Kernel Hypothesis Testing with Set-valued Data

Alexis Bellot[1,2], **Mihaela van der Schaar**[1,2,3]
[1]University of Cambridge, [2]The Alan Turing Institute, [3]University of California Los Angeles
[abellot,mschaar]@turing.ac.uk

Abstract

We present a general framework for hypothesis testing on distributions of *sets* of individual examples. Sets may represent many common data sources such as groups of observations in time series, collections of words in text or a batch of images of a given phenomenon. This observation pattern, however, differs from the common assumptions required for hypothesis testing: each set differs in size, may have differing levels of noise, and also may incorporate nuisance variability, irrelevant for the analysis of the phenomenon of interest; all features that bias test decisions if not accounted for. In this paper, we propose to interpret sets as independent samples from a collection of latent probability distributions, and introduce kernel two-sample and independence tests in this latent space of distributions. We prove the consistency of these tests and observe them to outperform in a wide range of synthetic experiments. Finally, we showcase their use in practice with experiments on healthcare and climate data, where previously heuristics were needed for feature extraction and testing.

1 Introduction

Hypothesis tests are used to answer questions about a specific dependency structure in data (e.g. independence between variables, equality of distributions between samples etc). They are used in applications across the sciences where they serve as an essential tool to summarize and quantify the evidence for structure in the distribution of data [20]. In consequence, a growing body of work is constantly revisiting established modelling assumptions to allow for consistent testing in increasingly heterogeneous data sources. Examples include non-parametric tests formulated as distances in Hilbert space [10, 9, 7, 45], tests based on neural network representations [21, 25, 1] and others that have significantly advanced the reach of hypothesis tests towards high-dimensional data of unknown distribution. Almost universally however, non-parametric tests require a *fixed* presentation of data (e.g. each instance living in $\mathbb{R}^d$) and do not account for *non-homogeneous noise* patterns across examples (e.g. such as found in medical data, each patient or instance having different level of variation). Many problems do exhibit these properties, including time series (e.g. multiple observations over time for each individual) and bagged data (e.g. multiple images of the same phenomenon) in domains such as medicine and climate science.

Intriguingly, there exists an appropriate representation of data that naturally encodes a more flexible observation pattern, namely each example represented as a *set* of observations (i.e. an unordered collection of multivariate observations), each set of potentially irregular length and sampled from potentially different distributions. In particular, sets do not presuppose a fixed representation of data (sets may be of different length) and each set may be associated with a unique distribution that encodes its particular variation pattern (potentially different from other sets). Testing on sets implicitly shifts the question of interest from a hypothesis on groups of actual observations to an hypothesis on groups of latent distributions assumed to represent each observed example or set. See Figure 1 for an illustration of this interpretation for the two sample problem. This set-up is common

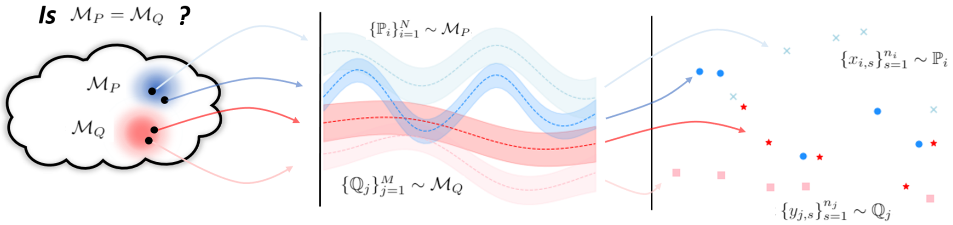
Is $\mathcal{M}_P = \mathcal{M}_Q$?
$\mathcal{M}_P$
$\mathcal{M}_Q$
$\{\mathbb{P}_i\}_{i=1}^N \sim \mathcal{M}_P$
$\{\mathbb{Q}_j\}_{j=1}^M \sim \mathcal{M}_Q$
$\{x_{i,s}\}_{s=1}^{n_i} \sim \mathbb{P}_i$
$\{y_{j,s}\}_{s=1}^{n_j} \sim \mathbb{Q}_j$

Figure 1: We consider an example from electronic health records to illustrate the proposed approach. Reading from right to left, we observe irregular, uncertain biomarker measurements over time in two groups of patients (treated and control) colored with different shades of red and blue, the question being whether these populations have the same trajectory in distribution. In the middle panel we encode the uncertainty in each patient trajectory by a probability distributions (middle panel) on the space of observations. The two-sample problem is to test for equality in distribution on the space of patient-specific distributions, rather than observations. This two-level hierarchy allows for noisy inputs and irregular input sizes. A description of the notation and more details can be found in Section 3.1.

in regression problems where one seeks to learn a mapping from distributions to associated labels [37, 38], but is unexplored in hypothesis testing.

The goal of this paper is to introduce kernel two-sample and kernel independence tests defined on *set*-valued examples. We will show that tests defined in this space appropriately encode individual-level heterogeneity, are much more flexible, do not require heuristic pre-processing of data, and are found to be more powerful than alternatives. We refer to the proposed approach as a framework, applicable to any kernel-based test that includes, in addition to two-sample and independence tests described here, conditional independence tests and three-variable interaction tests. The technical challenge to achieve consistency of test decisions is that latent distributions on which tests are defined are not available (and instead are approximated with each available set of observation). This introduces an additional layer of uncertainty that must be bounded to derive well-defined asymptotic distributions for the proposed test statistics. For this reason, we put emphasis also on the quality of finite-dimensional approximations of the proposed tests, with approaches to minimize test statistic variance and to tune hyperparameters for maximum power.

The contribution is three-fold: this paper for the first time formally describes tests on set-valued data, it demonstrates the consistency of these tests for two common problems, two-sample and independence testing, and it validates the proposed tests and optimization routines on simulated experiments that show that one may consistently discriminate between hypotheses on data that was previously not amenable to hypothesis testing (while guaranteeing consistency).

2 Background

The tests presented in this paper are defined on distributions. Testing on distributions is the problem of defining a test statistic that maps distributions to a scalar that quantifies the evidence for a hypothesis we might set on the relationships in data. However, we do not have access to probability distributions themselves, but rather distributions are observed only through samples,

$$\{x_{1,j}\}_{j=1}^{n_1}, ..., \{x_{N,j}\}_{j=1}^{n_N}. \tag{1}$$

Each $\{x_{i,j}\}_{j=1}^{n_i}$ is a *set* of n_i individual observations $x_{i,j}$ (typically in $\mathbb{R}^d$). We assume that $\{x_{i,j}\}_{j=1}^{n_i}$ are $i.i.d$ samples from an unobserved probability distribution $\mathbb{P}_i$. The probability distributions $\{\mathbb{P}_i\}_{i=1}^N$ themselves have inherent variability, such as can be expected for example from different medical patients. We assume each one of them to be drawn randomly from some unknown meta-distribution $\mathcal{M}_P$ defined over a set of probability measures $\mathcal{P}$. We illustrate this set-up in Figure 1 for the two-sample problem (more details in Section 3.1).

2.1 Embeddings of Distributions

Let $\mathcal{X}$ be a measurable space of observations. We use a positive definite bounded and measurable kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to represent distributions $\mathbb{P}_i$ on $\mathcal{X}$, and independent samples $\{x_{i,j}\}_{j=1}^{n_i}$, as two functions $\mu_{\mathbb{P}_i}$, and $\hat{\mu}_{\mathbb{P}_i}$ respectively, called kernel mean embeddings [28]. Both are defined in the corresponding Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ by,

$$\mu_{\mathbb{P}_i} := \int_{\mathcal{X}} k(x,\cdot)d\mathbb{P}(x), \qquad \hat{\mu}_{\mathbb{P}_i} := \frac{1}{n_i}\sum_{x\in\{x_{i,j}\}_{j=1}^{n_i}} k(x,\cdot)$$

To make inference on populations of distributions, the desiratum however is on defining useful representations of distributions $\mathcal{M}_P$ on the space probability measures, rather than on the space of observations. [4] showed that one may do so analogously to the definition of kernels on $\mathcal{X}$ by treating mean embeddings $\mu_{\mathbb{P}}$ themselves as inputs to kernel functions (replacing $x \in \mathcal{X}$ in the conventional learning setting as inputs to k). See eq. (2) below.

In practice, each set representation $\mu_{\mathbb{P}_i}$ is limited to be approximated by irregularly sampled observations $\{x_{i,j}\}_{j=1}^{n_i}$. Not all mean embeddings $\mu_{\mathbb{P}}$ are expected to provide the same amount information about their underlying distribution $\mathbb{P}$. Indeed, the empirical mean embeddings $\hat{\mu}_{\mathbb{P}_i}$ converge to their population counterpart at a rate $\mathcal{O}(1/\sqrt{n_i})$ (see e.g. Lemma 1 in the Appendix and also [34]) in their set size n_i. Rather than assuming access to a uniform sample of distributions $\{\mathbb{P}_i\}_{i=1}^{N}$ from $\mathcal{M}_P$, like we did with the raw observations $\{x_{i,j}\}_{j=1}^{n_i}$, we may account for this irregularity and uncertainty in approximation by interpreting the set of distributions as a weighted sample $\{(\mathbb{P}_i, w_i)\}_{i=1}^{N} \sim \mathcal{M}_P$. Each weight quantifying the accuracy of the approximation of each distribution with the limited samples available. The corresponding population and empirical mean embedding in this space may be written as,

$$\mu_{\mathcal{M}} := \int_{\mathcal{P}} K(\mu_{\mathbb{P}},\cdot)d\mathcal{M}(\mathbb{P}), \qquad \hat{\mu}_{\mathcal{M}} := \sum_{i=1}^{N} w_i K(\mu_{\mathbb{P}_i},\cdot) \tag{2}$$

We will make use of the Gaussian kernel between distributions defined $K(\mathbb{P},\mathbb{Q}) := \exp(-||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||^2_{\mathcal{H}_K}/2\sigma^2)$ [4, 27]. Note that for kernels on $\mathcal{X}$, their RKHS consists of functions $\mathcal{X} \to \mathbb{R}$, while the kernel K lives on the space of distributions on $\mathcal{X}$, $\mathcal{P}(\mathcal{X})$, and its RKHS consists of functions $\mathcal{P}(\mathcal{X}) \to \mathbb{R}$. We may use K to learn from samples that are individual distributions, rather than individual observations [4].

With this construction (i.e. kernels evaluated on mean embeddings) [37] investigated generalization performance in distributional regression: regressing to a real-valued response from a probability distribution. Results that were subsequently extended to study distributional regression for causal inference [23] and for transfer learning [3]. A technical contribution of this paper is to extend these results to demonstrate consistent hypothesis testing on distributions.

2.2 Hypothesis Testing with Kernels

The advantage for hypothesis testing of mapping distributions $\mathcal{M}$ and $\mathcal{M}'$ to functions in an RKHS is that we may now say that $\mathcal{M}$ and $\mathcal{M}'$ are close if the RKHS distance $||\mu_{\mathcal{M}} - \mu_{\mathcal{M}'}||_{\mathcal{H}_K}$ is small [7]. This distance depends on the choice of the kernel K and k; a crucial property of the embeddings is that for certain kernels the feature map is injective. These kernels are called characteristic [35]. Probability distributions may be distinguished exactly by their images in the RKHS, and also $||\mu_{\mathcal{M}} - \mu_{\mathcal{M}'}||_{\mathcal{H}_K}$ is zero if and only if the distributions coincide [7]. From the statistical testing point of view, this coincidence axiom is key as it ensures consistency of comparisons for any pair of different distributions. As a key property of the set-up we have introduced, [**?**,]in Theorem 2.2]christmann2010universal demonstrated that for well known kernels, such as the Gaussian kernel, if used in both levels of the embedding the resulting embedding is injective (i.e. kernels are characteristic).

The empirical version of the RKHS distance, however, will not necessarily be exactly zero even if the distributions do coincide. Some variability is to be expected due to the limited number of samples, and in contrast to conventional kernel tests, in the case considered here also due to the variability in the estimation of set embeddings. Instead of testing on an $i.i.d.$ sample $\{\mu_{\mathbb{P}_i}\}_{i=1}^{N}$, we are testing over the set $\{\hat{\mu}_{\mathbb{P}_i}\}_{i=1}^{N}$. There is an additional level of uncertainty which must be accounted for.

In practice, tests are constructed such that a certain hypothesis is rejected whenever a test statistic exceeds a certain threshold away from 0 [20]. Then, short from achieving perfect discrimination between two hypotheses, the goal of hypothesis testing is to derive a threshold such that false positives are upper bounded by a design parameter α and false negatives are as low as possible.

2.3 Related work

As a first observation, note that kernels defined on sets directly [17], measuring the similarity between sets by the average pairwise point similarities between the sets, are not known to be characteristic. Attempts have also been made to define kernels on the space of distributions, including probability product kernel [13], the Fisher kernel [12], diffusion kernels [18] and kernels arising from Kullback-Leibler divergences [26], none of them known to be characteristic and in this case with the shortcoming that many of the above are parametrized by a family of densities which may or may not hold in data.

Deep learning has emerged as an alternative for defining tests on structured objects. [25] define classifier two-sample tests and [21] use deep kernels to embed structured objects. Tests in these cases, however, are defined directly on the space of observations, it is not clear how to input examples of varying sizes, or how to account for the uncertainty in individual observations especially if these change across sets.

Accommodating for input uncertainty has connections with robust hypothesis testing. These tests attempt to explicitly enforce invariances in test statistics in a certain uncertainty ball to remove irrelevant sources of variation [6, 11]. Other types of invariances can also be enforced, for instance [19] use features designed to be invariant to additive noise and use distances between those representations for hypothesis testing. One may also use a model-based approach to capture this uncertainty, for instance [2] use Gaussian processes and compare posterior distributions. More generally, also work in the functional data analysis literature [44, 29] uses a model-based approach to testing sets that represent functions.

3 Hypothesis Tests for Uncertain Sets

In the following sections, we propose tests to evaluate two common hypotheses: the two sample problem of testing equality of distributions in two samples, and the independence problem of testing whether joint distributions in paired samples coincide with the product of their marginals. For both tests, the exposition mirrors well-known results in kernel hypothesis testing which we will only briefly describe (see [7, 10] for more background). The contribution of this paper is to show that tests defined with a second level of sampling are consistent and to show that correctly weighting representations according to their set size is most efficient.

We may summarize hypothesis testing in this context as follows:

1. Embed the distributions $\{\mathbb{P}_i\}_{i=1}^N$ into an RKHS using approximations of the mean embeddings $\{\hat{\mu}_{\mathbb{P}_i}\}_{i=1}^N$ computed with independent samples $\{x_{i,j}\}_{j=1}^{n_i} \sim \mathbb{P}_i$.
2. Define test statistics on this feature representations to test for a certain hypothesis or dependency structure in $\mathcal{M}$.

3.1 The two sample problem

Consider a first collection of sets of observations, each i-th set denoted $\{x_{i,s}\}_{s=1}^{n_i} \sim \mathbb{P}_i$, for a total of N such sets with distributions $\{\mathbb{P}_i\}_{i=1}^N \sim \mathcal{M}_P$, and define similarly a second collection of sets, each j-th set $\{y_{j,s}\}_{s=1}^{n_j} \sim \mathbb{Q}_j$, for $\{\mathbb{Q}_j\}_{j=1}^M \sim \mathcal{M}_Q$. The problem we consider is to test whether,

$$\mathcal{H}_0 : \mathcal{M}_P = \mathcal{M}_Q \quad \text{or else} \quad \mathcal{H}_1 : \mathcal{M}_P \neq \mathcal{M}_Q \tag{3}$$

holds on the basis of the observations available in each set. We illustrate this problem in Figure 1. The proposed test statistic approximates the square of the RKHS distance between densities $\mathcal{M}_P$ and $\mathcal{M}_Q$, also called Maximum Mean Discrepancy (MMD), which may be decomposed as follows [7],

$$\text{MMD}^2 := \mathbb{E}_{\mathbb{P},\mathbb{P}' \sim \mathcal{M}_P} K(\mathbb{P},\mathbb{P}') + \mathbb{E}_{\mathbb{Q},\mathbb{Q}' \sim \mathcal{M}_Q} K(\mathbb{Q},\mathbb{Q}') - 2\mathbb{E}_{\mathbb{P}\sim\mathcal{M}_P,\mathbb{Q}\sim\mathcal{M}_Q} K(\mathbb{P},\mathbb{Q}) \tag{4}$$

where K is the kernel on distributions given after equation (2). We denote $\widehat{\text{MMD}}^2$ the empirical estimator of the MMD^2 with expectations replaced by averages, obtained from independent samples $\{\mathbb{P}_i\}_{i=1}^N \sim \mathcal{M}_P$ and $\{\mathbb{Q}_j\}_{j=1}^M \sim \mathcal{M}_Q$. The proposed statistic is defined by considering approximate mean embeddings of each distribution and considering the weighted sample of their meta-distribution each of them represents,

$$\widehat{\text{RMMD}}^2 := \sum_{i,j=1}^{N} w_{\mathbb{P}_i} w_{\mathbb{P}_j} K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j}) + \sum_{i,j=1}^{M} w_{\mathbb{Q}_i} w_{\mathbb{Q}_j} K(\hat{\mu}_{\mathbb{Q}_i}, \hat{\mu}_{\mathbb{Q}_j}) - 2 \sum_{i,j=1}^{N,M} w_{\mathbb{P}_i} w_{\mathbb{Q}_j} K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{Q}_j})$$

R stands for robust. Weights in all cases are normalized, $\sum_i w_{\mathbb{P}_i} = \sum_j w_{\mathbb{Q}_j} = 1$, and will be assumed fixed. We return to the specification of weights in section 3.3. The asymptotic behaviour of $\widehat{\text{MMD}}^2$ is well understood [7] and the test itself extensively used in many applications [22, 31]. However, these results do not extend trivially if each independent set exhibits an additional source of variation due to the estimation of the mean embedding. In the following proposition, we bound the contribution of this additional source of variation and show that under the asymptotic regime where both the set sizes and number of sets grow larger, asymptotic distributions are well defined.

Proposition 1 ***(Asymptotic distribution)****. Let two samples of data be defined as above and let K be characteristic and L_K-Lipschitz continuous. Further, let $N = M$ for clarity of exposition. In the asymptotic regime where both the number of sets N and set sizes n_i tends to infinity,*

- *Under the null $\mathcal{M}_P = \mathcal{M}_Q$, the distributions of $\sqrt{N}\widehat{MMD}^2$ and $\sqrt{N}\widehat{RMMD}^2$ coincide.*
- *Under the alternative $\mathcal{M}_P \neq \mathcal{M}_Q$, the distributions of $N\widehat{MMD}^2$ and $N\widehat{RMMD}^2$ coincide.*

Proof. All proofs are given in the Appendix.

3.2 The independence problem

Independence tests are concerned with the question of whether two random variables are distributed independently of each other. For this problem, we start with a collection of *paired* distributions $\{(\mathbb{P}_i, \mathbb{Q}_i)\}_{i=1}^N$ drawn from a joint distribution we denote $\mathcal{M}_{PQ}$, and denote their marginals $\mathcal{M}_P$ and $\mathcal{M}_Q$. The hypothesis problem is to determine whether,

$$\mathcal{H}_0 : \mathcal{M}_{PQ} = \mathcal{M}_P \mathcal{M}_Q \quad \text{or else} \mathcal{H}_1 : \mathcal{M}_{PQ} \neq \mathcal{M}_P \mathcal{M}_Q \tag{5}$$

Example. *To illustrate this problem for set-valued data consider gene expression measurements associated with a corresponding human trait. A common problem is to identify dependencies between them for feature selection or to motivate causal discovery. Gene expression measurements however are known to be noisy and it is common practice to replicate experiments [5, 42, 27]. For each gene we typically observe multiple combinations of gene expressions and trait observations, one for each experiment, with slight variation due to different experimental conditions. Here distributions faithfully describe the uncertainty in underlying gene expression, and independence may be tested in this space directly.*

As in the two-sample test, we may quantify the difference between distributions using the RKHS distance $||\mu_{\mathcal{M}_{PQ}} - \mu_{\mathcal{M}_P} \otimes \mu_{\mathcal{M}_Q}||_{HS}^2$. Kernels K, L are assumed characteristic; $|| \cdot ||_{HS}$ is the norm on the space of $\mathcal{H}_K \rightarrow \mathcal{H}_L$ Hilbert-Schmidt operators, and $\otimes$ denotes the tensor product, such that $(a \otimes b)c = a\langle b, c\rangle$. This distance is called the Hilbert Schmidt Independence Criterion (HSIC) [8, 10].

Two empirical estimators can be written: one assuming access to independent samples $\mathcal{M}_{PQ}$ and one with independent samples from each of the paired distributions sampled from $\mathcal{M}_{PQ}$,

$$\widehat{\text{HSIC}} = \text{Tr}\,(KHLH)/N^2, \qquad \widehat{\text{RHSIC}} = \text{Tr}\,(\hat{K}H\hat{L}H) \tag{6}$$

for kernel matrices with (i, j) entries $K_{ij} = K(\mathbb{P}_i, \mathbb{P}_j) = \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle_{\mathcal{H}_K}$ and $L_{ij} = \langle \mu_{\mathbb{Q}_i}, \mu_{\mathbb{Q}_j} \rangle_{\mathcal{H}_L}$ for the population version and $\hat{K}_{ij} = w_{\mathbb{P}_i} w_{\mathbb{P}_j} \langle \hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j} \rangle_{\mathcal{H}_K}$ and $\hat{L}_{ij} = w_{\mathbb{Q}_i} w_{\mathbb{Q}_j} \langle \hat{\mu}_{\mathbb{Q}_i}, \hat{\mu}_{\mathbb{Q}_j} \rangle_{\mathcal{H}_L}$ with mean embeddings replaced by their weighted finite sample counterparts for the robust alternative. The centering matrix is defined by $H = I - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ and Tr is the trace operator. Independence testing with the $\widehat{\text{HSIC}}$ is well understood [10, 45, 14]. Approximations due to a second level of sampling are well behaved and mirror those of the robust statistic for the two-sample problem. In particular, as in the two-sample problem we show in the Appendix that asymptotic distributions coincide in the regime with increasing set size and increasing sample size, making hypothesis testing with the $\widehat{\text{RHSIC}}$ consistent for the independence problem in equation (5).

3.3 Practical considerations

Weights for high power. Set sizes in practice may be limited. In the asymptotic regime of increasing number of sets but finite set size, the properties of the estimator may depend on appropriately weighting sets for high power. The proposed weighting scheme addresses this point. Recall that each individual observation x_{ij} is drawn independently from their respective distributions $\mathbb{P}_i$. Other factors of variations assumed to be common across sets, the variance of the approximate embedding $\hat{\mu}_{\mathbb{P}_i}$ is therefore proportional to $1/n_i$ (i.e. the variation in approximation of mean embeddings is due solely to diverging set sizes). When mean embeddings have different variances, it is efficient to give less weight to mean embeddings that have high variances. By efficient in this context, we mean highest asymptotic power of tests based on mean embedding representations of sets. For V-statistics the asymptotic power function is well known, and an argument involving the delta method for differentiable kernels, expanded on in the Appendix, can be used to determine the optimal weights to be given by $w_{\mathbb{P}_i} := n_i / \sum_i n_i$ for each i.

Hyperparameters for high power. With a similar intuition, even though in theory we can expect high power for any alternative hypothesis and any choice of kernel, with finite sample size, some kernel hyperparameters will give higher power than others. The proposed tests optimize the choice of kernels by choosing hyperparameters that minimize the asymptotic variance under the alternative similarly to [36, 14]. But, in addition, we extend the optimization to tune both the mean embedding to represent sets and the kernel used for comparisons in Hilbert space. Please find more details in the Appendix.

Low-dimensional approximations for large scale data. Testing on distributions as described is not scalable for even modestly-sized datasets, as computing each of the entries of the relevant kernel matrices requires defining a high-dimensional mean embedding. To define test statistics on these representations we further embed the non-linear feature space $\mathcal{H}_k$ defined by k into a random low dimensional Euclidean space using their expansion in Hilbert space as a linear combination of the Fourier basis [32, 30]. If we draw m samples from the Gaussian spectral measure, we can approximate the Gaussian kernel k by,

$$k(x,y) \approx \frac{2}{m}\sum_{j=1}^{m} \cos(\langle \omega_j, x\rangle + b_j)\cos(\langle \omega_j, y\rangle + b_j) = \langle \phi(x), \phi(y)\rangle$$

where $\omega_1, ..., \omega_m \sim \mathcal{N}(0,\gamma)$, $b_1, ..., b_m \sim \mathcal{U}[0, 2\pi]$, and $\phi(x) = \sqrt{\frac{2}{m}}[\cos(\omega_1 x + b_1), ..., \cos(\omega_m x + b_m)] \in \mathbb{R}^m$ [30]. The mean embedding $\mu_{\mathbb{P}} = \mathbb{E}_{X\sim\mathbb{P}}\phi(X)$ can then be approximated with elements in the span of $(\cos(\langle \omega_j, x\rangle + b_j))_{j=1}^m$. By averaging over the available n_i samples in X_i from the distribution $\mathbb{P}_i$, the approximate finite-dimensional embedding is given by,

$$\hat{\mu}_{\mathbb{P}_i,m} = \frac{1}{n_i}\sum_{x\in\{x_{ij}\}_{j=1}^{n_i}} (\cos(\langle w_j, x\rangle + b_j))_{j=1}^m \in \mathbb{R}^m$$

4 Synthetic Data Experiments

The purpose of synthetic experiments will be to test power: the rate at which we correctly reject $\mathcal{H}_0$ when it is false, as we increase the difficulty of the testing problems; and Type I error: the rate at which we incorrectly reject $\mathcal{H}_0$ when it is true. In all experiments, α (the target Type I error) is set to 0.05, the number of time series is set to $N = 500$, the number of observations made on each time series is random between 5 and 50, and each problem is repeated for 500 trials.

Tests for empirical comparisons. To the best of our knowledge, no existing test naturally accommodates for set-valued data with irregular sizes. Our approach to empirical comparisons will be to coerce the data into a fixed dimensional vector in a well-defined manner, and evaluate existing tests on this representation. To do so, we focus on time-series -like data which we interpolate along the time axis with cubic splines and evaluate at a fixed number of time points. The following tests are evaluated for the two-sample problem. The **MMD** [7] with hyperparameters optimized for maximum power, two-sample classifier tests [25] which involve fitting a deep classifier. We considered a recurrent neural network with GRU cells for sequential data (**C2ST-GRU**) and the DeepSets approach of [43] modelling permutation invariance to be expected in sets (**C2ST-Sets**). We consider also the Gaussian

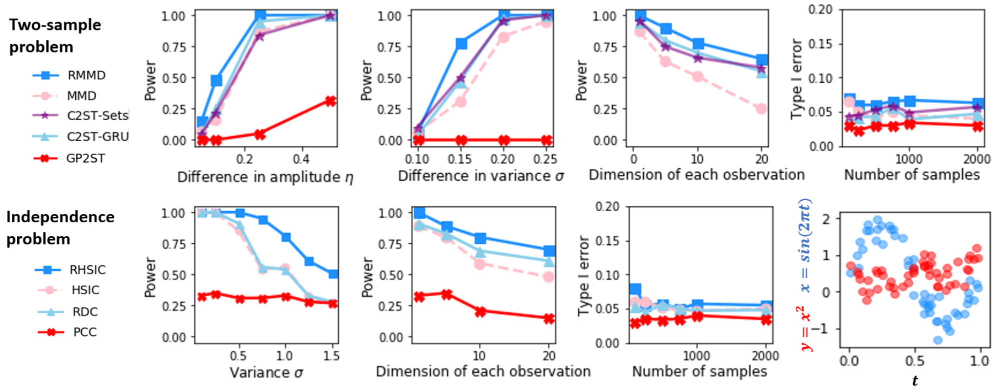

Figure 2: Power (higher better) and Type I error on synthetic data. The top panels, from left to right, evaluate power as we increase the difference in time series amplitude (with equal variance $\sigma = 0.1$) and observation variance (with equal amplitude $\eta = 1$) between the two populations. Power comparisons as the dimension of each time series increases (on data sampled with a difference in amplitude equal to 0.25) is shown next. Each new dimension is sampled as in the one-dimensional problem but with equal amplitude across the two populations, in other words only the distribution of the first dimension in each multivariate time series varies. The rightmost panel gives type I error with approximate control at the level $\alpha = 0.05$ for all methods. The bottom row considers the independence problem, we evaluate power as we increase the variance of paired time series, and consider increasing dimensionality for a fixed variance $\sigma = 0.5$. Finally, the bottom right plot shows a sample of two dependent noisy time series, colored blue and red respectively, for illustration.

process-based test (**GP2ST**) by [2]. For the independence problem we consider: the **HSIC** [10], the Randomized Dependence Coefficient (**RDC**) [23] and Pearson Correlation Coefficient (**PCC**). For all kernel-based tests, because their null distributions are given by an infinite sum of weighted χ^2 variables (no closed-form quantiles), in each trial we use 400 random permutations to approximate the null distribution. We give more details on the implementation of each of these tests in the Appendix.

4.1 Two-sample problem

Each one of the two samples is defined by a family of N distributions $\{\mathbb{P}_i\}_{i=1}^N$ we take to be Gaussian $\mathbb{P}_i = \eta \sin(2\pi t) + \mathcal{N}(0, \sigma_i + \sigma)$. The variability between the $\{\mathbb{P}_i\}_{i=1}^N$ is specified by σ_i, drawn from a one-parameter inverse gamma distribution, which mimics the behaviour of the meta-distribution and the observation pattern we may observe in heterogeneous data. The difference between two populations of sampled distributions is the mean amplitude η and/or shifts in *baseline* variance σ. Two-sample problems become harder whenever these parameters converge to the same value in the two samples and are easier when they diverge. Recall that the sampled Gaussian distributions themselves are not observable and, in turn, we have access to observations $x_{ij} \sim \mathbb{P}_i$. Each x_{ij} is obtained by fixing t to $t_j \sim \mathcal{U}[0,1]$ and subsequently sampling from the Gaussian. The result is two collections of noisy time series with non-linear dynamics. Each time series, or set of observations, is irregularly sampled with noise levels that vary between sets.

Results. We report power and type I error for the two sample problems in the top row of Figure 2. All tests approximately control for type I error at the desired threshold. In terms of power, we observe the RMMD to outperform across all experiments with an important contrast on the difference in performance with the MMD. Even though using similar test statistics, the RMMD much more faithfully captures the irregularity and uncertainty of every individual set of observations. RMMD similarly outperforms C2ST-based tests, the strongest baselines, with up to a two-fold increase in power in some cases.

4.2 Independence problem

Define the mean of each distribution $\mathbb{P}_i$ as $f_i(t) := \beta_i \sin(2\pi t) + \alpha_i t$. Differently than in the two-sample problem, the variability among the $\{\mathbb{P}_i\}$ appears in the amplitude and trend of the sine function, let these be $\beta_i \sim \mathcal{U}[0.5, 1.5]$ and $\alpha_i \sim \mathcal{U}[-0.5, 0.5]$. Once these parameters are sampled, paired distributions $(\mathbb{P}_i, \mathbb{Q}_i)$ are given by $\mathbb{P}_i = f_i(t) + \mathcal{N}(0,\sigma)$ and $\mathbb{Q}_i = g(f_i(t)) + \mathcal{N}(0,\sigma)$. Each

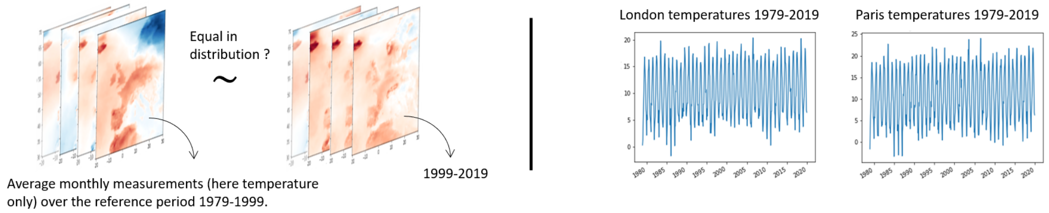

Figure 3: Illustration of the two-sample problem with *global* set-valued data versus *local* time series data.

observation from this pair is obtained as in the two sample problem by fixing a random t and sampling from the resulting distribution. The difficulty of the problem is governed by two factors: g and σ. g determines the dependency between the two functions. In every trial, $g(x)$ is randomly chosen from the set of functions $\{x^2, x^3, \cos(x), \exp(-x)\}$. Testing for dependency is hard also for increasing variance σ of observations, as this makes the dependent paired samples appear independent.

Results. Power and type I error are shown in the bottom row of Figure 2. The conclusions for this problem mirror the two-sample testing experiments, with however a much larger increase in power over alternatives, all using less flexible data representations as none of them avoids interpolating between observations before testing independence). This is made clear with data of increasing variance that significantly hampers interpolation performance.

5 Testing on Lung function Data of Cystic Fibrosis Patients

For people with Cystic Fibrosis (CF), mucus in the lungs is linked with chronic infections that can cause permanent damage, making it harder to breathe [15]. This condition is often measured over time using `FEV1% predicted`; the Forced Expiratory Volume of air in the first second of a forced exhaled breath we would expect for a person without CF of the same age, gender, height, and ethnicity [39]. For example, a person with CF who has `FEV1% predicted` equal to 50% can breathe out half the amount of air as we would expect from a comparable person without CF. In this experiment, we work with data from the UK Cystic Fibrosis Trust containing records from $10,980$ patients with approximately annual follow ups between 2008 and 2015, with the objective of better understanding the dependence of lung function over time with other biomarkers. For this problem we found a significant influence of Body Mass Index (BMI) over time and the number of days under intravenous antibiotics in a given year; both already known to be associated with lung function [41, 16].

We use this information to create a set of problems under the alternative $\mathcal{H}_1$ with an additional twist. We increase heterogeneity among patients by artificially removing a proportion p of densely sampled patients (here more than 4 recordings). The problem is to test for independence between a patients two-dimensional trajectory of BMI and antibiotics measurements over time, and their lung function trajectory over time. In this set-up, we expect the information content of the average patient to decrease, a scenario that lends itself to an importance-weighted approach (more weight on densely sampled trajectories), such as described in section 3.3. In this section we test this property, which we found advantageous for higher missingness data patterns, as shown in Figure 4. In this case, power tends to be higher after weighting (RHSIC) versus not weighting (RHSIC-weight). We report also type I errors, well controlled by all methods, evaluated after shuffling the lung function trajectories between patients, such as to break the associations between BMI and antibiotics, and lung function trajectories.

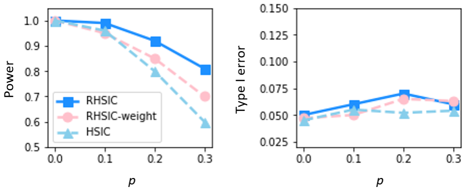

Figure 4: Power and Type I error on Cystic Fibrosis data.

6 Testing on Climate Data

This experiment explores the use of extensive weather data to determine whether the recent rapid changes in climate associated with human-induced activities significantly differ from natural climate variability. We usually think of temperature as characterizing climate, but in fact a number of variables are used to monitor the state of the climate including precipitation, wind patterns, and atmospheric composition among others. It depends on the latitude and longitude, and regions may vary and evolve differently. We think of the multivariate measurements in different locations across the globe at a given time as a set of data points. Each set sampled from a probability distribution that represents the global weather pattern of the climate. We follow standard descriptions to define the climate as a collection of these sets observed over a period of 20 years. The problem is to test for significant differences in climate, represented by the evolution of (multi-channel) images, over time (see Figure 3).

The data is publicly available, provided by the Copernicus Climate Change Service. We include a total of 12 climate variables identified as essential to characterize the climate[1], including temperature, atmospheric pressure, observed over monthly periods for the last 40 years across Europe. The available data thus consists of a two streams of sets $\{x_{i,j}\}_{j=1}^{n_i}$ and $\{y_{i,j}\}_{j=1}^{n_i}$ for $i = 1, \dots, 144$ (12 months over 20 years). The first describes the climate over the period $1979 - 1999$, and the second set over the period $1999 - 2019$. Both contain measurements $x_{i,j} \in \mathbb{R}^{12}$ ($y_{i,j}$ respectively) in *approximately* $n_i = 250$ different locations (approximately because not all locations are consistently observed over time) which makes the length of each set irregular. Existing tests would thus require some form of interpolation which is not trivial over space and time in this case.

RMMD rejects (with high significance, p-value 0.0002) the hypothesis of equally distributed climate data over the past 4 decades. We note that this result would be much weaker if only a particular location was considered (which could have been a viable reductionist strategy to use existing tests). For instance, we found that the RMMD applied to climate data over the same periods in London and Paris to not be significantly different (p-value 0.21). This experiment demonstrates the potential benefits of using more flexible tests that better represent available data to faithfully investigate complex phenomena such as climate that involve multiple measurements over time and space.

7 Conclusions

By way of conclusion, we emphasize the importance of modelling assumptions for the consistency of hypothesis tests. Many problems are not naturally amenable to hypothesis testing without previously coercing the data, a first step that we have shown to hamper performance and that voids existing theoretical guarantees.

In this paper we extended the toolkit of applied statisticians to do hypothesis testing on uncertain *set*-valued data. We show that by appropriately representing each set of observations in a Hilbert space, kernel-based hypothesis testing may be applied consistently. Specifically, we introduced tests for the two-sample and the independence problem, derived their asymptotic distributions and provided efficient algorithms and optimization schemes to analyse a wide range of scenarios in an automatic fashion.

References

[1] Alexis Bellot and Mihaela van der Schaar. Conditional independence testing using generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 2202–2211, 2019.

[2] Alessio Benavoli and Francesca Mangili. Gaussian processes for bayesian hypothesis tests on regression functions. In *Artificial intelligence and statistics*, pages 74–82, 2015.

[3] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv preprint arXiv:1711.07910*, 2017.

[4] Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. In *Advances in neural information processing systems*, pages 406–414, 2010.

[1] https://public.wmo.int/en/programmes/global-climate-observing-system/essential-climate-variables

[5] Héctor Climente-González, Chloé-Agathe Azencott, Samuel Kaski, and Makoto Yamada. Block hsic lasso: model-free biomarker detection for ultra-high dimensional data. *bioRxiv*, page 532192, 2019.

[6] Rui Gao, Liyan Xie, Yao Xie, and Huan Xu. Robust hypothesis testing using wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, pages 7902–7912, 2018.

[7] 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.

[8] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[9] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pages 673–681, 2009.

[10] Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592, 2008.

[11] Gökhan Gül and Abdelhak M Zoubir. Robust hypothesis testing with $\backslash\alpha$-divergence. *IEEE Transactions on Signal Processing*, 64(18):4737–4750, 2016.

[12] Tommi Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Advances in neural information processing systems*, pages 487–493, 1999.

[13] Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *Journal of Machine Learning Research*, 5(Jul):819–844, 2004.

[14] Wittawat Jitkrittum, Zoltén Szabó, and Arthur Gretton. An adaptive test of independence with analytic kernel embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1742–1751. JMLR. org, 2017.

[15] Eitan Kerem, Joseph Reisman, Mary Corey, Gerard J Canny, and Henry Levison. Prediction of mortality in patients with cystic fibrosis. *New England Journal of Medicine*, 326(18):1187–1191, 1992.

[16] Eitan Kerem, Laura Viviani, Anna Zolin, Stephanie MacNeill, Elpis Hatziagorou, Helmut Ellemunter, Pavel Drevinek, Vincent Gulmans, Uros Krivec, and Hanne Olesen. Factors associated with fev1 decline in cystic fibrosis: analysis of the ecfs patient registry. *European Respiratory Journal*, 43(1):125–133, 2014.

[17] Risi Kondor and Tony Jebara. A kernel between sets of vectors. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 361–368, 2003.

[18] John Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6(Jan):129–163, 2005.

[19] Ho Chung Law, Christopher Yau, and Dino Sejdinovic. Testing and learning on distributions with symmetric noise invariance. In *Advances in Neural Information Processing Systems*, pages 1343–1353, 2017.

[20] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[21] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and DJ Sutherland. Learning deep kernels for non-parametric two-sample tests. *arXiv preprint arXiv:2002.09116*, 2020.

[22] James R Lloyd and Zoubin Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, pages 829–837, 2015.

[23] David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. The randomized dependence coefficient. In *Advances in neural information processing systems*, pages 1–9, 2013.

[24] David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pages 1452–1461, 2015.

[25] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2016.

[26] Pedro J Moreno, Purdy P Ho, and Nuno Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Advances in neural information processing systems*, pages 1385–1392, 2004.

[27] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *Advances in neural information processing systems*, pages 10–18, 2012.

[28] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *arXiv preprint arXiv:1605.09522*, 2016.

[29] Victor M Panaretos, David Kraus, and John H Maddocks. Second-order comparison of gaussian random functions and the geometry of dna minicircles. *Journal of the American Statistical Association*, 105(490):670–682, 2010.

[30] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[31] Anant Raj, Ho Chung Leon Law, Dino Sejdinovic, and Mijung Park. A differentially private kernel two-sample test. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 697–724. Springer, 2019.

[32] Walter Rudin. *Fourier analysis on groups*, volume 121967. Wiley Online Library, 1962.

[33] Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

[34] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, Gert RG Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[35] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

[36] Dougal J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*, 2016.

[37] Zoltán Szabó, Arthur Gretton, Barnabás Póczos, and Bharath Sriperumbudur. Two-stage sampled learning theory on distributions. In *Artificial Intelligence and Statistics*, pages 948–957, 2015.

[38] Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

[39] David Taylor-Robinson, Margaret Whitehead, Finn Diderichsen, Hanne Vebert Olesen, Tania Pressler, Rosalind L Smyth, and Peter Diggle. Understanding the natural progression in% fev1 decline in patients with cystic fibrosis: a longitudinal study. *Thorax*, 67(10):860–866, 2012.

[40] Aad W van der Vaart and Jon A Wellner. The delta-method. In *Weak Convergence and Empirical Processes*, pages 372–400. Springer, 1996.

[41] Jeffrey S Wagener, Michael J Williams, Stefanie J Millar, Wayne J Morgan, David J Pasta, and Michael W Konstan. Pulmonary exacerbations and acute declines in lung function in patients with cystic fibrosis. *Journal of Cystic Fibrosis*, 17(4):496–502, 2018.

[42] Yee Hwa Yang and Terry Speed. Design issues for cdna microarray experiments. *Nature Reviews Genetics*, 3(8):579, 2002.

[43] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in neural information processing systems*, pages 3391–3401, 2017.

[44] Jin-Ting Zhang. Statistical inferences for linear models with functional responses. *Statistica Sinica*, pages 1431–1451, 2011.

[45] Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018.

Appendices

This appendix provides additional material accompanying the paper "Kernel Hypothesis Testing with Set-valued Data".

Appendix A: Proofs

A.1. Asymptotic distribution of $\widehat{\text{RMMD}}^2$

Our proof strategy consists of demonstrating convergence in probability of each inner product $K(\hat{\mu}_{\mathbb{P}}, \hat{\mu}_{\mathbb{Q}})$ to its population counterpart $K(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}})$, and take also into account approximations to the embeddings themselves we might make such as with Fourier features. Given convergence in probability, the equivalence of their asymptotic distributions then follows by convergence results of random variables.

All results in this section consider the asymptotic regime of increasing sample size and increasing set size. We therefore make abstraction for notational purposes of our weighting mechanism, assumed fixed and each weight identical across sets asymptotically which is equivalent to reverting to the equal weight scenario for our asymptotic results.

We start by recalling some definitions. The empirical statistic of the RMMD is given by,

$$\widehat{\text{RMMD}}^2 := \frac{1}{n^2}\sum_{i,j=1}^{n} K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j}) + \frac{1}{m^2}\sum_{i=1}^{m} K(\hat{\mu}_{\mathbb{Q}_i}, \hat{\mu}_{\mathbb{Q}_j}) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{Q}_j}) \tag{7}$$

while the MMD with population mean embeddings is given by,

$$\widehat{\text{MMD}}^2 := \frac{1}{n^2}\sum_{i,j=1}^{n} K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}) + \frac{1}{m^2}\sum_{i=1}^{m} K(\mu_{\mathbb{Q}_i}, \mu_{\mathbb{Q}_j}) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{Q}_j}) \tag{8}$$

Let $n = \rho_n$ and $m = \rho_m t$ where $t = n + m$, and note that,

$$t\widehat{\text{RMMD}}^2 = t\widehat{\text{MMD}}^2 + (t\widehat{\text{RMMD}}^2 - t\widehat{\text{MMD}}^2)$$
$$\sqrt{t}\widehat{\text{RMMD}}^2 = \sqrt{t}\widehat{\text{MMD}}^2 + (\sqrt{t}\widehat{\text{RMMD}}^2 - \sqrt{t}\widehat{\text{MMD}}^2)$$

We are interested in bounding the contribution of the second term in each case under the null and alternative hypotheses asymptotically. The absolute differences we are interested in bounding then under the null hypothesis given by,

$$\begin{aligned}\left|t\widehat{\text{RMMD}}^2 - t\widehat{\text{MMD}}^2\right| \leq & \frac{t}{n^2}\sum_{i=1}^{n}\left|K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}) - K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j})\right| + \frac{t}{m^2}\sum_{i=1}^{m}\left|K(\mu_{\mathbb{Q}_i}, \mu_{\mathbb{Q}_j}) - K(\hat{\mu}_{\mathbb{Q}_i}, \hat{\mu}_{\mathbb{Q}_j})\right| \\ & - \frac{2t}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left|K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{Q}_j}) - K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{Q}_j})\right|\end{aligned}$$

and under the alternative hypothesis,

$$\left|\sqrt{t}\widehat{\text{RMMD}}^2 - \sqrt{t}\widehat{\text{MMD}}^2\right| \leq \frac{\sqrt{t}}{n^2}\sum_{i=1}^{n}\left|K(\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j}) - K(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j})\right| + \frac{\sqrt{t}}{m^2}\sum_{i=1}^{m}\left|K(\mu_{\mathbb{Q}_i},\mu_{\mathbb{Q}_j}) - K(\hat{\mu}_{\mathbb{Q}_i},\hat{\mu}_{\mathbb{Q}_j})\right|$$
$$- \frac{2\sqrt{t}}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left|K(\mu_{\mathbb{P}_i},\mu_{\mathbb{Q}_j}) - K(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{Q}_j})\right|$$

In both cases it suffices to show that inner products between population mean embeddings and empirical counterparts converge in probability. We will traverse this result in two steps, using results that show the convergence of empirical mean embeddings to their population counterparts and using a Lipschitz condition to extend this to inner products between mean embeddings. Assume K to be a real-valued, shift invariant ($K(x,x') = K(x-x',0)$), and L_K-Lipschitz kernel,

$$|K(x,0) - K(x',0)| \leq L_K|x - x'| \tag{9}$$

also satisfying the boundedness condition $|K(x,x')| < 1$ for all $x, x' \in \mathcal{X}$.

The following two Lemmas demonstrate our claim.

Lemma 1 (Bound on the empirical mean embedding [24]) *Let the kernel K satisfy the assumptions above. Then we have,*

$$|\mu_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_i}|_{\mathcal{H}_K} \leq 2\sqrt{\frac{\mathbb{E}_{x\sim\mathbb{P}_i}K(x,x)}{n_i}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n_i}} \tag{10}$$

with probability at least $1-\delta$ over the randomness in the empirical sample from $\mathbb{P}_i$. n_i is the number of samples from $\mathbb{P}_i$.

Lemma 2 (Bound on kernels computed on empirical mean embeddings) *Let K be defined as above. The it holds that for any $\epsilon > 0$,*

$$Pr(\left|K(\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j}) - K(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j})\right| > \epsilon) \leq \exp\left\{2 - \frac{\epsilon^2 n}{8L_K}\right\} \tag{11}$$

as $n := \min(n_i, n_j) \to \infty$ we get that the limit of the above probability is 0, that is, $K(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j})$ converges in probability to $K(\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j})$.

Proof. The proof is based on the Lipschitz condition and the error bound on empirical mean embeddings with respect to their population counterparts.

$$\begin{aligned}
|K(\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j}) - K(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j})| &= \left|K(\mu_{\mathbb{P}_i} - \mu_{\mathbb{P}_j},0) - K(\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_j},0)\right| &(12)\\
&\leq L_K\left|\mu_{\mathbb{P}_i} - \mu_{\mathbb{P}_j} - (\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_j})\right| &(13)\\
&\leq L_K\left|\mu_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_i}\right| + L_K\left|\mu_{\mathbb{P}_j} - \hat{\mu}_{\mathbb{P}_j}\right| &(14)\\
&\leq L_K\left(2\sqrt{\frac{\mathbb{E}_{x\sim\mathbb{P}_i}K(x,x)}{n_i}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n_i}} + 2\sqrt{\frac{\mathbb{E}_{x\sim\mathbb{P}_j}K(x,x)}{n_j}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n_j}}\right) &(15)\\
&\leq L_K\left(\sqrt{\frac{4+2\log\frac{1}{\delta}}{n_i}} + \sqrt{\frac{4+2\log\frac{1}{\delta}}{n_j}}\right) &(16)
\end{aligned}$$

where the last line follows from the inequality: $\sqrt{x} + \sqrt{y} \leq \sqrt{x+y}, \forall x, y > 0$. Moreover, we have for $n := \min(n_i, n_j)$, by letting $\epsilon = 2L_K\sqrt{\frac{4+2\log(1/\delta)}{n}}$ such that $\delta = \exp\{2 - \frac{\epsilon^2 n}{8L_K}\}$,

$$Pr(\left|k(\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j})\right| > \epsilon) \leq \exp\left\{2 - \frac{\epsilon^2 n}{8L_K}\right\} \tag{17}$$

as $n \to \infty$ we get that the limit of the above probability is 0 which means that $k(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j})$ converges in probability to $k(\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j})$.

As a consequence then, the asymptotic distributions of $t\widehat{\text{RMMD}}^2$ and $t\widehat{\text{MMD}}^2$, and, $\sqrt{t}\widehat{\text{RMMD}}^2$ and $\sqrt{t}\widehat{\text{MMD}}^2$ coincide.

Using random Fourier features. For completeness, in addition to considering convergence in distribution using empirical embeddings, we extend our analysis to include Fourier feature approximations in the empirical embeddings themselves and their asymptotic behaviour. To do so notice that we may write,

$$\begin{aligned}|k(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}) - &k(\hat{\mu}_{\mathbb{P}_i,m}, \hat{\mu}_{\mathbb{P}_j,m})| \leq \\ &\left|k(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j})\right| + \left|k(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_i,m}, \hat{\mu}_{\mathbb{P}_j,m})\right| \end{aligned} \tag{18}$$

by the triangle inequality.

The following two lemmas are similar to the first two above but instead related the empirical mean embedding $\hat{\mu}_{\mathbb{P}_i}$ with its random Fourier feature approximation $\hat{\mu}_{\mathbb{P}_i,m}$

Lemma 3 (Bound on the randomized empirical mean embedding [24]) *Let k be defined as above. For a fixed sample of size n_i from a probability distribution $\mathbb{P}_i$ on $\mathbb{R}^d$ and any $\delta > 0$, we have,*

$$|\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_i,m}|_{L^2(\mathbb{P})} \leq \frac{2}{\sqrt{m}}\left(1 + \sqrt{2\log n_i/\delta}\right) \tag{19}$$

with probability larger than $1 - \delta$ over the randomness of the samples $(\omega_i, b_i)_{i=1}^m$.

Lemma 4 (Bound on kernels computed on approximated empirical mean embeddings) *Let k be defined as above. Then for any $\epsilon > 0$ it holds that,*

$$Pr(\left|k(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_i,m}, \hat{\mu}_{\mathbb{P}_j,m})\right| > \epsilon) \leq n \exp\left\{-\frac{1}{8}\left(\frac{\epsilon\sqrt{m}}{2L_k} - 1\right)^2\right\} \tag{20}$$

m is the number of random features, n_i and n_j are the number of observations in time series X_i and X_j respectively, and $n := \min(n_i, n_j)$. If further we assume that $\min(n_i, n_j)\exp\{-m\} \to 0$ as $n_i, n_j, m \to \infty$, then $k(\hat{\mu}_{\mathbb{P}_i,m}, \hat{\mu}_{\mathbb{P}_j,m})$ converges in probability to $k(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j})$.

Proof. The proof strategy is similar to Lemma 3, but for with a different bound on the difference between mean embeddings. We proceed as follows,

$$\begin{aligned}\left|k(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_i,m}, \hat{\mu}_{\mathbb{P}_j,m})\right| &= \left|k(\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_j}, 0) - k(\hat{\mu}_{\mathbb{P}_i,m} - \hat{\mu}_{\mathbb{P}_j,m}, 0)\right| && (21)\\ &\leq L_k\left|\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_j} - (\hat{\mu}_{\mathbb{P}_i,m} - \hat{\mu}_{\mathbb{P}_j,m})\right| && (22)\\ &\leq L_k\left|\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_i,m}\right| + L_k\left|\hat{\mu}_{\mathbb{P}_j} - \hat{\mu}_{\mathbb{P}_j,m}\right| && (23)\\ &\leq \frac{2L_k}{\sqrt{m}}\left(2 + \sqrt{2\log(n_i/\delta)} + \sqrt{2\log(n_j/\delta)}\right) && (24)\\ &\leq \frac{2L_k}{\sqrt{m}}\left(2 + 2\sqrt{2\log(n/\delta)}\right) && (25)\end{aligned}$$

where we have written $n := \min(n_i, n_j)$ and the inequalities hold with probability at least $(1 - \delta)$ over the randomness of the samples $(\omega_i, b_i)_{i=1}^m$. Now, set $\epsilon := n\exp\left\{-\frac{1}{8}\left(\frac{\epsilon\sqrt{m}}{2L_k} - 2\right)^2\right\}$. Then,

$$Pr(\left|k(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_i,m}, \hat{\mu}_{\mathbb{P}_j,m})\right| > \epsilon) \leq n \exp\left\{-\frac{1}{8}\left(\frac{\epsilon\sqrt{m}}{2L_k} - 2\right)^2\right\} \tag{26}$$

With the condition that $n\exp(-m) \to 0$ as $n, m \to \infty$, $k(\hat{\mu}_{\mathbb{P}_i,m}, \hat{\mu}_{\mathbb{P}_j,m})$ converges in probability $k(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j})$.

A.2. Asymptotic distribution of $\widehat{\text{RHSIC}}$

Proposition 2 ***(Asymptotic distribution RHSIC).*** *Let a sample of paired sets be defined as in the main body of this paper and let K and L be characteristic and L_K and L_L-Lipschitz continuous respectively. In the asymptotic regime where both the number of paired sets N and set sizes n_i tends to infinity,*

- *Under the null $\mathcal{M}_{PQ} = \mathcal{M}_P\mathcal{M}_Q$, the distributions of $\sqrt{N}\widehat{HSIC}^2$ and $\sqrt{N}\widehat{RHSIC}^2$ coincide.*
- *Under the alternative $\mathcal{M}_{PQ} \neq \mathcal{M}_P\mathcal{M}_Q$, the distributions of $N\widehat{HSIC}^2$ and $N\widehat{RHSIC}^2$ coincide.*

Proof. We use a similar proof strategy to that used above. The $\widehat{\text{RHSIC}}$ may be written as a sum of V-statistics as follows [10],

$$\widehat{\text{RHSIC}} = \frac{1}{N^2}\sum_{i,j}^{N}\hat{K}_{ij}\hat{L}_{ij} + \frac{1}{N^4}\sum_{i,j,q,r}^{N}\hat{K}_{ij}\hat{l}_{qr} - \frac{2}{N^3}\sum_{i,j,q}^{N}\hat{K}_{ij}\hat{l}_{iq} \tag{27}$$

where to avoid cluttering the notation we have written $\hat{K}_{ij} := K(\hat{\mu}_{\mathbb{P}_i,m}, \hat{\mu}_{\mathbb{P}_j,m})$ and $\hat{L}_{ij} := L(\mu_{Y_i,m}, \mu_{Y_j,m})$. Sums with two summation indeces refer to double sums of all pairs of numbers drawn with replacement from $\{1, ..., N\}$, and similarly for three and four summation indeces [10]. Similarly to the two sample problem, equality in asymptotic distribution may be shown by considering the absolute differences in the product of population and empirical kernels. That is, we are interested in bounding the following,

$$|\hat{K}_{ij}\hat{L}_{qr} - K_{ij}L_{qr}| \tag{28}$$

for any quadruple of indeces i, j, q, r.

Assuming as above that kernels K and L are Lipschitz functions it follows that their product is also Lipschitz,

$$\begin{aligned}
|K(x,0)L(y,0) - &K(x',0)L(y',0)| \\
&\leq |(K(x,0) - K(x',0))L(y,0) + (L(y,0) - L(y',0))K(x',0)| \\
&\leq |K(x,0) - K(x',0)| \cdot ||L(y,0)||_{\mathcal{H}_L} + |L(y,0) - L(y',0)| \cdot ||K(x',0)||_{\mathcal{H}_K} \\
&\leq L_K|x - x'| + L_L|y - y'|
\end{aligned}$$

The same arguments and lemmas used in the two-sample case apply which proves the equivalence in asymptotic distributions of the $\widehat{\text{RHSIC}}$ and $\widehat{\text{HSIC}}$.

A.3. Approximations for high power: Kernel hyperparameters

For the two sample problem, let N be the number of samples in both groups, which simplifies the formulation of the asymptotic power of the $\widehat{\text{RMMD}}^2$. The following procedure mirrors [36].

Proposition 3 ***(Approximate power of $\widehat{RMMD}^2$).*** *Under $\mathcal{H}_1$, for large N and fixed r, the test power $Pr(N\widehat{\text{RMMD}}^2 > r) \approx 1 - \Phi(\frac{r}{\sqrt{N}\sigma_{\text{RMMD}}} - \sqrt{N}\frac{\text{RMMD}^2}{\sigma_{\text{RMMD}}})$ where Φ denotes the cumulative distribution function of the standard normal distribution, σ^2_{RMMD} is the asymptotic variance under $\mathcal{H}_1$ for the $\widehat{RMMD}^2$.*

Consider the terms inside the *cdf* of the normal. Observe that the first term $\frac{r}{\sqrt{N}\sigma_{\text{RMMD}}} = \mathcal{O}(N^{-1/2})$ goes to 0 as $N \to \infty$, while the second term, $\sqrt{N}\frac{\text{RMMD}^2}{\sigma_{\text{RMMD}}} = \mathcal{O}(N^{1/2})$, dominates the first one for large N. As an approximation, for sufficiently large N, the parameters that maximize the test power are given by $\theta^* = \text{argmax}_\theta \; Pr(N\widehat{\text{RMMD}}^2 > r) \approx \frac{\text{RMMD}^2}{\sigma_{\text{RMMD}}}$. In our case θ includes the bandwidth parameter used to compute the mean embeddings and the bandwidth parameter used to compute the test statistic. The empirical estimate of the variance $\hat{\sigma}_{\text{RMMD}}$ that appears in our objective is approximated up to second order terms, as in [36]. Similar derivations hold for the power optimization of the HSIC with the exception that the definition of the HSIC requires optimization of two kernels, one for each set in our paired samples: K and L.

Note that since RMMD and σ_{RMMD} are unknown, to maintain the validity of the hypothesis test we divide the sample into a training set, used to estimate the ratio with $\frac{\widehat{\text{RMMD}}^2}{\hat{\sigma}_{\text{RMMD}}}$ and choose the kernel parameters, and a testing set used to perform the final hypothesis test with the learned kernels.

An analogous result holds for the approximate power of $\widehat{\text{RHSIC}}$.

A.4. Approximations for high power: Weighting scheme

Under the alternative hypothesis, the asymptotic variance of the proposed test statistics is well defined and given by asymptotic theory of V-Statistics (up to scaling) equal to Var($\mathbb{E}K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j})$), see e.g. Theorem 5.5.1 [33]. To specify the set of weights that maximize power we may use the same reasoning to the section above and minimize the asymptotic variance.

With finite samples to approximate the mean embedding, assuming that all randomness comes from the number of samples available to estimate mean embeddings, its variance is proportional to $1/n_i$. The delta method (see e.g. [40]) may be applied on the bivariate sample $(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j})$ with the function K to conclude that the variance of each $K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j})$ is proportional to $1/(n_i \cdot n_j)$. Now, with a finite number of sets, or in other words a finite number of distributions, we approximate the expectation $\mathbb{E}K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j})$ with averages. Assuming that the covariance between any pair $K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j})$ and $K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_k})$ for any i, j, k does not vary by changing indeces, that is, is fixed, weighting each term $K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j})$ with the inverse of its variance gives the lowest attainable variance Var($\mathbb{E}K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j})$) in finite samples.

Appendix B: Additional details on experiments and implementation

B1. Details on the data generation mechanisms

The inverse gamma distribution has appeared parameterized by one and two parameters. We choose the one-parameter distributions with density,

$$f(x; \mu) = \frac{x^{-\mu-1}}{\Gamma(\mu)} \exp\left(-1/x\right) \tag{29}$$

where $x \geq 0$, $\mu > 0$ and Γ is the gamma function.

B2. RMMD and RHSIC

We create empirical kernel mean embeddings by concatenating data along each dimension. Each embedding has random features sampled to approximate a Gaussian kernel with length scale parameter σ^2. σ^2 is estimated by cross-validation on a grid of parameter values around the median of squared pairwise distances of the stacked data. In practice, we set the number of random features to $m = 50$ (larger amounts of random features show no significant performance improvements). The parameters of the kernel used for testing are similarly optimized via cross-validation by defining a grid of parameter values around the median of squared pairwise distances of computed random features. In summary, for each random feature length-scale we test with a number of test length-scales and choose the pair of parameters with best performance according to our power criterion. A summary of these tests' implementation is as follows.

1. For each observed set $\{x_{i,j}\}_{j=1}^{n_i} \sim \mathbb{P}_i$, compute its approximated mean embedding using a Fourier basis, with elements in the span of $(\cos(\langle \omega_j, x\rangle + b_j))_{j=1}^m$,

$$\hat{\mu}_{\mathbb{P}_i, m} = \frac{1}{n_i} \sum_{x \in \{x_{ij}\}_{j=1}^{n_i}} (\cos(\langle w_j, x\rangle + b_j))_{j=1}^m \in \mathbb{R}^m$$

2. Compute weights that describe the confidence we have in each of the above approximations, $w_{\mathbb{P}_i} := n_i / \sum_i n_i$ for each i, that result in posterior test statistics with lowest variance.

3. Compute two-sample or independence test statistics on this weighted representation of the data to obtain a real-valued scalar $\hat{t}$ that discriminates between the two hypotheses of interest.
4. In practice, a test decision will be made based on a comparison of the computed value $\hat{t}$ with an approximated null distribution obtained by repeated test statistic computation on permuted data representations. If $\hat{t}$ is greater than the α quantile of this approximated null distribution, reject the null hypothesis, otherwise fail to reject.

B3. GP2ST

The test developed by [2] was designed to test the equality of regression functions from observed two-dimensional data $(\mathbf{t}_1, \mathbf{y}_1)$ and $(\mathbf{t}_2, \mathbf{y}_2)$ from two samples. They assume a GP prior on the time series and compute posterior distributions by conditioning on each sample of observed data. Denote the posterior GPs by f_1 and f_2. With the assumption of gaussianity it follows that $\Delta f := f_1 - f_2$ is also a GP, and evaluations on a fine grid of regular times $\mathbf{t}$ in $[0, 1]$ will be multivariate Gaussian with mean denoted $\Delta\mu$ and covariance matrix $\Delta\Sigma$. The hypothesis of equality of data generating processes is then equivalent to testing departures of Δf from the zero function. As a result, the two functions are equal with posterior probability $1 - \alpha$ if the credible region for Δf includes the zero vector or, in other words, if:

$$\Delta\mu^T \Delta\Sigma^{-1} \Delta\mu \leq \chi^2_v(1 - \alpha) \tag{30}$$

$\chi^2_v(1 - \alpha)$ is the $(1 - \alpha)$-quantile of a χ^2 distribution with v degrees of freedoms and v is the number of positive eigenvalues of $\Delta\Sigma$.

B4. RDC

The Randomized Dependence Coefficient (RDC) measures the dependence between fixed-dimensional random samples X and Y as the largest canonical correlation between k randomly chosen nonlinear projections of their copula transformations. It is formally defined an analyzed in [23].

$$\hat{\rho}(\mathbf{x}, \mathbf{y}) := \sup_{\alpha,\beta} PCC(\alpha^T \Phi_{\mathbf{x}}, \beta^T \Phi_{\mathbf{y}})$$

where PCC is Pearson's correlation coefficient and Φ are nonlinear random projections, such as sine or cosine projections. To apply this function on irregularly observed data, we interpolate as we do with the MMD and HSIC.

We conduct a test using this measure of dependence by repeatedly shuffling the paired time series M times to induce an empirical distribution of $\{\hat{\rho}_m\}_{m=1}^M$ under the null hypothesis of independence. The p-value is then given by $\sum_{m=1}^M \mathbf{1}\{\hat{\rho}_m > \hat{\rho}\}/M$ where $\hat{\rho}$ is the statistic obtained from the observed data.

B5. PCC

The Pearson's correlation coefficient (PCC) is a measure of linear correlation between two variables. It is defined as,

$$\hat{\rho}(\mathbf{x}, \mathbf{y}) := \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})}\sqrt{\sum_i (y_i - \bar{y})}}$$

Similarly to the RDC, we conduct a test using this measure of dependence by repeatedly shuffling the paired time series M times to induce an empirical distribution of $\{\hat{\rho}_m\}_{m=1}^M$ under the null hypothesis of independence.

B6. C2ST

We implemented the C2ST with tensorflow in python. We used a RNN with GRU cells in one version and the deepset architecture in another. The number of samples in each mini-batch is set to 64 the hidden layer size to 10. We optimize model parameters with Adam and learning rate equal to 0.01, while all variables are initialized with Xavier initialization. We use sigmoid and tanh as the activation functions for each layer and use sigmoid activation for the output layer given that we perform classification. For the architecture designed for set-valued data we use the existing implementation provided by [43].

Both tests proceeds as follows [25]:

Let $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ be two samples of observed time series that include their corresponding time points in each case.

1. Construct the data set $\mathcal{D} = \{(x_i, 0)\}_{i=1}^n \cup \{(y_i, 1)\}_{i=1}^n =: \{(z_i, l_i)\}_{i=1}^{2n}$.
2. Shuffle $\mathcal{D}$ at random and partition into a training set $\mathcal{D}_{tr}$ and a testing set $\mathcal{D}_{te}$.
3. Fit a classifier g on the training set to predict the sample indicator l.
4. Compute test statistic as classification accuracy on $\mathcal{D}_{te}$: $\widehat{t} := \frac{1}{n_{te}} \sum_{(z_i,l_i)\in\mathcal{D}_{te}} \mathbf{1}\{\mathbf{1}\{g(z_i) > 1/2\} = l_i\}$
5. If $\widehat{t}$ is greater that the α quantile of a $\mathcal{N}(1/2, 1/(4n_{te}))$ reject $\mathcal{H}_0$; otherwise accept $\mathcal{H}_0$.

$\mathbf{1}$ is the indicator function.