Robust model-based clustering and mixture modeling by using trimming

Luis A. García-Escudero, Alfonso Gordaliza, Carlos Matrán and Agustín Mayo-Iscar

Abstract The use of trimming procedures is common in several statistical frameworks to achieve robustness against anomalous observations. Trimming can be also applied in model-based clustering and mixture modeling and this work reviews some trimming proposals aimed at robustifying these two widely used techniques. The consideration of computationally feasible constraints on the scatters of the clusters and mixture components also plays an important role in these proposals. This work also presents some recent extensions of these methods to related problems. Some new tools helping the user to fix the different tuning parameters are shown.

Key words: Model-based clustering; mixture models; robustness; trimming

1 Introduction

Trimming is surely the easiest and one of the most widely applied strategies to achieve robustness in statistical procedures and it simply means that certain fraction α of observations are not taking into account when applying the statistical proce-

Luis Angel García-Escudero
Departamento de Estadística e I.O. and IMUVA, Facultad de Ciencias. Universidad de Valladolid. 47002, Valladolid. Spain, e-mail: lagarcia@eio.uva.es

Alfonso Gordaliza
Departamento de Estadística e I.O. and IMUVA, Facultad de Ciencias. Universidad de Valladolid. 47002, Valladolid. Spain, e-mail: alfonsog@eio.uva.es

Carlos Matrán
Departamento de Estadística e I.O. and IMUVA, Facultad de Ciencias. Universidad de Valladolid. 47002, Valladolid. Spain, e-mail: matran@eio.uva.es

Agustín Mayo-Iscar
Departamento de Estadística e I.O. and IMUVA, Facultad de Ciencias. Universidad de Valladolid. 47002, Valladolid. Spain, e-mail: agustinm@eio.uva.es

dure. If this proportion α of discarded observation includes the most "outlying" ones then the harmful effects of these outlying observations on the results of the statistical procedure are naturally avoided. The detection of anomalous data can be directly done by using standard exploratory-graphical tools in the most simple cases but it is not an easy task depending on the statistical procedure to be applied or depending on the problem dimensionality. With this idea in mind, we are interested in the development of unsupervised or almost unsupervised procedures to trim outlying observations in model-based clustering and mixture modeling problems as a way to derive robust approaches in these two interesting problems. We will present some already developed methodologies within this framework and some still open research will be outlined.

2 Trimming and model-based clustering

As done in other positive breakdown point procedures as the Least Trimmed Squares (LTS), the Least Median of Squares (LMS), the MCD (Minimum Covariance Determinant) and the MVE (Minimum Volume Ellipsoid) covariance matrix estimator (see Rousseeuw and Leroy [1987]), a sensible way to perform trimming is letting the data themselves tell us which observations should be trimmed in order to find an optimal clustering for the untrimmed ones. The first application of this trimming approach in Cluster Analysis was the trimmed K-means method introduced in Cuesta-Albertos et al. [1997]. This approach is based on the minimization of the target function

$$\sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \mathbf{m}_k\|^2, \tag{1}$$

with $C_0, C_1, ..., C_K$ being a partition of the data set $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ such that C_0 includes $[n\alpha]$ observations. Note that summation goes from $k = 1$ to K (not from $k = 0$) and, thus, the observations in C_0 are not taken into account when evaluating the target function. As shown in García-Escudero and Gordaliza [1999], this approach provides a notably robustification with respect the classical K-means and with respect the K-medoids (which can still be broken down just by adding one single outlying observations). A feasible algorithm for its implementation can be found in García-Escudero et al. [2003].

Unfortunately, as happens with the classical K-means, this clustering approach is aimed at searching for spherical clusters with similar scatters and this is not always the case in many interesting applications. Thus, it is interesting to extend this trimming methodology to heterogeneous clustering problems. With this idea in mind, let $\varphi_d(\cdot; \mu, \Sigma)$ denote the probability density function of the d-variate normal distribution with mean μ and covariance matrix Σ and let us consider the maximization of the target function

$$\sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \log\left(p_k \varphi_d\left(\mathbf{x}_i; \mathbf{m}_k, \mathbf{S}_k\right)\right), \tag{2}$$

where maximization is in terms of:

(i) K centers $\mathbf{m}_k$ in $\mathbb{R}^d$,
(ii) K symmetric positive definite $d \times d$ scatter matrices $\mathbf{S}_k$,
(iii) K weights in $[0,1]$ satisfying $\sum_{k=1}^{K} p_k = 1$, and,
(iv) a partition $C_0, C_1, \ldots, C_K$ of the dataset where C_0 includes the $[n\alpha]$ the trimmed observations.

The TCLUST procedure introduced in García-Escudero et al. [2008] is defined throughout the constrained maximization of the classification trimmed likelihood in (2) under appropriate constraints on the eigenvalues of the $\mathbf{S}_k$ scatter matrices. Gallegos and Ritter [2005] consider a mathematical probabilistic framework, called the "spurious outliers model", that justifies the use of (2) and Neykov et al. [2007] has also considered this type of trimmed likelihoods but without imposing constraints on the scatter matrices.

Note that we are ideally assuming clusters with similar sizes if these p_k weights are not included in the trimmed likelihood (i.e., $p_1 = ... = p_K$) and, in the most constrained case, we recover the trimmed K-means method when assuming equal weights $p_1 = ... = p_K$ and $\mathbf{S}_1 = ... = \mathbf{S}_K = s^2 I_d$. Moreover, it is also important to note that the maximization of (2) without any constraints on the scatter (covariance) matrices $\mathbf{S}_k$ is a mathematically ill-posed problem. To see this, just take $C_k = \{\mathbf{x}_i\}$ for any $\mathbf{x}_i$ in the data set, $\mathbf{m}_k = \mathbf{x}_i$ and $\mathbf{S}_k$ with $\det(\mathbf{S}_k) \to 0$, which makes (2) unbounded.

To be more precise with the constraints, if we consider the eigenvalues system of the $\mathbf{S}_k$ matrices denoted by $\lambda_l(\mathbf{S}_k)$ and their maximum and minimum values

$$M_n = \max_{k=1,\ldots,K} \max_{l=1,\ldots,d} \lambda_l(\mathbf{S}_k) \text{ and } m_n = \min_{k=1,\ldots,K} \min_{l=1,\ldots,d} \lambda_l(\mathbf{S}_k), \tag{3}$$

it is proposed to constrain the ratio M_n/m_n to be smaller than a fixed constant c. The proposed constraints try to extend Hathaway [1985]'s ones to the heterogeneous robust cluster analysis framework. Ingrassia and Rocci [2007] have considered analogous type of constraints in a mixture modeling framework.

An efficient algorithm to approximately solve the constrained maximization is given in Fritz et al. [2012]. The algorithm is a Classification-EM type algorithm where "concentration" steps (Rousseeuw and Van Driessen [1999]) are also applied. An implementation of this algorithm is available through the `tclust` package (see Fritz et al. [2013]) for the R environment for statistical computing.

Finally, it is important to comment that the TCLUST approach admits a population version with a consistency result under mild assumptions.

3 Trimming in mixture modeling

The methodology presented in previous sections can be extended to mixture modeling by considering the maximization of the trimmed mixture likelihood defined as

$$\sum_{i=1}^{n} z(\mathbf{x}_i) \log \left(\sum_{k=1}^{K} p_k \varphi_d(\mathbf{x}_i; \mathbf{m}_k, \mathbf{S}_k) \right), \tag{4}$$

where $z(\cdot)$ is a 0-1 trimming indicator function that tell us whether observation $\mathbf{x}_i$ is trimmed off ($z(\mathbf{x}_i)$=0), or not ($z(\mathbf{x}_i)$=1). A fixed fraction α of observations are allowed to be unassigned by setting $\sum_{i=1}^{n} z(\mathbf{x}_i) = n - [n\alpha]$. The trimmed mixture likelihood have been also considered in Neykov et al. [2007] and Gallegos and Ritter [2009]. Again, the maximization of (4) without any constraint in the $\mathbf{S}_k$ scatter matrices is an ill-possed problem since it is again unbounded. It is proposed in García-Escudero et al. [2010a] the maximization of (4) under eigenvalue ratio constraints like in (3). The joint consideration of trimming and constraints certainly provides robustness and stability to the mixture modeling. In fact, it can be seen that trimming without constraints do not provide too much robustness due to the undesired detection of spurious solutions (a few proportion of almost collinear observations which results in a high values for (4)).

A computationally feasible algorithm is available by modifying the well-known EM algorithm in mixture fitting and by forcing the eigenvalues ratio constraints on the scatter matrices as was done in Fritz et al. [2012].

Cuesta-Albertos et al. [2008] proposed an alternative approach for robust mixture modeling where trimming plays a key role.

4 Extensions and open research lines

Derivation of proper choices for the number of groups K and for the trimming proportion α are two related problems that should be addressed simultaneously. Moreover, these choices clearly depend on the type of clusters we are searching for and on the allowed differences between cluster/components scatters. García-Escudero et al. [2011] proposed some graphical tools that helps the user to make sensible choices for the parameters. Trying to also develop numerical (and not only graphical) procedures, some modified BIC proposals have been recently introduced which take into account the model complexity introduced by the allowed differences between cluster/components scatters through constant c in the constraint for the eigenvalues in (3). This is an ongoing research project together with colleagues from the University of Parma.

The use of constrains, which extends the simpler eigenvalues ratio constraint, can be also considered as providing soft transitions between models consequence of well-known "parsimonious" parameterizations (Celeux and Govaert [1995]) of the $\mathbf{S}_k$ scatter matrices. We think that the just already commented modified BIC proposal

can be also applied when dealing with these new models for cluster analysis and mixture modeling.

Cluster-weighted modeling (CWM) introduced in Gershenfeld [1991] is also a very useful methodology that can be modified by including trimming and restrictions in its formulation as has been recently proposed in García-Escudero et al. [2014]. This approach is based on the maximization of

$$\sum_{i=1}^{n} z(\mathbf{x}_i, y_i) \log \left[\sum_{k=1}^{K} \varphi_1(y_i; \mathbf{b}_k' \mathbf{x}_i + b_k^0, \sigma_k^2) \varphi_d(\mathbf{x}_i; m_k, \mathbf{S}_k) p_k \right], \tag{5}$$

where the $\{\mathbf{x}_i\}_{i=1}^n$ corresponds to the values taken by the explanatory variables and $\{y_i\}_{i=1}^n$ to those taken by the response variable. Differentiated constraints on the $\mathbf{S}_k$ scatter matrices and on the σ_k^2 variances of the error terms are considered in this new approach. It can be seen that this methodology naturally provides a convenient mathematical formulation for the "second trimming" process described in García-Escudero et al. [2010b]. The consideration of trimming and constraints in mixtures of Factor Analyzers is also an interesting research topic which is now jointly addressed with S. Ingrassia and F. Greselin.

Finally, we would like to comment that the consideration of a dynamic range of α trimming proportions is clearly connected with the " forward search" methodology for data analysis (Atkinson et al. [2004]) and, thus, a common theoretical and computational treatment of both problems could be an interesting research line.

References

A. Atkinson, M. Riani, and A. Cerioli. *Exploring multivariate data with the forward search*. Springer Series in Statistics, Springer, New York, 2004.

G. Celeux and A. Govaert. Gaussian parsimonious clustering models. *Pattern Recogn*, 28:781–793, 1995.

J. Cuesta-Albertos, A. Gordaliza, and C. Matrán. Trimmed k-means: an attempt to robustify quantizers. *Ann Stat*, 25:553–576, 1997.

J. Cuesta-Albertos, C. Matran, and A. Mayo-Iscar. Robust estimation in the normal mixture model based on robust clustering. *J Roy Stat Soc, Ser. B*, 70:779–802, 2008.

H. Fritz, L. García-Escudero, and A. Mayo-Iscar. tclust: An R package for a trimming approach to cluster analysis. *J Stat Softw*, 47, 2012. URL http://www.jstatsoft.org/v47/i12.

H. Fritz, L. García-Escudero, and A. Mayo-Iscar. A fast algorithm for robust constrained clustering. *Comput Stat Data Anal*, (61):124–136, 2013.

M. Gallegos and G. Ritter. A robust method for cluster analysis. *Ann Stat*, 33: 347–380, 2005.

M. Gallegos and G. Ritter. Trimmed ML estimation of contaminated mixtures. *Sankhya (Ser. A)*, (71):164–220, 2009.

L. García-Escudero and A. Gordaliza. Robustness properties of k-means and trimmed k-means. *J Am Stat Assoc*, 94:956–969, 1999.

L. García-Escudero, A. Gordaliza, and C. Matrán. Trimming tools in exploratory data analysis. *J Comput Graph Stat*, 12:434–449, 2003.

L. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A general trimming approach to robust cluster analysis. *Ann Stat*, 36:1324–1345, 2008.

L. García-Escudero, A. Gordaliza, and A. Mayo-Iscar. A constrained robust proposal for mixture modeling avoiding spurious solutions. *Adv Data Anal Classif*, 8:27–43, 2010a.

L. García-Escudero, A. Gordaliza, A. Mayo-Iscar, and R. San Martín. Robust clusterwise linear regression through trimming. *Comput Stat Data Anal*, 54:3057–3069, 2010b.

L. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. Exploring the number of groups in robust model-based clustering. *Stat Comput*, 21:585–599, 2011.

L. García-Escudero, A. Gordaliza, F. Greselin, S. Ingrassia, and A. Mayo-Iscar. Trimmed cluster weighted restricted modeling. *Submited for publication*, 2014.

N. Gershenfeld. Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences*, 808:18–24, 1991.

R. Hathaway. A constrained formulation of maximum likelihood estimation for normal mixture distributions. *Ann Stat*, 13:795–800, 1985.

S. Ingrassia and R. Rocci. Constrained monotone em algorithms for finite mixture of multivariate gaussians. *Comput Stat Data Anal*, (51):5339–5351, 2007.

N. Neykov, P. Filzmoser, R. Dimova, and P. Neytchev. Robust fitting of mixtures using the trimmed likelihood estimator. *Comput Stat Data Anal*, 52:299–308, 2007.

P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. Wiley-Interscience, New York, 1987.

P. Rousseeuw and K. Van Driessen. A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics*, 41:212–223, 1999.