Appl. Comput. Harmon. Anal. • • • (• • • •) • • •–• • •

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Intrinsic modeling of stochastic dynamical systems using empirical geometry

Ronen Talmon *, Ronald R. Coifman

Department of Mathematics, Yale University, New Haven 06520, CT, United States

ARTICLE INFO

Article history:
Received 16 September 2013
Received in revised form 18 August 2014
Accepted 31 August 2014
Available online xxxx
Communicated by Radu Balan

Keywords:
Intrinsic model
Nonlinear inverse problem
Differential geometry
Information geometry
Non-parametric estimation
Nonlinear dynamical systems
Manifold learning
Graph-based method

ABSTRACT

In a broad range of natural and real-world dynamical systems, measured signals are controlled by underlying processes or drivers. As a result, these signals exhibit highly redundant representations, while their temporal evolution can often be compactly described by dynamical processes on a low-dimensional manifold. In this paper, we propose a graph-based method for revealing the low-dimensional manifold and inferring the processes. This method provides intrinsic models for measured signals, which are noise resilient and invariant under different random measurements and instrumental modalities. Such intrinsic models may enable mathematical calibration of complex measurements and build an empirical geometry driven by the observations, which is especially suitable for applications without a priori knowledge of models and solutions. We exploit the temporal dynamics and natural small perturbations of the signals to explore the local tangent spaces of the low-dimensional manifold of empirical probability densities. This information is used to define an intrinsic Riemannian metric, which in turn gives rise to the construction of a graph that represents the desired low-dimensional manifold. Such a construction is equivalent to an inverse problem, which is formulated as a nonlinear differential equation and is solved empirically through eigenvectors of an appropriate Laplace operator. We examine our method on two nonlinear filtering applications: a nonlinear and non-Gaussian tracking problem as well as a non-stationary hidden Markov chain scheme. The experimental results demonstrate the power of our theory by extracting the underlying processes, which were measured through different nonlinear instrumental conditions, in an entirely data-driven nonparametric way.

1. Introduction

Due to natural constraints, in a broad range of real-world applications, the accessible (high-dimensional) data exhibit typical structure and often lie on a (low-dimensional) manifold. In recent years, this observation gave rise to the development of manifold learning methods, which aim at finding parameterizations of the underlying low-dimensional structures of given data sets [1–5]. A similar observation applies to dynamical

* Corresponding author.
E-mail addresses: ronen.talmon@yale.edu (R. Talmon), coifman@math.yale.edu (R.R. Coifman).

http://dx.doi.org/10.1016/j.acha.2014.08.006
1063-5203/

systems: the measured output signal of the system is often controlled by few underlying drivers, whose temporal evolution can be compactly described by dynamical processes on a low-dimensional manifold [6–10]. This belief is naturally encoded in the standard state-space formalism used to describe dynamical systems: given signal measurements $\mathbf{z}_t$ in time t, we are interested in estimating the associated system state $\boldsymbol{\theta}_t$. We remark that the focus of this paper will be on the state estimation problem, which is typically extended to filtering, forecasting and prediction, as well as noise suppression problems by attaching statistical models in a Bayesian manner.

To support the analysis of time series and dynamical systems, the standard geometric setting of manifold learning needs to be extended. First, the mapping between the measured signal and the underlying processes is often stochastic and contains measurement noise. As a result, repeated measurements of the same phenomenon yield different measurement realizations. Furthermore, the measurements may be acquired using different instruments or sensors. Each set of related measurements of the same phenomenon will then have a different manifold, depending on the instrument and the specific realization. Thus, to provide meaningful information on the true state of the system, the geometric parameterization should be invariant to the measurement and instrumental modalities. Second, the dynamics of the signals carry essential information and should be encoded in the analysis results. Third, the ability to sequentially analyze streaming data is an important aspect of dynamical systems and signal processing. When a stream of new incoming signal samples becomes available, the manifold model needs to be efficiently extended.

In [11], Singer and Coifman addressed the problem in which the desired "interesting" data are accessible via an unknown nonlinear measurement function. To provide a model for the desired data, rather than the accessible measurements, the Mahalanobis distance was used to invert the measurement function locally, assuming that the function that maps the data into a set of measurements is deterministic and stably invertible on its range. As a result, their approach provides a model for the desired manifold, whereas classic manifold learning methods provide a model for the manifold of the observations.

In this paper, we extend [11] and propose a graph-based method for the parameterization of the underlying processes controlling stochastic dynamical systems, which we refer to as empirical intrinsic geometry (EIG). The primary focus of the paper is on the construction of an intrinsic Riemannian distance metric between measurement samples that exhibits the desired invariance to the measurement modality and noise. The construction of the metric is carried out in two scales of short-time windowing. In a micro-scale, local probability densities of the measurements are estimated in windows and viewed as descriptors, or features, where the sampling rate of the measurements is assumed to be sufficiently high to allow for an accurate densities estimation. We will motivate this particular choice of features by showing that any stationary measurement noise in the signal domain is translated to a linear operation in the probability densities domain. In a macro-scale, we exploit the temporal dynamics and natural small perturbations of the underlying process to explore the local tangent spaces to the manifold by estimating the covariances of the probability densities in time windows. Here we further assume that the signal is pseudo-stationary, i.e., its probability density is slowly changing with time. Based on the estimates of the local probability densities and their covariances, we compute the Mahalanobis distance [11]. Since the Mahalanobis distance is invariant to linear transformations, and since any measurement noise is translated to a linear transformation in the domain of probability densities, the constructed distance metric is invariant to moderate noise.

Despite drawing most of our attention in the present work, the Mahalanobis distance is an intermediate analysis result, since it provides merely a *local* Euclidean structure. Nevertheless, the availability of such an intrinsic distance metric enables us to construct diffusion geometry [5] via a graph that represents the intrinsic manifold of underlying processes. This construction is shown to be equivalent to an inverse problem, which is solved empirically through eigenvectors of an appropriate Laplace operator. Specifically, the eigenvectors provide an embedding (or a parametrization) of the underlying processes on the intrinsic manifold. We remark that the construction of the graph is implemented using a reference set of measurements

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

[12,13]. This allows for the parameterization of a training signal in advance, and in turn, for the extension of the parameterization to newly acquired signal samples in a sequential manner.

Our approach is tightly connected to two lines of studies. The short-time windowing implies that the obtained parameterization captures the slow drivers of the dynamical system, thereby establishing a close connection to slow feature analysis (SFA) [14]. In addition, the key element in our methodology is the intersection between the dynamics and the geometric structure of the data. This relationship has been extensively investigated in many information geometry studies [15]. However, as opposed to traditional information geometry, we present a *data-driven* methodology to learn the intrinsic manifold of empirical probability densities.

Experimental results of the application of this modeling method to two nonlinear filtering applications will be presented. One is a nonlinear and non-Gaussian tracking problem that has been inspired by a variety of nonlinear filtering studies in the areas of maneuvering targets and financial data processing [16,17]. We show that the obtained model represents the underlying process and is indeed noise resilient and invariant to the measurement function. Two is a non-stationary hidden Markov chain toy problem. In this case, we demonstrate the ability of our approach to provide appropriate modeling for Markovian measurements with memory.

We note that this work was presented in part in [18], where in addition a non-parametric Bayesian filtering framework was defined based on the obtained data-driven model. The Bayesian framework enables to optimally filter, estimate and predict the underlying process, demonstrating the effectiveness of this approach in providing empirical models for real signals without existing definitive models.

The remainder of this paper is organized as follows. Section 2 presents the problem formulation. In Section 3, the empirical probability densities of the signal are proposed as features and their properties are presented. In Section 4, the intrinsic distance metric is derived and its relationship to information geometry is established. In Section 5, we describe the graph-based method to parameterize the underlying processes. Finally, in Section 6, experimental results are presented, demonstrating the performance of this technique.

2. Problem formulation

We utilize the state-space approach to formulate the observed signal as the output of a dynamical system. The state-space formalism includes two models: the dynamics model, which consists of a stochastic differential equation describing the evolution of an underlying process (state) with time, and the measurement model, which relates the noisy observations to the underlying process.

Let $\boldsymbol{\theta}_t$ be a d-dimensional underlying process in time index t. The dynamics of the process are described by normalized stochastic differential equations (written in local coordinates) as follows[1]

$$d\theta_t^i = a^i\left(\theta_t^i\right)dt + dw_t^i, \quad i = 1, \ldots, d, \tag{1}$$

where a^i are unknown drift functions and w_t^i are standard Brownian motions.

Let $\mathbf{y}_t$ denote an n-dimensional observation process in time index t, drawn from a conditional probability density function (pdf) $f(\mathbf{y}|\boldsymbol{\theta})$. The statistics of the observation process are time-varying and depend upon the underlying process $\boldsymbol{\theta}_t$. We consider a model in which the "interesting" observation process (related to the underlying process $\boldsymbol{\theta}_t$) is accessible only via a noisy n-dimensional measurement process $\mathbf{z}_t$, given by

$$\mathbf{z}_t = g(\mathbf{y}_t, \mathbf{v}_t) \tag{2}$$

where g is an unknown (possibly nonlinear) measurement function and $\mathbf{v}_t$ is a corrupting n-dimensional measurement noise, drawn from an unknown stationary pdf $q(\mathbf{v})$ and independent of $\mathbf{y}_t$.

[1] x^i denotes access to the ith coordinate of a vector $\mathbf{x}$.

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

The underlying process $\boldsymbol{\theta}_t$ constitutes a manifold in d dimensions, and its dynamics drive the dynamical system. Hence, it is viewed as the "natural" intrinsic state of the system. Our goal in this work is to empirically discover $\boldsymbol{\theta}_t$ and its dynamics based on a sequence of measurements $\mathbf{z}_t$ without prior knowledge of the dynamics model, the measurement model, and the model of the underlying state. This distinguishes the present work from many Bayesian algorithms, e.g. the well-known Kalman filter and its extensions [19–21], and various sequential Monte Carlo algorithms [22–24], in which the knowledge of these models is required.

We remark that the dynamics model (1) includes locally independent coordinates and standard Brownian motions. This is a critical assumption in the presented analysis, which does not necessarily hold in real dynamical systems and practical applications. Thus, the inferred data-driven model will only approximate the observed dynamical system as a system with such underlying dynamics.

3. Empirical local densities

Let $p(\mathbf{z}|\boldsymbol{\theta})$ denote the conditional pdf of the measured process $\mathbf{z}_t$ controlled by $\boldsymbol{\theta}_t$, which satisfies the following property.

Lemma 1. *The pdf of the measured process $\mathbf{z}_t$, i.e., $p(\mathbf{z}|\boldsymbol{\theta})$, is given by a linear transformation of the pdf of the clean observation component $\mathbf{y}_t$, i.e., $f(\mathbf{y}|\boldsymbol{\theta})$.*

Proof. The proof is straightforward. By relying on the independence of $\mathbf{y}_t$ and $\mathbf{v}_t$, the pdf of the measured process is given by

$$p(\mathbf{z}|\boldsymbol{\theta}) = \int_{g(\mathbf{y},\mathbf{v})=\mathbf{z}} f(\mathbf{y}|\boldsymbol{\theta}) q(\mathbf{v}) d\mathbf{y} d\mathbf{v}. \qquad \square \tag{3}$$

For example, in case of additive measurement noise, i.e., $g(\mathbf{y},\mathbf{v}) = \mathbf{y} + \mathbf{v}$, only a single solution $\mathbf{v}(\mathbf{z}) = \mathbf{z} - \mathbf{y}$ exists. Thus, $p(\mathbf{z}|\boldsymbol{\theta})$ is given by convolution

$$p(\mathbf{z}|\boldsymbol{\theta}) = \int_{\mathbf{y}} f(\mathbf{y}|\boldsymbol{\theta}) q(\mathbf{z}-\mathbf{y}) d\mathbf{y} = f(\mathbf{z}|\boldsymbol{\theta}) * q(\mathbf{z}).$$

In other words, Lemma 1 states that any stationary measurement noise $\mathbf{v}$, which is independent of the observation part $\mathbf{y}$, is given by a linear transformation in the pdf domain.

Since the model of the underlying process, its dynamics, and the measurement model are unknown, the pdfs are unknown as well. In this work, we use histograms as empirical estimates of the pdfs; histograms have been shown to be powerful descriptors and have been used in a variety of fields, e.g., computer vision applications [25] and 3D object recognition [26]. Let $\mathbf{h}_t$ be the local histogram of the measured process $\mathbf{z}_t$ in a short-time window of length L_1 centered at time t. Let $\mathcal{Z}$ be the sample space of $\mathbf{z}_t$ and let $\mathcal{Z} = \bigcup_{j=1}^{m} \mathcal{H}_j$ be a finite partition of $\mathcal{Z}$ into m disjoint histogram bins. Thus, the value of each histogram bin is given by

$$h_t^j = \frac{1}{L_1} \sum_{s \in \mathcal{I}_t} \mathbf{1}_{\mathcal{H}_j}(\mathbf{z}_s), \tag{4}$$

where $\mathbf{1}_{\mathcal{H}_j}(\mathbf{z}_t)$ is the indicator function of the bin $\mathcal{H}_j$, and $\mathcal{I}_t$ is a discrete grid between $t - L_1/2 + 1$ and $t + L_1/2$. By assuming that the density of the samples in each histogram bin is uniform, the expected value of (4) is given by

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

$$\mathbb{E}_{\mathbf{z}}\left[h_t^j\right] = \mathbb{E}_{\mathbf{z}}\left[\mathbf{1}_{\mathcal{H}_j}(\mathbf{z})\right] = \int\limits_{\mathbf{z}\in\mathcal{H}_j} p(\mathbf{z}|\boldsymbol{\theta})d\mathbf{z}, \tag{5}$$

thereby implying that the histograms are *linear transformations* of the pdf. In addition, in the limit, when we shrink the bins of the histograms, the expected values of the histograms converge point-wise to the pdf

$$\frac{1}{|\mathcal{H}_j|}\mathbb{E}[\mathbf{h}_t] \xrightarrow{|\mathcal{H}_j|\to 0} p(\mathbf{z}|\boldsymbol{\theta}), \tag{6}$$

where $|\mathcal{H}_j|$ is the cardinality of the jth bin.

Combining Lemma 1 and (5) yields the following results.

Corollary 2. *The expected value of the histogram* $\mathbb{E}[\mathbf{h}_t]$ *is given by a linear transformation of the pdf of the clean observation component* $\mathbf{y}_t$.

In other words, any measurement noise is expressed as a linear transformation in the domain of the expected values of histograms.

Corollary 3. *The expected value of the histogram* $\mathbb{E}[\mathbf{h}_t]$ *is a deterministic nonlinear function of the underlying process* $\boldsymbol{\theta}_t$.

These properties of the histograms suggest that histograms are "good" observers (or features) of the measurements. A few important remarks are due at this point. First, these properties are not unique to histograms. In fact, any other linear transformation of the pdf (or expected value of a linear function of the measurements) may be viewed as an alternative observer that maintain the described properties. In this paper, we use histograms for simplicity. Second, since the computation of high-dimensional histograms is infeasible, we may preprocess high-dimensional data by applying random filters in order to reduce the dimensionality without corrupting the information [27]. In addition, each histogram bin can be viewed as an "independent observer", and hence, merely few histogram bins may be sufficient to convey the necessary information. Third, the empirical densities can be more accurately estimated using histograms with overlapping non-rectangle bins and kernels. And forth, the choice of the window of length L_1 in which the local histograms are estimated is of particular importance and represents a standard "bias-variance" tradeoff: a longer window yields a more accurate estimation at the expense of a bias caused by the time variation of the pdfs within the window. These remarks extend the scope of this paper. See [28] for more details.

4. Intrinsic metric computation using empirical information geometry

In information geometry [15], the parameters of the pdf of the observations confine the data to an underlying manifold. Thus, the pdf is usually required in a parametric form. In this section, we propose a data-driven approach to recover this underlying manifold without a prior knowledge of the pdf and its parameters.

4.1. Mahalanobis distance

We view the local histogram $\mathbf{h}_t$ as a feature vector for each measurement $\mathbf{z}_t$. As described in Section 3, in this feature domain, measurement noise is translated to a linear transformation. In this section, we define a Riemannian metric that is invariant to linear transformations, and hence, robust to the distortions imposed by measurement noise.

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

By combining the dynamics of the underlying process (1) and Corollary 3, the expected values of the histograms $\mathbb{E}[\mathbf{h}_t]$ can be seen as a high-dimensional random process that satisfies the dynamics given by Itô's lemma

$$d\mathbb{E}\big[h_t^j\big] = \sum_{i=1}^{d}\left(\frac{1}{2}\frac{\partial^2\mathbb{E}[h^j]}{\partial\theta^i\partial\theta^i} + a^i\frac{\partial\mathbb{E}[h^j]}{\partial\theta^i}\right)dt + \sum_{i=1}^{d}\frac{\partial\mathbb{E}[h^j]}{\partial\theta^i}dw_t^i, \quad j=1,\ldots,m. \tag{7}$$

For simplicity of notation, we omit the time index t from the partial derivatives. In addition, we emphasize that the expected value of the histogram is with respect to the conditional measurement density $p(\mathbf{z}|\boldsymbol{\theta})$, and hence, can be viewed as a random process that depends upon the random dynamics of $\boldsymbol{\theta}$.

According to (7), the (j,k)th element of the $m\times m$ covariance matrix $\mathbf{C}_t$ (with respect to the dynamics) of $\mathbb{E}[\mathbf{h}_t]$ is given by

$$C_t^{jk} = \mathrm{Cov}\big(\mathbb{E}\big[h_t^j\big],\mathbb{E}\big[h_t^k\big]\big) = \sum_{i=1}^{d}\frac{\partial\mathbb{E}[h^j]}{\partial\theta^i}\frac{\partial\mathbb{E}[h^k]}{\partial\theta^i}, \quad j,k=1,\ldots,m. \tag{8}$$

In matrix form, (8) can be rewritten as

$$\mathbf{C}_t = \mathbf{J}_t\mathbf{J}_t^T \tag{9}$$

where $\mathbf{J}_t$ is the $m\times d$ Jacobian matrix, whose (j,i)th element is defined by

$$J_t^{ji} = \frac{\partial\mathbb{E}[h^j]}{\partial\theta^i}, \quad j=1,\ldots,m,\ i=1,\ldots,d.$$

Thus, the covariance matrix $\mathbf{C}_t$ is a semi-definite positive matrix of rank d. The derivation above is taken from [11] and suggests a local linearization of the observers/features as

$$\mathbb{E}[\mathbf{h}_s] = \mathbb{E}[\mathbf{h}_t] + \mathbf{J}_t(\boldsymbol{\theta}_s - \boldsymbol{\theta}_t) + \boldsymbol{\epsilon}_{s,t}$$

where $\boldsymbol{\epsilon}_{s,t}$ represents the residual higher order terms.

The local covariance matrices represent the natural perturbations of the underlying process as manifested in the feature domain. This information is exploited to define a Riemannian metric; we define a symmetric $\mathbf{C}$-dependent squared distance between pairs of measurements as

$$d^2(\mathbf{z}_t,\mathbf{z}_s) = 2\big(\mathbb{E}[\mathbf{h}_t] - \mathbb{E}[\mathbf{h}_s]\big)^T(\mathbf{C}_t+\mathbf{C}_s)^{-1}\big(\mathbb{E}[\mathbf{h}_t] - \mathbb{E}[\mathbf{h}_s]\big). \tag{10}$$

Since the dimension d of the underlying process is usually smaller than the number of histogram bins m, the covariance matrix is singular and non-invertible. Consequently, we use the pseudo-inverse to compute the inverse matrix in (10). In addition, the expected values of the histograms and their covariance matrices need to be estimated from the data. This issue is described in more detail in Section 4.2.

The distance in (10) is known as the *Mahalanobis distance* with the property that it is invariant under linear transformations. Thus, by Lemma 1 and Corollary 2, it is invariant to stationary measurement noise (e.g., additive noise or multiplicative noise).

Assumption 4. *The expected value of the histogram $\mathbb{E}[\mathbf{h}_t]$ is a bi-Lipschitz function with respect to $\boldsymbol{\theta}_t$.*

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

This is a formulation of "good" observers/features that are both: (a) stable and smooth, i.e., small changes of the underlying process do not translate to large changes of the features, and (b) sensitive, i.e., small scale perturbations are detected in the features. For example, we note that the linear transformation employed by measurement noise on the observable pdf (3) may degrade the available information; an additive Gaussian noise employs a low-pass blurring filter on the clean observation component. In case the dependency on the underlying process is manifested in high-frequencies, the linear transformation derived by the noise significantly attenuates the sensitivity of the features to small perturbations of the underlying process. Therefore, we expect to perform well up to a certain noise level as long as the histograms can be viewed as bi-Lipschitz functions of the underlying process. Above this level, we expect to experience a sudden drop in performance. For more details, see [28].

Under Assumption 4, combining Lemma 3.1 in [29] and Corollary 3 yields that the Mahalanobis distance in (10) approximates the Euclidean distance between samples of the underlying process. Let $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_s$ be two samples of the underlying process. Then, the Euclidean distance between the samples is approximated to a second order by a local linearization of $\mathbb{E}[\mathbf{h}_t]$ with respect to $\boldsymbol{\theta}_t$, and is given by

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_s\|^2 = d^2(\mathbf{z}_t, \mathbf{z}_s) + O\big(\big\|\mathbb{E}[\mathbf{h}_t] - \mathbb{E}[\mathbf{h}_s]\big\|^4\big). \tag{11}$$

This approximation is further discussed in Section 4.2. For more details see [11] and [29].

Assumption 4 implies that there is an intrinsic map $i(\mathbb{E}[\mathbf{h}_t]) = \boldsymbol{\theta}_t$ from the features to the underlying process. Thus, finding the Euclidean distance between the underlying samples in (11) is equivalent to the inverse problem defined by the following nonlinear differential equation

$$\sum_{i=1}^{m} \frac{\partial \theta^j}{\partial \mathbb{E}[h^i]} \frac{\partial \theta^k}{\partial \mathbb{E}[h^i]} = \big[C_t^{-1}\big]^{jk}, \quad j,k = 1,\ldots,d. \tag{12}$$

In this work, this equation is empirically solved in Section 5 through the solution of an eigenvectors problem of an appropriate discrete Laplace operator.

4.2. Local covariance matrix estimation and inverse

Let t_0 be the time index of a "pivotal" histogram $\mathbf{h}_{t_0}$ of a "cloud" of histograms $\{\mathbf{h}_{t_0,s}\}_{s=1}^{L_2}$ of size L_2 taken from a local neighborhood. In this work, since we assume that a sequence of measurements is available, the neighborhoods can be simply short windows in time centered at time index t_0.

As described in the introduction, the histograms and the local clouds implicitly define two time scales of analysis. The fine time scale is defined by short-time windows of L_1 measurements, in which the pdfs are estimated (by computing histograms). The coarse time scale is defined by the local neighborhood of L_2 neighboring histograms (features) in time. We note that the approximation in (11) is valid as long as the statistics of the ambient measurement noise are locally constant in the short-time windows of length L_1, and the variations of the underlying process derived by the Brownian motion can be detected in the differences between the histograms in windows of length L_2.

Stemmed from the driving Brownian motion in the dynamics model in (1) and (7), the histograms in the local cloud can be seen as small perturbations of the pivotal histogram and may be used to estimate the local covariance matrix in (10).[2] The empirical covariance matrix of the cloud is estimated by

$$\widehat{\mathbf{C}}_{t_0} = \frac{1}{L_2} \sum_{s=1}^{L_2} (\mathbf{h}_{t_0,s} - \widehat{\boldsymbol{\mu}}_{t_0})(\mathbf{h}_{t_0,s} - \widehat{\boldsymbol{\mu}}_{t_0})^T \tag{13}$$

[2] We emphasize that we consider the covariance of the features and not the features themselves, which are estimates of the time varying pdfs of the "raw" measurements.

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

where $\widehat{\boldsymbol{\mu}}_{t_0}$ is the empirical mean of the set

$$\widehat{\boldsymbol{\mu}}_{t_0} = \frac{1}{L_2}\sum_{s=1}^{L_2} \mathbf{h}_{t_0,s}. \tag{14}$$

Thus, the Mahalanobis distance (10) can be empirically evaluated by

$$d^2(\mathbf{z}_t, \mathbf{z}_s) \simeq 2(\mathbf{h}_t - \mathbf{h}_s)^T (\widehat{\mathbf{C}}_t + \widehat{\mathbf{C}}_s)^{-1} (\mathbf{h}_t - \mathbf{h}_s). \tag{15}$$

Since the rank of the covariance matrix d is usually smaller than its dimension m, in order to compute the inverse matrix only the d principal components of the matrix are used. This operation "cleans" the matrix and filters out noise. In addition, when the empirical rank of the local covariance matrices of the features is lower than d, it indicates that the available features are insufficient and larger clouds should be used. Let $\mathbf{J}_{t_0} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ be the singular value decomposition (SVD) of the Jacobian $\mathbf{J}_{t_0}$ at t_0, where $\mathbf{U}$ is an $m \times d$ unitary matrix consisting of the left-singular vectors, $\boldsymbol{\Sigma}$ is a diagonal $d \times d$ matrix consisting of the singular values, and $\mathbf{V}$ is a $d \times d$ unitary matrix consisting of the right-singular vectors. From (9), using the SVD, we define the pseudo-inverse of the local covariance as

$$\mathbf{C}_{t_0}^{-1} \triangleq \mathbf{U}\boldsymbol{\Sigma}^{\dagger}\mathbf{U}^T \tag{16}$$

where $\boldsymbol{\Sigma}^{\dagger}$ is a $d \times d$ diagonal matrix consisting of the squared reciprocals of the singular values.

Let $\mathbf{h}_{t_0,t}$ and $\mathbf{h}_{t_0,s}$ be two samples from the same local neighborhood centered at time t_0. By (10), the Mahalanobis distance between the samples is given by

$$d^2(\mathbf{z}_{t_0,t}, \mathbf{z}_{t_0,s}) = \big(\mathbb{E}[\mathbf{h}_{t_0,t}] - \mathbb{E}[\mathbf{h}_{t_0,s}]\big)^T \mathbf{C}_{t_0}^{-1} \big(\mathbb{E}[\mathbf{h}_{t_0,t}] - \mathbb{E}[\mathbf{h}_{t_0,s}]\big).$$

Based on the local linearization $\mathbb{E}[\mathbf{h}_{t_0,t}] \simeq \mathbb{E}[\mathbf{h}_{t_0}] + \mathbf{J}_{t_0}(\boldsymbol{\theta}_{t_0,t} - \boldsymbol{\theta}_{t_0})$ and $\mathbb{E}[\mathbf{h}_{t_0,s}] \simeq \mathbb{E}[\mathbf{h}_{t_0}] + \mathbf{J}_{t_0}(\boldsymbol{\theta}_{t_0,s} - \boldsymbol{\theta}_{t_0})$ we have

$$d^2(\mathbf{z}_{t_0,t}, \mathbf{z}_{t_0,s}) \simeq (\boldsymbol{\theta}_{t_0,t} - \boldsymbol{\theta}_{t_0,s})^T \mathbf{J}_{t_0}^T \mathbf{C}_{t_0}^{-1} \mathbf{J}_{t_0} (\boldsymbol{\theta}_{t_0,t} - \boldsymbol{\theta}_{t_0,s}). \tag{17}$$

By substituting (16) and the SVD of $\mathbf{J}_{t_0}$ into (17) and by using the unitary property of $\mathbf{U}$ and $\mathbf{V}$, we get (11), i.e., the Mahalanobis distance between each pair of histograms in the local neighborhood approximates the Euclidean distance between the underlying samples of the intrinsic process. We remark that this approximation can be extended to any two histograms by considering the linearization around a middle sample [11].

4.3. Geometric interpretation

The input of most existing geometric methods is usually a data set of samples given in an arbitrary order, and hence, in case of time series, the time labels are often ignored. Here, the temporal cue is encoded through the Mahalanobis distance by exploiting the time labels to define local neighborhoods.

Fig. 1 illustrates the geometric interpretation of the Mahalanobis distance. The dynamics model of the underlying process (1) implies that in a short window in time, in which the drift term is assumed to be almost fixed, the samples of the underlying process $\boldsymbol{\theta}_t$ are normally distributed in a standard Gaussian cloud due to the driving Brownian motion. This cloud of samples is transformed into an elliptic cloud in the features domain according to the particular measurement modality and choice of features. In particular, the principal axes of the elliptic cloud represent the largest variation/distortion directions imposed by the

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

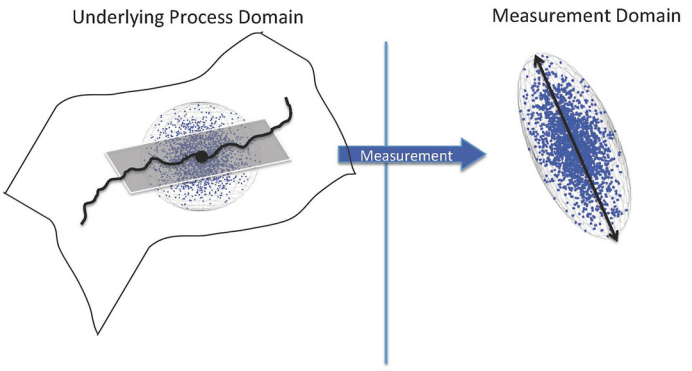


Fig. 1. Geometric illustration. The curve represents a trajectory of the underlying process on the intrinsic manifold, and the thick dot represents a pivotal sample. The time labels of the signal enable to identify the local neighborhood around the pivotal sample. Assuming that the drift is fixed, the samples of the underlying process in the local neighborhood are normally distributed around this sample, thereby creating the standard Gaussian cloud of blue points. This cloud of points is transformed into an elliptic cloud according to the measurement modality. In particular, the major axis of the ellipse (marked by a black double sided arrow) represents the local distortion imposed by the measurement. The principal component of the covariance matrix of the samples in the cloud estimates this distortion, and in turn, is used in the Mahalanobis distance.

function $\mathbb{E}[\mathbf{h}_t]$ with respect to $\boldsymbol{\theta}_t$. These axes can be estimated by the principal components of the covariance matrix of the samples in the cloud. Thus, the key point is that the time labels of the samples define such local neighborhoods/clouds, and hence, allow for the estimation of the local sources of variability/distortion through the eigenvectors of the covariance matrix of the samples in each cloud.

The geometric interpretation implies that the Mahalanobis distance (10), which consists of the inverse covariance matrices, inverts the "locally linear" distortion $\mathbb{E}[\mathbf{h}_s] \simeq \mathbb{E}[\mathbf{h}_t] + \mathbf{J}_t(\boldsymbol{\theta}_s - \boldsymbol{\theta}_t)$ and compares a pair of samples in the coordinate system of the feature domain according to the canonical coordinate system of the underlying process domain.

4.4. Relationship to information geometry

In this section we discuss the relationship between the presented analysis and information geometry. In particular, we show isometry between the Mahalanobis distance between empirical pdfs (10) and the Kullback–Liebler divergence, which is approximated by the Fisher metric on the manifold of the parameters of the pdf of measurements [15].

To show this isometry, we consider slightly different features of the signal. Let $\mathbf{l}_{t_0,t}$ be a new feature vector defined by

$$l^j_{t_0,t} = \sqrt{h^j_{t_0}} \log\big(h^j_{t_0,t}\big) \tag{18}$$

where t_0 is the time index of the pivotal sample of the cloud containing the sample at t. We note that this choice of features is no longer a linear transformation, and therefore, the Mahalanobis metric between the new features is no longer noise resilient. Thus, in practice we use $\mathbf{h}_t$ as features. In the following analysis we assume infinitesimal clouds with sufficient number vectors.

Theorem 5. *The matrix* $\mathbf{I}_{t_0} \triangleq \mathbf{J}^T_{t_0}\mathbf{J}_{t_0}$ *is an approximation of the Fisher Information matrix, i.e.,*

$$I^{ii'}_{t_0} \simeq \mathbb{E}\left[\frac{\partial}{\partial\theta^i_{t_0}} \log\big(p(\mathbf{z}|\boldsymbol{\theta}_{t_0})\big) \frac{\partial}{\partial\theta^{i'}_{t_0}} \log\big(p(\mathbf{z}|\boldsymbol{\theta}_{t_0})\big)\right]. \tag{19}$$

Proof. See Appendix A. □

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

By (9) and Theorem 5, we get that the SVD of the Jacobian $\mathbf{J}_{t_0}$ describes the relationship between the local covariance matrix and the Fisher Information matrix, when the features are defined to be the logarithm of the local density (18). Let $\{\rho_j, \boldsymbol{\upsilon}_j, \boldsymbol{\nu}_j\}_j$ be the singular values, left-singular vectors, and right-singular vectors of the Jacobian matrix $\mathbf{J}_{t_0}$. Then, by (9) and Theorem 5, $\mathbf{C}_{t_0}$ and $\mathbf{I}_{t_0}$ share the same eigenvalues. In addition, $\boldsymbol{\upsilon}_j$ are the eigenvectors of the local covariance matrix $\mathbf{C}_{t_0}$. According to (10), it is used to define an intrinsic metric between feature vectors (the histograms of the measurements) that reveals the underlying process (11). On the other hand, by Theorem 5, $\boldsymbol{\nu}_j$ are the eigenvectors of the Fisher Information matrix of the measurements. The Fisher Information matrix defines the Fisher metric, that locally approximates the Kullback–Liebler divergence between densities of measurements in the cloud, denoted by $\mathcal{D}$, i.e.,

$$\mathcal{D}\big(p(\mathbf{z}_{t_0,t}|\boldsymbol{\theta}_{t_0,t})\|p(\mathbf{z}_{t_0}|\boldsymbol{\theta}_{t_0})\big) \simeq (\boldsymbol{\theta}_{t_0,t} - \boldsymbol{\theta}_{t_0})^T \mathbf{I}_{t_0} (\boldsymbol{\theta}_{t_0,t} - \boldsymbol{\theta}_{t_0}). \tag{20}$$

Using Theorem 5, the divergence (20) can be computed based on the cloud of samples according to

$$J^{ji}_{t_0}\big(\theta^i_{t_0,t} - \theta^i_{t_0}\big) = \lim_{\theta^i_{t_0,t} \to \theta^i_{t_0}} \mathbb{E}\big[l^j_{t_0,t}\big] - \mathbb{E}\big[l^j_{t_0}\big]$$

which can be empirically estimated by samples in the cloud as follows

$$J^{ji}_{t_0}\big(\theta^i_{t_0,t} - \theta^i_{t_0}\big) \simeq \frac{1}{L}\sum_{t=1}^{L}\big(l^j_{t_0,t} - l^j_{t_0}\big).$$

Thus, we obtain an isometry between the "external" intrinsic metric of the measurements (10) and the "internal" metric of the pdfs (20).

To further demonstrate the difference between the "internal" Fisher metric and the "external" empirical metric, we present a simple 1-dimensional example. Consider the following measurement modality

$$z_t = y_t + v_t,$$

where $y_t \sim \mathcal{N}(0, \theta_t)$ is the "interesting" component controlled by θ_t, and $v_t \sim \mathcal{N}(0, \sigma_v^2)$ is an additive stationary noise. Consequently, the pdf of the measurements is a parametric family of normal distributions controlled by both the underlying process θ_t and the variance of the measurement noise, i.e., $z_t \sim \mathcal{N}(0, \theta_t + \sigma_v^2)$. In this case, the corresponding Fisher Information matrix is $\mathbf{I}_t = 1/2(\theta_t + \sigma_v^2)^2$, and the Fisher metric between $p(z_{t_0}|\theta_{t_0})$ and $p(z_t|\theta_t)$ is

$$(\theta_{t_0,t} - \theta_{t_0})^T \mathbf{I}_{t_0} (\theta_{t_0,t} - \theta_{t_0}) = \frac{|\theta_t - \theta_{t_0}|^2}{2(\theta_{t_0} + \sigma_v^2)^2}. \tag{21}$$

On one hand, the Fisher metric is invariant to change of variables applied to the measurements z_t. However, it depends on the measurement modality (e.g., the fact that the observation y_t is a Gaussian process whose variance is the underlying process) and on the measurement noise (e.g., the metric is a function of variance of the noise σ_v^2). On the other hand, the Mahalanobis distance (10) treats the pdf of z_t merely as a function of the underlying process θ_t. In particular, by Lemma 1, the pdf of z_t can be written as a linear transformation of the pdf of y_t (in this case, as a linear convolution). Since the "external" Mahalanobis distance is invariant to linear transformations, it is independent of the measurement noise. Moreover, unlike the Fisher metric in (21), it approximates the Euclidean distance $|\theta_t - \theta_{t_0}|^2$, as long as $\mathbb{E}[\mathbf{h}_t]$ is a bi-Lipschitz function of θ_t. We believe that these results support the choice of local empirical pdfs as appropriate features that convey the information on the measurements.

We remark that similarly to the role of the inverse covariance matrix in the Mahalanobis distance, the inverse of the Fisher Information matrix can be used to restrict the stochastic measurement process to the manifold of the pdf parameters [30,31].

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

5. Graph-based intrinsic embedding

In Section 4, the Euclidean distance between the samples of the underlying process is approximated via the Mahalanobis distance between histograms (10). In this section, we build diffusion geometry [5], which provides a parameterization of the underlying process itself from the pairwise Euclidean distances through the solution of an eigenvector problem of a Laplace operator, without assuming any particular statistical model of the measurements.

5.1. Laplace operator

Let $\{\mathbf{z}_t\}_{t=1}^N$ be a sequence of N measurements. We construct an $N \times N$ affinity matrix (kernel) $\mathbf{W}$ using a Gaussian based on the Mahalanobis distance between the measurements (10)

$$W^{t\tau} = \exp\left\{-\frac{d^2(\mathbf{z}_t, \mathbf{z}_\tau)}{\varepsilon}\right\} \tag{22}$$

where $\varepsilon > 0$ is a tunable kernel scale. Thus, $\mathbf{W}$ measures the affinity between the measurements $\mathbf{z}_t$ according to the distance between the corresponding samples of the underlying process $\boldsymbol{\theta}_t$. It is invariant to the observation modality and it is resilient to measurement noise. The kernel defines a weighted graph, in which the samples $\mathbf{z}_t$ are the nodes and the kernel sets the weights of the edges: node $\mathbf{z}_t$ and $\mathbf{z}_\tau$ are connected by an edge with weight $W^{t\tau}$. Thus, each sample is in effect connected to other samples that are within a neighborhood of size ε with respect to the corresponding Mahalanobis distance. In this work, we set the scale ε to be the median of the values of the kernel; it implies that every node is connected to approximately half of the nodes. Let $\mathbf{D}$ be a diagonal $N \times N$ matrix, whose diagonal elements are given by $D^{tt} = \sum_\tau W^{t\tau}$. Let $\mathbf{W}_{\text{norm}} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ be a normalized kernel that has the same eigenvectors as the normalized graph-Laplacian $\mathbf{I} - \mathbf{W}_{\text{norm}}$ [32]. The matrix $\mathbf{D}$ is often called a density matrix as D^{tt} approximates the local density in the vicinity of $\mathbf{z}_t$, and hence, the normalization can handle nonuniform sampling of the measurements [5]. The eigenvalue decomposition (EVD) is applied to $\mathbf{W}_{\text{norm}}$ and its eigenvectors are denoted by $\boldsymbol{\phi}_i$. By [11] and [29], the normalized graph-Laplacian converges to a Laplace (diffusion) operator on the intrinsic manifold, and its eigenvectors provide an approximate parameterization of the underlying intrinsic process.[3] Specifically, when the underlying process consists of independent coordinates, the leading d eigenvectors (except the trivial), which are a local canonical/intrinsic coordinate system for the manifold [33], recover d proxies for the underlying process coordinates up to a monotonic scaling [29]. In other words, they are empirical solutions to the inverse problem described by the differential equation in (12). In addition, the eigenvectors are independent in case the manifold is flat [11]. Thus, under these conditions, without loss of generality, the tth coordinate of the ith eigenvector parameterizes the ith coordinate of the sample of the underlying process at time t, i.e.,

$$\phi_i^t = \phi_i\big(\theta_t^i\big), \quad i = 1, \ldots, d, \ t = 1, \ldots, N,$$

where $\phi_i(\cdot)$ is a monotonic function and $\boldsymbol{\theta}_t$ is the sample of the underlying process associated with the measurement $\mathbf{z}_t$. Based on the eigenvectors, a d-dimensional parameterization of the measurement samples are defined by the following embedding

$$\boldsymbol{\Phi}(\mathbf{z}_t) \triangleq \big[\phi_1^t, \phi_2^t, \ldots, \phi_d^t\big], \quad t = 1, \ldots, N. \tag{23}$$

[3] While the described graph normalization isolates the independent variables as coordinates by incorporating the drift of the underlying process, a different normalization would yield, for example, the Laplace–Beltrami operator of the intrinsic manifold, which is independent of the drift and separates the geometry from the dynamics [5].

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

By the monotonicity of the eigenvectors, this embedding organizes the measurements according to the values of the underlying process. Consequently, we view this embedding as the *empirical intrinsic modeling* of the measurements representing the underlying process.

We remark that the embedding does not take explicitly into account the chronological order of the measurements. However, the Mahalanobis distance encodes the time dependency by using local covariance matrices, and the Laplace operator reveals the dynamics by integrating those distances over the entire set of samples.

5.2. Sequential processing

The construction of the intrinsic embedding in Section 5.1 is computationally heavy due to the application of the EVD. In many practical settings, the measurements are not available in advance, but rather become available sequentially. As a result, the computationally demanding procedure should be applied repeatedly. In this section, we describe a sequential procedure for extending the embedding, which circumvents the EVD applications.

Let $\{\bar{\mathbf{z}}_s\}_{s=1}^N$ be a sequence of N reference measurements that are assumed to be available in advance. The availability of the reference measurements allows for the computation of the local histograms and their corresponding covariance matrices as described in Section 3 and Section 4, respectively. Then, the $N \times N$ kernel of the reference measurements $\bar{\mathbf{z}}_s$, denoted now by $\mathbf{W}_{\text{ref}}$, can be constructed according to (22), and the embedding of the reference measurements can be computed in an "offline" batch manner.

Let $\{\mathbf{z}_t\}_{t=1}^M$ be a new sequence of M measurements, which are assumed to become available sequentially and should be processed in an "online" (or even realtime) manner. As proposed in [29,34,35], a nonsymmetric pairwise metric between any new measurement $\mathbf{z}_t$ and a reference measurement $\bar{\mathbf{z}}_s$ is defined, similarly to the Mahalanobis distance (10), by

$$a(\mathbf{z}_t, \bar{\mathbf{z}}_s) = (\mathbf{h}_t - \bar{\mathbf{h}}_s)^T \mathbf{C}_s^{-1} (\mathbf{h}_t - \bar{\mathbf{h}}_s), \tag{24}$$

and a corresponding nonsymmetric $M \times N$ affinity matrix $\mathbf{A}$ between the two sets of measurements is constructed, whose (t,s)th element is given by

$$A^{ts} = \exp\left\{-\frac{a(\mathbf{z}_t, \bar{\mathbf{z}}_s)}{\varepsilon}\right\}. \tag{25}$$

The construction of (25) requires the corresponding features of the measurements (histograms) and the local covariance matrix of merely the reference measurement $\bar{\mathbf{z}}_s$ and does not use the covariance matrix of the new measurement $\mathbf{z}_t$. The nonsymmetric kernel defines a bipartite graph [36], where $\{\bar{\mathbf{z}}_s\}$ and $\{\mathbf{z}_t\}$ are two disjoint sets of nodes, and each pair of nodes $\bar{\mathbf{z}}_s$ and $\mathbf{z}_t$ is connected by an edge with weight A^{ts}.

Define the normalized nonsymmetric kernel $\widetilde{\mathbf{A}} = \mathbf{D}_A^{-1}\mathbf{A}\mathbf{Q}^{-1}$, where $\mathbf{D}_A$ is a diagonal matrix whose diagonal elements are the sums of rows of $\mathbf{A}$, and $\mathbf{Q}$ is a diagonal matrix whose diagonal elements are the sums of rows of $\mathbf{D}_A^{-1}\mathbf{A}$. It is shown in [29] that

$$\mathbf{W}_{\text{ref}} = \widetilde{\mathbf{A}}^T \widetilde{\mathbf{A}}. \tag{26}$$

By definition, the (s,s')th element of the symmetric kernel is given by

$$W_{\text{ref}}^{ss'} = \sum_t \widetilde{A}^{ts} \widetilde{A}^{ts'},$$

which implies that the affinity metric (kernel) between any two reference measurements $\bar{\mathbf{z}}_s$ and $\bar{\mathbf{z}}_{s'}$ integrates the nonsymmetric affinities between these two reference measurements and all possible measurements $\mathbf{z}_t$.

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

The dual extended $M \times M$ kernel between the new samples is defined as $\mathbf{W}_{\text{ext}} = \widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^T$. Similarly to the interpretation of $\mathbf{W}_{\text{ref}}$, the (t,τ)th element of $\mathbf{W}_{\text{ext}}$ can be interpreted as an affinity metric between any pair of new measurements $\mathbf{z}_t$ and $\mathbf{z}_\tau$ via all the reference measurements $\bar{\mathbf{z}}_s$. It implies that two measurements are similar if they "see" the reference measurements in the same way. Furthermore, it is shown in [29] and [12] that the elements of the extended kernel are proportional to a Gaussian defined similarly to (22) with the corresponding Mahalanobis distances between pairs of new measurements.

The relationship between the definitions of the kernels $\mathbf{W}_{\text{ref}}$ and $\mathbf{W}_{\text{ext}}$ through $\widetilde{\mathbf{A}}$ yields: (a) the kernels share the same eigenvalues λ_i; (b) the eigenvectors $\boldsymbol{\varphi}_i$ of $\mathbf{W}_{\text{ref}}$ are the right-singular vectors of $\widetilde{\mathbf{A}}$; (c) the eigenvectors $\boldsymbol{\psi}_i$ of $\mathbf{W}_{\text{ext}}$ are the left-singular vectors of $\widetilde{\mathbf{A}}$. As discussed in Section 5.1, the right-singular vectors $\boldsymbol{\varphi}_i$ represent the underlying process of the reference measurements $\bar{\mathbf{z}}_s$, and by [12] and [29], the left-singular vectors $\boldsymbol{\psi}_i$ naturally extend this representation to the new measurements $\mathbf{z}_t$. Thus, we define a d-dimensional embedding, similarly to (23), by

$$\boldsymbol{\Psi}(\mathbf{z}_t) \triangleq \left[\boldsymbol{\psi}_1^t, \boldsymbol{\psi}_2^t, \ldots, \boldsymbol{\psi}_d^t\right], \tag{27}$$

for any new measurement $\mathbf{z}_t$. This embedding (27) is seen as the intrinsic modeling of the new measurements parameterizing the corresponding underlying process.

The relationship between the spectral representations of the kernels $\mathbf{W}_{\text{ref}}$ and $\mathbf{W}_{\text{ext}}$ is given by the singular value decomposition (SVD) of $\widetilde{\mathbf{A}}$:

$$\boldsymbol{\psi}_i = \frac{1}{\sqrt{\lambda_i}}\widetilde{\mathbf{A}}\boldsymbol{\varphi}_i, \tag{28}$$

and gives rise to a supervised sequential processing algorithm consisting of two stages: a training stage in which a sequence of reference measurements is assumed to be available in advance, and a test stage in which new incoming measurements are sequentially processed [13].

In the training stage, the reference measurements are processed to form a learned model. The histograms and their corresponding local covariance matrices associated with the reference samples are computed. The kernel $\mathbf{W}_{\text{ref}}$, defined on the reference measurements, is directly computed based on the Mahalanobis distance, and its EVD is calculated. The eigenvectors of the kernel form the learned model for the reference set. We store the histograms of the reference measurements along with their local covariance matrices and the EVD of the kernel.

In the test stage, as new incoming measurements $\mathbf{z}_t$ become available, we efficiently extend the model. The nonsymmetric kernels $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ are computed according to (25). Then, the extended representation is obtained by (28), and the embedding of the new samples are defined via (27).

We remark that the proposed extension scheme involves only the information associated with the reference measurements and, in particular, does not involve the local covariance matrices of the new measurements. Thus, it is particularly adequate to real-time processing, since it circumvents the lag required to collect a local neighborhood for each new measurement. In addition, the processing of new measurements is computationally efficient, since it does not require repeated EVD applications. For detailed analysis of the computational burden we refer the readers to [13].

5.3. Probabilistic interpretation

In this section, we revisit the construction of the intrinsic embedding presented in Section 5.1 and Section 5.2 from a probabilistic standpoint. Consider a probabilistic model consisting of a mixture of local statistical models defined by the set of reference measurements. Assume that the sample domain $\mathcal{Z}$ of possible measurements is given by a union of N disjoint subsets, i.e., $\mathcal{Z} = \bigcup_{s=1}^{N} \mathcal{Z}_s$, where each subset is represented by a corresponding reference sample $\bar{\mathbf{z}}_s$. Since usually the number of reference samples is much

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

larger than the number of histogram bins, i.e., $N \gg m$, this is a different, finer, partition than the partition described in Section 3. We further assume that the probability that any measurement $\mathbf{z}_t$ is associated with a particular subset is uniform, i.e., $\Pr(\mathbf{z}_t \in \mathcal{Z}_s) = 1/N$.

Let $\alpha(t,s)$ be the following conditional probability

$$\alpha(t,s) = \Pr(\mathbf{z}_t | \mathbf{z}_t \in \mathcal{Z}_s) \tag{29}$$

which describes the probability of a measurement $\mathbf{z}_t$ given it is associated with $\mathcal{Z}_s$. Define $\tilde{\alpha}(t,s)$ as

$$\tilde{\alpha}(t,s) = \alpha(t,s)/\omega(t),$$

where $\omega(t) = \sum_{s=1}^{N} \alpha(t,s)$.

Let $\mathbf{A}_\alpha$ be an $M \times N$ matrix whose elements are given by $A_\alpha^{ts} = \tilde{\alpha}(t,s)$, and let $\mathbf{W}_\alpha = \mathbf{A}_\alpha \mathbf{A}_\alpha^T$.

Theorem 6. *If the measurements $\mathbf{z}_t$ are statistically independent, then the elements of the $M \times M$ kernel matrix $\mathbf{W}_\alpha$ are the conditional probability that a pair of given measurements are associated with the same reference measurement, i.e.,*

$$W_\alpha^{t\tau} = \Pr(\mathbf{z}_t \in \mathcal{Z}_s, \mathbf{z}_\tau \in \mathcal{Z}_s | \mathbf{z}_t, \mathbf{z}_\tau)$$

for any s.

This result extends the result presented in [13], which was limited to normal distributions.

Proof. See Appendix B. □

By definition (24)–(25), $\mathbf{A}$ is a special case of $\mathbf{A}_\alpha$, when the conditional probability of a measurement $\mathbf{z}_t$ given it is associated with $\mathcal{Z}_s$ (29) is defined as a normal distribution with $\bar{\mathbf{z}}_s$ mean and $\mathbf{C}_s$ covariance matrix. Thus, the construction of the Mahalanobis metric and the associated Gaussian kernel induce an implicit multi-Gaussian mixture model in the measurements domain. Each reference measurement $\bar{\mathbf{z}}_s$ represents a local (infinitesimal) Gaussian model, and the metric defined by (24) and (25) computes the probability that any arbitrary measurement $\mathbf{z}_t$ is associated with the local model represented by the reference measurement $\bar{\mathbf{z}}_s$.

6. Experimental results

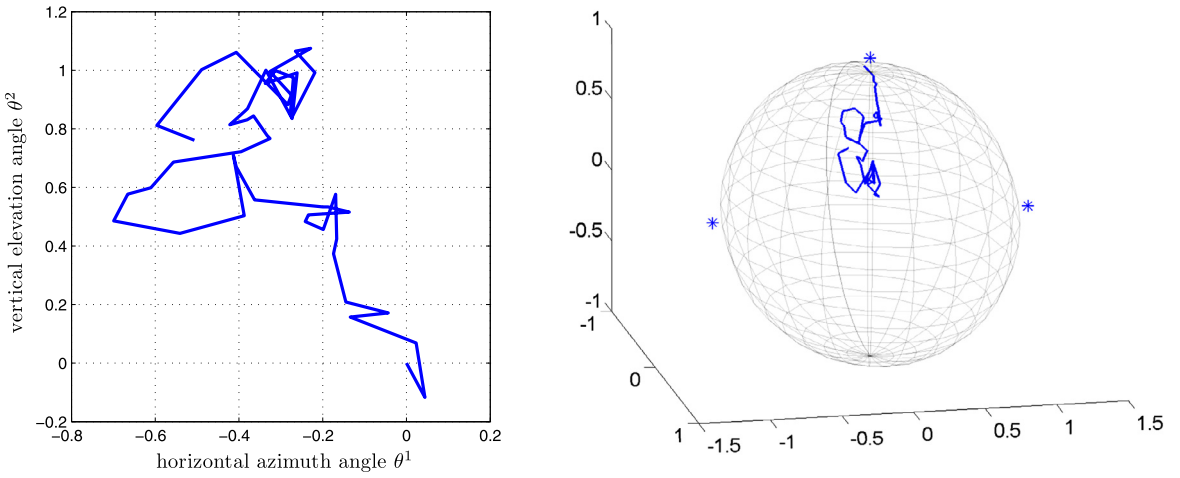
6.1. Simulated dynamics model

We simulate an underlying process whose temporal propagation mimics the motion of a particle in a potential field. Each coordinate of the process is independent and evolves in time according to the following Langevin equation

$$d\theta_t^i = -\nabla U^i\left(\theta_t^i\right) dt + dw_t^i \tag{30}$$

where w_t^i are independent standard Brownian motions, and $U^i(\theta_t^i)$ are the potential fields, fixed in time and varying according to the current position θ_t^i. The potential fields determine the drift of the underlying process and establish the intrinsic manifold. We note that this propagation model is chosen since it describes many natural signals and phenomena.

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006



(a) (b)

Fig. 2. (a) A segment of the 2-dimensional trajectory of the 2 coordinates of the underlying process: the horizontal and vertical angles. (b) The corresponding segment of the 3-dimensional movement of the source on the sphere. The locations of the 3 sensors are marked with $*$.

6.2. Nonlinear tracking

In this experiment, we aim to model the movement of a radioactive source on a 3-dimensional sphere. Since the radius of the sphere is fixed, we assume that the movement of the source is controlled by two independent processes $\boldsymbol{\theta}_t = [\theta_t^1; \theta_t^2]$: the horizontal azimuth angle θ_t^1 and the vertical elevation angle θ_t^2. Suppose the temporal propagation of the spherical angles evolves in time according to the Langevin equation (30). The potential field of each angle is a mixture of two Gaussians with 0 and $\pi/4$ means and 5 and 10 variances, respectively.

Let $\mathbf{x}(\boldsymbol{\theta}_t)$ denote the 3-dimensional coordinates of the source position on the sphere. By assuming that the center of the sphere is located at the origin of the coordinate system, the position of the source is given by

$$\begin{aligned} x^1(\boldsymbol{\theta}_t) &= r\cos\left(\theta_t^1\right)\sin\left(\theta_t^2\right) \\ x^2(\boldsymbol{\theta}_t) &= r\sin\left(\theta_t^1\right)\sin\left(\theta_t^2\right) \\ x^3(\boldsymbol{\theta}_t) &= r\cos\left(\theta_t^2\right), \end{aligned}$$

where r is the radius of the sphere. Fig. 2 illustrates a segment of a trajectory of the 2-dimensional underlying process and the corresponding segment of the 3-dimensional trajectory of the radiating source on the sphere.

To demonstrate the robustness of the proposed method (EIG) to different measurements, we consider three measurement modalities. In Modality 1, the radiation of the source is measured by 3 "Geiger counter" sensors positioned at $\mathbf{x}_j, j = 1, 2, 3$ outside the sphere (see Fig. 2). The sensors detect the radiation and fire "spikes" through spatial Poisson point processes y_t^j in varying rates, which depend on the proximity of the source to each of the sensors. The firing rate of each sensor is given by $\lambda^j(\boldsymbol{\theta}_t) = \exp\{-\|\mathbf{x}_j - \mathbf{x}(\boldsymbol{\theta}_t)\|\}$, where the firing rate is higher when the source is closer to the sensor and the amount of radiation reaching the sensor is higher. The output of each sensor is corrupted by additive noise and is given by

$$z_t^j = g^j(y_t, v_t) = y_t^j + v_t^j, \quad j = 1, 2, 3,$$

where v_t^j is a spike sequence drawn from a Poisson distribution with a fixed rate λ_v^j. Modality 2 is similar to Modality 1. Each sensor fires spikes randomly according to the proximity of the source. The difference

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

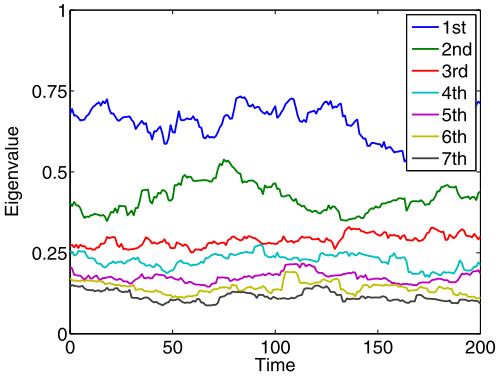


Fig. 3. The spectrum of the local covariance matrices of a sequence of 200 feature vectors as a function of time.

is that in this modality we simulate sensors with unreliable clocks. We measure the time interval between two consecutive spikes given by $z_t^j = y_t^j + v_t^j$. Suppose y_t^j is drawn from an exponential distribution with a rate parameter $\lambda^j(\boldsymbol{\theta}_t)$, and suppose v_t^j is drawn from a fixed normal distribution representing the clock inaccuracy. We note that in Modality 1 the noisy sequence of spikes has a Poisson distribution with rate $\lambda^j(\boldsymbol{\theta}_t) + v_t^j$, whereas the distribution of the measured signal in Modality 2 is not of particular type. In Modality 3, we consider a measurement of a different nature. Consider three sensors that measure the location of the source directly, i.e.

$$z_t^j = x_t^j + v_t^j, \quad j = 1, 2, 3,$$

where v_t^j is an additive white Gaussian noise. This case exhibits nonlinearity in the measurement stemmed from the nonlinear mapping of the two spherical coordinates to the measured Cartesian coordinates.

Under all the measurement modalities, the goal is to reveal the 2-dimensional trajectory $\boldsymbol{\theta}_t$ of the horizontal and vertical angles based on a sequence of noisy measurements $\mathbf{z}_t$. The dynamics model and the measurement model are unknown and the sequence of measurements is all the available information.

We simulate 2-dimensional trajectories of the two independent underlying processes according to (30) and the corresponding noisy measurements under the three measurement modalities. The first $N = 2000$ samples of measurements are used as the reference set, which empirically was shown to be a sufficient amount of data to represent the model of the underlying angles. For each reference measurement we compute a histogram in a short window of length $L_1 = 10$ (with full overlap) and obtain the features. Then, we estimate the local covariance matrix according to (13) setting $L_2 = 10$. Next, the empirical Mahalanobis distance (15), the associated Gaussian kernel (22), and the embedding of the reference measurements (23) are constructed as described in Section 5.

Fig. 3 presents the spectrum of the local covariance matrices of a sequence of 200 features obtained under Modality 1. Each curve shows one eigenvalue out of the seven largest eigenvalues as a function of time. We observe that the two largest eigenvalues are dominant compared to the others which implies that the empirical covariance matrices have rank $d = 2$. This reveals two degrees of freedom hidden in the data, which are indeed the two spherical angles.

Measurements following the reference sequence at times $t > 2000$ are sequentially processed, i.e., for each measurement, the kernel matrix (25) and the extended embedding (27) are computed, as described in Section 5.2.

Fig. 4 shows a scatter plot of the obtained coordinates of the principal eigenvector ψ_1^t as a function of the horizontal angle θ_t^1 (the ground truth) under Modality 1. In Fig. 4(a) the embedding is obtained using the Mahalanobis graph based on the measurements [11] and in Fig. 4(b) the embedding is obtained using the proposed EIG method. We observe that the principal eigenvector obtained via EIG is linearly

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

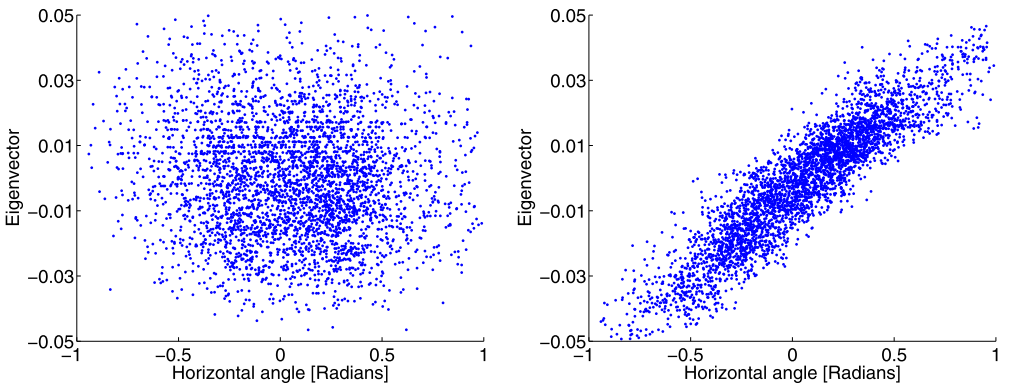


Fig. 4. Scatter plots of the coordinates of the principal eigenvector as a function of the horizontal angle (ground truth). (a) The principal eigenvector obtained by the Mahalanobis graph [11]. (b) The principal eigenvector obtained by EIG.

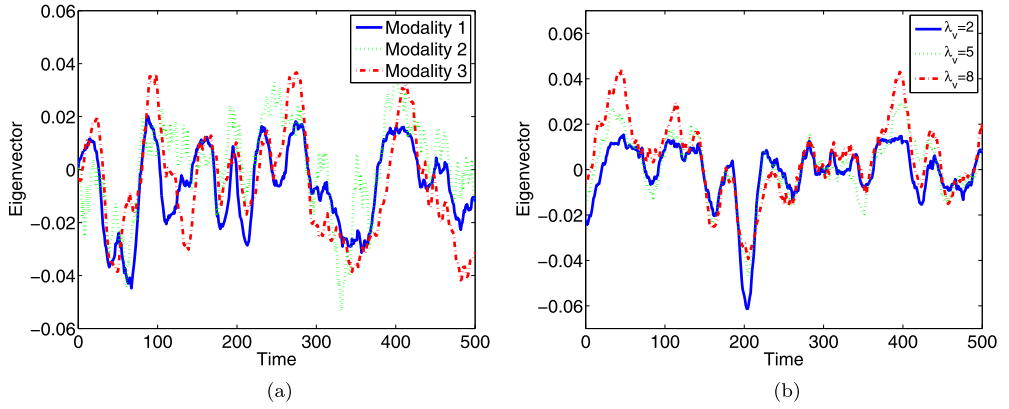


Fig. 5. A comparison of the obtained parameterization of the vertical angle. (a) A comparison between the obtained eigenvectors (corresponding to the vertical angle) under the three different measurement modalities (measuring the same movement). (b) A comparison between the obtained eigenvectors (corresponding to the vertical angle) under measurement Modality 1 with different noise levels ($\lambda_v = 2, 5, 8$).

correlated with the horizontal angle, whereas the leading eigenvector obtained using the Mahalanobis graph is uncorrelated with the angle. We note that no correlation is found between any other pair of an angle and an eigenvector. Thus, it shows that the obtained embedding is an accurate parameterization of the angle. Furthermore, the comparison to the embedding obtained by the Mahalanobis graph based on the measurements suggests that the use of histograms as features is essential as it removes noise interference.

In Fig. 5 we compare the modeling of the vertical angle obtained under different measurement modalities and noise. We note that the presented 500 coordinates of the eigenvectors are computed by extension at times $t > 2000$. Fig. 5(a) depicts three eigenvectors that correspond to the same movement of the source. Each eigenvector is obtained from measurement samples under a different modality through a separate application of the proposed EIG method. We observe that the three eigenvectors follow the same trend, which implies *intrinsic modeling* of the movement and demonstrates the invariance of EIG to the measurement modality. We emphasize that the measurements under the three modalities are very different in their nature: spike sequences in Modalities 1 and 2 and noisy 3-dimensional Cartesian coordinates in Modality 3. In order to further demonstrate the resilience of the modeling to measurement noise, we present in Fig. 5(b) three eigenvectors obtained from measurement samples under Modality 1 (through separate applications of the method) with 3 realizations of noise sequences of spikes $\mathbf{v}_t$ in three different rates λ_v. We observe that the three eigenvectors follow the same trend, and hence, conclude that they *intrinsically* model the movement of

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

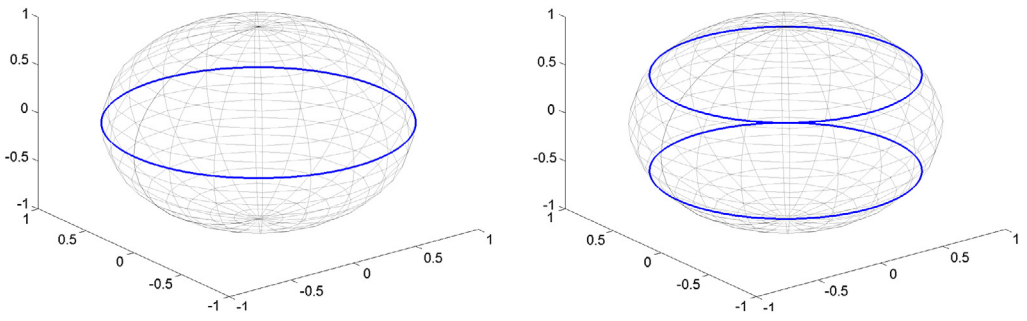


Fig. 6. An illustration of the movement of the two sources confined to rings on a sphere. (a) The solid curve is the ring confining the simulated movements of both sources. (b) The solid curves are the two rings, each confining the simulated movement of one of the two sources.

the source. We note that the relation between the pdf estimates (histograms) and the underlying process in higher noise levels becomes very weak, and thus, as discussed in Section 4, we experience in our experimental study a sudden drop in the correlation between the obtained eigenvectors and the underlying angles when λ_v is set to be greater than a certain value.

In a second experiment, we alter the experimental setup as follows. We now consider *two* radiating sources moving on the sphere. The movement of each source $i = 1, 2$ is confined to a horizontal ring and controlled solely by its azimuth angle θ_t^i, which is the single underlying degree of freedom of the movement. We simulate two movement scenarios. In the first, the two sources move on the same ring (Fig. 6(a)), and in the second, the two sources move on two different rings (Fig. 6(b)). The total amount of radiation from the two sources is measured in the 3 sensors similarly to the previous experiment and the locations of the sensors remain the same.

The obtained experimental results are similar to the results obtained in the previous experiment, i.e., the two underlying angles θ_t^1 and θ_t^2 are accurately parameterized by the two principal eigenvectors. In this experiment, each source has a different structure, which is separated and recovered by the proposed method. In terms of the problem formulation, the difference between the two experiments stems from different measurement models. Hence, this experiment further demonstrates the robustness of the proposed nonparametric blind method in recovering the intrinsic sources of variability. Furthermore, it illustrates the potential of the proposed approach to yield good performance in blind source separation applications.

6.3. Non-stationary hidden Markov chain

In this experiment, the observation process y_t is a 2-states Markov chain with time-varying transition probabilities, which are determined by a 1-dimensional underlying process θ_t. We simulate trajectories of θ_t according to the Langevin equation (30) with a potential field corresponding to a single Gaussian with 0.5 mean and 5 variance. To view θ_t as probability, we clip values outside $[0, 1]$ using hard-thresholding. The clean observation process is measured with additive zero-mean white Gaussian noise v_t, i.e.,

$$z_t = y_t + v_t.$$

The objective is to reveal the underlying process θ_t (determining the time-varying transition probabilities) given a sequence of measurements z_t. The entire interval of $N = 4000$ measurement samples is processed and their embedding is computed directly without extension.

We examine two Markov chain configurations: a Bernoulli scheme and a scheme of order 1 in which the transition depends solely on the current state. In the latter case, the current measurement depends on the underlying process in the current time step and on the measurement in the previous time step.

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

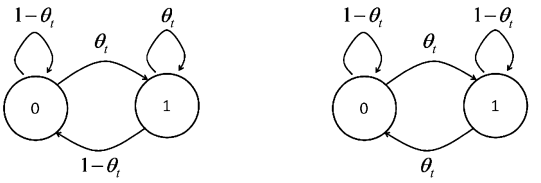


Fig. 7. Two Markov chains configurations. (a) A Bernoulli scheme. (b) A Markov scheme of order 1.

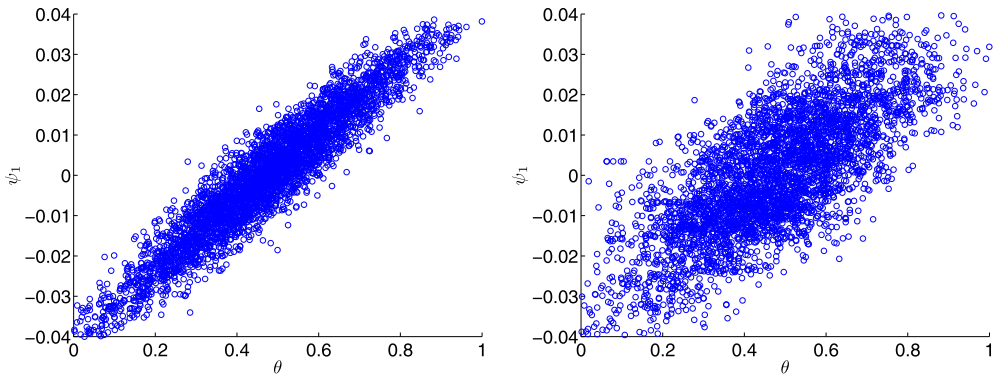


Fig. 8. Scatter plots of the principal eigenvectors as functions of the underlying process under the Bernoulli scheme. (a) The coordinates of the principal eigenvector obtained by short-time averaging and the Mahalanobis graph. (b) The coordinates of the principal eigenvector obtained by EIG.

This context-dependency makes it different from the former scheme and from the experiments described in Section 6.2.

The Bernoulli scheme is illustrated in Fig. 7(a), and the observation process is given by

$$y_t = \begin{cases} 0, & w.p.\ 1-\theta_t \\ 1, & w.p.\ \theta_t. \end{cases}$$

We note that in this specific scenario, revealing the underlying process θ_t from z_t can be simply done by short-time averaging, since

$$\mathbb{E}[z_t|\theta_t] = \theta_t.$$

Fig. 8 presents scatter plots of the obtained embedding as a function of the underlying process θ_t (the ground truth). In Fig. 8(a) we show the embedding obtained by the Mahalanobis graph based on the means of the measurements in short-time windows. In Fig. 8(b) we show the embedding based on the histograms of the measurements in short-time windows obtained by the proposed method. As expected, exploiting the prior knowledge that the underlying process can be revealed by averaging yields good performance; the embedding based on the means in short windows is highly correlated with the underlying process. Moreover, the correlation is stronger than the correlation obtained using EIG, which does not use any a-priori knowledge.

The second scheme is illustrated in Fig. 7(b). In this case, the underlying process is more difficult to recover without any prior knowledge on the measurement model. We process the difference signal (discrete derivative) $\tilde{z}_t = z_t - z_{t-1}$ to convey the first-order dependency. Alternatively, we could process pairs of consecutive measurements. We note that a higher order dependency would require the processing of higher order derivatives or several consecutive measurements together.

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

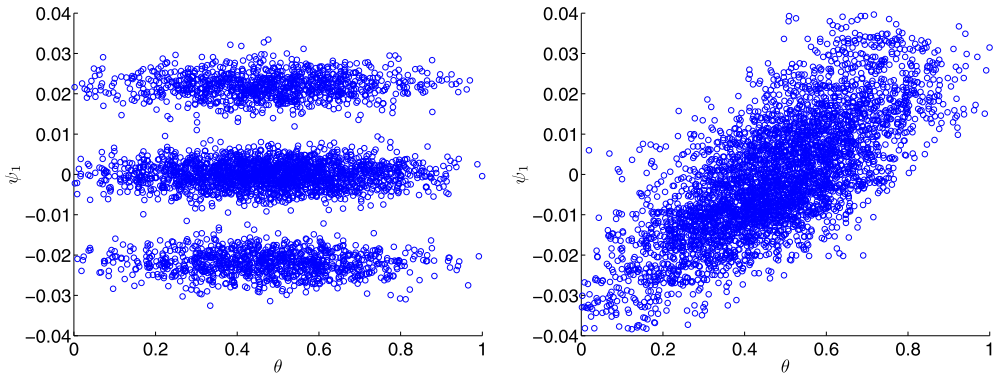


Fig. 9. Scatter plots of the principal eigenvectors as functions of the underlying process under the Markov chain scheme of order 1. (a) The coordinates of the principal eigenvector obtained by the Mahalanobis graph. (b) The coordinates of the principal eigenvector obtained by EIG.

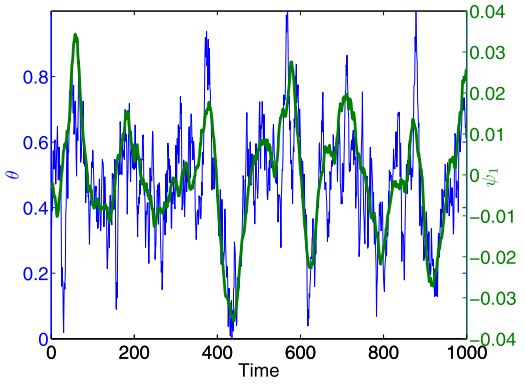


Fig. 10. The principal eigenvector (green curve) obtained by EIG and the underlying process (blue curve) as functions of time. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 9 presents scatter plots of the obtained embedding and the underlying process θ_t. In Fig. 9(a) we show the embedding obtained by the Mahalanobis graph based on short-time averages of the difference signal $\tilde{z}_t$ in short-time windows. In Fig. 9(b) we show the embedding obtained by EIG based on the histograms of the difference signal $\tilde{z}_t$ in short-time windows. We observe that the embedding based on the short-time averaging is degenerate and does not parameterize the underlying process. On the other, the embedding obtained using EIG exhibits high correlation with the underlying process. To further illustrate the obtained modeling of the time series, we present in Fig. 10 the coordinates of the principal eigenvector ψ_1^t obtained using EIG (green curve) along with the samples of underlying process θ_t (blue curve) as functions of time. It can be seen that the eigenvector tracks accurately the trend/drift of the underlying process, whereas the small (diffusion) perturbations are disregarded as expected.

7. Conclusions

In this paper, we propose a probabilistic kernel method for data-driven modeling of stochastic dynamical systems using empirical geometry. It enables us to empirically learn the intrinsic manifold of local probability densities of noisy measurements and provides a parameterization of the underlying processes governing the system. We show that the obtained data-driven parameterization is intrinsic, i.e., invariant under different measurement modalities as well as noise resilient, thereby suggesting that intrinsic filters that eliminate the need to adapt the configuration and to calibrate the measurement instruments may be built. In addition,

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

our modeling method can be implemented in a sequential manner, and hence, it can be applied to online signal processing tasks. Experimental results of two nonlinear and non-Gaussian filtering applications are presented.

The results presented in this paper will be extended in a future work to propose a data-driven Bayesian filtering framework for nonparametric estimation and prediction of signals without the prior knowledge of their probabilistic models. In addition, the continuity of the signals in time will be used to address the proper choice of parameters and time scales, and in particular, to improve the estimation of the empirical probability densities and local covariance matrices. Future work will also address real signals acquired from dynamical systems in various fields, e.g., biomedical applications and chemical engineering systems.

Acknowledgments

The authors would like to thank Stephane Mallat and Amit Singer for helpful discussions. This work was supported by the NSF Award No. 1309858, ARO-MURI W911NF-09-1-0383, and ONR Award No. N000141210797.

Appendix A. Proof of Theorem 5

Proof. The Jacobian elements using the new features (18) are given by

$$J_{t_0}^{ji} = \frac{\partial \mathbb{E}[l_{t_0}^j]}{\partial \theta_{t_0}^i} = \lim_{\theta_{t_0,t}^i \to \theta_{t_0}^i} \frac{\mathbb{E}[l_{t_0,t}^j] - \mathbb{E}[l_{t_0}^j]}{\theta_{t_0,t}^i - \theta_{t_0}^i}. \tag{A.1}$$

By definition, (A.1) is explicitly expressed as

$$\begin{aligned} J_{t_0}^{ji} &= \sqrt{\mathbb{E}\big[h_{t_0}^j\big]} \lim_{\theta_{t_0,t}^i \to \theta_{t_0}^i} \frac{\log(\mathbb{E}[h_{t_0,t}^j]) - \log(\mathbb{E}[h_{t_0}^j])}{\theta_{t_0,t}^i - \theta_{t_0}^i} \\ &= \sqrt{\mathbb{E}\big[h_{t_0}^j\big]} \frac{\partial}{\partial \theta_{t_0}^i} \log\big(\mathbb{E}\big[h_{t_0}^j\big]\big). \end{aligned}$$

Thus, the elements of the matrix $\mathbf{I}_{t_0} = \mathbf{J}_{t_0}^T \mathbf{J}_{t_0}$ are given by

$$\begin{aligned} I_{t_0}^{ii'} &= \sum_{j=1}^{m} \sqrt{\mathbb{E}\big[h_{t_0}^j\big]} \frac{\partial}{\partial \theta_{t_0}^i} \log\big(\mathbb{E}\big[h_{t_0}^j\big]\big) \sqrt{\mathbb{E}\big[h_{t_0}^j\big]} \frac{\partial}{\partial \theta_{t_0}^{i'}} \log\big(\mathbb{E}\big[h_{t_0}^j\big]\big) \\ &= \sum_{j=1}^{m} \frac{\partial}{\partial \theta_{t_0}^i} \log\big(\mathbb{E}\big[h_{t_0}^j\big]\big) \frac{\partial}{\partial \theta_{t_0}^{i'}} \log\big(\mathbb{E}\big[h_{t_0}^j\big]\big) \mathbb{E}\big[h_{t_0}^j\big]. \end{aligned} \tag{A.2}$$

If we additionally assume that the histograms converge to the true pdf of the measurements under the conditions that led to (6), we have

$$\begin{aligned} I_{t_0}^{ii'} &\simeq \int_{\mathbf{z}} \frac{\partial}{\partial \theta_{t_0}^i} \log\big(p(\mathbf{z}|\boldsymbol{\theta}_{t_0})\big) \frac{\partial}{\partial \theta_{t_0}^{i'}} \log\big(p(\mathbf{z}|\boldsymbol{\theta}_{t_0})\big) p(\mathbf{z}|\boldsymbol{\theta}_{t_0}) d\mathbf{z} \\ &= \mathbb{E}\left[\frac{\partial}{\partial \theta_{t_0}^i} \log\big(p(\mathbf{z}|\boldsymbol{\theta}_{t_0})\big) \frac{\partial}{\partial \theta_{t_0}^{i'}} \log\big(p(\mathbf{z}|\boldsymbol{\theta}_{t_0})\big)\right] \end{aligned} \tag{A.3}$$

Thus, by definition $\mathbf{I}_{t_0}$ is an approximation of the Fisher Information matrix. □

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

Appendix B. Proof of Theorem 6

Proof. Let s be an arbitrary index of a sample from the reference set. By definition and since the measurements are independent, we have

$$W_{\alpha}^{t\tau} = \frac{\sum_{s'=1}^{N} \Pr(\mathbf{z}_t, \mathbf{z}_\tau | \mathbf{z}_t \in \mathcal{Z}_{s'}, \mathbf{z}_\tau \in \mathcal{Z}_{s'})}{\sum_{s''=1}^{N} \Pr(\mathbf{z}_t | \mathbf{z}_t \in \mathcal{Z}_{s''}) \sum_{s''=1}^{N} \Pr(\mathbf{z}_\tau | \mathbf{z}_\tau \in \mathcal{Z}_{s''})}. \tag{B.1}$$

Using the uniform distribution, we can rewrite (B.1) as

$$W_{\alpha}^{t\tau} = \frac{\Pr(\mathbf{z}_t \in \mathcal{Z}_s, \mathbf{z}_\tau \in \mathcal{Z}_s) \sum_{s'=1}^{N} \Pr(\mathbf{z}_t \in \mathcal{Z}_{s'}) \Pr(\mathbf{z}_t, \mathbf{z}_\tau | \mathbf{z}_t \in \mathcal{Z}_{s'}, \mathbf{z}_\tau \in \mathcal{Z}_{s'})}{\sum_{s''=1}^{N} \Pr(\mathbf{z}_t \in \mathcal{Z}_{s''}) \Pr(\mathbf{z}_t | \mathbf{z}_t \in \mathcal{Z}_{s''}) \sum_{s''=1}^{N} \Pr(\mathbf{z}_\tau \in \mathcal{Z}_{s''}) \Pr(\mathbf{z}_\tau | \mathbf{z}_\tau \in \mathcal{Z}_{s''})}. \tag{B.2}$$

By the law of total probability and since the measurements are independent, we obtain

$$W_{\alpha}^{t\tau} = \frac{\Pr(\mathbf{z}_t \in \mathcal{Z}_s, \mathbf{z}_\tau \in \mathcal{Z}_s) \Pr(\mathbf{z}_t, \mathbf{z}_\tau | \mathbf{z}_t \in \mathcal{Z}_s, \mathbf{z}_\tau \in \mathcal{Z}_s)}{\Pr(\mathbf{z}_t, \mathbf{z}_\tau)}.$$

Finally, Bayes' theorem yields

$$W_{\alpha}^{t\tau} = \Pr(\mathbf{z}_t \in \mathcal{Z}_s, \mathbf{z}_\tau \in \mathcal{Z}_s | \mathbf{z}_t, \mathbf{z}_\tau). \qquad \square$$

References

[1] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 260 (2000) 2319–2323.
[2] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 260 (2000) 2323–2326.
[3] D.L. Donoho, C. Grimes, Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data, Proc. Natl. Acad. Sci. USA 100 (2003) 5591–5596.
[4] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (2003) 1373–1396.
[5] R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (2006) 5–30.
[6] I. Kevrekidis, C. Gear, G. Hummer, Equation-free: the computer-aided analysis of complex multiscale systems, AIChE J. 50 (7) (2004) 1346–1355.
[7] A. Rahimi, T. Darrell, B. Recht, Learning appearance manifolds from video, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 868–875.
[8] R. Lin, C. Liu, M. Yang, N. Ahuja, S. Levinson, Learning nonlinear manifolds from time series, in: Computer Vision, ECCV, 2006, pp. 245–256.
[9] R. Li, T.-P. Tian, S. Sclaroff, Simultaneous learning of nonlinear manifold and dynamical models for high-dimensional time series, in: IEEE 11th International Conference on Computer Vision, ICCV-2007, 2007, pp. 1–8.
[10] J. Macke, J. Cunningham, M. Byron, K. Shenoy, M. Sahani, Empirical models of spiking in neural populations, in: J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F. Pereira, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 24, 2011, pp. 1350–1358.
[11] A. Singer, R. Coifman, Non-linear independent component analysis with diffusion maps, Appl. Comput. Harmon. Anal. 25 (2008) 226–239.
[12] A. Haddad, D. Kushnir, R.R. Coifman, Texture separation via a reference set, Appl. Comput. Harmon. Anal. 36 (2) (2014) 335–347.
[13] R. Talmon, I. Cohen, S. Gannot, R. Coifman, Supervised graph-based processing for sequential transient interference suppression, IEEE Trans. Audio, Speech Language Process. 20 (9) (2012) 2528–2538.
[14] L. Wiskott, T.J. Sejnowski, Slow feature analysis: unsupervised learning of invariances, Neural Comput. 14 (2002) 715–770.
[15] S. Amari, H. Nagaoka, Methods of Information Geometry, American Mathematical Society, 2007.
[16] G. Storvik, Particle filters for state-space models with the presence of unknown static parameters, IEEE Trans. Signal Process. 50 (2002) 281–289.
[17] S. Godsill, J. Vermaak, W. Ng, J. Li, Models and algorithms for tracking of maneuvering objects using variable rate particle filters, Proc. I.E.E.E. 95 (2007) 925–952.
[18] R. Talmon, R. Coifman, Empirical intrinsic geometry for nonlinear modeling and time series filtering, Proc. Natl. Acad. Sci. USA 110 (31) (2013) 12535–12540.
[19] R. Kalman, A new approach to linear filtering and prediction problems, Trans. ASME J. Basic Eng. 82 (1960) 34–45.
[20] Y. Bar-Shalom, Tracking and Data Association, Academic Press Professional, 1987.

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006

[21] S. Julier, J. Uhlmann, Unscented filtering and nonlinear estimation, Proc. I.E.E.E. 92 (2004) 401–422.
[22] A. Doucet, S. Godsill, C. Andrieu, On sequential Monte Carlo sampling methods for Bayesian filtering, Stat. Comput. 10 (3) (2000) 197–208.
[23] M. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, IEEE Trans. Signal Process. 50 (2003) 174–188.
[24] O. Cappé, S. Godsill, E. Moulines, An overview of existing methods and recent advances in sequential Monte Carlo, Proc. I.E.E.E. 95 (5) (2007) 899–924.
[25] S. Belomgie, J. Malik, J. Puzicha, Shape matching and object recognition using shape context, IEEE Trans. Pat. Anal. Mach. Intell. 24 (4) (2002) 509–522.
[26] A.E. Johnson, M. Hebert, Using spin images for efficient object recognition in cluttered 3D scenes, IEEE Trans. Pat. Anal. Mach. Learn. 21 (5) (1999) 433–449.
[27] E.J. Candes, T. Tao, Near-optimal signal recovery from random projections: universal encoding strategies?, IEEE Trans. Inf. Theory 52 (12) (2006) 5406–5425.
[28] R. Talmon, S. Mallat, Z. Hitten, R. R. Coifman, Manifold learning for latent variable inference in dynamical systems, Tech. report YALEU/DCS/TR1491, 2014, submitted for publication, http://cpsc.yale.edu/sites/default/files/files/tr1491.pdf.
[29] D. Kushnir, A. Haddad, R. Coifman, Anisotropic diffusion on sub-manifolds with application to earth structure classification, Appl. Comput. Harmon. Anal. 32 (2) (2012) 280–294.
[30] G. Roberts, O. Stramer, Langevin diffusions and Metropolis–Hastings algorithms, Methodol. Comput. Appl. Probab. 4 (2003) 337–358.
[31] M. Girolami, B. Calderhead, Riemann manifold Langevin and Hamiltonian Monte Carlo methods, J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (2011) 123–214.
[32] F.R.K. Chung, Spectral Graph Theory, CBMS-AMS, 1997.
[33] P. Jones, M. Maggioni, R. Schul, Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels, Proc. Natl. Acad. Sci. USA 105 (6) (2008) 1803–1808.
[34] R. Talmon, D. Kushnir, R.R. Coifman, I. Cohen, S. Gannot, Parametrization of linear systems using diffusion kernels, IEEE Trans. Signal Process. 60 (3) (2012) 1159–1173.
[35] R. Talmon, I. Cohen, S. Gannot, Supervised source localization using diffusion kernels, in: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA'11, 2011.
[36] A. Bondy, U. Murty, Graph Theory, Springer, 2008.

Please cite this article in press as: R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. (2014), http://dx.doi.org/10.1016/j.acha.2014.08.006