The International Journal of Informatics, Media and Communication Technology (IJIMCT)
Available online: https://ijimct.journals.ekb.eg/
ISSN: Online : 2682-2881 Print: : 2682-2105

Principal Component Regression for Egyptian Stock Market Prediction

Dr. Heba M. Ezzat

Assistant Professor in Department of Socio-Computing, Faculty of Economics and Political Science, Cairo University, Cairo, Egypt

ABSTRACT

Financial markets are very rich with information and variables. In contradiction to the Efficient Market Hypothesis, much research has been conducted to predict asset prices with promising accuracy. However, ensuring good models requires extracting important information from given data sets. This paper investigates the main Egyptian Stock Exchange index (EGX 30) and constructs some alternative portfolios by identifying important linear combinations of EGX 30 constituents. This could be approached by a dimensionality reduction technique, which is performed following the principal components analysis (PCA). The results show that the first three Principal Components (PCs) could summarize 83% of data variability. Each one of the first three PCs highlights the most contributed individual stocks. These three PCs provide investors with alternative portfolios. Moreover, a Principal Component Regression (PCR) model is built to predict the future behavior of the EGX 30. The performance of the obtained PCR model is very well. This result is reached by comparing observed values of EGX 30 with the predicted ones (R-squared estimated as 0.98).

ARTICLE INFO

Article history:

Received 29 nov.2020

Accepted 4 feb 2021

Keywords:

Dimensionality reduction, EGX 30, principal component regression, multiple imputation

1. Introduction

Financial markets play significant role in any economy (Mishkin 2010). Contrary to the Efficient Market Hypothesis (EMH), many researchers built different models to predict financial markets (Fama 1970, Chen and Yeh 2002). Predicting financial markets is not impossible any more especially with the increasing computational power (Hargreaves 2019, Zhong and Enke 2019, Cavalcante, et al. 2016).

However, to predict financial markets we need first to identify important factors affecting asset prices. Much research was carried out to predict financial markets using dimensionality reduction techniques (Cao and Wang 2020, Ghorbani and Chong 2020, Zhang 2018, Waqar, et al. 2017). To the best of our knowledge, no research as run to predict the Egyptian Exchange (EGX) following the principal Component Analysis (PCA) as a dimensionality reduction technique.

For this purpose, the main Egyptian stock index (EGX 30) is investigated. EGX 30 contains the major 30 individual stocks in terms of activity and liquidity traded in the EGX. The main aim of this research is to find lower dimensions that can capture most of the data variability. However, there are two main challenges; (i) most of the variables (individual stock prices) are multicollinear and (ii) there are huge number of dimensions (30).

To approach research objectives, the dimensionality reduction algorithm PCA is applied. A data set containing p variables can be converted into a new set of p Principal Components (PCs) following PCA, where each principal component is a linear combination of all original variables. The original large number of variables may be substituted by a much smaller number of PCs that explain most of the

variation in the data. Eigenvalues of PCs typically decay rapidly, and the higher numbered PCs have comparatively small eigenvalues.

It is very interesting to construct portfolios based on the PCs to get a disclosure to all the risk sources. Thereafter, it seems excessive to allocate risk budget to higher PCs which are not major risk sources. This manuscript aims to help identifying important individual stocks to predict the main index of the EGX. To the best of the knowledge, no research has been conducted to extract important individual stocks and their contribution to the EGX 30 following PCA.

The paper is constructed as following. Section 2 presents data description and preprocessing. Section 3 is devoted to explaining the research methodology. Main results and analyses are explained in Section 4. Finally, Section 5 concludes the paper.

2. Data description and preprocessing

This research investigates the EGX 30 index, which is a market capitalization weighted index of the 30 top companies in terms of activity and liquidity. However, Egyptian economy and political systems witnessed a period of instability since 2011 revolution till the presidential elections conducted in 2014. Additionally, COVID-19 caused economic instability all over the world since late October 2019. Thereafter, we investigated the EGX 30 price index and its constituents from June 2014 to October 2019 (1311 observations). The total number of stocks registered during this period is 42. The set of stocks that had been registered in EGX 30 for the whole study period was extracted, and there were 20 such stocks. The remaining 22 stocks were either listed after August 2014 or delisted before August 2019. Description of individual stocks' names, symbols, and sectors is provided in Table 1.

As depicted in Table 1, some data were missing. To delete their respective records, the Missing Completely At Random (MCAR) Test should be run. The p-value for the Hawkins test of normality and homoscedasticity is 1.75e-29. This indicates that either the test of multivariate normality or homoscedasticity (or both) is rejected. Provided that normality can be assumed, the hypothesis of MCAR is rejected at 0.05 significance level.

Table 1. Description of the 20 individual stocks listed in EGX 30 during the investigated period.

company	Stock	Number of missing data (Percentage)	Sector
Arab Cotton Ginning	ACGC	2 (0.15%)	Personal and Household Products
Oriental Weavers	ORWE	2 (0.15%)	
Eastern Company	EAST	71 (5.41%)	
Citadel Capital – Common Shares	CCAP	7 (0.53%)	Financial Services excluding Banks
Egyptian Financial Group-Hermes Holding Company	HRHO	—	
Egyptian Kuwaiti Holding	EKHO	—	
Pioneers Holding	PIOH	—	
Commercial International Bank (Egypt)	COMI	—	Banks
Credit Agricole Egypt	CIEB	59 (4.50%)	
Egyptian Chemical Industries (Kima)	EGCH	—	Chemicals
Juhayna Food Industries	JUFO	47 (3.58%)	Food and Beverage
Egyptian for Tourism Resorts	EGTS	—	Travel & Leisure
Egyptian Iron & Steel	IRON	1 (0.08%)	Basic Resources
Ezz Steel	ESRS	13 (0.99%)	
ELSWEDY ELECTRIC	SWDY	—	Industrial Goods and Services and Automobiles
Heliopolis Housing	HELI	1 (0.08%)	Real Estate
Medinet Nasr Housing	MNHD	—	

Palm Hills Development Company	**PHDC**	**—**
Six of October Development & Investment (SODIC)	**OCDI**	**2 (0.15%)**
T M G Holding	**TMGH**	**—**

Thereafter, deletion of missed values could produce hugely biased data set. Missing values were replaced following multiple imputation by chained equations (MICE) approach that uses Classification And Regression Trees (CART) (Burgette and Reiter 2010). It does not set parametric assumptions or data transformations to fit nonlinear relations and complex distributions. The algorithm was run for 1000 times and the median of the resulted 1000 imputed numbers was used for replacement. Figure 1 shows stock closing prices after imputing missing values in individual stocks.

Financial time series are characterized by heteroskedasticity which would influence the results of the PCA. Therefore, analyzing stock returns is preferred over stock prices. Stock returns, r_{tj}, are calculated such that;

$$r_{tj} = \left(\frac{p_{tj} - p_{(t-1)j}}{p_{(t-1)j}}\right) \quad (1)$$

where $p_{tj} = \log(P_{tj})$ and P_{tj} is the closing price at time t, where $t = 1,2,\dots,T$, for stock j, where j=1, 2,…, J.

3. Research Methodology

Stock market data are highly correlated (see Figure 2). PCA is suitable to explore such data set as it can address multicollinearity and high dimensionality. After preprocessing the data, the PCA algorithm is applied. Data adequacy for PCA is tested by Kaiser-Meyer-Olkin

(KMO) Statistics. The lowest (highest) value of the KMO is 0.79 (0.95) for HELI (PIOH) stock. The KMO-criterion for the whole data set is 0.89 revealing the adequacy to perform the PCA.

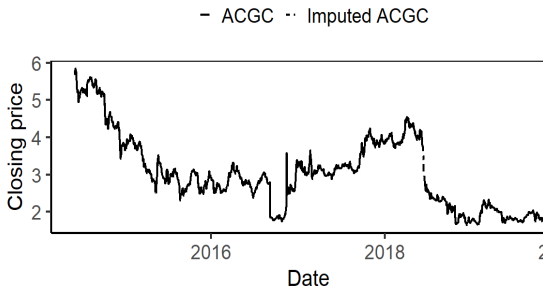
PCA is one of the most used techniques for Dimensionality reduction. The main purpose of this technique is to project the data in new directions. Thus, PCA aims to find lower dimensions (less than the number of original variables included in the model) that can explain most of the data variability. For instance, the data set under investigation contains 20 variables. PCA will output, hopefully, lower number of PCs (less than 20) than can explain most of the data variability (at least 60%). The new dimensions or PCs are orthogonal, so they are uncorrelated.

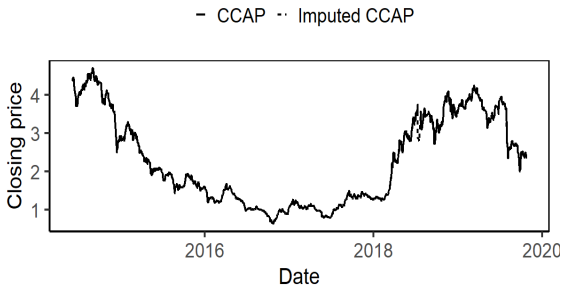
To find the PCs, we need first to standardize the data, such as

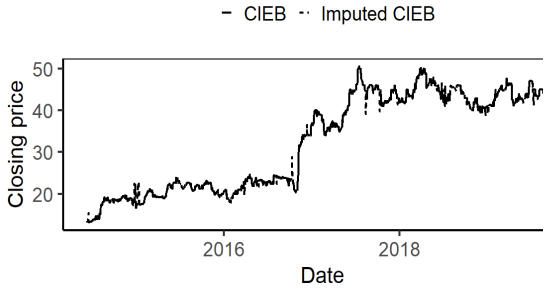
$$z_{tj} = \frac{r_{tj} - \bar{r}_j}{s_j} \quad (2)$$

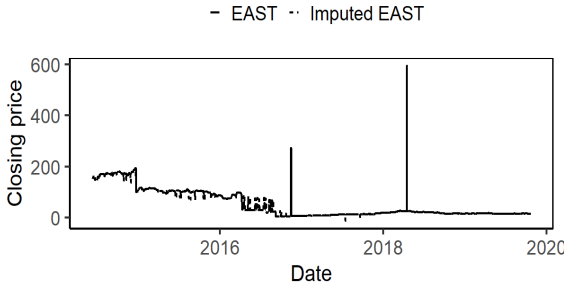
where $\bar{r}_j$ is the mean and s_j is the standard deviation for stock j, $where\ j = 1,2, \dots, J$.

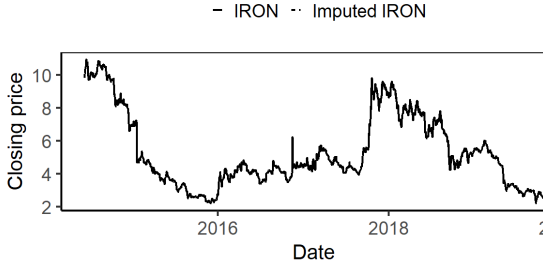
So, variables contribute equally to the PCA. Then, we calculate the covariance or correlation matrix for standardized values. However, correlation matrix is preferred if stock prices are reported with different currencies. Correlation matrix is considered as a safer option. Eigenvectors of the correlation matrix are the directions that explain most variance or simply the PCs. Eigenvalues are the coefficients assigned to eigenvectors. Eigenvalues provide the amount of variance captured by each PC. The final step to compute the PCs, eigenvectors and eigenvalues are computed for the correlation matrix.

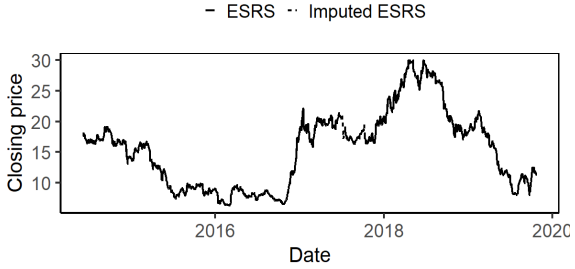

ACGC
Imputed ACGC
Closing price
Date
2016
2018

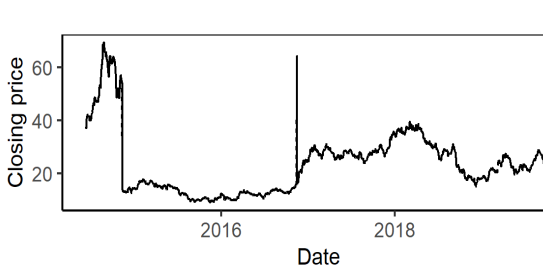

CCAP
Imputed CCAP
Closing price
Date
2016
2018
2020

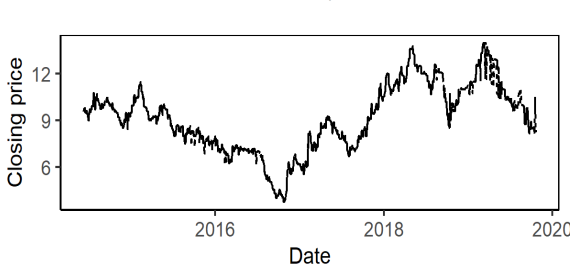

CIEB
Imputed CIEB
Closing price
Date
2016
2018


EAST
Imputed EAST
Closing price
Date
2016
2018
2020


IRON
Imputed IRON
Closing price
Date
2016
2018


ESRS
Imputed ESRS
Closing price
Date
2016
2018
2020


HELI
Imputed HELI
Closing price
Date
2016
2018


JUFO
Imputed JUFO
Closing price
Date
2016
2018
2020

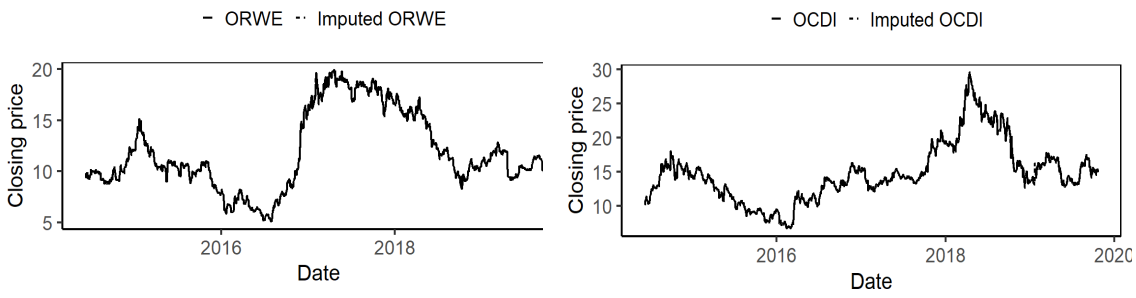


Figure 1. Multiple imputation for individual stock prices with missing data.

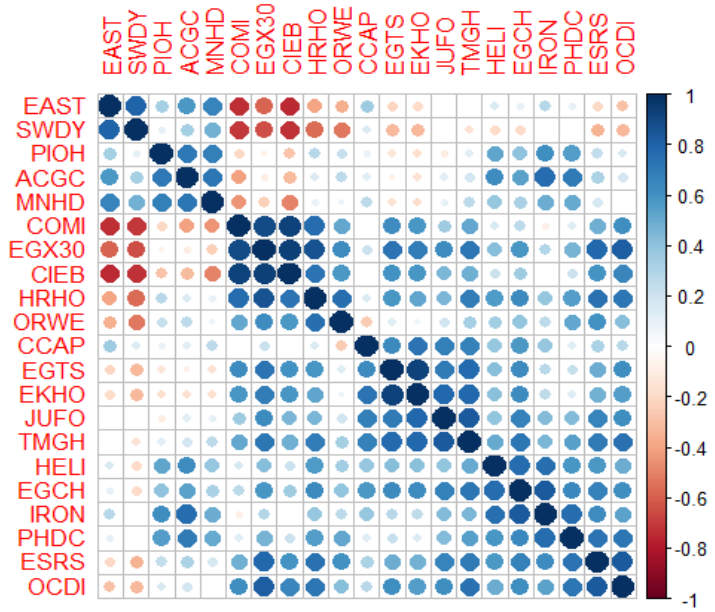


Figure 2. Correlation estimates at 0.01 significance level.

4. Results and Analyses

This section is devoted to display the results of implementing the PCA. Table 2 reports the eigenvalue, percentage of explained variance, and cumulative percentage of variance for each PC. To identify the number of PCs that could be considered, there are three rules in the literature to be followed (Jolliffe 2002); (i) the first PCs that explain the required percentage of cumulative variance are included, (ii) the elbow

in scree plot, and/or (iii) choosing PCs with eigenvalues that are greater than one, this is known as Kaiser's rule (Kaiser 1960). Table 2 shows that the first three PCs could explain about 83% of data variability. This cumulative percentage of variance is considered very satisfactory. Figure 3 displays the scree plot, which depicts that after the fourth PC the added explanation for the variance is very low. Finally, Table 2 reports that eigenvalues attached to the first three PCs are greater than one. This reveals that these PCs can explain variance more than the variance that could be explained by single original variables. So, the three rules indicate considering the first three PCs in the analysis.

Each panel in Figure 4 depicts the highest ten individual stocks contributing to the first three PCs. For instance, TMGH, ESRS, EGCH, OCDI, HRHO, EGTS, JUFO, EKHO, HELI, and PHDC are the highest stocks contributing to the first PC, ACGC, MNHD, EAST, COMI, PIOH, CIEB, SWDY, and IRON are the highest stocks contributing to the second PC, and CCAP, ORWE, EKHO, JUFO, EGTS, PIOH, and SWDY are the highest stocks contributing to the third PC. The relationship between the variables and the first three PCs could be depicted by the following three equations:

$$PC1 = 0.3ACGC + 0.49CCAP + 0.67CIEB + 0.62COMI - 0.2EAST + 0.86EGCH + 0.76EGTS + 0.74EKHO + 0.87ESRS + 0.71HELI + 0.86HRHO + 0.65IRON + 0.76JUFO + 0.16MNHD + 0.86OCDI + 0.56ORWE + 0.69PHDC + 0.26PIOH - 0.41SWDY + 0.87TMGH \quad (3)$$

$$PC2 = 0.86ACGC + 0.25CCAP - 0.69CIEB - 0.71COMI + 0.84EAST + 0.31EGCH - 0.27EGTS - 0.28EKHO + 0.04ESRS + 0.40HELI - 0.19HRHO + 0.61IRON + 0.01JUFO + 0.85MNHD - 0.08OCDI - 0.13ORWE + 0.46PHDC + 0.69PIOH + 0.65SWDY + 0.04TMGH \quad (4)$$

$$PC3 = -0.27ACGC + 0.76CCAP - 0.16CIEB - 0.13COMI + 0.35EAST + 0.09EGCH + 0.46EGTS + 0.57EKHO - 0.18ESRS - 0.10HELI - 0.31HRHO - 0.11IRON + 0.47JUFO - 0.07MNHD - 0.08OCDI - 0.66ORWE - 0.35PHDC - 0.41PIOH + 0.38SWDY + 0.31TMGH \quad (5)$$

By considering (3), (4), and (5), 83% of data variability could be explained.

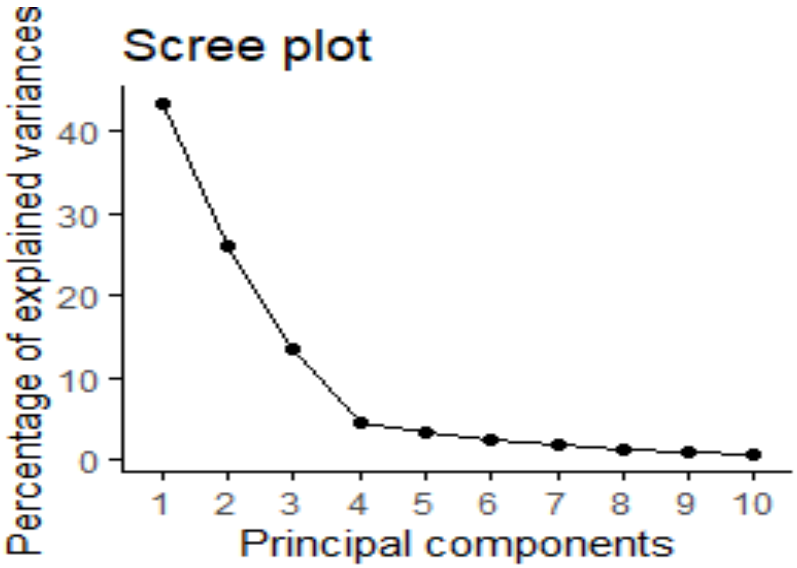


Figure 3. Scree plot depicting the percentage of variance explained by each PC.

Figure 5 depicts the three Principal Portfolios (PPs) against the EGX 30. To predict the future behavior of EGX 30, Principal Component Regression (PCR) is built. PCR is a regression model, which is regressing the dependent variable (EGX 30) on the PCs. The main advantage of the PCR is reducing dimensionality and then fit a linear regression model to the set with lower dimension. This is applied while keeping most of the variability of the original predictors. Another important benefit of the PCR is avoiding multicollinearity between variables in the data set.

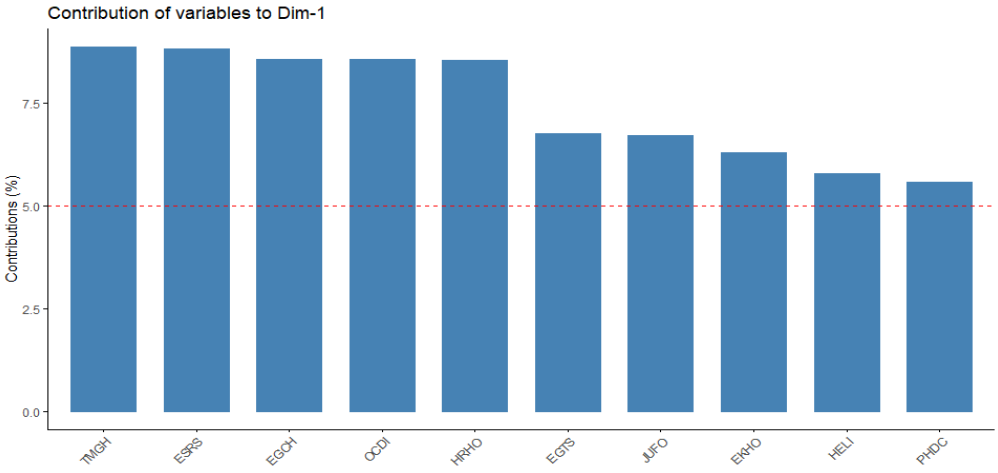
Table 2. Summary of main results of the PCA.

PCs	Eigenvalue	% of var.	Cumulative % of var.	PCs	Eigenvalue	% of var.	Cumulative % of var.
Dim.1	**8.64**	**43.21**	**43.21**	**Dim.11**	**0.13**	**0.63**	**98.10**
Dim.2	**5.16**	**25.83**	**69.05**	**Dim.12**	**0.09**	**0.45**	**98.55**
Dim.3	**2.71**	**13.56**	**82.61**	**Dim.13**	**0.06**	**0.32**	**98.87**
Dim.4	**0.90**	**4.50**	**87.11**	**Dim.14**	**0.05**	**0.27**	**99.14**
Dim.5	**0.67**	**3.37**	**90.47**	**Dim.15**	**0.04**	**0.22**	**99.36**
Dim.6	**0.50**	**2.50**	**92.97**	**Dim.16**	**0.04**	**0.19**	**99.55**
Dim.7	**0.35**	**1.77**	**94.74**	**Dim.17**	**0.03**	**0.15**	**99.69**
Dim.8	**0.24**	**1.22**	**95.96**	**Dim.18**	**0.03**	**0.13**	**99.82**
Dim.9	**0.17**	**0.84**	**96.81**	**Dim.19**	**0.02**	**0.12**	**99.94**
Dim.10	**0.13**	**0.67**	**97.47**	**Dim.20**	**0.01**	**0.06**	**100.00**

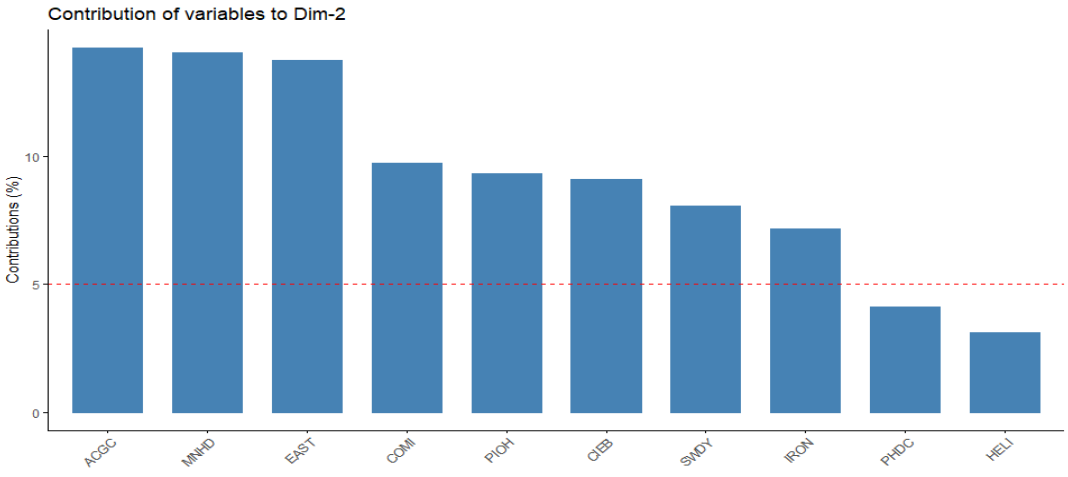
Figure 6 displays the results of applying cross-validation method to verify the best model. The figure illustrates that, the minimal Mean Squared Error of Prediction (MSEP) is reached at the second PC. Henceforth, the first two PCs are sufficient for predicting EGX 30 (97% of variability is explained with the first two PCs). The PCR model containing the first two PCs in terms of the original predictors could be displayed in (6) and (7) as follows.

$$PC1 = 0.03ACGC + 0.05CCAP + 0.07CIEB + 0.06COMI - 0.02EAST + 0.09EGCH + 0.08EGTS + 0.07EKHO + 0.09ESRS + 0.07HELI + 0.09HRHO + 0.07IRON + 0.08JUFO + 0.02MNHD + 0.09OCDI + 0.06ORWE + 0.07PHDC + 0.03PIOH - 0.04SWDY + 0.09TMGH \quad (6)$$

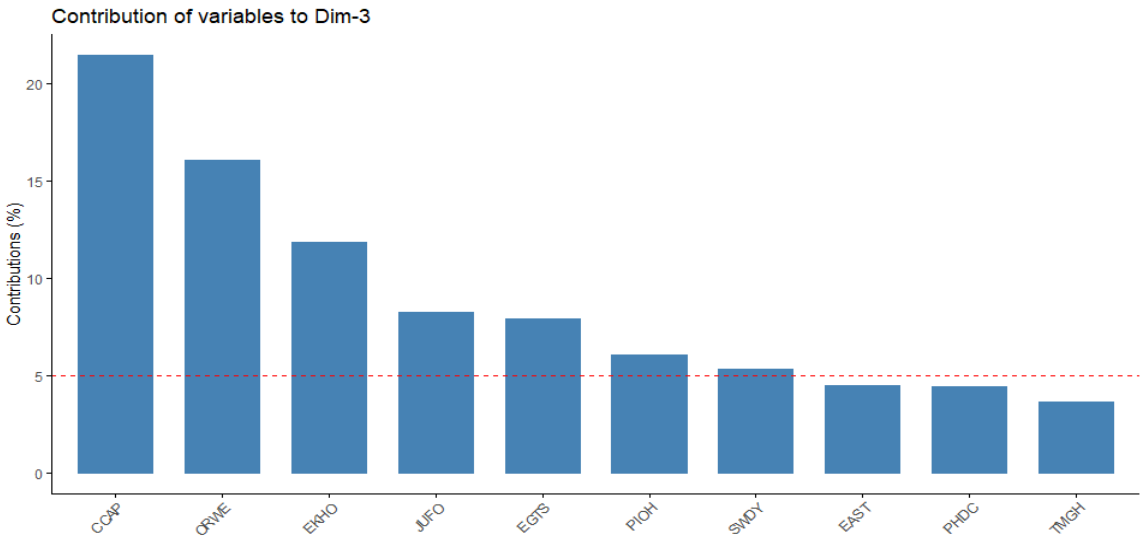
$$PC2 = -0.04ACGC + 0.03CCAP + 0.13CIEB + 0.13COMI - 0.09EAST + 0.06EGCH + 0.10EGTS + 0.10EKHO + 0.08ESRS + 0.04HELI + 0.10HRHO + 0.01IRON + 0.08JUFO - 0.06MNHD + 0.09OCDI + 0.07ORWE + 0.03PHDC - 0.03PIOH - 0.10SWDY + 0.09TMGH \quad (7)$$



(a)



(b)



(c)

Figure 4. The most ten contributing stocks to the first three PCs, respectively.

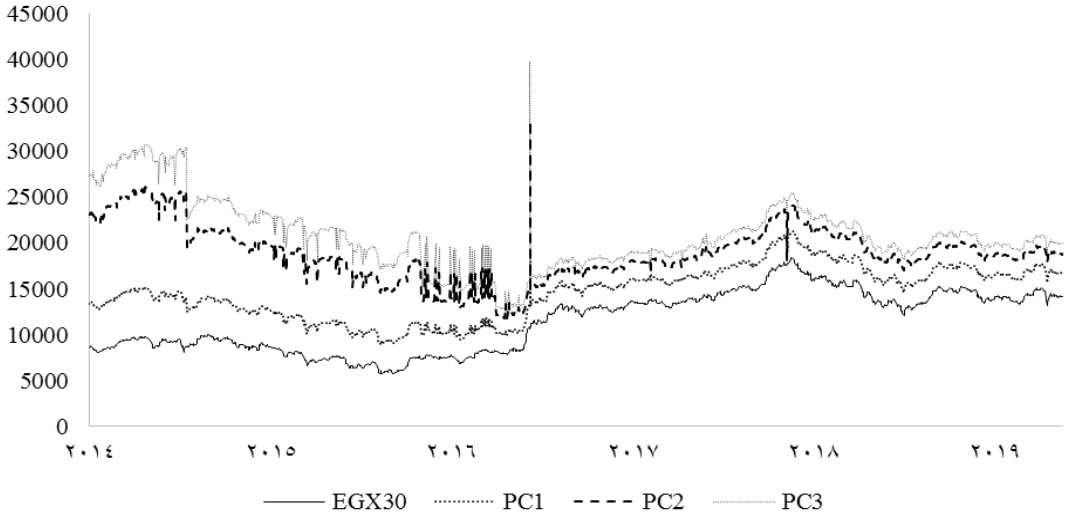


Figure 5. Plots of principal portfolios 1 to 3 with the EGX30 index.

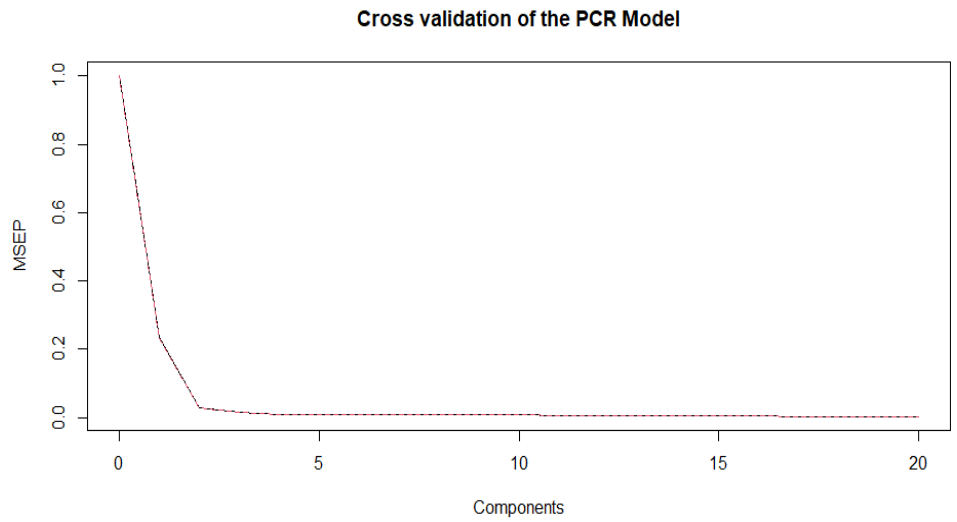


Figure 6. Cross-validation method to identify the best model.

Figure 7 graphs the observed EGX 30 values against the predicted values by the PCR model. The figure depicts that the PCR performs very well in predicting EGX 30 values.

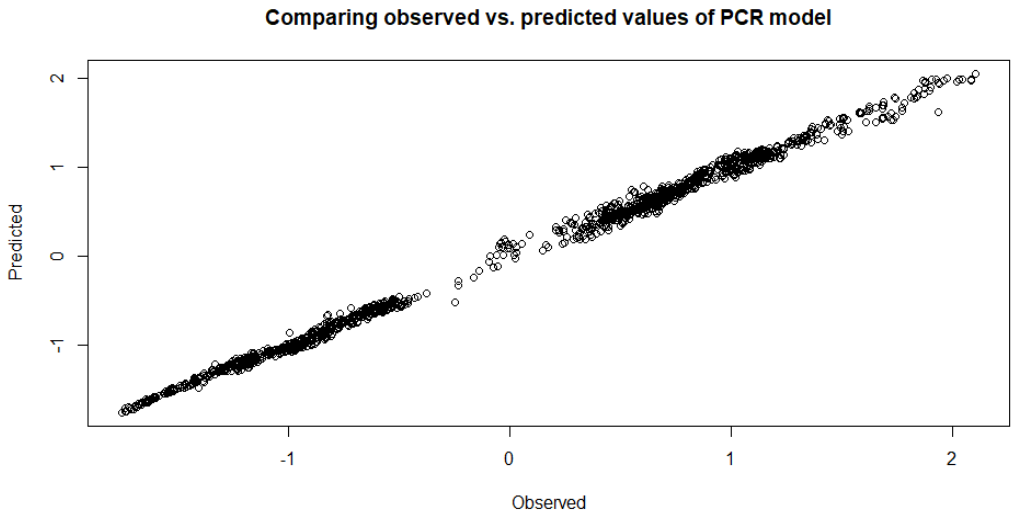


Figure 7. Observed EGX 30 Vs. predicted values by the PCR.

5. Conclusion

The Egyptian Stock Exchange is investigated in this paper to determine the most contributing stocks. To avoid periods of instability, the behavior of EGX 30 index is investigated between 2014 and 2019. There are 20 individual stocks continuously constituted in the index. Principal Component Analysis (PCA) is used to reduce data dimensionality. The results show that the first three Principal Components (PCs) are sufficient to explain 83% of data variability. A Principal Component Regression (PCR) model was built to predict the future prices for the EGX 30. PCR is a regression model built for the reduced dimension set of variables. Results of the Cross Validation (CV) for the PCR show that the first two PCs are sufficient to capture 97% of data variability. Also, the comparison between expected and observed values of the EGX 30 proved that the PCR model is performing fairly well (R-squared estimated as 0.98).

6. References

Burgette, L. F., and J. P. Reiter. 2010. "Multiple imputation for missing data via sequential regression trees." American Journal of Epidemiology 172 (9): 1070–1076. doi:https://doi.org/10.1093/aje/kwq260.

Cao, J., and J. Wang. 2020. "Exploration of stock index change prediction model based on the combination of principal component analysis and artificial neural network." Soft Comput 24: 7851–7860. doi:https://doi.org/10.1007/s00500-019-03918-3.

Ghorbani, M., and E. Chong. 2020. "Stock price prediction using principal component." PloS one 15 (3): 1 - 20. doi:https://doi.org/10.1371/journal.pone.0230124.

Hargreaves, C. A. 2019. "An automated stock investment system using machine learning techniques: An application in Australia." World Academy of Science, Engineering and Technology International Journal of Mathematical and Computational Sciences 13 (10): 189 - 192.

Jolliffe, I. T. 2002. Principal Component Analysis, Second Edition. New York: Springer.

Kaiser, H. F. 1960. "The application of electronic computers to factor analysis." Educational and Psychological Measurement 20 (1): 141-151.

Waqar, M., H. Dawood, P. Guo, M. B. Shahnawaz, and M. A. Ghazanfar. 2017. "Prediction of Stock Market by Principal Component Analysis." 2017 13th International Conference on Computational Intelligence and Security (CIS). Hong Kong, China: IEEE. 599 - 602. doi:10.1109/CIS.2017.00139.

Zhang, H. L. 2018. "The forecasting model of stock price based on PCA and BP neural network." Journal of Financial Risk Management 7: 369 - 385. doi:10.4236/jfrm.2018.74021 .

Zhong, X., and D. Enke. 2019. "Predicting the daily return direction of the stock market using hybrid machine learning algorithms." Financ Innov 24 (5): 1 - 20. doi:https://doi.org/10.1186/s40854-019-0138-0.