A time series two-sample test based on comparing distributions of pairwise distances

Pablo Montero-Manso and José A. Vilar

Research Group on Modeling, Optimization and Statistical Inference (MODES),
Department of Mathematics, Computer Science Faculty, University of A Coruña

Abstract. A new test statistic based on comparing empirical distributions of pairwise distances is proposed to check the homogeneity of two groups of time series. By considering distributions instead of specific features of the pairwise distances, the proposed method facilitates the identification of a suitable time series distance in order to increase the discriminatory power of the test. An extensive simulation study shows the performance of the proposed test compared to test statistics based on other distance features such as means or nearest neighbor.

1 Introduction

This paper deals with the problem of testing homogeneity of generating processes for two sets of observed time series. Specifically, let $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ be n and m independent realizations of real-valued processes $\mathbf{X} \equiv \{X_t : t \in \mathbb{Z}\}$ and $\mathbf{Y} \equiv \{Y_t : t \in \mathbb{Z}\}$, respectively. Our goal is to test the null hypothesis of equality of $\mathbf{X}$ and $\mathbf{Y}$, i.e. these time series exhibit the same behavior. Assuming equal length T for all observed series and denoting by F_X and F_Y the T-dimensional distributions of random vectors $(X_1, \ldots, X_T)$ and $(Y_1, \ldots, Y_T)$, respectively, the hypothesis test can be formally stated as

$$\begin{cases} H_0 : F_X = F_Y \\ H_1 : F_X \neq F_Y \end{cases} \tag{1}$$

Therefore, the problem is initially posed in terms of a two-sample test for equal distribution in high dimension. Note that two-sample problems involving partial realizations of time series frequently arise in applications from different fields, ranging from financial settings, such as testing for Monday effect in stock markets [11], to genetics [16] or brain signal analysis [13]. Nevertheless, classical multivariate approaches to handle (1) could face serious drawbacks in a time series framework due to the specific characteristics associated to series. Overall, dimensionality is much greater than the sample size ($p \gg n$). Frequently, the observed time series have different length. In many problems the interest lies in the temporal behavior while differences in terms of location or dispersion tend to be considered as nuisance factors, thus series are normalized prior to analysis. Other relevant issues such as phase difference tend to be ignored too. In short, this kind of particularities limit the effectiveness of commonly used multivariate methods, requiring specific time series approaches, as can be seen in related problems such as time series clustering or classification.

Likewise distances or dissimilarities between time series objects play a key role in many successful methods of time series classification [10], intuitively one would also expect that knowledge on the distances between time series should be useful to address the two-sample problem (1). In fact, distances are also the base of successful nonparametric multivariate two-sample tests, (mainly because they are suited for the $p \gg n$ scenario) [3, 6, 9, 17], and an increasing research effort has been put on generalizing the class of distances that may be used with these methods [9, 15, 18]. We argue that different distances are expected to perform better than others, e.g. the supremum/infimum norm outperforms the Euclidean norm in multivariate scenarios where the difference lies only in the location of one of the variables (see Section 3.2 in [7]).

These evidences and the fact that distances are often used to convey invariance to time series nuisance factors (see [2] for a review of invariance types required in several time series domains) directed us to study the behavior in two-sample problems of different combinations of distances and test statistics. There are two main contributions in this work. First, a new test statistic based on comparing the empirical distributions of pairwise distances between series coming from the same sample and series from the different samples is proposed. The statistic takes advantage of considering the whole distributions instead of only a few specific moments, and exhibits a good performance in different simulation scenarios compared to other alternative test procedures. Second, supported by an extensive numerical study, it is also shown the effect of choosing suitable distances to increase the power of the test. In particular, it is observed that the proposed method produces good results for a large set of distances.

2 Two-sample distance based methods

The methods to be compared are all based on distances. Other methods that do not incorporate distances such as [19] are expected to underperform when extra domain knowledge is introduced through a distance.

The nearest neighbors [8] method is selected among variants and similar methods such as [7] due to its direct association with the k-nearest neighbors method used in classification, which makes easier its interpretation. The so-called “energy” two-sample test [17] is one of the most popular distance-based two-sample methods and its core concept is also used to develop a dependency test, all part of the class of ε-statistics. Including the energy test is especially interesting due to its similarity with the proposed method, both of them based on distributional properties of interpoint distances. A short description of these procedures is given below.

Nearest neighbors. This test, introduced in [8], is based on counting, for each element i, the amount of elements belonging to the same sample as i among its r nearest neighbors.

Let $\mathcal{Z} \equiv \{\mathbf{Z}_i,\, i = 1, \ldots, n+m\}$ be the pooled sample including all $n+m$ series $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$. Define I_i as the indicator function such as $I_i(r) = 1$, if $\mathbf{Z}_i$ and the r-th nearest neighbor to $\mathbf{Z}_i$ belong to the same sample, and $I_i(r) = 0$ otherwise. The statistic of the nearest neighbors test is given by

$$\mathbb{T} = \sum_{i=1}^{n+m} \sum_{j=1}^{r} I_i(j)$$

The statistic $\mathbb{T}$ has an asymptotic normal distribution, but in practice parameters of the distribution are difficult to obtain analytically [17], thus the null distribution of the statistic is approximated via the permutation procedure. Under the alternative hypothesis, values of the statistic are expected to be greater than under the null, so the test rejects the null for large values of the statistic. The restriction to use the Euclidean distance is removed by Henze [9] so other distances can be used. The test depends on the selection of the r parameter, failing to achieve nominal significance levels in some scenarios, as reported in [17] and confirmed in our experiments. It also assumes continuity of distributions. Weighted versions of the statistic have been considered to tune the sensibility of the test against certain alternatives [14].

Energy test. This test is based on Euclidean interpoint distances between different samples and within the same sample, proposed by Székely and Rizzo [17] and by Baringhaus and Franz [1]. The term energy comes from the similitude with Newton's gravitational potential energy. The statistic takes the form

$$E = \frac{nm}{n+m} \left(\frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \|\mathbf{X}_i - \mathbf{Y}_j\| - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|\mathbf{X}_i - \mathbf{X}_j\| - \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \|\mathbf{Y}_i - \mathbf{Y}_j\| \right)$$

with $\|\cdot\|$ denoting the Euclidean norm. The restriction of Euclidean distance is later reduced to the class of continous negative definite functions [18].

The null distribution of the statistic is calculated using the permutation procedure and the test rejects the null for large values of E.

It is worthy to point out the strong similarity between the energy statistic and the *kernel maximum mean discrepancy* (kernel MMD) statistic proposed by Gretton et al. [6], which is based on comparing mean kernel embeddings of distributions. Sedjinovic *et al.* [15] state the equivalence between distance-based and kernel-based statistics for the two-sample problem.

3 A new two-sample test based on pairwise distances

Let $b_{11}, \ldots, b_{nm}$ be the interpoint distances between samples, $b_{ij} = \|\mathbf{X}_i - \mathbf{Y}_j\|$, $1 \leq i \leq n$, $1 \leq j \leq m$, and $w_{11}, \ldots, w_{nn}, w_{n+1,n+1}, \ldots, w_{n+m,n+m}$ be the joint interpoint distances within samples, $w_{ij} = \|\mathbf{X}_i - \mathbf{X}_j\|$, $1 \leq i, j \leq n$ and $w_{ij} = \|\mathbf{Y}_{i-n} - \mathbf{Y}_{j-n}\|$, $n+1 \leq i, j \leq n+m$. The value of the proposed statistic D is the Cramer-von Mises type statistic between the empirical distribution functions of b and w. In the pooled sample of b and w, define $r_1, \ldots, r_{nm}$ as the ranks of b and $s_1, \ldots, s_{n^2+m^2}$ as the ranks of w, then the statistic is defined as:

$$D = \frac{U}{nm(n^2+m^2)(nm+(n^2+m^2))} - \frac{4nm(n^2+m^2)-1}{6(nm+(n^2+m^2))},$$

with

$$U = nm \sum_{i=1}^{nm} (r_i - i)^2 + (n^2+m^2) \sum_{j=1}^{n^2+m^2} (s_i - i)^2.$$

The null distribution of D is obtained via the permutation procedure.

As mentioned, there is a strong similarity between the energy statistic and the proposed statistic, D. To motivate our approach, an intuition of the difference in performance between the energy statistic and the proposed method is given.

In some cases, the test statistics based on the means of within- and between-groups interpoint distances can perform worse than the ones based on the whole distributions of these distances. Figure 1a shows within- and between samples interpoint densities for series generated from AR(0.3) and AR(0.8) processes. Densities instead of distributions are depicted for a clearer view of the effect.

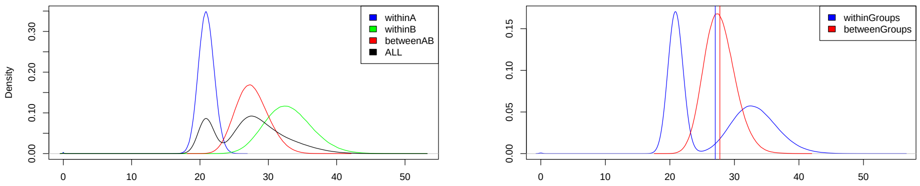
The energy statistic compares the average of the means of the within-groups distances against the mean of the between-groups distances. In some scenarios such as the one presented here, the value of this statistic may not be large enough to reject the null hypothesis, and larger sample sizes would be required. Even though the mean-based energy statistic is small, comparing interpoint densities can produce better discriminatory power. This effect is illustrated in Figure 1b, where the whitin-groups distances are joint. The energy statistic essentially compares the means of these two densities while our method directly compares the distributions. It can be seen that when comparing against betweenGroups distances, joining withinA and withinB distances may result in an very small mean difference while the densities difference can still be significative.

Scenarios where the mean-based approach works better than the distribution method can indeed appear, for instance when sampling distributions only differ in location. In these scenarios, the energy statistic outperforms the D statistic in the same way that a two-sample mean test is superior to a general distribution test when only the means are different.

When we introduce a complex distance, the resulting interpoint distances may separate groups in more ways than the different location scenario, so a test robust against this possibility is desirable, even at the cost of discriminatory power in the previously metioned scenario.

4 Simulation study

In this section, the empirical power of the nearest neighbors test with parameter $r = 3$ (3-NN), the energy test (Energy) and the proposed test based on the distributions of interpoint distances (D-D) are compared under the following scenarios:



(a) (b)

Fig. 1: (a) Densities of within and between samples Euclidean distances for AR(0.3) (A) and AR(0.8) (B) processes. (b) Comparing joint within group and between group interpoint distances densities with the means used by the energy statistic.

- E1: First order autoregressive processes with coefficients -0.9, -0.6, -0.3, 0, 0.3, 0.6, 0.9. In all cases, the error term follows a standard normal distribution, the length of the series is 50, and the distance used is the Euclidean.
- E2: Practitioners may be only interested in the temporal behavior of the series, thus they standardize data to have mean 0 and variance 1. In order to measure the sensibility of the tests in this setting, we have considered the same experiment as E1, but adjusting the models to have the same marginal variance.
- E3: The third experiment is conducted to assess the effect of using different dissimilarities specifically designed for time series. Besides considering particularly interesting comparisons of AR(1) models, some GARCH models are also included to examine more complex dependence structures.
- E4: Real data examples taken from the UCR Time Series archive [5] using different dissimilarities. The selected datasets cover a range of series' lengths from 96 to 720 and do not require either very small or very large sample sizes to produce meaningful results.

A detailed description of the dissimilarities used in the experiments can be seen in [12]. The dissimilarities are selected as representatives of broad groups, such as PDC for the Complexity-based dissimilarities group, AR.LPC.CEPS for the Autoregressive Model-based, CID and DTWARP for their complexity and time warping invariances, etc. Note that this study is not intended as a comparison of dissimilarities but as evidence of the benefits of introducing specialized dissimilarities in the time series context. Sample size was $n = m = 20$ for E1, E2 and E3. E4 has sample sizes of $n = m = 15$ except for the ScreenType and TwoPatterns datasets, both with $n = m = 25$. Experiments were replicated 500 times to approximate the rejection proportion. Real datasets experiments were replicated using permutations from the pooled training and testing subsets. The compared classes were always the first and second one when the datasets had more than 2 classes.

Experiments E1 and E2 are summarized in Table 1, where rejection rates at $\alpha = 0.05$ are shown. Results from E1 show that the proposed method is the best in this scenario, except for the comparison AR(-0.6) vs AR(-0.3), where 3-NN takes a slight advantage, and AR(-0.3) vs AR(0.3), where 3-NN fairly outperforms D-D. The energy test achieves less power than D-D in all considered scenarios. One representative example is the comparison of AR(0.6) vs AR(-0.3) and AR(0.6) vs AR(-0.6). Despite being more separate in their coefficients, the first scenario achieves greater power for both energy and D-D methods. This is explained by the difference of variance of the processes, greater in the first scenario. Results from E2 are free of this effect, and it is observed that the changes in variance drastically reduce the power of the tests, illustrating the difficulty of comparing normalized time series. The results of the nearest neighbor approach are superior on average to the ones obtained with the proposed test, although there exist comparisons where D-D performs better than 3-NN. The energy statistic leads to the lowest powers. Note that the rejection rates under the null are very close to the nominal size for all statistics and experiments.

Experiments E3 and E4 are summarized in Table 2, rejection rates are also for $\alpha = 0.05$. The effect of considering different dissimilarities in the normalized autoregressive scenario (AR(0.2) vs AR(0)) is especially illustrative. When using the Euclidean distance, all methods achieve very low discriminatory power. The power is greatly improved by including distances like SPEC.LLR and AR.LPC.CEPS, which measure dissimilarity in terms of spectra and cepstral coefficients, respectively. The optimal properties of these distances to deal with these models are inherited by the test statistics, resulting in higher powers. It is worthy to remark that, if the Euclidean distance is considered, the 3-NN method is the best one, but Energy and D-D statistics outperform 3-NN when SPEC.LLR and AR.LPC.CEPS are used. This result, which can be extrapolated to the whole of E2, shows that a dissimilarity may differentiate samples in ways which some tests are not sensible to. Note that energy outperforms D-D, for instance the standarization coupled with the coefficients extracted by the AR.LPC.CEPS dissimilarity creates the ideal location-separated scenario metioned in section 3.

The D-D method shows better performance in the non-normalized GARCH scenarios. Series in the experiment marked with ' in Table 2 have length 200. In this scenario, 3-NN fails to achieve the nominal value, requiring fine-tuning of the parameter r. This effect occurs for other GARCH experiments of length 200, not shown here. In the normalized, length 1000 GARCH scenario, marked with * in Table 2, using other distances increases the discriminatory power. D-D is the method taking the most advantage of the new distances. Notably, 3-NN benefits more than Energy with SPEC.LLR, while the opposite happens with AR.LPC.CEPS.

On the real data scenarios, the use of non-Euclidean distances improves the rejection rates with few exceptions. The results also show that the effect of a

Scenario	3-NN	Energy	D-D	*3-NN	*Energy	*D-D
AR(-0.9) vs AR(-0.9)	0.05	0.05	0.06	0.05	0.06	0.05
AR(-0.9) vs AR(-0.6)	0.97	0.96	1.00	0.18	0.09	0.14
AR(-0.9) vs AR(-0.3)	1.00	1.00	1.00	0.29	0.13	0.33
AR(-0.9) vs AR(0)	1.00	1.00	1.00	0.50	0.16	0.61
AR(-0.9) vs AR(0.3)	1.00	1.00	1.00	0.84	0.22	0.81
AR(-0.9) vs AR(0.6)	1.00	1.00	1.00	1.00	0.26	0.97
AR(-0.9) vs AR(0.9)	1.00	0.36	1.00	1.00	0.38	1.00
AR(-0.6) vs AR(-0.6)	0.05	0.06	0.06	0.06	0.05	0.05
AR(-0.6) vs AR(-0.3)	0.26	0.11	0.82	0.11	0.06	0.07
AR(-0.6) vs AR(0)	0.68	0.27	0.99	0.39	0.08	0.17
AR(-0.6) vs AR(0.3)	0.98	0.23	0.99	0.91	0.10	0.47
AR(-0.6) vs AR(0.6)	1.00	0.10	0.90	1.00	0.15	0.90
AR(-0.6) vs AR(0.9)	1.00	1.00	1.00	1.00	0.27	0.96
AR(-0.3) vs AR(-0.3)	0.06	0.03	0.04	0.06	0.06	0.06
AR(-0.3) vs AR(0)	0.11	0.06	0.14	0.10	0.06	0.08
AR(-0.3) vs AR(0.3)	0.43	0.07	0.16	0.43	0.07	0.16
AR(-0.3) vs AR(0.6)	0.98	0.27	0.99	0.92	0.10	0.48
AR(-0.3) vs AR(0.9)	1.00	1.00	1.00	0.86	0.19	0.81
AR(0) vs AR(0)	0.07	0.06	0.06	0.05	0.05	0.05
AR(0) vs AR(0.3)	0.10	0.06	0.12	0.10	0.07	0.08
AR(0) vs AR(0.6)	0.72	0.25	0.99	0.40	0.08	0.16
AR(0) vs AR(0.9)	1.00	1.00	1.00	0.52	0.18	0.64
AR(0.3) vs AR(0.3)	0.05	0.04	0.04	0.05	0.05	0.05
AR(0.3) vs AR(0.6)	0.28	0.14	0.83	0.11	0.06	0.07
AR(0.3) vs AR(0.9)	1.00	1.00	1.00	0.33	0.14	0.35
AR(0.6) vs AR(0.6)	0.06	0.07	0.06	0.04	0.05	0.05
AR(0.6) vs AR(0.9)	0.96	0.96	1.00	0.21	0.09	0.13
AR(0.9) vs AR(0.9)	0.04	0.06	0.06	0.06	0.05	0.05

Table 1: E1 and E2 rejection rates results, * columns have the normalized versions of the series.

dissimilarity is not uniform across the studied methods, e.g., while all methods benefit from the PDC dissimilarity [4] in the RefrigeratorDevices dataset, the 3-NN with PDC decreases its power in the TwoPatterns dataset, when compared with the Euclidean.

The UCR time series datasets are used for classification, classes are presumed to be separable. Nevertheless the two-sample problem may include different scenarios, such as comparing groups consisting of the same two classes,but in different proportions. In the last row of Table 2, mixtures of the first classes of two datasets in different proportions are compared, showing that 3-NN method is not particularly suited for this scenario.

5 Conclusions and further research

We have shown the effect of the chosen dissimilarity when testing homogeneity of generating processes for two sets of time series. Compared to the standard mul-

Scenario	Dissimilarity	3-NN	Energy	D-D
(AR(0.2) vs AR(0))	EUCL	0.06	0.03	0.03
(AR(0.2) vs AR(0))	SPEC.LLR	0.62	0.92	0.83
(AR(0.2) vs AR(0))	AR.LPC.CEPS	0.76	0.95	0.91
(GARCH ω=0.2 α=0.05 β=0.7) vs (GARCH ω=0.1 a=0.1 β=0.8)	EUCL	0.05	0.07	0.28
(GARCH ω=0.2 α=0.1 β=0.7) vs (GARCH ω=0.1 α=0.15 β=0.7)	EUCL	0.1	0.27	0.88
'(GARCH ω=0.2 α=0.1 β=0.7) vs (GARCH ω=0.1 α=0.15 β=0.7)	EUCL	0.00	0.60	1.00
*(GARCH ω=0.1 α=0.7 β=0.2) vs (GARCH ω=0.05 α=0.65 β=0.15)	EUCL	0.04	0.03	0.04
*(GARCH ω=0.1 α=0.7 β=0.2) vs (GARCH ω=0.05 α=0.65 β=0.15)	SPEC.LLR	0.14	0.08	0.39
*(GARCH ω=0.1 α=0.7 β=0.2) vs (GARCH ω=0.05 α=0.65 β=0.15)	AR.LPC.CEPS	0.10	0.30	0.43
*(GARCH ω=0.1 α=0.7 β=0.2) vs (GARCH ω=0.05 α=0.65 β=0.15)	INT.PER	0.09	0.06	0.09
ScreenType	EUCL	0.1	0.1	0.08
ScreenType	INT.PER	0.21	0.09	0.26
RefrigerationDevices	EUCL	0.06	0.04	0.04
RefrigerationDevices	PDC	0.95	0.97	0.99
ShapeletSim	EUCL	0.05	0.04	0.04
ShapeletSim	CID	0.86	0.98	0.97
ToeSegmentation1	EUCL	0.12	0.06	0.05
ToeSegmentation1	DTWARP	0.86	0.73	0.78
TwoPatterns	EUCL	0.45	0.2	0.1
TwoPatterns	PDC	0.15	0.22	0.21
TwoPatterns	SPEC.LLR	0.4	0.58	0.60
ElectricDevices	EUCL	0.1	0.21	0.48
ElectricDevices	PDC	1	1	1
ShapeletSim & Twopatterns Mixture (0.6,0.4) vs (0.8,0.2)	SPEC.LLR	0.33	1	1

Table 2: E3 and E4 rejection rates results.

tivariate approach based on the Euclidean distance, our experiments show that a well-selected time series dissimilarity can substantially improve the discriminatory power. Dissimilarities can separate samples in ways which existing two-sample methods do not take full advantage of. The proposed test compares the whole empirical distributions of pairwise distances within- and between-groups, thus providing a more complete information than simply using a number of specific moments of these distributions. Compared to other alternative statistics, our proposal attained the best performance in many of the simulated scenarios, and produced competitive results in the rest. This good behavior is expected to ease the selection of a proper dissimilarity for the purpose of testing.

The results presented here open the possibility of performing time series two-sample tests in situations where the discriminatory power is low.

Possible lines of research include studying the different ways of comparing interpoint distributions, such using nonparametric densities instead of empirical distributions, and the automatic selection of dissimilarities in the context of hypothesis testing.

References

1. Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of multivariate analysis*, 88(1):190–206, 2004.
2. Gustavo EAPA Batista, Eamonn J Keogh, Oben Moses Tataw, and Vinícius MA de Souza. Cid: an efficient complexity-invariant distance for time series. *Data Mining and Knowledge Discovery*, 28(3):634–669, 2014.

3. Munmun Biswas, Minerva Mukhopadhyay, and Anil K Ghosh. A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika*, pages 913–926, 2014.
4. Andreas M Brandmaier. pdc: An r package for complexity-based clustering of time series. *Journal of Statistical Software*, 67(5):1–23, 2015.
5. Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The ucr time series classification archive, July 2015. `www.cs.ucr.edu/~eamonn/time_series_data/`.
6. Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.
7. Peter Hall and Nader Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002.
8. Norbert Henze. On the number of random points with nearest neighbor of the same type and a multivariate two-sample test. *Metrika*, 31:259–273, 1984.
9. Norbert Henze. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, pages 772–783, 1988.
10. Jason Lines and Anthony Bagnall. Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery*, 29(3):565–592, 2015.
11. Yingying Ma, Wei Lan, and Hansheng Wang. A high dimensional two-sample test under a low dimensional factor structure. *Journal of Multivariate Analysis*, 140:162–170, 2015.
12. Pablo Montero and José A Vilar. Tsclust: An r package for time series clustering. *Journal of Statistical Software*, 62(1):1–43, 2015.
13. Emanuele Olivetti, Danilo Benozzo, Seyed Mostafa Kia, Marta Ellero, and Thomas Hartmann. The kernel two-sample test vs. brain decoding. In *Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on*, pages 128–131. IEEE, 2013.
14. Mark F Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806, 1986.
15. Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, et al. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.
16. Oliver Stegle, Katherine J Denby, Emma J Cooke, David L Wild, Zoubin Ghahramani, and Karsten M Borgwardt. A robust bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *Journal of Computational Biology*, 17(3):355–367, 2010.
17. Gábor J Székely and Maria L Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5:1–6, 2004.
18. Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.
19. Susan Wei, Chihoon Lee, Lindsay Wichers, and JS Marron. Direction-projection-permutation for high-dimensional hypothesis tests. *Journal of Computational and Graphical Statistics*, 25(2):549–569, 2016.