A PROPOSAL FOR THE JOINT AUTOMATED DETECTION OF CLUSTERS AND ANOMALIES

Luis A. García-Escudero [1] , Christian Hennig[2] , Agustín Mayo-Iscar[1] , Gianluca Morelli[3] and Marco Riani[3]

[1] Department of Statistics and Operation Research and IMUVA, University of Valladolid, (e-mail: lagarcia@uva.es, agustin.mayo.iscar@uva.es)

[2] Department of Statistical Sciences "Paolo Fortunati", University of Bologna, (e-mail: christian.hennig@unibo.it)

[3] Dipartimento di Scienze Economiche e Aziendali, Universita' degli Studi di Parma, (e-mail: gianluca.morelli@unipr.it, marco.riani@unipr.it))

ABSTRACT: It is known that outliers can be problematic when statistical techniques are applied. This is also the case in Cluster Analysis and, with this in mind, the TCLUST method was introduced as a robust clustering alternative. Given a fixed trimming level α, TCLUST attempts to detect the fraction α of observations that should best be discarded after assuming k normally distributed components. However, the main problem is how to determine reasonable values for k and α for a given data set. An approach was introduced to choose k and α through visual inspection of "classification trimmed likelihood" curves. Theoretical background will be provided for a better understanding of that approach, along with a parametric bootstrap method to reduce subjectivity and produce a small list of sensible robust clustering partitions.

KEYWORDS: clustering, robustness, trimming, outliers

1 Robust clustering and TCLUST

It is well known that outliers can be problematic when applying statistical methods for data analysis, and this also happens in the case of Cluster Analysis. Outliers can affect clustering methods in such a way that main clusters can be joined artificially or clusters formed of few outlying observations are detected (see, e.g., García-Escudero & Gordaliza, 1999). Moreover, it is interesting to apply clustering techniques to deal with outliers since clustered sets of outliers are known to be particularly harmful for many (even robust) statistical procedures. Consequently, different robust clustering methods have been introduced that can be used successfully to jointly deal with clusters and outliers (Ritter, 2014, García-Escudero *et al.*, 2016).

One such approach to robust clustering is based on applying impartial trimming. Given a fixed trimming level α, the term "impartial" means that is the data set itself that indicates what fraction α of observations should be trimmed. The TCLUST method introduced in García-Escudero *et al.*, 2008 is a robust clustering procedure based on that impartial trimming principle and where elliptically contoured clusters are allowed.

Given a sample $\mathcal{X} = \{x_1, ..., x_p\}$ in $\mathbb{R}^p$, the TCLUST method is defined by maximizing

$$\sum_{j=1}^{k} \sum_{i \in R_j} \log(\pi_j \phi(x_i; \mu_j, \Sigma_j)), \tag{1}$$

where $\phi(\cdot; \mu, \Sigma)$ is the density function of the p-variate normal distribution, $\{R_0, R_1, ..., R_k\}$ is a partition of the indexes $\{1, 2, ..., n\}$ such that $\#R_0 = [n\alpha]$. Also, in that maximization, we enforce

$$M_n / m_n \leq c$$

for $M_n = \max_{j=1,...,k} \max_{l=1,...,p} \lambda_l(\Sigma_j)$ and $m_n = \min_{j=1,...,k} \min_{l=1,...,p} \lambda_l(\Sigma_j)$ being, respectively, the largest and the smallest of the eigenvalues of the Σ_j scatter matrices. The constant $c \geq 1$ plays an important role by avoiding uninteresting "spurious clusters" and providing well-defined mathematical problems. The $\pi_j \geq 0$ weights also satisfy $\sum_{j=1}^{k} \pi_j = 1$.

The TCLUST procedure can be implemented using the `tclust` package in R (Fritz *et al.*, 2012) and the `FSDA` Matlab toolbox (Riani *et al.*, 2012). However, TCLUST requires the simultaneous specification of the number of clusters k and the trimming fraction α. Choosing correctly those two parameters for a given data set is not always an easy task because, for instance, a set of close outliers could be considered as "noise" to be trimmed (requiring a higher α) or, alternatively, as an additional cluster (requiring a higher k). Therefore, the determination of k and α is a clearly interrelated problem that requires an unified treatment. Even choosing the number of groups k in Cluster Analysis, without trimming, is already well known to be a very complex problem.

2 Classification trimmed likelihood curves

A graphical procedure for selecting sensible values for k and α for TCLUST (when c is fixed) was introduced in García-Escudero *et al.*, 2011. The procedure was based on the visual inspection of the so-called "classification trimmed likelihood" curves. These curves are defined through

$$(k, \alpha) \mapsto \mathcal{L}^{\Pi}(\alpha, k; \mathcal{X}), \tag{2}$$

where $\mathcal{L}^{\Pi}(\alpha,k;\mathcal{X})$ denotes the maximum value reached in the constrained maximization of (1). García-Escudero *et al.*, 2011 explained that

$$t^n_{k,\alpha} = \mathcal{L}^{\Pi}(\alpha,k+1;\mathcal{X}) - \mathcal{L}^{\Pi}(\alpha,k;\mathcal{X})$$

should not be too small when there is a clear benefit in increasing k to $k+1$ for a trimming level α. This heuristic led to a graphical exploratory tool for choosing reasonable values for k and α.

Given a probability measure P, we can define a population version of the TCLUST problem (García-Escudero *et al.*, 2008). We can also define population versions of the classification trimmed likelihoods appearing in (2), which are denoted as $\mathcal{L}^{\Pi}_{\alpha,k}(P)$. We have that $\mathcal{L}^{\Pi}_{\alpha,k}(P_n) = \mathcal{L}^{\Pi}(\alpha,k;\mathcal{X})$, where P_n denotes the empirical measure corresponding to $\mathcal{X}$ ($\mathcal{X}$ seem as the realization of an i.i.d. sample from P). Given the consistency

$$\mathcal{L}^{\Pi}_{\alpha,k}(P_n) \to \mathcal{L}^{\Pi}_{\alpha,k}(P),$$

and the fact that $t^n_{k,\alpha} = \mathcal{L}^{\Pi}_{\alpha,k+1}(P_n) - \mathcal{L}^{\Pi}_{\alpha,k}(P_n)$, it makes sense to analyse the behaviour of $\mathcal{L}^{\Pi}_{\alpha,k}(P)$ to see under what circumstances $t^n_{k,\alpha}$ should be small. Theoretical have been obtained on the expected changes in $\mathcal{L}^{\Pi}_{\alpha,k}(P)$, when increasing k to $k+1$, depending on the underlying distribution P. These results provide some theoretical background to better understand the key ingredients involved in the classification trimmed likelihood curve and how these curves should be interpreted.

3 Parametric bootstrap automated procedure

In practical applications, it is not always easy to determine sensible values for k and α just from that visual inspection of the classification trimmed likelihood curves. The user must make rather subjective decisions about whether or not $t^n_{k,\alpha}$ can be considered small due to sample variability. A parametric bootstrap procedure will be presented trying to overcome that trouble.

By applying TCLUST to compute $t^n_{k,\alpha}$, we also obtain parameter estimates for the k fitted normal components. These parameters are used to draw B parametric bootstrap samples $\{\mathcal{X}^{*b}\}_{b=1}^{B}$, but also trying to emulate the mechanism generating the fraction α of contaminating observations in $\mathcal{X}$. If k and α are reasonable parameters, then $\{\mathcal{L}(\alpha,k+1;\mathcal{X}^{*b}) - \mathcal{L}(\alpha,k;\mathcal{X}^{*b})\}_{b=1}^{B}$ would allow us to "mimic" the sampling distribution of $t^n_{k,\alpha}$ and compute bootstrap p-values as

$$p_{k,\alpha} = \frac{\#\{b : \mathcal{L}(\alpha,k+1;\mathcal{X}^{*b}) - \mathcal{L}(\alpha,k;\mathcal{X}^{*b}) > t^n_{k,\alpha}\}}{B}.$$

We can use these bootstrap p-values to finally get a reduced list of reasonable k and α values for applying TCLUST in an fully automated way. Users can use this reduced list to choose the robust cluster partition that best meets their ultimate cluster and outlier detection goals, by applying standard cluster validation/visualization tools.

Illustrative and real data examples, together with a simulation study, also seem to justify the interest of the automated selection proposal. Therefore, we consider that the proposal is clearly valuable since it can certainly help the user in the detection of anomalies.

References

FRITZ, H., GARCÍA-ESCUDERO, L.A., & MAYO-ISCAR, A. 2012. tclust: An R Package for a Trimming Approach to Cluster Analysis. *Journal of Statistical Software*, **47**(12).

GARCÍA-ESCUDERO, L.A., & GORDALIZA, A. 1999. Robustness properties of k-means and trimmed k-means. *Journal of the American Statistical Association*, **94**, 956–969.

GARCÍA-ESCUDERO, L.A, GORDALIZA, A., MATRÁN, C., & MAYO-ISCAR, A. 2008. A general trimming approach to robust Cluster Analysis. *Annals of Statistics*, **36**, 1324–1345.

GARCÍA-ESCUDERO, L.A, GORDALIZA, A., MATRÁN, C., & MAYO-ISCAR, A. 2011. Exploring the number of groups in robust model-based clustering. *Statistics and Computing*, **21**, 585–599.

GARCÍA-ESCUDERO, L.A., GORDALIZA, A., MATRÁN, C., MAYO-ISCAR, A., & HENNIG, C.M. 2016. Robustness and Outliers. *Pages 653 – 678 of:* C. HENNIG, M. MEILA, F. MURTAGH, & R. ROCCI (eds), *Handbook of Cluster Analysis*. Serie Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

RIANI, M., PERROTTA, D., & TORTI, F. 2012. FSDA: A MATLAB toolbox for robust analysis and interactive data exploration,. *Chemometrics and Intelligent Laboratory Systems*, **116**, 17–32.

RITTER, G. 2014. *Cluster Analysis and Variable Selection*. Boca Raton: CRC Press.