Statistics and Probability Letters 83 (2013) 2088–2093

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

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

Dimension reduction for model-based clustering via mixtures of shifted asymmetric Laplace distributions

Katherine Morris, Paul D. McNicholas *

Department of Statistics, University of Guelph, 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada

ARTICLE INFO

Article history:
Received 20 December 2012
Received in revised form 27 March 2013
Accepted 15 April 2013
Available online 1 May 2013

Keywords:
Asymmetric Laplace
Dimension reduction
Mixture models
Model-based clustering

ABSTRACT

A dimension reduction method for model-based clustering via a finite mixture of shifted asymmetric Laplace distributions is introduced. The approach is based on existing work within the Gaussian paradigm and relies on identification of a reduced subspace. This subspace contains linear combinations of the original data, ordered by importance using the associated eigenvalues. This clustering approach is illustrated on simulated and real data, where it performs favourably compared to its Gaussian analogue.

1. Introduction

A dimension reduction method for model-based clustering via shifted asymmetric Laplace (SAL) mixtures is introduced. This approach is analogous to an existing approach for Gaussian mixtures (Scrucca, 2010), and works by looking for a subspace that captures most of the clustering structure within the data. This subspace is found based on the variation in group means and group covariances, and contains linear combinations of the original data ordered by importance through the associated eigenvalues.

The 'model-based' approach to clustering assumes an underlying finite mixture model. A p-dimensional random vector $\boldsymbol{X}$ is said to arise from a parametric finite mixture distribution if we have $f(\boldsymbol{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\boldsymbol{x}|\boldsymbol{\theta}_g)$, where G is the number of components, π_g are mixing proportions, so that $\sum_{g=1}^{G} \pi_g = 1$ and $\pi_g > 0$, and $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is the parameter vector. Until recently, Gaussian mixtures have dominated the model-based clustering literature. The likelihood for $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from a Gaussian mixture model is $\mathcal{L}(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, where $\phi(\boldsymbol{x}_i|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$.

Over the last few years, work on non-Gaussian model-based clustering has gained momentum. Space does not permit an exhaustive listing here, but suffice it to say that the quantity of work on non-Gaussian model-based clustering is approaching, or has perhaps recently exceeded, that on its Gaussian precursors. Of particular relevance to work described herein is the mixture of SAL distributions (Franczak et al., 2012): the density is $f(\boldsymbol{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \, \xi(\boldsymbol{x}|\boldsymbol{\alpha}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\mu}_g)$, where

$$\xi(\boldsymbol{x}|\boldsymbol{\alpha}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\mu}_g) = \frac{2 \exp\{(\boldsymbol{x}-\boldsymbol{\mu}_g)^\top \boldsymbol{\Sigma}_g^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_g)\}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_g|^{1/2}} \left(\frac{\delta(\boldsymbol{x}, \boldsymbol{\mu}_g|\boldsymbol{\Sigma}_g)}{2+\boldsymbol{\alpha}_g^\top \boldsymbol{\Sigma}_g^{-1}\boldsymbol{\alpha}_g}\right)^{\nu/2} K_\nu(u) \tag{1}$$

* Corresponding author.
E-mail address: pmcnicho@uoguelph.ca (P.D. McNicholas).

0167-7152/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.spl.2013.04.011

is the density of a multivariate SAL distribution, $\boldsymbol{\mu}_g$ is the location, $\boldsymbol{\Sigma}_g$ is the scale matrix, and $\boldsymbol{\alpha} \in \mathbb{R}^p$ denotes the skewness. Here, $\delta(\boldsymbol{x}, \boldsymbol{\mu}_g|\boldsymbol{\Sigma}_g)$ is the squared Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}_g$, $u = \sqrt{(2 + \boldsymbol{\alpha}_g^\top \boldsymbol{\Sigma}_g^{-1}\boldsymbol{\alpha}_g)\delta(\boldsymbol{x}, \boldsymbol{\mu}_g|\boldsymbol{\Sigma}_g)}$, and K_ν is the modified Bessel function of the third kind with index $\nu = (2-p)/2$. Note that although $\boldsymbol{\Sigma}_g$ is a covariance matrix, it is not the covariance matrix of the random variable $\boldsymbol{X}$ with the density in (1), save for the special case where $\boldsymbol{\alpha} = \boldsymbol{0}$; the covariance of $\boldsymbol{X}$ is $\tilde{\boldsymbol{\Sigma}}_g := \boldsymbol{\Sigma}_g + \boldsymbol{\alpha}\boldsymbol{\alpha}^\top$. Further details on the role of $\boldsymbol{\Sigma}_g$ in the SAL density are given by Franczak et al. (2012). Parameter estimation for SAL models is carried out *via* the expectation–maximization (EM) algorithm (Dempster et al., 1977).

SAL mixtures are amongst a few methods that show promise for clustering data with asymmetric clusters. Franczak et al. (2012) give real and simulated data examples showing that SAL mixtures can outperform Gaussian mixtures when applied to such data; furthermore, they illustrate that the inferior performance of Gaussian mixtures in these cases cannot necessarily be mitigated by merging components. This point will be also be illustrated in relation to the methodology introduced herein (cf. Section 3).

The remainder of the paper is laid out as follows. Our dimension reduction clustering method is presented (Section 2). In Section 3, we apply our algorithm to simulated and real data sets, and compare the performance of our method to its Gaussian analogue and several other clustering methods. The paper concludes with discussion and suggestions for future work (Section 4).

2. Methodology

We introduce a dimension reduction for model-based clustering via SAL mixtures. This is analogous to the Gaussian mixture modelling and dimension reduction (GMMDR) approach of Scrucca (2010), which uses the MCLUST (Fraley and Raftery, 1999) family of models. In short, following the sliced inverse regression work of Li (1991, 2000), GMMDR looks for the smallest subspace that captures the clustering information contained within the data. To do this, we seek those directions where the cluster means $\tilde{\boldsymbol{\mu}}_g := \boldsymbol{\mu}_g + \boldsymbol{\alpha}_g$ and the cluster covariances $\tilde{\boldsymbol{\Sigma}}_g$ vary the most, provided that each direction is $\boldsymbol{\Sigma}$-orthogonal to the others. These variations are captured by $\boldsymbol{M}_{\text{I}}$ and $\boldsymbol{M}_{\text{II}}$ below and we find the subspace via the generalized eigendecomposition of the kernel matrix $\boldsymbol{M}$:

$$\boldsymbol{M}\boldsymbol{v}_i = l_i \boldsymbol{\Sigma}\boldsymbol{v}_i, \tag{2}$$

where $l_1 \geq l_2 \geq \cdots \geq l_d > 0$ and $\boldsymbol{v}_i^\top \boldsymbol{\Sigma}\boldsymbol{v}_j = 1$ if $i = j$ and $\boldsymbol{v}_i^\top \boldsymbol{\Sigma}\boldsymbol{v}_j = 0$ otherwise. Here, $\boldsymbol{M} = \boldsymbol{M}_{\text{I}}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_{\text{I}} + \boldsymbol{M}_{\text{II}}$, $\boldsymbol{M}_{\text{I}} = \sum_{g=1}^G \pi_g(\tilde{\boldsymbol{\mu}}_g - \boldsymbol{\mu})(\tilde{\boldsymbol{\mu}}_g - \boldsymbol{\mu})^\top$, and $\boldsymbol{M}_{\text{II}} = \sum_{g=1}^G \pi_g(\tilde{\boldsymbol{\Sigma}}_g - \bar{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{\Sigma}}_g - \bar{\boldsymbol{\Sigma}})^\top$. Note that $\boldsymbol{\mu} = \sum_{g=1}^G \pi_g \tilde{\boldsymbol{\mu}}_g$ is the global mean, $\boldsymbol{\Sigma} = (1/n)\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$ is the global covariance matrix, and $\bar{\boldsymbol{\Sigma}} = \sum_{g=1}^G \pi_g \tilde{\boldsymbol{\Sigma}}_g$ is the pooled within-cluster covariance matrix.

Our approach will use SAL mixtures instead of the Gaussian mixtures used in GMMDR; accordingly, we will refer to it as SALMMDR. Outline details of the SALMMDR approach follow; it is analogous to the GMMDR approach and the extensive details given by Scrucca (2010) are not repeated here.

Definition 2.1. The SALMMDR directions are the eigenvectors $[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d] \equiv \boldsymbol{\beta}$, which form the basis of the dimension reduction subspace $\mathcal{S}(\boldsymbol{\beta})$ and are ordered based on eigenvalues.

The projections of the mean and covariance onto $\mathcal{S}(\boldsymbol{\beta})$ are then given by $\boldsymbol{\beta}^\top \tilde{\boldsymbol{\mu}}_g$ and $\boldsymbol{\beta}^\top \tilde{\boldsymbol{\Sigma}}_g \boldsymbol{\beta}$, respectively.

Definition 2.2. The SALMMDR variables, $\boldsymbol{Z}$, are the projections of the $n \times p$ data matrix $\boldsymbol{X}$ onto the subspace $\mathcal{S}(\boldsymbol{\beta})$ and can be computed as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\beta}$.

The estimation of the SALMMDR variables is akin to extraction of features, where the components are reduced through a set of linear combinations of the original variables. As with GMMDR, this set of features may contain estimated SALMMDR variables that provide no clustering information but require parameter estimation, and thus need to be removed.

Scrucca (2010) employed a modified version of the variable selection method of Raftery and Dean (2006) to filter the GMMDR features. We use the same approach to select the most appropriate SALMMDR features. This is carried out using the Bayesian information criterion (BIC; Schwarz, 1978):

$$\text{BIC} = 2l(\boldsymbol{x}, \hat{\boldsymbol{\vartheta}}) - r \log n,$$

where $l(\boldsymbol{x}, \hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood, $\hat{\boldsymbol{\vartheta}}$ is the maximum likelihood estimate of $\boldsymbol{\vartheta}$, r is the estimated number of free parameters, and n is the number of observations. While alternatives to the BIC exist, it remains the most popular mixture model selection criterion within the literature.

Thus, we compare two subsets of features, s and $s' = \{s \setminus i\} \subset s$, using

$$\begin{aligned} \text{BIC}_{\text{diff}}(Z_{i\in s}) &= \text{BIC}_{\text{clust}}(Z_s) - \text{BIC}_{\text{not clust}}(Z_s) \\ &= \text{BIC}_{\text{clust}}(Z_s) - [\text{BIC}_{\text{clust}}(Z_{s'}) + \text{BIC}_{\text{reg}}(Z_i|Z_{s'})], \end{aligned} \tag{3}$$

where $\text{BIC}_{\text{clust}}(Z_s)$ is the BIC value for the best clustering model fitted using features in s, $\text{BIC}_{\text{clust}}(Z_{s'})$ is the BIC value for the best clustering model fitted using features in s', and $\text{BIC}_{\text{reg}}(Z_i|Z_{s'})$ is the BIC value for the regression of the ith feature on the