Robust Clustering of Banks in Argentina |
José M. Vargas, Margarita Díaz y Fernando García |
Revista de Economía y Estadística | Vol. LVI | N° 1 | 2018 | pp. 21 - 41 | ISSN 0034-8066 | e-ISSN 2451-7321|
Instituto de Economía y Finanzas | Facultad de Ciencias Económicas | Universidad Nacional de Córdoba |
http://www.revistas.unc.edu.ar/index.php/REyE

Robust Clustering of Banks in Argentina

Agrupación robusta de Bancos en Argentina

José M. Vargas
Facultad de Ciencias Económicas, Universidad Nacional de Córdoba (Argentina)
jose.vargas@unc.edu.ar.

Margarita Díaz
Facultad de Ciencias Económicas, Universidad Nacional de Córdoba (Argentina)

Fernando García
Facultad de Ciencias Económicas, Universidad Nacional de Córdoba (Argentina)

Abstract

The purpose of this paper is to classify and characterize 64 banks, active as of 2010 in Argentina, by means of robust techniques used on information gathered during the period 2001-2010. Based on the strategy criteria established in (Wang 2007) and (Werbin 2010), seven variables were selected. In agreement with bank theory, four "natural" clusters were obtained, named "Personal", "Commercial", "Typical" and "Other banks". In order to understand this grouping, projection pursuit based robust principal component analysis was conducted on the whole set showing that essentially three variables can be attributed the formation of different clusters. In order to reveal each group inner structure, we used R package `mclust` *to fit a finite Gaussian mixture to the data. This revealed approximately a similar component structure, granting a common principal components analysis as in (Boente and Rodrigues, 2002). This allowed us to identify three variables which suffice for grouping and characterizing each cluster. Boente's influence measures were used to detect extreme cases in the common principal components analysis.*

* The authors gratefully acknowledge partial funding from Secretaría de Ciencia y Técnica (SeCyT) from Universidad Nacional de Córdoba, Argentina.

Keywords: Robust clustering, Projection pursuit, Common Principal Components, Robust K-means, Influence measures, Theory of the firm.

JEL Codes: C23, G21, L2

Resumen

El propósito de este documento es clasificar y caracterizar 64 bancos, activos en 2010 en la Argentina, mediante técnicas robustas utilizadas con información para el período 2001-2010. En base a los criterios de estrategia establecidos en (Wang 2007) y (Werbin 2010), se seleccionaron siete variables. De acuerdo con la teoría bancaria, se obtuvieron cuatro conglomerados "naturales", denominados "Personal", "Comercial", "Típico" y "Otros bancos". Para comprender este agrupamiento, se utilizó el todo el conjunto de banco y se realizó un análisis de los componentes principales basado en la proyección, que mostró que esencialmente tres variables pueden atribuirse a la formación de diferentes agrupaciones. A fin de revelar la estructura interna de cada grupo, utilizamos el paquete R `mclust` *para ajustar una mezcla gaussiana finita a los datos. Esto reveló aproximadamente una estructura de componentes similar, lo que garantiza un análisis de componentes principales comunes como en (Boente y Rodrigues, 2002). Esto nos permitió identificar tres variables que son suficientes para agrupar y caracterizar cada cluster. Las medidas de influencia de Boente se utilizaron para detectar casos extremos en el análisis de componentes principales comunes.*

Palabras clave: Agrupación robusta, Búsqueda de proyecciones, Componentes principales comunes, K-media robusta, Medidas de influencia, Teoría de la empresa.

Códigos JEL: C23, G21, L2

I. Introduction and some literature review

It has become evident only recently that banking segmentation might be the starting point of studies on different aspects of the banking industry, such as efficiency, cost structure, determinants of profitability or market strategy.

Sørensen and Puigvert Gutiérrez, 2006, use hierarchical cluster analysis with the objective of detecting some basic patterns in the euro area financial system in terms of the degree of homogeneity of countries. They focus on the degree of integration of the banking sector in the euro area countries over time in the period 1998-2004. They show that despite of the tendency to homogeneity induced by the euro differences still remain that tend to cluster banks together according to well defined geographical regions.

Dan Wang in two essays on US banking industry (Wang, 2007) reveals how market niches are created by selection of different market strategies to gain competitiveness, instead of the more traditional assumption of homogeneous technology from the cost function approach. For the same data set, he shows that those similarities among banks can be revealed by clustering techniques based on proxies of how banks create value.

In Kassani et al. (2015) branch performance and efficiency is analyzed on some 589 branches of a particular bank in Iran. They do so first by some hierarchical clustering (HCA) on efficiency scores obtained with data envelopment analysis (DEA) based on knowledge of bank management. At that stage, they perform a Reduced Multivariate Polynomial Pattern Classifier to model the class of the branches.

Ercan and Sayaseng (2016) conduct an exploratory study on the European banking sector by gathering ranges of consolidated banking indicators from the European Central Bank. They explore whether the foreign ownership of the banks contribute to the characteristic or clustering of these banks or it is a country specific composition. The data in their study is comprised of consolidate data from 26 countries in the European Union (EU) zone covering the period from 2008 to 2013. In their study they employ a Hierarchical Cluster Analysis to identify the clusters in EU Banking Sector. The variables used are Leverage, ROA, Tier 1, Capital requirement and equity/asset ratios.

Farné and Vouldis (2017) identify business models of the banks in the euro area by adopting an enhanced version of Vichi (2001) factorial k-means algorithm which incorporates a procedure to identify outliers within clusters. This approach combines dimensionality reduction and clustering. Their sample consists of 365 banks residing in the 19 euro area coun-

tries. They focus on Financial Reporting (FINREP) variables, providing a detailed decomposition of the balance sheet.

Werbin's study of Argentine banks during 2005-2007 period (Werbin, 2010), replicates Wang's analysis concluding that the ideal number of banks clusters is four. She used Hartigan (1975) F-test of variability reduction to determine the optimal number of clusters and K-means. However the methods used were not statistically robust.

The present work intends to group banks in Argentina adopting a criterion of offered products, following Wang (2007), which is characterized by a specific set of variables that define a particular market strategy on which decisions in the firm are made. In our work we use the same variables as in Werbin' work but on a larger period, comprising financial states from 2005 to 2010. This clustering is recovered by use of robust K-means R function RSKC. Being aware of work by Ding and He (2004), by which principal components of variance matrix are essentially the continuous solutions to cluster membership indicator functions of the K-means algorithm, as they span the same subspace as cluster centroids, we look into the principal components structure of the whole set of banks finding that the data is essentially bi-dimensional. As a byproduct, three variables can be selected among the seven to reproduce Werbin's clustering of banks with minimum loss. We further look into the principal components of each group finding approximately a similar structure, granting the assumptions necessary for a common principal components analysis of the four clusters. This allowed us to characterize each cluster membership in terms of three variables containing the previous two variables that suffice for clustering. Both PCA and CPC are performed in their robust projection pursuit version following Croux and Ruiz-Gazen (2005) and Boente et al. (2002).

The rest of this paper is organized as follows, in section 2 we make a somewhat detailed description of all the methods used specially for robust common principal components model as this topic might be novel to our readership; in section 3 we present and discuss our results paying attention as to how each method grants the next, in section 4 we make some concluding remarks highlighting the main findings of this research.

II. Methodology

II.1 Data

The sixty four active banks in Argentina as to the year 2010 were considered for this study. For each one of them the financial states corresponding to December months during the years from 2005 to 2010 were obtained from the Argentine central bank (BCRA).

In this work cluster analysis aims at identifying relatively homogeneous groups based on certain variables pertaining to a market strategy or offered products, following Wang (2007). The resulting groups should be characterized by a set of "strategic" variables that affect the decision making of the firm. In banking industry, these market strategies manifest themselves in several dimensions which include product mixtures, client perspective, size, geographical reach and sources of funding among others. As in Werbin (2010), we use seven variables:

1.Service revenues / Total income

2.Titles / Total assets

3.Deposits / Total assets

4.Implicit passive rate

5.Personal loans and credit cards / Total loans

6.Advances in checking accounts, document discounts and other commercial loans / Total loans

7.Net worth / Total assets

In contrast with Werbin's, we use a larger period base information and robust techniques.

II.2. Robust statistical methods

At its minimum, robust statistical methods solve similar problems than classical methods but are less affected by unusual cases. An advantage of robust estimation is a better detection of atypical cases through some form of influence measures. Frequently, classical methods perform poorly

in presence of atypical values. In order to classify banks, we choose robust versions of K-means and to characterize them robust PCA.

II.2.1 K-means

The algorithm *K*-means finds a partition $\pi=\{C_1,...,C_K\}$ of a given finite data set into *K* clusters such that within cluster variance is minimized. Some distance among pairs of data points is used and the number of clusters *K* is assumed to be known. If the euclidean distance is used, cluster variance within can be expressed in terms of cluster centroids (Kondo, 2011, p. 11) facilitating algorithmic computation of local optimal partitions. If π^* is an optimal partition,

$$\pi^* = \underset{\pi:|\pi|=K}{\operatorname{argmin}} SSW(\pi)$$

where

$$\mathrm{SSW}(\pi)=\sum_{k=1}^{K}\frac{1}{2n_k}\sum_{i,j\in C_k}\|x_i-x_j\|^2=\sum_{k=1}^{K}\sum_{j\in C_k}\|x_j-\bar{x}_k\|^2$$

No analytical solutions is known for an optimal partition. To solve the problem, several algorithms have been proposed, being Lloyd (1982) the most commonly used. Gordaliza (1991a and 1991b) introduced a robust version of Lloyd's algorithm known as "trimmed K-means" which is less influenced by outliers by trimming certain small proportion of most distanced cases in the computation of centroids before cluster reassignment is made. As minimizing within cluster dissimilarities can be viewed as maximizing between cluster dissimilarities, finding an optimal partition can be posed as a maximization problem. An alternative that solves a maximizing objective, known as "sparse K-means", was proposed by Witten and Tibshirani (2010) in order to make K-means algorithm less affected by certain type of noise. (Kondo, 2011) introduces a robust version of sparse K-means, known as "robust sparse K-means", by combining the idea behind trimmed K-means with that of sparse K-means. This algorithm is implemented in the R package RSKC and is the one we use in this work. For further details and a thorough review of the K-means algorithms, the reader is encouraged to read Kondo (2011).

2.2.2 Robust principal components

Classical principal component analysis seeks an orthogonal transformation of a set of observations of correlated variables into a set of linearly uncorrelated ones, named principal components. The first component is taken so that the set of observations has the largest possible variance in that direction; successively taking directions orthogonal to the previous ones that maximize the variance, a new set of variables is obtained which are independent if the original data is jointly normally distributed. In the original approach, the principal components are eigenvectors of the empirical covariance matrix giving each eigenvalue the variance in each component. Following Croux et al. (2005), this procedure can be made robust in two possible ways, one by robust estimation of the covariance matrix and the other by direct estimation of eigenvectors and eigenvalues with no use of the covariance matrix through a technique called projection pursue (PP) developed by Li and Chen (1985). The algorithm can be easily described. If $x_1,\ldots,x_n$ is the sample data in R^p, let $\hat{\mu}$ be a location center of the sample and S_n a scale estimator; then use the directions provided by the data in the hope that it will be dense in the principal components directions and define $\Gamma_1=\{(x_i-\hat{\mu})/\| x_i-\hat{\mu}\| \mid 1 \leq i \leq n\}$. Compute the first "eigenvector" as

$$\hat{\alpha}_1 = \underset{z \in \Gamma_1}{\operatorname{argmax}} S_n^2(z^t \mathbf{x_1}, \ldots, z^t \mathbf{x_n})$$

The scores on the first component are $S_{i,1} = \hat{\alpha}_1^t x_i$ for $1 \leq i \leq n$. Define recursively for $k=2,\ldots,p$, the scores on the previous components $S_{i,k-1} = \hat{\alpha}_{k-1}^t x_{i,k-1}$, the projected data on the orthogonal complement of the previous components by $x_{i,k} = x_{i,k-1} - s_{i,k-1}\, \hat{\alpha}_{k-1}$ and

$$\Gamma_{1,k}=\{(x_{1,k} - \hat{\mu})/\| x_{1,k} - \hat{\mu} \| \mid 1 \leq i \leq n\}$$

then the k-th component is

$$\hat{\alpha}_k = \underset{z \in \Gamma_k}{\operatorname{argmax}} S_n^2(z^t \mathbf{x_1}, \ldots, z^t \mathbf{x_n})$$

Estimation of the *k*-th eigenvalue is $\hat{\lambda}_k = S^2 (\hat{\alpha}_k^t \quad x_1,\ldots,\hat{\alpha}_k^t \quad x_n)$ and the covariance matrix can be estimated from $C = \sum_{k=1}^{p} \hat{\lambda}_{kk} \hat{\alpha}_k \hat{\alpha}_k^t$. In R language, this is implemented as PCAproj in the package `pcaPP` by (Filzmoser, 2012).

II.2.3 Common Principal Components

The model of Common Principal Components (CPC) generalizes the model of principal components to a given number of subpopulations which share the same principal components but with possibly different variances in each of the principal directions. In more detail, *K* sub-populations $x_{1j}, \ldots, x_{n_j j}$, $1 \leq j \leq K$, constituting the sample data in R^p, have a common dispersion structure according to a common principal components model if the covariance matrix of each subpopulation admits an orthogonal decomposition as follows

$$\Sigma_i = \beta^t \Lambda_i \beta; \quad 1 \leq i \leq K$$

where Σ_i is the covariance matrix of the i-th sub-population, Λ_i is the diagonal matrix with the variances of each principal direction corresponding to the *i*-th sub-population, and β is the orthogonal matrix whose columns are the principal components common to all the sub populations. The model CPC was introduced by Flury (1984) for the special case in which all Λ_i are assumed proportional among them, and by Flury (1988) for the more general case where maximum likelihood estimators of the model were studied. Croux and Ruiz-Gazen (1996) used the Projection-Pursuit algorithm to estimate the parameters of the usual model PCA. Simple and fast, this algorithm easily lends itself for robust estimation of PCA model parameters simply by considering robust estimators of position and dispersion in one dimension. A first implementation of this algorithm was written in Matlab and is still available in Croux (n.d.). R language implementations exist through the packages rrcov and pcaPP. Boente and Orellana (2001) introduced the Projection-Pursuit algorithm in the case of the CPC model by maximization of an aggregate variance obtained as a weighted sum by sub-population proportions. The first principal direction $\widehat{\beta_1}$ is selected to maximize the aggregate variance:

$$\hat{\beta}_1 = \underset{\|\beta\|=1}{\operatorname{argmax}} \sum_{j=1}^{K} \tau_j s^2(\beta^t \mathbf{x_{1j}}, \ldots, \beta^t \mathbf{x_{n_j j}})$$

where $\tau_j=n_j/n$ is the proportion of the j-th sub-population. As in the case of projection-pursue for PCA, data is orthogonally projected on the orthogonal complement of $\widehat{\beta_1}$, $\tilde{x}_{ij} = x_{ij} - \widehat{\beta_1}' x_{ij}$; the procedure is repeated to select the second principal direction $\widehat{\beta_2}$:

$$\hat{\beta}_2 = \underset{\|\beta\|=1}{\operatorname{argmax}} \sum_{j=1}^{K} \tau_j s^2(\beta^t \tilde{\mathbf{x}}_{\mathbf{1j}}, \ldots, \beta^t \tilde{\mathbf{x}}_{\mathbf{nj}})$$

In p steps, p mutually orthogonal principal directions are obtained. Individual variances are then obtained as the univariate variances of the projections of each sub-population along the principal directions. The implementation of this algorithm in Boente and Orellana (2001) is based on a modified version of Croux and Ruiz-Gazen algorithm in Matlab. In Boente and Rodrigues (2002) and Boente et al. (2010), the influence measures iml, for eigenvalues, and imb, for eigenvectors, were obtained. These functions are defined by:

$$IML_i(x,\hat{\beta},\hat{\lambda})^2 = \frac{1}{2}\sum_{r=1}^{p}\left(\frac{(\hat{\beta}_r^t x)^2 - \hat{\lambda}_{ir}}{\hat{\lambda}_{ir}}\right)^2$$

to measure the influence of a point x when estimating lambdas (eigenvalues) and:

$$IMB_i(x,\hat{\beta},\hat{\lambda})^2 = \sum_{r=1}^{p}\sum_{s\square r}\frac{(\hat{\beta}_r^t x \hat{\beta}_s^t x)^2}{\hat{\lambda}_{ir}\,\hat{\lambda}_{is}}$$

to measure the influence of a point x when estimating principal directions beta (eigenvectors). For this work we wrote R language implementations of the projection-pursue CPC algorithm and the corresponding influence functions iml and imb. Under assumption of normality, their asymptotic distributions are

$$G_\lambda^2 = \frac{1}{2}\sum_{r=1}^{p}(z_r^2 - 1)^2$$

for *IML*, and

$$G_\beta^2 = \sum_{r=1}^{p}\sum_{s\neq r} z_r^2 z_s^2$$

for imb, where $z_1,\ldots,z_p$ independent, N(0,1) distributed, variables. In order to compute critical values for these distributions, Monte Carlo simulations were conducted in R.

II.2.4 Model based clustering

Finally we pay some attention to model based clustering through the R package `mclust` (Fraley and Raftery, 2007 and Fraley et al., n.d.)). In `mclust` data is treated as coming from a finite mixture of Gaussian distributions. Each cluster is modeled as a multivariate normal distribution and an EM algorithm is used to fit the model and estimate the model parameters. Covariance matrices of each component are parameterized through eigenvalue decomposition as follows:

$$\Sigma_k = \lambda_k D_k A_k D_k^t,$$

Table 1. Model options available in the R package `mclust`.
K is the number of Gaussian components, or groups, and d is the dimension of the underlying space. Best model fitted by `mclust` VVI, on 4 components.

Identifier	Model	Qty. of covariance parameters	Distribution
EII	λI	1	Spherical
VII	$\lambda_k I$	K	Spherical
EEI	λA	d	Diagonal
VEI	$\lambda_k A$	$K + (d-1)$	Diagonal
EVI	λA_k	$1 + K(d-1)$	Diagonal
VVI	$\lambda_k A_k$	Kd	Diagonal
EEE	$\lambda D A D^t$	$d(d+1)/2$	Ellipsoidal
EEV	$\lambda D_k A D_k^t$	$1 + (d-1) + K[d(d-1)/2]$	Ellipsoidal
VEV	$\lambda_k D_k A D_k^t$	$K + (d-1) + K[d(d-1)/2]$	Ellipsoidal
VVV	$\lambda_k D_k A_k D_k^t$	$K[d(d+1)/2]$	Ellipsoidal

K is the number of Gaussian components, or groups, and d is the dimension of the underlying space. Best model fitted by mclust VVI, on 4 components.

III. RESULTS

As a first step we investigated the optimal number of clusters. We looked into several methods, within sum of squares and an F statistics proposed by Hartigan (1975), partitioning around medoids to estimate the number of clusters using the pamk function of the R package `fpc` (Hennig, n.d.), Bayesian Information Criterion for expectation-maximization for parameterized Gaussian mixture models using the function `mclust` of the R package `mclust` (Fraley et al., n.d.)), and the gap-statistic from (Tibshirani, at al 2001) (see also Chen (2010) for code implementing the gap-statistic.) All

Table 2: F statistic of Hartigan showing significance for 4 clusters.

Clusters	3	4	5
WithinSS	8.83619	6.35087	5.49846
F test		23.48	9.15

of them agree on four clusters as the most reasonable estimation for cluster dimension. Hartigan's F statistic is less than 10 at five clusters, suggesting four clusters as the optimal number. A summary is in Table 2 and Figure 1.

It is particularly important to draw attention to the best model estimated by `mclust`: VVI. This implies two things: first, the best fit is attained at a model with equal orientation (Identity) for all clusters, and, second, the common eigenvectors should be very close to being coordinate axis (variables themselves). This is the hypothesis necessary for using Common Principal Components (CPC) and its results will be in agreement with those of `mclust`.

In a second step, we used robust sparse k-means implemented in R package `RSKC` (Kondo, 2011) allowing for 10% trimming of outliers for the computation of cluster centers. In agreement with bank theory, we recover four clusters, named "Typical", "Other Banks", "Commercial" and "Personal". `RSKC` identified six banks as atypical; see Table 3.

Table 3. Cluster size and atypical per cluster.

Cluster	Typical	Other Banks	Commercial	Personal	Total
Size	24	12	14	14	64
Atypical	0	1	1	4	6

A close look into the variable values, shows which values are affecting those six atypical banks. We summarized this in Table 4. For comparison, medians per variable in each group are shown in Table 5.

An analysis of the medians allows to point to some specific characteristics of each group. Cluster "Other Banks" is characterized by the variable PN, which exhibits the largest median index. These are banks with an important proportion of capital which tend to invest in titles or inter-banks

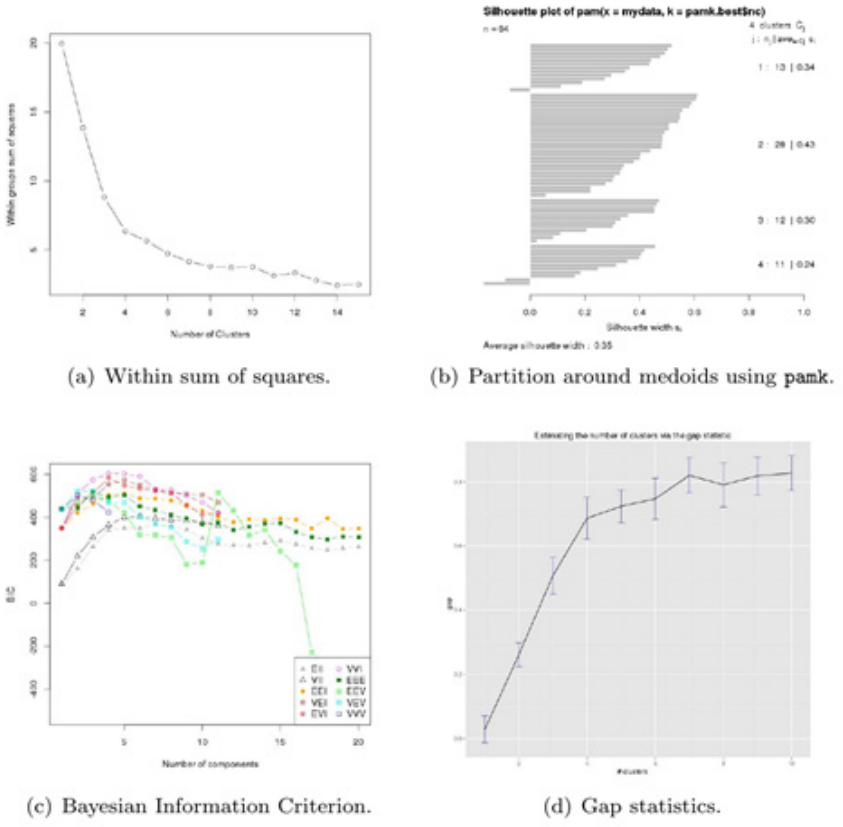
Table 4. Standardized values for atypical banks detected by RSKC.

	Other Banks	Commercial	Personal			
Atypical banks (codes)	(044)	(339)	(306)	(310)	(331)	(332)
Total income (ING)	0,8558	-1,7441	**3,6135**	**-3,6898**	-1,5170	**6,0935**
Titles (TIT)	0,0283	-1,2447	-0,6046	-1,4643	1,4738	-1,4738
Deposits (DEP)	0,2756	**-3,5217**	**-0,6681**	**-6,1929**	**-7,4279**	**-9,5485**
Implicit passive rate (TPI)	**5,7339**	-1,2500	**3,8088**	0,8000	**6,6765**	-1,4353
Personal loans (PER)	0,0000	-1,0000	1,6970	1,0396	1,3416	1,6118
Commercial loans (COM)	-0,3641	1,5611	-1,1703	-0,1566	-1,2482	-1,2482
Net worth / Total assets (PN)	**-2,5257**	**6,8237**	-1,6053	**7,8566**	0,1888	0,1581

Table 5. Medians of each variable per cluster

Variable	Typical Banks	Other Banks	Commercial	Personal
ING	0,2846	0,145	0,1782	0,2759
TIT	0,1891	0,2151	0,0931	0,1266
DEP	0,7631	0,1523	0,5277	0,7085
TPI	0,0386	0,0109	0,0547	0,0244
PER	0,3052	0,0000	0,0115	0,7567
COM	0,4067	0,2451	0,7878	0,1056
PN	0,1017	0,6044	0,12093	0,125

Figure 1. Optimal number of clusters by different methods.



(a) Within sum of squares.

(b) Partition around medoids using `pamk`.

(c) Bayesian Information Criterion.

(d) Gap statistics.

loans. One bank identified as atypical in this group shows unusually large value in the variable TPI and less than the group median in PN. "Commercial" banks are characterized by large values in variable COM, given that these banks have in their portfolios a majority of commercial loans. The atypical bank detected in this group shows unusually large value in the variable PN while a low value for DEP. "Personal" banks are characterized by large values in the variable PER, where this kind of loans represent most of their portfolio. The four atypical banks identified do not show a common pattern. Two of them exhibit larger values in TPI than the median of the group, while the opposite is true respect of DEP. Unlike the previous groups, "Typical Banks" do not show any particular variable to distinguish them from the rest. Even though high values are observed for DEP and ING, these

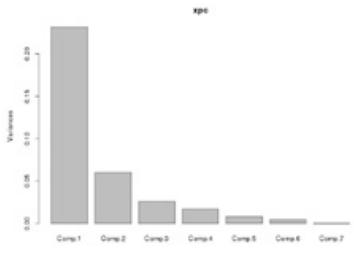
values are also high for "Personal" banks. These are banks having large proportion of deposits that intermediates as personal and commercial loans. No atypical cases are observed in this group.

In a third step, we investigated the principal components of the whole data set to obtain the smallest subset of the original variables that could reproduce the classification with little or no errors. We pay attention to those variables with highest loadings against the principal directions (see Table 6). We used R package `pcaPP` for robust estimation of principal components. The two largest eigenvalues represent 83.5% of the total variance and the three largest 91% (see Figure 2(a)).

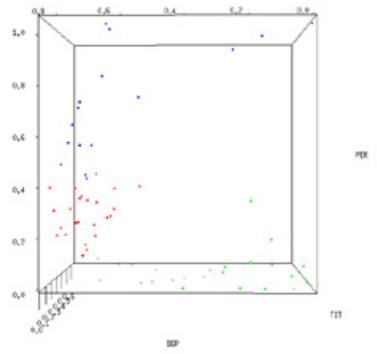
Table 6. Loadings for robust PCA.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
ING	0.162	0.102		-0.216	-0.378	0.876
TIT			**0.818**	0.490	0.251	0.152
DEP	0.394	**0.576**		-0.379	0.603	
TPI					-0.116	
PER	**0.639**		-0.424	0.627		
COM	-0.591	0.560	-0.331	0.412		0.212
PN	-0.245	-0.576	-0.189		0.636	0.392

Figure 2. Robust PCA after pcaPP



(a) Eigenvalues after pcaPP.



(b) A view of data in variables PER, DEP and TIT.

Table 6 suggests that a classification is possible considering just three variables: PER, DEP and TIT. Even PER and DEP should suffice: PER separates "Personal" and "Typical" banks, assigning them high and moderate values respectively, from the other two clusters to which assigns them low values; while DEP separates "Commercial" from "Other Banks", by assigning higher values to the first. Afterwards, this becomes apparent from Table 6. Observe that, as far as the first three components, variable PN is almost opposite to DEP and it could be an alternative choice to DEP. We preferred DEP as its loading over the third component is near zero. Several reclassification runs with RSKC were made based on these three variables. With errors around 10%, the reassignment of banks to each cluster was almost the same as using the seven original variables. A view of the whole data set in these variables is in Figure 2(b), where Typical Banks are in red color, Personal Banks in blue, Other Banks in grey and Commercial Banks in green.

Finally, in a fourth step, in order to understand the structure of each cluster itself, we performed common principal components analysis to the entire data assuming each cluster as a subpopulation, expecting different spectra to identify each group. In each group, the three greatest eigenvalues are concentrated in the first three components, except for "Typical Banks" where the third largest eigenvalue is in the fourth component (see Table 7).

Table 7. Explained variance per group in the first three and four components for robust CPC.

	Typical Banks	Other Banks	Commercial	Personal
Three	64%	68%	87%	83%
Four	80%	83%	91%	88%

This suggests paying attention at the first three common components to explain variability in each group. Table 8 shows the spectra per group. Clearly the cluster "Personal Banks" is dominated by the first component, "Other Banks" by the second, and "Commercial" by the third; while "Typical Banks" show, as expected by bank theory, a rather spherical behavior.

Observe, Table 9, that the variables with greatest loadings to the first three components are again PER, DEP and TIT, on the first, third and second

Table 8: Eigenvalues per group and component for robust CPC

Cluster	β_1	β_2	β_3	β_4	β_5	β_6	β_7
1 Typical	0.017370821	0.013988463	0.008056388	0.01004638	0.008245744	0.003699692	0.0005975094
2 Other	0.017501731	**0.135473283**	0.038569694	0.04119667	0.035492990	0.011448319	0.0008632282
3 Commercial	0.007653666	0.021799816	**0.039045141**	0.00339211	0.002715343	0.002319646	0.0017347925
4 Personal	**0.127270866**	0.008729003	0.019920530	0.01030537	0.005329917	0.014645502	0.0024193412

Table 9: Eigenvectors for robust CPC

Variable	β_1	β_2	β_3	β_4	β_5	β_6	β_7
ING	-0.14353786	-0.34380395	-0.001988733	0.35534239	-0.14109858	-0.76793038	-0.353974122
TIT	0.19461777	**-0.61955159**	0.344908978	-0.58867978	-0.30862426	0.07476819	-0.109244879
DEP	0.47598782	-0.16643094	**-0.846316955**	-0.15075104	-0.01361848	-0.05016366	-0.063688420
TPI	-0.05116156	-0.06427680	-0.019018891	-0.19890461	0.03904271	-0.45791108	0.861463484
PER	**-0.65027730**	0.08347308	-0.351258534	-0.16220699	-0.63704264	0.12010890	0.015117275
COM	0.51546941	0.52053777	0.202244779	0.03829549	-0.61698167	-0.20069756	0.004041191
PN	-0.15417396	0.43381629	-0.011585638	-0.66115152	0.31087417	-0.37119812	-0.341097442

component, respectively. And in general, each component tends to weigh heavily on one particular variable; for example, PN dominates the fourth component, PER and COM dominate the fifth, ING the sixth and TPI the seventh. This explains why `mclust` prefers a model with the Identity matrix as common orientation for all clusters (model VVI).

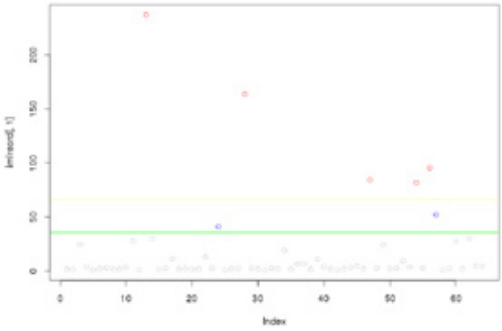
It is particularly important to note that the three variables producing the four clusters are also explaining the dispersion within each cluster.

Influence measures are an appreciated byproduct of robust CPC showing more atypical banks than `RSKC` and with some disagreement. All outliers detected by `RSKC` were detected by both, iml and imb, at 0.01 critical level, save one bank, coded (306), belonging to the "Personal" cluster, which was not detected by neither iml nor imb. This could be due to the way `RSKC` trims: it does it in a spherical way rather than considering the elongated dispersion natural to each cluster. Every outlier detected by iml was detected by imb (in general, this inclusion need not be the case). At a 0.001 critical level, iml detects three more banks as outliers than `RSKC`; and imb detects a total of eleven. This is an expected result given the relative low density of the data set. Yet, both influence measures pinpoint those banks which have special characteristics that deviate from their cluster of origin in a more sensitive way than `RSKC` does. In Figure 3, both influence measures are shown together with critical lines at 0.01 and 0.001. A summary of detected outliers is in Table 10.

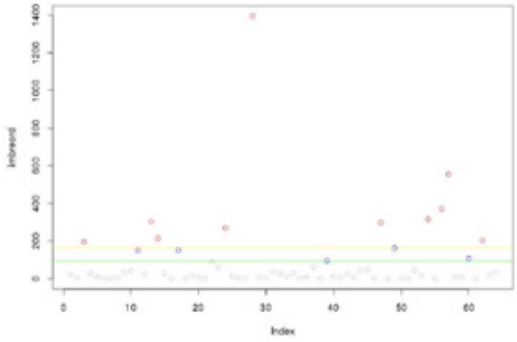
Table 10: Outliers detected by RSKC, iml and imb, at 0.001 level.

	Typical Banks	Other Banks	Commercial	Personal
RSKC		044	339	306, 310, 331, 332
iml		044, 147	198, 325	310, 331, 332
imb	011, 045, 341	044, 147	198, 312, 325	310, 331, 332.

Figure 3: Influence measures after robust CPC.



(a) Atypical banks detected by iml.



(b) Atypical banks detected by imb.

IV. CONCLUDING REMARKS

This research shows clear agreement of Wang's bank theory in the case of argentine banks; in particular, there is not one homogeneous group as classical bank theory predicts, but four well differentiated groups which can be characterized by a rather small set of economic variables defining strategic decisions of the firm. Correct assignment of banks to each group has been greatly improved by use of robust techniques, obtaining as a byproduct detection of outliers. Furthermore, robust spectral analysis through PCA and CPC, show that the phenomena producing the clusters are in fact dispersing

each cluster internally (except for the Typical Banks cluster, behaving in a rather spherical manner). Boente's influence measures associated to CPC have proved to be far more sensitive in detecting outliers than other usual methods.

V. References

Boente, G., and L. Orellana (2001). "A robust approach to common principal components". *Statistics in Genetics and in the Environmental Sciences*. Ed. by Birkhauser Basel et al., pp. 117–147.

Boente, G., A. M. Pires, and I. M.. Rodrigues (2002). "Influence functions and outlier detection under the common principal components model: A robust approach". *Biometrika* 89.4, pp. 861–875.

Boente, G., A. M. Pires, and I. M. Rodrigues (2010). "Detecting influential observations in principal components and common principal components". *Computational Statistics and Data Analysis* 54, pp. 2967–2975.

Chen, Edwin (2010). R Implementation of gap-statistics. [Retrieved from https://github.com/echen/gap-statistic/blob/master/gap-statistic.R].

Croux, C. personal website. [Retrieved from http://www.econ.kuleuven.be/public/NDBAE06/programs/\#pca].

Croux, C., P. Filzmoser, and M. R. Oliveira (2005). "Algorithms for projection-pursuit robust principal component analysis". Department of Decision Sciences an Information Management (KBI) KBI 0624.

Croux, C., and A. Ruiz-Gazen (1996). "A fast algorithm for robust principal components based on projection pursuit". *Compstat: Proceedings in Computational Statistics.* Ed. by A. Prat. Physica-Verlag, Heidelberg, pp. 211-217.

Croux, C., and A. Ruiz-Gazen(2005). "High breakdown estimators for principal components: The projection-pursuit approach revisited". *Journal of Multivariate Analysis* 95, pp. 206–226.

Ding, Chris, and Xiaofeng He (2004). "K-means clustering via principal component analysis". *Proceedings of the 21 St International Conference on Machine Learning*. Canada, 2004: Banff.

Ercan, H and S. Sayaseng (2016). "The cluster analysis of the banking sector in Europe". *Economics and Management of Global Value Chains,* pp. 111–127.

Farnè, M. and A. Vouldis (2017). "Business models of the banks in the euro area". Working Paper Series European Central Bank 2070.

Filzmoser, P. et al. (2012). Package "pcaPP".Peter Filzmoser, Heinrich Fritz and Klaudius Kalcher. [Retrieved from http://www.statistik.tuwien.ac.at/public/filz/]

Flury, B. K. (1984). "Common principal components in K Groups".*J. Amer. Statist. Assoc.* 79, pp. 892–898.

Flury, B. K. (1988). *Common principal components and related multivariate models*, Wiley, New York.

Fraley, C., and A. Raftery (2007). "Model-based methods of classification: Using mclust software in chemometrics". *Journal of Statistical Software* 18.6. [Retrieved from: http://www.jstatsoft.org/].

Fraley, C., A. Raftery, and Scrucca L. R package mclust Mantainer Luca Scrucca luca@stat.unipg.it.

Gordaliza, A. (1991a). "Best approximations to random variables based on trimming procedures". *Journal of Approximation Theory* 64.2, pp. 162–180.

Gordaliza, A. (1991b). "On the breakdown point of multivariate location estimators based on trimming procedures". *Statistics & Probability Letters* 11.5, pp. 387–394.

Hartigan, J. A. (1975). *Clustering Algorithms*. Inc: John Wiley & Sons.

Hennig, C. R Package fpc. c.hennig@ucl.ac.uk ucakche/ [Retrieved from http://www.homepages.ucl.ac.uk/].

Kassani, S.H., P. H. Kassani, and S. E. Najafi (2015). "Introducing a hybrid model of DEA and data mining in evaluating efficiency. Case study: Bank Branches". *Academic Journal of Research in Economics and Management* 3.2, pp. 72–80.

Kondo, Y. (2011). "Robustification of the sparse K-means clustering algorithm". University of British Columbia.

Kondo, Yumi, Matias Salibian-Barrera, and Ruben Zamar (2016). "RSKC: An R package for a robust and sparse K-means clustering algorithm". *Journal of Statistical Software* 72.5, pp. 1–26. [Retrieved from https://doi.org/10.18637/jss.v072.i05].

Li, G., and Z. Chen (1985). "Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo". J*ournal of the American Statistical Association* 80.391, pp. 759–766.

Lloyd, S. P. (1982). "Least squares quantization in PCM". *IEEE Transactions on Information Theory* 28.2, pp. 129–136.

Sørensen, C. K. and J. M. Puigvert Gutiérrez (2006). "Euro area banking sector integration using hierarchical cluster analysis techniques". Working Paper Series EuropeanCentral Bank 627.

Tibshirani, R., G. Whalter, and T. Hastie (2001). "Estimating the number of clusters in a data set via the gap statistic". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 Part 2, pp. 411–423.

Wang, D. (2007). "Three Essays on Bank Technology, cost Structure, and Performance". PhD Dissertation. State University of New York at Binghamton.

Werbin, Eliana (2010). "Los determinantes de la rentabilidad de los bancos en Argentina (2005 – 2007)". PhD thesis, Universidad Nacional de Córdoba.

Witten, D. M., and R. A Tibshirani (2010). "Framework for feature selection in clustering". *Journal of the American Statistical Association* 105.490, pp. 713–726.