Multivariate Time-Series Analysis and Diffusion Maps

[1]Wenzhao Lian, [2]Ronen Talmon, [3]Hitten Zaveri, [1]Lawrence Carin and [2]Ronald Coifman

[1]Department of Electrical & Computer Engineering, Duke University
[2]Department of Mathematics, Yale University
[3]School of Medicine, Yale University

***Abstract*—Dimensionality reduction in multivariate time series has broad applications, ranging from financial-data analysis to biomedical research. However, high levels of ambient noise and random interference result in nonstationary signals, which may lead to inefficient performance of conventional methods. In this paper, we propose a nonlinear dimensionality-reduction framework, using diffusion maps on a learned statistical manifold. This yields a low-dimensional representation of the high-dimensional time series. We show that diffusion maps, with affinity kernels based on the Kullback-Leibler divergence between the *local statistics* of samples, allow for efficient approximation of pairwise geodesic distances. To construct the statistical manifold, we estimate time-evolving parametric distributions, by designing a family of Bayesian generative models. The proposed framework can be applied to problems in which the time-evolving distributions (of temporally localized data), instead of the samples themselves, are driven by a low-dimensional underlying process. We provide efficient parameter estimation and a dimensionality reduction methodology, and apply it to two applications: music analysis and epileptic-seizure prediction.**

I. Introduction

In the study of high-dimensional data, it is often of interest to embed the high-dimensional observations in a low-dimensional space, where hidden parameters may be discovered, noise suppressed, and interesting and significant structure revealed. Due to high dimensionality and nonlinearity in many real-word applications, nonlinear dimensionality reduction techniques have been increasingly popular [1], [2], [3]. These manifold-learning algorithms build data-driven models, organizing data samples according to local affinities on a low-dimensional manifold. Such methods have broad applications to, for example, analysis of financial data, computer vision, hyperspectral imaging, and biomedical engineering [4], [5], [6].

The notion of dimensionality reduction is useful in multivariate time series analysis. In the corresponding low-dimensional space, hidden states may be revealed, change points detected, and temporal trajectories visualized [7], [8], [9]. Recently, various nonlinear dimensionality reduction techniques have been extended to time series, including spatio-temporal Isomap [10] and temporal Laplacian eigenmap [11]. In these methods, besides local affinities in the space of the data, available temporal covariate information is incorporated, leading to significant improvements in discovering the latent states of the series.

The basic assumption in dimensionality reduction is that the observed *data samples* do not fill the ambient space uniformly, but rather lie on a low-dimensional manifold. Such an assumption does not hold for many types of signals, for example, data with high levels of noise [4], [12], [13], [14]. In [13] and [14], the authors consider a different, relaxed dimensionality reduction problem on the domain of the underlying probability distributions. The main idea is that the varying *distributions*, rather than the samples themselves, are driven by few underlying controlling processes, yielding a low-dimensional smooth manifold in the domain of the distribution parameters. An information-geometric dimensionality reduction (IGDR) approach is then applied to obtain an embedding of high-dimensional data using Isomap [1], thereby preserving the geodesic distances on the manifold *of distributions*.

Two practical problems arise in these methods, limiting their application. First, in [13], [14] multiple data sets were assumed, with the data in each set drawn from the same distributional form, with set-dependent distribution parameters. The embedding was inferred in the space of the distribution parameters. In this setting a large number of data sets are required, and time dependence in the evolution of the distribution parameters is not considered. By considering time evolution of the distribution from a single time-evolving dataset, we here substantially reduce the amount of needed data, and we extend the method to analysis of time series. A second limitation of previous work concerns how geodesic distances were computed. In [13], [14] the approximatation of the geodesic distance between all pairs of samples was computed using a step-by-step walk on the manifold, requiring $\mathcal{O}(N^3)$ operations, which may be intractable for large-N problems.

In this paper, we present a dimensionality-reduction approach using diffusion maps for nonstationary high-dimensional time series, which addresses the above shortcomings. Diffusion maps constitute an effective data-driven method to uncover the low-dimensional manifold, and provide a parametrization of the underlying process [15]. The main idea in diffusion maps resides in aggregating local connections between samples into a global parameterization, via a kernel. Many kernels implicitly induce a mixture of local statistical models in the domain of the measurements. In particular, it is shown that using distributional information outperforms using sample information when the distributions are available [13].

We exploit this assumption and articulate that the observed multivariate time series $\boldsymbol{X}_t \in \mathbb{R}^N, t = 1, \cdots, T$, is generated from a smoothly varying parametric distribution $p(\boldsymbol{X}_t|\boldsymbol{\beta}_t)$, where $\boldsymbol{\beta}_t$ is a local parameterization of the time evolving distribution. We propose to construct a Bayesian generative model with constraints on $\boldsymbol{\beta}_t$, and use Markov Chain Monte Carlo (MCMC) to estimate $\boldsymbol{\beta}_t$. Diffusion maps are then applied to reveal the statistical manifold (of the estimated distributions), using a kernel with Kullback-Leibler (KL) divergence as the distance metric. Noting that the parametric form of distributions significantly affects the structure of the mapped data, the Bayesian generative model should avoid using a strong informative prior without substantial evidence.

Diffusion maps rely on constructing a Laplace operator, whose eigenvectors approximate the eigenfunctions of the backward Fokker-Planck operator. These eigenfunctions describe the dynamics of the system [16]. Hence, the trajectories embedded in the coordinate system formulated by the principal eigenvectors can be regarded as a representation of the controlling underlying process $\boldsymbol{\theta}_t$ of the time series $\boldsymbol{X}_t$.

One of the main benefits of embedding the time series samples into a low-dimensional domain is the ability to define meaningful distances. In particular, diffusion-map embedding embodies the property that the Euclidean distance between the samples in the embedding domain corresponds to a diffusion distance in the distribution domain. Diffusion distance measures the similarity between two samples according to their connectivity on the low-dimensional manifold [3], and has a close connection to the geodesic distance. Thus, diffusion maps circumvent the step-by-step walk on the manifold [13], computing an approximation to the geodesic distance in a single low-cost operation. Another practical advantage of the proposed method is that we may first reveal the low-dimensional coordinate system based on reference data, and then in an online fashion extend the model for newly acquired data with low computational cost. This is demonstrated further when considering applications in Section IV.

The proposed framework is applied to two applications in which the data are best characterized by temporally evolving local statistics, rather than based on measures directly applied to the data itself: music analysis and epileptic seizure prediction based on electroencephalography (EEG) recordings. In the first application, we show that using the proposed approach, we can uncover the key underlying processes: human voice and instrumental sounds. In particular, we exploit the efficient computation of diffusion distances to obtain intra-piece similarity measures, applied to well-known music, which are compared to state-of-the-art techniques.

In EEG, one goal is to map the recordings to the unknown underlying "brain activity states". This is especially crucial in epileptic seizure prediction, where preseizure (dangerous) states can be distinguished from interictal (safe) states, so that patients can be warned prior to seizures [17]. In this application, the observed time series is the EEG recordings and the underlying process is the brain state: preseizure or interictal. EEG recordings tend to be noisy, and hence, the mapping between the state of the patient's brain and the available measurements is not deterministic, and the measurements do not lie on a smooth manifold. Thus, the intermediate step of mapping the observations to a time-evolving parametric family of distributions is essential to overcome this challenge. We use the proposed approach to infer a parameterization of the signal, viewed as a model summarizing the signal's distributional information. Based on the inferred parameterization, we show that preseizure state intervals can be distinguished from interictal state intervals. In particular, we show the possibility of predicting seizures by visualization and simple detection algorithms, tested on an anonymous patient.

This paper makes three principal contributions. We first present an efficient diffusion map approach based on distributional information of time-series data, which preserves the geodesic distances between samples on a statistical manifold, and uncovers an underlying process that consists of the controlling factors. Second, we propose a class of Bayesian models with various prior specifications, to learn the time-evolving statistics. We finally apply the proposed framework to two applications: music analysis and analysis of EEG recordings.

The remainder of the paper is organized as follows. In Section II we review the diffusion-maps technique, propose an extended construction and examine its theoretical and practical properties. We propose in Section III multiple approaches to model multivariate time series with time-evolving distributions. In Section IV, results on two real-world applications are discussed. Conclusions and future work are outlined in Section V.

II. Diffusion maps using Kullback-Leibler divergence

A. Underlying parametric model

Let $\boldsymbol{X}_t \in \mathbb{R}^N$ be the raw data or extracted features at time t. The key concept is that the high-dimensional representation of $\boldsymbol{X}_t$ exhibits a characteristic geometric structure. This structure is assumed to be governed by an underlying process on a low-dimensional manifold, denoted by $\boldsymbol{\theta}_t$, that propagates over time as a diffusion process according to the following stochastic differential equation (SDE)[1]

$$d\theta_t^i = a_i(\theta_t^i)dt + dw_t^i \tag{1}$$

where θ_t^i is component i of $\boldsymbol{\theta}_t$, a_i are (possibly nonlinear) drift functions and $\boldsymbol{w}_t$ is a Brownian motion. In particular, we assume that the underlying process induces a parametric family of probability distributions in the measurable domain, i.e., $p(\boldsymbol{X}_t|\boldsymbol{\theta}_t)$. In other words, the underlying process controls the statistics of the measured signals.

Note that $\boldsymbol{\theta}_t$ controls the time-evolution of the underlying *distribution* of the data, rather than directly the data itself. We do not assume *a priori* knowledge of the form of the distribution $p(\cdot|\boldsymbol{\theta}_t)$.

[1]In this paper, superscripts represent access to elements in vectors, i.e., x^i is the i-th element of the vector $\boldsymbol{x}$.

B. Local models and the Kullback-Leibler divergence

We use *empirical* densities to represent the local statistics of the signal. In particular, as an intermediate step, we assume a parametric family of local distributions. Let $p(\boldsymbol{X}_t|\boldsymbol{\beta}_t)$ denote the local density of $\boldsymbol{X}_t$. We emphasize that the assumed parameterization of the local distribution $\boldsymbol{\beta}$ is considerably different than $\boldsymbol{\theta}$: $\boldsymbol{\theta}$ is the fundamental parametrization of the low-dimensional manifold, that represents the intrinsic state governing the signal; $\boldsymbol{\beta}$ is merely used within a *chosen* local distribution, employed as an intermediate step, with the goal of inferring $\boldsymbol{\theta}$.

The key thing to note is that because the data are assumed noisy, we do *not* assume the data itself lives on a low-dimensional manifold. Rather, we assume that there is an underlying and unknown distribution $p(\boldsymbol{X}_t|\boldsymbol{\theta}_t)$, that evolves with time, and that is responsible for the data. To uncover the time evolution of $\boldsymbol{\theta}$ (and to infer the dimension of the parameter vector $\boldsymbol{\theta}$), we *assume* a form of a generally different distribution $p(\boldsymbol{X}_t|\boldsymbol{\beta}_t)$, typically selected to balance accuracy with computational simplicity. We then compute distances between data at time t and t', based on $p(\boldsymbol{X}_t|\boldsymbol{\beta}_t)$ and $p(\boldsymbol{X}_{t'}|\boldsymbol{\beta}_{t'})$, using an appropriate kernel. From the resulting distance matrix we seek to uncover $\boldsymbol{\theta}_t$ for all t. In the end we still do not know the responsible distribution $p(\boldsymbol{X}_t|\boldsymbol{\theta}_t)$, but we uncover how the (typically low-dimensional) parameters $\boldsymbol{\theta}_t$ evolve with time, manifesting a useful embedding. In [13], [14] the authors also estimated distributions $p(\boldsymbol{X}_n|\boldsymbol{\beta}_n)$ for multilple *data sets*, here with $\boldsymbol{X}_n$ representing all the data in dataset $n \in \{1,\dots,N\}$; they estimate the associated $\{\boldsymbol{\theta}_n\}$ via a similar embedding procedure. By leveraging time, we effectively infer N *local* datasets, characterized by time-evolving distributions.

We propose to use the Kullback-Leibler (KL) divergence as a metric between the parametric probability density functions (pdfs). For any pair of measurements $\boldsymbol{X}_t$ and $\boldsymbol{X}_{t'}$, the KL divergence between the corresponding parametric pdfs is defined as

$$\mathcal{D}(p(\boldsymbol{X}|\boldsymbol{\beta}_t)||p(\boldsymbol{X}|\boldsymbol{\beta}_{t'})) = \int_{\boldsymbol{X}} \ln\left(\frac{p(\boldsymbol{X}|\boldsymbol{\beta}_t)}{p(\boldsymbol{X}|\boldsymbol{\beta}_{t'})}\right) p(\boldsymbol{X}|\boldsymbol{\beta}_t) d\boldsymbol{X}. \quad (2)$$

Let $\boldsymbol{\beta}_{t_0}$ and $\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t_0} + \delta\boldsymbol{\beta}_t$ be two close samples in the intermediate parametric domain. It can be shown [18] that the KL divergence is locally approximated by the Fisher information metric, i.e.,

$$\mathcal{D}(p(\boldsymbol{X}|\boldsymbol{\beta}_t)||p(\boldsymbol{X}|\boldsymbol{\beta}_{t_0})) \simeq \delta\boldsymbol{\beta}_t^T \boldsymbol{I}(\boldsymbol{\beta}_{t_0})\delta\boldsymbol{\beta}_t \quad (3)$$

where $\boldsymbol{I}(\boldsymbol{\beta}_{t_0})$ is the Fisher information matrix.

We then define the Riemannian manifold $(\mathcal{M}, \boldsymbol{g})$, where the Fisher metric in (3) is associated with the inner product on the manifold tangent plane $\boldsymbol{g}$ between the local distributions,

$$g_{ij}(\boldsymbol{\beta}_t) = \sum_{i,j} \int \frac{\partial \log p(\boldsymbol{X}|\boldsymbol{\beta}_t)}{\partial \beta_t^i} \frac{\partial \log p(\boldsymbol{X}|\boldsymbol{\beta}_t)}{\partial \beta_t^j} p(\boldsymbol{X}|\boldsymbol{\beta}_t) d\boldsymbol{X} \quad (4)$$

The points residing on $\mathcal{M}$ are parametric probability density functions $p(\boldsymbol{X}|\boldsymbol{\beta}_t)$. Thus, for $p(\boldsymbol{X}|\boldsymbol{\beta}_{t_0+\delta t})$ in a local neighborhood of $p(\boldsymbol{X}|\boldsymbol{\beta}_{t_0})$, the Fisher metric between these two points can be approximated by $\mathcal{D}(p(\boldsymbol{X}|\boldsymbol{\beta}_t)||p(\boldsymbol{X}|\boldsymbol{\beta}_{t_0}))$. Therefore, we use the KL divergence to construct the affinity kernel and build the graph for diffusion maps, thus obtaining diffusion distance approximating a Riemannian metric on the manifold of local distributions. This will be addressed in detail in Section II-C.

For the signal types reported in this paper (music and EEG), we have empirically found that a simple local Gaussian model (Gaussian mixture model) with zero mean and time evolving covariance matrices can effectively describe the local empirical densities of the selected feature sets of the signals. Thus, the intermediate parameterization $\boldsymbol{\beta}_t$ is a local covariance matrix $\boldsymbol{\Sigma}_t$, and the *local* distribution of $\boldsymbol{X}_t$ is approximated by $\mathcal{N}(0, \boldsymbol{\Sigma}_t)$. In this case, the KL divergence can be explicitly written as

$$\mathcal{D}(p(\boldsymbol{X}|\boldsymbol{\beta}_t)||p(\boldsymbol{X}|\boldsymbol{\beta}_{t'})) = \frac{1}{2}\mathrm{Tr}(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\Sigma}_{t'} - \boldsymbol{I}_N). \quad (5)$$

Based on the KL divergence, we define a symmetric pairwise affinity function using a Gaussian kernel

$$k(\boldsymbol{X}_t, \boldsymbol{X}_{t'}) = \exp\left\{-\frac{\mathcal{D}(p(\boldsymbol{X}|\boldsymbol{\beta}_t)||p(\boldsymbol{X}|\boldsymbol{\beta}_{t'})) + \mathcal{D}(p(\boldsymbol{X}|\boldsymbol{\beta}_{t'})||p(\boldsymbol{X}|\boldsymbol{\beta}_t))}{\varepsilon}\right\} \quad (6)$$

The decay rate of the exponential kernel implies that only pdfs $p(\boldsymbol{X}|\boldsymbol{\beta}_t)$ within an ε-neighborhood of $p(\boldsymbol{X}|\boldsymbol{\beta}_{t'})$ are taken into account and have non negligible affinity. Thus, we can use the approximation of the KL divergence using the Fisher metric (3) and obtain that

$$k(\boldsymbol{X}_t, \boldsymbol{X}_{t'}) \simeq \exp\left\{-\frac{(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t'})^T(\boldsymbol{I}(\boldsymbol{\beta}_t) + \boldsymbol{I}(\boldsymbol{\beta}_{t'}))(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t'})}{\varepsilon}\right\} \quad (7)$$

C. Laplace operator and diffusion maps

Let $\boldsymbol{W}$ be a pairwise affinity matrix between the set of measurements $\boldsymbol{X}_t$, whose (t, t')-th element is given by

$$W_{t,t'} = k(\boldsymbol{X}_t, \boldsymbol{X}_{t'}). \quad (8)$$

Based on the kernel, we form a weighted graph, where the measurements $\boldsymbol{X}_t$ are the graph nodes and the weight of the edge connecting node $\boldsymbol{X}_t$ to node $\boldsymbol{X}_{t'}$ is $W_{t,t'}$. The particular choice of the Gaussian kernel exhibits a notion of locality by defining a neighborhood around each measurement $\boldsymbol{X}_t$ of radius ε, i.e., measurements $\boldsymbol{X}_{t'}$ such that $(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t'})^T(\boldsymbol{I}(\boldsymbol{\beta}_t) + \boldsymbol{I}(\boldsymbol{\beta}_{t'}))(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t'}) > \varepsilon$ are weakly connected to $\boldsymbol{X}_t$. In practice, we set ε to be the median of the elements of the kernel matrix. According to the graph interpretation, such a scale results in a well connected graph because each measurement is effectively connected to half of the other measurements. For more details, see [19], [20].

Using the KL divergence (the Fisher information metric) as an affinity measure has the effect of an adaptive scale. Consider the parametric family of normal distributions. In particular, assume that the process $\boldsymbol{X}_t$ is one dimensional and is given by

$$X_t = Y_t + V_t \quad (9)$$

where $Y_t \sim \mathcal{N}(0, \theta_t)$ and V_t is an adaptive white Gaussian noise with zero mean and fixed σ_v^2 variance. According to the parametric model (Section II-A), θ_t follows a diffusion process propagation model which results in time varying distributions.

Consequently, the parametric pdf of the measurements is given by

$$p(X|\beta_t) = \mathcal{N}(0, \beta_t) \tag{10}$$

where $\beta_t = \theta_t + \sigma_v^2$, and the corresponding Fisher Information matrix is

$$\boldsymbol{I}(\beta_t) = \frac{1}{2\beta_t^2}. \tag{11}$$

In this case, the corresponding kernel based on the KL divergence is

$$k(X_t, X_{t'}) = \exp\left\{-\frac{\|\theta_t - \theta_{t'}\|^2}{\varepsilon(\beta_t, \beta_{t'})}\right\} \tag{12}$$

where

$$\varepsilon(\beta_t, \beta_{t'}) = \frac{\varepsilon}{2}\left(\frac{1}{(\theta_t + \sigma_v^2)^2} + \frac{1}{(\theta_{t'} + \sigma_v^2)^2}\right)^{-1} \tag{13}$$

is a locally adapted kernel scale with the following interpretation: when the noise rate σ_v^2 increases, a larger scale (neighborhood) is used in order to see "beyond the noise" and capture the geometry and variability of the underlying parameter θ. We remark that in this specific case, the adaptive scale is the local covariance of the measurements. Thus, this metric is equal to the Mahalanobis distance between the measurements [9].

Let $\boldsymbol{D}$ be a diagonal matrix whose elements are the sums of rows of $\boldsymbol{W}$, and let $\boldsymbol{W}^{\text{norm}} = \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ be a normalized kernel that shares its eigenvectors with the normalized graph-Laplacian $\boldsymbol{I} - \boldsymbol{W}^{\text{norm}}$ [21]. It was shown [3] that $\boldsymbol{W}^{\text{norm}}$ converges to a diffusion operator that reveals the low-dimensional manifold and a subset of its eigenvectors give a parameterization of the underlying process. We assume that these eigenvectors are the principal eigenvectors associated with the largest eigenvalues, although there is no guarantee. Thus, the eigenvectors of $\boldsymbol{W}^{\text{norm}}$, denoted by $\widetilde{\boldsymbol{\varphi}}_j$, reveal the underlying structure of the data. Specifically, the t-th coordinate of the j-th eigenvector can be associated with the j-th coordinate of the underlying process $\boldsymbol{\theta}_t$ of measurement $\boldsymbol{X}_t$. The eigenvectors are ordered such that $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{T-1} > 0$, where λ_j is the eigenvalue associated with eigenvector $\widetilde{\boldsymbol{\varphi}}_j$. Because $\boldsymbol{W}^{\text{norm}} \sim \boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{W}$, and $\boldsymbol{D}^{-1}\boldsymbol{W}$ is row-stochastic, $\lambda_0 = 1$ and $\widetilde{\boldsymbol{\varphi}}_0$ is the diagonal of $\boldsymbol{D}^{1/2}$. In addition, $\boldsymbol{W}^{\text{norm}}$ is positive semidefinite, and hence, the eigenvalues are positive. The matrix $\boldsymbol{P}$ may be interpreted as a transition matrix of a Markov chain on the graph nodes. Specifically, the states of the Markov chain are the graph nodes and $P_{t,t'}$ represents the probability of transition in a single Markov step from node $\boldsymbol{X}_t$ to node $\boldsymbol{X}_{t'}$. Propagating the Markov chain n steps forward corresponds to raising $\boldsymbol{P}$ to the power of n. We also denote the probability function from node $\boldsymbol{X}_t$ to node $\boldsymbol{X}_{t'}$ in n steps by $p_n(\boldsymbol{X}_t, \boldsymbol{X}_{t'})$.

The eigenvectors are used to obtain a new data-driven description of the measurements $\boldsymbol{X}_t$ via a family of mappings that are called diffusion maps [3]. Let $\boldsymbol{\Psi}_{\ell,n}(\boldsymbol{X}_t)$ be the diffusion mappings of the measurements into the Euclidean space $\mathbb{R}^\ell$, defined as

$$\boldsymbol{\Psi}_{\ell,n}(\boldsymbol{X}_t) = \left(\lambda_1^n\widetilde{\varphi}_1^t, \lambda_2^n\widetilde{\varphi}_2^t, \ldots, \lambda_\ell^n\widetilde{\varphi}_\ell^t\right)^T \tag{14}$$

where ℓ is the new space dimensionality ranging between 1 and $T-1$. Diffusion maps have therefore two parameters: n and ℓ. Parameter n corresponds to the number of steps of the Markov process on the graph, since the transition matrices $\boldsymbol{P}$ and $\boldsymbol{P}^n$ share the same eigenvectors, and the eigenvalues of $\boldsymbol{P}^n$ are the eigenvalues of $\boldsymbol{P}$ raised to the power of n. Parameter ℓ indicates the intrinsic dimensionality of the data. The dimension may be set heuristically according to the decay rate of the eigenvalues, as the coordinates in (14) become negligible for a large ℓ. In practice, we expect to see a distinct "spectral gap" in the decay of the eigenvalues. Such a gap is often a good indicator of the intrinsic dimensionality of the data and its use is a common practice in spectral clustering methods. The mapping of the data $\boldsymbol{X}_t$ into the low-dimensional space using (14) provides a parameterization of the underlying manifold and its coordinates represent the underlying processes $\boldsymbol{\theta}_t$ (1). We note that as n increases, the decay rate of the eigenvalues also increases (they are confined in the interval $[0, 1]$). As a result, we may set ℓ to be smaller, enabling to capture the underlying structure of the measurements in fewer dimensions. Thus, we may claim that a larger number of steps usually brings the measurements closer in the sense of the affinity implied by $\boldsymbol{P}^n$, and therefore, a more "global" structure of the signal is revealed.

The Markov process aggregates information from the entire set into the affinity metric $p_n(\boldsymbol{X}_t, \boldsymbol{X}_{t'})$, defining the probability of "walking" from node $\boldsymbol{X}_t$ to $\boldsymbol{X}_{t'}$ in n steps. For any n, the following metric

$$D_n^2(\boldsymbol{X}_t, \boldsymbol{X}_{t'}) = \int_{\boldsymbol{X}_s} [p(\boldsymbol{X}_t, \boldsymbol{X}_s) - p_n(\boldsymbol{X}_{t'}, \boldsymbol{X}_s)]^2 w(\boldsymbol{X}_s) d\boldsymbol{X}_s \tag{15}$$

is called *diffusion distance*, with $w(\boldsymbol{X}_s) = 1/\widetilde{\boldsymbol{\varphi}}_0(\boldsymbol{X}_s)$. It describes the relationship between pairs of measurements in terms of their graph connectivity, and as a consequence, local structures and rules of transitions are integrated into a global metric. If the integral is evaluated on the points of the observed data, it can be shown that the diffusion distance (15) is equal to the Euclidean distance in the diffusion maps space when using all $\ell = T - 1$ eigenvectors [3], i.e.,

$$D_n(\boldsymbol{X}_t, \boldsymbol{X}_{t'}) = \|\boldsymbol{\Psi}_{\ell,n}(\boldsymbol{X}_t) - \boldsymbol{\Psi}_{\ell,n}(\boldsymbol{X}_{t'})\|_2 \tag{16}$$

Thus, comparing between the diffusion mappings using the Euclidean distance conveys the advantages of the diffusion distance stated above. In addition, since the eigenvalues tend to decay fast, for large enough n, the diffusion distance can be well approximated by only the first few eigenvectors, setting $\ell << T - 1$.

D. Sequential implementation

The construction of the diffusion maps embedding is computationally expensive due to the application of the eigenvector decomposition (EVD). In practice, the measurements are not always available in advance. Thus, the computationally demanding procedure should be applied repeatedly whenever a new set of measurements become available. In this section, we describe a sequential procedure for extending the embedding, which circumvents the EVD applications and may be suitable for supervised techniques [22], [23], [24].

Let $\boldsymbol{X}_t$ be a sequence of T reference measurements that are assumed to be available in advance. The availability of these measurements enables one to estimate the local densities and the corresponding kernel based on the KL divergence. Then, the embedding of the reference measurements can be computed.

Let $\boldsymbol{X}_s$ be a new sequence of S measurements, which are assumed to become available sequentially. As proposed in [23], [24], we define a nonsymmetric pairwise metric between any new measurement $\boldsymbol{X}_s$ and any reference measurement $\boldsymbol{X}_t$, similarly to (7) as

$$a(\boldsymbol{X}_s, \boldsymbol{X}_t) = \exp\left\{-\frac{(\boldsymbol{\beta}_s - \boldsymbol{\beta}_t)^T \boldsymbol{I}(\boldsymbol{\beta}_t)(\boldsymbol{\beta}_s - \boldsymbol{\beta}_t)}{\varepsilon}\right\} \quad (17)$$

where $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}_s$ are the parametrization of the local densities of the measurements at t and s, respectively, and a corresponding nonsymmetric kernel

$$A_{s,t} = a(\boldsymbol{X}_s, \boldsymbol{X}_t). \quad (18)$$

The construction of the nonsymmetric kernel requires the feature vectors of the measurements and the Fisher Information matrix of merely the reference measurements.

Let $\widetilde{\boldsymbol{A}} = \boldsymbol{D}_A^{-1}\boldsymbol{A}\boldsymbol{Q}^{-1}$, where $\boldsymbol{D}_A$ is a diagonal matrix whose diagonal elements are the sums of rows of $\boldsymbol{A}$, and $\boldsymbol{Q}$ is a diagonal matrix whose diagonal elements are the sums of rows of $\boldsymbol{D}_A^{-1}\boldsymbol{A}$. It was shown by [22], [23] that

$$\boldsymbol{W} = \widetilde{\boldsymbol{A}}^T \widetilde{\boldsymbol{A}} \quad (19)$$

where $\boldsymbol{W}$ is the pairwise affinity matrix on the T reference measurements $\boldsymbol{X}_t$ as defined in Section II-C.

We define now the dual extended $S \times S$ kernel between the new samples as $\boldsymbol{W}^{\text{ext}} = \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{A}}^T$. It is shown in [24] that the elements of the extended kernel are proportional to a Gaussian defined similarly to (7) between a pair of new measurements. Combining the relationship between the kernels $\boldsymbol{W}$ and $\boldsymbol{W}^{\text{ext}}$ yields: (1) the kernels share the same eigenvalues λ_j; (2) the eigenvectors $\boldsymbol{\varphi}_j$ of $\boldsymbol{W}$ are the right singular vectors of $\widetilde{\boldsymbol{A}}$; (3) the eigenvectors $\boldsymbol{\psi}_j$ of $\boldsymbol{W}^{\text{ext}}$ are the left singular vectors of $\widetilde{\boldsymbol{A}}$. As discussed in Section II-C, the right singular vectors represent the underlying process of the reference measurements, and by [22], [23], the left singular vectors naturally extend this representation to the new measurements. In addition, the relationship between the eigenvectors of the two kernels is given by the singular value decomposition (SVD) of $\widetilde{\boldsymbol{A}}$ and is explicitly expressed by

$$\boldsymbol{\psi}_j = \frac{1}{\sqrt{\lambda_j}}\widetilde{\boldsymbol{A}}\boldsymbol{\varphi}_j. \quad (20)$$

Now, the extended eigenvectors $\boldsymbol{\psi}_j$ can be used instead of $\widetilde{\boldsymbol{\varphi}}_j$ to construct the embedding of the new measurements $\boldsymbol{\Psi}_{\ell,n}(\boldsymbol{X}_s)$ in (14).

III. Modeling time evolving covariance matrices

To calculate the KL divergence, we need to estimate the local/intermediate parametric distribution $p(\boldsymbol{X}_t|\beta_t)$ at each time. The amount of data in each time window is limited, and therefore, assumptions have to be made to constrain the parameteric space. In this paper, we assume that the signal sample at each time is drawn from a multivariate Gaussian distribution with time evolving parameters. For simplicity, we focus on zero mean Gaussian distributions. We assume that the time evolving covariance matrices characterize the dynamics of the time series. Such a time evolving covariance model can be applied to many multivariate time series, including volatility analysis in finance [4] and EEG activity in neurology [5]. Popular approaches for estimating smoothly varying covariance matrices include the exponentially weighted moving average (EWMA) model [25] and multivariate generalized autoregressive conditional heteroscedasticity (GARCH) models [26]. The former captures the smoothly varying trends, but fails to handle missing data, and requires long series to achieve high estimation accuracy [27]. The latter handles missing data at the expense of over-restricting the flexibility of the dynamics of the covariance matrices [4]. Most moving average type approaches can be simplified as $\hat{\boldsymbol{\Sigma}}_t = \pi_t\hat{\boldsymbol{\Sigma}}_{t-1} + (1-\pi_t)\bar{\boldsymbol{\Sigma}}_t$, where $\bar{\boldsymbol{\Sigma}}_t$ is the covariance matrix of the sample at t and π_t is a smoothing parameter. $\hat{\boldsymbol{\Sigma}}_t$ can be considered as the posterior mean estimate for $\boldsymbol{\Sigma}_t$ using an inverse-Wishart prior with mean proportional to $\hat{\boldsymbol{\Sigma}}_{t-1}^{-1}$. Given their broad use in practice, moving average type approaches are also tested in our experiments.

In this paper, we present a latent variable model to infer the time evolving covariance matrices, inspired by the idea proposed in [12]. This generative model allows for handling missing data and aims to capture the dynamics of the evolving structure of the covariance matrices. We assume at each time $t, t = 1, \cdots, T$, a factor analyzer (FA) model fits the observed data sample $\boldsymbol{X}_t \in \mathbb{R}^N$:

$$\begin{aligned} \boldsymbol{X}_t &\sim \mathcal{N}(\boldsymbol{F}^t\boldsymbol{S}_t, \alpha^{-1}\boldsymbol{I}_N) && (21)\\ \boldsymbol{S}_t &\sim \mathcal{N}(0, \boldsymbol{I}_K) && (22)\\ \alpha &\sim \text{Ga}(e_0, f_0) && (23) \end{aligned}$$

where $\boldsymbol{F}^t \in \mathbb{R}^{N\times K}$ formulates a time evolving factor loading matrix, with a prior that will be described next. $\boldsymbol{S}_t$ is a K dimensional variable in the latent space, and α models the noise level in the observation space. This model constrains the high-dimensional data to locally reside in a K dimensional space, but does not assume local stationarity due to the time evolving factor-loading matrix. Thus, at time t, the marginal distribution of $\boldsymbol{X}_t$ is

$$\begin{aligned} \boldsymbol{X}_t &\sim \mathcal{N}(\boldsymbol{0}_N, \boldsymbol{\Sigma}_t) && (24)\\ \boldsymbol{\Sigma}_t &= (\boldsymbol{F}^t)(\boldsymbol{F}^t)^T + \alpha^{-1}\boldsymbol{I}_N && (25) \end{aligned}$$

Therefore, we could use posterior estimates of $\boldsymbol{F}^t$ and α to estimate $\boldsymbol{\Sigma}_t$. Because we are only interested in $(\boldsymbol{F}^t)(\boldsymbol{F}^t)^T$, instead of $\boldsymbol{F}^t$ or $\boldsymbol{S}_t$, this latent variable model circumvents the identifiability problem encountered in common FA models [28]. To impose smoothness and facilitate parameter estimation, two types of priors for $\boldsymbol{F}^t$ are proposed: a Gaussian process (GP) [29] and non-stationary autoregressive (AR) process [7].

In the GP model, elements of $\boldsymbol{F}^t$ are constructed as

$$\begin{aligned} \boldsymbol{F}_{ij} &\sim \mathcal{GP}(0, \boldsymbol{\Omega}^{ij}) && (26)\\ \boldsymbol{\Omega}^{ij}_{t_1,t_2} &= \sigma^{-1}(k(t_1,t_2) + \sigma_n \delta(t_1 - t_2)) && (27)\\ \sigma &\sim \text{Ga}(c_0, d_0) && (28) \end{aligned}$$

where σ^{-1} represents the variance of the factor loadings over time. To infer σ^{-1} from the data, a broad gamma distribution prior is proposed. The hyperparameter σ_n represents the noise variance for the factor loadings, which is kept fixed at a small value (10^{-3} in our case). Each time varying factor loading $\boldsymbol{F}^t_{ij}, t = 1, \cdots, T$, is constructed from a GP. Various kernel functions $k(t_1, t_2)$ can be used for the GP, including the radius basis function (RBF) $k(t_1, t_2) = \exp(-\frac{(t_1-t_2)^2}{2\tau^2})$. We have tested different kernel functions and RBF is chosen to allow for simple interpretation in our experiments. τ is the length-scale parameter, which determines globally the shape of autocorrelation function [12]. This facilitates strong correlation between $\boldsymbol{F}^{t_1}_{ij}$ and $\boldsymbol{F}^{t_2}_{ij}$ if $|t_1 - t_2| < \tau$ and inhibits the correlation otherwise. τ can be estimated from the data by putting a discrete uniform prior over a library of candidate atoms [30]. However, in practice, the correlation length might be available *a priori* and used as the appropriate value, which often works effectively. Standard MCMC sampling can be used to infer model parameters, as summarized in the Appendix. Sampling $\boldsymbol{F}_{ij}$ from the GP has a $\mathcal{O}(T^3)$ complexity, which can be alleviated using the random projection idea in [30].

In the non-stationary AR process prior model, elements of $\boldsymbol{F}^t$ are constructed as

$$\begin{aligned} \boldsymbol{F}^t_{ij} &= \boldsymbol{F}^{t-1}_{ij} + \boldsymbol{\xi}^t_{ij} && (29)\\ \boldsymbol{\xi}^t_{ij} &\sim \mathcal{N}(0, \boldsymbol{\eta}^{-1}_{ij}) && (30)\\ \boldsymbol{\eta}_{ij} &\sim \text{Ga}(g_0, h_0) && (31) \end{aligned}$$

The time varying factor loading matrix $\boldsymbol{F}^t, t = 1, \cdots, T$, is a random walk whose smoothness is determined by $\boldsymbol{\xi}^t$. A shrinkage prior [31] favoring $\boldsymbol{\xi}^t$ to be sparse is added to encourage smoothness. Other kinds of sparseness-promoting priors, including spike-slab [32] and generalized double Pareto [33], can also be considered. The zero mean distribution for $\boldsymbol{\xi}_{ij}$ models the difference between consecutive covariance matrices as a stationary process, which captures the drift of factor loadings over time. In this model, the trend of each factor loading is assumed independent. However, correlated trends of $\boldsymbol{F}^t$ and group sparsity of $\boldsymbol{\xi}^t$ can be considered for more sophisticated data, as a future direction. A forward filtering backward sampling (FFBS) [34] method is used to sample $\boldsymbol{F}^t$ (summarized in the Appendix).

Choosing which estimation method to use is data specific and depends on the available prior knowledge. If the covariance structures are highly correlated for nearby samples, while they are highly uncorrelated for faraway samples, the model with GP prior should be adopted. If the difference between consecutive covariance matrices is approximately a stationary process, the model with non-stationary AR process prior should be considered. If the covariance matrices are evolving smoothly over time, local stationarity approximately holds, and a large number of data samples are provided, moving average type approaches can be easily implemented. This is considered in Section IV. The generalized framework can be summarized in Algorithm 1.

Algorithm 1 Diffusion maps using time evolving statistics

Input: Observations $\boldsymbol{X}_t \in \mathbb{R}^N, t = 1, \cdots, T$, diffusion step n, neighbourhood size ϵ, embedding dimension l

Output: Embeddings $\{\boldsymbol{\Psi}_{l,n}(\boldsymbol{X}_t), t = 1, \cdots, T\}$

1: At each time t, estimate a distribution $p(\boldsymbol{X}_t|\boldsymbol{\beta}_t)$ (using either moving average (MA), Bayesian generative model with AR process prior (BAR), or Gaussian Process prior (BGP) to infer $\boldsymbol{\beta}_t$)

2: Compute the $T \times T$ matrix $\boldsymbol{W}$, where $W_{t_1,t_2} = \exp\left(-\frac{\mathcal{D}(p(\boldsymbol{X}|\boldsymbol{\beta}_{t_1})||p(\boldsymbol{X}|\boldsymbol{\beta}_{t_2})) + \mathcal{D}(p(\boldsymbol{X}|\boldsymbol{\beta}_{t_2})||p(\boldsymbol{X}|\boldsymbol{\beta}_{t_1}))}{\epsilon}\right)$

3: Set $\boldsymbol{D} = \text{diag}\left\{\sum_{\tau=1}^{T} W_{t,\tau}\right\}$ and compute the normalized kernel $\boldsymbol{W}^{\text{norm}} = \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$

4: Keep the top ℓ non-trivial eigenvalues λ_j and eigenvectors $\tilde{\boldsymbol{\varphi}}_j \in \mathbb{R}^T$ of $\boldsymbol{W}^{\text{norm}}$, $j = 1, \cdots, l$, and construct the corresponding diffusion maps embeding $\boldsymbol{\Psi}_{\ell,n}(\boldsymbol{X}_t) = (\lambda_1^n \tilde{\boldsymbol{\varphi}}^t_1, \lambda_2^n \tilde{\boldsymbol{\varphi}}^t_2, \ldots, \lambda_\ell^n \tilde{\boldsymbol{\varphi}}^t_\ell)^T$

IV. Applications

The proposed framework is applied to a toy example and two real-world applications. In the synthetic example, we show that the estimated diffusion distance between data points recovers the geodesic distance on the statistical manifold, where IGDR [13] is used as a baseline method. In the first real-world application, we analyze a well-know music piece, where we estimate the diffusion distance between time points to discover the intra-piece similarities as a function of time. In the second application, the proposed framework is used to discover the different brain states of an epileptic patient based on EEG recordings.

A. Synthetic Experiment

We assume that the data $\boldsymbol{X}_t$ is generated by a zero-mean multivariate Gaussian distribution, with time-evolving covariance matrix $\boldsymbol{\Sigma}_t$, constructed from a GP prior as defined in (24)-(27). We consider observation length $T = 500$, dimension $N = 5$, local latent dimension $K = 3$, and GP length-scale parameter $\tau = 0.02$. The goal is to recover the geodesic distance between data points on the statistical manifold, defined by their corresponding covariance matrices. Because the pairwise symmetric KL distance matrix is needed in both the proposed framework and IGDR, we let both methods know the true $\boldsymbol{\Sigma}_t$ therefore we can focus on the comparison between the two dimensionality reduction schemes. In other words, in this experiment we do not estimate $\boldsymbol{\Sigma}_t$ from the data, but simply assume it is known. The purpose of this test is to compare the diffusion-distance method with IGDR, on the same time-evolving density function.

We apply diffusion maps with the pairwise affinity kernel defined in (6). We consider $n = 200$ and obtain the

low-dimensional embeddings of $\boldsymbol{X}_t$ as defined in (14). The pairwise geodesic distance between data points can be approximated by the Euclidean distances between the corresponding embeddings, as shown in Fig. 1(a). On the other hand, using IGDR, a shortest-path algorithm is executed to find approximate pairwise geodesic distances, followed by classical multidimensional scaling (MDS) methods ($i.e.$, Laplacian eigenmaps [2]) for dimensionality reduction. The approximated pairwise geodesic distances are presented in Fig. 1(b) and used as a comparison. As illustrated in Fig. 1, both methods yield similar distances. However, the running time of the proposed method is a couple of seconds, whereas the running time of IGDR is approximately 400 seconds, because of the $\mathcal{O}(T^3)$ complexity required to compute the pairwise shortest-path distances. These computations were performed on a computer with 2.2GHz CPU and 8GB RAM, with all software written in Matlab.

B. Music analysis

In music information retrieval, automatic genre classification and composer identification are of increasing interest. Thus, one goal is to compute similarities between short intervals at different times of a music piece [35]. For comparison with a previous method, we test our framework on the same music piece used in [35], "A Day in the Life" from the Beatles' album Sgt. Pepper's Lonely Hearts Club Band. The song is 5 minutes 33 seconds long and sampled at 22.05 KHz. The music piece is divided into 500ms contiguous frames to obtain Mel Frequency Cepstral Coefficients (MFCCs), which leads to 667 available frames overall. As depicted in Fig. 2 (a), 40 normalized ($i.e.$, with zero mean) MFCCs are used as features, yielding $\boldsymbol{X}_t \in \mathbb{R}^N, t = 1, \ldots, T$, where $N = 40$ and $T = 667$.

In this application we compare music analysis based on distances computed directly between MFCC feature vectors, and based upon the statistics of MFCC features within a local temporal window. The motivation of this work is that the local statistics of MFCC features within time moving windows constitutes a better representation of the similarities and differences in the music than distances computed directly on the MFCC features. For the former, we must compute distances between time-evolving distributions, which motivates the methods discussed in Section II.

In Fig. 2(a) we plot the frequency content of the frames of music, as captured via a spectrogram; the spectrogram frequency content is closely related to the MFCC features. By modeling the evolution in the statistics of multiple contiguous frames of frequency content, the hope is that we capture more specific aspects of the music, than spectral content at one point in time. Specifically, the frequency spectrum at one time point may miss statistical aspects of the music captured by frequency content at neighboring times.

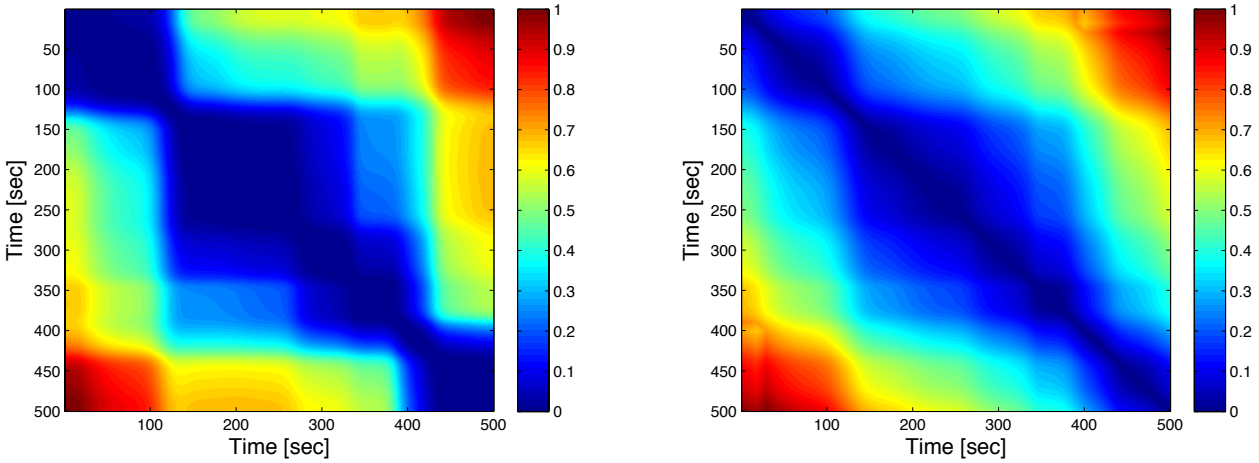
In the proposed model, we assume that the frames have a time evolving distribution parametrized by $\boldsymbol{\Sigma}_t$, as in (24). The diagonal elements of $\boldsymbol{\Sigma}_t$ denote the energy content in each frequency band at time t, and the off-diagonal elements represent the correlation between different frequency bands. In music, we observe that nearby frames in time usually tend to have similar $\boldsymbol{\Sigma}_t$, whereas, temporally distant frames tend to have different $\boldsymbol{\Sigma}_t$. Thus, the time evolution of $\boldsymbol{\Sigma}_t$ is smooth and modeled with the GP prior. As described in (21)-(23) and (26)-(28), the GP piror explicitly encodes this belief and is utilized to estimate $\boldsymbol{\Sigma}_t$. The length-scale parameter is set to $\tau = 5$, the number of local factors is set to $K = 5$, and the other hyperparamters are set to $c_0 = d_0 = e_0 = f_0 = \sigma_n = 10^{-3}$. Empirically, we find that this model fits the data well and the performance is relatively insensitive to the parameter settings (many different parameter settings yielded similar results). A total of 4000 MCMC iterations are performed, with the first 2000 samples discarded as burn-in. The covariance matrix $\boldsymbol{\Sigma}_t$ is calculated by averaging across the collected samples. Then the pairwise affinity kernel is computed according to (5)-(6), the diffusion-map algorithm is applied, and the diffusion distances are calculated according to (14)-(16).

To demonstrate the advantage of organizing the music intervals based on local statistics, we compare the obtained results to diffusion maps using pairwise affinity kernel constructed with Euclidean distances between the MFCC features directly. Additionally, we compare our method to results from the kernel beta process (KBP) [35]. The KBP does not explicitly learn an embedding, but rather represents the MFCCs at each time in terms of a learned dictionary, yielding a low-dimensional representation. The statistical relationships between pieces of music are defined by the similarity of dictionary usage. The KBP approach models each MFCC frame, but imposes that temporally nearby frames are likely to have similar dictionary usage. The proposed method explicitly utilizes a time-evolving covariance matrix; the subspace defined by that covariance matrix is related to the subspace defined by the KBP dictionary-based method, but the GP model does not explicitly impose dictionary structure (the covariance is allowed to vary more freely).

In Figs. 2(b)-(d) we illustrate the intra-piece relationships as a function of time, based on the three approaches considered. The results in Figs. 2(b)-(c) were computed via the diffusion-based embedding, where (b) used the proposed method of computing distances between data at two points, and (c) used Euclidian distance between MFCC feature vectors. The results in Fig. 2(d) were computed by KBP.

In Figs. 2(b)-(c), the relationship between data at times t and t' is represented as $f(d_{tt'}) = \exp(-\frac{d_{tt'}}{\delta})$, where $d_{tt'}$ represents diffusion distance, and $\delta = 10^{-4}$. This plots the relationship between the data on a scale of [0,1]. In Fig. 2(d) the correlation is shown between the probability of dictionary usage at times t and t', as done in [35].

At the website http://youtu.be/XhDz0npyHmg, one may listen to the music, and examine how the music maps onto the segmentations and relationships in Figs. 2(b)-(d). It is evident that diffusion maps based on Euclidian distance applied directly to the MFCC features, in Fig. 2(c), does not capture the detailed temporal relational information of the proposed approach in Fig. 2(b) and of the KBP approach in Fig. 2(d). Figs. 2(b) and (d) are in good agreement with regard to large-scale behavior, but the proposed method in (b) appears to capture more-detailed temporal structure.



(a) (b)

Fig. 1. (a) Normalized diffusion distance between time points. (b) Normalized approximated geodesic distance between time points.

For example, interval $[10, 100]$ seconds consists of a solo, dominated by human voice, whereas the subinterval $[30, 42]$ seconds contains a rhythm different from other parts of the piece. The KBP analysis in Fig. 2(d) is unable to capture the latter detailed structure, but it is clearly evident in the results of the proposed algorithm, in Fig. 2(b). The KBP approach seems to be effective in inferring large-scale temporal relationships, but not as good at distinguishing localized, fine-scale temporal differences. The method by which the diffusion analysis is performed appears to be important, as the results in Fig. 2(c), in which a simple Euclidian distance was used in the diffusion kernel, yield relatively poor results, missing most large-scale and fine-scale details.

To further illustrate this point, we analyze the performance of the three methods in the short interval $[65, 75]$ seconds in Fig. 3. As shown in Fig. 3(a), in the interval $[68.5, 71.5]$ seconds, the human voice harmonics disappear, indicating a break in the singing. Comparing the corresponding distances in Figs. 3(b)-(d), we find that diffusion maps using time evolving statistics capture this break, whereas KBP fails to capture this break. Although diffusion maps based on Euclidean distances between the features capture the break, other false breaks are captured as well. Similar trends and performance can be also found in the intervals $[97, 99]$ seconds and $[223, 225]$ seconds.

The diffusion mapping formulates an embedding space that discovers the low-dimensional manifold of the data. Figure 4 depicts two coordinates of the mapping (14) corresponding to the eigenvectors associated with the largest two non-trivial eigenvalues. For evaluation, we annotated the song with markers indicating whether human voice appears. As depicted in the figure, we find that the second coordinate correlates with human voice. For instance, we observe that the second coordinate is significantly large during the interval $[10, 100]$ seconds, which consists of the voice solo. In addition, the first coordinate correlates with the overall background sounds: it takes small values for the first 312 seconds of the song, and then exhibits a sudden increase when peaky humming appears. Such information can be utilized to interpret the similarity between frames and may be used for other music-analysis tasks. This suggests that the coordinates of the embedding, *i.e.*, the eigenvectors of the graph, indeed represent the underlying factors controlling the music. See http://youtu.be/4uPaLgbMSQw for an audio-visual display of these results.

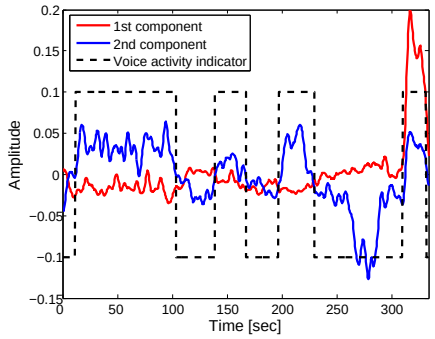


Fig. 4. The two principal eigenvectors as a function of time compared to human voice activity indicator

C. Epileptic seizure prediction

We now consider epileptic-seizure prediction based on EEG recordings. It is important and desirable to predict seizures so that patients can be warned a few minutes prior. Many studies have been conducted in order to devise reliable methods that can distinguish interictal and preseizure states [17]. Recent literature suggests that the correlation between frequency components of EEG signals indicate the underlying brain state [36], [37], [38]. However, because EEG recordings tend to be very noisy, and because the brain activity states and their relationship to the EEG activities are unknown, it is considered a difficult problem without existing solutions [9], [39].

In the present work, we consider intracranial EEG (icEEG) data collected from a patient at the Yale-New Haven Hospital. Multiple electrode contacts were used during the recording; in this work, we focus on the contacts located at the seizure onset area in the right occipital lobe. Discovering the relationships between different areas will be a subject of future research. We study four 60-minute long icEEG recordings with a

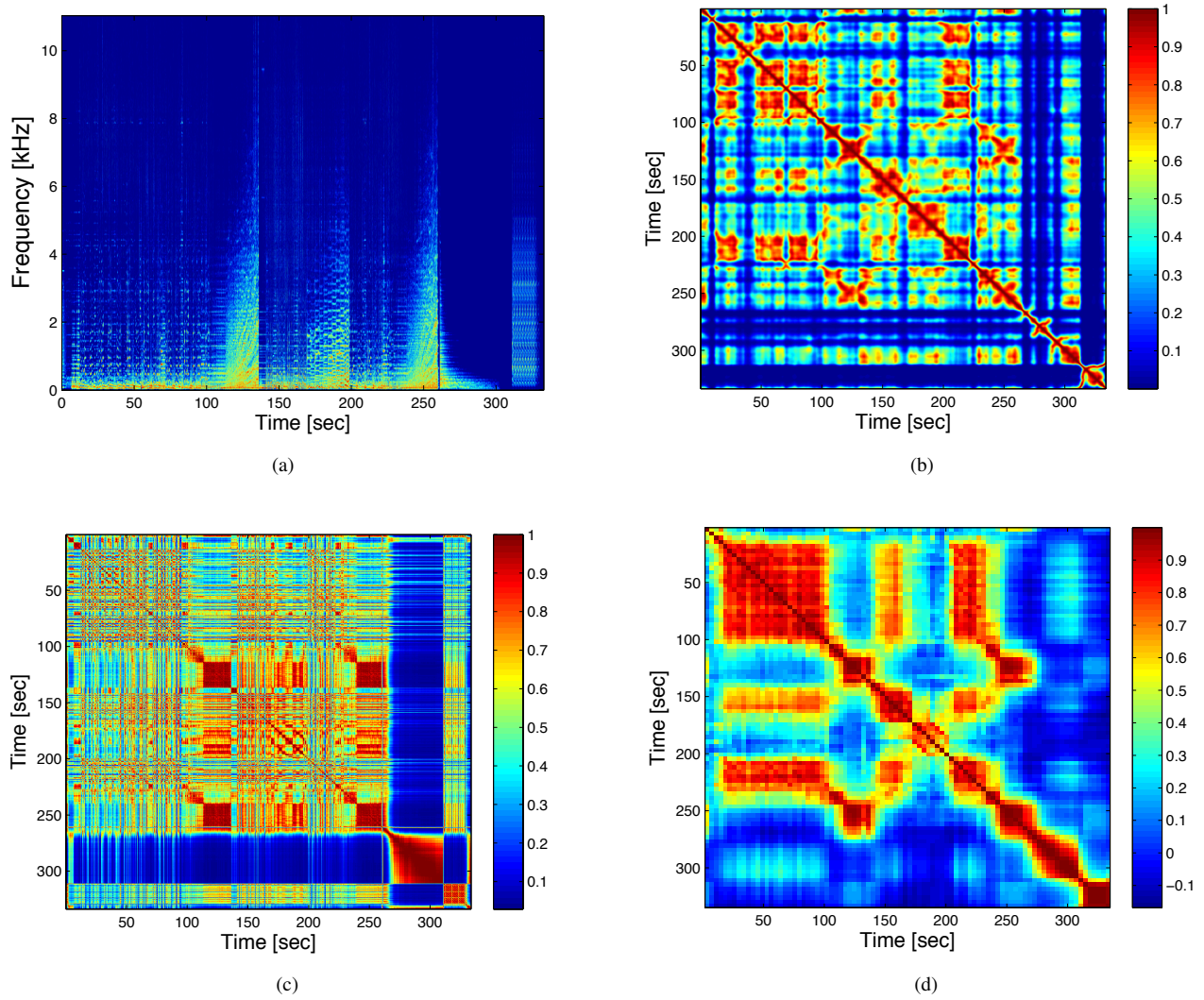


Fig. 2. Analysis results of the song "A Day in the Life". (a) Spectrogram of the song. (b)-(d) Comparison of the pairwise similarity matrices obtained by (b) diffusion maps using time evolving statistics, (c) diffusion maps using Euclidean distances between features, and (d) kernel beta process [35].

sampling rate of 256 Hz, each containing a seizure. A detailed description of the collected dataset can be found in [39].

Figure 5 presents a 60-minute EEG recording (Fig. 5(a)) and its short time spectrogram (Fig. 5(b)). As shown in Fig. 5(a), the seizure occurs after about 56 minutes in this recording, and is visible. However, it is difficult by observation to notice differences between recording parts that immediately precede the seizure and parts that are located well before the seizure. Our goal is, first, to infer a low-dimensional representation of the recordings, which discovers the intrinsic states ($i.e.$, the brain activities) of the signal. By relying on such a representation, we detect anomaly patterns prior to the seizure onsets (preseizure states) and distinguish them from samples recorded during resting periods (interictal state), thereby enabling prediction of seizures.

The short time Fourier transform (STFT) with a 1024 sample window and 512 sample overlap is applied to obtain features in the frequency domain. The amplitude of frequency components in Delta (0.5-4 Hz) and Theta (4-8 Hz) bands, with 0.25 Hz spacing, are collected for each time frame and used as feature vectors in the following experiments. Thus, the feature vectors $\boldsymbol{X}_t \in \mathbb{R}^{32}$ of the data in the frequency domain are obtained, as shown in Fig. 5(b). The Beta (13-25 Hz) and Gamma (25-100 Hz) bands were also included but empirically showed no significant contribution. In this experiment, two 5-minute intervals from each recording are analyzed: one is the 5-minute interval immediately preceding the seizure onset (preseizure state), and the other is a 5-minute interval located 40 minutes before the seizure (interictal state). Therefore, for each 5-minute interval, we have a set of vectors $\boldsymbol{X}_t \in \mathbb{R}^N, t = 1, \ldots, T$, where $N = 32$ and $T = 149$. The obtained features are centered, and hence, each $\boldsymbol{X}_t$ is considered as a sample from a zero-mean multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{0}_N, \boldsymbol{\Sigma}_t)$.

In this application, unlike the prior knowledge we have in the music analysis case, we do not have a reliable notion of the way the covariance matrices are correlated at different times. Thus, the only belief we can incorporate into the generative

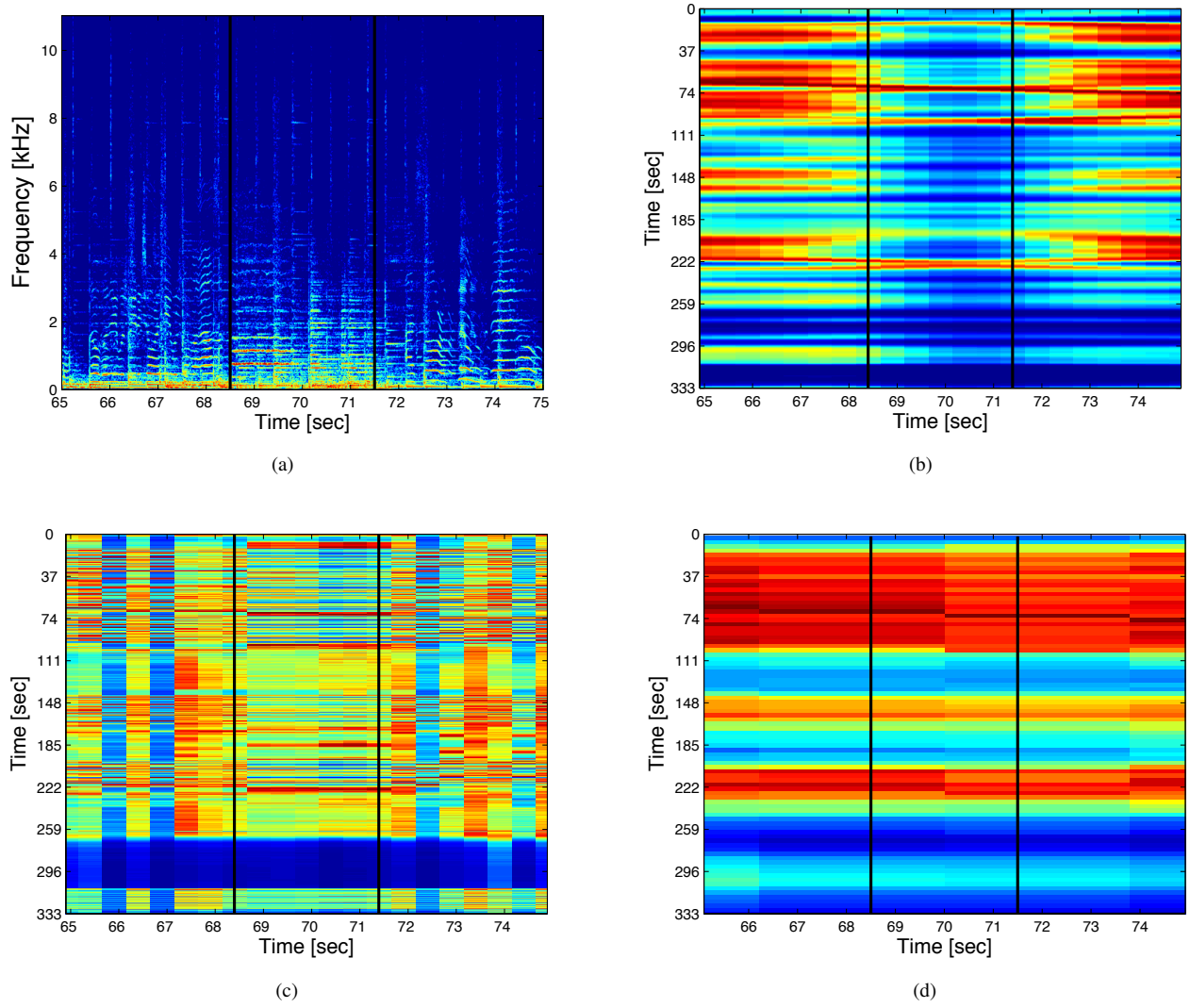
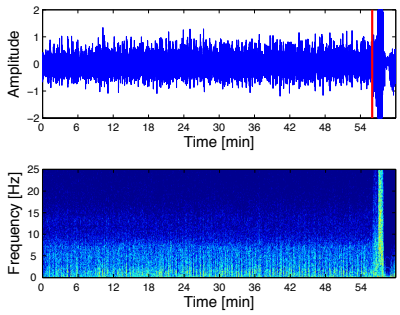


Fig. 3. Analysis results in the subinterval $[65, 75]$ seconds of the song "A Day in the Life". (a) Spectrogram of the song. (b)-(d) Comparison of the pairwise similarity matrices obtained by (b) diffusion maps using time evolving statistics, (c) diffusion maps using Euclidean distances between features, and (d) kernel beta process [35].



(a)

Fig. 5. (a) A sample recording. The red line marks seizure's onset, at approximate 56 minutes. (b) The STFT features of the recording.

model is that $\mathbf{\Sigma}_t$ are smoothly varying. This information is encoded in two priors, which are used to estimate $\mathbf{\Sigma}_t$: A Bayesian approach with a latent variable model using a non-stationary AR process prior (BAR) and an empirical approach using the moving averaging (MA). The former assumes stationarity of the differences between consecutive covariance matrices. The latter assumes local stationarity, *i.e.*, feature vectors within a small window are generated from the same Gaussian distribution.

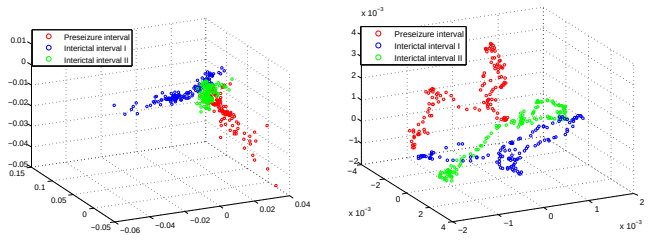
In our experiments, both models (BAR and MA) are used for covariance matrix estimation in each 5-minute interval. The BAR model is implemented according to (21)-(23) and (29)-(31), where the number of local factors is set to $K = 5$ and the hyperparameters are set to $e_0 = f_0 = g_0 = h_0 = 10^{-3}$; 4000 MCMC iterations are performed, with the first 2000 discarded as burn-in. The collection samples are averaged to estimate $\mathbf{\Sigma}_t$. Under the MA model, we simply estimate the covariance directly based on the features in a local neighborhood in time, *i.e.*, $\hat{\mathbf{\Sigma}}_t^M = \sum_{s=t-M}^{t+M} \boldsymbol{X}_s \boldsymbol{X}_s^T / (2M+1)$, where $2M+1$ is the length of the window in which local stationarity is assumed to hold. We need to choose M large enough to provide an accurate estimate of $\mathbf{\Sigma}_t$ while small enough to

avoid smoothing out the local varying statistics. In practice, we set $M = 32$ according to an empirical criterion of being the smallest value that yields sufficient smoothness, formulated by the admittedly *ad hoc* criterion: $\frac{\sum_{t=1}^{T} ||\hat{\boldsymbol{\Sigma}}_t^M - \hat{\boldsymbol{\Sigma}}_t^{M+1}||_2}{\sum_{t=1}^{T} ||\hat{\boldsymbol{\Sigma}}_t^M||_2} \leq 0.05$. Circumventing the need for such criteria is one reason the BAR approach is considered.

In Fig. 6, we present the obtained embedding (by applying Algorithm 1) of multiple intervals of an EEG recording (indexed as recording 1) using diffusion maps based on the time-evolving statistics. We embed three 5-minute intervals: one preseizure interval and a couple of interictal intervals (located 30 and 40 minutes prior to the seizure onset, respectively). The presented scatter plot shows the embedded samples in the space formulated by the 3 principal eigenvectors (setting $\ell = 3$ in (14)), *i.e.*, each 3 dimensional point corresponds to the diffusion map of a single feature vector $\boldsymbol{X}_t$.

We observe that under both models (BAR and MA) the low-dimensional representation separates samples recorded in the preseizure state (colored red) from samples recorded in interictal states (colored blue and green). In both Figs. 6(a) and (b), the embedded samples of the two interictal intervals are located approximately in the same region, with the embedded samples of one of the interictal intervals (colored green) tend slightly towards the embedded samples of the preseizure interval. This result exemplifies the ability of the proposed method to discover the underlying states. Indeed, without prior knowledge of the labeling of the different intervals (preseizure or interictal) and based merely on the recorded signal, the proposed method organizes the signal according to the intrinsic state of the patient.

We remark that the embeddings obtained under MA and BAR models, in Figs. 6(a) and (b), respectively, are different. Under the BAR modeling, the method tends to recover trajectories with similar evolving patterns due to the strong time correlation assumption. On the other hand, under the MA modeling, the method tends to form point clouds resulting from the local Gaussian distribution assumption.

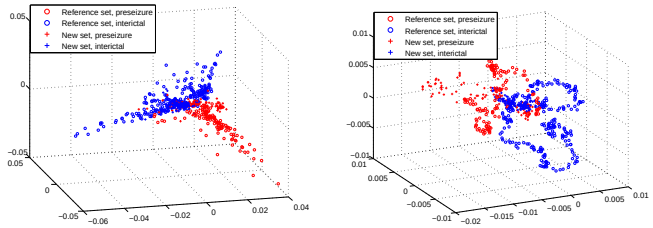


(a) (b)

Fig. 6. Scatter plots of the embedded samples of three 5-minute intervals of EEG recording 1. Each point is the diffusion map of the features of each time frame in the recording, setting $\ell = 3$. The colors indicate the different intervals. (a) The embedded samples under the MA modeling. (b) The embedded samples under the BAR modeling.

We now test the consistency and extensibility of the obtained low-dimensional representation. In practice, it is desirable to learn the mapping from reference recordings, and then, when new recordings become available to embed them into the low-dimensional space in an online fashion in order to identify and predict seizures.

Figure 7 depicts the embedded samples obtained by applying Algorithm 2 using recording 1 as the reference set and recording 2 as the new incoming set. As observed, the new incoming samples are embedded into the same regions as the reference samples from corresponding interval types. For example, we observe in Fig. 7(a) that the new samples of the interictal state interval are embedded close to the reference samples of the interictal state interval. In addition, we observe that the samples of the interictal state intervals are embedded around the origin, whereas the samples of the preseizure intervals are embedded further away from the origin, suggesting that the preseizure state intervals are "anomalies" that tend to stick out from the learned "normal state" model. These properties allow for a simple identification of preseizure samples: preseizure labels can be assigned to new samples when their corresponding embedded points are near the region of preseizure reference samples and far from the origin. As shown in Fig. 7(b), under the BAR modeling, the embedded points have a different representation. In this case, the embedded samples from different intervals form different trajectories. Nevertheless, we can assign labels to new incoming points in a similar manner - based on their distance to the reference trajectories of each state.



(a) (b)

Fig. 7. Scatter plots of the embedded samples of four 5-minute intervals of EEG recordings: two reference intervals (preseizure and interictal) from recording 1 and two new incoming intervals (preseizure and interictal) from recording 2. The representation is obtained using the reference samples and then extended to the new incoming samples according to Algorithm 2. The dimension of the embedding is set to $\ell = 3$ for visualization. (a) The obtained embedding under the MA modeling. (b) The obtained embedding under the BAR modeling.

From the presented results in Figs. 6 and 7, we conclude that the proposed method indeed discovers the brain activity and enables, merely by visualization in 3 dimensions, to distinguish preseizure states from interictal states. Furthermore, the results imply that the coordinates of the diffusion embedding, *i.e.*, the eigenvectors of the constructed kernel, have a real "meaning" in this application. Similarly to the music application, where the coordinates indicate, for example, the human voice, here they capture different aspects of the intrinsic state of the patient. In the present work, we exploit this representation and devise a simple classification procedure to identify preseizure states, which enables to predict seizures. We remark that a simple algorithm is sufficient since the embedding already encodes the required information and enables to distinguish the different states.

We repeat the experiment described above, and extend the model to three unseen recordings according to Algorithm 2.

Algorithm 2 Sequential implementation of diffusion maps based on time evolving statistics

Input: Reference observations $\{\boldsymbol{X}_t\}, t = 1, \cdots, T$, new incoming observations $\{\boldsymbol{X}_s\}, s = 1, \cdots, S$, diffusion step n, neighbourhood size ϵ, dimension ℓ

Output: Embeddings $\{\boldsymbol{\Psi}_{l,n}(\boldsymbol{X}_t)\}, t = 1, \cdots, T$ and $\{\boldsymbol{\Psi}_{l,n}(\boldsymbol{X}_s)\}, s = 1, \cdots, S$

1: Estimate distribution $p(\boldsymbol{X}_t|\boldsymbol{\beta}_t)$ for all time points (using moving average (MA), Bayesian model with AR process prior (BAR) or Bayesian model with Gaussian process prior (BGP) to infer $\boldsymbol{\beta}_t$)

2: Compute the $T \times T$ matrix $\boldsymbol{W}$, where $W_{t_1,t_2} = \exp\left(-\frac{\mathcal{D}(p(\boldsymbol{X}|\boldsymbol{\beta}_{t_1})||p(\boldsymbol{X}|\boldsymbol{\beta}_{t_2}))+\mathcal{D}(p(\boldsymbol{X}|\boldsymbol{\beta}_{t_2})||p(\boldsymbol{X}|\boldsymbol{\beta}_{t_1}))}{\epsilon}\right)$

3: Apply eigenvalue decomposition to $\boldsymbol{W}$ and keep ℓ principle eigenvalues λ_j and eigenvectors $\boldsymbol{\varphi}_j$

4: For new incoming data $\boldsymbol{X}_s, s = 1, \cdots, S$, estimate the distribution $p(\boldsymbol{X}_s|\boldsymbol{\beta}_s)$ (using MA, BAR or BGP).

5: Compute the $S \times T$ nonsymmetric kernel matrix $\boldsymbol{A}$, where $A_{s,t} = \exp\left(-\frac{\mathcal{D}(p(\boldsymbol{X}|\boldsymbol{\beta}_s)||p(\boldsymbol{X}|\boldsymbol{\beta}_t))}{\epsilon}\right)$

6: Construct $\boldsymbol{R} = \boldsymbol{D}_A^{-1}\boldsymbol{A}$, where $\boldsymbol{D}_A = \text{diag}\left\{\sum_{t=1}^{T} A_{s,t}\right\}$

7: Construct $\tilde{\boldsymbol{A}} = \boldsymbol{R}\boldsymbol{Q}^{-1}$, where $\boldsymbol{Q} = \text{diag}\left\{\sum_{t=1}^{T} R_{s,t}\right\}$

8: Calculate the diffusion maps embeddings of the new incoming samples $\boldsymbol{\Psi}_{\ell,n}(\boldsymbol{X}_s) = (\psi_1^s, \psi_2^s, \ldots, \psi_\ell^s)^T$, where $\boldsymbol{\psi}_j = \frac{1}{\sqrt{\lambda_j^n}}\tilde{\boldsymbol{A}}\boldsymbol{\varphi}_j$

9: Calculate the diffusion maps embeddings of the reference samples $\boldsymbol{\Psi}_{\ell,n}(\boldsymbol{X}_t) = (\lambda_1^n\varphi_1^t, \lambda_2^n\varphi_2^t, \ldots, \lambda_\ell^n\varphi_\ell^t)^T$

As before, recording 1 is used as reference set to learn the model, which in turn is extended to the other three recordings (2-4). In each recording, there is a preseizure state interval whose T samples are labeled as "preseizure" and an interictal interval whose T samples are labeled as "interictal". By using the labels of the reference samples as training data, we train standard linear classifiers in the low-dimensional embedding space to distinguish samples recorded in preseizure states from samples recorded in interictal states. In our algorithm, the classification boundary is the hyperplane lying in the middle of the two empirical means of the embedded reference samples of each state.

Table I summarizes the detection rate and false alarm rate of the samples from all the recordings. As can be seen, both implementations relying on the MA and BAR modeling perform well in classifying samples into preseizure and interictal states. In general, a larger portion of data samples in the preseizure state interval are correctly identified, while only a small fraction of data samples in the interictal state are identified as preseizure state. Further, we find BAR modeling has an overall higher detection rate while MA has a lower false alarm rate. One reason is in MA, we assume local Gaussian distribution of data samples and smooth variation of controlling parameters, which inhibit sudden changes, thus reducing the probability of detecting anomaly samples. The other reason is the Nystrom method used in Algorithm 2 (Step 8) to embed new incoming data samples has the effect of shrinking coordinates' amplitude (in the reference set, more embeddings lie close to the origin than far away from it). In MA, this causes identifying more new incoming samples as interictal state because of the reference set, interictal state samples are embedded around origin (see Fig. 7). While in BAR, we assume similar evolving patterns of data samples, organizing samples from two state intervals into two trajectories. Therefore, identifying states of new incoming data samples is not effected by the shrinking effect of Nystrom method.

An interesting result is that the large portion of embedded samples from a preseizure interval actually reside in the interictal state region. Such a result was observed in other experiments on human collected data. It implies that during the preseizure interval, the subject's brain tries to maintain the normal state and resist the seizure. As a result, the states of the samples alternate rapidly between normal and anomaly states. Thus, to effectively predict seizures, relying on single samples/time frames is insufficient, and an online algorithm that aggregates a consecutive group of samples is required. For example, it is evident from the results in Table I that tracking the percentage of anomaly samples within a 5-minute interval may be adequate: if the percentage of anomalies is greater than a predefined threshold of 0.35, a prediction of seizure is reported. Designing more sophisticated classification algorithms and testing them on larger dataset with multiple subjects will be addressed in future work.

TABLE I
DETECTION RATE AND FALSE ALARM RATE (REPRESENTED AS PERCENTAGE) OF THE EEG SAMPLES FROM ALL THE RECORDINGS

Model	MA		BAR	
Recording	Detection	False Alarm	Detection	False Alarm
1	94.3	26.2	100.0	29.5
2	41.8	20.5	56.0	24.0
3	62.3	19.7	62.3	24.6
4	48.4	19.7	82.8	23.8

V. CONCLUSIONS

A dimensionality-reduction method for high-dimensional time series is presented. The method exhibits two key components. First, multiple approaches to estimate time evolving covariance matrices are presented and compared. Second, using the Kullback-Leibler divergence as a distance metric, diffusion maps are applied to the probability distributions estimated from samples, instead of samples themselves, to obtain a low-dimensional embedding of the high-dimensional time series. Theoretical and experimental results show that the embedding inferred by this method discovers the underlying factors, which govern the observations, and preserves the geodesic distance between samples on the corresponding statistical manifold. Encouraging results are obtained in two real-world applications: music analysis and epileptic seizure prediction. Especially for the latter application, an online seizure identification system is developed, showing the possibility of predicting epileptic seizures based on time evolving statistics of EEG recordings. In future work, we will propose models capturing higher order time evolving statistics beyond the covariance matrices.

REFERENCES

[1] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[4] D. Durante, B. Scarpa, and D. B. Dunson. Locally adaptive Bayesian covariance regression. *ArXiv e-prints*, October 2012.

[5] E. B. Fox and M. West. Autoregressive Models for Variance Matrices: Stationary Inverse Wishart Processes. *ArXiv e-prints*, July 2011.

[6] R. R. Coifman and M. J. Hirn. Diffusion maps for changing data. *ArXiv e-prints*, September 2012.

[7] A. F. Zuur, R. J. Fryer, I. T. Jolliffe, R. Dekker, and J. J. Beukema. Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics*, 14(7):665–685, 2003.

[8] R. Talmon and R. R. Coifman. Empirical intrinsic geometry for intrinsic modeling and nonlinear filtering. *Proc. Nat. Acad. Sci.*, 110(31):12535–12540, 2013.

[9] R. Talmon and R. R. Coifman. Empirical intrinsic modeling of signals and information geometry. *Technical Report YALEU/DCS/TR-1467*, 2012.

[10] O. C. Jenkins and M. J. Matarić. A spatio-temporal extension to isomap nonlinear dimension reduction. In *Proceedings of the twenty-first international conference on Machine learning*, page 56. ACM, 2004.

[11] M. Lewandowski, J. Martinez-del Rincon, D. Makris, and J. C. Nebel. Temporal extension of laplacian eigenmaps for unsupervised dimensionality reduction of time series. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 161–164. IEEE, 2010.

[12] E. Fox and D. Dunson. Bayesian nonparametric covariance regression. *Arxiv preprint arXiv:1101.2017*, 2011.

[13] K. M. Carter, R. Raich, W.G. Finn, and A.O. Hero. Information-geometric dimensionality reduction. *Signal Processing Magazine, IEEE*, 28(2):89–99, 2011.

[14] K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero. Fine: Fisher information nonparametric embedding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):2093–2098, 2009.

[15] A. Singer and R. R. Coifman. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226–239, 2008.

[16] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

[17] 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, J. Jefferys, G. Worrell, M. Le Van Quyen, S. J. Schiff, and F. Mormann. Controversies in epilepsy: Debates held during the fourth international workshop on seizure prediction. *Epilepsy and Behavior*, 19(1):4 – 16, 2010.

[18] R. Dahlhaus. On the kullback-leibler information divergence of locally stationary processes. *Stochastic Processes and their Applications*, 62(1):139–168, 1996.

[19] 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*, pages 289–296, 2005.

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

[21] F. R. K. Chung. *Spectral Graph Theory*. CBMS - American Mathematical Society, Providence, RI, 1997.

[22] D. Kushnir, A. Haddad, and R. Coifman. Anisotropic diffusion on sub-manifolds with application to earth structure classification. *Appl. Comput. Harmon. Anal.*, 32(2):280–294, 2012.

[23] A. Haddad, D. Kushnir, and R. R. Coifman. Texture separation via a reference set. *to appear in Appl. Comput. Harmon. Anal.*, 2013.

[24] R. Talmon, I. Cohen, S. Gannot, and R.R. Coifman. Supervised graph-based processing for sequential transient interference suppression. *IEEE Trans. Audio, Speech Lang. Process.*, 20(9):2528–2538, Nov. 2012.

[25] C. Alexander. Moving average models for volatility and correlation, and covariance matrices. *Handbook of Finance*, 2008.

[26] L. Bauwens, S. Laurent, and J. V. Rombouts. Multivariate garch models: a survey. *Journal of applied econometrics*, 21(1):79–109, 2006.

[27] R. S. Tsay. *Analysis of financial time series*, volume 543. Wiley-Interscience, 2005.

[28] P. Richard Hahn, C. M. Carvalho, and J. G. Scott. A sparse factor analytic probit model for congressional voting patterns. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(4):619–635, 2012.

[29] C. E. Rasmussen. Gaussian processes for machine learning. 2006.

[30] A. Banerjee, D. Dunson, and S. Tokdar. Efficient Gaussian Process Regression for Large Data Sets. *ArXiv e-prints*, June 2011.

[31] B. Chen, M. Chen, J. Paisley, A. Zaas, C. Woods, G. Ginsburg, A. Hero, J. Lucas, D. Dunson, and L. Carin. Bayesian inference of the number of factors in gene-expression analysis: application to human virus challenge studies. *BMC bioinformatics*, 11(1):552, 2010.

[32] H. Ishwaran and J. S. Rao. Spike and slab variable selection: frequentist and bayesian strategies. *Annals of Statistics*, pages 730–773, 2005.

[33] A. Armagan, D. Dunson, and J. Lee. Generalized double pareto shrinkage. *arXiv preprint arXiv:1104.0861*, 2011.

[34] Früwirth-Schnatter. Data augmentation and dynamic linear models. *J. Time Ser. Anal.*, 15:183–202, 1994.

[35] L. Ren, Y. Wang, D. B. Dunson, and L. Carin. The kernel beta process. In *NIPS*, pages 963–971, 2011.

[36] K. Gadhoumi, J. M. Lina, and J. Gotman. Seizure prediction in patients with mesial temporal lobe epilepsy using {EEG} measures of state similarity. *Clinical Neurophysiology*, (0):–, 2013.

[37] C. Alvarado-Rojas, M. Valderrama, A. Witon, V. Navarro, and M.L. Van Quyen. Probing cortical excitability using cross-frequency coupling in intracranial eeg recordings: A new method for seizure prediction. In *Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE*, pages 1632–1635, 2011.

[38] A. Shoeb and J. Guttag. Application of machine learning to epileptic seizure detection. In *Proc. of int. conf. on machine learning*, 2010.

[39] 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*, 10(3):579 – 590, 2013.