Application of Kernel Hypothesis Testing on Set-valued Data (Supplementary material)

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

This appendix provides additional material to supplement the main body of this paper. It is outlined as follows:

- Section A provides the proofs for all statements made in the main body of this paper.
  - Section A.1 gives the proof of the consistency of the RMMD.
  - Section A.2 gives the proof of the consistency of the RHSIC.
- Section B gives details on the approximations used to deal with irregular set sizes and high-dimensional data.
- Section C gives details on the implementation of baseline tests.

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 (at a fast enough rate), the equivalence of their asymptotic distributions then follows by convergence results of random variables.

A.1.1 Background

All results in this section consider the asymptotic regime of increasing sample size N and increasing set size n_i for each i. We therefore make abstraction for notational simplicity 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,

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

while the MMD with population mean embeddings is given by,

$$\widehat{\text{MMD}}^2 := \frac{1}{N^2}\sum_{i,j=1}^{N} K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}) + \frac{1}{M^2}\sum_{i,j=1}^{M} K(\mu_{\mathbb{Q}_i}, \mu_{\mathbb{Q}_j}) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{Q}_j}). \tag{2}$$

We assume without loss of generality that $N = M$ for notational simplicity.

Let us recall also the asymptotic distributions under the null and alternative of the $\widehat{\text{MMD}}^2$ given by (Gretton *et al.*, 2012).

Theorem (Gretton *et al.*, 2012). Assume that K has finite second moments. Then, the following statements hold.

1. Under $\mathcal{H}_0$, $N\widehat{\text{MMD}}^2 \overset{d}{\rightarrow} \sum_{l=1}^{\infty} \lambda_l (z_l^2 - 2)$. z_l is a sequence of Gaussian random variables and λ_l are the eigenvalues solution to a certain eigenvalue problem.
2. Under $\mathcal{H}_1$, $N^{1/2}\left(\widehat{\text{MMD}}^2 - \text{MMD}^2\right) \overset{d}{\rightarrow} \mathcal{N}\left(0, \sigma^2_{\mathcal{H}_1}\right)$.

Please find the details of the eigenvalues and asymptotic variance in (Gretton *et al.*, 2012).

Now note that,

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

The first term relates to the asymptotic distribution of the RMMD under the null and the second term relates to the distribution of the RMMD under the alternative hypothesis.

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,

$$\begin{aligned}\left|N\widehat{\text{RMMD}}^2 - N\widehat{\text{MMD}}^2\right| \leq & \frac{1}{N}\sum_{i,j=1}^{N}\left|K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}) - K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j})\right| + \frac{1}{N}\sum_{i,j=1}^{N}\left|K(\mu_{\mathbb{Q}_i}, \mu_{\mathbb{Q}_j}) - K(\hat{\mu}_{\mathbb{Q}_i}, \hat{\mu}_{\mathbb{Q}_j})\right| \\ & - \frac{2}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{Q}_j}) - K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{Q}_j})\right|, \end{aligned} \tag{3}$$

and under the alternative hypothesis,

$$\begin{aligned}\left|\sqrt{N}\widehat{\text{RMMD}}^2 - \sqrt{N}\widehat{\text{MMD}}^2\right| \leq & \frac{1}{N\sqrt{N}}\sum_{i,j=1}^{N}\left|K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}) - K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j})\right| + \frac{1}{N\sqrt{N}}\sum_{i,j=1}^{N}\left|K(\mu_{\mathbb{Q}_i}, \mu_{\mathbb{Q}_j}) - K(\hat{\mu}_{\mathbb{Q}_i}, \hat{\mu}_{\mathbb{Q}_j})\right| \\ & - \frac{2}{N\sqrt{N}}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{Q}_j}) - K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{Q}_j})\right|. \end{aligned} \tag{4}$$

In both cases it suffices to show that inner products between population mean embeddings and empirical counterparts converge in probability at a rate fast enough such that a union bound over all terms in the summation scaled by $1/N$ and $1/(N\sqrt{N})$ converges to 0. We note here that we are considering two asymptotic regimes, once in the size of each set n_i that is relevant in the convergence of $K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j})$ to $K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j})$ and one in N which is the number of sets. Each may vary independently, and here we will assumed that the rate of growth of n_i is sufficient to ensure the weighted sums converge as $N \rightarrow \infty$.

A.1.2 Results

We will traverse the convergence of empirical kernels to their population counterparts in two steps, first using results that show the convergence of empirical mean embeddings to their population counterparts (Lemma 1) and second, using a Lipschitz condition to extend this to inner products between mean embeddings (Lemma 2).

For this we will assume K to be a real-valued, shift invariant ($K(x, x') = K(x - x', 0)$), and L_K-Lipschitz kernel,

$$|K(x, 0) - K(x', 0)| \leq L_K |x - x'|, \tag{5}$$

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 (Lopez-Paz *et al.*, 2015)) *Let the kernel K satisfy the assumptions above. Then we have,*

$$|\mu_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_i}|_{\mathcal{H}_K} \leq 2\sqrt{\frac{\mathbb{E}_{x\sim\mathbb{P}_i} K(x, x)}{n_i}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n_i}}, \tag{6}$$

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,*

$$|K(\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j}) - K(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j})| \leq L_K\left(4\sqrt{\frac{1}{\eta}} + 2\sqrt{\frac{2\log\frac{1}{\delta}}{\eta}}\right), \tag{7}$$

with probability at least $1-\delta$. As $\eta := \min(n_i,n_j) \to \infty$ we get that $K(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j})$ converges in probability to $K(\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j})$.

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

$$\begin{aligned}
|K(\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j}) - K(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j})| &= \left|K(\mu_{\mathbb{P}_i}-\mu_{\mathbb{P}_j},0) - K(\hat{\mu}_{\mathbb{P}_i}-\hat{\mu}_{\mathbb{P}_j},0)\right| && (8)\\
&\leq L_K\left|\mu_{\mathbb{P}_i}-\mu_{\mathbb{P}_j} - (\hat{\mu}_{\mathbb{P}_i}-\hat{\mu}_{\mathbb{P}_j})\right| && (9)\\
&\leq L_K\left|\mu_{\mathbb{P}_i}-\hat{\mu}_{\mathbb{P}_i}\right| + L_K\left|\mu_{\mathbb{P}_j}-\hat{\mu}_{\mathbb{P}_j}\right| && (10)\\
&\leq L_K\left(2\sqrt{\frac{\mathbb{E}_{x\sim\mathbb{P}_i}K(x,x)}{n_i}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n_i}} + 2\sqrt{\frac{\mathbb{E}_{x\sim\mathbb{P}_j}K(x,x)}{n_j}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n_j}}\right) && (11)\\
&\leq L_K\left(4\sqrt{\frac{1}{\eta}} + 2\sqrt{\frac{2\log\frac{1}{\delta}}{\eta}}\right), && (12)
\end{aligned}$$

where $\eta := \min(n_i,n_j)$ and we have use the boundedness condition on K, $\mathbb{E}_{x\sim\mathbb{P}_i}K(x,x) \leq 1$.

The for a rate of of increase of n_i fast enough in comparison to n, each term in equations (3) and (4) converges to zero which implies that the asymptotic distributions of $N\widehat{\text{RMMD}}^2$, $\sqrt{N}\widehat{\text{RMMD}}^2$ and $\sqrt{N}\widehat{\text{MMD}}^2$, $N\widehat{\text{MMD}}^2$, coincide respectively.

A.1.3 Extension to approximations 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,

$$\begin{aligned}
|k(\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j}) - &k(\hat{\mu}_{\mathbb{P}_i,m},\hat{\mu}_{\mathbb{P}_j,m})| \leq \\
&\left|k(\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j})\right| + \left|k(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_i,m},\hat{\mu}_{\mathbb{P}_j,m})\right|,
\end{aligned} \tag{13}$$

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 (Lopez-Paz *et al.*, 2015)) *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,*

$$|\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_i,m}|_{L^2(\mathbb{P})} \leq \frac{2}{\sqrt{m}}\left(1 + \sqrt{2\log n_i/\delta}\right), \tag{14}$$

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,*

$$\left|k(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_i,m},\hat{\mu}_{\mathbb{P}_j,m})\right| \leq \frac{2L_k}{\sqrt{m}}\left(2 + 2\sqrt{2\log(\eta/\delta)}\right). \tag{15}$$

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 $\eta := \min(n_i,n_j)$. 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,

$$\begin{aligned}\left|k(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_i,m},\hat{\mu}_{\mathbb{P}_j,m})\right| &= \left|k(\hat{\mu}_{\mathbb{P}_i}-\hat{\mu}_{\mathbb{P}_j},0) - k(\hat{\mu}_{\mathbb{P}_i,m}-\hat{\mu}_{\mathbb{P}_j,m},0)\right| && (16)\\ &\leq L_k\left|\hat{\mu}_{\mathbb{P}_i}-\hat{\mu}_{\mathbb{P}_j}-(\hat{\mu}_{\mathbb{P}_i,m}-\hat{\mu}_{\mathbb{P}_j,m})\right| && (17)\\ &\leq L_k\left|\hat{\mu}_{\mathbb{P}_i}-\hat{\mu}_{\mathbb{P}_i,m}\right| + L_k\left|\hat{\mu}_{\mathbb{P}_j}-\hat{\mu}_{\mathbb{P}_j,m}\right| && (18)\\ &\leq \frac{2L_k}{\sqrt{m}}\left(2+\sqrt{2\log(n_i/\delta)}+\sqrt{2\log(n_j/\delta)}\right) && (19)\\ &\leq \frac{2L_k}{\sqrt{m}}\left(2+2\sqrt{2\log(\eta/\delta)}\right), && (20)\end{aligned}$$

where we have written $\eta := \min(n_i, n_j)$ and the inequalities hold with probability at least $(1-\delta)$ over the randomness of the samples $(\omega_i, b_i)_{i=1}^m$.

A.2 ASYMPTOTIC DISTRIBUTION OF $\widehat{\text{RHSIC}}$

The asymptotic distribution of the RHSIC follows a very similar procedure since it can similarly bee decomposed in sums of kernels.

Proof. The $\widehat{\text{RHSIC}}$ may be written as a sum of V-statistics as follows (Gretton *et al.*, 2008),

$$\widehat{\text{RHSIC}} = \frac{1}{N^2}\sum_{i,j}^{N}\hat{K}_{ij}\hat{L}_{ij} + \frac{1}{N^4}\sum_{i,j,q,r}^{N}\hat{K}_{ij}\hat{L}_{qr} - \frac{2}{N^3}\sum_{i,j,q}^{N}\hat{K}_{ij}\hat{L}_{iq}, \tag{21}$$

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 indices refer to double sums of all pairs of numbers drawn with replacement from $\{1,...,N\}$, and similarly for three and four summation indices (Gretton *et al.*, 2008). 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,

$$|\hat{K}_{ij}\hat{L}_{qr} - K_{ij}L_{qr}|, \tag{22}$$

for any quadruple of indices 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 $\widehat{\text{HSIC}}$.

B APPROXIMATIONS FOR HIGH POWER

B.1 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 (Sutherland *et al.*, 2016).

Proposition 3 ***(Approximate power of*** $\widehat{\textbf{RMMD}}^2$***).*** *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 (Sutherland *et al.*, 2016). 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 $\widehat{\text{RHSIC}}$.

B.2 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 (Serfling, 2009). 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. (van der Vaart & Wellner, 1996)) 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 indices, 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.

C ADDITIONAL DETAILS ON EXPERIMENTS AND IMPLEMENTATION

C.1 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,

$$f(x;\mu) = \frac{x^{-\mu-1}}{\Gamma(\mu)} \exp\left(-1/x\right), \tag{23}$$

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

C.2 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.

C.3 GP2ST

The test developed by (Benavoli & Mangili, 2015) 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:

$$\Delta\mu^T \Delta\Sigma^{-1} \Delta\mu \leq \chi_v^2(1 - \alpha). \tag{24}$$

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

C.4 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 (Lopez-Paz *et al.*, 2013).

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

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.

C.5 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.

C.6 C2ST

We implemented the C2ST with tensorflow in python. We used a RNN with GRU cells in one version and the deepset architecture (with the author's implementation (Zaheer *et al.*, 2017)) in the other. The number of samples in each mini-batch is set to 64 the hidden layer size to 10. We optimize model parameters with Adam, learning rate equal to 0.01, and all variables are initialized with Xavier initialization. We use the elu activation functions for each layer and use sigmoid activation for the output layer given that we perform classification.

Both tests proceeds as follows (Lopez-Paz & Oquab, 2016):

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.

REFERENCES

Benavoli, Alessio, & Mangili, Francesca. 2015. Gaussian Processes for Bayesian hypothesis tests on regression functions. *Pages 74–82 of: Artificial intelligence and statistics*.

Gretton, Arthur, Fukumizu, Kenji, Teo, Choon H, Song, Le, Schölkopf, Bernhard, & Smola, Alex J. 2008. A kernel statistical test of independence. *Pages 585–592 of: Advances in neural information processing systems*.

Gretton, Arthur, Borgwardt, Karsten M, Rasch, Malte J, Schölkopf, Bernhard, & Smola, Alexander. 2012. A kernel two-sample test. *Journal of Machine Learning Research*, **13**(Mar), 723–773.

Lopez-Paz, David, & Oquab, Maxime. 2016. Revisiting classifier two-sample tests. *In: International Conference on Learning Representations*.

Lopez-Paz, David, Hennig, Philipp, & Schölkopf, Bernhard. 2013. The randomized dependence coefficient. *Pages 1–9 of: Advances in neural information processing systems*.

Lopez-Paz, David, Muandet, Krikamol, Schölkopf, Bernhard, & Tolstikhin, Iliya. 2015. Towards a learning theory of cause-effect inference. *Pages 1452–1461 of: International Conference on Machine Learning*.

Serfling, Robert J. 2009. *Approximation theorems of mathematical statistics*. Vol. 162. John Wiley & Sons.

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

van der Vaart, Aad W, & Wellner, Jon A. 1996. The delta-method. *Pages 372–400 of: Weak Convergence and Empirical Processes*. Springer.

Zaheer, Manzil, Kottur, Satwik, Ravanbakhsh, Siamak, Poczos, Barnabas, Salakhutdinov, Russ R, & Smola, Alexander J. 2017. Deep sets. *Pages 3391–3401 of: Advances in neural information processing systems*.