Journal of Computational Mathematics
Vol.36, No.3, 2018, 426–440.

http://www.global-sci.org/jcm
doi:10.4208/jcm.1708-m2016-0830

REDUCED-RANK MODELING FOR HIGH-DIMENSIONAL MODEL-BASED CLUSTERING*

Lei Yang
Department of Environmental Medicine, New York University, New York, NY, USA
Email: ly888@nyu.edu
Junhui Wang
Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong
Email: j.h.wang@cityu.edu.hk
Shiqian Ma
Department of Mathematics, University of California, Davis, CA 95616, USA
Email: sqma@math.ucdavis.edu

Abstract

Model-based clustering is popularly used in statistical literature, which often models the data with a Gaussian mixture model. As a consequence, it requires estimation of a large amount of parameters, especially when the data dimension is relatively large. In this paper, reduced-rank model and group-sparsity regularization are proposed to equip with the model-based clustering, which substantially reduce the number of parameters and thus facilitate the high-dimensional clustering and variable selection simultaneously. We propose an EM algorithm for this task, in which the M-step is solved using alternating minimization. One of the alternating steps involves both nonsmooth function and nonconvex constraint, and thus we propose a linearized alternating direction method of multipliers (ADMM) for solving it. This leads to an efficient algorithm whose subproblems are all easy to solve. In addition, a model selection criterion based on the concept of clustering stability is developed for tuning the clustering model. The effectiveness of the proposed method is supported in a variety of simulated and real examples, as well as its asymptotic estimation and selection consistencies.

Mathematics subject classification: 62-07, 90C30
Key words: Clustering, Gaussian mixture model, Group Lasso, ADMM, Reduced-rank model.

1. Introduction

Cluster analysis is to assign observations into a number of clusters so that observations within the same cluster are more similar compared with those in different clusters. In literature, the similarity is often measured by certain pre-specified forms of distance, leading to various clustering algorithms such as K-means, hierarchical clustering, and spectral clustering [17]. On the contrary, model-based clustering [28] approaches cluster analysis from a probabilistic perspective. It models the data as a finite mixture of parametric distributions, and converts cluster analysis into estimation of the unknown parameters.

The model-based clustering is popular in literature, especially when the data dimension is small [13]. However, challenges arise when the data dimension becomes relatively large,

* Received January 8, 2017 / Revised version received May 3, 2017 / Accepted August 9, 2017 / Published online March 28, 2018 /

due to the large number of unknown parameters [8]. To circumvent the difficulty, several remedial treatments have been proposed. For instances, [14] and [16] propose to first conduct principal component analysis to reduce the dimension and then conduct model-based clustering on the low-dimensional space. [20] and [11] propose to conduct a forward stepwise or prescreening variable selection for the model-based clustering based on a BIC criterion. [18] restricts the covariance matrix of each Gaussian mixture component to be diagonal. Such treatments can substantially alleviate the computation cost, yet may be too restrictive to capture all the important information. More recently, [7] developed a sparse Fisher-EM algorithm, which combines the model-based clustering and Fisher's linear discriminant analysis and can conduct clustering in high-dimensional space.

In this paper, reduced-rank model and group-sparsity regularization are proposed to equip with the model-based clustering to tackle the high-dimensional cluster analysis and variable selection simultaneously. The key motivation is to project the original data onto a low-dimensional space so that the projected data follows a finite mixture of Gaussian distributions. A group Lasso penalty is enforced on each row of the projection matrix to assure that the non-informative variables can be automatically identified if they do not contribute to any of the reduced dimension. We propose an EM algorithm for this task. One of the steps in the EM algorithm involves both nonsmooth function and nonconvex constraint, and thus we propose a linearized ADMM for solving it. This leads to an efficient algorithm whose subproblems are all easy to solve. To optimally determine the tuning parameters, a stability-based selection criterion [24, 26] is employed, which searches for the tuning parameters that lead to the most stable clustering algorithm. The advantages of the proposed methods are demonstrated in a variety of the numerical examples, as well as their asymptotic estimation and selection consistencies.

The rest of the article is organized as follows. Section 2 introduces the reduced-rank Gaussian mixture model, as well as the computing algorithms including the expectation-maximization (EM) algorithm and the linearized ADMM algorithm. Section 3 presents the stability-based selection criterion for optimally determining the tuning parameters. Section 4 establishes the asymptotic estimation and selection consistencies of the proposed methods. Numerical simulation and real applications are presented in Section 5. Section 6 gives conclusions. The appendix is devoted to technical proofs and a counter example for the likelihood-based criteria.

2. Reduced-rank Model-based Clustering

In a typical clustering setup, the training set consists of n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ independently and identically distributed. The primary goal is to cluster $\mathbf{x}_i$'s into a set of subgroups so that the observations within the same subgroup are more similar than those in different subgroups.

2.1. Reduced-rank Gaussian mixture model

Classical model-based clustering models each subgroup by a distinct probability distribution, say Gaussian distribution, and assumes that the data follows

$$\mathbf{X} \sim \sum_{k=1}^{K} \pi_k N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$