ON THE USE OF CLUSTER ANALYSIS FOR INDIVIDUATING VARIABLE INFLUENCE ON SPREAD VARIATION IN LARGE DATASETS

Gabriella Schoier, Adriana Monte

Introduction

In a global financial market with an ever increasing number of financial products, investors and financial organs faced an ever increasing risk associated with their asset allocation or their investment strategies. In this context the problem of the behavior of portfolios of credit risk corporate assets such as bonds has become very important as the probability of default for a company can be estimated from the prices of bonds it has issued. In particular we consider the influence of different variables on spread variation in a portfolio of bonds. We want to obtain clusters of units (bonds) which must be homogeneous inside and heterogeneous outside.

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been considered in many contexts. It is useful in several exploratory pattern-analysis, grouping, decision-making and machine-learning situations. Different clustering algorithm have been proposed (see e.g. Fung, 2001), however several clustering methods have been criticized due to the lack of theoretical robustness both from a mathematical and a probabilistic point of view. For this reason model based clustering which can be defined as a set of clustering procedures based on finite mixture models are being increasingly preferred over heuristic methods (McLachlan, 2010, Ingrassia et al., 2012). This type of models can be used in different fields concerning clustering to high-dimensional data too (see e. g. Fraley, 2002, McLachlan, 2010).

In this paper we apply the mixture model approach and compare it with the classical K-means approach for analyzing the influence of some financial variables on spread variation in a portfolio of bonds. In the second paragraph we defined the problem, in the third the used methodology while in the fourth we present the application.

1. Problem definition

1.1. The dataset

In order to build up clusters of bonds homogeneous according to spread behavior, a data set of 7100 records (one for each bond) referred to the year 2012 has been considered .The information are both qualitative and quantitative, here is the list: *Bond type*, *Type of bond sector*, *Subordination level*, *Government coverage*, *Date of maturity*, *Bucket of maturity*, *Coupon frequency*, *Rating of the bond*, *Rating of the issuer*, *Country of the bond*, *Currency*, *Market price*, *Yield to maturity*, *Spread*. As the *Spread* is the "key" variable we have not considered records with missing values for it, so 6400 records have been used in the analysis. Some variables have required a preliminary transformation, in particular the dichotomous variables *Bond type* and *Government coverage* have been transformed in binary variables and the ordinal variables *Rating of the bond* and *Rating of the issuer* have been recoded into discrete variables so to assume constant distance between two contiguous rating levels. The variable *Subordination level* has been transformed from a qualitative into a quantitative variable using information about the probability of payment in relation to each subordination level. Another quantitative variable is *Coupon frequency* that is the number of coupons into the year. From the information codified in *Date of maturity* (day:month:year) we used only the year; *Bucket of maturity* is treated as an interval variable taking account of the central values of the classes. In addition to the *Spread* there are other two continuous variables: *Market price* and *Yield to maturity*. In the dataset there are also three categorical variables, that is *Type of bond sector*, *Country of the bond* and *Currency*; all these variables present a large number of categories.

The variables *Spread* and *Yield to maturity* are strong correlated; this is due to the fact that the *Yield* is the rate of return anticipated on a bond if it is held until the maturity date. The other variables are not so related to the *Spread.*

2. The methodology: model based clustering versus K-means algorithm

The base assumption of model based clustering methods is that the data are generated by an underlying mixture of a finite number of distributions. The objective is to identify the parameters of each of them and their number. Usually the assumption is to take the component distributions to be multivariate normal (Banfield *et al.*, 1993). The basic concept of model based clustering is that of mixture model (Lindsay, 1995, McLachlan, 2007).

Given $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ a random sample of size n, dove $\boldsymbol{Y}_j$ is a p-dimensional a random vector with density probability function $f(\boldsymbol{y}_j)$ on $\mathbb{R}^p$. Let $\boldsymbol{Y}$ be a random vector consisting of p features $\boldsymbol{Y} = (\boldsymbol{Y}_1{}^T, \dots, \boldsymbol{Y}_n{}^T)^T$ while let $\boldsymbol{y} = (\boldsymbol{y}_1{}^T, \dots, \boldsymbol{y}_n{}^T)^T$ be an observed random sample of size n on $\boldsymbol{Y}$.

With the finite mixture model based approach to density and clustering (McLachlan *et al.*, 2000), the density $f(\boldsymbol{y}_j)$ of $\boldsymbol{Y}_j$ (one of the g density components of the mixture) can be written as:

$$f(\boldsymbol{y}_j) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}_j) \qquad (1)$$

where $f_i(\boldsymbol{y}_j)$ are the component densities of the mixture and π_i are some unknown proportions such as:

$$0 \leq \pi_i \leq 1 \quad (i = 1, \dots, g),\ \sum_{i=1}^{g} \pi_i = 1.$$

The number of components g can be taken sufficiently large to provide accurate estimate of the underlying density function. For clustering purpose each of the g components correspond to a cluster.

The posterior probability that an observation, on which $\boldsymbol{y}_j$ has been observed, belongs to the i-th component of the mixture is

$$\tau_i(\boldsymbol{y}_j) = \pi_i f_i(\boldsymbol{y}_j) / f(\boldsymbol{y}_j) \quad \text{for g=1,…,g; j=1,…,n} \qquad (2)$$

A probabilistic clustering of the data in g clusters can be obtained in terms of the fitted posterior probabilities of component membership for the data as given in (2). It is possible to obtain a partition of the observations in g nonoverlapping clusters $C_1, \dots, C_g$ assigning each observation to the component to which it has the highest estimated posterior probability of belonging. In this way the i-th cluster C_i contains all the observations assigned to group G_i.

Formally C_i contains those observations j with $\hat{z}_{ij} = (\hat{\boldsymbol{z}}_j)_i = 1$, where

$$\hat{z}_{ij} = \begin{cases} 1\ if\ \hat{\tau}_i(\boldsymbol{y}_j) \geq \hat{\tau}_h(\boldsymbol{y}_j) & (i, h = 1, \dots, g; h \neq i\ j = 1, \dots, n) \\ \quad 0 & otherwise \end{cases}$$

with $\hat{\tau}_i(\boldsymbol{y}_j)$ an estimate of $\tau_i(\boldsymbol{y}_j)$.

According to this notation $\hat{z}_{ij}$ can be viewed as an estimate of z_{ij} which, under the hypothesis that the observations come from a mixture of g groups $G_1, \dots, G_g$, is defined to be one or zero accordingly to the fact that the j-th observation does or does not come from G_i $(i = 1, \dots, g; j = 1, \dots, n)$.

The model can be fitted to the data using the maximum likelihood estimation method implemented via the *EM* (*Expectation Maximization*) algorithm (Dempster et al., 1977, McLachlan et al., 1997). Different models can be obtained.

On the other side K-means algorithm is the simplest and the most popular in the class of partitional algorithms for cluster analysis. The method allows to find a partition into *k* clusters that minimizes the square error between the empirical mean of a cluster and the points in the same cluster. It is necessary to specify “a priori” the number of clusters (*k*), the initial centers (seeds) and a distance metric (the most used metric in K-means method is the Euclidean metric). The procedure is iterative and the steps are:

1. *Choose k points (seeds) into the dataset, to use as initial group centers.*
2. *Assign each unit of the dataset to the group that has the closest center.*
3. *For all the k groups recalculate the centers.*
4. *Repeat from step 2 until the centers get stable.*

The algorithm was first proposed over 50 years ago (Jain, 2010). The advances in storage technology and the developments of Data Mining techniques produced a lot of extensions of the K-means in order to cluster large data sets containing both numerical and categorical variables (Huang, 1998). In this paper we use the K-means algorithm in the standard form on a large dataset with mixed variables, then we test the algorithm comparing the results with those obtained from the model based clustering method.

3. The Application

3.1. Variable choice: the regression method

In this study the regression analysis is used in order to choice the variables for the subsequent cluster analysis. As the aim is to identify homogenous clusters regarding the *Spread*, this variable is used as the dependent variable of a multiple regression model while the independent variables are individuated within the dataset by a stepwise regression. In this approach we have considered all the quantitative variables and two binary variables. We have also explored the influence of the categorical variables *Subordination level*, *Country of the bond*, *Currency*, for each of these variables, a reclassification of the categories, in order to reduce their number, has been carried out, then each of these categories has been transformed into binary variables. No significance has been obtained for these categorical variables, for this reason they are not involved in the clustering. At this point we have leaved out the variables with multicollinearity problems and so we have obtained the final regression model described in tab. 1. As one can see this regression is highly explicative.

3.2. Cluster Identification

Both the algorithms identify eight clusters formed more or less by the same bonds. The BIC criterion is used to compare the different models in the model based clustering (see fig. 1 and tab. 2) while the Anova is considered in the K-means algorithm (see tab. 3). About the R^2 index the best value of it turns up for k=8. Referring to the model based clustering the Mclust package of the *R* language has been used for the analysis (Fraley, 2012). The best values for the BIC criterium regards model "VEV" with seven or eight components.

Table 1 – *Regression model description.*

R^2 = 0.952 Adjusted R^2= 0.906 F=12374.258 (Sig.=0.000)	Unstandardized coefficients		T	Sig.	Collinearity statistics	
	B	Std. error			Tolerance	VIF
Constant	-97.555	5.222	-18.681	0.000		
Bond type	-20.418	4.236	-4.820	0.000	0.756	1.323
Government coverage	-21.032	2.673	-7.868	0.000	0.934	1.071
Coupon frequency	-4.888	0.942	-5.186	0.000	0.764	1.309
Rating of the bond	1.452	0.420	3.457	0.001	0.571	1.751
Yield to maturity	89.896	0.475	189.306	0.000	0.596	1.678

Figure 1 – *BIC values for the different models.*

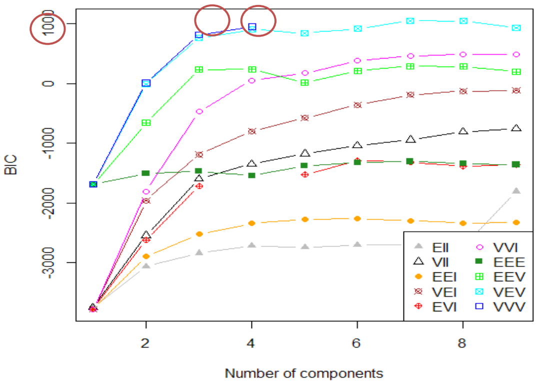

Table 2 – *BIC values for the different models.*

	1	2	3	4	5	6	7	8	9
EII	-3748.9	-3061.0	-2838.6	-2716.1	-2743.7	-2704.9	-2693.6	-2731.3	-1807.6
VII	-3748.9	-2543.5	-1592.2	-1345.2	-1176.8	-1039.7	-942.2	-806.9	-752.0
EEI	-3775.9	-2893.0	-2516.9	-2340.2	-2272.6	-2259.9	-2293.2	-2330.9	-2329.1
VEI	-3775.9	-1957.0	-1189.4	-796.7	-575.0	-358.3	-187.9	-1126.5	-116.7
EVI	-3775.9	-2617.4	-1714.9	NA	-1522.3	-1288.7	-1314.7	-1387.3	-1357.2
VVI	-3775.9	-1803.4	-472.6	55.3	174.5	381.8	461.3	493.4	484.4
EEE	-1686.8	-1501.7	-1469.2	-1541.6	-1373.0	-1325.1	-1303.8	-1341.5	-1355.4
EEV	-1686.8	-654.2	235.6	238.2	12.3	214.9	289.2	283.7	195.2
VEV	-1686.8	-10.6	764.7	906.1	842.7	920.4	1053.0	1044.8	937.8
VVV	-1686.8	11.8	819.1	952.1	NA	NA	NA	NA	NA

Table 3 – *Clusters Identification on the basis of the ANOVA.*

	Variance Between	Df	Variance Within	df	F	Standard deviation	R^2
Bond type	2.248	7	0.099	6392	22.63	63.25	0.99
Government coverage	7.281	7	0.199	6392	36.59	54.10	0.99
Coupon frequency	29.366	7	2.004	6392	14.66	73.45	0.99
Rating of the bond	5613.342	7	7.576	6392	740.91	112.02	0.99
Yield to maturity	8253.320	7	1.241	6392	665.14	137.16	0.99

As we want to identify clusters homogeneous regarding the *Spread*, first of all we describe its behavior among clusters and inside the clusters. In tab. 4 clusters are ordered for increasing values of the *Spread*: the number of bonds inside the clusters decreases with the increasing of the *Spread* and bonds with small values of the *Spread* belong to the same cluster (Cluster 7) that is also the larger one. The last cluster (Cluster 8) is formed only by three bonds with very large values for the *Spread*. The nonoverlapping of the values of the *Spread* derives from the method used in the variables choice *i.e.* the regression analysis.

Table 4 – *Distribution of the Spread into clusters.*

	Number of units	Minimum	Maximum	Mean	Standard deviation
Cluster 7	2794	-543.87	89.342	12.23	63.25
Cluster 1	1930	89.57	287.30	166.45	54.10
Cluster 5	1223	287.78	574.20	408.07	73.45
Cluster 6	313	575.14	981.38	739.60	112.02
Cluster 2	91	983.07	1596.86	1219.88	177.23
Cluster 4	33	1651.81	2367.76	1996.45	198.68
Cluster 3	13	2427.86	3172.69	2770.78	261.30
Cluster 8	3	3634.43	3885.09	3791.94	137.16

In order to describe the eight clusters we examine the variables used for the analysis and the others belonging to the dataset. We note that in the clusters where the values of the *Spread* are smaller the bonds are mainly fixed rate bonds. The *Yield to maturity* increases as the *Spread* increases.

As regard the variables not used in the clustering, the *Market price* mean decreases from Cluster 7 to the last one (Cluster 8) where the *Spread* has the highest values. In the clusters which present a smaller risk measured by the Spread there is a variety of currencies, in particular in Cluster 7 there are all the types of currencies. The currency of 72% of the bonds in portfolio is *euro* and its presence increases with the increasing of the *Spread*, this is due to the euro volatility. As regards the *Bucket of maturity* in Cluster 7 (that shows the best performance for

Spread values) the bonds have a lower distance to *Date of maturity* than the bonds in the remaining clusters. This distance increases with the *Spread*, on the contrary the *Subordination level* reveals decreasing values with respect to the probability of payment. Most of the bonds belong to the financial sector.

4. Conclusions

In this analysis we consider the model based clustering on mixture models and compare it with the classical K-means approach. The application regards the influence on some financial variables on spread variations in a portfolio of bonds and the subsequent clustering of bonds. The results obtained, after running the model based algorithm, are consistent with the K-means approach. Moreover the choice of model based clustering is supported from a theoretical point of view. In fact several clustering methods have been criticized due to the lack of theoretical robustness while the model based clustering, which can be defined as clustering procedures based on finite mixture models, have a strong mathematical and a probabilistic background. This type of models can be used in different fields concerning clustering to high-dimensional data too.

References

BANFIELD, J.D., RAFTERY A.E. 1993. Model-Based Gaussian and Non-Gaussian Clustering, *Biometrics*, Vol. 49, No. 3, pp. 803-821.

DEMPSTER A.P., LAIRD N.M., RUBIN D.B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol.39, No.1, pp.1-38.

FRALEY C., RAFTERY A.E., MURPHY T.B., SCRUCCA L. 2012. MCLUST Version 4 for R: normal mixture modeling and model-based clustering, classification and density estimation, *Technical Report*, No. 597, Department of Statistics, University of Washington.

FRALEY C., RAFTERY A.E. 2002. Model-Based Clustering, Discriminant Analysis and Density Estimation, *Journal of the American Statistical Association*, Vol. 97, No. 458, pp. 611-631.

FUNG G. 2001. A Comprehensive Overview of Basic Clustering Algorithms, http://pages.cs.wisc.edu/~gfung/, cited 06/2013.

INGRASSIA S., MINOTTI S.C., VITTADINI G., 2012. Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions, *Journal of Classification,* Vol. 29. No.3, pp. 363-401.

JAIN A.K. 2010. Data Clustering: 50 Years Beyond K-Means, *Pattern Recognition Letters*, Vol. 31, No. 8, pp. 651-666 .

LINDSAY B.G., 1995. *Mixture Models: Theory, Geometry and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics Volume 5,* Hayward, California: Institute of Mathematical Statistics.

McLACHLAN G.J., Ng S.K., WANG K. 2010. Clustering of High-Dimensional and Correlated Data. In PALUMBO F., LAURO C., GREENACRE M.J. (Eds), *Data Analysis and Classification: Proceedings of the 6th Conference of the Classification and Data Analysis Group of the Società italiana di Statistica, Macerata, Italy 12-14 September 2007,* Berlin; Heidelberg, Germany: Springer-Verlag, pp. 3-11.

MCLACHLAN G.J. 2007. Model-Based Clustering, http://www.maths.uq.edu.au/~gjm/, cited 06/2013.

McLACHLAN G.J., Peel D. 2000. *Finite Mixture Models*. New York: Wiley.

McLACHLAN G.J., KRISHNAN T. 1997. *The EM Algorithm and Extensions*. New York: Wiley.

RALAMBONDRAINY H. 1995. A Conceptual Version of K-means Algorithm, *Pattern Recognition Letters,* Vol. 16, No. 11, pp. 1147-1157.

SUMMARY

In this paper we consider the influence of different variables on spread variation in a portfolio of bonds. In order to choose the most relevant variables a preliminary regression analysis has been considered. In order to obtain cluster of units on the base of the variables selected by using a preliminary regression analysis two clustering methods have been considered: a classical *k* means cluster analysis and a model based clustering. As it is well known different clustering algorithm have been proposed in literature, however several clustering methods have been criticized due to the lack of theoretical robustness both from a mathematical and a probabilistic point of view. For this reason model based clustering - which can be defined as clustering procedures based on finite mixture models are being increasingly preferred over heuristic methods. This type of models can be used in a lot of fields . The application regards a portfolio of bonds on which a set of variables has been collected.

Gabriella SCHOIER, Dipartimento di Scienze Economiche Aziendali Matematiche e Statistiche, Università degli Studi di Trieste, gabriella.schoier@econ.units.it
Adriana MONTE, Dipartimento di Scienze Economiche Aziendali Matematiche e Statistiche, Università degli Studi di Trieste, adriana.monte@econ.units.it