Neurocomputing 000 (2017) 1–8

Contents lists available at ScienceDirect

Neurocomputing

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

Robust subspace clustering via penalized mixture of Gaussians

Jing Yao[1], Xiangyong Cao[1], Qian Zhao*, Deyu Meng, Zongben Xu

School of Mathematics and Statistics and Ministry of Education Key Lab of Intelligent Networks and Network Security, Xi'an Jiaotong University, Xian 710049, PR China

ARTICLE INFO

Article history:
Received 12 October 2016
Revised 5 April 2017
Accepted 21 May 2017
Available online xxx

Keywords:
Subspace clustering
Low-rank representation
Mixture of Gaussians
Expectation maximization

ABSTRACT

Many problems in computer vision and pattern recognition can be posed as learning low-dimensional subspace structures from high-dimensional data. Subspace clustering represents a commonly utilized subspace learning strategy. The existing subspace clustering models mainly adopt a deterministic loss function to describe a certain noise type between an observed data matrix and its self-expressed form. However, the noises embedded in practical high-dimensional data are generally non-Gaussian and have much more complex structures. To address this issue, this paper proposes a robust subspace clustering model by embedding the Mixture of Gaussians (MoG) noise modeling strategy into the low-rank representation (LRR) subspace clustering model. The proposed MoG-LRR model is capitalized on its adapting to a wider range of noise distributions beyond current methods due to the universal approximation capability of MoG. Additionally, a penalized likelihood method is encoded into this model to facilitate selecting the number of mixture components automatically. A modified Expectation Maximization (EM) algorithm is also designed to infer the parameters involved in the proposed PMoG-LRR model. The superiority of our method is demonstrated by extensive experiments on face clustering and motion segmentation datasets.

1. Introduction

With dramatic development of techniques on data collection and feature extraction in recent years, the data obtained from real applications are often with high dimensionality. Such high-dimensional characteristic of data not only tends to bring large computation burden to the later data processing implementation, but also might possibly degenerates the performance of the utilized data process technique due to the curse of dimensionality issue.

An effective strategy to alleviate this problem is to find the intrinsic low-dimensional subspace where such high-dimensional data intrinsically reside, and then make implementations on the low-dimensional projections of data. This is always feasible in real scenarios since features of practically collected data are generally with evident correlations. Such a "subspace learning" methodology has attracted much attention in the past decades and various related methods have been proposed for different machine learning, computer vision and pattern recognition tasks [1–3].

In the recent years, a new trend on subspace learning, called subspace clustering, has appeared by simultaneously clustering data and extracting multiple subspaces, each corresponding to one data cluster [4–6]. Compared with traditional methods which assume data lie on a unique low-dimensional subspace, such "subspace clustering" model better complies with many real scenarios where data are located on multiple subspace clusters. Typical applications include face clustering [7], image segmentation [8], metric learning [9], feature grouping [10] and image representation [11]. Accordingly this research has been attracting increasing attention in the recent years.

Now let's introduce the formal definition for the subspace clustering problem as below [4]:

Definition 1.1 (Subspace Clustering)**.** Given a set of sampled data $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_k] = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ drawn from a union of k subspaces $\{\mathcal{S}_i\}_{i=1}^{k}$, where $\mathbf{X}_i$ denotes a collection of n_i samples drawn from the subspace $\mathcal{S}_i$, and $n = \sum_{i=1}^{k} n_i$. The task of subspace clustering is to cluster the samples according to the underlying subspaces they are drawn from.

The main assumption underlying the subspace clustering is that each datum is sampled from one of several low-dimensional subspaces, and hence can be well represented as a linear combination of the other data from the same subspace. The representation matrix composed of all coefficients of such combinations is

* Corresponding author.
E-mail addresses: jasonyao92@gmail.com (J. Yao), caoxiangyong45@gmail.com (X. Cao), timmy.zhaoqian@mail.xjtu.edu.cn, timmy.zhaoqian@gmail.com (Q. Zhao), dymeng@mail.xjtu.edu.cn (D. Meng), zbxu@mail.xjtu.edu.cn (Z. Xu).
[1] 1 indicates equal contribution

http://dx.doi.org/10.1016/j.neucom.2017.05.102
0925-2312/

firstly taken as the encoding representation for the intrinsic subspace clusters and then is utilized to extract subspace knowledge from data. This subspace learning task is mathematically formulated as follows:

$$\min_{\mathbf{C}} \mathcal{R}(\mathbf{C}) + \lambda \mathcal{L}(\mathbf{X} - \tilde{\mathbf{X}}\mathbf{C}), \quad (1)$$

where $\tilde{\mathbf{X}}$ is a predefined dictionary, the first term regularize the representation matrix $\mathbf{C}$ to encode the subspace clustering prior knowledge in it, the second term is the loss function to fit the additive noise and λ is a positive trade-off parameter.

A natural choice for the loss function $\mathcal{L}(\cdot)$ in Eq. (1) is the $\|\cdot\|_F$ norm, which, from the probabilistic perspective, mainly characterize Gaussian noise. However, real noises in applications are generally non-Gaussian, and thus other types of loss functions were considered, such as $\|\cdot\|_1$ norm and $\|\cdot\|_{2,1}$ norm[2], corresponding to Laplacian noise and sample specific Gaussian noise, respectively [12]. These noise assumptions, however, still have limitations, since data noise in practical applications often exhibits much more complex statistical structures [13–16]. Therefore, a pre-fixed simple loss term is generally incapable of well fitting practical noise in data. The clustering accuracy of the utilized subspace clustering method [6] thus tends to be negatively influenced by this improper assumption. In this sense, it's crucial to propose a robust subspace clustering model to tackle complex noise.

To address this issue, this paper presents a novel subspace clustering method that is robust against a wider range of noise distributions beyond traditional Gaussian, Laplacian or sample Gaussian noises. Our basic idea is to encode data noise as Mixture of Gaussians (MoG) in the subspace clustering model, and simultaneously extract subspace cluster knowledge and adapt data noise. Here we prefer to employ MoG as the noise model due to its universal approximation property to any continuous distribution [17]. Such idea is inspired by some recent noise modeling methodology designed on some typical machine learning and computer vision tasks, such as low-rank matrix factorization (LRMF) [18], robust principle component analysis (RPCA) [19] and tensor factorization [20], and has been substantiated to be effective in the complicated noise scenarios. There exist multiple subspace clustering approaches recently, such as Sparse Subspace Clustering (SSC) [5], Low-Rank Representative (LRR) [6] and Least Square Regression (LSR) [21]. In this work we readily adopt the LRR model [6,22] due to its utilization of clean data as the dictionary matrix, instead of taking original corrupted ones as most others, which better complies with the insight of subspace clustering. In addition, regarding the automatic selection of mixture Gaussian components, we propose a penalized MoG-LRR model through adopting a penalized likelihood technique inspired by [23,24]. We further design an effective Expectation Maximization (EM) algorithm to infer all parameters involved in this model. The superiority of the proposed method is substantiated on face clustering and motion segmentation problems as compared with the current state-of-the-art methods on subspace clustering.

Specifically, the contribution of this work can be summarized as follows:

- On subspace clustering, we integrated MoG noise modeling methodology into the LRR model, which enhances a robust subspace clustering strategy with capability of adaptively fitting a wide range of data noises beyond current methods.
- On MoG noise modeling methodology, through employing the penalized likelihood technique, the EM algorithm designed on the proposed MoG-LRR model is capable of automatically selecting a proper Gaussian mixture component number as well as other involved parameters in this model. This also prompts the frontier of noise modeling and makes it easy for the selection of this important parameter.

[2] The three norms are calculated as $\|\cdot\|_F = \sum_{j=1}^N(\sum_{i=1}^D(\cdot)_{ij}^2)$, $\|\cdot\|_1 = \sum_{j=1}^N(\sum_{i=1}^D|(\cdot)_{ij}|)$ and $\|\cdot\|_{2,1} = \sum_{j=1}^N(\sum_{i=1}^D(\cdot)_{ij}^2)^{\frac{1}{2}}$, respectively.

Table 1
Some utilized notations in this paper.

Notation	Definition
k	Number of subspaces
N	Data size
D	Data dimensionality
$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$	Observed data
$\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N]$	Clean data
$\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_N]$	Representation matrix
$\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_N]$	Sample-wise noise
$\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$	Mixture proportions
$\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$	Covariance matrices

Our paper is organized as follows. Related work is introduced in Section 2. Section 3 proposes our model called *Penalized Mixture of Gaussians Low-Rank Representation* (PMoG-LRR) and then presents a modified EM algorithm for solving this model. Section 4 presents experimental results implemented on synthetic and real data sets to substantiate the superiority of our proposed method over other state-of-the-arts. Finally, a brief conclusion is drawn in Section 5. Throughout the paper, we denote scalars, vectors, matrices as the non-bold, bold lower case and bold upper case letters, respectively. Some notations used in this paper are summarized in Table 1.

2. Related work

The past two decades has witnessed a rapid development in the field of subspace clustering. The related methods can be roughly classified into four categories: algebraic methods, iterative methods, statistical methods, and spectral-clustering-based methods [4].

Algebraic methods, typically represented by Matrix Factorization-based methods [25–27], first find a permutation matrix and calculate the multiplication matrix of data and the permutation matrix, and then factorize this multiplication matrix into two rank-r matrices, a base-matrix and a block diagonal matrix, respectively. But these methods are generally sensitive to noise and require the knowledge of the rank r of data matrix. The iterative methods, e.g., K-subspaces [28] use an iterative way to model and segment data. Specifically, such methods first assign data to pre-defined multiple subspaces, and then update the subspaces and reassign each data point to the closest subspace. The drawback of above two methods is that they incline to be sensitive to initialization and outliers. Besides, they need to know the number of subspace and their corresponding dimensions in advance. The statistical methods, e.g., Mixtures of Probabilistic PCA (MPPCA) [29], assume that the sample data are generated from a Mixture of Gaussians (MoG) distribution and then uses the Expectation Maximization (EM) algorithm to update the data segmentation and model parameters alternatively under the Maximum Likelihood Estimation (MLE) framework. One disadvantage of these methods is that the model always cannot fit the cases that the intrinsic distributions of the data inside each subspace are not Gaussian.

Recently, spectral-clustering-based subspace clustering methods has been attracting more attention [5,6,21] due to its rational methodology and successful performance in applications [30]. The fundament of these methods is to assume that each data point can be linearly represented by all the other data points from the same subspace cluster. These methods generally contain two steps. Firstly, an affinity matrix is built to capture the similarity between pairs of data points. Then, the segmentation of data is obtained by applying spectral clustering algorithm [31] to the affinity matrix.