JMLR: Workshop and Conference Proceedings 44 (2015) 103-115 NIPS 2015

The 1st International Workshop "Feature Extraction: Modern Questions and Challenges"

Spatiotemporal Feature Extraction with Data-Driven Koopman Operators

Dimitrios Giannakis DIMITRIS@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012-1185, USA

Joanna Slawinska JOANNA.SLAWINSKA@ENVSCI.RUTGERS.EDU
Center for Environmental Prediction
Rutgers University
New Brunswick, NJ 08901-8551, USA

Zhizhen Zhao JZHAO@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012-1185, USA

Editor: Dmitry Storcheus

Abstract

We present a framework for feature extraction and mode decomposition of spatiotemporal data generated by ergodic dynamical systems. Unlike feature extraction techniques based on kernel operators, our approach is to construct feature maps using eigenfunctions of the Koopman group of unitary operators governing the dynamical evolution of observables and probability measures. We compute the eigenvalues and eigenfunctions of the Koopman group through a Galerkin scheme applied to time-ordered data without requiring a priori knowledge of the dynamical evolution equations. This scheme employs a data-driven set of basis functions on the state space manifold, computed through the diffusion maps algorithm and a variable-bandwidth kernel designed to enforce orthogonality with respect to the invariant measure of the dynamics. The features extracted via this approach have strong timescale separation, favorable predictability properties, and high smoothness on the state space manifold. The extracted features are also invariant under weakly restrictive changes of observation modality. We apply this scheme to a synthetic dataset featuring superimposed traveling waves in a one-dimensional periodic domain and satellite observations of organized convection in the tropical atmosphere.

Keywords: Feature extraction, ergodic dynamical systems, Koopman operators, kernel methods, spatiotemporal data

1. Introduction

An important problem in data science is to perform feature extraction from spatiotemporal data. When the data are generated by ergodic dynamical systems (as is the case in many science and engineering applications) they acquire an important property, namely that the system's state space can be densely explored by long time series of snapshots of sufficiently high resolution. In many cases of interest, the sampled data lies in a low-dimensional subset

of the ambient data space (an attractor) with a nonlinear geometric structure. Modern data analysis techniques, such as kernel and manifold learning methods (e.g., Blanchard et al., 2007; Storcheus et al., 2015; Lee and Verleysen, 2007), take advantage of such geometric structures to perform efficient feature extraction, but oftentimes do not take into consideration the temporal structure of the data which is a direct manifestation of the underlying dynamical system.

In the context of ergodic theory, an alternative viewpoint (introduced by Koopman (1931)) is to characterize a nonlinear dynamical system operating on a nonlinear state space through a group of linear operators acting on vector spaces of observables, i.e., functions on the state space (Budisić et al., 2012; Mezić, 2013, and references therein). In this operator-theoretic viewpoint, spaces of observables are generally equipped with a Hilbert space structure, and dynamical evolution is represented by a group of unitary operators on those Hilbert spaces which translate observables along the orbits of the dynamics. Specifically, let $\Phi_t : M \mapsto M$ be the dynamical evolution map on the state space such that $a_t = \Phi_t(a_0)$ is the point on M after dynamical evolution for time t starting from the point a_0. Let also μ be a Φ_t-invariant probability measure, and f an arbitrary observable in the Hilbert space $L^2(M, \mu)$ of complex-valued square-integrable functions on M with respect to μ. Assuming that Φ_t is invertible with $\Phi_t^{-1} = \Phi_{-t}$, the Koopman operator for time t is the unitary operator $U_t : L^2(M, \mu) \mapsto L^2(M, \mu)$ such that $U_t(f) = f \circ \Phi_t$ (the unitarity of U_t is a consequence of the measure-preserving property of Φ_t). The set $\{U_t \mid t \in \mathbb{R}\}$ forms a group under composition of operators, called Koopman group, which is generated by a vector field $v = dU_t/dt|_{t=0}$ on M with vanishing divergence with respect to μ. If Φ_t is non-invertible, then $\{U_t \mid t \geq 0\}$ forms a semigroup of isometric operators on $L^2(M, \mu)$.

In general, the space $L^2(M, \mu)$ is infinite-dimensional even if the dimension of M is finite, but due to its intrinsically linear structure, the Koopman formalism opens up the possibility to use finite-dimensional operator approximation methods (such as pseudospectral and Galerkin methods) for nonlinear dynamical modeling and nonparametric forecasting (Berry et al., 2015; Giannakis, 2015). In the applications of interest here, we assume that we do not have explicit knowledge of the Koopman group or the evolution map. Instead, we will construct approximations of v from a time series $\{y_t\}$ of data generated by the dynamical system. In particular, we consider that y_t is a snapshot lying in a Hilbert space H (details of which will be made precise below), and it is the result $y_t = F(a_t)$ of a mapping $F : M \mapsto H$ where a_t is the dynamical state in M given by $a_t = \Phi_t(a_0)$. We also assume that the y_t and a_t are in one-to-one correspondence, i.e., that F is invertible on its image (note that this assumption can be relaxed using Takens delay-coordinate maps as described in section 4). We refer to F and H as the observation map and the ambient data space, respectively. Due to our invertibility assumption, the sets $\{y_t\}$ and $\{a_t\}$ are equivalent, but we prefer to keep these objects distinct to be able to deal with situations where one has access to data generated by the same dynamical system but acquired through different observation maps.

Besides their use in a predictive setting, Koopman operators have several desirable properties for feature extraction. In particular, if v has eigenfunctions $\{z_i\}$ in $L^2(M, \mu)$, then these eigenfunctions can be used to construct mappings of M to low-dimensional Euclidean spaces as in kernel eigenmaps, but with the difference that the feature spaces are complex (the z_i are complex orthogonal eigenfunctions and the corresponding eigenvalues λ_i lie on the imaginary line by unitarity of U_t). That is, given an appropriately selected set

$\{z_1, \ldots, z_m\}$ of Koopman eigenfunctions, we construct the feature map $\pi : M \mapsto \mathbb{C}^m$, where

$$\pi(a) = (z_1(a), \ldots, z_m(a)). \tag{1}$$

The key properties of feature extraction with Koopman eigenfunctions are as follows (Mezić, 2005; Budisić et al., 2012; Mezić, 2013; Giannakis, 2015).

1. The Koopman eigenfunctions are intrinsic to the dynamical system generating the data, in the sense that they depend only on U_t and not on the observation map F. Thus, (1) can be used to construct a universal low-dimensional Euclidean space to represent data generated by the dynamical system and acquired via different sensors corresponding to distinct observation maps.

2. The dynamics are projectible under π, in the sense that for any two points on M mapping to the same point in $\mathbb{C}^m$ the corresponding images of the generator under the tangent map $T\pi$ agree, i.e., if $\pi(a) = \pi(a')$ then $T_a\pi(v) = T_{a'}\pi(v)$. Thus, the dynamics in $\mathbb{C}^m$ are Markovian and closure issues are avoided.

3. The eigenfunctions and eigenvalues have a group structure under multiplication, in the sense that if z_1 and z_2 are eigenfunctions with corresponding eigenvalues λ_1 and λ_2, then $z_1 z_2$ is also an eigenfunction corresponding to the eigenvalue $\lambda_1\lambda_2$. This means that the full spectrum of can be generated recursively from a finite set of eigenfunctions corresponding to rationally independent eigenvalues.

4. The dynamics in the feature space $\mathbb{C}^m$ are simple harmonic oscillations with frequencies given by the Koopman eigenvalues, even if the dynamical system is nonlinear. Specifically, if $\zeta_i(t) = z_i(\Phi_t(a))$ is the time series of the values of the eigenfunction z_i sampled along an orbit starting at an arbitrary point $a \in M$, then $d\zeta_i/dt = \mathrm{i}\omega_i t$, where $\omega_i = \operatorname{Im}\lambda_i$. Thus, feature maps constructed from Koopman eigenfunctions produce a decomposition of the dynamics into uncoupled simple harmonic oscillators, whose frequencies are intrinsic to the dynamical system. In the feature space, the multiple timescales that may present in the observed data $\{y_t\}$ are separated into distinct coordinates with known temporal evolution which can be used for nonparametric (model-free) forecasting of observables and probability measures.

The spatial patterns $\hat{F}_i = \langle z_i, F\rangle = \int_M z_i^*(a)F(a)\,d\mu(a)$ obtained by projecting the data onto the Koopman eigenfunctions using the Hilbert space inner product are referred to as Koopman modes and the decomposition of the observed data into the triplets $\{(z_i, \lambda_i, \hat{F}_i)\}$ is called dynamic mode decomposition (Mezić, 2005).

Traditionally, methods for computing Koopman eigenfunctions are based on iterative algorithms (such as Arnoldi algorithms) operating directly in the ambient data space (Rowley et al., 2009; Williams et al., 2015). A drawback of these approaches is high computational cost due to the dimension of ambient data space (which can far exceed the intrinsic dimension of M) and risk of numerical instabilities. Moreover, the computed eigenfunctions generally have no smoothness guarantees, e.g., with respect to the Dirichlet form inherited by M through the observation map. Recently, Giannakis (2015) (hereafter, G15) developed an alternative approach for the computation of Koopman eigenfunctions which is based on

a finite-dimensional representation of the generator v of the Koopman group in a smooth orthonormal basis of $L^2(M,\mu)$. This basis is constructed from the observed data $\{y_t\}$ using a kernel method developed earlier by Berry et al. (2015) which generates orthonormal functions on M with respect to the correct ergodic measure μ (as opposed to the ambient-space Riemannian measure). The method of Berry et al. (2015) employs a variable-bandwidth kernel (Berry and Harlim, 2015) and the diffusion maps algorithm (Coifman and Lafon, 2006) to construct the basis from the eigenfunctions of the generator of a gradient flow on M. G15 employs this basis in a spectral Galerkin formulation of the eigenvalue problem for v whose computational cost depends on the bandwidth of the Koopman eigenfunctions in the diffusion maps basis (as opposed to the ambient space dimension), and isolates the eigenfunctions with maximal smoothness on $L^2(M,\mu)$ through Tikhonov regularization. The applications in G15 demonstrated the utility of this scheme for feature extraction and nonparametric forecasting of dynamical systems on tori with multiple timescales, but were limited to synthetic data of low ambient-space dimension.

In this paper, we further develop and apply the framework of G15 to time-evolving data in spatially extended domains; e.g., a line interval or the surface of a sphere. Mathematically, this situation corresponds to the case that the ambient data space H has the structure of an infinite-dimensional Hilbert space $L^2(X)$ on a set X (the spatial domain) of sufficient smoothness. For example, X could be a Lipschitz domain in $\mathbb{R}^n$, or a compact Riemannian manifold. Our applications are motivated by a current open problem in the atmospheric sciences, namely the extraction of a multiscale hierarchy of traveling convective waves in the tropical atmosphere (Dias et al., 2013). These waves collectively exert global influences on the weather climate, yet are poorly represented by numerical models and in some cases are incompletely understood theoretically. As a result, their objective detection in observational data is important for both model guidance and improvement of existing scientific understanding and theories.

The plan of this paper is as follows. In section 2, we describe our approach for computing Koopman eigenfunctions from spatiotemporal data. In sections 3 and 4, we present applications of this technique to synthetic traveling-wave data in 1D and atmospheric convection data from satellite observations, respectively.

2. Kernel methods for Koopman eigenfunctions

Let $\{y_0, y_1, \ldots, y_{n-1}\}$ be a dataset consisting of n time-ordered snapshots $y_i = F(a_i)$ sampled from the orbit $a_i = \Phi_{t_i}(a_0)$ of the dynamics on the state space M at times $t_i = (i-1)\,\delta t$ for a uniform timestep δt. Here, we assume that M is a smooth, compact, orientable m-dimensional manifold without boundary, and we also assume that the dynamics Φ_t are smooth. Moreover, as stated in section 1, we are interested in the case that the snapshots are real-valued, square-integrable scalar fields on a spatial domain X. Thus, $y_i(x)$ is a real number corresponding to the evaluation of y_i at $x \in X$, and we have the Riemannian inner product $\langle y_i, y_j\rangle = \int_X y_i(x) y_j(x)\, dx$. Numerically, we approximate such inner products by quadrature on a finite set $\{x_1, x_2, \ldots, x_d\}$ of nodes on M with corresponding weights $\{w_1, w_2, \ldots, w_d\}$. That is, we have $\langle y_i, y_j\rangle \simeq \langle \vec{y}_i, \vec{y}_j\rangle$, where $\vec{y}_i = (y_i(x_1), \ldots, x_i(x_d))$ are d-dimensional vectors and $\langle \vec{y}_i, \vec{y}_j\rangle = \sum_{k=1}^{d} w_k y_i(x_k) y_j(x_k)$. Similarly, we represent complex-valued functions $f \in L^2(M,\mu)$ by n-dimensional vectors $\vec{f} = (f(a_0), \ldots, f(a_{n-1}))$ with

components equal to the function values at the sampled states on M. By ergodicity, the inner product on $L^2(M,\mu)$ can be approximated by time averages along orbits of the dynamics, i.e., $\langle f_1, f_2\rangle \simeq \langle \vec{f_1}, \vec{f_2}\rangle := \sum_{i=0}^{n-1} f_1^*(a_i) f_2(a_j)/n$. In what follows, we assume that all expressions involving inner products on $L^2(X)$ and/or $L^2(M,\mu)$ are evaluated in practice using the corresponding discrete formulas.

To construct an orthonormal basis of $L^2(M,\mu)$ we start from the kernel $K : M \times M \mapsto \mathbb{R}_+$ given by (Berry and Harlim, 2015)

$$K(a_i, a_j) = \exp\left(-\frac{\|y_i - y_j\|^2}{\epsilon \sigma_\epsilon^{-1/\hat{m}}(a_i)\sigma_\epsilon^{-1/\hat{m}}(a_j)}\right). \tag{2}$$

In (2), $\|\cdot\|$ is the Hilbert space norm of $L^2(X)$, ϵ is a positive bandwidth parameter, σ_ϵ is a function approximating the sampling density σ at $O(\epsilon)$ accuracy, and $\hat{m}$ is an estimate of the dimension of M. The function σ_ϵ can be computed using any suitable density-estimation technique, and in what follows we employ the kernel method described in Berry and Harlim (2015) and Berry et al. (2015). This method uses an updated formulation of an automatic bandwidth-selection procedure originally developed in Coifman et al. (2008), which also provides an estimate $\hat{m}$ for the dimension of M. Alternatively, this parameter can be set using one of the dimension estimation techniques available in the literature (e.g., Hein and Audibert (2005); Little et al. (2009)).

Assuming that for every $a \in M$ the function $K_a(b) = K(a,b)$ is in $L^2(M,\mu)$, the kernel in (2) induces an integral operator $\mathcal{K} : L^2(M,\mu) \mapsto L^2(M,\mu)$ such that $\mathcal{K}(f)(a) = \langle K_a, f\rangle$. By performing the sequence of normalizations introduced in the diffusion maps algorithm, and further developed by Berry and Sauer (2015), we use $\mathcal{K}$ to construct the Markov operator $\mathcal{P}(f) = \mathcal{K}(f)/\mathcal{K}(1)$, where in the last expression 1 denotes the function equal to one everywhere on M. For the choice of kernel in (2), and as the bandwidth parameter ϵ tends to zero, the eigenfunctions of $\phi_0, \phi_1, \ldots$ of $\mathcal{P}$ converge to the eigenfunctions of a Laplace-Beltrami operator Δ associated with a Riemannian metric h on M whose volume has uniform density relative to the invariant measure of the dynamics (G15); i.e., the ϕ_i provide an orthonormal basis of $L^2(M,\mu)$. Moreover, the Δ-eigenvalues corresponding to ϕ_i, denoted here by η_i, can be estimated from the logarithms of the corresponding $\mathcal{P}$-eigenvalues. In applications, we tune the kernel bandwidth parameter ϵ using the tuning procedure of Berry and Harlim (2015). Note that the η_i are also equal to the Dirichlet energies $E(\phi_i) = \int_M \|\mathrm{grad}_h\, \phi_i\|^2\, d\mu$ measuring the roughness of the corresponding eigenfunctions in the Riemannian metric h.

Next, consider the eigenvalue problem for the generator v of the Koopman group, $v(z_i) = \lambda_i z_i$. Instead of solving this eigenvalue problem directly, we solve the eigenvalue problem for the advection-diffusion operator $L = v + \nu\Delta$, where ν is a positive regularization parameter. Here, the role of Δ is to suppress highly oscillatory eigenfunctions from the spectrum of L, and one can verify through simple perturbation expansions that the effect of Δ to the eigenfunctions and eigenvalues of v is $O(\nu)$ and $O(\nu^2)$, respectively (G15). We solve the eigenvalue problem for L in weak form, setting the trial and test spaces to the Sobolev space $H^1(M,h,\mu)$ associated with the Riemannian metric h and the invariant measure on the dynamics. Our basis for this space consists of the rescaled diffusion eigenfunctions $\varphi_i = \phi_i/\eta_i^{1/2}$, ordered in order of increasing eigenvalue (i.e., Dirichlet energy). This basis is tailored to the $H^1(M,h,\mu)$ regularity of the problem in the sense that for any sequence

$(c_1, c_2, \ldots) \in \ell^2$ the function $\sum_{i=1}^{\infty} c_i \varphi_i$ is in $H^1(M, h, \mu)$ (this would not be the case in the $\{\phi_i\}$ basis since the basis elements exhibit unbounded growth of Dirichlet energy). Moreover, in finite-dimensional Galerkin schemes constructed in this basis, the diffusion operator Δ is represented by the identity matrix, $\langle \varphi_i, \Delta \varphi_j \rangle = \delta_{ij}$, and remains well conditioned at large spectral orders. Restricting the trial and test spaces to the l-dimensional subspaces of $H^1(M, h, \mu)$ spanned by $\{\varphi_1, \ldots, \varphi_l\}$, the weak form of the eigenvalue problem for L becomes the matrix generalized eigenvalue problem

$$Ac_i = \hat{\lambda}_i B c_i, \tag{3}$$

where A and B are $l \times l$ matrices with elements $A_{ij} = \langle \varphi_i, v(\varphi_j) \rangle + \nu \delta_{ij}$ and $B_{ij} = \langle \varphi_i, \varphi_j \rangle$, respectively, and $c_i = (c_{1i}, c_{2i}, \ldots, c_{li})^\top$ are l-dimensional column vectors such that $\hat{z}_i = \sum_{j=1}^{l} c_{ji} \varphi_j$ approximates the Koopman eigenfunction z_i. Also, the imaginary part of the generalized eigenvalue $\hat{\lambda}_i$ provides an approximation to λ_i.

Note that the time ordering of the data is crucial for the evaluation of the matrix elements A_{ij}. In particular, let $\{\tilde{\varphi}_{1i}, \tilde{\varphi}_{2i}, \ldots, \tilde{\varphi}_{ni}\}$ be the time series of the φ_i sampled along the dynamical trajectory. For instance, $\tilde{\varphi}_{ji} = \varphi_i(a_j) = \phi_i(a_j)/\eta_j$, where the eigenfunction values $\phi_i(a_j)$ are determined by diffusion maps as described above. Because v is the infinitesimal generator of the Koopman group, $v(\varphi_i)$ can be approximated by finite differences of the $\{\tilde{\varphi}_{ji}\}$ time series. For instance,

$$v(\phi_i)(a_j) = (\tilde{\varphi}_{j+1,i} - \tilde{\varphi}_{j-1,i})/2 + O(\delta t^2) \tag{4}$$

is a second-order centered approximation which we will use in sections 3 and 4 ahead.

After solving the generalized eigenvalue problem in (3), we sort the computed eigenfunctions in order of increasing Dirichlet energy $E(\hat{z}_i)$. The latter can be conveniently computed from the 2-norm of the vector of the expansion coefficients of the $\hat{z}_i$ in the $\{\varphi_i\}$ basis. That is, we have $E(\hat{z}_i) = \|c_i\|^2$, where we have assumed that $\hat{z}_i$ has been normalized to unit norm on $L^2(M, \mu)$. We then construct the feature map in (1) by selecting the first m eigenfunctions corresponding to rationally independent eigenvalues in that sequence. Note that ordering the eigenfunctions with respect to the Dirichlet energy is important because the set of the eigenvalues can be dense (yet countable) on the imaginary line (G15).

In summary, the algorithm described above builds a feature map with projectible dynamics and timescale separation, and also endows this map with high smoothness for the given observation modality. The Koopman modes are given by taking the $L^2(M, \mu)$ inner product between the eigenfunctions and the observation map as stated in section 1, i.e., $\hat{F}_i = \langle z_i, F \rangle$. Using the Koopman modes and the eigenfunctions, we can reconstruct spatiotemporal patterns in data space though

$$Y_{it} = \mathrm{Re}(\hat{F}_i z_i(a_t)), \tag{5}$$

and in the limit $l = n$, the sum $\sum_i Y_{it}$ recovers the input data y_t exactly.

3. Demonstration for traveling-wave synthetic data

We begin by demonstrating the feature extraction technique described in section 2 for a spatiotemporal signal

$$u(x, t) = (0.5 + \sin x)[2 \cos(k_1 x - \theta_1(t)) + 0.5 \cos(k_2 x - \theta_2(t))] \tag{6}$$

defined in a 1D periodic domain, $x \in [0, 2\pi)$. In (6), k_1 and k_2 are integer-valued wavenumbers set to $k_1 = 2$ and $k_2 = 10$, and $\theta_1(t)$ and $\theta_2(t)$ are time-dependent phases such that $\theta_i(t) = \omega_i t$ for the rationally independent frequencies $\omega_1 = 2\pi/45$ and $\omega_2 = \sqrt{10}\omega_1$. This signal was chosen following Kikuchi and Wang (2010) as a simple model for three basic features in the convective variability of the tropical atmosphere as a function of longitude (x): (1) a time-independent profile, $0.5 + \sin x$, representing enhanced convective activity over warm oceans such as the Indian and western Pacific Oceans and suppressed activity over cold oceans such as the eastern Pacific and continental land; (2) a long-wavelength eastward-propagating wave, $\cos(k_1 x - \theta_1(t))$, representing a large-scale mode of organized convection called Madden-Julian oscillation (MJO); (3) a short-wavelength westward-propagating wave representing the building blocks of the MJO (so-called convectively coupled equatorial waves). The natural time units in (6) are days so that the long wave has a period of 45 days and the period of the short wave is approximately 14 days. These periods are comparable to the timescales observed in nature (see section 4 ahead).

From a dynamical systems standpoint, this signal can be described as the outcome of an ergodic linear flow on the 2-torus. In this case, the state space is $M = \mathbb{T}^2$ and the generator of the Koopman group is the vector field $v = v_1 \frac{\partial}{\partial \theta_1} + v_2 \frac{\partial}{\partial \theta_2}$ such that $v_i = \omega_i$ and $\dot{\theta}_i = v(\theta_i)$. The signal in (6) is generated by observing this dynamical system through the observation map $F : M \mapsto L^2(X)$ where X is the circle equipped with the canonical arclength metric, and for the point $a_t \in \mathbb{T}^2$ with coordinates $(\theta_1(t), \theta_2(t))$ we have $F(a_t) = y_t$ where $y_t(x) = u(x, t)$. With this embedding and the kernel in (2), M inherits the Riemannian metric h with components

$$h_{ij} = C \int_0^{2\pi} \frac{\partial u}{\partial \theta_i} \frac{\partial u}{\partial \theta_j} \, dx,$$

where C is a normalization constant chosen such that the volume element of h is compatible with the equilibrium measure of the dynamics. It is straightforward to check that $h_{ii} = r_i^2 k_i^2$ for positive constants r_i, and that $h_{ij} = 0$ for $j \neq i$. Thus, h is a flat metric assigning radii equal to $k_i r_i$ to the 2-torus.

In this example, v has constant coefficients and therefore the Koopman eigenfunctions are Fourier exponential functions on $\mathbb{T}^2$; i.e., $z_{ij}(a) = e^{\mathrm{i}(i\theta_1 + j\theta_2)}$ for two integers i and j, and the corresponding eigenvalues are $\lambda_{ij} = \mathrm{i}(i\omega_1 + j\omega_2)$. Note that because ω_1 and ω_2 are rationally independent, the set $\{\lambda_{ij}\}_{i,j \in \mathbb{Z}}$ is dense on the imaginary line. As stated in section 2, eigenfunctions with low roughness can be selected from this set by examining the corresponding Dirichlet energy in the h metric. In this case, we have $E(z_{ij}) = i^2 k_1^2 r_1^2 + j^2 k_2^2 r_2^2$, and therefore the smoothest eigenfunctions corresponding to rationally independent frequencies are $z_{10}(a) = e^{\mathrm{i}\theta_1}$ and $z_{01}(a) = e^{\mathrm{i}\theta_2}$. Clearly, the feature map in (1) constructed from these eigenfunctions recovers the 2-torus in $\mathbb{C}^2$ from the infinite-dimensional spatiotemporal dataset. Moreover, the corresponding Koopman modes, given by

$$\hat{F}_{10} = \langle z_{10}, F \rangle = A_{10}(1 + 0.5 \cos x) e^{\mathrm{i}k_1 x}, \quad \hat{F}_{01} = \langle z_{01}, F \rangle = A_{01}(1 + 0.5 \cos x) e^{\mathrm{i}k_2 x}$$

for constant A_{10}, A_{01}, recover the spatial profiles associated with the two traveling waves.

To verify that our algorithm is consistent with these results, we generated a spatiotemporal signal from (6) which we sampled uniformly at temporal and spatial intervals $\delta t = 0.14$ days and $\delta x = 2\pi/70$, respectively. We collected a total of $n = 62{,}354$ temporal samples

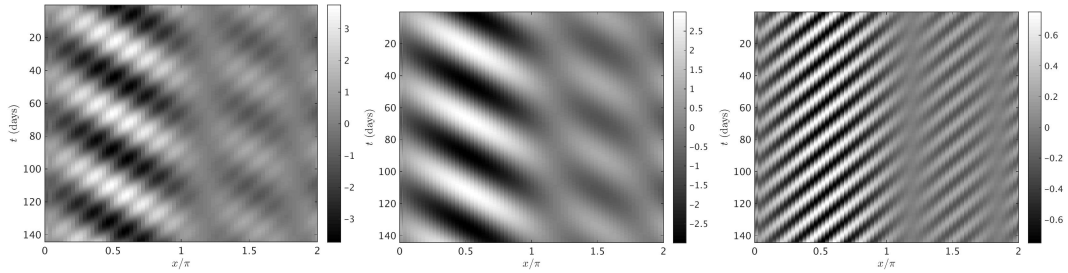

Figure 1: (a) Spatiotemporal dataset on a 1D periodic domain from (6). (b, c) Reconstructed data via (5) for the Koopman eigenfunctions z_{10} and z_{01} respectively.

(i.e., approximately 100 multiples of the long-wave period) to produce the dataset shown in figure 1(a). We solved the generalized eigenvalue problem for the Koopman eigenfunctions in (3) using $l = 100$ basis functions and the regularization parameter value $\nu = 10^{-6}$. Selecting the eigenfunctions with the minimal Dirichlet energy and rationally independent eigenfunctions as described in section 2 leads to the reconstructed data shown in figure 1(b, c). Visually, it is evident that the reconstructions successfully separate the two traveling waves in the input signal, while retaining the (non-dynamical) amplitude profile in x. The numerical eigenvalues, 0.13961i and 0.44122i, agree with the exact values, $\mathrm{i}\omega_1 \approx 0.13963\mathrm{i}$ and $\mathrm{i}\omega_2 \approx 0.44154\mathrm{i}$, up to a relative error of 1.4×10^{-5} and 3.1×10^{-4}, respectively. Comparably accurate eigenvalues can also be obtained using $l = 20$ basis functions.

4. Application to convectively coupled waves in the atmosphere

As a real-world application of our technique, we analyze brightness temperature (T) data from the CLAUS multi satellite archive (Hodges et al., 2000). Brightness temperature is a measure of the Earth's infrared emission in terms of the temperature of a hypothesized blackbody emitting the same amount of radiation at the same wavelength. In the tropics, atmospheric convection produces deep, cold clouds and an associated strong negative T anomaly. For this reason, and due to the availability of high-resolution data dating back to the early 1980s, satellite observations of brightness temperature are widely used to study the dynamics of tropical atmospheric convection.

While part of the tropical convective variability has a stochastic, red-noise character, it also exhibits a significant traveling wave-like coherent component. These disturbances propagate parallel to the equator, and are organized in a hierarchy of scales. Prominent coherent structures in this hierarchy are cloud clusters and mesoscale convective systems with horizontal scales of the order a few hundred kilometers and lifetimes of a few hours. These objects form the building blocks of the larger and longer-lived convectively coupled equatorial waves (CCEWs), which in turn organize into the planetary-scale MJO mentioned in section 3. Through various teleconnection mechanisms, the MJO exerts global influences on weather and climate variability, linking short-term weather forecasts and long-term climate projections. Currently, efforts to extract this hierarchy from observational data are

consistent only on the coarse features of large-scale structures such as the MJO (but with important differences in the details), and evidence of preferred organized patterns for the MJO building blocks has been lacking (Dias et al., 2013). Here, we demonstrate that, applied to brightness temperature data, the Koopman eigenfunction approach described in sections 1 and 3 reveals a multiscale hierarchy of modes on timescales spanning years to days, including the MJO but also traveling waves on timescales characteristic of CCEWs. While our results are still preliminary, to our knowledge, this is the first time that evidence of such structures has been detected via comparable eigendecomposition techniques applied to brightness temperature observational data.

In these experiments, we study brightness temperature data sampled on a uniform longitude-latitude grid of 0.5° resolution, and observed every 3 hours for the period July 1, 1983 to June 30, 2006. Following Tung et al. (2014), we average the data over the tropical belt between 15°S and 15°N antisymmetrically about the equator to produce a 1D spatiotemporal signal $T(x,t)$. Portions of the raw data for longitudes over the Indian and Western Pacific Oceans (where strong MJO activity takes place) are shown in Fig. 3(a, e).

A major challenge for analyzing these data via Koopman techniques is that the evolution of $T(x,t)$ is not governed by an autonomous dynamical system; this is because the tropical atmosphere interacts with a multitude of other degrees of freedom of the Earth's climate system and knowledge of $T(x,t)$ at a given time is not sufficient to uniquely determine its evolution at a later time. One way of partially overcoming this obstacle (which was adopted by Tung et al. (2014) and we also adopt here) is to embed the data in a higher dimensional space via delay-coordinate maps (Sauer et al., 1991). Specifically, selecting an integer parameter q we construct the time series

$$u(x,t) = (T(x,t), T(x,t-\delta t), \ldots, T(x,t-(q-1)\,\delta t)). \tag{7}$$

Due to a theorem of Takens, for sufficiently large q, the signal $u(x,t)$ is expected to be more Markovian than the individual snapshots. Here, we set q to 512, corresponding to a time interval of 64 days for our 3-hour sampling interval. Tung et al. (2014) found that this value is sufficient to recover the MJO and other physically meaningful patterns using kernel algorithms. After delay-coordinate mapping, the number of samples available for analysis is $n = 66{,}693$ and the ambient space dimension is $q \times d = 368{,}640$, where $d = 720$ is the number of sampled longitude points at the 0.5° resolution. We have computed diffusion eigenfunctions for this data using the kernel algorithm described in section 2 with an intrinsic dimension parameter $\hat{m} = 3$ (this is an underestimate of the true intrinsic dimension of the data, but we found that the results are not too sensitive in the choice of $\hat{m}$).

Figures 2 and 3 show representative Koopman eigenfunctions and the associated spatiotemporal reconstructions computed through the eigenvalue problem in (3) using $l = 110$ diffusion eigenfunctions and the regularization parameter value $\nu = 0.02$. Among the modes shown here one is periodic (figure 2(a, f)) and does not exhibit propagation in space (figure 3(b, f)), and the other modes are amplitude-modulated traveling waves. The periodic mode represents the seasonal cycle (a significant component of brightness temperature variability), and the amplitude-modulated modes correspond to various types of eastward (figure 3(c, g)) and westward (figure 3(d, h)) traveling convective organization, which we now describe.

The mode in figures 2(c, h) and 3(c) represents the MJO—a planetary-scale ($\sim$ 20,000 km wavelength) envelope of convective activity forming over the Indian Ocean at $\sim 60°$E latitudes and propagating eastward until it dissipates upon reaching the central Pacific Ocean at $\sim 160°$W latitudes. The MJO is mainly active in the boreal winter (November–March), and its dominant frequency is in the intraseasonal scale of one cycle per $\sim$ 60 days. During the boreal summer (May–September), the MJO is replaced by the so-called boreal summer intraseasonal oscillation (BSISO), shown in figures 2(b, g) and 3(g). This pattern originates over the Indian Ocean and propagates northeastward towards the Indian subcontinent where it affects the variability of the Indian Monsoon on $\sim$ 50 day timescales.

The modes in figures 2(d, e, i, j) and 3(d, h) are westward-propagating traveling waves with periods in the 10–20 day range. Spatially, these modes are mainly active at longitudes associated with the Western Pacific warm pool, a region of high sea surface temperature and moisture that favors deep convection. Based on their timescales and qualitative spatial features, we believe that these modes represent CCEWs and may act as the building blocks of the larger-scale intraseasonal oscillations. These waves have so far alluded detection by means of objective data analysis algorithms (i.e., algorithms that do not apply ad hoc preprocessing to the data to isolate the temporal and spatial scales of interest).

Despite the success of our method to extract a multiscale hierarchy of convective organization, it is important to note a feature of our results which is not theoretically compatible with the framework of Koopman eigenfunctions, namely strong amplitude modulations. Such modulations, which are evident in figure 2, are theoretically precluded from the fact that the Koopman operators are unitary with respect to the invariant measure of the dynamics—this forces the eigenfunctions to lie on a circle in the complex plane, i.e., $|z|$ must be constant, and clearly this is not the case in figure 2. A likely source of this discrepancy is non-Markovianity of the data. That is, the delay-coordinate mapping in (7) is highly unlikely to have resolved all unobserved dynamical degrees of freedom of the climate system, meaning that knowledge of the signal $u(x,t)$ at a given time is insufficient to uniquely determine its evolution at later time. In this scenario, a stochastic description of the dynamics would be more appropriate, where the Koopman group of unitary operators is replaced by a semigroup generated by a Fokker-Planck operator. Indeed, Chen et al. (2014) show that MJO time series such as those in figure 2(c) are well described by stochastic oscillators, and in the stochastic case time shift methods related to (4) can be used to approximate the solution semigroup (Berry et al., 2015). We plan to study the spectral properties of these operators and their application to feature extraction in future work.

Acknowledgments

We acknowledge support from the National Science Foundation (grant DMS-1521775), Office of Naval Research (DRI grant N00014-14-1-0150 and MURI grant 25-74200-F7112), and Center for Prototype Climate Modeling at New York University Abu Dhabi. Part of this research was carried out on the high performance computing resources at New York University Abu Dhabi.

References

T. Berry and J. Harlim. Variable bandwidth diffusion kernels. Appl. Comput. Harmon. Anal., 2015. In press.

T. Berry and T. Sauer. Local kernels and the geometric structure of data. Appl. Comput. Harmon. Anal., 2015. In press.

T. Berry, D. Giannakis, and J. Harlim. Nonparametric forecasting of low-dimensional dynamical systems. *Phys. Rev. E.*, 91:032915, 2015. doi: 10.1103/PhysRevE.91.032915.

G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Mach. Learn.*, 66:259–294, 2007. doi: 10.1007/s10994-006-6895-9.

M. Budisić, R. Mohr, and I. Mezić. Applied Koopmanism. *Chaos*, 22:047510, 2012. doi: 10.1063/1.4772195.

N. Chen, A. J. Majda, and D. Giannakis. Predicting the cloud patterns of the Madden-Julian Oscillation through a low-order nonlinear stochastic model. *Geophys. Res. Lett.*, 41(15):5612–5619, 2014. doi: 10.1002/2014gl060876.

R. R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21:5–30, 2006. doi: 10.1016/j.acha.2006.04.006.

R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer. Graph laplacian tomography from unknown random projections. *IEEE Trans. Image Process.*, 17(10):1891–1899, 2008. doi: 10.1109/tip.2008.2002305.

J. Dias, S. Leroux, S. N. Tulich, and G. N. Kiladis. How systematic is organized tropical convection within the MJO? *Geophys. Res. Lett.*, 40:1420–1425, 2013. doi: 10.1002/grl.50308.

D. Giannakis. Data-driven spectral decomposition and forecasting of ergodic dynamical systems. *Appl. Comput. Harmon. Anal.*, 2015. In review. arXiv:1507.02338.

M. Hein and J.-Y. Audibert. Intrinsic dimensionality estimation of submanifolds in Euclidean space. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 289–296, 2005.

K. Hodges, D.W. Chappell, G.J. Robinson, and G. Yang. An improved algorithm for generating global window brightness temperatures from multiple satellite infrared imagery. *J. Atmos. Oceanic Technol.*, 17:1296–1312, 2000. doi: 10.1175/1520-0426(2000)017<1296:aiafgg>2.0.co;2.

K. Kikuchi and B. Wang. Spatiotemporal wavelet transform and the multiscale behavior of the Madden-Julian oscillation. *J. Climate*, 23:3814–3834, 2010. doi: 10.1175/2010jcli2693.1.

B. O. Koopman. Hamiltonian systems and transformation in Hilbert space. *Proc. Natl. Acad. Sci.*, 17(5):315–318, 1931.

J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Information Science and Statistics. Springer, New York, 2007.

A. V. Little, J. Lee, Y.-M. Jung, and M. Maggioni. Estimation of intrinsic dimensionality of samples from noisy low-dimensional manifolds in high dimensions with multiscale SVD. In *Proceedings of the 15th IEEE/SPWorkshop on Statistical Signal Processing*, pages 85–88, 2009. doi: 10.1109/SSP.2009.5278634.

I. Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dyn.*, 41:309–325, 2005. doi: 10.1007/s11071-005-2824-x.

I. Mezić. Analysis of fluid flows. *Analysis of Fluid Flows via Spectral Properties of the Koopman Operator*, 45:357–378, 2013. doi: 10.1146/annurev-fluid-011212-140652.

C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson. Spectral analysis of nonlinear flows. *J. Fluid Mech.*, 641:115–127, 2009. doi: 10.1017/s0022112009992059.

T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *J. Stat. Phys.*, 65(3–4):579–616, 1991. doi: 10.1007/bf01053745.

D. Storcheus, M. Mohri, and A. Rostamizadeh. Foundations of coupled nonlinear dimensionality reduction, 2015.

W.-w. Tung, D. Giannakis, and A. J. Majda. Symmetric and antisymmetric signals in MJO deep convection. Part I: Basic modes in infrared brightness temperature. *J. Atmos. Sci.*, 71:3302–3326, 2014. doi: 10.1175/jas-d-13-0122.1.

M. O. Williams, I. G. Kevrekidis, and C. W. Rowley. A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. J. Nonlinear Sci., 2015.

Appendix A.

This Appendix contains plots of representative Koopman eigenfunctions (figure 2) and spatiotemporal reconstructions (figure 3) for the CLAUS brightness temperature data studied in section 4. The eigenfunctions are shown for a two-year portion of the 23-year dataset from January 1, 1992 to December 31, 1993. The reconstructions are for three-month portions of this interval highlighting the active MJO (January 1, 1992 to March 31, 1992; figure 3(a–d)) and active BSISO periods (August 1, 1993 to October 31, 1993; figure 3(e–h)).

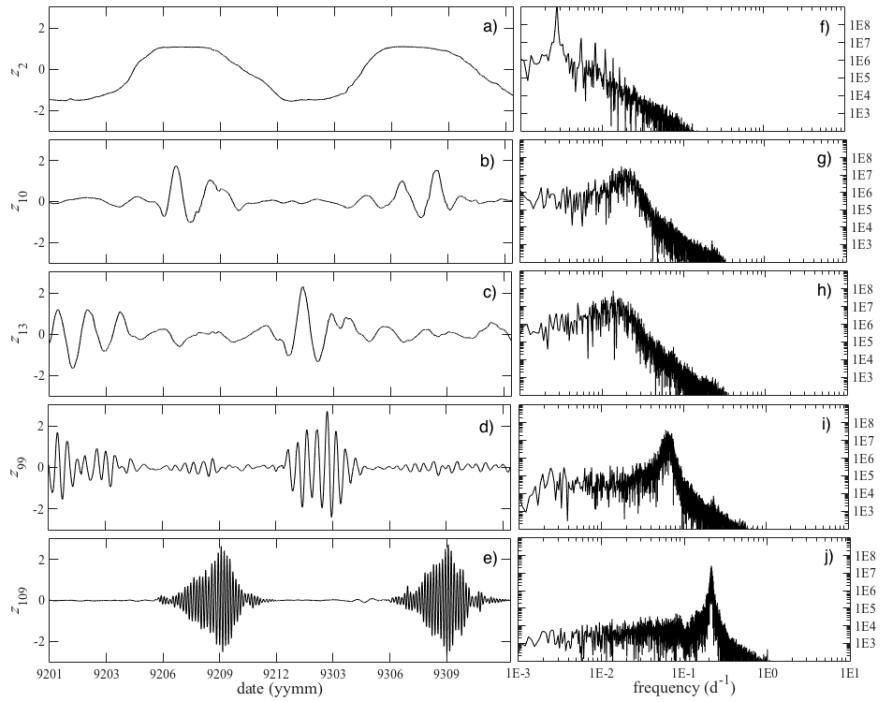

Figure 2: Time series (a–e) and power spectral densities (f–j) of Koopman eigenfunctions for brightness temperature data. (a, f) Seasonal cycle; (b, g) boreal summer intraseasonal oscillation; (c, h) Madden-Julian oscillation; (d–j) westward-propagating convectively coupled equatorial waves.

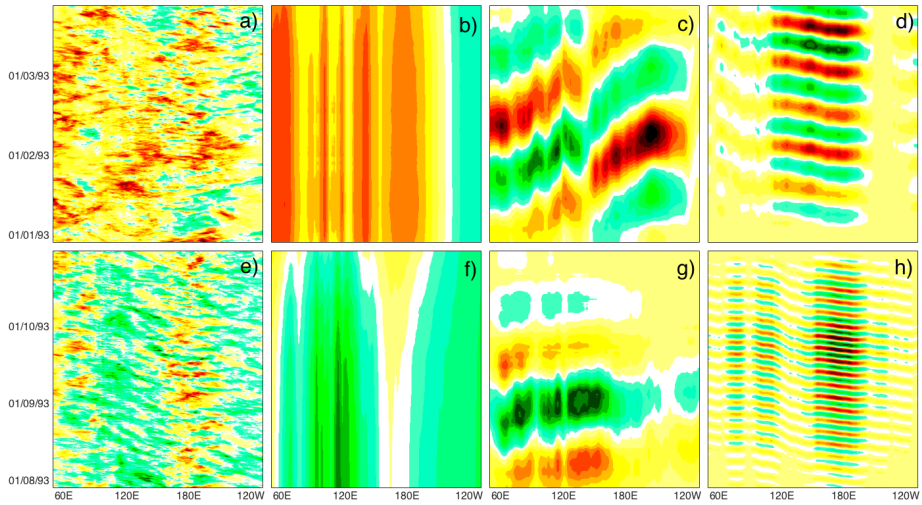

Figure 3: Raw data (a, e) and spatiotemporal reconstructions (b–d, f–h) of brightness temperature for the Koopman eigenfunctions in figure 2. Panels (a–d) and (e–h) show reconstructions for three-month intervals in the boreal winter (January–March) and boreal summer (August–October), respectively. The modes depicted here are the annual cycle (b, f), the Madden-Julian oscillation (c), the boreal summer intraseasonal oscillation (g), and westward-propagating convectively coupled equatorial waves (d, h).