Journal of Multivariate Analysis 125 (2014) 100–120

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Assessment of the number of components in Gaussian mixture models in the presence of multiple local maximizers

Daeyoung Kim [a], Byungtae Seo [b,*]

[a] Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003, USA
[b] Department of Statistics, Sungkyunkwan University, Seoul 110-745, Republic of Korea

ARTICLE INFO

Article history:
Received 18 October 2012
Available online 18 December 2013

AMS subject classifications:
62H30
62F10
62F12

Keywords:
Gaussian mixture
Maximum likelihood
Spurious local maximizer
Unbounded likelihood

ABSTRACT

Gaussian mixtures are very flexible in representing the underlying structure in the data. However, the likelihood inference for Gaussian mixtures with unrestricted covariance matrices is theoretically and practically challenging because the likelihood function is unbounded and often has multiple local maximizers. As shown in the numerical studies of this paper, the presence of multiple local maximizers including spurious local maximizers affects the performances of model selection criteria used to choose the number of components. In this paper we propose a new type of likelihood-based estimator, a gradient-based k-deleted maximum likelihood estimator, for Gaussian mixture models. The proposed estimator is designed to avoid spurious local maximizers and choose a statistically desirable local maximizer in the presence of multiple local maximizers. We first prove the consistency of the proposed estimator and then examine, by a real-data example and simulation studies, the performance of the proposed method in the likelihood-based model selection criteria commonly used to assess the number of components in Gaussian mixture models.

1. Introduction

Modern computing power and algorithms have greatly increased interest in the Gaussian mixture model as an effective tool for density estimation and model based clustering (see Lindsay [22], McLachlan and Peel [23], Fraley and Raftery [15] and Hennig [19] for an overview and references). Nevertheless, it is well known that the unbounded likelihood of the Gaussian mixture with unrestricted covariance matrices results in a few theoretical and practical difficulties for statistical inference. First, theoretical questions concerning the asymptotic properties of the maximum likelihood estimator (MLE) need to be carefully addressed as the usual regularity conditions for the consistency of the MLE do not hold. The expectation–maximization (EM) algorithm [11] for the parameter estimation is likely to diverge towards a singular solution with singular covariance matrix estimate(s). Second, the mixture likelihood has multiple local maximizers including spurious local maximizers [10,23,26] that would complicate statistical inferences if they were not dealt with properly.

Several researchers have proposed methods to avoid singularities, either by constraining the parameter space or by modifying the likelihood function. Hathaway [18] and Tanaka and Takemura [30] suggested a constrained ML estimation with constraint on scale parameters so that the estimators of scale parameters stay within a reasonable range. Ciuperca et al. [9], Chen et al. [8] and Chen and Tan [7] proposed a penalized maximum likelihood estimation method which modifies the likelihood instead of the parameter space to avoid singularity. It can also be considered as a Bayesian regularization method [16]. Seo and Lindsay [27] modified the likelihood by using double smoothing techniques to regularize the mixture likelihood.

All of these can resolve the singularity problem. However, a more challenging problem in real data analysis arises from the existence of spurious local maximizers in the likelihood for a Gaussian mixture model. McLachlan and Peel [23, p. 99]

* Corresponding author.
E-mail addresses: daeyoung@math.umass.edu (D. Kim), seobt@skku.edu (B. Seo).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.11.018

described a spurious local maximizer as "a local maximizer that lies very close to the edge of the parameter space, but with component-covariance matrices that are not actually singular, although they may be close to singular for some components". The spurious local maximizer often has a high likelihood and has one or more components which overfits a small random localized pattern in the data rather than any other underlying group structure. All methods to avoid singularity may lessen the issues associated with spurious local maximizers, but generally they rely too much on the degrees of tuning parameters.

Seo and Kim [26] extensively studied the problems associated with spurious local maximizers and proposed a new likelihood-based estimator called the *k-deleted MLE*, which is a new way of selecting a local maximizer in the presence of multiple local maximizers. The basic idea of their proposed methods is as follows. Suppose one obtains multiple local maximizers by employing multiple starting values in the EM algorithm. Given all found local maximizers, one first detects a few, say k, log-likelihood terms that are disproportionately influential on the formation of each local maximizer. Then, one recomputes a *k-deleted log-likelihood* at each local maximizer after removing the detected k log-likelihood terms from the ordinary log-likelihood, and defines the *k-deleted MLE* to be the local maximizer having the highest k-deleted log-likelihood value. They showed, by simulation studies and real examples, that using k-deleted MLE can greatly reduce the risk of choosing a problematic local maximizer that has the same features as spurious solutions.

Based on how one determines the unduly influential likelihood terms, Seo and Kim [26] suggested two methods, the likelihood- and score-based deletions. The former is using individual log-likelihood terms to identify the observations disproportionately contributing to the formation of a given local maximizer. That is, one constructs the *likelihood-based k-deleted log-likelihood* by removing the k largest log-likelihood terms from the ordinary log-likelihood and then compute the *likelihood-based k-deleted MLE*. The latter method uses the information of the individual score functions in the likelihood equations. In other words, one removes k log-likelihood terms whose corresponding observations have excessive influences on making the sum of all score functions evaluated at the local maximizer being equal to zero, obtain the *score-based k-deleted log-likelihood* and then compute the *score-based k-deleted MLE*. Their simulation studies showed that the k-deleted MLEs obtained from the two methods solve the problems associated with singularities, and the score-based k-deleted MLE is considerably superior to the likelihood-based k-deleted MLE when one concerns spurious local maximizers.

However, there are two issues to be addressed in the methodologies and results presented by Seo and Kim [26]. First, computing the score-based k-deleted MLE requires the derivation of the standardized score statistic at each observation given a local maximizer. This may not considerably increase the computing time compared to that of the EM iteration. However, such derivation would be painful especially for high dimensional data or mixtures with non-normal component densities. Moreover, the computing time with a large sample size or many local maximizers is not negligible. Second, the works reported therein assumed that the number of components is known a priori. When the number of components is unknown, we may use some likelihood-based information criteria to determine a suitable number of components. However, since likelihood-based information criteria rely heavily on the likelihood value at a chosen local maximizer, an incorrect choice of the local maximizer can also considerably influence those information criteria.

The contribution of this paper is twofold. First, we propose a new k-deleted MLE, *gradient-based k-deleted MLE*, which is intuitively more natural and computationally more efficient than the score-based k-deleted MLE while preserving similar performances as the score-based one. Second, to the best of our knowledge, there is no study to rigorously address the effects of spurious local maximizers on the likelihood-based model selection procedure used to choose the number of components in Gaussian mixture models. We investigate this along with the performances of the proposed gradient-based k-deleted MLE as well as other k-deleted MLEs in Seo and Kim [26] and the penalized maximum likelihood estimator (PMLE) in Chen and Tan [7].

This article is organized as follows: In Section 2 we provide detailed reviews on two k-deleted MLEs proposed by Seo and Kim [26]. Section 3 proposes a new k-deleted MLE in the Gaussian mixture model and asymptotic properties are studied in Section 4. In Section 5 we carry out a real data analysis as well as simulation studies to evaluate the performances of a newly proposed k-deleted MLE, two existing k-deleted MLEs [26] and the PMLE [7] in the four commonly used likelihood-based model selection criteria for the number of components in Gaussian mixtures with unrestricted covariance matrices, the Akaike's information criterion (AIC) [1], the Bayesian information criterion (BIC) [24], the consistent Akaike's information criterion (CAIC) [5] and the Integrated Completed Likelihood (ICL)-BIC [2]. We then end this article with a discussion in Section 6.

2. Background on *k*-deleted MLEs in the Gaussian mixture model

Let $\mathbf{X}$ be a p-dimensional random vector having a m-component p-variate Gaussian mixture distribution with unrestricted covariance matrices,

$$f(\mathbf{X};\boldsymbol{\theta}) = \sum_{j=1}^{m} p_j \phi\left(\mathbf{X};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right), \tag{2.1}$$

where $p_1,\ldots,p_m$ are mixing weights with $0 < p_j < 1$ and $\sum_{j=1}^{m} p_j = 1$, and $\phi(\mathbf{X};\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the p-dimensional Gaussian density with $p\times 1$ mean vector $\boldsymbol{\mu}$ and $p\times p$ covariance matrix $\boldsymbol{\Sigma}$:

$$\phi(\mathbf{X};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right\}.$$