In many natural and real-world applications, the measured signals are controlled by underlying processes or drivers. As a result, these signals exhibit highly redundant representations and their temporal evolution can be compactly described by a dynamical process on a low-dimensional manifold. In this paper, we propose a graph-based method for revealing the low-dimensional manifold and inferring the underlying process. This method provides intrinsic modeling for signals using empirical information geometry. We construct an intrinsic representation of the underlying parametric manifold from noisy measurements based on local density estimates. This construction is shown to be equivalent to an inverse problem, which is formulated as a nonlinear differential equation and is solved empirically through eigenvectors of an appropriate Laplace operator. The learned intrinsic nonlinear model exhibits two important properties. We show that it is invariant under different observation and instrumental modalities and is noise resilient. In addition, the learned model can be efficiently extended to newly acquired measurements in a sequential manner. We examine our method on two nonlinear filtering applications: a nonlinear and non-Gaussian tracking problem and 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.

Empirical Intrinsic Modeling of Signals and Information Geometry

Ronen Talmon and Ronald R. Coifman
Research Report YALEU/DCS/TR-1467
Yale University
November 9, 2012

Keywords: *Intrinsic model, nonlinear inverse problem, differential geometry, information geometry, non-parametric estimation, nonlinear dynamical systems*

1 Introduction

In many natural and real-world applications, the measured signals are controlled by underlying processes or drivers. As a result, the signals are often highly structured and lie on a manifold. These signals exhibit highly redundant representations and their temporal evolution can be compactly described by a dynamical process on a low-dimensional manifold, e.g. [1, 2, 3, 4, 5]. In recent years, there has been a growing effort to develop analysis methods based on the geometry of the acquired raw data [6, 7, 8, 9, 10]. These manifold learning techniques imply a completely different perspective than the classical approach in signal processing. Instead of relying on predefined models, this nonparametric approach aims to capture adaptively the geometry of the signal at hand and view its parametrization as a data-driven model. The nonlinear independent component analysis (NLICA), proposed in [11], is especially useful in signal processing, since the data are assumed to be inaccessible and can be observed only via unknown nonlinear measurement functions. The NLICA approach provides a parametrization of the manifold of the underlying processes, whereas classical manifold learning methods provide a parametrization of the observations. The main idea in [11] is to empirically invert the measurement function by solving local differential equations assuming that the function that maps the underlying process into a subset of observations is deterministic. This poses a major limitation since the measurements in many cases are noisy and related to the underlying process via a probabilistic model.

In this work, we utilize the widespread state-space approach to represent such cases, which are naturally formulated on a manifold. The state-space formalism includes two models: the dynamical model which consists of a stochastic differential equation describing the evolution of the underlying process (state) with time, and the measurement model that relates the noisy observations to the underlying process. The prior knowledge of the two models is necessary in many dynamic estimation problems. Specifically, it is required in Bayesian algorithms, e.g. the well-known Kalman filter and its extensions [12, 13, 14], and various sequential Monte Carlo algorithms [15, 16, 17]. Unfortunately, these models might be unknown and difficult to reveal. For example, Electroencephalography (EEG) recordings translate brain activity into sequences of electrical pulses [18]. We assume that the statistics of the pulses are controlled by an underlying process that characterizes the brain activity in an unknown way. The significance of revealing the model of the underlying process in EEG recordings is demonstrated in epilepsy research. In an ongoing work that will appear in a future publication, we attempt to detect and predict epileptic seizures based on the model of the recovered underlying process. Since samples of such an underlying process cannot be obtained, traditional Bayesian algorithms cannot be employed and nonparametric data-driven estimation methods are required.

In this paper, we propose a graph-based method using empirical information geometry for revealing the low-dimensional manifold and inferring the underlying process. The observation that the temporal statistics of the measurements convey the geometric information on the underlying process rather than the specific realization at-hand naturally leads to information geometry [19]. Unlike traditional information geometry analysis, we *empirically* learn the underlying manifold of local probability densities and recover their model. We propose to estimate the local probability densities of the measurements and view them as descriptors or features of the measurements that convey the desired information on the

underlying process. Then, we construct an intrinsic model of the underlying parametric manifold based on these features. This construction is shown to be equivalent to an inverse problem, which is formulated as a nonlinear differential equation and is solved empirically through eigenvectors of an appropriate Laplace operator. The role of the Laplace operator is to quantify the connections between the measurements and to integrate all the information. Specifically, its eigenvectors provide an embedding (or a parametrization) of the measurements, which is viewed as a representation of the underlying process on the parametric manifold. We discuss the relationship between the proposed embedding and information geometry. In particular, we show isometry between the proposed metric in the obtained intrinsic model and the Kullback Liebler divergence, which is defined on the parametric manifold using the Fisher Information matrix [20].

The proposed method exhibits two key properties that may be highly beneficial in a wide variety of real-world applications. We show that the learned intrinsic model is invariant under different observation modalities and is noise resilient. In addition, the construction of the graph is described with respect to a reference set of measurements [21, 22]. This property enables to obtain a graph-based model of a training signal in advance, and then, to extend the model to newly acquired measurements in a sequential manner.

We apply the proposed modeling method on two nonlinear filtering applications. The first 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, e.g. [23, 24]. We show that the obtained model represents the underlying process and is indeed noise resilient and invariant to the measurement function. The second application consists of a measurement modality described by a non-stationary hidden Markov chain. 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 [25], where in addition we proposed to define a Bayesian framework based on the obtained model. The Bayesian framework enables to filter, estimate and predict the underlying process, demonstrating the effectiveness of this approach in providing an empirical model for filtering tasks.

This paper is organized as follows. Section 2 presents the problem formulation, in which the parametric manifold and the measurement modality are described using the state-space formalism. In Section 3, we present the local density estimates of the signal as features of the measurements, and we describe their relationship to the underlying process. In Section 4, we derive an intrinsic metric and establish a relationship to the classical information geometry. In Section 5, we propose a graph-based algorithm to recover the underlying process. In addition, we address the probabilistic interpretation implied by the proposed method. Finally, in Section 6, experimentation results are presented, demonstrating the performance and the usefulness of the algorithm.

2 Problem Formulation

In this section, we adopt the state-space formalism to provide a generic problem formulation that may be adapted to a wide variety of applications. 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 as follows[1]

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

where a^i are unknown drift functions and $\dot{w}_t^i$ are independent white noises. For simplicity, we consider here normalized processes with unit variance noises. Since a^i are any drift functions, we may first apply normalization without effecting the following derivation. See [11] for details. We note that the underlying process is equivalent to the system state in the classical terminology of the state-space approach.

Let $\mathbf{y}_t$ denote an n-dimensional observation process in time index t, drawn from a probability density function (pdf) $f(\mathbf{y};\boldsymbol{\theta})$. The statistics of the observation process are time-varying and depend on the underlying process $\boldsymbol{\theta}_t$. We consider a model in which the clean observation process 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 $\mathbf{y}_t$.

The description of $\boldsymbol{\theta}_t$ constitutes a parametric manifold that controls the accessible measurements at-hand. Our goal in this work is to reveal the underlying process $\boldsymbol{\theta}_t$ and its dynamics based on a sequence of measurements $\{\mathbf{z}_t\}$.

3 Local Densities and Histograms

Let $p(\mathbf{z};\boldsymbol{\theta})$ denote the 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 $\mathbf{z}_t$ is a linear transformation of the pdf of the clean observation component $\mathbf{y}_t$.*

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

$$p(\mathbf{z};\boldsymbol{\theta}) = \int\limits_{g(\mathbf{y},\mathbf{v})=\mathbf{z}} f(\mathbf{y};\boldsymbol{\theta})q(\mathbf{v})d\mathbf{y}d\mathbf{v}. \tag{3}$$

□

We note that in the common 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, $p(\mathbf{z};\boldsymbol{\theta})$ in (3) becomes a linear convolution

$$p(\mathbf{z};\boldsymbol{\theta}) = \int\limits_{\mathbf{y}} f(\mathbf{y};\boldsymbol{\theta})q(\mathbf{z}-\mathbf{y})d\mathbf{y} = f(\mathbf{z};\boldsymbol{\theta}) * q(\mathbf{z}).$$

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

The dynamics of the underlying process are conveyed by the time-varying pdf of the measured process. Thus, this pdf may be very useful in revealing the desired underlying process and its dynamics. Unfortunately, the pdf is unknown since the underlying process and the dynamical and measurement models are unknown. Assume we have access to a class of estimators of the pdf over discrete bins which can be viewed as linear transformations. Let $\mathbf{h}_t$ be such an estimator with m bins which is viewed as an m-dimensional process and is given by

$$p(\mathbf{z};\boldsymbol{\theta}_t) \overset{\mathcal{T}}{\mapsto} \mathbf{h}_t, \tag{4}$$

where $\mathcal{T}$ is a linear transformation of the density $p(\mathbf{z};\boldsymbol{\theta})$ from the infinite sample space of $\mathbf{z}$ into a finite interval space of dimension m. By Lemma 1 and by definition (4) we get the following results.

Corollary 1. *The process $\mathbf{h}_t$ is a linear transformation of the pdf of the clean observation component $\mathbf{y}_t$.*

Corollary 2. *The process $\mathbf{h}_t$ can be described as a deterministic nonlinear map of the underlying process $\boldsymbol{\theta}_t$.*

In this work, we use histograms as estimates of the pdf, and we assume that a *sequence* of measurements is available. Accordingly, let $\mathbf{h}_t$ be the empirical local histogram of the measured process $\mathbf{z}_t$ in a short-time window of length L_1 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

$$h_t^j = \frac{1}{|\mathcal{H}_j|}\frac{1}{L_1}\sum_{s=t-L_1+1}^{t} \mathbf{1}_{\mathcal{H}_j}(\mathbf{z}_s), \tag{5}$$

where $\mathbf{1}_{\mathcal{H}_j}(\mathbf{z}_t)$ is the indicator function of the bin $\mathcal{H}_j$ and $|\mathcal{H}_j|$ is its cardinality. By assuming (unrealistically) that infinite number of samples are available and that their density in each histogram bin is uniform, (5) can be expressed as

$$h_t^j = \frac{1}{|\mathcal{H}_j|}\int\limits_{\mathbf{z}\in\mathcal{H}_j} p(\mathbf{z};\boldsymbol{\theta})d\mathbf{z}. \tag{6}$$

Thus, ideally the histograms are *linear transformations* of the pdf. In addition, if we shrink the bins of the histograms as we get more and more data, the histograms converge to the pdf

$$\mathbf{h}_t \xrightarrow[|\mathcal{H}_j|\to 0]{L_1\to\infty} p(\mathbf{z};\boldsymbol{\theta}). \tag{7}$$

In practice, since the computation of high-dimensional histograms is challenging, we propose to preprocess high-dimensional data by applying random filters in order to reduce the dimensionality without corrupting the information. However, this extends the scope of this paper and will appear in a future publication.

4 Intrinsic Metric Computation Using Empirical Information Geometry

In classical information geometry, the parameters of the distribution of the observations confine the data to an underlying manifold. Thus, the distribution is usually required in an analytic form. In this section, we propose a data-driven approach to recover the manifold without a prior knowledge of the distribution of the observations. Instead, we propose to rely on the histograms or the local density estimates described in Section 3.

4.1 Mahalanobis Distance

We view $\mathbf{h}_t$ (the linear transformation of the local densities, e.g. the local histograms) as feature vectors for each measurement $\mathbf{z}_t$. By Corollary 2 and (1), the process $\mathbf{h}_t$ satisfies the dynamics given by Itô's lemma

$$\begin{aligned} h_t^j &= \sum_{i=1}^{d} \left(\frac{1}{2} \frac{\partial^2 h^j}{\partial \theta^i \partial \theta^i} + a^i \frac{\partial h^j}{\partial \theta^i} \right) dt \\ &+ \sum_{i=1}^{d} \frac{\partial h^j}{\partial \theta^i} dw_t^i, \ j = 1, \ldots, m. \end{aligned} \tag{8}$$

For simplicity of notation, we omit the time index t from the partial derivatives. According to (8), the (j,k)th element of the $m \times m$ covariance matrix $\mathbf{C}_t$ of $\mathbf{h}_t$ is given by

$$C_t^{jk} = \mathrm{Cov}(h_t^j, h_t^k) = \sum_{i=1}^{d} \frac{\partial h^j}{\partial \theta^i} \frac{\partial h^k}{\partial \theta^i}, \ j,k = 1, \ldots, m. \tag{9}$$

In matrix form, (9) can be rewritten as

$$\mathbf{C}_t = \mathbf{J}_t \mathbf{J}_t^T \tag{10}$$

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

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

Thus, the covariance matrix $\mathbf{C}_t$ is a semi-definite positive matrix of rank d.

We define a nonsymmetric $\mathbf{C}$-dependent squared distance between pairs of measurements as

$$a_{\mathbf{C}}^2(\mathbf{z}_t, \mathbf{z}_s) = (\mathbf{h}_t - \mathbf{h}_s)^T \mathbf{C}_s^{-1} (\mathbf{h}_t - \mathbf{h}_s) \tag{11}$$

and a corresponding symmetric distance as

$$d_{\mathbf{C}}^2(\mathbf{z}_t, \mathbf{z}_s) = 2(\mathbf{h}_t - \mathbf{h}_s)^T \left(\mathbf{C}_t + \mathbf{C}_s \right)^{-1} (\mathbf{h}_t - \mathbf{h}_s). \tag{12}$$

Since usually the dimension d of the underlying process is smaller than the number of histogram bins m, the covariance matrix is singular and non-invertible. Thus, in practice we use the pseudo-inverse to compute the inverse matrices in (11) and (12).

The distance in (12) is known as the *Mahalanobis distance* with the property that it is invariant under linear transformations. Thus, by Lemma 1 and Corollary 1, it is invariant to the measurement noise and function (e.g., additive noise or multiplicative noise). We note however that the linear transformation employed by the measurement noise on the observable pdf (3) may degrade the available information. For example, 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 employed by the noise significantly attenuates the connection between the measurements and the underlying process. Therefore, we expect to exhibit noise resilience up to a certain noise level as long as the observable pdfs can be regarded as functions of the underlying process. Above this level, we expect to experience a sudden drop in the performance.

In addition, by Lemma 3.1 in [26] and by Corollary 2, the Mahalanobis distance in (12) 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 the nonlinear map of $\boldsymbol{\theta}_t$ to $\mathbf{h}_t$, and is given by

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_s\|^2 = d^2_{\mathbf{C}}(\mathbf{z}_t, \mathbf{z}_s) + O(\|\mathbf{h}_t - \mathbf{h}_s\|^4). \tag{13}$$

For more details see [11] and [26]. Assuming there is an intrinsic map $i(\mathbf{h}_t) = \boldsymbol{\theta}_t$ from the feature vector to the underlying process, the approximation in (13) is equivalent to the inverse problem defined by the following nonlinear differential equation

$$\sum_{i=1}^{m} \frac{\partial \theta^j}{\partial h^i} \frac{\partial \theta^k}{\partial h^i} = \left[C_t^{-1}\right]^{jk}, \ j,k = 1, \ldots, d. \tag{14}$$

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

4.2 Local Covariance Matrix Estimation

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

The pdf estimates and the local clouds implicitly define two time scales on the sequence of measurements. The fine time scale is defined by short-time windows of L_1 measurements to estimate the temporal pdf. The coarse time scale is defined by the local neighborhood of L_2 neighboring feature vectors in time. Accordingly, we note that the approximation in (13) is valid as long as the statistics of the noise are locally fixed in the short-time windows of length L_1 (i.e., slowly changing compared to the fast variations of the underlying process) and the fast variations of the underlying process can be detected in the difference between the feature vectors in windows of length L_2.

According to the dynamical model in (1) and (8), the samples in the local cloud can be seen as small perturbations of the pivot sample created by the noise $\mathbf{w}_t$. Thus, we assume

that the samples share similar local probability densities[2]and may be used to estimate the local covariance matrix, which is required for the construction of the Mahalanobis metric (12). The empirical covariance matrix of the cloud is estimated by

$$\begin{aligned}\hat{\mathbf{C}}_{t_0} &= \frac{1}{L_2}\sum_{s=1}^{L_2}\left(\mathbf{h}_{t_0,s}-\hat{\boldsymbol{\mu}}_{t_0}\right)\left(\mathbf{h}_{t_0,s}-\hat{\boldsymbol{\mu}}_{t_0}\right)^T \\ &\simeq \mathbb{E}\left[\left(\mathbf{h}_{t_0}-\mathbb{E}[\mathbf{h}_{t_0}]\right)\left(\mathbf{h}_{t_0}-\mathbb{E}[\mathbf{h}_{t_0}]\right)^T\right]=\mathbf{C}_{t_0}\end{aligned} \tag{15}$$

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

As the rank of the matrix d is usually smaller than the covariance matrix dimension m, in order to compute the inverse matrix we use only the d principal components of the matrix. This operation "cleans" the matrix and filters out noise. In addition, when the empirical rank of the local covariance matrices of the feature vectors is lower than d, it indicates that the available feature vectors are insufficient and a larger cloud should be used.

4.3 Relationship to Information Geometry

In this section we consider different features to convey the statistical information of the data. Let $\mathbf{l}_{t_0,t}$ be a new feature vector defined by

$$l^j_{t_0,t}=\sqrt{h^j_{t_0}}\log\left(h^j_{t_0,t}\right) \tag{16}$$

where t_0 is the index of the pivot sample of the cloud of t. We note that this choice of features is no longer a linear transformation, and therefore, the metric in (12) 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 1. *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.,*

$$I^{ii'}_{t_0}\simeq\mathbb{E}\left[\frac{\partial}{\partial\theta^i_{t_0}}\log\left(p(\mathbf{z};\boldsymbol{\theta}_{t_0})\right)\frac{\partial}{\partial\theta^{i'}_{t_0}}\log\left(p(\mathbf{z};\boldsymbol{\theta}_{t_0})\right)\right]. \tag{17}$$

Proof. See Appendix 8. □

By (10) and Theorem 1, we get that the singular value decomposition 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 (16). Let $\{\rho_j,\boldsymbol{\upsilon}_j,\boldsymbol{\nu}_j\}_j$ be the singular values, singular left vectors, and singular right vectors of the Jacobian matrix $\mathbf{J}_{t_0}$. Then, by (10) and Theorem 1, $\mathbf{C}_{t_0}$ 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 (11) and (12) it is used to define an intrinsic metric between feature vectors (pdf estimates of the

[2]We emphasize that we consider the statistics of the feature vectors and not the feature vectors themselves, which are estimates of the varying statistics of the raw measurements.

measurements) that reveals the underlying process (13). On the other hand, by Theorem 1, $\boldsymbol{\nu}_j$ are the eigenvectors of the Fisher Information matrix of the measurements. According to [20], it approximates the Kullback Liebler divergence between densities of measurements in the cloud, i.e.,

$$\mathcal{D}(p(\mathbf{z}_{t_0,t};\theta_{t_0,t})\|p(\mathbf{z}_{t_0};\theta_{t_0})) = \delta\boldsymbol{\theta}_t^T \mathbf{I}_{t_0} \delta\boldsymbol{\theta}_t \tag{18}$$

where $\delta\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t_0,t} - \boldsymbol{\theta}_{t_0}$. Using Theorem 1, the divergence (18) can be computed based on the cloud of samples according to

$$J_{t_0}^{ji}\delta\theta_{t_0}^i = \lim_{\theta_{t_0,t}^i \to \theta_{t_0}^i} l_{t_0,t}^j - l_{t_0}^j$$

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

$$J_{t_0}^{ji}\delta\theta_{t_0}^i \simeq \frac{1}{L}\sum_{t=1}^{L}\left(l_{t_0,t}^j - l_{t_0}^j\right).$$

Thus, we obtain an isometry between the "external" intrinsic metric of the measurements (11)-(12) and the "internal" metric of the pdfs (18).

In addition, according to information geometry approaches [27, 20], the inverse of the Fisher Information matrix can be used to restrict the stochastic measurement process to the parametric lower-dimensional manifold that is imposed by the underlying process[3].

These results support the choice of local density estimates as appropriate features that convey the measurements information. In Section 5, we describe a constructive method to empirically reveal the parametric manifold by recovering the underlying process without assuming any particular statistical model of the measurements.

5 Graph-based Algorithm for Intrinsic Modeling

5.1 Intrinsic Embedding

Let $\{\bar{\mathbf{z}}_s\}_{s=1}^N$ be a sequence of N reference measurements. The availability of a sequence of measurements with corresponding time labels enables us to estimate the local densities and their covariance matrices as described in Section 3 and Section 4. Let $\{\mathbf{z}_t\}_{t=1}^M$ be another sequence of M arbitrary measurements. As proposed in [26, 28, 29], we define a "one-sided" kernel consisting of an affinity measure between the two sets of measurements. We construct an $M \times N$ nonsymmetric affinity matrix $\mathbf{A}$, whose (t,s)th element is given by

$$A^{ts} = \exp\left\{-\frac{a_{\boldsymbol{\Sigma}}^2(\mathbf{z}_t, \bar{\mathbf{z}}_s)}{\varepsilon}\right\}, \tag{19}$$

where $\varepsilon > 0$ is a tunable scale. The construction of (19) requires the corresponding feature vectors of the measurements and the local covariance matrix of merely the reference measurement $\bar{\mathbf{z}}_s$ and does not use the covariance matrix of the measurement $\mathbf{z}_t$. The significance of the distinction to two sets of measurements and the significance of the latter comment will

[3]Recall that the inverse of the local covariance matrix is used to locally invert the measurement function and compute intrinsic metric on the manifold.

become apparent in Section 5.2, where we describe the extension of the following derivation to support sequential processing. The one-sided kernel defines a bipartite graph [30], 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}.

Let $\mathbf{W}_r$ be a "two-sided" symmetric kernel of size $N \times N$, defined as

$$\mathbf{W}_r = \mathbf{A}^T\mathbf{A}. \tag{20}$$

By definition, the (s, s')th element of the two-sided kernel is given by

$$W_r^{ss'} = \sum_t A^{ts}A^{ts'},$$

which implies that the two-sided kernel can be interpreted as an affinity metric between any two reference measurements $\bar{\mathbf{z}}_s$ and $\bar{\mathbf{z}}_{s'}$ via all the measurements in the set $\{\mathbf{z}_t\}$.

The kernel $\mathbf{W}_r$ is then normalized by

$$\mathbf{W}_{r,norm} = \mathbf{D}_r^{-1}\mathbf{W}_r\mathbf{D}_r^{-1}, \tag{21}$$

where $\mathbf{D}_r$ is a diagonal matrix, whose sth diagonal term is given by

$$D_r^{ss} = \sum_{s'=1}^{N} W_r^{ss'}.$$

$\mathbf{D}_r$ is often called a density matrix as D_r^{ss} approximates the local density in the vicinity of $\bar{\mathbf{z}}_s$, and hence, the normalization handles nonuniform sampling of the measurements [21]. The kernel $\mathbf{W}_{r,norm}$ is made row stochastic by $\mathbf{W}_{r,rs} = \mathbf{D}_{rs}^{-1}\mathbf{W}_{r,norm}$, where $\mathbf{D}_{rs}$ is a diagonal matrix with diagonal elements

$$D_{rs}^{ss} = \sum_{s'=1}^{N} W_{r,norm}^{ss'}.$$

Next, we compute the eigenvalues $\{\lambda_i\}_{i=1}^N$ and eigenvectors $\{\boldsymbol{\varphi}_i\}_{i=1}^N$ of $\mathbf{W}_{r,rs}$. The eigenvalues, which are nonnegative and bounded by 1 due to the normalization, are sorted such that $\lambda_1 = 1$. The corresponding first eigenvector is trivial and equals to a column vector of ones $\boldsymbol{\varphi}_1 = \mathbf{1}$. By [11] and [26], $\mathbf{I} - \mathbf{W}_{r,rs}$ converges to a diffusion operator (Laplace-Beltrami) that reveals the low-dimensional manifold, and the eigenvectors give an approximate parametrization of the parametric manifold. Specifically, the leading d eigenvectors (except the trivial), which are a local coordinate system for the manifold [31], recover d proxies for the underlying process up to a monotonic scaling [26]. In other words, they are empirical solutions to the inverse problem described by the differential equation in (14). In addition, the eigenvectors are independent in case the manifold is flat [11]. Thus, without loss of generality, we may write

$$\varphi_i^s = \varphi_i(\bar{\theta}_s^i),\ i = 1, \ldots, d; s = 1, \ldots, N,$$

where $\varphi_i(\cdot)$ is a monotonic function and $\bar{\boldsymbol{\theta}}_s$ is the sample of the underlying process corresponding to $\bar{\mathbf{z}}_s$. Based on the eigenvectors, we define a d-dimensional representation of any sample in the reference set by the following embedding

$$\boldsymbol{\Phi}(\bar{\mathbf{z}}_s) \triangleq [\varphi_1^s, \varphi_2^s, \ldots, \varphi_d^s],\ s = 1, \ldots, N. \tag{22}$$

By the monotonicity of the eigenvectors, the embedded domain organizes the reference measurements according to the values of the underlying process. Accordingly, we view the proposed embedding as *empirical modeling* of the measurements representing the underlying process.

The kernel $\mathbf{W}_{r,rs}$ is similar to the following symmetric kernel

$$\tilde{\mathbf{W}}_r = \mathbf{D}_{rs}^{-1/2}\mathbf{W}_{r,norm}\mathbf{D}_{rs}^{-1/2}$$

via $\tilde{\mathbf{W}}_r = \mathbf{D}_{rs}^{1/2}\mathbf{W}_{r,rs}\mathbf{D}_{rs}^{-1/2}$. Thus, $\mathbf{W}_{r,rs}$ and $\tilde{\mathbf{W}}_r$ share the same eigenvalues, and the eigenvectors of $\tilde{\mathbf{W}}_r$ are given by

$$\tilde{\boldsymbol{\varphi}}_j = \mathbf{D}_{rs}^{1/2}\boldsymbol{\varphi}_j.$$

In addition, by Lemma 3.1 in [26], the kernel $\tilde{\mathbf{W}}_r$ is approximated to a second order by

$$\tilde{W}_r^{ts} = \frac{c}{D_r^{tt}D_r^{ss}} \exp\left\{-\frac{d_{\mathbf{C}}^2(\bar{\mathbf{z}}_t, \bar{\mathbf{z}}_s)}{\varepsilon}\right\} \tag{23}$$

where c is a constant that can be estimated from the measurements (See details in [26]). Combining the relationship between the kernels and the approximations in (13) and (23) yields that $\mathbf{W}_{r,rs}$ measures the affinity between the measurements according to the distance between the corresponding samples of the underlying process. It is invariant to the observation modality and it is resilient to measurement noise. This property has a key role since it enables to reveal the underlying process that represents the parametric manifold rather than the manifold of the measurements. In addition, it enables to compute the kernel $\mathbf{W}_{r,rs}$ directly from the reference measurements.

Next, we define a row stochastic $M \times N$ matrix $\mathbf{A}_{rs}$ as

$$\mathbf{A}_{rs} = \mathbf{D}^{-1}\mathbf{A}\mathbf{D}_{rs}^{-1} \tag{24}$$

where $\mathbf{D}$ is a diagonal matrix whose diagonal terms are given by

$$D^{tt} = \sum_{s=1}^{N} \frac{A^{ts}}{D_r^{ss}}.$$

Let $\mathbf{W}$ be an $M \times M$ symmetric kernel matrix of the set $\{\mathbf{z}_t\}$, defined as

$$\mathbf{W} = \mathbf{A}_{rs}\mathbf{A}_{rs}^T.$$

Similarly to the interpretation of $\mathbf{W}_r$, the (t,τ)th element of $\mathbf{W}$ can be interpreted as an affinity metric between any pair of measurements $\mathbf{z}_t$ and $\mathbf{z}_\tau$ via all the reference measurements in $\{\bar{\mathbf{z}}_s\}$. It further implies that two measurements are similar if they "see" the reference measurements in the same way.

By the construction of the kernel $\mathbf{W}_{r,rs}$ in (20) and (21), it can be shown that the eigenvectors $\{\boldsymbol{\varphi}_i\}$ of $\mathbf{W}_{r,rs}$ are the singular right vectors of $\mathbf{A}_{rs}$ [21]. Furthermore, the eigenvectors $\{\boldsymbol{\psi}_i\}$ of $\mathbf{W}$ are the singular left vectors of $\mathbf{A}_{rs}$. As discussed in [21] and [26], the right singular vectors of $\mathbf{A}_{rs}$ represent the underlying process of the reference measurements $\{\bar{\mathbf{z}}_s\}$, and the left singular vectors of $\mathbf{A}_{rs}$ naturally extend the representation

to the measurements $\{\mathbf{z}_t\}$. Then, the spectral representation of the kernel can be efficiently extended from the reference measurements to arbitrary measurements by the following relationship between the singular vectors

$$\boldsymbol{\psi}_i = \frac{1}{\sqrt{\lambda_i}} \mathbf{A}_{rs} \boldsymbol{\varphi}_i, \tag{25}$$

which involves only the information associated with the reference measurements. In particular, as mentioned before, the calculation does not involve the local covariance of the arbitrary measurements.

We define a d-dimensional representation similarly to (22) by

$$\boldsymbol{\Psi}(\mathbf{z}_t) \triangleq \left[\boldsymbol{\psi}_1^t, \boldsymbol{\psi}_2^t, \ldots, \boldsymbol{\psi}_d^t\right], \tag{26}$$

for each $\mathbf{z}_t$. Then, the embedding (26) is seen as the obtained modeling of the measurements revealing the corresponding underlying process.

We note 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 reference set.

5.2 Sequential Processing

The construction of the embedding described in Section 5.1 is especially suitable for sequential processing [22]. Here, we describe a supervised algorithm consisting of two stages: a training stage in which a sequence of training measurements is assumed to be available in advance, and a test stage in which new incoming measurements are sequentially processed.

In the training stage, reference measurements taken from the training sequence are processed to form a learned model. The feature vectors and the corresponding local covariance matrices associated with the reference samples are computed. The row-stochastic kernel $\mathbf{W}_{r,rs}$, defined on the reference measurements, is directly computed from the training measurements, and its eigen-decomposition is calculated. The eigenvectors of the kernel form a learned model for the reference set. Then, we are able to construct the map (22) which provides an embedding for the reference set and reveals the underlying process. We store the feature vectors of the reference set along with their local covariance matrices and the corresponding embedding.

In the test stage, as new incoming measurements $\{\mathbf{z}_t\}$ become available, we construct $\mathbf{A}$ and $\mathbf{A}_{rs}$ according (19) and (24), respectively. Finally, we compute the extended representation by (25), and acquire the embedding of the new samples by (26). It is worthwhile noting that the covariance matrices of new measurements are not required for the extension. Thus, this scheme 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 involve low computational complexity. For detailed analysis of the computational burden we refer the readers to [22].

5.3 Probabilistic Interpretation

We consider a probabilistic model consisting of a mixture of local statistical models implied by the reference set. 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 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 of any measurement $\mathbf{z}_t$ to be 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

$$\alpha(t, s) = \Pr(\mathbf{z}_t | \mathbf{z}_t \in \mathbf{Z}_s) \tag{27}$$

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 2. *Assuming statistically independent measurements, the elements of the $M \times M$ kernel matrix $\mathbf{W}_\alpha$ are the conditional probability of a pair of given measurements to be 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.

Thus, the kernel matrix $\mathbf{W}_\alpha$ measures the probability of any two measurements to be associated with the same reference measurement. This result extends the result presented in [22] obtained for Gaussian kernels.

Proof. See Appendix 9 □

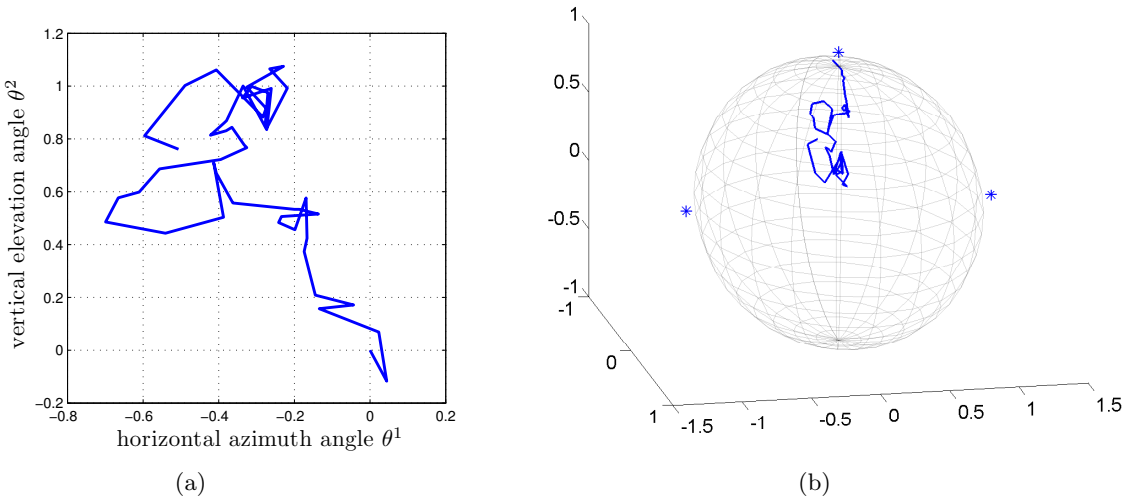
The definition of $\mathbf{A}_{rs}$ in (11) and (19) implies that this kernel is a special case of $\mathbf{A}_\alpha$, when the conditional probability of a measurement $\mathbf{z}_t$ given it is associated with $\mathcal{Z}_s$ (27) is defined as a normal distribution with $\bar{\mathbf{z}}_s$ mean and $\mathbf{C}_s$ covariance matrix. Thus, the construction of the metric and the kernel in Section 4 and Section 5.1 naturally implies an implicit multi-Gaussian mixture model in the measurements domain. Each reference measurement at time s represents a local (infinitesimal) Gaussian model, and the metric defined by (11) and (19) computes the probability of any measurement at time t to be associated with the local model represented by the reference measurement at s.

6 Experimental Results

6.1 Parametric Model

We simulate an underlying process whose temporal propagation mimics a 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

$$d\theta_t^i = -\nabla U^i(\theta_t^i)dt + dw_t^i \tag{28}$$



(a) (b)

Figure 1: (a) A segment of the 2-dimensional trajectory of the 2 coordinates of the underlying process: the horizontal and vertical angles. (b) A corresponding segment of the 3-dimensional movement of the object on the sphere. The locations of the 3 sensors are marked with $*$.

where $\dot{w}_t^i$ are independent white Gaussian noises, 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 low-dimensional manifold. We note that this propagation model is chosen for demonstration since it is general and may be used to describe many natural signals and phenomena.

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 object 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 evolve in time according to the Langevin equation (28). 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 object 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 object is given by

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

where r is the radius of the sphere. Figure 1 illustrates a segment of the 2-dimensional underlying process and a corresponding segment of the 3-dimensional trajectory of the radiating object on the sphere.

To examine the robustness of the proposed method to different measurements, we consider three measurement schemes. In Scheme 1, the radiation of the object is measured by 3 "Geiger counter" sensors positioned at $\mathbf{x}_j, j = 1, 2, 3$ outside the sphere (see Fig. 1). The sensors detect the radiation and fire spikes through a spatial point process in a varying rate which depends on the proximity of the object to the sensors (the closer the object is, the higher the amount of radiation reaching the sensor). Each sensor fires spikes according to a Poisson distribution with rate $\lambda^j(\boldsymbol{\theta_t}) = \exp\{-\|\mathbf{x}_j - \mathbf{x}(\boldsymbol{\theta_t})\|\}$. We obtain three spike sequences y_t^j in which the firing rate is higher when the object is closer to the sensor. The output of each sensor is corrupted by additive noise and is given by

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

where v_t^j is a spike sequence drawn from a Poisson distribution with a fixed rate λ_v^j. Scheme 2 is similar to Scheme 1. Each sensor fires spikes randomly according to the proximity of the object. The difference is that in this scheme 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 Scheme 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 value in Scheme 2 is of unknown type. In Scheme 3, we consider a measurement of a different nature. We use three sensors that measure the location of the source directly, i.e.

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

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

Under all the measurement schemes, 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 dynamical 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 (28) and the corresponding noisy measurements under the three measurement schemes. 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 $L_1 = 10$ (with full overlap) and obtain the feature vectors. Then, we estimate the local covariance matrix according to (16) with $L_2 = 10$. Next, the kernels and the embedding of the reference measurements (22) is constructed as described in Section 5.

Figure 2 presents the spectrum of the local covariance matrices of a sequence of 200 feature vectors. Each curve describes 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 covariance matrices have rank $d = 2$. This reveals the two degrees of freedom stored in the data, which are the two spherical angles.

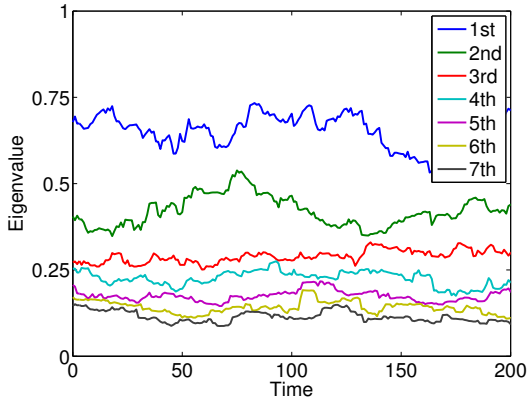
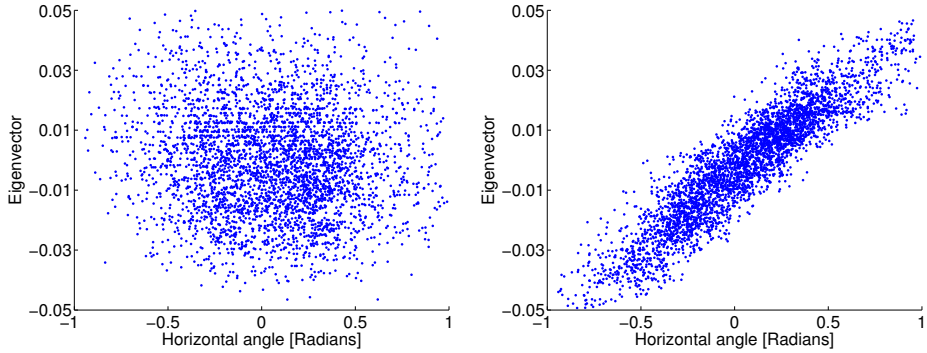


Figure 2: The spectrum of the local covariance matrices of a sequence of 200 feature vectors.



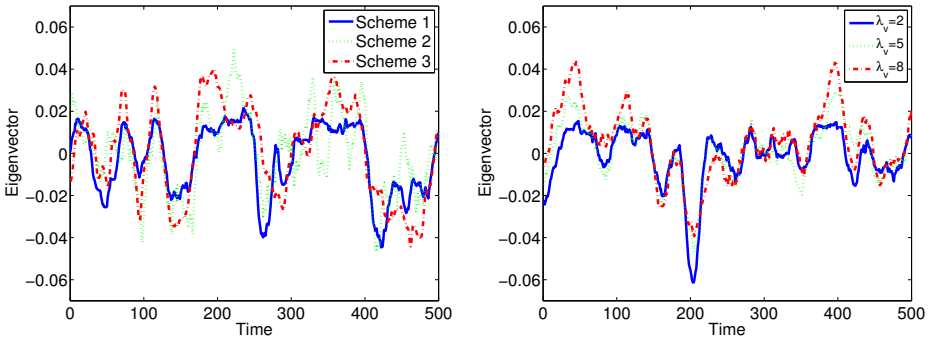
(a) (b)

Figure 3: Scatter plots of the leading eigenvector and the horizontal angle. (a) The leading eigenvector obtained by the NLICA. (b) The leading eigenvector obtained by the proposed method.

Measurements following the reference sequence at times $t > 2000$ are sequentially processed, i.e., for each measurement the kernel matrix (19) and the extended embedding (26) are computed, as described in Section 5.2.

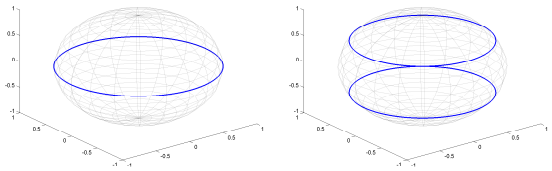
Figure 3 shows a scatter plot of the leading eigenvector and the horizontal angle under Scheme 1. In Fig. 3(a) the embedding is obtained using the NLICA [11] and in Fig. 3(b) the embedding is obtained using the proposed method. We observe that the leading eigenvector obtained via the proposed method is linearly correlated with the horizontal angle, whereas the leading eigenvector obtained using the NLICA is uncorrelated with the angle. We note that no correlation is found between any other pair of an angle and an eigenvectors. Thus, it shows that the obtained embedding is a good representation of the angle. Furthermore, the comparison to the embedding obtained by the NLICA suggests that the use of histograms as feature vectors is an essential component to convey the local statistical information.

In Fig. 4 we compare the modeling of the vertical angle obtained under different measurement schemes and noise. We note that the presented 500 coordinates of the eigenvectors are computed by extension. Figure 4(a) depicts three eigenvectors that correspond to the



(a) (b)

Figure 4: A comparison of the obtained modeling of the vertical angle. (a) A comparison between the obtained eigenvectors (corresponding to the vertical angle) under the three different measurement schemes (measuring the same movement). (b) A comparison between the obtained eigenvectors (corresponding to the vertical angle) under the first measurement scheme with different noise levels ($\lambda_v = 2, 5, 8$).

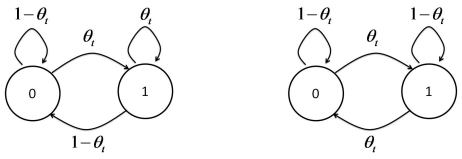


(a) (b)

Figure 5: An illustration of the movement of the two objects confined to rings on a sphere. (a) The solid curve represents the simulated movements of the two objects on the same ring. (b) The solid curves represent the simulated movements of the two objects on two different rings.

same movement of the object. Each eigenvector is obtained under a different measurement scheme. We observe that the three eigenvectors follow the same trend, which implies intrinsic modeling of the movement and demonstrates the invariability of the proposed approach to the measurement scheme. We emphasize that the measurements under the three schemes are very different in their nature: spike sequences in Schemes 1 and 2 and noisy 3-dimensional coordinates in Scheme 3. In order to further demonstrate the resilience of the modeling to measurement noise, we present in Fig. 4(b) three eigenvectors obtained under Scheme 1 with noise sequence of spikes $\mathbf{v}_t$ in three different rates. We observe that the three eigenvectors follow the same trend and hence conclude that they *intrinsically* model the movement of the object. We note that the connection between the pdf estimates and the underlying process in higher noise levels becomes very weak, and thus, as discussed in Section 4, we experience a sudden drop in the correlation between the obtained eigenvectors and the underlying angles.

In a second experiment, we alter the experimental setup as follows. We now consider *two*



(a) (b)

Figure 6: Two Markov chains configurations. (a) A Bernoulli scheme. (b) A Markov scheme of order 1.

radiating objects moving on the sphere. The movement of each object $i = 1, 2$ is confined to a horizontal ring and controlled by the azimuth angle θ_t^i, which is the single underlying parameter of the movement. We simulate two movement scenarios. In the first, the two objects move on the same ring (Fig. 5(a)), and in the second, the two objects move on two different rings ((Fig. 5(b)). The total amount of radiation from the two objects 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 recovered by the eigenvectors. In this experiment, each object has a different structure, which is separated and recovered by the proposed algorithm. In terms of the problem formulation, the difference between the two experiments is conveyed by different measurement schemes. Hence, this experiment further demonstrates the robustness of the proposed nonparametric blind algorithm. Furthermore, it illustrates the potential of the proposed approach to yield good performance in blind source separation applications.

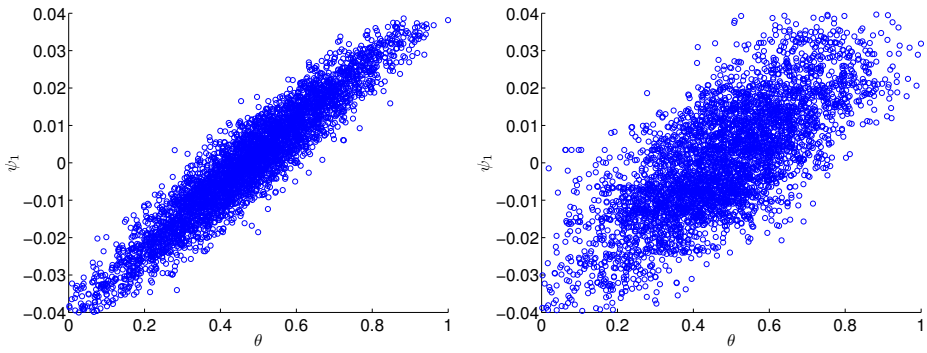
6.3 Non-stationary Hidden Markov Chain

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 a potential field corresponding to a single Gaussian with 0.5 mean and 5 variance and clip values outside $[0, 1]$. The clean observation process is measured with additive zero-mean 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. We process the entire interval of $N = 4000$ measurement samples and compute their embedding directly without extension.

We examine two Markov chain configurations: a Bernoulli scheme and a scheme of order 1 in which the transition depends merely 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. This context-dependency makes it different from the former scheme and from the experiments described in Section 6.2.



(a) (b)

Figure 7: Scatter plots of the leading eigenvector and the underlying process under the Bernoulli scheme. (a) The leading eigenvector obtained by the NLICA. (b) The leading eigenvector obtained by the proposed method.

The Bernoulli scheme is illustrated in Fig. 6(a), and the observation process is given by

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

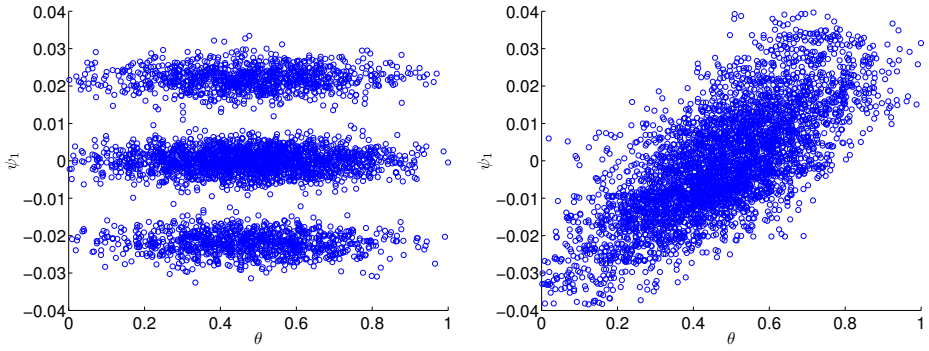
We note that in this specific scenario, revealing the underlying process can be easily done by short-time averaging, since

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

Figure 7 presents scatter plots of the obtained embedding and the underlying process θ_t. In Fig. 7(a) we show the embedding based on the means of the measurements in short-time windows obtained via the NLICA. In other words, the means of short-time windows are viewed and feature vectors and processed using the NLICA. In Fig. 7(b) we show the embedding based on the histograms of the measurements in short-time windows obtained via the proposed method. As expected, exploiting the prior knowledge that the underlying process can be revealed by short-time averaging yields good performance. The embedding based on the means in short windows is highly correlated to the underlying process. Moreover, the correlation is stronger than the correlation obtained using the proposed method, which does not use any a-priori knowledge.

The second scheme is illustrated in Fig. 6(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 first order 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 higher order dependency would require the processing of higher order derivatives or several consecutive measurements together.

Figure 8 presents scatter plots of the obtained embedding and the underlying process θ_t. In Fig. 8(a) we show the embedding based on the means of the difference signal in short-time windows obtained via the NLICA. In Fig. 8(b) we show the embedding based on the histograms of the measurements in short-time windows obtained via the proposed method. We observe that the embedding based on the means in short windows is degenerate and



(a) (b)

Figure 8: Scatter plots of the leading eigenvector and the underlying process under the Markov scheme of order 1 . (a) The leading eigenvector obtained by the NLICA. (b) The leading eigenvector obtained by the proposed method.

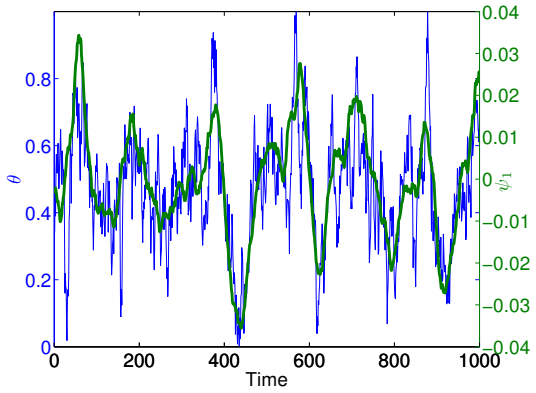


Figure 9: The leading eigenvector obtained using the proposed method and the underlying process as a function of time.

does not convey any information on the underlying process. On the other, the embedding obtained using the proposed method exhibits high correlation with the underlying process. To further illustrate the obtained modeling of the time series, we present in Fig. 9 the leading eigenvector obtained using the proposed method and the underlying process as a function of time. It can be seen that the eigenvector tracks accurately the drift of the underlying process, whereas the small (noisy) perturbations are disregarded as expected.

7 Conclusions

In this paper, we propose a probabilistic kernel method for intrinsic modeling of signals using differential geometry. It enables to empirically learn the underlying manifold of local probability densities of noisy measurements and provides a compact and efficient representation of the signals. We show that the obtained data-driven representation is intrinsic, i.e., invariant under different measurement modalities and noise resilient. In addition, our

modeling method is designed in a sequential manner, which is common in many signal processing tasks. The experimental results for two nonlinear and non-Gaussian filtering applications show that the obtained data-driven models are indeed independent of the observation modalities and instruments. Thereby suggesting that intrinsic filters that eliminate the need to adapt the configuration and to calibrate the measurement instruments may be built. This result will be harnessed in a future work to propose a data-driven Bayesian filtering framework for estimation and prediction of signals without providing a-priori probabilistic models. In addition, future work will also address real-world applications: modeling of EEG recordings for epileptic seizure identification and molecular dynamics simulations.

8 Appendix 1: Proof of Theorem 1

Proof. The Jacobian elements using the new features (16) are given by

$$J_{t_0}^{ji} = \frac{\partial l_{t_0}^j}{\partial \theta_{t_0}^i} = \lim_{\theta_{t_0,t}^i \to \theta_{t_0}^i} \frac{l_{t_0,t}^j - l_{t_0}^j}{\theta_{t_0,t}^i - \theta_{t_0}^i}. \tag{29}$$

By definition, (29) is explicitly expressed as

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

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

$$\begin{aligned} I_{t_0}^{ii'} &= \sum_{j=1}^{m} \sqrt{h_{t_0}^j} \frac{\partial}{\partial \theta_{t_0}^i} \log\left(h_{t_0}^j\right) \sqrt{h_{t_0}^j} \frac{\partial}{\partial \theta_{t_0}^{i'}} \log\left(h_{t_0}^j\right) \\ &= \sum_{j=1}^{m} \frac{\partial}{\partial \theta_{t_0}^i} \log\left(h_{t_0}^j\right) \frac{\partial}{\partial \theta_{t_0}^{i'}} \log\left(h_{t_0}^j\right) h_{t_0}^j. \end{aligned} \tag{30}$$

If we additionally assume that the histograms converge to the actual pdf of the measurements under the conditions that led to (7), we have

$$\begin{aligned} I_{t_0}^{ii'} &\simeq \int_{\mathbf{z}} \frac{\partial}{\partial \theta_{t_0}^i} \log\left(p(\mathbf{z}; \boldsymbol{\theta}_{t_0})\right) \frac{\partial}{\partial \theta_{t_0}^{i'}} \log\left(p(\mathbf{z}; \boldsymbol{\theta}_{t_0})\right) p(\mathbf{z}; \boldsymbol{\theta}_{t_0}) d\mathbf{z} \\ &= \mathbb{E}\left[\frac{\partial}{\partial \theta_{t_0}^i} \log\left(p(\mathbf{z}; \boldsymbol{\theta}_{t_0})\right) \frac{\partial}{\partial \theta_{t_0}^{i'}} \log\left(p(\mathbf{z}; \boldsymbol{\theta}_{t_0})\right)\right] \end{aligned} \tag{31}$$

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

9 Appendix 2: Proof of Theorem 2

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

$$W_\alpha^{t\tau} = \frac{\sum_{s'=1}^{N} \Pr(\mathbf{z}_t, \mathbf{z}_\tau | \mathbf{z}_t \in \mathcal{Z}_{s'}, \mathbf{z}_\tau \in \mathcal{Z}_{s'})}{\sum_{s''=1}^{N} \Pr(\mathbf{z}_t | \mathbf{z}_t \in \mathcal{Z}_{s''}) \sum_{s''=1}^{N} \Pr(\mathbf{z}_\tau | \mathbf{z}_\tau \in \mathcal{Z}_{s''})}. \tag{32}$$

Using the uniform distribution, we can rewrite (32) as (33).

$$W_\alpha^{t\tau} = \frac{\Pr(\mathbf{z}_t \in \mathcal{Z}_s, \mathbf{z}_\tau \in \mathcal{Z}_s) \sum_{s'=1}^{N} \Pr(\mathbf{z}_t \in \mathcal{Z}_{s'}) \Pr(\mathbf{z}_t, \mathbf{z}_\tau | \mathbf{z}_t \in \mathcal{Z}_{s'}, \mathbf{z}_\tau \in \mathcal{Z}_{s'})}{\sum_{s''=1}^{N} \Pr(\mathbf{z}_t \in \mathcal{Z}_{s''}) \Pr(\mathbf{z}_t | \mathbf{z}_t \in \mathcal{Z}_{s''}) \sum_{s''=1}^{N} \Pr(\mathbf{z}_\tau \in \mathcal{Z}_{s''}) \Pr(\mathbf{z}_\tau | \mathbf{z}_\tau \in \mathcal{Z}_{s''})}. \tag{33}$$

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

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

□

Acknowledgment

The authors would like to thank Prof. Amit Singer for helpful discussions and suggestions.

References

[1] I.G. Kevrekidis, C.W. Gear, and G. Hummer, "Equation-free: The computer-aided analysis of complex multiscale systems," *AIChE Journal*, vol. 50, no. 7, pp. 1346–1355, 2004.

[2] A. Rahimi, T. Darrell, and B. Recht, "Learning appearance manifolds from video," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 868–875, 2005.

[3] R.S. Lin, C.B. Liu, M.H. Yang, N. Ahuja, and S. Levinson, "Learning nonlinear manifolds from time series," *Computer Vision (ECCV)*, pp. 245–256, 2006.

[4] R. Li, T.-P. Tian, and S. Sclaroff, "Simultaneous learning of nonlinear manifold and dynamical models for high-dimensional time series," *IEEE 11th International Conference on Computer Vision (ICCV-2007)*, pp. 1–8, 2007.

[5] J.H. Macke, J.P. Cunningham, M.Y. Byron, K.V. Shenoy, and M. Sahani, "Empirical models of spiking in neural populations," *In J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, eds., Advances in Neural Information Processing Systems*, vol. 24, pp. 13501358, 2011.

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

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

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

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

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

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

[12] R.E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME J. Basic Eng.*, vol. 82, pp. 3445, 1960.

[13] Y. Bar-Shalom, *Tracking and data association*, Academic Press Professional, 1987.

[14] S.J. Julier and J.K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. of the IEEE*, vol. 92, pp. 401–422, 2004.

[15] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, , no. 10, pp. 197–208, 2000.

[16] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, pp. 174–188, 2003.

[17] O. Cappé, S.J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential Monte Carlo," *Proc. of the IEEE*, vol. 95, no. 5, pp. 899–924, May 2007.

[18] E. Niedermeyer and F.H.L. Da Silva, *Electroencephalography: basic principles, clinical applications, and related fields*, Lippincott Williams & Wilkins, 2005.

[19] S. Amari and H. Nagaoka, *Methods of information geometry*, American Mathematical Society, 2007.

[20] M. Girolami and B. Calderhead, "Riemann manifold langevin and hamiltonian monte carlo methods," *Journal of the Royal Statistical Society: Series B*, vol. 73, pp. 123–214, 2011.

[21] A. Haddad, D. Kushnir, and R. R. Coifman, "Filtering via a reference set," *Technical Report YALEU/DCS/TR-1441*, Feb. 2011.

[22] R. Talmon, I. Cohen, S. Gannot, and R.R. Coifman, "Supervised graph-based processing for sequential transient interference suppression," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 9, pp. 2528–2538, Nov. 2012.

[23] G. Storvik, "Particle filters for state-space models with the presence of unknown static parameters," *IEEE Trans. Signal Process.*, vol. 50, pp. 281–289, 2002.

[24] S.J. Godsill, J. Vermaak, W. Ng, and J.F. Li, "Models and algorithms for tracking of maneuvering objects using variable rate particle filters," *Proc. of the IEEE*, vol. 95, pp. 925–952, 2007.

[25] R. Talmon and R.R. Coifman, "Differential stochastic sensing: Intrinsic modeling of random time series with applications to nonlinear tracking," *to appear in Proc. Nat. Acad. Sci.*, 2012.

[26] D. Kushnir, A. Haddad, and R. Coifman, "Anisotropic diffusion on sub-manifolds with application to earth structure classification," *Appl. Comput. Harmon. Anal.*, vol. 32, no. 2, pp. 280–294, 2012.

[27] G. Roberts and O. Stramer, "Langevin diffusions and Metropolis-Hastings algorithms," *Methodology and Computing in Applied Probability*, vol. 4, pp. 337358, 2003.

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

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

[30] A. Bondy and U.S.R. Murty, *Graph Theory*, Springer, 2008.

[31] P.W. Jones, M. Maggioni, and R. Schul, "Manifold parametrizations by eigenfunctions of the laplacian and heat kernels," *Proc. Nat. Acad. Sci.*, vol. 105, no. 6, pp. 1803–1808, Feb. 2008.