Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Clustering, classification, discriminant analysis, and dimension reduction via generalized hyperbolic mixtures

Katherine Morris [a], Paul D. McNicholas [b,*]

[a] Department of Mathematics & Statistics, University of Guelph, Ontario, Canada, N1G 2W1

[b] Department of Mathematics & Statistics, McMaster University, 1280 Main St. W., Hamilton, Ontario, Canada, L8S 4L8

ARTICLE INFO

Article history:
Received 7 September 2013
Received in revised form 16 October 2015
Accepted 17 October 2015
Available online xxxx

Keywords:
Dimension reduction
Generalized hyperbolic distribution
Mixture models
Model-based clustering
Model-based classification
Model-based discriminant analysis

ABSTRACT

A method for dimension reduction with clustering, classification, or discriminant analysis is introduced. This mixture model-based approach is based on fitting generalized hyperbolic mixtures on a reduced subspace within the paradigm of model-based clustering, classification, or discriminant analysis. A reduced subspace of the data is derived by considering the extent to which group means and group covariances vary. The members of the subspace arise through linear combinations of the original data, and are ordered by importance via the associated eigenvalues. The observations can be projected onto the subspace, resulting in a set of variables that captures most of the clustering information available. The use of generalized hyperbolic mixtures gives a robust framework capable of dealing with skewed clusters. Although dimension reduction is increasingly in demand across many application areas, many applications are biological and so some of the real data examples are within that sphere. Simulated data are also used for illustration.

1. Introduction

A method for estimating a projection subspace basis derived from the fit of a generalized hyperbolic mixture (HMMDR) is introduced within the paradigms of model-based clustering, classification, and discriminant analysis. This is the most general case of work in this direction over the last few years, starting with an analogous approach based on Gaussian mixtures (GMMDR; Scrucca, 2010). Many dimension reduction methods summarize the information available through a reduced combination of the original variables. However, in terms of visualization, they do not always provide adequate information on the potential structure of the data at hand. The method proposed herein addresses this issue by revealing the underlying data clusters. At the same time, using heavy-tailed distributions, such as the generalized hyperbolic distribution, to model data can be advantageous because they assign appropriate weights to more extreme points (McNeil et al., 2005). The goal is to estimate a subspace that captures most of the clustering structure contained in the data. At the core of the method lies the sliced inverse regression (SIR) work of Li (1991, 2000), which reduces data dimensionality by considering the variation in group means to identify the subspace. Scrucca (2010) extended the SIR ideas to also include variation of group covariances. The members of the subspace arise through linear combinations of the original data, and are ordered by importance via their associated eigenvalues. The original observations in the data can be projected onto the subspace, resulting in a set of variables that captures most of the clustering information available.

The remainder of the paper is outlined as follows. Sections 2 and 3 present the background material. We then outline our dimension reduction method for selecting a reduced combination of the variables while retaining most of the clustering

* Corresponding author. Tel.: +1 905 525 9140x23419.
E-mail address: mcnicholas@math.mcmaster.ca (P.D. McNicholas).

http://dx.doi.org/10.1016/j.csda.2015.10.008
0167-9473/

Please cite this article in press as: Morris, K., McNicholas, P.D., Clustering, classification, discriminant analysis, and dimension reduction via generalized hyperbolic mixtures. Computational Statistics and Data Analysis (2015), http://dx.doi.org/10.1016/j.csda.2015.10.008

information contained within the data (Section 4). Then, the algorithm is applied to simulated (Section 5) and real (Section 6) data sets and the performance of our method is compared with its Gaussian and non-Gaussian analogues as well as with other subspace clustering techniques. Section 7 provides conclusions and suggestions for future work. Note that all computational work herein was carried out using R (R Core Team, 2013).

2. Finite mixture models

Modern data sets used in many practical applications have grown in size and complexity, compelling the use of clustering and classification algorithms based on probability models. The model-based approach assumes that data are generated by a finite mixture of probability distributions. A p-dimensional random vector $\mathbf{X}$ is said to arise from a parametric finite mixture distribution if its density is a convex set of probability densities, i.e.,

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x} \mid \boldsymbol{\theta}_g),$$

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. The $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ are called component densities and $f(\mathbf{x} \mid \boldsymbol{\vartheta})$ is formally referred to as a G-component parametric finite mixture distribution. The use of mixture models in clustering applications can be traced back a half-century to an application of Gaussian mixture models (Wolfe, 1963). Gaussian mixture model-based approaches have been very popular due to their mathematical tractability and, until recently, they dominated literature in the field. Extensive details on finite mixture models are given by Everitt and Hand (1981), McLachlan and Basford (1988), and McLachlan and Peel (2000).

In the past several years, non-Gaussian approaches to model-based clustering, classification, and discriminant analysis have flourished. This includes work on mixtures of multivariate t-distributions (Peel and McLachlan, 2000; Greselin and Ingrassia, 2010; Andrews et al., 2011; Steane et al., 2012; Andrews and McNicholas, 2012a; McNicholas, 2013; Lin et al., 2014), shifted asymmetric Laplace distributions (Franczak et al., 2014), skew-normal distributions (Lin, 2010), skew t-distributions (Vrbik and McNicholas, 2012, 2014; Lee and McLachlan, 2013, 2014; Murray et al., 2014a,b), variance-gamma distributions (McNicholas et al., 2014), multivariate normal-inverse Gaussian distributions (Karlis and Santourian, 2009; O'Hagan et al., 2016), and other approaches (e.g., Browne et al., 2012; Tortora et al., in press; Dang et al., in press). Mixtures of generalized hyperbolic distributions (Browne and McNicholas, 2015) are particularly relevant to work described herein. While it is not feasible to provide an exhaustive listing here, suffice it to say that the breadth of research on non-Gaussian model-based clustering and classification is becoming as rich as that of its Gaussian precursor.

Generalized hyperbolic distributions were introduced by Barndorff-Nielsen (1977) and used to model eolian sand deposits, i.e., sand deposits arising from the action of wind. The name of the distribution was derived from the fact that its log-density has the shape of a hyperbola. Properties of generalized hyperbolic densities were discussed in Barndorff-Nielsen and Halgreen (1977), Blæsild (1978) and, more recently, mixtures of these distributions appeared in McNeil et al. (2005) and Härdle and Simar (2011). Generalized hyperbolic distributions can effectively model extreme values, making them very useful in the context of financial and risk management applications, where the normal distribution does not offer a good description of reality. The multivariate generalized hyperbolic family is extremely flexible and contains many special and limiting cases, such as the inverse Gaussian, Laplace, and skew-t distributions.

Browne and McNicholas (2015) propose a multivariate generalized hyperbolic mixture model (HMM),

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_{\mathrm{h}}(\mathbf{x} \mid \lambda_g, \omega_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\alpha}_g), \tag{1}$$

where $\pi_g > 0$, with $\sum_{g=1}^{G} \pi_g = 1$, are the mixing proportions and the gth component density is

$$f_{\mathrm{h}}(\mathbf{x} \mid \lambda_g, \omega_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\alpha}_g) = \left[\frac{\omega_g + \delta(\mathbf{x}, \boldsymbol{\mu}_g \mid \boldsymbol{\Sigma}_g)}{\omega_g + \boldsymbol{\alpha}_g^{\top}\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\alpha}_g}\right]^{(\lambda_g - p/2)/2} \times \frac{K_{\lambda_g - p/2}\left(\sqrt{[\omega_g + \boldsymbol{\alpha}_g^{\top}\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\alpha}_g)(\omega_g + \delta(\mathbf{x}, \boldsymbol{\mu}_g \mid \boldsymbol{\Sigma}_g)]}\right)}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}_g\right|^{1/2} K_{\lambda_g}(\omega_g)\exp(-(\mathbf{x} - \boldsymbol{\mu}_g)^{\top}\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\alpha}_g)}, \tag{2}$$

with index parameter λ_g, concentration parameter $\boldsymbol{\omega}_g$, skewness parameter $\boldsymbol{\alpha}_g$, location $\boldsymbol{\mu}_g$, and scale matrix $\boldsymbol{\Sigma}_g$. Here, $\delta(\mathbf{x}, \boldsymbol{\mu}_g \mid \boldsymbol{\Sigma}_g) = (\mathbf{x} - \boldsymbol{\mu}_g)^{\top}\boldsymbol{\Sigma}_g^{-1}(\mathbf{x} - \boldsymbol{\mu}_g)$ is the squared Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}_g$ and K_{λ_g} denotes the modified Bessel function of the third kind with index λ_g.

The evaluation of modified Bessel functions in the density (2) sometimes leads to numerical overflow or underflow. To avoid these issues, we use asymptotic expansions from Abramowitz and Stegun (1972), i.e., for large x or λ,

$$K_{\lambda}(\lambda x) = \sqrt{\frac{\pi}{2\lambda}}\frac{\exp\{-\lambda\rho\}}{(1 + x^2)^{1/4}}\left[1 + \sum_{k=1}^{\infty}(-1)^k\frac{u_k(\tau)}{\lambda^k}\right],$$

Please cite this article in press as: Morris, K., McNicholas, P.D., Clustering, classification, discriminant analysis, and dimension reduction via generalized hyperbolic mixtures. Computational Statistics and Data Analysis (2015), http://dx.doi.org/10.1016/j.csda.2015.10.008