Signal Processing 104 (2014) 232–247

Contents lists available at ScienceDirect

Signal Processing

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

Tangent-based manifold approximation with locally linear models

Sofia Karygianni*, Pascal Frossard

Ecole Polytechnique Fédérale de Lausanne (EPFL), Signal Processing Laboratory (LTS4), CH-1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 27 August 2013
Received in revised form
26 February 2014
Accepted 28 March 2014
Available online 8 April 2014

Keywords:
Manifold approximation
Tangent space
Affine subspaces
Flats

ABSTRACT

In this paper, we consider the problem of manifold approximation with affine subspaces. Our objective is to discover a set of low dimensional affine subspaces that represent manifold data accurately while preserving the manifold's structure. For this purpose, we employ a greedy technique that partitions manifold samples into groups, which are approximated by low dimensional subspaces. We start by considering each manifold sample as a different group and we use the difference of local tangents to determine appropriate group mergings. We repeat this procedure until we reach the desired number of sample groups. The best low dimensional affine subspaces corresponding to the final groups constitute our approximate manifold representation. Our experiments verify the effectiveness of the proposed scheme and show its superior performance compared to state-of-the-art methods for manifold approximation.

1. Introduction

The curse of dimensionality is one of the most fundamental issues that researchers have to face across various data processing disciplines. High dimensional data is often difficult to manipulate: it might belong to huge parametric spaces that are challenging to exploit while the corresponding models can be complex enough to make learning challenging and prone to over-fitting. However, it is not rare that the data follows some underlying structure, which can lead to more efficient data representation and analysis if modeled properly.

The underlying structure of signals of a given family can often be described adequately by a manifold model that has a smaller dimensionality than the signal space. Prominent examples are signals that are related by transformations, like images captured under different viewpoints in a 3D scene, or signals that represent different observations of the same physical phenomenon like EEG and ECG data. Manifold models have been successfully used in many different applications like transformation-invariant classification, recognition and dimensionality reduction [1–3].

In general, manifolds are topological spaces that locally resemble a Euclidean space. Therefore, although they might be extremely complicated structures, they have locally, i.e., in the neighborhood of a point, the same characteristics as the usual Euclidean space. In this work, we are going to consider d-dimensional, differentiable manifolds that are embedded into a higher dimensional Euclidean space, $\mathbb{R}^N, N \gg d$. Intuitively, one can think of a d-dimensional manifold embedded into $\mathbb{R}^N$ as the generalization of a surface in N dimensions: it is a set of points that locally seem to live in $\mathbb{R}^d$ but that macroscopically synthesize a structure living into $\mathbb{R}^N$. For example, a sphere in $\mathbb{R}^3$ and a circle in $\mathbb{R}^2$ are both manifolds of dimensions 2 and 1 respectively. Although manifolds are appealing for effective data representation, their unknown and usually strongly non-linear structure makes their

* Corresponding author.
E-mail addresses: sofia.karygianni@epfl.ch (S. Karygianni), pascal.frossard@epfl.ch (P. Frossard).

http://dx.doi.org/10.1016/j.sigpro.2014.03.047
0165-1684/

manipulation quite challenging. There are cases where an analytical model can represent the manifold, like a model built on linear combinations of atoms coming from a predefined dictionary [4]. However, an analytical model is unfortunately not always available. A workaround consists in trying to infer a global, data-driven parametrization scheme for the manifold by mapping the manifold data from the original space to a low-dimensional parametric space. The problem of unveiling such a parametrization is called manifold learning [1,2].

However, it is in general hard to compute a universal manifold representation that is accurate for all data in the datasets. In general, it is not possible to represent all the non-linearities of the manifold by one single mapping function. Therefore, instead of using just one global scheme, it is often preferable to employ a set of simpler structures to approximate the manifold's geometry. This can be done in the original space of the manifold. The objective of the approximation is to create a manifold model that is as simple as possible while preserving the most crucial characteristic of a manifold, namely its geometrical shape. An example of such an approximation for a 1D manifold is shown in Fig. 1a, where a set of lines approximates the spiral shape.

In this paper, we approximate generic manifolds with simple models that are affine subspaces (flats). Such a choice is motivated by the locally linear character of manifolds as well as the simplicity and efficiency of flats for performing local computations like projections. Our objective is to compute a set of low dimensional flats that represent the data as accurately as possible, and at the same time preserves the geometry of the underlying manifold. We formulate the manifold approximation problem as a constrained clustering problem for manifold samples. The constraints are related to the underlying geometry of the manifold, which is represented by the neighborhood graph of the data samples. We borrow elements of the constrained clustering theory to motivate the use of a greedy scheme for manifold approximation. We first propose to relate the capability of a set of points to be represented by a flat, with the variance of the tangents at these points. Then, we use the difference of tangents to uncover groups of points that comply with the low dimensionality of flats. The partitioning is done in a bottom-up manner where each manifold sample is considered as a different group at the beginning. Groups are then iteratively merged until their number reduces to the desired value. We have tested our algorithm on both synthetic and real data where it gives a superior performance compared to state-of-the-art manifold approximation techniques.

The rest of the paper is organized as follows. In Section 2, we discuss the related work in manifold approximation and other relevant fields like manifold learning and hybrid linear modeling. In Section 3, we give some mathematical definitions related to manifolds and tangent spaces, which are essential for the work presented in this paper. In Section 4, we motivate the use of a greedy strategy with concepts from constrained clustering theory and we present our novel problem formulation for the manifold approximation. We present our approximation algorithm in detail in Section 5. In Section 6, we describe the experimental setup and the results of our experiments. Finally, in Section 7, we provide concluding remarks.

2. Related work

Data representation with affine models has received quite some attention lately. Relative approaches usually fall under the name of either subspace clustering or hybrid linear modeling. Their objective is to find a set of affine models explaining the different data sources, i.e., to cluster the data into groups so that each group can be well represented by a low-dimensional affine space. A common approach is to use an iterative scheme to alternate between steps of data segmentation and subspace estimation aiming at either

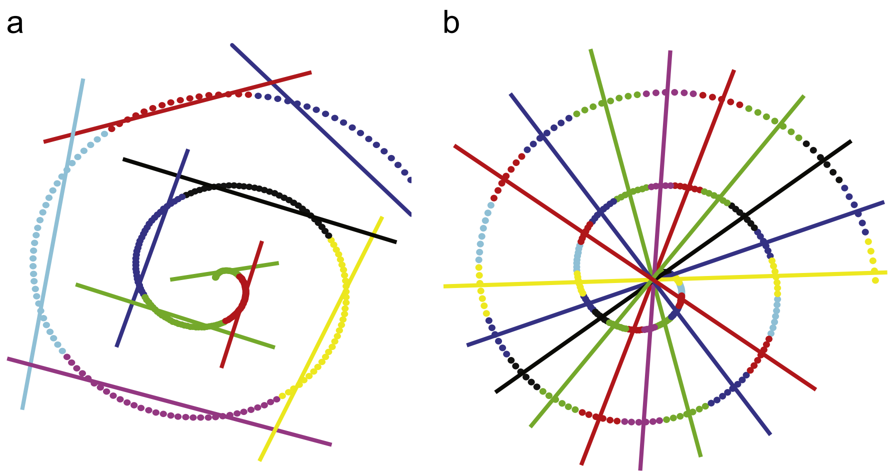

Fig. 1. Manifold approximation illustration. On the left, we have an example of a valid approximation by lines of a 1D manifold embedded into $\mathbb{R}^2$. The different colors represent the different groups of samples, each approximated by a line. On the right, we have an example where the approximation does not align well with the manifold structure, as a result of the median k-flats algorithm [6]. (a) Good manifold approximation example. (b) Bad manifold approximation example. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)