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 data 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 domain of the data. 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 variables 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]–[7].

The graph affinity matrix, if properly normalized, can be interpreted as the transition probability matrix of a Markov chain on the graph [2], [8], which converges to a diffusion process on the corresponding manifold [8]–[10]. 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], [11]. This is the underlying idea behind *directed diffusions* [9], *self-tuning kernels* [12], and other kernels with a data-driven distance metric [13] which are successfully applied to many applications in the past decade. These applications include analysis of dynamical systems [14]–[16], multimodal data analysis [17], [18], non-linear independent component analysis [19], and system identification [4].

In this paper, we address the problem of representation learning from measurements, which besides the phenomenon of interest, contain other undesired sources of variability that lie on a different low-dimensional manifold. We propose data-driven kernels, whose corresponding Markov chains (or

Manuscript received May 22, 2019; revised August 29, 2019 and October 28, 2019; accepted November 20, 2019. Date of publication December 13, 2019; date of current version January 10, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Olivier Lezoray. This work was supported in part by the Research Foundation Flanders under Grant 12X6719 N, in part by the Minerva Stiftung short-term research grant, Israel Science Foundation under Grant 1490/16, in part by the KU Leuven Internal Funds C2-16-00449 and VES/19/004, and in part by the European Research Council under the European Union's Horizon 2020 Research and Innovation Program/ERC Consolidator Grant SONORA 773268. *(Corresponding author: Maja Taseska.)*

M. Taseska and T. van Waterschoot are with the Department of Electrical Engineering (ESAT-STADIUS/ETC), Katholieke Universiteit Leuven, 3000 Leuven, Belgium (e-mail: maja.taseska@esat.kuleuven.be; toon.vanwaterschoot@esat.kuleuven.be).

E. A. P. Habets is with the International Audio Laboratories Erlangen (a joint institution between the University of Erlangen-Nuremberg and Fraunhofer IIS), 91058 Erlangen, 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).

Digital Object Identifier 10.1109/TSP.2019.2959190

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 find a low-dimensional representation of the measurements, that recovers relevant geometric properties of the phenomenon of interest, while reducing undesired variability in the data. To reach this objective, we require prior information that enables us to approximate the distance metric on the undesired manifold. Although the requirement of such prior 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 the proposed kernels can be applied to enhance small-scale sources of variability 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 compact Riemannian manifolds $(\mathcal{X}, g_x)$ and $(\mathcal{V}, g_v)$, respectively. X and V are related to an observable variable S by an unknown deterministic function h which embeds the product manifold $\mathcal{X} \times \mathcal{V}$ into a l_s-dimensional Euclidean space, as follows:[1]

$$S = h(X, V), \quad h : \mathcal{X} \times \mathcal{V} \to \mathcal{S}, \quad \mathcal{S} \subset \mathbb{R}^{l_s}. \tag{1}$$

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, such as time-frequency transform of a time series, lag-map, pixels of an image, etc. The function h comprises the sensor mechanism, and possibly, application-specific preprocessing transforms.

If d_x and d_v are distance functions on $\mathcal{X}$ and $\mathcal{V}$, induced by the corresponding metric tensors g_x and g_v, a distance on $\mathcal{X} \times \mathcal{V}$ can be defined as [21, Ch 1]

$$d_{xv}\left((\boldsymbol{x}_1, \boldsymbol{v}_1), (\boldsymbol{x}_2, \boldsymbol{v}_2)\right) = \left(d_x(\boldsymbol{x}_1, \boldsymbol{x}_2)^p + d_v(\boldsymbol{v}_1, \boldsymbol{v}_2)^p\right)^{\frac{1}{p}}, \tag{2}$$

for any $1 \leq p < \infty$. We assume that the measurement function h is a monotonic and locally isometric embedding of $\mathcal{X} \times \mathcal{V}$ into $\mathbb{R}^{l_s}$. Namely, if d_s denotes the Euclidean distance on $\mathbb{R}^{l_s}$, then for all $(\boldsymbol{x}, \boldsymbol{v})$ in the neighborhood of $(\boldsymbol{x}_1, \boldsymbol{v}_1)$ it holds that

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

[1]As $\mathcal{X}$ and $\mathcal{V}$ are smooth Riemannian manifolds, the product $\mathcal{X} \times \mathcal{V}$ is also a smooth manifold [20, page 5].

In applications where the local isometry might be restrictive, the Euclidean distance d_s can be replaced by a data-driven Mahalanobis distance. It was shown in [19] that by computing a data-driven Mahalanobis distance, one can approximate small distances on the underlying manifold, thereby satisfying (3) to the second order.

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

$$S^{(a)} = h^{(a)}(V, Z), \quad h^{(a)} : \mathcal{V} \times \mathcal{Z} \to \mathcal{S}^{(a)}, \tag{4}$$

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 $h^{(a)}$, which may be different from h. The auxiliary sensor is endowed with an analogous metric structure as (2) and (3). The proposed data model might be relevant in various multi-sensor medical applications [22]–[24], as well as audio-visual applications with multiple microphones and/or cameras [18].

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 variables 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 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, i.e.,

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

that approximately preserves local distances among $\{\boldsymbol{x}_i\}_{i=1}^N$ as defined by the metric d_x on the manifold of interest $\mathcal{X}$, while contracting distances as defined by d_v on the undesired manifold $\mathcal{V}$. Under certain circumstances, it has been shown that embeddings which preserve local distances suffice to approximately reconstruct the latent points $\{\boldsymbol{x}_i\}_{i=1}^N$ [25]. We note that construction of manifold embeddings with a small local bi-Lipschitz distortion has been discussed in [7], when the measurements are sampled directly from a manifold of interest $\mathcal{X}$.

III. Diffusion Kernels for Manifold Learning: A Brief Overview

Manifold learning approaches are often used for data processing by modeling the measurement samples $\{\boldsymbol{s}_i\}_{i=1}^N \in \mathcal{S}$ as points on or near a smooth and compact low-dimensional manifold $\mathcal{X}$, embedded in the ambient space $\mathcal{S}$ [26]. 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 $\mathcal{X}$ [27], [28]. This setting is simpler than the signal model we introduced in Section II, where the measurements are samples

from a product manifold $\mathcal{X} \times \mathcal{V}$. 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.,

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

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]. The probability of the Markov chain that started at $\boldsymbol{s}_i$, to be at $\boldsymbol{s}_j$ at step t is given by

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

The Markov chain on the graph leads to a natural definition of distance between points based on their connectivity, known as the *diffusion distance* [8], [9]. 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

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

It is shown in [8] that the diffusion distance can be computed using the eigenvectors $\{\boldsymbol{\psi}_i\}_{i=0}^{N-1}$ and eigenvalues $1 = \lambda_0 > \lambda_1 \geq \ldots > 0$ of $\boldsymbol{P}$. Generally, due to the intrinsic low-dimensionality of the manifold, the spectrum of $\boldsymbol{P}$ has a rapid decay and the diffusion distance can be approximated using only the first few eigenvectors. An l-dimensional diffusion maps embedding $\Psi_t : \mathcal{S} \to \mathbb{R}^l$, for given t and l is defined as

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

where the constant eigenvector $\boldsymbol{\psi}_0$ is not included. Hence, an l-dimensional diffusion map with $l < l_s$, embeds the data approximately isometrically with respect to d_t [9], [29]. The dimensionality l is chosen by identifying the number of significant eigenvalues of $\boldsymbol{P}^t$, e.g., by setting a threshold depending on the desired accuracy [8, Sec. 2.5].

If the measurements are sampled uniformly on the manifold, the eigenvectors of the the isotropic diffusion kernel constructed by (6)-(8) converge to the eigenfunctions of the LBO[2] in the limits $N \to \infty$ and $\epsilon \to 0$ [5]. To maintain this property for an arbitrary sampling density, an additional normalization of the kernel $\boldsymbol{K}$ is required as follows [9], [29]

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

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

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 as follows

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

where $\boldsymbol{M}$ is a task-driven matrix approximating the metric tensor on the manifold. Such kernel construction has been often used in the past decade for dynamical system analysis [14], [30], data fusion [17], non-linear independent component analysis [19], and other applications [4], [26].

Other approaches for informed metric construction based on prior information about the problem at hand have been proposed in [31]–[33].

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 accurately captures 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. 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.[3] 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 [36]. Our objective is to design

[2]The LBO on compact Riemannian manifolds has a discrete spectrum.

[3]The second problem is related to the crucial property of the LBO eigenfunctions, namely, that different eigenfunctions may encode the same source of variability on the manifold. See [34], [35] for more details about this property and its implications in practice.

suitable diffusion kernels which warp the data geometry in a way that information about the variable 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) kernels [13], where the bandwidth is prescribed by a real-valued, non-negative, and symmetric 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 [12] and dynamical system modeling [15], [16]. 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 on $\mathcal{S}$, according to

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

If a kernel implemented with d'_s is to have more leading eigenvectors that are smooth with respect to the $\mathcal{X}$, compared to a kernel implemented with d_s, the distance d'_s should be less sensitive to the undesired variable than the observable Euclidean distance d_s. To achieve such behavior, we propose the following 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. It should be mentioned that although the resulting d'_s is used as a distance, it is not guaranteed to obey the triangle inequality.

The bandwidth function in (17) was chosen such that distances in the kernel-induced geometry 1) are less sensitive to undesired sources of variability than the Euclidean distances in the measurement space, 2) are robust to estimation errors in d_v, and 3) preserve the local geometry of the desired manifold $\mathcal{X}$. Note that the proposed bandwidth in (17) is not the only function that satisfies these properties. In fact, any smooth monotonic transformation of d_v that is locally bi-Lipschitz has the potential to provide non-linear filtering capability in the resulting kernels. Theoretical justification for our bandwidth function in (17) is provided 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

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

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

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

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

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

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

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

Furthermore, from the distance properties in (3), (2) we have

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

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 $|d'_s(\boldsymbol{s}_i, \boldsymbol{s}_j) - \hat{d}'_s(\boldsymbol{s}_i, \boldsymbol{s}_j)| \leq \epsilon_v\, d_s(\boldsymbol{s}_i, \boldsymbol{s}_j)$

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

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

Omitting the distance function arguments for brevity, we have

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

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

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 $d_x(\boldsymbol{x}_i, \boldsymbol{x}_j) < d_x(\boldsymbol{x}_k, \boldsymbol{x}_l) \Longrightarrow d'_s(\boldsymbol{s}_i, \boldsymbol{s}_j) < d'_s(\boldsymbol{s}_k, \boldsymbol{s}_l)$.

Proof: If $d_x(\boldsymbol{x}_i, \boldsymbol{x}_j) < d_x(\boldsymbol{x}_k, \boldsymbol{x}_l)$, then from (2) it follows that

$$d_{xv}\left((\boldsymbol{x}_1, \boldsymbol{v}_1), (\boldsymbol{x}_2, \boldsymbol{v}_2)\right) < d_{xv}\left((\boldsymbol{x}_1, \boldsymbol{v}_1), (\boldsymbol{x}_2, \boldsymbol{v}_2)\right). \tag{24}$$

As the measurement function h is assumed to be monotonic, then (24) implies that

$$d_s(\boldsymbol{s}_i, \boldsymbol{s}_j) < d_s(\boldsymbol{s}_k, \boldsymbol{s}_l). \tag{25}$$

The proposition follows immediately from (19) and (25), i.e.,

$$d_s(\boldsymbol{s}_i, \boldsymbol{s}_j)\left(1 + \xi\right)^{-1} < d_s(\boldsymbol{s}_k, \boldsymbol{s}_l)\left(1 + \xi\right)^{-1}$$
$$\Leftrightarrow d'_s(\boldsymbol{s}_i, \boldsymbol{s}_j) < d'_s(\boldsymbol{s}_k, \boldsymbol{s}_l). \tag{26}$$

■

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 source of variability, corrupted by sensor-specific variables [37], [38]. In our problem, the undesired source of variability is captured by the primary and the auxiliary sensor. Hence, the AD algorithm can be used to find an embedding that provides an estimate of the pairwise distances $d_v(\boldsymbol{v}_i, \boldsymbol{v}_j)$. The key object of AD is the AD kernel $\boldsymbol{P}_{\rm ad}$ [37], defined as

$$\boldsymbol{P}_{\rm ad} = \boldsymbol{P}\,\boldsymbol{P}^{(a)}. \tag{27}$$

where $\boldsymbol{P}$ and $\boldsymbol{P}^{(a)}$ are the standard sensor-specific diffusion kernels. The diffusion kernel $\boldsymbol{P}$ of the primary sensor is computed by (6), (7), and (13). The diffusion kernel $\boldsymbol{P}^{(a)}$ is computed following the same steps for the auxiliary sensor measurements, i.e.,

$$\boldsymbol{K}^{(a)}[i,j] = \exp\left(-\frac{\|\boldsymbol{s}_i^{(a)} - \boldsymbol{s}_j^{(a)}\|_2^2}{\varepsilon}\right), \tag{28a}$$

$$\boldsymbol{D}^{(a)}[i,i] = \sum_{j=1}^{N} \boldsymbol{K}^{(a)}[i,j], \tag{28b}$$

$$\boldsymbol{K}_o^{(a)} = [\boldsymbol{D}^{(a)}]^{-1}\,\boldsymbol{K}^{(a)}\,[\boldsymbol{D}^{(a)}]^{-1}, \tag{28c}$$

$$\boldsymbol{D}_o^{(a)}[i,i] = \sum_{j=1}^{N} \boldsymbol{K}_o^{(a)}[i,j], \tag{28d}$$

$$\boldsymbol{P}^{(a)} = [\boldsymbol{D}_o^{(a)}]^{-1}\,\boldsymbol{K}_o^{(a)}. \tag{28e}$$

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 [38]

$$\Psi_{\rm ad}(\boldsymbol{s}_i, \boldsymbol{s}_i^{(a)}) = \left[\,\sigma_1 \boldsymbol{v}_1[i],\, \sigma_2 \boldsymbol{v}_2[i],\, \ldots, \sigma_d \boldsymbol{v}_l[i]\,\right]^{\rm T}. \tag{29}$$

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

$$d_{\rm ad}(i,j) = \|\Psi_{\rm ad}(\boldsymbol{s}_i, \boldsymbol{s}_i^{(a)}) - \Psi_{\rm ad}(\boldsymbol{s}_j, \boldsymbol{s}_j^{(a)})\|_2. \tag{30}$$

According to [37], $\Psi_{\rm ad}$ 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

$$\begin{aligned} d'_s(\boldsymbol{s}_i, \boldsymbol{s}_j) &= \frac{d_s(\boldsymbol{s}_i, \boldsymbol{s}_j)}{1 + d_{\rm ad}(i,j)} \\ &= \frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2}{1 + \|\Psi_{\rm ad}(\boldsymbol{s}_i, \boldsymbol{s}_i^{(a)}) - \Psi_{\rm ad}(\boldsymbol{s}_j, \boldsymbol{s}_j^{(a)})\|_2}, \end{aligned} \tag{31}$$

which corresponds to a kernel with the bandwidth function

$$b(i,j) = (1 + d_{\rm ad}(i,j))^2. \tag{32}$$

We note that the dimensionality l of $\Psi_{\rm ad}$ is not very critical. In theory, all eigenvectors of the AD kernel are smooth with respect to 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 (29). In addition, if only the first few coordinates are used, the AD can be applied to estimate d_v even if the auxiliary sensor captures the desired variable. This is an important advantage in practice, that doesn't allow leakage of the desired variable from the auxiliary sensor to disrupt the filtering capability of the kernel.

2) Estimating d_v without an Auxiliary Sensor: If only the measurements $\{\boldsymbol{s}_i\}_{i=1}^N$ from the primary sensor are given, the pairwise distances $\hat{d}_v(\boldsymbol{v}_i, \boldsymbol{v}_j)$ can be estimated, provided that the undesired variable V represents the largest scale source of variability on the product manifold $\mathcal{X} \times \mathcal{V}$. Recall the structure of the diffusion spectrum: the largest-scale source of variability corresponds to the slowest relaxation processes of the Markov chain, which in turn, correspond to the largest eigenvalues of the kernel [2]. From the manifold perspective, these correspond to the LBO eigenfunctions with maximal smoothness. Hence, if we consider the one-dimensional diffusion map obtained with a standard kernel as described in Section III-A,

$$\Psi_1(\boldsymbol{s}_i) = \lambda_1\,\boldsymbol{\psi}_1[i], \tag{33}$$

the Euclidean distance $|\Psi_1(\boldsymbol{s}_i) - \Psi_1(\boldsymbol{s}_j)|$ can be used as an approximation of $d_v(\boldsymbol{v}_i, \boldsymbol{v}_j)$. Consequently, we propose to implement the following metric transform for our kernel

$$d'_s(\boldsymbol{s}_i, \boldsymbol{s}_j) = \frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|_2}{1 + |\Psi_1(\boldsymbol{s}_i) - \Psi_1(\boldsymbol{s}_j)|}, \tag{34}$$

which corresponds to a kernel with the bandwidth function

$$b(i,j) = (1 + |\Psi_1(\boldsymbol{s}_i) - \Psi_1(\boldsymbol{s}_j)|)^2. \tag{35}$$

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 [35] and nonlinear dynamical system analysis [34].

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 [37], where for each point i, a local scale ε_i is introduced that is equal to the median of the squared distances from a chosen number of 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 number of significant eigenvalues is not suitable to determine the dimensionality l_x of the embedding f. Although the eigenvectors that parametrize the variable of interest are higher in the spectrum of the proposed kernel compared to an isotropic one, eigenvectors that parametrize the undesired variable may have large eigenvalues as well. This is the case if the undesired variable remains sufficiently smooth with respect to the product manifold $\mathcal{X} \times \mathcal{V}$ even after warping its kernel-induced geometry by the proposed bandwidth function. Instead, relevant eigenvectors (and hence l_x) can for instance be identified by calculating

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: For each pair, (i,j) compute the local scale ε_{ij} as described in Section IV-C.
3: Construct an exponential kernel matrix $\boldsymbol{K}$ with the VB kernel $\boldsymbol{K}[i,j] = \exp(-\frac{d(\boldsymbol{s}_i,\boldsymbol{s}_j)^2}{\varepsilon_{ij}\, b(i,j)})$.
4: Apply density normalization to $\boldsymbol{K}$, according to (13a).
5: Compute the diffusion kernel $\boldsymbol{P}$, according to (13b).
6: 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 $\boldsymbol{s}_i$
$\triangleright\ f(\boldsymbol{s}_i) = [\,\lambda_1\,\boldsymbol{\psi}_1[i],\ \lambda_2\,\boldsymbol{\psi}_2[i],\ \ldots,\ \lambda_{l_x}\,\boldsymbol{\psi}_{l_x}[i]]^{\mathrm{T}}$.

the mutual information between the leading eigenvector of the AD kernel (which provides reference information about the undesired variable), and the proposed kernel eigenvectors.

V. Illustrative Examples

With the toy examples in this section, we investigate the effect of the proposed bandwidth function on the kernel eigenvectors. We illustrate that the resulting Markov chain is biased to propagate faster along the directions of variation that correspond to the undesired variable, compared to an isotropic Markov chain. As a result, the leading eigenvector of the diffusion kernel parametrizes the desired variable X. The local kernel scales ε_{ij} are computed as described in Section IV-C, by considering 100 nearest neighbors at each point.

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 $L_1 > L_2$. Assuming that the measurement function h is the identity, the strip coordinates correspond to the random variables X and V in our data model. The strip represents the product manifold $\mathcal{X} \times \mathcal{V}$, which is a flat manifold with standard Euclidean metric. The distances d_x and d_v correspond to the coordinate-wise distances on the strip, i.e.,

$$d_1(s_{i1}, s_{j1}) = |s_{i1} - s_{j1}| \tag{36a}$$

$$d_2(s_{i2}, s_{j2}) = |s_{i2} - s_{j2}|. \tag{36b}$$

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 properties we desire for a data-driven kernel are i) the leading eigenvector should parametrize the coordinate 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 isotropic kernel. 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

$$b_1(i,j) = (1 + d_1(s_{i1}, s_{j1}))^2, \tag{37a}$$

$$b_2(i,j) = (1 + d_2(s_{i1}, s_{j2}))^2. \tag{37b}$$

Note that the eigenvalues of the Laplace-Beltrami operator on the strip (with Neumann boundary conditions) can be computed analytically as

$$\mu_{k_1,k_2} = \left(\frac{k_1\pi}{L_1}\right)^2 + \left(\frac{k_2\pi}{L_2}\right)^2, \tag{38}$$

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

$$\rho_{k_1,k_2}(l_1, l_2) = \cos\left(\frac{k_1 l_1 \pi}{L_1}\right)\cos\left(\frac{k_2 l_2 \pi}{L_2}\right). \tag{39}$$

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 always correspond to the two largest eigenvalues. As the ratio L_1/L_2 increases, the more eigenfunctions $\rho_{k_1,0}(l_1)$ appear before $\rho_{0,1}(l_2)$ in the spectrum. In our experiment, we uniformly sampled $N = 2880$ points from a strip with lengths $L_1 = L$ and $L_2 = 0.4\,L$. From (38), and (39), 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.

The first four diffusion map coordinates obtained with the bandwidths in (37) are shown in Figure 1(b) and 1(c). In Figure 1(b), the bandwidth $b_2(i,j)$ shrinks vertical variations. As a result all four eigenvectors parametrize 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. Torus 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 $\mathbb{R}^3$. If the random variables X and V in our data model correspond to the major and minor angle on the torus, and R and r are the major and minor radii, the function h in (1) is given by

$$S = h(X,V) = \begin{bmatrix} (R + r\,\cos(2\pi V))\,\cos(2\pi X) \\ (R + r\,\cos(2\pi V))\,\sin(2\pi X) \\ r\,\sin(2\pi V) \end{bmatrix}, \tag{40}$$

Similarly as in the strip experiment, we assume that the following distance on $\mathcal{V}$ is accessible

$$d_v(v_i, v_j) = \|\boldsymbol{n}_{v_i} - \boldsymbol{n}_{v_j}\|_2, \quad \boldsymbol{n}_v = [\cos(v), \sin(v)]^{\mathrm{T}}. \tag{41}$$

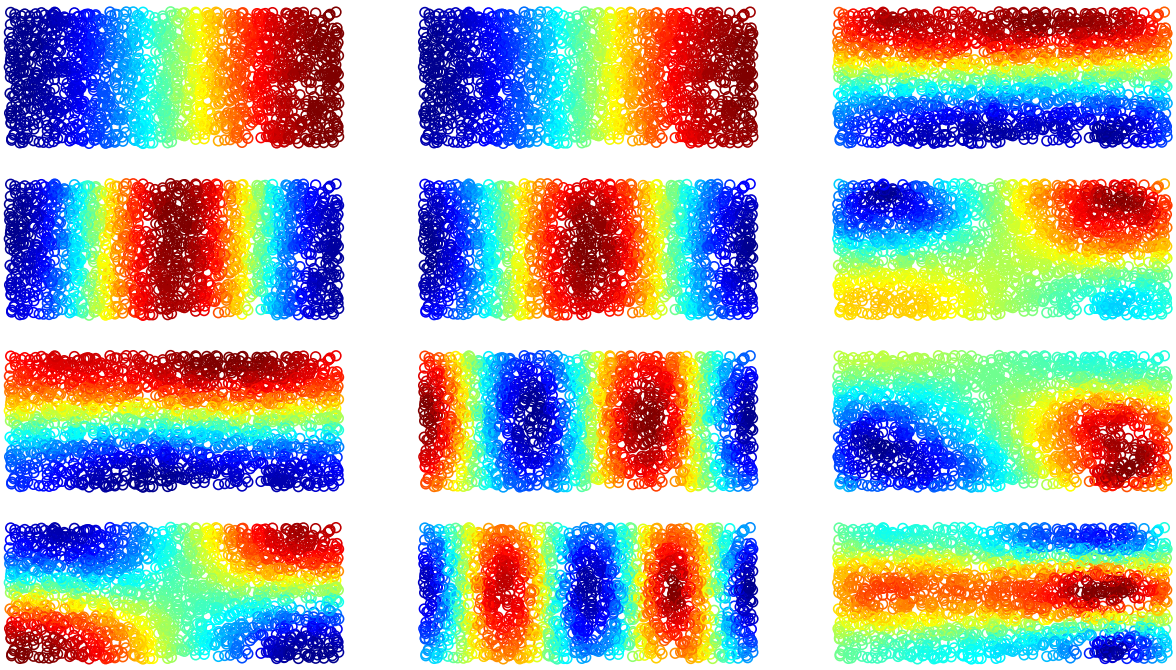


Fig. 1. Points sampled from a strip. The first four diffusion eigenvectors are shown coded in color (top to bottom: 1st to 4th). **(a)** Isotropic kernel. **(b)** Proposed kernel, when the horizontal coordinate is the desired signal X and the vertical coordinate is the noise V. **(c)** Proposed kernel,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

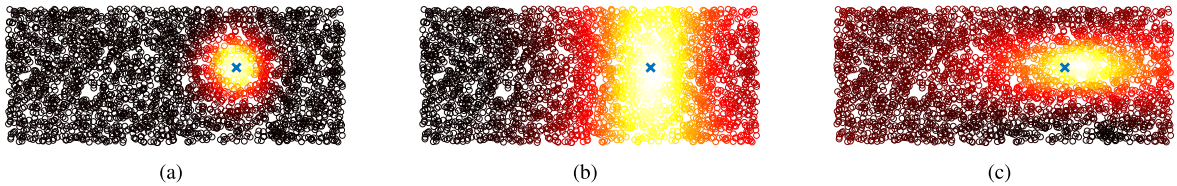


Fig. 2. Heat diffusion on the strip after 5 steps of the Markov chain, starting from the point denoted by $\times$. **(a)** Isotropic kernel; **(b)** Proposed kernel when the horizontal coordinate is the desired signal. The diffusion is then faster along the vertical coordinate; **(c)** Proposed kernel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

If the minor angle is the variable of interest, the kernel is constructed analogously to (41), using $d_x(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The diffusion on the torus resulting from the different kernels is shown in Figure 3. While the standard kernel leads to an isotropic diffusion, the proposed kernels induce a directed diffusion that is faster along one of the angles.

C. Discussion

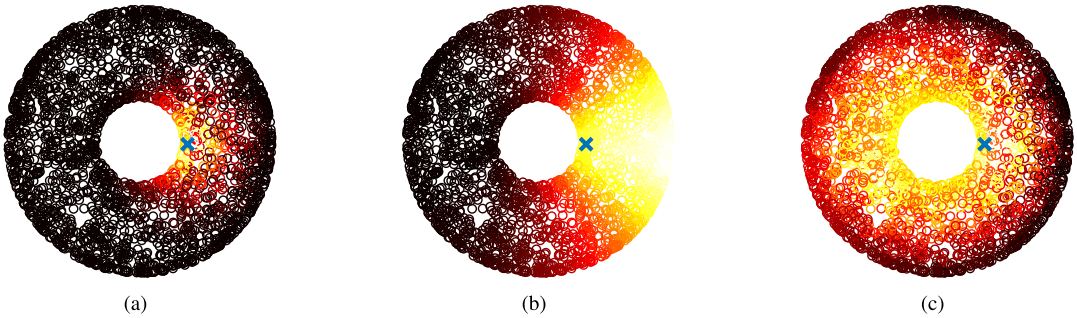
In contrast to the presented toy examples, the metric d_v is typically unobservable in practice. If we consider these toy examples as single-sensor scenarios, then we need to estimate d_v using the principal eigenvector of an isotropic diffusion kernel computed from $\{\boldsymbol{s}_i\}_{i=1}^N$. Clearly, the horizontal coordinate of the strip and the major angle of the torus are the largest scale sources of variability in these examples. Therefore, the proposed variable-bandwidth kernel with estimated d_v leads to a faster diffusion precisely along those directions, as shown in Figure 4.

To quantitatively compare the distance metric d'_s used in the proposed kernel and the Euclidean distance metric d_s used in a standard isotropic kernel, we consider the following local stress values for d_s and d'_s

$$S = \sum_{i \neq j} w_{ij} \; |d_s(\boldsymbol{s}_i, \boldsymbol{s}_j) - d_x(\boldsymbol{x_i}, \boldsymbol{x}_j)| \,, \tag{42a}$$

$$S' = \sum_{i \neq j} w_{ij} \; |d'_s(\boldsymbol{s}_i, \boldsymbol{s}_j) - d_x(\boldsymbol{x_i}, \boldsymbol{x}_j)| \,, \tag{42b}$$

where $w_{ij} = \frac{k_\varepsilon(\boldsymbol{x}_i, \boldsymbol{x}_j)}{K}$, and $K = \sum_{i \neq j} k_\varepsilon(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The Gaussian kernel $k_\varepsilon(\boldsymbol{x}_i, \boldsymbol{x}_i)$ on $\mathcal{X} \times \mathcal{X}$ is computed in the same manner as $k_\varepsilon(\boldsymbol{s}_i, \boldsymbol{s}_i)$ for the measurements (using 100 nearest neighbors for the local bandwidths ε_{ij}, as discussed in Section IV-C). The stress values in (42) quantify the extent to which the distance metrics d_s and d'_s preserve local neighborhoods as defined by d_x. Smaller stress values indicate better neighborhood preservation. The stress values of d_s and d'_s for the



(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.

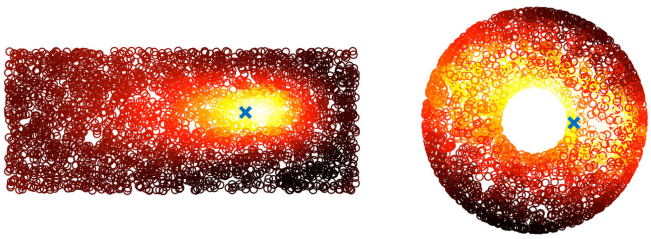


Fig. 4. Heat diffusion on the strip (left) and the torus (right) when the distance metric d_v for the proposed kernel bandwidth is estimated from the first eigenvector of an isotropic diffusion kernel (as discussed in Section IV-B2).

TABLE I
LOCAL STRESS VALUES OF THE EUCLIDEAN DISTANCE d_s (DENOTED BY S) AND THE PROPOSED DISTANCE d'_s (DENOTED BY S')

	strip		torus	
	x-coord.	y-coord	major	minor
S	0.80	2.09	0.61	1.44
S' (oracle)	0.24	0.34	0.39	0.67
S' (estimated)	n/a	0.41	n/a	0.68

strip and the torus (and their different coordinates as the desired source of variability X) are summarized in Table I. Note that S' (oracle) corresponds to experiments where d_v is assumed to be known and S' (estimated) corresponds to experiments where d_v is estimated from the first eigenvector of an isotopic kernel.

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) [22], [39], [40]. We use electrocardiogram (ECG) signals from the PhysioNet collection [41]. 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.

Estimation of fIHR from signals that contain mECG corresponds to a multiple frequency detection problem. A non-linear time-frequency transform for this type of problems, known as the de-shape short-time Fourier transform (dsSTFT), was recently proposed in [42]. The dsSTFT was employed for fIHR estimation in [39], by first estimating the mECG and then subtracting this estimate from the abdominal signal. Note that identification of the mECG is a common first step in various fECG estimation algorithms in the signal processing community (cf. [40] and references therein). In the following, we show that using the proposed kernels, the fIHR can be elegantly obtained without estimating the mECG waveform first. In fact, the fIHR can be directly identified in the dsSTFT of the proposed kernel eigenvectors. Implementation details of the dsSTFT and the procedure to extract an instantaneous heart rate (IHR) from a dsSTFT representation are provided in [39].

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 20 seconds signal excerpt from a given patient, resulting in $N = 1975$ data points per experiment. The following preprocessing 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.

A. Evaluation With a Direct Fetal ECG Reference

In this experiment, we use the *Abdominal and Direct Fetal Electrocardiogram Database (adfecgdb)* from PhysioNet [22], 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. Note that the normal maternal heart rate ranges from 60 to 90 beats-per-minute (bpm), while the normal fetal heart rate ranges from 110 to 150 bpm. Even if in some signals the fetal heart rate is detected, the maternal heart rate is always the dominant spectral line. Our objective is to apply

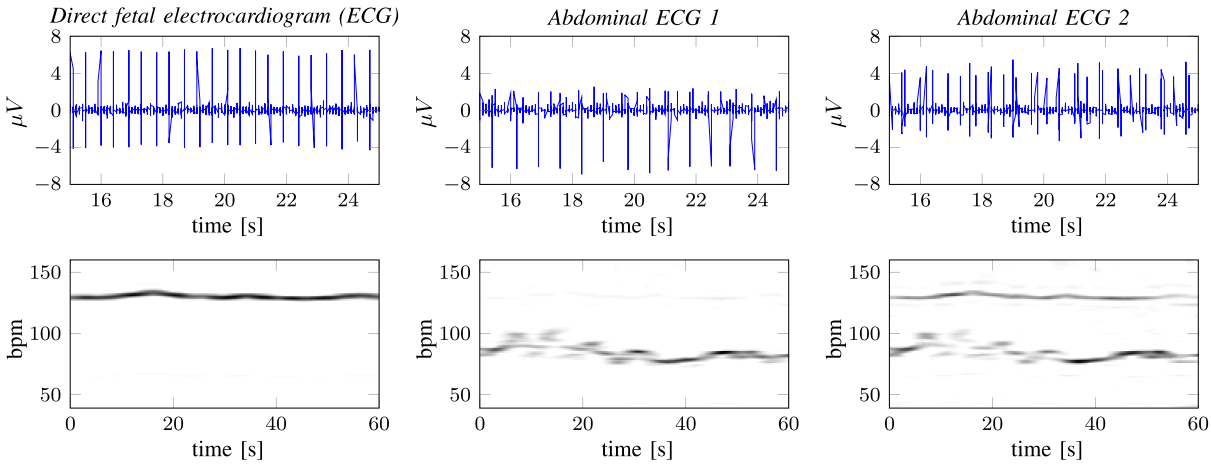


Fig. 5. Example signals from Patient 1 in the *adfecgdb* database. Top: time-domain signals. In order to clearly see the peaks due to the heart beat, we only plot 10 second signal excerpts in the time-domain. Bottom: dsSTFT representation of the complete recordings.

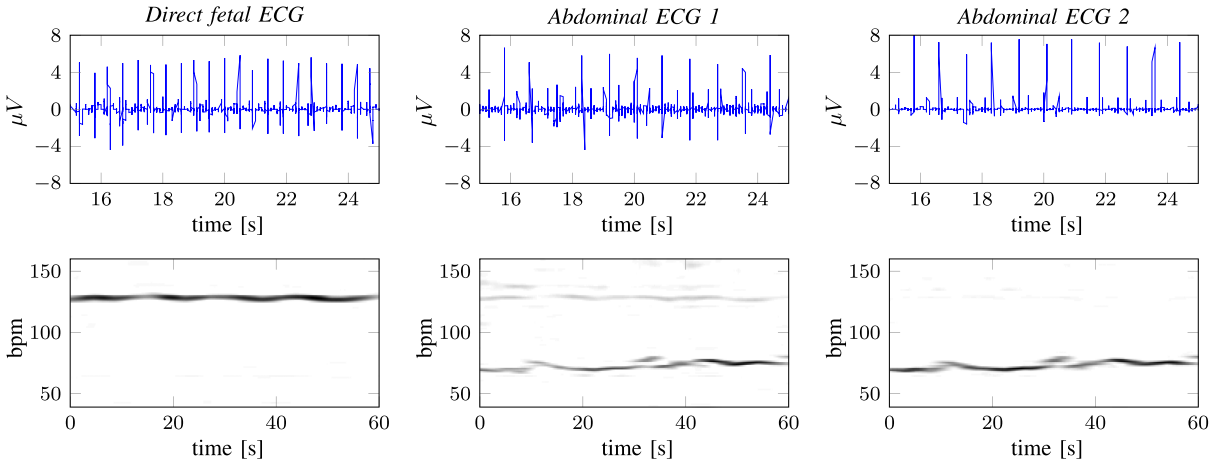


Fig. 6. Example signals from Patient 2 in the *adfecgdb* database. Top: time-domain signals. In order to clearly see the peaks due to the heart beat, we only plot 10 second signal excerpts in the time-domain. Bottom: dsSTFT representation of the complete recordings.

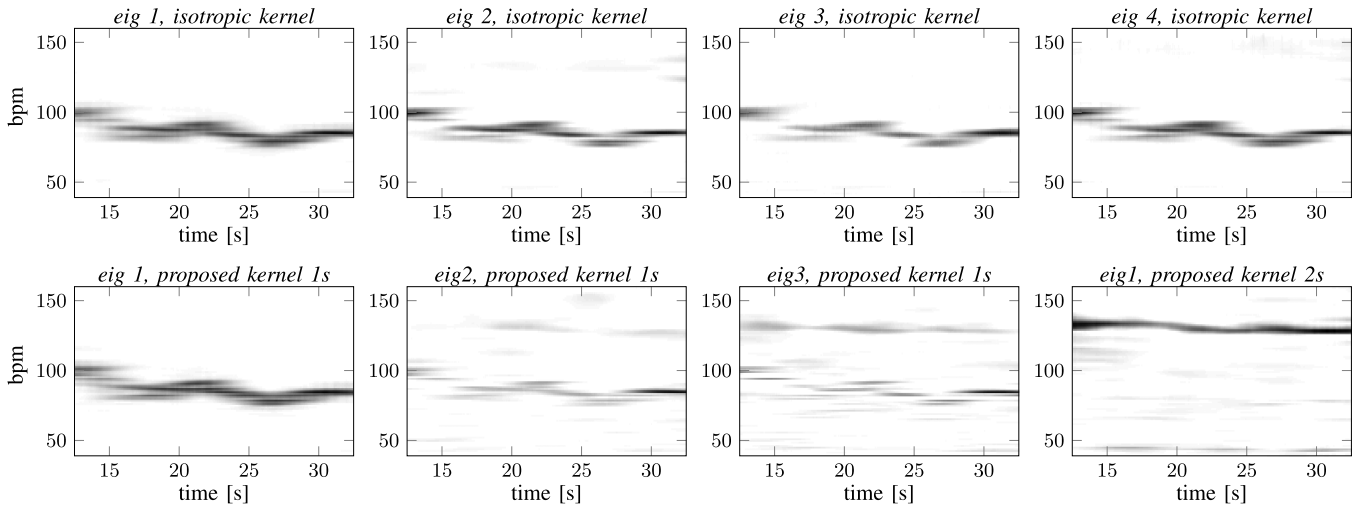


Fig. 7. Eigenvectors from the different diffusion kernels for Patient 1. Top: isotropic kernel. Bottom: from proposed kernels without and with an auxiliary sensor (indicated by *1 s* and *2 s* respectively). The time axis corresponds to the signal excerpt considered in this experiment.

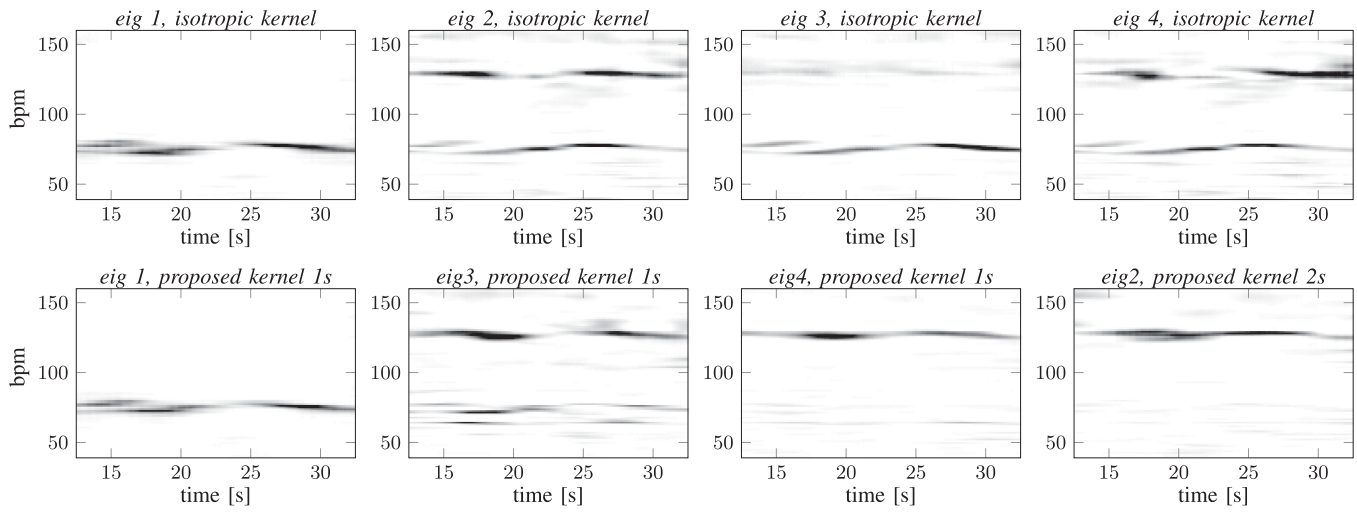
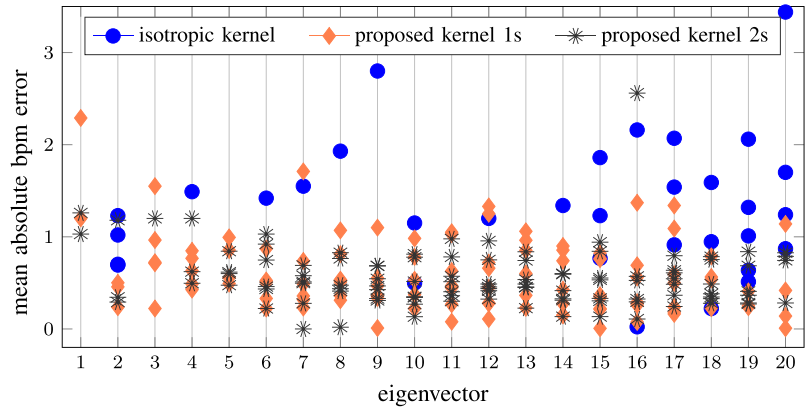


Fig. 8. Eigenvectors from the different diffusion kernels for Patient 2. Top: isotropic kernel. Bottom: from proposed kernels without and with an auxiliary sensor (indicated by *1 s* and *2 s* respectively). The time axis corresponds to the signal excerpt considered in this experiment.



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

Fig. 9. Summary of quantitative results over 9 experiments, without and with an auxiliary sensor (denoted by 1 s and 2 s, respectively). 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.

the proposed kernels and obtain a filtered signal whose dsSTFT recovers the fetal heart rate, while suppressing or completely removing the maternal one.

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, with one sensor, is denoted by *1 s*. 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, with two sensors, is denoted by *2 s*. In both cases, the early eigenvectors of the proposed kernels recover the fIHR. In particular, in the *2 s* 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

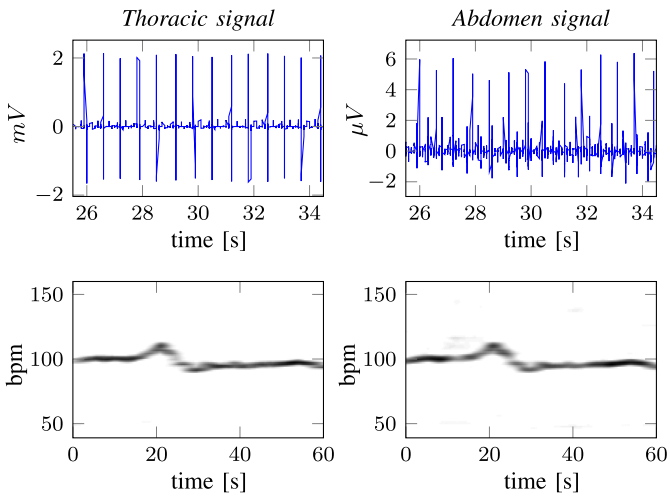


Fig. 10. Signal examples from the *nifecgdb* dataset. Top row: time-domain waveforms. Bottom row: dsSTFTs representations.

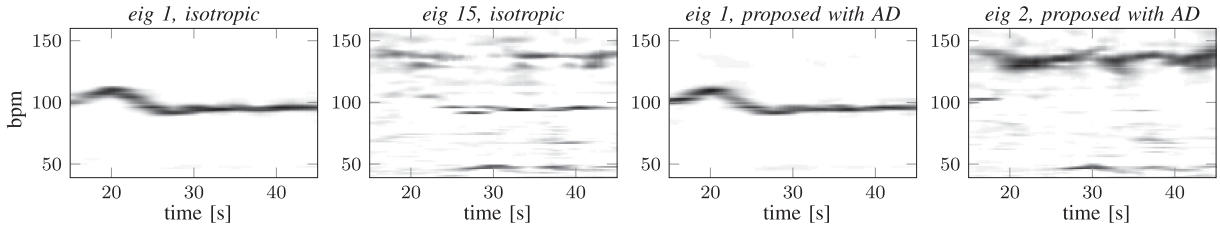


Fig. 11. Results from the *nifecgdb* dataset. The AD distance extracted with the thoracic auxiliary sensor provides auxiliary information about the maternal instantaneous heart rate (mIHR), allowing for its complete suppression in eigenvector 2 (rightmost figure). The isotropic kernel does not recover the fECG in early eigenvectors.

ECG, shown in Figure 8. However, application of our algorithm ensures that the fECG is the dominant spectral line.

For a quantitative evaluation, we extracted IHR curves (using the method in [39]) from each of the first 20 eigenvectors in 9 different experiments, using signal excerpts from four patients. Similarly, ground truth fIHR was extracted from the dsSTFT representation from the direct fECG signal. From the total of 180 analyzed eigenvectors for each kernel (across all experiments), we only kept the eigenvectors that successfully captured the fIHR. The scatter plot in Figure 9 shows the mean error in beats-per-minute (bpm) for each of these eigenvectors. The percentage from the total of 180 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% from the total of 90 eigenvectors. Importantly, these tend to be higher in the spectrum than the eigenvectors of an isotropic kernel. We once again emphasize that multiple eigenvectors of the diffusion kernels can encode the same direction of variability in the data [34]. 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 *2 s* scenario is 0.5 bpm. The error in the *1 s* scenario is 0.6 bpm, while the isotropic kernel is inferior with an error of 1.2 bpm.

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.

B. 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 distance 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. We used a 30 seconds signal segment and the same lag map parameters as in the previous experiments. It can be seen that the second eigenvector recovers the fIHR, while removing the mIHR from the spectrum. In this experiment, we only present a qualitative result, as without a direct fECG we were unable to perform quantitative analysis as in Section VI-A, since we do not have the ground truth.

VII. Conclusion

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.

Acknowledgment

We thank the reviewers and the associate editor for their helpful comments and constructive suggestions. This article reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

References

[1] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*. Boston, MA, USA: Springer, 2010.
[2] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI, USA: Amer. Math. Soc., 1997.

[3] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007.
[4] 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.
[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] P. H. Berard, *Spectral Geometry: Direct and Inverse Problems*, A. Dold and B. Eckmann, Eds. Berlin, Germany: Springer-Verlag, 1986.
[7] 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, 2008.
[8] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, pp. 5–30, 2006.
[9] S. Lafon, "Diffusion maps and geometric harmonics," Ph.D. dissertation, Yale University, New Haven, CT, USA, 2004.
[10] B. Nadler, S. Lafon, R. Coifman, and I. G. Kevrekidis, "Diffusion maps—A probabilistic interpretation for spectral embedding and clustering algorithms," in *Lecture Notes in Computational Science and Enginee.* Berlin, Germany: Springer-Verlag, 2008, ch. 10, pp. 238–260.
[11] B. Nadler and M. Galun, "Fundamental limitations of spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1017–1024.
[12] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," *Adv. Neural Inf. Process. Syst.*, 2004, pp. 1601–1608.
[13] T. Berry and J. Harlim, "Variable bandwidth diffusion kernels," *Appl. Comput. Harmon. Anal.*, vol. 40, no. 1, pp. 68–96, 2016.
[14] 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.*, vol. 15, pp. 1327–1351, 2016.
[15] D. Giannakis and A. J. Majda, "Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability," *Proc. Nat. Acad. Sci.*, vol. 109, no. 7, pp. 2222–2227, 2012.
[16] D. Giannakis, "Dynamics-adapted cone kernels," *SIAM J. Appl. Dyn. Syst.*, vol. 14, no. 2, pp. 556–608, 2015.
[17] V. Papyan and R. Talmon, "Multimodal latent variable analysis," *Signal Process.*, vol. 142, pp. 178–187, 2018.
[18] D. Dov, R. Talmon, and I. Cohen, "Kernel-based sensor fusion with application to audio-visual voice activity detection," *IEEE Trans. Signal Process.*, vol. 64, no. 24, pp. 6406–6416, Dec. 2016.
[19] A. Singer and R. R. Coifman, "Non-linear independent component analysis with diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 25, pp. 226–239, 2008.
[20] V. Guillemin and A. Pollack, *Differential Topology*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1974.
[21] E. Kreyszig, *Introductory Functional Analysis With Applications*. Hoboken, NJ, USA: Wiley, 1978.
[22] 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.
[23] H.-T. Wu, R. Talmon, and Y.-L. Lo, "Assess sleep stage by modern signal processing techniques," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1159–1168, Apr. 2015.
[24] T. Shnitzer, M. Rapaport, N. Cohen, N. Yarovinsky, R. Talmon, and J. Aharon-Peretz, "Alternating diffusion maps for dementia severity assessment," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 831–835.
[25] Y. Terada and U. von Luxburg, "Local ordinal embedding," in *Proc. Int. Conf. Mach. Learn.*, 2014, vol. 32, pp. 847–855.
[26] R. Talmon, I. Cohen, S. Gannot, and R. R. Coifman, "Diffusion maps for signal processing: A deeper look at manifold-learning techniques based on kernels and graphs," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 75–86, Jul. 2013.
[27] M. Belkin, "Problems of learning on manifolds," Ph.D. dissertation, The University of Chicago, Chicago, IL, USA, 2003.
[28] D. Ting, L. Huang, and M. I. Jordan, "An analysis of the convergence of Graph Laplacians," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 1079–1086.
[29] R. R. Coifman *et al.*, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proc. Nat. Acad. Sci.*, vol. 102, no. 21, pp. 7426–7431, 2005.
[30] R. Talmon, S. Mallat, H. Zaveri, and R. R. Coifman, "Manifold learning for latent variable inference in dynamical systems," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 3843–3856, Aug. 1, 2015.
[31] 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.
[32] O. Yair and R. Talmon, "Local canonical correlation analysis for nonlinear common variables discovery," *IEEE Trans. Signal Process.*, vol. 65, no. 5, pp. 1101–1115, Mar. 1, 2017.
[33] A. Holiday, M. Kooshkbaghi, J. M. Bello-Rivas, C. W. Gear, A. Zagaris, and I. G. Kevrekidis, "Manifold learning for parameter reduction," *J. Comput. Phys.*, vol. 392, pp. 419–431, 2019.
[34] 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.
[35] 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.
[36] 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.
[37] 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.
[38] R. Talmon and H.-T. Wu, "Latent common manifold learning with alternating diffusion: Analysis and applications," *Appl. Comput. Harmon. Anal.*, vol. 47, no. 3, pp. 848–892, 2018.
[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. Statist.*, vol. 3, pp. 1–26, 2017.
[40] G. J. J. Warmerdam, R. Vullings, L. Schmitt, J. O. E. H. Van Laar, and J. W. M. Bergmans, "Hierarchical probabilistic framework for fetal R-Peak detection, Using ECG waveform and heart rate information," *IEEE Trans. Signal Process.*, vol. 66, no. 16, pp. 4388–4397, Aug. 15, 2018.
[41] A. L. Goldberger *et al.*, "PhysioBank, physiotoolkit, and physionet : Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
[42] 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.

Maja Taseska received the B.Sc. degree in electrical engineering from the Jacobs University, Bremen, Germany, in 2010, the M.Sc. degree (*summa cum laude*) from the Friedrich-Alexander-University, Erlangen, Germany, and the Ph.D. degree (*summa cum laude*) from the International Audio Laboratories Erlangen, a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer II, in 2017. From 2018 to 2019, she was an FWO Postdoctoral Fellow with KU Leuven, Belgium (no. 12X6719N).

Toon van Waterschoot (S'04–M'12) received the M.Sc. and Ph.D. degrees in electrical engineering in 2001 and 2009, respectively, from Katholieke Universiteit Leuven, Leuven, Belgium, where he is currently an Associate Professor and Consolidator Grantee of the European Research Council. He has previously also held teaching and research positions with Delft University of Technology, Delft, The Netherlands, and the University of Lugano, Switzerland. His research interests are in signal processing, machine learning, and numerical optimization, applied to acoustic signal enhancement, acoustic modeling, audio analysis, and audio reproduction.

Dr. Waterschoot has been serving as an Associate Editor for the *Journal of the Audio Engineering Society* and for the *EURASIP Journal on Audio, Music, and Speech Processing*, and as a Guest Editor for *Elsevier Signal Processing*. He is a Director of the European Association for Signal Processing (EURASIP), a Member of the IEEE Audio and Acoustic Signal Processing Technical Committee, a Member of the EURASIP Special Area Team on Acoustic, Speech and Music Signal Processing, and a Founding Member of the EAA Technical Committee in Audio Signal Processing. He was the General Chair of the 60th AES International Conference in Leuven, Belgium, in 2016, and has been serving on the Organizing Committee of the European Conference on Computational Optimization (EUCCO 2016), the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2017), and the 28th European Signal Processing Conference (EUSIPCO 2020). He is a member of EURASIP, ASA, and AES.

Emanuël A. P. Habets (S'02–M'07–SM'11) received the B.Sc. degree in electrical engineering from Hogeschool Limburg, The Netherlands, in 1999, and the M.Sc. and Ph.D. degrees in electrical engineering from the Technische Universiteit Eindhoven, Eindhoven, The Netherlands, in 2002 and 2007, respectively. He is currently an Associate Professor with the International Audio Laboratories Erlangen (a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer IIS), and Head of the Spatial Audio Research Group, Fraunhofer IIS, Germany.

From 2007 to 2009, he was a Postdoctoral Fellow with the Technion—Israel Institute of Technology and at the Bar-Ilan University, Israel. From 2009 to 2010, he was a Research Fellow with the Communication and Signal Processing Group, Imperial College London, London, U.K. His research activities center around audio and acoustic signal processing, and include spatial audio signal processing, spatial sound recording and reproduction, speech enhancement (dereverberation, noise reduction, echo reduction), and sound localization and tracking.

Dr. Habets was a member of the organization committee of the 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC) in Eindhoven, The Netherlands, a General Co-Chair of the 2013 International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in New Paltz, New York, and General Co-Chair of the 2014 International Conference on Spatial Audio (ICSA) in Erlangen, Germany. He was a member of the IEEE Signal Processing Society Standing Committee on Industry Digital Signal Processing Technology (2013–2015), a Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING and the *EURASIP Journal on Advances in Signal Processing*, an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (2013–2017), and Editor-in-Chief of the *EURASIP Journal on Audio, Speech, and Music Processing* (2016–2018). Currently, he is a member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, a member of the EURASIP Technical Activities Board, and Chair of the EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing. He is the recipient, with S. Gannot and I. Cohen, of the 2014 IEEE SIGNAL PROCESSING LETTERS Best Paper Award.

Ronen Talmon received the B.A. degree (*cum laude*) in mathematics and computer science from the Open University in 2005, and the Ph.D. degree in electrical engineering from the Technion—Israel Institute of Technology, Haifa, Israel, in 2011. He is currently an Associate Professor of electrical engineering with the Technion—Israel Institute of Technology. From 2000 to 2005, he was a Software Developer and Researcher with a technological unit of the Israeli Defense Forces. From 2005 to 2011, he was a Teaching Assistant with the Department of Electrical Engineering, Technion. From 2011 to 2013, he was a Gibbs Assistant Professor with the Mathematics Department, Yale University, New Haven, CT, USA. In 2014, he joined the Department of Electrical Engineering of the Technion. His research interests are geometry-based data analysis and modeling, applied harmonic analysis, diffusion geometry, biomedical signal processing, audio and speech signal processing, and computational neuroscience.

Dr. Talmon is the recipient of the Irwin and Joan Jacobs Fellowship, the Andrew and Erna Fince Viterbi Fellowship, and the Horev Fellowship.