Manifold Learning With Contracting Observers for Data-Driven Time-Series Analysis

Tal Shnitzer, Ronen Talmon, *Member, IEEE*, and Jean-Jacques Slotine

***Abstract*—Analyzing signals arising from dynamical systems typically requires many modeling assumptions. In high dimensions, this modeling is particularly difficult due to the "curse of dimensionality." In this paper, we propose a method for building an intrinsic representation of such signals in a purely data-driven manner. First, we apply a manifold learning technique, diffusion maps, to learn the intrinsic model of the latent variables of the dynamical system, solely from the measurements. Second, we use concepts and tools from control theory and build a linear contracting observer to estimate the latent variables in a sequential manner from new incoming measurements. The effectiveness of the presented framework is demonstrated by applying it to a toy problem and to a music analysis application. In these examples, we show that our method reveals the intrinsic variables of the analyzed dynamical systems.**

***Index Terms*—Intrinsic modeling, manifold learning, linear observer, diffusion maps, non-parametric filtering.**

I. Introduction

HIGH dimensional signals generated by dynamical systems arise in many fields of science. For example, biomedical signals such as EEG and EMG can be modeled by few latent processes measured by a large set of noisy sensors. In such applications the goal is to identify the latent intrinsic variables which describe the true, intrinsic state of the system.

Analyzing such signals typically requires vast modeling assumptions. For example, Bayesian filtering methods require a priori knowledge of the statistical model and often rely on parameter estimation [1], [2]. Finding appropriate models and estimating their parameters from high dimensional data is challenging, since the "curse of dimensionality" leads to failure of many data analysis techniques that perform well for low dimensional data.

We approach the problem of high dimensional signal analysis in dynamical systems from a geometric modeling standpoint, by applying manifold learning techniques. From this standpoint, the main assumption is that the accessible high dimensional data (the observations of the system) lie on an underlying nonlinear manifold of lower dimensions. In the past decade, various manifold learning methods have been introduced, e.g., [3]–[6], in which the geometry of the underlying manifold is captured in a data-driven manner, and the data are embedded in a low dimensional space, thereby attaining dimensionality reduction. In classical manifold learning, time series are processed as data sets of samples, ignoring their embodied dynamics and temporal order. Recently, several methods have addressed this problem and incorporated the time dependency of consecutive samples into the manifold learning framework [7]–[12]. For example, in [9]–[11] non-parametric frameworks are presented, based on manifold learning. In these papers, Berry and Harlim suggest a probabilistic approach for filtering and forecasting in dynamical systems governed by gradient flows in [10] and by more general drift-diffusion equations in [9] and [11]. This is performed by applying diffusion maps [6] and projecting the filtering problem onto a basis created by the diffusion maps coordinates, which are the eigenfunctions of the backward Fokker-Planck operator [13]. These coordinates are used to represent the probability density of the evolution of the system state which allows for estimation and forecasting. However, in their setting, the state of the system is assumed to be known and accessible (up to additive noise). In our work, we consider a setting in which the state is unknown and it is revealed using a state-space framework in a purely data-driven manner, given measurements of the state through an unknown, non-linear measurement function. In addition, we suggest a framework for the extension of the state representation to unseen measurements which allows for the analysis of large data sets.

In another work [7], a non-parametric, Bayesian framework is presented, which incorporates the dependency of consecutive time samples into diffusion maps. However, this framework assumes a Gaussian setting, i.e. given the underlying state the measurements are assumed to be locally Gaussian, and it requires the estimation of the mean value and the local covariance matrices of the observations.

In this paper, we use geometric analysis tools to capture the inherent structure of the observations and their dynamics. We exploit the recovered geometry, along with the smoothness in time, to construct a representation of the underlying state. For this purpose, a filtering framework is introduced, based on diffusion maps [6], [14] in a setting which is non-parametric and non-Gaussian. We particularly address time series analysis and propose an approach consisting of two steps: (i) learning the intrinsic model of the latent variables of the signal solely from measurements, and then, (ii) estimating the latent variables in a sequential manner from new incoming measurements. Such an approach allows us to analyze and process real signals without

Manuscript received April 22, 2016; revised August 24, 2016 and September 28, 2016; accepted September 30, 2016. Date of publication October 11, 2016; date of current version December 5, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. Morten Mørup. This work was supported in part by the European Union's Seventh Framework Programme (FP7) under Marie Curie Grant 630657.

T. Shnitzer and R. Talmon are with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: shnitzer@campus.technion.ac.il; ronen@ee.technion.ac.il).

J.-J. Slotine is with the Nonlinear Systems Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: jjs@mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2016.2616334

existing adequate models in the current literature. More specifically, we show that our method reveals the intrinsic state, its dynamics and its relationship to the observations based on the measurements. Furthermore, we show that the drift (the deterministic dynamics) of the constructed diffusion maps coordinates is *linear*, even when modeling highly *nonlinear* systems. This allows us to devise a framework, which incorporates a standard processing technique for time series – an observer, and devise a method, which is especially designed to handle noisy data. Finally, we apply our framework to a toy example and to a practical application of music analysis, in which we show that our method reveals intrinsic variables describing dominant musical notes as well as different musical instruments.

The paper is organized as follows. In Section II the general setting and formulation of the problem are presented. In Section III the diffusion maps framework is introduced as a method to recover the dimension and dynamics of the system. Section IV presents a contracting observer, which is constructed based on the recovered information and learned dynamics, in order to reconstruct the state of the dynamical system. Section V illustrates the advantages of the constructed observer by applying our framework to a toy example and to a music analysis application. Section VI offers brief concluding remarks.

II. Problem Formulation

Many problems involving high dimensional signals can be modeled using a state-space formulation. In this framework we are given a set of high-dimensional measurements $\boldsymbol{z}(t) \in \mathbb{R}^n$. We assume that the measurements are samples from a dynamical system of the form

$$\dot{\boldsymbol{\theta}}(t) = f(\boldsymbol{\theta}(t), \dot{\boldsymbol{\omega}}(t)) \tag{1}$$

$$\boldsymbol{z}(t) = g(\boldsymbol{\theta}(t), \boldsymbol{v}(t)) \tag{2}$$

where $\boldsymbol{v} \in \mathbb{R}^n$ is a white noise process, $\boldsymbol{\omega} \in \mathbb{R}^d$ is Brownian motion and $\dot{\boldsymbol{\omega}}$ is the time derivative, $\boldsymbol{\theta}(t) \in \mathbb{R}^d$ is the true state of the system and f, g are non-linear functions which represent the system dynamics and measurement function respectively.

We focus on the case in which f is modeled as the following autonomous SDE:

$$\dot{\boldsymbol{\theta}}(t) = a(\boldsymbol{\theta}(t)) + b(\boldsymbol{\theta}(t))\dot{\boldsymbol{\omega}}(t) \tag{3}$$

$$\boldsymbol{z}(t) = h(\boldsymbol{\theta}(t)) + \boldsymbol{v}(t) \tag{4}$$

where $a(\boldsymbol{\theta})$ and $b(\boldsymbol{\theta})$ are the drift and diffusion coefficients. In this particular case, the drift and diffusion terms are given by

$$a(\boldsymbol{\theta}) = -\nabla U(\boldsymbol{\theta}) \tag{5}$$

$$b(\boldsymbol{\theta}) = \sqrt{\frac{2}{\beta}} \tag{6}$$

where $U(\boldsymbol{\theta})$ is referred to as a potential field and $\sqrt{2/\beta}$ is a constant diffusion coefficient. The SDE in (3) is known as the Langevin equation, which describes the evolution of the state $\boldsymbol{\theta}(t)$ according to the potential $U(\boldsymbol{\theta})$ at an inverse temperature β. This terminology is derived from the physical problem of particle motion in fluid, where the potential $U(\boldsymbol{\theta})$ represents the motion component which drives the particles to high density areas. We assume that the potential $U(\boldsymbol{\theta})$ is smooth and bounded (relaxation of this assumption is described in [15]), consequently, it is reasonable to assume that (3) describes a diffusion process, confined to a finite, compact, connected region $\mathcal{M} \subseteq \mathbb{R}^d$ with smooth reflecting boundaries [15].

Our goal here is to reveal the underlying dynamical process, $\boldsymbol{\theta}(t)$, given the noisy measurements, $\boldsymbol{z}(t)$, in a data-driven setting without additional model assumptions. In this setting, all the parameters of the dynamical system are unknown, i.e., the potential function $U(\boldsymbol{\theta})$, the thermal factor β, the measurement function $h : \mathbb{R}^d \to \mathbb{R}^n$, as well as the dimensionality d and the coordinate system of the state $\boldsymbol{\theta}(t)$.

We present the Ornstein Uhlenbeck equation, a simple case of (3), as an example, which is implemented in Section V to demonstrate the properties and efficacy of our proposed method. The Ornstein Uhlenbeck equation is described as follows:

$$\dot{\boldsymbol{\theta}}(t) = k(\mu - \boldsymbol{\theta}(t)) + \sigma\dot{\boldsymbol{\omega}}(t)$$

where μ is the long term mean of the process, $k > 0$ is the rate that the process reverts to its mean value, and σ is the diffusion coefficient. This equation describes a noisy relaxation process, e.g. over damped spring in the presence of thermal fluctuations.

We note that for simplicity, we omit the notation of (t) in the following sections. However, all presented coordinates and processes are still time dependent.

The model described in (3) can represent a vast group of physical phenomena, e.g. thermal noise processes, non-ideal harmonic oscillators, diffusive particle motions, and financial processes [16]. In addition, it was previously used for empirical modeling of EEG signals [17] and of audio signals [18], [19]. Due to its generality, most dynamical systems can be roughly described by this model, however, our proposed method is not confined to the model and we present results on a real application in which the model can only be assumed. We note that, here, the main purpose of the Langevin equation model (3) is to provide a solid theoretical foundation.

III. Discovering the System Model With Diffusion Maps

In this section we present the basic analysis showing how diffusion maps can be used to reveal the model of a dynamical system from measurements in a data-driven manner. This analysis is carried out assuming that there is no measurement noise. When noise is present, the following analysis no longer holds and the ability of the diffusion maps to accurately recover the model parameters is hampered. In Section IV, we extend this framework for noisy systems by explicitly incorporating the system dynamics through the implementation of a contracting observer.

A. Revealing the Dynamics of the Intrinsic State

Consider the state dynamics (3). Under the assumptions outlined in Section II, the local equilibrium transition probability of this process is given by $p_{eq}(\boldsymbol{\theta}) = e^{-U(\boldsymbol{\theta})}$. For such a differential equation, the transition probability density $p(\boldsymbol{\theta}, t|\boldsymbol{\theta}_0, t_0)$

of finding the system at time t and at location $\boldsymbol{\theta}$, given an initial location $\boldsymbol{\theta}_0$ at time t_0, satisfies the backward Fokker-Planck equation:

$$\frac{\partial p}{\partial t} = \mathcal{L}p = \frac{1}{\beta}\Delta p - \nabla U \cdot \nabla p \tag{7}$$

where Δ is the Laplacian operator.

Since we assume that the potential U is smooth, it can be shown [13], [20] that the operator $\mathcal{L}$ has a discrete spectrum of non-positive decreasing eigenvalues $\{-\lambda_\ell\}_{\ell=0}^{\infty}$ with associated eigenfunctions which satisfy:

$$\mathcal{L}\psi_\ell = -\lambda_\ell \psi_\ell \tag{8}$$

Based on Itô's lemma, each of these eigenfunctions also evolve according to the following stochastic differential equation [15]:

$$\dot{\psi}_\ell = -\lambda_\ell \psi_\ell + \tilde{b}_\ell\left(\psi_\ell, t\right)\dot{\omega}_\ell,\ \ \ell = 0, 1, 2 \ldots \tag{9}$$

where $\tilde{b}_\ell(\psi_\ell, t)$ has a known closed-form expression, and $-\lambda_\ell$ are the eigenvalues of $\mathcal{L}$. See details in Appendix A.

The eigenvalues and eigenfunctions of $\mathcal{L}$ along with their dynamics (9) play a pivotal role in this work; their importance is three-fold. First, the eigenfunctions of the Fokker-Planck operator form a parametrization of the state, since the solution of (7) can be written based on these eigenfunctions as $p\left(\boldsymbol{\theta}, t|\boldsymbol{\theta}_0, 0\right) = \sum_{\ell=0}^{\infty} c_\ell e^{-\lambda_\ell t}\psi_\ell\left(\boldsymbol{\theta}\right)$, where the coefficients c_ℓ are determined by the initial conditions at $t = 0$ [15]. Second, the dynamics of this parametrization is revealed in (9), which illustrates that the resulting eigenfunctions, describing the long term behavior of the given diffusion process, evolve according to a *linear drift, determined by the eigenvalues of the operator*, with some additional noise process. Third, in Section III-B, we show that both the parametrization and the dynamics, which are based on the eigenfunctions and eigenvalues of the Fokker-Planck operator, can be approximated from the data without prior knowledge of the system.

B. Data-Driven Manifold Learning and Diffusion Maps

Recall that the objective is to recover the underlying state given the measurement process $\boldsymbol{z}\left(t\right) \in \mathbb{R}^n$. For this purpose, as described in Section III-A, we wish to compute the eigenfunctions of the backward Fokker-Planck operator describing the diffusion process, which are used to represent the underlying state. In this section, we first present the diffusion maps framework [6], [14], a computational method to obtain these eigenfunctions from the underlying state $\boldsymbol{\theta}\left(t\right)$, without prior information on the components comprising the right-hand side of (7). Second, in Section III-C, we present a method for obtaining the required information on $\boldsymbol{\theta}\left(t\right)$ based on the measurements $\boldsymbol{z}\left(t\right)$.

Based on the underlying state $\boldsymbol{\theta}\left(t\right)$, we build a pairwise affinity kernel $k_\epsilon\left(t, s\right)$ according to

$$k_\epsilon\left(t, s\right) = \exp\left\{-\frac{\left\|\boldsymbol{\theta}\left(t\right) - \boldsymbol{\theta}\left(s\right)\right\|^2}{\epsilon}\right\} \tag{10}$$

where $\epsilon > 0$. Here, $\left\|\cdot\right\|^2$ denotes the squared Euclidean norm, and ϵ is the kernel scale which denotes a characteristic distance within the data set. In other words, ϵ induces a notion of locality: if $\left\|\boldsymbol{\theta}\left(t\right) - \boldsymbol{\theta}\left(s\right)\right\|^2 \gg \epsilon$, then $k_\epsilon\left(t, s\right)$ is negligible.

The kernel is normalized as follows:

$$p_\epsilon\left(t, s\right) = \frac{k_\epsilon\left(t, s\right)}{d_\epsilon\left(t\right)} \tag{11}$$

where $d_\epsilon\left(t\right) = \int k_\epsilon\left(t, s\right) p_{eq}\left(s\right) ds$ and $p_{eq}\left(s\right) = e^{-U\left(\boldsymbol{\theta}(s)\right)}$ is the equilibrium density of the underlying state parameter $\boldsymbol{\theta}\left(s\right)$.

Define the operator P_ϵ on any real function $g\left(\cdot\right)$ on $\mathcal{M}$ by

$$\left(P_\epsilon g\right)\left(t\right) = \int p_\epsilon\left(t, s\right) g\left(s\right) p_{eq}\left(s\right) ds \tag{12}$$

In the limit $\epsilon \to 0$, the operator

$$L_\epsilon = \frac{1}{\epsilon}(P_\epsilon - \mathrm{I}) \tag{13}$$

converges to the backward Fokker-Planck operator (7) [15], [21], where I is the identity operator. As a result, the eigenfunctions of L_ϵ approximate the eigenfunctions of the backward Fokker-Planck operator, and therefore, inherit all the properties described in the remainder of this subsection.

The eigenvalue decomposition of the Fokker-Planck operator generates a discrete spectrum of eigenvalues $\{-\lambda_\ell\}_{\ell=0}^{\infty}$ containing several dominant eigenvalues. In addition, these eigenvalues are decreasing, therefore, we can assume that the first m eigenvalues are the dominant ones and approximate the dynamical process (3) by a finite set of eigenfunctions and eigenvalues $\ell \in \{0, 1, ..., m\}$. We construct a representation of the state based on these eigenfunctions, creating an embedded space, i.e. a new coordinate system:

$$\left[\psi_1(t), \psi_2(t), \ldots, \psi_m(t)\right]^T \tag{14}$$

In addition, as presented in (9), the dynamics of these constructed coordinates can be approximated based on the corresponding eigenvalues $\{-\lambda_\ell\}_{\ell=0}^{m}$.

C. Nonlinear Measurement Mapping

In Section III-B we show that the eigenfunctions of the backward Fokker-Planck operator give rise to a new coordinate system, with linear dynamics given by the eigenvalues, which appropriately describes the latent intrinsic state of the observed system. Since the underlying state $\boldsymbol{\theta}\left(t\right)$ is inaccessible and we are only given a set of measurements $\boldsymbol{z}(t) = h(\boldsymbol{\theta}\left(t\right))$, to construct the affinity kernel (10) we approximate the required Euclidean distances of $\boldsymbol{\theta}\left(t\right)$ from the measurements by applying a modified version of the Mahalanobis distance presented in [21].

The modified Mahalanobis distance between two measurements, $\boldsymbol{z}\left(t\right)$ and $\boldsymbol{z}\left(s\right)$, is given by:

$$\begin{aligned} &d\left(\boldsymbol{z}\left(t\right), \boldsymbol{z}\left(s\right)\right) \\ &= \frac{1}{2}\left(\boldsymbol{z}\left(t\right) - \boldsymbol{z}\left(s\right)\right)\left(C^{-1}\left(t\right) + C^{-1}\left(s\right)\right)\left(\boldsymbol{z}\left(t\right) - \boldsymbol{z}\left(s\right)\right)^T \end{aligned} \tag{15}$$

where $C\left(t\right)$ is the covariance matrix of the measurements $\boldsymbol{z}$ at time t and $C\left(s\right)$ is the covariance at time s. We note that when n, the dimensionality of the measurements $\boldsymbol{z}\left(t\right)$, is larger than the state dimensionality d, the covariance matrices are not of

full rank and pseudo-inverse is used in (15). Singer *et al.* show that this form of the Mahalanobis distance approximates the Euclidean distance between two corresponding samples of the underlying process, $\boldsymbol{\theta}(t)$:

$$d(\boldsymbol{z}(t), \boldsymbol{z}(s)) = \|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(s)\|^2 + O\left(\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(s)\|^4\right)$$

The derivation of this modified Mahalanobis distance is presented in Appendix B. This holds assuming that the Brownian motions of different coordinates, $\theta_i,\ i = 1, ..., d$, in (3) are independent. Therefore, by constructing the diffusion maps based on this distance, we can recover the parameters of the underlying state, i.e., its dynamics and diffusion maps coordinate system, instead of those describing the measurements.

We can now construct a mapping from the measurements, $\boldsymbol{z}$, to the embedded space, based on the representation in (14):

$$\boldsymbol{z}(t) \mapsto [\psi_1(t), \psi_2(t), \ldots, \psi_m(t)]^T \tag{16}$$

The above setting describes a purely data-driven scheme which provides a representation of the underlying state based solely on the given measurements. However, this framework has two main shortcomings. First, system dynamics are not explicitly expressed in this mapping. Second, the convergence of the constructed operator to the backward Fokker-Plank operator is attained only when noiseless measurements are available. We address these weaknesses in our observer setting presented in Section IV.

D. Modeling the Lift Function

In the presented setting in Section II, the dynamical system is modelled by (3) and (4). The state and its parameters in (3) are unknown and we have access only to the measurements $\boldsymbol{z}(t)$. By applying the modified Mahalanobis distance (15) we gain access to the Euclidean distances between the samples of the underlying state, from the given measurements. Based on these approximated Euclidean distances we can recover the parametrization and dynamics of the underlying state by applying diffusion maps as described in Section III-B. At this point, for the construction of a complete representation of the system, we are still missing (4), even in the noiseless case. In (4), h is a function mapping the latent intrinsic state $\boldsymbol{\theta}(t)$ into the domain of the measurements $\boldsymbol{z}(t)$. Let $g(\cdot)$ be the lift function which is analogous to h, mapping the new recovered coordinate system to the domain of the measurements.

In this section we propose such a lift function, based on the above eigenfunctions. Since the Fokker-Planck operator of the Langevin equation (3) is Hermitian, its eigenfunctions form a basis for all real functions defined on the underlying diffusion process [22]. Based on this concept we define a linear reconstruction function between the new coordinate system and the measurements.

We expand each coordinate of the measurements, $z_j(t)\ \ j = 1, ..., n$, using the eigenfunction basis,

$$z_j(t) = \sum_{\ell=1}^{\infty} \alpha_{j,\ell}\psi_\ell(t),\ \ j = 1, \ldots, n \tag{17}$$

where the expansion coefficients $\alpha_{j,\ell}$ are given by

$$\alpha_{j,\ell} = \langle z_j, \psi_\ell \rangle_q = \int_{-\infty}^{\infty} z_j(t)\psi_\ell(t)p_{eq}(t)\,dt \tag{18}$$

and $p_{eq}(t)$ is the equilibrium density of the underlying state parameter $\boldsymbol{\theta}(t)$.

By assuming that the spectrum of L_ϵ in (13) decays fast, and that for a finite m, the m eigenfunctions associated with the m largest eigenvalues capture most of its energy, z_j can be well approximated by

$$z_j(t) \simeq \sum_{\ell=1}^{m} \alpha_{j,\ell}\psi_\ell(t) \tag{19}$$

Let $\Psi(t)$ denote the parametrization of the state at time t consisting these m eigenfunctions,

$$\Psi(t) = [\psi_1(t), \psi_2(t), \ldots, \psi_m(t)]^T$$

We define $g(\cdot)$ as the mapping from this new coordinate system to the measurements

$$\boldsymbol{z}(t) = g(\Psi(t)). \tag{20}$$

Therefore, in matrix form, (19) can be rewritten as

$$\boldsymbol{z}(t) = g(\Psi(t)) \simeq \boldsymbol{\alpha}\Psi(t) \tag{21}$$

where $\boldsymbol{\alpha}$ is an $n \times m$ matrix whose elements are given by $(\boldsymbol{\alpha})_{j,\ell} = \alpha_{j,\ell}$.

Choosing the optimal dimensionality m is a widely studied problem [23], [24]. Here we apply a heuristic approach, as described in [15], by determining the dimensionality based on the existence and location of a spectral gap in the eigenvalues. This is shown to attain a meaningful, low dimensional parametrization.

To conclude this section, we emphasize that the presented approach yields a data-driven coordinate system of the latent intrinsic state that does not require prior knowledge of the system. The main benefit of this approach is that the constructed coordinate system comprises both linear dynamics as well as a linear lift function. Therefore, it enables us to provide linear solutions, as presented in Section IV, to highly nonlinear problems. We revisit this setting in Section IV-B, where we describe the relation of our approach to the Koopman operator.

IV. The Observer Framework

The derivation presented in Section III is based on a setup, which does not include measurement noise. When considering real, noisy systems, the recovered parametrization is merely an approximation of the results above, leading to inaccurate representations of the underlying system state. In this section, we propose a framework that improves the constructed parametrization for noisy systems by facilitating the dynamics embodied in the measurements, which did not take part in the construction of the embedding. For this purpose, we use concepts and tools from control theory and propose to build a linear contracting observer [25].

Implementing an observer requires the model parameters, i.e. the dynamics and lift function, and a parametrization of the

underlying state. Since we assume that these components are unknown we approximate them using the manifold learning approach as described in Section III. In particular, the drift component describing system dynamics is shown to be linear in (9), and therefore a linear observer can be constructed.

A. Contracting Observer

Let $\widehat{\Psi}(t) \in \mathbb{R}^m$ denote the estimated state, which is built based on the following standard recursive observer equation

$$\dot{\widehat{\Psi}} = \Lambda\widehat{\Psi} + \boldsymbol{\kappa}(\boldsymbol{z} - \hat{\boldsymbol{z}}). \tag{22}$$

The components comprising the observer equation are as follows. First, $\Lambda \in \mathbb{R}^{m\times m}$ is a diagonal matrix with the m largest eigenvalues of the Fokker-Planck operator on its diagonal: $\Lambda_{\ell\ell} = -\lambda_\ell$, $\ell = 1, \ldots, m$, since, as discussed in Section III-A, they approximate the dynamics of the system. Second, $\hat{\boldsymbol{z}} = g(\hat{\Psi})$ is the measurement associated with the current value of the estimated state, where g is the lift function defined in (21). Finally, $\boldsymbol{\kappa} \in \mathbb{R}^{m\times n}$ is an "adjustable" gain.

Note that the constructed coordinates, $\hat{\Psi}$, defined by the observer equation, are neither orthogonal nor a basis, however, we assume that the extension of the function $g(\cdot)$ to $\hat{\Psi}$ approximates the lift function from these coordinates to the measurements. A related extension is presented in [22], which describes a scheme for the extension of functions, defined on a given set, to additional elements.

By substituting (21) into (22), the observer becomes :

$$\dot{\widehat{\Psi}} = \Lambda\widehat{\Psi} + \boldsymbol{\kappa}(\boldsymbol{z} - \boldsymbol{\alpha}\widehat{\Psi}) = (\Lambda - \boldsymbol{\kappa}\boldsymbol{\alpha})\widehat{\Psi} + \boldsymbol{\kappa}\boldsymbol{z} \tag{23}$$

To obtain a contracting observer, we set the Jacobian $\mathbf{J} = (\Lambda - \boldsymbol{\kappa}\boldsymbol{\alpha})$ of the system to be negative. Specifically, by setting

$$\boldsymbol{\kappa} = \gamma\Lambda\boldsymbol{\alpha}^\dagger \tag{24}$$

where $\boldsymbol{\alpha}^\dagger$ denotes pseudo-inverse, we have

$$\mathbf{J} = (\Lambda - \boldsymbol{\kappa}\boldsymbol{\alpha}) = (\Lambda - \gamma\Lambda\boldsymbol{\alpha}^\dagger\boldsymbol{\alpha}) = (1-\gamma)\Lambda \tag{25}$$

which guarantees contraction for any $\gamma < 1$, since Λ is negative. We remark that by setting this particular gain $\boldsymbol{\kappa}$ (24), we take advantage of the fact that in our particular setup arising from the manifold learning standpoint, and in contrast to the common practice in dynamical systems, the dimension m of the state (in our case, the inferred state $\hat{\Psi}$) is assumed to be significantly smaller than the dimension n of the measurement. This implies the existence of a left pseudo-inverse satisfying $\boldsymbol{\alpha}^\dagger\boldsymbol{\alpha} = \mathbf{I}$. By using such a constant value of $\boldsymbol{\kappa}$, the recursive equation of the observer (22) becomes

$$\dot{\widehat{\Psi}} = \Lambda\widehat{\Psi} + \gamma\Lambda\boldsymbol{\alpha}^\dagger(\boldsymbol{z} - \boldsymbol{\alpha}\widehat{\Psi}) = (1-\gamma)\Lambda\widehat{\Psi} + \gamma\Lambda\boldsymbol{\alpha}^\dagger\boldsymbol{z} \tag{26}$$

where the value of γ enables us to control the relative weighting between the revealed dynamics ($\Lambda\hat{\Psi}$ in the first term in the right-hand side of (26)), and the correspondence to the measurements $\boldsymbol{z}$ (second term in the right-hand side of (26)). We note that the above form of $\boldsymbol{\kappa}$ is not optimal and is chosen mainly for inversion of the lift ("measurement") function. The optimal choice of $\boldsymbol{\kappa}$ will be addressed in future work.

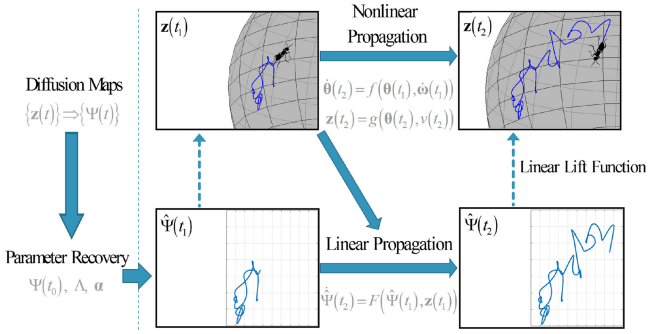

Fig. 1. A schematic view of the observer framework.

B. Relation to the Koopman Operator

Based on the coordinate dynamics described in Section III, our approach is related to Koopman spectral analysis [26]. Applying diffusion maps to the given measurements generates a parametrization of the state space (intrinsic embedded coordinates) which evolves according to known dynamics. These dynamics can be approximated by a linear operator Λ as described in Section (IV-A). In addition, in the presented linear case the mapping between the embedded coordinates and the measurements, denoted by the function g, is also linear.

Therefore, similarly to the Koopman operator, we create a linear operator by applying Λ to the observer's estimated state, describing the evolution of the constructed coordinates on the state space of the dynamical system. However, in contrast to common methods, which approximate the required quantities for the construction of the Koopman operator, in our approach no additional information is required and the constructed coordinates are purely data driven. For example, the extended dynamic mode decomposition (EDMD) [26] requires additional information in the form of dictionary elements.

Fig. 1 presents a schematic view of the observer framework. The observer is constructed based on the dynamics parameter Λ and lift function $\boldsymbol{\alpha}$, revealed by diffusion maps. Fig. 1 shows that similarly to the Koopman operator, the attained observer propagates according to a linear function (26), marked by $F(\cdot)$ in the figure, even for nonlinear systems in which the underlying process and measurements are governed by nonlinear functions (1), (2), marked by f and g in the figure.

C. Implementation in Discrete Setting

The derivations up to this point are presented for a continuous setting, however, in most cases only a finite set of discrete measurements is available. In this section we present a discrete setting in which the eigenvalues and eigenfunctions of the Fokker-Planck operator introduced in Section III can be approximated, and we describe the discrete version of the constructed observer.

Given a set of N measurements, $\{\boldsymbol{z}(t_i)\}_{i=0}^{N-1}$, where $t_i = i\Delta t$ and $\Delta t > 0$ is some time step, the discrete diffusion maps algorithm is given in Algorithm 1. Based on this algorithm we compute the eigenvalues $\mu_0, \ldots, \mu_{N-1}$ and eigenvectors $\psi_0, \ldots, \psi_{N-1}$ of the row-stochastic matrix $\boldsymbol{W}$,

Algorithm 1: Diffusion maps discrete setting.

1) Construct the affinity matrix, $\boldsymbol{K}$,
$$K\left(i,j\right)=\exp\left(-\frac{d\left(\boldsymbol{z}\left(t_i\right),\boldsymbol{z}\left(t_j\right)\right)}{\epsilon}\right)$$
where $d\left(\boldsymbol{z}\left(t_i\right),\boldsymbol{z}\left(t_j\right)\right)$ is the modified Mahalanobis distance in (15).
2) Create the row stochastic matrix:
$$W\left(i,j\right)=\frac{1}{D\left(i\right)}K\left(i,j\right)$$
where $D\left(i\right)=\sum_{j=0}^{N-1}K\left(i,j\right)$.
3) Compute the eigenvalues, $\mu_0,...,\mu_{N-1}$, and eigenvectors, $\psi_0,...,\psi_{N-1}$, of the matrix $\boldsymbol{W}$.

Algorithm 2: Discrete observer framework.

1) Construct the lift function based on the computed eigenvectors and given measurements
$$\alpha_{j,\ell}=\langle z_j,\psi_\ell\rangle=\sum_{i=0}^{N-1}z_j(t_i)\psi_\ell(t_i)\,,\ \ell=1,...,m$$
2) The discrete observer is given by
$$\begin{aligned}\widehat{\Psi}\left(t_{i+1}\right)&=\widehat{\Psi}\left(t_i\right)+\Lambda\widehat{\Psi}\left(t_i\right)+\boldsymbol{\kappa}(\boldsymbol{z}\left(t_i\right)-\hat{\boldsymbol{z}}\left(t_i\right))\\&=\left[\mathrm{I}+(1-\gamma)\Lambda\right]\widehat{\Psi}\left(t_i\right)+\gamma\Lambda\boldsymbol{\alpha}^{\dagger}\boldsymbol{z}\left(t_i\right)\end{aligned}\quad(28)$$
where $\Lambda\in\mathbb{R}^{m\times m}$ is a diagonal matrix with $\Lambda_{\ell\ell}=-\lambda_\ell$, I is the identity matrix and $\boldsymbol{\kappa}$ is chosen as in (26).

and order them such that $1=\mu_0\geq\mu_1\geq\cdots\geq\mu_{N-1}\geq 0$. Note that $\psi_0=[1\ 1\ \cdots\ 1]^T$ is the trivial eigenvector associated with $\mu_0=1$; the next few eigenvectors provide a coordinate system for the data, so that $\psi_\ell(i)$, the i-th entry of ψ_ℓ, provides the ℓ-th coordinate for $\boldsymbol{z}(t_i)$.

As in Section III-B we view the m eigenvectors associated with the m largest eigenvalues (except the trivial ψ_0) as a parametrization of the state of the system [7]. These m eigenvectors form empirical embedding coordinates of the data by the mapping:
$$\boldsymbol{z}(t_i)\mapsto[\psi_1(i),\psi_2(i),\ldots,\psi_m(i)]^T,\ i=0,\ldots,N-1\quad(27)$$
These eigenvectors are shown to be discrete approximations of the eigenfunctions of the Fokker-Planck operator $\mathcal{L}$ (7) in [6], [20], [27], [28]. In addition, the corresponding eigenvalues of the Fokker-Planck operator can be approximated by $-\lambda_\ell=\frac{2}{\beta\epsilon}\log\mu_\ell$, where μ_ℓ are the eigenvalues of the row-stochastic matrix $\boldsymbol{W}$ [29]. This adaptation is necessary since the accrued variance in each time step of (3) is $2/\beta$ and the matrix $\boldsymbol{W}$ represents a Markov chain with time steps of ϵ.

In [10], a comprehensive approach for the estimation of β is derived. However, note that in our setting $\beta=1$ due to the use of the Mahalanobis distance, as described in Appendix B and in [29].

Based on these eigenvalues and eigenvectors, the discrete observer is constructed as described in Algorithm 2.

For simplicity, the discrete observer (28) is presented with $\Delta t=t_{i+1}-t_i$. However, we note that in general, the time step can be altered from the time step of the given measurements, $\Delta t=t_{i+1}-t_i$, by multiplying $\Lambda\widehat{\Psi}\left(t_i\right)+\boldsymbol{\kappa}(\boldsymbol{z}\left(t_i\right)-\hat{\boldsymbol{z}}\left(t_i\right))$ with the desired $\Delta t^*<\Delta t$ and fixing the measurement at time t_i for all $t_i+n\Delta t^*<t_{i+1}$, $n\in\mathbf{N}$.

D. Diffusion Filtering

In the discrete setting, the interpretation in (26) can be extended further by using the approximation (21), this yields:
$$\widehat{\Psi}\left(t_{i+1}\right)-\widehat{\Psi}\left(t_i\right)=(1-\gamma)\Lambda\widehat{\Psi}\left(t_i\right)+\gamma\Lambda\Psi\left(t_i\right)\quad(29)$$
We now see that the evolution in time of the observer's estimated state is a weighted sum of Ψ and $\hat{\Psi}$ controlled by γ. Since the eigenvectors can be viewed as solutions of the Fokker-Planck equation, they describe the propagation in time of the density evolution due to diffusion. Accordingly, we can interpret the two terms in the recursive equation. Both terms $\Lambda\Psi$ and $\Lambda\hat{\Psi}$ can be viewed as an approximation of the propagation of the probability density of $\boldsymbol{\theta}\left(t\right)$ one step forward as presented in (9). On the one hand, $\Lambda\Psi$ is one step forward from Ψ (which was constructed merely from the data and is independent of time). On the other hand, $\Lambda\hat{\Psi}$ is one step forward from $\hat{\Psi}$ which is a "trajectory of evolving densities" from the initial point, and hence, it is time dependent.

To make the explanation above more clear, we write the differential equation (29) explicitly in the discrete form by recursively telescoping $\hat{\Psi}\left(t_i\right)$, which yields
$$\begin{aligned}\widehat{\Psi}\left(t_{n+1}\right)=&\left[\mathrm{I}+(1-\gamma)\Lambda\right]^{n+1}\Psi\left(t_0\right)+\\&+\gamma\Lambda\sum_{i=0}^{n}\left[\mathrm{I}+(1-\gamma)\Lambda\right]^{(n-i)}\Psi\left(t_i\right)\end{aligned}\quad(30)$$
where $\widehat{\Psi}\left(t_0\right)=\Psi(t_0)$.

The explicit form in (30) highlights the two terms comprising an effective filtering applied by the observer. The first term is the propagation of the density $n+1$ steps forward from the initial point. This term encapsulates merely the diffusion propagation of the densities as captured by diffusion maps (with γ weighting), such that it is enhanced by small values of γ. However, it does not explicitly take into account the samples along the given trajectory. For example, for $\gamma=0$:
$$\widehat{\Psi}(t_{n+1})=(\mathrm{I}+\Lambda)^{n+1}\Psi(t_0)\quad(31)$$
The second term is the propagation of a single step forward from each point along the trajectory. This term can be viewed as the correction of the general propagation of the densities, stemming from the particular realization at hand. The term is enhanced for large values of γ, for example, for $\gamma=1$:
$$\widehat{\Psi}(t_{n+1})=\Psi(t_0)+\Lambda\sum_{i=0}^{n}\Psi(t_n)\quad(32)$$

Algorithm 3: Out-of-sample extension.

1) Given an initial set of $N-1$ measurements, apply diffusion maps to reveal system dynamics Λ_{N-1} and reconstruction matrix $\boldsymbol{\alpha}_{N-1}$.
2) Construct the observer for the initial set of measurements.
3) Given a new measurement $\boldsymbol{z}(t_N)$ apply

$$\widehat{\Psi}(t_{N+1}) = [\mathrm{I} + (1-\gamma)\Lambda_{N-1}]\,\widehat{\Psi}(t_N) + + \gamma\Lambda_{N-1}\boldsymbol{\alpha}_N^{\dagger}\boldsymbol{z}(t_N) \quad (35)$$

We note that (30) takes the form of a filter: $\gamma\Lambda\Psi(t_n)$ convolved with the filter $[\mathrm{I}-(1-\gamma)\Lambda]^n$, which weighs past samples exponentially.

Moreover, if for simplicity we initiate the observer with zero and place $\Psi(t_i) = \boldsymbol{\alpha}^{\dagger}\boldsymbol{z}(t_i)$, the observer can be written as a data driven, moving average filter on the transformed measurements, $\boldsymbol{\alpha}^{\dagger}\boldsymbol{z}$, in which the window size is determined by the dynamics coefficient, Λ, and the parameter γ:

$$\widehat{\Psi}(t_{n+1}) = \gamma\Lambda\sum_{i=1}^{n}[\mathrm{I} + (1-\gamma)\Lambda]^{(n-i)}\,\boldsymbol{\alpha}^{\dagger}\boldsymbol{z}(t_i) \quad (33)$$

E. Observer Based Out-of-Sample Extension

One of the main shortcomings of addressing dynamical systems using manifold learning techniques concerns the handling of streaming data. When a new measurement $\boldsymbol{z}(t_i)$ is acquired, we wish to extend the learned coordinate system. This is commonly performed by the Nyström extension [30], which is an extension scheme for eigenvectors ψ_ℓ:

$$\psi_\ell(t_i) = \frac{1}{\lambda_\ell}\sum_{j=0}^{N-1} W(i,j)\,\psi_\ell(t_j) \quad (34)$$

where $\boldsymbol{W}$ is constructed as described in Algorithm 1, based on the affinity kernel between the new measurement $\boldsymbol{z}(t_i)$ and the measurements with known eigenvectors.

In our case, the new coordinate system constructed by diffusion maps consists of eigenvectors, and therefore, the Nyström extension can be applied. This extension allows extrapolation of the learned parametrization to new, unseen measurements. However, it is accurate only for measurements which are closely related to known data. In this section we present an out-of-sample extension scheme which naturally arises from the observer's structure.

As presented in Section IV-A, the observer is constructed based on the dynamics matrix Λ and the lift function $\boldsymbol{\alpha}$, both revealed by the diffusion maps framework. We assume that the dynamics and reconstruction function do not change significantly in time and therefore describe the system for new measurements as well. We present the proposed out-of-sample extension framework in Algorithm 3.

V. Experimental Results

A. Demonstrating the Estimation of the Underlying Process

We base our toy example on the setting presented in [7] which describes a radiating object moving on a 3D sphere. The process is described by its elevation and azimuth angles, as a 2D Langevin equation with a parabolic potential:

$$\dot{\theta}_1 = \left(\frac{\pi}{2}\cdot c - c\cdot\theta_1\right) + b\dot{\omega}_1 \quad (36)$$

$$\dot{\theta}_2 = \left(\frac{\pi}{10}\cdot c - c\cdot\theta_2\right) + b\dot{\omega}_2 \quad (37)$$

where b is the diffusion coefficient (set to 0.005) and c is the drift rate parameter which we vary between 0.08 and 0.024.

The resulting 3D process is given by:

$$x_1(t) = \cos(\theta_2)\sin(\theta_1)$$
$$x_2(t) = \sin(\theta_2)\sin(\theta_1)$$
$$x_3(t) = \cos(\theta_1)$$

We mark the 3D location of the object at time t by $\boldsymbol{x}(t) = [x_1(t), x_2(t), x_3(t)]$. The position of the object is measured by 3 sensors located at $\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3$ as presented in Fig. 4. Each sensor $\boldsymbol{s}_j$ is modeled as a "Geiger Counter" and fires spikes according to a Poisson distribution in a rate, r_j, which depends on the proximity of the object to the sensor: $r_j(t) = \exp(-\|\boldsymbol{s}_j - \boldsymbol{x}(t)\|)$, $j = 1,2,3$.

The output of each sensor is described by:

$$z_j(t) = y_j(t) + v_j(t)$$
$$y_j(t) \sim Pois(r_j(t)) \quad j = 1,2,3$$

where $v_j(t)$ is a spike train drawn from a Poisson distribution with a fixed rate parameter. Here, $y_j(t)$ represents the measurement modality of the intrinsic process and $v_j(t)$ represents an additive, independent measurement noise. The available measurements $z_j(t)$ are the sum of $y_j(t)$ and $v_j(t)$. Note that this Poisson setting presents a more challenging scenario compared with the typical Gaussian setting. In addition, common practice in many denoising procedures is to exploit the difference in the distributions of the signal and the noise. Thus, considering a scenario in which both have the same distribution poses an additional challenge.

Our goal in this example is to recover the underlying 2D dynamical process of the angles based solely on the measurements in this non-linear, non-Gaussian setting.

In this section the diffusion maps framework is implemented similarly to the one described in [7]. In this framework, histograms are constructed as feature vectors and are shown to estimate the probability density function of the clean observation process $y_i(t)$. The histograms are calculated based on non-overlapping time frames, each containing 60 time samples, $z_j(t_i : t_{i+59})$. We note that we have found empirically that different choices of histogram binning led to comparable results. After histogram construction, the Mahalanobis distance, presented in (15), is calculated based on these histograms. Mahalanobis distance is then used to construct the pairwise affinity kernel as described in (10). Diffusion maps coordinates and the

observer's estimated state are constructed based on the description in Section III-B and Section IV. To summarize, the analysis is performed as follows:

1) Given N measurements, construct histograms based on time frames of 60 samples, resulting in $\frac{N}{60}$ histograms.
2) Calculate the Mahalanobis distance based on these histograms and build the affinity matrix of size $\frac{N}{60} \times \frac{N}{60}$.
3) Continue as described in Algorithm 1 and Algorithm 2.

Note that the first step of Algorithm 1 requires knowledge of ϵ. Common practice is to set the scale ϵ to be of the order of the median of the pairwise distances $d\left(\boldsymbol{z}\left(t_i\right), \boldsymbol{z}\left(t_j\right)\right), \forall i, j$. Here, we set it to be the median itself, since empirically it was shown to attain good performance. We note that in all examples presented in this section, despite the typical great influence of ϵ on the obtained diffusion maps coordinates, our empirical examinations showed that the observer's estimated state consistently provides good results for a large range of ϵ values and is less sensitive to the choice of ϵ.

We generated 240,000 time samples of $\theta_i\left(t\right),\ i = 1, 2$, in time steps of $\Delta t = 0.1$ and constructed the measurements $z_j\left(t\right)$ based on the presented setting. As described above, this correspond to 4000 histograms of $z_j\left(t\right)$ and therefore to 4000 recovered coordinates (embedding coordinates). The state estimate of the observer was created as described in Section IV, with $\gamma = 0.85$. This choice of γ provided good empirical results. For a description of the considerations in the choice of γ, we refer the reader to Section IV-D. We applied the proposed observer and compared its performance to the diffusion maps embedding in recovering the underlying 2D dynamical process, $\theta_i\left(t\right)$. We emphasize that no information on the true underlying angles was used in the construction of the observer's estimated state and in the construction of the diffusion maps coordinates.

In order to evaluate our results more accurately, we initially perform linear regression on the first 4 diffusion maps coordinates. This is carried out since one of the shortcomings of the diffusion maps coordinates is that, while they can provide a good representation of the underlying state, they are not necessarily separable [31]. Therefore, the linear regression is performed merely for the purpose of comparing between representations of the underlying state by the observer and by the diffusion maps coordinates. We note that without this linear regression the representation errors are larger, however, the observer still improves the representation of the underlying state, compared with the diffusion maps coordinates.

In Fig. 2, a short interval of the normalized and centered vertical elevation angle process is displayed (blue plot), along with the first diffusion map coordinate (green plot) and the first coordinate of the observer's estimated state (red plot). Each point in the plots of the observer coordinate and diffusion maps coordinate represents one histogram of 60 time samples and therefore, the true angle, $\theta_1\left(t\right)$, is down-sampled. At the top of the figure, the absolute errors between the true elevation angle and both coordinates are displayed in gray. Note that we refer to the recovered state based on the observer equation, as the observer coordinates, in order to avoid confusion with the recovered state based on diffusion maps.

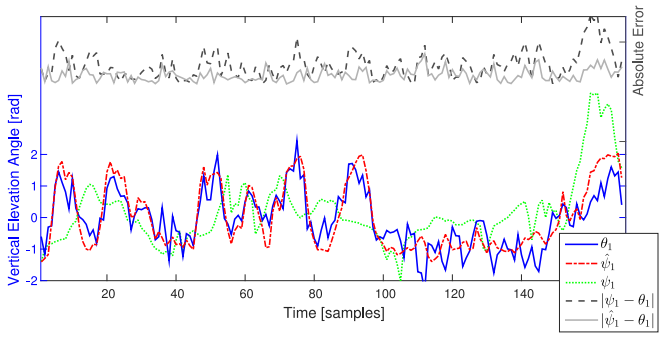

Fig. 2. Vertical elevation angle recovery. Comparison between the true vertical angle (blue), created with $c = 0.024$, the first coordinate of the constructed observer (red) and the diffusion maps embedding (green). The absolute errors of the observer coordinate (light gray) and the diffusion maps coordinate (dashed gray) are presented at the top of the plot.

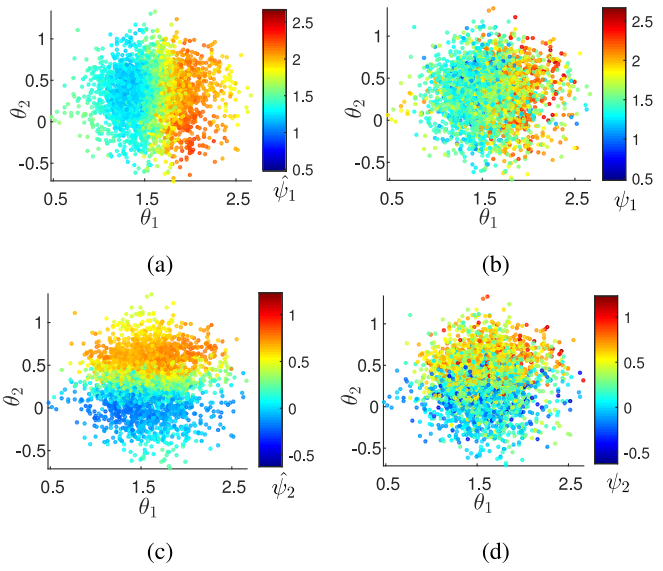

Fig. 3. Scatter plots of the vertical and horizontal angles created with $c = 0.024$. The plots are colored by values of the first and second observer coordinates (plots a, c) and the values of the first and second diffusion maps coordinates (plots b, d).

It is noticeable in Fig. 2 that both coordinates follow the general trend of the true angle, however the observer coordinate represents the true angle more accurately as can be seen in the coordinate plots and in the absolute error plot.

This improvement in correlation is emphasized in Fig. 3. The figure contains four identical scatter plots of the two underlying angles i.e., $\theta_2\left(t\right)$ as a function of $\theta_1\left(t\right)$, which differ only in their color schemes. Each of the four plots (denoted by (a), (b), (c), (d)) is colored based on either the first and second observer coordinates (plots (a) and (c) respectively) or the first and second diffusion maps coordinates (plots (b) and (d) respectively). The color schemes of these scatter plots depict that both underlying angles are represented better by the first and second observer coordinates, since the color gradient is smoother in plots (a) and (c), than the color gradient in plots (b) and (d). Therefore, as expected in the presented toy example which suffers from noise, both additive and model based (due to the stochastic nature of

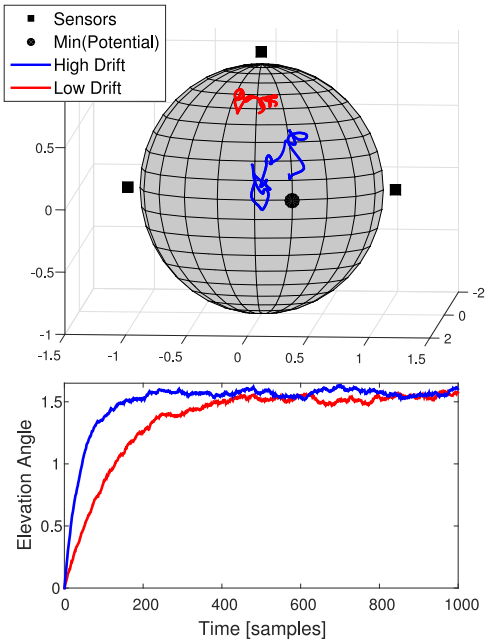

Fig. 4. Toy example setup with different drift rates c. (top) Two 300 point segments of the 3D movement on the sphere with different drift rates. (bottom) Two realizations of the elevation angle over 1000 time samples with different drift rates: $c = 0.008$ (red) and $c = 0.024$ (blue).

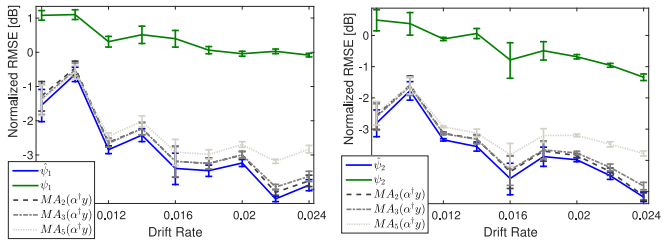

Fig. 5. Normalized RMSE [dB] between the true angles and their estimations in different drift rates. The angles are represented by the diffusion maps embedding (green), observer coordinates (blue) and moving average filters on $\boldsymbol{\alpha}^\dagger \boldsymbol{z}$ with window sizes of 2 (dashed gray), 3 (dot-dashed gray) and 5 (dotted gray) samples.

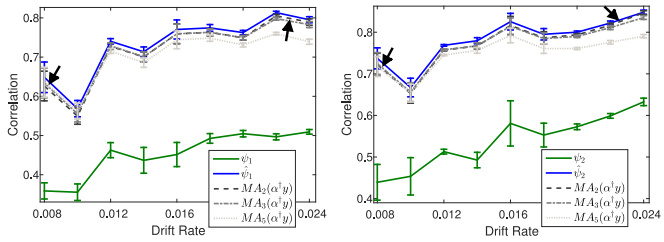

Fig. 6. Correlation between the true angles and their estimations in different drift rates. The angles are represented by the diffusion maps embedding (green), observer coordinates (blue) and moving average filters on $\boldsymbol{\alpha}^\dagger \boldsymbol{z}$ with window sizes of 2 (dashed gray), 3 (dot-dashed gray) and 5 (dotted gray).

the sensors), the observer describes the underlying dynamical process more accurately.

We note that although the observer provides better representations of the underlying state than diffusion maps, it suffers from inaccuracies at the boundaries of the data. This is visible in Fig. 3, for example in plot (a) when $\theta_1 < 1$. This is due to the inaccuracy of the linear lift function at the boundaries and will be addressed in future work.

As described in (33), the observer is a data-driven filter applied to $\boldsymbol{\alpha}^\dagger \boldsymbol{z}$ with varying filter length which is determined by Λ and γ. Therefore, we compared the result of the observer (with a constant parameter γ) to several moving average filters with varying window sizes applied to $\boldsymbol{\alpha}^\dagger \boldsymbol{z}$, in different parabolic potentials (with different parameter c values). Our motivation for varying the drift rate arises from Fig. 4 which displays the effect of different parabolic potentials on the resulting path. In this figure, the top plot displays two exemplary paths on the 3D sphere, one with a high drift rate $c = 0.024$ (blue plot) and one with a low drift rate $c = 0.008$ (red plot). The sensor locations are marked with black rectangles and the location of the minimum of the parabolic potential is denoted by a black circle. The bottom plot displays the first coordinate of the 2D underlying state (elevation angle) for both drift rates. These plots depict the differences in convergence rate and in the step size, mainly before convergence. The high drift rate process converges faster, causing higher step sizes until convergence and its perturbations from the minimum potential are smaller.

Fig. 5 shows the normalized RMSE, between the true angles and their estimates, where the plot on the left presents the the first coordinate (elevation angle) and the plot on the right presents the second coordinate (azimuth angle). These plots contain average values of the normalized RMSE, over 50 iterations, with error bars of one standard deviation, for varying drift rates between $c = 0.008$ and $c = 0.024$. We compared estimates based on the observer (blue line), diffusion maps embedding (green line) and moving average filters on $\boldsymbol{\alpha}^\dagger \boldsymbol{z}$ with 3 windows of lengths 2, 3, and 5 (dashed gray lines). Similarly, Fig. 6 presents the average correlation between the true angles and their estimates, over 50 realizations, with error bars of one standard deviation.

It is clear from these plots that the observer's estimated state and moving average filters outperform the diffusion maps embedding, as their correlation to the true angles is significantly higher. In addition, it is worth noting that different averaging filters perform best in different settings. This is due to the change in path characteristics as the drift rate parameter c varies, e.g. slow drift rates (wide parabolic potential) correspond to slower convergence rates and smaller step sizes, and therefore, a wider averaging filter performs best. Examples of this effect are marked by black arrows in Fig. 6, where at a low drift rate ($c = 0.008$) the moving average filter with window size of 5 samples performs best (out of the 3 moving average filters) and at a high drift rate, ($c = 0.024$), the shortest moving average filter performs best. In all instances, the proposed observer closely follows the best filter in each setting, demonstrating its superiority as a data driven filter in which the window size is determined based on the given measurements without prior information, particularly without information on the drift.

Finally, we examine the extension scheme presented in Subsection IV-E. We simulated a trajectory of length 240,000, corresponding to 4000 histograms. We constructed the diffusion maps coordinates and recovered the required observer parameters, i.e. dynamics and reconstruction function, based on 3000

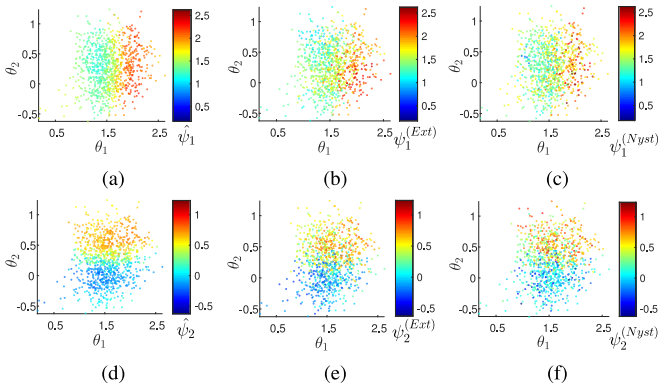

Fig. 7. Extension scheme: Scatter plots of the vertical and horizontal angles created with $c = 0.024$, colored by the extended coordinates. The plots are colored by values of the first and second observer extension coordinates (plots a,d), the values of the first and second diffusion maps extension coordinates in [7] (plots b,e) and the first and second coordinates of the Nyström extension (plots c,f).

of these histograms as described above. We then applied the observer extension scheme, as described in Algorithm 3, to the remaining 1000 histograms, in order to estimate the 2D underlying process at these time points. We compared our results to the Nyström extension (34) as a baseline and to the extension scheme proposed in [7]. Fig. 7 presents six identical scatter plots containing these remaining 1000 time frames (histograms) of the two angles i.e., $\theta_2\,(t)$ as a function of $\theta_1\,(t)$, which differ only in their color schemes, similarly to Fig. 3. In this figure, plots (a) and (d) are colored by the first and second observer's estimated state coordinates respectively, plots (b)and (e) are colored by the first and second coordinates of the extension scheme in [7] and plots (c) and (f) are colored by the first and second coordinates of the Nyström extension. It is visible that the observer-based extension coordinates are superior in describing the elevation and azimuth angles, since increasing angle values are represented by increasing observer coordinate values as seen by the distinct coloring in plots (a) and (d). In addition, the color gradient is smoother in these plots, which shows higher correlation between the observer coordinates and the underlying angles.

B. Demonstrating the Dynamics Estimation

In this subsection, we test the accuracy of the dynamics parameter estimation Λ from (22). Here, the system is based on a toy example similar to the one presented in Section V-A with the following modification. Instead of a radiating object moving on a 3D sphere, we examine the above scheme in 2D, which depicts a radiating object moving on a 2D circle, where the underlying 1D process is described by an azimuth angle. We use the following dynamics equation, which has known eigenvalues, to generate the angle:

$$\dot{\theta} = -\theta + \sqrt{2}\dot{w} \tag{38}$$

and the 2D process is given by:

$$\begin{aligned} x_1\,(t) &= \cos\,(\theta) \\ x_2\,(t) &= \sin\,(\theta) \end{aligned}$$

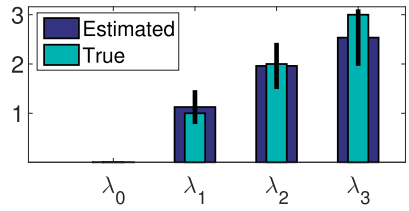

Fig. 8. Estimation of eigenvalues $-\lambda_\ell$ in the 2D framework, averaged over 50 realizations.

The eigenfunctions of the backward Fokker-Plank operator of this process satisfy the probabilists Hermite equation [32] which has solutions with eigenvalues $\lambda_\ell = -\ell$, $\ell \in \{0, 1, 2, ...\}$. The known eigenvalues provide ground truth, and as a result, we can verify our estimation of the dynamics $\Lambda_{\ell\ell} = -\lambda_\ell$, which are determined by these eigenvalues. Here, the scale ϵ was set to $\epsilon = 0.16\zeta$ where ζ denotes the median of the pairwise distances. This value of ϵ was chosen since empirically it was shown to attain best performance.

These modifications were performed since in the 1D setting more complex stochastic differential equations with known solutions are available. In addition, the effects of different components in the proposed framework on the quality of the estimation are more noticeable in this simpler case.

Since the accuracy of the estimation is highly dependent on the correct calculation of the covariance matrices in the modified Mahalanobis distance (15), *in this subsection only* we simulated short bursts at each time point for the covariance calculations. This is performed in order to acquire a sufficient number of samples from the same distribution at each small neighborhood and achieve a sufficiently small error due to finite sampling. We emphasize that here, the data are realizations of the SDE in (38) and only the covariance matrices for the modified Mahalanobis distance are calculated based on the simulated short bursts. Our goal in this example is to present the contribution of the proposed state-space framework isolated from the problem of covariance estimation and therefore we assume that we have the true local covariances at hand. To obtain local covariance matrices as accurate as possible, we simulate the short bursts to estimate a covariance at each time sample. When such bursts are not available, the basic assumption is that points in each small neighborhood are drawn from approximately the same distribution. We note that when applying our method to this example without access to the short bursts, the estimations of λ_1 and λ_2 are comparable, however, higher eigenvalues contain larger estimation errors. An analysis of the accuracy of the covariance estimation extends the scope of this paper and is addressed in [33].

Fig. 8 presents the estimation of the eigenvalues $-\lambda_\ell$ based on 1600 time frames (histograms), each containing 60 time points which were used to construct one histogram. The plot displays the ground truth values (green) and the average estimates of $-\lambda_\ell$ (blue), over 50 realizations, for the four smallest eigenvalues (in absolute value), along with error-bars of one standard deviation (black lines). It shows that the average dynamics estimation is close to the true value, however, the estimation

variance increases for larger eigenvalues (in absolute value). Empirical analysis revealed that the choice of ϵ in (10) has a significant impact on the accuracy of λ_ℓ estimates. Furthermore, additive noise might also increase the estimation error, however it mostly affects larger eigenvalues ($j = 3, 4, ...$). Note that this behaviour depicts one of the advantages of the presented approach. Roughly, small eigenvalues represent signal components with distinct "structures", whereas large eigenvalues tend to represent noise.

Other methods for estimating the deterministic dynamics, Λ, exist. For example in [34] and [35], a method for learning PDEs describing the evolution of the input data to a desired output is presented. In these papers, the coefficients of a second order PDE are estimated by solving an optimization problem based on a training set. However, this method requires access to the true system state for the construction of the training set, and, in addition, it contains assumptions regarding the general form of the estimated PDE. In contrast, the method we present is completely data-driven and does not require prior knowledge of the true state.

C. Music Analysis

In this subsection we show that the proposed observer, when applied to music, reveals meaningful underlying processes describing different characteristics of the data such as its dominant musical notes and a distinction between intervals with different musical instruments. This is achieved without explicit pitch tracking or modeling. For this purpose we applied Algorithm 1 and Algorithm 2 (with $\gamma = 0.1$) to two songs: (i) the theme song of "Once upon a time in the west" performed by Yo-Yo Ma (Yo-Yo Ma plays Ennio Morricone) and (ii) the theme song of "The good, the bad and the ugly" by Ennio Morricone. Both songs are sampled at 44.1 KHz and were analyzed in 15-25 second segments. In each segment we first performed short time Fourier transform (STFT) on 23 millisecond time frames and constructed a kernel based on the modified Mahalanobis distance (15) between the resulting spectrograms (absolute value of the STFT). Note that due to the use of STFT, this application of music analysis contains high-dimensional data and therefore, will be harder to address using standard processing techniques. We then applied the diffusion maps framework as described in Section III-B and constructed the observer based on the learned dynamics and linear lift function (26). In both songs we examined the first three coordinates of the observer's estimated state.

Fig. 9 shows a 25 second segment of the first song (Yo-Yo Ma plays Ennio Morricone), which contains a long segment of a single instrument, playing distinctive notes (with a quiet background melody). The top plot of the figure (plot (a)) displays the musical notes of the examined segment. In plot (b), the spectrogram of the music segment is presented with marked musical notes, along with the corresponding waveform of the music signal (below the spectrogram). Different colors on the spectrogram represent different musical notes and correspond to the coloring of plots (c) and (d). Plot (c) displays a 3D scatter plot of the first three observer coordinates. Each point in this plot

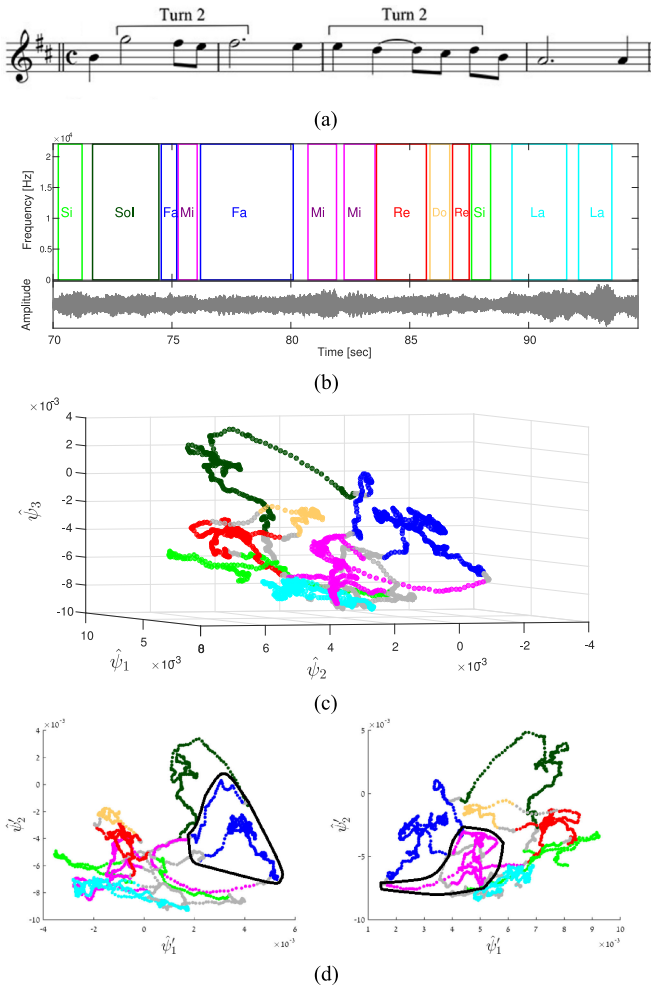

Fig. 9. Musical note identification in "Once upon a time in the west" theme. (a) Actual musical notes at 1:10 to 1:40 minutes. (b) Spectrogram of the music segment with marked musical notes. (c) 3D scatter plot of the first three observer coordinates. (d) Two different views of the 3D scatter plot.

represents one time sample of the spectrogram and is colored according to the marked notes. Gray points represent transitions between notes which are unmarked on the spectrogram. The two bottom plots (d) contain different 2D views of the 3D scatter plot, where $\hat{\psi}'_1, \hat{\psi}'_2$ represent rotated axes for better view angle.

In this figure, the 3D scatter plot depicts that the first three coordinates of the observer's estimated state create an embedding which separates different musical notes to different locations in the 3D space. This is expressed as differently colored points, representing different notes, which appear in distinct locations in the 3D space. In addition, the two bottom scatter plots illustrate that identical musical notes, which appear at different times, are represented by the same color and are grouped together in the new coordinate system created by the observer. In these bottom plots, black polygons are marked to emphasize two such examples.

Another example for musical notes identification is shown in Fig. 10, which presents an 11 second segment of the second song (6.2-17.2 seconds). This song segment contains

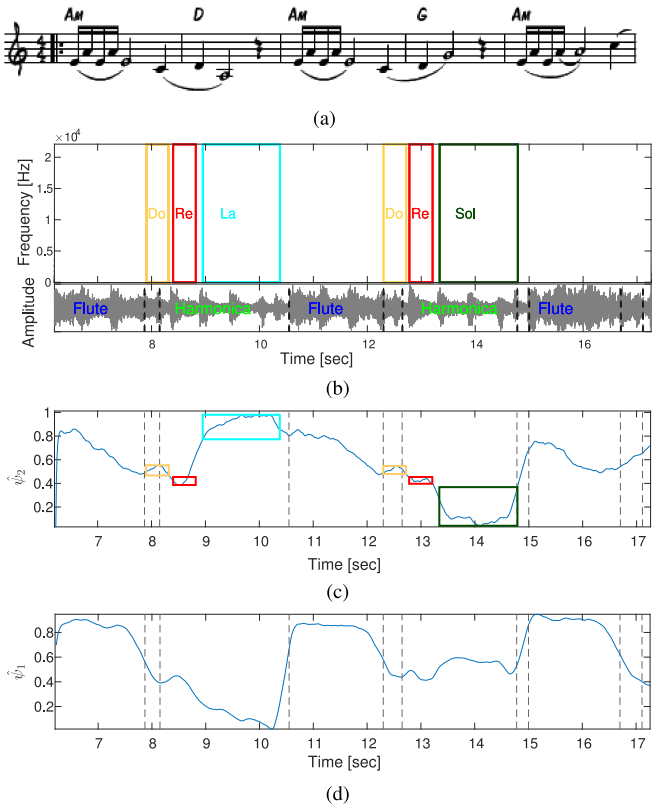

Fig. 10. Musical note and instrument identification in "The good, the bad and the ugly" theme. (a) Actual musical notes at 6.2 to 17.2 seconds. (b) Spectrogram and waveform with marked musical notes and instruments. (c) The second observer coordinate with marked note locations. (d) The first observer coordinate with locations of instrument transitions marked by adjacent dashed lines.

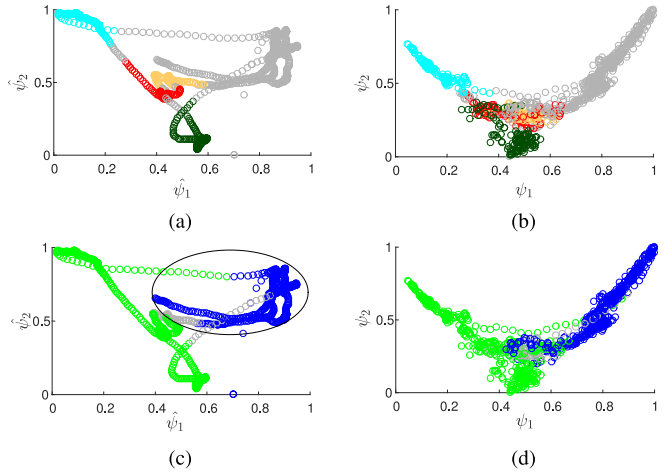

Fig. 11. Comparison of observer coordinates and diffusion maps coordinates for "The good, the bad and the ugly" theme. (a, b) Scatter plot of the observer and diffusion maps coordinates respectively, colored by musical notes. (c, d) Scatter plot of the observer and diffusion maps coordinates respectively, colored by instruments: flute (blue) and harmonica (green).

frequent transitions between different instruments (a flute and an harmonica). The top plot (a) in this figure contains the musical notes of the examined segment. In plot (b) the spectrogram of the music segment is presented with marked musical notes in the harmonica sections, along with the waveform below. On the waveform plot, the different instruments and transitions between them are marked by dashed lines i.e., intervals where both instruments are playing. The different colors on the spectrogram represent different musical notes and correspond to the coloring of the plot (c). Plot (c) presents the second observer coordinate as a function of time with marked notes and marked transitions between the instruments. The bottom plot (d) displays the first observer coordinate as a function of time with marked instrument transitions.

The two bottom plots of Fig. 10 illustrate that based on the observer, a good classification of both musical notes and instruments can be attained. For example, in plot (c), the values and shape of the curves in the harmonica intervals are consistent for similar notes at different times, e.g. Do at 7.8 seconds and at 12.1 seconds. In addition, different notes are represented by different values, e.g. La at 8.9 seconds and Sol at 13.3 seconds. We note that, in the flute intervals, such a distinction is not possible since they contain rapid note changes. This leads to different time scales in the data which are not represented properly by the embedded coordinates (both observer coordinates and diffusion maps coordinates). The problem can be alleviated by a different choice of parameters, ϵ and Δt, in the diffusion maps and observer frameworks. In the example presented in Fig. 10, similarly to the results shown in Fig. 9, the second coordinate of the observer captures the musical notes. However, since the second song includes another factor (different instruments), the first coordinate describes the instruments in this case. An example for such instrument distinction can be seen in the plot (d), where a simple threshold value can be used to separate between the flute intervals and the harmonica intervals.

In Fig. 11 we compare the coordinate systems attained by the observer and by diffusion maps for the song segment of "The good, the bad and the ugly" theme (6.2-17.2 seconds), presented in Fig. 10. Fig. 11 contains two identical scatter plots of the observer coordinates in (a) and (c), $\hat{\psi}_2$ as a function of $\hat{\psi}_1$, and two identical scatter plots of the diffusion maps coordinates in (b) and (d), ψ_2 as a function of ψ_1. Each point in the plots corresponds to one time sample of the spectrogram depicted in Fig. 10. The presented scatter plots differ in their color schemes. Plots (a) and (b) are colored according to the musical notes, which are marked on the spectrogram in Fig. 10. In addition, here, gray points represent unmarked time samples which are either flute intervals or transitions between notes in the harmonica intervals. Plots (c) and (d) are colored according to the true instrument segmentation which is presented on the audio waveform in Fig. 10. This color scheme marks flute intervals in blue, harmonica intervals in green and transitions between them, in which both instruments are playing, in gray. In the displayed plots, the diffusion maps coordinates and the observer coordinates were scaled for comparison only. Fig. 11 depicts that the coordinate system attained by the observer represents the data more accurately than the one attained by diffusion maps. This is visible in plots (a) and (b) where the coloring of the notes in plot (a) (observer coordinates) implies a better

TABLE I
MUSICAL NOTES CLASSIFICATION SUCCESS RATES OF THE DIFFERENT COORDINATE SYSTEMS CONSTRUCTED BASED ON A SEGMENT OF "THE GOOD, THE BAD AND THE UGLY" THEME (6.2-45 SECONDS)

$\hat{\psi}$	$\hat{\psi}_{EUC}$	ψ	ψ_{EUC}	ψ_{PCA}
0.97 ± 0.17	0.9 ± 0.29	0.76 ± 0.43	0.61 ± 0.49	0.69 ± 0.46

distinction compared to plot (b). Furthermore, the two separate segments containing Do and Re in the spectrogram in Fig. 10 are correctly grouped together in plot (a). In addition, plots (c) and (d) depict that the observer coordinates perform better in the classification of different music instruments as the blue and green colored points are highly separable.

Finally, in order to quantitatively assess our method and present the benefits of its various steps, we perform musical notes classification using K-Nearest Neighbors (K-NN) based on several coordinate systems. In this context, the construction of the different coordinate systems is essentially feature extraction, used as an input for the K-NN classifier. In Table I, we present results of a leave-one-out cross validation test using K-NN, with $K = 5$, applied to the coordinate systems attained by the observer with and without the use of the modified Mahalanobis distance (15), namely $\hat{\psi}$ and $\hat{\psi}_{EUC}$, respectively, by the diffusion maps coordinates with and without Mahalanobis distance, namely ψ and ψ_{EUC}, respectively, as well as by the coordinate system attained by applying Principal Component Analysis (PCA) directly to the STFT features, denoted by ψ_{PCA}. These coordinate systems were constructed based on a segment of "The good, the bad and the ugly" theme (6.2-45 seconds). Table I presents the mean and standard deviation of the classification success rates using the different coordinate systems. The classification results depict both the benefit of using the observer compared with diffusion maps and PCA, and the benefit of using the modified Mahalanobis distance.

VI. CONCLUSION

In this work we presented a purely data driven framework, in which an intrinsic representation is derived for highly nonlinear noisy systems. The framework consists of a combination between a manifold learning technique, which is used to reveal the hidden parameters of the system, and a linear observer, that incorporates dynamics and attains the intrinsic representation. We showed that our method indeed reveals intrinsic hidden features in data. In particular, when applied to music, it discovers both a segmentation to different musical notes as well as an indication of various musical instruments.

In future work we plan to address two main aspects. First, we plan to implement a Kalman filtering scheme that will provide an optimal adaptive choice of the gain parameter κ. Such a scheme will have many advantages, for example, it will circumvent the need to empirically tune the weighting parameter γ. The second issue we will address concerns the dynamics estimation. As presented in the experimental results section, the estimation of the dynamics based on diffusion maps might be influenced by measurement noise. In addition, in our work, we focused on the linear component of the diffusion maps dynamics and compensated for the stochastic component as part of the measurement correspondence element. We plan to extend our observer framework by adding adaptive dynamics which will account for these errors.

ACKNOWLEDGMENT

The authors would like to thank Prof. R. Coifman for fruitful discussions.

APPENDIX A
EIGENFUNCTION DYNAMICS

In this appendix we show the derivation of the stochastic differential equation for the eigenfunctions of the backward Fokker-Planck operator presented in Section III-A. Recall the state equation presented in Section II:

$$\dot{\boldsymbol{\theta}}(t) = -\nabla U\left(\boldsymbol{\theta}\left(t\right)\right) + \sqrt{\frac{2}{\beta}}\dot{\boldsymbol{\omega}}\left(t\right) \tag{39}$$

Since the eigenfunctions of the Fokker-Planck operator which describes this system are smooth functions of the underlying state, based on Itô calculus they also evolve according to a stochastic differential equation of the form

$$\dot{\psi}_\ell\left(\boldsymbol{\theta}\left(t\right)\right) = \mu_\ell\left(\boldsymbol{\theta}\left(t\right)\right) + \sigma_\ell\left(\boldsymbol{\theta}\left(t\right)\right)\dot{\omega}_\ell\left(t\right) \tag{40}$$

where $\mu_\ell\left(\boldsymbol{\theta}\left(t\right)\right)$, $\sigma_\ell\left(\boldsymbol{\theta}\left(t\right)\right)$ are the drift and diffusion coefficients respectively and $\omega_\ell\left(t\right)$ is Brownian motion. According to Itô's lemma, assuming that the Brownian motions of different coordinates of $\boldsymbol{\theta}\left(t\right)$ are independent, the drift and diffusion parameters are given by

$$\mu_\ell\left(\boldsymbol{\theta}\left(t\right)\right) = \frac{\partial\psi_\ell}{\partial t} + \frac{1}{\beta}\triangle_\theta\psi_\ell - \nabla_\theta U\cdot\nabla_\theta\psi_\ell \tag{41}$$

$$\sigma_\ell\left(\boldsymbol{\theta}\left(t\right)\right) = \sqrt{\frac{2}{\beta}}\left\|\nabla_\theta\psi_\ell\right\| \tag{42}$$

The partial derivative of ψ_ℓ in time is zero, since the eigenfunctions depend only on the state parameter and describe the system in steady state. Therefore, we are left with the two terms in (41) which are exactly the left-hand side of the backward Fokker-Planck operator (7). Since ψ_ℓ are the eigenfunctions of this operator with corresponding eigenvalues $-\lambda_\ell$, the resulting drift is described by $\mu_\ell\left(\boldsymbol{\theta}\left(t\right)\right) = -\lambda_\ell\psi_\ell\left(\boldsymbol{\theta}\left(t\right)\right)$.

Therefore, the stochastic differential equation, describing the dynamics of the eigenfunctions is

$$\dot{\psi}_\ell\left(\boldsymbol{\theta}\left(t\right)\right) = -\lambda_\ell\psi_\ell\left(\boldsymbol{\theta}\left(t\right)\right) + \sqrt{\frac{2}{\beta}}\left\|\nabla_\theta\psi_\ell\left(\boldsymbol{\theta}\left(t\right)\right)\right\|\dot{\omega}_\ell\left(t\right) \tag{43}$$

APPENDIX B
MAHALANOBIS DISTANCE

We present the derivation of the modified Mahalanobis distance [29] described in Section III-C, which approximates the Euclidean distances of $\boldsymbol{\theta}\left(t\right)$ from the measurements $\boldsymbol{z}\left(t\right)$.

For the noiseless case, consider the Taylor expansion at τ of the measurement function (4)

$$\begin{aligned} \boldsymbol{z}(t) &= h(\boldsymbol{\theta}(\tau)) + J_h(\boldsymbol{\theta}(\tau))(\boldsymbol{\theta}(t) - \boldsymbol{\theta}(\tau)) \\ &\quad + O\left(\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(\tau)\|^2\right) \end{aligned} \tag{44}$$

where $J_h(\boldsymbol{\theta}(\tau))$ is the Jacobian matrix of $h(\boldsymbol{\theta}(\tau))$. Applying the squared Euclidean norm and taking the inverse of the Jacobian we get

$$\begin{aligned} \|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(\tau)\|^2 &= \left\|J_h^{-1}(\boldsymbol{z}(\tau))(\boldsymbol{z}(t) - \boldsymbol{z}(\tau))\right\|^2 \\ &\quad + O\left(\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(\tau)\|^4\right) \end{aligned} \tag{45}$$

Similarly, we can derive the Taylor expansion at t and average the two resulting terms. This can be written as:

$$\begin{aligned} \|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(\tau)\|^2 &= \frac{1}{2}(\boldsymbol{z}(t) - \boldsymbol{z}(\tau)) \cdot M(t,\tau) \cdot (\boldsymbol{z}(t) - \boldsymbol{z}(\tau))^T \\ &\quad + O\left(\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(\tau)\|^4\right) \end{aligned} \tag{46}$$

where $M(t,\tau) = \left(J_h J_h^T\right)^{-1}(\boldsymbol{z}(t)) + \left(J_h J_h^T\right)^{-1}(\boldsymbol{z}(\tau))$.

Lastly, to obtain the modified Mahalanobis distance in [29], we show that $\left(J_h J_h^T\right)(\boldsymbol{z}(t))$ is in fact the covariance matrix of the measurements at time t, $C(\boldsymbol{z}(t))$.

Since the state of the system $\boldsymbol{\theta}(t)$ satisfies the stochastic differential equation in (3) and $\boldsymbol{z}(t) = h(\boldsymbol{\theta}(t))$, based on Itô's lemma the measurements satisfy

$$dz_j(t) = \sum_{i=1}^{d}\left(\frac{1}{2}(b_i)^2\frac{\partial^2 h_j}{\partial\theta_i^2} + a_i\frac{\partial h_j}{\partial\theta_i}\right)dt + \sum_{i=1}^{d} b_i\frac{\partial h_j}{\partial\theta_i}d\omega_i$$

where $j = 1, ..., n$, a_i and b_i are the drift function and diffusion coefficient of coordinate $\theta_i(t)$, as presented in (3) and ω_i is Brownian motion. Assuming that the Brownian motions of different coordinates $\theta_i(t)$ are independent, the covariance matrix of the measurements is given by

$$C_{jk}(\boldsymbol{z}(t)) = \sum_{i=1}^{d}(b_i)^2\frac{\partial h_j}{\partial\theta_i}\frac{\partial h_k}{\partial\theta_i}, \quad j,k = 1, ..., n \tag{47}$$

Note that the diffusion coefficients in our setting are constant, therefore, we can first apply a scaling transformation to eliminate b_i as described in [29]. Finally, after this scaling, the covariance can be written using the Jacobian matrix J_h:

$$C(\boldsymbol{z}(t)) = J_h J_h^T(\boldsymbol{z}(t)) \tag{48}$$

By placing (48) in (46) we attain the desired form of the modified Mahalanobis distance.

REFERENCES

[1] R. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME J. Basic Eng.*, vol. 82, pp. 34–45, 1960.

[2] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statist. Comput.*, vol. 10, no. 3, pp. 197–208, 2000.

[3] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 260, pp. 2319–2323, 2000.

[4] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, pp. 1373–1396, 2003.

[5] D. L. Donoho and C. Grimes, "Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data," *Proc. Nat. Acad. Sci.*, vol. 100, pp. 5591–5596, 2003.

[6] R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, pp. 5–30, Jul. 2006.

[7] R. Talmon and R. Coifman, "Empirical intrinsic geometry for nonlinear modeling and time series filtering," *Proc. Nat. Acad. Sci.*, vol. 110, no. 31, pp. 12 535–12 540, 2013.

[8] T. Berry and J. Harlim, "Semiparametric forecasting and filtering: correcting low-dimensional model error in parametric models," *J. Comput. Phys.*, vol. 308, pp. 305–321, 2016.

[9] T. Berry, D. Giannakis, and J. Harlim, "Nonparametric forecasting of low-dimensional dynamical systems," *Phys. Rev. E*, vol. 91, no. 3, 2015, Art. no. 032915.

[10] T. Berry and J. Harlim, "Nonparametric uncertainty quantification for stochastic gradient flows," *SIAM/ASA J. Uncertainty Quantification*, vol. 3, no. 1, pp. 484–508, 2015.

[11] T. Berry and J. Harlim, "Forecasting turbulent modes with nonparametric diffusion models: Learning from noisy data," *Phys. D: Nonlinear Phenom.*, vol. 320, pp. 57–76, 2016.

[12] R. Talmon and R. R. Coifman, "Intrinsic modeling of stochastic dynamical systems using empirical geometry," *Appl. Comput. Harmon. Anal.*, vol. 39, no. 1, pp. 138–160, 2015.

[13] B. Nadler, S. Lafon, R. Coifman, and I. G. Kevrekidis, "Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators," in *Proc. Neural Inf. Process. Syst.*, 2005, vol. 18, pp. 955–962.

[14] R. Coifman *et al.*, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proc. Nat. Acad. Sci.*, vol. 102, no. 21, pp. 7426–7431, May 2005.

[15] R. Coifman, I. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, "Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems," *Multiscale Model. Simul.*, vol. 7, no. 2, pp. 842–864, 2008.

[16] W. T. Coffey and Y. P. Kalmykov, *The Langevin Equation: With Applications to Stochastic Problems In Physics, Chemistry and Electrical Engineering*, vol. 27. Singapore: World Scientific, 2012.

[17] R. Talmon, S. Mallat, H. Zaveri, and R. Coifman, "Manifold learning for latent variable inference in dynamical systems," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 3843–3856, Aug. 2015.

[18] R. Talmon, I. Cohen, S. Gannot, and R. R. Coifman, "Diffusion maps for signal processing: A deeper look at manifold-learning techniques based on kernels and graphs," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 75–86, Jul. 2013.

[19] R. Talmon, D. Kushnir, R. R. Coifman, I. Cohen, and S. Gannot, "Parametrization of linear systems using diffusion kernels," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1159–1173, Mar. 2012.

[20] B. Nadler, S. Lafon, R. Coifman, and I. G. Kevrekidis, "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems," *Appl. Comput. Harmon. Anal.*, vol. 21, pp. 113–127, 2006.

[21] A. Singer and R. Coifman, "Non-linear independent component analysis with diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 25, pp. 226–239, 2008.

[22] R. Coifman and S. Lafon, "Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions," *Appl. Comput. Harmon. Anal.*, vol. 21, pp. 31–52, Jul. 2006.

[23] M. Hein and J. Audibert, "Intrinsic dimensionality estimation of submanifolds in R^d," in *Proc. 22nd Int. Conf. Mach. Learning*, 2005, pp. 289–296.

[24] R. Coifman, Y. Shkolnisky, F. Sigworth, and A. Singer, "Graph laplacian tomography from unknown random projections," *IEEE Trans. Image Process.*, vol. 17, no. 10, pp. 1891–1899, Oct. 2008.

[25] W. Lohmiller and J. Slotine, "On contraction analysis for non-linear systems," *Automatica*, vol. 34, no. 6, pp. 683–696, 1998.

[26] 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.*, vol. 25, no. 6, pp. 1307–1346, Dec. 2015.

[27] M. Belkin and P. Niyogi, "Convergence of Laplacian eigenmaps," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, vol. 19, pp. 129–136.

[28] A. Singer, "From graph to manifold Laplacian: The convergence rate," *Appl. Comput. Harmon. Anal.*, vol. 21, pp. 135–144, 2006.

[29] A. Singer and R. Coifman, "Non-linear independent component analysis with diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 25, pp. 226–239, 2007.

[30] Y. Bengio, J. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, "Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, vol. 16, pp. 177–184.
[31] C. J. Dsilva, R. Talmon, R. R. Coifman, and I. J. Kevrekidis, "Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study," *Appl. Comput. Harmon. Anal.*, 2015, to be published.
[32] H. Risken, *Fokker-Planck Equation*. New York, NY, USA: Springer-Verlag, 1984.
[33] C. J. Dsilva, R. Talmon, C. W. Gear, R. R. Coifman, and I. G. Kevrekidis, "Data-driven reduction for a class of multiscale fast-slow stochastic dynamical systems," *SIAM J. Appl. Dynam. Syst.*, vol. 15, no. 3, pp. 1327–1351, 2016.
[34] Z. Lin, W. Zhang, and X. Tang, "Learning partial differential equations for computer vision," Peking Univ., Chin. Univ. of Hong Kong, 2008.
[35] Z. Lin, W. Zhang, and X. Tang, "Designing partial differential equations for image processing by combining differental invariants," Microsoft Res., Berkeley, CA, USA, Tech. Rep. MSR-TR-2009-192, 2009.

Tal Shnitzer received the B.Sc. degree *(summa cum laude)* in electrical engineering and biomedical engineering from the Technion—Israel Institute of Technology, Haifa, Israel, in 2013, where she is currently working toward the Ph.D. degree in electrical engineering.

From 2012 to 2013, she worked in the field of Signal Processing and Algorithms at the Israeli defense industry. Since 2014, she has been a Teaching Assistance in the Department of Electrical Engineering, Technion—Israel Institute of Technology. Her main areas of interest include signal processing, biomedical signals, and geometric methods for time series analysis.

Ms. Shnitzer received the Meyer Fellowship and Zipers Award for 2014 and the Diane and Leonard Sherman Interdisciplinary Fellowship for 2015.

Ronen Talmon (M'11) received the B.A. degree *(cum laude)* in mathematics and computer science from the Open University in 2005, and the Ph.D. degree in electrical engineering from the Technion—Israel Institute of Technology, Haifa, Israel, in 2011. He is an Assistant Professor of electrical engineering at the Technion–Israel Institute of Technology.

From 2000 to 2005, he was a Software Developer and Researcher in a technological unit of the Israeli Defense Forces. From 2005 to 2011, he was a Teaching Assistant in the Department of Electrical Engineering, Technion—Israel Institute of Technology. From 2011 to 2013, he was a Gibbs Assistant Professor in the Mathematics Department, Yale University, New Haven, CT, USA. In 2014, he joined the Department of Electrical Engineering in Technion—Israel Institute of Technology.

His research interests are statistical signal processing, analysis and modeling of signals, speech enhancement, biomedical signal processing, applied harmonic analysis, and diffusion geometry.

Dr. Talmon received the Irwin and Joan Jacobs Fellowship, the Andrew and Erna Fince Viterbi Fellowship, and the Horev Fellowship.

Jean-Jacques Slotine was born in Paris in 1959. He received the Ph.D. from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1983. After working at Bell Labs in the computer research department, in 1984 he joined the faculty at MIT, where he is now a Professor of mechanical engineering and information sciences, Professor of brain and cognitive sciences, and the Director of the Nonlinear Systems Laboratory. He is the coauthor of the textbooks *Robot Analysis and Control* (Wiley, 1986) and *Applied Nonlinear Control* (Prentice-Hall, 1991). Prof. Slotine was a member of the French National Science Council from 1997 to 2002, and of Singapore's A*STAR SigN Advisory Board from 2007 to 2010. He is currently on the Scientific Advisory Board of the Italian Institute of Technology.