Appl. Comput. Harmon. Anal. 37 (2014) 271–287

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Reduced row echelon form and non-linear approximation for subspace segmentation and high-dimensional data clustering

Akram Aldroubi [a,*], Ali Sekmen [b]

[a] Department of Mathematics, Vanderbilt University, Nashville, TN 37212, United States
[b] Department of Computer Science, Tennessee State University, Nashville, TN 37209, United States

ARTICLE INFO

Article history:
Received 10 May 2012
Received in revised form 9 December 2013
Accepted 15 December 2013
Available online 17 December 2013
Communicated by Jared Tanner

Keywords:
Subspace segmentation
Data clustering
High dimensional data

ABSTRACT

Given a set of data $\mathbf{W} = \{w_1, \ldots, w_N\} \in \mathbb{R}^D$ drawn from a union of subspaces, we focus on determining a nonlinear model of the form $\mathscr{U} = \bigcup_{i\in I} S_i$, where $\{S_i \subset \mathbb{R}^D\}_{i\in I}$ is a set of subspaces, that is nearest to $\mathbf{W}$. The model is then used to classify $\mathbf{W}$ into clusters. Our approach is based on the binary reduced row echelon form of data matrix, combined with an iterative scheme based on a non-linear approximation method. We prove that, in absence of noise, our approach can find the number of subspaces, their dimensions, and an orthonormal basis for each subspace S_i. We provide a comprehensive analysis of our theory and determine its limitations and strengths in presence of outliers and noise.

1. Introduction

In many engineering and mathematics applications, data lives in a union of low dimensional subspaces [1–6]. For instance, the set of all two dimensional images of a given face i, obtained under different illuminations and facial positions, can be modeled as a set of vectors belonging to a low dimensional subspace S_i living in a higher dimensional space $\mathbb{R}^D$ [4,7,8]. A set of such images from different faces is then a union $\mathscr{U} = \bigcup_{i\in I} S_i$. Similar nonlinear models arise in sampling theory where $\mathbb{R}^D$ is replaced by an infinite dimensional Hilbert space $\mathscr{H}$, e.g., $L^2(\mathbb{R}^D)$ [1,9–12].

The goal of subspace clustering is to identify all of the subspaces that a set of data $\mathbf{W} = \{w_1, \ldots, w_N\} \in \mathbb{R}^D$ is drawn from and assign each data point w_i to the subspace it belongs to. The number of subspaces, their dimensions, and a basis for each subspace are to be determined even in presence of noise, missing data, and outliers. The subspace clustering or segmentation problem can be stated as follows: Let $\mathscr{U} = \bigcup_{i=1}^M S_i$ where $\{S_i \subset \mathscr{B}\}_{i=1}^M$ is a set of subspaces of a Hilbert space or Banach space $\mathscr{B}$. Let $\mathbf{W} = \{w_j \in \mathscr{B}\}_{j=1}^N$ be a set of data points drawn from $\mathscr{U}$. Then,

1. determine the number of subspaces M,
2. determine the set of dimensions $\{d_i\}_{i=1}^M$,

* Corresponding author.

1063-5203/$ – see front matter
http://dx.doi.org/10.1016/j.acha.2013.12.001

3. find an orthonormal basis for each subspace S_i,
4. collect the data points belonging to the same subspace into the same cluster.

Note that often the data may be corrupted by noise, may have outliers or the data may not be complete, e.g., there may be missing data points. In some subspace clustering problems, the number M of subspaces or the dimensions of the subspaces $\{d_i\}_{i=1}^M$ are known. A number of approaches have been devised to solve the problem above or some of its special cases. They are based on sparsity methods [13–18], algebraic methods [19,20], iterative and statistical methods [2,3,10,21–24], and spectral clustering methods [14,15,25–32].

1.1. Paper contributions

- We develop an algebraic method for solving the general subspace segmentation problem for noiseless data. For the case where all the subspaces are four dimensional, Gear observed, without proof, that the reduced echelon form can be used to segment motions in videos [33]. In this paper, we develop this idea and prove that the reduced row echelon form can completely solve the subspace segmentation problem in its most general version. This is the content of Theorem 3.7 in Section 3.1.
- For noisy data, the reduced echelon form method does not work, and a thresholding must be applied. However, the effect of the noise on the reduced echelon form method depends on the noise level and the relative positions of the subspaces. This dependence is analyzed in Section 3.2 and is contained in Theorems 3.9 and 3.11.
- When the dimensions of the subspaces is equal and known, we relate the subspace segmentation problem to the non-linear approximation problem (Problem 1). The existence of a solution as well as an iterative search algorithm for finding the solution is presented in Theorem 2.1. This algorithm works well with noisy data but requires a good initial condition to locate the global minimum.
- The reduced echelon form together with the iterative search algorithm are combined together: A thresholded reduced echelon form algorithm provides the initial condition to the iterative search algorithm. This is done in Section 4.
- In Section 5, the algorithms are tested on synthetic and real data to evaluate the performance and limitations of the methods.

2. Non-linear approximation formulation of subspace segmentation

When M is known, the subspace segmentation problem, for both the finite and infinite dimensional space cases, can be formulated as follows:

Let $\mathscr{B}$ be a Banach space, $\mathbf{W} = \{w_1, \ldots, w_N\}$ a finite set of vectors in $\mathscr{B}$ that correspond to observed data. For $i = 1, \ldots, M$, let $\mathscr{C} = C_1 \times C_2 \times \cdots \times C_M$ be the Cartesian product of M families C_i of closed subspaces of $\mathscr{B}$ each containing the trivial subspace $\{0\}$. Thus, an element $\mathbf{S} \in \mathscr{C}$ is a sequence $\{S_1, \ldots, S_M\}$ of M subspaces of $\mathscr{B}$ with $S_i \in C_i$. For example, when each C_i is the family of all subspaces of dimensions less than or equal to d in the ambient space $\mathscr{B} = \mathbb{R}^D$, then an element $\mathbf{S} \in \mathscr{C}$ is a set of M subspaces $S_i \subset \mathbb{R}^D$, with dimensions $\dim S_i \leqslant d$. Another example is the infinite dimensional case in which $\mathscr{B} = L^2(\mathbb{R})$ and each C_i is a family of closed, shift-invariant subspaces of $L^2(\mathbb{R})$ that are generated by at most $r < \infty$ generators. For example if $r = 1$, $M = 2$, an element $\mathbf{S} \in \mathscr{C}$ may be the subspace S_1 of all bandlimited functions (generated by integer shifts of the generator function $\operatorname{sinc}(x) = \sin(x)/x$), and S_2 the shift invariant space generated by the B-spline functions β^n of degree n. In these cases the subspaces in $S_i \in C_i$ are also infinite dimensional subspaces of L^2.

Problem 1.

1. Given a finite set $\mathbf{W} \subset \mathscr{B}$, a fixed p with $0 < p \leqslant \infty$, and a fixed integer $M \geqslant 1$, find the infimum of the expression