ROBUST MODEL-BASED CLUSTERING WITH COVARIANCE MATRIX CONSTRAINTS

Pietro Coretto\textsuperscript{1} and Christian Hennig\textsuperscript{2}

\textsuperscript{1} Dipartimento di Scienze Economiche e Statistiche, Università di Salerno, (e-mail: pcoretto@unisa.it)
\textsuperscript{2} Department of Statistical Science, University College London, (e-mail: c.hennig@ucl.ac.uk)

\textbf{ABSTRACT}: The Optimally Tuned Robust Improper Maximum Likelihood Estimator (OTRIMLE) for robust model-based clustering is introduced. It is based on a ML-type procedure for a pseudo model in which clusters are represented by a finite mixture of Gaussian distributions, while noise is represented with the addition of an improper constant density (ICD). The OTRIMLE requires constraints on the underlying covariance matrices that prevent spurious solutions. These constraints may have strong impact on the final clustering and alternative algorithms are provided with the OTRIMLE software.

\textbf{KEYWORDS}: Cluster analysis, covariance matrix constraints, Gaussian mixtures models, EM-algorithm, robustness.

1 OTRIMLE clustering

Suppose \(\{x_1, x_2, \ldots, x_n\}\) are observed datapoints, here \(x_i \in \mathbb{R}^p\) for all \(i = 1, 2, \ldots, n\). These observed vectors have to be assigned to \(G\) clusters. When clusters have approximately Gaussian shapes one can use model-based clustering methods based on finite Gaussian mixture models. In this framework it is assumed that each component of a Gaussian mixtures generates a cluster. Clustering is performed based ML estimate of the mixture parameters (see Fraley & Raftery, 2002). Kiefer & Wolfowitz, 1956 showed that the ML for finite mixture models does not exists due the unboundedness of the likelihood function. Hennig, 2004 showed that the MLE can be strongly affected by outliers and noise. By noise here we mean points in the dataset that forms density regions far from being Gaussian shaped. The OTRIMLE aims at solving both issues. The main idea of the OTRIMLE is to approximate the data generating process (DGP) with a pseudo density model where the noise is represented the
ICD over the whole Euclidean space:

$$\psi_\delta(x, \theta) = \pi_0 \delta + \sum_{j=1}^{G} \pi_j \phi(x; \mu_j, \Sigma_j),$$  \hspace{1cm} (1)

with $\pi_0, \pi_j \in [0, 1]$ for $j = 1, 2, \ldots, G$, $\pi_0 + \sum_{j=1}^{G} \pi_j = 1$, $\phi(\cdot; \mu, \Sigma)$ is multivariate Gaussian density function with mean $\mu$ and covariance matrix $\Sigma$. $\delta > 0$ is the ICD, and it is meant to represent data points in low density regions that cannot be adequately fitted by Gaussian components.

The parameter vector $\theta$ contains all Gaussian parameters plus all proportion parameters including $\pi_0$. Define pseudo posterior probabilities

$$\tau_j(x_i, \theta) := \begin{cases} \frac{\pi_0 \delta}{\psi_\delta(x_i, \theta)} & \text{if } j = 0 \\ \frac{\pi_j \phi(x_i; \mu_j, \Sigma_j)}{\psi_\delta(x_i, \theta)} & \text{if } j = 1, 2, \ldots, G; \end{cases}$$

for a given $\theta$ points are assigned to clusters according to the usual maximum posterior rule. Candidate values of $\theta$ are chosen among those that maximize the associated pseudo log-likelihood function, i.e.

$$\theta_\theta(\delta) = \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log \psi_\delta(x_i, \theta),$$  \hspace{1cm} (2)

where the constrained parameter set is defined as

$$\Theta := \left\{ \theta : \pi_j \geq 0 \forall j \geq 1; \pi_0 + \sum_{j=1}^{G} \pi_j = 1; \frac{1}{n} \sum_{i=1}^{n} \tau_0(x_i, \theta) \leq \pi_{\max}; \frac{\lambda_{\max}(\theta)}{\lambda_{\min}(\theta)} \leq \gamma \right\},$$  \hspace{1cm} (3)

$\theta_\theta(\delta)$ is called the RIMLE. $\lambda_{\max}(\theta)/\lambda_{\min}(\theta)$ is the ratio between the largest and the smallest eigenvalue among all those of $\Sigma_1, \Sigma_2, \ldots, \Sigma_G$. $\gamma$ defines what is called “eigenratio constraint” (ERC). $\pi_{\max}$ puts an upper bound on the maximum number of points that can be assigned to the noise component ($j = 0$), this is called “noise proportion constraint” (NPC). Both ERC and NPC are discussed in the next Section. Theoretical properties of the RIMLE are investigated in Coretto & Hennig, 2013. The OTRIMLE is a procedure to select an adequate value of $\delta$. In practice this is done by solving the following optimization problem

$$\delta_\theta = \arg \min_{\delta \in [0, \delta_{\max}]} D(\delta) + \beta \pi_{0,\theta}.$$

$$\lambda_{\max}(\theta)/\lambda_{\min}(\theta) \leq \gamma.$$
where $D(\delta)$ measures the departure of clusters from the Gaussian shape. $\beta$ is a penalty term that has the role to prevent that too many points are assigned to the noise component in the tentative to enforce Gaussianity when in fact the DGP departs from it. Details about the OTRIMLE are given in Coretto & Hennig, 2014 where it is shown its competitive performance with respect to alternative methods. Alternative methods include: ML-type estimators for Gaussian mixtures with the addition of a “proper” uniform noise component (see Banfield & Raftery, 1993 and Coretto & Hennig, 2011); ML for mixture models based on the Student $t$-distributions due to McLachlan & Peel, 2000; robust partitioning methods based on ML for classification likelihood such as the TCLUST method of García-Scudero et al., 2008, and the the “k–parameters clustering” (kPC) of Gallegos & Ritter, 2013. Both TCLUST and kPC solve the robustness issue by trimming a fixed proportion of data points. The advantage of the OTRIMLE over its competitor is twofold: (i) being based on a mixture likelihood function it allows a smooth classification between clusters, and a smooth transition between clustered regions and the outliers/noise region; (ii) it is the only method with a data-driven decision about the amount of noise.

2 Constraints and Algorithms

Computation of $\delta_n$ is performed by a line search algorithm solving (2) at each iteration. (2) is solved by the EM algorithm developed in Coretto & Hennig, 2013. Here we focus on ERC and NPC. ERC was first proposed in Hathaway, 1985 and Ingrassia, 2004. These constraints are also used in TCLUST. For the OTRIMLE one needs to prevent degeneration of the pseudo-likelihood also in the case that almost all observations are assigned to the noise component. For the OTRIMLE one needs ERC together with NPC. This may look like a disadvantage, but for the methods cited above such problems are only avoided by fixing the trimming rate. Contrary to ERC the NPC has a much easier interpretation. One can fix $\pi_{\text{max}}=50\%$ in order to implement a standard condition in robust statistics that no more than half of the data should be classified as outliers/noise. The ERC may have strong impact on the final clustering, and results may vary with their implementation in algorithms. In practice robustness and the spurious maximum problems exhibit an interplay. Often the emergence of spurious solutions happens when there are few concentrated points lying in a region that does not overlap with any of the clusters, and often these points may well be qualified as outliers/noise. The difficulty here is not only computational, but also conceptual. The question is, in fact, to which extend one has to consider these points as small clusters with an “atypical” scatter rather than
outliers/noise. Software implementation of the OTRIMLE takes into consideration these effects and allows for several alternative uses of the ERC.

References


