Accepted Manuscript

On model-based clustering of skewed matrix data

Volodymyr Melnykov, Xuwen Zhu

PII: S0047-259X(17)30512-2
DOI: https://doi.org/10.1016/j.jmva.2018.04.007
Reference: YJMVA 4354

To appear in: *Journal of Multivariate Analysis*

Received date : 24 August 2017

Please cite this article as: V. Melnykov, X. Zhu, On model-based clustering of skewed matrix data, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.04.007

On model-based clustering of skewed matrix data

Volodymyr Melnykov[a,*], Xuwen Zhu[b]

[a]*Department of Information Systems, Statistics, and Management Science, The University of Alabama, Tuscaloosa, AL 35487, USA*
[b]*Department of Mathematics, University of Louisville, Louisville, KY 40292, USA*

Abstract

The existing finite mixture modeling and model-based clustering literature focuses primarily on the analysis of multivariate data observed in the form of vectors, with each element representing a specific feature. In this setting, multivariate Gaussian mixture models have been the most commonly used. Due to severe modeling issues observed when normal components cannot provide adequate fit to groups, much attention has been paid to developing models capable of accounting for skewness in data. In our work, we target the problem of mixture modeling with components that can handle skewness in matrix-valued data. The proposed developments open a wide range of possible modeling capabilities, with numerous applications, as illustrated in this paper. A novel matrix mixture model is proposed. Its skewness parameters enjoy appealing interpretability. The corresponding estimation procedure and various ways of parameterization are discussed. Comprehensive simulation studies and applications to real-life datasets illustrate the efficiency of the proposed developments, supported by good results.

Keywords: Cluster analysis, Matrix transformation, Mixture model, Skewness
2000 MSC: 62H30

1. Introduction

Finite mixture modeling [28] provides a powerful way of modeling heterogeneity in data by means of weighted averaging of several probability distributions. The specific form of the contributing distributions as well as the size of the associated weights define the resulting mixture. Model-based clustering assumes that each cluster can be seen as a sample from an underlying mixture component, i.e., there exists a one-to-one correspondence between components and data groups. The assignment of observations to clusters is usually done by Bayes' decision rule. One of the most popular choices for components is the multivariate Gaussian distribution. Unfortunately, the attractive one-to-one correspondence can cease to hold if the underlying distribution is misspecified and cannot provide an adequate fit for a data group. Hence, it is highly desirable to avoid model misspecifications.

As a result, one of the most actively developing directions of finite mixture modeling is related to distributions capable of modeling various data shapes, especially skewness. For this task, mixtures of multivariate skew-normal [9, 23] and skew-*t* [21, 22, 24] distributions are the most popular choices. Some other alternatives include mixtures of generalized hyperbolic distributions [8], shifted asymmetric Laplace distributions [15], normal-inverse Gaussian mixtures [32], and Manly mixture models [38]. It is worth mentioning that overly complex models often pose a problem of potential overparameterization which can lead to the severe underestimation of the mixture order [30]. To address this concern, researchers consider a variety of parsimonious models capable of preserving the desired modeling flexibility, with fewer parameters [4, 29].

Nearly all developments in finite mixture modeling involve vector-valued observations. However, matrix- and tensor-valued data occur in a wide variety of settings. Just a few examples of such cases include studies with repeated vector measurements, multivariate time series, spatial data, and multifactor data with several levels. Although some matrix distributions have been developed [2, 11, 20], many of them have rather complicated functional forms or have a limited use. The most recent proposal in this field is the matrix skew-*t* distribution proposed in [16].

*Corresponding author