Signal Processing 130 (2017) 403–411

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Dynamical system classification with diffusion embedding for ECG-based person identification

Jeremias Sulam [a,*], Yaniv Romano [b], Ronen Talmon [b]

[a] Department of Computer Science, Technion – Israel Institute of Technology, Israel
[b] Department of Electrical Engineering, Technion – Israel Institute of Technology, Israel

ARTICLE INFO

Article history:
Received 9 May 2016
Received in revised form
6 July 2016
Accepted 24 July 2016
Available online 26 July 2016

Keywords:
Classification
Manifold learning
Person identification
ECG

ABSTRACT

The problem of system classification consists of identifying the source system corresponding to a certain output signal. In the context of dynamical systems, the outputs are usually given in the form of time series, and this identification process includes determining the underlying states of the system or their intrinsic set of parameters. In this work we propose a general framework for classification and identification based on a manifold learning algorithm. This data-driven approach provides a low-dimensional representation of the system's intrinsic variables, which enables the natural organization of points in time as a function of their dynamics. By leveraging the diffusion maps algorithm, a particular manifold learning method, we are not only able to distinguish between different states of the same system but also to discriminate different systems altogether. We construct a classification scheme based on a notion of distance between the distributions of embedded samples for different classes, and propose three ways of measuring such separation. The proposed method is demonstrated on a synthetic example and later applied to the problem of person identification from ECG recordings. Our approach obtains a 97.25% recognition accuracy over a database of 90 subjects, the highest accuracy reported for this database.

1. Introduction

The problem of system classification or recognition deals with identifying the source system of a particular output. In this work we study those cases where the output signal is given in terms of a time series, as this is often the case in dynamical systems. This classification problem has shown to be of interest in a range of applications [1–6], and it usually involves the estimation of the intrinsic state of a system, or of its intrinsic parameters. This task can be challenging, in particular when analyzing complex biological systems where we know little about their underlying variables.

System recognition can be viewed as a multi-class classification problem, where machine learning algorithms are trained to identify the source of a test signal. Under this framework, one builds a representation of the training data which efficiently represents certain features of the underlying system. These observations, or feature vectors, are then used to train a classifier in a supervised learning fashion, adapting its parameters to best fit the training data. While efficient, the design of these features is a key factor in the overall performance of the classifier. Traditional approaches employ hand-crafted features, given often in the form of filters, which are effective for a specific problem, yet are hard to design, tune and generalize to other cases.

Other methods rely solely on the data and its dynamics to extract relevant features for classification. The method presented in this paper is based on manifold learning, and it falls into this data-driven category. Manifold learning is an active area of research in which the signal under analysis is assumed to belong to a low-dimensional manifold [7–9]. By seeking for a low dimensional representation of such manifolds, many problems in signal and image processing become more tractable. More specifically, recent manifold learning methods rely on the construction of a diffusion distance between observations, and build an intrinsic representation based on the notion of similarity induced by this distance [10]. The resulting diffusion maps algorithm embeds signal samples in time to a low-dimensional domain where the Euclidean distance between the mapped samples approximates the diffusion distance in the original domain [11].

In the context of dynamical systems, the diffusion maps algorithm exploits the dynamics of the system in order to determine the distance in the diffusion embedding [12] and has shown to be successful in a number of applications. In [13], a related method was employed to distinguish between healthy and acidotic fetuses from intrapartum heart rate variability signals. In a different work [14], the authors showed that a similar approach can be used to

* Corresponding author.
E-mail address: jsulam@cs.technion.ac.il (J. Sulam).

http://dx.doi.org/10.1016/j.sigpro.2016.07.026
0165-1684/

predict epileptic seizures in intracranial EEG recordings, by relying on the evolution of the assumed underlying latent variable.

When dealing with real applications, and especially with biological or physiological data, the measured signal might have several sources of variability other than the one of interest. In these cases, these irrelevant sources prevent us from finding a low-dimensional representation of the latent variable under study. A simple solution is the introduction of a proper observation operator which is invariant or robust to such nuisance sources of variability. The scattering transform, introduced recently in [15], provides a representation which is stable to deformation and has been very useful in a variety of applications [13,16,17].

In this work, we go one step further by formalizing a multiclass classification framework based on the diffusion maps algorithm with the scattering transform as a nonlinear observation operator. We propose the construction of class-specific proxies by applying the scattering transform to the training data. Then, given a test signal, we construct virtual observations by joining every class-specific training signals with the testing samples. The diffusion maps algorithm is then applied to every such virtual observation, obtaining a low-dimensional representation in the embedded space. By studying the distance between the distributions of the obtained training and testing samples, we are able to identify the system to which the test signal belongs as the system whose embedded samples diffuse the least from those of a given set.

In particular, we apply this classification algorithm to the challenging problem of person identification through electrocardiographic (ECG) signals [18]. The ECG is a signal that reflects mainly the electrical activity of the heart, and it has been shown to represent important recognition capabilities due to the very personal aspects that influence the generation of these signals [19]. These features make ECG as a robust candidate for biometric recognition and verification [18,20]. Most previous studies dealing with this problem have employed a traditional classification scheme where certain ad hoc morphological features (such as QRS-complex information, duration and amplitude of the different waves within each cardiac cycle, etc.) are extracted and fed into a classifier [21–23]. These methods usually require a pre-processing stage in order to detect the R peak in the ECG, and the segmentation and alignment of the corresponding segments [23–25]. Other works have employed some spectral characterization of the ECG signal [26] and its analysis through the wavelet transform [27]. Yet, we are not aware of any data-driven or manifold learning-based method which has attempted to tackle this problem. Our results indicate that the proposed approach, which follows this recent line of work, obtains state-of-the-art results in a publicly available database.

The paper is organized as follows. In Section 2, we provide a brief description of the Scattering Transform and the diffusion maps algorithm, and their relevance for the dynamics of a system. In Section 3 we introduce our classification scheme. In Section 4, experimental results are presented. First we present the study of a synthetic example, demonstrating the proposed approach. We then move to the study of electrocardiographic (ECG) signals in the context of a classification problem, particularly, person identification. Lastly, we conclude in Section 5.

2. Diffusion maps for dynamics inference

In this section we will briefly review the diffusion maps algorithm in the context of latent variable inference. We only describe those elements that relate to the problem addressed in this paper, and the reader is referred to [11,28] for a thorough review. Consider a high-dimensional time series $z(t) \in \mathbb{R}^N$, which is the output of an unknown dynamical system $\mathcal{S}_\theta$. We assume that this system is controlled by a hidden (and possibly time dependent) variable $\boldsymbol{\theta}(t)$. Manifold learning methods attempt to obtain a low dimensional embedding from $z(t)$ to $\boldsymbol{\theta}(t)$, thus discovering the intrinsic state of the system. To accomplish this, most (if not all) manifold learning methods rely on the construction of some notions of distance between samples $z(t)$ and $z(\tau)$. However, this distance is not computed directly on samples from the time series. Instead, it is measured through a nonlinear operator, which we address next.

2.1. Scattering moments as observations

As shown in [14], when applying manifold learning algorithms to complex data such as that originated from biological and physiological systems, many sources of variability might prevent these methods from finding the low dimensional latent variable $\boldsymbol{\theta}(t)$. The solution is not to use the data $z(t)$ directly but rather to employ a nonlinear observation operator $\boldsymbol{\Phi}$ which is robust or stable to such sources of variability.

A common source of variability in real applications is time deformations. Consider the time-shifted (or time-deformed) signal $z_g(t)$, governed by the intrinsic variable $\boldsymbol{\theta}_g(t) = \boldsymbol{\theta}(t - g(t))$, where the deformation is controlled by the time-varying variable $g(t)$. The operator $\boldsymbol{\Phi}$, which is applied to the time series obtaining the observation vector $\boldsymbol{\Phi}\mathbf{z}(t)$, is said to be stable to deformation if exists a constant C such that

$$\| \boldsymbol{\Phi}\mathbf{z}(t) - \boldsymbol{\Phi}\mathbf{z}_g(t)\|_2 \leq C \ \| \boldsymbol{\theta}(t) - \boldsymbol{\theta}_g(t)\|_2. \tag{1}$$

In words, the operator is stable if small time deformations in the intrinsic variable are not translated into arbitrarily large changes in the observation domain.

The recently proposed scattering transform is a bi-Lifshitz nonlinear transformation which presents this kind of stability, in contrast to the traditional Short-Time Fourier Transform [15]. These properties make this transform appealing for analyzing real data. Indeed, the scattering transform has shown to be very useful in a variety of applications, e.g., in [16,29,30].

Consider a complex mother wavelet $\psi(t)$, and the corresponding dilated versions at scales 2^j, $\psi_j(t)$. Given the corresponding scaling function (or a low-pass filter) $\phi(t)$, the first order scattering transform of a signal $z(t)$ is defined as the average of the absolute value of the wavelet transform of the signal over a time window of length $T = 2^J$, and it is formally given by

$$\boldsymbol{\Phi}_S z(k, j_1) = |z(t) * \psi_{j_1}(t)| * \phi_J(t), \tag{2}$$

where $1 \leq j_1 \leq J$ and $t = k2^{J-1}$. In words, the scattering transform first applies a convolution with a wavelet function, at different scales, and then applies a (contractive) modulus operation. These coefficients are then convolved with the function $\phi_J(t)$, which averages coefficients within an interval of length 2^J. The second order transform is obtained by employing another level of convolutions, formally expressed as

$$\boldsymbol{\Phi}_S z(k, j_1, j_2) = \| z(t) * \psi_{j_1}(t)| * \psi_{j_2}(t)| * \phi_J(t). \tag{3}$$

This cascading decomposition presents in fact a convolutional network interpretation, and it is capable of extracting useful features from complex signals [29]. We compute the first and second order scattering coefficients, which typically capture most of the content of the analyzed signal [15]. We employ the ScatNet Matlab package,[1] and collect the coefficients into the Scattering Transform observation vectors by

[1] Available online at http://www.di.ens.fr/data/software/scatnet.

$$\mathbf{\Phi}z(t) = \left(\{ \mathbf{\Phi}_S z(j,k) \}_{1 \le j_1 \le J}, \{ \mathbf{\Phi}_S z(j,k) \}_{1 \le j_1 \le j_2 \le J} \right). \tag{4}$$

This way, each sample $z(t)$ yields an observation vector $\mathbf{\Phi}z(t) \in \mathbb{R}^n$, where n is the number of scattering coefficients.

2.2. Diffusion maps

The above presented observations, being stable to time deformations and other nuisance factors, can now be used by a manifold learning algorithm with the objective of recovering the intrinsic variables of the system generating the signal $z(t)$. The diffusion maps algorithm is therefore applied to the data obtained by $\mathbf{\Phi z}(t)$. This method depends on a notion of distance between samples in time $z(t)$ and $z(\tau)$, which should include information about the dynamics of the system.

When analyzing time-dependent data, the Mahalanobis distance is very useful as it considers dynamics when providing a notion of similarity [28]. This distance incorporates information from the time series around times t and τ, and it approximates the distance between the latent *unknown* variables $\boldsymbol{\theta}(t)$ and $\boldsymbol{\theta}(\tau)$. The local variability of the data is characterized by the sample covariance matrices $\hat{\mathbf{C}}(t)$, given by

$$\hat{\mathbf{C}}(t) = \sum_{l=t-L}^{t+L} \left(\mathbf{\Phi z}(t) - \hat{\boldsymbol{\mu}}(t) \right) \left(\mathbf{\Phi z}(t) - \hat{\boldsymbol{\mu}}(t) \right)^T, \tag{5}$$

where $\hat{\boldsymbol{\mu}}(t)$ is the empirical mean of the observation vectors in a time window of length $2L-1$. A modified version of the Mahalanobis distance [14] between points $z(t)$ and $z(\tau)$ (in the observation domain) is then constructed as

$$d(z(t), z(\tau)) = \frac{1}{2} \left(\overline{\mathbf{\Phi z}}(t) - \overline{\mathbf{\Phi z}}(\tau) \right)^T \ldots \left(\hat{\mathbf{C}}^{\dagger}(t) + \hat{\mathbf{C}}^{\dagger}(\tau) \right) \left(\overline{\mathbf{\Phi z}}(t) - \overline{\mathbf{\Phi z}}(\tau) \right), \tag{6}$$

where $\hat{\mathbf{C}}^{\dagger}(t)$ denotes the pseudo-inverse of the sample covariance matrix $\hat{\mathbf{C}}(t)$, and $\overline{\mathbf{\Phi z}}(t)$ denotes the zero-mean observation vector given by $\mathbf{\Phi z}(t) - \hat{\boldsymbol{\mu}}(t)$.

Once the distances between all N observations have been obtained, we can construct the affinity matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ from the similarities between every pair of observations. Formally, this matrix is constructed using a Gaussian kernel given by

$$W_{t,\tau} = \exp\left(-\frac{d(z(t), z(\tau))}{\epsilon} \right), \tag{7}$$

where ϵ is the kernel scale. This parameters define the extent of the local neighborhood of each sample $z(t)$, so that if $d(z(t), z(\tau)) > \epsilon$, then $W_{t,\tau} \approx 0$. In our work, we choose this parameter as the median of the pairwise distances $d(z(t), z(\tau))$, as this provides a reasonable connection between all samples [14].

Let $\mathbf{D}$ be a diagonal $N \times N$ matrix such that $D_{t,t} = \sum_j W_{j,t}$. Then, we can define the normalized affinity matrix as $\mathbf{W}_{\text{norm}} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$. Its eigenvectors, denoted by $\boldsymbol{\varphi}_i$, $i = 0, \ldots, N-1$, are the same as those of the corresponding Graph Laplacian, given by $\mathbf{L} = \mathbf{I} - \mathbf{W}_{\text{norm}}$. The key observation in the diffusion maps algorithm is to define an embedding based on the concept of diffusion distance between the variables $\boldsymbol{\theta}(t)$ and $\boldsymbol{\theta}(\tau)$, denoted by $\gamma(\boldsymbol{\theta}(t), \boldsymbol{\theta}(\tau))$. Intuitively, this distance measures the degree of connectivity between points $\boldsymbol{\theta}(t)$ and $\boldsymbol{\theta}(\tau)$ in terms of the graph corresponding to $\mathbf{W}_{\text{norm}}$. The reader is referred to [11] for more details on this algorithm. This way, the low-dimensional embedding is defined as a mapping between the observations $\mathbf{\Phi z}(t)$ and the low-dimensional embedded vectors $\boldsymbol{\zeta}(t) \in \mathbb{R}^d$ $(d < n)$ given by

$$\mathbf{\Phi z}(t) \mapsto \boldsymbol{\zeta}(t) = \left(\boldsymbol{\varphi}_1(t), \boldsymbol{\varphi}_2(t), \ldots, \boldsymbol{\varphi}_d(t) \right). \tag{8}$$

Note that because $\mathbf{D}^{-1}\mathbf{W}$ is row-stochastic, $\boldsymbol{\varphi}_0$ is the diagonal of $\mathbf{D}^{1/2}$ (with corresponding eigenvalue $\lambda_0 = 1$), and it is therefore ignored in the mapping. This embedding provides a low dimensional representation of each observation, and takes into account global information by depending on the entire matrix $\mathbf{W}_{\text{norm}}$ (or $\mathbf{L}$). However, it also conveys an interesting local interpretation, as samples which are close in the embedded space correspond to samples with a low diffusion distance. Concretely, the Euclidean distance in the embedded space provides an approximation to the diffusion distance $\gamma(\boldsymbol{\theta}(t), \boldsymbol{\theta}(\tau))$ in the original domain[2][11]; i.e., $\gamma(\boldsymbol{\theta}(t), \boldsymbol{\theta}(\tau)) \approx \| \boldsymbol{\zeta}(t) - \boldsymbol{\zeta}(\tau) \|_2$. In the context of the study of dynamical systems, these neighboring embedded samples correspond to similar dynamics.

2.3. Latent variable inference in ECG

Before moving on to the description of the classification framework, we present here a case study of intrinsic variable inference of a real biological signal. The diffusion maps algorithm has been shown to be efficient in recovering the underlying states of synthetic systems [12,14]. It has also been employed in several studies to analyze real complex systems [31,32]. In particular, the work in [14] presented an application of the ideas described in this section to the prediction of epileptic seizures from intracranial electroencephalographic (iEEG) signals. The authors demonstrated how the intrinsic variable of the underlying system (the brain cortex) presents a natural organization in the embedded space depicting a transition from a normal to a pre-ictal state. However, in this and many other cases, one can only assume to recover the latent variable of these very complex systems, as there is no ground truth or other physical indication that can be used for comparison.

In the case of the analysis of ECG signals, however, we are able to present more concrete evidence of the recovery of the intrinsic variable involved. Even though many factors influence the evolution of a cardiac cycle and its encoding into an ECG signal, it is known that these phenomena repeat in a cyclic (though not strictly periodic) manner. In Fig. 1 (left) we present a typical ECG signal, where each sample has been colored according to the corresponding time within the respective cardiac cycle.[3] We then apply the diffusion maps algorithm, with the Scattering Transform as an observer. We take highly overlapping windows (90%) employing an averaging time of 256 samples (roughly half a second, given a sampling frequency is 500 Hz), and we set $L=40$ samples for the estimation of the covariance matrices. We recover the first two eigenvectors from the affinity matrix, depicted in Fig. 1 (right), where we color each embedded sample with the color of its corresponding sample in the time series. As can be seen, the embedded samples are naturally organized into a cycle. Moreover, observations corresponding to samples from similar stages within the cardiac cycle remain close in the embedded space. Thus, these samples are organized into an angular-varying variable in completely data-driven way, demonstrating the accurate recovery of the underlying cardiac state.

3. System classification

Manifold learning methods provide a useful representation of observations from a time-varying system, which can be exploited

[2] This approximation becomes an equality if $d=N$ in Eq. (8).

[3] We denote a cardiac cycle from a QRS complex to the next one.

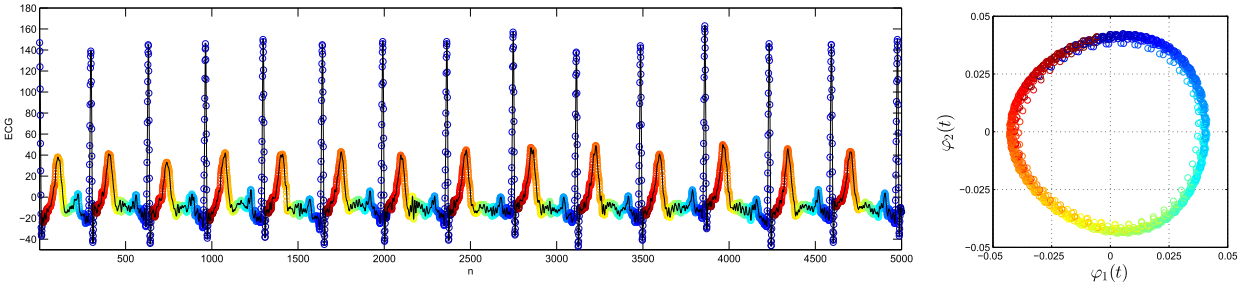

Fig. 1. Analysis of an ECG signal (left) with the diffusion maps algorithm and the Scattering Transform as an observer. The signal is colored according to the stage within each cardiac cycle, and each embedded point (right) is colored accordingly. The embedded points are naturally organized as a function of the stage within each cycle, demonstrating the recovery of the intrinsic variable of the underlying system. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

for classification tasks. More specifically, we leverage the ability of the above discussed algorithm to discriminate different dynamics in a class-specific manner, obtaining a notion of distance between a test signal and the different possible classes in the embedded space.

3.1. Classification framework

Generalizing the above formulation, assume now signals $z_k(t)$ belonging to the systems $\mathcal{S}_{\boldsymbol{\theta}_k}$, controlled by latent variables $\boldsymbol{\theta}_k(t)$, for $k = 1, \ldots, K$, where K is the number of classes. Consider now the problem of assigning a class to a test signal given by $\tilde{z}(t)$. Naturally, we assume $\tilde{z}(t)$ to be generated by a system with latent variable $\tilde{\boldsymbol{\theta}}(t)$, and that $\tilde{\boldsymbol{\theta}}(t) \approx \boldsymbol{\theta}_j(t)$ for some class j.

Consider further that all nonlinear observations through the Scattering Transform are ordered column-wise in the matrix $\boldsymbol{\Phi}\mathbf{z}_k$, for each training signal, and $\boldsymbol{\Phi}\tilde{\mathbf{z}}$ for the test measurements. To employ the diffusion maps algorithm, we construct virtual observation matrices $\boldsymbol{\Lambda}_k$ such that $\boldsymbol{\Lambda}_k = \left[\boldsymbol{\Phi}\mathbf{z}_k, \boldsymbol{\Phi}\tilde{\mathbf{z}}\right] \in \mathbb{R}^{n\times(N_k+N)}$, where N_k and N are the number of observations from class k and from the test signal, respectively. By concatenating the observations in this manner, we are implicitly assuming that there is a corresponding underlying latent variable $\lambda_k(t)$ such that

$$\lambda_k(t) = \begin{cases} \boldsymbol{\theta}_k(t) & \text{if } t \le N_k \\ \tilde{\boldsymbol{\theta}}(t - N_k) & \text{if } t > N_k. \end{cases} \tag{9}$$

Such a construction corresponds to a system governed by the latent variable $\lambda_k(t)$, which behaves according to the dynamics given by $\boldsymbol{\theta}_k(t)$ for $t \le N_k$. Once $t > N_k$, the latent variable changes to $\tilde{\boldsymbol{\theta}}(t - N_k)$, the one of the testing signal, and whose dynamics we want to study. From a different perspective, there is phase transition in the dynamics of λ_k around $t = N_k$, which reflects the evolution of the system from $\boldsymbol{\theta}_k(t)$ to $\tilde{\boldsymbol{\theta}}(t - N_k)$.

We then apply the diffusion maps algorithm to the observations in $\boldsymbol{\Lambda}_k$, obtaining the embedded vectors $\boldsymbol{\zeta}_k(t)$. Denote as $\mathcal{M}_k$ the set of embedded points from the known class $\{\boldsymbol{\zeta}_k(t)\}_{t<N_k}$, and by $\widetilde{\mathcal{M}}_k$ the set of testing points we want to compare to that class, $\{\tilde{\boldsymbol{\zeta}}_k(t)\}_{t>N_k}$. Furthermore, denote by δ_k the distance between these sets, such that $\delta_k = dist\left(\mathcal{M}_k, \widetilde{\mathcal{M}}_k\right)$. This notion of distance, which we will address in the following subsection, serves as an indication of the similarity between the dynamics of a given class and that of the test signal. Indeed, δ_k could be seen as a generalization of the diffusion distance between two samples to the distance between two sets of samples.

The key observation in this approach is that if the diffusion distance $\gamma(\boldsymbol{\theta}_k, \tilde{\boldsymbol{\theta}})$ is small, then $\boldsymbol{\zeta}_k \approx \tilde{\boldsymbol{\zeta}}_k$. In other words, if the latent variables of both systems are similar, the set of corresponding points do not diffuse much from each other, and the distance between the given sets in the embedded space is small. In contrast, if the dynamics between the training and test observations differ significantly, the embedded vectors $\tilde{\boldsymbol{\zeta}}_k$ will diffuse more, indicating the dissimilarity between both dynamics. This way, the classification of a testing signal $\tilde{\mathbf{z}}(t)$ into a specific class is done by repeating the above process for each class, yielding a set of distances δ_k, $i = 1, \ldots, K$. Once done, the correct class is chosen as the one with the smallest distance δ_k. We summarize the proposed approach in Algorithm 1.

Algorithm 1. Dynamical system classification based on diffusion embeddings.

Data: Set of training samples $\left\{z_k(t)\right\}_{1\le k\le K}$ for each of the K classes. Testing signal $\tilde{z}(t)$.

1 Apply the nonlinear operator (the scattering transform) to each signal, obtaining the set of observations

$$\left\{\boldsymbol{\Phi}z_k(t)\right\}_{1\le k\le K} \quad \text{and} \quad \boldsymbol{\Phi}\tilde{z}(t);$$

2 **for** *each class* k **do**

3 Construct the virtual observation matrices

4 $\boldsymbol{\Lambda}_k = \left[\boldsymbol{\Phi}\mathbf{z}_k, \boldsymbol{\Phi}\tilde{\mathbf{z}}\right]$;

5 Apply the diffusion maps algorithm to $\boldsymbol{\Lambda}_k$ obtaining the sets of low–dimensional embedded vectors:

$\mathcal{M}_k = \{\boldsymbol{\zeta}_k(t)\}_{t<N}$ and $\widetilde{\mathcal{M}}_k = \{\tilde{\boldsymbol{\zeta}}_k(t)\}_{t>N}$;

Compute

$\delta_k = dist\left(\mathcal{M}_k, \widetilde{\mathcal{M}}_k\right)$

6 **end**

Result: Chosen class $k^* = \arg\min_k \delta_k$.

3.2. Embedded distances

This method relies on the diffusion maps algorithm to separate the embedded points according to the similarity of their dynamics. Yet, the final classification performance depends on the robustness of the distance according to which we define δ_k. In order to demonstrate the flexibility of this approach, we propose three alternatives to defining this distance, each having its own interpretation and implications.

The first alternative is motivated by an information theory perspective and regards the embedded points as being random processes drawn from a certain distribution. In this context, we propose a method to measure the distance between the distributions of $\mathcal{M}_k$ and $\widetilde{\mathcal{M}}_k$ with the Kullback–Leibler (KL) divergence [33]. This measure is not symmetric, and it is not a formal distance measure. Nevertheless, the KL divergence is often employed to obtain a distance between distributions. Assuming that samples in the training set $\mathcal{M}_k$ come from a distribution Q_k and that the testing samples in $\widetilde{\mathcal{M}}_k$ are drawn from a distribution P_k, the KL

divergence is given by

$$\delta_k^{KL}(P_k \parallel Q_k) = \sum_t P_k(t) \log \frac{P_k(t)}{Q_k(t)}. \tag{10}$$

This measure can be understood from the amount of information loss when samples from P_k are approximated with the prior distribution or model Q_k. In our context, we measure the ability of the training samples in $\mathcal{M}_k$ to model the testing samples in $\widetilde{\mathcal{M}}_k$. In practice, we might not have a sufficient amount of samples to estimate the distributions P_k and Q_k appropriately, and therefore we turn to an estimation of this quantity. For the sake of simplicity, we assume that these distributions are Gaussian, and then we compute $\delta_k^{KL}(P_k \parallel Q_k)$ explicitly by using the corresponding covariance matrices, Σ_Q and Σ_P. In this case, the KL divergence is given by

$$\delta_k^{KL}(P_k \parallel Q_k) = \frac{1}{2}\left[\mathrm{tr}\left(\Sigma_Q^{-1}\Sigma_P\right) + \log\frac{|\Sigma_Q|}{|\Sigma_P|} - d + (\mu_Q - \mu_P)^T\Sigma_Q^{-1}\left(\mu_Q - \mu_P\right)\right], \tag{11}$$

where μ_Q and μ_P are the respective mean vectors.

The second measure is based on a simple conditional probability interpretation. Indeed, given the distribution of the samples $\zeta(t)_k$, we want to address the probability of the samples $\tilde{\zeta}_k(t)$ belonging to such distribution. Here again, the amount of samples motivates us to model the distribution of $\mathcal{M}_k$ using a Gaussian function with covariance Σ_k and mean μ_k, and then to compute the likelihood $p(\tilde{\zeta}_k(t)|\Sigma_k, \mu_k)$. Once the probability of all samples has been computed, several ways of quantifying the distance between the distributions could be proposed. We have observed that the median of these conditional probabilities provides a more robust measure than other characterizations such as the mean. We therefore employ this measure and define the following distance:

$$\delta_k^P = \left(\underset{t}{\mathrm{median}}\left\{p(\tilde{\zeta}_k(t)|\Sigma_k, \mu_k)\right\}\right)^{-1}. \tag{12}$$

Lastly, another perspective is to construct a distance measure based on the (in)ability of a classifier to discriminate between the different sets, in a more pragmatic approach. Under this point of view, if a classifier cannot discriminate samples from different distributions, we might conclude that these distributions are very close. On the contrary, if a perfect classification is obtained, we can define the distance between them as infinite. Formally, given the linear classifier $\mathbf{w}_k$, we define

$$\delta_k^{Cls} = \left(\frac{1}{N_c}\sum_t |y_t - \mathrm{sign}(\mathbf{w}_k^T\zeta_k(t) + b_k)|\right)^{-1}, \tag{13}$$

where y_t denotes the class and N_c is the number of vectors $\zeta_k(t)$. Recall that in this context, the class refers to either $\mathcal{M}_k$ or $\widetilde{\mathcal{M}}_k$. Also, the reader should keep in mind that this is not a classical classification setup, and that the objective is to evaluate the ability of a linear classifier to discriminate between the two groups. For this reason we train and evaluate the classifier on all samples ζ_k and $\tilde{\zeta}_k$, and then compute the classification error on the same data. In particular, for the sake of simplicity, we use Fisher's Linear Discriminant [34,35] as a classifier $\mathbf{w}_k$. It is worth noting that this classifier makes the same assumptions that we used in modeling the distributions of the training and testing samples for the computation of the KL divergence; i.e., that the conditional probability distributions are normally distributed.

4. Experimental results

In this section, we first consider a toy problem to demonstrate our approach in a controlled setup, and then move to the more challenging problem of subject identification from ECG data.

4.1. Synthetic example

Let us first consider the toy problem of a time variant first order autoregresive system with time deformations, which has been used previously in a variety of applications, e.g., in [36,37]. The system is characterized by the latent variable θ_k, and outputs the time series $z_k(t)$. In discrete time $t = 1, \ldots, T$, this system is given by

$$\begin{cases} x_k(t) = v(t)\ u(t) + (\theta_k + w_1(t))\ x_k(t-1) \\ z_k(t) = x_k(t - g(t)), \end{cases} \tag{14}$$

where $v(t)$ is a nuisance factor given by

$$v(t) = 0.95 + 0.1\sin(2\pi t/T), \tag{15}$$

$u(t)$ is a white Gaussian driving process and the variable $g(t)$ controls the time deformation of the system. We set this variable to be

$$g(t) = 0.1 + 0.4*(t/T) + 0.05*w_2(t). \tag{16}$$

The system includes measurement (white Guassian) noise given by $w_1(t)$ and $w_2(t)$. We obtain 5 realizations of this system for $T = 10,000$, each with a different latent variable θ_k in the range $(-1,1)$. In particular, we employ $\theta_k = [-0.8, -0.6, 0, 0.6, 0.8]$. The training signals, for each value of the intrinsic parameter, are depicted on the left side of Fig. 2. The testing signals are obtained through 5 other different realizations. We then apply Algorithm 1, employing an averaging time of 32 samples, and a $L=5$ samples for the estimation of the covariance matrix.

In this synthetic example all distance measures, δ_k^{KL}, δ_k^{P} and δ_k^{Cls}, manage to identify the corresponding latent variable in all cases. To provide some insight into how the proposed method works, on the right side of Fig. 2, we present the embeddings obtained for a test signal corresponding to θ_2. Specifically, we plot the embedded samples ζ_k (in blue) and $\tilde{\zeta}_k$ (in red) for each $k = 1, \ldots, 5$, where the testing signal is taken with $\tilde{\theta} = \theta_2$. We observe that the further the value of the latent variable of the training signal from that of θ_2, the clearer the separation becomes in the embedded space. In contrast, when the test signal is compared to the training signal corresponding to the same parameter, the distributions are practically equal.

These results provide a visual explanation of the distances measured between the samples $\tilde{\zeta}_k$ and ζ_k, shown in Fig. 3. There, we see how all distances present a minimum at $\theta_k = \theta_2$, indicating the correct match. Moreover, these distances increase as the difference between the training and the testing latent variable becomes larger. In this scenario, the δ_k^{KL} seems to provide a relatively more robust measure, judging by the difference between its minimum and its other values. This reflects the visual interpretation of the difference between the distributions observed in the embedded space in Fig. 2. In contrast, δ_k^{Cl} and δ_k^{Cl} present a milder decrease, though still they achieve 100% identification accuracy in all cases.

4.2. Subject identification through ECG

We now apply the proposed approach to the problem of person identification by performing recognition on electrocardiographic (ECG) signals.

We examine the ECG-ID database,[4] previously used for this task by several authors [23,38,39]. This database comprises 310 ECG

[4] This database is made freely available through http://physionet.org/.

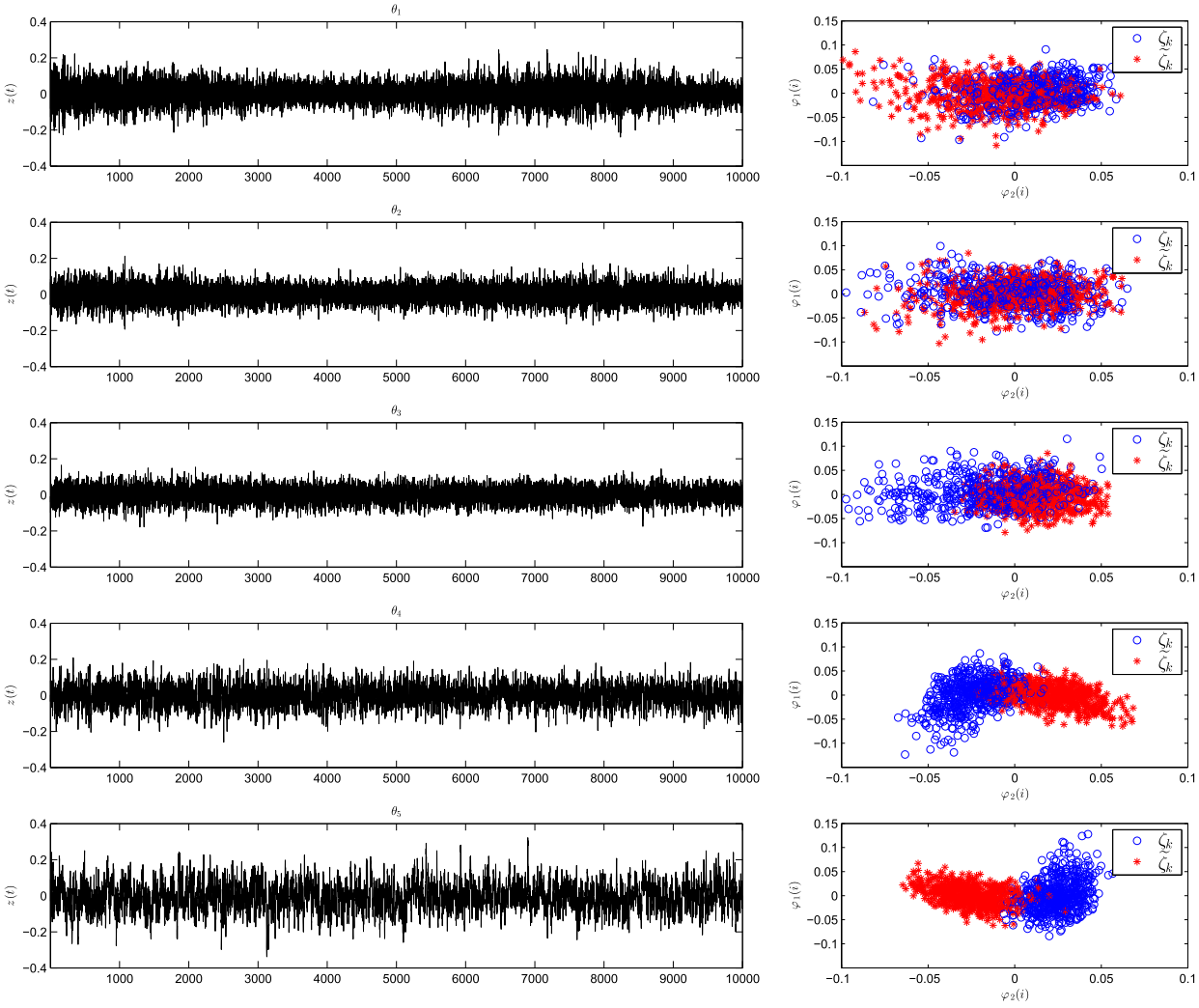

Fig. 2. Synthetic experiment. Left: 5 realizations of the system in (14) for different values of the parameter θ. Right: The embeddings resulting from comparing a test signal (not shown here) corresponding to θ_2 against the training signal for each case. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

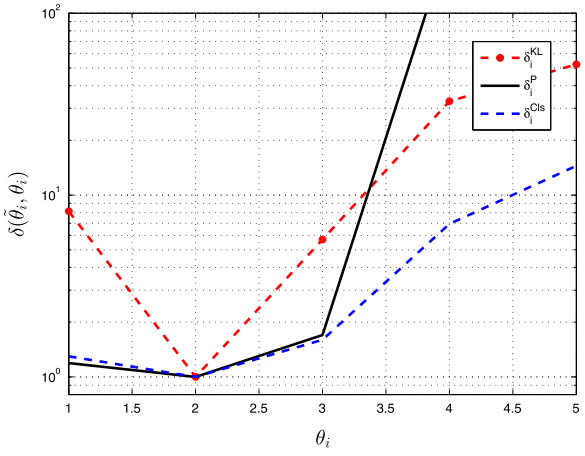

Fig. 3. Distances δ_k^{KL}, δ_k^{P} and δ_k^{Cls} obtained in the Synthetic experiment for the case when the testing signal corresponds to θ_2.

recordings, obtained from 90 subjects at different times and different conditions. The records were digitalized at 500 Hz, and they are 20 seconds long. The number of records for each person varies from 2 (collected during one day) to 20 (collected periodically over 6 months). The signals have gone through some basic preprocessing to remove baseline-drift, high frequency and power-line noise. However, even after this basic filtering many signals contain substantial noise and high-amplitude artifacts.

To provide some intuition on the proposed approach applied to ECG identification, we first present a binary classification example to show how in this case as well, where the underlying system is very complex and with unknown parameters, the proposed approach correctly discriminates between different dynamics. In Fig. 4(a) and (c) we present the ECG of two subjects, together with a testing signal in Fig. 4(b), which belongs to the second subject. We employ these records in particular for this demonstration since the noise in the signal of Subject 2 degrades the training data significantly, making the classification task challenging.

Even though it is hard or even impossible to distinguish visually the subject to which the testing signal belongs, the diffusion maps algorithm, together with the invariance of the Scattering Transform, manages to discriminate between them quite well. When the observations from the test signal are embedded with those corresponding to Subject 1, in Fig. 4(d), there is a natural separation between the low-dimensional samples $\boldsymbol{\zeta}(t)$ (in blue asterisks) and $\tilde{\boldsymbol{\zeta}}(t)$ (in black circles). Conversely, the separation between the test embedded samples and those from Subject 2 (in red asterisks), in Fig. 4(e), is clearly smaller, indicating that this is in fact the correct original *system*, or subject.

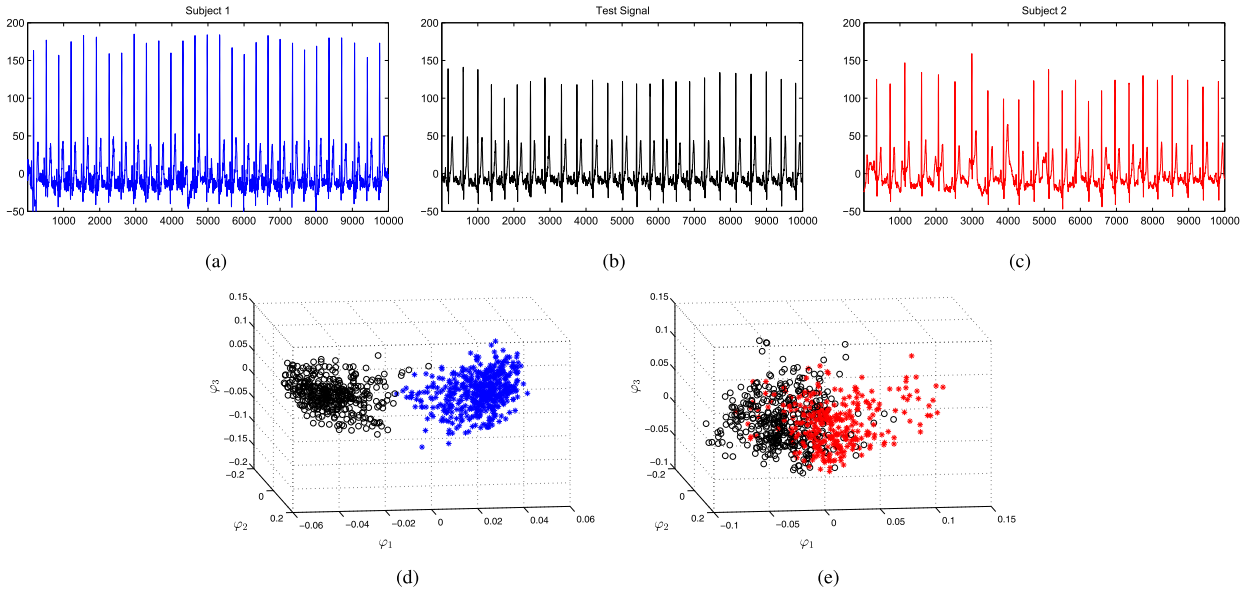

Fig. 4. ECG person identification demonstration. (a) and (c) are two training signals from Subject 1 and Subject 2, respectively. (b) is a test signal from Subject 2. (d) shows the first three components of the embedded samples $\xi_1(t)$ and $\tilde{\xi}_1(t)$, and (e) shows the corresponding samples $\xi_2(t)$ and $\tilde{\xi}_2(t)$. While identifying the correct subject from the test signal might seem difficult (or even impossible) by visual assessment, the proposed approach yields a clear result which indicates that the testing embedded samples (in black circles) are closer to the dynamics of Subject 2 compared to Subject 1 (in asterisks). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

When analyzing the entire dataset, few additional considerations are needed. Due to the inhomogeneity in the number of signals per subject, we employ the following procedure to select the training and testing data, which we believe is a feasible scenario in practical applications: if there are 6 or more records for that subject, we select at random up to 5 training signals per subject. Otherwise we take less than 5 to guarantee that at least one signal per subject is not included in the training set, and it is used as a testing signal for that patient. This results in a training set comprising 180 records. This sampling procedure is repeated 20 times in order to introduce variability in the training and testing conditions. We report the average accuracy, defined as the number of correct classification results over the total number of testing signals, together with their corresponding statistics.

We note that many signals contain blank intervals (where the electrodes were probably disconnected). In addition, severe artifacts might occlude substantial information in other recordings. To avoid such corrupted data, we implement a basic version of the Pan–Tompkins algorithm [40] to identify QRS complexes, and only employ data where such complexes are detected.[5] Note, however, that we do not perform a QRS segmentation as done in other works [22,23]. The QRS detection is just a mean to select informative data in an automated way, and the signal is analyzed as one complete signal vector. Each training signal is analyzed by applying the Scattering Transform, with 50% overlapping windows of 128 samples (roughly a quarter of a second). Each observation vector results in a dimension of $n=157$. We employ $L=10$ samples to compute the covariance matrices, and use a dimension $d=10$ for the diffusion embedding.

Regarding computational time, note that once the observations vectors have been gathered, the classification procedure amounts to applying the diffusion maps algorithm to the virtual observation matrices $\mathbf{\Lambda}_k$ for each class (or subject), and computing the respective distances. In the current setup, with 90 classes, this takes approximately 4 minutes per testing signal with unoptimized code in Matlab, in a PC with an Intel Core i7 CPU and 16 Gb of RAM. Note, however, that our method is highly parallelizable, as the diffusion maps algorithm can be applied to each $\mathbf{\Lambda}_k$ independently.

We compare our approach with a classifier composed of a multi-class (linear) support vector machine (SVM). We train such a classifier on the observation vectors obtained through the Scattering Transform, exhibiting the benefit of shift and deformation invariance but not the discriminative power of the diffusion maps algorithm. We train the SVM on the same collection of training data described above, and then run the classifier on the same testing signals employed by our method. To provide a more complete picture of the performance that can be achieved by traditional classifiers, we include the results obtained by a multi-class kernel SVM, where we have used a polynomial kernel of degree 3. These nonlinear classifiers are able to provide more complex hyperplanes and significantly better classification accuracy.

Several points can be drawn from the results, presented in Fig. 5. First of all, the Scattering Transform seems to be very effective in providing a good representation for the ECG signals. This enables a simple linear SVM classifier to obtain an accuracy of 77.64%, in average. If instead we use a multi-class kernel SVM, this accuracy is boosted to 91.12%. Our algorithm achieves the highest classification results when used with the δ_k^{Cls} distance, reaching a classification accuracy of 97.25%. Note that this is remarkable given the very simple way of computing the distance in the embedded space – especially when compared to kernel SVM. Moreover, this is better than the best reported result for this database of 96% [23].

We also report that the distance based on the conditional likelihood performs just as good as the Kernel SVM. Surprisingly, the KL divergence does not enable a very effective classification in this case, yielding similar results to that of the linear SVM. We conjecture that the amount of samples in the embedded space seriously limits the performance of these two probabilistic measures, in particular the KL divergence.

We conclude this section with a comment regarding the results reported in [23] for the same database. Briefly, the authors employ a QRS segmentation algorithm per subject, extracting these

[5] If QRS complexes are not detected, we filter the signal discarding the approximation and last (highest frequency) detail coefficients of a 10 scales wavelet decomposition in order to attenuate possible artifacts. If QRS complexes are not detected after this basic filtering, the record is discarded.

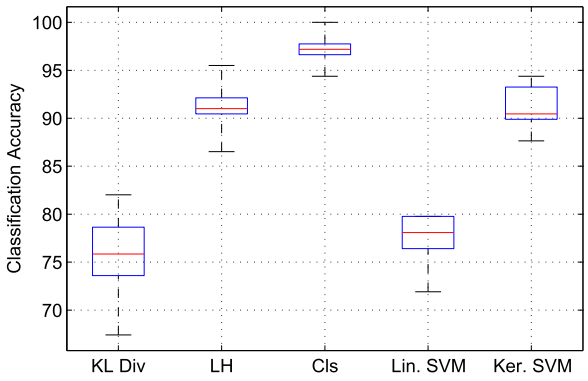

Fig. 5. General classification results for the problem of subject identification. First three methods are versions of the framework presented in this work. The first one corresponds to the KL divergence measure, while the second and third to the Conditional Likelihood and Fisher Linear Discriminant, respectively. The last two are multi-class SVM (linear and polynomial kernel) classifiers. The whiskers denote maximum and minimum values, the blue box limits the 25th and 75th percentile, and the red line corresponds to the median for $n=20$ realizations. (For interpretation of the references to color in this figure caption,the reader is referred to the web version of this paper.)

segments and discarding those that deviate from the average behavior. They then construct a number of morphological features from the processed QRS complexes after a correction or modulation that depends on the measured the heart rate. In the final classification stage, their method employs a Linear Discriminant Analysis after a dimensionality reduction. While training on a slightly higher number of samples (195 records over all, whereas we train on 180), the authors in [23] report a classification accuracy of 96%. In contrast to [23], our method receives the input signals *as is*; i.e., they are not pre-processed (segmented, normalized, etc.) and they do not undergo a feature-extraction process. Instead, it is up to our algorithm to extract relevant features or information to aid the classification process. This demonstrates the great benefit and power of data-driven algorithms, enabling to perform just as well and even better than other, significantly more elaborate and complex, classification schemes.

5. Conclusions

We have presented a classification method based on the diffusion maps algorithm and the Scattering Transform. Leveraging the ability of such manifold learning method to aggregate or separate signal samples as a function of their dynamics, we employ these ideas to the problem of dynamical system identification. The proposed approach is general to the extent that different concepts or notions can be used to quantify the distance between the low-dimensional embedded samples, in a class-specific manner. We demonstrate this by employing three distance measures.

The proposed algorithm is applied first to a synthetic example, showing how these distance measures correlate with the distance of the underlying latent variable. Moreover, we show its applicability to real signals in the problem of subject identification from ECG signals. When employed with the classification-based distance, our approach outperforms popular classification algorithms and achieves the highest reported results for the database employed. Other distance measures provide a lower classification accuracy, probably suffering from the relatively small number of samples in the embedded space in this particular application. We believe that other, more sophisticated, definitions or ways to quantify the similarity between the embedded vectors will boost the performance of the proposed scheme, not only in the problem of ECG identification but also in other complex classification applications. These, and other ideas, are the subject of current ongoing work.

References

[1] H. Park, S. Yun, S. Park, J. Kim, C.D. Yoo, Phoneme classification using constrained variational Gaussian process dynamical system, in: Advances in Neural Information Processing Systems, 2012, pp. 2006–2014.
[2] X.-W. Wang, D. Nie, B.-L. Lu, Emotional state classification from EEG data using machine learning approach, Neurocomputing 129 (2014) 94–106.
[3] C. Lainscsek, T.J. Sejnowski, Electrocardiogram classification using delay differential equations, Chaos: Interdiscip. J. Nonlinear Sci. 23 (2) (2013) 023132.
[4] K.Q. Shen, C.J. Ong, X.P. Li, Z. Hui, E.P.V. Wilder-Smith, A feature selection method for multilevel mental fatigue EEG classification, IEEE Trans. Biomed. Eng. 54 (7) (2007) 1231–1237.
[5] K. Samiee, P. Kovács, M. Gabbouj, Epileptic seizure classification of EEG time-series using rational discrete short-time Fourier transform, IEEE Trans. Biomed. Eng. 62 (2) (2015) 541–552.
[6] Y. Hu, D. Wu, A. Nucci, Fuzzy-clustering-based decision tree approach for large population speaker identification, IEEE Trans. Audio Speech Lang. Process. 21 (4) (2013) 762–774.
[7] J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[8] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[9] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[10] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, S.W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps, Proc. Natl. Acad. Sci. U. S. A. 102 (21) (2005) 7426–7431.
[11] R.R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (2006) 5–30.
[12] A. Singer, R. Erban, I.G. Kevrekidis, R.R. Coifman, Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps, Proc. Natl. Acad. Sci. 106 (38) (2009) 16090–16095.
[13] V. Chudacek, R. Talmon, J. Anden, S. Mallat, R. Coifman, P. Abry, M. Doret, Low dimensional manifold embedding for scattering coefficients of intrapartum fetale heart rate variability, in: Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, 2014, pp. 6373–6376.
[14] R. Talmon, S. Mallat, H. Zaveri, R.R. Coifman, Manifold learning for latent variable inference in dynamical systems, IEEE Trans. Signal Process. 63 (15) (2015) 3843–3856.
[15] S. Mallat, Group invariant scattering, Commun. Pure Appl. Math. 65 (10) (2012) 1331–1398.
[16] V. Chudácek, J. Andén, S. Mallat, P. Abry, M. Doret, Scattering transform for intrapartum fetal heart rate characterization and acidosis detection, in: Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, IEEE, Osaka, 2013, pp. 2898–2901.
[17] C. Baugé, M. Lagrange, J. Andén, S. Mallat, Representing environmental sounds using the separable scattering transform, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Vancouver, BC, 2013, pp. 8667–8671.
[18] I. Odinaka, P.-H. Lai, A.D. Kaplan, J.A. O'Sullivan, E.J. Sirevaag, J.W. Rohrbaugh, ECG biometric recognition: a comparative analysis, IEEE Trans. Inf. Forensics Secur. 7 (6) (2012) 1812–1824.
[19] L. Biel, O. Pettersson, L. Philipson, P. Wide, ECG analysis: a new approach in human identification, IEEE Trans. Instrum. Meas. 50 (3) (2001) 808–812.
[20] J.C. Sriram, M. Shin, T. Choudhury, D. Kotz, Activity-aware ECG-based patient authentication for remote health monitoring, in: Proceedings of the 2009 International Conference on Multimodal Interfaces, ACM, New York, NY, USA, 2009, pp. 297–304.
[21] S.A. Israel, J.M. Irvine, A. Cheng, M.D. Wiederhold, B.K. Wiederhold, ECG to identify individuals, Pattern Recognit. 38 (1) (2005) 133–142.
[22] J.M. Irvine, S.A. Israel, A sequential procedure for individual identity verification using ECG, EURASIP J. Adv. Signal Process. 2009 (1) (2009) 1–13.
[23] T. Lugovaya, Biometric human identification based on electrocardiogram (Master's thesis), Faculty of Computing Technologies and Informatics, Electrotechnical University "LETI", Saint-Petersburg, Russian Federation, 2005.
[24] G. Wübbeler, M. Stavridis, D. Kreiseler, R.-D. Bousseljot, C. Elster, Verification of humans using the electrocardiogram, Pattern Recognit. Lett. 28 (10) (2007) 1172–1175, ISSN 0167-8655, , URL ⟨http://www.sciencedirect.com/science/article/pii/S0167865507000463⟩.
[25] M. Li, S. Narayanan, Robust ECG biometrics by fusing temporal and cepstral information, in: 2010 20th International Conference on Pattern Recognition (ICPR), IEEE, Istanbul, 2010, pp. 1326–1329.
[26] A.D. Chan, M.M. Hamdy, A. Badre, V. Badee, Wavelet distance measure for person identification using electrocardiograms, IEEE Trans. Instrum. Meas. 57 (2) (2008) 248–253.
[27] S.Z. Fatemian, D. Hatzinakos, A new ECG feature extractor for biometric recognition, in: 2009 16th International Conference on Digital Signal Processing, IEEE, Santorini-Hellas, 2009, pp. 1–6.
[28] A. Singer, R.R. Coifman, Non-linear independent component analysis with

diffusion maps, Appl. Comput. Harmon. Anal. 25 (2008) 226–239.
[29] J. Bruna, S. Mallat, Invariant scattering convolution networks, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1872–1886.
[30] J. Anden, S. Mallat, Deep scattering spectrum, IEEE Trans. Signal Process. 62 (16) (2014) 4114–4128.
[31] C.J. Dsilva, R. Talmon, R.R. Coifman, I.G. Kevrekidis, Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study, Appl. Comput. Harmon. Anal., 2015.
[32] W. Lian, R. Talmon, H. Zaveri, L. Carin, R. Coifman, Multivariate time-series analysis and diffusion maps, Signal Process. 116 (2015) 13–28.
[33] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86.
[34] R.A. Fisher, The statistical utilization of multiple measurements, Ann. Eugen. 8 (4) (1938) 376–386.
[35] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, 2013.
[36] R.W. Schafer, L.R. Rabiner, Digital representations of speech signals, Proc. IEEE 63 (4) (1975) 662–667.
[37] T.F. Quatieri, Discrete-time Speech Signal Processing: Principles and Practice, Pearson Education, India, 2002.
[38] N. Belgacem, R. Fournier, A. Nait-Ali, F. Bereksi-Reguig, A novel biometric authentication approach using ECG and EMG signals, J. Med. Eng. Technol. 39 (4) (2015) 226–238.
[39] G. Altan, Y. Kutlu, ECG based human identification using logspace grid analysis of second order difference plot, in: Signal Processing and Communications Applications Conference (SIU), 2015 23th, IEEE, Malatya, 2015, pp. 1288–1291.
[40] J. Pan, W.J. Tompkins, A real-time QRS detection algorithm, IEEE Trans. Biomed. Eng. (3) (1985) 230–236.