ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Manly transformation in finite mixture modeling

Xuwen Zhu, Volodymyr Melnykov *

Department of Information Systems, Statistics, and Management Science, The University of Alabama, Tuscaloosa, AL 35487, USA

ARTICLE INFO

Article history:
Received 6 June 2015
Received in revised form 26 January 2016
Accepted 27 January 2016
Available online xxxx

Keywords:
Finite mixture model
Manly transformation
Model-based clustering
Skewness
Classification

ABSTRACT

Finite mixture modeling is one of the most rapidly developing areas of statistics due to its modeling flexibility and appealing interpretability. Gaussian mixture models have been popular among researchers for decades proving their usefulness in various applications. However, when Gaussian mixture components do not provide an adequate fit for the data, more general models must be considered. Traditional remedies for deviation from normality include employing a more appropriate distribution as well as transforming data to near-normality. Merging both approaches by introducing a mixture model with components derived from the multivariate Manly transformation is proposed. Such mixture models show good performance in modeling skewness and have excellent interpretability. Forward and backward model selection algorithms are proposed to choose an appropriate multivariate transformation. At each step of these algorithms, a model with the specific combination of skewness parameters is estimated by means of the expectation–maximization algorithm. The developed technique is carefully illustrated on synthetic data and applied to several well-known datasets, with promising results.

1. Introduction

Finite mixture models assume a linear combination of several probability distributions, usually called components, with weights following some finite discrete distribution. Nowadays, mixture models are common in finance (Dinov, 2008), medical science (Schlattmann, 2009), transportation (Park and Lord, 2009), web-analytics (Melnykov, 2016), and many other areas. Their first documented use was in 1894 by Pearson (1894) who employed a two-component Gaussian mixture to analyze data collected on 1000 crabs. Since then, finite mixtures have attracted attention of many researchers but their golden age began several decades ago, at least partially due to studies devoted to the relation between mixture models and cluster analysis (Banfield and Raftery, 1993; Bensmail et al., 1997). Model-based clustering is an application of finite mixture modeling that assumes the existence of a one-to-one correspondence between mixture components and data groups. This appealing association implies that each probability distribution involved in the mixture is responsible for modeling a particular cluster. If this is not the case and more than one component is needed to fit each group adequately, the application of model-based clustering techniques encounters serious issues. To provide a remedy, some work has been done in the area of merging mixture components for clustering (Baudry et al., 2010; Hennig, 2010; Melnykov, in press; Scrucca, 2016) but a more intuitive approach is based on finding appropriate mixture components, capable of establishing the one-to-one association.

Among various mixture models, mixtures with all components assumed to be normally distributed are the most commonly used by practitioners. This can be explained by the existence of well-developed statistical theory for Gaussian distributions as well as the fact that normal distributions arise naturally in many applications. Despite these facts, there are many

* Correspondence to: The University of Alabama, Alston Hall 346, Tuscaloosa, AL 35487, USA. Tel.: +1 2053486292.
E-mail address: vmelnykov@ua.edu (V. Melnykov).

http://dx.doi.org/10.1016/j.csda.2016.01.015
0167-9473/

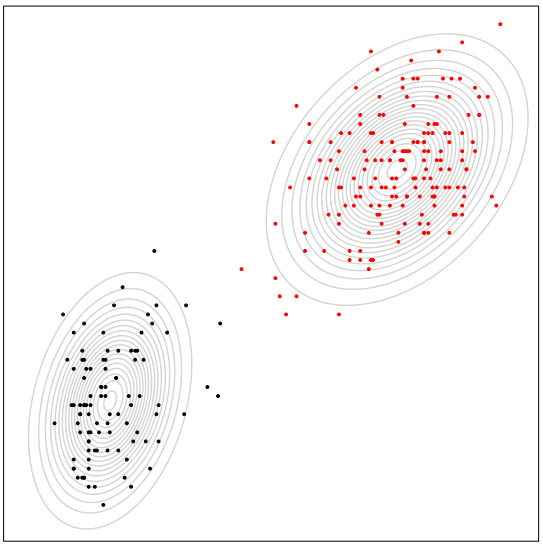

Fig. 1. A contour plot representing a Gaussian mixture model fit for the dataset *Faithful*. The clustering result is displayed through black and red data points. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

situations when observed data deviate from normality, for example, due to skewness. In such cases, a choice of components capable of modeling skewness can provide a better fit. Mixtures of skew-normal (Lin, 2009; Cabral et al., 2012) and skew-*t* (Lee and McLachlan, 2013b; Lee and MacLachlan, 2014) distributions are immediate candidates for this task. For more information on the families of these distributions and their application to model-based clustering, the reader can be referred to Lee and McLachlan (2013a) and Vrbik and McNicholas (2014). Some other models capable of accounting for skewness include mixtures of shifted asymmetric Laplace distributions (Franczak et al., 2014) and mixtures of generalized hyperbolic distributions (Giorgi and McNeil, 2014; Browne and McNicholas, 2015). Another possible alternative is based on finding a transformation that leads to near-normality in transformed data.

To motivate the problem discussed in this paper, the dataset *Faithful* (Azzalini and Bowman, 1990), publicly available in R, is considered. It consists of the duration of eruptions and waiting times between eruptions of the Old Faithful geyser in the Yellowstone National Park, WY. Fig. 1 provides a Gaussian mixture fit along with the corresponding model-based clustering solution shown through black and red points. It can be seen that the Gaussian mixture model fit does not capture the spread of data points very well. Numerous points gather at the left-hand side of the black cluster. At the same time, the black cluster is stretched toward the red data group located in the top right-hand corner. A similar effect can be observed for the red cluster. This example illustrates the situation when Gaussian components do not provide an entirely satisfactory fit to the data.

Merging two ideas of handling non-normality in data, *i.e.*, finding a more appropriate distribution and transforming the original data to near-normality, is proposed. Among all transformations, the Box–Cox power transformation (Box and Cox, 1964) proposed in 1964 is the most well-known for being effective in getting approximately normal distributions. It has great value to the history of statistics and has proven to be useful in many fields such as social science (Osborne, 2010), medical research (Hou et al., 2011), as well as computer science (Sylvia and Santoso, 2014). Some criticism related to the Box–Cox transformation is due to its incapability of handling left-skewed data and restriction to the set of positive numbers as the range of possible values. Since the invention of the Box–Cox transformation, some alternative approaches have been proposed. The Manly exponential transformation (Manly, 1976) has several advantages over the Box–Cox procedure. It can be applied to data in the range from $-\infty$ to ∞ and is proven to be equally effective with right- and left-skewed data (Sakia, 1992; Chortirat et al., 2011).

A novel mixture component based on the Manly back-transformation of a multivariate normal distribution is developed. Additional transformation parameters allow modeling skewness effectively, with great interpretability of parameters and components due to the direct connection to Gaussian distributions. Similar ideas have been studied in the context of the Box–Cox transformation (Lo et al., 2008; Lo and Gottardo, 2012; Lee et al., 2009). The authors assumed a global transformation parameter for all coordinates and components. In addition to the shortcomings of the Box–Cox transformation mentioned earlier, the immediate criticism of such an approach is the fact that transformations of different components in various dimensions cannot be identical. While the authors successfully applied the global transformation parameter to the flow cytometry data, its general application is limited due to restrictive assumptions.

In Section 2, the probability density function of the proposed model is introduced and the corresponding maximum likelihood estimator (MLE) is derived. The forward and backward transformation selection methods are proposed to relax possible overparameterization. The proposed techniques are illustrated, evaluated, and compared with their competitors in Section 3. In Section 4, several real-world datasets are analyzed. Section 5 concludes the paper with a discussion that summarizes main results.

Please cite this article in press as: Zhu, X., Melnykov, V., Manly transformation in finite mixture modeling. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.01.015