

#### AUTHOR QUERIES

|        |                                                                                                                                                |
|--------|------------------------------------------------------------------------------------------------------------------------------------------------|
| [AQ01] | As per the style guide, abstract should not exceed 200 words. Please check.                                                                    |
| [AQ02] | Please confirm the citation of Tables 9–11 inserted in the text.                                                                               |
| [AQ03] | Reference citations and the corresponding references have been renumbered for the orderly citation in the text. Please check the changes made. |
| [AQ04] | Kindly provide complete details for reference [3].                                                                                             |
| [AQ05] | Please provide the publisher name, publisher location for reference [5].                                                                       |
| [AQ06] | Please provide the publisher name, publisher location for reference [6].                                                                       |
| [AQ07] | All figures are low resolution. Please provide high resolution figures with 300dpi or more.                                                    |

02 Principal Component Analysis in  
03 Financial Data Science04 *Stefana Janićijević, Vule Mizdraković and Maja Kljajić*

## 05 Abstract

06 Numerous methods exist aimed at examining patterns in structured and  
07 unstructured financial data. Applications of these methods include fraud detection,  
08 risk management, credit allocation, assessment of the risk of default, customer ana-  
09 lytics, trading prediction, and many others, creating a broad field of research named  
10 Financial data science. A problem within the field that remains significantly under-  
11 researched, yet very important, is that of differentiating between the three major  
12 types of business activities—merchandising, manufacturing, and service based on the  
13 structured data available in financial reports. A reliable method for fuzzy classification  
14 of business entities into the three categories of activities would assist numerous other  
15 methods in achieving their tasks. It can be argued that, due to the inherent idiosyn-  
16 crasies of the three types of business activities, methods for assessment of the risk of  
17 default, methods for credit allocation, and methods for fraud detection would all see  
18 an improved performance if reliable information on the percentage of entities' busi-  
19 ness activities allocated to the three major activities would be available. To this end, in  
20 this paper, we propose a clustering procedure that relies on Principal Component  
21 Analysis (PCA) for dimensionality reduction and feature selection. The procedure is  
22 presented using a large empirical data set comprising complete financial reports for  
23 various business entities operating in the Republic in Serbia, that pertain to the  
24 reporting period 2019.

AQ01

25 **Keywords:** data science, principal component analysis, random forest algorithm,  
26 financial data, financial reporting

## 27 1. Introduction

28 The established financial reporting system within an entity is the basic source of  
29 information on its financial position and results. The economic and financial globali-  
30 zation of the world market has emphasized the importance of high quality financial  
31 reporting. For the business decision-making process, financial and audit reports are  
32 the main source of information, as they contain information on financial position,  
33 business results, changes in equity, cash-flows and other reliable information [1].  
34 Development of the capital market and the increase in the number of interested  
35 parties (investors) created even higher demand of reliable, on time and fair financial  
36 statements as the main results of financial reporting. The regulation of the relationship

between the state and society, owners of capital and management, various stakeholders and society, and others; has been further improved by a quality financial reporting and audit process. However, in order to fulfill their main purpose for all interested parties, financial statements must provide information that is true, objective, comprehensible, comparable and uniform [2]. In the first place, financial statements have to be publicly available, which is usually regulated by law. For example, Law on Accounting of the Republic of Serbia prescribes that all business entities have to submit their financial reports to the competent institution which later publishes them on the official internet site [3]. Information contained in financial statements can be used for numerous purposes. For example, other business entities can use them in the process of making business, financial, investment and other decisions. Likewise, banks and financial institutions can use them in order to approve loans or assess investment risks related to the certain business entity. However, financial information contained in financial statements are not processed and represent a raw data that should be analyzed in order to assess the performance of a certain business entity. Aside Notes to financial statements, as one of the qualitative statements that business entities prepare and report, all other statements are quantitative in nature and offer hundreds of pieces of data. Therefore, it is of great importance to perform certain type of analysis on the collected data in order to gain a solid basis for business decision making process. Analysis of financial statements is one of the most common methods of assessing business performance. The main goal of conducting the analysis of financial statements is to obtain information on the performance of the observed company, i.e. liquidity, profitability and solvency. Measuring financial performance using compiled and disclosed financial statements is a quantitative analysis of the position of the observed company, including the way in which the company uses the capital invested in business. High quality analysis of the performance of the observed entity provides a comprehensive image of the business, including meeting the information needs of stakeholders. The authors [4] point out in their paper that the analysis of financial performance is crucial in determining the efficiency in terms of the use of available resources. Likewise, an entity owners will be able to assess management skills and decisions that have been made in previous, as well as in current reporting period, so that they could analyze entities strengths, weaknesses and therefore improve their overall performance [5–7].

Some pieces of data disclosed in financial statements have informational power to be used on their own, such as Total assets, Sales revenue, or Net result. However, informational power of data increases when they are put into relation with other pieces of data. Therefore, financial statements analysis using ratios has been one of the most commonly used methods of assessing business performance. Financial ratio is a relative magnitude of two (or more) selected numerical values taken from financial statements. For example, relation between Net result and Equity will provide information on how much dollars of profit an entity earns for each dollar invested in equity. Results of financial statements analysis can be used to compare performance of a certain entity over a period of time, or for comparison with other entities within the industry. However, since financial statements analysis takes time and there are numerous financial ratios that analysts could use (and the fact that most of these ratios are correlated), the number of ratios that are being calculated and assessed should be reduced so that an analyst could focus on several of them without losing data that could be relevant for the analysis [8]. One of the methods that can be used is Principal Component Analysis (PCA), which reduces number of observed variables for any further, regression, or any other type of analysis [9]. PCA analysis has found its

numerous purposes in different industries, for example, in image compressing [9–11], as well as in biometrics or “bioimaging” where physical characteristics of a person are used for its identification with application on communication devices and security systems.

The significance of PCA results is reflected in the fact that they can be used for more effective and efficient analysis of performance of certain entity, or for all business entities within a certain industry, or if analyzed financial data is related to whole economy, than results could be used for the analysis of all entities within it. The main advantages of PCA are precision of results; reduction of time needed for the analysis and evaluation of results; as well as reduction of related costs and efforts of the analyst.

With the development of technology, we have gained the ability to generate massive amounts of data. The use of correct methodologies for data analysis has become essential when dealing with complex financial challenges. In this paper, we discuss the theory underlying PCA. This type of analysis is one of the most used statistical tools in the field of financial data analysis. To ensure that the proper method is used for the analysis, theoretical knowledge and an comprehension of statistical methods are essential.

## 1.1 General postulates of PCA

PCA is primarily designed as a statistical technique that selectively reduces the dimensionality of data in complex data sets while preserving maximum variance. Since research in the financial sector involves both a large amount of data and a large number of variables simultaneously, it is difficult for us to perform analysis for this type of data.

Visualization techniques are only useful in two or three dimensional spaces, and single-variable analysis does not provide precise results due to overlapping variance. To achieve dimensionality reduction, it is necessary to generate principal components, i.e., a new set of variables containing a linear combination of the original variables. PCA can be used for a variety of tasks. A very small number of components are sufficient to cope with the variability of a data set. Since the number of components is reduced by using principal components, the complexity of the analysis itself is also reduced by avoiding analyzing a large number of output variables.

The standard PCA procedure takes as its starting point a data set in which  $m$  numerical variables are observed for each  $n$  individuals. These data are defined by the vectors  $x_1, \dots, x_m$  or  $n \times m$  of the data matrix  $X$ . The  $j^{th}$  column is the vector  $x_j$  resulting from the  $j^{th}$  variable. Linear combinations of columns for an  $X$  matrix with maximum variance are calculated as  $\sum_{j=1}^m c_j x_j = Xc$ . Here  $c$  stands for the vector of constants  $c_1, c_2, \dots, c_m$ . The variants of such a linear combination are obtained as  $var(Xc) = c'Mc$ . Here  $M$  stands for an exemplary covariance matrix. Finding a linear combination with maximum variance is the same as finding a  $m$  dimensional vector  $c$  that maximizes the quadratic form  $c'Mc$ . For this reason, it is necessary to enter another constraint, which is usually unit norm vectors. Such vectors require  $c'c = 1$ . This problem is the same as maximizing  $c'Mc - \lambda(c'c - 1)$ , where  $\lambda$  represents the Lagrange multiplier. Equating it to the zero vector gives the following equation:

$$Mc - \lambda c = 0 \Leftrightarrow Mc = \lambda c \quad (1)$$

This equation is valid even when the eigenvectors are multiplied by  $-1$ . Here,  $c$  is the eigenvector and  $\lambda$  is the corresponding eigenvalue for the covariance matrix  $M$ . We need the largest  $\lambda_1$ , the largest eigenvalue, and the corresponding eigenvector  $c_1$ . Eigenvalues are defined by the corresponding eigenvector  $c$  :  $var(Xc) = c'Ma = \lambda c'c = \lambda$ . The covariance matrix  $M$  is a symmetric  $m \times m$  matrix and has exactly  $m$  real eigenvalues.  $\lambda_k (k = 1, \dots, m)$  can be defined together with the corresponding eigenvectors to form a set of vectors that are orthonormal. An example of this is  $c'_m c_m = 1$  if  $m = m'$ . The eigenvectors of  $M$  are used to obtain up to  $m$  linear combinations of  $Xc_k = \sum_{j=1}^m c_{jk} x_j$  that maximize the variances. The fact that the covariance between the two linear combinations of  $Xc_k$  and  $Xc_{k'}$  is obtained from  $c'_k M c_{k'} = \lambda_k c'_k c_{k'} = 0$  if  $k' \neq k$ , leads to results of uncorrelatedness [12]. Linear combinations of  $Xc_k$  represent the principal component of a data set. There are several PCA terms used for specific values. Elements of linear combinations  $Xc_k$  are called principal component scores (PCA scores) and eigenvectors  $c_k$  are also called principal component loads (PCA loads). These contain a generic element  $x_{ij}^* = x_{ij} - \bar{x}_j$ , where  $x_j^*$  represents the observed value for variable  $j$ .

The  $n \times m$  matrix labeled  $X^*$  contains columns with centered variables  $x_j^*$ , resulting in the following equation:

$$(n - 1)M = X^{*'} X^* \quad (2)$$

## 1.2 Premises of PCA

For the final outcome of the PCA assessment to be successful and significant, numerous conditions must be met. Initially, it is crucial that the data entered are uninterrupted and that variables should be measured on an interval or ratio scale. This condition must be met because PCA tests important correlation patterns for these variables.

Another crucial requirement is that the relationships between the individual pairs of variables are linear. If there are nonlinear relationships between the individual pairs of variables, appropriate data transformation techniques, such as logarithmic transformations, should be considered. Presumptions for PCA are filling missing values with not null values, outliers handling, and normalization scaling. All outliers should be filtered out prior to analysis, as they can bias the results by affecting the magnitude of the correlation.

To obtain more accurate estimates for the correlation population parameters, a large sample size is required. The data sets must be linear in order to be formed. The basic principle of PCA is that high variance must be taken into account, while variables with lower variance can be considered noise and are not taken into account. All variables must be processed at the same level of measurement.

## 1.3 Features extraction in PCA

Eq. (2) associates the eigenvalue decomposition of the covariance matrix  $M$  and the singular value decomposition of the matrix  $X^*$  with the centered column data. For dimension  $n \times m$  and rank  $r$ , where it must be  $r \leq \min \{n, m\}$ , the matrix  $Y$  can be calculated as follows:

$$Y = ULA' \quad (3)$$

Where  $U$  and  $A$  represent the matrices  $n \times r$  and  $m \times r$  containing orthonormal columns  $U'U = I_r = A'A$ , where  $I_r$  represents the identity matrix  $r \times r$ .  $L$  is the  $r \times r$  diagonal matrix. The columns  $A$  are also called right singular vectors and represent eigenvectors for the  $m \times m$  matrix  $Y'Y$  associated with its non-zero eigenvalues. Columns  $U$  are also called left singular vectors and represent eigenvectors for the  $n \times n$  matrix  $YY'$  associated with its non-zero eigenvalues. Singular values of  $Y$  represent diagonal elements of the matrix, denoted by  $L$ . These elements are non-negative square roots for the non-zero eigenvalues of the two matrices  $Y'Y$  and  $YY'$ . We consider that the diagonal elements are sorted from the largest to the smallest element, which determines the order of the columns  $U$  and  $A$ , except for singular values that are equal [12]. This is true in all cases except when the singular values are equal. If we assume that  $Y = X^*$ , then the right singular vectors for the matrix  $X^*$  are vectors  $c_k$  of principal component loads. Because of the orthogonality of columns  $A$ , columns  $X^*A = ULA'A = UL$  are the principal components for  $X^*$ . The types of these principal components are obtained by squaring the singular values of  $X^*$  and dividing by  $n - 1$ . This results in the following equation:

$$(n - 1)M = X^{*'}X^* = (ULA')'(ULA') = ALU'ULA' = AL^2A' \quad (4)$$

Here  $L^2$  stands for a diagonal matrix with one square of the singular values. With this equation we get the eigenvalue decomposition for the matrix  $(n - 1)M$ . The singular value decomposition for the  $X^*$  matrix with the data centered in the column is equivalent to PCA. Taking the rank  $r$  in the matrix  $Y$ , which has the magnitude  $n \times m$ , the matrix  $Y_q$ , which has the same magnitude but the second rank  $q < R$  and whose elements reduce the sum of squared differences with the corresponding elements of  $Y$ , is obtained as:

$$Y_q = U_q L_q A_q' \quad (5)$$

Here  $L_q$  stands for the diagonal matrix of dimensions  $q \times q$ , which contains the first largest diagonal element  $q$  of  $L$  and  $U_q$ .  $A_q$  stands for the matrices  $n \times q$  and  $m \times q$  obtained by keeping the  $q$  columns in  $U$  and  $A$ . The number of rows  $n$  from the rank  $r$  of the matrix  $X^*$  defines the scatter plot from the number  $n$  of points in the  $r$  dimensional subspace  $\mathbb{R}^m$ , where the beginning of the gravity center for the scatter plot is located. It follows that the best approximation of the  $n$  points in this scatterplot in the  $q$  dimensional subspace, obtained by using  $X_q^*$  rows, is given by this equation. That means that the sum of the squared distances between the given points in each scatterplot is minimal, as in Pearson's original approach [13]. The  $q$  axis system defines the main subspace. It can be concluded that PCA is a dimensionality reduction method where a set of  $m$  original variables can be replaced by a given set of  $q$  variables. In the case of  $q = 2$  or  $Q = 3$ , it is possible to make a graphical approximation for  $n$  points in the scatter plot, and it is very often used to visualize the whole data set. A very important aspect is that the results are incremental in their dimensions.

The variability associated with the set of retained principal components can be used to ensure the quality of any  $q$  dimensional approximation. The trace, i.e. the sum of the diagonal elements, of the covariance matrix  $M$  is equal to the sum of the variances of the  $m$  variables. It is possible to achieve this with the help of matrix theory results. It is easy to prove that this number is also the sum of the variances of all  $m$  principal components. Consequently, the proportion of the overall variation

01 accounted for by a given principal component is a standard measurement of its quality  
02 and it's equal to:

$$\pi_j = \frac{\lambda_j}{\sum_{j=1}^m \lambda_j} = \frac{\lambda_j}{\text{tr}(M)'} \quad (6)$$

03 The trace of  $M$  is labeled  $\text{tr}(M)$ . Due to the incremental behavior of principal  
04 components, we can speak of a proportion of the total variance explained by a set  $M$  of  
05 principal components, which is usually expressed as a percentage of the total variance  
06 and is accounted for:

$$\sum_{j \in M} \pi_j \times 100\% \quad (7)$$

07 It is a common approach to use a pre-specified percentage of the total variance to  
08 determine how many principal components to keep, but graphical constraints often  
09 lead to keeping only the first two or three principal components. The percentage of  
10 total variance is a basic tool for measuring the quality of these low-dimensional  
11 graphical representations of the data set.

12 The biggest problem is the number of components needed to obtain a sufficient  
13 number of variances while achieving a reduction in dimensionality. There are several  
14 ways to determine the components, and one of them is to set a threshold.

15 The next very popular approach is the "Scree Plot" [14], where the components are  
16 arranged on the  $X$ -axis from largest to smallest with respect to their eigenvalues. In  
17 this way, we can see a very large difference between important and less important  
18 components. The only drawback to this approach is that it is subjective in determining  
19 the correct number of components.

20 The most popular method is parallel analysis [15], where PCA is performed with as  
21 many variables as the original data set includes. The average eigenvalues between the  
22 original data set and the simulated data set are measured. Any values from the original  
23 data that are lower than the data in the simulated set are discarded.

## 24 1.4 Sparse PCA

25 PCA has many advantages. In terms of maximizing variance in  $Q$  dimensions, PCA  
26 provides the best possible representation of a  $m$  dimensional data set in  $q$  dimensions  
27  $q < m$ . However, the new variables it defines are often linear functions of all the  $m$   
28 original variables, which is a downside. Multiple variables with not so simple  
29 coefficients are common for larger  $m$ , making the components difficult to read. A  
30 number of PCA adjustments have been proposed to facilitate interpretation of the  $q$   
31 dimensions while limiting the loss of variance that results from not using the principal  
32 components themselves. There is a compromise between interpretability and  
33 variance. Two types of adjustments are briefly outlined below.

34 Factor analysis is a method that is often combined with PCA and it inspires the  
35 concept of rotating principal components [16]. Assume that  $A_q$  is the  $m \times q$  matrix  
36 whose columns are the loadings of the first  $q$  of the principal components. Then  $XA_q$  is  
37 the  $n \times qnq$  matrix whose columns are the scores of the first  $q$  of the principal  
38 components for the  $n$  observations. Let us assume that  $T$  is an orthogonal  $q \times q$  matrix.  
39 Multiplying  $A_q$  by  $T$  causes orthogonal rotation of the axes within the space spanned  
40 by the first  $q$  of principal components, resulting in  $B_q = A_q T$ , a  $m \times q$  matrix whose

columns are the charges of the  $q$  rotated principal components.  $XB_q$  is an  $n \times q$  matrix containing the associated values of the rotated principal components. Any orthogonal matrix  $T$  can be used to rotate the components, but it is preferable to make the rotated components easy to understand. For this reason,  $T$  is chosen to maximize simplicity. A variety of such criteria have been proposed, some of which involve non-orthogonal rotation. The criterion where an orthogonal matrix  $T$  is chosen for maximizing

$$Q = \sum_{k=1}^q \left[ \sum_{j=1}^m b_{jk}^4 - \left( \frac{1}{m} \right) \left( \sum_{j=1}^m b_{jk}^2 \right)^2 \right], \text{ where } b_{jk} \text{ is the } (j, k)^{th} \text{ member of } B_q, \text{ is}$$

probably the most commonly used. No variance is lost when considering the rotated  $q$  dimensional space, since the sum of the variances of the  $q$  rotated components is the same as the sum of the variances of the unrotated components. Successive maximization of the non-rotated principal components is lost, which means that the sum of the variances of the  $q$  rotated components is the same as the sum of the variances of the non-rotated components. A disadvantage of rotation is the necessary choice between different rotation criteria, although this choice often makes less difference than the choice of the number of components to rotate. If  $q$  is increased by 1, the rotated components may look substantially different. That is because this does not happen in principal components with defined non-rotated nature.

Another method of simplifying the principal components is to limit the charges of the new variables. This is called adding a constraint. There are several variants of this strategy, one of which uses LASSO linear regression [17], that represents least absolute shrinkage and selection operator. In this approach, SCoTLASS components are discovered, solving the same optimization problem as PCA, but with the additional constraint  $\sum_{j=1}^m |c_{jk}| \leq \tau$ , where tuning parameter is  $\tau$ . The constraint has no effect for  $\tau > \sqrt{m}$ , and principal components are generated; however, more charges are pushed to zero at a lower value, which simplifies the interpretation. These simplified components must have less variation than the corresponding number of principal components, and multiple values of  $\tau$  are often examined to find a reasonable compromise between added simplicity and loss of variance. One distinction between rotation and constraint techniques is that the second has the advantage that some loadings in linear functions are set exactly to zero for interpretation, whereas this is usually not the case with rotation. Sparse variants of PCA are type of adjustments in which many coefficients are zero, and numerous studies of such principal components have been conducted in recent years. Hastie et al. [18] provides a good overview of this work.

## 1.5 Robust PCA

PCA is inherently sensitive to the occurrence of outliers and thus to large errors in data sets [19]. As a result, efforts have been made to define robust variants of PCA, and the terminology RPCA has been used to refer to several approaches to this problem. Huber's early work focused on robust alternatives to covariance or correlation matrices and how they could be used to generate robust principal components [20]. The demand for methods to process very large data sets sparked renewed interest in robust PCA variants. This led to PCA research lines, especially in areas such as machine learning, image processing, web data analysis, and many others.

Wright et al. [21] defined RPCA as the sum of two  $n \times m$  components, a low-rank component  $L$  and a sparse component  $S$  in an  $n \times m$  data matrix  $X$ . Identifying the matrix components of  $X = L + S$  that minimize a linear combination of two separate component norms was defined as a convex optimization task and calculated as:

$$\min_{L,S} \|L\|_* + \lambda \|S\|_1 \quad (8)$$

where  $\|L\|_* = \sum_r \sigma_r(L)$  is the nuclear norm of  $L$ , and  $\lambda \|S\|_1 = \sum_i \sum_j |s_{ij}|$  is the  $l_1$  norm of matrix  $S$ .

## 2. Related work

PCA was first introduced into mechanics by [22], as an analogue of the axis theorem. It was later named “PCA” by [23]. The range of applications in finance and economics is extensive. Take as an example [24], who used PCA to document three factor structures. Stock and Watson [25] used PCA to monitor economic development and activity, as well as the inflation index. Egloff et al. [26] used PCA as a way to analyze the dimensions of inconsistent dynamics. Volatility is a statistical measure that can be used to determine these inconsistencies using a two-factor volatility model. This includes long-term and short-term fluctuations in the volatility structure. Baker and Wurgler [27] used PCA to measure investors sentiment, i.e., their positive or negative view. This was done according to the principle of the number of sentiment proxies before Baker, [28] created the policy uncertainty index. This index represents potential risks in the near future.

The most important item in the construction of PCA is the estimation of the eigenvalues of the covariance matrix sample. Anderson and Weeks [29] and Anderson [30] showed that sample eigenvalues were consistent when dealing with asymptomatic sentiment proxy results. Waternaux [31] proved that similar results are obtained with simple eigenvalues as long as there is a fourth moment in the data. In addition to the discussions in the [32] book, [33] was able to establish the asymptotic distribution of eigenvectors using generalized assumptions.

However, this PCA approach to eigenvalues has some downsides. The first problem is certainly dimensionality, which can be noticed when the cross sectional dimension grows simultaneously with the sample in the same period. Then inconsistencies occur. Another problem arises from linear data types that do not include nonlinear patterns. A third problem [34] arises from the dependence of the asymptotic theory on fixed assumptions for the analysis. For these reasons, we have a problem when we use PCA for reimbursement data. Most of the time, we need years of data to make an assumption, which in turn leads to other problems, such as permanence and consistency of non-fixed parameters. This type of data has backlogs and volatility times often vary.

These problems stimulate the improvement in this field and motivate the development of tools for PCA methods. The approach to the problem, where the number of occurrences grows in fixed time periods, touches all the listed downsides. Theoretically, it is known that as the frequency of the sample increases, the estimated variance and covariance increase. This is true until the microstructure of the market begins to take effect. Incidentally, this is not a serious problem if we choose a sampling frequency of minutes, which we use as opposed to the below one second time interval most often used for liquid stocks. A high frequency asymptotic analysis with the cross-sectional dimension is expected as the time interval increases sharply. This high frequency asymptotic framework allows us to perform non-parametric analysis as well as independent, non-static and analysis without underlying parameters as is the case with low frequency processes.

Asymptotic theory is very common in many contexts. Jacod et al. [13] and Jacod and Podolskij [35] also dealt with one problem that we deal with in this paper, where the cross sectional dimensions are invariant and the process is continuous. Mykland and Zhang [36] designed an alternative theory to the one put forward by [37], that discuss inference for volatility function dependence. It is based on the aggregation of local estimates and uses a finite number of blocks. Saha et al. [38] considered the expected values of the integrated covariance matrix under conditions where there is an error measure and the matrix is large containing high frequency data. Tao et al. [39] addressed work on the convergence rate. Jacod and Rosenbaum [40] analyzed estimators, composed of aggregating functions of estimates. They did so using integrated quarticity estimation. Heinrich and Podolskij [41] discussed empirical covariate matrices of Brownian integrals. Here is discussed the measurement of the leverage effect and its evaluation by the integrated correlation method [42].

PCA analysis can be used in analysis of financial data for different purposes. For example [43] used it to identify the type of impact on grouped impact factors, such as assessing the quality of accounting information and facilitating the process of financial analysis conducted by different users. On the other hand, [44] used PCA to assess the impact of the evolution of Finnish standards on IFRS (International Financial Reporting Standards). Finally [45] used PCA analysis to determine the macroeconomic impact on the profitability of Romanian listed companies, using data from 1997 to 2007, and identified following indicators: liquidity, solvency, and firm's dimension.

When it comes to the use of PCA analysis in financial statements analysis, four papers that focus on Romanian listed companies will be reviewed first. All papers emphasize the importance of using PCA analysis in the analysis of key financial ratios. In the first paper author [46] analyzed the data of 16 initial variables which he grouped into 3 new variables (general efficiency indicator, indicator in correlation with historical debts of companies and development indicator (given long-term debt and deferred income). Those three variables were able to explain 96.72% of initial variability. In the second paper, [47] analyzed data for 2010 including initially seven indicators of standard financial analysis and they reduced them to only two (which explain 94% of initial variability). In third paper, [48] used data from the stock exchange in the period 2006–2011 to identify the main components of financial statements which explain 79.08% of initial variability. The same group of indicators has been used by [43] on research sample that consisted of 111 companies from Madrid stock exchange and 32 companies from Eurostoxx50 for reporting periods 2005–2007. Research results showed that those six indicators explained 87% of total variance, with the first two indicators at app 44% of total variance.

### **3. Case study—PCA and cluster analysis in financial accounting data**

#### **3.1 Research methodology**

In order to provide an answer on defined research question, 3,013 medium and large business entities were selected by random and used as a research sample. Financial statements for 2019 reporting period have been downloaded manually from the official website of the Business Registers Agency (BRA). BRA is a state administrative

body that collects financial statements and corresponding audit reports of business entities that operate within the territory of the Republic of Serbia. Information published by BRA is used for financial analysis of business entities and as a basis of decision-making process. Afterwards, data from the pdf files containing financial statements have been copied and recorded in pre-set up tables in Excel files. Namely, medium and large business entities in the Republic of Serbia have an obligation to prepare and disclose full set of financial statements, consisting of balance sheet, income statement, cash-flow statement, statement of changes in equity and notes to financial statements. Since all previously mentioned statement, except notes to financial statements, are quantitative in nature, they were used for this research. Values originally disclosed in RSD, as the reporting currency, were converted into euros by using the average exchange rate of euros on the balance sheet date (31st December). Values of each financial statement line is presented in thousands, and therefore they are presented as such in this research [49].

Financial statement item lines in official financial statements are marked by corresponding automatic data processing number (in Serbian: Automatska obrada podataka—AOP), that belongs to the national nomenclature system. These markings are used in order to perform control of mathematical calculations before each financial statement is accepted for publishing by BRA. They also serve as an instrument of connecting data and information regarding the same financial statement item presented in financial statements. Balance sheet items cover automatic data processing numbers from 0001 to 0465; income statement from 1001 to 1071; statement of cash-flows from 3001 to 3047; and statement of changes in equity from 4001 to 4252.

**Table 1** shows the formulas used for the calculation of the selected financial indicators that will be used in this research. Having in mind that these variables will be used in order to differentiate business entities to three major types of business activities, these variables have been selected by a common sense.

| Variables                                                         | Derived from            |
|-------------------------------------------------------------------|-------------------------|
| Fixed assets in total assets                                      | AOP2/AOP71              |
| Percent sales of merchandise in total operating revenue           | AOP1002/AOP1001         |
| Percent sales of products and services in total operating revenue | AOP1009/AOP1018         |
| Percent cost of merchandise sold in total operating expenses      | AOP1019/AOP1018         |
| Percent cost of material in total operating expenses              | AOP1023/AOP1018         |
| Percent fuel and energy cost in total operating expenses          | AOP1024/AOP1018         |
| Percent wage cost in total operating expenses                     | AOP1025/AOP1018         |
| Percent productive service cost in total operating expenses       | AOP1026/AOP1018         |
| Percent depreciation cost in total operating expenses             | AOP1027/AOP1018         |
| Percent raw material in total assets                              | AOP45/AOP71             |
| Percent WIP in total assets                                       | AOP46/AOP71             |
| Percent finished products in total assets                         | AOP47/AOP71             |
| Percent WIP and finished products in total assets                 | $(AOP46 + AOP47)/AOP71$ |
| Percent merchandise in total assets                               | AOP48/AOP71             |

**Table 1.**  
*Calculation of selected financial indicators.*

## 01 **3.2 Algorithm**

02 Data preparation is a key process in data analysis. The basic preparation and  
03 cleaning procedures are:

- 04 • Preparing a copy of the table
- 05 • Adding new attributes
- 06 • Conversion of column types
- 07 • General data cleaning and adjustment

08 Specifically, the cleaning includes the following items:

- 09 • Editing date variables—the most common formatting problems
- 10 • Recoding of zeros/missing values
- 11 • Decoding categorical variables using labels and hot encoding
- 12 • Arranging outliers
- 13 • Application of normalization/standardization/ log transformation
- 14 • Calculating descriptive statistics—mean, median, mode, standard deviation,  
15 variance, rank, etc.
- 16 • Calculating inferential statistics - distributions, t-value, p-value, frequencies,  
17 cross-tabulations, correlation, covariance, etc.

18 More advanced techniques include:

- 19 • Coding:

20 Categorical variables are labeled as character variables and must be converted to a  
21 factor type for modeling purposes. Queues perform this task.

- 22 • Outliers:

23 For numeric variables, we can identify deviations numerically by the value of the bias.

- 24 • Normalization/logarithmic transformation:

25 One of the techniques to normalize the biased distribution is logarithmic transfor-  
26 mation. First, a new variable is created, while later the value of the bias of this new  
27 variable is calculated and printed.

- 28 • Standardization:

29 One of the standardization techniques is that all characteristics are centered  
30 around zero and have approximately the variance of one unit. Scaling is used so that

the variable is converted. The result is that these variables are standardized with a mean of zero.

As part of the preparation for PCA, firstly missing values from the dataset were filled with zeros. After that, the data was scaled by using a standard scaler, which standardizes features by removing the mean and scaling to unit variance. The preprocessed dataset, was then used for:

- PCA
- Sparse PCA
- Robust PCA

All three of the PCA methods were instantiated with the number of components set to 7. After PCA, the now transformed data went through several clustering methods for the purpose of comparing results. The clustering methods that were used for each PCA are:

- K-means clustering
- Agglomerative clustering
- BIRCH clustering
- Gaussian Mixture
- Spectral clustering

Furthermore, each of the clustering methods were executed with just the preprocessed data, without PCA, also for the purpose of comparing results.

---

**Algorithm 1:** Principal Component Analysis.

---

**procedure:**

    Data preparation:  $X \leftarrow X^*$   
    Compute dot product matrix:  $X^{*'}X^* \leftarrow (n-1)M$   
    Eigenanalysis:  $AL^2A' \leftarrow X^{*'}X^*$   
    Compute eigenvectors:  $U \leftarrow X^*AL$   
    Keep first 7 components:  $U_7 \leftarrow [u_1 \cdots u_7]$   
    Compute 7 features:  $Y \leftarrow U_d'X$   
end procedure.

---

## 4. Results and discussion

### 4.1 Comparative results—total variance explained

This chapter discusses the outcomes of PCA and cluster analysis. The initial variables that load on the principal components are studied. Correlations or covariances

between the original variables and the principal components correlate with the loadings. The variable loadings are contained in a loading matrix, which is created by multiplying the eigenvector matrix by a diagonal matrix containing the square root of each eigenvalue. The entries are determined by the component extraction method used. Non-standardized loadings show the covariance between mean-centered variables and standardized component values, regardless of whether the extraction is based on the singular value decomposition of the matrix or the eigenvalue decomposition of the covariance matrix.

The eigenvalue decomposition of the correlation matrix results in the standardized charges. The correlations between the original variables and the component scores are represented by these loadings. Because they always vary between  $-1$  and  $1$  and are independent of the scale used, standardized charges are easy to read. In most cases, a threshold is set and only variables with loadings above this threshold are examined.

The total variance presents sum of variances of principal components. The ratio between the variance of principal component and the total variance is the fraction of variance explained by a principal component.

**Figure 1** shows total variance explained by using three methods of PCA. The steepest increase belongs to the PCA line, which cumulative explained variance is app. 87%. This line is almost parallel to the line from Sparse PCA which cumulative explained variance is 83%. However, when it comes to Robust PCA line it has been noticed that cumulative explained variance is only app. 26% and the increase of values is minimal.

PCA: The highest fraction of explained variance among these variables is 32%, and the lowest one is 5%. Cumulative explained variance is 86% (see **Table 2**).

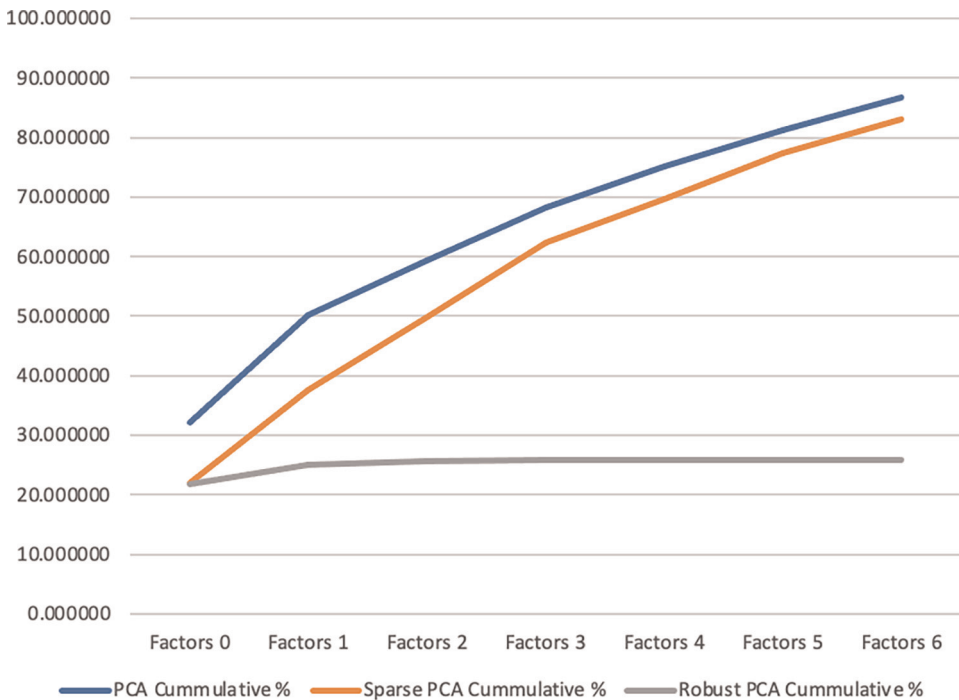


**Figure 1.**  
Total variance explained.

AQ07

| Factors  | Total    | % of variance | Cumulative % |
|----------|----------|---------------|--------------|
| Factor 0 | 4.491515 | 32.082248     | 32.082248    |
| Factor 1 | 2.540717 | 18.147978     | 50.230226    |
| Factor 2 | 1.269778 | 9.069843      | 59.300069    |
| Factor 3 | 1.243867 | 8.884762      | 68.184831    |
| Factor 4 | 0.961330 | 6.866641      | 75.051473    |
| Factor 5 | 0.867145 | 6.193891      | 81.245364    |
| Factor 6 | 0.760536 | 5.432398      | 86.677761    |

**Table 2.**  
*PCA total variance explained.*

Sparse PCA: The highest fraction of explained variance among these variables is 21%, and the lowest one is 5%. For instance, variables together explain 83% of the total variance (see **Table 3**).

Robust PCA: The highest fraction of explained variance among these variables is 21%, and the lowest one is 0%. For instance, variables together explain 25% of the total variance (see **Table 4**).

PCA is the best approach for this kind of data, regarding number of features.

| Factors  | Total    | % of variance | Cumulative % |
|----------|----------|---------------|--------------|
| Factor 0 | 3.078591 | 21.989939     | 21.989939    |
| Factor 1 | 2.186255 | 15.616108     | 37.606047    |
| Factor 2 | 1.698036 | 12.128828     | 49.734874    |
| Factor 3 | 1.757003 | 12.550022     | 62.284897    |
| Factor 4 | 1.047037 | 7.478832      | 69.763729    |
| Factor 5 | 1.062211 | 7.587224      | 77.350953    |
| Factor 6 | 0.809469 | 5.781923      | 83.132875    |

**Table 3.**  
*Sparse PCA total variance explained.*

| Factors  | Total    | % of variance | Cumulative % |
|----------|----------|---------------|--------------|
| Factor 0 | 3.035926 | 21.685184     | 21.685184    |
| Factor 1 | 0.454951 | 3.249650      | 24.934834    |
| Factor 2 | 0.108168 | 0.772628      | 25.707462    |
| Factor 3 | 0.020284 | 0.144884      | 25.852346    |
| Factor 4 | 0.006630 | 0.047355      | 25.899701    |
| Factor 5 | 0.000018 | 0.000128      | 25.899829    |
| Factor 6 | 0.000000 | 0.000000      | 25.899829    |

**Table 4.**  
*Robust PCA total variance explained.*

4.2 Communalities

The amount of variance in each variable considered is represented by the communalities. The variance in each variable explained by all components or factors is estimated using the initial communalities.

The percent fuel and energy cost in total operating expenses is given here with 88% variance. The percent productive service cost in total operating expenses is given here with 75% variance. The percent finished products in total assets here is 75% of the estimated variance (see **Table 5**).

The percent fuel and energy cost in total operating expenses here is 91% variance. The percent finished products in total assets here is 80% of the estimated variance. The percent productive service cost in total operating expenses here is 74% variance (see **Table 6**).

| Columns                                                           | Communality |
|-------------------------------------------------------------------|-------------|
| Percent merchandise in total assets                               | 0.159427    |
| Percent sales of merchandise in total operating revenue           | 0.222216    |
| Percent cost of merchandise sold in total operating expenses      | 0.224299    |
| Percent sales of products and services in total operating revenue | 0.236318    |
| Fixed assets in total assets                                      | 0.347415    |
| Percent cost of material in total operating expenses              | 0.411423    |
| Percent raw material in total assets                              | 0.426201    |
| Percent WIP and finished products in total assets                 | 0.449704    |
| Percent depreciation cost in total operating expenses             | 0.683213    |
| Percent wage cost in total operating expenses                     | 0.729997    |
| Percent WIP in total assets                                       | 0.731771    |
| Percent finished products in total assets                         | 0.745349    |
| Percent productive service cost in total operating expenses       | 0.752027    |
| Percent fuel and energy cost in total operating expenses          | 0.880639    |

**Table 5.**  
*PCA communalities.*

| Columns                                                           | Communality |
|-------------------------------------------------------------------|-------------|
| Percent merchandise in total assets                               | 0.191833    |
| Percent sales of products and services in total operating revenue | 0.227810    |
| Percent sales of merchandise in total operating revenue           | 0.260545    |
| Percent cost of merchandise sold in total operating expenses      | 0.263888    |
| Fixed assets in total assets                                      | 0.354743    |
| Percent cost of material in total operating expenses              | 0.407825    |
| Percent raw material in total assets                              | 0.417451    |
| Percent WIP and finished products in total assets                 | 0.451553    |
| Percent depreciation cost in total operating expenses             | 0.555661    |

| Columns                                                     | Communality |
|-------------------------------------------------------------|-------------|
| Percent wage cost in total operating expenses               | 0.695148    |
| Percent WIP in total assets                                 | 0.719447    |
| Percent productive service cost in total operating expenses | 0.742714    |
| Percent finished products in total assets                   | 0.800108    |
| Percent fuel and energy cost in total operating expenses    | 0.911274    |

**Table 6.**  
*Sparse PCA communalities.*

| Columns                                                           | Communality |
|-------------------------------------------------------------------|-------------|
| Percent WIP in total assets                                       | 0.200472    |
| Percent merchandise in total assets                               | 0.317793    |
| Percent finished products in total assets                         | 0.333984    |
| Percent depreciation cost in total operating expenses             | 0.345393    |
| Percent fuel and energy cost in total operating expenses          | 0.349862    |
| Percent sales of products and services in total operating revenue | 0.365996    |
| Percent raw material in total assets                              | 0.433737    |
| Percent WIP and finished products in total assets                 | 0.444081    |
| Percent cost of material in total operating expenses              | 0.519423    |
| Fixed assets in total assets                                      | 0.651365    |
| Percent productive service cost in total operating expenses       | 0.680299    |
| Percent cost of merchandise sold in total operating expenses      | 0.745842    |
| Percent sales of merchandise in total operating revenue           | 0.789024    |
| Percent wage cost in total operating expenses                     | 0.822730    |

**Table 7.**  
*Robust PCA communalities.*

The percent wage cost in total operating expenses here is 82% variance. The percent sales of merchandise in total operating revenue here is 79% of the estimated variance. The percent cost of merchandise sold in total operating expenses here is 74% variance (see **Table 7**).

**Figure 2** presents the amount of variance for each considered variable represented by the communalities. From the aspect of PCA and Sparse PCA it can be noticed that variable Percent fuel and energy cost in total operating expenses and variable Percent finished products in total assets have significant estimated variance. When it comes to Robust PCA, variance of 82% refers to the variable Percent wage cost in total operating expenses. From the economic point of view first two variables could be used to distinguish type of three major business activities. Mainly, the amount of fuel and energy cost will differ between business activities. It is expected that production entities will have higher values of fuel and energy costs because plant, machinery and equipment will require energy to operate. Also, merchandise entities will probably have higher values of fuel and energy costs compared to other services having in mind

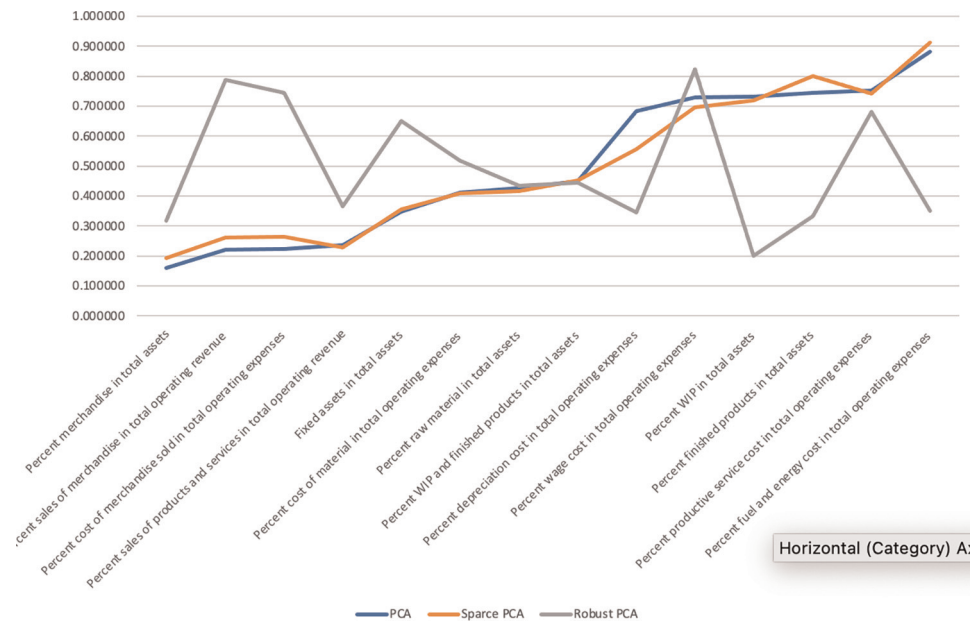


**Figure 2.**  
*Amount of variance represented by the communalities.*

fuel spent for transportation of merchandise and energy needed for operation of their facilities. Second variable Percent finished products in total assets is also expected to be used for differentiation since only production entities will have this balance sheet line in their financial statements. Main surprise might be third variable Percent wage cost in total operating expenses, since most entities have very similar share of total wage costs in total operating expenses. Namely, although official state records showed that average wages differ across industries, management of companies usually plan operating expenses and their structure.

### 4.3 Clustering

The best approach for the PCA/Clustering combination regarding high level of Silhouette Index and Cluster Sizes are: K-means/Robust PCA and Spectral/Robust PCA. The Davies Bouldin Index implies that a smaller value gives better clustering. This produces the idea that no cluster has to be similar to another, and that object inside clusters are very uniformly distributed (see **Table 8**).

| Clustering/PCA method | Cluster sizes    | Silhouette index    | Davies bouldin index |
|-----------------------|------------------|---------------------|----------------------|
| K-means/No PCA        | (1345, 932, 733) | 0.30208710358306756 | 1.5444364169813884   |
| K-means/PCA           | (1353, 934, 723) | 0.3637346841903855  | 1.3405097768944103   |
| K-means/Sparse PCA    | (1356, 939, 715) | 0.36307616530243575 | 1.3418713066940657   |
| K-means/Robust PCA    | (1209, 944, 857) | 0.5193200382282146  | 0.7834359567299072   |
| Agglomerative/no PCA  | (1151, 935, 924) | 0.27839422485839554 | 1.7150687814273013   |
| Agglomerative/ PCA    | (1225, 962, 823) | 0.31642069773357084 | 1.4995739243069988   |

| Clustering/PCA method       | Cluster sizes     | Silhouette index    | Davies bouldin index |
|-----------------------------|-------------------|---------------------|----------------------|
| Agglomerative/sparse PCA    | (1888, 893, 229)  | 0.31642069773357084 | 1.4995739243069988   |
| Agglomerative/robust PCA    | (1311, 878, 821)  | 0.4593880561940543  | 0.9274868826361716   |
| Birch/no PCA                | (1151, 935, 924)  | 0.27839422485839554 | 1.7150687814273013   |
| Birch/ PCA                  | (1225, 962, 823)  | 0.31642069773357084 | 1.4995739243069988   |
| Birch/sparse PCA            | (1225, 962, 823)  | 0.31642069773357084 | 1.4995739243069988   |
| Birch/robust PCA            | (1317, 867, 826)  | 0.45631070311567473 | 0.9348852316431389   |
| Gaussian mixture/no PCA     | (1336, 992, 682)  | 0.17495781525891207 | 2.1078218204567496   |
| Gaussian mixture/ PCA       | (1161, 1155, 694) | 0.2539355374019169  | 1.6227017939395394   |
| Gaussian mixture/sparse PCA | (1161, 1155, 694) | 0.2539355374019169  | 1.6227017939395394   |
| Gaussian mixture/robust PCA | (1467,784, 759)   | 0.28455634384131373 | 1.1919962215015028   |
| Spectral/no PCA             | (2994, 8, 8)      | 0.460433642421337   | 0.9718901349784725   |
| Spectral/PCA                | (3001, 7, 2)      | 0.5399338738262545  | 0.6856986473871954   |
| Spectral/sparse PCA         | (3001, 7, 2)      | 0.5399338738262545  | 0.6856986473871954   |
| Spectral/robust PCA         | (1346, 920, 744)  | 0.5146721760042233  | 0.7917964357887189   |

**Table 8.**  
PCA with different clustering methods.

## 01 5. Conclusion

02 This chapter was focused on the use of Principle component analysis in financial  
03 data science. Research has been conducted that included 3013 medium and large  
04 business entities and their financial statements from 2019 reporting period. PCA has  
05 been used in order to differentiate between the three major types of business activities  
06 - merchandising, manufacturing, and service. Therefore, 14 financial ratios have been  
07 selected by common sense and further analyzed according to their significance in  
08 dimensionality reduction. Results of clustering gave 7 new variables: 1. cost of mer-  
09 chandise sold in total operating expenses, and cost of material in total operating  
10 expenses; 2. fuel and energy cost in total operating expenses, and sales of product and  
11 services in total operating revenue; 3. wage costs in total operating expenses, and sales  
12 on merchandise in total operating revenue; 4. productive service cost in total operat-  
13 ing expanses, and fixed assets in total assets; 5. depreciation cost in total operating  
14 expenses, and merchandise in total assets; 6. raw material in total assets, and WIP and  
15 finished products in total assets; 7. finished products in total assets, and WIP in total  
16 assets. These groups of variables were able to explain 86.7% of initial variability.  
17 Compared to the results of authors previously mentioned in literature review, it can  
18 be concluded that percentage is within the range of reached results. When it comes to  
19 initial communalities which estimated the variance in each variable, three financial  
20 ratios that had the highest percentage were: fuel and energy cost in total operating  
21 expenses (original PCA—88%, sparse PCA—91%); productive service cost in total  
22 operating expenses (original PCA—75%, sparse PCA—74%); and finished products in  
23 total assets (original PCA 75%, sparse PCA—80%). Although these ratios showed the  
24 best results, it has to be mentioned that there is a correlation between all of financial  
25 ratios used in analysis and therefore results would be different when ratios are used.

**Acknowledgements**

We would like to express our gratitude to Prof. Nemanja Stanišić, Ph.D. from the Singidunum University for supporting this research through valuable suggestions, and assignment of a research database.

**Author contribution**

Author Stefana Janićijević contributed to the design and implementation of the research and analysis of the results. Authors Vule Mizdraković and Maja Kljajić prepared sections of the chapter that refers to the financial data science and financial reporting: introduction, related work, research methodology and analysis of discussion and result. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

**Conflicts of interest**

Authors declare no conflict of interest.

**Nomenclature**

|           |                                                        |
|-----------|--------------------------------------------------------|
| $m$       | number of numerical variables                          |
| $n$       | individuals                                            |
| $x$       | vector                                                 |
| $X$       | data matrix                                            |
| $j$       | number of columns                                      |
| $Xc$      | linear combinations                                    |
| $c$       | vector of constants                                    |
| $M$       | covariance matrix                                      |
| $\lambda$ | lagrange multiplier                                    |
| $U$       | matrix with orthonormal columns—eigenvectors           |
| $A$       | matrix with singular vectors                           |
| $L$       | diagonal elements of the matrix                        |
| $L^2$     | diagonal matrix with one square of the singular values |
| $r$       | rank of the matrix                                     |
| $q$       | dimensional subspace                                   |
| $tr$      | trace of matrix                                        |

A. Appendix

See Tables 9–11.

AQ02

| Columns/factors                                                   | Factor 0  | Factor 1  | Factor 2  | Factor 3  | Factor 4  | Factor 5  | Factor 6  |
|-------------------------------------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| Fixed assets in total assets                                      | 0.178413  | −0.326641 | 0.415221  | −0.102354 | −0.025277 | −0.072754 | −0.141675 |
| Percent sales of merchandise in total operating revenue           | −0.436002 | 0.152729  | 0.080029  | −0.025182 | 0.028728  | 0.016652  | −0.025519 |
| Percent sales of products and services in total operating revenue | 0.398117  | 0.022315  | −0.270570 | −0.046006 | 0.031509  | 0.012930  | −0.028959 |
| Percent cost of merchandise sold in total operating expenses      | −0.432559 | 0.162995  | 0.080296  | −0.035352 | 0.022542  | 0.026778  | −0.041260 |
| Percent cost of material in total operating expenses              | 0.269688  | 0.303749  | −0.078000 | −0.386050 | −0.243323 | −0.066637 | −0.166323 |
| Percent fuel and energy cost in total operating expenses          | 0.150958  | −0.217356 | 0.283494  | −0.008988 | 0.096350  | 0.822637  | −0.210100 |
| Percent wage cost in total operating expenses                     | 0.210317  | −0.224488 | −0.145697 | 0.058149  | 0.719422  | −0.304476 | −0.022063 |
| Percent productive service cost in total operating expenses       | 0.137457  | −0.048360 | −0.397081 | 0.585499  | −0.374661 | 0.172006  | 0.245674  |
| Percent depreciation cost in total operating expenses             | 0.095815  | −0.269993 | 0.484683  | 0.000289  | −0.400868 | −0.359862 | 0.275725  |
| Percent raw material in total assets                              | 0.190490  | 0.245296  | −0.165694 | −0.526444 | −0.137683 | 0.071830  | 0.032101  |
| Percent WIP in total assets                                       | 0.158335  | 0.359273  | 0.200936  | 0.383609  | −0.059087 | −0.175372 | −0.596528 |
| Percent finished products in total assets                         | 0.174390  | 0.375283  | 0.278149  | 0.015943  | 0.252221  | 0.151402  | 0.640266  |
| Percent WIP and finished products in total assets                 | 0.214830  | 0.474174  | 0.309621  | 0.255892  | 0.126328  | −0.013709 | 0.034896  |
| Percent merchandise in total assets                               | −0.355975 | 0.151166  | −0.014079 | −0.013559 | 0.088827  | 0.039431  | 0.005508  |

Table 9.  
PCA component matrix.

| Columns/factors                                                   | Factor 0  | Factor 1  | Factor 2  | Factor 3  | Factor 4  | Factor 5  | Factor 6  |
|-------------------------------------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| Fixed assets in total assets                                      | 0.000000  | -0.020910 | -0.504987 | 0.000000  | -0.254639 | -0.185601 | -0.002365 |
| Percent sales of merchandise in total operating revenue           | 0.435472  | 0.000000  | 0.246576  | -0.085566 | 0.000000  | 0.052803  | 0.000000  |
| Percent sales of products and services in total operating revenue | -0.433624 | 0.008108  | 0.000000  | 0.199284  | 0.000000  | 0.000000  | 0.000000  |
| Percent cost of merchandise sold in total operating expenses      | 0.438993  | 0.000000  | 0.254341  | -0.067395 | 0.000000  | 0.044065  | 0.000000  |
| Percent cost of material in total operating expenses              | -0.027509 | 0.085834  | 0.000000  | 0.630267  | 0.000000  | 0.045436  | -0.019978 |
| Percent fuel and energy cost in total operating expenses          | 0.000000  | 0.000000  | 0.000000  | 0.000000  | 0.000000  | -0.954607 | 0.000000  |
| Percent wage cost in total operating expenses                     | -0.453726 | 0.000000  | 0.000000  | -0.325539 | -0.594326 | 0.173439  | 0.000000  |
| Percent productive service cost in total operating expenses       | -0.333679 | 0.000000  | 0.000000  | -0.222344 | 0.762847  | 0.000000  | 0.000000  |
| Percent depreciation cost in total operating expenses             | 0.108694  | 0.000000  | -0.726249 | 0.000000  | 0.000000  | 0.128099  | 0.000000  |
| Percent raw material in total assets                              | 0.000000  | -0.007293 | 0.083372  | 0.624407  | -0.000201 | 0.000000  | 0.143399  |
| Percent WIP in total assets                                       | 0.000000  | 0.460374  | 0.000000  | 0.000000  | 0.000000  | 0.000000  | -0.712393 |
| Percent finished products in total assets                         | 0.000000  | 0.573218  | 0.000000  | 0.000000  | 0.000000  | 0.000000  | 0.686680  |
| Percent WIP and finished products in total assets                 | 0.000000  | 0.671977  | 0.000000  | 0.000000  | 0.000000  | 0.000000  | 0.000000  |
| Percent merchandise in total assets                               | 0.315974  | 0.000000  | 0.291738  | -0.076742 | 0.000000  | 0.031515  | 0.000000  |

**Table 10.**  
*Sparse PCA component matrix.*

| Columns/factors                                                   | Factor 0  | Factor 1  | Factor 2  | Factor 3  | Factor 4  | Factor 5  | Factor 6  |
|-------------------------------------------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| Fixed assets in total assets                                      | -0.173467 | 0.275000  | 0.507855  | -0.499215 | 0.156525  | -0.088965 | 0.078114  |
| Percent sales of merchandise in total operating revenue           | 0.525938  | -0.128407 | 0.122190  | -0.093645 | -0.109748 | -0.078270 | 0.673835  |
| Percent sales of products and services in total operating revenue | -0.444479 | -0.119035 | -0.307493 | -0.162452 | -0.114015 | -0.009416 | -0.142249 |
| Percent cost of merchandise sold in total operating expenses      | 0.510523  | -0.126489 | 0.118665  | -0.111772 | -0.087492 | 0.088045  | -0.653626 |
| Percent cost of material in total operating expenses              | -0.204519 | -0.558377 | 0.030438  | -0.282440 | -0.265091 | 0.084341  | 0.087885  |
| Percent fuel and energy cost in total operating expenses          | -0.119620 | 0.103472  | 0.453140  | 0.056594  | -0.245465 | 0.200560  | -0.125820 |
| Percent wage cost in total operating expenses                     | -0.204794 | 0.179552  | 0.159958  | 0.368300  | -0.273464 | -0.715834 | -0.010912 |
| Percent productive service cost in total operating expenses       | -0.131802 | 0.032733  | 0.012123  | 0.533088  | -0.234838 | 0.537258  | 0.183658  |
| Percent depreciation cost in total operating expenses             | -0.098392 | 0.135731  | 0.478665  | 0.038015  | -0.137017 | 0.259630  | -0.023300 |
| Percent raw material in total assets                              | -0.120240 | -0.430495 | 0.139923  | -0.141176 | -0.424059 | -0.117921 | 0.026698  |
| Percent WIP in total assets                                       | -0.048543 | -0.251299 | 0.164302  | 0.160983  | 0.282957  | -0.044589 | -0.000473 |
| Percent finished products in total assets                         | -0.062609 | -0.324119 | 0.211913  | 0.207631  | 0.364950  | -0.057519 | 0.022279  |
| Percent WIP and finished products in total assets                 | -0.071814 | -0.371772 | 0.243069  | 0.238158  | 0.418607  | -0.065970 | -0.072968 |
| Percent merchandise in total assets                               | 0.289968  | -0.108389 | 0.082710  | 0.230665  | -0.308640 | -0.196863 | -0.167039 |

**Table 11.**  
Robust PCA component matrix.

01 **Author details**

02 Stefana Janićijević\*, Vule Mizdraković and Maja Kljajić  
03 Singidunum University, Belgrade, Republic of Serbia

04 \*Address all correspondence to: [sjanicijevic@singidunum.ac.rs](mailto:sjanicijevic@singidunum.ac.rs)

**IntechOpen**

---

05 © 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms  
06 of the Creative Commons Attribution License ([http://creativecommons.org/licenses/](http://creativecommons.org/licenses/by/3.0)  
07 [by/3.0](http://creativecommons.org/licenses/by/3.0)), which permits unrestricted use, distribution, and reproduction in any medium,  
08 provided the original work is properly cited. 

## AQ03 01 References

- [1] Mrvaljevic M, Dobricanin S, Djuricanin J. Finansijski izveštaji u funkciji menadžment odlučivanja. *Ekonomski signali*. 2014;**9**(2):85-103
- [2] Djukic T, Pavlovic M. Kvalitet finansijskog izveštavanja u republici srbiji. *Ekonomske teme*. 2014;**52**(1):101-116
- [3] Off. herald of rs”, nos. 62/2013, 30/2018 and 73/2019 - other law. 2021
- [4] Bhunia A, Mukhuti SS, Roy SG. Financial performance analysis—A case study. *Current Research Journal of Social Sciences*. 2011;**3**(3):269-275
- [5] Abraham A. A Model of Financial Performance Analysis Adapted for Nonprofit Organisations. Wollongong, Australia: University of Wollongong; 2004
- [6] Bhargava P. Financial analysis of information and technology industry of India (A Case Study of Wipro Ltd and Infosys Ltd). *Journal of Accounting, Finance and Auditing Studies*. 2017;**3**(3):13
- [7] Schönbohm A. Performance Measurement and Management with Financial Ratios: The Basf se Case. Technical report, Working Paper. London: IEEE; 2013
- [8] Taylor SL. Analysing financial statements: How many variables should we look at? *JASSA*. 1986;**1**(1):19-21
- [9] Karamizadeh S, Abdullah SM, Manaf AA, Zamani M, Hooman A. An overview of principal component analysis. *Journal of Signal and Information Processing*. 2013;**4**(3B):173
- [10] Abbas AH, Arab A, Harbi J. Image compression using principal component analysis. *Mustansiriyah Journal of Science*. 2018;**29**(2):01854
- [11] Polyak BT, Khlebnikov MV. Principle component analysis: Robust versions. *Automation and Remote Control*. 2017;**78**(3):490-506
- [12] Duntelman GH. *Principal Components Analysis*. Thousand Oaks, California, United States: Sage; 1989. p. 69
- [13] Jacod J, Lejay A, Talay D. Estimation of the brownian dimension of a continuous itô process. *Bernoulli*. 2008;**14**(2):469-498
- [14] Rummel RJ. *Applied Factor Analysis*. Evanston, Illinois, United States: Northwestern University Press; 1988
- [15] Demmel JW. *Applied Numerical Linear Algebra*, SIAM. Philadelphia, Pennsylvania, United States: Society for Industrial and Applied Mathematics; 1997
- [16] Abdi H, Williams LJ. *Principal component analysis*. Wiley Interdisciplinary Reviews: Computational Statistics. 2010;**2**(4):433-459
- [17] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996;**58**(1):267-288
- [18] Hastie T, Tibshirani R, Wainwright M. The lasso for linear models. *Statistical Learning with Sparsity: The LASSO and Generalization*. 2015;**7–28**
- [19] Xu H, Caramanis C, Sanghavi S. Robust PCA Via Outlier Pursuit. arXiv preprint arXiv:1010.4237. 2010
- [20] Candès EJ, Li X, Ma Y, Wright J. Robust principal component analysis?

|    |                                                   |                                               |    |
|----|---------------------------------------------------|-----------------------------------------------|----|
| 01 | Journal of the ACM (JACM). 2011; <b>58</b> (3):   | [30] Anderson OL. A simplified method         | 42 |
| 02 | 1-37                                              | for calculating the debye temperature         | 43 |
| 03 | [21] Wright J, Ganesh A, Rao S, Peng Y,           | from elastic constants. Journal of Physics    | 44 |
| 04 | Ma Y. Robust principal component                  | and Chemistry of Solids. 1963; <b>24</b> (7): | 45 |
| 05 | analysis: Exact recovery of corrupted             | 909-917                                       | 46 |
| 06 | low-rank matrices via convex                      | [31] Waternaux CM. Asymptotic                 | 47 |
| 07 | optimization. Advances in Neural                  | distribution of the sample roots for a        | 48 |
| 08 | Information Processing Systems. 2009;             | nonnormal population. Biometrika.             | 49 |
| 09 | <b>58</b> :289-298                                | 1976; <b>63</b> (3):639-645                   | 50 |
| 10 | [22] Pearson K. LIII. On lines and planes         | [32] Jolliffe D. Whose education matters      | 51 |
| 11 | of closest fit to systems of points in            | in the determination of household             | 52 |
| 12 | space. The London, Edinburgh, and                 | income? Evidence from a developing            | 53 |
| 13 | Dublin Philosophical Magazine and                 | country. Economic Development and             | 54 |
| 14 | Journal of Science. 1901; <b>2</b> (11):559-572   | Cultural Change. 2002; <b>50</b> (2):287-312  | 55 |
| 15 | [23] Hotelling H. Analysis of a complex of        | [33] Tyler WG. Growth and export              | 56 |
| 16 | statistical variables into principal              | expansion in developing countries: Some       | 57 |
| 17 | components. Journal of Educational                | empirical evidence. Journal of                | 58 |
| 18 | Psychology. 1933; <b>24</b> (6):417               | Development Economics. 1981; <b>9</b> (1):    | 59 |
| 19 | [24] Litterman R, Scheinkman J.                   | 121-130                                       | 60 |
| 20 | Common factors affecting bond returns.            | [34] Brillinger DR. Time Series: Data         | 61 |
| 21 | Journal of Fixed Income. 1991; <b>1</b> (1):54-61 | Analysis and Theory. Philadelphia,            | 62 |
| 22 | [25] Stock JH, Watson MW. Forecasting             | Pennsylvania, United States: SIAM.            | 63 |
| 23 | inflation. Journal of Monetary                    | Society for Industrial and Applied            | 64 |
| 24 | Economics. 1999; <b>44</b> (2):293-335            | Mathematics; 2001                             | 65 |
| 25 | [26] Egloff D, Leippold M, Wu L. The              | [35] Jacod J, Podolskij M. A test for the     | 66 |
| 26 | term structure of variance swap rates             | rank of the volatility process: The           | 67 |
| 27 | and optimal variance swap investments.            | random perturbation approach. The             | 68 |
| 28 | Journal of Financial and Quantitative             | Annals of Statistics. 2013; <b>41</b> (5):    | 69 |
| 29 | Analysis. 2010; <b>45</b> (5):1279-1310           | 2391-2427                                     | 70 |
| 30 | [27] Baker M, Wurgler J. Investor                 | [36] Mykland PA, Zhang L. Inference for       | 71 |
| 31 | sentiment and the cross-section of stock          | continuous semimartingales observed at        | 72 |
| 32 | returns. The Journal of Finance. 2006;            | high frequency. Econometrica. 2009;           | 73 |
| 33 | <b>61</b> (4):1645-1680                           | 77(5):1403-1445                               | 74 |
| 34 | [28] Baker SR, Bloom N, Davis SJ.                 | [37] Li J, Todorov V, Tauchen G.              | 75 |
| 35 | Measuring economic policy uncertainty.            | Volatility occupation times. The Annals       | 76 |
| 36 | The Quarterly Journal of Economics.               | of Statistics. 2013; <b>41</b> (4):1865-1891  | 77 |
| 37 | 2016; <b>131</b> (4):1593-1636                    | [38] Saha S, Moorthi S, Pan H-L, Wu X,        | 78 |
| 38 | [29] Anderson DL, Weeks WF. A                     | Wang J, Nadiga S, et al. The ncep climate     | 79 |
| 39 | theoretical analysis of sea-ice strength.         | forecast system reanalysis. Bulletin of       | 80 |
| 40 | Eos, Transactions American Geophysical            | the American Meteorological Society.          | 81 |
| 41 | Union. 1958; <b>39</b> (4):632-640                | 2010; <b>91</b> (8):1015-1058                 | 82 |

- 01 [39] Tao M, Wang Y, Yao Q, Zou J. Large 43  
02 volatility matrix inference via combining 44  
03 low-frequency and high-frequency 45  
04 approaches. Journal of the American  
05 Statistical Association. 2011;**106**(495):  
06 1025-1040
- 07 [40] Jacod J, Rosenbaum M. Quarticity  
08 and other functionals of volatility:  
09 Efficient estimation. The Annals of  
10 Statistics. 2013;**41**(3):1462-1484
- 11 [41] Heinrich C, Podolskij M. On Spectral  
12 Distribution of High Dimensional  
13 Covariation Matrices. arXiv preprint  
14 arXiv:1410.6764. 2014
- 15 [42] Kalnina I, Xiu D. Nonparametric  
16 estimation of the leverage effect: A  
17 trade-off between robustness and  
18 efficiency. Journal of the American  
19 Statistical Association. 2017;**112**(517):  
20 384-396
- 21 [43] Jara EG, Ebrero AC, Zapata RE.  
22 Effect of international financial reporting  
23 standards on financial information  
24 quality. Journal of Financial Reporting  
25 and Accounting. 2011;**9**(2):176-196
- 26 [44] Lantto A-M, Sahlström P. Impact of  
27 international financial reporting  
28 standard adoption on key financial  
29 ratios. Accounting & Finance. 2009;  
30 **49**(2):341-361
- 31 [45] Triandafil C, Brezeanu P, Badea L.  
32 Impactul macroeconomic asupra  
33 profitabilității sectorului corporativ:  
34 analiza la nivelul companiilor listate la  
35 bursa de valori bucurești. Economie  
36 teoretică și aplicată. 2010;**17**(10):551
- 37 [46] Tudor E. Metode de recunoaștere a  
38 formelor în analiza economico-  
39 financiară. Bucharest: Academy of  
40 Economic Studies; 2009
- 41 [47] Armeanu D, Negru A. Aplicarea  
42 analizei componentelor principale în  
managementul portofoliului de investiții. Internal Auditing & Risk Management. 2011;**6**(3):01865
- [48] Robu IB, Istrate C. The analysis of the principal components of the financial reporting in the case of Romanian listed companies. Procedia Economics and Finance. 2015;**20**:553-561
- [49] Mizdrakovic V, Stanic N, Mitic V, Obradovic A, Kljajic M, Obradovic M, et al. Empirical data on financial and audit reports of Serbian business entities. In: FINIZ 2020-People in the Focus of Process Automation. Belgrade, Serbia: Singidunum University International Scientific Conference; 2020. pp. 193-198