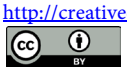
Singular Valued Decomposition and Principal Component Analysis to Compare Market Indexes

Fernando Vadillo[1], Nerea Vadillo[2]

[1]Department of Mathematics, University of the Basque Country, Leioa, Spain
[2]École des Ponts Paris Tech, Paris, France
Email: fernando.vadillo@ehu.es, nerea.vadillo-fernandez@eleves.enpc.fr

How to cite this paper: Vadillo, F. and Vadillo, N. (2021) Singular Valued Decomposition and Principal Component Analysis to Compare Market Indexes. *Journal of Mathematical Finance*, **11**, 484-494.
https://doi.org/10.4236/jmf.2021.113027

Received: July 14, 2021
Accepted: August 10, 2021
Published: August 13, 2021



Abstract

In this paper, we used the Singular Value Decomposition (SVD) to find the relationships in the fluctuation of the six market indexes CAC 40, DAX, DOW JONES 30, FTSE 100, IBEX35 and NIKKEI 225 during the year 2018. This technique allows relating several indexes in a very similar way the classical Principal Component Analysis (PCA). In fact, we will just use the statistical software to confirm some results.

Keywords

Singular Value Decomposition, Principal Component Analysis, Computational Aspects of Data Analysis

1. Introduction

It is assumed that there are six indexes: CAC 40, DAX, DOW JONES 30, FTSE 100, IBEX35 and NIKKEI 225 with $n = 254$ trading days, in fact not all indexes have the same number of days, when a value was missing the value of the index has been repeated. In this order, let be

$$\boldsymbol{Q} = \left(q_{ij}\right) = \left(\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3, \boldsymbol{q}_4, \boldsymbol{q}_5, \boldsymbol{q}_6\right) \in \mathcal{R}^{254\times 6}. \tag{1}$$

In Figure 1, you can see these values, although the difference in size prevents any idea of their possible relationships, in other words, the data needs to be normalized. The first is centering the values in each column using the mean value for that column, *i.e.*

$$Q_{ij} = q_{ij} - \overline{q}_j, \quad \text{with } \overline{q}_j = \frac{1}{254}\sum_{i=1}^{254} q_{ij}, \tag{2}$$

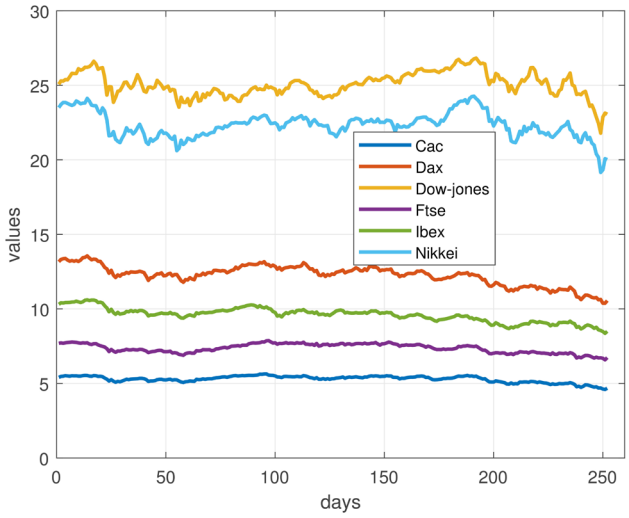


Figure 1. Variation of the index values.

and the second step is to scale the values in each column using a characteristic value S_j for that column, in this case we used the maximum entry in the jth column in absolute value, so

$$\boldsymbol{P} = \left(\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3, \boldsymbol{p}_4, \boldsymbol{p}_5, \boldsymbol{p}_6\right) = \left(p_{ij}\right) \quad \text{with} \quad p_{ij} = \frac{Q_{ij}}{S_j}, \tag{3}$$

for $i = 1, \cdots, 254$; $j = 1, \cdots, 6$. In Figure 2, we have plotted the columns of the matriz $\boldsymbol{P}$. In this graphic, it already seems to detect possible relations between the variations in the indices.

The question considered is, is there a connection between the movements in the indexes of the markets? and if there are, which are the relationships and which are not?

This paper is organized as follows. In Section 2, we describe the principal components and possible linear approaches. In Sections 2 and 5, we study the one and two dimensional approximations respectively and we compare our results with those obtained with the software in Matlab. In Section 8, we consider the five dimensional approximation with a result maybe a little surprising. Finally in Section 11, we analyze the numerical results and draw the main conclusions.

Our numerical methods were implemented in Matlab, the codes are available on request. The experiments were carried out in an Intel(R) Core(TM)2 Duo CPU U9300 @ 1.18 GHz, 1.91 GB of RAM.

2. Principal Component

The goal of Principal Component Analysis is to find the principal directions of the normalized data matrix $\boldsymbol{P} \in \mathcal{R}^{254\times 6}$. This technique has widely be used in computer science for data reduction as it enables to summarise the main directions of the data set. However, we will use an alternative option to Component Analysis called singular value decomposition. Before properly entering data

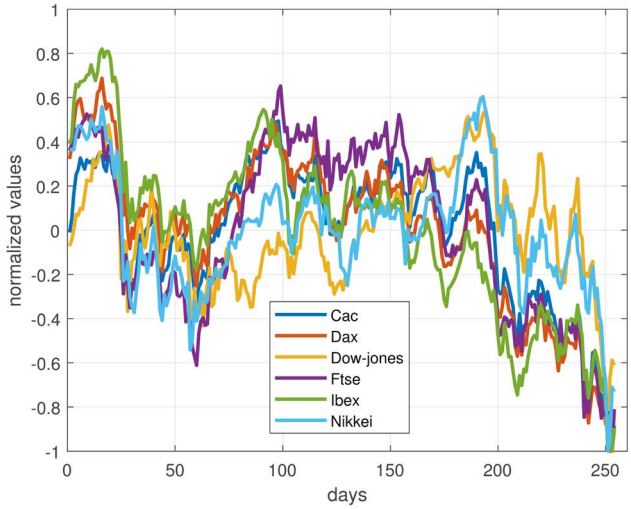


Figure 2. Variation of the normalized index values.

processing we split our data set into a training and testing set, respectively $\boldsymbol{P}1 \in \mathcal{R}^{127\times 6}$ and $\boldsymbol{P}2 \in \mathcal{R}^{127\times 6}$. $\boldsymbol{P}1$ is used to find the principal components while $\boldsymbol{P}2$ is used to test the accuracy of the approximations. In practice, we select the even rows of the dataset for training and the odd for testing.

Singular Values Composition (SVD) gives: $\boldsymbol{P}1 = \boldsymbol{U}\Sigma\boldsymbol{V}^{\mathrm{T}}$ where $\boldsymbol{U}$ is an 127×127 orthogonal matrix, $\boldsymbol{V}$ is an 6×6 orthogonal matrix and $\Sigma = \operatorname{diag}(\sigma_1, \cdots, \sigma_6)$ is an 127×6 diagonal matrix. The columns of $\boldsymbol{V}$ are called the principal components. This technique has widely been applied in applied mathematics. For further analysis we suggest some literature references [1]-[10]. In particular, SVD can be seen as a basis transformation that enables us to go from an initial six dimension corresponding to the initial data basis into a second six-dimension space with orthogonal basis. $\boldsymbol{U}$ and $\boldsymbol{V}$ correspond to the transformation matrices. In Matlab the relevant command is svd. In this particular case, the singular values of $\boldsymbol{P}1$, that is the diagonal elements of the matrix Σ are

$$\sigma_1 = 8.0164, \quad \sigma_2 = 3.3344, \quad \sigma_3 = 1.9863,$$
$$\sigma_4 = 1.2209, \quad \sigma_5 = 0.9312, \quad \sigma_6 = 0.5284.$$

and

$$\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4, \boldsymbol{v}_5, \boldsymbol{v}_6)$$
$$= \begin{pmatrix} 0.4129 & -0.0473 & 0.1955 & 0.3034 & -0.7363 & -0.3934 \\ 0.4820 & 0.1892 & -0.1485 & 0.1407 & -0.1590 & 0.8153 \\ 0.1561 & -0.7115 & -0.2788 & 0.5518 & 0.2950 & -0.0157 \\ 0.5053 & -0.0008 & 0.7078 & -0.0484 & 0.4868 & -0.0663 \\ 0.4869 & 0.3991 & -0.5855 & -0.0815 & 0.2817 & -0.4181 \\ 0.2837 & -0.5445 & -0.1345 & -0.7581 & -0.1709 & 0.0316 \end{pmatrix}$$

Because we want to determine the best linear fit of the normalized data

$$\boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3, \boldsymbol{p}_4, \boldsymbol{p}_5, \boldsymbol{p}_6),$$

one need select one of the following linear functions

$$\boldsymbol{p} = \alpha_1 \boldsymbol{v}_1, \tag{4}$$

$$\boldsymbol{p} = \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2, \tag{5}$$

$$\boldsymbol{p} = \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \alpha_3 \boldsymbol{v}_3, \tag{6}$$

$$\boldsymbol{p} = \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \alpha_3 \boldsymbol{v}_3 + \alpha_4 \boldsymbol{v}_4, \tag{7}$$

$$\boldsymbol{p} = \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \alpha_3 \boldsymbol{v}_3 + \alpha_4 \boldsymbol{v}_4 + \alpha_5 \boldsymbol{v}_5, \tag{8}$$

the first choice (4) corresponds a one-dimensional approximation or linear case, (5) is a two-dimensional approximation, etc.

3. Linear Case

In this section, we only focus in the first linear relation defined in the previous section: $\boldsymbol{p} = \alpha \boldsymbol{v}_1$ that written by components is

$$p_j = \alpha v_{j1}, \quad \text{for } j = 1, \cdots, 6.$$

Our goal is to assess the predictive power of this first approximation, in other words, if we knew an index, could we predict another index? For example, if we select the French market index (Cac) ($j = 1$) we take $\alpha = p_1 / v_{11}$, then the best fits for the other indexes are

$$p_j = \alpha\, v_{j1} = \frac{v_{j1}}{v_{11}} p_1, \quad \text{for } j = 2, \cdots, 6.$$

This procedure applied in each day $i = 1, \cdots, 127$ is the prediction

$$p_{ij} = \frac{v_{j1}}{v_{11}} p_{i1}, \quad \text{for } j = 2, \cdots, 6. \tag{9}$$

The resulting lines are shown in **Figure 3** in red for three example, from the left to right we predict the values of the indexes Dax, Ftse and Dow Jones respectively. The blue points are the real values versus from the testing set in the matrix $\boldsymbol{P}2$; we could conclude that the Cac index can be used to predict the indexesDax and Ftse but the Dow Jones doesn't. In the **Figure 4** you can see other examples, in the center we can observe the best approximation that is obtained with the Dax and Ibex indices.

On the other hand, the Principal Component Analysis (PCA) is a classical technique with a very wide bibliography, see for example ([11], Chap. 8), [12] [13] ([14], Chap. 12), ([15], Chap. 3) ([16], Chap. 4), ([17], Chap. 6), ([18], Chap. 1). This technique looks for connections between quantities, even though there is no obvious reason why they have to be connected. The Principal Component Analysis of raw data with Matlab is the command pca with the following sintaxis:

- >> [coefs,score] = pca(data);
- >> vbls = {'Cac','Dax','Dow','Ftse','Ibex','Nikkei'}
- >> biplot(coefs(:,1:2),'Scores',score(:,1:2),'VarLabels',vbls)

where the each column of the matrix

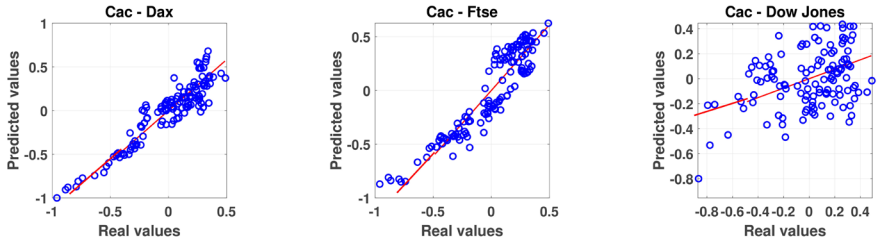


Figure 3. Prediction using Cac index. The red line come from the training set, while the data comes from the testing set.

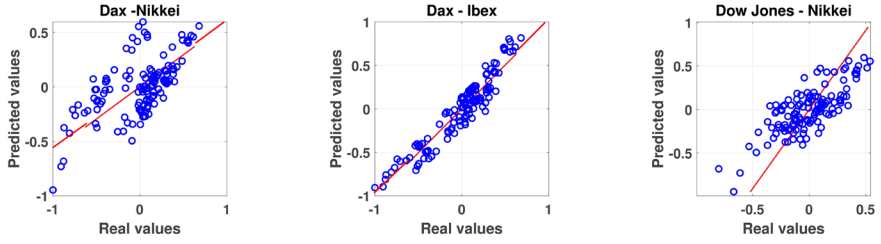


Figure 4. Other predictions.

$$\text{coefs} = \left(\boldsymbol{v}_1, -\boldsymbol{v}_2, \boldsymbol{v}_3, -\boldsymbol{v}_4, -\boldsymbol{v}_5, \boldsymbol{v}_6\right)$$

contains coefficients for one principal component in descending order of component variance, and the matrix $\text{score} \in \mathcal{R}^{127\times 6}$ correspond to observations. All six variables are represented in the biplot **Figure 5** and the direction and length of the vector indicate how each variable contributes to the two principal components, *i.e.* the first principal component, which is on the horizontal axis, has positive coefficients for the six variables and the largest coefficient in the first principal component corresponding: Ftse, Dax, Ibex and Cac Indexes. Morevover, the second principal component, which is on the vertical axis, has positive coefficients for the indexes Dow and Nikkei and negative for Ibex and Dax. The red points for each of the 176 observations indicates the score of each observation for the two principal components in the plot. On the other hand, note that the nearest indexes are the British Ftse with the French Cac and German Dax with Spanish Ibex. Moreover, the American Dow Jones is the furthest away.

4. Two Dimensional Approximation

In this subsection we will try to answer the following question: knowing two indices are we able to predict the other four ones? For simplification, let assume we only have the first two variables; the French CAC index and the German DAX index. We would like to predict the other indices: Dow Jones ($j=3$), Ftse ($j=4$), Ibex ($j=5$) or Nikkei with ($j=6$). For each index *j*, because $\boldsymbol{p} = \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2$, in components

$$p_j = \alpha_1 v_{j1} + \alpha_2 v_{j2}, \quad \text{for } j = 2, \cdots, 6.$$

and we must compute the parameters α_1, α_2 using the data, *i.e.*, for each $i = 1, \cdots, 127$ solving the two dimensional system

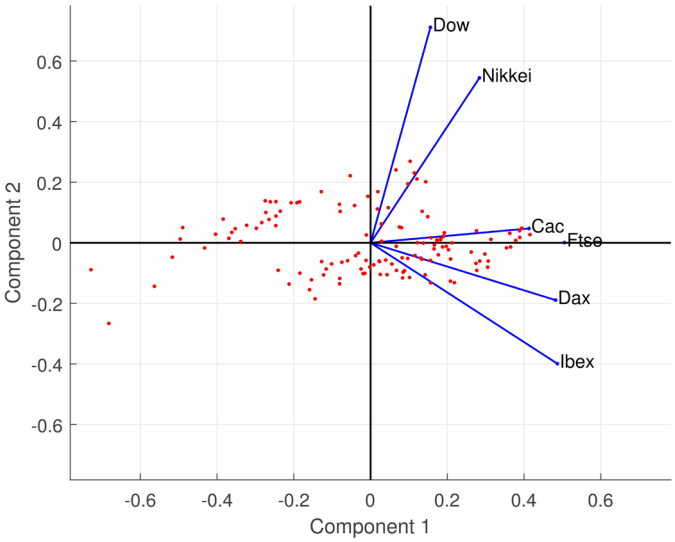


Figure 5. Representation of two principal components with biplot.

$$\begin{pmatrix} p_{i1} \\ p_{i2} \end{pmatrix} = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix} \begin{pmatrix} \alpha_{i1} \\ \alpha_{i2} \end{pmatrix}, \tag{10}$$

and the i-th prediction for the j index is

$$p_{ij} = \alpha_{i1} v_{j1} + \alpha_{i2} v_{j2}. \tag{11}$$

In **Figure 6** and **Figure 7** we have plotted the four cases. For each one, the red line is what would be obtained if the predicted values and real values are equal and the blue points are the real values versus the values predicted using the training set. From these four graphs two predictions are reasonable (Ftse and Ibex) and the other two are quite bad (Dow Jones and Nikkei).

On the other hand, using the before software in Matlab we can represent three components by typing:

- >> biplot(coefs(:,1:3),'Scores',score(:,1:3),'VarLabels',vbls)

with the result in **Figure 8** a picture similar to the previous **Figure 5** but with three principal components. Maybe the most remarkable detail is that the vectors for indexes Cac, Dax, Ftse and Ibex are almost on a single plane.

5. Five Dimensional Approximation

The question that we now consider is whether known five indices, how well can predict the sixth.

Now the fit is (8) *i.e.*

$$\boldsymbol{p} = \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \alpha_3 \boldsymbol{v}_3 + \alpha_4 \boldsymbol{v}_4 + \alpha_5 \boldsymbol{v}_5, \tag{12}$$

which in matrix form is

$$\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_6 \end{pmatrix} = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{15} \\ v_{21} & v_{22} & \cdots & v_{25} \\ \vdots & \vdots & \ddots & \vdots \\ v_{61} & v_{62} & \cdots & v_{65} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_5 \end{pmatrix}. \tag{13}$$

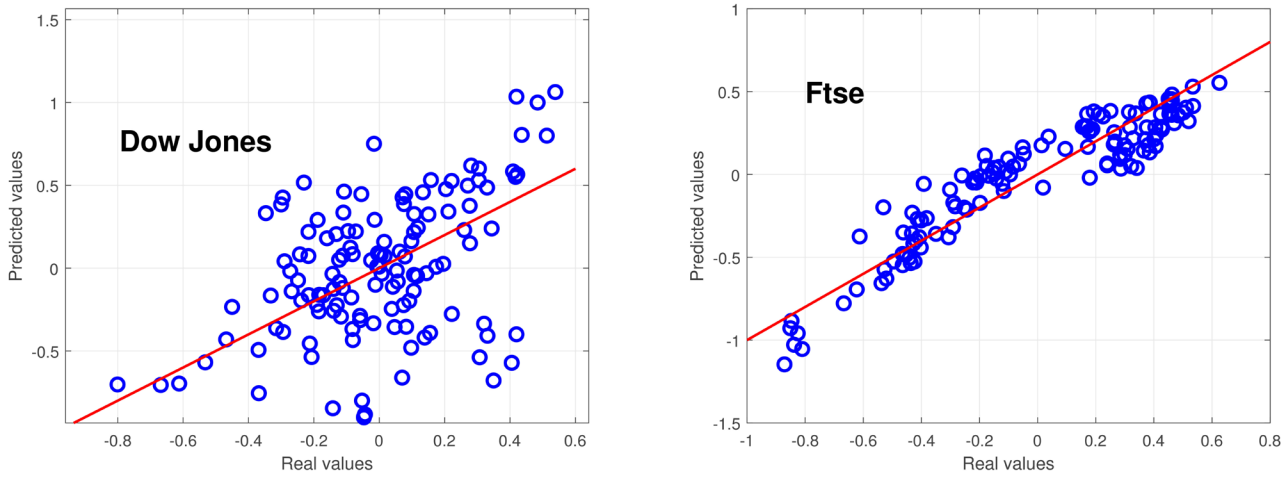


Figure 6. Test of the predictions for Dow Jones and Ftse using Cac and Dax.

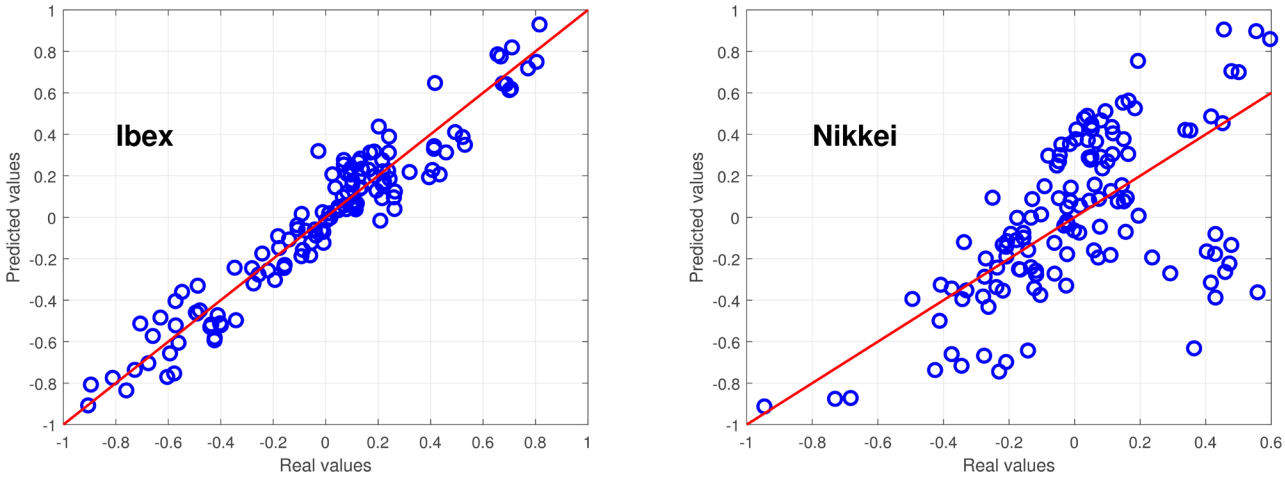


Figure 7. Test of the predictions for Ibex and Nikkei using Cac and Dax.

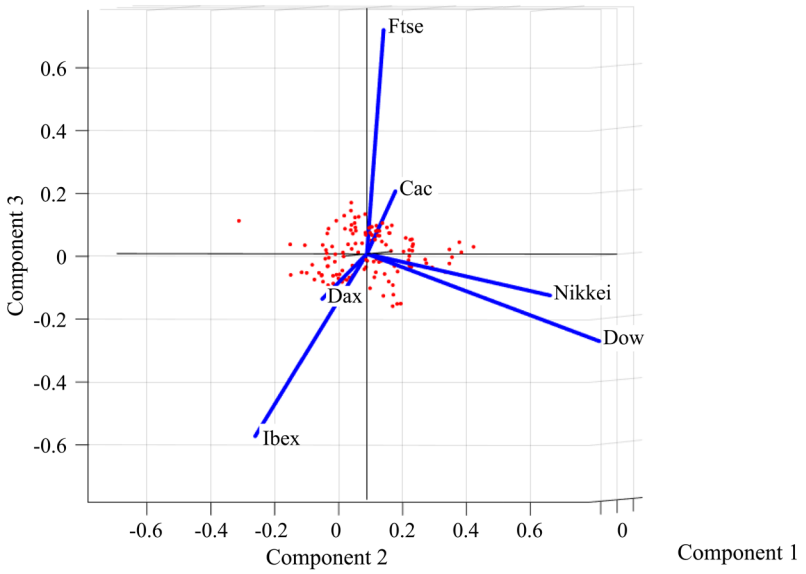


Figure 8. Representation of three principal components with biplot.

In order to make the mathematical formulation easier, we assume that the interesting variable is p_1, then we need to find the α_k's in terms of $p_2,\cdots,p_6$ what is achieved by solving the linear system

$$\begin{pmatrix} p_2 \\ p_3 \\ \vdots \\ p_6 \end{pmatrix} = \begin{pmatrix} v_{21} & v_{22} & \cdots & v_{25} \\ v_{31} & v_{32} & \cdots & v_{35} \\ \vdots & \vdots & \ddots & \vdots \\ v_{61} & v_{62} & \cdots & v_{65} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_5 \end{pmatrix}, \tag{14}$$

If we write $\alpha_k = \sum_{j=2}^{6} a_{kj} p_j$ for $k = 1,\cdots,5$

$$\begin{aligned} p_1 &= \alpha_1 v_{11} + \alpha_2 v_{12} + \alpha_3 v_{13} + \alpha_4 v_{14} + \alpha_5 v_{15} \\ &= a_2 p_2 + a_3 p_3 + a_4 p_4 + a_5 p_5 + a_6 p_6, \end{aligned}$$

with $a_k = \sum_{j=1}^{5} v_{1j} a_{jk}$.

Finally, this procedure applied in each data $i = 1,\cdots,127$ result prediction

$$p_{i1} = a_2 p_{i2} + a_3 p_{i3} + a_4 p_{i4} + a_5 p_{i5} + a_6 p_{i6}. \tag{15}$$

In Figures 9-11 we have represented the six cases. Similarly to the previous Figure 6 and Figure 7, the red line is what would be obtained if the real values and predicted values are equal and the blue points are the real versus the predicted values using the training set. From these figures we could highlight the following comment:

There are two different behaviors, while for the indexes Dax, Cac and Ibex the predictions are reasonable, the other three indexes have quite bad predictions.

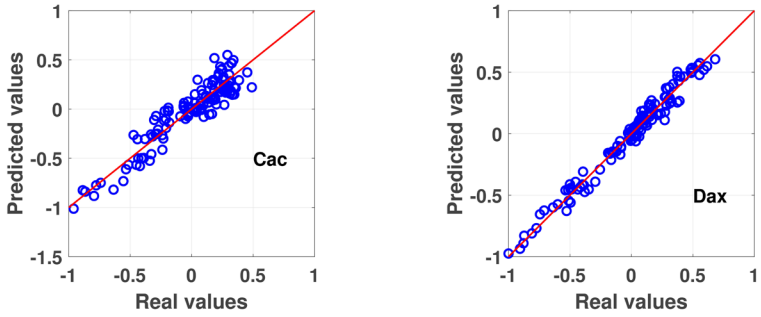


Figure 9. The real values versus the predicted values using the training set for Cac and Dax respectively.

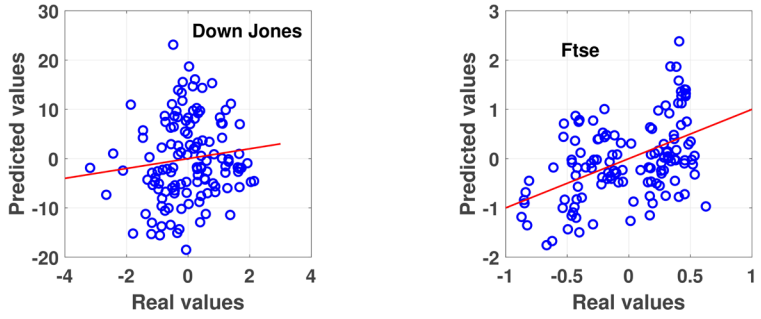


Figure 10. The real values versus the predicted values using the training set for Dow Jones and Ftse respectively.

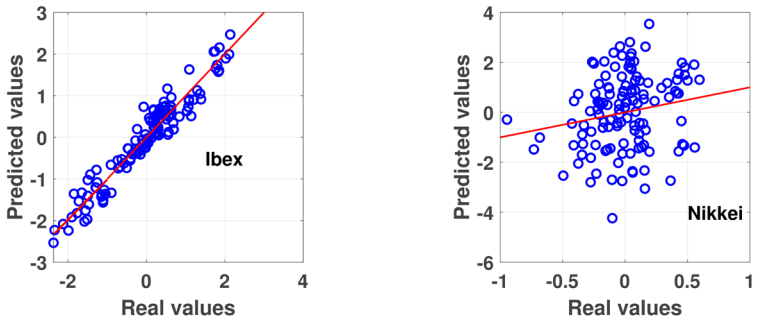


Figure 11. The real values versus the predicted values using the training set for Ibex and Nikkei respectively.

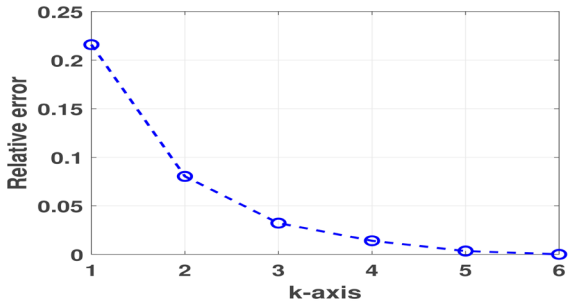


Figure 12. Relative error (16).

6. Conclusions and Discussion

In this paper, we have considered six indexes: CAC 40, DAX, DOW JONES 30, FTSE 100, IBEX35 and NIKKEI 225 in the year 2018, and we have tried to link their movements using the principal components of the matrix of their fluctuations. Our numerical results are very close to the numerical software with Matlab, however, the result of the last Section 8 might be a little surprising, especially because of the deviation in the English index Ftse.

In this point, a question we might ask is: how many columns of $\boldsymbol{V}$ is needed for our analysis? ([19] [20]). If we use the first k columns from $\boldsymbol{V}$, the relative error is given in ([10], p. 415) by

$$R_k = \frac{\sigma_{k+1}^2 + \cdots + \sigma_6^2}{\sigma_1^2 + \cdots + \sigma_6^2}, \tag{16}$$

for $k = 1, \cdots, 5$ representing in Figure 12. Usually, this information is used to decide on how many components to use for a PCA, the most used is based on a noticeable change in this plot, applying this to Figure 12, one would again decide on taking one or two columns and no more. There are those who use a criterion of the form $R_k < tol$, where tol is a number chosen somewhere between 0.25 and 0.05, here the result is similar. Moreover, after the results of Figure 5 and Figure 8 in this research it would seem that the linear case with one column is the best.

This small academic exercise does not conclude important results, however, we believe that it is quite easy to extend this kind of analysis to longer series of

data or also apply it to other indexes, for example in Investing.com there are 45 indexes. This would need to prepare the data files, a routine but important job, but with little academic interest. The main philosophy of all this is written by Yuval Noah Harari in the introduction of [21]: *In a wold deluged by irrelevant information, clarity is power*.

Fund

This work was supported by Spanish Ministry of Sciences Innovation and Universities with the project PGC2018-094522-B-100 and by the Basque Government with the project IT1247-19.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Bulirsch, R. and Stoer, J. (1980) Introduction to Numerical Analysis. Springer, Berlin.

[2] Golub, G.H. and Van Loan, C.F. (1989) Matrix Computations. The Johns Hopkins University Press, Baltimore.

[3] Walkins, D.S. (1991) Fundamentals of Matrix Computations. John Wiley, Hoboken.

[4] Higham, N.J. (1996) Accuracy and Stability of Numerical Algorithms. SIAM, Philadelphia.

[5] Trefethen, L.N. and Bau, D. (1997) Numerical Linear Algebra. SIAM, Philadelphia. https://doi.org/10.1137/1.9780898719574

[6] Meyer, C.D. (2000) Matrix Analysis and Applied Linear Algebra. SIAM, Philadelphia. https://doi.org/10.1137/1.9780898719512

[7] Moler, C.B. (2004) Numerical Computing with Matlab. SIAM, Philadelphia. https://doi.org/10.1137/1.9780898717952

[8] Yang, W.Y., Cao, W., Chung, T.S. and Morris, J. (2005) Applied Numerical Methods Using Matlab. Wiley Interscience, Hoboken. https://doi.org/10.1002/0471705195

[9] Ascher, U.M. and Greif, C. (2011) A Firts Course in Numerical Methods. SIAM, Philadelphia. https://doi.org/10.1137/9780898719987

[10] Holmes, M.H. (2016) Introduction to Scientific Computing and Data Analysis. Springer, Berlin. https://doi.org/10.1007/978-3-319-30256-0

[11] Marques de Sa, J.P. (2007) Applied Statistics Using SPSS, STATISTICA, MATLAB and R. Second Edition, Springer, Berlin. https://doi.org/10.1007/978-3-540-71972-4

[12] Jolliffe, I.T. (2010) Principal Component Analysis. Second Edition, Springer, Berlin. https://doi.org/10.1007/978-3-642-04898-2_455

[13] Abdi, H. and Williams, L.J. (2010) Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 433-459. https://doi.org/10.1002/wics.101

[14] Rencher, A.C. and Christensen, W.F. (2012) Methods of Multivariate Analysis. 3rd Edition, Wiley, Hoboken. https://doi.org/10.1002/9781118391686

[15] Dolliffe, I.T. (2016) Computational and Statistical Methods for Analysing Big Data

with Applications. Elsevier, Amsterdam.

[16] Chalfileld, C. and Collins, A.J. (2017) Introduction to Multivariate Analysis. Chapman and Hall/CRC, London.

[17] Olive, D.J. (2017) Robust Multivariate Analysis. Springer, Berlin. https://doi.org/10.1007/978-3-319-68253-2

[18] Husson, F., Lê, S. and Pag, J. (2017) Exploring Multivariate Analysis by Examples Using R. CRC Press, Boca Raton. https://doi.org/10.1201/b21874

[19] Peres-Neto, P.R., Jackson, D.A. and Somers, K.M. (2005) How Many Principal Components? Stopping Rules for Determining the Number of Non-Trivial Axes Revisited. *Computational Statistic & Data Analysis*, **49**, 974-997. https://doi.org/10.1016/j.csda.2004.06.015

[20] Josse, J. and Husson, F. (2012) Selecting the Number of Components in Principal Component Analysis Using Cross-Validation Approximations. *Computational Statistic & Data Analysis*, **56**, 1869-1879. https://doi.org/10.1016/j.csda.2011.11.012

[21] Harari, Y.N. (2018) 21 Lessons for the 21st Century. Vintage Books, New York. https://doi.org/10.17104/9783406727795-21