1. Record Nr. TD18048016

Autore	DOTTO, FRANCESCO
Titolo	Advances in robust clustering methods with applications [Tesi di dottorato]
Lingua di pubblicazione	Inglese
Formato	Tesi di dottorato
Livello bibliografico	Monografia
Note	diritti: info:eu-repo/semantics/openAccess In relazione con info:eu-repo/semantics/altIdentifier/hdl/11573/953915
Sommario	Robust methods in statistics are mainly concerned with deviations from model assumptions. As already pointed out in Huber (1981) and in Huber & Ronchetti (2009) hese assumptions are not exactly true since they are just a mathematically convenient rationalization of an often fuzzy knowledge or belief". For that reason a minor error in the mathematical model should cause only a small error in the nal conclusions". Nevertheless it is well known that many classical statistical procedures are excessively sensitive to seemingly minor deviations from the assumptions". All statistical methods based on the minimization of the average square loss may suer of lack of robustness. Illustrative examples of how outliers' in uence may completely alter the nal results in regression analysis and linear model context are provided in Atkinson & Riani (2012). A presentation of classical multivariate tools' robust counterparts is provided in Farcomeni & Greco (2015). The whole dissertation is focused on robust clustering models and the outline of the thesis is as follows. Chapter 1 is focused on robust methods. Robust methods are aimed at increasing the eciency when contamination appears in the sample. Thus a general denition of such (quite general) concept is required. To do so we give a brief account of some kinds of contamination we can encounter in real data applications. Secondly

we introduce the Spurious outliers model" (Gallegos & Ritter 2009a) which is the cornerstone of the robust model based clustering models. Such model is aimed at formalizing clustering problems when one has to deal with contaminated samples. The assumption standing behind the Spurious outliers model" is that two dierent random mechanisms generate the data: one is assumed to generate the clean" part while the another one generates the contamination. This idea is actually very common within robust models like the Tukey-Huber model" which is introduced in Subsection 1.2.2. Outliers' recognition, especially in the multivariate case, plays a key role and is not straightforward as the dimensionality of the data increases. An overview of the most widely used (robust) methods for outliers detection is provided within Section 1.3. Finally, in Section 1.4, we provide a non technical review of the classical tools introduced in the Robust Statistics' literature aimed at evaluating the robustness properties of a methodology. Chapter 2 is focused on model based clustering methods and their robustness' properties. Cluster analysis, he art of nding groups in the data" (Kaufman & Rousseeuw 1990), is one of the most widely used tools within the unsupervised learning context. A very popular method is the k-means algorithm (MacQueen et al. 1967) which is based on minimizing the Euclidean distance of each observation from the estimated clusters' centroids and therefore it is aected by lack of robustness. Indeed even a single outlying observation may completely alter centroids' estimation and simultaneously provoke a bias in the standard errors' estimation. Cluster's contours may be in ated and the eal" underlying clusterwise structure might be completely hidden. A rst attempt of robustifying the k- means algorithm appeared in Cuesta-Albertos et al. (1997), where a trimming step is inserted in the algorithm in order to avoid the outliers' exceeding in uence. It shall be noticed that k-means algorithm is ecient for detecting spherical homoscedastic clusters. Whenever more exible shapes are desired the procedure becomes inecient. In order to overcome this problem Gaussian model based clustering methods should be adopted instead of k-means algorithm. An example, among the other proposals described in Chapter 2, is the TCLUST methodology (Garca- Escudero et al. 2008), which is the cornerstone of the thesis. Such methodology is based on two main characteristics: trimming a xed proportion of observations and imposing a constraint on the estimates of the scatter matrices. As it will be explained in Chapter 2, trimming is used to protect the results from outliers' in uence while the constraint is involved as spurious maximizers may completely spoil the solution. Chapter 3 and 4 are mainly focused on extending the TCLUST methodology. In particular, in Chapter 3, we introduce a new contribution (compare Dotto et al. 2015 and Dotto et al. 2016b), based on the TCLUST approach, called reweighted TCLUST or RTCLUST for the sake of brevity. The idea standing behind such method is based on reweighting the observations initially agged as outlying. This is helpful both to gain eciency in the parameters' estimation process and to provide a reliable estimation of the true contamination level. Indeed, as the TCLUST is based on trimming a xed proportion of observations, a proper choice of the trimming level is required. Such choice, especially in the applications, can be cumbersome. As it will be claried later on, RTCLUST methodology allows the user to overcome such problem. Indeed, in the RTCLUST approach the user is only required to impose a high preventive trimming level. The procedure, by iterating through a sequence of decreasing trimming levels, is

aimed at reinserting the discarded observations at each step and provides more precise estimation of the parameters and a nal estimation of the true contamination level ^. The theoretical properties of the methodology are studied in Section 3.6 and proved in Appendix A.1, while, Section 3.7, contains a simulation study aimed at evaluating the properties of the methodology and the advantages with respect to some other robust (reweigthed and single step procedures). Chapter 4 contains an extension of the TCLUST method for fuzzy linear clustering (Dotto et al. 2016a). Such contribution can be viewed as the extension of Fritz et al. (2013a) for linear clustering problems, or, equivalently, as the extension of Garca-Escudero, Gordaliza, Mayo-Iscar & San Martn (2010) to the fuzzy clustering framework. Fuzzy clustering is also useful to deal with contamination. Fuzziness is introduced to deal with overlapping between clusters and the presence of bridge points, to be dened in Section 1.1. Indeed bridge points may arise in case of overlapping between clusters and may completely alter the estimated cluster's parameters (i.e. the coecients of a linear model in each cluster). By introducing fuzziness such observations are suitably down weighted and the clusterwise structure can be correctly detected. On the other hand, robustness against gross outliers, as in the TCLUST methodology, is guaranteed by trimming a xed proportion of observations. Additionally a simulation study, aimed at comparing the proposed methodology with other proposals (both robust and non robust) is also provided in Section 4.4. Chapter 5 is entirely dedicated to real data applications of the proposed contributions. In particular, the RTCLUST method is applied to two dierent datasets. The rst one is the Swiss Bank Note" dataset, a well known benchmark dataset for clustering models, and to a dataset collected by Gallup Organization, which is, to our knowledge, an original dataset, on which no other existing proposals have been applied yet. Section 5.3 contains an application of our fuzzy linear clustering proposal to allometry data. In our opinion such dataset, already considered in the robust linear clustering proposal appeared in Garca-Escudero, Gordaliza, Mayo-Iscar & San Martn (2010), is particularly useful to show the advantages of our proposed methodology. Indeed allometric quantities are often linked by a linear relationship but, at the same time, there may be overlap between dierent groups and outliers may often appear due to errors in data registration. Finally Chapter 6 contains the concluding remarks and the further directions of research. In particular we wish to mention an ongoing work (Dotto & Farcomeni, In preparation) in which we consider the possibility of implementing robust parsimonious Gaussian clustering models. Within the chapter, the algorithm is briefly described and some illustrative examples are also provided. The potential advantages of such proposals are the following. First of all, by considering the parsimonious models introduced in Celeux & Govaert (1995), the user is able to impose the shape of the detected clusters, which often, in the applications, plays a key role. Secondly, by constraining the shape of the detected clusters, the constraint on the eigenvalue ratio can be avoided. This leads to the removal of a tuning parameter of the procedure and, at the same time, allows the user to obtain ane equivariant estimators. Finally, since the possibility of trimming a xed proportion of observations is allowed, then the procedure is also formally robust.