Signal Processing 142 (2018) 178–187

Contents lists available at ScienceDirect

Signal Processing

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

Multimodal latent variable analysis

Vardan Papyan[a,*], Ronen Talmon[b]

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

ARTICLE INFO

Article history:
Received 21 December 2016
Revised 17 July 2017
Accepted 19 July 2017
Available online 20 July 2017

Keywords:
Manifold learning
Diffusion maps
Sensor fusion
Alternating diffusion
Fetal ECG

ABSTRACT

Consider a set of multiple, multimodal sensors capturing a complex system or a physical phenomenon of interest. Our primary goal is to distinguish the underlying sources of variability manifested in the measured data. The first step in our analysis is to find the common source of variability present in all sensor measurements. We base our work on a recent paper, which tackles this problem with alternating diffusion (AD). In this work, we suggest to further the analysis by extracting the sensor-specific variables in addition to the common source. We propose an algorithm, which we analyze theoretically, and then demonstrate on three different applications: a synthetic example, a toy problem, and the task of fetal ECG extraction.

1. Introduction

The analysis of a physical phenomenon or some complex system at hand can often be made easier through the use of several sensors instead of a single complex one. The hope is that each of the sensors captures a different part of the convoluted system, while the fusion of all the information captures the global picture. This line of thinking has led to the abundance of multimodal and multi-sensory data in recent years and to an increased demand for algorithms that enable its processing and analysis [1]. A prime example for the above is medical diagnosis based on collected bedside data, where one monitors a patient using various basic sensors, such as heart rate, pulse, blood pressure and oxygen level just to name a few, and attempts to diagnose the complex system at hand, that is the patient state, using the collected data.

Elaborate systems, such as the one mentioned above, are usually governed by many sources of variability. A central problem is then the analysis of latent sources, given measurements originating from several sensors of various types. Naturally, analyzing the measured data in terms of its underlying sources of variability requires their extraction. Unfortunately, driving sources are often hidden in nonlinear unknown manners, thereby posing a true challenge to the analysis and to the extraction.

In order to facilitate the extraction of the different sources of variability, we divide them into two conceptual categories: (i) sources of variability common to all sensors; and (ii) variables unique to a specific sensor. In our work, we focus on a two step implementation where we first reveal the common variable. Once it is found, we extract the remaining sources of variability, i.e, the sensor-specific ones. Intuitively, our approach marginalizes the common variable, which is found in the first step, and then continues to extract the sources of variability left in the filtered data. This simplifies our task, since we do not attempt to extract all the sources manifested in the data at once.

In this paper, we use an unsupervised manifold learning approach to address the problem. Various manifold learning algorithms were proposed in the literature over the years, [2–4]. The reader is referred to [5] for a thorough review of existing approaches and their advantages. However, most of these classical methods assume that the data is captured by a single sensor, rather than in the multimodal multi-sensory setting we consider here. In this work, we focus on a particular paradigm – Diffusion Geometry, as presented in [6,7]. Within this framework, the alternating diffusion(AD) algorithm was recently proposed in [8,9] for the purpose of extracting the source of variability common to multiple sensors. AD follows a recent line of papers that propose to use multiplications and manipulations of kernels for the purpose of fusing data from different sensors, e.g., [10–13]. Similarly to recently presented nonlinear methods, e.g., [14,15], AD is shown to reveal only the common components among all processed sensors. Successful applications of AD to real measured data were demonstrated, e.g., in [16] for the task of sleep stage identification. Herein, we rely on AD and aim to extend it by further analyzing the measurements and finding the sensor-specific variables. Our main motivation is that in some applications the sensor

* Corresponding author.
E-mail addresses: vardanp@campus.technion.ac.il, vardanp91@gmail.com (V. Papyan).

http://dx.doi.org/10.1016/j.sigpro.2017.07.016
0165-1684/

specific variables are far more important than the common variable. Indeed, we show one real-life example of such an application – fetal Electrocardiography (ECG) extraction.

Our main contribution in this work is a new algorithm, attempting to recover all the sources of variability manifested in a set of multi-sensory multimodal measurements. This operates by first extracting the common variable and then leveraging it in order to extract the remaining sensor-specific variables. We justify our proposed scheme theoretically, showing that it is guaranteed to find the underlying parametrizations under certain prescribed conditions. In addition, we demonstrate our method on a synthetic example, a toy problem and a real-life application.

The strength of our algorithm is stemming from first extracting the common variable, a task which is easier to handle since the common variable is measured by both sensors, and only then trying to extract the sensor-specific variables. Other methods, such as those mentioned above, do not implement such a two-stage procedure. This is also the weakness of our approach, since it relies on the successful extraction of the common variable. Other holistic methods, which aim to extract all the variables jointly, might not have this drawback.

Herein, we focus on applications in which the readings are from two sensors. However, our algorithm can be readily extended to a multi-sensor scenario due to the capability of AD to extract the common variable, even when several sensors are involved. In this case, one would first extract the common variable and then proceed to the sensor-specific variables by operating on each of the sensors independently. Recent works [17–19] extend the problem definition for more than two sensors by explicitly searching for not just the common and the sensor-specific variables, but also variables common to every possible subset of sensors. One could envision an extension of our algorithm, where AD is applied to every subset of the sensors, enabling the extraction of the corresponding sensor-specific variables. However, this extension is beyond the scope of this work.

In recent years many approaches were proposed for the analysis of multi-modal multi-sensory data. For instance, the works of [20,21] suggested to learn non-parametric mapping functions that transform different modalities into a shared latent space. The work of [22] considered the problem where one is given multiple unlabeled views of some data and the task is to learn some useful representations, using deep learning, that could be used in test stage when only one view is available. In [23], the authors studied the problem of semantic retrieval, where documents from different modalities need to be ranked according to their relevance to a certain query. All of these methods tackle different problems that arise in the context of multi-modal multi-sensor data analysis. However, none of these focus on the setting we consider in this work – the extraction of the sensor-specific variable.

This paper is organized as follows. In Section 2 we introduce formally the problem we address, and in Section 3 we review the diffusion maps and AD algorithms. In Section 4 we present the proposed method and in Section 5 we analyze it theoretically. In Section 6 we test our method on a synthetic example, a toy problem and a real-life application – the extraction of fetal ECG. We conclude this paper in Section 7.

2. Problem formulation

Consider three latent random variables X, Y and Z in $\mathbb{R}^{d_x}$, $\mathbb{R}^{d_y}$ and $\mathbb{R}^{d_z}$, respectively, which are jointly distributed according to some probability density function(PDF) denoted by $P(X, Y, Z)$. Following the work in [8], we assume that the variables Y and Z are independent given X, i.e., the joint PDF can be written as follows:

$$P(X,Y,Z) = P(Y|X)P(Z|X)P(X), \tag{1}$$

where $P(X)$ is the marginal PDF of X, and $P(Y|X)$ and $P(Z|X)$ are the conditional PDFs of Y and Z given X, respectively. When measuring a system of interest, a measurement instance is defined by the triplet $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)$, which is a realization sampled from $P(X, Y, Z)$. We do not have access to the latent variables; instead, we have two sensors observing the system at hand through two unknown observation functions given by $g(\mathbf{x}_i, \mathbf{y}_i)$ and $h(\mathbf{x}_i, \mathbf{z}_i)$. We assume g and h are smooth and locally invertible bilipschitz functions. Let $\{\mathbf{s}_i^{(1)}\}_{i=1}^N$ and $\{\mathbf{s}_i^{(2)}\}_{i=1}^N$ denote two sets of N measurement samples, taken simultaneously from the two sensors, such that $\mathbf{s}_i^{(1)} = g(\mathbf{x}_i, \mathbf{y}_i) \in \mathbb{R}^{d_1}$ and $\mathbf{s}_i^{(2)} = h(\mathbf{x}_i, \mathbf{z}_i) \in \mathbb{R}^{d_2}$, where $\{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)\}_{i=1}^N$ are N realizations of the system's hidden variables. In other words, we have hidden realizations $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)$ of three underlying variables and two sensor observations $\mathbf{s}_i^{(1)}$ and $\mathbf{s}_i^{(2)}$; $\mathbf{x}_i$ is the common latent variable between the two observations, whereas $\mathbf{y}_i$ and $\mathbf{z}_i$ are two sensor-specific variables.

Given the two sets of measurement samples, the work in [8] showed that a method based on AD operators extracts a parameterization of the common variable X. In this work, we aim to further the analysis and extract a parametrization of the variables Y and Z as well. Such a complementing capability enables us to fully parameterize all the hidden variables underlying the measurements of the system of interest.

Although the analysis and methods used in this paper will be carried out from a different standpoint, the factorization in (1) can be used to explain the main concept. Intuitively, the extraction of the common variable X in [8] can be viewed as a marginalization operator applied to the joint probability $P(X, Y, Z)$ obtaining $P(X)$. In this work, we devise another operator which uses $P(X)$ to construct the conditional probabilities $P(Y|X)$ and $P(Z|X)$. Then, given $P(Y|X)$ and $P(Z|X)$, it marginalizes the variable X and obtains a parametrization of the sensor-specific variables Y and Z.

3. Preliminaries

3.1. Diffusion maps

Diffusion maps [6,7] is a data-driven nonlinear dimensionality reduction algorithm. Given a set of N measurements $\{\mathbf{u}_i\}_{i=1}^N$, the method constructs an affinity matrix $\mathbf{W}$ of size $N \times N$, whose (i, j)th entry is given by

$$W_{i,j} = \exp\left(-\frac{\left\|\mathbf{u}_i - \mathbf{u}_j\right\|^2}{\epsilon}\right), \quad \forall i, j = 1, \ldots, N. \tag{2}$$

Intuitively, $\mathbf{W}$ can be interpreted as a weight matrix of a graph with N vertices, where the coefficient $\epsilon > 0$ dictates the sparsity of the edges. If ϵ is small, most edges have a negligible, close to zero weight and the graph is effectively sparse, whereas if ϵ is large, most edges are assigned with non negligible weights and the graph is dense. The constant ϵ is usually chosen according to the data at hand, and in this work we set it using the method suggested in [8]. Therein, the constant was chosen to be $\epsilon = \sqrt{\epsilon_i \epsilon_j}$, where ϵ_i a scaling constant corresponding to the ith vertex. In particular, ϵ_i is chosen to be the mean squared distance from the ith vertex to its k nearest neighbors.

The next step is to normalize the affinity matrix $\mathbf{W}$, which results in the matrix $\mathbf{K}$. Various normalization procedures have been suggested in the literature [24,25], each having a different interpretation when analyzed theoretically. In this work, $\mathbf{K}$ is constructed by dividing each column of $\mathbf{W}$ by its sum, yielding a column-stochastic matrix. As a result, $\mathbf{K}$ can be viewed as a transition probability matrix of a Markov chain on the graph. An example for a different approach using such a construction is spectral clustering [26], where a similar kernel normalization is used. Specifically, di-

viding each row by its sum results in a stochastic matrix, whose first eigenvector is the solution of the normalized cut problem [26].

Once the affinity matrix is constructed and normalized, a d-dimensional embedding $\{\hat{\mathbf{u}}_i\}_{i=1}^N$ is formed according to the following nonlinear map:

$$f\hat{\mathbf{u}}_i = \left[\lambda_1^m \phi_1^i, \ldots, \lambda_d^m \phi_d^i\right]^T, \tag{3}$$

where $\boldsymbol{\phi}_j$ is the jth left eigenvector of the matrix $\mathbf{K}$ and ϕ_j^i is its ith entry, λ_j^m is the jth eigenvalue (when the eigenvalues are denoted in descending order) raised to the power of m, and $m > 0$ is a constant. Typically, d is set to be much smaller than $\min(d_1, d_2)$, thereby attaining dimensionality reduction. In addition to providing compact representation, this nonlinear map attempts to reveal the essence of the data in few dimensions, accurately representing their underlying intrinsic variables. In the context of diffusion maps, special attention is given to the Euclidean distance between the embedded samples $\hat{\mathbf{u}}_i$. Specifically, the Euclidean distance between the embedded samples $\hat{\mathbf{u}}_i$ approximates the Euclidean distance between the corresponding columns of $\mathbf{K}^m$. This distance is termed the *diffusion distance*, since it takes into account transition probabilities on the constructed graph consisting of m Markov chain steps. We note that diffusion distance plays a large role in the algorithm presented in this paper. For more details, as well as the motivation behind this particular dimensionality reduction method, we refer the reader to [6].

3.2. Alternating diffusion

Given two sets of measurement samples originating from two sensors, i.e., $\{\mathbf{s}_i^{(1)}\}_{i=1}^N$ and $\{\mathbf{s}_i^{(2)}\}_{i=1}^N$, the first step in the AD algorithm is constructing two pairwise affinity matrices, $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ based on the Gaussian kernel

$$W_{i,j}^{(1)} = \exp\left(-\left\|\mathbf{s}_i^{(1)} - \mathbf{s}_j^{(1)}\right\|^2 / \epsilon^{(1)}\right) \tag{4}$$

$$W_{i,j}^{(2)} = \exp\left(-\left\|\mathbf{s}_i^{(2)} - \mathbf{s}_j^{(2)}\right\|^2 / \epsilon^{(2)}\right). \tag{5}$$

The constants $\epsilon^{(1)}$ and $\epsilon^{(2)}$ have a similar interpretation to the one presented in Section 3.1. The algorithm proceeds by normalizing $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ to be column-stochastic, yielding two matrices $\mathbf{K}^{(1)}$ and $\mathbf{K}^{(2)}$, where the sum of each of their columns equals one. As a result, each stochastic matrix can be interpreted as a transition probability matrix of a Markov chain on a graph whose vertices are the samples (as described in Section 3.1). In other words, the (i, j)th entry in $\mathbf{K}^{(1)}$ or in $\mathbf{K}^{(2)}$ represents the probability of transition to the ith vertex from the jth vertex in the graph. Importantly, by construction (4), $\mathbf{K}^{(1)}$ describes a Markov chain that jumps in high probability from the jth vertex to the ith vertex if the underlying values of both X and Y are similar (namely, $\mathbf{x}_i$ is similar to $\mathbf{x}_j$ and $\mathbf{y}_i$ is similar to $\mathbf{y}_j$). Analogously, by construction (5), $\mathbf{K}^{(2)}$ describes a Markov chain that jumps in high probability from the jth vertex to the ith vertex if the underlying realizations of both X and Z are similar.

Given the normalized $\mathbf{K}^{(1)}$ and $\mathbf{K}^{(2)}$, an AD kernel is then defined by

$$\mathbf{K} = \mathbf{K}^{(2)}\mathbf{K}^{(1)}. \tag{6}$$

This corresponds to a transition probability matrix $\mathbf{K}$ consisting of two consecutive, *alternating* steps – the first step is employed according to $\mathbf{K}^{(1)}$ and second according to $\mathbf{K}^{(2)}$. Next, by raising $\mathbf{K}$ to the power of m, we obtain a transition matrix $\mathbf{K}^m$ that corresponds to $2m$ steps, where the odd steps correspond to $\mathbf{K}^{(1)}$ and the even steps correspond to $\mathbf{K}^{(2)}$. Consequently, the odd steps jump (in high probability) to a vertex where both X and Y values are similar, while the even steps jump (in high probability) to a vertex where both X and Z are similar. As a result, after many (odd and even) steps, we maintain similarity only to the X value whereas the Y and Z may vary significantly.

In order to obtain an affinity matrix in terms of the common variable X between pairs of samples $(\mathbf{s}_i^{(1)}, \mathbf{s}_i^{(2)})$ and $(\mathbf{s}_j^{(1)}, \mathbf{s}_j^{(2)})$ (recalling that each pair of samples shares the same X value according to the model assumptions presented in Section 2), the method computes the ℓ_2 distance between the corresponding columns in the matrix $\mathbf{K}^m$. In [8] a rigorous analysis is provided justifying this statement. Moreover, it was suggested to use a refinement step consisting of an additional diffusion maps application where the columns of $\mathbf{K}^m$ are the new graph vertices, resulting in a low dimensional embedding as defined in (3). In the AD setting, since the underlying variable of the affinity matrix $\mathbf{K}^m$ is X, we shall denote the resulting embedding by $\hat{\mathbf{x}}$ instead of the general notation of $\hat{\mathbf{u}}$, which was used in (3).

4. Proposed method

The first step towards a full parametrization of all the latent variables underlying the measurements is finding the common latent variable X, as previously suggested, using the AD algorithm. Once the common variable is extracted, we proceed to analyzing the measurements from the first and second sensors separately. Hereafter, for the sake of brevity, we will focus on the analysis of the first sensor only, while the analysis of the second sensor is analogous. For each sample $\mathbf{s}_i^{(1)}$, let $\mathcal{N}_i^{(1)}$ be a neighborhood of samples consisting of samples j with a similar common variable. Formally, define

$$\mathcal{N}_i^{(1)} = \left\{j \mid \|\hat{\mathbf{x}}_j - \hat{\mathbf{x}}_i\|_2 < \eta_i\right\}, \tag{7}$$

where $\eta_i > 0$ is a small tunable threshold. In practice, instead of fixing a threshold η_i for every signal, we choose all the neighborhoods $\mathcal{N}_i$ to be of the same size q. In other words, the η_i are fixed implicitly such that the size of each neighborhood $\mathcal{N}_i$ is equal to q. The key point in defining these neighborhoods relies on the assumption that the AD algorithm is able to successfully recover the common variable X and to suppress the sensor specific variable Y. That is, the measurements in these neighborhoods, i.e., $\left\{\mathbf{s}_j^{(1)} | j \in \mathcal{N}_i^{(1)}\right\}$, share equal, or close, values of X. As a result, the only remaining variability in such neighborhoods of samples stems from variations in Y.

For each such neighborhood, we propose to compute its sample mean

$$\boldsymbol{\mu}_i^{(1)} = \frac{1}{\left|\mathcal{N}_i^{(1)}\right|} \sum_{j \in \mathcal{N}_i^{(1)}} \mathbf{s}_j^{(1)}, \tag{8}$$

and its sample Covariance

$$\mathbf{C}_i^{(1)} = \frac{1}{\left|\mathcal{N}_i^{(1)}\right|} \sum_{j \in \mathcal{N}_i^{(1)}} (\mathbf{s}_j^{(1)} - \boldsymbol{\mu}_i^{(1)})(\mathbf{s}_j^{(1)} - \boldsymbol{\mu}_i^{(1)})^T \tag{9}$$

both in the domain of the measurements $\{\mathbf{s}_j^{(1)}\}$. Thus, $(\boldsymbol{\mu}_i^{(1)}, \mathbf{C}_i^{(1)})$ can be seen as a Gaussian representation of the local variability of the sensor-specific variable Y around every sample, and hence, in light of the discussion above, we have a local representation of Y.

In order to get a global parametrization, we compare the local neighborhoods by means of "registration" of point clouds or Gaussian distributions. Consider the following affinity kernel

$$\widetilde{W}_{i,j}^{(1)} = \exp\left\{-\frac{1}{\epsilon}\left((\mathbf{s}_i^{(1)} - \boldsymbol{\mu}_i^{(1)}) - (\mathbf{s}_j^{(1)} - \boldsymbol{\mu}_j^{(1)})\right)^T \left(\mathbf{C}_i^{(1)^\dagger} + \mathbf{C}_j^{(1)^\dagger}\right) \left((\mathbf{s}_i^{(1)} - \boldsymbol{\mu}_i^{(1)}) - (\mathbf{s}_j^{(1)} - \boldsymbol{\mu}_j^{(1)})\right)\right\}. \tag{10}$$

We have denoted by $\mathbf{C}_i^{(1)^\dagger}$ the Moore–Penrose pseudo inverse, which is employed since the rank of the Covariance matrix is lower than its dimension. This follows the underlying assumption that the dimension of the measurements d_1 is larger than the dimension of the sensor-specific variable[1] d_y. We note that the omission of the low eigenvalues and eigenvectors, as done by the pseudo inverse, results both in denoising possible ambient noise and estimation inaccuracies, and also in the attenuation of the common variable X remainders, thereby enhancing the desired variation – that of the sensor-specific variable only.

The distance in the Gaussian kernel in (10) is a modified Mahalanobis distance between the signal samples $\mathbf{s}_i^{(1)}$ and $\mathbf{s}_j^{(1)}$, which was presented in [27], with the exception that the local Covariance matrices are computed based on neighborhoods in the extracted common variable domain. In [28–30], this distance was used in the context of manifold learning and diffusion maps to determine an intrinsic representation of (single) sensor data, invariant to interferences and measurement modalities. Such a manipulation of the Mahalanobis distance via the neighborhood choice was suggested in [31] for the task of sea mine detection in sonar images, and in [32] of reduction of stochastic dynamical systems. There, by controlling the locality within a pre-defined training set, a new metric was presented, which is invariant to perturbations in the appearance of the target. In our work, rather than building invariances, we use a similar approach of choosing the neighborhoods in a multi-sensor setting in order to obtain a full parametrization of all the underlying sources of variability.

Once the affinity kernel $\widetilde{\mathbf{W}}^{(1)}$ is constructed, given that it captures only the variability of the (desired) sensor-specific variable Y, we apply the standard diffusion maps algorithm to find the parametrization of the underlying variable Y, denoted by $\hat{\mathbf{y}}$. The proposed algorithm is summarized in Algorithm 1 .

Algorithm 1: The proposed algorithm.

Input: signals $\{\mathbf{s}_i^{(1)}\}_{i=1}^N$ and $\{\mathbf{s}_i^{(2)}\}_{i=1}^N$ originating from both sensors.
Output: parametrizations of the sensor specific variables Y and Z.

1. Compute the parametrization $\hat{\mathbf{x}}$ of the common variable X using the alternation-diffusion algorithm.
2. For each signal $\mathbf{s}_i^{(1)}$:
   (a) Find the local neighborhood of $\mathbf{s}_i^{(1)}$ denoted by $\mathcal{N}_i^{(1)}$ in terms of the parametrization found in the previous step.
   (b) Compute the local mean, using Equation (8), and the local Covariance, using Equation (9).
3. Compute the affinity matrix using the Mahalanobis distance between the signals $\mathbf{s}_i^{(1)}$ and $\mathbf{s}_j^{(1)}$, as done in Equation (10).
4. Apply the standard diffusion maps algorithm on the above matrix to obtain a parametrization $\hat{\mathbf{y}}$ for the variable Y.
5. Repeat the above steps for the second sensor.

5. Theoretical analysis

In this section, we provide a theoretical analysis, showing that indeed the proposed algorithm approximates the distance between two signal samples in terms of the sensor-specific variable. As above, without loss of generality, we focus on signal samples arising from the first sensor, and therefore, our goal is to extract the variable Y. For simplicity, in this section we omit the sensor index. A similar derivation to the one presented in this section was done in [27,31]. Here, we highlight the significant differences both in terms of the analysis and in terms of the underlying assumptions.

Assumption 1. If the measurement sample $\mathbf{s}_j = g(\mathbf{x}_j, \mathbf{y}_j)$ belongs to the neighborhood of $\mathbf{s}_i = g(\mathbf{x}_i, \mathbf{y}_i)$, i.e., $j \in \mathcal{N}_i^{(1)}$, then $\|\mathbf{x}_j - \mathbf{x}_i\|_2 = O(\|\mathbf{y}_j - \mathbf{y}_i\|_2^2)$.

This assumption relies on the ability of AD to capture the common variable X, as was proven in [8]. By definition, if a signal sample $\mathbf{s}_j$ is in the neighborhood of $\mathbf{s}_i$, then the distance between their extracted values of common variable X is small (which in practice, is controlled by the tunable threshold η_i). Here we further assume that, if a signal sample $\mathbf{s}_j$ is in the neighborhood of $\mathbf{s}_i$, then the distance between their respective X values (which is small by definition) is also smaller than the distance between their associated Y values by at least one order of magnitude.

Assumption 2. Locally, for every signal sample $\mathbf{s}_i$, the empirical Covariance matrix of the sensor-specific variable Y given the extracted common variable X is isotropic, i.e., it is given by

$$\sum_{j \in \mathcal{N}_i^{(1)}} (\mathbf{y}_j - \mathbf{y}_i)(\mathbf{y}_j - \mathbf{y}_i)^T = \mathbf{I}, \tag{11}$$

where $\mathbf{I}$ is the identity matrix.

While Assumption 2 may seem to be artificial and restrictive, in Section 6 we present experimental results supporting it empirically, by showing the successful deployment of our algorithm in three different applications. In addition, we note that it was used in slightly different contexts in [27,30,33] and successfully applied in many applications with real measured data. The following result follows Assumption 1 and Assumption 2.

Theorem 1. *For any signal sample $\mathbf{s}_i = g(\mathbf{x}_j, \mathbf{y}_j)$, if Assumptions 1 and 2 are satisfied, then*

$$\sum_{j \in \mathcal{N}_i} (\mathbf{s}_j - \mathbf{s}_i)(\mathbf{s}_j - \mathbf{s}_i)^T = \mathbf{J}_i^y \mathbf{J}_i^{yT} + O\left(\left\|\mathbf{y}_j - \mathbf{y}_i\right\|_2^3\right), \tag{12}$$

where $\mathbf{J}_i^y$ is the Jacobian of the function g with respect to the variables Y, computed at the ith sample $\mathbf{s}_i$.

Proof. Using Taylor expansion, we can linearly approximate the observation function $g(\mathbf{x}_j, \mathbf{y}_j)$ around the point $(\mathbf{x}_i, \mathbf{y}_i)$, obtaining

$$\begin{aligned}\mathbf{s}_j - \mathbf{s}_i = \mathbf{J}_i^x(\mathbf{x}_j - \mathbf{x}_i) + \mathbf{J}_i^y(\mathbf{y}_j - \mathbf{y}_i) \\ + O\left(\left\|\mathbf{x}_j - \mathbf{x}_i\right\|_2^2 + \left\|\mathbf{y}_j - \mathbf{y}_i\right\|_2^2 + \left(\mathbf{x}_j - \mathbf{x}_i\right)^T\left(\mathbf{y}_j - \mathbf{y}_i\right)\right),\end{aligned} \tag{13}$$

where $\mathbf{J}_i^x$ and $\mathbf{J}_i^y$ are the Jacobians at the ith sample, $\mathbf{s}_i$, with respect to the variables X and Y, respectively. The last term in (13) encapsulates all the higher order derivatives that do not appear in this linear approximation. Under Assumption 1, (13) can be rewritten as

$$\mathbf{s}_j - \mathbf{s}_i = \mathbf{J}_i^y(\mathbf{y}_j - \mathbf{y}_i) + O\left(\left\|\mathbf{y}_j - \mathbf{y}_i\right\|_2^2\right). \tag{14}$$

for any $j \in \mathcal{N}_i$. Using (14) and by Assumption 2, the empirical Covariance around the sample $\mathbf{s}_i$ is given by

$$\begin{aligned}\sum_{j \in \mathcal{N}_i} (\mathbf{s}_j - \mathbf{s}_i)(\mathbf{s}_j - \mathbf{s}_i)^T &= \sum_{j \in \mathcal{N}_i} \mathbf{J}_i^y(\mathbf{y}_j - \mathbf{y}_i)(\mathbf{y}_j - \mathbf{y}_i)^T \mathbf{J}_i^{yT} + O\left(\left\|\mathbf{y}_j - \mathbf{y}_i\right\|_2^3\right) \\ &= \mathbf{J}_i^y \mathbf{J}_i^{yT} + O\left(\left\|\mathbf{y}_j - \mathbf{y}_i\right\|_2^3\right).\end{aligned} \tag{15}$$

[1] Although the Covariance matrix is constructed as a sum of q outer products, the rank of this matrix is lower than q since we assume that the dimension of the sensor-specific variables is $dy < q$. This is a reasonable assumption once the neighborhood is large enough.

concluding our proof. □

Theorem 1 shows that in order to estimate empirically the Gram matrix $\mathbf{J}_i^y \mathbf{J}_i^{yT}$ of the Jacobian $\mathbf{J}_i^y$, we can simply compute the empirical Covariance matrix of the samples in the neighborhood of $\mathbf{x}_i$, where the neighborhood is defined by the AD metric. This is accomplished without the knowledge of the function g itself. In the next result, we use this Gram matrix for estimating the distance between a pair of signal samples in terms of the sensor-specific Y.

Theorem 2. *For any two signal sample* $\mathbf{s}_i = g(\mathbf{x}_i, \mathbf{y}_i)$ *and* $\mathbf{s}_j = g(\mathbf{x}_j, \mathbf{y}_j)$, *the Euclidean distance between the corresponding realizations of the sensor-specific variable is given by*

$$\left\|\mathbf{y}_j - \mathbf{y}_i\right\|_2^2 = (\mathbf{s}_j - \mathbf{s}_i)^T \left(\mathbf{J}_i^y \mathbf{J}_i^{yT}\right)^{\dagger} (\mathbf{s}_j - \mathbf{s}_i) \tag{16}$$

$$+ O(\left\|\mathbf{s}_j - \mathbf{s}_i\right\|_2^3). \tag{17}$$

Proof. In the proof of Theorem 1, we considered the Taylor expansion of the observation function g from the domain of the latent variables X and Y to the range of the measured signal. Similarly, consider the Taylor expansion of its inverse function g^{-1} (recalling that g is assumed bilipschitz), which is given by

$$\begin{bmatrix} \mathbf{x}_j - \mathbf{x}_i \\ \mathbf{y}_j - \mathbf{y}_i \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_i^x \\ \mathbf{Q}_i^y \end{bmatrix} (\mathbf{s}_j - \mathbf{s}_i) + O\left(\left\|\mathbf{s}_j - \mathbf{s}_i\right\|_2^2\right), \tag{18}$$

where $\mathbf{Q}_i^x$ and $\mathbf{Q}_i^y$ are the Jacobian matrices of g^{-1} with respect to the variables X and Y, respectively. Isolating the Y variable yields

$$\mathbf{y}_j - \mathbf{y}_i = \mathbf{Q}_i^y (\mathbf{s}_j - \mathbf{s}_i) + O\left(\left\|\mathbf{s}_j - \mathbf{s}_i\right\|_2^2\right). \tag{19}$$

By applying the ℓ_2 norm to both sides, we obtain

$$\left\|\mathbf{y}_j - \mathbf{y}_i\right\|_2^2 = (\mathbf{s}_j - \mathbf{s}_i)^T \mathbf{Q}_i^{yT} \mathbf{Q}_i^y (\mathbf{s}_j - \mathbf{s}_i) + O\left(\left\|\mathbf{s}_j - \mathbf{s}_i\right\|_2^3\right). \tag{20}$$

The Taylor expansions in (14) and (19) correspond to g and g^{-1}, respectively. Thus, due to the inverse function theorem, we have

$$\mathbf{Q}_i^{yT} \mathbf{Q}_i^y = \left(\mathbf{J}_i^y \mathbf{J}_i^{yT}\right)^{-1}. \tag{21}$$

Typically, the dimension of the measurements is larger than the sum of the dimensions of the common and sensor-specific variables, i.e., $d_1 > d_x + d_y$. As a result, the number of rows in $\mathbf{J}_i^y$ is larger than the number of columns, and hence, the Gram matrix $\mathbf{J}_i^y \mathbf{J}_i^{yT}$ is not full-rank. Consequently, it is not invertible and one needs to employ a pseudo-inverse operator instead. By substituting (21) into (20), we obtain

$$\left\|\mathbf{y}_j - \mathbf{y}_i\right\|_2^2 = (\mathbf{s}_j - \mathbf{s}_i)^T \left(\mathbf{J}_i^y \mathbf{J}_i^{yT}\right)^{\dagger} (\mathbf{s}_j - \mathbf{s}_i) + O\left(\left\|\mathbf{s}_j - \mathbf{s}_i\right\|_2^3\right), \tag{22}$$

as required. □

Two important notes are due at this point. One is that the above analysis is based on the Taylor expansion around the sample $\mathbf{s}_i$. If we repeat the derivations with the Taylor expansion around the sample $\mathbf{s}_j$ as well, then, the mean of the two resulting expressions is given by

$$\left\|\mathbf{y}_j - \mathbf{y}_i\right\|_2^2 = \frac{1}{2}(\mathbf{s}_j - \mathbf{s}_i)^T \left(\left(\mathbf{J}_i^y \mathbf{J}_i^{yT}\right)^{\dagger} + \left(\mathbf{J}_j^y \mathbf{J}_j^{yT}\right)^{\dagger}\right)(\mathbf{s}_j - \mathbf{s}_i) + O\left(\left\|\mathbf{s}_j - \mathbf{s}_i\right\|_2^3\right). \tag{23}$$

Once such symmetrization is employed, further analysis presented in [27] improves the order of the error term in (23) to $\left\|\mathbf{s}_j - \mathbf{s}_i\right\|_2^4$. Two is that, for simplicity, the above analysis assumes (unrealistically) that every signal has a local mean equal to zero, i.e., $\boldsymbol{\mu}_i = \mathbf{0}$. This simplifies the Taylor approximation, which would otherwise have to be of the signal $\mathbf{s}_j - \boldsymbol{\mu}_j$ around the point $\mathbf{s}_i - \boldsymbol{\mu}_i$ in order to align with our algorithm that subtracts the mean from every local neighborhood of the signal. We refer the reader to [30], where a similar derivation was done without such an assumption. Combining these two notes results in the expression presented in (10), without a rigorous proof.

To conclude, the analysis presented in this section coincides with the proposed method. Indeed, in Algorithm 1, we begin by seeking for signals close in terms of the common variable using the parametrization obtained from the AD algorithm. Once such a neighborhood is found, we compute its local empirical Covariance matrix (9), which is then used in the modified Mahalanobis distance (10) to approximate the desired Euclidean distance. Theorem 1 proves that the aforementioned empirical Covariance approximates the Gram matrix $\mathbf{J}_i^y \mathbf{J}_i^{yT}$. According to Theorem 2, this approximation of the Gram matrix can be used to approximate the Euclidean distance in terms of the desired sensor-specific variable Y via the modified Mahalanobis distance (23) (which is used in Algorithm 1).

Final remark concerns the accuracy of the Euclidean distances approximation. Theorem 2 implies that when the distances are large, the error terms are large and the (local) approximation via the linear terms is poor. This problem is "automatically" alleviated by the standard usage of the Gaussian kernel in (10); due to its fast decay, large distances are implicitly attenuated.

6. Experimental results

6.1. Synthetic example

Consider three independent and identically distributed random variables, X, Y, Z, sampled uniformly in [0, 1]. We generate from these variables 3000 triplets of $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)$. Assume two sensors observing these hidden samples through the following nonlinear functions g and h

$$\mathbf{s}_i^{(1)} = g(\mathbf{x}_i, \mathbf{y}_i) = \begin{bmatrix} R + r^{(1)} cos(2\pi \mathbf{y}_i) cos(2\pi \mathbf{x}_i) \\ R + r^{(1)} cos(2\pi \mathbf{y}_i) sin(2\pi \mathbf{x}_i) \\ r^{(1)} sin(2\pi \mathbf{y}_i) \end{bmatrix} \tag{24}$$

$$\mathbf{s}_i^{(2)} = h(\mathbf{x}_i, \mathbf{z}_i) = \begin{bmatrix} R + r^{(2)} cos(2\pi \mathbf{z}_i) cos(2\pi \mathbf{x}_i) \\ R + r^{(2)} cos(2\pi \mathbf{z}_i) sin(2\pi \mathbf{x}_i) \\ r^{(2)} sin(2\pi \mathbf{z}_i) \end{bmatrix}, \tag{25}$$

so that we obtain 3000 pairs of signal measurements $(\mathbf{s}_i^{(1)}, \mathbf{s}_i^{(2)})$. We set $R = 10$, $r^{(1)} = 4$, $r^{(2)} = 2$.

Notice that g and h correspond to two tori, with major angle X, serving as their common hidden variable, and with minor angles, Y or Z, serving as their respective sensor-specific variables. We apply Algorithm 1 to these samples where the size of the neighborhoods in the common variable domain is $q = 11$. In Fig. 1 we color both tori according to the extracted parametrization of the common variable and also according to the obtained parametrizations of the two sensor-specific variables. Indeed, we observe that our method accurately extracts the three hidden variables. The coloring of the tori according to the common variable X is highly coherent with the major angle, while the coloring with respect to the sensor-specific variables, Y and Z, are consistent with the minor angles. In order to test the robustness of our algorithm with respect to the parameter q, we apply the proposed method for different values of q and present the extracted sensor-specific variable of one torus in Fig. 2. One can observe that the algorithm is robust to the choice of q, as it successfully extracts the minor angle for a wide range of values. Despite this robustness, one might still wonder how to set q in practice. Following the heuristics suggested

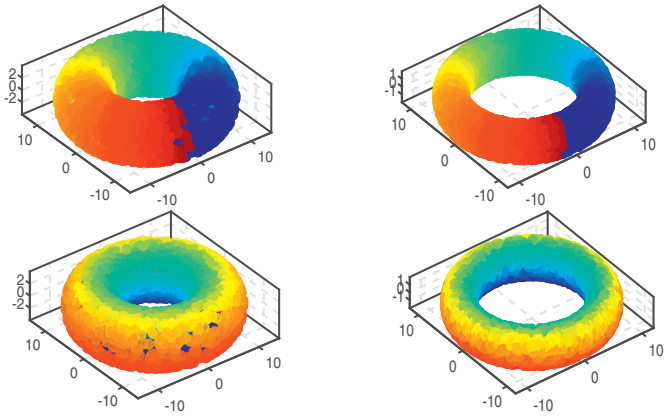


Fig. 1. On the left we plot the samples $\mathbf{s}_i^{(1)}$ and on the right the samples $\mathbf{s}_i^{(2)}$. In the top row, the samples are colored according to the obtained parametrization of the common variable, X. In the bottom row, the samples are colored according to the respective parametrization obtained for the sensor-specific variable, Y and Z.

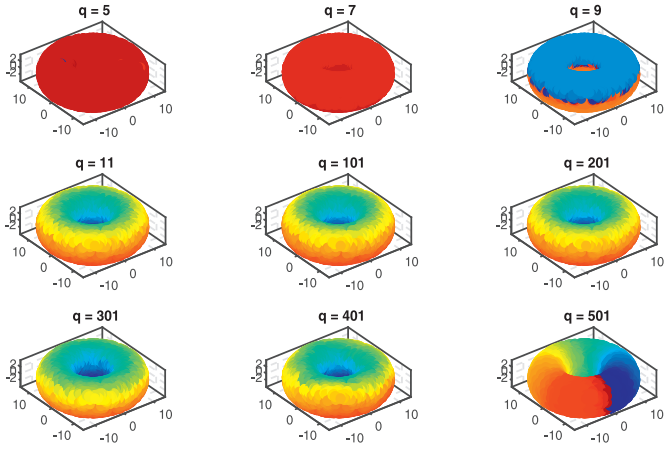
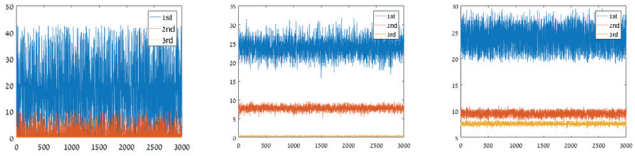


Fig. 2. The sensor-specific variable obtained for different values of q.



(a) $q = 5$. (b) $q = 101$. (c) $q = 501$.

Fig. 3. The eigenvalues of the covariance matrix in (9), $\mathbf{C}_i^{(1)}$, for $i = 1, \ldots, N$. The best $q = 101$ has three separated eigenvalues: the first corresponds to the sensor-specific variable, the second to remainders of the common variable and the third is simply zero.

in [29], in Fig. 3, we plot the eigenvalues of the local covariance matrices in (9) as functions of the neighborhood index i, for three choices of q – showing how these can help in choosing the correct value of q. The best q, in this case $q = 101$, has three separated eigenvalues. The first corresponds to the sensor-specific variable, which we aim to extract. The second corresponds to the "residual" of the common variable, which was not marginalized perfectly in the first step of the algorithm. The third and final is simply zero, since the points are intrinsically two dimensional. For worse values of q, the eigenvectors behave differently. For $q = 5$, we obtain degenerate parametrization of the sensor-specific variable, which is not clearly separated from the "residual" of the common variable. While for $q = 501$ the intrinsic dimension is equal to three, instead of two.

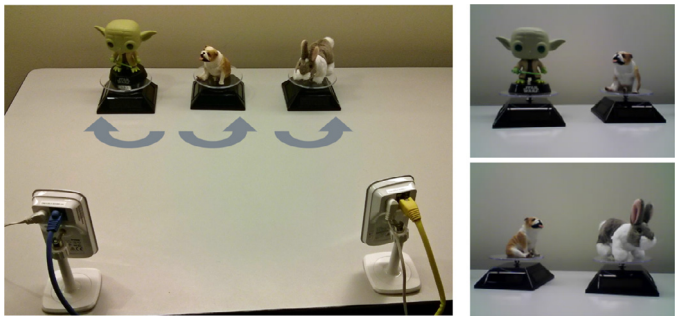


Fig. 4. The experimental setup of the toy problem (left), and examples of images captured simultaneously by the two cameras (right).

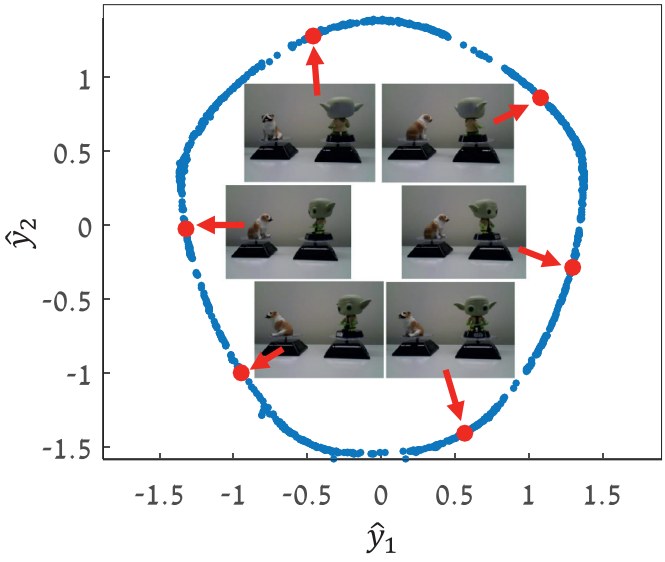


Fig. 5. The parametrization obtained by the proposed algorithm, displaying the first two components from $\hat{\mathbf{y}}$. It is demonstrated that Algorithm 1 enables us to capture the sensor-specific variable – the angle of Yoda.

6.2. Playing with Toys: Yoda, Bulldog and Rabbit

In this experiment, we consider the toy problem presented in [8]. The setting of the problem includes three objects: a figure of Yoda (green alien), a Bulldog, and a Rabbit, which were placed on rotating platforms. The three figures rotate in different speeds, and one (Yoda) in a different direction. This entire scene was captured by two cameras, as demonstrated in Fig. 4 (left). The view of the first camera included both the figures of Yoda and Bulldog (Fig. 4 (top-right)), while the view of the second camera included the Bulldog and the Rabbit (Fig. 4 (bottom-right)). The two cameras were synchronized, i.e., they were taking simultaneous snapshots. In this problem, the latent variables are the orientation angles of the three figures, where the angle of the Bulldog is the common variable X, and the angles of Yoda and of the Rabbit are the sensor specific-variables Y and Z, respectively. The data at hand consist of images (snapshots) of the rotating figures, captured simultaneously by the two cameras.

In [8], it was shown that the AD algorithm attains a parametrization of the angle of the Bulldog X, namely, the common variable hidden in the sets of images. In this work we infer a parametrization of the angle of the sensor-specific Yoda Y. In Fig. 5 we present the result of applying Algorithm 1 in this setup where the size of the neighborhoods in the common variable domain is $q = 15$. We scatter plot the first two coordinates in the obtained parametrization, $\hat{\mathbf{y}}$, and observe that the parametrization

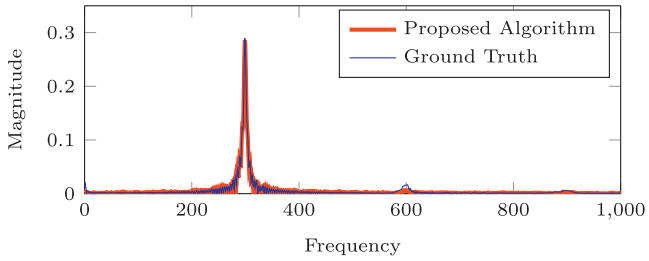


Fig. 6. The discrete Fourier transform (magnitude only) of the first components in $\tilde{\mathbf{y}}$ and $\hat{\mathbf{y}}$. The parametrization $\tilde{\mathbf{y}}$ (ground truth) is obtained by diffusion maps applied to the cropped images, and the parametrization $\hat{\mathbf{y}}$ is obtained by Algorithm 1.

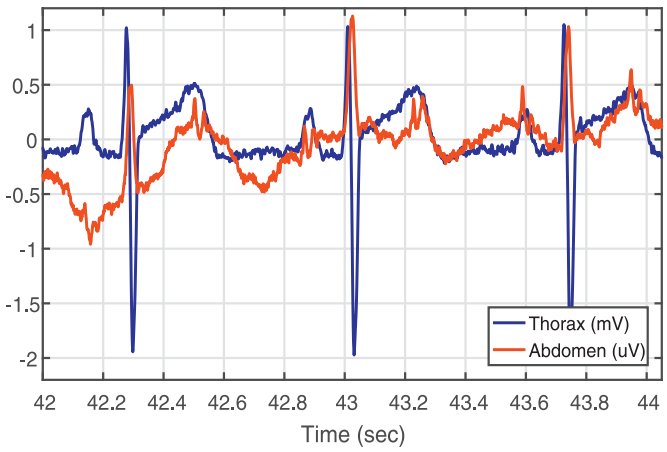


Fig. 7. A short interval of the abdomen and thorax signals. The large spikes correspond the QRS complexes of the maternal ECG, and the small spikes in the abdomen signal correspond to the QRS complexes of the fetal ECG.

takes the shape of a circle, correctly representing a rotating angle. To show that this angle is indeed associated with the rotating angle of Yoda, we overlay several images corresponding to the embedded points.[2] It can be seen that the orientation angle of Yoda corresponds to angles on the obtained circle. For further evaluation, we compute the ground truth parametrization of the angle of Yoda, denoted by $\tilde{\mathbf{y}}$. To this end, we crop all images captured by the first camera, discarding Bulldog and maintaining only Yoda, and then we apply diffusion maps to the set of cropped images. We emphasize that the information of how to appropriately crop the image is not available to the proposed algorithm, which does not use any prior knowledge on the experimental setting, and it was done for illustration and evaluation purposes only. In Fig. 6, we present the discrete Fourier transform of both the first components of $\tilde{\mathbf{y}}$ and $\hat{\mathbf{y}}$; the parametrization $\tilde{\mathbf{y}}$ is obtained by diffusion maps applied to the cropped images, and the parametrization $\hat{\mathbf{y}}$ is obtained by Algorithm 1. We observe a sharp peak around the frequency 310, which according to [8] corresponds exactly to the rotation speed of Yoda.[3] Moreover, the curves are similar, implying on the successful recovery of the sensor-specific variable by the proposed algorithm.

6.3. Non-invasive fetal ECG

Fetal heart rate monitoring [34,35] is widely-used for the assessment of the fetal health both during pregnancy and during delivery. The most accurate method, relying on the placement of electrodes on the fetal scalp, carries many risks. Consequently, non-invasive measurements are usually carried out by placing electrodes on the abdomen of the mother. The reader is referred to [36] for a review of this topic. Naturally, the measured signal contains, in addition to the fetal's heart beats, the maternal ECG, masking the desired information. In order to suppress the maternal ECG and to extract the fetal ECG, another (reference) electrode is often placed on the mother's thorax for the purpose of measuring only the maternal ECG. In practice, in addition to being occluded by the maternal ECG, the fetal ECG is also contaminated by noise, such as power line disturbance and maternal muscle movements [37].

As reported in [37], due to its time-varying statistical character, the ECG of the fetal is a highly nonstationary signal. Moreover, the relation between the measured abdomen signal and the fetal ECG is nonlinear. As such, standard approaches, e.g., the adaptive least mean squares (LMS) algorithm [38], provide only coarse estimations in recovering the fetal ECG, and the solution for this problem is not trivial, and it is still considered an open problem.

We use the fetal ECG extraction problem as a testbed for our algorithm not only to demonstrate its applicability, but also to show the relevance of the problem setting we present in this paper to real measured data. Let us now return to the problem formulation, as defined in Section 2. In our context, the variable common to both the abdomen and thorax signals is the maternal ECG, while the sensor-specific variable in the abdomen signal is the desired fetal ECG. We demonstrate that our proposed method is capable of not only recovering the maternal ECG (common variable), but also factoring out the mother's pulse from the measured abdomen signal. This results in revealing the fetal ECG (sensor-specific variable), which is relatively weak when compared to the maternal signal. More specifically, we show that our method builds a parameterization of the fetal ECG, which, in turn, could aid in detecting fetal QRS[4] complexes which are used in measuring the fetal's heart rate.

In our experiments we use the "Non-invasive Fetal ECG Database" from PhysioNet [39], without applying any preprocessing to the raw data. In Fig. 7 we present a short 2 s interval from both the thorax and the abdomen signals. The length of the entire measurement is 5 min and 20 s and in the following experiments we analyze a subinterval of length (approximately) 16 s. The sampling rate of the ECG signals is 1 kHz.

Prior to employing our proposed algorithm, we begin by trying a much simpler method, namely nonlinear ICA. Broadly speaking, this attempts to unmix the maternal ECG from the fetal ECG by transforming the two signals at hand into two components that are as statistically independent from each other as possible. We use an online available implementation of ICA [40] and depict the results in Fig. 8 (the nonlinearity used was $g(x) = x^3$). Clearly, this method does not yield satisfactory results in separating the fetal and maternal ECG signals. We suspect that some nontrivial preprocessing must be employed in order to facilitate the employment of this algorithm. Conversely, we emphasize that our proposed method does not rely on any preprocessing of the data.

Throughout this work, we relied on the AD algorithm for extracting the common source of variability. A different approach, tackling the same problem, is Canonical Correlation Analysis (CCA) [41]. In this framework, one seeks for linear projections of the input signals that would lead to the highest possible correlation. Note that in contrast to AD, this algorithm is linear and hence cannot handle the non-linearities in the fetal ECG extraction task. In order to support this claim, we run CCA on the thorax and ab-

[2] We flipped the images captured by the camera horizontally for easier viewing of the figure.

[3] The frequency is given in terms of the number of cycles completed in the duration of the experiment.

[4] QRS complex is a name for the three graphical deflections seen on a typical ECG diagram.

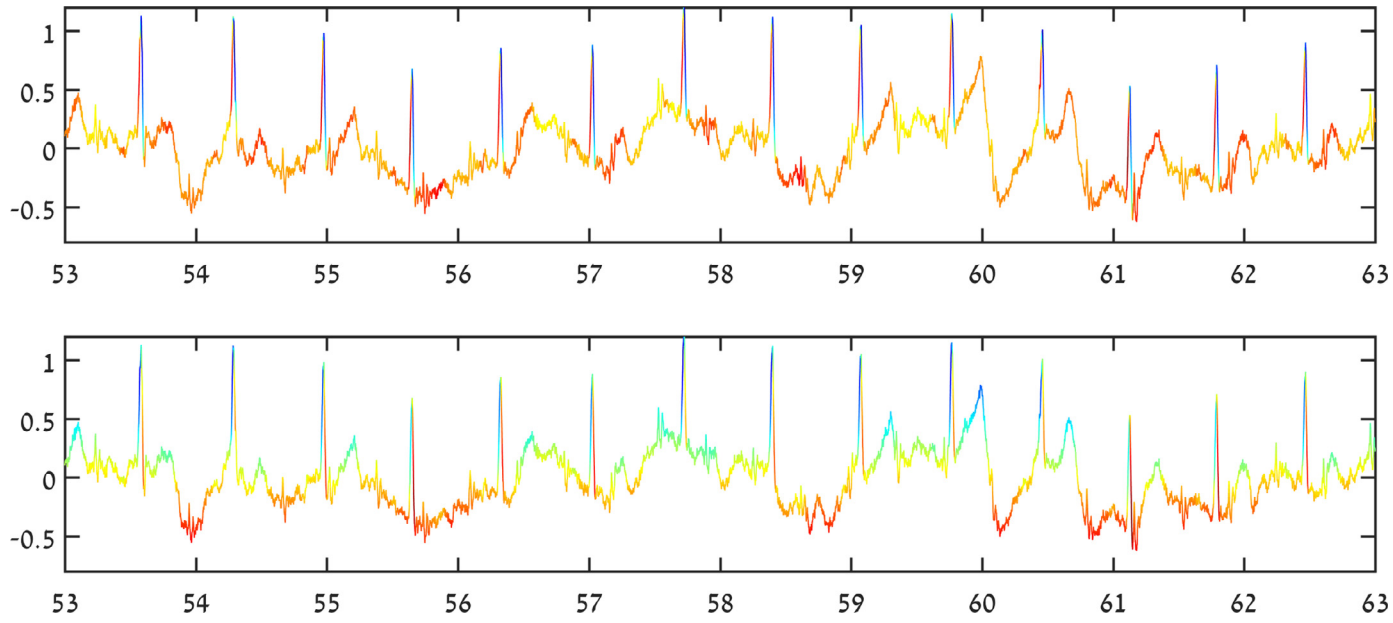


Fig. 8. The abdomen signal colored according to the first and second independent components obtained from ICA. The first component captures the peaks in the maternal ECG, as can be seen from their blue color, while the second captures local maxima and minima in the maternal ECG, coloring them in cyan and red, correspondingly. In both cases, the fetal QRS complexes do not have a consistent color and as such cannot be detected. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

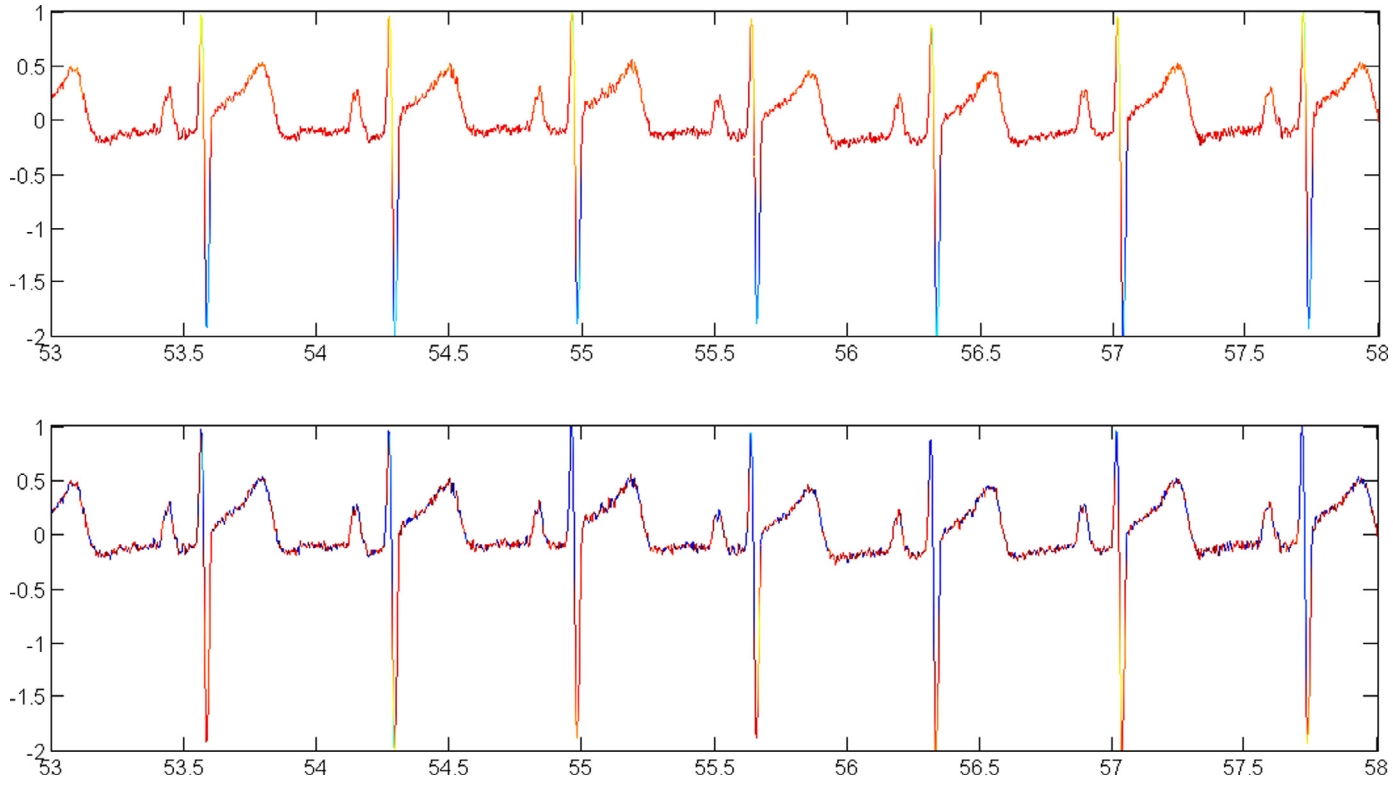


Fig. 9. The thorax signal colored according to the first two components of CCA.

domen[5] signals. Fig. 9 depicts the thorax signal colored according to the first two components of CCA. We observe that both CCA components are non-informative and do not correspond to the common source of variability, the maternal ECG cycle.

We now move to employing our proposed algorithm. Given the abdomen and thorax signals, we first apply the AD algorithm to extract the common variable and present the obtained parametrization in Fig. 10. The algorithm is applied to time segments of length 256 samples (lag map) with 16 samples overlap, which are ex-

[5] The CCA is applied to two matrices that contain in their columns non-overlapping segments of length 8 samples taken from the thorax and abdomen signals. Varying sizes were tested, however none resulted in satisfactory results.

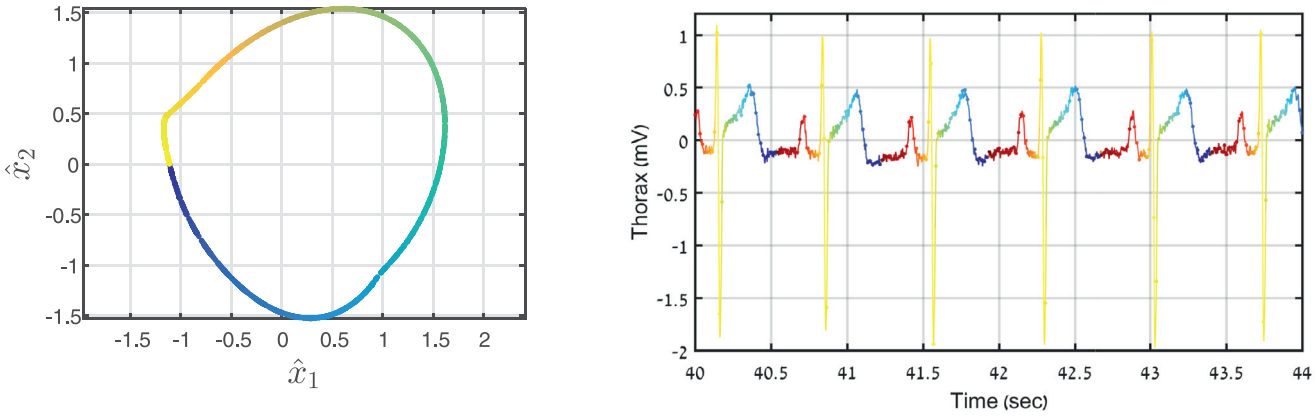


Fig. 10. Left: The parametrization obtained by taking the first two components from the AD algorithm applied to the abdomen and thorax signals. Right: The thorax signal colored according to the common variable.

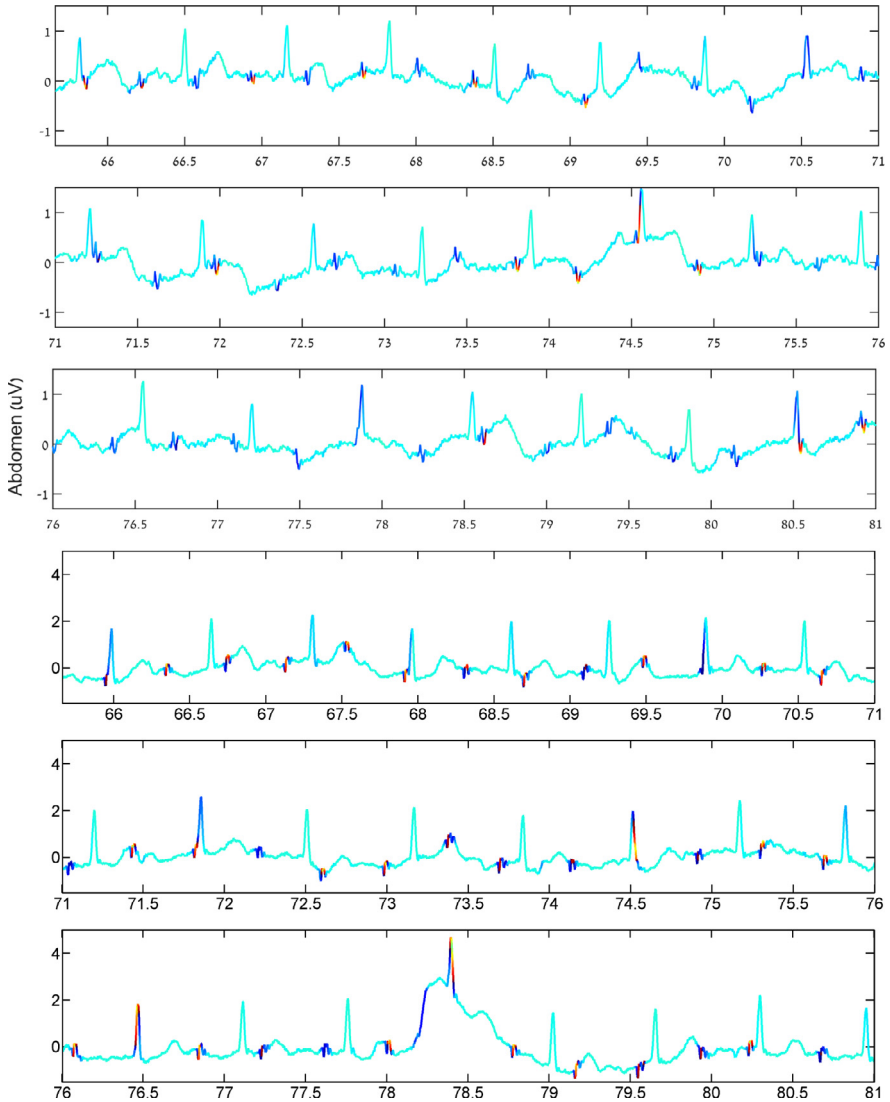


Fig. 11. The abdomen signal as a function of time (in seconds). Top three plots correspond to one patient, while the bottom three to another. The signal is colored according to the parametrization of the sensor-specific variable. As such, in places where there is a fetal QRS complex, one expects to see the color of the graph change. Indeed, we can observe a distinct color change, where points in dark-blue correspond to the fetal QRS complexes. This result implies that the obtained parametrization can be used to define an indicator of the fetal ECG signal. Importantly, at $t = 70.5$ s (top) , $t = 77.8$ s (top), $t = 69.9$ (bottom) sec and $t = 78.5$ (bottom) we observe that the fetal's heart beat is detected, even in pathological cases where it is completely "buried" in the maternal heart beat. Moreover, at times $t = 78.2$ s (bottom) a possibly redundant fetal heart peak is detected. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tracted from both the abdomen and thorax signals.[6] Since the sampling rate is 1 kHz, each segment is of duration of 256 ms. In the context of this paper, these segments are viewed as the sensor samples, and thus, are denoted by $\mathbf{s}_i^{(1)}$ and $\mathbf{s}_i^{(2)}$. Each 2D point in the scatter plot in Fig. 10 (left), representing a pair of segments, is colored according to the angle created between the axis origin and the embedded point, using the first two components obtained by AD. We choose this method of coloring since the obtained parametrization is equivalent to a circle, representing the common variable's periodicity). To emphasize the validity of our assumption that the common variable is indeed related to the maternal ECG, we present in Fig. 10 (right) the thorax signal and color its samples according to color of the left figure. Clearly, the common variable coincides with the cardiac cycle of the mother. We can also compare this parametrization with the one obtained from the first (top) component of ICA in Fig. 8. Both capture the maternal ECG cycle, however the result of AD is clearly superior.

Next, we proceed by extracting the sensor-specific variable from the abdomen signal using our proposed algorithm. In this experiment we set the size of the neighborhoods in the common variable domain to be $q = 21$. In the top three plots of Fig. 11, we present the abdomen signal colored according to the obtained parametrization of the sensor-specific variable. The results imply that indeed the sensor-specific variable is related to the ECG of the fetal. To demonstrate the generality of our method, we repeat the experiment on signals measured from another patient. The result of this experiment are presented in the bottom three plots of Fig. 11, showing that the parametrization of the sensor-specific variable manages to capture the ECG of the fetus in this case as well. We tried our method on several additional patients and the results were comparable to those presented here.

7. Conclusions

Given a set of measurements, originating from several sensors, the AD algorithm extracts a parametrization of a variable common to all sources. In this work, leveraging on AD, we proposed a method which further analyzes the signals by extracting the sensor-specific variables. We provided a theoretical justification as well as various applications. A shortcoming of our method is the need to extract the intermediate common variable parametrization. Proposing a method that could skip this stage is a promising future direction.

References

[1] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: an overview of methods, challenges, and prospects, Proc. IEEE 103 (9) (2015) 1449–1477.
[2] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering., in: NIPS, 14, 2001, pp. 585–591.
[3] D.L. Donoho, C. Grimes, Hessian eigenmaps: locally linear embedding techniques for high-dimensional data, Proc. Natl. Acad. Sci. 100 (10) (2003) 5591–5596.
[4] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[5] J.A. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, Springer Science & Business Media, 2007.
[6] R.R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (1) (2006) 5–30.
[7] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, S.W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps, Proc. Natl. Acad. Sci. U.S.A. 102 (21) (2005) 7426–7431.
[8] R.R. Lederman, R. Talmon, Learning the geometry of common latent variables using alternating-diffusion, Appl. Comput. Harmon. Anal. (2015).
[9] R. Talmon, H.-t. Wu, Latent common manifold learning with alternating diffusion: analysis and applications, arXiv preprint arXiv:1602.00078 (2016).
[10] V.R. de Sa, Spectral clustering with two views, ICML Workshop on Learning with Multiple Views, 2005.
[11] V.R. de Sa, P.W. Gallagher, J.M. Lewis, V.L. Malave, Multi-view kernel construction, Mach. Learn. 79 (1–2) (2010) 47–71, doi:10.1007/s10994-009-5157-z.
[12] B. Boots, G.J. Gordon, Two-manifold problems with applications to nonlinear system identification, Proceedingsof the 29th International Conference on Machine Learning (ICML), 2012.
[13] O. Lindenbaum, A. Yeredor, M. Salhov, A. Averbuch, Multiview diffusion maps, arXiv preprint arXiv:1508.05550 (2015).
[14] T. Michaeli, W. Wang, K. Livescu, Nonparametric canonical correlation analysis, arXiv preprint arXiv:1511.04839 (2015).
[15] O. Yair, R. Talmon, Local canonical correlation analysis for nonlinear common variables discovery, IEEE Trans. Signal Process. (2016). arXiv:1606.04268. arXiv preprint
[16] R.R. Lederman, R. Talmon, H.-T. Wu, Y.-L. Lo, R.R. Coifman, Alternating diffusion for common manifold learning with application to sleep stage assessment, in: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, IEEE, 2015, pp. 5758–5762.
[17] Y. Jia, M. Salzmann, T. Darrell, Factorized latent spaces with structured sparsity, in: Advances in Neural Information Processing Systems, 2010, pp. 982–990.
[18] M.W. Johns, Reliability and factor analysis of the epworth sleepiness scale, Sleep 15 (4) (1992) 376–381.
[19] A. Damianou, C. Ek, M. Titsias, N. Lawrence, Manifold relevance determination, arXiv preprint arXiv:1206.4610 (2012).
[20] G. Song, S. Wang, Q. Tian, et al., Multimodal similarity gaussian process latent variable model, IEEE Trans. Image Process. (2017).
[21] G. Song, S. Wang, Q. Huang, Q. Tian, Similarity Gaussian process latent variable model for multi-modal data analysis, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4050–4058.
[22] W. Wang, R. Arora, K. Livescu, J. Bilmes, On deep multi-view representation learning, in: Proceedings of the 32nd International Conference on Machine Learning (ICML-15), 2015, pp. 1083–1092.
[23] Y. Hua, S. Wang, S. Liu, A. Cai, Q. Huang, Cross-modal correlation learning by adaptive hierarchical semantic aggregation, IEEE Trans. Multimedia 18 (6) (2016) 1201–1216.
[24] B. Nadler, S. Lafon, R.R. Coifman, I.G. Kevrekidis, Diffusion maps, spectral clustering and reaction coordinates of dynamical systems, Appl. Comput. Harmon. Anal. 21 (1) (2006) 113–127.
[25] S.S. Lafon, Diffusion Maps and Geometric Harmonics, Ph.D. thesis, Yale University, 2004.
[26] U. Von Luxburg, A tutorial on spectral clustering, Stat. Comput.. 17 (4) (2007) 395–416.
[27] A. Singer, R.R. Coifman, Non-linear independent component analysis with diffusion maps, Appl. Comput. Harmon. Anal. 25 (2) (2008) 226–239.
[28] R. Talmon, R.R. Coifman, Empirical intrinsic geometry for nonlinear modeling and time series filtering, Proc. Natl. Acad. Sci. 110 (31) (2013) 12535–12540.
[29] R. Talmon, R.R. Coifman, Intrinsic modeling of stochastic dynamical systems using empirical geometry, Appl. Comput. Harmon. Anal. 39 (1) (2015) 138–160.
[30] R. Talmon, S. Mallat, H. Zaveri, R.R. Coifman, Manifold learning for latent variable inference in dynamical systems, Signal Process., IEEE Trans. 63 (15) (2015) 3843–3856.
[31] G. Mishne, R. Talmon, I. Cohen, Graph-based supervised automatic target detection, Geosci. Remote Sens., IEEE Trans. 53 (5) (2015) 2738–2754.
[32] C.J. Dsilva, R. Talmon, C.W. Gear, R.R. Coifman, I.G. Kevrekidis, Data-driven reduction for a class of multiscale fast-slow stochastic dynamical systems, SIAM J. Appl. Dyn. Syst. 15 (3) (2016) 1327–1351.
[33] M.E. Tipping, C.M. Bishop, Probabilistic principal component analysis, J. R. Stat. Soc. 61 (3) (1999) 611–622.
[34] R.K. Freeman, T.J. Garite, M.P. Nageotte, L.A. Miller, Fetal Heart Rate Monitoring, Lippincott Williams & Wilkins, 2012.
[35] F. Rochard, B.S. Schifrin, F. Goupil, H. Legrand, J. Blottiere, C. Sureau, Nonstressed fetal heart rate monitoring in the antepartum period., Am. J. Obstetrics Gynecol. 126 (6) (1976) 699–706.
[36] E.W. Abdulhay, R.J. Oweis, A.M. Alhaddad, F.N. Sublaban, M.A. Radwan, H.M. Almasaeed, Review article: non-invasive fetal heart rate monitoring techniques, Biomed. Sci. Eng. 2 (3) (2014) 53–67.
[37] E.R. Ferrara, B. Widraw, Fetal electrocardiogram enhancement by time-sequenced adaptive filtering, IEEE Trans. Biomed. Eng. (6) (1982) 458–460.
[38] B. Widrow, J.M. McCool, M.G. Larimore, C.R. Johnson, Stationary and nonstationary learning characteristics of the LMS adaptive filter, Proc. IEEE 64 (8) (1976) 1151–1162.
[39] 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, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220, doi:10.1161/01.CIR.101.23.e215. (June 13) Circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215PMID:1085218.
[40] A. Hyvärinen, Fast and robust fixed-point algorithms for independent component analysis, Neural Netw., IEEE Trans. 10 (3) (1999) 626–634.
[41] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3/4) (1936) 321–377.

[6] The mean of every segment was subtracted.