KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Maja Taseska, Toon van Waterschoot, Emanuël A. P. Habets, Ronen Talmon (2019) **Nonlinear filtering with variable-bandwidth exponential kernels**
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	
Journal homepage	https://signalprocessingsociety.org/publications-resources/ieee-transactions-signal-processing
Author contact	your email maja.taseska@esat.kuleuven.be
IR	

(article begins on next page)

Nonlinear Filtering with Variable-Bandwidth Exponential Kernels

Maja Taseska, Toon van Waterschoot, *Member, IEEE,* Emanuël A. P. Habets *Senior Member, IEEE,* and Ronen Talmon, *Member, IEEE,*

***Abstract*—Frameworks for efficient and accurate signal processing often rely on a suitable representation of measurements that capture phenomena of interest. Typically, such representations are high-dimensional vectors obtained by a transformation of raw sensor signals such as time-frequency transform, lag-map, etc. In this work, we focus on representation learning approaches that consider the measurements as the nodes of a weighted graph, with edge weights computed by a given *kernel*. If the kernel is chosen properly, the eigenvectors of the resulting graph affinity matrix provide suitable representation coordinates for the measurements. Consequently, tasks such as regression, classification, and filtering, can be done more efficiently than in the original signal domain. In this paper, we address the problem of representation learning from measurements, which besides the phenomenon of interest contain undesired sources of variability. We propose data-driven kernels to learn representations that accurately parametrize the phenomenon of interest, while reducing variations due to other sources of variability. This is a non-linear filtering problem, which we approach under the assumption that certain geometric information about the undesired sources can be extracted from the measurements, e.g., using an auxiliary sensor. The applicability of the proposed kernels is demonstrated in toy problems and in a real signal processing task.**

***Index Terms*—manifold learning, non-linear filtering, metric learning, diffusion kernels**

I. Introduction

In many applications, high-dimensional measured data arise from physical systems with a small number of degrees of freedom. Consequently, the number of parameters required to fully describe the data is much smaller than the data dimensionality [1]. This insight justifies learning of low-dimensional representations of the data, before addressing tasks such as function approximation, clustering, signal prediction, etc. An important class of algorithms in this context, based on spectral graph theory [2], start by interpreting the high-dimensional measurements as nodes of a weighted graph, where the edge weights of the graph are computed by a suitably chosen kernel. Subsequently, the leading eigenvectors of the resulting graph affinity matrix provide coordinates that faithfully represent information about the underlying physical system [3], [4]. The spectral graph-theoretic view on representation learning is closely related to manifold learning in Riemannian geometry [2]. In the former, the measurements represent nodes of a graph, while in the latter, they represent samples from a low-dimensional Riemannian manifold, smoothly embedded in the high-dimensional measurement space. The graph can then be viewed as a discrete approximation of the manifold and the eigenvectors of the graph affinity matrix converge to the eigenfunctions of the Laplace-Beltrami Operator (LBO) on the manifold [5], [6], [7], [8].

The graph affinity matrix, if properly normalized, can be interpreted as the transition probability matrix of a Markov chain on the graph [2], [9], [10], which converges to a diffusion process on the corresponding manifold [10], [11], [12]. The Markov chain / diffusion perspective provides a theoretically sound framework for constructing application-dependent and data-driven kernels. In practice, the measurements are rarely clean observations of a phenomenon of interest, and often contain undesired sources of variability. Considering a Markov chain on the graph, it is intuitively clear that in order to obtain suitable representation by spectral analysis of the Markov chain, one needs to construct the transition probability matrix in such a way that the slowest relaxation processes capture the geometry of the phenomenon of interest [2], [13]. This is the underlying idea behind *directed diffusions* [11], *self-tuning kernels* [14], and other kernels with a data-driven distance metric [15] which are successfully applied to many applications in the past decade. These applications include multiscale analysis of dynamical systems [16], [17], [18], multimodal data analysis [19], [20], and non-linear independent component analysis [21].

In this paper, we address the problem of representation learning from measurements, which besides the phenomenon of interest (*signal*), contain undesired sources of variability (*noise*). We propose data-driven kernels, whose corresponding Markov chains (or diffusion processes) behave as if the data were sampled from a manifold whose geometry is mainly determined by the phenomenon of interest. In other words, our objective is to recover a noise-robust low-dimensional representation of the measurements, that recovers relevant

M. Taseska and T. van Waterschoot are with KU Leuven, Dept. of Electrical Engineering (ESAT-STADIUS/ETC), Leuven, Belgium (e-mail: maja.taseska@esat.kuleuven.be; toon.vanwaterschoot@esat.kuleuven.be). M. Taseska is a Postdoctoral Fellow of the Research Foundation Flanders (no. 12X6719N).

E. A. P. Habets is with the International Audio Laboratories Erlangen (a joint institution between the University of Erlangen-Nuremberg and Fraunhofer IIS), Erlangen 91058, Germany (e-mail:emanuel.habets@audiolabs-erlangen.de).

R. Talmon is with the Viterbi Faculty of Electrical Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel (e-mail: ronen@ee.technion.ac.il).

The research leading to these results has received funding from the Research Foundation Flanders (Grant 12X6719N), the Minerva Stiftung short-term research grant, the Israel Science Foundation (Grant 1490/16), KU Leuven Internal Funds C2-16-00449 and VES/19/004, and the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

geometric properties of the desired signal. To reach this objective, we require prior information in the form of a distance metric that is consistent with the noise component. Although the requirement of such information might seem restrictive, we propose a purely data-driven approach to estimate the required distance metric using an auxiliary sensor. In addition, we demonstrate that under certain conditions, the proposed kernels can be applied to enhance weak signals in single-sensor scenarios, without the need for an auxiliary sensor.

The paper is organized as follows. In Section II, we define the data model and formulate the problem. In Section III, we describe the relevant concepts from manifold learning. Section IV presents the main contribution of this paper, where we propose data-driven kernels for non-linear filtering. In Section V, we illustrate the properties of the proposed kernels with several toy experiments. The non-linear filtering capability of the kernels is demonstrated in Section VI in a real signal processing task. Section VII concludes the paper.

II. Problem formulation

A. Data model

Consider two hidden random variables X and V, whose codomains are the metric spaces $(\mathcal{X}, g_x)$ and $(\mathcal{V}, g_v)$, respectively. X and V are related to an observable variable S by an unknown deterministic function g as follows

$$S = g(X, V), \qquad g : \mathcal{X} \times \mathcal{V} \to \mathcal{S}. \tag{1}$$

A realization of S, denoted by $\boldsymbol{s}$, models a single measurement from a sensor that captures a variable of interest $\boldsymbol{x}$ (a realization of X) and a nuisance variable $\boldsymbol{v}$ (a realization of V). In practice, the measurements are often vectors in a high-dimensional Euclidean space $\mathcal{S} \subset \mathbb{R}^{l_s}$, where l_s is the dimensionality (e.g. time-frequency transform of a time series, lag-map, pixels of an image, etc). The function g comprises the sensor mechanism, and possibly, application-specific preprocessing transforms. In the following, we refer to $\boldsymbol{x}$ and $\boldsymbol{v}$, as *signal* and *noise*, respectively.

In modern applications, data is often captured by multiple sensors. Of interest in this work are auxiliary sensors that can serve as a noise reference. We model the measurements from such a sensor by a random variable $S^{(a)}$

$$S^{(a)} = g^{(a)}(V, Z), \qquad g^{(a)} : \mathcal{V} \times \mathcal{Z} \to \mathcal{S}^{(a)}, \tag{2}$$

where Z is a nuisance variable. Note that in contrast to the classical data model in signal processing literature, the second sensor does not provide a clean reference of V: it contains an additional nuisance variable and an unknown measurement function $g^{(a)}$, which may be different from g.

We assume that g embeds the product space $\mathcal{X} \times \mathcal{V}$ into $\mathbb{R}^{l_s}$ in an approximately isometric fashion. Namely, if d_s denotes the Euclidean distance on $\mathbb{R}^{l_s}$, and d_{xv} is a distance on $\mathcal{X} \times \mathcal{V}$, then for any $(\boldsymbol{x}_1, \boldsymbol{v}_1)$ and $(\boldsymbol{x}_2, \boldsymbol{v}_2)$

$$d_s(\boldsymbol{s}_1, \boldsymbol{s}_2) \approx d_{xv}\left((\boldsymbol{x}_1, \boldsymbol{v}_1), (\boldsymbol{x}_2, \boldsymbol{v}_2)\right). \tag{3}$$

A distance on the product $\mathcal{X} \times \mathcal{V}$ can be defined as [22, Ch 1]

$$d_{xv}\left((\boldsymbol{x}_1, \boldsymbol{v}_1), (\boldsymbol{x}_2, \boldsymbol{v}_2)\right) = \left(d_x(\boldsymbol{x}_1, \boldsymbol{x}_2)^p + d_v(\boldsymbol{v}_1, \boldsymbol{v}_2)^p\right)^{\frac{1}{p}}, \tag{4}$$

for any $1 \leq p < \infty$, where d_x and d_v are distance functions on $\mathcal{X}$ and $\mathcal{V}$, respectively, induced by the corresponding metrics g_x and g_v. The data model of the auxiliary sensor can be endowed with an analogous distance structure.

For the purpose of our analysis, we assume that the metric spaces $\mathcal{X}$ and $\mathcal{V}$ are smooth Riemannian manifolds. In this case, the product $\mathcal{X} \times \mathcal{V}$ is also a smooth manifold [23, Ch. 1].

B. Problem statement

In the considered two-sensor model, a single realization of the latent variable triplet $(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{z})$ is associated to a pair of measurements $(\boldsymbol{s}, \boldsymbol{s}^{(a)})$. Then, given N measurement pairs $(\boldsymbol{s}_1, \boldsymbol{s}_1^{(a)}), \ldots (\boldsymbol{s}_N, \boldsymbol{s}_N^{(a)})$, we wish to recover the latent signals of interest $\{\boldsymbol{x}_i\}_{i=1}^N$ in the primary sensor.

In our non-parametric and unsupervised setting, classical estimation of $\{\boldsymbol{x}_i\}_{i=1}^N$ from the noisy measurements is an unfeasible task. Instead, we seek to recover a parametrization of $\{\boldsymbol{x}_i\}_{i=1}^N$ by a low-dimensional embedding f

$$f : \mathcal{S} \to \mathcal{E}, \qquad \mathcal{E} \subseteq \mathbb{R}^{l_x}, \ \text{ where } \ l_x << l_s, \tag{5}$$

that approximately preserves the local distance relationships among $\{\boldsymbol{x}_i\}_{i=1}^N$, as defined by the distance d_x on $\mathcal{X}$. Under certain circumstances, it has been shown that such embeddings suffice to approximately reconstruct the latent points $\{\boldsymbol{x}_i\}_{i=1}^N$ [24]. We note that construction of manifold embeddings with a small local bi-Lipschitz distortion has been discussed in [8], when the measurements are sampled from a manifold of interest $\mathcal{X}$ without the presence of noise. In our work, we seek to obtain such embeddings when the measurements contain an unknown noise component.

III. Diffusion kernels for manifold learning: a brief overview

Manifold learning approaches are often used for signal processing by modeling the measurements (signal samples) $\{\boldsymbol{s}_i\}_{i=1}^N \in \mathcal{S}$ as points on or near a low-dimensional manifold $\mathcal{X}$, embedded in the ambient space $\mathcal{S}$ [25]. To learn a meaningful low-dimensional representation, the samples $\{\boldsymbol{s}_i\}_{i=1}^N$ are interpreted as the nodes of a graph, where a kernel function $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ assigns the edge weights (pairwise similarities). The graph represents a discrete approximation of the manifold $\mathcal{X}$ [6], [26], [27]. This setting is simpler than the signal model we introduced in Section II, where the measurements are samples from a product manifold $\mathcal{X} \times \mathcal{V}$ that contains a noise component. Nevertheless, as kernel-based manifold learning lays the theoretical basis for our work, we briefly discuss the main concepts in this section.

A. Diffusion distance and diffusion maps

Consider a positive semi-definite kernel function k, and let $\boldsymbol{K}$ denote the $N \times N$ kernel matrix with entries $\boldsymbol{K}[i, j] = k(\boldsymbol{s}_i, \boldsymbol{s}_j)$. A common choice for k is an exponentially decaying homogeneous and isotropic Gaussian kernel, given by

$$k_\varepsilon(\boldsymbol{s}_i, \boldsymbol{s}_j) = \exp\left(-\frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2^2}{\varepsilon}\right), \tag{6}$$

where $\varepsilon > 0$ is the kernel bandwidth. Let a diagonal matrix $\boldsymbol{D}$ contain the degree of each graph node, i.e.,

$$\boldsymbol{D}[i,i] = \sum_{j=1}^{M} k(\boldsymbol{s}_i, \boldsymbol{s}_j) = \sum_{j=1}^{M} \boldsymbol{K}[i,j]. \tag{7}$$

A Markov chain on the graph can be constructed by considering the following normalized kernel matrix, referred to as a *diffusion kernel*,

$$\boldsymbol{P} = \boldsymbol{D}^{-1}\,\boldsymbol{K}, \tag{8}$$

where $\boldsymbol{P}$ represents the transition probability matrix of the Markov chain [2], [9]. The probability of the Markov chain that started at $\boldsymbol{s}_i$, to be at $\boldsymbol{s}_j$ at step t is given by

$$p_t\left(\boldsymbol{s}_j \mid \boldsymbol{s}_i\right) = \boldsymbol{P}^t[j,i]. \tag{9}$$

The Markov chain on the graph leads to a natural definition of distance between points based on their connectivity, known as the *diffusion distance* [10], [11]. If the graph is connected and non-bipartite, the Markov chain has a unique stationary distribution given by [2, Ch.1]

$$\pi_o(\boldsymbol{s}_i) = \frac{\boldsymbol{D}[i,i]}{\sum_j \boldsymbol{D}[j,j]}. \tag{10}$$

The diffusion distance at step t is then defined as

$$d_t^2(\boldsymbol{s}_i, \boldsymbol{s}_j) = \sum_{l=1}^{N} \frac{\left(\boldsymbol{P}^t[l,i] - \boldsymbol{P}^t[l,j]\right)^2}{\pi_o(\boldsymbol{s}_l)}. \tag{11}$$

An embedding that is consistent with d_t can be constructed from the eigenvectors of $\boldsymbol{P}^t$ [10]. Let $\{\boldsymbol{\psi}_i\}_{i=0}^{M-1}$ denote the right eigenvectors of $\boldsymbol{P}$, with eigenvalues $1 = \lambda_0 > \lambda_1 \geq \ldots > 0$. Then, an l-dimensional diffusion maps embedding $\Psi_t : \mathcal{S} \to \mathbb{R}^l$, for given t and l is defined as

$$\Psi_t(\boldsymbol{s}_i) = \left[\, \lambda_1^t\, \boldsymbol{\psi}_1[i],\ \lambda_2^t\, \boldsymbol{\psi}_2[i],\ \ldots,\ \lambda_l^t\, \boldsymbol{\psi}_l[i] \right]^{\mathrm{T}}. \tag{12}$$

The constant eigenvector $\boldsymbol{\psi}_0$ is excluded from the embedding. Due to the intrinsic low-dimensionality of the manifold, an l-dimensional diffusion map with $l << l_s$, embeds the data approximately isometrically with respect to d_t [11], [28]. The dimensionality l is chosen by identifying the spectral gap, i.e., the number of significant eigenvalues of $\boldsymbol{P}^t$.

The eigenvectors of the isotropic diffusion kernel constructed by (6)-(8) are consistent with the manifold geometry only if the measurements are sampled uniformly on the manifold. In this case, the eigenvectors converge to the eigenfunctions of the LBO [5]. To maintain this property for an arbitrary sampling density, an additional normalization of the kernel $\boldsymbol{K}$ is required as follows [11], [28]

$$\boldsymbol{K}_o = \boldsymbol{D}^{-1}\,\boldsymbol{K}\,\boldsymbol{D}^{-1}, \tag{13a}$$

$$\boldsymbol{P} = \boldsymbol{D}_o^{-1}\,\boldsymbol{K}_o, \tag{13b}$$

where $\boldsymbol{D}_o$ is a diagonal matrix with $\boldsymbol{D}_o[i,i] = \sum_{j=1}^{N} \boldsymbol{K}_o[i,j]$.

B. Directed diffusion and data-driven kernels

The theory of diffusion maps with isotropic exponential kernels such as (6), and their ability to recover the manifold geometry, is valid when the measurements are sampled from the manifold of interest. In most applications, including the non-linear filtering problem considered in our work, this is not the case. Hence, learning a suitable representation of a quantity of interest in the measurements requires design of data-driven diffusion kernels. This can be achieved by employing a data-driven distance function in the kernels.

In the literature, several approaches to metric learning have been proposed for this task. A class of approaches replace the Euclidean distance in the kernel with a quadratic forms defined by a task-driven metric tensor $M(i,j)$ as follows

$$k_{\varepsilon,M}(\boldsymbol{s}_i, \boldsymbol{s}_j) = \exp\left(-\frac{\left(\boldsymbol{s}_i - \boldsymbol{s}_j\right)^{\mathrm{T}} M(i,j)\left(\boldsymbol{s}_i - \boldsymbol{s}_j\right)}{\varepsilon}\right). \tag{14}$$

Such kernel construction has been often used in the past decade for analysis of dynamical systems [16], image processing [29], non-linear independent component analysis [21], and other applications [25]. Quadratic form distances have also been used with alternating diffusion kernels in multi-sensor applications [20]. Other approaches for informed metric construction based on prior information about the problem at hand have been proposed in [30], [31], [32].

IV. Proposed noise-informed diffusion kernels for nonlinear filtering

According to the diffusion maps theory discussed in Section III, if the data lie on a manifold, the diffusion distance associated with a suitable Markov chain is consistent with the manifold geometry. However, in our problem, the data is sampled from the product manifold $\mathcal{X} \times \mathcal{V}$, while the objective is to recover the geometry of $\mathcal{X}$ alone. Two problems arise if we apply the diffusion maps algorithm with a standard Gaussian kernel to learn parametrization of X. First, we cannot identify whether a given diffusion maps coordinate corresponds to X, V, or a combination thereof. Second, even if we could identify the relevant coordinates, they might not correspond to leading eigenvectors of the kernel.[1] The second problem is relevant for implementation of manifold learning algorithms in practice, as efficient large-scale eigensolvers compute the eigenvectors of matrices consecutively, starting from the largest ones [35]. Our objective is to design suitable diffusion kernels which warp the data geometry in a way that information about the signal of interest concentrates higher in the spectrum (i.e., in eigenvectors that correspond to larger eigenvalues), compared to a standard diffusion kernel on $\mathcal{X} \times \mathcal{V}$.

A. Kernel construction with noise-informed bandwidth

The type of data-driven kernels that we consider for non-linear filtering are known as variable-bandwidth (VB) ker-

[1]The second problem is related to the crucial property of the LBO eigenvectors, namely, that different eigenvectors may encode the same source of variability on the manifold. See [33], [34] for more details about this property and its implications in practice.

nels [15], where the bandwidth is prescribed by a location-dependent scalar function $b(i,j)$ as follows

$$k_{\varepsilon,b}(\boldsymbol{s}_i,\boldsymbol{s}_j)=\exp\left(-\frac{\|\boldsymbol{s}_i-\boldsymbol{s}_j\|^2}{\varepsilon\, b(i,j)}\right). \tag{15}$$

VB kernels have been used for robust spectral clustering [14] and dynamical system modeling [17], [18]. Here, we show that with suitably defined bandwidth, they can be applied for non-linear filtering on product manifolds.

We start by noting that a bandwidth $b(i,j)$ defines a transformation of the Euclidean distances in $\mathcal{S}$, according to

$$d_s(\boldsymbol{s}_i,\boldsymbol{s}_j)=\|\boldsymbol{s}_i-\boldsymbol{s}_j\| \;\longmapsto\; \frac{\|\boldsymbol{s}_i-\boldsymbol{s}_j\|}{b(i,j)}=\hat{d}_s(\boldsymbol{s}_i,\boldsymbol{s}_j). \tag{16}$$

If a kernel implemented with the distance $\hat{d}_s$ is to have more of its leading eigenvectors consistent with the geometry of $\mathcal{X}$ compared to a kernel implemented with d_s, the transformed distance $\hat{d}_s$ should be less sensitive to noise than the observable Euclidean distance d_s. To achieve such behavior, we propose the following noise-informed bandwidth function

$$b(i,j)=\left(1+d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)\right)^2. \tag{17}$$

Clearly, the pairwise distances $d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)$ are unobservable in practice. In Section IV-B, we discuss data-driven methods to estimate $d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)$ for each pair of observations.

With the non-linear filtering problem in mind, the bandwidth function in (17) was chosen such that distances in the kernel-induced geometry 1) are less sensitive to noise than the Euclidean distances in the measurement space, 2) are robust to estimation errors in d_v, and 3) preserve the geometry of the desired signal to a certain extent. We formalize these properties in the following three propositions.

Proposition 1. *If $(\boldsymbol{x}_i,\boldsymbol{v}_i)$ and $(\boldsymbol{x}_j,\boldsymbol{v}_j)$ are the hidden variables corresponding to $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, respectively, then*

$$d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)>0 \implies \hat{d}_s(\boldsymbol{s}_i,\boldsymbol{s}_j)<d_s(\boldsymbol{s}_i,\boldsymbol{s}_j) \tag{18a}$$

$$d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)=0 \implies \hat{d}_s(\boldsymbol{s}_i,\boldsymbol{s}_j)=d_x(\boldsymbol{x}_i,\boldsymbol{x}_j). \tag{18b}$$

Proof. The bandwidth function induces a locally scaled Euclidean distance between the measurements, given by

$$\hat{d}_s(\boldsymbol{s}_i,\boldsymbol{s}_j)=d_s(\boldsymbol{s}_i,\boldsymbol{s}_j)\left(1+d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)\right)^{-1}. \tag{19}$$

It is straightforward that the scaling $\left(1+d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)\right)^{-1}$ depends on d_v as follows

$$d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)>0 \implies \left(1+d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)\right)^{-1}<1 \tag{20a}$$

$$d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)=0 \implies \left(1+d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)\right)^{-1}=1. \tag{20b}$$

Furthermore, from the distance properties in (3), (4) we have

$$d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)=0 \implies d_s(\boldsymbol{s}_i,\boldsymbol{s}_j)=d_x(\boldsymbol{x}_i,\boldsymbol{x}_j). \tag{21}$$

The proof follows by substituting (19), (20), and (21) in (18). □

From (18), it follows that if the noise contributes to the measured distance $d_s(\boldsymbol{s}_i,\boldsymbol{s}_j)$, then the distance in the kernel-induced geometry is smaller than $d_s(\boldsymbol{s}_i,\boldsymbol{s}_j)$. In this sense, the proposed noise-informed bandwidth results in a distance measure that is less sensitive to noise, compared to $d_s(\boldsymbol{s}_i,\boldsymbol{s}_j)$.

As d_v has to be estimated from the data, the bandwidth function needs to be stable under small estimation errors of d_v. Let $\hat{d}_v(\boldsymbol{v}_i,\boldsymbol{v}_j)$ denote the estimate and $\hat{d}'_s(\boldsymbol{s}_i,\boldsymbol{s}_j)$ the resulting scaled Euclidean distance.

Proposition 2. *If $|d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)-\hat{d}_v(\boldsymbol{v}_i,\boldsymbol{v}_j)|<\epsilon_v$, then $|\hat{d}_s(\boldsymbol{s}_i,\boldsymbol{s}_j)-\hat{d}'_s(\boldsymbol{s}_i,\boldsymbol{s}_j)|\le\epsilon\, d_s(\boldsymbol{s}_i,\boldsymbol{s}_j)$*

Proof. To describe the behavior of the scaling factor $\left(1+d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)\right)^{-1}$, consider the function $f(u)=(1+u)^{-1}$. The following holds

$$|f(u)-f(w)|\le|u-w|. \tag{22}$$

Omitting the distance function arguments for brevity, we have

$$|\hat{d}_s-\hat{d}'_s|=d_s\left|\frac{1}{1+d_v}-\frac{1}{1+\hat{d}_v}\right|\le d_s\,(d_v-\hat{d}_v), \tag{23}$$

where the inequality follows from (22). Thus, we conclude $|\hat{d}_s(\boldsymbol{s}_i,\boldsymbol{s}_j)-\hat{d}'_s(\boldsymbol{s}_i,\boldsymbol{s}_j)|\le\epsilon\, d_s(\boldsymbol{s}_i,\boldsymbol{s}_j)$. □

Proposition 3. *Consider the set of ordered pairs $\mathcal{L}_\xi=\{(i,j)\mid d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)=\xi\}$, for some constant $\xi>0$. $\mathcal{L}_\xi$ represents a set of measurement pairs for which the pairwise distance due to noise is constant. Let $(i,j),(k,l)\in\mathcal{L}_\xi$. Then $d_x(\boldsymbol{x}_i,\boldsymbol{x}_j)<d_x(\boldsymbol{x}_k,\boldsymbol{x}_l) \implies \hat{d}_s(\boldsymbol{s}_i,\boldsymbol{s}_j)<\hat{d}_s(\boldsymbol{s}_k,\boldsymbol{s}_l)$.*

Proof. From the definition of the distance, and the scaling function, it follows

$$\begin{aligned}\hat{d}_s(\boldsymbol{s}_i,\boldsymbol{s}_j)&=\left(d_x(\boldsymbol{x}_i,\boldsymbol{x}_j)^p+\xi^p\right)^{\frac{1}{p}}\,(1+\xi)^{-1}\\ \hat{d}_s(\boldsymbol{s}_k,\boldsymbol{s}_l)&=\left(d_x(\boldsymbol{x}_k,\boldsymbol{x}_l)^p+\xi^p\right)^{\frac{1}{p}}\,(1+\xi)^{-1}\end{aligned} \tag{24}$$

It is immediate that for a fixed ξ, $d_x(\boldsymbol{x}_i,\boldsymbol{x}_j)<d_x(x_k,x_l)$ implies $\hat{d}_s(\boldsymbol{s}_i,\boldsymbol{s}_j)<\hat{d}_s(\boldsymbol{s}_k,\boldsymbol{s}_l)$. □

Finally, note that the proposed bandwidth in (17) is not the only function that satisfies these propositions. In fact, any smooth monotonic transformation of d_v that is locally bi-Lipschitz has the potential to provide good non-linear filtering capabilities in the resulting kernels.

B. Estimating the noise distance metric d_v

To implement the proposed bandwidth function in (17), the pairwise distances $d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)$ need to be estimated from the measurements. Although scenarios with an auxiliary sensor are our main target, we also discuss a special case where estimation is possible with a single sensor.

*1) **Estimating d_v with an auxiliary sensor**:* The recently proposed alternating diffusion (AD) algorithm extends the diffusion framework to multiple sensors that capture a common signal, corrupted by sensor-specific variables [36], [37]. In our problem, the noise is a common signal at the primary and the auxiliary sensor. Hence, the AD algorithm can be used to find an embedding that is consistent with the geometry of V, and provide an estimate of the pairwise distances $d_v(\boldsymbol{v}_i,\boldsymbol{v}_j)$. The key object of AD is the AD kernel $\boldsymbol{P}_{\text{ad}}$ [36], defined as

$$\boldsymbol{P}_{\text{ad}}=\boldsymbol{P}\,\boldsymbol{P}^{(a)}. \tag{25}$$

where $\boldsymbol{P}$ and $\boldsymbol{P}^{(a)}$ are the standard sensor-specific diffusion kernels discussed Section III-A.

Let $\boldsymbol{P}_{\rm ad} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^{\rm T}$, where the columns $\{\boldsymbol{v}_i\}_{i=1}^N$ of $\boldsymbol{V}$, are the right singular vectors, and the entries $\{\sigma_i\}_{i=1}^N$ of the diagonal matrix $\boldsymbol{\Lambda}$, are the singular values (in decreasing order). Then, an l-dimensional AD embedding $\Psi_{\rm ad} : \mathcal{S} \times \mathcal{S}^a \to \mathbb{R}^l$ is given by [37]

$$\Psi_{\rm ad}(\boldsymbol{s}_i, \boldsymbol{s}_i^{(a)}) = \left[\, \sigma_1 \boldsymbol{v}_1[i],\, \sigma_2 \boldsymbol{v}_2[i],\, \ldots, , \sigma_d \boldsymbol{v}_l[i] \,\right]^{\rm T}. \quad (26)$$

The AD distance $d_{\rm ad}((\boldsymbol{s}_i, \boldsymbol{s}_i^{(a)}), (\boldsymbol{s}_j, \boldsymbol{s}_j^{(a)}))$, denoted by $d_{\rm ad}(i,j)$ for brevity, is defined as

$$d_{\rm ad}(i,j) = \|\Psi_{\rm ad}(\boldsymbol{s}_i, \boldsymbol{s}_i^{(a)}) - \Psi_{\rm ad}(\boldsymbol{s}_j, \boldsymbol{s}_j^{(a)})\|_2. \quad (27)$$

According to [36], $\Psi_{\rm ad}$ approximates a diffusion maps embedding that would be obtained if data was sampled directly from $\mathcal{V}$. As a result, $\Psi_{\rm ad}$ provides a parametrization of the noise samples $\{\boldsymbol{v}_i\}_{i=1}^N$, and $d_{\rm ad}(i,j)$ can be used to approximate the pairwise distances $d_v(\boldsymbol{v}_i, \boldsymbol{v}_j)$.

Using the AD distance, we implement the following distance transform for our proposed kernel

$$\begin{aligned}\hat{d}_s(\boldsymbol{s}_i, \boldsymbol{s}_j) &= \frac{d_s(\boldsymbol{s}_i, \boldsymbol{s}_j)}{1 + d_{\rm ad}(i,j)} \\ &= \frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2}{1 + \|\Psi_{\rm ad}(\boldsymbol{s}_i, \boldsymbol{s}_i^{(a)}) - \Psi_{\rm ad}(\boldsymbol{s}_j, \boldsymbol{s}_j^{(a)})\|_2}, \end{aligned} \quad (28)$$

which corresponds to a kernel with the bandwidth function

$$b(i,j) = (1 + d_{\rm ad}(i,j))^2. \quad (29)$$

We note that the dimensionality l of $\Psi_{\rm ad}$ is not very critical. In theory, all coordinates obtained by the AD algorithm are consistent with the geometry of $\mathcal{V}$. However, our experiments suggested that due to estimation errors in practice, it is preferable to only use the first one or two coordinates in (26).

*2) **Estimating** d_v **without an auxiliary sensor**:* If only the measurements $\{\boldsymbol{s}_i\}_{i=1}^N$ from the primary sensor are given, we claim that pairwise distance estimates $\hat{d}_v(\boldsymbol{v}_i, \boldsymbol{v}_j)$ can be obtained, if the signal-to-noise ratio (SNR) of the measurements is lower than 0 dB. Recall the structure of the diffusion spectrum: the strongest sources of variability correspond to the slowest relaxation processes of the Markov chain, which in turn, correspond to the largest eigenvalues of the kernel [2]. Hence, if we consider the one-dimensional diffusion map obtained with a standard kernel as described in Section III-A,

$$\Psi_1(\boldsymbol{s}_i) = \lambda_1 \boldsymbol{\psi}_1[i], \quad (30)$$

it follows that the Euclidean distance $|\Psi_1(\boldsymbol{s}_i) - \Psi_1(\boldsymbol{s}_j)|$ is consistent with $d_v(\boldsymbol{v}_i, \boldsymbol{v}_j)$. Consequently, we propose to implement the following metric transform for our kernel

$$\hat{d}_s(\boldsymbol{s}_i, \boldsymbol{s}_j) = \frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2}{1 + |\Psi_1(\boldsymbol{s}_i) - \Psi_1(\boldsymbol{s}_j)|}, \quad (31)$$

which corresponds to a kernel with the bandwidth function

$$b(i,j) = \left(1 + |\Psi_1(\boldsymbol{s}_i) - \Psi_1(\boldsymbol{s}_j)|\right)^2. \quad (32)$$

We note that the idea of using the first eigenvector of the diffusion kernel to uncover other sources of variability, has been previously used for dimensionality reduction [34] and nonlinear dynamical system analysis [33].

Algorithm 1 Diffusion maps with a noise-informed VB kernel

Input: Measurements $\{\boldsymbol{s}_i\}_{i=1}^N$, and estimated pairwise distances $\hat{d}_v(\boldsymbol{v}_i, \boldsymbol{v}_j)$ (described in Section IV-B).

1: For each pair (i,j) compute the bandwidth function $b(i,j)$ in (17), using $\hat{d}_v(\boldsymbol{v}_i, \boldsymbol{v}_j)$
2: Construct an exponential kernel matrix $\boldsymbol{K}$ with the VB kernel $\boldsymbol{K}[i,j] = \exp\left(-\frac{d(\boldsymbol{s}_i, \boldsymbol{s}_j)^2}{\varepsilon_{ij}\, b(i,j)}\right)$
3: Apply density normalization to $\boldsymbol{K}$, according to (13a)
4: Compute the diffusion kernel $\boldsymbol{P}$, according to (13b)
5: Compute the principal l_x eigenvectors $\{\boldsymbol{\psi}_i\}_{i=1}^{l_x}$ with eigenvalues $\{\lambda\}_{i=1}^{l_x}$ (exclude $\boldsymbol{\psi}_0$).

Output: The new representation $f(\boldsymbol{s}_i)$ for each $\boldsymbol{s}_i$
▷ $f(\boldsymbol{s}_i) = \left[\, \lambda_1 \boldsymbol{\psi}_1[i],\ \lambda_2 \boldsymbol{\psi}_2[i],\ \ldots,\ \lambda_{l_x} \boldsymbol{\psi}_{l_x}[i] \right]^{\rm T}$.

C. Summary and practical considerations

In real datasets, distances from nearest neighbors may differ significantly for different points. As a result, if ε is fixed, some vertexes of the graph can be isolated, while others highly connected. To take this into account, the scale ε can be location-dependent as well. We used the method suggested in [36], where for each point i, a local scale ε_i is introduced that is equal to the median of the squared distances from the 200 nearest neighbors. Then, the scale for a pair of points (i,j) is set to $\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$. The complete algorithm that implements diffusion maps with the proposed data-driven kernels, is summarized in Algorithm 1.

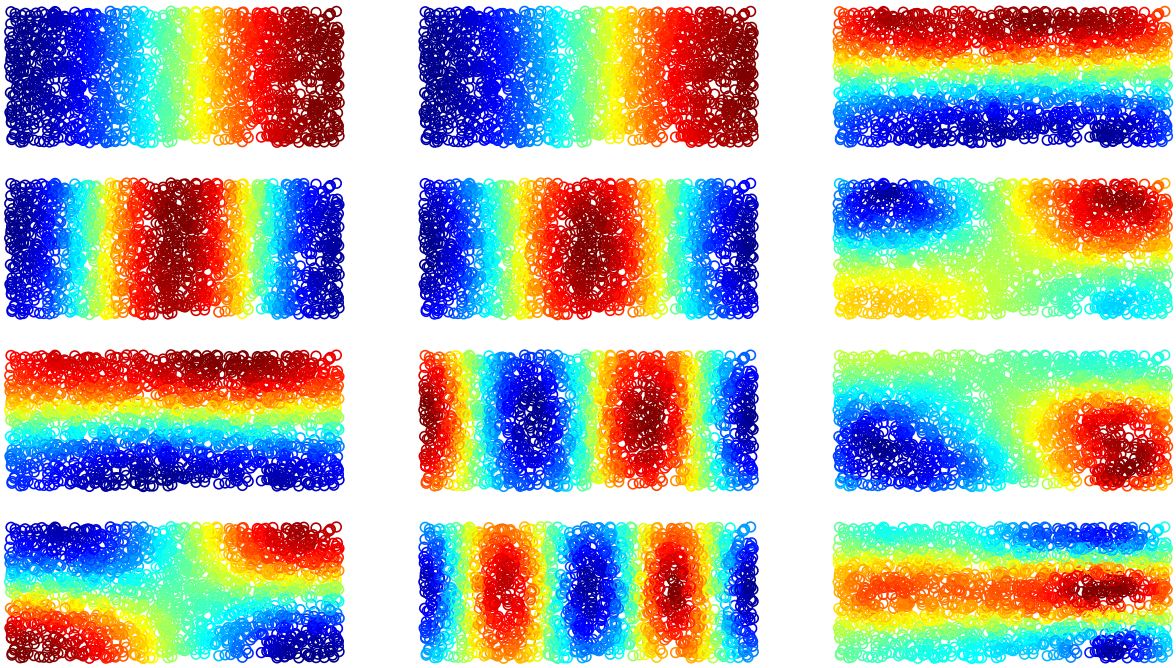
Note that in contrast to the standard diffusion maps algorithm, the spectral gap is not suitable to determine the dimensionality l_x of the embedding f. Although the eigenvectors that parametrize the desired signal are higher in the spectrum of the proposed kernel, compared to an isotropic kernel, there is no guarantee that all eigenvectors before the spectral gap parametrize the desired signal. The relevant eigenvectors, and hence, the dimensionality l_x, could be identified for instance by calculating the mutual information between the leading eigenvector of the AD kernel (which provides a noise reference signal), and the eigenvectors of the proposed kernel.

V. Illustrative examples

With the toy examples in this section, we investigate the effect of the noise-informed bandwidth function on the diffusion kernel eigenvectors. The goal is to illustrate that the resulting Markov chain is biased to propagate faster along the directions of variation that correspond to the noise variable (compared to an isotropic Markov chain). As a result, the leading eigenvector of the diffusion kernel provides a representation consistent with the desired variable X, regardless of the SNR.

A. Two-dimensional strip

Let the measurements $\{\boldsymbol{s}_i = [s_{i1}, s_{i2}]\}_{i=1}^N$ be samples from a two-dimensional rectangular strip, with lengths $L_1 > L_2$.



(a) (b) (c)

Fig. 1. Points sampled from a strip. The first four diffusion eigenvectors are shown coded in color (top to bottom: 1st to 4th). **(a)** Isotropic kernel. **(b)** Proposed kernel, when the horizontal coordinate is the desired signal X and the vertical coordinate is the noise V. **(c)** Proposed kernel, when the vertical coordinate is the desired signal X and the horizontal coordinate is the noise V.

The eigenvalues of the Laplace-Beltrami operator (with Neumann boundary conditions) can be computed analytically as

$$\mu_{k_1,k_2} = \left(\frac{k_1\pi}{L_1}\right)^2 + \left(\frac{k_2\pi}{L_2}\right)^2, \tag{33}$$

for $k_1, k_2 = 0, 1, 2\ldots$, with the corresponding eigenfunctions

$$\rho_{k_1,k_2}(l_1, l_2) = \cos\left(\frac{k_1 l_1 \pi}{L_1}\right) \cos\left(\frac{k_2 l_2 \pi}{L_2}\right). \tag{34}$$

Although the eigenfunctions $\rho_{1,0}(l_1) = \cos(l_1\pi/L_1)$ and $\rho_{0,1}(l_2) = \cos(l_2\pi/L_2)$ fully parametrize the strip, they do not necessarily correspond to the two largest eigenvalues. As the ratio L_1/L_2 increases, the more eigenfunctions of the form $\rho_{k_1,0}(l_1)$ appear before $\rho_{0,1}(l_2)$ in the spectrum. We uniformly sampled $N = 2880$ points from a strip with lengths $L_1 = L$ and $L_2 = 0.4\,L$. From (33), and (34), it follows that the leading four eigenfunctions are $\rho_{0,1}, \rho_{0,2}, \rho_{1,0}$, and $\rho_{1,1}$. The first four coordinates obtained by a standard diffusion maps algorithm with an isotropic kernel, illustrated in Figure 1(a), correspond to these four eigenfunctions.

If the two strip coordinates represent X and V, the properties we desire for a data-driven kernel are i) the leading eigenvector should parametrize the desired signal X, even if X corresponds to the shorter strip dimension, and / or ii) the number of eigenvectors among the leading ones that parametrize X, is larger than the same number for the standard kernel. Consider the coordinate-wise distances on the strip

$$d_1(s_{i1}, s_{j1}) = |s_{i1} - s_{j1}| \tag{35a}$$
$$d_2(s_{i2}, s_{j2}) = |s_{i2} - s_{j2}|. \tag{35b}$$

If s_{i1} is the coordinate of interest, then d_2 corresponds d_v, and if s_{i2} is the coordinate of interest, then d_1 corresponds to d_v. The associated bandwidth functions are

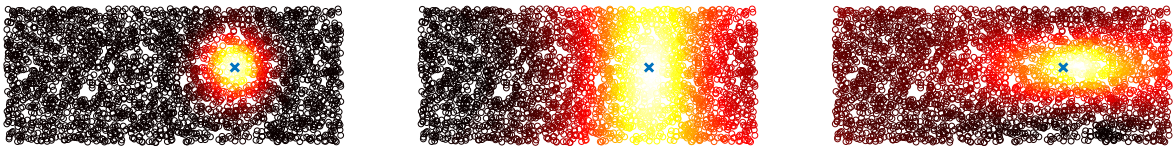
$$b_1(i,j) = (1 + d_1(s_{i1}, s_{j1}))^2, \tag{36a}$$
$$b_2(i,j) = (1 + d_2(s_{i2}, s_{j2}))^2. \tag{36b}$$

In these examples, we wish to demonstrate the behavior of proposed kernels with an ideal estimate of d_v. Therefore, we assume that d_1 and d_2 are accessible. The first four diffusion map coordinates obtained with the bandwidths in (36) are shown in Figure 1(b) and 1(c). In Figure 1(b), bandwidth is $b_2(i,j)$ shrinks vertical variations. As a result all four eigenvectors are consistent with the horizontal coordinate. Similarly, in Figure 1(c), the bandwidth $b_1(i,j)$ shrinks horizontal variations, and the principal eigenvector parametrizes the vertical coordinate.

We can visualize the evolution of the Markov chains as follows. We start from an arbitrary point on the strip by defining a unit probability vector centered at that point. Propagating the chain forward corresponds to multiplying the probability vector from the right by the transition probability matrix. The probability evolution (the *heat diffusion*), can be visualized by a scatter plot of all measurements, with each point colored by the probability of the Markov chain to be at that point, at a given step. While the standard kernel is characterized by an isotropic diffusion, the proposed kernels induce diffusion that is faster along the undesired coordinate, as seen in Figure 2.

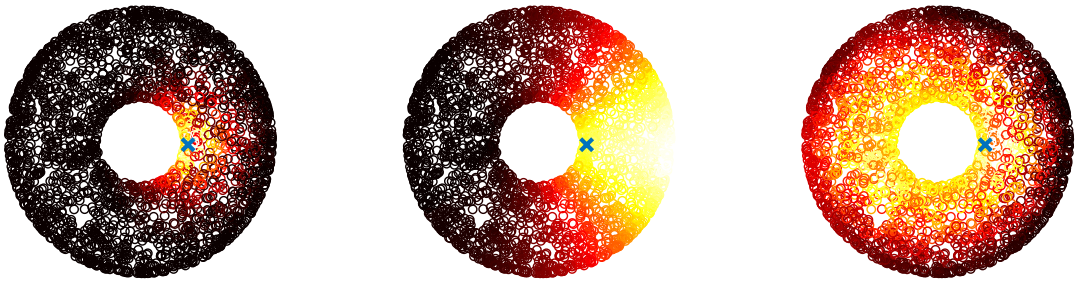
B. Manifolds embedded in $\mathbb{R}^3$

In this example, the measurements $\{\boldsymbol{s}_i\}_{i=1}^N$ are $N = 2500$ points sampled from the surface of a torus embedded in $\mathbb{R}^3$.



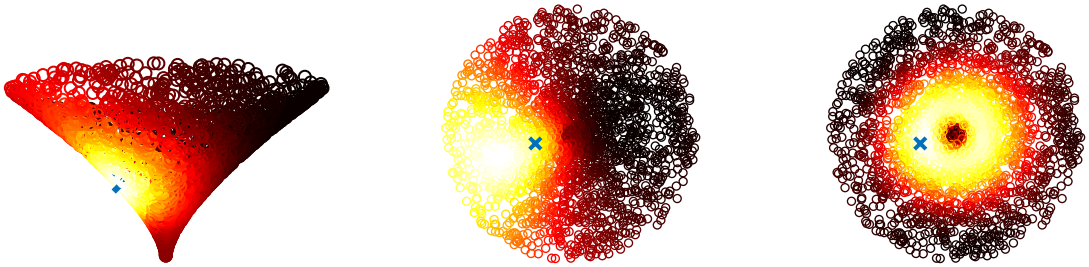
(a) (b) (c)

Fig. 2. Heat diffusion on the strip after 5 steps of the Markov chain, starting from the point denoted by $\times$. **(a)** Isotropic kernel; **(b)** Proposed kernel when the horizontal coordinate is the desired signal. The diffusion is then faster along the vertical coordinate; **(c)** Proposed kernel when the vertical coordinate is the desired signal. The diffusion is then faster along the horizontal coordinate.



(a) (b) (c)

Fig. 3. Heat diffusion on the torus after 5 steps of the Markov chain, starting from the point denoted by $\times$. **(a)** Isotropic kernel; **(b)** Proposed kernel when the major angle is the desired signal. The diffusion is then faster along the minor angle; **(c)** Proposed kernel when the minor angle is the desired signal.



(a) Isotropic diffusion (b) Directed diffusion faster in radial direction (top view) (c) Directed diffusion faster in angular direction (top view)

Fig. 4. Heat diffusion on the cone-like surface after 5 steps of the Markov chain, starting from the point denoted by $\times$. **(a)** Isotropic kernel; **(b)** Proposed kernel when the angular position is the desired signal (top view); **(c)** Proposed kernel when the distance from the cone tip is the desired signal (top view).

Each point on the torus is parametrized as

$$\boldsymbol{s}(x,v) = \begin{bmatrix} (R + r\cos(2\pi v))\,\cos(2\pi x) \\ (R + r\cos(2\pi v))\,\sin(2\pi x) \\ r\sin(2\pi v) \end{bmatrix}, \qquad (37)$$

where R and r are the major and minor radius, and x and v are the major and minor angle of the torus, respectively. Similarly as in the strip experiment, we assume that the following distance is accessible

$$d_v(v_i, v_j) = \|\boldsymbol{n}_{v_i} - \boldsymbol{n}_{v_j}\|_2, \quad \boldsymbol{n}_v = [\cos(v), \sin(v)]^{\mathrm{T}}. \quad (38)$$

If the minor angle is a desired signal, the kernel is constructed using $d_x(x_i, x_j)$, defined analogously to (38). The diffusion on the torus surface resulting from the different kernels is shown in Figure 3. While the standard kernel is characterized by an isotropic diffusion, the proposed kernels induce directed diffusions that are faster along one of the angles.

As a last illustration, we consider points sampled on a concave cone-like surface in $\mathbb{R}^3$, illustrated in Figure 4(a). Each point is parametrized by the distance r from the cone tip, and the angle θ as follows

$$\boldsymbol{s}(r,\theta) = \left[r\cos(\theta),\ \ r\sin(\theta),\ \ r^{0.2} + r^{0.8}\right]^{\mathrm{T}}. \qquad (39)$$

If the angular location of a point is the desired signal, we construct a kernel that induces a heat diffusion that is faster along the radial direction. This is achieved with the bandwidth

$$b_r(i,j) = (1 + d_r(r_i, r_j))^2. \qquad (40)$$

If the distance from the tip is the desired signal, the kernel bandwidth can be computed similarly as in the torus example. The heat diffusion induced by the proposed kernels is illustrated in Figures 4(b) and 4(c). We show top view of the point cloud to better illustrate the direction of diffusion.

VI. Experiments with real data: fetal ECG extraction

In this section, we apply the proposed VB kernels to estimate the fetal instantaneous heart rate (fIHR) non-invasively, from abdominal maternal electrocardiogram (mECG) [38],

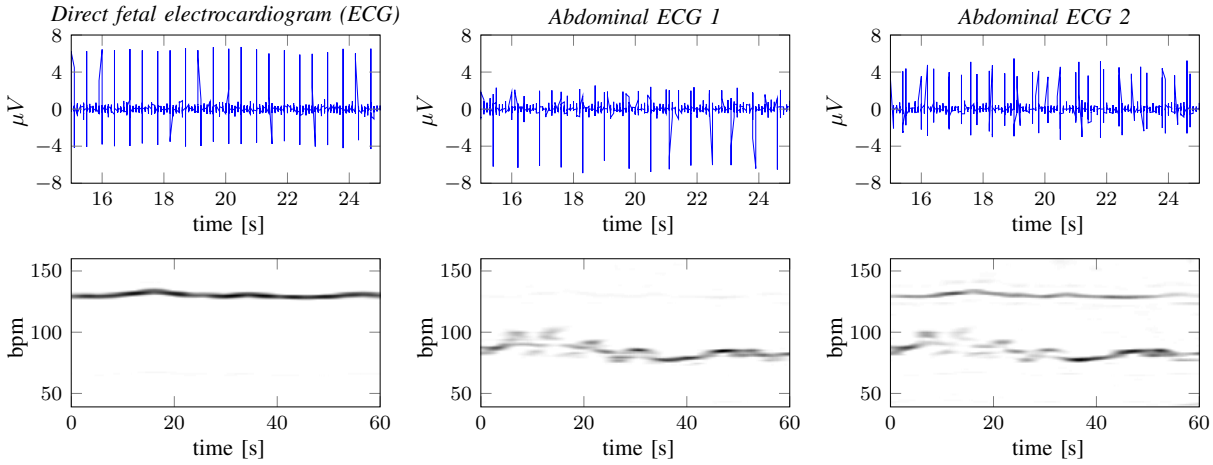


Fig. 5. Example signals from Patient 1 in the *adfecgdb* database. Top: time-domain signals. Bottom: dsSTFT representations.

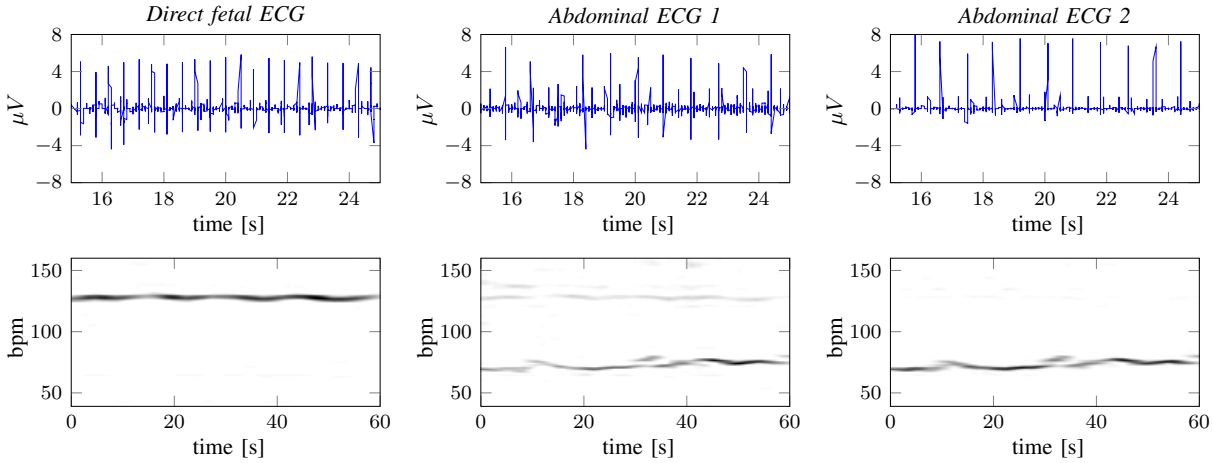


Fig. 6. Example signals from Patient 2 in the *adfecgdb* database. Top: time-domain signals. Bottom: dsSTFT representations.

[39]. We use ECG signals from the PhysioNet collection [40]. The fIHR extraction problem is suitable to demonstrate the non-linear filtering capability of the proposed kernels with and without an auxiliary sensor. We note that recovery of the fetal electrocardiogram (fECG) waveform involves additional steps after fIHR extraction: beat tracking and median filtering [39]. However, as the overall scheme relies on the fIHR, we only consider the latter in our experiments.

A. Experiment description

Estimation of fIHR from signals that contain mECG corresponds to a multiple frequency detection problem. A time-frequency transform for this type of problems, known as the de-shape short-time Fourier transform (dsSTFT), was recently proposed in [41]. As the fECG in the abdominal signals is an order of magnitude weaker than the mECG, it is not dominant in the dsSTFT spectrum, and often not detected at all. The dsSTFT was employed for fIHR estimation in [39], by first estimating the mECG and then subtracting this estimate from the abdominal signal. In the following, we show that using the proposed kernels, the fIHR can be obtained without estimating the mECG waveform first. All ECG signals are sampled at 1 kHz with 16-bit resolution. Measurement vectors $\boldsymbol{s}$ are obtained using a lag-map, by concatenating 256 consecutive signal samples, with a hop of 10 samples between measurements. Each experiment consists of a 25 seconds signal excerpt from a given patient, resulting in $N = 2500$ data points per experiment. The following pre-processing steps are applied to the waveforms [39]: low-pass filtering with 100 Hz cut-off to suppress noise, median filtering with a window length of 0.1 seconds to subtract trends, and normalization to unit variance.

B. Evaluation with a direct fetal ECG reference

In this experiment, we use the *Abdominal and Direct Fetal Electrocardiogram Database (adfecgdb)* from PhysioNet [38], which contains abdominal ECGs from five women between 38 and 40 weeks of pregnancy. A direct fetal ECG recorded with a fetal scalp lead is included for each patient. Signal excerpts from two patients with the corresponding dsSTFTs are shown in Figures 5 and 6. Even if in some signals the fetal heart rate is detected, the maternal instantaneous heart rate (mIHR) is always the dominant spectral line. Our objective is to apply the proposed kernels and obtain a signal representation where the fIHR is the dominant spectral line.

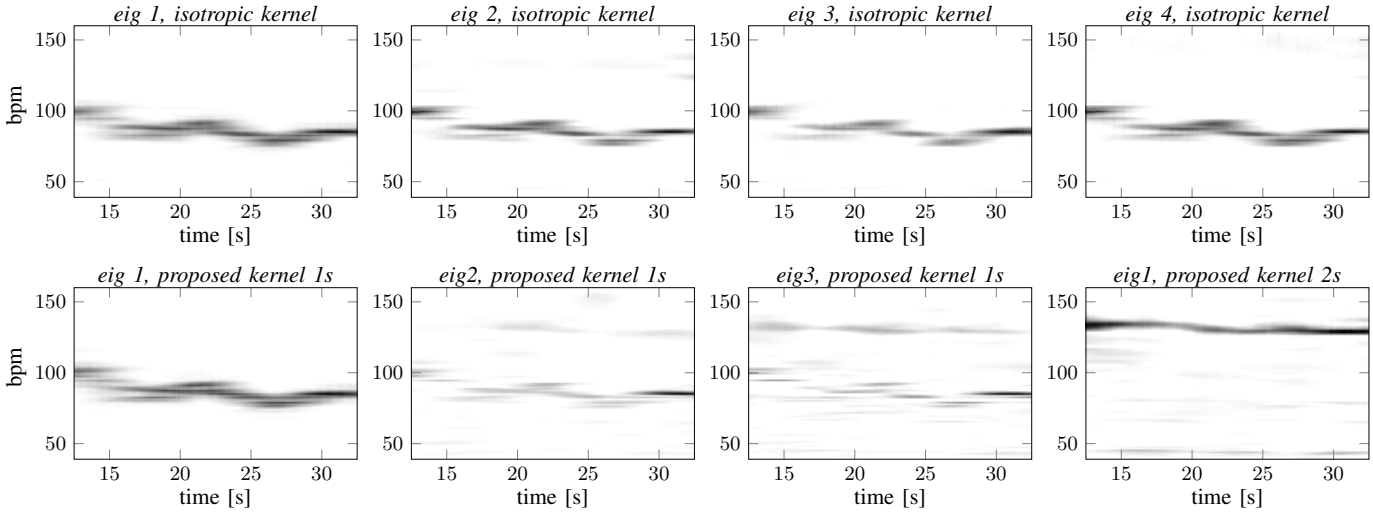


Fig. 7. Eigenvectors from the different diffusion kernels for Patient 1. **Top:** isotropic kernel. **Bottom:** from proposed kernels without and with an auxiliary sensor (indicated by *1s* and *2s* respectively).

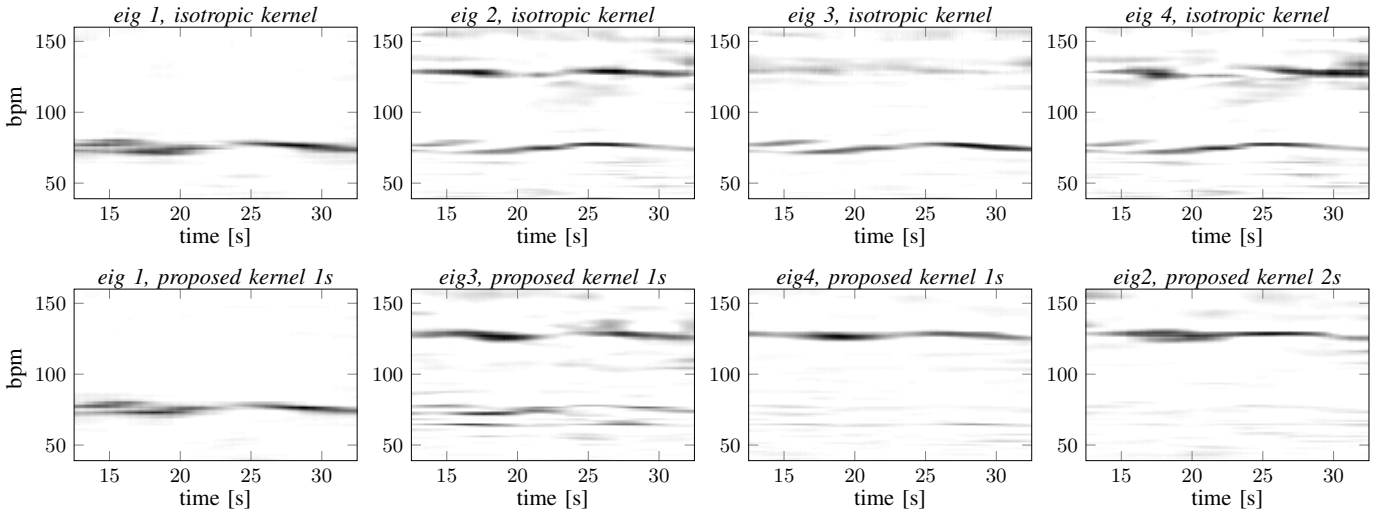
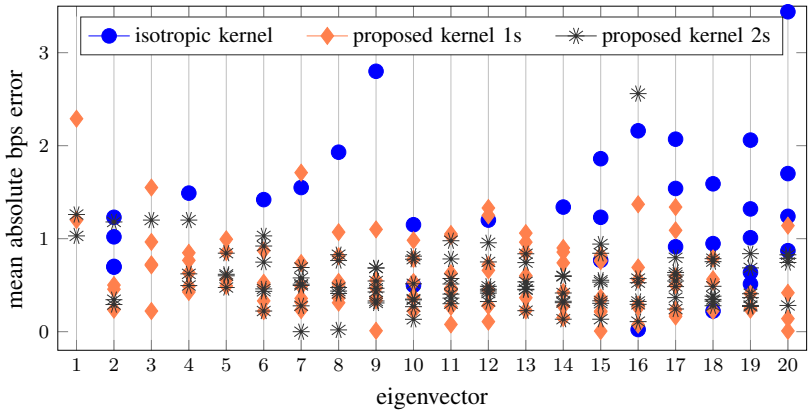


Fig. 8. Eigenvectors from the different diffusion kernels for Patient 2. **Top:** isotropic kernel. **Bottom:** from proposed kernels without and with an auxiliary sensor (indicated by *1s* and *2s* respectively).

The dsSTFT of the first few leading eigenvectors from two experiments are shown in Figures 7 and 8. The proposed kernel was first implemented without an auxiliary sensor, where the noise distances are estimated as proposed in Section IV-B2. This scenario is denoted by *1s*. Then, the kernel was implemented using a second abdominal ECG as an auxiliary sensor, where the noise distances are estimated using AD, as proposed in Section IV-B1. This scenario is denoted by *2s*. In both cases, the early eigenvectors of the proposed kernels recover the fIHR. In particular, in the *2s* scenario, a complete suppression of the maternal ECG is observed already in the first or second eigenvector. It is important to mention that the effectiveness of the proposed kernels is influenced by the fECG strength in the abdominal ECGs. For instance, for the second patient, the fECG appears in the dsSTFT of the unprocessed ECG, shown in Figure 8. However, application of our algorithm ensures that the fECG is the dominant spectral line.

To obtain a quantitative evaluation, we extracted instantaneous heart rate (IHR) curves from each of the first 20 eigenvectors in 9 different experiments, using signal excerpts from four patients. The IHR was computed from the dsSTFTs representation using the method presented in [39], and compared to the reference fIHR. From the total of 180 analyzed eigenvectors for each kernel, we only kept the eigenvectors that successfully extracted the fIHR. The scatter plot in Figure 9 shows the mean error in beats-per-second (bps) for each of these eigenvector. The percentage of eigenvectors that extracted the fIHR is shown in the accompanying table. Notice that the percentage is by more than three times larger for the proposed kernels than for the isotropic one. Even by considering only the first 10 eigenvectors per experiment, the fIHR is recovered in more than 50% of the cases. Importantly, these tend to be higher in the spectrum than the eigenvectors of an isotropic kernel. We once again emphasize that multiple



	isotr.	prop-1s	prop-2s
% for 20	20	62	61
avg. error [bps]	1.2	0.7	0.5
% for 10	13	56	50
avg. error [bps]	1.26	0.8	0.6

Fig. 9. Summary of quantitative results over 9 experiments (4 patients in total, but multiple electrodes per patient available). The scatter plot illustrates all eigenvectors that detect the fetal ECG. The table summarizes the percentage of eigenvectors detect the fetal ECG (when considering the first 20 - top row, and the first 10 - bottom row), as well as the average absolute error in fIHR estimate compared to the reference fIHR.

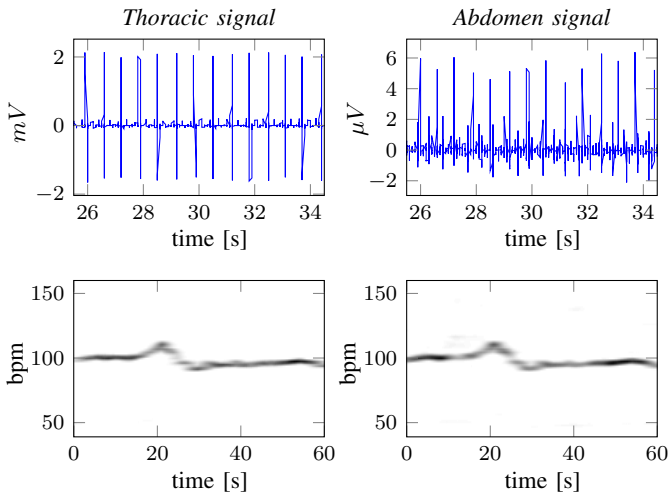


Fig. 10. Signal examples from the *nifecgdb* dataset. Top row: time-domain waveforms. Bottom row: dsSTFTs representations.

eigenvectors of the diffusion kernels can encode the same direction of variability in the data [33]. In fact, as visible in Figure 9, a large proportion of the leading eigenvectors can used to estimate the fIHR with a good accuracy. The average fIHR estimation error (averaged across eigenvectors) of the proposed kernel in the *2s* scenario is 0.5 bps. The error in the *1s* scenario is 0.6 bps, while the isotropic kernel is inferior with an error of 1.2 bps.

It would be of interest to compare the accuracy of the fIHR to the related dsSTFT-based approach in [39]. However, the results only report performance from later stages of the fECG extraction pipeline, after the fIHR estimation takes place. An in-depth analysis of the influence of the proposed fIHR estimation method on the complete pipeline is a topic for future application-dedicated research.

C. Qualitative evaluation without a fetal ECG reference

In this experiment, we use the *Non-Invasive Fetal Electrocardiogram Database (nifecgdb)* from PhysioNet, which consists of abdominal ECGs of women between 21 and 40 weeks of pregnancy. The recordings include a thoracic signal which provides a good reference of the maternal ECG. Sample waveforms from the database are shown in Figure 10.

In most recordings, we noticed an extremely weak fetal ECG compared to the mECG, which is visible in Figure 10. Consequently, the fIHR was not recovered among the top eigenvectors of an isotropic kernel. However, given the thoracic mECG signal as a reference, the proposed kernel is particularly suited for this scenario: the metric d_v can be accurately estimated with the AD algorithm applied with an abdominal sensor and the thoracic sensor. The results for one patient are shown in Figure 11. It can be seen that the second eigenvector recovers the fIHR, while removing the mIHR from the spectrum. For this experiment, we only present a qualitative result, as without a fECG we were unable to perform quantitative analysis as in Section VI-B, since we do not have the ground truth.

VII. Conclusions

In this paper, we developed a non-linear filtering framework based on diffusion kernels. Distinguishing properties of the proposed kernels are their non-homogeneity and anisotropy, determined by a noise-informed kernel bandwidth. Our algorithmic concept is that by extracting geometric information about the noise signal from the measurements, one can define a suitable kernel bandwidth function, which is equivalent to defining a metric that is less sensitive to noise variations that the Euclidean distance in the measurement space. The findings in this paper open a few interesting questions for future research. These include characterization of a broader family of possible bandwidth functions with filtering capabilities, and extending the signal representation to new measurements. We note that extension to new measurements is a weak point of kernel-based approaches in general, and certain techniques have already been investigated in the literature. However, the applicability of these techniques in combination with a data-driven kernel bandwidth is an important open question. Finally, the proposed bandwidth function can be combined with other task-driven metric transforms to devise new kernels for a wider range of applications.

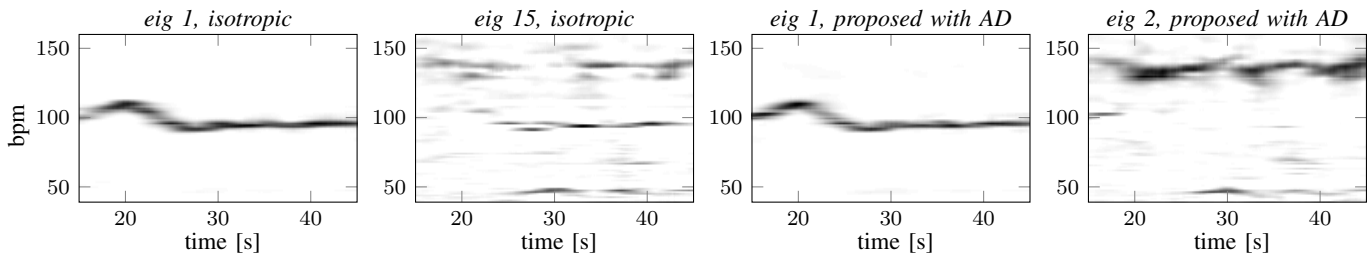


Fig. 11. Results from the *nifecgdb* dataset. The AD distance extracted with the thoracic auxiliary sensor is consistent with the mIHR, and allows for its complete suppression in eigenvector 2 (rightmost figure). The isotropic kernel does not recover the fECG in early eigenvectors.

REFERENCES

[1] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, 2010.

[2] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI: American Mathematical Society, 1997.

[3] Y. Weiss, "Segmentation using eigenvectors: a unifying view," in *Proc. Seventh IEEE Int. Conf. Comput. Vis.*, 1999, pp. 975–982.

[4] U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, dec 2007.

[5] M. Belkin and P. Niyogi, "Towards a Theoretical Foundation for Laplacian-Based Manifold Methods," *J. Comput. Syst. Sci.*, vol. 74, no. 8, pp. 1289–1308, 2008.

[6] M. Hein, J. Audibert, and U. von Luxburg, "From graphs to manifolds - Weak and strong pointwise consistency of graph Laplacians," in *Proc. 18th Conf. Learn. Theory (COLT), Lect. Notes Comput. Sci., vol. 3559*, P. Auer and R. Meir, Eds. Berlin: Springer-Verlag, 2005, pp. 470–485.

[7] P. H. Berard, *Spectral Geometry: Direct and Inverse Problems*, A. Dold and B. Eckmann, Eds. Springer-Verlag Berlin Heidelberg, 1986.

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

[9] M. Meila and J. Shi, "A random walks view of spectral segmentation," in *AI Stat.*, 2001.

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

[11] S. Lafon, "Diffusion Maps and Geometric Harmonics," Ph.D. dissertation, Yale University, 2004.

[12] B. Nadler, S. Lafon, R. Coifman, and I. G. Kevrekidis, "Diffusion maps - A probabilistic interpretation for spectral embedding and clustering algorithms," in *Lect. Notes Comput. Sci. Eng.*, A. N. Gorban, B. Kegl, D. C. Wunsch, and Z. Andrei, Eds. Springer-Verlag Berlin Heidelberg, 2008, ch. 10, pp. 238–260.

[13] B. Nadler and M. Galun, "Fundamental Limitations of Spectral Clustering," in *Adv. Neural Inf. Process. Syst. 19*, 2007, pp. 1017–1024.

[14] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," *Adv. Neural Inf. Process. Syst.*, pp. 1601–1608, 2004.

[15] T. Berry and J. Harlim, "Variable bandwidth diffusion kernels," *Appl. Comput. Harmon. Anal.*, vol. 40, no. 1, pp. 68–96, 2016.

[16] C. J. Dsilva, R. Talmon, C. W. Gear, R. R. Coifman, and I. G. Kevrekidis, "Data-Driven Reduction for a Class of Multiscale Fast-Slow Stochastic Dynamical Systems," *SIAM J. Appl. Dyn. Syst.*, pp. 1327–1351, 2016.

[17] D. Giannakis and A. J. Majda, "Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability," *Proc. Natl. Acad. Sci.*, vol. 109, no. 7, pp. 2222–2227, 2012.

[18] D. Giannakis, "Dynamics-adapted cone kernels," *SIAM J. Appl. Dyn. Syst.*, vol. 14, no. 2, pp. 556–608, 2015.

[19] O. Yair and R. Talmon, "Multimodal metric learning with local CCA," in *IEEE Stat. Signal Process. Work.*, jun 2016, pp. 1–5.

[20] V. Papyan and R. Talmon, "Multimodal latent variable analysis," *Signal Processing*, vol. 142, pp. 178–187, 2018.

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

[22] E. Kreyszig, *Introductory functional analysis with applications*. John Wiley & Sons, Ltd, 1978.

[23] V. Guillemin and A. Pollack, *Differential Topology*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1974.

[24] Y. Terada and U. von Luxburg, "Local Ordinal Embedding," in *Int. Conf. Mach. Learn.*, vol. 32, 2014, pp. 847–855.

[25] R. Talmon, I. Cohen, S. Gannot, and R. R. Coifman, "Diffusion Maps for Signal Processing," *IEEE Signal Process. Mag.*, vol. 30, no. july, pp. 75–86, 2013.

[26] M. Belkin, "Problems of Learning on Manifolds," Ph.D. dissertation, The University of Chicago, 2003.

[27] D. Ting, L. Huang, and M. I. Jordan, "An analysis of the convergence of Graph Laplacians," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010.

[28] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proc. Natl. Acad. Sci.*, vol. 102, no. 21, pp. 7426–7431, 2005.

[29] A. Haddad, D. Kushnir, and R. R. Coifman, "Texture separation via a reference set," *Appl. Comput. Harmon. Anal.*, vol. 36, no. 2, pp. 335–347, 2014.

[30] A. D. Szlam, M. Maggioni, and R. R. Coifman, "Regularization on graphs with function-adapted diffusion processes," *J. Mach. Learn. Res.*, vol. 9, pp. 1711–1739, 2008.

[31] O. Yair, R. Talmon, R. R. Coifman, and I. G. Kevrekidis, "Reconstruction of normal forms by learning informed observation geometries from data," in *Proc. Natl. Acad. Sci.*, vol. 114, no. 38, 2017, pp. E7865–E7874.

[32] A. Holiday, M. Kooshkbaghi, J. M. Bello-Rivas, C. W. Gear, A. Zagaris, and I. G. Kevrekidis, "Manifold learning for parameter reduction," *arXiv:1807.08338*, 2018.

[33] C. J. Dsilva, R. Talmon, R. R. Coifman, and R. Kevredikis, "Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study," *Appl. Comput. Harmon. Anal.*, vol. 44, no. 3, pp. 759–773, 2018.

[34] S. Gerber, T. Tasdizen, and R. Whitaker, "Robust non-linear dimensionality reduction using successive 1-dimensional Laplacian Eigenmaps," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 1–8.

[35] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*. SIAM, Software, Environments, and Tools, 1998.

[36] R. R. Lederman and R. Talmon, "Learning the geometry of common latent variables using alternating-diffusion," *Appl. Comput. Harmon. Anal.*, vol. 44, no. 2018, pp. 509–536, 2015.

[37] R. Talmon and H.-T. Wu, "Latent common manifold learning with alternating diffusion: Analysis and applications," *Appl. Comput. Harmon. Anal.*, 2018.

[38] J. Jezewski, A. Matonia, T. Kupka, D. Roj, and R. Czabanski, "Determination of the fetal heart rate from abdominal signals: evaluation of beat-to-beat accuracy in relation to the direct fetal electrocardiogram," *Biomed. Eng. / Biomed. Tech.*, vol. 57, no. 5, pp. 383–394, 2012.

[39] L. Su and H.-T. Wu, "Extract Fetal ECG from Single-Lead Abdominal ECG by De-Shape Short Time Fourier Transform and Nonlocal Median," *Front. Appl. Math. Stat.*, vol. 3, feb 2017. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fams.2017.00002/full

[40] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet : Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[41] C.-Y. Lin, L. Su, and H.-T. Wu, "Wave-Shape Function Analysis," *J. Fourier Anal. Appl.*, vol. 24, no. 2, pp. 451–505, apr 2018.