Model-Based Clustering IV

Prof. Sharon McNicholas

STATS 780/CSE 780

Introduction

- We have looked at clustering using Gaussian mixture models.
- This includes the GPCM family, the MFA model and PGMM family, and an approach for clustering longitudinal data in a mixture framework.
- In this "lecture", we will look at some non-Gaussian mixtures.
- Some of the material is taken from McNicholas (2016a,b), and references are given at the end.

Mixtures of Non-Gaussian Dists.

- Recall that the density of a G-component (parametric finite) Gaussian mixture model is

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g).$$

- Using Gaussian component densities is, in itself, quite restrictive, e.g.,
  - components are symmetric, and
  - have "light" tails.

- Some have argued that a Gaussian mixture can be used and components then merged after the fact. . . but this is not a reliable approach (example later).

Mixtures of Non-Gaussian Dists. cont.

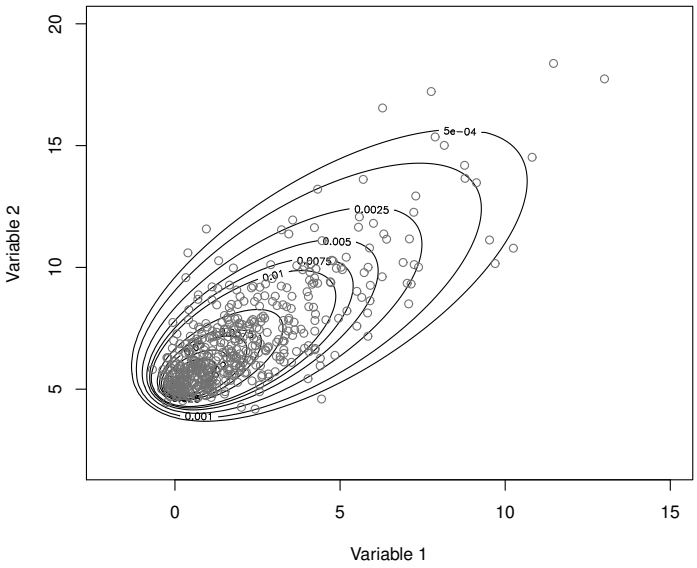
- Suppose that a component in a mixture corresponds to a cluster.

- A mixture of densities with a concentration parameter, e.g., a mixture of multivariate t-distributions (e.g., Peel and McLachlan, 2000), can be used for "fatter" clusters.

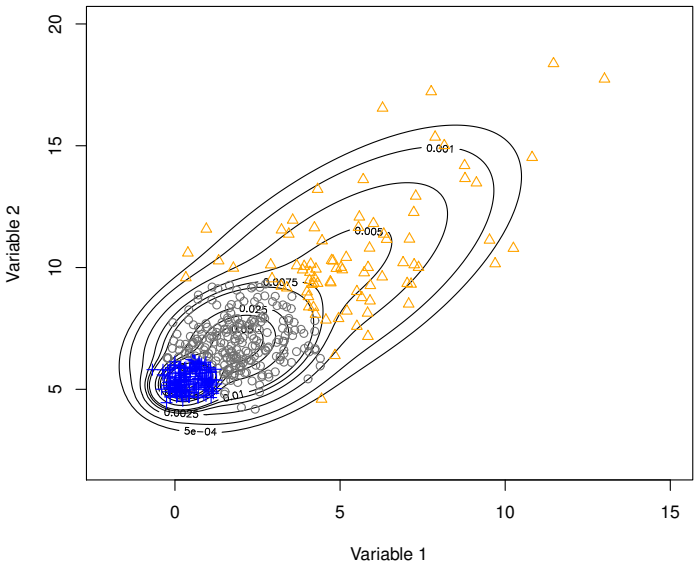
- A mixture of densities with a skewness parameter, e.g., a mixture of SAL distributions (Franczak et al., 2014) or a mixture of skew-normal distributions (e.g., Vrbik and McNicholas, 2014), can be used for asymmetric clusters.

- A mixture of densities with skewness and concentration parameters, e.g., a mixture of skew-t distributions (e.g., Murray et al., 2014), can be used for asymmetric, "fatter" clusters.

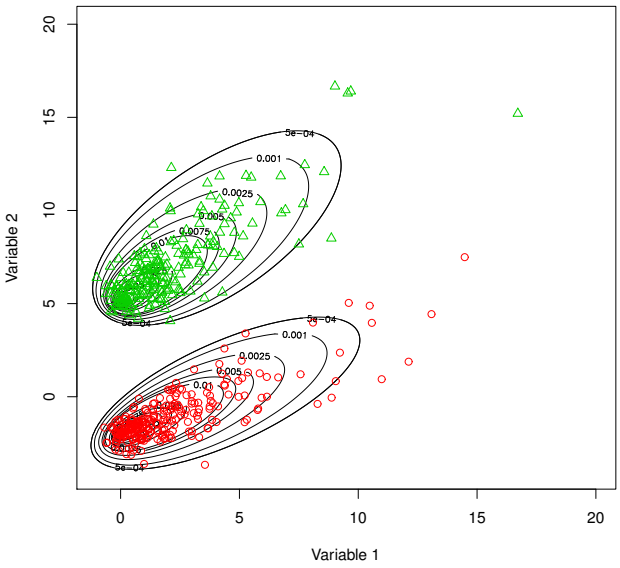
$G = 1$: SAL



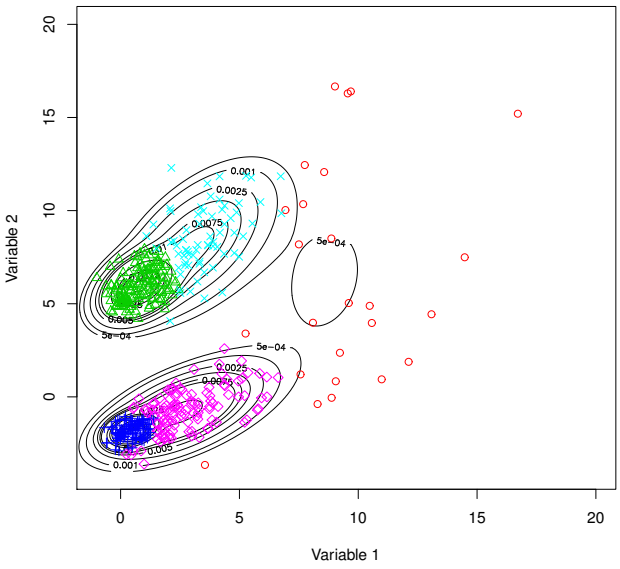
$G = 1$: (Merged) Gaussian



$G = 2$: SAL Mixture



$G = 2$: (Merged) Gaussian Mixture



Comments

- This simple two-component example illustrates the important role of mixtures of non-Gaussian distributions.
- In a few minutes, we will look at some examples in R.
- For further details on non-Gaussian mixtures, with a focus on component densities that are normal mean-variance mixtures, you can look at McNicholas (2016a) or McNicholas (2016b).

References

- Franczak, B.C., R.P. Browne, and P.D. McNicholas (2014). 'Mixtures of shifted asymmetric Laplace distributions'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(6), 1149–1157.
- McNicholas, P.D. (2016a). *Mixture Model-Based Classification*. Boca Raton: Chapman & Hall/CRC Press.
- McNicholas, P.D. (2016b). 'Model-based clustering'. *Journal of Classification* **33**(3).
- Murray, P.M., R.B. Browne, and P.D. McNicholas (2014). 'Mixtures of skew-t factor analyzers'. *Computational Statistics and Data Analysis* **77**, 326–335.
- Peel, D. and G.J. McLachlan (2000). 'Robust mixture modelling using the t-distribution'.*Statistics and Computing* **10**(4), 339–348.
- Vrbik, I. and P.D. McNicholas (2014). 'Parsimonious skew mixture models for model-based clustering and classification'. *Computational Statistics and Data Analysis* **71**, 196–210.