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

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-