Journal of Machine Learning Research 26 (2025) 1-47 Submitted 7/24; Revised 10/25; Published 11/25

Infinite-dimensional Mahalanobis Distance with Applications to Kernelized Novelty Detection

Nikita Zozoulenko N.ZOZOULENKO23@IMPERIAL.AC.UK
Department of Mathematics
Imperial College London
UK

Thomas Cass THOMAS.CASS@IMPERIAL.AC.UK
Department of Mathematics
Imperial College London
UK
Institute for Advanced Study
Princeton, USA

Lukas Gonon LUKAS.GONON@UNISG.CH
School of Computer Science
University of St. Gallen
Switzerland
Department of Mathematics
Imperial College London
UK

Editor: Aryeh Kontorovich

Abstract

The Mahalanobis distance is a classical tool used to measure the covariance-adjusted distance between points in $\mathbb{R}^d$. In this work, we extend the concept of Mahalanobis distance to separable Banach spaces by reinterpreting it as a Cameron-Martin norm associated with a probability measure. This approach leads to a basis-free, data-driven notion of anomaly distance through the so-called variance norm, which can naturally be estimated using empirical measures of a sample. Our framework generalizes the classical $\mathbb{R}^d$, functional $(L^2[0,1])^d$, and kernelized settings; importantly, it incorporates non-injective covariance operators. We prove that the variance norm is invariant under invertible bounded linear transformations of the data, extending previous results which are limited to unitary operators. In the Hilbert space setting, we connect the variance norm to the RKHS of the covariance operator, and establish consistency and convergence results for estimation using empirical measures with Tikhonov regularization. Using the variance norm, we introduce the notion of a kernelized nearest-neighbour Mahalanobis distance, and study some of its finite-sample concentration properties. In an empirical study on 12 real-world data sets, we demonstrate that the kernelized nearest-neighbour Mahalanobis distance outperforms the traditional kernelized Mahalanobis distance for multivariate time series novelty detection, using state-of-the-art time series kernels such as the signature, global alignment, and Volterra reservoir kernels.

Keywords: Mahalanobis distance; covariance operator; kernel methods; nearest neighbours; multivariate time series

1. Introduction

The Mahalanobis distance (Mahalanobis, 1936) is a classical tool used to measure the covariance-adjusted distance between points in space on $\mathbb{R}^d$. Given a random vector X in $\mathbb{R}^d$ with non-singular covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ and mean $\mathfrak{m} \in \mathbb{R}^d$, the Mahalanobis distance of a sample point $y \in \mathbb{R}^d$ can be defined in the following three equivalent ways:

$$\begin{aligned} d_M(y;X) &:= \sqrt{(y-\mathfrak{m})^T\Sigma^{-1}(y-\mathfrak{m})} \\ &= \left\|\Sigma^{-\frac{1}{2}}(y-\mathfrak{m})\right\|_{\mathbb{R}^d} \\ &= \sqrt{\sum_{i=1}^{d}\frac{1}{\lambda_i}\langle y-\mathfrak{m}, e_i\rangle^2}, \end{aligned} \tag{1}$$

where $(e_n, \lambda_n)_{n=1}^N$ are the eigenvector-eigenvalue pairs of the covariance matrix Σ. Initially proposed by Mahalanobis (1936) for classification, the Mahalanobis distance has since become a cornerstone technique in multivariate analysis (De Maesschalck et al., 2000). It is particularly valued for outlier detection, but its utility extends broadly, finding applications in diverse fields such as medicine (Wang et al., 2011), cybersecurity (Daneshgadeh Çakmakçı et al., 2020), chemometrics (De Maesschalck et al., 2000), unmanned vehicle detection (Lin et al., 2010), supervised classification (Xiang et al., 2008), data clustering (Brown et al., 2022), and financial market anomaly detection (Akyildirim et al., 2022), to name a few.

In this article, we propose a novel framework for Mahalanobis-type outlier detection on separable Banach and Hilbert spaces, based on a generalized notion of variance norms (Shao et al., 2023) and ideas from Cameron-Martin spaces (see e.g. Bogachev, 2015; Lifshits, 2012; Hairer, 2023). Our extended framework includes the classical Mahalanobis distance on $\mathbb{R}^d$ (Mahalanobis, 1936), the functional Mahalanobis distance on $(L^2[0,1])^d$ (Pedro Galeano and Lillo, 2015; Berrendero et al., 2020), and the kernelized Mahalanobis distance (Ruiz and López-de Teruel, 2001) as special cases. Notably, our formulation includes the general case of non-injective covariance operators, which is not addressed in the current literature.

Our work is motivated by the lack of theory surrounding outlier detection on general infinite-dimensional spaces, and more specifically work on novelty detection for time series data using the signature transform (Shao et al., 2023; Akyildirim et al., 2022; Arrubarrena et al., 2024), an object originating from the theory of rough paths (Lyons, 1998). Existing methods for signature-based outlier detection have been limited to low-dimensional time series due to the exponential $\mathcal{O}(Td^m)$ time complexity in the path dimension d when computing m-level truncated signatures of time series of length T. Our unified framework addresses this bottleneck, allowing for efficient computations of signature Mahalanobis distances in linear time with respect to d through the use of signature kernels (Kiraly and Oberhauser, 2019; Salvi et al., 2021a), without truncating the signature, allowing for true infinite-dimensional outlier detection in $\mathcal{O}(T^2d)$ time. This improvement in time complexity enables these methods to be applied to high-dimensional time series data.

1.1 Previous Infinite-dimensional Proposals

The first extension of the finite-dimensional Mahalanobis distance to finite-dimensional, non-linear data was via the kernelized Mahalanobis distance (Ruiz and López-de Teruel,

2001). It is defined by replacing the implicit dot products in (1) with inner products of a feature map, or equivalently, by positive definite kernel evaluations. This method has been successfully used in applications such as supervised classification (Wang et al., 2007; Pekalska and Haasdonk, 2009; Chang et al., 2020) and outlier detection (Lahdhiri et al., 2017; Shang et al., 2018; Daneshgadeh Çakmakçı et al., 2020).

Recently, the Mahalanobis distance was generalized to the Hilbert space $L^2[0,1]$ in the context of functional data analysis (Pedro Galeano and Lillo, 2015; Berrendero et al., 2020). This extension uses Hilbert-Schmidt covariance operators to define functional analogues of the Mahalanobis distance. In this setting, we consider a stochastic process $\big(X(t)\big)_{t\in[0,1]}$ in $L^2[0,1]$ with continuous covariance function $a(s,t) := \mathrm{Cov}[X(s), X(t)]$ and functional mean $\mathfrak{m}(t) := \mathrm{E}[X(t)] \in L^2[0,1]$. The covariance operator $\mathcal{K}$, defined by $\mathcal{K}f(t) := \int_0^1 a(s,t)f(s)ds$ for $f \in L^2[0,1]$, is symmetric, positive, compact, and hence diagonalizable by the spectral theorem via the eigenvector-eigenvalue pairs $(e_n, \lambda_n)_{n=1}^{\infty}$ with non-negative eigenvalues. The naive definition of the functional Mahalanobis distance d_{FM} reads

$$d_{FM}(f; X) := \|\mathcal{K}^{-\frac{1}{2}} f\|_{L^2[0,1]}, \tag{2}$$

and the difficulty in this infinite-dimensional setting is the non-invertibility of $\mathcal{K}^{\frac{1}{2}}$. When the inverse exists, it is given by $\mathcal{K}^{-\frac{1}{2}} f = \sum_{n=1}^{\infty} \frac{1}{\sqrt{\lambda_n}} \langle e_n, f\rangle e_n$, but since $\mathcal{K}$ is of trace class, we have that $\sum_{n=1}^{\infty} \lambda_n < \infty$. This restricts the set of elements for which the inverse is well defined. In fact, if X is a Gaussian process and f is a sample path of X, then a classical result from Gaussian probability theory states that $\mathcal{K}^{-\frac{1}{2}} f$ will almost surely not exist (see e.g. Bogachev, 2015, Theorem 2.4.7). The first paper to use d_{FM} resolved this issue by approximating $\mathcal{K}$ via its M biggest eigenvalues, where M was determined via cross-validation (Pedro Galeano and Lillo, 2015). Further theoretical advances were later made to the functional theory under the assumption that $\mathcal{K}$ is injective (Berrendero et al., 2020), using the RKHS $\mathcal{H}(\mathcal{K}) := \mathcal{K}^{1/2}(L^2[0,1])$ to regularize d_{FM} by considering the minimization problem

$$f_\alpha := \operatorname*{argmin}_{h\in\mathcal{H}(\mathcal{K})} \|f-h\|^2 + \alpha\|\mathcal{K}^{-\frac{1}{2}}h\|^2 = (\mathcal{K}+\alpha I)^{-1}\mathcal{K}f = \sum_{n=1}^{\infty} \frac{\lambda_n}{\lambda_n+\alpha}\langle f, e_n\rangle e_n, \tag{3}$$

for some $\alpha > 0$. The regularized functional Mahalanobis distance is defined by replacing f with f_α in (2), or equivalently by considering Tikhonov regularization on the operator $\mathcal{K}^{\frac{1}{2}}$. This effectively bypasses the previous invertibility issues, allowing for a well-behaved anomaly distance on $L^2[0,1]$ with theoretical guarantees like consistency of the sample estimator, and well-understood distributional properties under Gaussian assumptions on X.

1.2 Limitations in the current Functional Theory

In this work we want to address two major limitations in the current theory. The first limitation of the functional Mahalanobis theory is that the sample estimator of d_{FM} is special to the $L^2[0,1]$ setting, and reverts back to finite-dimensional Euclidean theory. The procedure involves discretizing d-dimensional sample paths on a grid of T time steps, and computing the Mahalanobis distance in $\mathbb{R}^{Td}$ (Pedro Galeano and Lillo, 2015; Ramsay and

Silverman, 2005). While this method works for $L^2[0,1]$, it fails for other inner products that require different infinite-dimensional geometry, such as anomaly detection using the signature transform from rough path theory (Akyildirim et al., 2022; Shao et al., 2023; Arrubarrena et al., 2024; Cass and Salvi, 2024). Furthermore, the theoretical guarantees of Berrendero et al. (2020) were developed for the special case $V = L^2[0,1]$, while we need these properties in the general Hilbert space setting for applications.

The second key limitation we address is the injectivity assumptions in the current functional theory. This is problematic because, when working with sample data, the empirical covariance operator is by definition of finite rank, and therefore non-injective in infinite-dimensional settings. In the functional case, a separate finite-dimensional construction was used for the sample estimator. Our unified framework does not require injectivity, and overcomes these issues by showing that the sample Mahalanobis estimator arises naturally by considering Cameron-Martin spaces with respect to empirical measures. Our framework encompasses both the general infinite-dimensional case and sample estimators within a single theory, eliminating the need for separate constructions.

1.3 Overview of the Unified Framework and Contributions

In our unified framework we work on a separable Banach space $(V, \|\cdot\|)$ with continuous dual denoted by V^*. Our main object of study is Borel probability measures μ on V of finite second moment, denoted $\mu \in \mathcal{M}_V$ as per Definition 1. For such measures μ, the vector-valued mean $\mathfrak{m} \in V$ and covariance operator $\mathcal{K} : V^* \to V$

$$\mathfrak{m} := \int_V x d\mu(x), \qquad \mathcal{K}f := \int_V (x - \mathfrak{m}) f(x - \mathfrak{m}) d\mu(x),$$

are well-defined as Bochner integrals (Chobanyan, 1987). The fundamental object we will work with is the μ-variance norm defined by

$$\|x\|_{\mu\text{-cov}} := \sup_{\substack{f \in V^*, \\ \mathrm{Cov}^{\mu}[f,f] \leq 1}} f(x), \tag{4}$$

which is well-defined for all $x \in V$, but is allowed to be infinite. The set of points for which $\|x\|_{\mu\text{-cov}} < \infty$ is called the Cameron-Martin space of μ, which we denote by H_μ. The measure-theoretic notion (4) was first suggested in a pre-print of Shao et al. (2023), but the authors provided no formal theory for the infinite-dimensional case, and the idea was subsequently reworked into a variance norm with respect to a finite sample only, without the use of probability measures or laws. In our extended setting, we define the Banach space Mahalanobis distance as

$$d_M(x; \mu) := \|x - \mathfrak{m}\|_{\mu\text{-cov}},$$

which coincides with the classical $\mathbb{R}^d$, functional $(L^2[0,1])^d$, and kernelized Mahalanobis distances, with the added benefit that our definition supports the use of non-injective covariance operators. More importantly, this entails that we no longer need one theory for the covariance operator of a random process, and a different theory for the finite sample estimator. Our framework allows to use the same results, theorems, and definitions for the

sample estimator and the underlying random process by considering the variance norm with respect to empirical measures.

An important property of the classical Mahalanobis distance in $\mathbb{R}^d$ is its invariance under invertible linear transformations of the data. Whether this remains true in the infinite-dimensional setting has been an open question. Berrendero et al. (2020) was able to prove that invariance holds for unitary operators in the functional case $V = L^2[0,1]$. Using our framework based on variance norms, we fully extend this result in Proposition 13 to invertible bounded linear operators on Banach spaces.

When specializing to Hilbert spaces, the Cameron-Martin space H_μ becomes the RKHS of the covariance operator $\mathcal{K}$, and we are able to express $\|x\|_{\mu\text{-cov}}$ in terms of the eigenvectors and eigenvalues of $\mathcal{K}$. For applications, we show that the sample estimator of the variance norm is obtained by considering empirical measures of the form $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$. The empirical μ^N-variance norm can then be computed via the procedure outlined in Theorem 22, which is based on an SVD decomposition of the inner product Gram matrix, and is closely related to kernel PCA (Schölkopf et al., 1998). Specifically, this framework allows us to define a kernelized nearest-neighbour Mahalanobis distance, which we show can be computed with the same time complexity as the classical kernelized Mahalanobis distance. This is $\mathcal{O}(N^3 + N^2K)$ time for fitting the model, and $\mathcal{O}(N(K + M))$ time for inference, where N is the number of data points, K is the time complexity of a single inner product evaluation, and $M \leq N$ is the number of eigenvalues considered.

A Tikhanov-regularized variance norm can also be obtained in the general Hilbert space setting, similar to the functional $L^2[0,1]$ setting introduced by Berrendero et al. (2020) and in (3). This allows us to extend the consistency, speed of convergence, and Gaussian distributional results from the functional case to arbitrary separable Hilbert spaces using variance norms. More specifically, we show that the sample estimator based on empirical regularized variance norms converges almost surely to the actual regularized variance norm, with a speed of convergence in probability of $O_P(N^{-\frac{1}{4}})$. Moreover, when μ is a Gaussian measure, the regularized Mahalanobis distance is equal in distribution to an infinite series of independent standard chi-squared random variables.

We further study the finite-sample properties of the nearest-neighbour distance and its regularized Mahalanobis variant to justify their use in infinite-dimensional settings, where one might expect random points to be almost equidistant. To demonstrate the difficulty of this problem, we show for any set of linearly independent points $\{x_1, ..., x_N\} \subset V$ defining the empirical measure μ^N, that the unregularized empirical Mahalanobis distance satisfies $\|x_i - x_j\|_{\mu^N} = 2\sqrt{N}$ for all $i \neq j$. This highlights the need for a more nuanced analysis and provides additional justification for regularization. To address this, we establish finite-sample concentration bounds for the difference between the nearest- and furthest-neighbour distances under the Hilbert and regularized Mahalanobis norms. Our analysis is based on a Hilbert space Hanson-Wright inequality (Chen and Yang, 2021), and concentration properties of the finite sample covariance operator (Koltchinskii and Lounici, 2017). Importantly, we show that the nearest neighbour concentration phenomenon is not governed by the ambient dimension of the space V, but rather by the *effective dimensionality* of the covariance operator of the underlying data measure, as given in Definition 30.

1.4 Organization of the Paper

Section 2 introduces the covariance operator and the Cameron-Martin space of a probability measure μ in the Banach space setting. We prove that the variance norm coincides with the classical Cameron-Martin norm, and show that it is invariant under invertible bounded linear transformations of the data. The Mahalanobis and nearest-neighbour Mahalanobis distance is defined, and several important properties are proved.

In Section 3.1 we specialize to Hilbert spaces, and connect the Cameron-Martin space to the RKHS of the covariance operator μ. We derive computational formulas based on empirical measures with applications to kernel learning. We then define a Tikhonov-regularized variance norm, and derive consistency, speed of convergence, and Gaussian distributional results for the regularized variance norm.

Section 4 studies finite-sample properties of the nearest-neighbour Mahalanobis distance and establishes concentration bounds that justify its use in infinite-dimensional settings.

We conclude the paper with an application to kernelized multivariate time series novelty detection in Section 5, where we apply our developed framework to various state of the art time series kernels and compare their effectiveness.

2. Theoretical Foundations of Variance Norms

Throughout this section, we consider a separable Banach space $(V, \|\cdot\|)$ with continuous dual V^*, and a Borel probability measure μ defined on V. The primary objective of this section is to develop a comprehensive theory of variance norms on Banach spaces by extending the concepts of Cameron-Martin spaces and norms to non-Gaussian measures. This will allow us to extend the definition of Mahalanobis distance to the Banach space setting.

2.1 Covariance Operators

Covariance operators serve as the natural generalization of covariance matrices to infinite-dimensional spaces (Chobanyan, 1987; Tailen Hsing, 2015). These are classical objects in probability theory, and can be defined for random measures — or equivalently, probably measures — with finite second moment.

Definition 1 *Let $p \geq 1$. A measure μ on V is said to have finite p-th moment if $\|\cdot\| \in L^p(V, \mu)$. We denote by $\mathcal{M}_V$ the set of all Borel probability measures μ of finite second moment.*

In particular, empirical measures of the form $\frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ always belong to $\mathcal{M}_V$, which is essential for computations with observed data. The fundamental object of study in our framework is the covariance operator of μ, which is defined using the classical notion of Bochner integration (see e.g. Ledoux and Talagrand, 1991).

Definition 2 *Using the Bochner integral, we defined the mean of $\mu \in \mathcal{M}_V$ as the expectation $\mathfrak{m} = \int_V x d\mu(x) = \mathrm{E}^{x\sim\mu}[x]$. On the continuous dual V^* we define the functional covariance quadratic form $q : V^* \times V^* \to \mathbb{R}$ by*

$$q(f,g) := \mathrm{Cov}^{\mu}[f,g] = \mathrm{E}^{x\sim\mu}\Big[f(x-\mathfrak{m})g(x-\mathfrak{m})\Big] = \langle f, g\rangle_{L^2(\mu_{\mathfrak{m}})},$$

for $f, g \in V^*$. Here $\mu_{\mathfrak{m}}$ is the measure obtained by shifting μ by the mean $\mathfrak{m}$, i.e. the pushforward of μ under the map $x \mapsto x - \mathfrak{m}$. The above quantities are well-defined since μ is assumed to have finite second moment, which allows for the inclusion $V^* \subset L^2(\mu_{\mathfrak{m}})$.

One observes that q defines a positive quadratic form on V^*, but may fail to be an inner product if $q(f, f) = 0$ for $f \neq 0$. An alternative characterization of the functional covariance q is through the so-called covariance operator of μ. This is a natural functional-analytic object to study when we no longer have access to the Gaussian tools from the classical theory of Cameron-Martin spaces.

Definition 3 *We define the covariance operator of $\mu \in \mathcal{M}_V$ to be the bounded linear operator $\mathcal{K} : V^* \to V$ defined via*

$$\mathcal{K}f := \int_V x f(x) d\mu_{\mathfrak{m}}(x),$$

for $f \in V^*$.

The covariance operator of $\mu \in \mathcal{M}_V$ is well-defined due to the bound $\|xf(x)\| \leq \|f\|_{V^*}\|x\|^2$, which additionally implies that $\mathcal{K}$ indeed is a bounded linear operator. The following lemma establishes a useful relationship between the covariance operator $\mathcal{K}$ and the functional covariance q, which will be required in the subsequent analysis. These basic properties are well-known, but we include a short proof here for completeness.

Lemma 4 *Let $\mu \in \mathcal{M}_V$. The covariance operator $\mathcal{K} : V^* \to V$ is the unique operator satisfying*

$$q(f, g) = f(\mathcal{K}g),$$

for all $f, g \in V^*$. Moreover, $\mathcal{K}$ is compact, and in particular bounded.

Proof Suppose that $\mathcal{K}$ is the covariance operator of μ, and fix $f, g \in V^*$. Using the Bochner integral representation of $\mathcal{K}$ we obtain that

$$q(f, g) = \int_V f(x)g(x)d\mu_{\mathfrak{m}}(x) = \int_V f(xg(x))d\mu_{\mathfrak{m}}(x) = f\bigg(\int_V xg(x)d\mu_{\mathfrak{m}}(x)\bigg) = f(\mathcal{K}g),$$

where the second to last equality follows from the fact that bounded operators commute with Bochner integrals (see e.g. Aliprantis and Border, 2007, Lemma 11.45). Conversely, if $q(f, g) = f(\widetilde{\mathcal{K}}g)$ for all $f, g \in V^*$ and some operator $\widetilde{\mathcal{K}}$, then $0 = f(\widetilde{\mathcal{K}}g - \mathcal{K}g)$. Consequently $\widetilde{\mathcal{K}}g = \mathcal{K}g$ by the Hahn-Banach theorem, for each $g \in V^*$.

As for compactness, suppose that f_n is a bounded sequence in V^*, say $\|f_n\|_{V^*} \leq 1$. By Alaoglu's Theorem (Lax, 2014, Theorem 12.3) there exists a weak*-convergent subsequence f_{n_k} converging to some $f \in V^*$, that is $\lim_k f_{n_k} = f$ pointwise. Since $\|xf_{n_k}(x)\| \leq \|x\|^2$ for all $x \in V$, it follows by the dominated convergence theorem for Bochner integrals (Aliprantis and Border, 2007, Theorem 11.46) that

$$\lim_k \mathcal{K}f_{n_k} = \int_V \lim_k x f_{n_k}(x) d\mu_{\mathfrak{m}}(x) = \mathcal{K}f,$$

which concludes the proof that $\mathcal{K}$ is compact. ■

Remark 5 *An alternative way to define $\mathcal{K}$ is by using the quadratic form q to first define a linear operator $\mathcal{K} : V^* \to V^{**}$ via $(\mathcal{K}f)(g) = q(f, g)$. One then realizes that $\mathcal{K}f$ actually is an evaluation functional of the vector $\int_V xf(x)d\mu_{\mathfrak{m}}(x) \in V$ using the Lemma above, from which the first definition of $\mathcal{K}$ is recovered.*

2.2 The Cameron-Martin Space and Extended Covariance Operators

A key challenge when working with the covariance operator $\mathcal{K} : V^* \to V$ is that $\mathcal{K}$ may be non-injective. We address this by introducing what we term the extended covariance operator, which is injective in the $L^2(\mu_{\mathfrak{m}})$ topology. Our proposed approach of defining Cameron-Martin spaces via this extended covariance operator is to the best of our knowledge novel, and leads to an elegant Gaussian-free approach to variance norms.

Definition 6 *Let $\mu \in \mathcal{M}_V$. We define the space $\mathcal{R}_\mu$ to be the closure of V^* in the $L^2(\mu_{\mathfrak{m}})$ topology.*

The space $\mathcal{R}_\mu$ plays a crucial role throughout this section. As a closed subset of a Hilbert space, $\mathcal{R}_\mu$ inherits a Hilbert space structure under the $L^2(\mu_{\mathfrak{m}})$ norm. Our goal is to extend $\mathcal{K}$ to an operator $\mathcal{C} : \mathcal{R}_\mu \to V$, where the image $H_\mu = \mathcal{C}(\mathcal{R}_\mu)$ will be defined as the Cameron-Martin space of μ. By an *extension*, we mean that $\mathcal{C}$ coincides with $\mathcal{K}$ on V^*. This extension is what enables our subsequent results to apply to empirical measures of a sample, which by definition gives rise to finite-rank, and in particular non-injective, covariance operators. The following proposition shows that by changing topologies from the operator norm on V^* to the $L^2(V, \mu_m)$ topology, we obtain a well-defined extended injective operator. The existence of this extension is not immediately obvious, since the natural estimate $\|f\|^2_{L^2(\mu_{\mathfrak{m}})} = \int_V f(x)^2 d\mu_{\mathfrak{m}}(x) \leq \|f\|^2_{V^*} \int_V \|x\|^2 \mu_{\mathfrak{m}}(x)$ goes in the wrong direction.

Proposition 7 *The covariance operator $\mathcal{K} : V^* \to V$ extends to a bounded linear operator $\mathcal{C} : \mathcal{R}_\mu \to V$, where $\mathcal{R}_\mu$ is the $L^2(V, \mu_{\mathfrak{m}})$-closure of V^*, via the limit*

$$\mathcal{C}k := \lim_{n\to\infty} \mathcal{K}f_n = \int_V xk(x)d\mu_{\mathfrak{m}}(x),$$

where $(f_n)_{n=1}^\infty \subset V$ is any sequence converging to $k \in \mathcal{R}_\mu \subset L^2(V, \mu_{\mathfrak{m}})$.

Proof Let $k \in \mathcal{R}_\mu$. Since V^* is dense in $\mathcal{R}_\mu$, there exists a sequence $f_n \in V^*$ such that $\|f_n - k\|_{L^2(\mu_{\mathfrak{m}})} \to 0$ as $n \to \infty$. By Hölders inequality we have that

$$\begin{aligned}\left\|\mathcal{K}f_n - \int_V xk(x)d\mu_{\mathfrak{m}}(x)\right\| &\leq \int_V \|x\|\big|(k - f_n)(x)\big|d\mu_{\mathfrak{m}}(x) \\ &= \left(\int_V \|x\|^2 d\mu_{\mathfrak{m}}(x)\right)^{\frac{1}{2}} \left(\int_V \big|(k - f_n)(x)\big|^2 d\mu_{\mathfrak{m}}(x)\right)^{\frac{1}{2}},\end{aligned}$$

which goes to 0 as $n \to \infty$. This holds for any such sequence, and the conclusion follows. ■

We can now define the Cameron-Martin space of a general measure $\mu \in \mathcal{M}_V$. The Cameron-Martin space will be a Hilbert space isometrically isomorphic to $\mathcal{R}_\mu$, whose norm

will naturally be given by the covariance-adjusted distance through the extended covariance operator. This will provide the natural generalization of the Mahalanobis distance for any separable Banach space.

Definition 8 *Let $\mu \in \mathcal{M}_V$ with extended covariance operator $\mathcal{C} : \mathcal{R}_\mu \to V$. We define the Cameron-Martin space H_μ of μ to be the set $H_\mu := \mathcal{C}(R_\mu)$.*

Proposition 9 *The operator $\mathcal{C} : \mathcal{R}_\mu \to H_\mu$ is invertible. Hence H_μ is a Hilbert space under the norm*

$$\|h\|_{H_\mu} := \|\mathcal{C}^{-1}h\|_{L^2(\mu_\mathfrak{m})}, \qquad \langle h, l\rangle_{H_\mu} := \left\langle \mathcal{C}^{-1}h, \mathcal{C}^{-1}l\right\rangle_{L^2(\mu_\mathfrak{m})},$$

where $h, l \in H_\mu$.

Proof We need to prove that $\mathcal{C}$ is injective. To this end, assume that $\mathcal{C}k = 0$ for some $k \in \mathcal{R}_\mu$. By definition there exists a sequence $f_n \in V^*$ such that $f_n \to k$ in $\mathcal{R}_\mu$. Lemma 4 then implies that

$$0 = \lim_n g\big(\mathcal{K}f_n\big) = \lim_n \langle g, f_n\rangle_{L^2(\mu_\mathfrak{m})} = \langle g, k\rangle_{L^2(\mu_\mathfrak{m})},$$

for all $g \in V^*$. By continuity we obtain that $\langle g, k\rangle_{L^2(\mu_\mathfrak{m})} = 0$ for all $g \in \mathcal{R}_\mu$. Consequently we find that $k = 0$ ($\mu_\mathfrak{m}$-a.e.), which shows that $\mathcal{C} : \mathcal{R}_\mu \to H$ is injective. The latter statement of the proposition follows from the Hilbert space structure of $\mathcal{R}_\mu \subset L^2(V, \mu_\mathfrak{m})$ and the linearity of $\mathcal{C}$. ∎

The following fundamental result shows that the μ-variance norm $\|\cdot\|_{\mu\text{-cov}}$ is a genuine norm on a subspace of V, and infinite otherwise. More precisely, this subspace is the Cameron-Martin space $H_\mu \subset V$, and the μ-variance norm coincides with the Cameron-Martin Hilbert norm when restricted to this space. This result provides a solid theoretical foundation for variance-adjusted norms in the general infinite-dimensional setting, bridging the gap between classical infinite-dimensional Gaussian probability theory and the Mahalanobis distance literature. Recall that the μ-variance norm for $x \in V$ is defined as

$$\|x\|_{\mu\text{-cov}} := \sup_{f \in V^*,\, q(f,f) \le 1} f(x),$$

where q is the functional covariance of μ.

Theorem 10 *The Cameron-Martin space of $\mu \in \mathcal{M}_V$ is characterized by*

$$H_\mu = \{h \in V : \|h\|_{\mu\text{-}cov} < \infty\}.$$

Furthermore, the Cameron-Martin norm $\|\cdot\|_{H_\mu}$ and the variance norm $\|\cdot\|_{\mu\text{-}cov}$ coincide on H_μ, or in other words

$$\|h\|_{H_\mu} := \|\mathcal{C}^{-1}h\|_{L^2(\mu_\mathfrak{m})} = \|h\|_{\mu\text{-}cov},$$

for all $h \in H_\mu$.

Proof Suppose that $h = \mathcal{C}k$ for some $k \in \mathcal{R}_\mu$. We want to show that the variance norm $\|h\|_{\mu\text{-cov}}$ is finite and equal to $\|h\|_{H_\mu}$. To this end, observe that

$$\begin{aligned}\|h\|_{\mu\text{-cov}} &= \sup_{f\in V^*,\, q(f,f)\le 1} f(\mathcal{C}k) = \sup_{f\in V^*,\, q(f,f)\le 1} \langle f, k\rangle_{L^2(\mu_{\mathfrak{m}})} \\ &= \sup_{l\in\mathcal{R}_\mu,\, \|l\|_{L^2(\mu_{\mathfrak{m}})}\le 1} \langle l, k\rangle_{L^2(\mu_{\mathfrak{m}})} = \|\mathcal{C}^{-1}h\|_{L^2(\mu_{\mathfrak{m}})},\end{aligned}$$

where the third equality follows by the fact that V^* is dense in $\mathcal{R}_\mu$.

Conversely, assume that $\|x\|_{\mu\text{-cov}} < \infty$ for some $x \in V$. Let $T_x : V^* \to \mathbb{R}$ denote the evaluation functional $T_x g = g(x)$. For $g \in V^*$ with $\|g\|_{L^2(\mu_{\mathfrak{m}})} > 0$ we have the bound

$$|T_x g| = |g(x)| \le \|g\|_{L^2(\mu_{\mathfrak{m}})} \sup_{f\in V^*,\, q(f,f)\le 1} f(x),$$

hence T_x extends to a linear operator $\mathcal{T}_x : \mathcal{R}_\mu \to \mathbb{R}$ by continuity. More specifically, $\mathcal{T}_x k$ for $k \in \mathcal{R}_\mu$ can be defined via $\mathcal{T}_x k := \lim_n T_x f^{(n)} = \lim_n f^{(n)}(x)$ where $f^{(n)} \in V^*$ is any sequence such that $\|k - f^{(n)}\|_{L^2(\mu_{\mathfrak{m}})} \to 0$. The operator norm for a general bounded operator $\mathcal{T} \in \mathcal{R}_\mu^*$ is given by

$$\|\mathcal{T}\|_{\mathcal{R}^*} := \sup_{k\in\mathcal{R}_\mu,\, \|k\|_{L^2(\mu_{\mathfrak{m}})}\le 1} \mathcal{T}k \;=\; \sup_{f\in V^*,\, q(f,f)\le 1} \mathcal{T}f,$$

where the last equality follows by the fact that V^* is dense in $\mathcal{R}_\mu$. Restricting this to extended evaluation functionals $\mathcal{T}_x$ we obtain that

$$\|\mathcal{T}_x\|_{\mathcal{R}^*} = \sup_{f\in V^*,\, q(f,f)\le 1} f(x) = \|x\|_{\mu\text{-cov}}.$$

Since $\mathcal{T}_x \in \mathcal{R}_\mu^*$ if and only if the operator norm is finite, we may use the fact that $\mathcal{R}_\mu$ is a Hilbert space to identify $\mathcal{T}_x$ with an element of $\mathcal{R}_\mu$ itself, say k_x, such that $\mathcal{T}_x l = \langle l, k_x\rangle$ for all $l \in \mathcal{R}_\mu$. If $f \in V^*$, then

$$f\big(\mathcal{C}k_x - x\big) = \langle k_x, f\rangle_{L^2(\mu_{\mathfrak{m}})} - f(x) = f(x) - f(x) = 0,$$

and it follows by Hahn-Banach that $\mathcal{C}k_x = h$. This concludes the proof. ■

Remark 11 *In the above theorem we proved that $h \in H_\mu$ if and only if the evaluation functional T_h extends to a continuous linear functional on $\mathcal{R}_\mu \subset L^2(V, \mu_{\mathfrak{m}})$. This also proves that H_μ is a reproducing kernel Hilbert space.*

Remark 12 *The literature on Gaussian measures and infinite-dimensional Gaussian probability theory is rich in examples of Cameron-Martin spaces and norms. A classical example is the Wiener measure on $C[0,1]$, the space of continuous functions, where the Cameron-Martin space is the set of all absolutely continuous functions with square integrable derivative, with Cameron-Martin norm $\|h\|_{H_\mu} = \int_0^1 |\dot{h}(t)|^2 dt$. More generally, there exist expressions for Cameron-Martin spaces and norms for Gaussian measures on $C[0,1]$ in the case where the underlying Gaussian process can be written as an integral with respect to Gaussian white noise. We refer to Lifshits (2012) for further details.*

One important property of the classical Mahalanobis distance on $\mathbb{R}^d$ is invariance with respect to non-singular linear transformations of the data. The infinite-dimensional case is more difficult, as Berrendero et al. (2020) noted in the special case $V = L^2[0,1]$ in the Hilbert space setting of functional data analysis. They were able to prove that their functional Mahalanobis distance is invariant with respect to unitary transformations of the data. Using our proposed framework based on variance norms, we are able to extend this result to the Banach space setting for general invertible bounded linear operators. The following proposition comes as a natural consequence of the Cameron-Martin perspective we take in this paper, with a short and elegant proof.

Proposition 13 *The μ-variance norm is invariant under bounded invertible linear transformations of the data. More specifically, if $A : V \to V$ is an invertible bounded linear operator, and $\nu = \mu \circ A^{-1}$, then $\|Ax\|_{\nu\text{-}cov} = \|x\|_{\mu\text{-}cov}$ for all $x \in V$.*

Proof Denote by q^ν and q the functional covariance of ν and μ respectively. First, we observe by change of variables that

$$q^\nu(f,f) = \int f(x - A\mathfrak{m})^2 d\nu = \int f(Ax - A\mathfrak{m})^2 d\mu = q(f \circ A, f \circ A).$$

Next, note that $\{g \in V^* : g = f \circ A\} = V^*$, which follows from the fact that the adjoint operator A^* is invertible if and only if A is. Combining the above, we obtain that

$$\|Ax\|_{\nu\text{-cov}} = \sup_{f \in V^*,\, q(f\circ A, f\circ A) \le 1} f(Ax) = \sup_{g \in V^*,\, q(g,g)\le 1} g(x) = \|x\|_{\mu\text{-cov}}.$$

■

2.3 Mahalanobis Distance and Conformance Score

Having introduced the necessary theoretical background in the previous subsections, we are now ready to define the Mahalanobis distance on any separable Banach space V. We claim that a natural definition of an anomaly distance on V with respect to a law μ is the μ-variance norm of a new sample x against the mean $\mathfrak{m}$, as outlined in the following definition:

Definition 14 *We define the Mahalanobis distance $d_M(x;\mu)$ of the element $x \in V$ with respect to the measure $\mu \in \mathcal{M}_V$ to be*

$$d_M(x;\mu) := \|x - \mathfrak{m}\|_{\mu\text{-}cov},$$

where $\mathfrak{m}$ is the mean of μ.

For real world applications, μ is often unknown, and an estimator has to be used. This fits naturally within our proposed framework via working with empirical measures: Given a corpus of data $\{x_1, ..., x_N\} \subset V$, the empirical measure of the data is given by $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$, leading to the sample Mahalanobis distance $d_M(\cdot, \mu^N)$. The following examples demonstrate that our definition coincides with, and in fact extends the Mahalanobis distance in $\mathbb{R}^d$ to random variables with possibly degenerate covariance matrices Σ. Furthermore, the estimator of the Mahalanobis distance will be given by the case where μ is the empirical measure of the underlying data.

Example 1 (Finite-Dimensional Case) *Let μ be a measure on $\mathbb{R}^d$ with covariance matrix $\Sigma = \mathrm{E}^{x\sim\mu}[(x-\mathfrak{m})(x-\mathfrak{m})^T]$ and mean $\mathfrak{m} = \mathrm{E}^{x\sim\mu}[x]$. The functional covariance q with respect to μ is*

$$q(a,b) = \mathrm{E}^{x\sim\mu}\Big[\langle a, x-\mathfrak{m}\rangle\langle b, x-\mathfrak{m}\rangle\Big] = \mathrm{E}^{x\sim\mu}\Big[\sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j (x-\mathfrak{m})_i (x-\mathfrak{m})_j\Big]$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \Sigma_{i,j} = \langle a, \Sigma b\rangle = \langle \Sigma a, b\rangle = a^T \Sigma b,$$

for all $a, b \in \mathbb{R}^d$. Lemma 4 implies that the covariance operator of μ is simply Σ, from which Theorem 18 gives that the Cameron-Martin space is $H_\mu = \mathrm{Im}(\Sigma)$. Therefore, by Theorem 10, the μ-variance norm, and thus the Mahalanobis distance, is:

$$\|x-\mathfrak{m}\|^2_{\mu\text{-}cov} = \sup_{a\in\mathbb{R}^d} \frac{\langle a, x-\mathfrak{m}\rangle^2}{a^T\Sigma a} = \begin{cases} (x-\mathfrak{m})^T\Sigma^{-1}(x-\mathfrak{m}) & \text{if } x-\mathfrak{m}\in \mathrm{Im}(\Sigma), \\ +\infty & \text{otherwise.} \end{cases}$$

In the specific case of an empirical measure $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$, Σ becomes the sample covariance matrix $\hat{\Sigma}$, and $\mathfrak{m}$ becomes the sample mean $\hat{\mathfrak{m}}$. This recovers the classical (potentially degenerate) Mahalanobis distance used in finite dimensions (Shao et al., 2023).

While the classical Mahalanobis distance is a widely used metric for outlier detection, measuring distance to the mean may perform poorly in high dimensions. For instance, consider i.i.d. standard Gaussian data in $\mathbb{R}^d$: the covariance operator in this case is the identity, so the Mahalanobis distance reduces to the Euclidean norm, which concentrates around the sphere of radius $\sqrt{d}$. As a result, the likelihood of a new normal sample being close to the origin is very small if d is large. Consequently, if the Mahalanobis distance is used directly as an anonmaly score, such samples may incorrectly be classified as being outliers. An alternative approach in such scenarios is to use the k-nearest-neighbour distance to the normal corpus (Hautamaki et al., 2004; Verdier and Ferreira, 2011; Shao et al., 2023). This approach requires choosing a metric for calculating the nearest-neighbours. In the finite-dimensional setting the Euclidean, Minkowski, Manhattan, or even the Mahalanobis distance itself are commonly used. Shao et al. (2023) coined the term *conformance score* for the case when the Mahalanobis distance is used in conjunction with the 1-nearest-neighbour Mahalanobis distance (see also Verdier and Ferreira, 2011). Below, we generalize this notion to the Banach space setting for laws $\mu \in \mathcal{M}_V$. Note that the notion of variance norm by Shao et al. (2023) is restricted to empirical measures only, while our unified framework considers any law μ. In Sections 3.3 and 5, we derive computational formulas for the infinite-dimensional conformance score and evaluate these anomaly metrics in the context of time series novelty detection.

Definition 15 *Let $\mu \in \mathcal{M}_V$ for a Banach space V, and let $\{x_1, \cdots, x_N\} \subset V$ be a corpus of observed data. We define the conformance score $d_C(x;\mu)$ of x with respect to μ and the corpus as*

$$d_C(x;\mu) := \min_{1\le i\le N} \|x - x_i\|_{\mu\text{-}cov}.$$

The following proposition extends Example 1 to the case of empirical measures on a Banach space. In this setting, the Cameron-Martin space H_{μ^N} associated with the empirical measure μ^N is finite-dimensional. Nevertheless, a challenge lies in the fact the variance norm depends non-trivially on all of V^*, which is infinite-dimensional. The result also establishes the basis-independence of the variance norm with respect to the basis in which the data is observed. This extends the results of Shao et al. (2023) from the finite-dimensional setting to the Banach space setting.

Proposition 16 *Let $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ be an empirical measure. Write $y_i := x_i - \widehat{\mathfrak{m}}$, $i \in \{1, \cdots, N\}$ for the centered data, where $\widehat{\mathfrak{m}} = \frac{1}{N}\sum_{i=1}^N x_i$ is the empirical mean. Then the following statements hold:*

(i) The Cameron-Martin space is $H_{\mu^N} = \operatorname{span}\{y_1, \cdots, y_N\}$.

(ii) Let $\{e_1, \cdots, e_M\}$ be a basis of H_{μ^N}, and $A : V \to H_{\mu^N}$ be a surjective projection. Denote by $a^{(x)} \in \mathbb{R}^M$ for $x \in V$ the coordinates of Ax with respect to said basis, that is $Ax = \sum_{i=1}^M a_i^{(x)} e_i$. Then the μ^N-variance norm is given by

$$\|x\|_{\mu^N\text{-}cov} = \begin{cases} (a^{(x)})^T \Sigma^{-1} a^{(x)} & \text{if } a^{(x)} \in \operatorname{Im}(\Sigma), \\ +\infty & \text{otherwise,} \end{cases}$$

where $\Sigma \in \mathbb{R}^{M\times M}$ is the empirical covariance matrix of the coordinates $a^{(y_1)}, \cdots, a^{(y_N)}$.

Proof We begin by proving (ii), from which (i) will follow. Since the covariance operator $\mathcal{K} : V^* \to V$ is a finite rank operator, we have that $\mathcal{C}(\mathcal{R}_{\mu^N}) = \mathcal{K}(V^*)$, where $\mathcal{C} : \mathcal{R}_{\mu^N} \to V$ is the extended covariance operator. Hence it follows from the expression

$$\mathcal{K}f = \frac{1}{N}\sum_{i=1}^N y_i f(y_i),$$

that $H_{\mu^N} = \mathcal{K}(V^*) \subset \operatorname{span}\{y_1, \cdots, y_N\}$, which by Theorem 10 implies that $\|x\|_{\mu^N\text{-cov}}$ is infinite for all $x \notin \operatorname{span}\{y_1, \cdots, y_N\}$. Consequently we will only need to consider this finite span in the subsequent analysis.

Next, observe that by writing $y_i = \sum_{m=1}^M a_m^{(y_i)} e_m$, we obtain the following expression for the functional covariance of μ^N for all $f \in V^*$

$$q(f,f) = \frac{1}{N}\sum_{i=1}^N f(y_i)^2 = \frac{1}{N}\sum_{i=1}^N \left(\sum_{m=1}^M a_m^{(y_i)} f(e_m)\right)^2 = \frac{1}{N}\sum_{i=1}^N \langle a^{(y_i)}, b\rangle_{\mathbb{R}^M}^2 = \langle b, \Sigma b\rangle_{\mathbb{R}^M},$$

with $b \in \mathbb{R}^M$ given by $b_m = f(e_m)$, and where $\Sigma \in \mathbb{R}^{M\times M}$ is the empirical covariance matrix of the coordinates $a^{(y_1)}, \cdots, a^{(y_N)}$. Conversely, if $b \in \mathbb{R}^M$ is fixed, then $f_b(x) := \langle a^{(x)}, b\rangle_{\mathbb{R}^M}$ defines a continuous linear functional on $\operatorname{span}\{y_1, \cdots, y_N\}$, which extends continuously to V via Hahn-Banach. Consequently we obtain that

$$\|x\|_{\mu^N\text{-cov}}^2 = \sup_{f\in V^*} \frac{f(x)^2}{q(f,f)} = \sup_{b\in\mathbb{R}^M} \frac{\langle a^{(x)}, b\rangle_{\mathbb{R}^M}^2}{\langle b, \Sigma b\rangle_{\mathbb{R}^M}} = \begin{cases} (a^{(x)})^T \Sigma^{-1} a^{(x)} & \text{if } a^{(x)} \in \operatorname{Im}(\Sigma), \\ +\infty & \text{otherwise,} \end{cases} \tag{5}$$

where the last equality follows from Example 1. The equality $H_\mu = \text{span}\{y_1, \cdots, y_N\}$ then follows by Theorem 10 since $a^{(x)} \in \text{Im}(\Sigma)$ if and only if $x \in \text{span}\{y_1, \cdots, y_N\}$, which proves (i). ■

We want to stress that the choice of basis and projection map is purely for computational convenience, and will lead to the same result since the definition of the variance norm is basis-independent. The following corollary is a direct consequence of Eq. (5), and relates the μ-variance norm with respect to empirical measures on Banach spaces to the classical Mahalanobis distance in $\mathbb{R}^M$.

Corollary 17 *Under the assumptions of Proposition 16, we have that*

$$\|x\|_{\mu^N\text{-}cov} = \|a^{(x)}\|_{\nu-cov},$$

where $\nu = \frac{1}{N}\sum_{i=1}^N \delta_{a^{(y_i)}}$ *is an empirical measure on* $\mathbb{R}^M$.

While Corollary 17 provides a way to compute variance norms using coordinates relative to a basis of the Cameron-Martin space H_{μ^N}, the construction of such a basis depends heavily on the structure of the underlying space V. This is evident, for example, in the functional $L^2[0,1]$ Mahalanobis literature, where the Mahalanobis distance for a d-dimensional time series of length T is computed by flattening the data and applying the standard Mahalanobis distance in $\mathbb{R}^{Td}$ (Pedro Galeano and Lillo, 2015; Berrendero et al., 2020). From the perspective of Proposition 16, this corresponds to constructing a basis for the discretized paths. This approach works well due to the specific structure of the $L^2[0,1]$ inner product. However, it does not generalize to settings where the geometry of V is fundamentally different. In such cases, it is unclear how to obtain a tractable algorithm without explicitly relying on a Hilbert space structure to facilitate computations.

3. Specialization to Hilbert Spaces

In this section we specialize to the case where V is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$. In doing so, we are able to diagonalize the covariance operator $\mathcal{K}$ of $\mu \in \mathcal{M}_V$, to express the variance norm and Cameron-Martin space in terms of the eigenvalues of $\mathcal{K}$. This generalizes the results of Berrendero et al. (2020) from the setting $L^2[0,1]$ of functional data analysis to any separable Hilbert space, without any assumptions of continuity of a stochastic process or injectivity of covariance operator. This generalization is necessary to obtain a theory consistent for empirical measures, which by definition gives rise to non-injective covariance operators, and to obtain a general algorithm for computations which takes into account the infinite-dimensional properties of the chosen space V.

3.1 Hilbert Space Characterization

Recall from Lemma 4 that $\langle g, \mathcal{K}f \rangle = q(g,f) = q(f,g) = \langle f, \mathcal{K}g \rangle$, for all $f, g \in V$, and that $\mathcal{K}$ is compact. Consequently, this implies that $\mathcal{K}$ is a symmetric, positive, compact operator, hence by the spectral theorem (see e.g. Lax, 2014, Theorem 28.3) there exists an orthonormal sequence of eigenvectors $(e_n)_{n=1}^\infty$ and non-negative eigenvalues $(\lambda_n)_{n=1}^\infty$ such

that

$$\mathcal{K}f = \sum_{n=1}^{\infty} \lambda_n \langle f, e_n \rangle e_n, \quad \forall f \in V. \tag{6}$$

If V is finite-dimensional, we instead replace $(e_n)_{n=1}^{\infty}$ by a finite collection. Our ultimate goal is to derive an explicit computational formula for the variance norm $\|\cdot\|_{H_\mu}$. To achieve this, we first need to characterize the Cameron-Martin space H_μ, as Theorem 10 states that this is the subspace of V where the variance norm is finite. Understanding H_μ will also be crucial when we later consider empirical measures constructed from observed data in Section 3.3.

Theorem 18 *Let $\mu \in \mathcal{M}_V$ for a Hilbert space V, and let $(e_n)_{n=1}^{\infty}$ and $(\lambda_n)_{n=1}^{\infty}$ be orthonormal eigenvectors and eigenvalues that diagonalize the covariance operator $\mathcal{K} : V \to V$. Then the Cameron-Martin space H_μ is given by*

$$H_\mu = \left\{ h \in V : \sum_{n=1,\, \lambda_n \neq 0}^{\infty} \frac{\langle h, e_n \rangle^2}{\lambda_n} < \infty \text{ and } \left(\forall n \geq 1,\, \lambda_n = 0 \implies \langle h, e_n \rangle = 0 \right) \right\} = \mathcal{K}^{\frac{1}{2}}(V),$$

where $\mathcal{K}^{\frac{1}{2}}$ is the square root operator of $\mathcal{K}$. The variance norm is given by

$$\|h\|_{\mu\text{-}cov}^2 = \begin{cases} \|\mathcal{K}^{-\frac{1}{2}} h\|^2 = \sum_{n=1,\, \lambda_n \neq 0}^{\infty} \frac{\langle h, e_n \rangle^2}{\lambda_n} & \text{if } h \in \mathcal{K}^{\frac{1}{2}}(V), \\ \infty & \text{otherwise.} \end{cases} \tag{7}$$

Proof Let $h \in H_\mu$, that is $h = \mathcal{C}k = \lim_n \mathcal{K} f^{(n)}$ for some $k \in \mathcal{R}_\mu$ and a sequence $f^{(n)} \in V^*$ such that $\|k - f^{(n)}\|_{L^2(\mu_{\mathfrak{m}})} \to 0$. Using the symmetry of $\mathcal{K}$ we obtain that

$$\langle h, e_j \rangle = \lim_n \langle \mathcal{K} f^{(n)}, e_j \rangle = \lim_n \langle f^{(n)}, \lambda_j e_j \rangle,$$

for all $j \geq 1$, and hence $\langle h, e_j \rangle = 0$ whenever $\lambda_j = 0$. Moreover, Lemma 4 implies that $\|f^{(n)}\|_{L^2(\mu_{\mathfrak{m}})} = \|\mathcal{K}^{\frac{1}{2}} f^{(n)}\|$ for all $n \geq 1$, and consequently we find that

$$\begin{aligned} \|h\|_{H_\mu}^2 &= \|k\|_{L^2(\mu_{\mathfrak{m}})}^2 = \lim_n \|f^{(n)}\|_{L^2(\mu_{\mathfrak{m}})}^2 = \lim_n \|\mathcal{K}^{\frac{1}{2}} f^{(n)}\|^2 \\ &= \lim_n \sum_{j=1}^{\infty} \lambda_j \langle f^{(n)}, e_j \rangle^2 \geq \sum_{j=1}^{\infty} \lim_n \lambda_j \langle f^{(n)}, e_j \rangle^2 \\ &= \sum_{j=1,\, \lambda_j \neq 0}^{\infty} \lim_n \frac{\langle \mathcal{K} f^{(n)}, e_j \rangle^2}{\lambda_j} = \sum_{j=1,\, \lambda_j \neq 0}^{\infty} \frac{\langle h, e_j \rangle^2}{\lambda_j}, \end{aligned}$$

where the inequality follows by Fatou's lemma. This shows that $h \in \mathcal{K}^{\frac{1}{2}}(V)$ since $\|h\|_{H_\mu}^2 < \infty$.

Conversely, let $h \in \mathcal{K}^{\frac{1}{2}}(V)$. We want to show that $h = \mathcal{C}k$ for some $k \in \mathcal{R}_\mu$. We do this by defining the sequence $f^{(n)} := \sum_{j=1, \lambda_j \neq 0}^{n} \frac{\langle h, e_j \rangle}{\lambda_j} e_j$, and noting that $\mathcal{K} f^{(n)} = \sum_{j=1}^{n} \langle h, e_j \rangle e_j \to h$ as $n \to \infty$. Furthermore, $f^{(n)}$ converges in $L^2(V, \mu_{\mathfrak{m}})$ to some element $k \in R_\mu$ since we have the bound

$$\|f^{(n)} - f^{(m)}\|_{L^2(\mu_{\mathfrak{m}})}^2 = \sum_{j=n,\, \lambda_j \neq 0}^{m} \frac{\langle h, e_j \rangle^2}{\lambda_j} \leq \sum_{j=n,\, \lambda_j \neq 0}^{\infty} \frac{\langle h, e_j \rangle^2}{\lambda_j},$$

for all $n < m$. Consequently we obtain that $h = \mathcal{C}k = \lim_n \mathcal{K}(f^{(n)})$, from which (7) follows by definition. ■

Remark 19 *Note that the expression for the variance norm is* $\|h\|_{\mu\text{-}cov} = \sqrt{\langle h, \mathcal{K}^{-1}h\rangle} = \|\mathcal{K}^{-\frac{1}{2}}h\|$*, analogous to the finite-dimensional* $\mathbb{R}^d$ *case.*

In the functional data analysis literature Berrendero et al. (2020) argued that the naive functional Mahalanobis distance $\|\mathcal{K}^{-\frac{1}{2}}h\|$ fails to be defined due to the non-invertibility of the square root operator $\mathcal{K}^{\frac{1}{2}}$ in the $L^2[0,1]$ setting. However, when viewed through the lens of variance norms and Cameron-Martin spaces as per Theorem 18, we can see how such a notion can still be made precise despite the difficulties present in the infinite-dimensional and singular settings by allowing the anomaly distance to be infinite if the covariance structure of the underlying distribution does not match the new samples. This point of view was for instance taken by Shao et al. (2023) for their conformance score anomaly distance in the finite-dimensional setting.

3.2 Regularized Variance Norms

A classical result from Gaussian probability theory states that $\mu(H_\mu) = 0$ whenever μ is a Gaussian measure and $\dim(H_\mu) = \infty$ (see e.g. Bogachev, 2015, Theorem 2.4.7). This means that the sample outcomes of a V-valued random variable will almost surely not lie in the Cameron-Martin space H_μ, making the variance norm infinite with probability one. This issue was addressed in the functional data analysis literature in the special case $V = L^2[0,1]$ by regularizing the functional Mahalanobis distance, under the assumptions of a continuous covariance function and an injective covariance operator (Berrendero et al., 2020). Using our framework we are able to extend these results to the general Hilbert space setting without these restrictive assumptions.

There are two equivalent viewpoints for how to obtain said regularization. First, recall by Theorem 18 that the μ-variance norm of $x \in V$ is given by $\|x\|_{\mu\text{-cov}} = \|\mathcal{K}^{-\frac{1}{2}}x\|$ if $x \in \mathrm{Im}(\mathcal{K}^{\frac{1}{2}})$, and infinity otherwise. The first definition of a regularized norm is obtained by replacing the inverse $\mathcal{K}^{-\frac{1}{2}}$ with the *Tikhonov regularized operator* $R_\alpha = (\mathcal{K} + \alpha I)^{-1}\mathcal{K}^{\frac{1}{2}}$ with smoothing parameter $\alpha > 0$. Tikhonov regularization is a classical tool used in statistics (e.g. ridge regression) and functional analysis to deal with ill-posed equations (see e.g. Kress, 2013). In contrast to the inverse $\mathcal{K}^{-\frac{1}{2}}$, the Tikhonov operator R_α is well-defined on all of V and is an approximation of the pseudo-inverse of $\mathcal{K}^{\frac{1}{2}}$.

Definition 20 *Let* V *be a Hilbert space, and let* $\mu \in \mathcal{M}_V$*. We define the* α*-regularized* μ*-variance norm with smoothing parameter* $\alpha > 0$ *as*

$$\|x\|_{\mu,\alpha} := \left\|(\mathcal{K} + \alpha I)^{-1}\mathcal{K}^{\frac{1}{2}}x\right\|, \tag{8}$$

for $x \in V$*, where* $\mathcal{K}$ *is the covariance operator of* μ*.*

The alternative definition, following Berrendero et al. (2020), is based on the idea to approximate each $x \in V$ by an element $x_\alpha \in H_\mu$, $\alpha > 0$, and then take the μ-variance norm

of x_α, which is finite by construction. Since no closest element in H_μ to x exists in the infinite-dimensional case (since H_μ might not closed in V), x_α is chosen by minimizing

$$x_\alpha := \operatorname*{argmin}_{h \in H_\mu} \|x - h\|^2 + \alpha \|h\|^2_{H_\mu}. \tag{9}$$

Because H_μ is a reproducing kernel Hilbert space as discussed in the remark proceeding Theorem 10, it follows from Cucker and Zhou (2007, Theorem 8.4) that the unique solution x_α to (9) is given by

$$x_\alpha = (\mathcal{K} + \alpha I)^{-1} \mathcal{K} x = \sum_{n=1}^{\infty} \frac{\lambda_n}{\lambda_n + \alpha} \langle e_n, x \rangle e_n,$$

where $(e_n)_{n=1}^\infty$ and $(\lambda_n)_{n=1}^\infty$ are the eigenvectors and eigenvalues of $\mathcal{K}$. Moreover, Theorem 18 implies that the squared μ-variance norm of x_α is

$$\|x_\alpha\|^2_{\mu\text{-cov}} = \|\mathcal{K}^{-\frac{1}{2}} x_\alpha\|^2 = \|(\mathcal{K} + \alpha I)^{-1} \mathcal{K}^{\frac{1}{2}} x\|^2 = \sum_{n=1}^{\infty} \frac{\lambda_n}{(\lambda_n + \alpha)^2} \langle e_n, x \rangle^2, \tag{10}$$

coinciding with the Tikhonov perspective of Definition 20. Note that the element x_α depends not only on α but also on $\mathcal{K}$, which depends on μ. We will use the notation $\|x\|_{\mu,\alpha}$, rather than $\|x_\alpha\|_\mu$, which better highlights this dependence. This is important when working with empirical variance norms $\|x\|_{\mu^N,\alpha}$ based on a finite sample of data drawn from μ.

3.3 Computing Variance Norms and Kernelization

For most machine learning applications, the underlying probability measure μ is not explicitly known, and we must base our models on finite samples assumed to be drawn from μ. A natural estimator of the underlying distribution is the empirical measure $\mu^N = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}$. In this subsection, we derive computational formulas for the variance norm with respect to μ^N, and show how it is directly related to kernelization via Reproducing Kernel Hilbert Spaces (RKHS).

In this subsection, we denote by μ^N the empirical measure of a sample, and $\mathcal{K}_N$ the covariance operator of μ^N, which we call the empirical covariance operator. The following result follows directly from Proposition 16 and Theorem 18 given the fact that $\mathcal{K}_N = \frac{1}{N} \sum_{i=1}^N (x_i - \widehat{\mathfrak{m}}) \langle \cdot, x_i - \widehat{\mathfrak{m}} \rangle$ is a finite rank operator.

Proposition 21 *The Cameron-Martin space of μ^N is given by*

$$H_{\mu^N} = \operatorname{span}\{x_1 - \widehat{\mathfrak{m}}, \cdots, x_N - \widehat{\mathfrak{m}}\} = \operatorname{span}\{e_1, \cdots, e_M\},$$

where $\widehat{\mathfrak{m}} = \frac{1}{N} \sum_{i=1}^N x_i$ is the empirical mean of the data, and $e_1, \cdots, e_M$ are the eigenvectors of $\mathcal{K}_N$ with positive eigenvalues. The empirical α-regularized variance norm is given by

$$\|z\|^2_{\mu^N,\alpha} = \sum_{m=1}^{M} \frac{\lambda_m}{(\lambda_m + \alpha)^2} \langle z, e_m \rangle^2,$$

for $z \in V$.

The following theorem presents an algorithm for computing the eigenvalues and eigenvectors of the empirical covariance operator, based on an SVD decomposition of the inner product Gram matrix. This concept is closely related to the techniques used for kernel PCA (Schölkopf et al., 1998). We however give our own original proof, and relate the results back to the Cameron-Martin space of the empirical measure.

Theorem 22 *Let $A \in \mathbb{R}^{N\times N}$ be defined by $A_{i,j} = \langle f_i, f_j\rangle$, where $f_i = \frac{x_i - \widehat{\mathfrak{m}}}{\sqrt{N}}$, with SVD decomposition $A = U\Sigma U^T$. Let $v^{(n)}$ be the n-th column of U, and $\lambda_n = \Sigma_{n,n}$. Define $M = \max\{m \leq N : \lambda_m > 0\}$. Then the elements defined by*

$$e_n = \sum_{i=1}^{N} v_i^{(n)} f_i, \tag{11}$$

are orthogonal eigenvectors of $\mathcal{K}_N$ with eigenvalues λ_n and norms $\|e_n\| = \sqrt{\lambda_n}$ for $n \leq M$, and $e_n = 0$ for $M < n \leq N$. Moreover, we have that $\operatorname{span}\{e_1, \cdots, e_M\} = H_{\mu^N}$.

Proof To prove that e_m is an eigenvector of $\mathcal{K}_N$ with eigenvalue λ_m for $m \in \{1, \cdots, M\}$, observe that

$$\mathcal{K}_N\left(\sum_{i=1}^{N} v_i^{(m)} f_i\right) = \sum_{j=1}^{N} f_j \left\langle \sum_{i=1}^{N} v_i^{(m)} f_i, f_j \right\rangle = \sum_{j=1}^{N} f_j \sum_{i=1}^{N} v_i^{(m)} \langle f_i, f_j\rangle = \sum_{j=1}^{N} f_j \lambda_m v_j^{(m)},$$

where we used that $\sum_{i=1}^{N} v_i^{(m)} \langle f_i, f_j\rangle = (Av^{(m)})_j = \lambda_m v_j^{(m)}$.

Next, we verify that the vectors $\{e_1, ..., e_N\}$ are linearly independent. We see that

$$\begin{aligned}\langle e_n, e_m\rangle &= \left\langle \sum_{i=1}^{N} v_i^{(m)} f_i, \sum_{j=1}^{N} v_j^{(n)} f_j \right\rangle = \sum_{i=1}^{N}\sum_{j=1}^{N} v_i^{(m)} v_j^{(n)} \langle f_i, f_j\rangle \\ &= \langle v^{(n)}, Av^{(m)}\rangle_{\mathbb{R}^N} = \lambda_m \langle v^{(n)}, v^{(m)}\rangle_{\mathbb{R}^N},\end{aligned}$$

for $n, m \in \{1, \cdots, N\}$. When $n \neq m$ we have that $\langle v^{(n)}, v^{(m)}\rangle_{\mathbb{R}^N} = 0$, hence $\langle e_n, e_m\rangle = 0$. For $1 \leq m \leq M$ the element e_m is non-zero since $\langle e_m, e_m\rangle = \lambda_m \langle v^{(m)}, v^{(m)}\rangle_{\mathbb{R}^N} > 0$. On the other hand, if $M < n \leq N$ then $\lambda_n = 0$, hence $\langle e_n, e_n\rangle = 0$ and consequently $e_n = 0$.

Finally, we want to use Proposition 21 to conclude that $H_{\mu^N} = \operatorname{span}\{e_1, \cdots, e_M\}$. To this end, define $f = (f_1, \cdots, f_N)$ and $e = (e_1, \cdots, e_N)$ as column vectors. Using this notation, (11) can be written as $f = Ue \iff e = U^T f$, from which it follows that $\operatorname{span}\{f_1, \cdots, f_N\} = \operatorname{span}\{e_1, \cdots, e_N\} = \operatorname{span}\{e_1, \cdots, e_M\}$. This concludes the proof. ■

Theorem 18 implies that the variance norm depends only on the choice of inner product on V, as well as the eigenvectors of the covariance operator. For empirical measures, Theorem 22 provides an algorithm for computing these based on a finite sample of data, given an inner product. For applications, the special case where V is an RKHS (Cucker and Smale, 2001; Schölkopf, 2009) is of great interest, which we briefly introduce below before we present our final algorithm for computing variance norms.

Definition 23 *Let $\mathcal{X}$ be a set, and let $\mathcal{H}$ be a Hilbert space of functions of $\mathcal{X}$. A reproducing kernel is defined as a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying*

(i) $\forall x \in \mathcal{X},\ k(\cdot, x) \in \mathcal{H}$,

(ii) $\forall x \in \mathcal{X},\ \forall f \in \mathcal{H},\ f(x) = \big\langle f, k(\cdot, x)\big\rangle_{\mathcal{H}}$.

Furthermore, $\mathcal{H}$ is said to be a reproducing kernel Hilbert space (RKHS) if there exists a reproducing kernel for $\mathcal{H}$.

From a machine learning perspective, it is often helpful to view RKHSs through the lens of *feature maps*. A feature map is defined as a function $\phi : \mathcal{X} \to \mathcal{F}$, where $\mathcal{F}$ is a Hilbert space. Every such feature map induces a positive definite kernel via $k(x, y) := \langle \phi(x), \phi(y)\rangle_{\mathcal{F}}$. Conversely, given a RKHS with reproducing kernel k, the canonical feature map $\phi(x) := k(\cdot, x)$ reproduces k in the sense that $\langle \phi(x), \phi(y)\rangle_{\mathcal{H}} = \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}} = k(x, y)$. A kernelized variance norm is obtained by lifting an initial measure μ using a feature map ϕ via $\nu := \mu \circ \phi^{-1}$. The ν-variance norm is then computed in the RKHS associated with ϕ. This formulation is advantageous for applications because it allows for the use of *kernel tricks*: inner products in $\mathcal{F}$ can be computed via kernel evaluations $k(x, y)$, even when the explicit form of $\phi(x)$ is unavailable or infinite-dimensional.

The generality of Theorem 22, which only assumes that V is a Hilbert space, allows us to apply it directly in the kernelized setting. Let $x_1, \ldots, x_N \subset \mathcal{X}$ be a dataset and $\phi : \mathcal{X} \to \mathcal{F}$ a feature map into a Hilbert space $\mathcal{F}$. We define the empirical measure $\mu^N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\phi(x_i)}$ and let V be the RKHS induced by ϕ. In this setting, the Gram matrix of inner products becomes the kernel Gram matrix. This gives a solid theoretical foundation for the kernelized Mahalanobis distance (Ruiz and López-de Teruel, 2001) within our unified framework. Note that the non-kernelized (linear) setting is recovered by taking $\phi = I$, the identity map, and we refer to this case as the *linear kernel*.

Because the feature map ϕ may be non-linear and possibly infinite-dimensional, direct computation of inner products between normalized elements (e.g., $\langle f_n, f_m\rangle$) may not be feasible. To address this, we express these inner products as linear combinations of kernel evaluations $\langle x_i, x_j\rangle = k(x_i, x_j)$. Specifically, let N, f_i, M, $v_i^{(m)}$, e_m, and λ_m be as defined in Theorem 22. Then by bilinearity of the inner product:

$$\langle f_m, f_n\rangle = \frac{1}{N}\left(\langle x_m, x_n\rangle - \frac{1}{N}\sum_{j=1}^{N}\langle x_m, x_j\rangle - \frac{1}{N}\sum_{j=1}^{N}\langle x_n, x_j\rangle + \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\langle x_i, x_j\rangle\right), \tag{12}$$

and for any test point h:

$$\left\langle \frac{e_m}{\sqrt{\lambda_m}}, h\right\rangle = \sum_{i=1}^{N} \frac{v_i^{(m)}}{\sqrt{\lambda_m N}}\left(\langle x_i, h\rangle - \frac{1}{N}\sum_{j=1}^{N}\langle x_j, h\rangle\right). \tag{13}$$

Furthermore, we use a simple dynamic programming procedure to avoid a naive $\mathcal{O}(N^4)$ and $\mathcal{O}(N^2)$ time complexity when computing (12) and (13), respectively. The full procedure for computing the kernelized Mahalanobis distance and its nearest-neighbour variant (conformance score) is detailed in Algorithms 1 to 3. The fitting procedure described in Algorithm 1 has time complexity $\mathcal{O}(N^2(K+N))$, where K is the time complexity of a single kernel evaluation. For inference, detailed in Algorithms 2 and 3, both the kernelized Mahalanobis distance and conformance score can be computed in $\mathcal{O}(N(K+M))$ time, where $M \leq N$ is the number of eigenvalues.

Algorithm 1: Kernelized Gram matrix w.r.t. $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{\phi(x_i)}$.

Input: Data $\{x_1, \cdots, x_N\}$.

// Compute normalized Gram matrix via kernel trick

$B_{i,j} \leftarrow \langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j)$ for $i, j \in \{1, \cdots, N\}$

Column mean $a_i \leftarrow \frac{1}{N}\sum_{j=1}^N B_{i,j}$ for $i \in \{1, \cdots, N\}$

Matrix mean $b \leftarrow \frac{1}{N}\sum_{i=1}^N a_i$

$A_{i,j} \leftarrow \frac{1}{N}(B_{i,j} - a_i - a_j + b)$ for $i, j \in \{1, \cdots, N\}$ // $A_{i,j} = \langle f_i, f_j \rangle$

// Compute inner products of eigenvectors and data

Compute SVD decomposition $U\Sigma U^t = A$

Set $M \leftarrow \max\{m \leq N : \Sigma_{m,m} > \lambda\}$

for $n = 1$ **to** N **do**

  for $m = 1$ **to** M **do**

    $E_{n,m} \leftarrow \sum_{i=1}^N \frac{U_{i,m}}{\sqrt{N\Sigma_{m,m}}}(B_{i,n} - a_n)$ // $E_{n,m} = \langle \frac{e_m}{\sqrt{\lambda_m}}, \phi(x_n) \rangle$

Output: Matrix E, and SVD decomposition $A = U\Sigma U^t$.

Algorithm 2: Kernelized Mahalanobis distance w.r.t. $\mu = \frac{1}{N}\sum_{i=1}^N \delta_{\phi(x_i)}$.

Input: Data $\{x_1, \cdots, x_N\}$. SVD decomposition matrices $U, \Sigma \in \mathbb{R}^{N\times N}$ and $E \in \mathbb{R}^{N\times M}$ as per Algorithm 1. Regularization $\alpha > 0$. A new sample y.

// Compute inner product of eigenvectors and sample

Use kernel trick $s_i \leftarrow \langle \phi(y), \phi(x_i) \rangle = k(y, x_i)$ for $i \in \{1, \cdots, N\}$

Average $r \leftarrow \frac{1}{N}\sum_{i=1}^N s_i$

$p_m \leftarrow \frac{1}{\sqrt{N\Sigma_{m,m}}}\sum_{i=1}^N U_{i,m}(s_i - r)$ for $m \in \{1, \cdots, M\}$ // $p_m = \langle \frac{e_m}{\sqrt{\lambda_m}}, \phi(y) \rangle$

// Calculate Mahalanobis distance

Average $c_m \leftarrow \frac{1}{N}\sum_{i=1}^N E_{i,m}$ for $m \in \{1, \cdots, M\}$ // $c_m = \langle \frac{e_m}{\sqrt{\lambda_m}}, \frac{1}{N}\sum_{n=1}^N \phi(x_n) \rangle$

$d \leftarrow \sqrt{\sum_{m=1}^M \frac{\Sigma_{m,m}}{(\Sigma_{m,m}+\alpha)^2}(p_m - c_m)^2}$ // $d = \|\phi(y) - \frac{1}{N}\sum_{n=1}^N \phi(x_n)\|^2_{\mu^N,\alpha}$

Output: Kernelized Mahalanobis distance d with α-regularization.

Algorithm 3: Kernelized conformance score w.r.t. $\mu = \frac{1}{N}\sum_{i=1}^N \delta_{\phi(x_i)}$.

Input: Data $\{x_1, \cdots, x_N\}$. SVD decomposition matrices $U, \Sigma \in \mathbb{R}^{N\times N}$ and $E \in \mathbb{R}^{N\times M}$ as per Algorithm 1. Regularization $\alpha > 0$. A new sample y.

// Compute inner product of eigenvectors and sample

Use kernel trick $s_i \leftarrow \langle \phi(y), \phi(x_i) \rangle = k(y, x_i)$ for $i \in \{1, \cdots, N\}$

Average $r \leftarrow \frac{1}{N}\sum_{i=1}^N s_i$

$p_m \leftarrow \frac{1}{\sqrt{N\Sigma_{m,m}}}\sum_{i=1}^N U_{i,m}(s_i - r)$ for $m \in \{1, \cdots, M\}$ // $p_m = \langle \frac{e_m}{\sqrt{\lambda_m}}, \phi(y) \rangle$

// Calculate nearest-neighbour Mahalanobis distance

$d_n \leftarrow \sum_{m=1}^M \frac{\Sigma_{m,m}}{(\Sigma_{m,m}+\alpha)^2}(p_m - E_{n,m})^2$ for $n \in \{1, \cdots, N\}$ // $d_n = \|\phi(y) - \phi(x_n)\|^2_{\mu^N,\alpha}$

$c \leftarrow \sqrt{\min_n d_n}$

Output: Kernelized conformance score c with α-regularization.

3.4 Consistency, Speed of Convergence, and Distributional Results

We now turn to the statistical properties of the sample estimator $\|x\|_{\mu^N,\alpha}$. Similar results were obtained by Berrendero et al. (2020) in the special case $V = L^2[0,1]$ under the assumptions that $\mathcal{K}$ is injective. Most of their proofs easily generalize to our setting which we instead base on variance norms of measures $\mu \in \mathcal{M}_V$ for general separable Hilbert spaces V. The consistency analysis and speed of convergence relies on the following lemma, which follows from standard properties of covariance operators on Hilbert spaces (see e.g. Tailen Hsing, 2015, Theorems 8.1.1 and 8.1.2).

Lemma 24 *Let $\mu \in \mathcal{M}_V$ and μ^N be an empirical measure of μ, with covariance operators $\mathcal{K}$ and $\mathcal{K}_N$, respectively. Then $\|\mathcal{K}_N - \mathcal{K}\|_{op} \to 0$ as $N \to \infty$ almost surely. Furthermore, if μ has finite fourth moment, then $\|\mathcal{K}_N - \mathcal{K}\|_{op} = O_P(N^{-\frac{1}{2}})$.*

The following theorem proves that the empirical regularized variance norm converges to the actual regularized variance norm almost surely. Furthermore, if μ has finite fourth moment, we obtain a speed of convergence of $N^{-\frac{1}{4}}$.

Theorem 25 *Let $x \in V$ and $\mu \in \mathcal{M}_V$. Then the empirical regularized μ-variance norm is consistent almost surely, that is,*

$$\|x\|_{\mu^N,\alpha} \to \|x\|_{\mu,\alpha},$$

as $N \to \infty$. Additionally, if μ has finite fourth moment, then the speed of convergence in probability is

$$\|x\|_{\mu^N,\alpha} - \|x\|_{\mu,\alpha} = O_P(N^{-\frac{1}{4}}).$$

Proof Fix $x \in V$ and $\alpha > 0$. For brevity we write $T_N^\alpha = (\mathcal{K}_N + \alpha I)^{-1}$ and $T^\alpha = (\mathcal{K} + \alpha I)^{-1}$. By the reverse triangle inequality we obtain that

$$\begin{aligned}\left| \|x\|_{\mu^N,\alpha} - \|x\|_{\mu,\alpha} \right| &= \left| \|T_N^\alpha \mathcal{K}_N^{\frac{1}{2}} x\| - \|T^\alpha \mathcal{K}^{\frac{1}{2}} x\| \right| \leq \left\| T_N^\alpha \mathcal{K}_N^{\frac{1}{2}} x - T^\alpha \mathcal{K}^{\frac{1}{2}} x \right\| \\ &\leq \|T_N^\alpha\|_{op} \|\mathcal{K}_N^{\frac{1}{2}} x - \mathcal{K}^{\frac{1}{2}} x\| + \|T_N^\alpha - T^\alpha\|_{op} \|\mathcal{K}^{\frac{1}{2}} x\|. \end{aligned} \tag{14}$$

First note that T_N^α and T^α are bounded by $\|T_N^\alpha\|_{op} \leq \frac{1}{\alpha}$. Lemma 24 implies that $\|\mathcal{K}_N - \mathcal{K}\|_{op} \to 0$ almost surely, from which it follows that the first term of (14) goes to 0 as $N \to \infty$. The second term also goes to 0, by Gohberg et al. (2012, Corollary 8.3), since $\mathcal{K} - \mathcal{K}_N = (\mathcal{K} + \alpha I) - (\mathcal{K}_N + \alpha I)$. This proves consistency.

As for the speed of convergence, Gohberg et al. (2012, Corollary 8.2) implies that

$$\|T_N^\alpha - T\|_{op} \leq \frac{\|T^\alpha\|_{op}^2 \|\mathcal{K}_N - \mathcal{K}\|_{op}}{1 - \|T^\alpha\|_{op} \|\mathcal{K}_N - \mathcal{K}\|_{op}}$$

which is of order $O(\|\mathcal{K}_N - \mathcal{K}\|_{op})$ as $N \to \infty$. Furthermore, since $\mathcal{K}_N$ and $\mathcal{K}$ are positive operators, we have that $\|\mathcal{K}_N^{\frac{1}{2}} - \mathcal{K}^{\frac{1}{2}}\|_{op} \leq \|\mathcal{K}_N - \mathcal{K}\|_{op}^{\frac{1}{2}}$ (see e.g. Bhatia, 1997, Theorem X.1.1). Combining this with Lemma 24 we obtain a speed of convergence in probability of $O_P(N^{-\frac{1}{4}})$. ■

When μ is a Gaussian measure, we are able to obtain an explicit distribution of the Mahalanobis distance as an infinite sum of independent chi-squared random variables. This generalizes the classical Hotelling's T-statistic and the Gaussian functional Mahalanobis distance case (Berrendero et al., 2020) to the general Hilbert space setting.

Theorem 26 *Let μ be a Gaussian measure on a Hilbert space V, and let $\alpha > 0$. If $X \sim \mu$ is drawn from the measure μ, then the squared α-regularized Mahalanobis distance has distribution*

$$\|X - \mathfrak{m}\|^2_{\mu,\alpha} \overset{d}{=} \sum_{n=1}^{\infty} \left(\frac{\lambda_n}{\lambda_n + \alpha}\right)^2 Y_n,$$

where $\mathfrak{m}$ is the mean of μ, λ_n are the eigenvalues of the covariance operator $\mathcal{K}$ of μ, and $Y_1, Y_2, \cdots$ is a sequence of i.i.d. standard χ^2_1 random variables.

Proof Let $(e_n)_{n=1}^{\infty}$ be an orthonormal sequence of eigenvectors of $\mathcal{K}$. It follows by (10) that

$$\|X - \mathfrak{m}\|^2_{\mu,\alpha} = \sum_{n=1}^{\infty} \frac{\lambda_n}{(\lambda_n + \alpha)^2} \langle e_n, X - \mathfrak{m}\rangle^2.$$

Since μ is a Gaussian measure, the vectors e_n are by definition Gaussian distributed when acting as continuous linear functionals on V. Moreover, we have that

$$\mathrm{E}\big[\langle e_n, X - \mathfrak{m}\rangle \langle e_m, X - \mathfrak{m}\rangle\big] = \langle e_n, \mathcal{K} e_m\rangle = \begin{cases} \lambda_n & \text{if } n = m, \\ 0 & \text{otherwise,} \end{cases}$$

and $\mathrm{E}\big[\langle e_n, X - \mathfrak{m}\rangle\big] = \langle e_n, \mathrm{E}[X]\rangle - \langle e_n, \mathfrak{m}\rangle = 0$, from which the result follows. ■

4. Nearest Neighbour Properties

As discussed in Section 2.3, for some applications it may be advantageous to measure outlier distances via nearest-neighbours rather than the Mahalanobis distance to the mean. In this section, we study some useful properties of the infinite-dimensional nearest- and furthest-neighbour μ-variance norm in the Hilbert space setting. Let $X_0, ..., X_N \sim \mu$ be i.i.d. samples. We are interested in studying the random empirical variance norm nearest-neighbour distance, defined as

$$\min_{1 \le i \le N} \|X_0 - X_i\|_{\mu^N,\alpha}. \tag{15}$$

We adopt a stepwise approach by analyzing three progressively more complex cases:

1. The reference point X_0 belongs to the corpus defining the empirical measure μ^N.
2. X_0 is out of corpus and the norm measured w.r.t. μ; this requires α-regularization.
3. X_0 is out of corpus and the norm measured w.r.t. the empirical measure μ^N.

4.1 Case 1: Reference point in corpus

We begin by considering deterministic sample points $\{x_1, ..., x_N\}$ defining an empirical measure μ^N, and simplify (15) by setting $x_0 = x_1$ and leaving out x_1 from the calculation of the minimum. In this case, one can work with the non-regularized variance norm, since x_0 will lie in the empirical Cameron-Martin space. In the high-dimensional setting, an interesting property emerges: if all $x_1, ..., x_N$ are linearly independent, then without regularization every pair of distinct points in the corpus is equidistant under the empirical variance norm, at distance exactly $\sqrt{2N}$. The use of the nearest-neighbour Mahalanobis distance is uninformative in this case. We formalize this in the following:

Proposition 27 *Let $\{x_1, ..., x_N\} \subset V$ be linearly independent, and let $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ be the empirical measure. Then for any $i \neq j$ we have*

$$\sqrt{2N}\left(\frac{\lambda_{N-1}}{\lambda_{N-1}+\alpha}\right) \leq \|x_j - x_i\|_{\mu^N,\alpha} \leq \sqrt{2N}\left(\frac{\lambda_1}{\lambda_1+\alpha}\right),$$

where λ_m, $m \geq 1$, are the eigenvalues of $\mathcal{K}_N$ in decreasing order. In particular, when $\alpha = 0$, we obtain that $\|x_j - x_i\|_{\mu^N} = \sqrt{2N}$.

Proof Recall from Proposition 21 that $\|\cdot\|^2_{\mu^N,\alpha} = \sum_{m=1}^M \frac{\lambda_m}{(\lambda_m+\alpha)^2}\langle \cdot, z_m\rangle^2$, where λ_m and z_m are the eigenvalues and (normalized) eigenvectors of the covariance operator $\mathcal{K}_N$ of μ^N. Theorem 22 says these are given by an SVD decomposition: Let $A \in \mathbb{R}^{N\times N}$ be defined by $A_{i,j} = \langle f_i, f_j\rangle$, where $f_i = \frac{x_i - \widehat{\mathfrak{m}}}{\sqrt{N}}$, with SVD decomposition $A = U\Sigma U^T$. Let $v^{(n)}$ be the n-th column of U, and $\lambda_n = \Sigma_{n,n}$. Define $M = \max\{m \leq N : \lambda_m > 0\}$, and $e_n = \sum_{i=1}^N v_i^{(n)} f_i$. The vectors $z_m = e_m/\sqrt{\lambda_m}$ are orthonormal eigenvectors of $\mathcal{K}_N$. First note for all j, m, that

$$\langle f_j, e_m\rangle = \left\langle f_j, \sum_{k=1}^N v_k^{(m)} f_k \right\rangle = \sum_{k=1}^N v_k^{(m)}\langle f_j, f_k\rangle = (Av^{(m)})_j = \lambda_m v_j^{(m)},$$

hence

$$\begin{aligned}\|x_j - x_i\|^2_{\mu^N,\alpha} = N\|f_j - f_i\|^2_{\mu^N,\alpha} &= N\sum_{m=1}^M \frac{\lambda_m}{(\lambda_m+\alpha)^2}\langle f_j - f_i, \frac{e_m}{\sqrt{\lambda_m}}\rangle^2 \\ &= N\sum_{m=1}^M \frac{\lambda_m^2}{(\lambda_m+\alpha)^2}\left(v_j^{(m)} - v_i^{(m)}\right)^2.\end{aligned}$$

Next, we have that $M = N - 1$ due to linear independence and mean-centering. Consequently, we have that $v^{(N)} = (\frac{1}{\sqrt{N}}, ..., \frac{1}{\sqrt{N}})$ which follows from the unit vector belonging to the null space of A. Thus

$$N\sum_{m=1}^M \left(v_j^{(m)} - v_i^{(m)}\right)^2 = N\left(\|v_j - v_i\|^2 - (v_j^{(N)} - v_i^{(N)})^2\right) = N(2-0) = 2N,$$

where v_j denotes the j-th row of U, and the conclusion follows from the monotonicity of $\lambda \mapsto \frac{\lambda}{\lambda+\alpha}$. ∎

In the finite-dimensional $V = \mathbb{R}^d$ case, Proposition 27 is only applicable when the sample size N satisfies $N \leq d$. However, when V is infinite-dimensional, for many relevant distributions, sample points will be linearly independent with probability one, and all corpus points will be equidistant without regularization. This suggests that the in-corpus case is ill-suited for studying the infinite-dimensional setting, and provides additional justification for using α-regularization. As such, we consider an alternative approach in the sequel.

4.2 Case 2: Reference point out of corpus, sub-Gaussian measure

We now consider the *out-of-corpus* case, where the random reference point X_0 and corpus points $X_1, ..., X_N$ are drawn identically distributed from a sub-Gaussian measure μ. Our goal is to analyse the concentration properties of the difference between the nearest- and furthest-neighbor distances in the infinite-dimensional setting. For ease of notation, we use the symbol $\lesssim$ to denote inequality up to a universal constant. We use the following infinite-dimensional definition of sub-Gaussianity, which is a special case of an *R-sub-Gaussian* random variable with respect to covariance operators (Antonini, 1997).

Definition 28 *A random variable $X \sim \mu \in \mathcal{M}_V$ with covariance operator $\mathcal{K}$ is said to be sub-Gaussian with respect to $\mathcal{K}$ if there exists a $\beta \geq 0$ such that for all $z \in V$*

$$\mathrm{E}\left[e^{\langle z, X - \mathrm{E}X\rangle}\right] \leq e^{\beta^2 \langle \mathcal{K}z, z\rangle}. \tag{16}$$

Moreover, the sub-Gaussian norm of X with respect to $\mathcal{K}$ is defined as the smallest constant $\beta \geq 0$ such that (16) *holds, denoted $\|X\|_{\psi_2,\mathcal{K}}$. X is said to be $\mathcal{K}$-Gaussian if* (16) *is an equality with $\beta = 1$.*

Our analysis relies on a Hilbert-space version of the classical Hanson-Wright inequality (Chen and Yang, 2021). To apply this result, we need to furthermore impose a mild Bernstein-like tail condition on the squared norm of our random variables, as follows:

Definition 29 *A sub-Gaussian random variable $X \sim \mu \in \mathcal{M}_V$ with covariance operator $\mathcal{K}$ and sub-Gaussian norm $\beta = \|\mu\|_{\psi_2,\mathcal{K}}$ is said to satisfy a Bernstein condition on the squared norm with respect to $\mathcal{K}$ if*

$$\mathrm{E}\left|\|X\|^2 - \mathrm{E}\|X\|^2\right|^k \lesssim k!\beta^{k-2}\|\mathcal{K}\|_{op}^{k-2}\|\mathcal{K}\|_{HS}^2 \tag{17}$$

for all $k \geq 3$.

The following definitions of effective rank $\mathbf{r}(\mathcal{K})$ and dimension $\mathbf{d}(\mathcal{K})$ for covariance operators $\mathcal{K}$, defined below, will play an important role in our analysis. The effective rank $\mathbf{r}(\mathcal{K})$ was previously used by Koltchinskii and Lounici (2017) in the infinite-dimensional setting, see also Vershynin (2012). Another well-studied measure of rank in matrix theory is the so-called *stable rank* of a matrix (see Ipsen and Saibaba (2025) and references therein). The effective dimension $\mathbf{d}(\mathcal{K})$ below can be obtained as the squared quotient of the stable rank and effective dimension, and has previously been used in the context of particle systems in physics under the name of *participation ratio* (Recanatesi et al., 2022; Kramer

and MacKinnon, 1993). These quantities naturally appear in our analysis as a byproduct of the Hanson-Wright inequality (Chen and Yang, 2021). Note that if X is a d-dimensional isotropic Gaussian with covariance matrix $\sigma^2 I_d$, then the effective dimension and rank is exactly d.

Definition 30 *Let $\mu \in \mathcal{M}_V$ with covariance operator $\mathcal{K}$. We define the effective dimension $\boldsymbol{d}(\mathcal{K})$ and effective rank $\boldsymbol{r}(\mathcal{K})$ by*

$$\boldsymbol{d}(\mathcal{K}) = \frac{\mathrm{Tr}(\mathcal{K})^2}{\|\mathcal{K}\|_{HS}^2} = \frac{(\sum_{i=1}^{\infty} \lambda_i)^2}{\sum_{i=1}^{\infty} \lambda_i^2}, \qquad \boldsymbol{r}(\mathcal{K}) = \frac{\mathrm{Tr}(\mathcal{K})}{\|\mathcal{K}\|_{op}} = \frac{\sum_{i=1}^{\infty} \lambda_i}{\lambda_1}$$

where λ_m are the eigenvalues of $\mathcal{K}$ in decreasing order.

We first consider the general Hilbert space case with norm $\|\cdot\|$, and then specialize to the μ-variance norm via a transformation. The following result shows that with probability $1-\delta$, the relative difference between the furthest- and nearest-neighbour distance is bounded above by a term proportional to $\sqrt{\log(N/\delta)}$, divided by the effective rank or dimension of the underlying data. Therefore, if $\log(N/\delta) \lesssim \mathbf{r}(\mathcal{K})$ and $\log(N/\delta) \lesssim \mathbf{d}(\mathcal{K})$, then the difference between the furthest- and nearest-neighbour distance is small. This result provides a theoretical justification for why nearest-neighbour methods can remain effective in infinite-dimensional settings where one might expect random points to become nearly equidistant: The concentration phenomenon is not governed by the ambient infinite dimension of V, but rather by the effective dimensionality of the covariance operator $\mathcal{K}$.

Proposition 31 *Let $\delta \in (0,1)$. If $X_0, X_1, ..., X_N$ are drawn i.i.d. from a centered sub-Gaussian measure μ satisfying the Bernstein condition* (17)*, then with probability $1-\delta$*

$$\frac{\max_{1\le i\le N} \|X_0 - X_i\|^2 - \min_{1\le i\le N} \|X_0 - X_i\|^2}{\mathrm{E}\|X_0 - X_1\|^2} \lesssim \beta^2 \epsilon(N, \delta, \mathcal{K})$$

where

$$\epsilon(N, \delta, \mathcal{K}) = \max\left\{ \sqrt{\frac{\log(2N/\delta)}{\boldsymbol{d}(\mathcal{K})}},\ \frac{\log(2N/\delta)}{\boldsymbol{r}(\mathcal{K})} \right\}.$$

Proof Let $D_i^2 = \|X_0 - X_i\|^2$. First note that $\mathrm{E}[D_i^2] = \mathrm{E}\left(\|X_0\|^2 + \|X_i\|^2 - 2\langle X_0, X_i\rangle\right) = 2\,\mathrm{Tr}(\mathcal{K})$. By the Hanson-Wright inequality in Hilbert spaces (Chen and Yang, 2021, Theorem 2.8), there exists a universal constant $C > 0$ such that for any $t > 0$

$$P\left(\left|D_i^2 - \mathrm{E}[D_i^2]\right| \ge t\right) \le 2\exp\left(-C\min\left\{\frac{t^2}{\beta^4\|\mathcal{K}\|_{HS}^2}, \frac{t}{\beta^2\|\mathcal{K}\|_{op}}\right\}\right).$$

We want to use a union bound to obtain a concentration result. Let $\mathcal{A}_i = \{|D_i^2 - \mathrm{E}[D_i^2]| \ge t\}$. Consider the inequality

$$P(\bigcup_{i=1}^{N} \mathcal{A}_i) \le N P(\mathcal{A}_1) \le N 2\exp\left(-C\min\left\{\frac{t^2}{\beta^4\|\mathcal{K}\|_{HS}^2}, \frac{t}{\beta^2\|\mathcal{K}\|_{op}}\right\}\right) \le \delta.$$

We want this probability to be at most δ, so by rearranging the terms we obtain this happens when

$$t \gtrsim \beta^2 \|\mathcal{K}\|_{HS} \sqrt{\log \frac{2N}{\delta}}, \qquad \text{and} \qquad t \gtrsim \beta^2 \|\mathcal{K}\|_{op} \log \frac{2N}{\delta}$$

This gives, with probability $1-\delta$ and for all $1 \leq i \leq N$, that

$$|D_i^2 - \mathrm{E}[D_i^2]| \lesssim \beta^2 \max\left\{ \|\mathcal{K}\|_{HS} \sqrt{\log \frac{2N}{\delta}},\ \|\mathcal{K}\|_{op} \log \frac{2N}{\delta} \right\}.$$

The result then immediately follows by considering the furthest-neighbour distance minus the nearest-neighbour distance, divided by the relative expected size $E[D_i^2] = 2\,\mathrm{Tr}(\mathcal{K})$. ■

For the Tikhonov-regularized variance norm, we have the identity $\|z\|_{\mu,\alpha} = \|\mathcal{S}_\alpha z\|$ for all $z \in V$, where $\mathcal{S}_\alpha = (\mathcal{K}+\alpha I)^{-1}\sqrt{\mathcal{K}}$ (see Section 3.2). This allows us to reframe the problem: instead of considering a random variable $X \sim \mu$ with a regularized norm, we can equivalently study the transformed variable $Y = \mathcal{S}_\alpha X$ with the standard Hilbert space norm. The distribution of Y is given by the pushforward measure $\mu_\alpha = \mu \circ \mathcal{S}_\alpha^{-1}$, and Y is $\mathcal{K}_\alpha$-sub-Gaussian with respect to its own covariance operator $\mathcal{K}_\alpha$ if X is $\mathcal{K}$-sub-Gaussian (simply apply (16) to the transformed measure). The covariance operator $\mathcal{K}_\alpha$ is given by

$$\mathcal{K}_\alpha = \mathcal{S}_\alpha \mathcal{K} \mathcal{S}_\alpha^* = (\mathcal{K}+\alpha I)^{-2}\mathcal{K}^2,$$

which has the spectral decomposition $\mathcal{K}_\alpha = \sum_{i=1}^\infty \left(\frac{\lambda_i}{\lambda_i+\alpha}\right)^2 \langle e_i, \cdot \rangle$. The effective dimension and rank of this operator are then

$$\mathbf{d}(\mathcal{K}_\alpha) = \frac{\mathrm{Tr}(\mathcal{K}_\alpha)^2}{\|\mathcal{K}_\alpha\|_{HS}^2} = \frac{\left(\sum_{i=1}^\infty \left(\frac{\lambda_i}{\lambda_i+\alpha}\right)^2\right)^2}{\sum_{i=1}^\infty \left(\frac{\lambda_i}{\lambda_i+\alpha}\right)^4}, \qquad \mathbf{r}(\mathcal{K}_\alpha) = \frac{\mathrm{Tr}(\mathcal{K}_\alpha)}{\|\mathcal{K}_\alpha\|_{op}} = \frac{\sum_{i=1}^\infty \left(\frac{\lambda_i}{\lambda_i+\alpha}\right)^2}{\left(\frac{\lambda_1}{\lambda_1+\alpha}\right)^2}.$$

Corollary 32 *Let $\alpha, \delta > 0$. If $X_0, X_1, ..., X_N$ are drawn identically distributed from a centered sub-gaussian measure μ, and if μ_α satisfies the Bernstein condition* (17) *with sub-Gaussian constant β_α, then with probability $1-\delta$*

$$\frac{\max_{1\leq i\leq N} \|X_0 - X_i\|_{\mu,\alpha}^2 - \min_{1\leq i\leq N} \|X_0 - X_i\|_{\mu,\alpha}^2}{\mathrm{E}\|X_0 - X_1\|_{\mu,\alpha}^2} \lesssim \beta_\alpha^2 \epsilon(N, \delta, \mathcal{K}_\alpha).$$

Remark 33 *If μ is $\mathcal{K}$-Gaussian, then μ_α is $\mathcal{K}_\alpha$-Gaussian, and hence both satisfy the Bernstein conditions, and $\beta = \beta_\alpha = 1$ (see e.g. Chen and Yang, 2021).*

4.3 Case 3: Reference point out of corpus, empirical measure

We now proceed to study the case where we take the variance norm with respect to the empirical measure μ^N. In our analysis, we will use the following concentration result from Koltchinskii and Lounici (2017, Theorem 9) applied to centered square integrable Hilbert space valued random variables.

Lemma 34 *[Koltchinskii and Lounici (2017)] Let $X, X_1, ..., X_N$ be i.i.d. square integrable centered random vectors with covariance operator $\mathcal{K}$. If X is $\mathcal{K}$-sub-Gaussian, then for all $t \geq 1$, with probability at least $1 - e^{-t}$,*

$$\|\mathcal{K} - \mathcal{K}_N\|_{op} \lesssim \|\mathcal{K}\|_{op} \max\left(\sqrt{\frac{\boldsymbol{r}(\mathcal{K})}{N}}, \frac{\boldsymbol{r}(\mathcal{K})}{N}, \sqrt{\frac{t}{N}}, \frac{t}{N}\right).$$

In the next proposition, we obtain a result analogous to Proposition 31, but with an additional error term due to the use of empirical measures. This error term is of order $\mathcal{O}\left(\sqrt{\frac{\log(N)}{N}}\right)$ as $N \to \infty$, assuming δ and α constant.

Proposition 35 *Let $\alpha > 0$, $\delta \in (0,1)$, and let $X_0, X_1, ..., X_N$ be drawn i.i.d. from a measure μ satisfying the assumptions of Proposition 31 and Corollary 32. Then with probability $1-\delta$ the empirical μ^N-variance norm satisfies*

$$\frac{\max_{1\leq i\leq N} \|X_0 - X_i\|^2_{\mu^N,\alpha} - \min_{1\leq i\leq N} \|X_0 - X_i\|^2_{\mu^N,\alpha}}{\mathrm{E}\|X_0 - X_1\|^2_{\mu,\alpha}} \lesssim \beta_\alpha^2 \epsilon(N, \frac{\delta}{3}, \mathcal{K}_\alpha) + \frac{\Delta(N, \frac{\delta}{3}, \mathcal{K}) \operatorname{Tr}(\mathcal{K})}{\alpha^2 \operatorname{Tr}(\mathcal{K}_\alpha)} \left(1 + \beta^2 \epsilon(N, \frac{\delta}{3}, \mathcal{K})\right),$$

where $\Delta(N, \delta, \mathcal{K}) = \|\mathcal{K}\|_{op} \max\left(\sqrt{\frac{\boldsymbol{r}(\mathcal{K})}{N}}, \frac{\boldsymbol{r}(\mathcal{K})}{N}, \sqrt{\frac{\log(1/\delta)}{N}}, \frac{\log(1/\delta)}{N}\right)$.

Proof Let $D_i^2 = \|X_0 - X_i\|^2_{\mu,\alpha}$ and $D_{i,N}^2 = \|X_0 - X_i\|^2_{\mu^N,\alpha}$. Consider the inequality

$$\max_{1\leq i\leq N} D_{i,N}^2 - \min_{1\leq i\leq N} D_{i,N}^2 \leq \left(\max_{1\leq i\leq N} D_i^2 - \min_{1\leq i\leq N} D_i^2\right) + 2 \max_{1\leq i\leq N} \left|D_{i,N}^2 - D_i^2\right|. \tag{18}$$

The first term of (18) can be bounded by Corollary 32 with probability $1 - \delta/3$:

$$\max_{1\leq i\leq N} D_i^2 - \min_{1\leq i\leq N} D_i^2 \lesssim \beta_\alpha^2 \epsilon(N, \delta/3, \mathcal{K}_\alpha) \operatorname{Tr}(\mathcal{K}_\alpha). \tag{19}$$

The second term of (18) captures the error from using the empirical covariance operator $\mathcal{K}_N$ instead of $\mathcal{K}$. Let $d_i = X_0 - X_i$. The squared norms are quadratic forms:

$$D_i^2 = \langle d_i, \mathcal{K}(\mathcal{K} + \alpha I)^{-2} d_i \rangle \quad \text{and} \quad D_{i,N}^2 = \langle d_i, \mathcal{K}_N(\mathcal{K}_N + \alpha I)^{-2} d_i \rangle.$$

Their difference is bounded by

$$\begin{aligned} |D_{i,N}^2 - D_i^2| &= \left|\left\langle d_i, \left(\mathcal{K}_N(\mathcal{K}_N + \alpha I)^{-2} - \mathcal{K}(\mathcal{K} + \alpha I)^{-2}\right) d_i \right\rangle\right| \\ &\leq \|d_i\|^2 \left\|\mathcal{K}_N(\mathcal{K}_N + \alpha I)^{-2} - \mathcal{K}(\mathcal{K} + \alpha I)^{-2}\right\| \\ &\leq \frac{3}{\alpha^2} \|d_i\|^2 \|\mathcal{K} - \mathcal{K}_N\|_{op}, \end{aligned}$$

where the last inequality is due to the following bound: We first use the resolvent identity for invertible squared operators, $A^{-2} - B^{-2} = A^{-2}(B - A)B^{-1} + A^{-1}(B - A)B^{-2}$, where $A = \mathcal{K}_N + \alpha I$ and $B = \mathcal{K} + \alpha I$, to write

$$\begin{aligned} &\mathcal{K}_N(\mathcal{K}_N + \alpha I)^{-2} - \mathcal{K}(\mathcal{K} + \alpha I)^{-2} \\ &\quad = \mathcal{K}_N\left((\mathcal{K}_N + \alpha I)^{-2} - (\mathcal{K} + \alpha I)^{-2}\right) + (\mathcal{K}_N - \mathcal{K})(\mathcal{K} + \alpha I)^{-2} \\ &\quad = \mathcal{K}_N(\mathcal{K}_N + \alpha I)^{-2}(\mathcal{K} - \mathcal{K}_N)(\mathcal{K} + \alpha I)^{-1} + \mathcal{K}_N(\mathcal{K}_N + \alpha I)^{-1}(\mathcal{K} - \mathcal{K}_N)(\mathcal{K} + \alpha I)^{-2} \\ &\qquad\qquad + (\mathcal{K}_N - \mathcal{K})(\mathcal{K} + \alpha I)^{-2}. \end{aligned}$$

Taking operator norms and using $\|(\mathcal{K}+\alpha I)^{-1}\|_{op} \le \frac{1}{\alpha}$ together with $\|\mathcal{K}(\mathcal{K}+\alpha I)^{-1}\|_{op} \le 1$, we obtain

$$\begin{aligned}
&\left\|\mathcal{K}_N(\mathcal{K}_N+\alpha I)^{-2} - \mathcal{K}(\mathcal{K}+\alpha I)^{-2}\right\|_{op} \\
&\le \|\mathcal{K}_N(\mathcal{K}_N+\alpha I)^{-2}\|_{op}\frac{1}{\alpha}\|\mathcal{K}-\mathcal{K}_N\|_{op} + \|\mathcal{K}_N(\mathcal{K}_N+\alpha I)^{-1}\|_{op}\frac{1}{\alpha^2}\|\mathcal{K}-\mathcal{K}_N\|_{op} + \frac{1}{\alpha^2}\|\mathcal{K}-\mathcal{K}_N\|_{op} \\
&\le \frac{1}{\alpha}\cdot\frac{1}{\alpha}\|\mathcal{K}-\mathcal{K}_N\|_{op} + 1\cdot\frac{1}{\alpha^2}\|\mathcal{K}-\mathcal{K}_N\|_{op} + \frac{1}{\alpha^2}\|\mathcal{K}-\mathcal{K}_N\|_{op} \\
&= \frac{3}{\alpha^2}\|\mathcal{K}-\mathcal{K}_N\|_{op}.
\end{aligned}$$

Thus, the second term of (18) is bounded by

$$\max_{1\le i\le N}|D_{i,N}^2 - D_i^2| \le \frac{3}{\alpha^2}\|\mathcal{K}-\mathcal{K}_N\|_{op}\max_{1\le i\le N}\|X_0 - X_i\|^2.$$

Next, we apply concentration inequalities to bound $\|\mathcal{K}-\mathcal{K}_N\|_{op}$ and $\max_i \|X_0 - X_i\|^2$. By Lemma 34, with probability at least $1-\delta/3$,

$$\|\mathcal{K}-\mathcal{K}_N\|_{op} \lesssim \Delta(N, \delta/3, \mathcal{K}).$$

For the maximum norm term, first recall that $\mathrm{E}\|X_0 - X_i\|^2 = 2\,\mathrm{Tr}(\mathcal{K})$. By the same argument as in Proposition 31 (Hanson-Wright inequality and a union bound over the N vectors $d_i = X_0 - X_i$), we have with probability at least $1-\delta/3$ that

$$\max_{1\le i\le N}\|X_0 - X_i\|^2 \lesssim 2\,\mathrm{Tr}(\mathcal{K}) + \beta^2\epsilon(N, \delta/3, \mathcal{K})\,\mathrm{Tr}(\mathcal{K}).$$

Combining these bounds via a union bound (total probability $1-\frac{2}{3}\delta$), we get

$$\max_{1\le i\le N}|D_{i,N}^2 - D_i^2| \lesssim \frac{\Delta(N, \frac{\delta}{3}, \mathcal{K})\,\mathrm{Tr}(\mathcal{K})}{\alpha^2}\left(1 + \beta^2\epsilon(N, \frac{\delta}{3}, \mathcal{K})\right). \tag{20}$$

Combining (19) and (20) in (18) with yet another union bound (total probability $1-\delta$) and dividing by $\mathrm{E}\|X_0 - X_1\|_{\mu,\alpha}^2$ yields the final result. ∎

5. Applications to Multivariate Time Series Novelty Detection

In this section, we apply the theory developed in Sections 2 to 3 to novelty detection of multivariate time series. Given a collection of non-anomalous multivariate time series, referred to as the normal corpus, we are presented with new time series samples that we want to classify as either belonging to the normal class or as outliers. We achieve this by defining an anomaly distance with respect to the corpus using either the Mahalanobis distance or the conformance score (see Definitions 14 and 15), assuming that our data originates from a suitable Hilbert space. For time series, a natural choice of Hilbert space is the classical space $V = (L^2[0,1])^d$, as well as letting V be the RKHS induced by our choice of kernel or feature map. The choice of V directly affects how we measure similarity

between elements in our normal corpus. For instance, for $V = L^2[0,1]$ two time series will be compared in a linear fashion by their $L^2[0,1]$ inner products. On the other hand, if V is for example the RKHS of the signature kernel (Salvi et al., 2021a), then the two time series are compared as *rough paths* (Lyons, 1998) in a non-linear fashion. Different choices of RKHS will lead to different ways to measure similarity between points, guided by the practitioner, by domain-specific knowledge or by empirical validation through techniques like cross-validation.

For a given normal corpus $\{x_1, \cdots, x_N\} \subset V$ we form either the standard empirical measure $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$, or the kernelized empirical measure $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{\phi(x_i)}$ where ϕ is a feature map corresponding to a positive definite kernel whose kernel trick is known. The empirical measure μ^N can be interpreted as an estimator of the underlying distribution of the normal corpus. Given a new sample y, we proceed by calculating either the Mahalanobis distance

$$d_M(y; \mu^N) = \|y - \widehat{\mathfrak{m}}\|_{\mu^N\text{-cov}},$$

where $\widehat{\mathfrak{m}} = \frac{1}{N}\sum_{i=1}^N x_i$ is the mean of the normal corpus, or the conformance score

$$d_C(y; \mu^N) = \min_{1 \le n \le N} \|y - x_n\|_{\mu^N\text{-cov}},$$

using Algorithms 1 to 3, which defines an anomaly distance to the normal corpus. We then classify each new sample y as an outlier or as belonging to the normal class based on a threshold $\gamma > 0$.

The threshold γ can be determined through various approaches. A data-driven method involves splitting the normal corpus into training and validation sets, then choosing γ as an empirical quantile of the anomaly distances in the validation set. Alternatively, a theoretical approach uses the distribution of the Mahalanobis distance as outlined in Theorem 26, assuming the data follows a Gaussian distribution. If a labelled subset of outliers is available, a supervised learning approach can be employed, using k-fold cross-validation to determine an optimal threshold; however, this requires access to a supervised data set of outliers. In our experiments, we evaluate each anomaly distance using Precision-Recall (PR) AUC and ROC-AUC metrics, which consider sensitivity across all positive thresholds, thus eliminating the need to select a fixed threshold explicitly.

Although semi-supervised anomaly detection using the nearest-neighbour Mahalanobis distance, as opposed to the classical Mahalanobis distance, has been successfully employed in the finite-dimensional $\mathbb{R}^d$ setting (Verdier and Ferreira, 2011; Sarmadi and Karamodin, 2020; Shao et al., 2023; Arrubarrena et al., 2024), to our knowledge no comprehensive comparison of these two anomaly distances has been carried out in the literature. In this section we carry out an extensive comparison of our newly introduced kernelized conformance score (including the non-kernelized linear case) against the kernelized Mahalanobis distance for the task of semi-supervised time series novelty detection, using the infinite-dimensional framework developed in the previous sections.

5.1 Time Series Kernels

We begin by giving a brief summary of the time series kernels considered in our experimentation. These consist of the linear kernel given by the $(L^2[0,1])^d$ inner product, a family of

generalized integral-class kernels related to linear time warping (Shimodaira et al., 2001), and a collection of time-dynamic state-of-the-art time series kernels including the global alignment kernel (Cuturi et al., 2007; Cuturi, 2011), the Volterra reservoir kernel (Gonon et al., 2022), and signature kernels (Kiraly and Oberhauser, 2019; Salvi et al., 2021a).

5.1.1 Static and Integral Class Kernels

Consider first the natural Hilbert space $(L^2[0,1])^d$, where the inner product is given by

$$\langle x, y\rangle = \int_{[0,1]} \langle x_t, y_t\rangle_{\mathbb{R}^d} dt, \tag{21}$$

for $x, y \in (L^2[0,1])^d$. If x and y are discretized on a regular time grid of size T, and consequently can be viewed as d-dimensional time series of length T, then the inner product (21) can simply be computed by flattening x and y into vectors in $\mathbb{R}^{Td}$, and then calculating their Euclidean dot product. Instead of using the Euclidean dot product, or in other words the linear kernel, one could replace this with any static kernel defined on $\mathbb{R}^{dT}$ as per Definition 36 below. Similarly, we can also replace the linear static kernel in (21) to obtain a class of integral-type kernels with respect to a static kernel. In our experimentation we consider both flattened and integral-type kernels. We give the following definitions:

Definition 36 *We define a static kernel on $\mathbb{R}^d$ as a positive definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Given such k, we define the time series integral class kernel of k to be the kernel*

$$K_k(x, y) = \int_{[0,1]} k(x_t, y_t) dt,$$

defined for d-dimensional time series x and y.

The positive definiteness of K_k follows trivially from that of k. By replacing the Euclidean dot product with a possibly non-linear static kernel, an algorithm may be able to take certain non-linearities of the data into account to increase classification accuracies. The integral type kernels can in fact be seen as a variant of linear time warping kernels, which were first introduced by Shimodaira et al. (2001). The static kernels we consider in this paper are:

(i) The linear kernel $k_{linear}(x, y) = \langle x, y\rangle$, which does not have any hyperparameters.

(ii) The polynomial kernel $k_{poly}(x, y) = (c + \langle x, y\rangle)^p$ with hyperparameters $c \in \mathbb{R}$ and $p \in \mathbb{Z}_+$.

(iii) The RBF kernel $k_{RBF}(x, y) = e^{-\frac{|x-y|^2}{2\sigma^2}}$ with hyperparameter $\sigma > 0$.

We thus have two distinct classes of time series kernels parametrized by static kernels: One is to consider flattened time series and static kernels in $\mathbb{R}^{Td}$, and the other is integral class kernels with respect to static kernels on $\mathbb{R}^d$. Note that the integral and static class kernels coincide when k is the linear kernel, and are distinct otherwise. These time series kernels can be computed in $\mathcal{O}(Td)$ time.

5.1.2 Global Alignment Kernel

The global alignment kernel (GAK) (Cuturi et al., 2007; Cuturi, 2011) is a dynamic-time kernel which is able to take non-linear time lags into account when measuring the similarity of two time series via dynamic time warping. While the classical dynamic time warping fails to define positive definite kernels due to failing to satisfy the triangle inequality, the global alignment kernel is able to overcome this by summing over all possible global alignments of the time series.

Definition 37 *Let $x = (x_1, \cdots, x_T)$ and $y = (y_1, \cdots, x_L)$ be two time series of length T and L respectively. An alignment π, denoted $\pi \in \mathcal{A}(T, L)$, is defined as a pair $\pi = (\pi_1, \pi_2)$ of vectors of length $p \leq T + L - 1$ such that $1 = \pi_1(1) \leq \cdots \leq \pi_1(p) = T$ and $1 = \pi_2(1) \leq \cdots \leq \pi_2(p) = L$. Given a similarity measure $\varphi : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$, the cost $D_{x,y}(\pi)$ is defined as*

$$D_{x,y}(\pi) := \sum_{i=1}^{|\pi|} \varphi(x_{\pi_1(i)}, y_{\pi_2(i)}),$$

and the global alignment kernel is defined as

$$K_{GA}(x, y) = \sum_{\pi \in \mathcal{A}(T,L)} e^{-D_{x,y}(\pi)} = \sum_{\pi \in \mathcal{A}(T,L)} \prod_{i=1}^{|\pi|} \kappa(x_{\pi_1(i)}, y_{\pi_2(i)}), \tag{22}$$

where $\kappa = e^{-\varphi}$ is the local similarity.

Cuturi et al. (2007) proved that K_{GA} defined via a local kernel κ is positive definite if $\frac{\kappa}{1+\kappa}$ is positive definite. A sufficient condition for this is for κ to be geometrically or infinitely divisible. In practice the local kernel $\kappa = \frac{k_{RBF}}{2-k_{RBF}}$ is often used, and due to the exponential nature of (22) the GAK kernel is always made to be normalized in feature space via $\frac{K_{GA}(x,y)}{\sqrt{K_{GA}(x,x)K_{GA}(y,y)}}$. The GAK kernel has a single hyperparameter $\sigma > 0$ inherited from the static RBF kernel, and $K_{GA}(x, y)$ can be computed in $\mathcal{O}(TLd)$ time using dynamic programming.

5.1.3 Volterra Reservoir Kernel

The Volterra Reservoir Kernel (VRK) (Gonon et al., 2022) is a universal dynamic kernel designed for sequences of arbitrary length. The kernel is built by constructing a state-space representation of the classical Volterra series expansions (Wiener, 1958; Sandberg, 1983; Boyd and Chua, 1985), a series representation for analytic maps between sequences. As discussed in detail in Gonon et al. (2022); Cuchiero et al. (2022) this idea is closely related to the principle of reservoir computing (Maass et al., 2002; Jaeger and Haas, 2004) and associated kernels (Grigoryeva and Ortega, 2021). The VRK kernel was recently shown to outperform the RBF, GAK, and signature kernels in a market forecasting task (Gonon et al., 2022).

For sequences of length T the VRK kernel with hyperparameters $\tau \in \mathbb{R}$ and $\lambda \in (0,1)$ is defined as

$$K^{\text{Volt}}(x,y) = 1 + \sum_{k=1}^{T} \lambda^{2k} \prod_{t=0}^{k-1} \frac{1}{1 - \tau^2 \langle x_{T-t}, y_{T-t} \rangle},$$

for time series x and y of equal length such that $\tau^2 \|x\| \|y\| < 1$, where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product on $\mathbb{R}^d$. The solution can be computed in $\mathcal{O}(Td)$ time using a recursive relation between the kernel at different time steps.

5.1.4 Signature Kernels

The signature kernel (Kiraly and Oberhauser, 2019; Salvi et al., 2021a) is a positive definite kernel for sequential data based on tools originating from stochastic analysis and rough path theory (Lyons, 1998). It has many desirable theoretical properties such as invariance to time-reparametrization, universality, and characteristicness on compact sets. Algorithms using signature kernels have successfully been applied to a wide variate of fields since their inception, for instance in Bayesian forecasting (Toth and Oberhauser, 2020), hypothesis testing (Salvi et al., 2021b), and for support vector machines (Salvi et al., 2021a; Tóth et al., 2025) achieveing state-of-the-art accuracies.

Below we give a very brief construction of the signature kernel, which is defined as an inner product in the extended tensor algebra via the so-called *signature transform*. For a detailed introduction to signature kernels we refer to the seminal papers by Kiraly and Oberhauser (2019); Salvi et al. (2021a), the review article by Lee and Oberhauser (2023), and the book by Cass and Salvi (2024).

Definition 38 *Let $\mathcal{H}$ be a Hilbert space. The m-fold iterated integral of a bounded variation path $x \in BV([0,1], \mathcal{H})$ is recursively defined as*

$$S_0(x) := 1, \qquad S_{m+1}(x) = \int_0^1 S_m(x) \otimes dx_t.$$

We define the signature transform as the map

$$\begin{aligned} S : BV([0,1], \mathcal{H}) &\to \prod_{m=0}^{\infty} \mathcal{H}^{\otimes m} \\ x &\mapsto \big(S_m(x)\big)_{m=0}^{\infty}, \end{aligned}$$

and similarly, we define the truncated signature as the map $S_{0:n}(x) := \big(S_m(x)\big)_{m=0}^{n}$. Here we use the convention that $\mathcal{H}^{\otimes 0} = \mathbb{R}$.

Given a path $x \in BV([0,1], \mathbb{R}^d)$ and a static kernel k on $\mathbb{R}^d$, we may canonically lift k to a path k_x taking values in its RKHS $\mathcal{H}$ via $t \mapsto k(x_t, \cdot) \in \mathcal{H}$ using the reproducing kernel property of k. If k is the linear kernel $k(x,y) = \langle x, y \rangle_{\mathbb{R}^d}$, then k_x is simply the original path x. However, if we choose k to be a non-linear kernel such as the RBF kernel, then k_x would genuinely be different to x, and in this particular case k_x would take values in an infinite-dimensional Hilbert space where direct computations of truncated signature features are

impossible. The main idea behind the signature kernel is to define the sequential kernel k^{sig} w.r.t. a static kernel k as the inner product of signature transforms $S(k_x)$ and $S(k_y)$ given two paths x and y.

Definition 39 *Let k be a static kernel on $\mathbb{R}^d$. We define the k-lifted signature kernel as the mapping $k^{sig} : BV([0,1],\mathbb{R}^d) \times BV([0,1],\mathbb{R}^d) \to \mathbb{R}$,*

$$k^{sig}(x,y) = \sum_{m=0}^{\infty} \left\langle S_m(k_x), S_m(k_y) \right\rangle_{\mathcal{H}^{\otimes m}},$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}^{\otimes m}}$ is the Hilbert-Schmidt inner product defined as $\langle a, b \rangle_{\mathcal{H}^{\otimes m}} = \sum_{i=1}^{n} \langle a_i, b_i \rangle_{\mathcal{H}}$ for elements $a = a_1 \otimes \cdots \otimes a_m$ and $b = b_1 \otimes \cdots \otimes b_m$. The truncated singature kernel $k^{sig}_{0:n}$ is defined similarly using the truncated signature.

There are currently three main algorithms for computing signature kernels, each of which come with their separate advantages and disadvantages. We list the methods below:

1. If $\dim \mathcal{H} = d < \infty$, then the truncated signature $S_{0:m}(x)$ can be computed exactly in $\mathcal{O}(Td^m)$ time, when treating the time series x as a piecewise linear path of length T. The truncated signature can then be computed by taking the inner product of the truncated signature. In practice this method is only applicable when k is the trivial linear kernel, and when d is very small (e.g. $d < 5$) due to the exponential time complexity.

2. The second method is due to Kiraly and Oberhauser (2019, Algorithms 3 and 6), and takes advantage of the kernel trick to compute $k^{sig}_{0:m}(x,y)$ via a Horner-type scheme. This can be computed in $\mathcal{O}(LTmd)$ time using a non-geometric approximation of $k^{sig}(x,y)$, or exactly in $\mathcal{O}(LT(md + m^3))$ time when viewing x and y as piecewise linear paths of lengths T and L with state-space $\mathbb{R}^d$.

3. The last method is due to Salvi et al. (2021a), who proved that the signature kernel $k^{sig}(x|_{[0,s]}, y|_{[0,t]})$ solves the Goursat PDE

$$k^{sig}(x|_{[0,s]}, y|_{[0,t]}) = 1 + \int_0^s \int_0^t k^{sig}(x|_{[0,u]}, y|_{[0,v]}) \langle dk_{x_u}, dk_{x_v} \rangle_{\mathbb{R}^d}, \tag{23}$$

   where $x|_{[0,s]}$ denotes the restriction of x to the interval $[0,s]$. Equation (23) can be solved for piecewise linear paths using numerical PDE methods in $\mathcal{O}(LTd)$ time (Salvi et al., 2021a), but is often much slower than the truncated approaches.

Generally the RBF-lifted signature kernel is preferred over the vanilla signature kernel. This is partly due to the latter having a tendency to blow up when the underlying time series are not properly normalized, something which is particularly pronounced for the PDE signature kernel which essentially acts as an inner product of tensor exponentials. In our experimentation we use the truncated signature kernel with the linear and RBF static kernels, as well as the RBF-lifted PDE signature kernel.

Another recently introduced variant of the signature and its signature kernel is the so-called *randomized signature* (Cuchiero et al., 2021), which we now define.

Definition 40 *Let $M \geq 1$ be an integer. Fix an initial condition $z_0 \in \mathbb{R}^M$, random matrices $A_1, \cdots, A_d \in \mathbb{R}^{M \times M}$, random biases $b_1, \cdots, b_d \in \mathbb{R}^M$ and an activation function σ. The randomized signature Z of $x \in BV([0,1], \mathbb{R}^d)$ is defined as the solution of the controlled differential equation (CDE)*

$$dZ_t = \sum_{i=1}^{d} \sigma(A_i Z_t + b_i) dx_t^{(i)}, \qquad Z_0 = z_0, \tag{24}$$

where $x^{(i)}$ denotes the i'th component of x. The randomized signature kernel is defined as the inner product of two randomized signatures.

The randomized signature was first constructed by Cuchiero et al. (2021) as a random projection of the signature, with an argument based on a non-trivial application of the Johnson-Lindenstrauss lemma. Randomized signatures have recently been successfully used for market anomaly detection (Akyildirim et al., 2022), graph conversion (Schäfl et al., 2023), optimal portfolio selection (Akyildirim et al., 2023; Cuchiero and Möller, 2025), generative time series modelling (Biagini et al., 2024), and for learning rough dynamical systems (Compagnoni et al., 2023). The CDE (24) has since been studied from the perspective of randomly initialized ResNets (Cirone et al., 2023, 2024), and path developments on compact Lie groups (Lou et al., 2023, 2024; Cass and Turner, 2024). In our experiments, we use Gaussian random matrices and biases, with tanh activation function.

5.2 Experiments

In this section we present an empirical study comparing the (potentially kernelized) Mahalanobis distance to the conformance score for semi-supervised multivariate time series novelty detection. Our primary objective is to validate Algorithms 1 to 3 presented in this paper for this infinite-dimensional setting. For comparisons of the finite-dimensional conformance score against other established methods like isolation forests, shapelets, and local outlier factors, we refer readers to Shao et al. (2023). Within the functional data analysis literature, the $(L^2[0,1])^d$ Mahalanobis distance has been evaluated against other common functional anomaly detection methodologies such as boxplots, outliergrams, and depth-based trimming (Arribas-Gil and Romo, 2014; Berrendero et al., 2020).

We will use UEA multivariate time series repository (Bagnall et al., 2018; Ruiz et al., 2021) in our experimentation, which in recent years has become a standard benchmark for multivariate time series classification. The repository contains 30 real world data sets consisting of multivariate time series, 26 of which are of equal lengths ranging from 8 to 2500 time steps, with state-space dimension ranging from 2 to 1345, see Table 1 for a summary. For the task of semi-supervised anomaly detection task we employ a one-versus-rest approach. In each experiment, we designate a single class label as the normal corpus, while considering all other classes as outliers. We evaluate the performance using both PR-AUC and ROC AUC. The results are then averaged across all class labels for a comprehensive assessment. Our experiment code is publicly available at `https://github.com/nikitazozoulenko/kernel-timeseries-anomaly-detection`.

Code	Name	Train size	Test size	Dims	Length	Classes	Avg. Corpus Size
AWR	ArticularyWordRecognition	275	300	9	144	25	11
AF	AtrialFibrillation	15	15	2	640	3	5
BM	BasicMotions	40	40	6	100	4	10
CR	Cricket	108	72	6	1197	12	9
DDG	DuckDuckGeese	50	50	1345	270	5	10
EW	EigenWorms	128	131	6	17,984	5	26
EP	Epilepsy	137	138	3	206	4	34
EC	EthanolConcentration	261	263	3	1751	4	65
ER	ERing	30	270	4	65	6	5
FD	FaceDetection	5890	3524	144	62	2	2945
FM	FingerMovements	316	100	28	50	2	158
HMD	HandMovementDirection	160	74	10	400	4	40
HW	Handwriting	150	850	3	152	26	6
HB	Heartbeat	204	205	61	405	2	102
LIB	Libras	180	180	2	45	15	12
LSST	LSST	2459	2466	6	36	14	176
MI	MotorImagery	278	100	64	3000	2	139
NATO	NATOPS	180	180	24	51	6	30
PD	PenDigits	7494	3498	2	8	10	749
PEMS	PEMS-SF	267	173	963	144	7	38
PS	PhonemeSpectra	3315	3353	11	217	39	85
RS	RacketSports	151	152	6	30	4	38
SRS1	SelfRegulationSCP1	268	293	6	896	2	134
SRS2	SelfRegulationSCP2	200	180	7	1152	2	100
SWJ	StandWalkJump	12	15	4	2500	3	4
UW	UWaveGestureLibrary	120	320	3	315	8	15

Table 1: Summary of the 26 equal length UEA multivariate time series data sets. In our empirical study we consider all data sets of total size less than 8000, where the average corpus size is greater than 30.

5.2.1 Experimental Setup

In total we will consider two different anomaly distances, namely the Mahalanobis distance and the conformance score, together with 11 different time series kernels. This includes the linear Euclidean kernel corresponding to the non-kernelized $(L^2[0,1])^d$ setting. In our open-source code we provide efficient PyTorch implementations of each kernel on both GPU and CPU, as well as an implementation of Algorithms 1 to 3 for computing the kernelized Mahalanobis distance and the kernelized conformance score. We consider the following time series kernels in our experimentation, all of which were defined in Section 5.1:

1. The family of time series kernels obtained by flattening a given time series of length T into a vector in $\mathbb{R}^{Td}$, and then applying a static kernel. We will use the RBF, polynomial, and linear kernels as our choices of static kernels, the latter of which corresponds to the $(L^2[0,1])^d$ inner product.
2. The family of integral-class kernels (linear time warping), with RBF and polynomial static kernels.
3. The global alignment kernel (GAK).
4. The Volterra reservoir kernel (VRK).
5. Four different variants of the signature kernel: The truncated signature kernel, the RBF-lifted truncated signature kernel, the RBF-lifted PDE signature kernel, and randomized signatures with tanh activation.

Our work adds to the growing body of literature on anomaly detection using signature features, which was first studied in Shao et al. (2023). This was done by explicitly computing m-level truncated signature features, which has $\mathcal{O}(Td^m)$ time complexity. Truncated signatures were later successfully used for market anomaly detection (Akyildirim et al., 2022), and radio astronomy (Arrubarrena et al., 2024). The use of signatures has however been limited to low-dimensional time series due to the exponential time complexity of explicitly computing truncated signatures. Our unified framework addresses this bottleneck, allowing for efficient computations of both signature conformance scores and signature Mahalanobis distances in $\mathcal{O}(T^2d)$ time. This significant improvement in d opens up the use of these methods for high-dimensional time series data.

5.2.2 Hyper-parameter Selection

For each kernel, data set, and class label, we run an extensive grid search on the designated training set using repeated k-fold cross-validation with 4 folds and 10 repeats to find the optimal kernel hyper-parameters. Let $\mathbb{R}^d$ be the state-space, and let T be the length of the time series for a given data set. For each method using the RBF static kernel we use the range $\sigma \in \frac{1}{\sqrt{d}}\{e^{-2}, e^{-1}, 1, e^1, e^2\}$, and similarly for the polynomial kernel we use $p \in \{2,3,4\}$, and $c \in \{\frac{1}{4}, \frac{1}{2}, 1, 2, 4\}$. For the GAK kernel we use the previously specified σ without the $\sqrt{d}$ term, multiplied by $\sqrt{T} \cdot \text{med}(\|x-y\|)$ as is recommended by Cuturi (2011). For the VRK kernel we use $\tau \in \frac{1}{\sqrt{d}}\{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1\}$, and we let λ vary from 0.25 to 0.999 on an inverse logarithmic grid of size 10.

The signature kernels inherit hyper-parameters from their respective static kernels. We additionally scale the kernel-lifted paths by $s \in \frac{1}{\sqrt{d}}\{\frac{1}{4}, \frac{1}{2}, 1, 2, 4\}$ for the truncated signature, and by $s \in \frac{1}{\sqrt{d}}\{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1\}$ for the untruncated PDE kernel. We use lower values for the PDE signature kernel since untruncated signatures essentially can be viewed as tensor exponentials, whose inner products will blow up if the input values are too big. For the truncated signature kernel we let the truncation level be in $\{1, 2, 3, 4, 5, 6, 7\}$. For the randomized signature we use the tanh activation function with number of features in $\{10, 25, 50, 100, 200\}$, and random matrix variances taken from a logarithmic grid of size 8 from 0.00001 to 1. Since the randomized signature is a randomized kernel, we perform the cross validation with 5 different random seeds for the random matrix initializations, and take the best performing model (using the training set only), as is common practice for randomized methods.

Furthermore, for each method we also cross-validate over the Tikhonov regularization parameter $\alpha \in \{10^{-8}, 10^{-5}, 10^{-2}\}$, whether to concatenate time as an additional dimension to each time series, and the eigenvalue threshold λ in Algorithm 1. We set an upper limit of 50 eigenvalues for the computation of the variance norm. For numerical stability, and to make the choices of α and λ be comparable across all kernels and data sets, we normalize all time series kernels K in feature space via $\frac{K(x,y)}{\sqrt{K(x,x)K(y,y)}}$.

5.2.3 Pre-processing

For each data set and each class label, we normalize the data to have mean zero and standard deviation one, using the statistics of the normal corpus. Average-pooling is then performed to reduce the maximum length of all time series to 100 time steps. After this, we concatenate the zero vector to each time series to allow each dynamic kernel to be translation-sensitive, and we clip all values to be in $[-5, 5]$ for additional numerical stability. Furthermore, in our cross-validation we also include the choice of adding time as an additional dimension to all time series. For the VRK kernel specifically, we perform further clipping of the data based on the τ hyper-parameter, which is required to make the VRK kernel well-defined.

5.2.4 Data and Results

Due to the high computational cost of evaluating 11 time series kernels on all 26 UEA data sets, with up to 40 experiments per data set-kernel combination, each of which goes through an extensive repeated k-fold cross-validation, we focus our analysis on UEA data sets with a total size under 8,000 entries (see Table 1). This excluded PenDigits and FaceDetection. Additionally, to ensure a sufficient statistical sample size, we only considered data sets where the average corpus size exceeded 30 entries, resulting in a final selection of 12 data sets.

The anomaly distances were computed as described by Algorithms 1 to 3. The optimal kernel hyper-parameters were obtained separately for the Mahalanobis distance and the conformance score, via a 10 times repeated 4-fold cross-validation on the training data for each data set. The objective score used in the cross-validation was the sum of ROC-AUC and PR-AUC. When calculating the precision-recall metric, we let the non-outlier class be the positive class. The final model was then evaluated on the out-of-sample test set to obtain the final results, presented in Table 2 and Table 3.

Data set		ROC AUC										
		linear	RBF	poly	I_{RBF}	I_{poly}	GAK	VRK	S_{lin}	S_{RBF}	S^{∞}_{RBF}	$S^{\text{rand}}_{\tanh}$
EP	C	.89	.95	.91	.94	.91	.97	.94	**.98**	**.98**	.92	.94
	M	.70	.81	.80	.80	.81	.88	.90	**.98**	.97	.91	.95
EC	C	.55	.56	.55	.55	.55	.56	.58	**.59**	.55	.56	.55
	M	.57	.56	.58	.55	.57	.57	**.60**	.56	.56	.56	.56
FM	C	.58	.52	.54	.54	.53	**.60**	.53	.49	.49	.48	.54
	M	**.58**	.51	.53	.50	.52	.54	.48	.49	.48	.51	.56
HMD	C	.55	.43	.54	.53	.49	.46	.52	.53	.48	.50	**.57**
	M	.45	.50	.47	.46	.51	**.55**	.54	.50	.50	.52	.52
HB	C	.63	.64	.60	.61	.61	.59	.67	.70	**.72**	.62	.61
	M	.61	.59	.61	.62	.59	.61	.62	**.69**	.60	.59	.60
LSST	C	.54	.61	.53	.61	.56	**.68**	.53	.57	.67	.62	.62
	M	.62	**.68**	.66	.67	.66	.67	.67	.67	.65	.63	.67
MI	C	.51	.54	.57	.57	.53	.50	.54	.43	.57	**.60**	.45
	M	.51	.52	.47	.46	.49	.50	**.54**	.47	.49	.43	.46
PEMS	C	.91	.92	.90	.91	.89	**.93**	.90	**.93**	.92	**.93**	.87
	M	.48	.69	.53	.66	.52	.77	**.90**	.80	.79	.71	.72
PS	C	.62	.65	.65	.64	.63	.66	.67	**.70**	.69	.56	.67
	M	.65	.67	.65	.65	.64	.65	.68	**.71**	.70	.54	.69
RS	C	.79	.73	.74	.80	**.81**	.77	.46	.73	.77	.68	.76
	M	.34	.58	.48	.60	.42	**.83**	.61	.79	.73	.75	.69
SRS1	C	.68	**.81**	.79	.80	**.81**	.77	**.81**	.61	.77	.71	.77
	M	.73	.60	.70	.59	.58	.62	.72	**.77**	**.77**	**.77**	.75
SRS2	C	**.57**	.51	.53	.53	.53	.54	.50	.49	.48	.53	.50
	M	.57	.55	.54	.55	**.59**	.55	.53	.54	.50	.50	.52
Avg. AUC	C	.65	.66	.65	**.67**	.65	**.67**	.64	.65	**.67**	.64	.66
	M	.57	.60	.58	.59	.58	.65	.65	**.66**	.65	.62	.64
Avg. Rank	C	11.0	10.7	10.9	10.0	11.8	**8.8**	9.9	10.4	**8.8**	12.4	11.1
	M	13.5	13.2	14.1	14.8	14.8	9.6	10.2	**8.8**	12.1	15.0	11.3

Table 2: One-versus-rest ROC-AUC for the semi-supervised anomaly detection experiments on the UEA multivariate time series repository. The conformance and Mahalanobis methods are denoted by C and M, respectively. The symbols I, S, S^{∞} and S^{rand} represent the integral, truncated signature, PDE signature, and randomized signature kernels, respectively.

5.2.5 Discussion

For the Mahalanobis distance, there seems to be a clear advantage to working in the kernelized setting, as the results show that the linear $(L^2[0,1])^d$ inner product achieves the lowest average test scores out of all methods, with ROC-AUC and PR-AUC scores of 0.57 and 0.39, respectively. The GAK, VRK and truncated signature kernels on the other hand perform best in this regard, obtaining AUC scores of 0.65-0.66 and 0.46-0.49, respectively.

The average test scores for the conformance score (nearest-neighbour Mahalanobis distance) do not differ much between the choices of kernels, but can have significant differences

Data set		Precision-Recall AUC										
		linear	RBF	poly	I_{RBF}	I_{poly}	GAK	VRK	S_{lin}	S_{RBF}	S^{∞}_{RBF}	$S^{\text{rand}}_{\text{tanh}}$
EP	C	.79	.88	.81	.87	.75	.92	.86	**.96**	**.96**	.80	.85
	M	.42	.58	.54	.57	.58	.73	.73	**.95**	.94	.78	.89
EC	C	.28	.28	.29	.28	.31	.29	**.32**	**.32**	.29	.30	.28
	M	.30	.32	.31	.30	.32	.32	**.33**	.28	.28	.30	.30
FM	C	.56	.52	.55	.57	.56	**.60**	.54	.54	.52	.53	.54
	M	**.55**	.51	.52	.50	.52	.52	.51	.52	.49	.53	**.55**
HMD	C	.30	.24	**.31**	.30	.28	.26	.27	.28	.29	.27	.29
	M	.25	.29	.26	.28	.27	.31	**.33**	.29	.29	**.33**	.27
HB	C	.58	.60	.53	.56	.55	.55	**.63**	**.63**	**.63**	.58	.60
	M	.58	.56	.57	.61	.55	.59	.61	**.64**	.55	.56	.58
LSST	C	.12	.12	.10	.15	.11	.14	.10	.12	**.19**	.14	.10
	M	.14	.15	.16	**.17**	.14	.14	.15	**.17**	**.17**	.14	.15
MI	C	.54	.54	.57	.57	.54	.49	.55	.47	.56	**.60**	.49
	M	.55	.54	.53	.49	.53	.53	**.57**	.51	.50	.46	.48
PEMS	C	.79	.82	.79	.81	.79	**.83**	**.83**	**.83**	.81	.82	.72
	M	.34	.41	.33	.39	.31	.48	**.64**	.50	.52	.38	.40
PS	C	.05	.06	.06	.05	.05	**.07**	.06	**.07**	**.07**	.03	.06
	M	.05	.06	.05	.05	.05	.05	.06	.07	**.08**	.03	.07
RS	C	.65	.64	.66	**.72**	.70	.62	.40	.56	.67	.52	.55
	M	.20	.33	.26	.34	.23	**.66**	.36	.61	.53	.57	.44
SRS1	C	.69	**.78**	.76	**.78**	.77	.74	.76	.63	.76	.70	.75
	M	**.77**	.67	.69	.66	.64	.69	.68	.75	.73	.74	.74
SRS2	C	**.57**	.51	.53	.54	.52	.55	.52	.51	.50	.54	.51
	M	.55	.55	.54	.56	**.58**	.55	.54	.55	.51	.51	.52
Avg. AUC	C	.49	.50	.50	**.52**	.49	.51	.49	.49	**.52**	.49	.48
	M	.39	.41	.40	.41	.39	.46	.46	**.49**	.47	.44	.45
Avg. Rank	C	11.0	11.1	10.6	8.0	11.4	10.3	9.8	10.6	**7.7**	11.2	12.8
	M	13.2	13.1	15.2	13.8	15.1	10.5	10.0	**8.8**	12.4	14.5	12.0

Table 3: One-versus-rest precision-recall AUC for the semi-supervised anomaly detection experiments on the UEA multivariate time series repository. The conformance and Mahalanobis methods are denoted by C and M, respectively. The symbols I, S, S^{∞} and S^{rand} represent the integral, truncated signature, PDE signature, and randomized signature kernels, respectively.

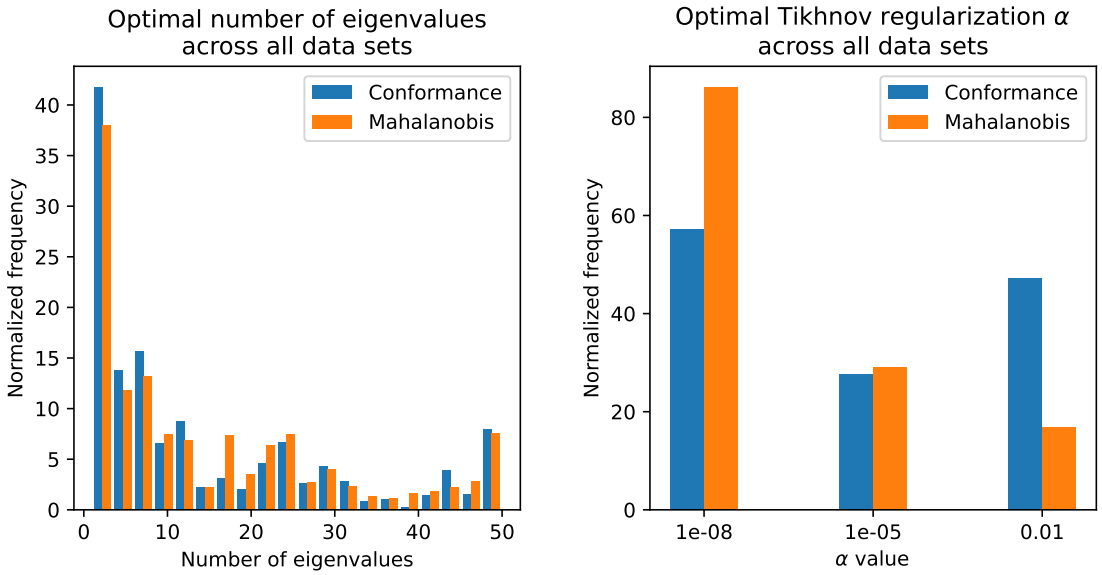

Figure 1: Optimal hyper-parameters for computing the anomaly distance as per Algorithms 1 to 3, sampled across all data sets and all kernels, normalized by the number of classes per data set. The results were obtained via a repeated k-fold cross-validation on the train set.

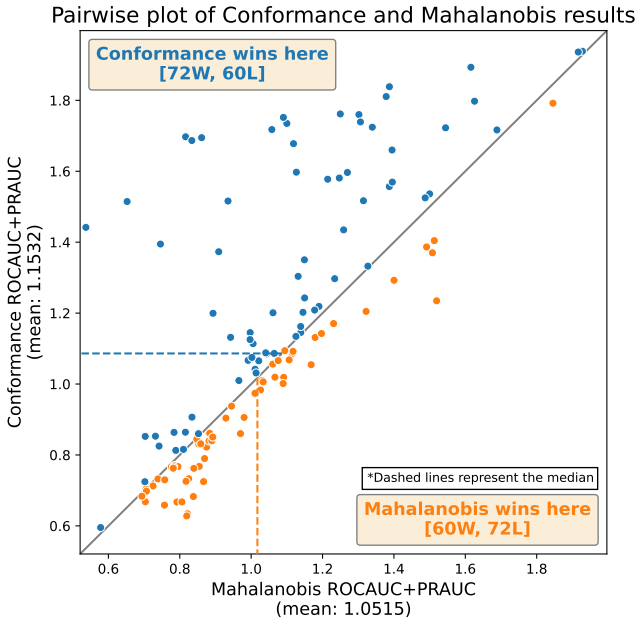

Figure 2: Pairwise comparison of one-versus-rest test scores for the Mahalanobis distance and the Conformance score. Each point represents one kernel and one data set.

within a single data set. The average ROC and PR AUC scores calculated over all data sets range from 0.64-0.67 and 0.49-0.52, respectively, with the best results obtained from the RBF integral, RBF signature, and GAK kernel.

When it comes to average rank, the Mahalanobis linear truncated signature and conformance RBF truncated signature take the number one spot, with the VRK and GAK kernels as close second place contenders. These kernels also have the most number of first places across all data sets, especially the linear truncated signature kernel. However, since the results are very data set dependent, the best performing model and kernel combination will vary on a case-by-case basis.

Fig. 2 shows a pairwise scatter plot of the Mahalanobis distance and conformance score test results for all kernels and all data sets. The results suggest that most of the time there is no significant advantage to using one anomaly distance over the other, except for a few cases seen in the upper left quadrant where the conformance score greatly outperforms the Mahalanobis distance. The difference in performance seem to be more pronounced for the simple flattened and integral-class kernels, where the average difference is 0.07 points, as opposed to the dynamic-time kernels where the average difference is 0.02 points. This difference is more pronounced for the PR-AUC metric, and two interesting examples are RacketSports and PEMS-SF where the PR-AUC doubles for select kernels when using the conformance method.

When it comes to computing the variance norm according to Algorithms 1 to 3, both the Mahalanobis and conformance methods on average obtained their highest cross validation scores using a low number of eigenvalues, as seen in Fig. 1. Furthermore, we see that both methods in general preferred a low regularization parameter α, with $\alpha =$ 1e-08 being most commonly used.

Acknowledgments

TC has been supported by the EPSRC Programme Grant EP/S026347/1 and acknowledges the support of the Erik Ellentuck Fellowship at the Institute for Advanced Study. NZ has been supported by the Roth Scholarship. We acknowledge computational resources and support provided by the Imperial College Research Computing Service (DOI: `10.14469/hpc/2232`). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising. We would like to thank the anonymous reviewers for their helpful comments on earlier versions of the manuscript which helped significantly improve the paper.

References

Erdinc Akyildirim, Matteo Gambara, Josef Teichmann, and Syang Zhou. Applications of signature methods to market anomaly detection. *Preprint, arXiv 2201.02441*, 2022.

Erdinc Akyildirim, Matteo Gambara, Josef Teichmann, and Syang Zhou. Randomized signature methods in optimal portfolio selection. *Preprint, arXiv 2312.16448*, 2023.

C.D. Aliprantis and K.C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2007.

Rita Giuliano Antonini. Subgaussian random variables in Hilbert spaces. *Rendiconti del Seminario Matematico della Università di Padova*, 98:89–99, 1997.

Ana Arribas-Gil and Juan Romo. Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15(4):603–619, 2014.

Paola Arrubarrena, Maud Lemercier, Bojan Nikolic, Terry Lyons, and Thomas Cass. Novelty detection on radio astronomy data using signatures. *Preprint, arXiv 2402.14892*, 2024.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The uea multivariate time series classification archive, 2018. *Preprint, arXiv 1811.00075*, 2018.

José R. Berrendero, Beatriz Bueno-Larraz, and Antonio Cuevas. On mahalanobis distance in functional settings. *J. Mach. Learn. Res.*, 21(1), 2020. ISSN 1532-4435.

Rajendra Bhatia. *Matrix Analysis*. Springer, 1997. ISBN 0387948465.

Francesca Biagini, Lukas Gonon, and Niklas Walter. Universal randomised signatures for generative time series modelling. *Preprint, arXiv 2406.10214*, 2024.

V.I. Bogachev. *Gaussian Measures*. Mathematical Surveys and Monographs. American Mathematical Society, 2015.

S. Boyd and L. Chua. Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems*, 32(11):1150–1161, 1985.

Paul O. Brown, Meng Ching Chiang, Shiqing Guo, Yingzi Jin, Carson K. Leung, Evan L. Murray, Adam G. M. Pazdor, and Alfredo Cuzzocrea. Mahalanobis distance based k-means clustering. In Robert Wrembel, Johann Gamper, Gabriele Kotsis, A. Min Tjoa, and Ismail Khalil, editors, *Big Data Analytics and Knowledge Discovery*, pages 256–262, Cham, 2022. Springer International Publishing.

Thomas Cass and Cristopher Salvi. Lecture notes on rough paths and applications to machine learning. *Preprint, arXiv 2404.06583*, 2024.

Thomas Cass and William F. Turner. Free probability, path developments and signature kernels as universal scaling limits. *Preprint, arXiv 2402.12311*, 2024.

Zhipeng Chang, Wenhe Chen, Yuping Gu, and Haoyue Xu. Mahalanobis-taguchi system for symbolic interval data based on kernel mahalanobis distance. *IEEE Access*, 8:20428–20438, 2020.

Xiaohui Chen and Yun Yang. Hanson–Wright inequality in Hilbert spaces with application to K-means clustering for non-Euclidean data. *Bernoulli*, 27(1):586 – 614, 2021.

N. N. Vakhania;V. I. Tarieladze;S. A. Chobanyan. *Probability distributions on Banach spaces*. Mathematics and its Applications. Springer Dordrecht, 1987.

Nicola Muca Cirone, Antonio Orvieto, Benjamin Walker, Cristopher Salvi, and Terry Lyons. Theoretical foundations of deep selective state-space models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Nicola Muça Cirone, Maud Lemercier, and Cristopher Salvi. Neural signature kernels as infinite-width-depth-limits of controlled resnets. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Enea Monzio Compagnoni, Anna Scampicchio, Luca Biggio, Antonio Orvieto, Thomas Hofmann, and Josef Teichmann. On the effectiveness of randomized signatures as reservoir for learning rough dynamics. In *IJCNN*, pages 1–8, 2023.

Christa Cuchiero and Janka Möller. Signature methods in stochastic portfolio theory. *SIAM Journal on Financial Mathematics*, 16(4):1239–1303, 2025.

Christa Cuchiero, Lukas Gonon, Lyudmila Grigoryeva, Juan-Pablo Ortega, and Josef Teichmann. Expressive power of randomized signature. In *Advances in Neural Information Processing Systems*, 2021.

Christa Cuchiero, Lukas Gonon, Lyudmila Grigoryeva, Juan-Pablo Ortega, and Josef Teichmann. Discrete-time signatures and randomness in reservoir computing. *IEEE Trans. Neural Netw. Learn. Syst.*, 33(11):6321–6330, 2022. ISSN 2162-237X,2162-2388.

Felipe Cucker and Stephen Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2001.

Felipe Cucker and Ding Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2007.

Marco Cuturi. Fast global alignment kernels. In *Proceedings of the 28th International Conference on Machine Learning*, page 929–936. Omnipress, 2011.

Marco Cuturi, Jean-Philippe Vert, Oystein Birkenes, and Tomoko Matsui. A kernel for time series based on global alignments. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages II–413–II–416, 2007.

Salva Daneshgadeh Çakmakçı, Thomas Kemmerich, Tarem Ahmed, and Nazife Baykal. Online ddos attack detection using mahalanobis distance and kernel-based learning algorithm. *Journal of Network and Computer Applications*, 168:102756, 2020.

R. De Maesschalck, D. Jouan-Rimbaud, and D.L. Massart. The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000.

I. Gohberg, S. Goldberg, and M. Kaashoek. *Basic Classes of Linear Operators*. Birkhäuser Basel, 2012.

Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega. Reservoir kernels and volterra series. *Preprint, arXiv 2212.14641*, 2022.

Lyudmila Grigoryeva and Juan-Pablo Ortega. Dimension reduction in recurrent networks by canonicalization. *Journal of Geometric Mechanics*, 13(4):647–677, 2021.

Martin Hairer. An introduction to stochastic pdes, 2023. URL `https://www.hairer.org/notes/SPDEs.pdf`.

V. Hautamaki, I. Karkkainen, and P. Franti. Outlier detection using k-nearest neighbour graph. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 430–433 Vol.3, 2004.

Ilse C. F. Ipsen and Arvind K. Saibaba. Stable rank and intrinsic dimension of real and complex matrices. *SIAM Journal on Matrix Analysis and Applications*, 46(3):1988–2007, 2025.

Herbert Jaeger and Harald Haas. Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science*, 304(5667):78–80, 2004.

Franz J. Kiraly and Harald Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20(31):1–45, 2019.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110 – 133, 2017.

B Kramer and A MacKinnon. Localization: theory and experiment. *Reports on Progress in Physics*, 56(12):1469, 1993.

R. Kress. *Linear Integral Equations*. Applied Mathematical Sciences. Springer New York, 2013.

Hajer Lahdhiri, Okba Taouali, Ilyes Elaissi, Ines Jaffel, Mohamed Faouzi Harakat, and Hassani Messaoud. A new fault detection index based on mahalanobis distance and kernel method. *The International Journal of Advanced Manufacturing Technology*, 91(5): 2799–2809, 2017.

P.D. Lax. *Functional Analysis*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2014.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Springer Berlin, 1991. ISBN 978-3-642-20211-7.

Darrick Lee and Harald Oberhauser. The signature kernel. *Preprint, arXiv 2305.04625*, 2023.

Mikhail Lifshits. *Lectures on Gaussian Processes*. Graduate Texts in Mathematics. Springer Berlin, Heidelberg, 2012.

Raz Lin, Eliyahu Khalastchi, and Gal A. Kaminka. Detecting anomalies in unmanned vehicles using the mahalanobis distance. In *2010 IEEE International Conference on Robotics and Automation*, pages 3038–3044, 2010.

Hang Lou, Siran Li, and Hao Ni. PCF-GAN: generating sequential data via the characteristic function of measures on the path space. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Hang Lou, Siran Li, and Hao Ni. Path development network with finite-dimensional lie group. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Terry J. Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2):215–310, 1998.

W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Computation*, 14:2531–2560, 2002.

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

Esdras Joseph Pedro Galeano and Rosa E. Lillo. The mahalanobis distance for functional data with applications to classification. *Technometrics*, 57(2):281–291, 2015.

Elzbieta Pekalska and Bernard Haasdonk. Kernel discriminant analysis for positive definite and indefinite kernels. *IEEE Trans Pattern Anal Mach Intell*, 31(6):1017–1032, 2009.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, 2005.

Stefano Recanatesi, Serena Bradde, Vijay Balasubramanian, Nicholas A. Steinmetz, and Eric Shea-Brown. A scale-dependent measure of system dimensionality. *Patterns*, 3(8): 100555, 2022. ISSN 2666-3899.

A Ruiz and P E López-de Teruel. Nonlinear kernel-based statistical pattern analysis. *IEEE Trans Neural Netw*, 12(1):16–32, 2001.

Alejandro Pasos Ruiz, Michael Flynn, James Large, Matthew Middlehurst, and Anthony Bagnall. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 35(2):401–449, 2021.

Cristopher Salvi, Thomas Cass, James Foster, Terry Lyons, and Weixin Yang. The signature kernel is the solution of a goursat pde. *SIAM Journal on Mathematics of Data Science*, 3(3):873–899, 2021a.

Cristopher Salvi, Maud Lemercier, Chong Liu, Blanka Horvath, Theodoros Damoulas, and Terry Lyons. Higher order kernel mean embeddings to capture filtrations of stochastic processes. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16635–16647. Curran Associates, Inc., 2021b.

I W Sandberg. Series expansions for nonlinear systems. *Circuits, Systems and Signal Processing*, 2(1):77–87, 1983.

Hassan Sarmadi and Abbas Karamodin. A novel anomaly detection method based on adaptive mahalanobis-squared distance and one-class knn rule for structural health monitoring under environmental effects. *Mechanical Systems and Signal Processing*, 140:106495, 2020.

Bernhard Schäfl, Lukas Gruber, Johannes Brandstetter, and Sepp Hochreiter. G-signatures: Global graph propagation with randomized signatures. *Preprint, arXiv 2302.08811*, 2023.

Alexander J Schölkopf, Bernhard;Smola. *Learning with kernels: support vector machines, regularization, optimization and beyond*. Adaptive computation and machine learning. MIT Press, 2009.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

Jun Shang, Maoyin Chen, and Hanwen Zhang. Fault detection based on augmented kernel mahalanobis distance for nonlinear dynamic processes. *Computers & Chemical Engineering*, 109:311–321, 2018.

Zhen Shao, Ryan Sze-Yin Chan, Thomas Cochrane, Peter Foster, and Terry Lyons. Dimensionless anomaly detection on multivariate streams with variance norm and path signature. *Preprint, arXiv 2006.03487*, 2023.

Hiroshi Shimodaira, Ken-ichi Noma, Mitsuru Nakai, and Shigeki Sagayama. Dynamic time-alignment kernel in support vector machine. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

Randall Eubank Tailen Hsing. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley, 2015.

Csaba Toth and Harald Oberhauser. Bayesian learning from sequential data using Gaussian processes with signature covariances. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 9548–9560, 2020.

Csaba Tóth, Harald Oberhauser, and Zoltán Szabó. Random fourier signature features. *SIAM Journal on Mathematics of Data Science*, 7(1):329–354, 2025.

Ghislain Verdier and Ariane Ferreira. Adaptive mahalanobis distance and k -nearest neighbor rule for fault detection in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 24(1):59–68, 2011.

Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press, 2012.

Defeng Wang, Daniel S. Yeung, and Eric C. C. Tsang. Weighted mahalanobis distance kernels for support vector machines. *IEEE Transactions on Neural Networks*, 18(5): 1453–1462, 2007.

Pa-Chun Wang, Chao-Ton Su, Kun-Huang Chen, and Ning-Hung Chen. The application of rough set and mahalanobis distance to enhance the quality of osa diagnosis. *Expert Systems with Applications*, 38(6):7828–7836, 2011.

Norbert Wiener. *Nonlinear Problems in Random Theory.* The Technology Press of MIT, 1958.

Shiming Xiang, Feiping Nie, and Changshui Zhang. Learning a mahalanobis distance metric for data clustering and classification. *Pattern recognition*, 41(12):3600–3612, 2008.