problems, including data mining, hypothesis testing, and inference.

The purpose of this work is threefold. First, we introduce the MMD estimator of Gretton et al. (2012a) and compare and contrast it with other popular methods in astronomy literature. With a set of controlled numerical experiments, we show the discriminating power of this estimator. Second, we provide an application example in astronomy data analysis. As an illustrative example, we compare the optically-selected galaxy clusters identified from the Sloan Digital Sky Survey (SDSS) DR8 data and the Dark Energy Survey (DES) SV data. This direct comparison at the population level allows us to perform a quality check on samples derived from different surveys. Last, we release the open-source software Two-sAmple TesT EstimatoR (TATTER,).[1] TATTER is an implementation of the estimators discussed in this paper. This tool allows the user to perform two-sample hypothesis test seamlessly. We hope that exposure to this powerful method encourages the astronomy community to develop novel algorithms and models for analyzing large, multi-dimensional data sets.

In Section 2, we begin with the problem setup and introduce the two-sample hypothesis test estimators employed in this work. In Section 3, we show the performance of these estimators in a set of numerical experiments; and then illustrate an application of the MMD method in Section 3.4. We conclude this work in Section 5. There are two appendices. In Appendix A, we argue why going to a higher dimensional "feature space" is propitious. Finally, we discuss the implemented estimators' convergence rate of the proposed estimators and computational cost in Appendix B.

Notation: Throughout this work, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is employed to denote the inner product in a Hilbert space, not an expectation value. The expectation value with respect to probability measure p is denoted by $\mathbb{E}_{x\sim p}[.]$.

2. Problem setup

In this section, we introduce notations and describe essential preliminaries required to understand the proposed kernel-based two-sample test estimator. The proof of most of the theorems mentioned in this work is provided in Gretton et al. (2012a); thus, we do not replicate them here. We adopt their notation, so the reader can easily follow the proofs without the need to switch between the notations.

2.1. Notation and problem setup

Suppose we have samples $\{x_i \mid x \in \mathcal{X}, i \in [1, \dots, m]\}$ and $\{y_i \mid y \in \mathcal{X}, i \in [1, \dots, n]\}$ drawn independently from distributions p and q, respectively. The domain of observables[2] x and y is denoted with $\mathcal{X}$. $p(x)$ and $q(y)$ describe the probability density function of the observable samples x and y, respectively. We assume a finite sample setting meaning that we do not have access to p or q directly; instead, we have random draws from each of them. The two-sample hypothesis testing problem asks if $p = q$ given the two sets of observed data.

A measure of Maximum Mean Discrepancy (MMD) squared is defined as

$$\begin{aligned}\mathrm{MMD}^2[k, p, q] = {} & \mathbb{E}_{x,x'\sim p}[k(x, x')] + \mathbb{E}_{y,y'\sim q}[k(y, y')] \\ & - 2\mathbb{E}_{x\sim p, y\sim q}[k(x, y)],\end{aligned} \tag{1}$$

where $\mathbb{E}_{x\sim p}[.]$ is the expected value under the probability measure $p(x)$, $k(x, x')$ is a kernel function defined in a reproducing kernel Hilbert space (RKHS), and x' and y' are independent copies of x and y. According to Gretton et al. (2012a), two probability densities are the same if and only if $\mathrm{MMD}^2[k, p, q] = 0$. There is no restriction on the dimensionality of the domain space $\mathcal{X}$ which automatically solves one of the key limitations of the K–S test. A few examples of RKHS and an interpretation of the MMD distance are provided in Appendix A.

In our setting, $\boldsymbol{x}_{1:m}$ and $\boldsymbol{y}_{1:n}$ are two observed samples from experiment A and experiment B, where x and y are d-dimensional

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

[2] Observable and random variables are used interchangeably.

data vectors with m and n samples, respectively.[3] x_i and y_j are the ith and jth observed data points from the samples x and y, respectively. For example, a sample of galaxies selected from an optical-survey can be described with a six-dimensional data vector $x_i = (M_{u,g,r}$, size, redshift, ellipticity), where M_u, M_g, and M_r are brightness in u, g and r bands, respectively.

MDD Estimator: An unbiased estimator of Eq. (1) in a finite sample setting is

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

where subscript u indicates the fact that this is an unbiased estimator of MMD^2 (Gretton et al., 2012a). We discuss the computational costs and convergence rate of this estimator in Appendix B. Convergence rate defines how fast the estimator's variance drops to zero as a function of the sample size.

The user has the freedom to choose the kernel function $k(.,.)$ and set its hyper-parameters. Representation of data with an appropriate choice of kernel $k(.,.)$ has been shown to preserve statistical information about the distribution (Song et al., 2013, and for a simple justification see Appendix A). However, the performance of the MMD estimator is sensitive to this choice.

The kernel employed in this work is a Gaussian Radial Basis Function (RBF) kernel (see Table 3.1 in Muandet et al., 2017, for a list of popular kernels and their properties). The Gaussian RBF kernel has the form of

$$k(x, y) = \exp(-\gamma\,\|x - y\|^2), \tag{3}$$

where γ a hyper-parameter set by the user.

Kernel Setup: The Gaussian kernel has been extensively used in the machine learning literature and has delivered promising results in a broad range of applications. One reason is that it maps the domain space into an infinite-dimensional space (see Appendix A). This allows the algorithm to capture complex, non-linear interactions between different observables. An astronomer might prefer to construct a kernel function motivated by the physics of their problem (e.g., a kernel function with periodic nature). The Gaussian kernel does not necessarily accommodate dimensions with radically different units. We recommend that the user renormalizes the units in practical applications. The Gaussian kernel has one hyper-parameter that needs to be tuned. The optimization of this hyper-parameter depends heavily on the application and is an active area of research. We do not attempt to optimize the results for this hyper-parameter. In this work, we set γ to

$$\gamma^{-1} = 2 \times \mathrm{median}[\{D^2(x_i, y_j) | \forall\, i \in m, j \in n\}], \tag{4}$$

where $D(x_i, y_j)$ is the pair-wise Euclidean distance between data points x_i and y_i. We find that this choice is performing reasonably well in our numerical examples.

The MMD method is non-parametric, as it does not require the form of p or q to be defined. As expected, this method is symmetric under interchange of p and q. Except in evaluating the kernel function, Eq. (2) is independent of the data dimension; hence, the algorithm is scalable to high-dimensional data. The MMD estimator has a fast convergence rate, with the convergence rate of $\mathcal{O}((m+n)^{-\frac{1}{2}})$. We compare the convergence rate of the MMD estimator with the K–S and K–L estimators in Appendix B.

2.2. Estimation of the statistical significance of the null hypothesis

In a two-sample test problem, the null hypothesis is $\mathcal{H}_0 : p = q$, and the alternative hypothesis is $\mathcal{H}_A : p \neq q$. If the probability of observing the data, given the test statistic distribution under the null hypothesis, exceeds a pre-determined threshold, then we can reject the null hypothesis. The estimator's variance under the null hypothesis directly impacts the Type II errors for a given significance threshold and other specifications. A Type I error occurs when the null hypothesis is rejected, while both data sets are generated from the same distribution (a false-positive error). It is equivalent to the chosen excess threshold. A Type II error is made when the distributions differ, but the null hypothesis is not rejected (a false-negative error). It is related to the power of the testing. There are recommended designed choices to minimize Type II error under special specifications (*e.g.*, Gretton et al., 2012b). Discussing these optimization strategies is beyond the scope of this work, but it is worth noting that there does not exist a universally good optimization algorithm, and often there is a bias–variance trade-off. In this work, we do not attempt to optimize for Type II error as (1) it would increase the chance of unintended confirmation bias, and (2) the specifications could vary from one problem to another.

A p-value is the probability of the null distribution exceeding the test statistic computed from the data. With this definition, the p-value in our problem is computed via

$$1 - p = \Pr(\mathrm{MMD}^2_{\mathrm{null}}[k, x, y] < \mathrm{MMD}^2_{\mathrm{data}}[k, x, y]), \tag{5}$$

where $\mathrm{MMD}^2_{\mathrm{null}}[k, x, y]$ and $\mathrm{MMD}^2_{\mathrm{data}}$ specify the null distribution and an estimation of the test statistic for actual data. Here, we employ a bootstrap algorithm to quantify the probability density function for the MMD^2 statistic under the null hypothesis. Algorithm 1 shows the steps of our bootstrap algorithm, which is illustrated in Fig. 1. First, an aggregated data vector is constructed and re-shuffled. The re-shuffled aggregated data vector is split into two data-like vectors of size m and n. Finally, the estimator in Eq. (2) is performed on this bootstrap realization. The above steps are repeated N_b times to get an estimation of the probability density function under the null hypothesis. The p-value is computed by comparing the estimated MMD with this null distribution.

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

```
Input: $x, y, N_b, k(.,.)$: observed samples, the number of bootstraps, and the kernel function.
Output: $\widehat{\mathrm{MMD}}^2$, Null, $p$-value: an estimation of the $\mathrm{MMD}^2$ for the observed samples, drawn from the null distribution, and $p$-value.
initialize the hyper parameters. ($\gamma$, see Eq. (4))
$\widehat{\mathrm{MMD}}^2 = \mathrm{MMD}^2(x, y, k)$: compute $\mathrm{MMD}^2(x, y, k)$ with Eq. (2)
Z $\leftarrow$ aggregate observed samples.

for $i$ in $\{1, \cdots, N_b\}$ do
    $x_{\mathrm{boot}} \leftarrow$ randomly draw $m$ data points from $Z$ (with replacements)
    $y_{\mathrm{boot}} \leftarrow$ randomly draw $n$ data points from $Z$ (with replacements)
    Null[$i$] $\leftarrow \mathrm{MMD}^2(x_{\mathrm{boot}}, y_{\mathrm{boot}}, k)$
end for

$p$-value = count(Null > $\widehat{MMD}^2$)/$N_b$
```

There are approximation methods to compute Type I and Type II errors (e.g., Gretton et al., 2009), for example, by fitting a

[3] Instead of experiments A and B, a sample can be constructed from a simulation, generated through a generative model, or drawn from a distribution. This method is not restricted to experimental settings.

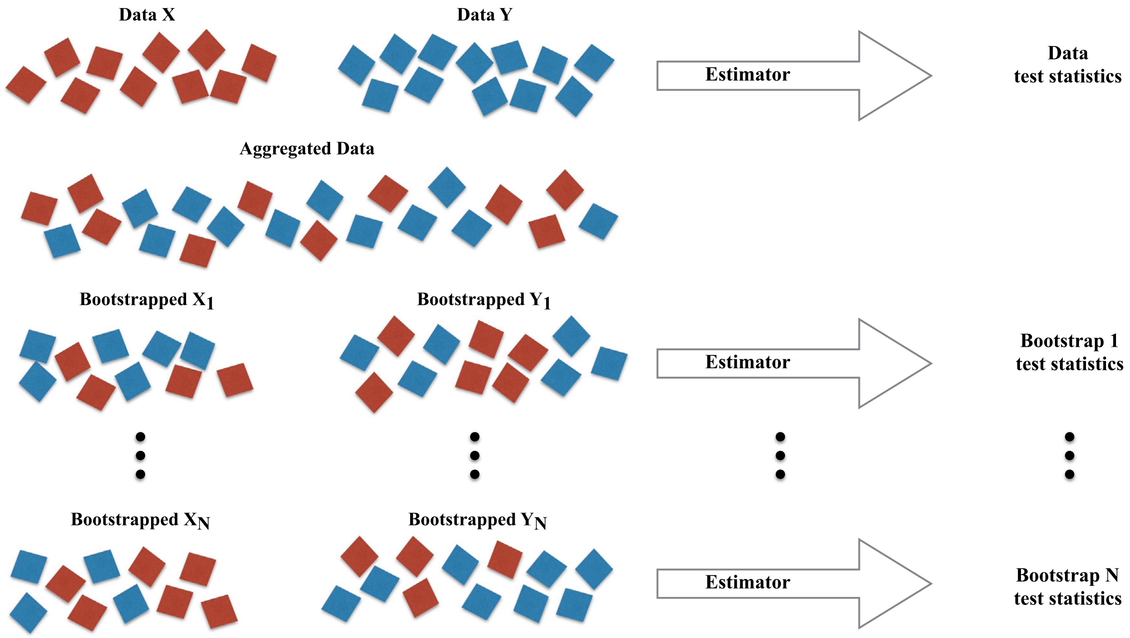

Fig. 1. An illustration of the bootstrap procedure. Each box represents one data point. First, an aggregated data vector is constructed. Then, the aggregated data vector is re-shuffled and split into two data-like vectors of size m and n, where m and n are the sizes of the original data vectors. We estimate the probability density function of the MMD^2 statistic under the null hypothesis from the bootstrap realizations.

Pearson distribution to its first four moments (Gretton et al., 2012a). These methods are inaccurate and can be computationally as expensive as the bootstrap method. Our results suggest that the bootstrap method performs quite well. Thus, for practical applications, the simulation-based bootstrap method might be favored, even with its computational overhead. An implementation of Algorithm 1 is provided through TATTER.

2.3. Alternative test statistics

The key ingredient of the two-sample test is the choice of the test statistic. In the astronomy literature, the K–S test statistic is prevalent and has been extensively exploited. This test, which is a non-parametric integral probability metric, characterizes the difference between the two CDFs. Specifically, the K–S statistic is the maximum value of the absolute difference between two CDFs

$$D_{KS} = \sup_{x\in\mathcal{X}} |\mathrm{CDF}(p(x)) - \mathrm{CDF}(q(y))|, \tag{6}$$

where $\sup_x$ is the supremum of the set of distances.

An alternative test statistic is the K–L divergence. The K–L divergence is an entropy-based approach that recently becomes popular among astronomers and high-energy physicists (Ben-David et al., 2015; De Simone and Jacques, 2019). The K–L divergence (D_{KL}),

$$D_{KL}(p \mid q) = \int_{\mathcal{X}} p(x) \ln\left[\frac{p(x)}{q(x)}\right] \mathrm{d}x\,, \tag{7}$$

computes a directional difference between a reference probability distribution p and a target probability distribution q. The K–L divergence is not invariant under exchange of p and q, thus is not a distance metric. Due to this directionality feature, the K–L divergence is inappropriate for the problem of two-sample hypothesis testing; however, it is employed in the literature to perform such a test (*e.g.*, Ben-David et al., 2015; De Simone and Jacques, 2019). This feature can be a limiting factor in certain applications that require directional symmetry. To address this limitation, we consider alternate ways of "averaging" the two K–L divergences. The so-called J-divergence is equal to the average of the two possible K–L divergence between two probability distributions

$$J(p, q) = \frac{D_{KL}(p \mid q) + D_{KL}(q \mid p)}{2}. \tag{8}$$

We compare the performance of Eq. (8) with the other two test statistics, and in the rest of this work, we refer to this symmetric test statistic as the K–L test statistic. Unlike the K–S test, the K–L test can handle multi-dimensional distributions.

A finite sample setting requires an estimator as the analytic forms of p and q are out of reach. The MMD estimator is introduced in Section 2.1. To estimate the K–S value, we use a Python implementation from the Scipy package.[4] We employ Wang et al. (2006) implementation to estimate the K–L divergence. Wang et al. (2006) proposed an unbiased k-nearest neighbors (kNN) estimator[5] of the K–L test statistic in a finite sample setting. Wang et al. (2006) also show that their estimator has a faster convergence rate with respect to the existing algorithms, and can handle multi-dimensional data. We set the hyper-parameter $k = 1$ in this work. We test the sensitivity of our results to the choice of this hyper-parameter and find that our results and conclusions remain unchanged. TATTER provides an implementation of the estimators discussed here and allows the user to perform a two-sample hypothesis test with Algorithm 1.

T-test is another popular testing procedure to perform hypothesis testing and is used to determine if two population means are equal. The t-test statistic is the optimal test statistic when we are dealing with two Gaussian distributions of same variance. We use it to benchmark the proposed test statistics in a Gaussian setting. The two sample t-test relies on the test statistic defined as

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}}, \tag{9}$$

[4] https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.ks_2samp.html.

[5] https://github.com/slaypni/universal-divergence.

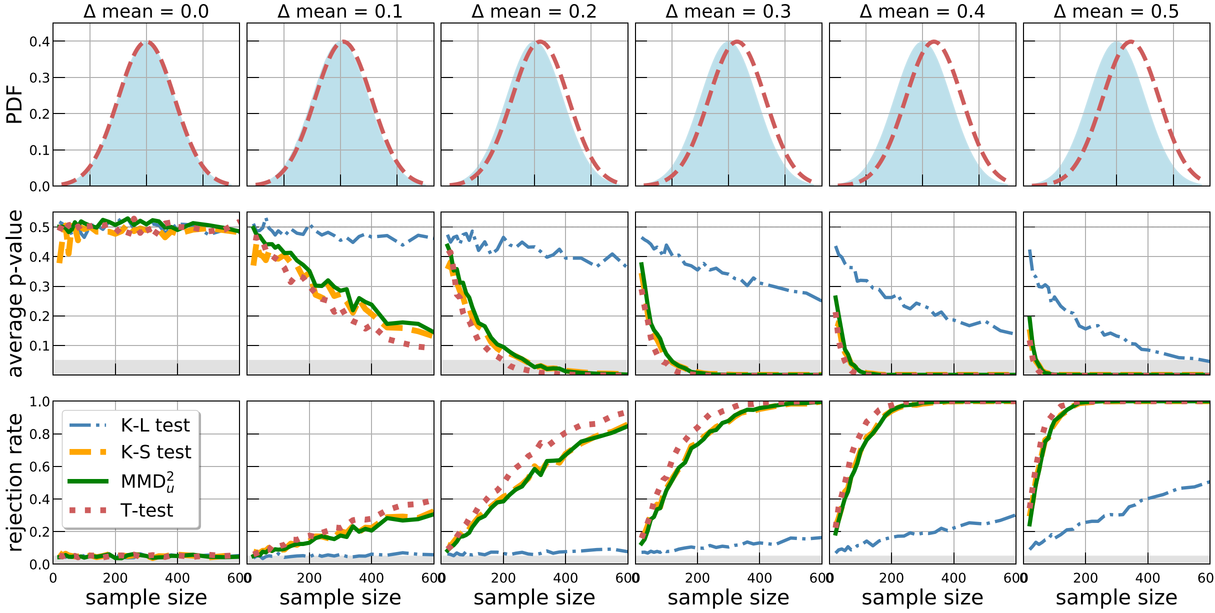

Fig. 2. The discriminating power of each test statistic as a function of sample size. **Top Panels:** The true Gaussian distributions used to draw samples from. The mean difference between the two Gaussian distributions of unity variance is given in the title of each column. **Middle Panels:** Average p-value vs. sample size per test statistic. Each sub-figure presents the average p-value over 1000 random data realizations. The gray shaded region shows the < 0.05 p-value significance region. **Bottom Panels:** Statistical power vs. sample size per test statistic. Each sub-figure presents the rejection rate of the null hypothesis given p-value threshold of 0.05. The gray region corresponds to the $< 0.05\%$ rejection rate region.

where $\bar{x}$ and $\bar{y}$ are the empirical means, s_x^2 and s_y^2 are the empirical variances, and m and n are the sample sizes. Under the null hypothesis, the t-statistic follows a Student-t distribution. To reject the null hypothesis, one can compute the probability of the tail to derive the p-value.

The statistical significance, i.e. p-value, of the K–S test and K–L divergence is computed via the bootstrap method proposed in Section 2.2. In Eq. (5), we replace the test statistic estimator and the rest remains the same. We note that the Scipy implementation of the K–S test outputs a p-value. Since this p-value relies on a two-sided asymptotic K–S distribution, we refrain using this precomputed p-value. Our proposed bootstrap method yields a more accurate estimation of the p-value when the data is not in asymptotic regime, and allows the user to have a consistent way of comparing different test statistics.

3. Controlled simulation setting

In this section, we study the performance of each test statistic discussed in the previous section. We perform a set of controlled numerical experiments where data points x and y are drawn from two known distributions. We also discuss the strengths and limitations of each estimator. Finally, we provide an application of the proposed MMD estimator in astronomy data analysis.

In the following, we consider two probability distributions, p and q, where $p \neq q$. Then, we draw m samples from each distribution, where $m = n$. Given these samples, we ask if a test statistic can reject the null hypothesis. To quantify the performance of each test statistic, the average p-value is estimated for each experimental specification. Type II error occurs when $p \neq q$, but the test statistic was not sufficient to rule it out. Smaller average p-value corresponds to a lower Type II error. For a finite sample size, the occurrence of Type II error is inevitable, but with a powerful test statistic, we can minimize it. An alternative way to show a Type II error is through the rejection rate. The rejection rate is the fraction of numerical experiments for which the null hypothesis is rejected. Here, Type II error is equivalent to $1 -$ rejection rate. When $p \neq q$, a rejection rate close to 1 is desired.

3.1. Two Gaussian distributions

In the first setting, we investigate the performance of MMD, K–S, and K–L tests with two samples drawn from two Gaussian distributions of same variance but with different mean values. The domain space is $\mathcal{X} = \mathbb{R}$ and p and q are two Gaussian distributions with unity variance. Their means are set to zero and $\triangle$Mean, respectively. Given p and q, m data points are randomly drawn from each distribution. Then, Algorithm 1 is employed to compute the p-value for each test statistic. We repeat these two steps 1000 times to get 1000 realizations and their corresponding p-values. We perform the steps above for multiple sample sizes and $\triangle$Mean's, and report the average p-value per test statistic and use the rejection rate as our performance metric.

Fig. 2 presents the performance of each metric for this setting. The Gaussian's mean difference, $\triangle$Mean, is given in the title of each sub-figure. Only when $\triangle\text{Mean} = 0$ the null hypothesis is correct, $p = q$, for the rest of the sub-figures the null hypothesis is not true, $p \neq q$. The middle panel shows the trend of average p-value for the MMD, K–S, K–L, and t-test as a function of sample size and $\triangle$ Mean. The gray shaded regions correspond to < 0.05 p-value area. The bottom panel shows the rejection rate of the null hypothesis assuming a p-value threshold of 0.05.

As expected, when the null hypothesis is true (first column), the average p-value for all test statistics is 0.5, and the Type-I error (rejection rate) is 5%. This finding implies that all our bootstrap procedure gives us an unbiased estimation of the null distribution. When the null hypothesis is false, as expected, the t-test outperforms the non-parametric methods in rejecting the null hypothesis and the K–L test has the worst performance. The t-test estimator is a parametric approach with strong, restrictive assumptions; namely, the two samples need to be drawn from two Gaussian distributions with the same variance. Under these assumptions, t-test is the optimal test statistic. These results also suggest that the employed K–L divergence estimator has the least discriminating power in a finite sample setting.

Fig. 3 presents the performance of MMD in a multivariate normal distribution setting. Like the one-dimensional setting,

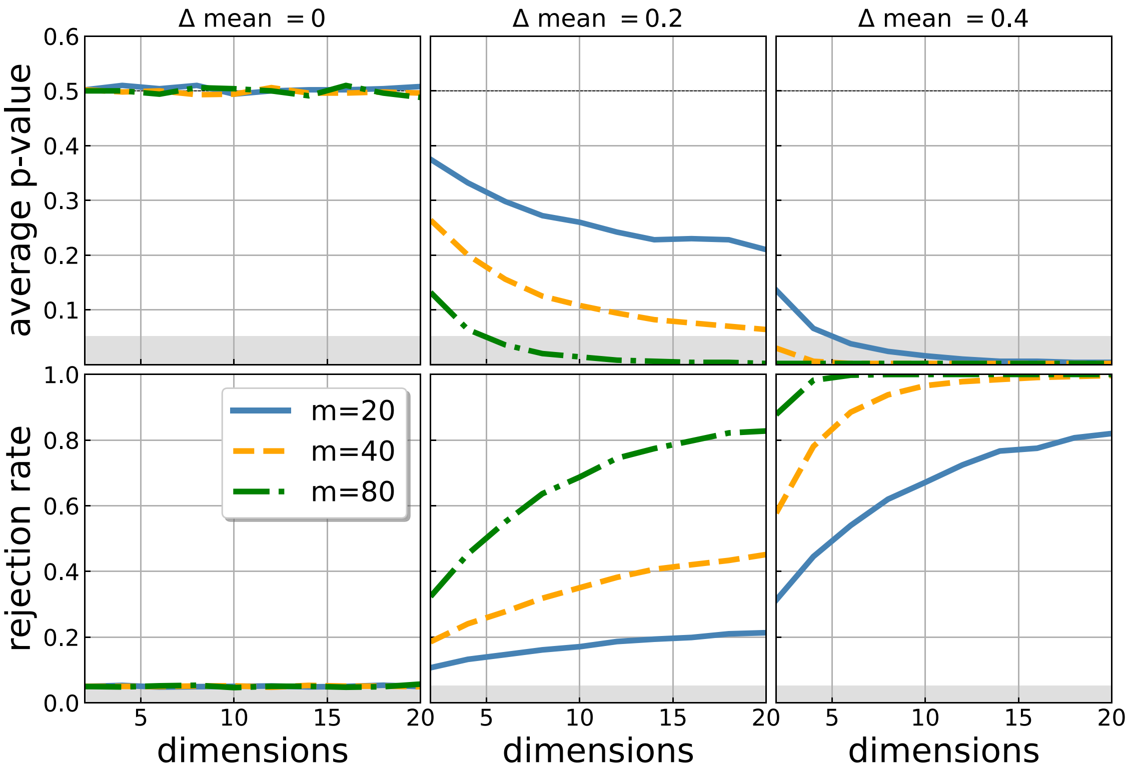

Fig. 3. Performance of the MMD test statistic in a multi-dimensional setting. Data are drawn from two multivariate normal distributions of diagonal unity covariance and with mean difference $\Delta\overrightarrow{\text{mean}} \in \{[0]^d, [0.2]^d, [0.4]^d\}$. The top panels are the average p-values and the bottom panels are the rejection rates for data of size $m \in \{20, 40, 80\}$.

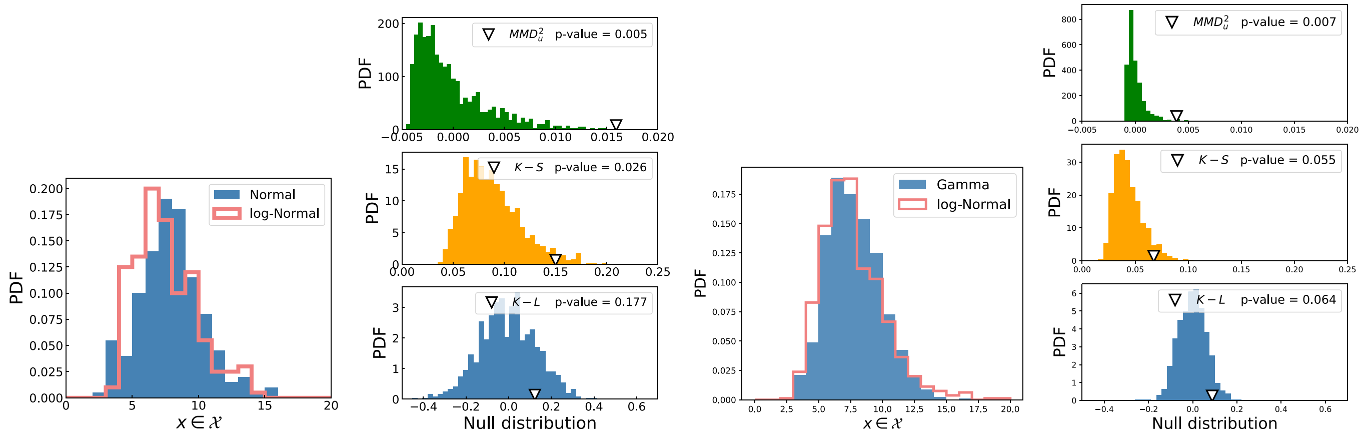

Fig. 4. **First and Third Panels**: The distribution of one random realization of data. **Second and Fourth Panels**: The null distribution and the estimated MMD, K–S, K–L divergence test statistics for data shown in their left panel. **Two Left (Right) Panels:** Compares a Gaussian (Gamma) distribution with a log-Normal distribution of matched mean and variance.

the difference between two multivariate normal distributions is in their mean vector. Furthermore, we assume no correlation between different Gaussian components. Fig. 3 shows the average p-value and the rejection rate of 2000 random realizations of two multivariate normal distributions with $\Delta\overrightarrow{\text{mean}} \in \{[0]^d, [0.2]^d, [0.4]^d\}$ where d is the data vector dimension. This example demonstrates the effectiveness of the MMD test statistic in a multi-dimensional setting. The user should check the performance of MMD under their model specification to ensure this is appropriate for their application.

It is important to note that the discriminating power of the MMD test statistic might improve with an appropriate choice of the kernel function and hyper-parameter. We do not attempt to optimize the kernel function for the reasons mentioned in Section 3.4. For example, since the only difference between p and q, in this example, is in their means, a Gaussian kernel might be too flexible (see Appendix A for an interpretation of the Gaussian kernel).

3.2. Asymmetric distributions

Next, we compare a Gaussian and a Gamma distribution with a log-Normal distribution. We match the means and variances of the distributions. Similar to the above experiments, first, m samples are drawn from each distribution. Then, the p-value is computed for each test statistic via Algorithm 1. Finally, the average p-value is computed for 1000 data realizations. An example of a single data realization is shown in Fig. 4, and performance

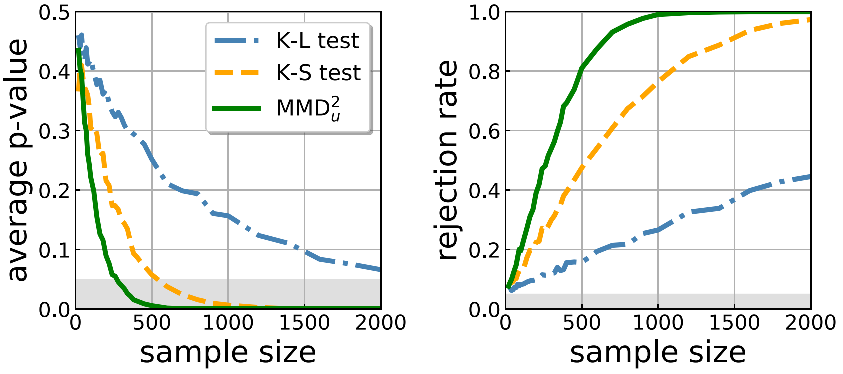

Fig. 5. Performance of the MMD, K–S, and K–L tests as discriminator of a log-Normal distribution and a Gaussian distribution of matched mean and variance. **Left Panel**: The average p-value as a function of the sample size. The gray shaded region shows < 0.05 significance region. **Right Panel**: Rejection rate as a function of sample size.

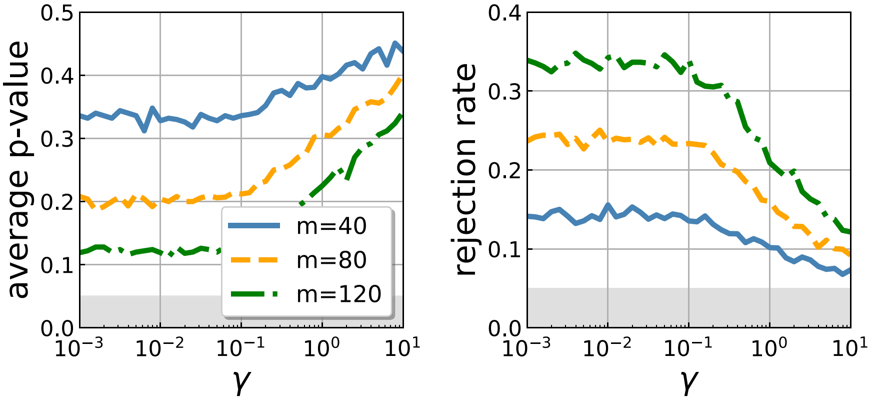

Fig. 6. Performance of MMD as a function of the kernel width inverse. **Left Panel**: The average p-value. **Right Panel**: Statistical power of MMD. Different colors (line-types) correspond to different sample sizes.

of each test statistic is presented in Fig. 5. Since the mean and variance of the two distributions are matched, only a test statistic sensitive to higher-order statistics can effectively reject the null hypothesis.

Fig. 4 presents one single realization with 200 random samples from each distribution. The left panels show the distribution of this realization. The right panels show the distribution of the null hypothesis with 5000 bootstrap re-sampling of the data for each test statistic. The MMD estimator achieves the lowest average p-value and the K–L test has the poorest performance in this setting. This experiment illustrates the fact that the MMD estimator can achieve good performance while the mean and variance of the two non identical distributions are the same. We note that Gamma and log-normal distributions are both asymmetric distributions with a lower bound of zero (the two right panels of Fig. 4).

Fig. 5 shows the rejection rate of the MMD (green line), K–L (blue dotted–dashed line), and K–S (dashed yellow line) test statistics. While the K–S test requires ∼2000 samples to reject the null hypothesis with a rejection rate close to $>$ 95%, the MMD estimator achieves a rejection rate of 95% with less than 1000 data points.

3.3. Sensitivity analysis to the kernel width

The MMD test statistic's performance is sensitive to the choice of kernel function and its hyperparameters. The kernel employed in this work has one free parameter, γ, which is equivalent to the inverse square of a Gaussian kernel width. While in this work, we do not attempt to optimize the two sample hypothesis test for this hyper-parameter, in this section, we explore the sensitivity of two sample test to the Gaussian kernel width. Similar to Section 3.1, we employ two Gaussian distributions of unity width and mean difference Δ mean $= 0.2$ to draw our data points.

Fig. 6 shows the average p-value (left panel) and the rejection rate (right panel) based on 2000 data realizations as a function of the inverse of kernel width. As the kernel width becomes smaller ($\gamma > 1$), MMD becomes less optimal and loses its statistical power. By increasing the kernel width, the average p-value and rejection rate improves and then plateau ($\gamma < 0.1$). Two Gaussian distributions of the same width have only one free parameter, and that is the difference between their means. Therefore, larger the kernel corresponds to less contribution from the higher-order statistics, and eventually, the test becomes dominated by the mean statistic (see Appendix A). Increasing the kernel width in a Gaussian setting makes it more optimal and its statistical power improves. But care should be taken in generalizing this argument to non-Gaussian distributions. By definition, a non-Gaussian distribution has higher-order statistics, implying that these curves do not plateau. There is an optimal kernel width which depends on the sample size and the density shape of distributions.

3.4. An application in astronomy

The two-sample hypothesis test is an essential tool in studying features of data derived from multiple surveys or comparing simulation outcomes with empirical data. Comparing properties of identified sources from multiple surveys and comparing their distributions allow us to assess the homogeneity of the survey products and learn about systematics. The two-sample hypothesis test may be used to quantify if combining two samples is

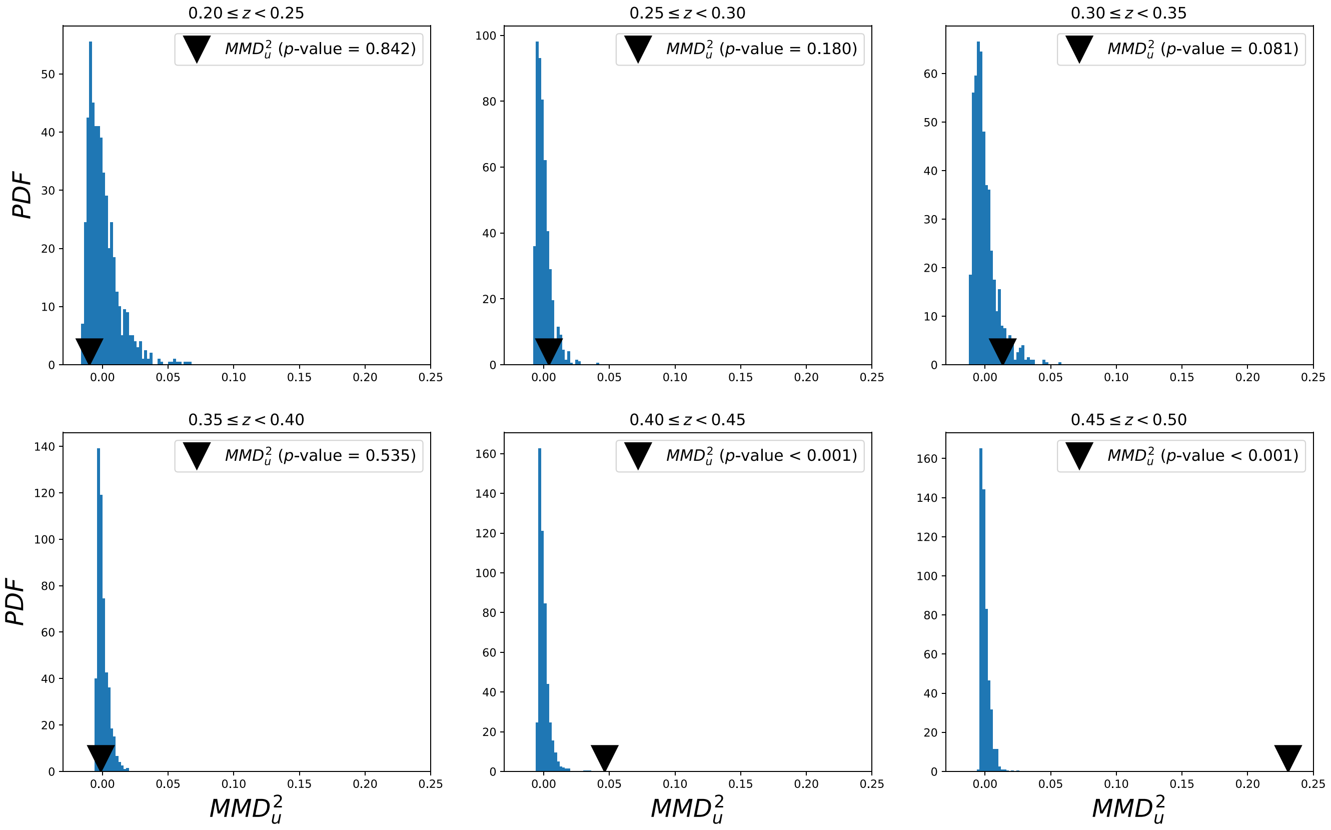

Fig. 7. Pairwise comparison of the observables of the redMaPPer cluster catalogs derived from the SDSS DR8 data and DES-SV galaxy catalogs for six redshift bins. The observables are optical-richness and redshift, $x = (\lambda_{\rm RM}, z)$. The blue histograms are the null distributions, $P_{\rm SDSS}(\lambda_{\rm RM}, z) = P_{\rm DES\ SV}(\lambda_{\rm RM}, z)$, and the markers show the estimated MMD distance.

statistically justified. For example, two data sets are additive if they are drawn from the same distribution. In this section, we provide one illustrative example of the two-sample hypothesis test with the MMD method in astronomy. The purpose of this section is to give the reader a flavor of applications of two-sample hypothesis test with the MMD method. In an online tutorial of the code, we provide extra examples such as comparing labeled images of handwritten digits.

Can we combine galaxy cluster samples derived from the SDSS and DES surveys? Here, we take galaxy clusters as an example. Clusters of galaxies, hosted by dark matter halos, are the most massive structures held together by gravity (Allen et al., 2011). The population statistic of these systems is a sensitive probe of the cosmological parameters. It is essential to compare and contrast the statistical properties of samples of clusters identified from large-scale surveys. Suppose there are surveys A and B. If the galaxies are selected in the same way and the same cluster-finding algorithm is applied to these two data sets, we do not expect the joint distribution of cluster observables to be systematically different. If we find that they are significantly different then there are potential systematics that have not been accounted for and cluster samples cannot be combined.

In Fig. 7, we evaluate the MMD distance between distributions of the redMaPPer cluster observables, optical-richness and redshift $x = (\lambda_{\rm RM}, z)$, identified from the Sloan Digital Sky Survey (SDSS, Rykoff et al., 2014) DR8 data and the Dark Energy Survey (DES, Rykoff et al., 2016) Science Verification (SV) data.[6] $\lambda_{\rm RM}$ is a measure of the number of red-sequence galaxies above a luminosity threshold. MMD is insensitive to the overall normalization or the total number of systems. Thus there is no need to correct for the difference in the survey area.

In Fig. 7, the blue histograms show the bootstrapped null distributions, i.e. $P_{\rm SDSS}(\lambda_{\rm RM}, z) = P_{\rm DES\ SV}(\lambda_{\rm RM}, z)$, and the markers show the estimated MMD distance. The estimated MMD distance between the DES and SDSS cluster samples for low redshift bins, $z < 0.4$, is consistent with the null hypothesis, while cluster statistics are significantly different for high redshift, $z > 0.4$, samples. The DES SV sample is deeper and has better photometry compared to the SDSS sample. The SDSS galaxy sample is not reliable for finding clusters with $z > 0.4$. This is consistent with the results presented in Fig. 7. Low (high) redshift clusters derived from SDSS and DES SV data have the same (different) underlying joint distribution. These results suggest that there are systematics in one or both data sets which have altered the joint distribution of observables for high redshift clusters.

4. Discussion

Our results suggest that the MMD distance estimator with a Gaussian kernel consistently outperforms a K–L divergence-based estimator in the context of the two-sample hypothesis test. For one-dimensional data, the K–S test and MMD test have similar performance. For multi-dimensional data, the MMD test is a more reliable choice than the divergence-based tests.

We note that the MMD estimator in Eq. (2) is not an optimal estimator, and its performance is sensitive to the kernel choice and its hyperparameters. Suppose the forms of p and q are known. In that case, one can develop a parametric but an optimal estimator (the Neyman–Pearson lemma, Neyman and Pearson, 1933). Distance based test statistics, such as MMD, are

[6] http://risa.stanford.edu/redmapper/.

not the only class of multivariate test statistics. Another class of test statistics employs rank statistics. Rank-based test statistics are also proposed to perform two-sample hypothesis test (*e.g*, Mondal et al., 2008; Modak and Bandyopadhyay, 2019).

In the past few years, there has been a proliferation of applications of the K–L divergence in astronomy (*e.g.*, Bovy, 2010; Ramos Almeida et al., 2011; Seehars et al., 2014; Ben-David et al., 2015; Sanderson et al., 2015; Charnock et al., 2017; Wang et al., 2017; Zhao et al., 2017; Nicola et al., 2019). Our code provides a complementary yet powerful approach for such studies in astronomy. The K–L divergence is a powerful tool in quantifying the information loss, for example, where a complex distribution is approximated with a simpler model. But in certain applications where a distance metric is more appropriate, the MMD test statistic might be a better choice.

Seehars et al. (2014) and Nicola et al. (2019) proposed the K–L divergence as a quantifier of tension between two posterior distributions. They dubbed this new quantifier the "surprise". They used a Gaussian statistic to calibrate and compute the statistical significance of the "surprise" values. Zhao et al. (2017) employed this approach and computed the "surprise" value for posteriors estimated with several cosmological probes. Zhao et al. (2017) then concluded that the employed observational data favor the dynamical dark energy model over the constant dark energy model. A limitation of this approach is that the K–L divergence is not a proper metric, i.e. asymmetric under interchanging experiments A and B.[7] More importantly, they assumed that the posteriors are multivariate normal distributions; hence they took a parametric approach. In this application, the MMD distance, without any modification, cannot be used as a tension metric either. MMD distance with a Gaussian kernel compares all moments of two distributions, and we do not expect two consistent posteriors have the same means, variances, and higher moments. The estimated distance is also not directly interpretable as the statistical significance of tension between two posteriors needs to be calibrated. In this application, however, an MMD-based test statistic can be employed to perform model comparison via posterior predictive checks.

While MMD test statistic is a powerful tool, it comes with its own challenges. The fact that the MMD method relies on hyper-parameters is both an opportunity and a challenge. These hyper-parameters give it flexibility and can be tuned to improve the performance of the test. The MMD method can benefit from developing physically motivated kernels and robust approaches in tuning the kernel hyper-parameter(s). However, extreme caution should be exercised in designing the kernel and the tuning procedure before applying it to data. One can tune the hyper-parameters until it reaches the desired results. To prevent deceptive interpretation or data dredging (Head et al., 2015), a blinding procedure is highly recommended. A blinding procedure decreases the chance of confirmation bias (MacCoun and Perlmutter, 2015), especially in applications of a two-sample test where the results might be sensitive to the choice of hyper-parameter(s). In our experiments, we decided to fix the model hyper-parameter in the beginning and did not attempt to tune, or "learn" it.

5. Conclusion

The two-sample hypothesis test is one of the most ubiquitous problems in statistical analysis and data mining. Despite its importance, there has not been significant progress in this direction in the astronomy literature. In this paper, we present the MMD method to perform the two-sample hypothesis test. MMD quantifies the distance between two multi-dimensional distributions. This novel method employs techniques from machine learning literature to construct a scalable, robust, and unbiased probability density distance estimator. This approach is non-parametric and distribution free. This work studies the properties of the MMD method and compares it with two other popular test statistics, the K–S and K–L tests. Our findings suggest that the MMD estimator with a Gaussian kernel outperforms the alternatives in a realistic finite-sample setting.

We release `TATTER` software as a part of this work. `TATTER` is an open-source software that allows user to perform two-sample hypothesis test employing the estimators discussed in this work. The source code, as well as a tutorial and examples, can be found in the public code repository. `TATTER` implementation supports multi-processing jobs and allows the user to run multiple parallel jobs on local computing machines or external clusters and supercomputers, without any effort on the user's side.

Astronomical data are multi-population, multi-dimensional, time-varying, and large-scale in nature. To extract pattern and information from such complex data sets, it is a necessity for the community to develop novel, scalable, and robust algorithms. Given the features of the MMD method, we are hopeful that exposure to this new method will empower the community to tackle new scientific problems and will be employed in addressing further problems in astronomy, ranging from issues in data mining, model comparison, and hypothesis testing to developing new inference algorithms.

CRediT authorship contribution statement

A. Farahi: Conceptualization, Analysis, Validation, Methodology, Software, Visualization, Writing - review & editing. **Y. Chen:** Supervision, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by a Michigan Institute for Data Science (MIDAS) Fellowship, United States and NSF, United States DMS 18–11083. We are grateful to August Evrard and Dragan Huterer for their insightful feedback on the early version of the draft. We thank Sergey Koposov, and Markus Michael Rau for providing helpful feedback. We thank Mattew Ho for reading the draft and providing constructive suggestions. We finally would like to thank Dhayaa Anbajagane for testing an early version of the software.

Appendix A. An interpretation of RKHS and the MMD distance

Suppose we have distributions p and q over a set $\mathcal{X}$. The MMD is defined by a feature map $\phi : \mathcal{X} \to \mathcal{H}$ where $\mathcal{H}$ is an RKHS. The MMD value can be computed via

$$\mathrm{MMD}[p, q] = \|\mathbb{E}_p[\phi(x)] - \mathbb{E}_q[\phi(x)]\|_{\mathcal{H}} = \|\mu_p - \mu_q\|_{\mathcal{H}}. \tag{A.1}$$

where $\mu_{p/q} = \mathbb{E}_{p/q}[\phi(x)]$. Now suppose $\phi(x) = x$ and $\mathcal{X} = \mathcal{H} = \mathbb{R}^n$. Then MMD becomes

$$\mathrm{MMD}[p, q] = \|\mathbb{E}_p[\phi(x)] - \mathbb{E}_q[\phi(x)]\|_{\mathbb{R}^n} = \|\mathbb{E}_p[x] - \mathbb{E}_q[x]\|_{\mathbb{R}^n}. \tag{A.2}$$

This MMD is simply the norm distance between the means of the two distributions. Therefore, this test cannot distinguish between

[7] Surprise(A, B) $\neq$ Surprise(B, A), where A and B are two different experiments.

two different distributions with similar mean. This is not ideal as this mapping is not general and has limited applications. Now, suppose there is a projection map $A_{n'\times n}$ where $n' < n$, i.e. $\phi(x) = Ax$, which maps the random variable x to a lower dimension. This MMD leads to

$$\mathrm{MMD}[p, q] = \|\mathbb{E}_p[\phi(x)] - \mathbb{E}_q[\phi(x)]\|_{\mathbb{R}^n} = \|A(\mathbb{E}_p[x] - \mathbb{E}_q[x])\|_{\mathbb{R}^n}. \tag{A.3}$$

This MMD is even less discriminate, as it just compares the projected means of distributions in a sub-space of the domain space. Mapping the random variable x to a higher dimensional space is more powerful. Suppose $\phi(x) = (x, x^2)$, the corresponding MMD is

$$\begin{aligned}\mathrm{MMD}[p, q] &= \|\mathbb{E}_p[\phi(x)] - \mathbb{E}_q[\phi(x)]\|_{\mathbb{R}^n}\\ &= \|\left(\mathbb{E}_p[x] - \mathbb{E}_q[x],\ \mathbb{E}_p[x^2] - \mathbb{E}_q[x^2]\right)\|_{\mathbb{R}^{2n}}\\ &= \sqrt{(\mathbb{E}_p[x] - \mathbb{E}_q[x])^2 + (\mathbb{E}_p[x^2] - \mathbb{E}_q[x^2])^2}.\end{aligned} \tag{A.4}$$

The above MMD is more powerful than the first two as it contains more information, mean and variance. But still this is not sufficient to distinguish between distributions with similar mean and variance but different higher moments statistics. Taking x into a higher dimensional space increases the number of moments that appears in the MMD, which leads to an enhancement in the discriminating power of MMD.

The Gaussian kernel is proposed as a robust kernel to compute the MMD test statistic. The corresponding feature map of this kernel has infinite dimension, which implies that its embedded RKHS contains information about all moments of these two distributions. In theory, with infinite sample size, MMD via the Gaussian kernel is able to distinguish between any two non-identical probability distributions. This kernel has the form of

$$k(x, y) = \exp(-\gamma\,\|x - y\|^2), \tag{A.5}$$

with γ as a hyper-parameter. Its corresponding feature map, assuming x is one-dimensional, is

$$\phi(x) = \exp(-\gamma x^2)\left[1, \sqrt{\frac{2\gamma}{1!}}x, \sqrt{\frac{4\gamma^2}{2!}}x^2, \sqrt{\frac{8\gamma^3}{3!}}x^3, \ldots\right]. \tag{A.6}$$

In the above equation, γ can be interpreted as weight of each moment in the MMD computation. Smaller γ implies that a fewer number of moments contribute to the MMD test statistic.

Appendix B. Notes on convergence rate and computational costs of the proposed estimators

In a finite sample setting, each estimator has an inherent variance that goes to zero as the sample size increases. We define "convergence rate" as how fast this inherent variance as a function of the sample size goes to zero. This gives us a hint on how much a signal can be enhanced by increasing the sample size. Fig. 8 shows the convergence rate of the three estimators discussed in this work.

To compute the convergence rate, we take three steps. First, two sets of data points of size m are drawn from a Gaussian distribution with mean zero and unity variance. Then, we estimate their MMD distance, K–S value, and K–L divergence. We repeat these two steps for 500 times. Finally, we compute the standard deviation of the estimated test statistics per sample size and estimator. Fig. 8 is the standard deviation of these estimators. Because the normalization is irrelevant, all normalized with the variance of the estimator at $m = 20$. Here, it is critical to distinguish between $\mathrm{MMD}^2_u[k, p, q]$ and $\mathrm{MMD}_u[k, p, q]$. The MMD estimator in Eq. (2) is an estimator of $\mathrm{MMD}^2_u[k, p, q]$ and not the original test statistic $\mathrm{MMD}_u[k, p, q]$. Therefore, one needs to be careful about how to interpret this figure. If the convergence rate of the test statistic concerned all have similar convergence rate, with a rate of $\propto m^{-0.5}$, this implies that if $p \neq q$ and a test statistic gives us the largest signal, i.e. lowest p-value, there is no $m' > m$ that another test statistic gives us a larger signal.

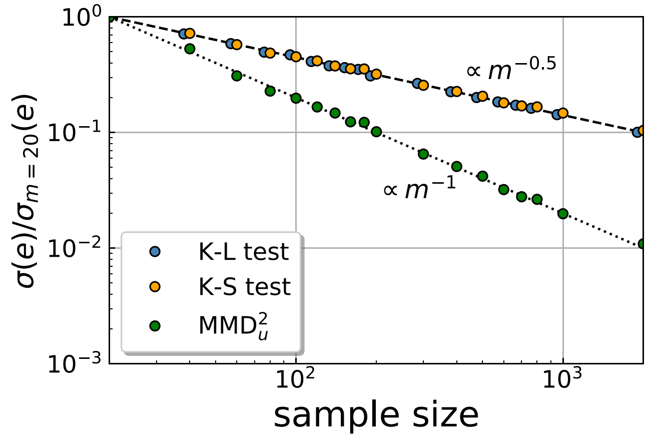

Fig. 8. The convergence rate of the MMD, K–S, and K–L divergence estimators as a function of the sample size. The y-axis is the normalized variance of the estimators per test statistic.

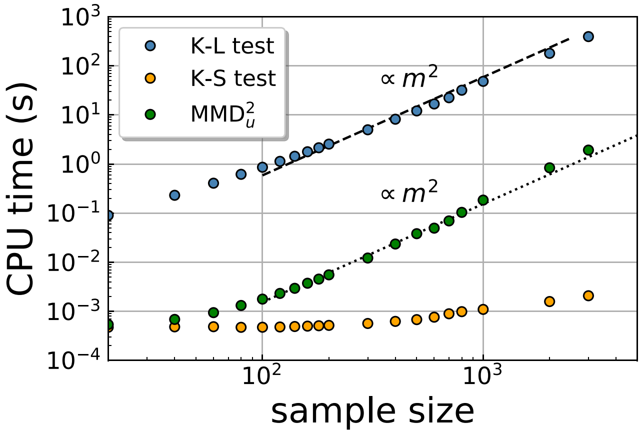

Fig. 9. The execution time of the MMD, K–S, and K–L divergence estimators as a function of sample size.

Given m data points from distribution p and n from distribution q, Eq. (2) implies that the computational cost of the proposed MMD estimator is $\mathcal{O}((m + n)^2)$. This expectation matches the experimental results shown in Fig. 9, where the execution time of the MMD estimator is compared with the K–S and K–L test estimators employed in this work. To compute the execution time, we take two steps. First, two sets of data points of size m are drawn from a Gaussian distribution with mean zero and unity variance. Then, we record the clock time of estimating MMD distance, K–S value, and K–L divergence. The K–L test estimator is computationally the most expensive algorithm of the three. Both MMD and K–L test estimators have a similar rate of increase in computational cost as a function of the sample size, $\propto m^2$. Zaremba et al. (2013) proposed an alternative MMD estimator which is computationally faster by a few orders of magnitude. However, the proposed estimator sacrifices variance for speed. For most practical applications in astronomy, the estimator in Eq. (2) seems to work the best, as it has the smallest variance and is computationally tractable, even with large sample size data. To manage the computational costs for a large data set, the sample can be divided into c equal chunks. The average MMD of all chunks is a good proxy of the MMD distance for this sample. It brings down the execution time to $\mathcal{O}((m+n)^2/c^2)$. To take advantage of machines with multi-processors, the TATTER implementation supports multi-processing jobs and allows the user to run several parallel jobs simultaneously, without any effort on the user's side.

References

Abbott, B.P., Abbott, R., Abbott, T.D., Abernathy, M.R., Acernese, F., Ackley, K., Adams, C., Adams, T., Addesso, P., Adhikari, R.X., et al., 2016. Binary black hole mergers in the first advanced LIGO observing run. Phys. Rev. X 6 (4), 041015.

Abbott, B.P., Abbott, R., Abbott, T.D., Acernese, F., Ackley, K., Adams, C., Adams, T., Addesso, P., Adhikari, R.X., Adya, V.B., et al., 2017. GW170817: Observation of gravitational waves from a binary neutron star inspiral. Phys. Rev. Lett. 119 (16), 161101.

Akeret, J., Refregier, A., Amara, A., Seehars, S., Hasner, C., 2015. Approximate Bayesian computation for forward modeling in cosmology. J. Cosmol. Astro-Part. Phys. 2015, 043.

Allen, S.W., Evrard, A.E., Mantz, A.B., 2011. Cosmological parameters from observations of galaxy clusters. Annu. Rev. Astron. Astrophys. 49, 409–470.

Baringhaus, L., Franz, C., 2004. On a new multivariate two-sample test. J. Multivariate Anal. 88 (1), 190–206.

Ben-David, A., Liu, H., Jackson, A.D., 2015. The Kullback-Leibler divergence as an estimator of the statistical properties of CMB maps. J. Cosmol. Astropart. Phys. 2015 (6), 051.

Bovy, J., 2010. Tracing the hercules stream around the galaxy. Astrophys. J. 725, 1676–1681.

Charnock, T., Battye, R.A., Moss, A., 2017. Planck data versus large scale structure: Methods to quantify discordance. Phys. Rev. D 95 (12), 123535.

Darling, D.A., 1957. The kolmogorov-smirnov, cramer-von mises tests. Ann. Math. Stat. 28 (4), 823–838.

De Simone, A., Jacques, T., 2019. Guiding new physics searches with unsupervised learning. Eur. Phys. J. C 79 (4), 289.

Fasano, G., Franceschini, A., 1987. A multidimensional version of the Kolmogorov-Smirnov test. Mon. Not. R. Astron. Soc. 225, 155–170.

Freeman, P.E., Kim, I., Lee, A.B., 2017. Local two-sample testing: a new tool for analysing high-dimensional astronomical data. Mon. Not. R. Astron. Soc. 471, 3273–3282.

Gosset, E., 1987. A three-dimensional extended Kolmogorov-Smirnov test as a useful tool in astronomy. Astron. Astrophys. 188, 258–264.

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A., 2012a. A kernel two-sample test. J. Mach. Learn. Res. 13 (Mar), 723–773.

Gretton, A., Fukumizu, K., Harchaoui, Z., Sriperumbudur, B.K., 2009. A fast, consistent kernel two-sample test. In: Advances in Neural Information Processing Systems. pp. 673–681.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., Sriperumbudur, B.K., 2012b. Optimal kernel choice for large-scale two-sample tests. In: Advances in Neural Information Processing Systems. pp. 1205–1213.

Harrison, D., Sutton, D., Carvalho, P., Hobson, M., 2015. Validation of Bayesian posterior distributions using a multidimensional Kolmogorov-Smirnov test. Mon. Not. R. Astron. Soc. 451, 2610–2624.

Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., Jennions, M.D., 2015. The extent and consequences of p-hacking in science. PLoS Biol. 13 (3), e1002106.

Herbel, J., Kacprzak, T., Amara, A., Refregier, A., Bruderer, C., Nicola, A., 2017. The redshift distribution of cosmological samples: a forward modeling approach. J. Cosmol. Astro-Part. Phys. 2017, 035.

Ishida, E.E.O., Vitenti, S.D.P., Penna-Lima, M., Cisewski, J., de Souza, R.S., Trindade, A.M.M., Cameron, E., Busti, V.C., COIN Collaboration, 2015. COSMOABC: Likelihood-free inference via population Monte Carlo approximate Bayesian computation. Astron. Comput. 13, 1–11.

Jennings, E., Madigan, M., 2017. astroABC : An Approximate Bayesian Computation Sequential Monte Carlo sampler for cosmological parameter estimation. Astron. Comput. 19, 16–22.

Justel, A., Peña, D., Zamar, R., 1997. A multivariate Kolmogorov-Smirnov test of goodness of fit. Statist. Probab. Lett. 35 (3), 251–259.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22 (1), 79–86.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., Póczos, B., 2017. MMD GAN: TOwards deeper understanding of moment matching network. In: Advances in Neural Information Processing Systems. pp. 2203–2213.

Lopes, R.H.C., Hobson, P.R., Reid, I.D., 2008. Computationally efficient algorithms for the two-dimensional Kolmogorov Smirnov test. J. Phys. Conf Ser. 119, 042019.

MacCoun, R., Perlmutter, S., 2015. Blind analysis: hide results to seek the truth. Nat. News 526 (7572), 187.

Mitrovic, J., Sejdinovic, D., Teh, Y.W., 2016. DR-ABC: approximate Bayesian computation with kernel-based distribution regression. J. Mach. Learn. Res.

Modak, S., Bandyopadhyay, U., 2019. A new nonparametric test for two sample multivariate location problem with application to astronomy. J. Stat. Theory Appl. 18, 136–146.

Mondal, S., Chattopadhyay, A.K., Chattopadhyay, T., 2008. Globular clusters in the milky way and dwarf galaxies: A distribution-free statistical comparison. Astrophys. J. 683 (1), 172–177.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al., 2017. Kernel mean embedding of distributions: A review and beyond. Found. Trends® Mach. Learn. 10 (1–2), 1–141.

Muandet, K., Sriperumbudur, B., Fukumizu, K., Gretton, A., Schölkopf, B., 2016. Kernel mean shrinkage estimators. J. Mach. Learn. Res. 17 (1), 1656–1696.

Neyman, J., Pearson, E.S., 1933. IX. On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 231 (694–706), 289–337.

Nicola, A., Amara, A., Refregier, A., 2019. Consistency tests in cosmology using relative entropy. J. Cosmol. Astropart. Phys. 2019 (1), 011.

Peacock, J.A., 1983. Two-dimensional goodness-of-fit testing in astronomy. Mon. Not. R. Astron. Soc. 202, 615–627.

Ramaswamy, H., Scott, C., Tewari, A., 2016. Mixture proportion estimation via kernel embeddings of distributions, in: International Conference on Machine Learning, pp. 2052–2060.

Ramos Almeida, C., Levenson, N.A., Alonso-Herrero, A., Asensio Ramos, A., Rodríguez Espinosa, J.M., Pérez García, A.M., Packham, C., Mason, R., Radomski, J.T., Díaz-Santos, T., 2011. Testing the unification model for active galactic nuclei in the infrared: Are the obscuring tori of type 1 and 2 seyferts different? Astrophys. J. 731, 92.

Rykoff, E.S., Rozo, E., Busha, M.T., Cunha, C.E., Finoguenov, A., Evrard, A., Hao, J., Koester, B.P., Leauthaud, A., Nord, B., Pierre, M., Reddick, R., Sadibekova, T., Sheldon, E.S., Wechsler, R.H., 2014. redMaPPer. I. Algorithm and SDSS DR8 catalog. Astrophys. J. 785, 104.

Rykoff, E.S., Rozo, E., Hollowood, D., Bermeo-Hernandez, A., Jeltema, T., Mayers, J., Romer, A.K., Rooney, P., Saro, A., Vergara Cervantes, C., Wechsler, R.H., Wilcox, H., Abbott, T.M.C., Abdalla, F.B., Allam, S., Annis, J., Benoit-Lévy, A., Bernstein, G.M., Bertin, E., Brooks, D., Burke, D.L., Capozzi, D., Carnero Rosell, A., Carrasco Kind, M., Castander, F.J., Childress, M., Collins, C.A., Cunha, C.E., D'Andrea, C.B., da Costa, L.N., Davis, T.M., Desai, S., Diehl, H.T., Dietrich, J.P., Doel, P., Evrard, A.E., Finley, D.A., Flaugher, B., Fosalba, P., Frieman, J., Glazebrook, K., Goldstein, D.A., Gruen, D., Gruendl, R.A., Gutierrez, G., Hilton, M., Honscheid, K., Hoyle, B., James, D.J., Kay, S.T., Kuehn, K., Kuropatkin, N., Lahav, O., Lewis, G.F., Lidman, C., Lima, M., Maia, M.A.G., Mann, R.G., Marshall, J.L., Martini, P., Melchior, P., Miller, C.J., Miquel, R., Mohr, J.J., Nichol, R.C., Nord, B., Ogando, R., Plazas, A.A., Reil, K., Sahlén, M., Sanchez, E., Santiago, B., Scarpine, V., Schubnell, M., Sevilla-Noarbe, I., Smith, R.C., Soares-Santos, M., Sobreira, F., Stott, J.P., Suchyta, E., Swanson, M.E.C., Tarle, G., Thomas, D., Tucker, D., Uddin, S., Viana, P.T.P., Vikram, V., Walker, A.R., Zhang, Y., DES Collaboration, 2016. The RedMaPPer galaxy cluster catalog from DES science verification data. Astrophys. J. Suppl. 224, 1.

Sanderson, R.E., Helmi, A., Hogg, D.W., 2015. Action-space clustering of tidal streams to infer the galactic potential. Astrophys. J. 801, 98.

Seehars, S., Amara, A., Refregier, A., Paranjape, A., Akeret, J., 2014. Information gains from cosmic microwave background experiments. Phys. Rev. D 90, 023533.

Song, L., Fukumizu, K., Gretton, A., 2013. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. IEEE Signal Process. Mag. 30 (4), 98–111.

Sriperumbudur, B.K., Fukumizu, K., Gretton, A., Schölkopf, B., Lanckriet, G.R., et al., 2012. On the empirical estimation of integral probability metrics. Electron. J. Stat. 6, 1550–1599.

Wang, Q., Kulkarni, S.R., Verdú, S., 2006. A nearest-neighbor approach to estimating divergence between continuous random vectors. In: Information Theory, 2006 IEEE International Symposium on. IEEE, pp. 242–246.

Wang, Y., Xu, L., Zhao, G.-B., 2017. A measurement of the hubble constant using galaxy redshift surveys. Astrophys. J. 849, 84.

Weiss, L., 1960. Two-sample tests for multivariate distributions. Ann. Math. Stat. 159–164.

Weyant, A., Schafer, C., Wood-Vasey, W.M., 2013. Likelihood-free cosmological inference with type ia supernovae: Approximate Bayesian computation for a complete treatment of uncertainty. Astrophys. J. 764, 116.

Zaremba, W., Gretton, A., Blaschko, M., 2013. B-test: A non-parametric, low variance kernel two-sample test. In: Advances in Neural Information Processing Systems. pp. 755–763.

Zhao, G.-B., Raveri, M., Pogosian, L., Wang, Y., Crittenden, R.G., Handley, W.J., Percival, W.J., Beutler, F., Brinkmann, J., Chuang, C.-H., Cuesta, A.J., Eisenstein, D.J., Kitaura, F.-S., Koyama, K., L'Huillier, B., Nichol, R.C., Pieri, M.M., Rodriguez-Torres, S., Ross, A.J., Rossi, G., Sánchez, A.G., Shafieloo, A., Tinker, J.L., Tojeiro, R., Vazquez, J.A., Zhang, H., 2017. Dynamical dark energy in light of the latest observations. Nature Astron. 1, 627–632.