Eigendecompositions of Transfer Operators in Reproducing Kernel Hilbert Spaces

Stefan Klus[1], Ingmar Schuster[1], and Krikamol Muandet[2]

[1]Department of Mathematics and Computer Science, Freie Universität Berlin, Germany
[2]Department of Mathematics, Faculty of Science, Mahidol University, Thailand

Abstract

Transfer operators such as the Perron–Frobenius or Koopman operator play an important role in the global analysis of complex dynamical systems. The eigenfunctions of these operators can be used to detect metastable sets, to project the dynamics onto the dominant slow processes, or to separate superimposed signals. We extend transfer operator theory to reproducing kernel Hilbert spaces and show that these operators are related to Hilbert space representations of conditional distributions, known as conditional mean embeddings in the machine learning community. Moreover, numerical methods to compute empirical estimates of these embeddings are akin to data-driven methods for the approximation of transfer operators such as extended dynamic mode decomposition and its variants. In fact, most of the existing methods can be derived from our framework, providing a unifying view on the approximation of transfer operators. One main benefit of the presented kernel-based approaches is that these methods can be applied to any domain where a similarity measure given by a kernel is available. We illustrate the results with the aid of guiding examples and highlight potential applications in molecular dynamics as well as video and text data analysis.

1 Introduction

Transfer operators such as the Perron–Frobenius or Koopman operator are ubiquitous in molecular dynamics, fluid dynamics, atmospheric sciences, and also control theory. The eigenfunctions of these operators can be used to decompose the system into fast and slow dynamics and to identify so-called metastable sets, which, in the molecular dynamics context, correspond to conformations of molecules. Compared to the fast vibrations of the atoms, the transitions between different conformations are much slower, the time scales typically differ by several orders of magnitude. We are in particular interested in the slow conformational changes of molecules and the corresponding transition probabilities and transition paths. However, the methods presented in this paper can be applied to data generated by any dynamical system and we will show potential novel applications pertaining to video and text data analysis.

Over the last decades, different numerical methods such as *Ulam's method* (Ulam, 1960), *extended dynamic mode decomposition* (EDMD) (Williams et al., 2015a,b, Klus et al., 2016), the *variational approach of conformation dynamics* (VAC) (Noé and Nüske, 2013, Nüske et al., 2014), and several extensions and generalizations have been developed to approximate transfer operators and their eigenvalues and eigenfunctions. The advantage of purely data-driven methods is that they can be applied to simulation or measurement data, information about the underlying system itself is not required. An overview and comparison of such methods can be found in Klus et al. (2017) and the recently published book Kutz et al. (2016). Applications and variants of these methods are also described in Rowley et al. (2009), Budišić et al. (2012), Tu et al. (2014), McGibbon and Pande (2015). Kernel-based reformulations of the aforementioned methods have been proposed in Williams et al. (2015b) and Schwantes and Pande (2015).

In this work, we construct representations of transfer operators using reproducing kernel Hilbert space (RKHS) theory. An RKHS $\mathbb{H}$ is a Hilbert space of real-valued functions in which all evaluation functionals are bounded (see Section 2). For any RKHS $\mathbb{H}$, there always exists a *reproducing kernel* $k\colon \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ such that $k(x,\cdot) \in \mathbb{H}$ for all $x \in \mathbb{X}$ and $f(x) = \langle f, k(x,\cdot)\rangle_{\mathbb{H}}$ for all $x \in \mathbb{X}$ and $f \in \mathbb{H}$. The latter is commonly known as the *reproducing property* of $\mathbb{H}$ and implies that $k(x,x') = \langle k(x,\cdot), k(x',\cdot)\rangle_{\mathbb{H}}$. In other words, the kernel evaluation $k(x,x')$ can be regarded as an inner product between *implicit* feature maps of x and x' in $\mathbb{H}$. As we will see, defining the transfer operators (see Section 3) in an RKHS enables us to model and analyze nonlinear dynamical systems without requiring an explicit data representation (see Section 4). In particular, we can directly express the kernel transfer operators in terms of covariance and cross-covariance operators in the RKHS. Existing kernel-based approximations such as kernel EDMD (Williams et al., 2015b) or kernel TICA (Schwantes and Pande, 2015) are special cases of our approach. The benefits of kernel-based methods are twofold: First, the basis functions need not be defined explicitly, which thereby allows us to handle infinite-dimensional feature spaces. Second, the proposed method can not only be applied to dynamical systems defined on Euclidean spaces, but also to systems defined on any domain that admits an appropriate kernel function such as images, graphs, or strings. In other words, our methods allow to characterize wide-sense stationary stochastic processes over many non-standard domains. We show that the kernel transfer operators are closely related to recently developed Hilbert space embeddings of probability distributions (Berlinet and Thomas-Agnan, 2004, Smola et al., 2007, Muandet et al., 2017).

Moreover, we propose an eigendecomposition technique for kernel transfer operators. As mentioned above, the eigenfunctions and eigenvalues of transfer operators provide insights into fast and slow dynamics of the system. For kernel transfer operators, we show that the corresponding eigenfunctions belong to the RKHS associated with the kernel function and can be expressed entirely in terms of the eigenvectors and eigenvalues of Gram matrices defined for training data. Therefore, our technique resembles several existing kernel-based component analysis techniques in machine learning. For example, kernel principal component analysis (KPCA) extends the well-known PCA to data mapped into an RKHS (Schölkopf et al., 1998). KPCA aims to find a low-dimensional projection which maximally preserves the variance of the data projected into the feature space. For a feature space corresponding to an RKHS $\mathbb{H}$, the basis of this projection can be expressed in terms of the eigenfunctions of the covariance operator in $\mathbb{H}$. Well-known applications of KPCA include

dimension reduction (Schölkopf et al., 1998) and image denoising (Mika et al., 1999). Similarly, our techniques can be used to reduce the dimension of high-dimensional dynamical systems. Given variables X and Y, kernel canonical correlation analysis (KCCA) aims to find projections of low-dimensional RKHS representations of each variable separately such that the projections are maximally correlated (Fukumizu et al., 2007). The purpose of KCCA is to find nonlinear projections that are important for explaining covariation between sets of variables. In the context of this work, X and Y represent two distinct observations of a stochastic process at time t and $t+\tau$, respectively. Our goal, on the other hand, is to find a low-dimensional projection of the process governing these observations. Additionally, an independent component analysis (ICA) is an important algorithm for the blind source separation problem. It aims to recover a latent random vector x whose components are *mutually independent* from observations of $y = Ax$ where A is a mixing matrix (Hyvärinen and Oja, 2000). Bach and Jordan (2003) proposed a class of efficient algorithms for ICA which use contrast functions based on canonical correlations defined in an RKHS. Due to the kernel functions, the contrast functions and their derivatives can be computed efficiently. Lastly, our work is also closely related to kernel-based functional principal component analysis (FPCA). FPCA aims to find the dominant modes of variation of functional data (Yao et al., 2005, Hall et al., 2006), which has applications in time series analysis, longitudinal data analysis, and functional regression/classification. It has been shown that an orthonormal basis which explains the most variation consists of the eigenfunctions of the autocovariance operator, which can be viewed as a particular transfer operator (see Section 3). For detailed exposition of the aforementioned techniques, we refer interested readers to some recent papers including Ramsay and Silverman (2005), Van Der Maaten et al. (2009), Burges (2010), for example.

Our work provides a unified framework for nonlinear component analysis of transfer operators pertaining to dynamical systems. Given that dynamical systems are ubiquitous in machine learning, we believe it could potentially lead to novel applications such as visualization of high-dimensional dynamics, dimension reduction, source separation and denoising, data summarization, and clustering based on sequence information (see Section 5). The main contributions of this work are:

1. We extend transfer operators, namely the Perron–Frobenius and the Koopman operator, to RKHSs and show that they can be expressed entirely in terms of covariance and cross-covariance operators defined by the underlying process (Proposition 4.1 and Corollary 4.2). Furthermore, we construct the empirical estimates of these operators (Proposition 4.3) which, as opposed to existing methods such as EDMD, do not require the basis functions to be given explicitly.

2. We propose an algorithm to obtain eigenfunctions and eigenvalues of the kernel transfer operators (Section 4.6). Existing methods to approximate transfer operator eigendecompositions such as TICA and DMD can be obtained as special cases of our algorithm by choosing a linear kernel function. It is also possible to obtain the non-linear counterparts including VAC and EDMD by using kernels with explicitly given finite-dimensional feature spaces. Analogously, kernel TICA and kernel EDMD can be derived with the aid of our kernel transfer operator framework (Section 4.7).

3. A particular kernel-based transfer operator, namely the embedded Perron–Frobenius

Table 1: Overview of notation.

random variable	X	Y
domain	$\mathbb{X}$	$\mathbb{Y}$
observation	x	y
kernel function	$k(x, x')$	$l(y, y')$
feature map	$\phi(x)$	$\psi(y)$
feature matrix	$\Phi = [\phi(x_1), \ldots, \phi(x_n)]$	$\Psi = [\psi(y_1), \ldots, \psi(y_n)]$
Gram matrix	$G_{XX} = \Phi^\top \Phi$	$G_{YY} = \Psi^\top \Psi$
RKHS	$\mathbb{H}$	$\mathbb{G}$

operator (Section 4.2), is indeed equivalent to the *conditional mean embedding* (CME) formulation (Song et al., 2009, 2013). The CME has applications ranging from probabilistic inference to reinforcement learning. Exploiting transfer operator theory will thus have an impact on the aforementioned applications.

4. Lastly, we demonstrate the use of kernel transfer operators in molecular dynamics as well as video and text data analysis (Section 5).

The remainder of this paper is organized as follows: In Section 2, we first introduce the notion of reproducing kernel Hilbert spaces, positive definite kernels, and Hilbert space embeddings of conditional distributions. Section 3 gives a brief introduction to transfer operators, followed by the kernel formulation of transfer operators in Section 4. We demonstrate the proposed methods in Section 5 using several illustrative and real-world examples and conclude with a short summary and future work in Section 6.

2 Reproducing Kernel Hilbert Spaces

In this section, we will introduce reproducing kernel Hilbert spaces and positive definite kernels (Schölkopf and Smola, 2001, Hofmann et al., 2008) as well as Hilbert space embeddings of probability distributions (Berlinet and Thomas-Agnan, 2004, Smola et al., 2007, Muandet et al., 2017), which will later on be used to reformulate the transfer operators defined below. Readers familiar with these concepts can skip this section. The notation and symbols, which we summarize in Table 1, are based on Song et al. (2009), Muandet et al. (2017).

Definition 2.1 (Reproducing kernel Hilbert space, (Schölkopf and Smola, 2001))**.** *Let* $\mathbb{X}$ *be a set and* $\mathbb{H}$ *a space of functions* $f \colon \mathbb{X} \to \mathbb{R}$*. Then* $\mathbb{H}$ *is called a* reproducing kernel Hilbert space *(RKHS) with corresponding scalar product* $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ *and induced norm* $\|f\|_{\mathbb{H}} = \langle f, f \rangle_{\mathbb{H}}^{1/2}$ *if there is a function* $k \colon \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ *such that*

(i) $\langle f, k(x, \cdot) \rangle_{\mathbb{H}} = f(x)$ *for all* $f \in \mathbb{H}$ *and*

(ii) $\mathbb{H} = \overline{\operatorname{span}\{k(x, \cdot) \mid x \in \mathbb{X}\}}$.

The first requirement, which is called *reproducing property* of $\mathbb{H}$, in particular implies $\langle k(x, \cdot), k(x', \cdot) \rangle_{\mathbb{H}} = k(x, x')$ for any $x, x' \in \mathbb{X}$. As a result, the function evaluation of f at a

given point x can be regarded as an inner product evaluation in $\mathbb{H}$ between the representer $k(x,\cdot)$ of x and the function itself. Furthermore, we may treat $k(x,\cdot)$ as a feature map $\phi(x)$ of x in $\mathbb{H}$ such that $k(x,x') = \langle \phi(x), \phi(x') \rangle_{\mathbb{H}}$. We refer to $k(x,\cdot)$ as the *canonical feature map* of x. Note that in most applications of kernel methods, we only require the kernel evaluation $k(x,x')$, so $k(x,\cdot)$ needs not be computed explicitly. For more details, see Schölkopf and Smola (2001), Song et al. (2009).

Example 2.2. Let $\mathbb{X} \subset \mathbb{R}^2$. If we want to use a polynomial nonlinearity, we have two options of constructing a Hilbert space and endowing it with an inner product that would result in methods that numerically give the same result:

- Given the polynomial kernel $k(x,x') = (1 + \langle x, x' \rangle)^2$, we use the *canonical feature map* $\phi_{\text{can}}(x) = k(x,\cdot)$ and the standard RKHS inner product satisfying the reproducing property of Definition 2.1, i.e., $\langle f, k(x,\cdot) \rangle_{\mathbb{H}} = f(x)$. The features are then a subset of the function space $\mathbb{H}$. Given factors $\alpha_i \in \mathbb{R}$ and points $x_i \in \mathbb{X}$, with $i = 1, \ldots, n$, a function f can be written as $f(\cdot) = \sum_{i=1}^{n} \alpha_i \, \phi_{\text{can}}(x_i)$.
- Alternatively, the *explicit feature map* $\phi_{\text{exp}}(x) = [1, \sqrt{2}x_1, \sqrt{2}x_2 x_1^2, \sqrt{2}x_1 x_2, x_2^2]^\top$ with the standard Euclidean inner product could be used. The features are then a subset of $\mathbb{R}^6$. Using the explicit feature map, the function f can be represented as $f(\cdot) = \langle f_{\text{exp}}, \phi_{\text{exp}}(\cdot) \rangle$ with $f_{\text{exp}} = \sum_{i=1}^{n} \alpha_i \, \phi_{\text{exp}}(x_i) \in \mathbb{R}^6$. ▲

The second point of view is equivalent and will often save storage space and computing time. Whenever the polynomial kernel is used, it might thus be preferred. However, the polynomial kernel does not have the theoretical advantages of so-called characteristic kernels, where an explicit feature map view typically does not exist. For this reason, we will mostly stick to the first point of view.

The kernel k in Definition 2.1 is called a *reproducing kernel* of $\mathbb{H}$. It fully characterizes the RKHS $\mathbb{H}$. That is, for every positive definite kernel k on $\mathbb{X} \times \mathbb{X}$, there exists a unique RKHS with k as its reproducing kernel. Conversely, the reproducing kernel of a given RKHS is unique and positive definite (Aronszajn, 1950).

Definition 2.3 (Positive definite kernel)**.** *Given a set $\mathbb{D}_X = \{x_1, \ldots, x_n\} \subset \mathbb{X}$, let $G_{XX} \in \mathbb{R}^{n \times n}$ be the Gram matrix, i.e., $[G_{XX}]_{ij} = k(x_i, x_j)$. A bivariate function k on $\mathbb{X} \times \mathbb{X}$ is* positive definite *if $k(x,y) = k(y,x)$ and it satisfies*

$$\mathbf{c}^\top G_{XX} \, \mathbf{c} = \sum_{i,j=1}^{n} c_i \, c_j \, k(x_i, x_j) \geq 0$$

for any $n \in \mathbb{N}$, any choice of $x_1, \ldots, x_n \in \mathbb{X}$, and any $\mathbf{c} = [c_1, \ldots, c_n] \in \mathbb{R}^n$. It is said to be strictly positive definite *if $\mathbf{c}^\top G_{XX} \, \mathbf{c} = 0$ implies $\mathbf{c} = 0$.*

Example 2.4. The following functions are positive definite kernels on $\mathbb{R}^d$:

(i) Linear kernel: $k(x,x') = x^\top x'$.

(ii) Polynomial kernel of degree p: $k(x,x') = (x^\top x' + c)^p$ with $c > 0$.

(iii) Gaussian kernel: $k(x,x') = \exp\left(-\frac{1}{2\sigma^2} \|x - x'\|_2^2\right)$ with a bandwidth $\sigma > 0$.

(iv) Laplacian kernel: $k(x, x') = \exp\left(-\frac{1}{\sigma}\|x - x'\|_2\right)$ with a bandwidth $\sigma > 0$. ▲

The positive definiteness of the kernel ensures that we can always find a feature map $\phi\colon \mathbb{X} \to \mathbb{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathbb{H}}$. For example, the canonical feature map $\phi(x) = k(x, \cdot)$ satisfies this property (cf. Definition 2.1). As a result, if all we need to evaluate is the inner product between $\phi(x)$ and $\phi(x')$ in $\mathbb{H}$, we need not construct ϕ explicitly, which can be computationally expensive in high dimensional feature spaces. In fact, some kernels such as the Gaussian kernel correspond to infinite-dimensional feature spaces which make it impossible to construct ϕ in practice. Most kernel-based learning algorithms rely on computations involving only Gram matrices. As we will see later, although our transfer operators are defined in terms of ϕ and may live in an infinite-dimensional space, all associated operations can be carried out in terms of the finite-dimensional Gram matrices obtained from training data.

2.1 Hilbert Space Embedding of Marginal Distributions

The idea of kernel mean embeddings is to extend feature maps to the space of probability distributions (Berlinet and Thomas-Agnan, 2004, Smola et al., 2007, Muandet et al., 2017).

Definition 2.5 (Mean embedding)**.** *Let $\mathbb{M}_+^1(\mathbb{X})$ be the space of all probability measures $\mathbb{P}$ on $\mathbb{X}$ and $k\colon \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a measurable real-valued kernel endowed with the RKHS $\mathbb{H}$ such that $\sup_{x \in \mathbb{X}} k(x, x) < \infty$. Then the* kernel mean embedding $\mu_{\mathbb{P}} \in \mathbb{H}$ *is defined by*

$$\mu_{\mathbb{P}} = \mathbb{E}_X[\phi(X)] = \int \phi(x)\, \mathrm{d}\mathbb{P}(x) = \int k(x, \cdot)\, \mathrm{d}\mathbb{P}(x).$$

Given a set of training data $\mathbb{D}_X = \{x_1, \ldots, x_n\}$ drawn i.i.d. from $\mathbb{P}(X)$, the empirical estimate of the mean embedding can be computed as

$$\widehat{\mu}_{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) = \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot) = \frac{1}{n}\Phi\mathbb{1},$$

where $\Phi = [\phi(x_1), \ldots, \phi(x_n)]$ is the feature matrix and $\mathbb{1} = [1, \ldots, 1]^\top$ the vector of ones.

Remark 2.6. It follows from the reproducing property of $\mathbb{H}$ that

$$\mathbb{E}_X[f(X)] = \mathbb{E}_X[\langle f, \phi(X) \rangle] = \langle f, \mathbb{E}_X[\phi(X)] \rangle_{\mathbb{H}} = \langle f, \mu_{\mathbb{P}} \rangle_{\mathbb{H}}$$

for any $f \in \mathbb{H}$ and, analogously, $\frac{1}{n}\sum_{i=1}^{n} f(x_i) = \langle f, \widehat{\mu}_{\mathbb{P}} \rangle_{\mathbb{H}}$. Thus, the computation of expectations with respect to $\mathbb{P}$ can be regarded as a scalar product in a Hilbert space.

Different choices of kernel functions result in different representations of the distribution $\mathbb{P}$. In particular, the kernel mean embedding $\mu_{\mathbb{P}}$ fully characterizes $\mathbb{P}$ if k is a *characteristic kernel* (Fukumizu et al., 2004, Sriperumbudur et al., 2008, 2010).

Definition 2.7 (Characteristic kernel)**.** *If $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathbb{H}} = 0$ if and only if $\mathbb{P} = \mathbb{Q}$, then the kernel k is defined to be* characteristic*. The Hilbert space $\mathbb{H}$ is said to be characteristic if the associated kernel is characteristic.*

In other words, we do not lose any information about $\mathbb{P}$ by embedding it into a characteristic RKHS. Characteristic kernels are closely related to *universal kernels* (Steinwart, 2002). Let $C_b(\mathbb{X})$ be the space of bounded continuous function on a compact metric space $\mathbb{X}$. A kernel k on $\mathbb{X}$ is said to be universal if the corresponding RKHS $\mathbb{H}$ is dense in $C_b(\mathbb{X})$, i.e., for any $f \in C_b(\mathbb{X})$ and $\epsilon > 0$, there exists a function $h \in \mathbb{H}$ such that $\|f - h\|_\infty < \epsilon$. Gaussian kernels and Laplacian kernels, for instance, are known to be characteristic. In fact, all universal kernels are characteristic (Gretton et al., 2012, Theorem 5).

Definition 2.8 (Integral operator)**.** *A kernel k gives rise to an* integral operator $\mathcal{E}_k$*, defined by*

$$(\mathcal{E}_k f)(\cdot) := \int_{\mathbb{X}} k(x, \cdot) f(x) \, \mathrm{d}x.$$

The operator can be viewed as a generalization of conventional matrix-vector multiplication. The eigenvalues and eigenfunctions of this operator are, for instance, used to construct the Mercer feature space, see Schölkopf and Smola (2001). We will sometimes omit the subscript if it is clear which kernel is meant. It was shown in Kato (1980) that if $\int_{\mathbb{X}} |k(x,y)| \, \mathrm{d}x \leq M_1$, $\int_{\mathbb{X}} |k(x,y)| \, \mathrm{d}y \leq M_2$, and $f \in L^r(\mathbb{X})$, with $1 \leq r \leq \infty$, then we obtain $\|\mathcal{E}_k f\| \leq \max(M_1, M_2) \|f\|$ and the operator is bounded. In particular, if $\mathbb{X}$ is compact and $k(x,y)$ continuous in x and y, this is satisfied. Furthermore, if the kernel is *Hilbert–Schmidt*, i.e., $\iint_{\mathbb{X}\times\mathbb{X}} |k(x,y)|^2 \, \mathrm{d}x \, \mathrm{d}y < \infty$, then $\mathcal{E}_k$ is bounded (Renardy and Rogers, 2006, Lemma 8.2) and compact (Bump, 1998, Theorem 2.3.2). Whenever $\mathbb{P}$ has a density p, this means $\mu_{\mathbb{P}} = \mathcal{E}_k p$. For certain combinations of basis functions and kernels, the embedding can be computed analytically, see Appendix A.

2.2 Covariance Operators

We now introduce the concept of covariance operators in Hilbert spaces (Baker, 1970, 1973). Let (X, Y) be a random variable on $\mathbb{X} \times \mathbb{Y}$ with corresponding marginal distributions $\mathbb{P}(X)$ and $\mathbb{P}(Y)$, respectively, and joint distribution $\mathbb{P}(X, Y)$. In what follows, we assume integrability, i.e., $\mathbb{E}_X[k(X,X)] < \infty$ and $\mathbb{E}_Y[l(Y,Y)] < \infty$ so that $\mathbb{H} \subset L^2(\mathbb{P}(X))$ and $\mathbb{G} \subset L^2(\mathbb{P}(Y))$, respectively, where $L^2(\nu)$ denotes the space of square-integrable functions with respect to ν. See Muandet et al. (2017) for details.

Definition 2.9 (Covariance operators)**.** *Let ϕ and ψ be feature maps associated with the kernels k and l, respectively. Suppose that $\mathbb{E}_X[k(X,X)] < \infty$ and $\mathbb{E}_Y[l(Y,Y)] < \infty$. Then the* covariance operator $\mathcal{C}_{XX} \colon \mathbb{H} \to \mathbb{H}$ *and the* cross-covariance operator $\mathcal{C}_{YX} \colon \mathbb{H} \to \mathbb{G}$ *are defined as*

$$\begin{aligned} \mathcal{C}_{XX} &= \int \phi(X) \otimes \phi(X) \, \mathrm{d}\mathbb{P}(X) \quad\;\; = \mathbb{E}_X[\phi(X) \otimes \phi(X)], \\ \mathcal{C}_{YX} &= \int \psi(Y) \otimes \phi(X) \, \mathrm{d}\mathbb{P}(Y, X) = \mathbb{E}_{YX}[\psi(Y) \otimes \phi(X)]. \end{aligned}$$

Remark 2.10. Note that $\psi(y) \otimes \phi(x)$ defines a rank-one operator from $\mathbb{H}$ to $\mathbb{G}$ via

$$\big(\psi(y) \otimes \phi(x)\big) f = \langle \phi(x), f \rangle_{\mathbb{H}} \, \psi(y) = f(x) \, \psi(y)$$

so that $\big\langle \big(\psi(y) \otimes \phi(x)\big) f, \, g \big\rangle_{\mathbb{G}} = f(x) \, \langle \psi(y), \, g \rangle_{\mathbb{G}} = f(x) \, g(y)$.

The centered counterparts of $\mathcal{C}_{XX}$ and $\mathcal{C}_{YX}$ are defined similarly using the mean-subtracted feature maps $\phi_c(X) = \phi(X) - \mu_{\mathbb{P}(X)}$ and $\psi_c(Y) = \psi(Y) - \mu_{\mathbb{P}(Y)}$, where $\mu_{\mathbb{P}(X)} := \mathbb{E}_X[\phi(X)]$ and $\mu_{\mathbb{P}(Y)} := \mathbb{E}_Y[\psi(Y)]$. Intuitively, one may think of $\mathcal{C}_{XX}$ and $\mathcal{C}_{YX}$ as a nonlinear generalization of covariance and cross-covariance matrices. We can express the cross-covariance of two functions $f \in \mathbb{H}$ and $g \in \mathbb{G}$ in terms of $\mathcal{C}_{XY}$ and $\mathcal{C}_{YX}$ as

$$\mathbb{E}_{XY}[f(X)\, g(Y)] = \langle f,\, \mathcal{C}_{XY} g \rangle_{\mathbb{H}} = \langle \mathcal{C}_{YX} f,\, g \rangle_{\mathbb{G}}\,. \tag{1}$$

Hence, $\mathcal{C}_{XY}$ is the adjoint of $\mathcal{C}_{YX}$. The following result, which is due to Fukumizu et al. (2004), shows the relation between $\mathcal{C}_{XX}$ and $\mathcal{C}_{XY}$. We will use it later to define RKHS transfer operators.

Proposition 2.11. *If* $\mathbb{E}_{Y|X}[g(Y) \mid X = \cdot\,] \in \mathbb{H}$ *for all* $g \in \mathbb{G}$*, then*

$$\mathcal{C}_{XX}\mathbb{E}_{Y|X}[g(Y) \mid X = \cdot\,] = \mathcal{C}_{XY} g.$$

For a proof, see Fukumizu et al. (2004). The covariance operator and cross-covariance operator can in general not be computed directly since the joint distribution $\mathbb{P}(X, Y)$ is typically not known. We can, however, estimate it from sampled data. Given n pairs of training data $\mathbb{D}_{XY} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d. from the probability distribution $\mathbb{P}(X, Y)$, we define the feature matrices

$$\Phi = \begin{bmatrix} \phi(x_1) & \ldots & \phi(x_n) \end{bmatrix} \quad \text{and} \quad \Psi = \begin{bmatrix} \psi(y_1) & \ldots & \psi(y_n) \end{bmatrix}.$$

The corresponding Gram matrices are given by $G_{XX} = \Phi^\top \Phi$ and $G_{YY} = \Psi^\top \Psi$ (see Table 1) and the empirical estimates of $\mathcal{C}_{XX}$ and $\mathcal{C}_{YX}$ by

$$\widehat{\mathcal{C}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) \otimes \phi(x_i) = \frac{1}{n}\Phi\Phi^\top,$$
$$\widehat{\mathcal{C}}_{YX} = \frac{1}{n}\sum_{i=1}^{n} \psi(y_i) \otimes \phi(x_i) = \frac{1}{n}\Psi\Phi^\top.$$

Analogously, the mean-subtracted counterparts of $\widehat{\mathcal{C}}_{XX}$ and $\widehat{\mathcal{C}}_{YX}$ can be obtained as $\frac{1}{n}\Phi H \Phi^\top$ and $\frac{1}{n}\Psi H \Phi^\top$, where H is the centering matrix given by $H = I_n - \frac{1}{n}\mathbb{1}_n \mathbb{1}_n^\top$. Note that if both k and l are linear kernels for which ϕ and ψ are identity maps, we obtain covariance and cross-covariance matrices as a special case.

2.3 Hilbert Space Embedding of Conditional Distributions

In Section 2.1, we showed how to embed any marginal distribution into the RKHS. We will now extend this idea to conditional distributions. Interested readers should consult Song et al. (2009, 2013), Muandet et al. (2017) for further details on this topic. First of all, note that the embedding of a Dirac distribution supported on a single point $x \in \mathbb{X}$ is simply $\int k(z, \cdot)\, \mathrm{d}\delta_x(z) = k(x, \cdot)$. Given some $x \in \mathbb{X}$, the embedding of $\mathbb{P}(Y \mid X = x)$ in $\mathbb{G}$ can be defined according to Definition 2.5 as $\mu_{Y|x} = \mathbb{E}_{Y|x}[\psi(Y) \mid X = x]$. Hence, the Hilbert space representation of $\mathbb{P}(Y \mid X)$ is not a single element in $\mathbb{G}$, but a mapping which takes x to the embedding of the associated conditional distribution.

Definition 2.12 (Conditional mean embedding, (Song et al., 2009)). *Let $\mathcal{C}_{XX}$ be the covariance operator for X and $\mathcal{C}_{YX}$ be the cross-covariance operator from X to Y, respectively. Then the conditional mean embedding of $\mathbb{P}(Y \mid X)$ is given by*

$$\mathcal{U}_{Y|X} = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}.$$

Under the assumption that $\mathbb{E}_{Y|X}[g(Y) \mid X = \cdot\,] \in \mathbb{H}$ for all $g \in \mathbb{G}$, it follows from the reproducing property of $\mathbb{H}$ and Proposition 2.11 that

$$\begin{aligned}\mathbb{E}_{Y|x}[g(Y) \mid X = x] &= \langle \mathbb{E}_{Y|x}[g(Y) \mid X],\, k(x,\cdot)\rangle_{\mathbb{H}} \\ &= \left\langle \mathcal{C}_{XX}^{-1}\mathcal{C}_{XY}\, g,\, k(x,\cdot)\right\rangle_{\mathbb{H}} \\ &= \left\langle g,\, \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}k(x,\cdot)\right\rangle_{\mathbb{G}}\end{aligned}$$

for all $g \in \mathbb{G}$. That is, the conditional mean embedding of $\mathbb{P}(Y \mid X = x)$ can be expressed as $\mu_{Y|x} = \mathcal{U}_{Y|X}k(x,\cdot) = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}k(x,\cdot)$. See Song et al. (2009, Theorem 4) and Muandet et al. (2017) for further details and Song et al. (2013) for applications of conditional mean embeddings.

Remark 2.13. Since $\mathcal{C}_{XX}$ is a compact operator, finding its inverse is an ill-posed problem, i.e., $\mathcal{C}_{XX}^{-1}$ may not exist. That is, the assumption that $\mathbb{E}_{Y|X}[g(Y) \mid X = \cdot\,] \in \mathbb{H}$ for all $g \in \mathbb{G}$ may not hold in general. Hence, the operator $\mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$ may not exist in the continuous domain. A common approach to alleviate this problem is to consider the regularized version $(\mathcal{C}_{XX} + \varepsilon\mathcal{I})^{-1}$ instead, where ε is a regularization parameter and $\mathcal{I}$ is the identity operator in $\mathbb{H}$. The empirical estimator of the conditional mean embedding is then given by

$$\widehat{\mathcal{U}}_{Y|X} = \widehat{\mathcal{C}}_{YX}(\widehat{\mathcal{C}}_{XX} + \varepsilon\mathcal{I})^{-1} = \Psi(G_{XX} + n\varepsilon I_n)^{-1}\Phi^{\top}.$$

In Fukumizu et al. (2013, Theorem 8), the consistency and convergence of this estimator are shown under some mild conditions. For a detailed derivation, we refer to Song et al. (2009), Muandet et al. (2017).

3 Transfer Operators

We now give a brief introduction to transfer operators and their applications. A detailed exposition on this topic can be found in the recent review paper Klus et al. (2017).

Let $\{X_t\}_{t\geq 0}$ be a stochastic process defined on the state space $\mathbb{X} \subset \mathbb{R}^d$. Then the *transition density function* p_τ of observing the process near y at a time τ after it has been at x is defined by

$$\mathbb{P}[X_{t+\tau} \in \mathbb{A} \mid X_t = x] = \int_{\mathbb{A}} p_\tau(y \mid x)\,\mathrm{d}y,$$

where $\mathbb{A}$ is any measurable set. That is, $p_\tau(y \mid x)$ is the conditional probability density of $X_{t+\tau} = y$ given that $X_t = x$. The transfer operators considered in this work will be defined in terms of the transition density function p_τ. In what follows, $L^r(\mathbb{X})$, with $1 \leq r \leq \infty$, denotes the spaces of r-Lebesgue integrable functions and $\|\cdot\|_{L^r}$ the corresponding norm.

Definition 3.1 (Transfer operators). *Let $p_t \in L^1(\mathbb{X})$ be a probability density and $f_t \in L^\infty(\mathbb{X})$ an observable of the system. For a given lag time τ:*

(i) The Perron–Frobenius operator $\mathcal{P}\colon L^1(\mathbb{X}) \to L^1(\mathbb{X})$ *is defined by*

$$\mathcal{P}p_t(y) = \int p_\tau(y \mid x)\, p_t(x)\, \mathrm{d}x.$$

(ii) The Koopman operator $\mathcal{K}\colon L^\infty(\mathbb{X}) \to L^\infty(\mathbb{X})$ *is defined by*

$$\mathcal{K}f_t(x) = \int p_\tau(y \mid x)\, f_t(y)\, \mathrm{d}y = \mathbb{E}[f_t(X_{t+\tau}) \mid X_t = x].$$

Note that the operators and the corresponding eigenvalues implicitly depend on the lag time τ. A density π that is invariant under the action of $\mathcal{P}$ is called *invariant, equilibrium* or *stationary density*. That is, it holds that $\mathcal{P}\pi = \pi$ and thus π is an eigenfunction of $\mathcal{P}$ with corresponding eigenvalue 1. If the expectation of the process is the same at any time, then $\mathbb{E}[X_{t_2}] = \mathbb{E}[X_{t_1}]$ and specifying τ completely characterizes the covariances between X_t and $X_{t+\tau}$, the process is called *wide-sense stationary* (see for example Definition 3.6.9 in Gallager, 2013). For the following definition, we assume that there is a unique invariant density $\pi > 0$, which for molecular dynamics problems is given by the Boltzmann distribution $\pi \sim \exp(-\beta V)$, see Schütte and Sarich (2013). We now define the Perron–Frobenius operator reweighted with respect to this invariant density. The advantage of the reweighted Perron–Frobenius operator is that it can easily be estimated from long equilibrated trajectories, while other densities require, for instance, generating many short trajectories whose starting points are sampled from the corresponding probability distribution. For more details and examples, see Nüske et al. (2014), Klus et al. (2017).

Definition 3.2 (Transfer operators cont'd)**.** *Let $u_t(x) = \pi(x)^{-1} p_t(x)$ be a probability density with respect to the equilibrium density π.*

(iii) The Perron–Frobenius operator with respect to the equilibrium density*, denoted by $\mathcal{T}$, is defined as*

$$\mathcal{T}u_t(y) = \frac{1}{\pi(y)} \int p_\tau(y \mid x)\, \pi(x)\, u_t(x)\, \mathrm{d}x.$$

Under certain conditions, the transfer operators can be defined on other spaces L^r and $L^{r'}$, with $r \neq 1$ and $r' \neq \infty$, see Baxter and Rosenthal (1995), Klus et al. (2016). The operators $\mathcal{P}$ and $\mathcal{K}$ are adjoint to each other with respect to $\langle \cdot, \cdot \rangle$, defined by $\langle f, g \rangle = \int_{\mathbb{X}} f(x)\, g(x)\, \mathrm{d}x$, while $\mathcal{T}$ and $\mathcal{K}$ are adjoint with respect to $\langle \cdot, \cdot \rangle_\pi$, defined by $\langle f, g \rangle_\pi = \int_{\mathbb{X}} f(x)\, g(x)\, \pi(x)\, \mathrm{d}x$ for $f \in L^r_\pi(\mathbb{X})$ and $g \in L^{r'}_\pi(\mathbb{X})$ where $\frac{1}{r} + \frac{1}{r'} = 1$. That is, we have $\langle \mathcal{K}f, g \rangle_\pi = \langle f, \mathcal{T}g \rangle_\pi$.

Definition 3.3 (Reversibility)**.** *A system is called reversible if the* detailed balance condition

$$\pi(x)\, p_\tau(y \mid x) = \pi(y)\, p_\tau(x \mid y)$$

holds for all $x, y \in \mathbb{X}$.

If the system is reversible, then $\mathcal{K} = \mathcal{T}$. Moreover, the operators' eigenvalues λ_ℓ are real and the eigenfunctions φ_ℓ form an orthogonal basis with respect to the corresponding scalar product. As a result, the eigenvalues can be sorted in descending order so that $1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \dots$. The eigenfunctions determine the metastable sets of the system and the eigenvalues describe how fast the eigenfunctions converge to the invariant density. See Noé and Nüske (2013), Nüske et al. (2014), Klus et al. (2017) for more details.

Example 3.4. As a guiding example, we will use a simple one-dimensional Ornstein–Uhlenbeck process, given by the stochastic differential equation

$$\mathrm{d}X_t = -\alpha D X_t \,\mathrm{d}t + \sqrt{2D}\,\mathrm{d}W_t,$$

where α is the friction coefficient, $D = \beta^{-1}$ the diffusion coefficient, and $\{W_t\}_{t\geq 0}$ a one-dimensional standard Wiener process. The parameter β is also called the inverse temperature. The transition density of the Ornstein–Uhlenbeck process is

$$p_\tau(y \mid x) = \frac{1}{\sqrt{2\,\pi\,\sigma^2(\tau)}} \exp\left(-\frac{(y - x\,\exp(-\alpha D\tau))^2}{2\,\sigma^2(\tau)}\right),$$

with $\sigma^2(\tau) = \alpha^{-1}\,(1 - \exp(-2\alpha D\tau))$. For this simple dynamical system, the eigenfunctions can be computed analytically. See Pavliotis (2014), Klus et al. (2017) for more details. ▲

Remark 3.5. Definition 3.1 introduces the stochastic Koopman operator. For a deterministic dynamical system of the form $\dot{x} = F(x)$, we obtain $p_\tau(y \mid x) = \delta_{\boldsymbol{\Phi}_\tau(x)}(y)$, where $\boldsymbol{\Phi}_\tau$ denotes the flow map and δ_x the Dirac distribution centered in x. Thus, $\mathcal{K}f = f \circ \boldsymbol{\Phi}_\tau$. For a discrete dynamical system of the form $x_{i+1} = F(x_i)$, we obtain $\mathcal{K}f = f \circ F$. Note that in this case there is no implicit dependence on τ, the Koopman operator simply determines the observable mapped forward by the dynamical system. In the same way, the Perron–Frobenius operator can be defined for deterministic systems, see, e.g., Lasota and Mackey (1994), Koltai (2010), Klus et al. (2016).

Example 3.6. Consider the discrete dynamical system $F\colon \mathbb{R}^2 \to \mathbb{R}^2$, taken from Tu et al. (2014), with

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \begin{bmatrix} a\,x_1 \\ b\,x_2 + (b - a^2)\,x_1^2 \end{bmatrix}.$$

For the numerical experiments, we set $a = 0.8$ and $b = 0.7$. The eigenvalues of the Koopman operator associated with the system are $\lambda_1 = 1$, $\lambda_2 = a$, and $\lambda_3 = b$ with corresponding eigenfunctions $\varphi_1(x) = 1$, $\varphi_2(x) = x_1$, and $\varphi_3(x) = x_2 + x_1^2$. Furthermore, products of eigenfunctions are again eigenfunctions, for instance, $\varphi_4(x) = \varphi_2(x)^2 = x_1^2$ with eigenvalue $\lambda_4 = \lambda_2^2 = a^2$. Note that the ordering of the eigenvalues and eigenfunctions depends on the values of a and b. ▲

Given the eigenvalues and eigenfunctions of the Koopman operator, we can predict the evolution of the dynamical system. To this end, let $g(x) = x$ be the *full-state observable*. We then write $g(x)$ in terms of the eigenfunctions as

$$g(x) = x = \sum_\ell \varphi_\ell(x)\,\eta_\ell.$$

The vectors η_ℓ are called *Koopman modes*. Defining the Koopman operator to act componentwise for vector-valued functions, we obtain

$$\mathcal{K}g(x) = \mathbb{E}[g(X_\tau) \mid X_0 = x] = \sum_\ell \lambda_\ell(\tau)\,\varphi_\ell(x)\,\eta_\ell.$$

Example 3.7. For the simple deterministic system introduced in Example 3.6, we obtain the Koopman modes $\eta_1 = [0, 0]^\top$, $\eta_2 = [1, 0]^\top$, $\eta_3 = [0, 1]^\top$, and $\eta_4 = [0, -1]^\top$ so that $g(x) = \sum_{\ell=1}^{4} \varphi_\ell(x)\,\eta_\ell$ and

$$\mathcal{K}g(x) = \sum_{\ell=1}^{4} \lambda_\ell\,\varphi_\ell(x)\,\eta_\ell = \begin{bmatrix} \lambda_2\,\varphi_2(x) \\ \lambda_3\,\varphi_3(x) - \lambda_4\,\varphi_4(x) \end{bmatrix} = F(x). \qquad \blacktriangle$$

The above example illustrates that with the aid of the Koopman decomposition into eigenvalues, eigenfunctions, and modes, we can now evaluate the dynamical system at any data point. This is particularly useful if the system is not known explicitly. The Koopman representation of the system can be learned from training data as shown in Budišić et al. (2012), Williams et al. (2015a).

4 Transfer Operators in RKHS

In this section, we express the transfer operators introduced in Section 3 in terms of the covariance and cross-covariance operators defined on some RKHS $\mathbb{H}$ (see Section 2.2). For the transfer operators, the input and output spaces and thus also the kernels and resulting Hilbert spaces are identical, i.e., $\mathbb{X} = \mathbb{Y}$, $k = l$, and $\mathbb{H} = \mathbb{G}$. However, note that X and Y may be distinct random variables. For example, if $X \sim \delta_x$, then $Y \sim p_\tau(\cdot \mid x)$. In addition to the standard transfer operators, we will derive transfer operators for embedded densities and observables and analyze the relationships between them. To this end, we define—similar to the standard Gram matrices G_{XX} and G_{YY}—the *time-lagged Gram matrices* $G_{XY} = \Phi^\top\Psi$ and $G_{YX} = \Psi^\top\Phi$. In what follows, we assume that the Gram matrices and time-lagged Gram matrices are invertible or regularized in such a way that they become invertible.

We want to stress that especially the embedded operators can be used to describe any wide-sense stationary stochastic process, even if the marginal distribution at any time t does not allow a density. This opens up the possibility to analyze stochastic processes over, for example, string and graph domains, which do not admit densities since the domain is discrete. The main assumption is that a positive definite kernel function exists which enables measuring similarity of domain elements.

4.1 Kernel Perron–Frobenius Operator

Recall that the Perron–Frobenius operator $\mathcal{P}$ pushes forward any density p_t at time t to the density after the system has evolved for time τ. This density is denoted by $p_{t+\tau}$. Now we consider the *kernel Perron–Frobenius operator* $\mathcal{P}_k$ defined on the RKHS $\mathbb{H}$ induced by the kernel k. That is, we assume that $p_t \in \mathbb{H}$ and $p_{t+\tau} \in \mathbb{H}$. In general, this will not be the case and the question is how and under which conditions this operator approximates the Perron–Frobenius operator defined on L^1. In this paper, we will develop the framework for representing transfer operators using RKHS theory and compare it with other existing methods. The convergence properties are beyond the scope of this paper and will be studied in future work. For kernels with explicitly given feature spaces, related convergence results can be found in Williams et al. (2015a), Klus et al. (2016), Korda and Mezić (2017). Also the relationships with the *projected transfer operators* defined in Schütte and Sarich (2013) will be studied subsequently. Notice that the assumption that the relevant densities are in

$\mathbb{H}$ is different from the embedding approach discussed in Sections 2.1 and 2.3. A Perron–Frobenius operator for embedded densities is derived in Section 4.2.

Proposition 4.1. *Let $p_{\mathbb{X}}$ be the reference density on $\mathbb{X}$ and let $\mathcal{A}_k \colon \mathbb{H} \to \mathbb{H}$ be the kernel transfer operator with respect to this density, i.e., $\mathcal{A}_k g(y) = \frac{1}{p_{\mathbb{X}}(y)} \int p_\tau(y \mid x)\, g(x)\, p_{\mathbb{X}}(x)\, \mathrm{d}x$ for $g \in \mathbb{H}$. Then*

$$\mathcal{C}_{XX} \mathcal{A}_k g = \mathcal{C}_{YX} g.$$

Proof. The proof is similar to the proof of Proposition 2.11, which can be found, e.g., in Fukumizu et al. (2004). Using (1), it holds that

$$\begin{aligned}
\langle f, \mathcal{C}_{XX} \mathcal{A}_k g \rangle_{\mathbb{H}} &= \mathbb{E}_X [f(X)\, \mathcal{A}_k g(X)] \\
&= \int f(y) \frac{1}{p_{\mathbb{X}}(y)} \int p_\tau(y \mid x)\, p_{\mathbb{X}}(x)\, g(x)\, \mathrm{d}x\, p_{\mathbb{X}}(y)\, \mathrm{d}y \\
&= \iint f(y)\, g(x)\, p_\tau(y \mid x)\, p_{\mathbb{X}}(x)\, \mathrm{d}x\, \mathrm{d}y \\
&= \iint f(y)\, g(x)\, p_\tau(x, y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \mathbb{E}_{XY} [g(X)\, f(Y)] \\
&= \langle f, \mathcal{C}_{YX} g \rangle_{\mathbb{H}}\,.
\end{aligned}$$

■

It follows that $\mathcal{A}_k = \mathcal{C}_{XX}^{-1} \mathcal{C}_{YX}$, where a regularized version of $\mathcal{C}_{XX}$ might be required as described above. Let $u_{\mathbb{X}}$ denote the uniform density on $\mathbb{X}$ and π the invariant density defined in Section 3. If $p_{\mathbb{X}} = u_{\mathbb{X}}$, then $\mathcal{A}_k = \mathcal{P}_k$ and if $p_{\mathbb{X}} = \pi$, then $\mathcal{A}_k = \mathcal{T}_k$, where $\mathcal{T}_k$ denotes the kernel Perron–Frobenius operator with respect to the invariant density. We will later generate training data using these densities. It is important to note here that $\mathbb{X}$ and $\mathbb{Y}$ as well as $\mathbb{H}$ and $\mathbb{G}$ have to be the same spaces, otherwise the operator would be undefined since $\mathcal{C}_{YX}$ is a mapping from $\mathbb{H}$ to $\mathbb{G}$ and $\mathcal{C}_{XX}^{-1}$ a mapping from $\mathbb{H}$ to $\mathbb{H}$.

Corollary 4.2. *For specific choices of $p_{\mathbb{X}}$, we obtain:*

(i) If $p_{\mathbb{X}} = u_{\mathbb{X}}$, then $\mathcal{P}_k = \mathcal{C}_{XX}^{-1} \mathcal{C}_{YX}$.

(ii) If $p_{\mathbb{X}} = \pi$, then $\mathcal{T}_k = \mathcal{C}_{XX}^{-1} \mathcal{C}_{YX}$.

This is consistent with the derivation of EDMD for the Perron–Frobenius operator in Klus et al. (2016), where—from a kernel point of view—explicitly given finite-dimensional feature spaces are considered. In this case, the empirical estimates of the operators converge to a Galerkin approximation, i.e., the operator projected onto the space spanned by the feature map functions. For the Koopman operator, this was shown in Williams et al. (2015a).

Proposition 4.3. *The empirical estimate $\widehat{\mathcal{P}}_k$ of the kernel Perron–Frobenius operator $\mathcal{P}_k$ can be written as $\widehat{\mathcal{P}}_k = \Psi A \Phi^\top$, where A is a real matrix. Estimates for A are given by $A_1 = G_{XY}^{-1} G_{XX}^{-1} G_{XY}$ and $A_2 = G_{XY}^{-1} G_{YX}^{-1} G_{YY}$.*

Proof. The idea is to simply solve the equation $\widehat{\mathcal{P}}_k = \widehat{\mathcal{C}}_{XX}^{-1} \widehat{\mathcal{C}}_{YX} = (\Phi\Phi^\top)^{-1} \Psi \Phi^\top = \Psi A \Phi^\top$ for A. Dropping the $\Phi^\top$, we multiply the equations from the left by $\Phi\Phi^\top$ and then by

(i) $\Phi^\top$, which leads to $G_{XY} = G_{XX} G_{XY} A$ yielding the estimator A_1,

(ii) $\Psi^\top$, which leads to $G_{YY} = G_{YX}\, G_{XY} A$ yielding the estimator A_2. ■

We conjecture that both estimators are identical when taking the number of data points to infinity, assuming common support of p_t and $p_{t+\tau}$. Using the reproducing property of $\mathbb{H}$ and assuming that $p \in \mathbb{H}$, we can write

$$\mathcal{P}_k\, p(x) = \left\langle \mathcal{C}_{XX}^{-1} \mathcal{C}_{YX}\, p,\, k(x,\cdot) \right\rangle_{\mathbb{H}} = \left\langle p,\, \mathcal{C}_{XY}\, \mathcal{C}_{XX}^{-1} k(x,\cdot) \right\rangle_{\mathbb{H}} = \left\langle p,\, \mathcal{K}_{\mathcal{E}}\, k(x,\cdot) \right\rangle_{\mathbb{H}},$$

where $\mathcal{K}_{\mathcal{E}} = \mathcal{C}_{XY}\, \mathcal{C}_{XX}^{-1}$. Thus, the action of the Perron–Frobenius operator can be interpreted as an inner product in a Hilbert space. We will call $\mathcal{K}_{\mathcal{E}}$ the *embedded Koopman operator* and discuss it in detail in Section 4.4.

4.2 Embedded Perron–Frobenius Operator

In the previous subsection, we assumed that the densities p_t and $p_{t+\tau}$ are elements of the RKHS $\mathbb{H}$. Now we first embed the densities into the RKHS $\mathbb{H}$ using the mean embedding and consider the corresponding embedded densities μ_t and $\mu_{t+\tau}$. Since the definition of the Perron–Frobenius operator resembles the sum rule, we can extend it to the RKHS using the kernel sum rule (Song et al., 2013, Fukumizu et al., 2013). Let $\mu_t = \mathbb{E}_{p_t}[k(X,\cdot)] = \mathcal{E}_k\, p_t$ be a Hilbert space embedding of the density p_t, then the Perron–Frobenius operator for embedded densities can be expressed in terms of the conditional mean embedding $\mathcal{U}_{Y|X}$ as

$$\mu_{t+\tau} = \mathcal{U}_{Y|X}\, \mu_t = \mathcal{C}_{YX}\, \mathcal{C}_{XX}^{-1}\, \mu_t,$$

where $\mu_{t+\tau}$ is the Hilbert space embedding of the density $p_{t+\tau}$. The above equality is guaranteed under the assumption that $\mathcal{C}_{XX}$ is injective, $\mu_t \in \mathrm{Range}(\mathcal{C}_{XX})$, and $\mathbb{E}_{Y|X}[g(Y) \mid X = \cdot\,] \in \mathbb{H}$ for all $g \in \mathbb{H}$ (see also Fukumizu et al. (2013, Theorem 2)). Thus, we define $\mathcal{P}_{\mathcal{E}} = \mathcal{U}_{Y|X}$ to be the embedded Perron–Frobenius operator. The empirical estimate of the embedded Perron–Frobenius operator is given by

$$\widehat{\mathcal{P}}_{\mathcal{E}} = \widehat{\mathcal{C}}_{YX}\, \widehat{\mathcal{C}}_{XX}^{-1} = (\Psi\Phi^\top)(\Phi\Phi^\top)^{-1} = \Psi\, G_{XX}^{-1} \Phi^\top.$$

If the Gram matrix G_{XX} is not invertible, we may resort to the regularized estimate, given by $\widehat{\mathcal{P}}_{\mathcal{E}} = \Psi\, (G_{XX} + n\varepsilon I_n)^{-1}\, \Phi^\top$.

Proposition 4.4. *Let $\mu_t := \mathcal{E}_k\, p_t$ be an embedded probability density. Then the diagram*

$$\begin{array}{ccc} L^1 \ni p_t & \xrightarrow{\;\mathcal{E}_k\;} & \mu_t \in \mathbb{H} \\ \Big\downarrow{\scriptstyle \mathcal{P}} & & \Big\downarrow{\scriptstyle \mathcal{P}_{\mathcal{E}}} \\ L^1 \ni p_{t+\tau} & \xrightarrow{\;\mathcal{E}_k\;} & \mu_{t+\tau} \in \mathbb{H} \end{array}$$

is commutative.

Proof. Applying $\mathcal{P}$ to p_t and then embedding the resulting density leads to

$$\mathcal{E}_k(\mathcal{P} p_t) = \int k(y,\cdot) \int p_\tau(y \mid x)\, p_t(x)\, \mathrm{d}x\, \mathrm{d}y,$$

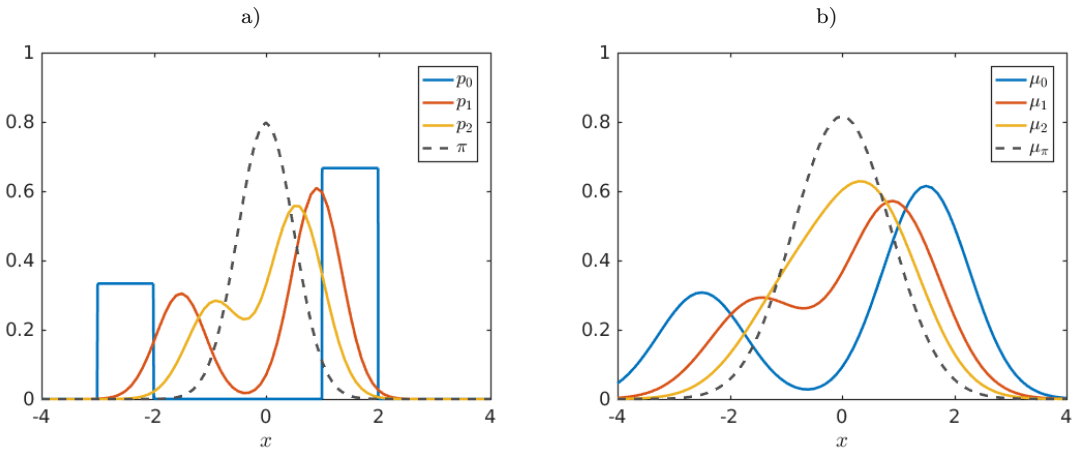

Figure 1: a) Propagation of the initial density p_0 by the Perron–Frobenius operator, where $p_1 = \mathcal{P}p_0$ and $p_2 = \mathcal{P}p_1$. b) Propagation of the embedded density μ_0 by the embedded Perron–Frobenius operator, where $\mu_1 = \mathcal{P}_\mathcal{E}\mu_0$ and $\mu_2 = \mathcal{P}_\mathcal{E}\mu_1$. The dashed black lines show the invariant and embedded invariant density, respectively.

embedding p_t and then applying the embedded Perron–Frobenius operator to

$$\begin{aligned}
\mathcal{P}_\mathcal{E}(\mathcal{E}_k\, p_t) &= \mathcal{P}_\mathcal{E} \int k(x,\cdot)\, p_t(x)\,\mathrm{d}x \\
&= \int \mathcal{P}_\mathcal{E}\, k(x,\cdot)\, p_t(x)\,\mathrm{d}x \\
&= \int \mathbb{E}_{Y|x}[\phi(Y) \mid X = x]\, p_t(x)\,\mathrm{d}x \\
&= \iint p_\tau(y \mid x)\, \phi(y)\,\mathrm{d}y\, p_t(x)\,\mathrm{d}x \\
&= \int k(y,\cdot) \int p_\tau(y \mid x)\, p_t(x)\,\mathrm{d}x\,\mathrm{d}y. \qquad \blacksquare
\end{aligned}$$

For the empirical estimates, the commutativity can be seen as follows: Let p_t be a probability density, then the empirical estimate of the kernel mean embedding is $\widehat{\mu}_t = \frac{1}{n}\Phi\mathbb{1}$. Applying $\widehat{\mathcal{P}}_\mathcal{E}$ yields

$$\widehat{\mathcal{P}}_\mathcal{E}\,\widehat{\mu}_t = \tfrac{1}{n}\widehat{\mathcal{C}}_{YX}\,\widehat{\mathcal{C}}_{XX}^{-1}\Phi\mathbb{1} = \tfrac{1}{n}\Psi\Phi^\top(\Phi\Phi^\top)^{-1}\Phi\mathbb{1} = \tfrac{1}{n}\Psi\mathbb{1} = \widehat{\mu}_{t+\tau}.$$

That is, we obtain the empirical estimate of the mean embedding of the density $p_{t+\tau}$. In the last step, we again used the identity $\Phi(\Phi^\top\Phi)^{-1} = (\Phi\Phi^\top)^{-1}\Phi$.

Example 4.5. Let us consider the Ornstein–Uhlenbeck process from Example 3.4. We choose $\tau = \frac{1}{2}$, $\alpha = 4$, $D = \frac{1}{4}$, and the Gaussian kernel with $\sigma^2 = \frac{1}{2}$. Figure 1a shows the piecewise constant initial probability density p_0 pushed forward by the Perron–Frobenius operator, Figure 1b the embedded initial density μ_0 pushed forward by the embedded Perron–Frobenius operator. ▲

The Perron–Frobenius operator $\mathcal{P}$ maps densities $p_t \in L^1$ to $p_{t+\tau} \in L^1$, while the embedded Perron–Frobenius operator $\mathcal{P}_\mathcal{E}$, given by the conditional mean embedding $\mathcal{U}_{Y|X}$, maps

embedded densities $\mu_t \in \mathbb{H}$ to $\mu_{t+\tau} \in \mathbb{H}$. Thus, the conditional mean embedding plays a similar role as the classical Perron–Frobenius operator in that it pushes forward—in this case: embedded—densities. Note that if we embed an eigenfunction of $\mathcal{P}$, we automatically obtain an eigenfunction of $\mathcal{P}_\mathcal{E}$. This is due to the linearity of the integral.

4.3 Kernel Koopman Operator

The Koopman operator $\mathcal{K}$ applied to an observable f evaluated at x results in the expectation of f when starting in x and evolving the system for time τ. Analogously to the kernel Perron–Frobenius operator, we now introduce the corresponding *kernel Koopman operator*, denoted by $\mathcal{K}_k$. That is, we assume that the observables and the observables mapped forward by the Koopman operator are elements of $\mathbb{H}$. From Proposition 2.11, it follows that

$$\mathcal{K}_k = \mathcal{C}_{XX}^{-1}\mathcal{C}_{XY}$$

and thus for $f \in \mathbb{H}$ that

$$\mathcal{K}_k f(x) = \left\langle \mathcal{C}_{XX}^{-1}\mathcal{C}_{XY} f,\, k(x,\cdot)\right\rangle = \left\langle f,\, \mathcal{P}_\mathcal{E}\, k(x,\cdot)\right\rangle.$$

Alternatively, the kernel Koopman operator can be derived using the reproducing property directly:

$$\begin{aligned}\mathcal{K}_k f(x) &= \int p_\tau(y \mid x)\, f(y)\,\mathrm{d}y = \int \langle f,\, k(y,\cdot)\rangle\, p_\tau(y \mid x)\,\mathrm{d}y \\ &= \left\langle f,\, \int k(y,\cdot)\, p_\tau(y \mid x)\,\mathrm{d}y \right\rangle = \langle f,\, \mathcal{P}_\mathcal{E}\, k(x,\cdot)\rangle\,.\end{aligned}$$

The empirical estimate of the Koopman operator is then given by

$$\widehat{\mathcal{K}}_k = \widehat{\mathcal{C}}_{XX}^{-1}\widehat{\mathcal{C}}_{XY} = (\Phi\Phi^\top)^{-1}(\Phi\Psi^\top) = \Phi\, G_{XX}^{-1}\Psi^\top.$$

Example 4.6. Let us approximate the kernel Koopman operator associated with the system defined in Example 3.6 using the kernel from Example 2.2. Generating 10000 test points x_i sampled from a uniform distribution on $\mathbb{X} = [-2,\,2] \times [-2,\,2]$ and the corresponding $y_i = F(x_i)$ values, we can compute the empirical estimator

$$\widehat{\mathcal{K}}_k = \widehat{\mathcal{C}}_{XX}^{-1}\widehat{\mathcal{C}}_{XY} = \left(\Phi\Phi^\top\right)^{-1}\left(\Phi\Psi^\top\right) = \left(\Psi\Phi^+\right)^\top \in \mathbb{R}^{6\times 6}.$$

Here, $^+$ denotes the pseudoinverse. The dominant eigenvalues and right eigenvectors as well as the corresponding eigenfunctions are given by

$$\begin{aligned}
&\lambda_1 = 1.0, \quad v_1 = [1 \quad 0 \quad\;\; 0 \quad\;\; 0 \quad 0 \quad 0]^\top, \quad \varphi_1(x) = \langle v_1,\, \phi(x)\rangle = 1,\\
&\lambda_2 = 0.8, \quad v_2 = [0 \quad 0.7071 \quad 0 \quad\;\; 0 \quad 0 \quad 0]^\top, \quad \varphi_2(x) = \langle v_2,\, \phi(x)\rangle \approx x_1,\\
&\lambda_3 = 0.7, \quad v_3 = [0 \quad 0 \quad\;\; 0.7071 \quad 1 \quad 0 \quad 0]^\top, \quad \varphi_3(x) = \langle v_3,\, \phi(x)\rangle \approx x_2 + x_1^2.
\end{aligned}$$

This is in good agreement with the analytically computed results. The eigenfunctions are clearly in $\mathbb{H}$ since they can be written as linear combinations of $\phi(x_i)$ for appropriately chosen x_i, e.g., $v_2 = \phi\left([\tfrac{1}{4}, 0]^\top\right) - \phi\left([-\tfrac{1}{4}, 0]^\top\right)$. The subsequent eigenvalues and eigenfunctions are simply products of the eigenvalues and eigenfunctions listed above, see Example 3.6. Note, however, that other than φ_4 further products of eigenfunctions cannot be represented as functions in $\mathbb{H}$ anymore since the feature space does not contain polynomials of order greater than two. Here, we used the explicit feature map of the polynomial kernel. ▲

The approach to obtain an approximation of transfer operators from data as described in the above example is also referred to as EDMD (Williams et al., 2015a, Klus et al., 2016). The data matrices are embedded into a typically high-dimensional feature space, given, for instance, by monomials up to a certain order, which corresponds to the feature space of a polynomial kernel. Details regarding the relationships with other methods can be found in Section 4.7 and further examples in Section 5.

4.4 Embedded Koopman Operator

If we want to embed the Koopman operator in the same way as the Perron–Frobenius operator, we need to introduce embedded observables first, which can be interpreted as the counterpart of the mean embedding of distributions. Let $f\colon \mathbb{X} \to \mathbb{R}$ be an observable of the system. We define $\nu := \mathcal{E}_k f$ to be the *embedded observable*. Given a set of training data, the empirical estimate $\widehat{\nu}$ of the embedded observable is given by

$$\widehat{\nu} = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)\, f(x_i) = \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot)\, f(x_i) = \frac{1}{n}\Phi\widehat{f},$$

where $\widehat{f} = [f(x_1), \ldots, f(x_n)]^\top$ contains the values of the observable evaluated at the training data points. Note that the data points do not have to be drawn from a particular probability distribution. Alternatively, we could perform regression to approximate the observable f by an element $\widetilde{f} \in \mathbb{H}$ and then compute the integral.

Remark 4.7. Let ν be the embedding of the observable f with respect to the kernel k and let p be a density lying in the RKHS $\mathbb{H}$ spanned by k. Then $\langle \nu, p\rangle_{\mathbb{H}} = \int f(x)\, \langle k(x,\cdot), p\rangle_{\mathbb{H}}\, \mathrm{d}x = \int p(x) f(x)\, \mathrm{d}x$.

This result is an analogue of what has been attained previously for the kernel mean embedding of distributions, which also allows the representation of integration as an inner product in the RKHS, see Remark 2.6. Note that when using embedded distributions we have to assume that the observable is in $\mathbb{H}$, while when using embedded observables we assume the relevant density to be in $\mathbb{H}$. One possible use of embedded observables arising from Remark 4.7 is when one is unsure of the RKHS an observable lies in, while the RKHS of the density of interest is given. Another might be that by embedding the observable instead of the distribution one can take advantage of a smoother RKHS.

Analogously to the embedded Perron–Frobenius operator, we define the *embedded Koopman operator* by $\mathcal{K}_\mathcal{E} = \mathcal{C}_{XY}\mathcal{C}_{XX}^{-1}$.

Proposition 4.8. *Let $\nu_t := \mathcal{E}_k f_t$ be the embedded observable. Then the diagram*

$$\begin{array}{ccc} L^\infty \ni f_t & \xrightarrow{\mathcal{E}_k} & \nu_t \in \mathbb{H} \\ \mathcal{K}\big\downarrow & & \big\downarrow \mathcal{K}_\mathcal{E} \\ L^\infty \ni f_{t+\tau} & \xrightarrow{\mathcal{E}_k} & \nu_{t+\tau} \in \mathbb{H} \end{array}$$

is commutative.

Proof. The proof is similar to the proof of Proposition 4.4. ∎

As for $\mathcal{P}$, embedding an eigenfunction of $\mathcal{K}$ results in an eigenfunction of $\mathcal{K}_\mathcal{E}$. For the kernel Koopman operator $\mathcal{K}_k$, the commutativity can be seen as follows: Assume that the observable is given by $f_t \in \mathbb{H}$. Thus, $\mathcal{K}_k f_t = \mathcal{C}_{XX}^{-1}\mathcal{C}_{XY} f_t$, while the embedding of f_t results in $\nu_t = \mathcal{C}_{XX} f_t$. Applying $\mathcal{K}_\mathcal{E}$, this results in

$$\mathcal{K}_\mathcal{E}(\mathcal{E}_k f_t) = \mathcal{C}_{XY}\mathcal{C}_{XX}^{-1}\mathcal{C}_{XX} f_t = \mathcal{C}_{XX}\mathcal{C}_{XX}^{-1}\mathcal{C}_{XY} f_t = \mathcal{E}_k(\mathcal{K}_k f_t).$$

Proposition 4.9. *The empirical estimate $\widehat{\mathcal{K}}_\mathcal{E}$ of the embedded Koopman operator $\mathcal{K}_\mathcal{E}$ can be written as $\widehat{\mathcal{K}}_\mathcal{E} = \Phi A \Psi^\top$, where A is a real matrix. Estimates for A are given by $A_1 = G_{YX}G_{XX}^{-1}G_{YX}^{-1}$ and $A_2 = G_{YY}G_{XY}^{-1}G_{YX}^{-1}$.*

Proof. The proof is analogous to the proof of Proposition 4.3. ■

Example 4.10. Let us consider again the system from Example 3.6 whose eigenfunctions φ_1, φ_2, and φ_3 we estimated numerically in Example 4.6 using the kernel from Example 2.2. Computing the corresponding embedded eigenfunctions analytically, we obtain the (properly rescaled) functions ν_1, ν_2, and ν_3 and associated vector representations w_1, w_2, and w_3, given by

$$\begin{aligned}
&\lambda_1 = 1.0, \quad w_1 = [\,3 \quad 0 \quad 0 \quad 4 \quad 0 \quad 4\,]^\top, \quad \nu_1(x) = 3 + 4x_1^2 + 4x_2^2,\\
&\lambda_2 = 0.8, \quad w_2 = [\,0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0\,]^\top, \quad \nu_2(x) = \sqrt{2}x_1,\\
&\lambda_3 = 0.7, \quad w_3 = [\,1 \quad 0 \quad \sqrt{2} \quad \tfrac{12}{5} \quad 0 \quad \tfrac{4}{3}\,]^\top, \quad \nu_3(x) = 1 + 2z_2 + \tfrac{12}{5}z_1^2 + \tfrac{4}{3}z_2^2.
\end{aligned}$$

The vectors w_1, w_2, and w_3 are indeed eigenvectors of the matrix $\widehat{\mathcal{K}}_\mathcal{E} = \widehat{\mathcal{C}}_{XY}\widehat{\mathcal{C}}_{XX}^{-1}$ corresponding to the eigenvalues λ_1, λ_2, and λ_3. ▲

4.5 Relationships between Operators

Overall, we derived four different operators that can be written in terms of the covariance and cross-covariance operators introduced in Section 2.

	Kernel operator	Embedded operator
Perron–Frobenius	$\mathcal{P}_k = \mathcal{C}_{XX}^{-1}\mathcal{C}_{YX}$ $\approx \Psi\, A\, \Phi^\top$	$\mathcal{P}_\mathcal{E} = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$ $\approx \Psi\, G_{XX}^{-1}\, \Phi^\top$
Koopman	$\mathcal{K}_k = \mathcal{C}_{XX}^{-1}\mathcal{C}_{XY}$ $\approx \Phi\, G_{XX}^{-1}\, \Psi^\top$	$\mathcal{K}_\mathcal{E} = \mathcal{C}_{XY}\mathcal{C}_{XX}^{-1}$ $\approx \Phi\, A^\top\, \Psi^\top$

We can express the kernel Perron–Frobenius operator using the embedded Koopman operator and the kernel Koopman operator using the embedded Perron–Frobenius operator—or mean embedding—via

$$\begin{aligned}
\mathcal{P}_k p(x) &= \langle \mathcal{P}_k p,\, k(x,\cdot)\rangle_\mathbb{H} = \langle p,\, \mathcal{K}_\mathcal{E}\, k(x,\cdot)\rangle_\mathbb{H}\,,\\
\mathcal{K}_k f(x) &= \langle \mathcal{K}_k f,\, k(x,\cdot)\rangle_\mathbb{H} = \langle f,\, \mathcal{P}_\mathcal{E}\, k(x,\cdot)\rangle_\mathbb{H}\,.
\end{aligned}$$

That is, $\mathcal{P}_k$ and $\mathcal{K}_\mathcal{E}$ as well as $\mathcal{K}_k$ and $\mathcal{P}_\mathcal{E}$ are adjoint to each other with respect to the inner product in $\mathbb{H}$. This is also reflected in the empirical estimators. Here, A is either $A_1 = G_{XY}^{-1}G_{XX}^{-1}G_{XY}$ or $A_2 = G_{XY}^{-1}G_{YX}^{-1}G_{YY}$.

4.6 Eigendecomposition of RKHS Operators

If the feature space is finite-dimensional and known explicitly, we can compute eigenfunctions as shown in Example 4.6, provided that the dimension of the feature space is small enough so that the resulting eigenvalue problem can still be solved numerically. The advantage of this approach is that the matrix size does not depend on the number of test points n. As described above, this approach converges to a Galerkin approximation of the respective operator for $n \to \infty$. The basis functions for the Galerkin ansatz are given by the feature map.

Now, we want to consider also the cases where the dimension of the feature space is larger than the number of test points or where the feature space is even infinite-dimensional. Let $\mathcal{S} = \Upsilon B \Gamma^\top$ be a Hilbert–Schmidt operator mapping from $\mathbb{H}$ to itself, with $\Upsilon = [k(z_1, \cdot), \ldots, k(z_n, \cdot)]$, $\Gamma = [k(z'_1, \cdot), \ldots, k(z'_n, \cdot)]$, and $B \in \mathbb{R}^{n \times n}$ for some n. Assume further that $\bigcup_{i=1}^n \{z_i, z'_i\}$ contains only pairwise different objects. Then the eigenvalues and eigenfunctions of $\mathcal{S}$ can be computed from eigenvalues and eigenvectors of $G_{\Gamma\Upsilon} B$ or $B G_{\Gamma\Upsilon}$, where $G_{\Gamma\Upsilon} = \Gamma^\top \Upsilon$.

Proposition 4.11. *The Hilbert–Schmidt operator $\mathcal{S} = \Upsilon B \Gamma^\top$ has an eigenvalue $\lambda \neq 0$ with corresponding eigenfunction*

$$\upsilon = \Upsilon \mathbf{v}$$

if and only if $\mathbf{v}$ is an eigenvector of $B G_{\Gamma\Upsilon}$ associated with λ. Similarly, $\mathcal{S}$ has an eigenvalue $\lambda \neq 0$ with corresponding eigenfunction

$$\gamma = \Gamma G_{\Gamma\Gamma}^{-1} \mathbf{v}.$$

if and only if $\mathbf{v}$ is an eigenvector of $G_{\Gamma\Upsilon} B$.

Proof. Let $\upsilon = \Upsilon \mathbf{v}$ be an eigenfunction of $\mathcal{S}$ associated with λ. Then

$$\begin{aligned} \mathcal{S}\upsilon = \lambda\upsilon \quad &\Leftrightarrow \\ \Upsilon B G_{\Gamma\Upsilon} \mathbf{v} = \lambda \Upsilon \mathbf{v} \quad &\Leftrightarrow \\ B G_{\Gamma\Upsilon} \mathbf{v} = \lambda \mathbf{v}. \end{aligned}$$

For the second part, let $\gamma = \Gamma G_{\Gamma\Gamma}^{-1} \mathbf{v}$ be an eigenfunction of $\mathcal{S}$. Then

$$\begin{aligned} \mathcal{S}\gamma = \lambda\gamma \quad &\Leftrightarrow \\ \Upsilon B G_{\Gamma\Gamma} G_{\Gamma\Gamma}^{-1} \mathbf{v} = \lambda \Gamma G_{\Gamma\Gamma}^{-1} \mathbf{v} \quad &\Leftrightarrow \\ G_{\Gamma\Upsilon} B \mathbf{v} = \lambda \mathbf{v}. \end{aligned}$$

■

While these eigendecomposition derivations are the most elegant, other derivations exist. We conjecture that the eigenfunction expressions would coincide when taking the infinite-dimensional limit in the number of data points n (and thus in the size of B, Υ, and Γ).

Setting $\Upsilon = \Phi$ and $\Gamma = \Psi$ or $\Upsilon = \Psi$ and $\Gamma = \Phi$ as well as $B = A$ or $B = A^\top$, we thus obtain eigendecomposition expressions for all empirical operator estimates listed in Section 4.5. In particular, let $\mathcal{P}_* = \Psi B \Phi^\top$. Then we need to solve the eigenvalue problem $G_{XY} B \mathbf{v} = \lambda \mathbf{v}$ (which reduces to $G_{XX}^{-1} G_{XY} \mathbf{v} = \lambda \mathbf{v}$ for $\mathcal{P}_k$ with the estimator $B = A_1$). We obtain an eigenfunction of $\mathcal{P}_*$ as $\varphi = \Phi G_{XX}^{-1} \mathbf{v}$. For the Koopman operators, let $\mathcal{K}_* = \Phi B \Psi^\top$.

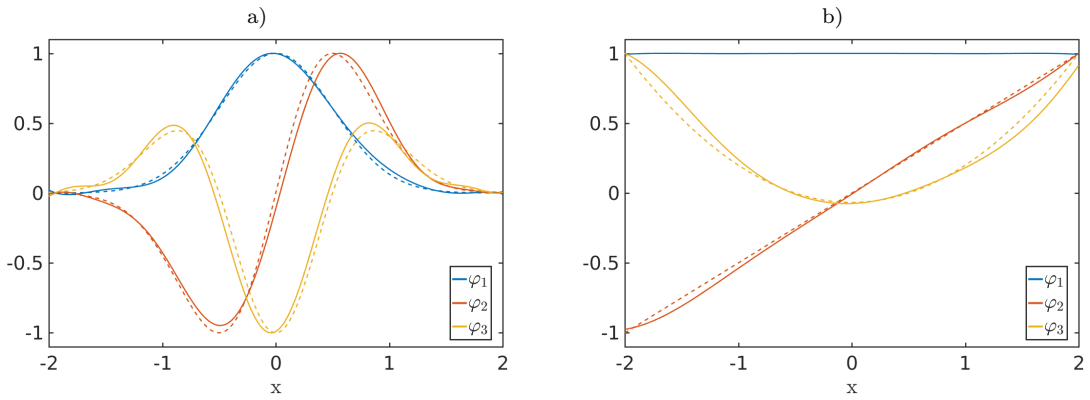

Figure 2: a) Dominant eigenfunctions of the Perron–Frobenius operator $\mathcal{P}$ associated with the Ornstein–Uhlenbeck process. b) Dominant eigenfunctions of the Koopman operator $\mathcal{K}$. The solid lines are the numerically computed and the dotted lines the analytically computed eigenfunctions. Also the numerically computed eigenvalues agree with the analytically computed values.

Then we solve $BG_{XY}\mathbf{v} = \lambda\mathbf{v}$ (i.e., $G_{XX}^{-1}G_{YX}\mathbf{v} = \lambda\mathbf{v}$ for $\mathcal{K}_k$). We get the eigenfunction $\varphi = \Phi\mathbf{v}$. We will use these methods, which are consistent with the derivations in Williams et al. (2015a), Klus et al. (2018), for the experiments in Section 5.

Example 4.12. Let us analyze the Ornstein–Uhlenbeck process introduced in Example 3.4. We use again $\tau = \frac{1}{2}$, $\alpha = 4$, and $D = \frac{1}{4}$ and generate 5000 uniformly distributed test points in $[-2, 2]$. Furthermore, we use the Gaussian kernel with $\sigma^2 = 0.3$. Applying our eigendecomposition result to the kernel Perron–Frobenius operator—since the test points are distributed uniformly, we obtain $\mathcal{P}_k$, see Corollary 4.2—and kernel Koopman operator yields the results shown in Figure 2. This special case is equivalent to the kernel EDMD method. A similar experiment using conventional EDMD with a basis comprising monomials is described in Klus et al. (2017). ▲

More complex examples from various application areas and different use cases will be discussed in Section 5.

4.7 Relationships with Other Methods

There are several existing methods such as *time-lagged independent component analysis* (TICA) (Molgedey and Schuster, 1994, Pérez-Hernández et al., 2013), *dynamic mode decomposition* (DMD) (Schmid, 2010, Tu et al., 2014), and their respective generalizations—the aforementioned VAC and EDMD—to approximate transfer operators and their eigenvalues, eigenfunctions, and eigenmodes. Although developed independently from each other, these methods are strongly related as shown in Klus et al. (2017). Our methods subsume existing ones and thereby provide a unified framework for transfer operator approximation using RKHS theory.

4.7.1 TICA and DMD

TICA can be used to separate superimposed signals (Molgedey and Schuster, 1994), solving the so-called *blind source separation* problem, and also for dimensionality reduction (Pérez-Hernández et al., 2013), by projecting a high-dimensional signal onto the main TICA coordinates (see Section 5 for an example). The method aims to find the time-lagged independent components that are uncorrelated and maximize the autocovariances at lag time τ. Given again training data $\mathbb{D}_{XY} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i = X_{t_i}$ and $y_i = X_{t_i+\tau}$, we define the associated data matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d\times n}$ by

$$\mathbf{X} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}.$$

By setting $k(x, x') = x^\top x'$ and $l(y, y') = y^\top y'$, the eigenvalue problem for the Koopman operator reduces to the standard eigenvalue problem

$$\widehat{\mathcal{C}}_{XX}^{-1} \widehat{\mathcal{C}}_{XY}\, \xi = \lambda \xi,$$

where $\widehat{\mathcal{C}}_{XX}$ and $\widehat{\mathcal{C}}_{XY}$ denote the covariance and cross-covariance matrices, respectively, defined by $\widehat{\mathcal{C}}_{XX} = \frac{1}{n}\mathbf{X}\mathbf{X}^\top$ and $\widehat{\mathcal{C}}_{XY} = \frac{1}{n}\mathbf{X}\mathbf{Y}^\top$. The resulting eigenvectors are defined to be the TICA coordinates.

DMD is frequently used for the analysis of high-dimensional fluid flow problems (Schmid, 2010). The DMD modes correspond to coherent structures in these flows. The derivation is based on the least-squares minimization problem $\|\mathbf{Y} - M\mathbf{X}\|_F$, whose solution is given by

$$M = \mathbf{Y}\mathbf{X}^+ = \left(\mathbf{Y}\mathbf{X}^\top\right)\left(\mathbf{X}\mathbf{X}^\top\right)^{-1} = \widehat{\mathcal{C}}_{YX}\widehat{\mathcal{C}}_{XX}^{-1}.$$

Eigenvectors of this matrix are then called DMD modes. Equivalently, the DMD modes can be interpreted as the left eigenvectors of the TICA matrix $\widehat{\mathcal{C}}_{XX}^{-1}\widehat{\mathcal{C}}_{XY}$. More details on the relationships between TICA and DMD can be found in Klus et al. (2017). As shown above, both TICA and DMD can be obtained as special cases of our algorithms.

4.7.2 VAC and EDMD

For a given set of basis functions $\phi_1, \ldots, \phi_r$, we define the vector-valued function $\phi = [\phi_1, \ldots, \phi_r]^\top : \mathbb{R}^d \to \mathbb{R}^r$. In the context of the kernel-based methods introduced above, the function ϕ corresponds to an explicitly defined feature map. This results in the feature matrices $\Phi, \Psi \in \mathbb{R}^{r\times n}$, given by

$$\Phi = \begin{bmatrix} \phi(x_1) & \cdots & \phi(x_n) \end{bmatrix} \quad \text{and} \quad \Psi = \begin{bmatrix} \phi(y_1) & \cdots & \phi(y_n) \end{bmatrix}.$$

Note that the same basis functions—and thus the same kernel—are used for x and y, a generalization of this approach can be found in Wu and Noé (2017). VAC and EDMD, which are equivalent for reversible dynamical systems, can be understood as nonlinear extensions of TICA and DMD, respectively. Both methods utilize the transformed data matrices Φ and Ψ for an explicitly given set of basis functions. VAC uses the matrix $\widehat{\mathcal{C}}_{XX}^{-1}\widehat{\mathcal{C}}_{XY}$ as an approximation of $\mathcal{T}$ (which is equivalent to $\mathcal{K}$ for a reversible system) to compute eigenfunctions. Similarly, EDMD considers the matrix $\widehat{\mathcal{C}}_{YX}\widehat{\mathcal{C}}_{XX}^{-1}$, which can be interpreted as a least-square approximation of the Koopman operator using the transformed data matrices. (In the same

way, we obtain $\widehat{\mathcal{C}}_{XY}\widehat{\mathcal{C}}_{XX}^{-1}$ for the Perron–Frobenius operator.) By defining the kernels k and l explicitly as $k(x, x') = \phi(x)^\top \phi(x')$ and $l(y, y') = \phi(y)^\top \phi(y')$ for some finite-dimensional feature spaces $\mathbb{H}$ and $\mathbb{G}$, we can also see the close relationship between the methods described in this paper and VAC and EDMD. Given a finite-dimensional feature space, $\widehat{\mathcal{C}}_{XX} = \frac{1}{n}\Phi\Phi^\top$ and $\widehat{\mathcal{C}}_{YX} = \frac{1}{n}\Psi\Phi^\top$ can be computed explicitly. See also Example 4.6 and Section 5.

4.7.3 Kernel TICA and Kernel EDMD

The advantage of our method compared to VAC and EDMD is that the eigenvalue problem can be expressed entirely in terms of the Gram matrices G_{XX}, G_{XY}, G_{YX}, and G_{YY}. The transformed data matrices Φ and Ψ need not be computed explicitly. This also allows us to work implicitly with infinite-dimensional feature spaces. Kernel-based variants, based on algebraic transformations of the conventional counterparts, of TICA and EDMD have also been proposed in Schwantes and Pande (2015), Williams et al. (2015b), Klus et al. (2018). Similar to VAC, kernel TICA is based on a variational approach and requires reversibility, whereas kernel EDMD is also defined for non-reversible systems. Although kernel TICA and kernel EDMD are generalizations of different methods—TICA is related to DMD and VAC to EDMD—, the resulting methods are strongly related again. In Schwantes and Pande (2015), conventional TICA is first implicitly extended to VAC and then to kernel TICA, whereas the derivation of kernel EDMD explicitly uses the EDMD feature space representation, see also Klus et al. (2018).

5 Experiments

We will briefly show how the methods introduced above can be used to analyze dynamical systems and time-series data. In the first example, we analyze two simple molecular dynamics related problems using methods that correspond to EDMD and TICA. The second example illustrates that the kernel-based reformulations can also be applied to high-dimensional video data. The third example shows another new application, the analysis of text data. Further molecular dynamics examples can be found in Nüske et al. (2014), McGibbon and Pande (2015), Schwantes and Pande (2015), Klus et al. (2016, 2017, 2018) and applications in fluid dynamics, e.g., in Budišić et al. (2012), Williams et al. (2015b), Rowley et al. (2009).

5.1 Molecular Dynamics

In this section, we apply the proposed techniques to extract meta-stable sets and to reduce the dimension of time series data.

5.1.1 Meta-stable sets

As a first example, let us illustrate how the eigendecomposition of $\mathcal{K}_k$—for an explicitly defined feature space, which corresponds to EDMD as described above—can be used for molecular dynamics applications. We consider a simple multi-well diffusion process given by a stochastic differential equation of the form

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sqrt{2D}\,\mathrm{d}W_t,$$

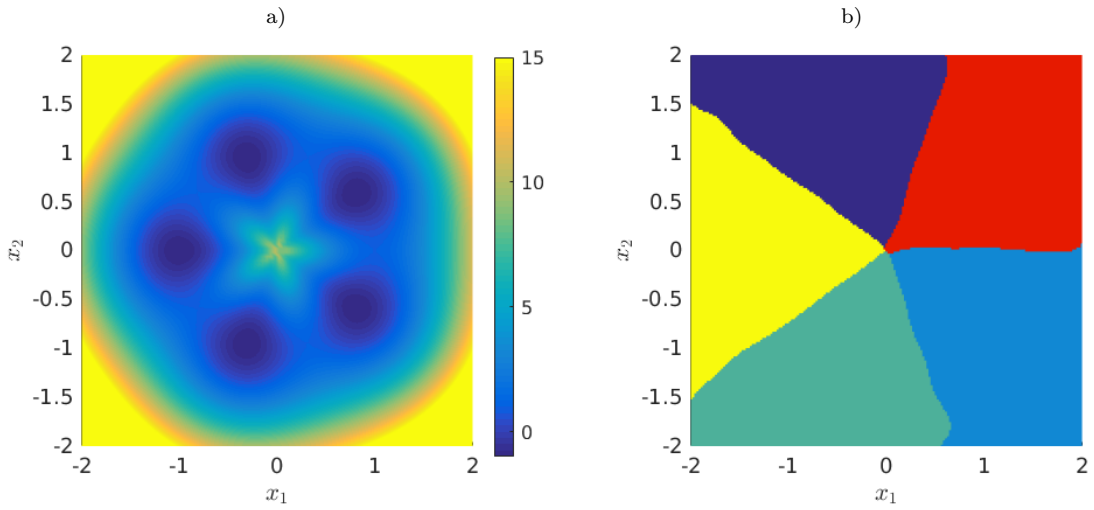

Figure 3: a) Potential V associated with the multi-well diffusion process. b) Partitioning of the state space based on the dominant eigenfunctions of the Koopman operator.

where V is the potential, $D = \beta^{-1}$ again the diffusion coefficient, and W_t a standard Wiener process. The potential, shown in Figure 3a, is given by

$$V(x) = \cos\left(k \arctan(x_2, x_1)\right) + 10\left(\sqrt{x_1^2 + x_2^2} - 1\right)^2,$$

see also Bittracher et al. (2017). We set $k = 5$. A particle will typically spend a long time in one of the wells and then jump to one of the adjacent wells. The transitions between the wells are rare events. Thus, this system exhibits metastable behavior and the five metastable sets—which are encoded in the five dominant eigenfunctions of the transfer operators associated with the system—correspond to the five wells of the potential.

We use a 50×50 box discretization of the domain $\mathbb{X} = [-2, 2] \times [-2, 2]$ to define a basis containing 2500 radial basis functions $k_i(x, c_i) = \exp(-\frac{1}{2\sigma^2}\|x - c_i\|^2)$ whose centers c_i are the centers of the boxes. This defines a kernel

$$k(x, x') = \sum_{i=1}^{2500} k_i(x, c_i)\, k_i(x', c_i) = \sum_{i=1}^{2500} \exp\left(-\frac{1}{2\sigma^2}\left(\|x - c_i\|^2 + \|x' - c_i\|^2\right)\right).$$

Furthermore, we choose the lag time $\tau = 0.2$ and $\sigma^2 = 0.9$. We generate 250000 uniformly distributed test points $x_i \in \mathbb{X}$ and solve the initial value problem with the Euler–Maruyama method to obtain the corresponding y_i values. We then compute the eigenvalues and eigenfunctions of the Koopman operator $\mathcal{K}_k$. There exist five dominant eigenvalues close to one and then there is a spectral gap between the fifth and sixth eigenvalue. We apply a k-means clustering to the dominant eigenfunctions to obtain the partitioning of the domain into the five metastable sets shown in Figure 3b. There are more sophisticated techniques to identify the metastable sets based on the eigenfunctions, but the example illustrates the basic workflow.

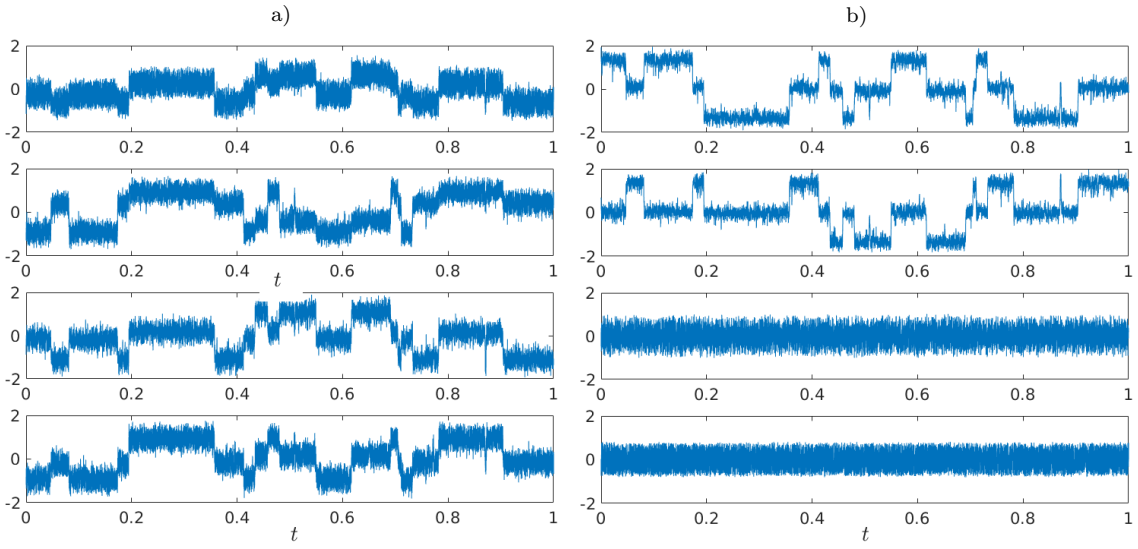

Figure 4: a) Original data set. b) Projection onto the TICA coordinates. Only the first two variables corresponding to the dominant eigenvalues exhibit metastable behavior.

5.1.2 Dimensionality reduction and blind source separation

Another use case of the methods introduced above is dimensionality reduction. Before methods to compute eigenfunctions of transfer operators such as EDMD or VAC can be applied to high-dimensional systems, the data often needs to be projected onto a lower-dimensional subspace first. This can be accomplished by approximating the eigenfunctions of the Koopman operator $\mathcal{K}_k$ using a linear kernel (i.e., TICA, see Section 4.7). Let us consider the simple data set $x \in \mathbb{R}^{4\times 10000}$ shown in Figure 4a. From this data set, we extract $X = [x_1, \ldots, x_{9999}]$ and $Y = [x_2, \ldots, x_{10000}]$, where x_i denotes the ith column vector of x. Applying TICA, we see that there are two dominant eigenvalues close to 1, the other two are close to 0. This indicates that two of the four variables exhibit metastable behavior. Projecting the data onto the TICA coordinates results in the trajectories shown in Figure 4b. The first two new variables corresponding to the dominant eigenvalues contain the metastability, while the other two variables contain just noise. (In fact, this is how the data set was constructed.) Since we are only interested in the slow metastable dynamics, we can neglect the last two variables and thus reduce the state space. At the same time, we can view the projection onto eigenfunction coordinates as the unmixing of previously mixed signal sources. Thus, our methods can be used for solving blind source separation problems.

5.2 Movie Data

Let us analyze a simple movie showing a pendulum[1]. We want to analyze this data set using the eigendecomposition of $\mathcal{K}_k$ for a Gaussian kernel k (corresponding to kernel EDMD). To this end, we convert each 576×720 RGB video frame to a grayscale intensity image—all intensities are between 0 and 1—and define a kernel $k(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|_F\right)$, with

[1]ScienceOnline: The Pendulum and Galileo (www.youtube.com/watch?v=MpzaCCbX-z4).

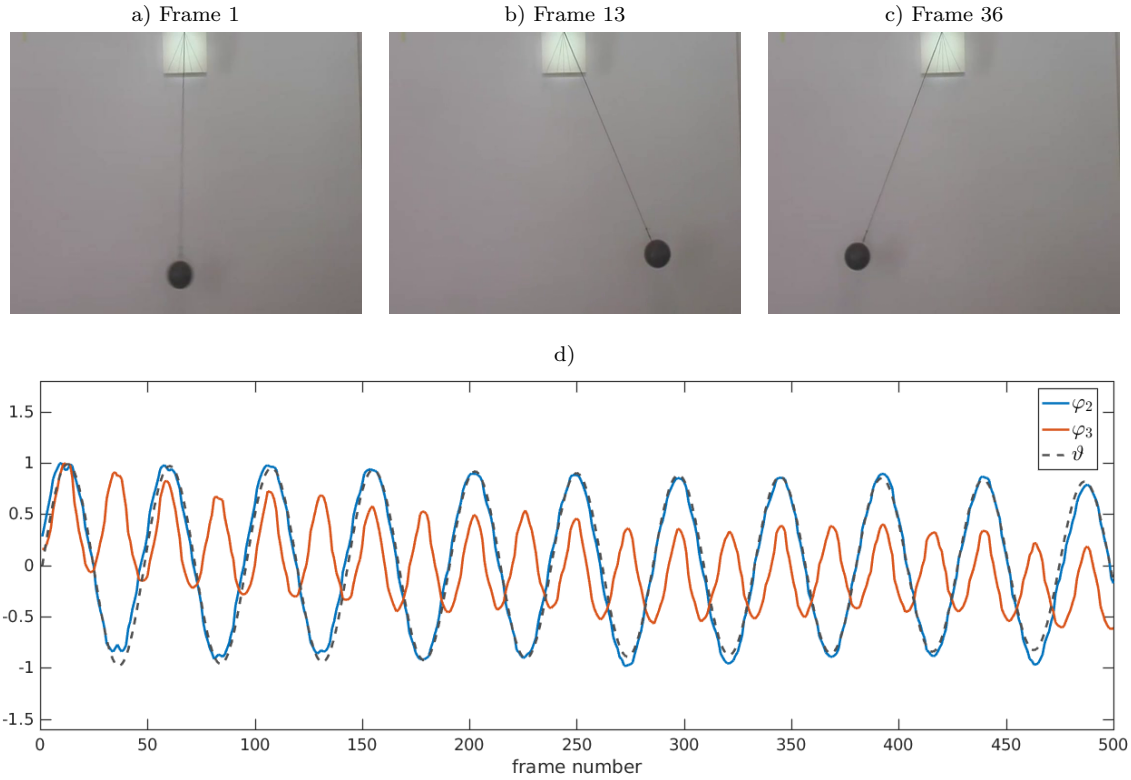

Figure 5: a) No angular displacement. b) Maximum displacement right-hand side. c) Maximum displacement left-hand side. d) Values of the normalized eigenfunctions φ_2 and φ_3 for each frame. The eigenfunctions encode the frequency of the pendulum. The frames 13 and 36 correspond to the first maximum and minimum of the eigenfunction φ_2. The period of φ_3 is twice the period of φ_2. The black dashed line shows the angular displacement ϑ (rescaled for the sake of comparison) obtained by a numerical simulation of the pendulum. The dominant eigenfunction parametrizes the angular displacement. The video snapshots are reproduced with the kind permission of *ScienceOnline*.

$\sigma^2 = 500$. Here, $\|\cdot\|_F$ denotes the Frobenius norm. It would also be possible to use the RGB signal directly, e.g., by defining $k_{\mathrm{RGB}}(x, y) = k(x_R, y_R) + k(x_G, y_G) + k(x_B, y_B)$, i.e., each primary color is compared separately. The video comprises 501 frames so that $X, Y \in \mathbb{R}^{576\times720\times500}$. That is, the data sets are now tensors of order three. Analogously, we could reshape the snapshot matrices into vectors. We choose the regularization parameter $\varepsilon = 0.05$. Thus, for our analysis, we have to solve the eigenvalue problem $(G_{XX} + \varepsilon I_n)^{-1} G_{YX}\, v = \lambda v$ to obtain eigenfunctions of $\mathcal{K}_k$.

The values of the resulting nontrivial dominant eigenfunctions φ_2 and φ_3 evaluated for each frame are shown in Figure 5. The first nontrivial eigenfunction encodes the frequency of the pendulum and the second eigenfunction twice the frequency. As a result, we could now sort the frames according to the angular displacement of the pendulum using the eigenfunctions. The example shows that even for high-dimensional problems the kernel-based methods are able to extract relevant information about the global dynamics. The frames attaining maxima and minima of the eigenfunctions provide a summary of the video using

typical (but maximally different) frames. This use of our methods resembles *determinantal point processes* as applied to data summarization (Kulesza and Taskar, 2012).

For this simple example, we used the raw video data. For more complex systems, preprocessing steps might be beneficial, e.g., mean subtraction, Sobel edge detection, or more sophisticated feature detection approaches such as SIFT or HOG (Bo et al., 2010). In this way, it would be possible to track features of images over time. Another potential application that we considered but did not include here is the analysis of persons walking or running. The eigenfunctions then describe the gait pattern and gait velocity.

5.3 Text Data

In this section, we show how the eigendecomposition of the kernel Perron–Frobenius operator with respect to the invariant density, denoted by $\mathcal{T}_k$, can be used for non-vectorial data. Consider the following scenario: Given a collection of text documents, erase all words not contained in a predefined vocabulary. Of the remaining words, one word (denoted by y_i) following another (denoted by x_i) is considered to be its time evolved version or successor. The lists of all such words x_i and y_i are denoted by $\mathbf{X}$ and $\mathbf{Y}$, respectively. We choose the vocabulary shown in Table 2 and collect 1000 word pairs from news articles. Typically, the same word or related words are used several times within one article, but words related to other topics are rarely mentioned. Since we consider the sequence of articles as one long document[2], transitions occur, for instance, when one article ends and the next one about a different topic starts, when different topics are mixed, or when words such as *state* or *cell* are used in a different context. These are the rare transitions that are similar to the jumps between the wells in the molecular dynamics example. Although this is a slightly artificial example, it illustrates how to extend transfer operator approaches to new domains where only a similarity measure given by a kernel is available.

Table 2: Predefined set of keywords.

browser	cell	computer	damage	department
disease	e-mail	election	hurricane	internet
midterm	president	rain	science	state
stem	storm	tablet	therapy	weather

We generate the Gram matrices G_{XY} and G_{XX} and compute eigenfunctions of the operator $\mathcal{T}_k$. Here, $[G_{XY}]_{ij} = k(x_i, y_j)$, where x_i is the ith word in $\mathbf{X}$ and y_j the jth word in $\mathbf{Y}$. Correspondingly, G_{XX} is the standard Gram matrix. Moreover, k is the text kernel proposed in Lodhi et al. (2002)[3]. We compute again the leading nontrivial eigenfunctions φ_2 and φ_3 and use the the eigenfunctions as coordinates. The results are shown in Figure 6. Note that the words are not clustered based on string kernel similarity but on proximity in the document collection. Words that often occur together are grouped into clusters. For this simple example, it would also have been possible to assign each word a distinct number and to generate a Markov state model by approximating the transition probabilities

[2]Parts of the same articles are used several times to increase the size of the data set, this is thus a synthetic example, mainly to illustrate the concept.

[3]We use the *String Kernel Software* implementation (`https://github.com/mmadry/string_kernel`).

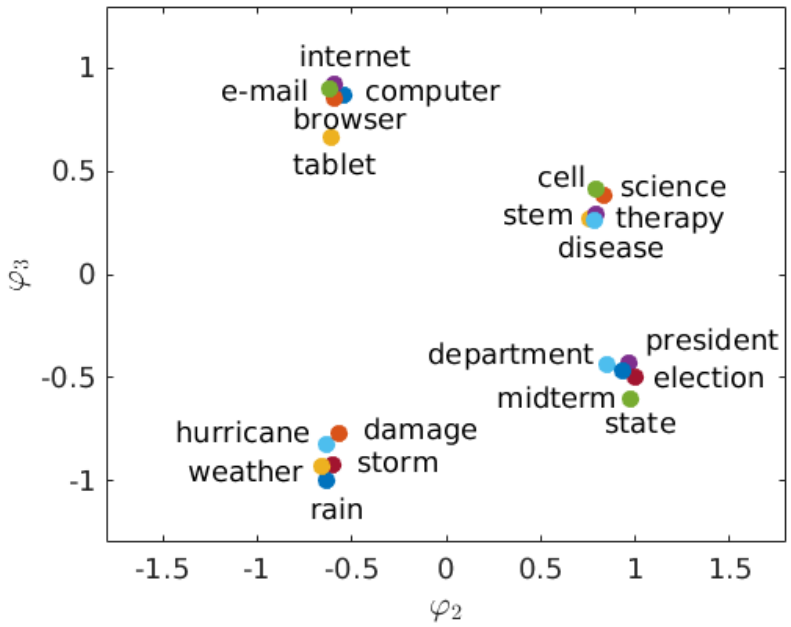

Figure 6: Topic modeling by clustering using on the dominant eigenfunctions. Words that are found in close proximity represent a topic. Our method identified four topic clusters: information technology, medicine, weather, and politics.

between words. The eigenvectors of the Markov matrix would then lead to a similar clustering. The text kernel, however, takes into account string similarity. This is important to account, for example, for grammatical variations reflected in word form (*green* vs. *greener*) and misspellings (*love* vs. *loove*) without necessarily resorting to lemmatizing, stemming, or other normalization techniques. Another possibility here would be to design linguistically informed string kernels. In German for example, a *Visumantrag* (visa application) is more similar to *Antrag* (application) than to *Visum*. A string kernel taking this into account would instantly be reflected in the word clusters discovered by our method, which could never be achieved when using a pure Markov state model.

Compared to *latent Dirichlet allocation* (LDA) (Blei et al., 2003), our method of uncovering topics in texts differs in several respects. First of all, LDA is derived as a Bayesian model, while our method can be considered frequentist. Second, LDA makes a bag-of-words assumption, i.e., the order of words in texts is not taken into account. Our method, on the other hand, relies first and foremost on word order. Third, the semantic content of words in LDA could be summarized as a real vector using their frequency in topics, while a clustering of words into topics is the primary object of interest. Our method, on the other hand, produces a semantic word representation, the eigenfunction values of a word, while a clustering into topics can be implemented as a postprocessing step. This is an interesting application warranting further research that follows directly from our general principle of decomposing an RKHS operator.

6 Conclusion

We have shown how to extend transfer operator theory to reproducing kernel Hilbert spaces and illustrated similarities with the conditional mean embedding framework. While the conventional transfer operator propagates densities, the kernel mean embedding can be viewed as an operator that propagates embedded densities. Moreover, we have highlighted relationships between the covariance and cross-covariance operator based methods to obtain empirical estimates of the conditional mean embedding and other well-known methods for the approximation of transfer operators developed by the dynamical systems, molecular dynamics, and fluid dynamics communities. One main benefit of purely kernel-based methods is that these methods can be applied to non-vectorial data such as strings or graphs, enabling the analysis of many (wide-sense) stationary processes. We demonstrated the efficiency and versatility of these methods using guiding examples as well as simple molecular dynamics applications, video data, and text data. Future work includes applying the aforementioned kernel-based methods to more realistic data sets. In particular the analysis of real-world text data and more complicated video data, potentially in combination with machine learning based preprocessing approaches, will be a challenging task. An open question is also the convergence of the RKHS operator to the actual transfer operator. Furthermore, the influence of the kernel itself, the regularization parameter, and the number of test points on the accuracy of the eigenfunction approximations is not clear yet. These questions will be addressed in future work. Another extension of the framework presented within this paper would be to use singular value decompositions instead of eigenvalue decompositions for the transfer operator representations. The resulting methods could then also be applied to problems where the spaces $\mathbb{X}$ and $\mathbb{Y}$ are different.

Acknowledgements

This research has been partially funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1114 *"Scaling Cascades in Complex Systems"*. Krikamol Muandet acknowledges fundings from the Faculty of Science, Mahidol University and the Thailand Research Fund (TRF).

A Analytical Mean Embeddings and Its Inversion

If the function is given by a sum of Gaussians and the kernel is Gaussian or Student-t, the embedding and its inverse can be computed analytically.

A.1 Gaussian functions, Gaussian kernels

Let $k(x, x')$ be the kernel given by the normalized d-dimensional Gaussian density with covariance Σ_k, mean x', and evaluation at point x. Let $g(\cdot) = \sum_{i=1}^{\infty} a_i \mathcal{N}(\cdot\,; \mu_i, \Sigma_i)$, where the $\mathcal{N}(\cdot\,; \mu_i, \Sigma_i)$ are d-dimensional multivariate Gaussian densities and $a_i \in \mathbb{R}$. Then

$$(\mathcal{E}_k\, g)(\cdot) = \sum_{i=1}^{\infty} a_i \mathcal{N}(\cdot\,; \mu_i, \Sigma_i + \Sigma_k).$$

Thus, we can analytically embed any function defined by a weighted sum of Gaussian densities. Furthermore, if g lies in the RKHS generated by k,

$$(\mathcal{E}_k^{-1} g)(\cdot) = \sum_{i=1}^{\infty} a_i \mathcal{N}(\cdot\,; \mu_i, \Sigma_i - \Sigma_k).$$

This means that for any function that is an embedding into a Gaussian RKHS with known Σ_k and given by a sum of Gaussian densities, we can find its pre-image in closed form.

A.2 Gaussian functions, Student-t kernels

A similar construction holds also for multivariate Student-t kernels. Let $l(x, x')$ be the kernel given by a multivariate Student-t density with scale matrix Σ_l, ν degrees of freedom, mean x', and evaluation at point x. Then for $f(\cdot) = \sum_{i=1}^{\infty} a_i \mathcal{N}(\cdot\,; \mu_i, \Sigma_i)$, we have

$$(\mathcal{E}_l f)(\cdot) = \sum_{i=1}^{\infty} a_i \,\mathrm{MVT}(\cdot\,; \mu_i, \Sigma_i + \Sigma_l, \nu)$$

and for $g(\cdot) = \sum_{i=1}^{\infty} a_i \,\mathrm{MVT}(\cdot\,; \mu_i, \Sigma_i, \nu)$, if g lies in the RKHS generated by l, we have

$$(\mathcal{E}_l^{-1} g)(\cdot) = \sum_{i=1}^{\infty} a_i \mathcal{N}(\cdot\,; \mu_i, \Sigma_i - \Sigma_l).$$

References

S. M. Ulam. *A Collection of Mathematical Problems*. Interscience Publisher NY, 1960.

M. O. Williams, I. G. Kevrekidis, and C. W. Rowley. A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015a.

M. O. Williams, C. W. Rowley, and I. G. Kevrekidis. A kernel-based method for data-driven Koopman spectral analysis. *Journal of Computational Dynamics*, 2(2):247–265, 2015b.

S. Klus, P. Koltai, and C. Schütte. On the numerical approximation of the Perron–Frobenius and Koopman operator. *Journal of Computational Dynamics*, 3(1):51–79, 2016.

F. Noé and F. Nüske. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Modeling & Simulation*, 11(2):635–655, 2013.

F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé. Variational approach to molecular kinetics. *Journal of Chemical Theory and Computation*, 10(4): 1739–1752, 2014.

S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé. Data-driven model reduction and transfer operator approximation. *ArXiv e-prints*, 2017.

J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor. *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. SIAM, 2016.

C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, 641:115–127, 2009.

M. Budišić, R. Mohr, and I. Mezić. Applied Koopmanism. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(4), 2012.

J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz. On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics*, 1 (2), 2014.

R. T. McGibbon and V. S. Pande. Variational cross-validation of slow dynamical modes in molecular kinetics. *The Journal of Chemical Physics*, 142(12), 2015.

C. R. Schwantes and V. S. Pande. Modeling molecular kinetics with tICA and the kernel trick. *Journal of Chemical Theory and Computation*, 11(2):600–608, 2015. doi: 10.1021/ct5007357.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer-Verlag, 2007.

K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10 (1–2):1–141, 2017.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

Sebastian Mika, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel pca and de-noising in feature spaces. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*, pages 536–542, Cambridge, MA, USA, 1999. MIT Press.

K. Fukumizu, F. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8:361–383, 2007.

A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Network*, 13(4-5):411–430, 2000.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2003.

F. Yao, H.-G. Müller, and J.-L. Wang. Functional linear regression analysis for longitudinal data. *Annals of Statistics*, 33(6):2873–2903, 2005.

P. Hall, H.-G. Müller, and J.-L. Wang. Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics*, 34(3):1493–1517, 2006.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, 2nd edition, 2005.

Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: a comparative review. *Journal of Machine Learning Research*, 10, 2009.

Christopher J. C. Burges. Dimension reduction: A guided tour. *Foundations and Trends in Machine Learning*, 2(4):275–365, 2010.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009. doi: 10.1145/1553374.1553497.

L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT press, Cambridge, USA, 2001.

T. Hofmann, B. Schölkopf, and A. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with Reproducing Kernel Hilbert Spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *The 21st Annual Conference on Learning Theory*, pages 111–122. Omnipress, 2008.

B. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 99:1517–1561, 2010.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2002.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

T. Kato. *Perturbation Theory for Linear Operators*. Springer, Berlin, 1980.

M. Renardy and R. C. Rogers. *An introduction to partial differential equations*, volume 13. Springer Science & Business Media, 2006.

D. Bump. *Automorphic forms and representations*, volume 55. Cambridge University Press, 1998.

C. Baker. Mutual information for Gaussian processes. *SIAM Journal on Applied Mathematics*, 19(2):451–458, 1970.

C. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14:3753–3783, 2013.

R. Gallager. *Stochastic processes: theory for applications*. Cambridge University Press, 2013.

C. Schütte and M. Sarich. *Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches*. Number 24 in Courant Lecture Notes. American Mathematical Society, 2013.

J. R. Baxter and J. S. Rosenthal. Rates of convergence for everywhere-positive Markov chains. *Statistics & probability letters*, 22(4):333–338, 1995.

G. A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, volume 60 of *Texts in Applied Mathematics*. Springer, 2014.

A. Lasota and M. C. Mackey. *Chaos, fractals, and noise: Stochastic aspects of dynamics*, volume 97 of *Applied Mathematical Sciences*. Springer, 2nd edition, 1994.

P. Koltai. *Efficient approximation methods for the global long-term behavior of dynamical systems – Theory, algorithms and examples*. PhD thesis, Technische Universität München, 2010.

M. Korda and I. Mezić. On convergence of Extended Dynamic Mode Decomposition to the Koopman operator. *ArXiv e-prints*, 2017.

S. Klus, P. Koltai, A. Bittracher, R. Banisch, and C. Schütte. Kernel-based spectral analysis of transfer operators. In preparation, 2018.

L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72:3634–3637, 1994.

G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé. Identification of slow molecular order parameters for Markov model construction. *The Journal of Chemical Physics*, 139(1), 2013.

P. J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010. doi: 10.1017/S0022112010001217.

H. Wu and F. Noé. Variational approach for learning Markov processes from time series data. *ArXiv e-prints*, 2017.

A. Bittracher, P. Koltai, S. Klus, R. Banisch, M. Dellnitz, and C. Schütte. Transition manifolds of complex metastable systems: Theory and data-driven computation of effective dynamics. *Accepted for publication in JNLS*, 2017.

A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012.

L. Bo, X. Ren, and D. Fox. Kernel descriptors for visual recognition. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 244–252. Curran Associates, Inc., 2010.

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002. doi: 10.1162/153244302760200687.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4–5):993–1022, 2003. doi: 10.1162/jmlr.2003.3.4-5.993.