Application of Kernel Hypothesis Testing on Set-valued Data

Alexis Bellot[1,2] **Mihaela van der Schaar**[1,2,3]

[1]University of Cambridge, Cambridge, UK.
[2]Alan Turing Institute, London, UK
[3]University of California, Los Angeles, Los Angeles, USA

Abstract

We present a general framework for kernel 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 and real data experiments, 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 experimental data and quantify the evidence for discoveries on the relationship of variables of interest, see e.g. [23] for a general introduction. As a 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, see e.g. [12, 11, 9, 47], tests based on neural network representations as developed in [24, 27, 3], 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. Many problems do exhibit these properties, for example with medical data, where each patient has different levels of variation and have observations irregularly measured over time. A similar pattern is observed in many other domains involving time series and bagged data (e.g. multiple images of the same phenomenon).

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 in regression problems where one seeks to learn a mapping from distributions to associated labels, see e.g. [40, 41], 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 propose an approach applicable to any kernel-based test that includes,

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

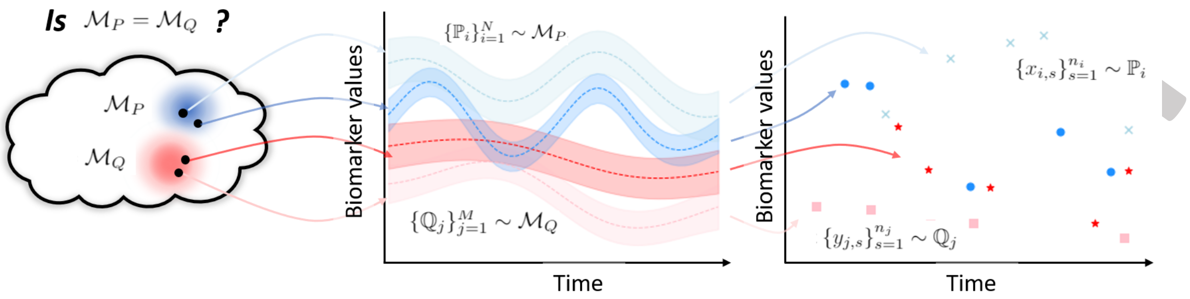
Is $\mathcal{M}_P = \mathcal{M}_Q$?
$\mathcal{M}_P$
$\mathcal{M}_Q$
$\{\mathbb{P}_i\}_{i=1}^N \sim \mathcal{M}_P$
$\{\mathbb{Q}_j\}_{j=1}^M \sim \mathcal{M}_Q$
Biomarker values
Time
$\{x_{i,s}\}_{s=1}^{n_i} \sim \mathbb{P}_i$
$\{y_{j,s}\}_{s=1}^{n_j} \sim \mathbb{Q}_j$
Biomarker values
Time

Figure 1: We consider an example from electronic health records to illustrate the proposed approach. **Right panel:** 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. **Middle panel**: we encode the uncertainty in each patient trajectory by a probability distributions on the space of observations. **Left panel:** The two-sample problem is to test for equality in distribution on the space of patient-specific distributions, rather than actual 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 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.

Our contributions are three-fold:

1. We formally describe tests on set-valued data, and to the best of our knowledge for the first time.
2. We demonstrate the consistency of these tests for the two-sample and independence testing problems.
3. We validate 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.

2 BACKGROUND

The tests presented in this paper are formally 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 *sets* of 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 [30]. Both are defined in the corresponding Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ by,

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

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. Christmann et al. [6] 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, $\mu_{\mathbb{P}}$ replacing $x \in \mathcal{X}$ in the conventional learning setting as inputs to k, see eq. (2) below.

Accounting for variance in embedding approximations. 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 of 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 [36]) 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 $K(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}}) := \exp(-||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||^2_{\mathcal{H}_K}/2\sigma^2)$ [6, 29]. Note that for kernels on $\mathcal{X}$, their RKHS consists of functions $\mathcal{X} \rightarrow \mathbb{R}$, while the kernel K lives on the space of distributions on $\mathcal{X}$, $\mathcal{P}(\mathcal{X})$, and its RKHS consists of functions $\mathcal{P}(\mathcal{X}) \rightarrow \mathbb{R}$. We may use K to learn from samples that are individual distributions, rather than individual observations, as described in [6].

Relationships with learning on distributions. With this construction (i.e. kernels evaluated on mean embeddings) [40] 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 in [26] and for transfer learning, see e.g. [5]. 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 [9]. 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 [37]. 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 [9]. 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 [6] demonstrated that for well known kernels, such as the Gaussian kernels, if used in both levels of the embedding and defined on a compact metric space the resulting embedding is injective (i.e. kernels are characteristic)[1].

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 [23]. 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

Distances on sets. As a first observation, note that kernels defined on sets directly, as done e.g. by [19], 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 [15], the Fisher kernel [14], diffusion kernels [20] and kernels arising from Kullback-Leibler divergences [28], none of them known to be characteristic and in this case with the shortcoming that many of the above are parameterized by a family of densities which may or may not hold in data.

Possible extensions to other tests. Deep learning has emerged as an alternative for defining tests on structured objects. [27] define classifier two-sample tests and [24] 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.

Other connections with hypothesis testing. 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 [8, 13]. Other types of invariances can also be enforced, for instance [21] 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

[1] Theorem 2.2 [6] technically shows that such kernels are universal, but universal kernels on a compact metric space are known to be characteristic, see e.g. Theorem 1 [9].

this uncertainty, for instance [4] use Gaussian processes and compare posterior distributions. More generally, tests in the functional data analysis literature, such as [46, 31, 44, 35, 7], may be consistently applied on regularly sampled time series data with a strong time-dependence. The set representation in (1) assumes instead each observation (and time stamp in the case of time series) to be drawn independently, a formalism that may be adequate for some problems (e.g. sufficiently sparse and irregular time series as observed in primary care electronic health records see e.g. [2, 1, 22] and other set-valued data) but not others (e.g. frequently sampled time series).

3 HYPOTHESIS TESTS ON 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 [9, 12] 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.

Algorithm. 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 [9],

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

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. Assume for now that all weights are fixed $w_{\mathbb{P}_i} = 1/N, w_{\mathbb{Q}_j} = 1/M$ for all i, j. We return to the specification of weights in section 3.3. The asymptotic behaviour of $\widehat{\text{MMD}}^2$ is well understood [9] and the test itself is extensively used in many applications [25, 33]. 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. Then, under the null and alternative and in the regime of increasing set size n_i and increasing sample size n, the asymptotic distributions of $\widehat{RMMD}^2$ coincides with that of $\widehat{MMD}^2$.*

Proof. All proofs are given in the Appendix.

In other words, the additional variability due to a second level of sampling converges to 0 asymptotically, and thus the asymptotic distribution coverges to that of the well known MMD two sample test of [9].

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,

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

Example. Consider an example from healthcare to illustrate this problem.

- A similar set-up as that given in Figure 1 may be used to illustrate independence testing with set-valued data. A common problem is identify dependencies between biomarkers, often observed irregularly over time in many patients. For instance cholesterol levels $\{x_{i,t_1}, \dots, x_{i,t_{n_i}}\}$ and blood pressure $\{y_{i,t_1}, \dots, y_{i,t_{n_i}}\}$ may be observed over times $t_1, \dots, t_{n_i}$ in N individuals $i = 1, \dots, N$. To formally test for dependencies between these samples one must account for the irregularity in observation time and uncertainty in biomarker reads. This can be done by considering instead distributions $\mathbb{P}_i$ and $\mathbb{Q}_i$ and testing for independence in this space directly.

As in the two-sample test, we may quantify the difference between distributions using the RKHS distance $||\mu_{\mathcal{M}_{PQ}} - \mu_{\mathcal{M}_P} \otimes \mu_{\mathcal{M}_Q}||^2_{HS}$. Kernels K, L are assumed characteristic; $|| \cdot ||_{HS}$ is the norm on the space of $\mathcal{H}_K \to \mathcal{H}_L$ Hilbert-Schmidt operators, and $\otimes$ denotes the tensor product, such that $(a \otimes b)c = a\langle b, c\rangle$ for a, b, c elements of a Hilbert space. This distance is called the Hilbert Schmidt Independence Criterion (HSIC) [10, 12].

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 $\mathcal{M}_{PQ}$,

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

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. Assume for now that all weights are fixed $w_{\mathbb{P}_i} = 1/N, w_{\mathbb{Q}_j} = 1/M$ for all i, j. The centering matrix is defined by $H = I - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ and Tr is the trace operator.

Here, similarly to the two sample problem, 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, that asymptotic distributions of the RHSIC and the HSIC 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).

Proposition 2 ***(Asymptotic distribution)****. Let two samples of data be defined as above and let K be characteristic and L_K-Lipschitz continuous. Then, under the null and alternative and in the regime of increasing set size n_i and increasing sample size n, the asymptotic distributions of $\widehat{RHSIC}$ coincides with that of $\widehat{HSIC}$.*

Independence testing with the $\widehat{\text{HSIC}}$ has been studied in [12, 47, 16].

3.3 PRACTICAL REMARKS

We make a number of remarks on the practical application of our tests.

- **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 [39, 16]. 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 often not scalable for even to large 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 as proposed by [34, 32]. If we draw m samples from the Gaussian spectral measure, we can approximate the Gaussian kernel k by,

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

where $\omega_1,...,\omega_m \sim \mathcal{N}(0,\gamma)$, $b_1,...,b_m \sim \mathcal{U}[0,2\pi]$, and $\phi(x) = \sqrt{\frac{2}{m}}[\cos(\omega_1 x + b_1),...,\cos(\omega_m x + b_m)] \in \mathbb{R}^m$ [32]. 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,

$$\hat{\mu}_{\mathbb{P}_i,m} = \frac{1}{n_i}\sum_{x\in\{x_{ij}\}_{j=1}^{n_i}}\sqrt{\frac{2}{m}}(\cos(\langle w_j,x\rangle+b_j))_{j=1}^{m} \in \mathbb{R}^m$$

All implementation details, including the above approximations, are given in the Appendix.

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** [9] with hyperparameters optimized for maximum power, two-sample classifier tests [27] 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 [45] modelling permutation invariance to be expected in sets (**C2ST-Sets**). We consider also the Gaussian process-based test (**GP2ST**) by [4].
- For the independence problem we consider: the **HSIC** [12], the Randomized Dependence Coefficient (**RDC**) [26] 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

Experiment design. 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. 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 nonlinear dynamics. Each time series, or set of observations, is irregularly sampled with noise levels that vary between sets.

Results. We report performance for the two sample problems in the **top row** of Figure 2. Power is measured in three experiments: *first*, as we increase the difference in time series amplitude (with equal variance $\sigma = 0.1$), second as we increase the observation variance (with equal amplitude $\eta = 1$) between the two populations, and third as the dimension of each time series increases (on data sampled with a single dimension with a difference in amplitude equal to 0.25 and other dimensions with no difference). Type I error is shown as a function of the number of samples.

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 for small differences in amplitude and variance.

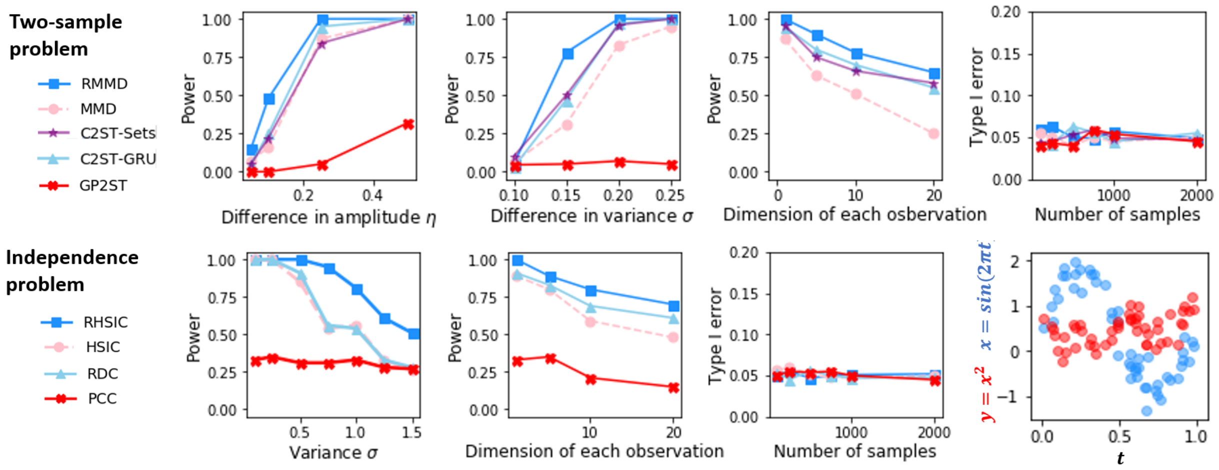

Figure 2: Power (higher better) and Type I error (at level 0.05) on synthetic data. The rightmost panel gives type I error with approximate control at the level $\alpha = 0.05$ for all methods. **Top row**: two-sample problem. **Bottom row**: independence problem. RMMD and RHSIC are the proposed tests.

4.2 INDEPENDENCE PROBLEM

Experiment design. We aim to construct pairs of distributions $(\mathbb{P}_i, \mathbb{Q}_i)$. 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 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. A sample of dependent sets of data using this data generating mechanism is given in the lower rightmost panel of Figure 2.

Results. Power and type I error are shown in the bottom row of Figure . The bottom row of Figure 2 gives performance results for the independence problem. In the first two leftmost panels we evaluate power as we increase the variance of paired time series and as we increase the dimensionality of each observation for a fixed variance $\sigma = 0.5$. The bottom rightmost plot shows a sample of two dependent noisy time series, colored blue and red respectively, for illustration.

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 which we hypothesize is one reason for their underperformance. This is consistent with the increasing variance experiment, in this case increasing variance worsens 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 [17]. 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 [42]. 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 [43, 18].

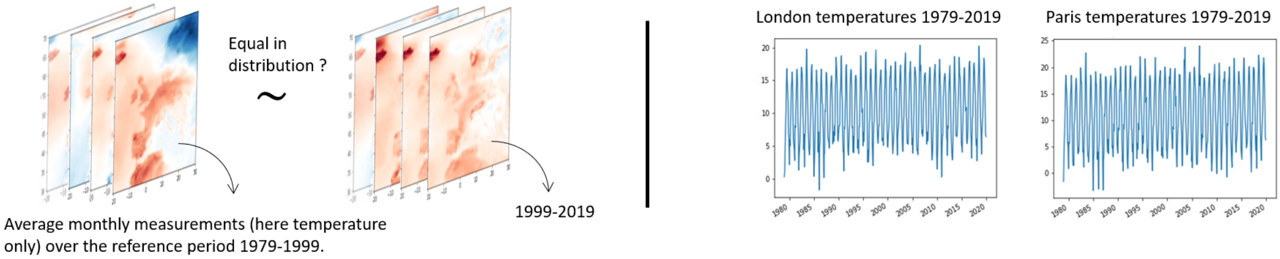
Equal in distribution ?
∼
Average monthly measurements (here temperature only) over the reference period 1979-1999.
1999-2019
London temperatures 1979-2019
Paris temperatures 1979-2019
1980
1985
1990
1995
2000
2005
2010
2015
2020

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

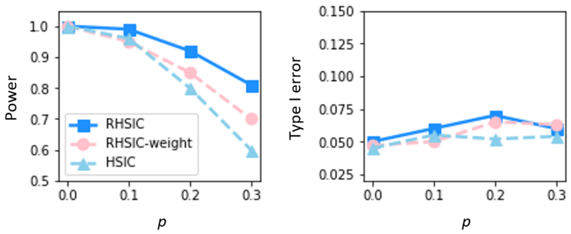

Figure 4: Power and Type I error on Cystic Fibrosis data.

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.

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. 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 over time [38].

Interpretation as set-valued data. We can 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 bags of (multi-channel) images, over time (see Figure 3).

Experiment design. The data is publicly available, provided by the Copernicus Climate Change Service[2]. We include a total of 12 climate variables identified as essential to characterize the climate[3], 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.

Problem. The problem is to test for the hypothesis of equally distributed climate data over the past 4 decades. We conduct 5 different tests: on data from the European, African, North American, South American and South-East Asian regions.

Results. RMMD rejects the hypothesis of equally distributed climate data over the past 4 decades in Europe (p-value 0.0002), Africa (p-value 0.0014), and South America (p-value 0.0001) but fails to reject at a level of 0.01 for North America (p-value 0.016) and South-East Asia (p-value 0.036).

In the case of Europe, we note that this result would be different 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

[2] https://climate.copernicus.eu/.

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

climate data over the same periods in London and Paris to not be significantly different (p-value 0.21). See Figure 3 for the time series of Paris and London temperature data. 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

In this paper we extended the toolkit of applied statisticians to do hypothesis testing on *set*-valued data. We have shown 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.

8 ACKNOWLEDGEMENTS

We thank the anonymous reviewers for valuable feedback and the UK Cystic Fibrosis Trust for access to data. All patient data was deidentified. This work was supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1, the ONR and the NSF grants number 1462245 and number 1533983.

REFERENCES

[1] Ahmed Alaa and Mihaela van der Schaar. Attentive state-space modeling of disease progression. In *Advances in neural information processing systems*, 2019.

[2] Alexis Bellot and Mihaela Van Der Schaar. Flexible modelling of longitudinal medical data: A bayesian nonparametric approach. *ACM Transactions on Computing for Healthcare*, 1(1):1–15, 2020.

[3] 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.

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

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

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

[7] Livio Corain, Viatcheslav B Melas, Andrey Pepelyshev, and Luigi Salmaso. New insights on permutation approach for hypothesis testing on functional data. *Advances in Data Analysis and Classification*, 8(3):339–356, 2014.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] Changhee Lee and Mihaela Van Der Schaar. Temporal phenotyping using deep predictive clustering of disease progression. In *International Conference on Machine Learning*, pages 5767–5777. PMLR, 2020.

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

[24] 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.

[25] 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.

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

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

[28] 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.

[29] 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.

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

[31] 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.

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

[33] 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.

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

[35] Łukasz Smaga and Jin-Ting Zhang. Linear hypothesis testing with functional data. *Technometrics*, 61(1):99–110, 2019.

[36] 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.

[37] 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.

[38] Thomas Stocker. *Introduction to climate modelling*. Springer Science & Business Media, 2011.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] George Wynne and Andrew B Duncan. A kernel two-sample test for functional data. *arXiv preprint arXiv:2008.11095*, 2020.

[45] 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.

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

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