Estimating the number of clusters in OTRIMLE robust Gaussian mixture clustering

C. Hennig[1*], P. Coretto[2]

[1] *Department of Statistical Science, University College London; c.hennig@ucl.ac.uk.*
[2] *Department of Economics and Statistics, University of Salerno; pcoretto@unisa.it.*
[*] *Presenting author*

Keywords. *Mixture model; parametric bootstrap; Optimally Tuned Robust Improper Maximum Likelihood.*

Coretto and Hennig (2015a, 2015b) developed the method of Robust Improper Maximum Likelihood (RIMLE; "OTRIMLE" stands for "Optimally Tuned RIMLE") for clustering based on a Gaussian mixture model but allowing for some observations that could not reasonably be assigned to any cluster.

RIMLE is based on modelling the data (from p-dimensional Euclidean space) as i.i.d. generated by a Gaussian mixture model with an additional improper mixture component, namely a uniform distribution with density level c over the whole Euclidean space ("noise component"), and to fit such an improper distribution by (pseudo-)maximum likelihood. This allows for a smooth classification of points as outliers/noise or being generated by one of the clusters, i.e., the Gaussian mixture components.

The OTRIMLE is defined by minimising a "Mahalanobis criterion". This amounts to choosing c in such a way that the distribution of Mahalanobis distances to the corresponding cluster mean of the portion of the data classified as non-outlying is optimally approximated by a χ^2-distribution, as should be the case if the non-outliers came indeed from a Gaussian mixture.

Coretto and Hennig (2015a, 2015b) assume the number of mixture components k as fixed.

In this presentation we explore model diagnostic and estimation of the number of clusters by the following parametric bootstrap principle: we generate many datasets from the Gaussian mixture (non-outlying) part of the estimated mixture, using the estimated parameters, and we compare the distribution of the values of the Mahalanobis criterion mentioned above to the value achieved for the dataset under study. This allows us to see which numbers of clusters yield models that are consistent

with the data. This is inspired by Davies's (1995) Data Features and can be applied to more general model selection and diagnostic problems in robust model-based clustering.

References

Coretto, P. & Hennig, C. (2015a). Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust Gaussian clustering. arXiv:1406.0808. To appear in *Journal of the American Statistical Association*.

Coretto, P. & Hennig, C. (2015b). A consistent and breakdown robust model-based clustering method. arXiv:1309.6895.

Davies, P. L. (1995). Data Features. *Statistica Neerlandica*, **49**, 185–245.