A Robust Two-Sample Test for Time Series data

Alexis Bellot, Mihaela van der Schaar
University of Cambridge and The Alan Turing Institute
Cambridge, United Kingdom
ab2583@cam.ac.uk, mschaar@turing.ac.uk

Abstract

We develop a general framework for hypothesis testing with time series data. The problem is to distinguish between the mean functions of the underlying temporal processes of populations of time series, which are often irregularly sampled and measured with error. Such an observation pattern can result in substantial uncertainty about the underlying trajectory, quantifying it accurately is important to ensure robust tests. We propose a new test statistic that views each trajectory as a sample from a distribution on functions and considers the distributions themselves to encode the uncertainty between observations. We derive asymptotic null distributions and power functions for our test and put emphasis on computational considerations by giving an efficient kernel learning framework to prevent overfitting in small samples and also showing how to scale our test to densely sampled time series. We conclude with performance evaluations on synthetic data and experiments on healthcare and climate change data.

1 Introduction

Two-sample tests capture the evidence in favour of two data samples being generated from the same underlying generating process. Such tests have been extensively studied for problems where data is naturally defined in a fixed dimensional space [15][1][9][8], yet many interesting questions fall beyond these assumptions. For instance, studies of disease progression, of economic and cognitive activity in the social sciences or in climate change, involve time series data which are irregularly sampled, measured with error, and with intervals between successive observation times often large (i.e. data is sparsely sampled). In this paper, we assume such observed samples are generated in an i.i.d manner from two random functions \( \{X(t) : t \in \mathbb{R}^+\} \) and \( \{Y(t) : t \in \mathbb{R}^+\} \), and the problem we consider is the two-sample test of means which we pose as the hypothesis testing problem,

\[
\mathcal{H}_0 : m_X(t) = m_Y(t) \quad \text{versus} \quad \mathcal{H}_1 : m_X(t) \neq m_Y(t) \quad \text{for all} \quad t \in T \subset \mathbb{R}^+ \quad (1)
\]

where \( m_X(t) := EX(t) \) and \( m_Y(t) := EY(t) \).

Almost all existing tests for the two-sample problem cannot be directly applied to such data [15][6][1][9][8]. While it might be tempting to approximate the mean function by interpolating between observations, evaluate it at equally spaced time points and resort to multivariate tests, we show that it is possible to construct tests that incorporate the uncertainty at and between observation times, and provide valid conclusions without assuming access to the true mean. This is important because, in the presence of nuisance variability in the data, ignoring the resulting uncertainty leads to tests rejecting \( \mathcal{H}_0 \) even when differences between the two processes are not due to differences in the underlying mean functions. This problem is exacerbated with sparsely sampled data where any interpolation is bound to mis-estimate the true curves.

We propose a test that represents each time series as a stochastic function in the form of a Gaussian process in a Hilbert space, and as a result, compares populations of distributions of functions for...
each trajectory, each one of them explicitly giving the likelihood of the underlying trajectory at each time point. A decision to reject or accept the null is then based on norms of the difference between mean processes in this Hilbert space. As different populations will have different correlation and smoothness properties, the choice of Hilbert space is important because it imposes prior structure of the resulting curves. Under or over-estimating the variability in the data leads to biased tests. The approach that we follow is to learn the Hilbert space in a data-driven manner with a two-stage approach that maximizes the marginal likelihood of the data at a population and individual level.

We turn our attention also to the case of densely-sampled time series, where working with Gaussian processes is known to be computationally challenging. We show how to ease computation while retaining model flexibility using the structured interpolation framework of Wilson and Nickisch [25], which allows us to make inference in large samples where uncertainty or noise is a concern.

Our contributions are three-fold:

- We propose a test for the two sample problem in [1], derive its null distribution and an explicit asymptotic power function as the number of samples increases.
- We propose approximations that ensure valid testing for a wide range of data generating scenarios and tractable evaluations for large time series.
- We analyze the performance of our test in a wide range of synthetic scenarios and illustrate the practical usefulness of our method on climate change data and healthcare data. The latter experiment is given in Appendix E.

Related work. Hypothesis testing in functional spaces, without considering irregular sampling or error, has received a lot of attention. We distinguish two main approaches. The first approach starts with a projection of the data onto a finite number of principal components and then considers high-dimensional multivariate tests (e.g. [3,13,15]). The second assumes access to curves in their original Hilbert space and base their tests on norms between functional samples (e.g. [28,19]). Studies that explicitly consider discrete sampling and error are rather scarce. Such a setting has only been considered in [10], that analyzes smoothing methods with regularly spaced observation times, and [20] that also considers projecting onto principal components and leverage multivariate tests. Both of these regard noise and sparsity on a population level, while we argue that uncertainty needs to be described at an individual level because irregularities can seldomly be assumed to be homogeneous. We are then able to describe the asymptotic distributions under the null and alternative of our uncertainty-aware representations, without finite dimensional restrictions, in contrast to [20].

2 Preliminaries

In this section we give a brief overview of the technical concepts we will be using to define our test.

Gaussian processes in functional spaces. Gaussian processes (GPs) are collections of random variables such that any finite number of them has a joint Gaussian distribution. Using a GP we can define a distribution over functions \( f \sim \mathcal{G}P(\mu, k) \), meaning that the collection of function evaluations at a vector \( t \) has a multivariate Gaussian distribution \( f(t) \sim \mathcal{N}(\mu(t), K(t, t)) \). To develop our testing procedure we will make use of gaussian processes \( f \) as random elements of the Hilbert space of square-integrable functions \( L^2(T) \) taking values in \( \mathbb{R} \), and indexed on a compact subset of the positive real line \( T \subset \mathbb{R}^+ \), such that \( L^2(T) \) is separable.

Covariance operators. Associated with the covariance kernel \( k \in L^2(T \times T) \), we define the covariance operator \( K : L^2(T) \rightarrow L^2(T) \), given by \( K(g)(t) = \int_T k(s, t)g(s)ds \). The operator \( K \) defines the sequence of orthonormal eigenfunctions \( \{\psi_i\}_{i=1}^\infty \), forming a basis in \( L^2(T) \), and such that its corresponding sequence of eigenvalues \( \{\lambda_i\}_{i=1}^\infty \) are non-negative, and decreasing in size with the index \( i \) satisfying \( \sum_{i=1}^\infty \lambda_i < \infty \). For \( K \) a trace-class operator in a separable Hilbert space we have additionally that \( \text{Tr}(K) = \sum_{i=1}^\infty \lambda_i < \infty \). In this case, any function \( f \) in \( L^2(T) \) can be expressed as a linear combination of \( \{\psi_i\}_{i=1}^\infty \). In particular, we will leverage the following representation of a centered Gaussian process \( f \) due to Karhunen and Loeve (see e.g. [24] for a recent review),

\[
f(t) = \sum_{i=1}^\infty \epsilon_i \psi_i(t)
\]  (2)
where \( \{e_i\}_{i=1}^{\infty} \) is an i.i.d sequence of Gaussian random variables with mean zero and variance \( \{\lambda_i\}_{i=1}^{\infty} \).

Our test involves mean processes, which can most easily be defined with respect to the usual squared distance in \( L^2(T) \), but this only accounts for linear variations in the amplitude of the data, ignoring other sources of variability such as phase variability [4]. A richer alternative to infer the centers of mass that preserve the structure of data in our context are barycenters in Wasserstein space. We overview this formalism below.

**Wasserstein distance for GPs.** The Wasserstein distance between probability measures can be understood as the minimal cost of mapping the unit mass of one probability measure into the unit mass of another probability measure, an optimal transport problem in its original formulation [23]. The \( L^2 \)-Wasserstein metric between Gaussian processes \( f_1 \) and \( f_2 \) [18] can be computed explicitly as,

\[
W^2_{L^2}(f_1, f_2) = \|m_1 - m_2\|^2 + \text{Tr}(K_1 + K_2 - 2(K_1^{1/2} K_2 K_1^{1/2})^{1/2})
\]

where \( \| \cdot \|_{L^2} \) denotes the norm in \( L^2 \).

**Barycenter GP.** Finding the barycenter of a population of \( n \) GPs \( f_i \) with respect to the Wasserstein metric in [4] amounts to solving the following minimization problem,

\[
\inf_{f \in L^2(T)} \sum_{i=1}^{n} \frac{1}{n} W^2_{L^2}(f_i, f)
\]

For a population of GPs \( \{f_i\}_{i=1}^{n} \) with \( f_i \sim \mathcal{GP}(m_i, K_i) \), [18] showed that there exists a unique barycenter \( \bar{f} \sim \mathcal{GP} (\bar{m}, \bar{K}) \) such that \( \bar{m} \) and \( \bar{K} \) satisfy,

\[
\bar{m} = \frac{1}{n} \sum_{i=1}^{n} m_i, \quad \bar{K} = \frac{1}{n} \sum_{i=1}^{n} (K_1^{1/2} K_2 K_1^{1/2})^{1/2}
\]

We can compute approximate solutions for mean kernel operator \( \bar{K} \) via an iterative fixed-point procedure [12], using finite dimensional restrictions for computations in practice. These approximations are well defined and consistent with infinite-dimensional processes [13].

### 3 Hypothesis test

Recall that each trajectory \( f_i \in L^2(T) \) is observed on a finite number of irregularly sampled discrete time points \( t_i = t_{i1}, ..., t_{in_i} \). Different time series are not necessarily observed at the same set of time points (i.e. \( t_i \neq t_j \) in general) nor the intervals between time points are assumed to be uniform. We assume each trajectory \( f_i \) to be generated from a Gaussian process observed with error,

\[
x_i = f_i(t_i) + \epsilon_i, \quad f_i(*) \sim \mathcal{GP}(m_i, k_i)
\]

where \( \epsilon_i \) are independent noise variables \( \mathcal{GP}(0, \sigma^2) \). The posterior distribution of the underlying process of each observed trajectory \( i \) is then also Gaussian with the following posterior mean function \( m_i \) and covariance function \( k_i \) (evaluated at any set of reference points \( \tau_1 \) and \( \tau_2 \)),

\[
m_i(\tau_1, \tau_2) := k_i(\tau_1, t_i)(k_i(t_i, t_i) + \sigma^2 I)^{-1} x_i
\]

\[
k_i(\tau_1, \tau_2) := k_i(\tau_1, \tau_2) - k_i(\tau_1, t_i)(k_i(t_i, t_i) + \sigma^2 I)^{-1} k_i(t_i, \tau_2)
\]

Here \( k_i(\cdot, \cdot) \) denotes a covariance matrix or a vector of pairwise similarities which should be clear from the dimensionality of the inputs.

As a result of using this representation, we compare populations of distributions over the time domain, each one of them explicitly giving the likelihood of the underlying trajectory. We propose to define a mean process \( \bar{f} \) using the Wasserstein distance. From a probabilistic perspective, this is equivalent to finding the measure that minimizes the "information distortion", a barycenter in this sense only removes the variability in each individual trajectory associated to the sample as a whole; thus preserving as much as possible the variability attributable to other individual factors [25]. In each one of the two populations \( X \) and \( Y \), we write

\[
\bar{f}_X \sim \mathcal{GP}(\bar{m}_X, \bar{K}_X) \quad \text{and} \quad \bar{f}_Y \sim \mathcal{GP}(\bar{m}_Y, \bar{K}_Y)
\]

for the population Barycenters, \( \bar{f}_X \) and \( \bar{f}_Y \) for their empirical counterparts, denote \( n_X \) and \( n_Y \) to be the number of time series observed in sample \( X \) and \( Y \) respectively, and let \( n := n_X + n_Y \).
3.1 Robust Test Statistic

In this subsection we propose a test for the hypothesis of equality of Wasserstein barycenters,

\[ \mathcal{H}_0 : \bar{m}_X = \bar{m}_Y, \]

Under this hypothesis, \( \sqrt{n}(\hat{f}_X - \hat{f}_Y) \) is itself a centered GP which can be written as an infinite expansion \( \sum_{i=1}^{\infty} e_i \psi_i(t) \) in terms of a common basis \( \{ \psi_i \} \) of \( X \) and \( Y \), where \( e_i \) are independently distributed centered Gaussian random variables with variance \( \lambda_i \), the corresponding eigenvalue for each \( i \). In this case the following holds,

\[
n \left\| \hat{f}_X - \hat{f}_Y \right\|_{L^2}^2 = \int_T \left( \sum_{j \geq 1} e_j \psi_j(t) \right) \left( \sum_{i \geq 1} e_i \psi_i(t) \right) dt = \sum_{i \geq 1} e_i^2 \quad (9) \]

since \( \{ \psi_i \}_{i>0} \) are pairwise orthonormal. A large value for this norm gives evidence against \( \mathcal{H}_0 \), and the sampling distribution of the above takes the form of an infinite mixture of uncorrelated \( \chi^2 \) random variables. \cite{27} proposed a method to approximate such a distribution by matching its first two moments with those of a scaled chi-square distribution \( \kappa \chi^2_n \) which in our case is given by,

\[
\kappa = \text{Tr} \left( K^* K \right) / \text{Tr} \left( K \right) \quad \text{and} \quad \eta = \text{Tr} \left( K \right)^2 / \text{Tr} \left( K^* K \right) \quad (10)
\]

where \( K^* \) is the adjoint of the common operator \( K \). In practice, simulating from the asymptotic null distribution can be challenging, since computing \( K \) requires a sample from \( \hat{f}_X - \hat{f}_Y \), which is not available. In the present formulation, as \( \text{Tr} \left( K \right) = \sum_{i=1}^{\infty} \lambda_i < \infty \), it follows that \( \lambda_i \to 0 \) as \( i \to \infty \). It is reasonable then to assume then that only a finite number \( q \) of eigenvalues meaningfully contribute to the variation in the GPs. Observe that in this set up, any finite-dimensional truncation of the above expansion optimally describes the variability in the function \( f \) as the \( e_i \) are uncorrelated and their variances \( \lambda_i \) decrease in size with their index. The approach we choose to estimate the null distribution considers the eigendecomposition of the pooled covariance operator \( \hat{K} := (n_X \hat{K}_X + n_Y \hat{K}_Y) / n \) \cite{7} which can be estimated consistently from data. We are now able to define our test statistic \( \text{RMD} : X \times Y \to \mathbb{R}^+ \), which we call the robust mean discrepancy (RMD), as,

\[
\text{RMD}_n(X, Y; q) := n \left( \sum_{i=1}^{q} \hat{\lambda}_i \right) \left( \sum_{i=1}^{q} \hat{\lambda}_i^2 \right)^{-1} \left\| \hat{f}_X - \hat{f}_Y \right\|_{L^2}^2 \quad (11)
\]

where \( \hat{\lambda}_i \) is the \( i \)th eigenvalue of the covariance operator \( \hat{K} \). The \( p \)-value is defined as the probability of observing a larger test statistic value under the null distribution \( \chi^2_\eta \):

\[
p = P(\chi^2_\eta > \text{RMD}_n) \quad (12)
\]

Rejecting the null if \( p < \alpha \) then defines an asymptotically valid level \( \alpha \) test.

3.2 Asymptotic power function

The power function \( \beta_n = P(p < \alpha | \mathcal{H}_1) \) determines how increased total sample size \( n \) translates to higher power (the proportion of correctly rejected hypotheses \( \mathcal{H}_0 \)) as sampling error is reduced. The faster the power function converges to 1 the less data is needed to correctly distinguish between \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \).

When comparing the difference between fixed hypotheses \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \), one can see easily that \( \beta_n \) tends to 1 asymptotically, and therefore it is typically of interest also to assess power against a sequence of local alternatives that vary as a function of the number of total samples \( n \). The idea behind this regime is that we impose the extra condition that \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \) become harder to distinguish as \( n \) grows. In our case, this means that the difference in trends between two samples becomes smaller. Suppose that,

\[
\mathcal{H}_{1n} : m_X(t) - m_Y(t) := n^{-\tau/2} \xi(t) \quad (13)
\]

with the assumption that \( 0 < \tau \leq 1, \int_T \xi(t)^2 dt < \infty \).
Theorem 1 Under the sequence of alternative hypotheses $\mathcal{H}_{1n}$ in (20), the asymptotic power function of RMD$_n$ at level $\alpha$, $\beta_n$, satisfies,

$$\beta_n = \Phi \left( \frac{n^{1/2} \mu^2 - \frac{c_\alpha}{2n(1-\tau)^2/2} \mu_\lambda}{2\mu_\lambda} \right) \to 0$$ (14)

Here $c_\alpha$ is the $\alpha$ quantile of RMD$_n$’s distribution under the null, $\mu_n := \int_T \psi_X(t) \xi(t) dt$, $\mu_n := \sum_{i=1}^\infty \mu_{r_i}^2$, $\lambda_\mu := \sum_{i=1}^\infty \lambda_i^2$, $\Phi$ is the cdf of a standard Gaussian random variable and $m$ is the number of positive eigenvalues.

Observe that as $n \to \infty$, $\beta_n$ tends to 1. The asymptotic power function indicates that the RMD$_n$ is able to detect departures from the null hypothesis of size greater than $n^{-1/2}$ in the direction of any square integrable function. This property is called root $n$-consistency and can be interpreted as the sampling error for recognizing alternatives of this form decreases at a rate of $n^{-1/2}$. Theorem 2 guarantees that the replacement in the computation of the asymptotic null distribution still yields a consistent test.

Theorem 2 Let $\hat{K} := (n_X \hat{K}_X + n_Y \hat{K}_Y)/n$, obtained from observed time series samples $X$ and $Y$. Suppose the test threshold is set to the upper $\alpha$ quantile of the distribution of the RMD under $\mathcal{H}_0$. Then, under $\mathcal{H}_0$, asymptotically the false positive rate of RMD is $\alpha$. Under $\mathcal{H}_{1n}$, the test power of RMD satisfies Theorem 1.

The proof is based on the consistency on $\hat{K}$ for the barycenter covariance operator and follows from the asymptotic results of Theorem 1.

Recall that we initially designed the mean process $\bar{f}$ to preserve the variation structure of the original distributional representation of each trajectory. It is natural to expect then, that the sparsity of available data, which we quantify as the trace of the barycenter covariance operators $\text{Tr} (\bar{K}) = \int_T \bar{k}(t, t) dt = \sum_{i=1}^\infty \lambda_i$, directly influences the power of our test. The following corollary describes this result.

Corollary 1 Let $\beta_{n,V}$ and $\beta_{n,W}$ be the power functions for the sequence of alternatives in (20) induced by populations of sparsely and densely sampled data respectively, that is with corresponding eigenvalues $\{\lambda_i^V\}$ and $\{\lambda_i^W\}$, such that $\sum_{i=1}^\infty \lambda_i^W < \sum_{i=1}^\infty \lambda_i^V$. Then, other sources of variation being equal,

$$\beta_{n,W} - \beta_{n,V} > 0$$ (15)

as $n \to \infty$.

More densely sampled time series results in higher power as we are able to remove one source of uncertainty from the testing procedure.

4 Approximations for valid testing and scalability

The approximations involved in the design of our test relate to the level of smoothness - via the choice of covariance kernel - and variability assumed in the observed trajectories - via the truncation level $q$ that determines the number of eigenfunctions that are retained in the GP expansion. It is natural to expect the performance of our test to depend upon these specifications, especially since different populations with potentially different correlation and smoothness properties are to be compared.

Prior covariance kernel. Prior covariance kernels impose structure on the space of resulting trajectories. We optimize kernel hyper-parameters $\theta_i$ and variance $\sigma^2_i$, for each trajectory, such that the marginal likelihood of the data under model (6) is maximized,

$$\text{argmax}_{\theta_i, \sigma^2} \left( -\frac{1}{2} x_i^T (k_0(t_i, t_i) + \sigma^2 I)^{-1} x_i - \frac{1}{2} \log |k_0(t_i, t_i) + \sigma^2 I| - \frac{n}{2} \log 2 \pi \right)$$ (16)

Choosing a kernel function $k$ and resulting function space that is too restrictive or too flexible will lead to an underestimation or overestimation of the variation and potentially biased hypothesis tests. For a sufficient number of samples, we propose a two stage approach. First we estimate the smoothness properties at a population level by choosing an appropriate family of kernels in each sample separately that optimizes the marginal likelihood of the sample as a whole evaluated out-of-sample. Then, in
a second stage, hyperparameters within this family are optimized on an individual level to capture differences in sparsity and noise. We give full details in Appendix C.

**Truncation parameter** $q$. Each principal component explains variation in an orthogonal direction, and thus the truncation of the Karhunen-Loeve expansion effectively induces smoothing upon the curves. This choice can have a significant effect on the resulting power of the test as higher order variations in the curves of two samples will only be apparent if $q$ is set large enough. However, for greedy choices of $q$ the accumulation of variance from a large number of eigenfunctions will incorporate exogenous noise not due to the underlying trends and lead to inflated type I error. The idea is to make the choice of smoothness $q$ in a given scenario data dependent, by trading off fidelity of the projected data into $q$ eigenfunctions and roughness. Let $\tilde{m}_i$ denote the true curve for a time series $t$ and $\hat{m}_i$ its truncated approximation, the objective is,

$$
\sum_{i=1}^{n_x} \left\| \tilde{m}_i - m_i \right\|_{L^2}^2 + \sum_{i=1}^{n_x} \left\| \hat{m}_i \right\|_{H_X}^2 = \sum_{i=1}^{n_x} \left\| \sum_{j=1}^{q} (m_i, \psi_j) \psi_j - m_i \right\|_{L^2}^2 + \sum_{i=1}^{n_x} \sum_{j=1}^{q} (m_i, \psi_j)^2
$$

(17)

where $\{\psi_i\}_i$ are the eigenfunctions of the operator $K_X$. This objective penalizes the mean-squared error in $L^2$ norm with an estimate of the roughness of the curves $\hat{m}_i$ as their norm in the Hilbert Space $H_X$ induced by the covariance operator $K_X$. An analogous objective is used for curves estimated in the $Y$ sample, and $q \in \mathbb{N}$ is chosen such as to minimize the weighted sum of the objectives for the $X$ and $Y$ samples.

**Remark.** Both procedures above are intended to better understand the complexity and smoothness of the trajectories themselves and not the complexity of the difference between the two samples. In this sense, these procedures are not related to the posterior hypothesis testing problem and do not encourage or discourage a specific test decision a priori.

### 4.1 Approximations for densely sampled Time series

Although we motivate our test with sparsely observed time series, in this section we show how to scale our tests to more densely sampled data. In this case, describing uncertainty on an individual level is still necessary because it is unlikely that all trajectories will be observed on a common set of observation times [14]. In practice we approximate trajectories on a set of reference points $\tau$ such that computing exact posterior mean and covariance matrices takes $O(n^3 + nd)$ and $O(n^3 + n^2d + nd^2)$ respectively, where $n$ is the number of observations in each trajectory and $d = |\tau|$ is a number of reference points used to approximate the posterior function for each trajectory.

The idea is to approximate the cross-covariance matrix for the kernel evaluated at the training times and inducing inputs $\Sigma_{t,u}$ by interpolating on the covariance matrix of inducing points $\Sigma_{u,u}$, such that $\Sigma_{t,u} \approx W_t \Sigma_{u,u} W_t^T$ as outlined in [25] and [14]. The key is that $W$ can be extremely sparse, only 4 non-zero entries in each row if cubic interpolation is used. The high-dimensional kernel matrix $\Sigma_{t,t} \in \mathbb{R}^{d \times d}$ can then be approximated by a kernel $\hat{\Sigma}$ defined in terms of a sparse decomposition $\hat{\Sigma} = W \Sigma_{u,u} W^T$, where $\Sigma_{u,u}$ is a collection of $m$ evenly spaced inducing points, and the posterior mean and covariance evaluated at $u$ is then given by,

$$
\hat{m}_u \approx W_x \Sigma_{u,u} W_t^T (W_t \Sigma_{u,u} W_t^T + \sigma^2 I)^{-1} x
$$

(18)

$$
\hat{\Sigma}_{u,u} \approx W_x \Sigma_{u,u} W_x^T - W_x \Sigma_{u,u} W_t^T (W_t \Sigma_{u,u} W_t^T + \sigma^2 I)^{-1} W_t \Sigma_{u,u} W_x^T
$$

(19)

The number of inducing points does not need to be small in comparison to the number of training times to improve computational speed which allows us to capture the structure of large data sets while retaining efficiency gains. If in addition we exploit Toeplitz structure in $\Sigma_{u,u}$ to efficiently solve linear systems involving its inverse, we achieve an approximation of $\mu$ and $\Sigma$ that scales in time complexity as $O(n + d + m \log m)$ [25] [14].

### 5 Experiments

The primary focus of our results will be on the comparison of **type I error**: the error made by rejecting $H_0$ when it is in fact true, and **power**: the rate at which we correctly reject $H_0$ when it is false. In all simulations we compute estimates of the $p$-value, which corresponds to type I error when
Figure 1: Type I error and power results for the synthetic simulations. RMD and RMD-A are the tests proposed in this paper without and with kernel approximations respectively. $T^2$ is Hotelling’s test and C2ST is the classifier two-sample test.

$H_0$ is true, and corresponds to test power when $H_1$ is true. We set $\alpha = 0.05$ and repeat each problem 1000 times.

For comparison, we consider ignoring the uncertainty about the underlying trajectories by interpolating between observations via kernel smoothing and evaluating at equal intervals such that each instance is summarized by a multidimensional vector, directly comparable between examples. The advantage of such an approach is that we can use any vector-valued two sample test for difference in means - in particular in this paper we will use the Hotelling $T^2$ test [15][11], which was shown to be uniformly most powerful for Gaussian data. In addition, an approach that can be adapted to our setting involves fitting a classifier that takes a trajectory as an input and attempts to predict which of the two samples it belongs to, the C2ST [16]. The accuracy on a held-out testing set serves as evidence against the null hypothesis of equal data generating process. Of course, building a classifier for irregularly sampled data is a research project in itself. In this analysis we built a simple recurrent neural network with GRU cells that takes the observation time as well as the observation value as an input in each step (see Appendix D for details). Note also that the hypothesis being tested with the C2ST is inherently stronger: equality in any moment of the distribution in each sample.

5.1 Synthetic data

We construct two samples $\{X_i\}_{i=1}^{n/2}$ and $\{Y_j\}_{j=1}^{n/2}$ by evaluating the stochastic processes at time indices uniformly sampled in the interval $[0, 1]$. Unless set otherwise, we consider 10 measurements in each trajectory and $n = 100$. Under $H_0$ and $H_1$ we use the following data generation mechanisms:

$H_0 : x \sim \epsilon_x, \quad y \sim \epsilon_y, \quad H_1 : x \sim \epsilon_x, \quad y \sim m + \epsilon_y$

where $m$ is a non-random mean function and $\epsilon_x$ and $\epsilon_y$ are independent noise processes $GP(0, 0.05I)$, unless set otherwise. In each experiment we evaluate performance with respect to two alternative mean functions: (1) $m(t) = 0.1 \sin(2\pi t)$ and (2) $m(t) = \exp(-(t - 0.5)^2/0.005)$; in the first case the mean function differs from 0 across most of the time domain in a non-linear manner while in the second case only around $t = 0.5$ the mean behaviour changes which we expect to be more difficult to detect, especially for sparsely sampled time series. Experiments for setting (2) are given in Appendix B with similar conclusions to setting (1).

**Test power and type I error vs. sparsity.** We begin by analyzing the rate of change of test power and type I error as we vary the number of observations in each trajectory, from very sparse to dense.

---

1We also experimented using point processes and different sample sizes but these did not result in different conclusions in our performance evaluations.
settings. The results are given in Figure 1. Observe that the RMD tests are the only ones successfully controlling type I error in both sparse and dense settings while the C2ST fails in sparse settings due to over-fitting the small number of observations, and $T^2$ fails in dense settings where the number of observations is larger than the number of samples. Similar conclusions can be made with regards to power.

**Test power and type I error vs. noise.** Our second experimental setting analyses the robustness of tests to increasing but differing noise variance $\sigma^2$ between the two samples. For this we draw independently in every run $\epsilon_x$ and $\epsilon_y$ from a Laplace distribution with variance $2\sigma^2$ and variance $\sigma^2$ respectively. As the lower panels of Figure 1 show, C2ST and $T^2$ underperform, unable to control type I error as noise increases, for two different reasons. The C2ST rejects $H_0$ incorrectly because it recognises differences in the second moment of their distributions while $T^2$ implicitly assumes Gaussian noise with equal variance which does not hold in this case. In turn, RMD tests do control Type I error well for the noise levels considered. RMD and the C2ST have comparably high power in this scenario.

We demonstrate the rate of increase of test power with the number $n$ of observed trajectories in Appendix B. We see that all tests have power converging to 1, but their efficiency differs: $T^2$ is most efficient followed closely by RMD while the C2ST has lower power because of the data requirements on neural network optimization and importantly also because it requires separate training and testing populations. In Appendix B we also provide running times for all methods. RMD-A is approximately twice as fast as RMD.

![Figure 2: Comparisons of temperature dynamics in Peru.](image)

5.2 Climate Data

Climate change is a major development whose consequences will transform every aspect of our life as we know it today. In this experiment we test for changes in temperatures across time in selected countries. The data was compiled and made available by Berkeley Earth\(^2\). It contains average monthly surface temperatures across the globe from 1740 to 2013 categorized by countries. We constructed yearly trajectories by concatenating 12 monthly measurements. We begin by making a comparison with C2ST using climate data from Peru. The left panel of Figure 2 shows the average temperature across the 12 monthly measurements in each year. Notice that the variability increases significantly over time, but the mean trend increase is fairly modest. We consider sets of randomly sampled yearly trajectories posterior to 1913 with increasing sample size, paired with data from the reference period 1893–1913. This allows us to create a collection of two-sample problems. We observe on the right panel of Figure 2 that while the C2ST clearly finds difference between the trajectories as we increase the number of samples, the evidence for the mean functions differing is much more modest. This suggests that for data with complex sources of variation, a non-parametric test like C2ST will likely reject the null given sufficient sample size, even though differences are not due to the mean functions. With the RMD we can specifically test for differences in mean trends while controlling for other sources of variations.

Table 1: RMD $p$-values for differences with reference period 1853–1873.

|                | 1873–′93 | 1893–′13 | 1913–′34 | 1933–′53 | 1953–′73 | 1973–′93 | 1993–′13 |
|----------------|----------|----------|----------|----------|----------|----------|----------|
| U.S.           | 0.18     | 0.33     | 0.29     | 0.068    | 0.032    | 0.0007   | <0.0001  |
| U.K.           | 0.22     | 0.14     | 0.76     | 0.14     | 0.16     | 0.13     | 0.12     |
| Vietnam        | 0.025    | 0.001    | 0.0001   | <0.0001  | <0.0001  | <0.0001  | <0.0001  |

\(^2\)http://berkeleyearth.org/data/
Next we evaluate the evidence for climate change in the U.S., U.K. and Vietnam. We choose 1853 – 1873 as a reference period and compared subsequent periods of 20 years for difference in mean yearly temperatures; our results (adjusted for multiple testing) can be found in Table I. Here we observe p-values consistently below 0.05 from 1953 suggesting that there are significant differences in mean temperatures with respect to the baseline temperatures of 1853 – 1873 in each country.

6 Conclusions

Time series collected for most real-world applications contain measurement noise and irregularly spaced observation times. We have proposed a test statistic for the problem of testing equality of mean functions that generalizes to arbitrarily observed time series by encoding an estimate of the uncertainty about the underlying true curve directly in our test. We have demonstrated theoretically and empirically that our test controls type I error and has good power even in challenging scenarios with noisy sparsely sampled data and proposed various approximations to extend our test to densely sampled data and a large number of trajectories.
Appendix A: Theoretical results

In this section we prove Theorem 1 and Corollary 1 given in section 3.2 of the main body of this paper. Recall that we consider the sequence of alternatives:

\[ H_{1n} : m_X(t) - m_Y(t) := n^{-\tau/2}\xi(t) \]  

(20)

with the assumptions that \(0 < \tau \leq 1\) and \(\int_T \xi(t)^2 dt < \infty\). We restate the theorem and corollary for convenience.

**Theorem 1** Under the sequence of alternative hypotheses \(H_{1n}\) in (20), the asymptotic power function of \(\text{RMD}_n\) at level \(\alpha, \beta_n\) satisfies,

\[ \beta_n = \Phi\left(\frac{n^{1-\tau} \mu^2}{2\mu_\lambda} - \frac{c_{\alpha}}{2\mu(1-\tau/2)\mu_\lambda}\right) \to 0 \]

(21)

Here \(c_{\alpha}\) is the \(\alpha\) quantile of the \(\text{RMD}_n\)'s distribution under the null, \(\mu_r := \int_T \psi_r(t)\xi(t) dt, \mu := \sum_{r=1}^{\infty} \mu_r^2, \mu_\lambda := \sum_{r=1}^{m} \lambda_r \mu_r^2\). \(\Phi\) is the cdf of a standard Gaussian random variable and \(m\) is the number of positive eigenvalues.

**Proof.** Let \(\\{\lambda_i\}\) be the eigenvalues of the operator \(n_X K_X/n + n_Y K_Y/n\). Then there exists an \(m\) such that \(\lambda_r > 0\) for \(r \leq m\) and \(\lambda_r = 0\) for \(r > m\). Under the alternative it then holds that,

\[ n ||f_X - f_Y||_{L^2}^2 = \sum_{r=1}^{m} \lambda_r e_r^2 + n^{1-\tau} \sum_{r=m+1}^{\infty} \mu_r^2 \]

(22)

where \(n^{-\tau} \mu_r^2 \to ||f_X - f_Y||_{L^2}^2\) is the squared norm of the representation of \(\xi\) in the basis \(\\{\psi_i\}\). This follows because for \(r > m\), \(e_r\) is deterministic. For \(r \leq m\), \(e_r^2\) is \(\chi^2\) with non-centrality parameter \(n^{1-\tau} \lambda_r^{-1} \mu_r^2\), which follows from the fact that the \(e_r\) are uncorrelated. Notice that we can write,

\[ e_r^2 = d (z_r + n^{1-\tau} \lambda_r^{-1/2} \mu_r)^2 \]

(23)

\[ = d z_r^2 + 2 n^{1-\tau} \sum_{r=1}^{m} z_r \lambda_r^{1/2} \mu_r + n^{1-\tau} \sum_{r=1}^{m} \mu_r^2 \]

(24)

where \(z_r\) is a standard Gaussian random variable and \(d\) means equality in distribution. Then,

\[ \sum_{r=1}^{m} \lambda_r e_r^2 = d \sum_{r=1}^{m} z_r^2 + 2n^{1-\tau} \sum_{r=1}^{m} z_r \lambda_r^{1/2} \mu_r + n^{1-\tau} \sum_{r=1}^{m} \mu_r^2 \]

(25)

As \(n \to \infty\), the last two terms in the above equation dominate the first one. Therefore we have asymptotically in \(n\),

\[ n ||f_X - f_Y||_{L^2}^2 \sim \mathcal{N}(n^{1-\tau} \mu^2, 4n^{1-\tau} \sum_{r=1}^{m} \lambda_r \mu_r^2) \]

(26)

where we write \(\mu = \sum_{r=1}^{\infty} \mu_r^2\). Then also approximately the \(\text{RMD}_n\) is distributed as (7) up to a constant finite factor \((\sum_i^{q} \lambda_i) (\sum_i^{q} \lambda_i^2)^{-1}\) and consistent estimation of \(f_X\) and \(f_Y\). Thus the power of \(\text{RMD}_n\) is given by,

\[ P(\text{RMD}_n > c_{\alpha}|H_{1n}) = 1 - \Phi\left(\frac{c_{\alpha} - n^{1-\tau} \mu^2}{2n^{1-\tau} \sum_{r=1}^{m} \lambda_r \mu_r^2}\right) \to 1 \]

(27)

as \(n \to \infty\) where \(\Phi\) is the cdf of a standard Gaussian random variable and \(c_{\alpha}\) is the \(\alpha\) quantile of the distribution of \(\text{RMD}_n\) under \(H_0\).

**Corollary 1** Let \(\beta_{n,V}\) and \(\beta_{n,W}\) be the power functions for the sequence of alternatives in (20) induced by populations of sparsely and densely sampled data, that is with corresponding eigenvalues \(\\{\lambda_i^0\}\) and \(\\{\lambda_i\}\), respectively, such that \(\sum_i^{\infty} \lambda_i < \sum_i^{\infty} \lambda_i^0\). Then, other sources of variation being equal,

\[ \beta_{n,W} - \beta_{n,V} > 0 \]

(28)
as } n \to \infty.

**Proof.** Notice that the eigenvalues appear only in the denominator of the power expression in (8). Therefore if we assume that the direction of variation } \{ \psi_i \} \text{ remain constant for processes of populations } V \text{ and } W, \text{ then}

\[
\sum_{i=1}^{\infty} \lambda_i^W < \sum_{i=1}^{\infty} \lambda_i^V \Rightarrow \sum_{i=1}^{\infty} \lambda_i^W \mu_i^2 < \sum_{i=1}^{\infty} \lambda_i^V \mu_i^2
\]

(29)

since } \mu_i^2 \geq 0, \text{ and therefore } \beta_{n,W} > \beta_{n,V}.
Appendix B: Additional synthetic simulations

In this section we demonstrate the rate on increase of test power with the number of trajectories \( n \) for setting (1), the empirical running times as a function of the number of trajectories and provide complete performance results for scenario (2), where \( m(t) = \exp(-(t - 0.5)^2/0.005) \).

**Test power and type I error vs number of trajectories \( n \)**

The performance results for all methods as a function of the number of trajectories \( n \) for setting (1) can be found in Figure 3. We observe that all tests have power converging to 1, but their efficiency differs: \( T^2 \) is most efficient followed closely by RMD while the C2ST has lower power because of the data requirements on neural network optimization and importantly also because it requires separate training and testing populations.

Figure 3: Power and type I error as a function of sample size for setting (1).

**Running times vs number of trajectories \( n \) and sparsity**

Figure 4 provides running times in seconds as a function of the number of trajectories. Notice the high computational cost of both RMD implementations (with and without approximations) because each additional trajectory requires fitting a new GP. However as the number of measurement increases RMD-A is much more efficient than RMD, with almost half the running time required to run a single two sample test.

Figure 4: Running time in seconds of a single hypothesis test:

**Performance results mean function (2)**

The performance results for all methods and all experiments with mean function (2) can be found in Figure 5.
Figure 5: Complete performance results for mean function (2).
Appendix C: Further details on prior covariance kernel optimization

Noisy datasets motivate the need for a kernel estimation strategy that is flexible enough such that the sources of variation in the mean trends can be captured accurately but also restrictive enough such that we do not overfit to the spurious noise and artificially inflate type I error. We propose to estimate the unknown data generating kernel $k$ in a two-stage approach.

**Stage 1** - The smoothness of the trajectories in a given population can reasonably be considered to be a property of the sample as a whole. From this observation, to be able to accurately model the smoothness in a given population we choose an appropriate kernel family $k_j$ from a library of kernels $\{k_j\}_j$ (specifically all kernels available in the sklearn library in python). Kernel hyperparameters are optimized for each member of the library by maximizing the marginal likelihood, and a choice is made by evaluating the fit with the marginal likelihood of data from a held-out set.

**Stage 2** - While smoothness of trajectories can be well approximated from the sample as a whole, we would expect noise and the uncertainty due to sparsity to be individual specific. In order to capture those features, we learn hyperparameters of our chosen kernel family and noise parameter optimized for each individual by again maximizing the marginal likelihood of each trajectory.

Appendix D: Implementation details

**RMD**

The RMD and all experiments were implemented in python. The pseudocode for the RMD is given in Algorithm 1 below.

**Algorithm 1: RMD**

1. Optimize kernel hyperparameters $\theta$ with two-stage approach.
2. Obtain posterior mean and variance over a set of reference points in each trajectory.
3. Optimize truncation level $q$ following criterion (17).
4. Compute RMD using the barycenters of each sample and their truncated KL expansions.
5. IF $\text{RMD} \geq c_\alpha$ reject $H_0$; otherwise accept $H_0$.

**C2ST**

We implemented the C2ST with tensorflow in python. We used a RNN with GRU cells. 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. The algorithm proceeds as follows [16]:

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 $\mathcal{D}_{te}$: $\hat{t} := \frac{1}{n_{te}} \sum_{(z_i, l_i) \in \mathcal{D}_{te}} 1\{g(z_i) > 1/2\} = l_t$.
5. IF $\hat{t}$ is greater that the $\alpha$ quantile of a $N(1/2, 1/(4n_{te}))$ reject $H_0$; otherwise accept $H_0$.

1 is the indicator function.
Appendix E: Cystic Fibrosis data

In this experiment we test for differences in the lung function of patient subgroups from the UK Cystic Fibrosis (CF) registry. In this case, as well as in most hospital systems, it remains rare to capture dense physiological data streams, and much more common for data streams to be both sparse and irregularly sampled. For people with CF, mucus in the lungs is linked with chronic infections that can cause permanent damage [12], making it harder to breathe. This condition is often measured 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 [22]. For example, a person with CF who has FEV1% predicted of 50% can breathe out half the amount of air in the first second of an exhaled breath as we would expect from a comparable person without cystic fibrosis [22]. As early as 1992, [12] demonstrated that patients with a FEV1% predicted below 30% had a 2-year mortality rate over 50%. The CF data contains 10,980 patients suffering from Cystic Fibrosis with approximately annual follow ups between 2008 and 2015.

![Figure 6: Three patient trajectories plotted together with the Barycenter mean GP in light gray and a naive mean, taken by averaging the standard deviations of interpolations of the three trajectories at each time, in dark gray.](image)

In Figure 6, we show the qualitative differences between the Wasserstein barycenter and the naive mean, taken by averaging the standard deviations of interpolations of the three trajectories at each time, for three selected trajectories. Observe that the variance is particularly underestimated in the [500, 700] interval where there is uncertainty because few measurements are observed.

Next compare subsets of the CF data partitioned based on the available binary features. Cystic Fibrosis has received a lot of attention from the medical literature and we can compare our findings with published medical results. The most significant variables for differing FEV1% predicted are: CFRD, Diabetes, Depression, Aspergillus, ALCA, Staph. Aureus, Osteopenia, Liver disease and Osteoporosis. CFRD, Diabetes, Depression and Liver disease are well known comorbidities for CF [21], and especially relevant are those related to the bones: Osteopenia and Osteoporosis [5]. Aspergillus, ALCA and Staph. Aureus are infections of the lungs which can precipitate the decline of lung function and thus directly affect the trajectory of FEV1% predicted. These infections are prevalent in CF patients because their thickened mucus in the lungs prevents bacteria from being cleared. They are further described in [17].
References

[1] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] Pedro C Álvarez-Esteban, E del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[3] Michal Benko, Wolfgang Härdle, Alois Kneip, et al. Common functional principal components. *The Annals of Statistics*, 37(1):1–34, 2009.

[4] Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Minimax convergence rate for estimating the wasserstein barycenter of random measures on the real line. *arXiv preprint arXiv:1606.03933*, 2016.

[5] SP Conway, AM Morton, B Oldroyd, JG Truscott, H White, AH Smith, and I Haigh. Osteoporosis and osteopenia in adults and adolescents with cystic fibrosis: prevalence and associated factors. *Thorax*, 55(9):798–804, 2000.

[6] 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.

[7] Jacques Dauxois, Alain Pousse, and Yves Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1):136–154, 1982.

[8] 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.

[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] Peter Hall and Ingrid Van Keilegom. Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica*, pages 1511–1531, 2007.

[11] Harold Hotelling. The generalization of student’s ratio. In *Breakthroughs in statistics*, pages 54–65. Springer, 1992.

[12] 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.

[13] Jun Li, Yumou Qiu, and Lingjun Li. A neighborhood-assisted hotelling’s t² test for high-dimensional means. *arXiv preprint arXiv:1712.01798*, 2017.

[14] Steven Cheng-Xian Li and Benjamin M Marlin. A scalable end-to-end gaussian process adapter for irregularly sampled time series classification. In *Advances in neural information processing systems*, pages 1804–1812, 2016.

[15] Miles Lopes, Laurent Jacob, and Martin J Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, pages 1206–1214, 2011.

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

[17] Jeffrey B Lyczak, Carolyn L Cannon, and Gerald B Pier. Lung infections associated with cystic fibrosis. *Clinical microbiology reviews*, 15(2):194–222, 2002.

[18] Anton Mallasto and Aasa Feragen. Learning from uncertain curves: The 2-wasserstein metric for gaussian processes. In *Advances in Neural Information Processing Systems*, pages 5660–5670, 2017.

[19] 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.
[20] Gina Maria Pomann, Ana-Maria Staicu, and Sujit K Ghosh. Two sample hypothesis testing for functional data. Technical report, North Carolina State University. Dept. of Statistics, 2013.

[21] Nicola J Ronan, Joseph Stuart Elborn, and Barry J Plant. Current and emerging comorbidities in cystic fibrosis. *La Presse Médicale*, 46(6):e125–e138, 2017.

[22] 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.

[23] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[24] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Review of functional data analysis. *arXiv preprint arXiv:1507.05135*, 2015.

[25] Andrew Wilson and Hannes Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *International Conference on Machine Learning*, pages 1775–1784, 2015.

[26] Hongkang Yang and Esteban G Tabak. Clustering through the optimal transport barycenter problem. *arXiv preprint arXiv:1902.10288*, 2019.

[27] Jin-Ting Zhang. Approximate and asymptotic distributions of chi-squared–type mixtures with applications. *Journal of the American Statistical Association*, 100(469):273–285, 2005.

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