We study the inference of latent intrinsic variables of dynamical systems from output signal measurements. The primary focus is the construction of an intrinsic distance between signal measurements, which is independent of the measurement device. This distance enables us to infer the latent intrinsic variables through the solution of an eigenvector problem with a Laplace operator based on a kernel. The signal geometry and its dynamics are represented with nonlinear observers. An analysis of the properties of the observers that allow for accurate recovery of the latent variables is given, and a way to test whether these properties are satisfied from the measurements is proposed. Scattering and window Fourier transform observers are compared. Applications are shown on simulated data, and on real intracranial Electroencephalography (EEG) signals of epileptic patients recorded prior to seizures.

Manifold Learning for Latent Variable Inference in Dynamical Systems

Ronen Talmon[1], Stéphane Mallat[2], Hitten Zaveri[3], and Ronald R. Coifman[4]

[1]Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa, Israel
[2]Ecole Normale Superieure, 45 rue dUlm, Paris, France
[3]Department of Neurology, Yale University, New Haven, CT
[4]Department of Mathematics, Yale University, New Haven, CT

Research Report YALEU/DCS/TR-1491
Yale University
June 10, 2014

Keywords: *Manifold learning, nonlinear observers, scattering transform, kernel methods*

1 Introduction

Given signal measurements $\mathbf{z}(t)$, our goal is to identify latent variables $\boldsymbol{\theta}(t)$. These latent variables may correspond to physical and natural variables, such as the state of a patient in medical diagnostic, brain activity in Electroencephalography (EEG) signal analysis, or the operational state (failure or success) of a machine, and hence, push forward our understanding of real recorded signals.

In this paper, we focus on signals without definitive ground truth for the latent variables. Thus, applying regression techniques is not possible and unsupervised analysis is required. For instance, EEG recordings translate processes that represent brain activity into sequences of electrical impulses. The significance of revealing the latent variables in EEG recordings will be demonstrated in epilepsy research [1, 2]. In this application, appropriate modeling of the brain activity may enable us to describe the measurements in their true physical intrinsic coordinates, and this, in turn, may allow for the detection and prediction of seizures.

Estimating latent variables from measurements has been heavily investigated in signal processing and statistics studies, e.g. using Bayesian learning [3], and graphical and topic models [4, 5, 6, 7, 8, 9, 10]. In the present work, we use manifold learning methods [11, 12, 13, 14, 15, 16]. These methods often analyze the signal samples "as is" by relying on the assumption that the measured signal samples $\mathbf{z}(t)$ do not fill the ambient space uniformly but rather lie on a low-dimensional manifold induced by physical and natural constraints. However, real recorded measurements typically have many sources of variability and do not belong to low-dimensional manifolds. Most such sources of variability usually do not provide crucial information on the latent variables and can thus be removed by an appropriate invariant observation operator Φ. Applying such an operator to the signal samples yields observables $\Phi z(t)$, which may then belong to a low-dimensional manifold. Finding a parameterization of this manifold allows for the computation of a coordinate system of the latent variables.

The dynamical system point of view is used for the problem formulation and signal analysis. We note that the notions of manifold and observers are central in dynamical systems research [17]. From the standpoint of dynamical systems, the problem of estimating hidden variables from measurements can be reformulated. The latent variables $\boldsymbol{\theta}(t)$ can be viewed as the hidden intrinsic state of a dynamical system, the measurements $\mathbf{z}(t)$ can be viewed as the system output signal, and then, the estimation of the hidden state variables from the output signal is at the core of dynamical systems theory. By revisiting the differential geometric approach [18, 19, 20], we give the necessary conditions for observability and stability, which allow for inferring the parameterization of the manifold of observations and computing the coordinate system of the latent intrinsic state variables.

We consider slowly varying state variables $\boldsymbol{\theta}(t)$ [21, 22]. As a consequence, the measured signal $\mathbf{z}(t)$ can be considered as locally stationary, and hence, we can restrict the scope to the problem of representing locally stationary processes. Often marginal statistics (such as histograms) are too poor to characterize complex processes. On the other hand, polynomial moments estimators of order larger than two are not precise because they have a large variance. Standard representations thus usually rely on second order moments, which are characterized by the Fourier power spectrum for stationary processes. Unfortunately, it suffers from few significant shortcomings. First, second order moments still have a relatively large variance. Second, it merely encodes the Gaussian properties, without characterizing intermittent behavior which is often very informative. Third, the Fourier power spectrum is not stable to deformations which often occur. In most nonlinear dynamical systems, the evolution of the system induces deformations or the creation of intermittent behavior in the signal. To overcome the shortcomings of the Fourier power spectrum, we propose to use the scattering transform to observe locally stationary processes. The scattering transform has a low variance because it is based on first order moments of contractive operators, it linearizes deformations, and it can represent effectively intermittent behavior [23, 24].

The main contribution of the paper is the introduction of an unsupervised data-driven method to infer slowly varying latent variables of locally stationary signals using nonlinear observers. An analysis of the properties of the observers is given, and a way to test whether they hold from the measurements is proposed. In particular, two observers are used: the common power spectrum based on the short time Fourier analysis, and the recently introduced scattering transform based on wavelet analysis. We will show that applying our method based on the latter observer to both simulation and real data enables to accurately estimate the latent intrinsic variables. Furthermore, for the real signal, we will show that the recovered latent variables have a true physical meaning, which is a remarkable result, since it is obtained implicitly by merely analyzing the measured signal, and may give rise to significant advancements in the field. In particular, we will show that the intrinsic variables recovered from intracranial EEG signals of epileptic patients, recorded just prior to seizures, exhibit a distinct trend related to the time to seizure onset.

The remainder of the paper is organized as follows. Section 2 presents the proposed manifold learning method. Section 3 addresses nonlinear observers. The observers' properties are presented, their estimation is described, and a test to empirically evaluate the validity/soundness of the properties from the measurements is given. In Section 4, the particular problem of deformations is addressed, which further motivates the introduction of the scattering transform that follows. Finally, in Section 5, experimental results are given on both simulated and real signals, which illustrate the power of the proposed method and its potential benefits.

2 The Proposed Manifold Learning Method

2.1 Problem Setting

Let $\mathbf{z}(t) \in \mathbb{R}^n$ denote a measured output signal of a dynamical system at time index t. Suppose the measurements are locally stationary and depend upon hidden variables $\boldsymbol{\theta}(t) \in \mathbb{R}^d$, which have slow variations in time. The dynamics of the underlying variables $\boldsymbol{\theta}(t)$ drive the dynamical system, and hence, $\boldsymbol{\theta}(t)$ is viewed as the natural/intrinsic state of the system. We emphasize that this state will be implicitly determined by the method (e.g. finding an adequate representation of brain activity in the EEG application), whereas in classical analysis, it is often predefined (e.g. as the position, velocity, and acceleration in tracking maneuvering targets problems).

Our goal in this work is to empirically discover the hidden intrinsic state of the system $\boldsymbol{\theta}(t)$ and its dynamics based on a sequence of measurements $\mathbf{z}(t)$, without prior knowledge on the system parameters or the description of the state. This will be done by applying a manifold learning methodology, and the intrinsic variables $\boldsymbol{\theta}(t)$ will be recovered through the eigenvectors of a graph Laplacian built from the measurements. The key component in manifold learning is to define a distance between the measurements, which in turn is used to construct the graph Laplacian. Consequently, the primary focus of the present work is to build a pairwise distance $d(\mathbf{z}(t), \mathbf{z}(\tau))$ between measurements, which satisfies the following property:

$$d(\mathbf{z}(t), \mathbf{z}(\tau)) \approx \|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(\tau)\|^2. \tag{1}$$

In this paper, we show how to construct a distance $d(\mathbf{z}(t), \mathbf{z}(\tau))$, which satisfies (1). Once we obtain such a distance, which properly compares the measurements in terms of the intrinsic state variables, we apply a standard manifold learning method.

In [25], we considered a different dimensionality reduction problem in the domain of probability distributions of $\mathbf{z}(t)$. The assumption there is that the time varying sample distribution of $\mathbf{z}(t)$, rather than the samples themselves, is driven by an intrinsic state $\boldsymbol{\theta}(t)$, yielding a low dimensional regular manifold. We showed that this domain of distributions exhibits a powerful property: nonlinear complex interferences are translated to linear operations in the domain of distributions. This property suggests that the time-varying distribution of the measurements may be of interest, especially in adverse conditions. In [25], for example, histograms were used as estimators. However, estimating the time-varying pdf from the measured signal is practically impossible because of the curse of dimensionality, i.e. there are usually not enough samples to densely cover the space, and hence, to estimate local probability densities. Although the probability density function of the measurements cannot be estimated, estimating inner products/projections of the densities with another function may be attainable. In addition, these projections maintain the linear behavior of the densities with respect to interferences. Computing estimators to such expected values, or "generalized moments", is therefore essential for the analysis of the signal.

Thus, in this paper, we present signal transforms as generalized moments that, on one hand, describe the densities well and convey sufficient information on the intrinsic state, and, on the other hand, can be accurately and efficiently estimated from measurements.

2.2 Local Analysis and the Mahalanobis Distance

Let $\Phi z(t) \in \mathbb{R}^m$ be a (possibly nonlinear) observer, which is an operator that associates an m-dimensional vector, which varies in time, to a signal $\mathbf{z}(t)$. Once the observables $\Phi z(t)$ are computed from the available signal $\mathbf{z}(t)$, the ultimate goal is to empirically invert the observation operator and recover the intrinsic state $\boldsymbol{\theta}(t)$. For example, given EEG measurements, it will enable us to recover the hidden variables representing the brain activity, allowing for a more accurate processing, and in particular, better understanding of the brain. Under the manifold learning setting, this goal can be relaxed, and it is sufficient to approximate the Euclidean distances between the hidden variables (1).

Several remarks on the statistical setting are due at this point. The intrinsic state $\boldsymbol{\theta}(t)$ is regarded as a realization of an unknown locally stationary random process, which is assumed to vary slowly compared to $\mathbf{z}(t)$. Since the hidden variables comprising the intrinsic state $\boldsymbol{\theta}$ are unknown, we further assume that locally, i.e. in a short time window, the intrinsic state at a fixed point in time t has a unit empirical variance

$$\frac{1}{L_o}\sum_{\tau\in\mathcal{I}_t}\left(\boldsymbol{\theta}(\tau)-\overline{\boldsymbol{\theta}}_t\right)\left(\boldsymbol{\theta}(\tau)-\overline{\boldsymbol{\theta}}_t\right)^T=\mathbf{I}, \tag{2}$$

where $\overline{\boldsymbol{\theta}}_t=\frac{1}{L_o}\sum_{\tau\in\mathcal{I}_t}\boldsymbol{\theta}(\tau)$, $\mathbf{I}$ is an identity matrix, and $\mathcal{I}_t$ is a sampling grid of size L_o in $[t-L_o/2,t+L_o/2]$. This assumption might not be respected in real signals. However, since the intrinsic state $\boldsymbol{\theta}(t)$ is unknown a-priori and will be empirically inferred, our method will approximate state that satisfies (2) in a way that best explains and fits the measurements. This assumption is made in many statistical and geometric methods, including Principal Component Analysis (PCA) where the search is for low dimensional uncorrelated variables [26]. The difference is that here it is made locally, and the mean of $\boldsymbol{\theta}(t)$ may vary with time.

The measured signal $\mathbf{z}(t)$ is a locally stationary random process with an unknown distribution. The observation operator Φ is applied to the random process $\mathbf{z}(t)$, which depends upon $\boldsymbol{\theta}(t)$. The result is thus a random process $\Phi z(t)$, whose values for a fixed t, are random vectors of size m. The key property is to use a local linearization of the observation operator at each time sample t in a short window, given $\boldsymbol{\theta}(t)$ according to

$$\Phi z(\tau)=\mathbb{E}[\Phi z(t)]+\mathbf{K}(t)\left(\boldsymbol{\theta}(\tau)-\boldsymbol{\theta}(t)\right)+\boldsymbol{\epsilon}(t,\tau),\ \forall\tau\in\mathcal{I}_t \tag{3}$$

where $\mathbf{K}(t)$ is a linear operator, $\boldsymbol{\epsilon}(t,\tau)$ is a random error containing higher order terms and random fluctuations. We will later show in more detail that $\mathbf{K}(t)$ entails the linearization of the dependency of the observables $\Phi z(t)$ in $\boldsymbol{\theta}$, i.e., $\mathbf{K}(t)=J_\theta(\mathbb{E}[\Phi z(t)])$, where J_θ denotes the Jacobian matrix with respect to $\boldsymbol{\theta}$.

The observables $\Phi z(t)$ will be computed by averaging in short time windows over nearly decorrelated random variables, since $\mathbf{z}(t)$ is assumed locally stationary (due to the slow variation of $\boldsymbol{\theta}(t)$). Thus, by the Central Limit Theorem, $\Phi z(t)$ may be approximately modeled by a Gaussian random process. As a result, from (2) and (3), the empirical local mean $\widehat{\boldsymbol{\mu}}(t)$ and covariance $\widehat{\mathbf{C}}(t)$ of the observables in a window of L_o observables centered at time t are approximately given by

$$\begin{aligned}\widehat{\boldsymbol{\mu}}(t) &= \frac{1}{L_o}\sum_{\tau\in\mathcal{I}_t}\Phi z(\tau)=\mathbb{E}[\Phi z(t)]-\mathbf{K}(t)\boldsymbol{\theta}(t)+\frac{1}{L}\sum_{\tau\in\mathcal{I}_t}\left(\mathbf{K}(t)\boldsymbol{\theta}(\tau)+\boldsymbol{\epsilon}(t,\tau)\right)\\ &\simeq \mathbb{E}[\Phi z(t)]-\mathbf{K}(t)\left(\boldsymbol{\theta}(t)-\overline{\boldsymbol{\theta}}(t)\right) \qquad (4)\\ \widehat{\mathbf{C}}(t) &= \frac{1}{L_o}\sum_{\tau\in\mathcal{I}_t}(\Phi z(\tau)-\widehat{\boldsymbol{\mu}}(t))(\Phi z(\tau)-\widehat{\boldsymbol{\mu}}(t))^T\\ &\simeq \mathbf{K}(t)\mathbf{K}(t)^T+\boldsymbol{\sigma}_\epsilon^2(t) \qquad (5)\end{aligned}$$

where $\boldsymbol{\sigma}_\epsilon^2(t)$ is a matrix comprising the residual terms. We remark that the two main sources that determine the "size" of $\boldsymbol{\epsilon}(t,\tau)$ are the accuracy of the representation of the expected values of observables $\mathbb{E}[\Phi z(t)]$ as a deterministic function of merely $\boldsymbol{\theta}(t)$ (i.e., $\boldsymbol{\epsilon}(t,\tau)$ comprises the affects of other nuisance factors), and the accuracy of the local linearization (3). Thus, we seek for observers that reduce $\boldsymbol{\sigma}_\epsilon^2(t)$ in light of these two aspects.

Since the measurements $\mathbf{z}(t)$ are governed by a latent state $\boldsymbol{\theta}(t)\in\mathbb{R}^d$, the manifold of the observables $\Phi z(t)\in\mathbb{R}^m$ is merely of dimension d. Indeed, the dimensions of the linear

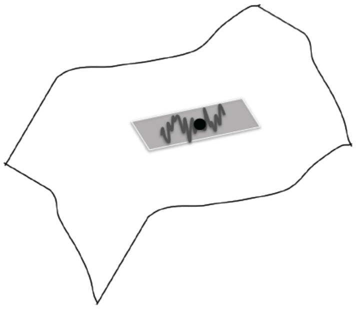


Figure 1: The black point illustrates an observable $\Phi z(t) \in \mathbb{R}^3$ for fixed t on a 2-dimensional manifold of observables. The trajectory of observables in a short time window around t, $(\Phi z(\tau),\ \tau \in \mathcal{I}_t)$, spans the tangent plane to the manifold at $\Phi z(t)$ (illustrated in gray). Therefore, the empirical covariance of the observables $\widehat{\mathbf{C}}(t)$ of this trajectory captures the shape of the tangent plane, and its principal components $\mathbf{V}_d$ are its principal directions.

operator $\mathbf{K}(t)$ are $m \times d$. Thus, by (5), assuming the elements of $\boldsymbol{\sigma}_\epsilon^2(t)$ are small, the rank of the $m \times m$ empirical covariance matrix $\widehat{\mathbf{C}}(t)$ is approximately d. In order to exploit this information, we apply the singular value decomposition (SVD) to $\mathbf{K}(t)$ and obtain its d non-zero singular values η_j and left and right singular vectors $\mathbf{v}_j$ and $\mathbf{u}_j$, respectively. From (5), by assuming that the local linearization (3) is accurate, the eigenvalue decomposition (EVD) of $\widehat{\mathbf{C}}(t)$ consists of the eigenvalues η_j^2 and eigenvectors $\mathbf{v}_j$. We use the d principal components to "filter" the covariance matrix (in a local PCA manner – by reconstructing the matrix from its principal components)

$$\tilde{\mathbf{C}}(t) = \mathbf{V}_d \boldsymbol{\Lambda}_d \mathbf{V}_d^T \tag{6}$$

where $\mathbf{V}_d$ is an $m \times d$ matrix whose columns are the d principal eigenvectors $\mathbf{v}_j$, and $\boldsymbol{\Lambda}_d$ is a $d \times d$ diagonal matrix, whose diagonal entries are the corresponding principal eigenvalues η_j^2. For simplicity, the time index is omitted from the eigenvalues and eigenvectors. Geometrically, the eigenvectors in $\mathbf{V}_d$ span the tangent plane to the manifold of the observations at $\Phi z(t)$. In addition, the different "lengths" of the principal directions, as conveyed by the eigenvalues η_j^2 of the local covariance matrix $\widehat{\mathbf{C}}(t)$, stem solely from the translation of the intrinsic state to the observation domain (depending on the measurement modality), since we assume in (2) that the intrinsic state is of unit variance. See Fig. 1 for a geometric illustration of the problem. In order to invert the effect of the observation, we apply a whitening procedure and build $\widehat{\mathbf{C}}^\dagger(t)$ as follows:

$$\widehat{\mathbf{C}}^\dagger(t) = \mathbf{V}_d \boldsymbol{\Lambda}_d^{-1} \mathbf{V}_d^T \tag{7}$$

We remark that in light of the last two steps, $\widehat{\mathbf{C}}^\dagger(t)$ can be defined as the pseudo-inverse of the local empirical covariance matrix $\widehat{\mathbf{C}}(t)$. In addition, the filtering through the EVD of the covariance matrix can be viewed as applying a local PCA procedure.

To construct a distance that satisfies (1), we use the Mahalanobis distance, as proposed by Singer and Coifman to define affinities that locally invert the observation [27]. The Mahalanobis distance often appears in the context of metric learning and leads to good performance in a broad range of applications [28, 29, 30]. Since the Mahalanobis distance compares two Gaussian, or nearly Gaussian, random vectors, it is an appropriate distance, given two realizations $\Phi z(t)$ and $\Phi z(\tau)$, which are assumed to be samples from nearly Gaussian distributions (due to the observation operator) and whose means are related through (3). The Mahalanobis distance is given by

$$\begin{aligned} d(\mathbf{z}(t), \mathbf{z}(\tau)) &= \frac{1}{2}\left((\Phi z(t) - \widehat{\boldsymbol{\mu}}(t)) - (\Phi z(\tau) - \widehat{\boldsymbol{\mu}}(\tau))\right)^T \\ &\times \left(\widehat{\mathbf{C}}^\dagger(t) + \widehat{\mathbf{C}}^\dagger(\tau)\right)\left((\Phi z(t) - \widehat{\boldsymbol{\mu}}(t)) - (\Phi z(\tau) - \widehat{\boldsymbol{\mu}}(\tau))\right). \end{aligned} \tag{8}$$

The local linearization of the observation operator that relates the means (3) allows to further justify the usage of the Mahalanobis distance. By assuming that the local linearization is accurate, i.e., $\boldsymbol{\sigma}_\epsilon^2(t)$ is negligible, substituting (3) and (7) into (8) and using the SVD of $\mathbf{K}(t)$ yields (1), thereby satisfying the main goal. For the approximation order and more details, we refer the readers to [27, 31]. We remark that minimizing the size of $\boldsymbol{\sigma}_\varepsilon^2(t)$ encapsulates a tradeoff in setting L_o. Small values of L_o yield an accurate linearization and a small "model mismatch" error at the expense of fewer samples and large estimation variance.

By further assuming the following local Gaussian model at time t, for $\tau \in \mathcal{I}_t$[1]:

$$\begin{aligned} \boldsymbol{\theta}(\tau) &\sim \mathcal{N}(\mathbb{E}[\boldsymbol{\theta}(t)], \mathbf{I}_d) && (9) \\ \Phi z(\tau)|\boldsymbol{\theta}(\tau) &\sim \mathcal{N}(\mathbb{E}[\Phi z(t)] + \mathbf{K}(t)\boldsymbol{\theta}(\tau), \sigma_\epsilon^2(t)\mathbf{I}_m) && (10) \end{aligned}$$

Tipping and Bishop [26] showed that (4) is the maximum likelihood (ML) estimate of $\mathbb{E}[\Phi z(t)]$,

$$\widehat{\sigma}_\epsilon^2(t) = \frac{1}{m-d}\sum_{i=d+1}^{m} \eta_i^2 \tag{11}$$

is the ML estimate of $\sigma_\epsilon^2(t)$, and

$$\widehat{\mathbf{K}}(t) = \mathbf{V}_d\left(\boldsymbol{\Lambda}_d - \widehat{\sigma}_\epsilon^2 \mathbf{I}_d\right)^{1/2} \tag{12}$$

is the ML estimate of $\mathbf{K}(t)$. Tipping and Bishop further showed that

$$\mathbb{E}[\boldsymbol{\theta}(\tau)|\Phi z(\tau)] = \left(\boldsymbol{\Lambda}_d - \widehat{\sigma}_\varepsilon^2 \mathbf{I}_d\right)^{1/2} \boldsymbol{\Lambda}_d^{-1} \mathbf{V}_d^T \left(\Phi z(\tau) - \widehat{\boldsymbol{\mu}}(t)\right). \tag{13}$$

It implies that under these local Gaussian models, the Mahalanobis distance (8) between two samples $\Phi z(\tau)$ and $\Phi z(\tau')$ in the same local neighborhood around time t, i.e., $\tau, \tau' \in \mathcal{I}_t$, corresponds to the Euclidean distance between the posterior expectations

$$d(\mathbf{z}(\tau), \mathbf{z}(\tau')) = \left\|\mathbb{E}[\boldsymbol{\theta}(\tau)|\Phi z(\tau)] - \mathbb{E}[\boldsymbol{\theta}(\tau')|\Phi z(\tau')]\right\|^2, \tag{14}$$

when assuming small error terms, i.e. $\sigma_\epsilon^2 \ll 1$. In addition to the statistical justification, this interpretation further supports the search for local near Gaussian observables. We remark that, on one hand, (13) includes a "denoising" procedure applied by subtracting the ML estimate of the variance of the error term $\sigma_\epsilon^2(t)$. On the other hand, it assumes that the error terms in (3) are independent among the coordinates of the observables, and it is restricted to the local neighborhood.

[1] (9) implies that the state is locally Gaussian, and (10) implies that $\boldsymbol{\epsilon}(t, \tau)$ in (3) is a Gaussian random vector of independent variables.

2.3 Manifold Learning

Suppose a finite sequence of T messurements $\mathbf{z}(t), t = 1, \dots, T$, is available. Let $\mathbf{W}$ be a pairwise $T \times T$ affinity matrix (kernel) between the measurements based on a Gaussian and the Mahalanobis distance (8), whose (t, τ)-th element is given by

$$W_{t,\tau} = \exp\left\{ -\frac{d(\mathbf{z}(t), \mathbf{z}(\tau))}{\varepsilon} \right\}, \tag{15}$$

where ε is the kernel scale, which can be set according to Hein and Audibert [32] and Coifman *et al.* [33]. Based on the kernel, we form a weighted graph, where the measurements $\mathbf{z}(t)$ are the graph nodes and the weight of the edge connecting node $\mathbf{z}(t)$ to node $\mathbf{z}(\tau)$ is $W_{t,\tau}$. In particular, such a Gaussian kernel exhibits a notion of locality by defining a neighborhood around each measurement $\mathbf{z}(t)$ of radius ε, i.e., measurements $\mathbf{z}(\tau)$ such that $d(\mathbf{z}(t), \mathbf{z}(\tau)) > \varepsilon$ are weakly connected to $\mathbf{z}(t)$. In the current implementation, we set ε to be the median of the pairwise distances. According to the graph interpretation, this implies a well-connected graph because each measurement is effectively connected to half of the other measurements.

Let $\mathbf{D}$ be a diagonal matrix whose elements are the row sums of $\mathbf{W}$, and let $\mathbf{W}^{\text{norm}} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ be a normalized kernel that shares its eigenvectors with the normalized graph-Laplacian, defined by $\mathbf{I} - \mathbf{W}^{\text{norm}}$ [34]. The eigenvectors of $\mathbf{W}^{\text{norm}}$, denoted by $\boldsymbol{\varphi}_j$, provide a new coordinate system for the measurements, which reveal their underlying structure [15]. The eigenvalues are ordered such that $|\lambda_0| \geq |\lambda_1| \geq \cdots \geq |\lambda_{T-1}|$, where λ_j is the eigenvalue associated with eigenvector $\boldsymbol{\varphi}_j$. Because $\mathbf{W}^{\text{norm}}$ is similar to $\mathbf{D}^{-1}\mathbf{W}$, and $\mathbf{D}^{-1}\mathbf{W}$ is row-stochastic, $\lambda_0 = 1$ and $\boldsymbol{\varphi}_0$ is the diagonal of $\mathbf{D}^{1/2}$. The next few eigenvectors are traditionally referred to as a parameterization (description of the geometry) of the underlying manifold [15]. In particular, based on the d principal eigenvectors (without the trivial one), a d-dimensional embedding of the signal $\mathbf{z}(t)$ is constructed as

$$\mathbf{z}(t) \mapsto (\varphi_1(t), \varphi_2(t), \dots, \varphi_d(t))^T. \tag{16}$$

This embedding defines an "inverse map" between the measurements and the intrinsic state, such that (without loss of generality) the t-th coordinate of the j-th eigenvector, i.e., $\varphi_j(t)$, represents the j-th coordinate of $\boldsymbol{\theta}(t)$.

To conclude this section, we summarize the proposed algorithm in Algorithm 1.

3 Nonlinear Observers

3.1 Observer Properties and Estimation

In this subsection we articulate the properties required by the algorithm for a small residual term $\boldsymbol{\epsilon}(t, \tau)$ in the key condition (3), such that (1) is achieved, using a dynamical systems approach.

Define the following properties:

Algorithm 1 The Proposed Algorithm

Input: a finite sequence of signal samples $\mathbf{z}(t) \in \mathbb{R}^n$.

Output: a low dimensional representation of the signal samples $\boldsymbol{\theta}(t) \in \mathbb{R}^d$ through eigenvectors of a kernel.

1. Compute the observables $\Phi z(t)$ by applying an observation operator Φ to the signal samples $\mathbf{z}(t)$.
2. For each observable $\Phi z(t)$, compute the empirical mean $\widehat{\boldsymbol{\mu}}(t)$ and empirical covariance matrix $\widehat{\mathbf{C}}(t)$ in a short window of L_o observables centered at t according to (4) and (5).
3. Compute the pseudo inverse matrices $\widehat{\mathbf{C}}^{\dagger}(t)$ of $\widehat{\mathbf{C}}(t)$.
4. Build a kernel $\mathbf{W}$ that consists of pairwise affinities between the observables $\Phi z(t)$ according to (15). The affinity function is based on a distance metric (8), which is constructed based on the empirical means $\widehat{\boldsymbol{\mu}}(t)$ and pseudo-inverse covariance matrices $\widehat{\mathbf{C}}^{\dagger}(t)$.
5. Build a normalized kernel $\mathbf{W}^{\text{norm}} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is a diagonal matrix whose elements are the sum of rows of $\mathbf{W}$.
6. Apply eigenvalue decomposition (EVD) to $\mathbf{W}^{\text{norm}}$ and obtain a set of d eigenvectors $\boldsymbol{\varphi}_j$ associated with the d largest eigenvalues.
7. View the eigenvectors as a low dimensional representation of the signal samples (16), i.e., the jth coordinate of $\boldsymbol{\theta}(t)$ is represented by $\varphi_j(t)$.

Observability The intrinsic state $\boldsymbol{\theta}(t)$ is *observable* through the observables $\Phi z(t)$ if there exists a constant $A > 0$ such that for any t and τ

$$A\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(\tau)\|^2 \leq \|\mathbb{E}[\Phi z(t)] - \mathbb{E}[\Phi z(\tau)]\|^2. \tag{17}$$

In a geometric context, where we can view the intrinsic state $\boldsymbol{\theta}(t)$ and the associated expected values $\mathbb{E}[\Phi z(t)]$ as points in d- and m-dimensional domains, respectively, this condition implies that small perturbations of the intrinsic state in dimension d are detected in the observation domain of dimension m.

Stability An observer Φ is *stable* if there exists a constant $B > 0$ such that for any t and τ

$$\|\mathbb{E}[\Phi z(t)] - \mathbb{E}[\Phi z(\tau)]\|^2 \leq B\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(\tau)\|^2. \tag{18}$$

In a geometric context, this condition implies that small perturbations of the intrinsic state are not translated to very large (infinite) perturbations in the observation domain.

In other words, an observer is informative and sensitive with respect to the intrinsic state $\boldsymbol{\theta}(t)$ if small variations of the these factors are detectable in the observation domain (i.e., discriminability of the states). Similarly, an observer is stable and regular with respect to the intrinsic state $\boldsymbol{\theta}(t)$ if small variations of these factors are translated to small variations of the observations. Under this setting, the observability and stability properties are equivalent to the condition that the observation $\mathbb{E}[\Phi z(t)]$ is bi-Lipschitz with respect to $\boldsymbol{\theta}$. In Section 3.2, we will show that testing whether these properties are satisfied can be done through the local covariance matrices of the observables, which are estimated and used to define the pivotal Mahalanobis distance (8).

Invariance to noise and nuisance factors Let $\boldsymbol{\nu}$ be a noise or nuisance variable. An observer Φ is invariant to $\boldsymbol{\nu}$ if

$$\frac{\|\mathbb{E}[\Phi z(t)] - \mathbb{E}[\Phi z(\tau)]\|^2}{\|\boldsymbol{\nu}(t) - \boldsymbol{\nu}(\tau)\|^2} \ll 1. \tag{19}$$

Since, the manifold of observations is determined by the problem, it could be very complex. Geometrically, due to high levels of noise, small perturbations of the intrinsic state $\boldsymbol{\theta}(t)$ may be considerably stretched and distorted when translated to the observable domain in directions that do not necessarily respect the shape of the manifold. In addition, the state of real dynamical systems may not be low dimensional. However, the number of state dimensions relevant to the task at hand are usually small. Thus, (19) implies that the observer is resilient to measurement noise and nuisance factors, thereby ensuring that the shape of the manifold induced by the intrinsic state coordinates $\boldsymbol{\theta}(t)$ can be detected by the observables.

Thus far in this section, the focus was on defining the desirable properties of the expected values of the observers, taking into account the time variability of the hidden intrinsic state. Now, we compute estimators calculated from the random signal realizations. These estimators rely on the local stationarity assumption. Assuming local ergodicity, the expected values are calculated with time empirical averages in short time windows of samples of length L_s. The choice of the window and its length L_s, in which the observers are estimated, is of particular importance and represents the "bias-variance" tradeoff: a longer window yields a more accurate estimation at the expense of a bias caused by the time variation of the intrinsic state, which hampers the local stationarity assumption. The length L_s of the window also introduces the "micro"/"fine" time scale of the proposed method. Namely, we assume that the estimation variance is smaller than the time variations of the expected values (originated by the variations of the intrinsic state) in time windows of length L_s. This assumption enables us to separate the scales of the dynamics and the estimation and compute the observables without including/discarding the variations/dynamics of the intrinsic state. On the other hand, the coarser time scale is defined by time windows of observables $\Phi z(t)$ of length L_o, in which we estimate their empirical mean $\widehat{\boldsymbol{\mu}}(t)$ and covariance matrix $\widehat{\mathbf{C}}(t)$ in (4) and (5), respectively, assuming a near Gaussian distribution of the observables.

We remark that the estimation variances in the different coordinates of the observer might not be identical due to the properties of the signal or the properties of the observer (e.g., a multiscale transform with a different time support in each coordinate). Therefore, we apply an additional standardization procedure; in each coordinate, the estimator is normalized/divided by the standard deviation of the empirical average over the samples in the window.

The observer at time τ can be rewritten as an estimator

$$\Phi z(\tau) = \mathbb{E}[\Phi z(\tau)] + \boldsymbol{\epsilon}_{\text{est}}(\tau), \tag{20}$$

where $\boldsymbol{\epsilon}_{\text{est}}(\tau)$ is the observer estimation error. Now, assuming the observer is invariant to $\boldsymbol{\nu}$ (property (c) holds), the first order Taylor expansion of $\mathbb{E}[\Phi z(\tau)]$ around $\boldsymbol{\theta}(t)$ for all $\tau \in \mathfrak{I}_t$ yields

$$\mathbb{E}[\Phi z(\tau)] = \mathbb{E}[\Phi z(t)] + (J_\theta(\mathbb{E}[\Phi z(t)]))(\boldsymbol{\theta}(\tau) - \boldsymbol{\theta}(t)) + \boldsymbol{\epsilon}_{\text{lin}}(t,\tau) \tag{21}$$

where $J_\theta(\mathbb{E}[\Phi z(t)])$ denotes the Jacobian of $\mathbb{E}[\Phi z(t)]$ with respect to $\boldsymbol{\theta}$, i.e., $(J_\theta(\mathbb{E}[\Phi z(t)]))_{ij} = \partial \mathbb{E}[\Phi z(t)]_i / \partial \theta_j$, and $\boldsymbol{\epsilon}_{\text{lin}}(t,\tau)$ consists of residual higher order terms. Finally, (21) gives a rigorous formulation of the linearization in (3), where $\boldsymbol{K}z(t) = (J_\theta \mathbb{E}[\Phi z(t)])$ and $\boldsymbol{\epsilon}(t,\tau) = \boldsymbol{\epsilon}_{\text{lin}}(t,\tau) + \boldsymbol{\epsilon}_{\text{set}}(t,\tau)$.

3.2 Observation Quality Empirical Test

The observability and stability properties, designated by the bi-Lipschitz condition applied to the observation function, suggest that the quality of an observer may be related to the ratio between the lower and upper bounds, A and B, respectively; as the bounds are tighter, the observation function is more regular and deforms less the intrinsic state, thereby allowing for a more accurate inversion of the observation.

Substituting the linearization (21) into (17) and (18) yields that the observability and stability conditions can be rewritten as

$$A \leq \|\boldsymbol{K}z(t)\|^2 \leq B. \tag{22}$$

In addition, the relation between the local covariance matrices and the Jacobians of the observers in (5) implies that the local covariance matrices of size $m \times m$ are of lower rank d. This implies that $\widehat{\mathbf{C}}(t)$ has approximately d nonzero positive eigenvalues, and each eigenvalue approximates the square of the corresponding singular value of the Jacobian matrix $Kz(t)$. Since the lower and upper bounds A and B are given by the smallest and largest singular values of the Jacobian matrix $Kz(t)$ over all times t, their ratio can be estimated empirically via

$$\widehat{\rho}(t) = \sqrt{\frac{\eta_d(t)}{\eta_1(t)}} \approx \frac{A}{B} \tag{23}$$

where $\eta_j(t)$ is the j-th largest eigenvalue of $\widehat{\mathbf{C}}(t)$. The empirical ratio ranges between $0 \leq \widehat{\rho}(t) \leq 1$, where 0 implies distorted observables and 1 implies a well represented signal (the observation operator as function of the hidden state is close to identity).

We remark that in case d is known, the eigenvalues $\eta_{d+1}, \ldots, \eta_m$ indicate the invariance of the observer to the nuisance factors. In case d is unknown, it can be determined by the spectral gap of the spectrum of the empirical covariance matrices.

4 Time Deformations and Scattering Moments

Nonlinearities in complex systems usually introduce deformations and an intermittent behavior. In the present work, we focus on a special type of such artifacts – time deformations, which are widely spread in real-life signals. Although the focus of the analysis in this section is on time deformations, many of the results can be extended to deformations and intermittencies in general [23, 24, 35].

4.1 Fourier Power Spectrum and Instability to Time Deformations

A common observer of (usually 1-dimensional) signals, which is also widely spread in manifold learning techniques [36, 37], is the Fourier power spectral density. Define the observables $\Phi_F z(t)$ as the vectors

$$\Phi_F z(t) = \left(\Phi_F z(t,\xi)\right)_\xi , \tag{24}$$

with

$$\Phi_F z(t,\xi) = \left| \int z(\tau) e^{j\xi\tau} w(t-\tau) d\tau \right|^2 * \phi(t) \tag{25}$$

where $w(t)$ is the short-time analysis window, t is the time frame index, and ξ is the frequency band. $\Phi_F z(t)$ is therefore the Fourier power spectrum estimate of time frame t obtained by averaging the square amplitudes of the Fourier transform of the signal in time using a smoothing window $\phi(t)$ of length L_s. The Fourier power spectrum itself is defined as

$$\mathbb{E}[\Phi_F z(t,\xi)] = \mathbb{E} \left| \int z(\tau) e^{j\xi\tau} w(t-\tau) d\tau \right|^2 . \tag{26}$$

Consider a special case in which the output signal $x(\tau)$ of a dynamical system undergoes time deformation, which is given by

$$z(\tau) = x(\tau + \theta(\tau)). \tag{27}$$

In this special case, for simplicity, we assume that the time deformation is the only hidden state variable controlling the measured signal, i.e., $d = 1$.

By applying linear approximation to $\theta(t)$ in each short time analysis window $w(t)$ around t with respect to t, the time deformation can be split into translation and scaling:

$$z(\tau) = x(\tau + \theta(\tau)) \simeq x(\tau + \theta(t) + \theta'(t)(\tau - t)) = x(\theta(t) - t\theta'(t) + \tau(1 + \theta'(t))) \tag{28}$$

where τ is the time index of the measured signal, t is the time frame index of the short time power spectral density, and $\theta'(t)$ is the first derivative of $\theta(t)$ with respect to t. Since the Fourier power spectrum is time shift invariant and by utilizing the smoothness of the short time analysis window (insensitive to small dilations), we get that

$$\mathbb{E}[\Phi_F z(t,\xi)] \simeq \mathbb{E}[\Phi_F x(t, \xi/(1+\theta'(t)))] \tag{29}$$

Thus, (29) implies that even small time deformations ($\theta'(t) \ll 1$) are translated to large distortions in high frequencies ($\xi \gg 1$). As a result, we need to look for a better observer which is stable with respect to the deformation.

Next, we provide an empirical test to identify time deformations. Differentiating (29) with respect to the time frame index yields

$$\frac{\partial \mathbb{E}[\Phi_F z(t,\xi)]}{\partial t} = \xi \frac{\theta''(t)}{(1+\theta'(t))^2} (\mathbb{E}[\Phi_F x(t, \xi/(1+\theta'(t)))])' \tag{30}$$

where $\theta''(t)$ denotes the second derivate of $\theta(t)$ with respect to t, and $(\mathbb{E}[\Phi_F x(t,\xi)])'$ denotes the first derivative of $\mathbb{E}[\Phi_F x(t,\xi)]$ with respect to the frequency band variable. Differentiating with respect to the frequency yields

$$\frac{\partial \mathbb{E}[\Phi_F z(t,\xi)]}{\partial \xi} = \frac{1}{1+\theta'(t)}(\mathbb{E}[\Phi_F x(t,\xi/(1+\theta'(t)))])'. \tag{31}$$

Let $\gamma(t,\xi)$ denote the ratio of the partial derivatives, which is given by

$$\gamma(t,\xi) = \frac{\partial \mathbb{E}[\Phi_F z(t,\xi)]}{\partial t} / \frac{\partial \mathbb{E}[\Phi_F z(t,\xi)]}{\partial \xi} = \xi \frac{\theta''(t)}{1+\theta'(t)} \tag{32}$$

and its logarithm separates the dependencies on time and frequency and can be expressed as

$$\log \gamma(t,\xi) = \log \xi + \alpha(t) \tag{33}$$

where $\alpha(t) = \theta''(t)/(1+\theta'(t))$. Finally, averaging over time yields

$$\int \log \gamma(t,\xi) dt = \log \xi + \int \alpha(t) dt. \tag{34}$$

Thus, to test time deformation presence, we propose to empirically compute the average log ratio $\int \log \widehat{\gamma}(t,\xi)dt$ over the signal samples in the available time interval for each frequency, where

$$\widehat{\gamma}(t,\xi) = \frac{\partial \Phi_F z(t,\xi)}{\partial t} / \frac{\partial \Phi_F z(t,\xi)}{\partial \xi}, \tag{35}$$

and test whether it is a linear function (the curve of the function is a line) of the frequency bin with slope 1.

We remark, that under our assumptions, the time deformation is the main source of variability and the other "nuisance" factors change slowly. Without time deformations, only slow "nuisance" factors remain to drive the dynamics of the system, and hence, the time derivative of the Fourier power spectrum should be close to zero.

4.2 Scattering Moments

Scattering moments are computed based on a cascade of wavelet transforms and modulus operators, and can be viewed as expected values of a transformation of the random signal [23, 24]. In this section, we briefly review their construction procedure. For simplicity, we will merely show here the construction of the first and second order scattering moments. In addition, we show it for 1-d signals, i.e. $z(t) \in \mathbb{R}$. For high dimensional signals, the same procedure is applied to each coordinate independently.

Let $\psi(t)$ be a complex wavelet, whose real and imaginary parts are orthogonal and have the same L_2 norm. Let $\psi_j(t)$ denote the dilated wavelet, defined as

$$\psi_j(t) = 2^{-j}\psi(2^{-j}t), \quad \forall j \in \mathbb{Z}. \tag{36}$$

Define the first and second order scattering transforms of $z(t)$ as

$$\begin{aligned} \Phi_S z(t,j_1) &= |z(t) * \psi_{j_1}(t)| * \phi(t) && (37) \\ \Phi_S z(t,j_1,j_2) &= ||z(t) * \psi_{j_1}(t)| * \psi_{j_2}(t)| * \phi(t), && (38) \end{aligned}$$

where $\phi(t)$ is the wavelet scaling (analysis) window of length L_s.

The first and second order scattering moments of $z(t)$ are defined as the expected values of the modulus of the wavelet transform of $z(t)$ and are given by

$$\begin{aligned} \mathbb{E}[\Phi_S z(t,j_1)] &= \mathbb{E}\left[|z(t)*\psi_{j_1}(t)|\right], \qquad (39) \\ \mathbb{E}[\Phi_S z(t,j_1,j_2)] &= \mathbb{E}\left[||z(t)*\psi_{j_1}(t)|*\psi_{j_2}(t)|\right] \qquad (40) \end{aligned}$$

Let $\Phi_S z(t)$ denote the observables computed from the signal samples $z(t)$ based on the estimates of the first and second scattering moments:

$$\Phi_S z(t) = (||z(t)*\psi_{j_1}(t)|*\psi_{j_2}(t)|*\phi(t) : \forall (j_1,j_2) \in \mathbb{Z}^m, m \in \{1,2\})_{j_1,j_2}. \qquad (41)$$

Scattering moments have been shown to be an observer that is especially suitable for deformations and intermittencies [23, 24]. In particular, it was shown that the scattering moments are stable (Lipschitz) with respect to time deformations. Therefore, we claim that the application of the scattering transform as an observer prior to the application of the manifold learning methodology is natural and useful. First, scattering moments have the properties of "good" observers as described in Section 3.1. Second, they can be accurately estimated from a single realization of the signal with a low estimation variance. Third, we will show in Section 5 that, indeed, for simulated and real signals, scattering moments outperform the commonly used Fourier power spectrum.

5 Experimental Results

5.1 Test Case - Autoregressive Process

In this section we examine a particular case of a linear system that is mathematically traceable and present the results on this synthetic example to illustrate our proposed methodology.

Consider a case in which we measure the output of a first order time variant autoregressive (AR) system with time deformation. Let $z(t)$ denote the output signal of the system, whose time evolution (in discrete time) is given by

$$\begin{aligned} x(t) &= \nu(t)*u(t)+\theta_1(t)x(t-1) \\ z(t) &= x(t-\theta_2(t)) \qquad (42) \end{aligned}$$

where $\boldsymbol{\theta}(t) = (\theta_1(t), \theta_2(t))$ is the hidden state that controls both the system temporal dynamics and time deformation, $\nu(t)$ is a nuisance factor, and $u(t)$ is a white Gaussian driving/excitation noise. Such an AR process is used in a broad range of applications to model signals. For example, it is widely used for modeling the human vocal tract in speech recognition tasks and for modeling financial time series [38, 39].

We remark that in [40], a similar task was presented, but merely one hidden variable (controlling the dynamics or the deformation) was recovered using model-based compressive sensing [41], given the other hidden variable. In this work, we will recover both variables simultaneously.

The Fourier power spectrum of $z(t)$ at time t can be written explicitly as the follows (assuming the slow varying nuisance factor $\nu(t)$ is unaffected by the time deformation)

$$\begin{aligned} \mathbb{E}[\Phi_F z(t,\xi)] &= \sigma_u^2 \sigma_\nu^2(t) \left| \frac{1}{1-\theta_1(t) e^{-j\xi/(1-\theta_2'(t))}} \right|^2 \\ &= \frac{\sigma_u^2 \sigma_\nu^2(t)}{1+\theta_1^2(t) - 2\theta_1(t)\cos(\xi/(1-\theta_2'(t)))}. \end{aligned} \tag{43}$$

If the autoregressive process is stable, i.e., the pole is in the unit circle $\theta_1(t) < 1-\epsilon$ for $\epsilon > 0$, then, a straight forward derivation yields that there exist a frequency ξ and constants A and B such that

$$A \leq |(J_{\theta_1} \mathbb{E}[\Phi_F z(t))_\xi| \leq B. \tag{44}$$

It implies that the Fourier power spectrum of the signal satisfies (at least in one frequency bin) the observability and stability conditions with respect to the hidden variable $\theta_1(t)$ that control the dynamics of the system. Indeed, in [42], we showed that in the special case, in which there is no time deformation ($\theta_2 = 0$), and the only controlling factor is the pole of the system ($\nu = 0$), the hidden variable θ_1 can be recovered effectively using the Fourier power spectrum.

On the other hand, the derivative of the Fourier power spectrum with respect to the derivative of the hidden variable θ_2' is proportional to ξ, i.e., $(J_{\theta_2'} \mathbb{E}[\Phi_F z(t)])_\xi \propto \xi$. This implies that the Fourier power spectrum is not a stable observation operator with respect to $\theta_2'(t)$.

To demonstrate our statements, we simulate $z(t), t = 1, \ldots, T$, where $T = 2^{16}$ is the number of simulated samples, according to (42). The hidden variables are simulated according to

$$\begin{aligned} \theta_1(t) &= 0.1 + 0.3\sin(\pi t/T) + 0.02 w_1(t) \\ \theta_2(t) &= 0.1 + 0.4(t/T)^4 + 0.05 w_2(t) \end{aligned}$$

where $w_1(t)$ and $w_2(t)$ are white Gaussian noise processes with standard normal distribution, and the slowly varying nuisance factor $\nu(t)$ is simulated according to

$$\nu(t) = 0.95 + 0.1\sin(2\pi t/T). \tag{45}$$

First, we empirically test the existence of time deformation in the simulated signal using the empirical test from Section 4.1. Figure 2 plots $\int \log \widehat{\gamma}(t,\xi) dt$ as a function of the frequency logarithm $\log \xi$. Indeed, as expected in (34), we obtain a line, whose slope is approximately 1. We remark that repeating the simulation with $\theta_2(t) \equiv 0$ yields a roughly constant line.

Figure 3 shows the obtained results of the application of Algorithm 1 to the simulated signal using the Fourier power spectrum as an observer. Figure 3(a) depicts the eigenvalues of the kernel λ_i. As seen, there is one dominant eigenvalue and the rest are much smaller. It implies that merely one hidden variable is identified. Figure 3(b) shows a scatter plot of the T coordinates of the obtained principal eigenvector φ_1 as a function of the corresponding T samples of the hidden variable $\theta_1(t)$, which controls the evolution of the AR system. We observe a strong correspondence (high correlation) between the values, suggesting that the

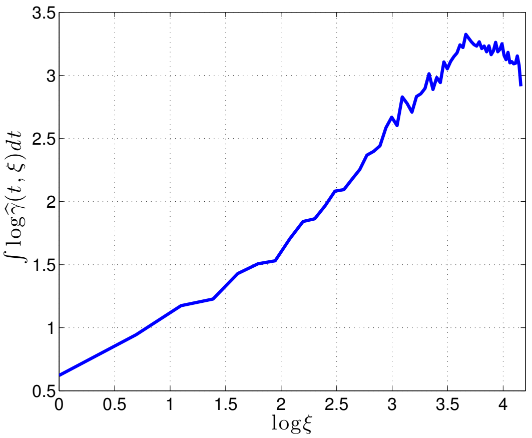


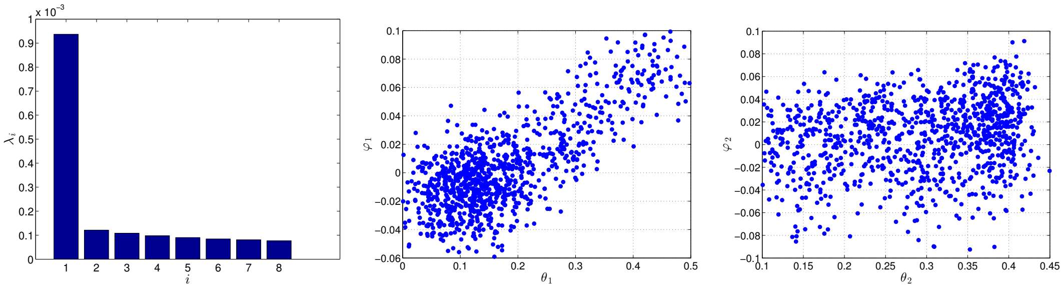
Figure 2: Empirical test for time deformation. A plot of $\int \log \widehat{\gamma}(t,\xi)dt$ as a function of the frequency logarithm $\log \xi$.

hidden variable is discovered and well represented by φ_1. Figure 3(c) shows a scatter plot of the T coordinates of φ_2 as a function of the corresponding T samples of the hidden variable $\theta_2(t)$, which governs the time deformation. As seen in the figure, the correspondence is weak, implying that φ_2 does not represent well $\theta_2(t)$. Indeed, Figure 3(a) indicates that only a single variable is recovered in this experiment. This demonstrates the analysis in Section 4.1 that shows that Fourier power spectrums are unstable observers in presence of time deformations.

Figure 4 is similar to Fig. 3 and shows the obtained results of the application of Algorithm 1 to the simulated signal using the scattering transform as an observer. Figure 3(a) depicts the eigenvalues of the kernel λ_i. As seen, compare to Fig 3(a), the spectrum decay is slower, indicating there are several dominant components. It implies that more than one hidden variable is identified. Figure 4(b) shows a scatter plot of the T coordinates of the obtained principal component φ_1 as a function of the corresponding T samples of the hidden variable $\theta_1(t)$. We observe a strong correspondence (high correlation) between the values (similar to Fig. 3(b)), suggesting that the hidden variable is discovered and well represented by the principal component in this case as well. Figure 4(c) shows a scatter plot of the T coordinates of φ_2 as a function of the corresponding T samples of the hidden variable $\theta_2(t)$. Here, unlike in Fig. 3(c), the correspondence is strong, implying that $\theta_2(t)$ is recovered and represented well by φ_2. These results correspond to the slower decay of the spectrum shown in Fig. 4(a) compared to Fig. 3(a) and to the fact that the scattering transform is a bi-Lipschitz observer with respect to time deformations.

5.2 Intracranial EEG Signal Analysis

In this section, we apply our method to intracranial EEG (icEEG) signals collected from a single epilepsy patient at the Yale-New Haven Hospital. The problem of identifying preseizure states in epilepsy patients has become a major focus of research during the last few decades [43, 1, 44]. Still, the question whether such states exist, and in particular, whether they can be detected in icEEG signals, is a controversy in the research community



(a) (b) (c)

Figure 3: The obtained results of the application of Algorithm 1 to the simulated signal using the Fourier power spectrum as an observer. (a) The eigenvalues (spectrum) of the kernel λ_i. Only one dominant eigenvalue exists, which implies that merely one hidden variable is identified. (b) A scatter plot of the T coordinates of φ_1 as a function of the corresponding T samples of the hidden variable $\theta_1(t)$, which controls the evolution of the AR system. We observe a strong correspondence, suggesting that the hidden variable is discovered and well represented by φ_1. (c) A scatter plot of the T coordinates of φ_2 as a function of the corresponding T samples of the hidden variable $\theta_2(t)$, which governs the time deformation. Since the correspondence is weak, φ_2 does not represent well $\theta_2(t)$.

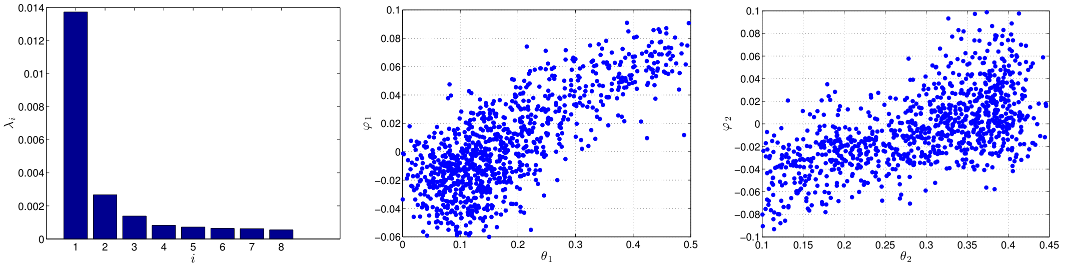
[45]. Thus, extracting in an unsupervised manner hidden variables from icEEG signals recorded prior to seizures, as well as showing that the extracted variables correspond to seizure indicators, are of great importance.

We process 3 contacts implanted in the right occipital lobe of the patient: Contact 1 and Contact 2 are located at the seizure onset area, and Contact 3 is located remotely from seizure onset area. We study recordings that immediately precede six epilepsy seizure episodes (excluding the seizures themselves), each 35-minutes long. The seizures were identified according to the analysis of a human expert, who marked the seizure time onset of each of the 6 seizures. The signals are sampled at rate of 256 Hz. A detailed description of the collected dataset can be found in [46]. We present the results obtained based on Seizure 1 and report that similar results are obtained for all six seizures.

Figure 5 presents the measured signal in Contact 1, which is located close to the seizure initiation location, that immediately precedes Seizure 1 . The figure depicts both (bottom) the signal in time and (top) its Fourier power spectrum. We observe no visible trend in both the signal or the power spectrum. In particular, it is difficult by observation to notice differences between the recording parts that immediately precede the seizure and parts that are located several minutes before the seizure.

We apply two observation operators to the signal. The first is the Fourier power spectrum, as described in Section 4.1, using Hamming analysis windows of length 1024 samples and 50% overlap. The second is the scattering transform, described in Section 4.2, with Morlet wavelet of length 1024 samples and 50% overlap.

In Fig. 6, we examine the quality of the computed observables according to the empirical test proposed in Section 3.2. The figure shows the log ratio between the largest and the k-th eigenvalues of the local covariance matrix as a function of time to seizure onset obtained



(a) (b) (c)

Figure 4: The obtained results of the application of Algorithm 1 to the simulated signal using the scattering transform as an observer. (a) The eigenvalues (spectrum) of the kernel λ_i. The spectrum decay is slower, indicating that more than one hidden variable is identified. (b) A scatter plot of the T coordinates of φ_1 as a function of the corresponding T samples of the hidden variable $\theta_1(t)$. The strong correspondence suggests that the hidden variable $\theta_1(t)$ is discovered and well represented by φ_1. (c) A scatter plot of the T coordinates of φ_2 as a function of the corresponding T samples of the hidden variable $\theta_2(t)$. The correspondence is strong, indicating that $\theta_2(t)$ is recovered and represented well by φ_2.

based on (a) the Fourier power spectrum and (b) the scattering transform. The presented ratios are based on the signal recorded in Contact 1 before Seizure 1. Similar results are obtained for all six seizures and three contacts. According to our analysis, as the ratio $\rho(t)$ (23) is closer to 1 (and stable over time), the Lipschitz bounds of the observation (and transform) are better. We observe that the ratios based on the scattering transform are closer to 1 and more stable over time compared to the ratios based on the Fourier power spectrum, thereby implying that the scattering moments are indeed better in terms of observability and stability for this signal.

Figure 7 presents the scatter plots of the 3 dimensional embedding (16) of the observables, setting $d = 3$. Figures 7 (a), (c), and (e) are based on the Fourier power spectrum and Figures 7 (b), (d), and (f) are based on the scattering moments. Figures 7 (a) and (b) depict the embedding of the 35 minutes prior to the seizure collected in Contact 1, (c) and (d) in Contact 2, and (e) and (f) in Contact 3. The color of the embedded samples represents the time to seizure onset (blue – 35 minutes prior to the seizure, red – at the seizure onset).

Remarkably, we observe that the embeddings of the observables of Contact 1 and Contact 2 based on the scattering moments follow the gradient of the color. On the other hand, the embedding of the observables of Contact 3 does not show correspondence to the time to seizure. This implies that our *unsupervised* data-driven method reveals a hidden state of the data, which corresponds to a true natural/physical variable that is closely related to the seizure. In this regard, we emphasize that the obtained correspondence between the time to seizure and the embeddings based on Contact 1 and 2, which are located near the seizure onset, and the lack of correspondence based on Contact 3, which is located remotely from the seizure onset, support the latter statement; it is reasonable to assume that the hidden states of the signals from Contact 1 and 2 bear more information on the seizure compared to the hidden state of the signal from Contact 3. We observe that when the contacts are

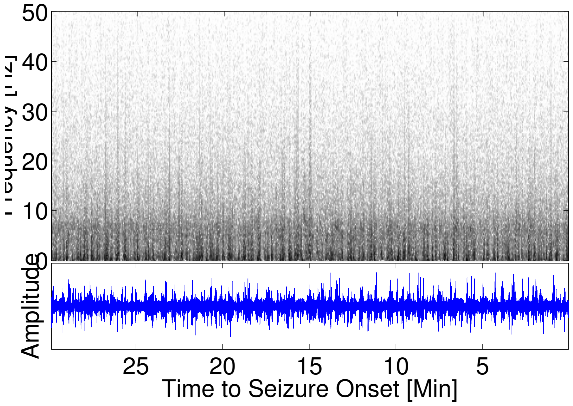
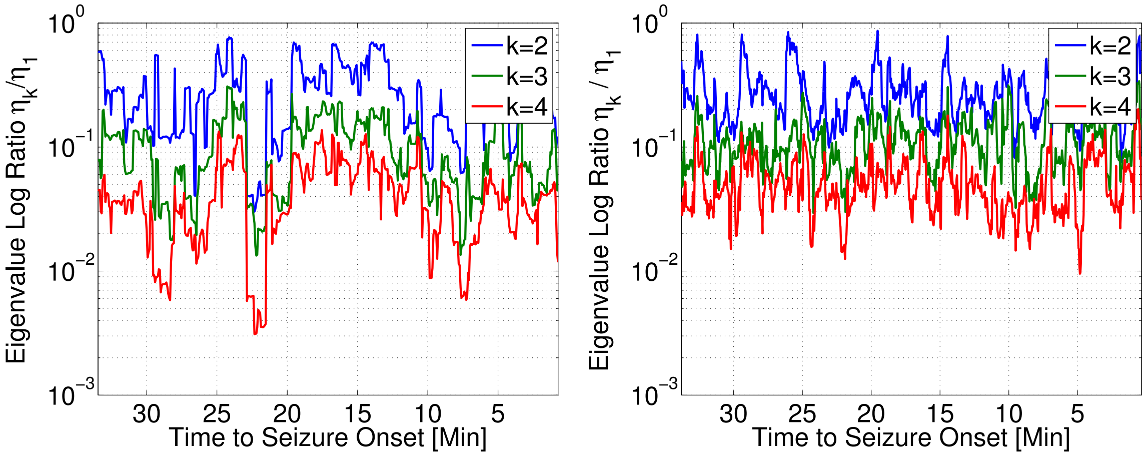


Figure 5: 35 minutes of the recorded signal in Contact 1 that precedes Seizure 1. Top: the signal Fourier power spectrum. Bottom: the signal in time.



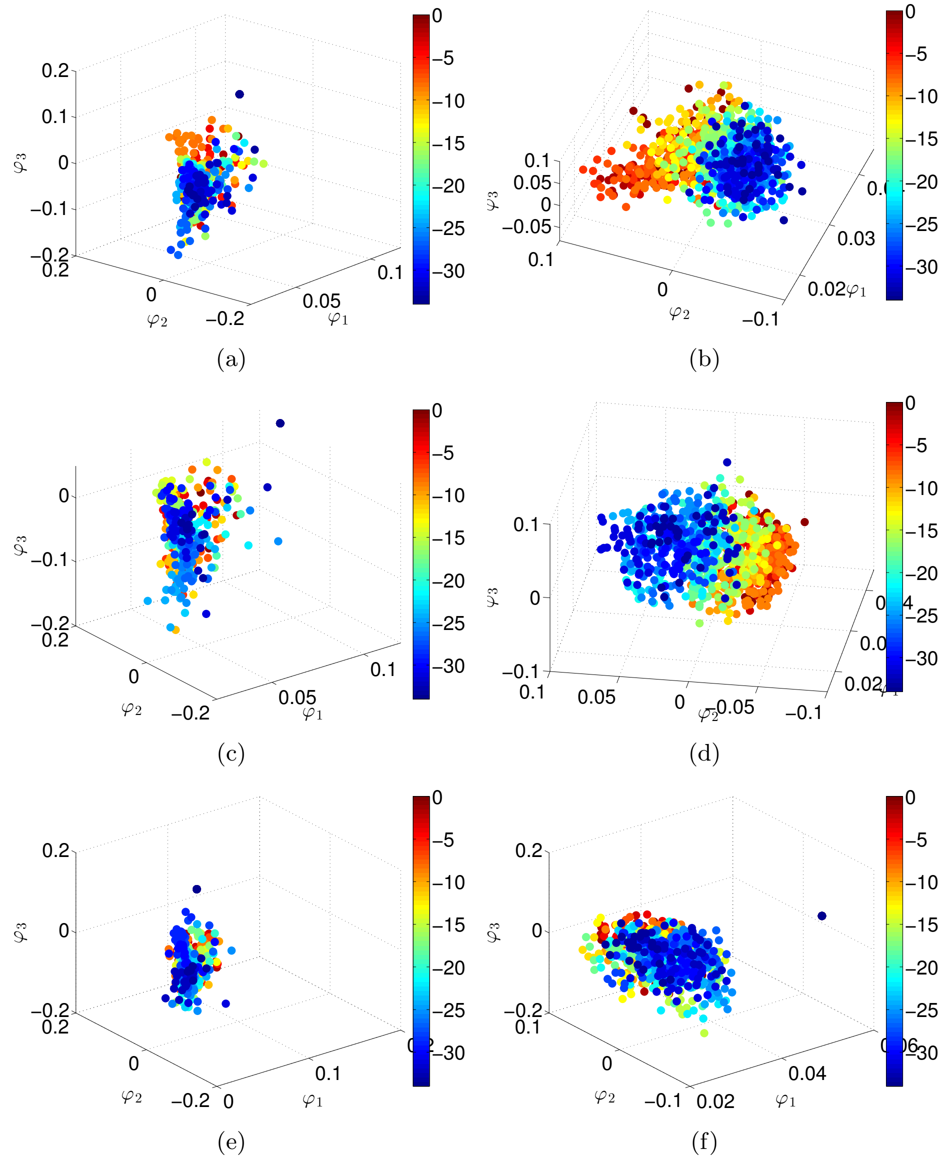
(a)

(b)

Figure 6: The log ratio between the largest and the k-th eigenvalues of the local covariance matrix as a function of time to seizure onset obtained based on (a) the Fourier power spectrum and (b) the scattering transform. The presented ratios are based on the signal recorded in Contact 1 before Seizure 1. The ratios based on the scattering transform are closer to 1 and more stable over time compared to the ratios based on the Fourier power spectrum, thereby implying that the scattering moments are indeed better in terms of observability and stability for this signal.

near the seizure onset, the method picks up the trend, and when it is located remotely from the seizure, the method does not recover it.

In addition, we observe that the embeddings of the observables based on the Fourier power spectrums does not exhibit any trend related to the time to seizure. Thus, the advantage of the scattering moments as observers for these signals over the Fourier power spectrums, as identified by the empirical test in Fig. 6, is respected is embedding results. Without knowing the ground truth in advance, namely, which trend will be recovers and to



(a) (b) (c) (d) (e) (f)

Figure 7: The scatter plots of the 3 dimensional embedding of the observables. Left column: the embedding computed from the Fourier power spectrum. Right column: the embedding computed from the scattering moments. Top row: the embedding of the 35 minutes prior to the seizure collected in Contact 1. Middle row: the embedding of the 35 minutes prior to the seizure collected in Contact 2. Bottom row: the embedding the 35 minutes prior to the seizure collected in Contact 3. The color of the embedded samples represents the time to seizure onset (blue – 35 minutes prior to the seizure, red – just at the seizure onset).

which physical variables it will correspond, we are able to choose the scattering moments over the Fourier power spectrum as observers for this signal.

To further explain the advantage of the scattering moments over the Fourier spectrum,

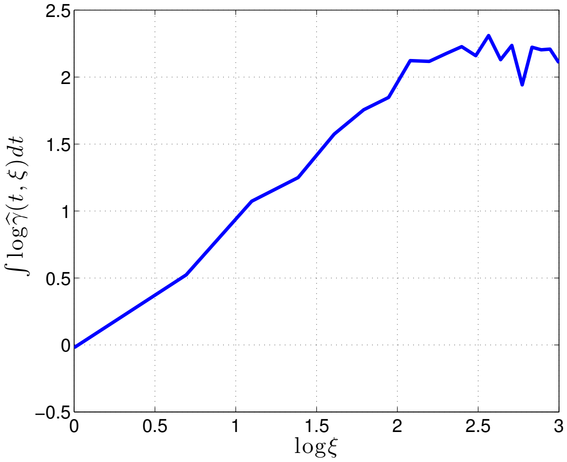


Figure 8: Empirical test for time deformation in the EEG signal. A plot of $\int \log \widehat{\gamma}(t,\xi)dt$ as a function of the frequency logarithm $\log \xi$.

we apply the proposed empirical test for time deformation. Figure 8 presents a plot of $\int \log \widehat{\gamma}(t,\xi)dt$ as a function of the frequency logarithm $\log \xi$. We observe a line whose slope is approximately 1, which suggests according to Section 4.1 the presence of time deformation in the EEG signal. Since the Fourier spectrum is not stable to deformations, it is not an appropriate observer. On the other hand, the scattering moments may be more adequate since they are stable to time deformations.

We note that seizure indication does not necessarily have to be linear in time. However, we used this assumption since there is no ground truth for the seizure indication, especially based on EEG signals.

In order to objectively evaluate the embedding, we apply two simple regression techniques. Several remarks are due at this point. First, we use the time to seizure as a ground truth although, as noted above, it might not be. This assumption helps to evaluate the embedding, however, further research is required. Second, we use standard regression methods to show that the trend is clearly evident in the embedding. If the time to seizure were to be estimated from the embedding, the design of regression techniques which further exploit the dynamics of the signal would have been required in order to obtain optimal performance.

For both regression methods, we randomly select 75% of the samples and use them for training, and then, we test the regression on the rest 25%. The first method is based on k-nearest neighbors (KNN). For each test sample, we find the $k = 5$ nearest training samples in the embedding and estimate the time to seizure at the test sample as a weighted interpolation of the time to seizure of the neighbors, using the Euclidean distance in the embedding as the weight. The second regression method is the Ridge linear regression. In order to account for the fluctuations observed in the embedding, the time to seizure of each sample is estimated as a linear combination of the current and 4 preceding (in time) embedded samples (15 samples in total). These two regression procedures are cross validated over 1000 repetitions.

Figure 9 presents the root mean square error and the estimation standard deviation obtained by the Ridge regression. We observe that the regression results respect the trend

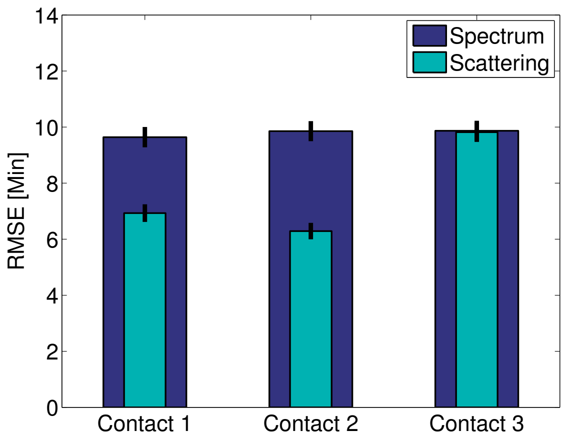


Figure 9: The root mean square error (bars) and the estimation standard deviation (vertical back lines) obtained by the Ridge regression. The regression results respect the trend identified in the embeddings. The embeddings of the observables based on the scattering moments of Contact 1 and Contact 2 indeed show simple correspondence to the time to seizure. On the other hand, the embeddings based on the Fourier power spectrums or based on Contact 3 show almost no correspondence.

identified in the embeddings. The embeddings of the observables based on the scattering moments of Contact 1 and Contact 2 indeed show simple correspondence to the time to seizure. On the other hand, the embeddings based on the Fourier power spectrums or based on Contact 3 show almost no correspondence and yield estimation error close to the degenerate constant estimator (using the mean time to seizure – 17.5 minutes as an estimator). We note that using the KNN regression, slightly inferior results were obtained. In addition, applying the two regression methods directly to the observables (rather than to the embeddings) does not show correspondence to the time to seizure as well.

The results show that there is a prior indication to a seizure in six epilepsy episodes collected from the same subject. To clinically establish our findings, we intend to test our method on multiple subjects. In addition, future work will include extensions to scalp EEG, which are more common and does not require surgery.

6 Conclusions

In this paper, we introduced an unsupervised data-driven method to infer slowly varying intrinsic latent variables of locally stationary signals. From a dynamical systems standpoint, the signals are viewed as the output of an unknown dynamical system, and the latent variables are viewed as the intrinsic state, which drives the system. The primary focus is on the construction of a distance metric between the available observables of the signal, which approximates the Euclidean distance between the corresponding samples of the unknown latent variables. For this construction, both the geometry of the observables and their dynamics are explicitly exploited. In addition, an analysis of the used observers of the signal is given, and an empirical test to evaluate their ability to properly recover the hidden variables is proposed.

The proposed inference method is unsupervised. Thus, unlike supervised regression and classification techniques, it allows for the recovery of intrinsic complex states of dynamical systems, and is not restricted to learning "labels". Indeed, experimental results on real biomedical signals show that the recovered variables have true physiological meaning, implying that some of the natural complexity of the signals was accurately captured.

Acknowledgment

This work is supported by the ERC InvariantClass 320959 grant.

References

[1] B. Litt, R. Esteller, J. Echauz, M. D'Alessandro, R. Shor, T. Henry, and G. Vachtsevanos, "Epileptic seizures may begin hours in advance of clinical onset: a report of five patients," *Neuron*, vol. 30, no. 1, pp. 51–64, 2001.

[2] P. E. McSharry, L. A. Smith, and L. Tarassenko, "Prediction of epileptic seizures: are nonlinear methods relevant?," *Nature medicine*, vol. 9, no. 3, pp. 241–242, 2003.

[3] A. S. Willsky, E. B. Sudderth, M. I. Jordan, and E. B. Fox, "Nonparametric Bayesian learning of switching linear dynamical systems," *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[4] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, *An introduction to variational methods for graphical models*, Springer Netherlands, 1998.

[5] D. M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research (JMLR)*, vol. 3, 2003.

[6] D. M. Blei and J.D. Lafferty, "Dynamic topic models," *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006.

[7] C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. Shawe-Taylor, "Variational inference for diffusion processes.," in *NIPS*, 2007.

[8] M. Opper, A. Ruttor, and G. Sanguinetti, "Approximate inference in continuous time gaussian-jump processes.," in *NIPS*, 2010, pp. 1831–1839.

[9] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *arXiv preprint arXiv:1210.7559*, 2012.

[10] A. Anandkumar, D. P. Foster, D. Hsu, S. Kakade, and Y. Liu, "A spectral algorithm for latent dirichlet allocation.," in *NIPS*, 2012, pp. 926–934.

[11] 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.

[12] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 260, pp. 2323–2326, 2000.

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

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

[15] R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "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.

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

[17] L. Arnold, *Random dynamical systems*, Springer, 1998.

[18] R. Hermann and A. J. Krener, "Nonlinear controllability and observability," *IEEE Transactions on automatic control*, vol. 22, no. 5, pp. 728–740, 1977.

[19] A. Isidori, *Nonlinear control systems*, vol. 1, Springer, 1995.

[20] A. J. Krener and W. Respondek, "Nonlinear observers with linearizable error dynamics," *SIAM Journal on Control and Optimization*, vol. 23, no. 2, pp. 197–216, 1985.

[21] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Computation*, vol. 14, pp. 715–770, 2002.

[22] A. Singer, R. Erban, I. G. Kevrekidis, and R. Coifman, "Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps," *PNAS*, vol. 106, no. 38, pp. 16090–1605, 2009.

[23] S. Mallat, "Group invariant scattering," *Pure and Applied Mathematics*, vol. 10, no. 65, pp. 1331–1398, 2012.

[24] J. Bruna, E. Mallat, S. Bacry, and J. F. Muzy, "Multiscale intermittent process analysis by scattering," *submitted, arXiv:1311.4104*, 2013.

[25] R. Talmon and R.R. Coifman, "Empirical intrinsic geometry for nonlinear modeling and time series filtering," *Proc. Nat. Acad. Sci.*, vol. 110, no. 31, pp. 12535–12540, 2013.

[26] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

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

[28] P. C. Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences (Calcutta)*, vol. 2, pp. 49–55, 1936.

[29] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," *Advances in neural information processing systems*, pp. 521–528, 2003.

[30] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th international conference on Machine learning.* ACM, 2007, pp. 209–216.

[31] R. Talmon and R. R. Coifman, "Empirical intrinsic modeling of signals and information geometry," *submitted, Tech. Report YALEU/DCS/TR1467*, 2012.

[32] M. Hein and J. Y. Audibert, "Intrinsic dimensionality estimation of submanifold in r^d," *L. De Raedt, S. Wrobel (Eds.), Proc. 22nd Int. Conf. Mach. Learn., ACM*, pp. 289–296, 2005.

[33] R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, "Graph laplacian tomography from unknown random projections," *IEEE Trans. Image Process.*, vol. 17, pp. 1891–1899, 2008.

[34] F. R. K. Chung, *Spectral Graph Theory*, CBMS-AMS, 1997.

[35] J. Andén and S. Mallat, "Deep scattering spectrum," *to appear in IEEE transactions of Signal Processing*, 2014.

[36] R. Talmon, I. Cohen, and S. Gannot, "Supervised source localization using diffusion kernels," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'11)*, 2011.

[37] R. Talmon, I. Cohen, and S. Gannot, "Single-channel transient interference suppression using diffusion maps," *to appear in IEEE Trans. Audio, Speech Lang. Process.*, 2012.

[38] R. Schafer and L. Rabiner, "Digital representations of speech signals," *Proc. IEEE*, vol. 63, no. 4, pp. 662–679, Apr. 1975.

[39] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*, Prentice-Hall signal processing series. Prentice Hall, 2005.

[40] C. Hedge, A. Sankaranarayanan, and R. Baraniuk, "Lie operators for compressive sensing," *Proc. 39rd IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-2014, Florence, Italy*, 2014.

[41] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[42] 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.

[43] W. A. Hauser, J. F. Annegers, and W. A. Rocca, “Descriptive epidemiology of epilepsy: contributions of population-based studies from rochester, minnesota,” in *Mayo Clinic Proceedings*. Elsevier, 1996, vol. 71, pp. 576–586.

[44] R. G. Andrzejak, F. Mormann, T. Kreuz, C. Rieke, A. Kraskov, C. E. Elger, and K. Lehnertz, “Testing the null hypothesis of the nonexistence of a preseizure state,” *Physical Review E*, vol. 67, no. 1, pp. 010901, 2003.

[45] M. G. Frei, H. P. Zaveri, S. Arthurs, G. K. Bergey, C. C. Jouny, K. Lehnertz, J. Gotman, I. Osorio, T. I. Netoff, W. J. Freeman, et al., “Controversies in epilepsy: debates held during the fourth international workshop on seizure prediction,” *Epilepsy & Behavior*, vol. 19, no. 1, pp. 4–16, 2010.

[46] D. Duncan, R. Talmon, H. P. Zaveri, and R. R. Coifman, “Identifying preseizure state in intracranial EEG data using diffusion kernels,” *Mathematical Biosciences and Engineering*, vol. 10, no. 3, pp. 579 – 590, 2013.