Assistant Professor Ana GARCÍA-GALLEGO, PhD
E-mail: ana.ggallego@unileon.es
Associate Professor María-Jesús MURES-QUINTANA, PhD
E-mail: mj.mures@unileon.es
Universidad de León, Spain

PRINCIPAL COMPONENTS AND CANONICAL CORRELATION ANALYSES AS COMPLEMENTARY TOOLS. APPLICATION TO THE PROCESSING OF FINANCIAL INFORMATION

Abstract. *The processing of huge amounts of data (big data) whose generation has been fostered by advances in the information and communications technologies involves a high cost of timing and resources, which can be simplified by the application of data reduction statistical methods, such as Principal Components Analysis (PCA). In this paper a PCA is applied in order to prove its usefulness in reducing financial information expressed as ratios. The achieved results, in terms of variable selection, are next justified by the application of a Canonical Correlation Analysis (CCA). The use of both methods shows they are complementary, since the ratios correlated to the extracted factors in PCA are also important in defining the canonical variates in CCA, showing the relationship between them.*

Keywords*: principal components analysis, canonical correlation analysis, financial information, ratios, big data.*

JEL Classification: O30

1. Introduction

Advances in the information and communications technologies over the last decades have fostered the generation of huge amounts of data, and their processing involves a high cost in terms of timing and resources. Moreover, the higher the quantity of available information, the more probable it is that information is redundant, which has negative consequences on the event of interest. Therefore, tools like data reduction statistical methods are needed in order to simplify and make the decision-making process easier.

Within these techniques, one of the main methods is Principal Components Analysis (PCA), which is appropriate when data are of quantitative nature, like for example accounting information that firms publish in their statements and that is usually defined as financial ratios. Due to the large number of relationships which can be set among accounting entries, there are a lot of financial ratios that can be computed. Thus, the previous application of a PCA allows us to reduce the number of potentially explanatory variables of the event of interest to a smaller group of

factors which explain a high percentage of the original variance, with the advantage of removing those ratios which do not contribute to explaining the event or contain redundant information (García-Ayuso, 1996) and trying to lose the least possible amount of information (Jiménez, 1996). Besides, it helps to solve the problem of multicollinearity among variables, since the extracted factors are uncorrelated (Lizarraga, 2002).

In order to show the possibilities of PCA in the field of financial analysis, we will focus on one of the most important research streams, such as the study of business insolvency or failure. Specifically, our target is to determine those financial factors that characterise business failure by selecting them through the application of PCA and which can be considered as the starting point for the development of failure prediction models. Complementary, a Canonical Correlation Analysis (CCA) is applied, whose objective is to analyse whether the group of selected ratios by PCA is related to the group of removed ratios, which would mean that the information supplied for both sets of variables is redundant and the elimination of the latter would be appropriate. Finally, a study of the relationship between both methods is carried out, by comparing the results obtained with each analysis in our empirical study.

To achieve our aims, the paper is organised as follows: first, we describe the methodology. Next, the empirical analysis is developed, which starts with the selection of a set of financial ratios as explanatory variables of business failure whose number is reduced through the application of a PCA; after that a CCA is applied, to conclude with a comparison of the results obtained with both methods. The paper finishes with the concluding remarks.

2. Methodology

In this section, a brief description of the theory on which each method is based is given. Nevertheless, as PCA is a well-known method due to its great applicability, we just describe certain features related to its interpretation[1] and focus on the description of CCA, which 'is less popular than many other methods' (Hair et al., 2010: 235).

2.1 Principal components analysis (PCA)

PCA is a data reduction statistical method that is widely applied when dealing with a large amount of quantitative information, in order to reduce the number of variables to consider. Its objective is to obtain new orthogonal and linearly independent variables, which are named factors or principal components, as linear combinations of the original variables.

[1] For a deeper development of the method, see Mardia et al. (1979), Rencher (2002) or Hair et al. (2010).

Regarding the number of factors to be extracted, it is necessary to take into account the percentage of variance which is explained by each factor and which also measures their relative importance. Mardia et al. (1979) point out as a practical rule to extract an enough number of factors so that they explain at least 90% of the original variance. However, Hair et al. (2010: 109) indicate this criterion should not be used 'in the social sciences, where information is often less precise', so they consider as satisfactory 'a solution that accounts for 60 percent of the total variance (and in some instances even less)'. In any case, as Mallo (1985) notes, a balance between the highest percentage of total variance and the summary of this variance in a reduced number of components must be achieved. Actually, in most research papers a percentage between 70 and 80% is usually enough, and this percentage is usually explained by 3 or 4 components[2].

The extracted factors are interpreted by their correlations with the original variables, which allow giving a description to the components.

2.2 Canonical correlation analysis (CCA)

CCA is considered by several researchers as the most generalised multivariate statistical technique, on which other statistical methods are based, such as multiple regression, ANOVA or PCA. In this sense, Tatsuoka (1971: 183) defines canonical analysis 'as a sort of "double-barrelled principal components analysis."'

The form of data for canonical analysis is a table T with n rows, which are the number of observations or individuals, and $p+q$ columns, which is partitioned into two subtables X and Y, with p and q columns, respectively, which represent two groups or sets of quantitative variables. Without loss of generality the variables are supposed to be centred and reduced, that is, they are divided by their standard deviation, so the total addition in each column in T is zero.The aim is to study the relationships between both sets of variables, by obtaining linear combinations of the original variables, called canonical variates, which present the highest correlation with each other and, simultaneously, are incorrelated to the rest of linear combinations that can be obtained.

- The canonical correlation coefficient measures the correlation between the linear combinations in each group of original variables, X and Y, and is obtained from the linear correlation coefficient. The squared coefficient corresponds to the eigenvalues or canonical roots of the square matrixes:

  $$(X'X)^{-1}X'Y(Y'Y)^{-1}Y'X \text{ or } (Y'Y)^{-1}Y'X(X'X)^{-1}X'Y \qquad (1)$$

  of order p and q, respectively.
- The first eigenvalue accounts for the highest correlation between the pairs of canonical variates, whereas the rest of eigenvalues are obtained in descending

[2] This number of factors is usually similar to the number of factors whose eigenvalues are higher than one, that is, the average eigenvalue or average variance of variables, which is another criterion used to decide the number of factors to extract (Mardia et al., 1979; Rencher, 2002; Hair et al., 2010).

order of correlation. The maximum number of non-zero canonical roots that are obtained is the number of variables in the smallest set, that is, min [*p*,*q*].

- The coefficients defining the canonical variates are obtained as the eigenvectors associated to the highest canonical roots in the above square matrixes, which is the first eigenvalue. The coefficients in the linear combinations for the set of variables *X* are obtained from the first matrix, whereas the second one allows obtaining the coefficients vectors for the linear combinations of variables *Y*.

In order to interpret the canonical variates, the following issues are used:

- **Standardised coefficients (canonical weights)**: the coefficients for the original variables in the linear combinations are their relative contribution to each canonical variate, but they can be influenced by the existence of multicollinearity, so it is more usual to interpret the structure coefficients.
- **Structure coefficients (canonical loadings)**: they measure the level in which the original variables in each set are represented by the respective canonical variates, as they are the correlation between them. The proportion of variance of the variable that is the explained by the canonical variate is the squared coefficient and the mean for each group is the total proportion of explained variance in the group. The total addition measures the proportion of variance in a set of variables that is explained by all canonical variates which are obtained for that set.
- **Canonical cross-loadings**: they measure the correlation between the original variables in a group and the canonical variates obtained for the other one. They are computed as the product of the above structure coefficients and the corresponding canonical correlation coefficient.
- **Redundancy index**: it was proposed by Stewart and Love (1968), as a measure of the correlation between both sets of original variables, since canonical correlations measure the correlation between the linear combinations for each set of variables, but not between the groups themselves. The squared canonical correlation –which accounts for the proportion of variance in a canonical variate of each pair that is explained by the other canonical variate in the pair– is multiplied by the variance explained by each canonical variate. Their total addition for all canonical variates is the redundancy index, which measures the proportion of variance in a set of variables that is explained by the other set.

3. Application to the processing of financial information

One of the most important applications in the field of financial analysis, where data reduction methods are applied because of the huge amount of available quantitative information, focuses on the evaluation of business failure or insolvency, due to the consequences that a firm's disappearance has on a diversity of agents

acting in the economic system and involved in the firm (Ravi Kumar & Ravi, 2007; Wu, 2010). To anticipate the financial difficulties firms can face, a large list of business failure prediction models have been developed, after Beaver's (1966) and Altman's (1968) models, which are considered to be pioneering in this field (Wu, 2010). Most of them have used financial ratios relating different accounting entries in financial statements as explanatory variables to be included in the models, as they allow processing all financial information in an easier way and making comparisons both among different firms and different periods (Laffarga & Mora, 1998).

Nevertheless, due to the lack of a theory of business failure that could be a guide for the selection of ratios to be included in the models (Scott, 1981), it have been basically empirical, based on their popularity in literature and their predictive success in previous research, as Beaver (1966) did (Balcaen & Ooghe, 2006). This criterion has resulted in a large list of ratios potentially explanatory of business failure (Labatut et al., 2009), which makes it necessary the previous application of a data reduction method, such as PCA.

In order to analyse the possibilities of PCA in the processing of financial information, in this paper we try to identify those characteristics which better explain business failure, due to their correlation with the extracted factors by PCA. The subsequent application of a CCA allows enriching the results, in the sense of analysing the relationship between the group of selected ratios by PCA and the group of removed ratios.

3.1 Data collection

To develop our empirical analysis, it is necessary to select a sample of firms, which has been derived from those companies with head offices in the region of Castilla y León (Spain). We have collected information from their financial statements in the way of ratios by using the database SABI (*Sistema de Análisis de Balances Ibéricos*), prepared by the company Informa D&B.

Both failed and non-failed firms were selected for the sample, defining business failure as the firm's formal declaration of a juridical proceeding, either bankruptcy or liquidation or whatever other concepts, attending the current legislation in each country, since this legal definition has been the most used in the previously developed models, due to the advantages of being a highly visible legal event that can be objectively dated (Keasey & Watson, 1991).

Unlike most of previous models, which have used a paired sample with the same number of failed and non-failed firms (Balcaen and Ooghe, 2006; Wu, 2010) –which Zmijewski (1984) names a state-based sample–, we have selected a random sample whose size is proportional to the firms' population in the database and on the basis of its composition.

Specifically, 41,584 firms met the requirement of information availability for at least one economic year (the last one before failure for the failed firms and the last available for the non-failed ones), 59 of which were failed. Due to the low

rate of failing firms in a normal economic situation (Sánchez-Lasheras et al., 2012) and in order to ensure a big enough size for this group, all of them were selected to derive the failed firms subsample. Regarding the non-failed firms, we used the formulae appropriate to calculate this subsample size, taking into account the population size and for a maximum sampling error of 5%. Considering the usual confidence level of 95%, it resulted in a size of 396 companies. To respect characteristics and peculiarities of different industries in the population, they were chosen from the same industry in which failed companies developed its activity, attending each industry population size, by applying a stratified sampling method with proportional allocation.

Table 1 shows a summary of the study sample, where firms are classified according to their activity, as it is coded in the Spanish Industrial Classification of Economic Activities (*CNAE-93*) by using two digits, although they are joined in the four main industries which are identified in the economic system.

Table 1. Firms'sample

Industry	CNAE–93 Code	Failed	Non-failed
Agriculture	01	5	14
Manufacturing	14–36	22	81
Building	45	12	97
Service	50-85	20	238
	Total	**59**	**396**

The next step forward in the empirical analysis is the information collection. The selection of financial ratios to be considered as variables has been basically empirical, based on the previous studies on business failure, focusing on the ratios used (and being significant) in several of the previous models[3] (Bellowary et al., 2007), especially those of Beaver (1966) and Altman (1968), since these studies are considered to be pioneering in this field (Wu, 2010) and their ratios have been used in a large list of models developed after them.

[3] A review of the different models developed in the United States and other countries, both European and from the rest of the world, can be found in Balcaen and Ooghe (2006) and Ravi Kumar and Ravi (2007), among others. Regarding Spain, Laffarga and Mora (1998) include a review of the research carried out in this field.

Table 2. Financial ratios used as independent variables

Category	Name	Definition
Liquidity	CACL	Current ratio: Current assets / Current liabilities
	AT	Acid test: (Current assets – Inventories) / Current liabilities
	CCL	Quick ratio: Cash / Current liabilities
	WCTA	Working capital / Total assets
	WCE	Working capital / Equity
Profitability	ROA	Return on assets: Net income / Total assets
	ROE	Return on equity: Net income / Equity
	EBTE	Earnings before taxes / Equity
	EBTTA	Earnings before taxes / Total assets
Leverage and solvency	TLTA	Total liabilities / Total assets
	CLTA	Current liabilities / Total assets
	FLTA	Fixed liabilities / Total assets
	ETA	Equity / Total assets
	ECL	Equity / Current liabilities
	EFLFA	(Equity + Fixed liabilities) / Fixed assets
	ORFE	Operating result / Financial expenses
	FES	Financial expenses / Sales
Turnover and activity	STA	Sales / Total assets
	VarSALES	$Sales_t$ / $Sales_{t-1}$
	WCS	Working capital / Sales
	CAOI	Current assets / Operating income
Cash-flow	CFTA	Cash flow / Total assets
	CFTL	Cash flow / Total liabilities
	CFCL	Cash flow / Current liabilities
Economic structure	CATA	Current assets / Total assets
	FATA	Fixed assets / Total assets
	CTA	Cash / Total assets

Taking into account the data availability (Alfaro et al., 2008) and the relationships among ratios, so the information was not redundant, a final list of 27 financial ratios was chosen. In order to consider those ratios that measure the key

dimensions of a firm, they were selected so they belonged to the following six groups: liquidity, profitability, leverage and solvency, turnover and activity, cash-flow and economic structure, as it can be observed in Table 2.

3.2 Variable selection: Principal Components Analysis

On the initial list of financial ratios, a PCA was applied in order to reduce their number to a smaller set of factors which reflect the main business activity issues, that is, which explain a high percentage of the original variance, so removing the redundant information they might contain. The procedure was carried out in several steps using the statistical software SPAD 6.0, as it is next described.

Table 3. Explained variance (PCA with 27 variables)

Factor	Eigenvalue	Percentage	Cumulative percentage
1	4.3274	16.03	16.03
2	2.9956	11.09	27.12
3	2.8765	10.65	37.77
4	1.9594	7.26	45.03
5	1.7672	6.55	51.58
6	1.7544	6.50	58.08

From the initial PCA on the 27 financial ratios, whose results are presented in Table 3, it is deduced that the first six extracted factors, with eigenvalues higher than one, explain a 58.08% of the variance in the original variables. This percentage decreases to less than 38% when considering only three components.

From the value of the correlation coefficients between the original variables and the extracted factors, 21 ratios present a high correlation with any of the factors, which is also confirmed by the communalities of the ratios with the factors, since the remaining six ratios present a value below 0.5. Therefore, they are removed from the analysis in the next step. These ratios are: variation in sales, ROE, return on shareholders, the percentage of fixed resources on total assets, the ratio of operating result on financial expenses and the percentage of cash on total assets.

The application of PCA on the 21 remaining ratios[4] results in an increase of the explained variance to 47.75% with the first three factors and to 72.75% with six. In this step, both economic profitability ratios are correlated to different factors. As

[4] Due to the importance of these variables according to their use in several of the previously developed models, they were included as illustrative variables in PCA. However, the percentages of explained variance did not increase or even decreased, so they were permanently removed from the analysis.

they are two measures of the same business issue, but in a different way, they should be correlated to the same component. Since this is not the case, it leads to contradictory results and it is necessary to eliminate one of them. Due to the higher popularity of ROA in the literature on business failure we remove the other ratio.

In the next step forward, the communalities of two of the remaining 20 ratios, WCE and CAOI, are low, so thay are not well explained by the factors extracted in this PCA. Therefore, we decide to remove both ratios and carry out the analysis with the remaining 18 ratios. The percentage of variance explained by three factors is 54.50%, which increases to 81.84% when six components are extracted.

From a statistical viewpoint, the 18 ratios in the analysis in this step are correlated to the extracted factors. Consequently, we analyse those ratios from a financial perspective, in order to determine whether the financial relationships among accounting entries could allow us to eliminate some ratios which can have a negative effect on PCA, even though they are correlated to the factors.

In this sense, a firm's total debt is classified into fixed and current liabilities, depending on the maturity, greater than one year or not, respectively, so these ratios could contain redundant information. Due to the fact that for most firms in the sample under study the total debt is divided, exclusively, into long and short term debt, there is a high correlation among these three ratios, so it is necessary to delete any of them. Specifically, since the most serious problems are usually related to the payment of current debts, we only work with the percentage of current liabilities on total assets.

Table 4. Explained variance (PCA with 15 variables)

Factor	Eigenvalue	Percentage	Cumulative percentage
1	3.6438	24.29	24.29
2	2.5104	16.74	41.03
3	2.0237	13.49	54.52
4	1.8142	12.09	66.61
5	1.5211	10.14	76.75
6	1.2396	8.27	85.02

Furthermore, the only assets entries in our sample firms' financial statements are fixed and current assets, so there is a perfect correlation between the percentages of these two entries on total assets. Thus, we remove the first ratio from the analysis, since the ratio of current assets is more important in order to measure a firm's ability to pay debts and, therefore, its future failure.

In the last step, PCA was carried out on the 15 selected ratios, resulting in a percentage of variance explained by the first three factors of 54.52%, which increases to 85.02% if six factors are extracted, as it is shown in Table 4.

Table 5. Correlacions between variable and factor (PCA with 15 variables)

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
STA	0.06	0.28	**–0.85**	0.41	–0.18	0.02
ROA	0.08	0.31	**–0.82**	0.42	–0.20	0.01
FES	0.01	–0.09	0.17	**0.67**	**0.61**	0.05
WCS	0.11	–0.18	0.20	**0.73**	0.49	0.06
CACL	**0.71**	–0.44	–0.05	0.07	–0.07	–0.23
CCL	**0.84**	–0.44	–0.09	–0.05	–0.05	–0.10
AT	**0.66**	–0.29	–0.07	–0.09	–0.05	–0.01
CATA	–0.16	–0.26	0.39	0.45	**–0.65**	0.13
ETA	0.46	**0.66**	0.31	0.12	–0.09	–0.31
CLTA	–0.47	**–0.69**	–0.26	–0.10	0.02	0.35
WCTA	–0.25	–0.30	0.34	0.49	**–0.61**	–0.01
CFTA	0.39	**0.68**	0.30	0.10	–0.14	–0.03
CFTL	0.47	0.29	0.12	–0.05	0.00	**0.73**
CFCL	**0.66**	0.09	0.04	–0.04	–0.06	**0.61**
ECL	**0.80**	–0.45	–0.09	0.01	–0.03	–0.20

According to the correlation coefficients shown in Table 5, the extracted factors, which are correlated to the 15 selected variables, can be described as:

- **Liquidity**: correlated to the ratios of current assets and cash on current liabilities, as well as the acid test, together with the firms' ability to pay current debts with their resources, both internal (cash-flow) and external (equity).
- **Liability structure**: correlated to the proportions of internal and external resources and current liabilities on total assets.
- **Economic profitability**: correlated to ROA and assets turnover, which are the two components of profit margin.
- **Turnover**: correlated to two ratios measured on sales (financial expenses and working capital).

- **Current position**: correlated to the ratio of current assets and working capital on total assets.
- **Cash-flow**, measured on total debt and current liabilities.

To conclude, the communalities these 15 ratios have with the extracted factors confirm the obtained results and justify the reduction of the 27 original ratios to the 15 that have been selected. All values are high, close to one for many variables, and none of them is smaller than 0.5, which involves the 15 selected variables are greatly explained by the six extracted factors.

3.3 Results from Canonical Correlacion Analysis

CCA was applied by using the software STATISTICA in order to prove whether the removed ratios supply the same information contained in the selected ones, justifying in this way their removal.

Its objective is to analyse the relationship between two sets of variables: in this case, the group of 15 ratios selected by PCA, due to their correlation with the extracted factors, and the group of 12 ratios removed from the analysis, because they were not correlated to the factors and these ones did not explain them well, so their communalities were low.

For each of the six business activity issues considered for the study, there are ratios which have been selected (first set) and other ones that have been removed from the analysis (second set). Since they measure the same issue in a different way, ratios in each group are supposed to be related, that is, they contain similar or redundant information (LeClere, 2006), which would justify the selection of the 15 ratios, to the detriment of the 12 removed ones.

Twelve canonical variates are obtained as linear combinations of ratios in each group, which is the number of variables in the smallest set. According to the canonical correlation coefficients in Table 6 that measure the relationship between pairs of canonical variates, only five are significant at a 5% level for the Chi-square test. Therefore, only the first five variates are necessary to explain the relationship between both sets of variables.

Furthermore, there is a practically perfect relationship between the first four pairs of canonical variates, since the coefficients are one or very close. The value decreases to 0.835 for the fifth coefficient, approaching zero for the last canonical correlations.

Table 6. Chi-square significance test for canonical correlacion coefficients in CCA

Step	Eigenvalue	Canonical correlation	Chi-square	df	p-value	Lambda
0	1.0000	1.0000	31.190.39	180	0.0000	0.0000
1	1.0000	1.0000	—	154	0.0000	0.0000
2	1.0000	1.0000	9.347.77	130	0.0000	0.0000
3	0.9910	0.9955	1.979.25	108	0.0000	0.0018
4	0.6973	0.8350	504.17	88	0.0000	0.1997
5	0.1740	0.4171	130.17	70	0.0000	0.6598
6	0.0931	0.3051	70.33	54	0.0671	0.7988
7	0.0562	0.2371	39.75	40	0.4815	0.8807
8	0.0362	0.1904	21.64	28	0.7976	0.9332
9	0.0269	0.1640	10.08	18	0.9291	0.9683
10	0.0027	0.0521	1.55	10	0.9988	0.9951
11	0.0022	0.0472	0.70	4	0.9515	0.9978

The different canonical variates are defined by the canonical function coefficients or weights given to the variables in each linear combination. Nevertheless, they are usually interpreted by using the structure coefficients or canonical loadings, which are presented in Table 7 for the first five canonical variates, since they measure their correlation with the ratios in each group. That is why they are 'direct analogs' to the correlations between variables and factors in PCA (Thomson, 1984: 23) and allow us to give a name or description to the corresponding canonical variates.

The fact that the canonical correlations coefficients between the first canonical variates are one or very close to it involves the so called canonical cross-loadings are very similar to the former. Their values mean that the ratios in each group (selected and removed) that are correlated to their respective canonical variate are also highly correlated to the linear combinations obtained with the ratios in the other set.

On the other hand, the twelve canonical variates explain 100% of the variance in the group of removed ratios, since the number of canonical variates equals the number of variables in this set, which is the smallest one. They also explain a high percentage of the variance in the group of selected variables, which in total is 86.03%.

Table 7. Structure coefficients or canonical loadings in CCA

Group	Ratio	CV 1	CV 2	CV 3	CV 4	CV 5
Selected ratios	STA	–0.0637	–0.1181	0.0119	–0.0050	–0.0120
	ROA	–0.0519	–0.1143	0.0220	–0.0052	–0.0129
	FES	–0.0039	–0.0863	–0.1316	–0.4570	–0.0090
	WCS	0.1174	0.0481	–0.0897	–0.9809	0.0363
	CACL	0.3083	–0.1301	0.0574	–0.0181	0.0748
	CCL	0.2189	–0.0970	0.0972	0.0268	0.4975
	AT	0.2644	–0.0961	0.1097	0.0203	0.2194
	CATA	0.6169	0.7870	0.0000	0.0000	0.0000
	ETA	0.5972	–0.5432	0.5902	0.0000	0.0000
	CLTA	–0.5907	0.8069	–0.0001	0.0000	0.0000
	WCTA	0.4390	0.5942	–0.0589	–0.0555	–0.6500
	CFTA	0.4142	–0.3256	0.3841	0.0034	0.0373
	CFTL	0.2692	–0.1551	0.2289	0.0311	0.3483
	CFCL	0.2098	–0.0659	0.1779	0.0080	0.1965
	ECL	0.1062	–0.2364	0.1531	–0.0434	0.0338
Removed ratios	VarSales	–0.1218	–0.0272	0.0527	0.1304	0.0661
	ROE	–0.0653	0.0577	–0.0764	–0.0259	–0.0654
	CTA	0.2700	0.1296	0.1067	0.0743	0.9397
	FATA	–0.6169	–0.7870	0.0000	0.0000	0.0000
	TLTA	–0.5972	0.5432	–0.5902	0.0000	0.0000
	FLTA	–0.0835	–0.3215	–0.9432	0.0000	0.0000
	EDTA	0.1764	0.0325	–0.0360	0.0315	0.2088
	ORFE	0.0367	0.0298	–0.0234	0.0267	0.1436
	CAOI	0.1084	0.0528	–0.0867	–0.9871	0.0451
	WCE	0.0787	0.0649	–0.0572	0.1834	–0.2154
	EBTE	0.0014	0.1551	0.0018	–0.0158	–0.0208
	EBTTA	0.3992	–0.3274	0.1359	0.0073	0.0514

The proportion of variance in each set that is explained by the other set is the redundancy index (Stewart and Love, 1968). This is a more appropriate measure to interpret the results because a high correlation between the linear combinations in

each group of ratios (as canonical correlation coefficients may indicate) does not mean the correlation between both sets of variables themselves is also high (Hermoso, 2012).

The set of removed ratios explains 44.82% of the variance in the group of selected ratios, whereas the former explains 46.59% of the variance in the latter. Even though there is no general criterion about the minimum value for the redundancy index (Hermoso, 2012), these indexes are at the same level as the ones obtained in other papers, such as LeClere (2006) and indicate that both sets of variables share some common information, since each group explains a proportion of the variance in the other one. Therefore, removing some ratios from the analysis and focusing on a smaller number to define the firms' key financial dimensions is justified.

3.4 Comparison of results

Due to the similarity in the procedure both methods follow to achieve their results, as it is based on obtaining linear combinations of the original variables, and by observing the results from PCA and CCA from an overall perspective, we proceed to analyse the relationship between them.

From an exhaustive analysis of the previous results, it is deduced that the variables correlated to the extracted factors in PCA are also the variables with a higher weight in the canonical variates obtained in CCA. In this case, we focus on the first group of variables, which are the ratios selected by PCA, since the second set consist of the ratios that are not correlated to the extracted factors, so they were removed from the analysis.

According to the correlations in Table 5, the first factor is correlated to the liquidity ratios and the proportion of equity and cash-flow on current liabilities, which are correlated to the ninth canonical variate, as the structure coefficient for this canonical variate shows. The only difference is that the ratio of cash-flow on total liabilities has a higher weight than the ratio measured on current liabilities in the definition of the canonical variate, which can be due to the relationship between both types of debt. In fact, both ratios are correlated to the sixth extracted factor.

The proportion of current liabilities and equity on total assets are correlated to the second factor in PCA and they contribute to defining the two first canonical variates, as it is shown by the canonical loadings in Table 7, although the ratio of current assets on total assets is also important in defining it. This ratio, together with the percentage of working capital whose structure coefficient for the second canonical variate is high, are correlated to the fifth factor extracted by PCA. The turnover of financial expenses on sales is also correlated with this factor and it is the ratio with the highest weight in the definition of the tenth canonical variate. Moreover, the percentage of cash-flow on total assets is correlated to the second and eight factors, and it contributes to obtaining the sixth canonical variate.

On the other hand, the third factor in PCA is correlated to both components of the profit margin, which are the variables defining the last canonical variate, whereas the fourth factor is correlated with the turnover of working capital and financial expenses on sales, which are the ratios contributing to defining the fourth canonical variate, especially the first one.

Therefore, it is observed a strong coincidence between the extracted factors in PCA and the canonical variates obtained in CCA for the group of selected ratios, which are the ones correlated to the former, showing the relationship between both methods.

Furthermore, the percentages of explained variance in every method can also be considered as similar. In PCA the percentage of variance explained by the extracted factors, which are the number of original variables to be reduced, is 100%. The percentage of variance explained by the canonical variates in CCA is also 100% for the smallest group of variables, whose number is exactly the number of canonical variates that are obtained. Taking into account the fact that the number of variables in the other set is higher, the percentage of variance in this group that is explained by the other one is smaller than 100%. The percentage of explained variance also decreases in PCA when we retain a number of factors lower than the number of variables, according to the available criteria for the factors extraction. In this sense, the percentage of explained variance when retaining six factors has been 85%, whereas the percentage of variance explained by the canonical variates in the group of selected ratios, which is the largest one, has been 86%.

4. Concluding remarks

The aim of this paper has been to prove the usefulness of data reduction statistical methods in the processing of financial information which is extracted from the statements that firms publish and which reflects the key dimensions of a firm. As an easier way of processing all that information, it is usual to use financial ratios.

One of the fields where financial ratios are considered as explanatory or predictor variables of an event is in the development of business failure prediction models. Due to the lack of an economic theory on which the selection of ratios could be based, the list of variables used during the period of nearly 50 years of this research stream is large. That is why it is necessary to apply methods like PCA, which allows us to reduce the large list of ratios to consider to a smaller group of factors with a high explanatory power of the event under study.

Results from PCA lead to the identification of six principal components or factors that explain a high percentage of the variance in the original variables (85.02%). The first factor is correlated to the liquidity ratios and the ability to pay current debts; the second one with both the own resources of a firm, either internal or external, and debt; the next factor is correlated to the two components of the profit margin (ROA and assets turnover); the fourth one measures the turnover of

financial expenses and working capital on sales; the fifth factor describes the current position; and the last one is correlated with the firm's ability to generate resources in order to pay debts.

This analysis is completed by the application of a CCA, where it is observed there is a high correlation between the canonical variates that are obtained as linear combinations of the selected and removed ratios, respectively. Moreover, the redundancy index for the set of removed ratios, given the set of selected ratios, involves that the latter explains 47% of the variance in the former, which means that the information contained in each group is similar, justifying therefore the reduction of ratios to consider as explanatory variables of business failure.

Furthermore, a high degree of relationship between both methods is observed. First, they are based on the same procedure of obtaining linear combinations of the original variables, which also implies that results are similar, since it has been proved that the ratios correlated to the extracted factors in PCA are important in defining the different canonical variates in CCA (obviously, in the group of selected ratios).

For all the above reasons, it can be concluded that PCA is appropriate in the analysis of financial information, as long as a smaller number of financial ratios allows explaining a high percentage of the original information, with the advantages it involves regarding the huge amount of data to process. Moreover, it is recommendable that the study is not limited to the application of one only statistical method, since the combined application of two or more methods leads to more valuable results. In this regard, CCA contributes to supporting the results from PCA, as long as both methods are related.

REFERENCES

[1] **Alfaro, E., Gámez, M., García, N. (2008),** ***Linear Discriminant Analysis versus Adaboost for Failure Forecasting***; *Revista Española de Financiación y Contabilidad*, XXXVII(137): 13-32;
[2] **Altman, E.I. (1968),** ***Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy;*** *The Journal of Finance*, XXIII(4): 589-609;
[3] **Balcaen, S., Ooghe, H. (2006),** ***35 Years of Studies on Business Failure: An Overview of the Classical Statistical Methodologies and their Related Problems***; *The British Accounting Review*, 38(1): 63-93;
[4] **Beaver, W.H. (1966),** ***Financial Ratios as Predictors of Failure***; *Journal of Accounting Research*, Supplement to Vol. 4: Empirical Research in Accounting: Selected Studies: 71-111;

[5] **Bellovary, J. L., Giacomino, D. E., Akers, M. D. (2007),** ***A Review of Bankruptcy Prediction Studies: 1930 to Present***; *Journal of Financial Education,* 33: 1-42;
[6] **García-Ayuso, M. (1996),** ***Técnicas de Análisis Factorial Aplicadas al Análisis de la Información Financiera (Clasificaciones a priori, Hallazgos y Evidencia Empírica Española)***; *Revista Española de Financiación y Contabilidad,* XXV(86): 57-101;
[7] **Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. (2010),** ***Multivariate Data Analysis: A Global Perspective, 7th ed***; *Pearson Education, Inc.*;
[8] **Hermoso, J. A. (2012),** ***Correlaciones Canónicas***. In: *T. Luque (coord.): Técnicas de análisis de datos en investigación de mercados, 2ª ed., Ediciones Pirámide*: 505-523;
[9] **Jiménez, S. M. (1996),** ***Una evaluación crítica de la investigación empírica desarrollada en torno a la solvencia empresarial***; *Revista Española de Financiación y Contabilidad,* XXV(87): 459-479;
[10] **Keasey, K., Watson, R. (1991),** ***Financial Distress Prediction Models: A Review of their Usefulness***; *British Journal of Management,* 2(2): 89-102;
[11] **Labatut, G., Pozuelo, J., Veres, E. J. (2009),** ***Modelización temporal de los ratios contables en la detección del fracaso empresarial de la PYME española***; *Revista Española de Financiación y Contabilidad,* XXXVIII(143): 423-447;
[12] **Laffarga, J., Mora, A. (1998),** ***Los modelos de predicción de la insolvencia empresarial: Un análisis crítico***. In: A. *Calvo-Flores and D. García Pérez de Lema (eds.), El riesgo financiero de la empresa*, Madrid: *AECA*, 11-58.
[13] **LeClere, M. J. (2006),** ***Bankruptcy Studies and Ad Hoc Variable Selection: A Canonical Correlation Analysis*****.** *Review of Accounting and Finance*, 5(4), 410-422.
[14] **Lizarraga, F. (2002),** ***La utilidad de los modelos de predicción de fracaso en la empresa española a lo largo de la última década***. In: *F. Doldán and M. Rodríguez (coords.), La Gestión del Riesgo de Crédito. Métodos y Modelos de Predicción de la Insolvencia Empresarial*; *AECA Monografías*: 219-252;
[15] **Mallo, F. (1985),** ***Análisis de componentes principales y técnicas factoriales relacionadas***; *Universidad de León, Servicio de Publicaciones*;
[16] **Mardia, K. V., Kent, J. T., Bibby, J. M. (1979),** ***Multivariate Analysis***; *Academic Press Inc. Ltd.*;
[17] **Ravi Kumar, P., Ravi, V. (2007),** ***Bankruptcy Prediction in Banks and Firms Via Statistical and Intelligent Techniques - A Review***; *European Journal of Operational Research*, 180(1): 1-28;
[18] **Rencher, A. C. (2002),** ***Methods of Multivariate Analysis, 2nd ed.***; *John Wiley & Sons, Inc.*;
[19] **Sánchez-Las Heras, F., de Andrés, J., Lorca, P., de Cos, F. J. (2012),** ***A Hybrid Device for the Solution of Sampling Bias Problems in the Forecasting of Firms Bankruptcy***; *Expert Systems With Applications*, 39(8): 7512-7523;

[20] **Scott, J. (1981),** ***The Probability of Bankruptcy. A Comparison of Empirical Predictions and Theoretical Models***; *Journal of Banking & Finance*, 5: 317-344;
[21] **Stewart, D., Love, W. (1968),** ***A General Canonical Correlation Index***; *Psychological Bulletin*, 70(3): 160-163;
[22] **Tatsuoka, M. M. (1971),** ***Multivariate Analysis: Techniques for Educational and Psychological Research***; *John Wiley & Sons*;
[23] **Thomson, B. (1984),** ***Canonical Correlation Analysis. Uses and Interpretation***; *Sage University Paper Series on Quantitative Applications in the Social Sciences*, No. 07-047; *Sage Publications*;
[24] **Wu, W. (2010),** ***Beyond Business Failure Prediction***; *Expert Systems with Applications*, 37(3): 2371-2376;
[25] **Zmijewski, M.E. (1984),** ***Methodological Issues Related to the Estimation of Financial Distress Prediction Models***; *Journal of Accounting Research*, Supplement to Vol. 22: Studies on Current Econometric Issues in Accounting Research: 59-82.