Computational Statistics and Data Analysis 93 (2016) 18–30

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Clustering with the multivariate normal inverse Gaussian distribution☆

Adrian O'Hagan [a,*], Thomas Brendan Murphy [a,b], Isobel Claire Gormley [a,b], Paul D. McNicholas [c], Dimitris Karlis [d]

[a] School of Mathematical Sciences, University College Dublin, Ireland
[b] INSIGHT: The National Centre for Big Data Analytics, University College Dublin, Ireland
[c] Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
[d] Department of Statistics, Athens University of Economics and Business, Greece

ARTICLE INFO

Article history:
Received 14 December 2013
Received in revised form 1 September 2014
Accepted 6 September 2014
Available online 19 September 2014

Keywords:
Model-based clustering
Multivariate normal inverse Gaussian distribution
mclust
Information metrics
Kolmogorov–Smirnov goodness of fit

ABSTRACT

Many model-based clustering methods are based on a finite Gaussian mixture model. The Gaussian mixture model implies that the data scatter within each group is elliptically shaped. Hence non-elliptical groups are often modeled by more than one component, resulting in model over-fitting. An alternative is to use a mean–variance mixture of multivariate normal distributions with an inverse Gaussian mixing distribution (MNIG) in place of the Gaussian distribution, to yield a more flexible family of distributions. Under this model the component distributions may be skewed and have fatter tails than the Gaussian distribution. The MNIG based approach is extended to include a broad range of eigendecomposed covariance structures. Furthermore, MNIG models where the other distributional parameters are constrained is considered. The Bayesian Information Criterion is used to identify the optimal model and number of mixture components. The method is demonstrated on three sample data sets and a novel variation on the univariate Kolmogorov–Smirnov test is used to assess goodness of fit.

1. Introduction

Mixture models are a commonly employed tool in statistical modeling, in particular the mixture of multivariate Gaussian distributions that forms the basis of the model-based clustering package **mclust** (Fraley and Raftery, 1998, 1999) in **R** (R Development Core Team, 2012). The Gaussian mixture model implies that the data, within each group, have an elliptical scatter. Hence non-elliptical groups are often modeled by more than one component, resulting in over-fitting. This may render the clustering rule ambiguous, meaning the correct (lower) number of groups is not identified. Ultimately this can result in higher misclassification rates. Also, the Gaussian mixture model can struggle to accommodate clusters with heavy tails or outliers.

One solution to this problem is to apply mixtures of t distributions, whose heavier tails can guard against the influence of outliers (McLachlan and Peel, 2000; Andrews and McNicholas, 2011). However this approach still implies that the data are elliptically contoured within each group (Banfield and Raftery, 1993). To address this issue, mixtures of skew-normal or skew-t

☆ Supplementary material for this paper has been added to the online version of the manuscript (see Appendix A).

* Corresponding author.
E-mail addresses: adrian.ohagan@ucd.ie, adrian.ohagan@hotmail.co.uk (A. O'Hagan).

http://dx.doi.org/10.1016/j.csda.2014.09.006
0167-9473/

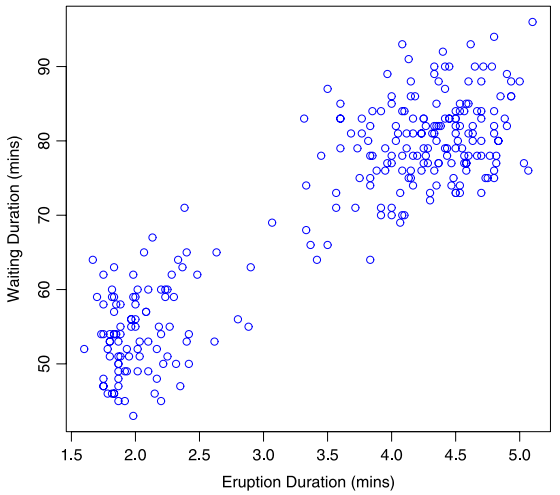

Fig. 1. Scatter plot of the *Old Faithful* data.

distributions can be used (Lin et al., 2007b,a; Cabral et al., 2012; Prates et al., 2013a; Vrbik and McNicholas, 2014). However, these distributions can prove numerically unstable in high-dimensional settings (Fruhwirth-Schnatter and Pyne, 2009).

Alternatively, data transformations prior to modeling can be used to reduce skew as much as possible (MacLean et al., 1976). Gottardo and Lo (2011) use the Box–Cox transformation as a precursor to fitting a mixture of t distributions to flow cytometry data. Unfortunately such a process is data and variable dependent, making automatic model fitting and selection difficult or impossible. Interpretability of results also suffers, since the data are no longer modeled in their original units.

A potential antidote to fitting a more complex model is to retain the Gaussian mixture model but merge mixture components from the initial model fit to produce an updated clustering solution (Hennig, 2010). This exploits the Gaussian mixture model's ability to provide a robust density estimate for the data while addressing its propensity for over-fitting.

The multivariate normal inverse Gaussian (MNIG) is a mean–variance mixture of multivariate Gaussians and is a special case of the generalized hyperbolic mixture (McNicholas et al., 2013). This yields a more flexible family of mixture distributions, which may be skewed and have fatter tails than a Gaussian distribution (Karlis and Santourian, 2008). The MNIG based approach is extended to the full range of eigenvalue decomposed covariance structures, as considered in **mclust**. Furthermore, the family of MNIG models where distributional parameters are constrained, is considered. This can improve model parsimony in cases where clusters have similar shape properties. The Bayesian Information Criterion (BIC) (Schwarz, 1978) is used to identify the optimal model and number of components. Disparities in clustering solutions under the mixture of MNIG and mixture of Gaussian (**mclust**) approaches are highlighted.

Section 2 presents the data sets used as motivating examples: the *Old Faithful* eruptions data, the *FLAME* flow cytometry data and the *Iris* data. Section 3 gives an account of the EM algorithm for fitting the mixture of MNIG distributions by maximum likelihood. Section 4 details a range of model diagnostics used to compare the competing clustering methods. BIC is used for model selection, metrics are used to compare clustering solutions and a goodness-of-fit test based on the Kolmogorov–Smirnov statistic is also detailed. Section 5 presents the results obtained for the motivating data sets, highlighting improvements in the clustering capability of a mixture of MNIG distributions over other mixture distributions. Section 6 summarizes the main findings and explores further avenues of investigation.

2. Illustrative data sets

Three data sets are used as motivating examples; these are described below.

Old Faithful data

The data are comprised of bivariate observations for 272 eruptions of the Old Faithful geyser in Yellowstone National Park (Azzalini and Bowman, 1990). Each observation records the eruption duration and the waiting duration until the next eruption, both measured in minutes. This is a classic test case for any clustering methodology because the data are multimodal. However, there are no "true" group labels available—different numbers of groups can be identified depending on the clustering rule applied (Hunter et al., 2007); see Fig. 1.

Flow cytometry FLAME data

The *FLAME* flow cytometry data is the *090806A0minLymphocytes* sample from the T cell phosphorylation data set as analyzed in Pyne et al. (2009). The data comprise 4669 observations and 4 cell surface marker variables: SLP76, ZAP70, CD4