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 Analysis for the Egyptian Economic Growth Under the Government's Vision 2030

Dr. Heba M. Ezzat

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

ABSTRACT

Under the Vision 2030, Egyptian economy is planned to be balanced, knowledge-based, competitive, and diversified. A knowledge-based economy depends on knowledge, information, and high skills that substantially contribute to economic growth and innovation in services in advanced economies. Economies characterized by knowledge-based activities support a significant share of the GDP growth. Thereby, identifying components of knowledge-based economy that could help approaching the Vision 2030 is very important. In this paper, six pillars for knowledge-based economy were selected: education, innovation, information and communication technology, development, employment, and human capital. As each pillar contains many indicators, we follow the Principal Component Analysis (PCA) to reduce predictors' dimensions. Results of the PCA show that, out of 35 variables, only 22 are highly affecting the GDP. The first principal component explains about 88% of the variance. Principal Component Regression (PCR) is built to predict the effect of these indicators on the GDP.

ARTICLE INFO

Article history:

Received 29 /11/2020

Accepted 30/07/2021

Keywords:

Education, innovation, employment, information and communication technology, principal component analysis

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

The predictive performance of the PCR model is assessed following the cross-validation technique. Results reveal that, the minimal Mean Squared Error of Prediction (MSEP) could be reached at the first PC. Additionally, the PCR model explains most of the variability of the response data around its mean (R-squared estimated as 0.99).

Introduction

Information and technology play significant role in knowledge-based economy. Vital factors like education and innovation would hugely contribute to smooth transformation to a knowledge-based economy. Knowledge workers would add to economy more than illiterate ones. Information and Communication Technology (ICT) is very important as they would enhance electronic payment systems and any economic electronic transaction (Pal, De', & Herath, 2020). Additionally, ICT supports different activities that eventually will add to the economy, such as e-learning and e-government (Zhang & Nunamaker, 2003). Moreover, economies providing gender-equally labor opportunities would perform better than others concentrating on male labors (Maceira, 2017). Different sources of income such as, remittances and investment would contribute to development which in turn will enhance the knowledge-based economy (Meyer & Shera, 2017). Finally, health is a crucial factor that supports transforming to a knowledge-based economy (Širá, Vavrek, Vozárová, & Kotuliˇ, 2020).

Thereby, this paper aims to investigate the contribution of these features to the current Egyptian economy. Many variables are considered as indicators for these features. To explore such high dimensionality set of variables, a dimensionality reduction algorithm is needed. Also, variables included in this study possess multicollinearity characteristic. High dimensionality and multicollinearity impose

Principal Component Analysis (PCA) to be the most suitable exploratory method that fits the purposes of this study. Despite the major benefits of the PCA, there are some limitations. For instance, choosing the proper number of the PCs to be included. Also, PCA (and all dimensional reduction algorithms) are data dependent. However, PCA is preferred due to its extreme low computational cost (Jolliffe, 2002).

To provide a prediction model, Principal Component Regression (PCR) model is applied on the resulted principal components (PCs) from the PCA. These components are used instead of the original predictors. This will address the multicollinearity and reduce the original data set to a new lower dimension one.

The structure of the paper is depicted as follows. In section 2, a detailed description of the data set is provided. In machine learning algorithms, preprocessing of the data is very crucial and it is a basic step at the very beginning. Thereafter, treating missed data is explained in this section. Section 3 presents the unsupervised learning method PCA and explains its steps. Section 4 is devoted to display the results and analyses of implementing the PCA. Additionally, the results of the PCR model and its accuracy in prediction are presented in this section. Finally, Section 5 concludes the paper.

2. Data description and preprocessing

Many studies suggested different features and indicators to measure the transformation to a knowledge-based economy (Asongu, Tchamyou, & Acha-Anyi, 2017; Tchamyou, 2015). This research focuses on the following six features: (i) education, (ii) innovation, (iii) information and communication technology, (iv) development, (v) employment, and (vi) human capital. Annual data comprise 36

variables from 1987 to 2018 were collected from the World Bank database. Table 1 summarizes the indicators according to the selected components of a knowledge-based economy (Asongu, Tchamyou, & Acha-Anyi, 2017; Tchamyou, 2015). Table 1 contains the 35 variables under investigation to study their effects on the GDP, which will be considered as an indicator for the economy ('ECO').

Before deleting missing data records, Missing Completely At Random (MCAR) test was run (Hawkins, 1981; Jamshidian & Jalal, 2010). The results of Hawkins test could be summarized as follows. The p-value for the Hawkins test of normality and homoscedasticity is1.52e-21. This implies 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. Knowledge-based economy indicators.

Pillar	Indicator	Symbol	Missing data (%)
Education	Education index	EDI	9
	School enrollment, primary (% gross) WB	$EDU1$	12
	School enrollment, primary, female (% gross)	$EDU2$	9
	School enrollment, primary, male (% gross)	$EDU3$	9
	Out-of-school children of primary school age, both sexes (number)	$EDU4$	6

Innovation	Patent applications, residents	$INN1$	6
	Patent applications, nonresidents	$INN2$	0
Information & communication technology	Mobile cellular subscriptions (per 100 people)	$ICT1$	0
	Fixed telephone subscriptions (per 100 people)	$ICT2$	0
	Communications, computer, etc. (% of service exports, BoP)	$ICT3$	0
	Communications, computer, etc. (% of service imports, BoP)	$ICT4$	0
	Computer, communications and other services (% of commercial service exports)	$ICT5$	0
	Computer, communications and other services (% of commercial service imports)	$ICT6$	0
Development	Foreign direct investment, net outflows (% of GDP)	$DEV1$	0
	Foreign direct investment, net inflows (% of GDP)	$DEV2$	0
	Primary income on FDI (current US$)	$DEV3$	0
	Current account balance (% of GDP)	$DEV4$	0
	Trade in services (% of GDP)	$DEV5$	6
	Personal remittances, paid	$DEV6$	0

	(current US$)		
	Personal remittances, received (% of GDP)	$DEV7$	0
	Total reserves (includes gold, current US$)	$DEV8$	0
	Total reserves minus gold (current US$)	$DEV9$	0
	Total reserves (% of total external debt)	$DEV10$	9
Employment	Employers, total (% of total employment) (modeled ILO estimate)	$EMP1$	9
	Vulnerable employment, total (% of total employment) (modeled ILO estimate)	$EMP2$	9
	Wage and salaried workers, total (% of total employment) (modeled ILO estimate)	$EMP3$	9
	Labor force, female (% of total labor force)	$EMP4$	9
	Unemployment, total (% of total labor force) (modeled ILO estimate)	$EMP5$	9
	Labor force participation rate, total (% of total population ages 15-64) (modeled ILO estimate)	$EMP6$	9
	Employment in agriculture (% of total employment) (modeled ILO estimate)	$EMP7$	9

	Employment in industry (% of total employment) (modeled ILO estimate)	$EMP8$	9
	Employment in services (% of total employment) (modeled ILO estimate)	$EMP9$	6
Human capital	Human Development Index (HDI)	$HCPI$	9
	Survival to age 65, female (% of cohort)	$HCP1$	6
	Survival to age 65, male (% of cohort)	$HCP2$	9
	Mortality rate, infant (per 1,000 live births)	$HCP3$	0

Thereby, deleting missed values may 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 & Reiter, 2010). It is flexible enough to fit interactions, nonlinear relations, and complex distributions without parametric assumptions or data transformations. The algorithm was run for 1000 times and the median of the resulted 1000 imputed numbers was used for replacement. Figure 1 shows the values of the variables under investigation after implementing multiple imputation.

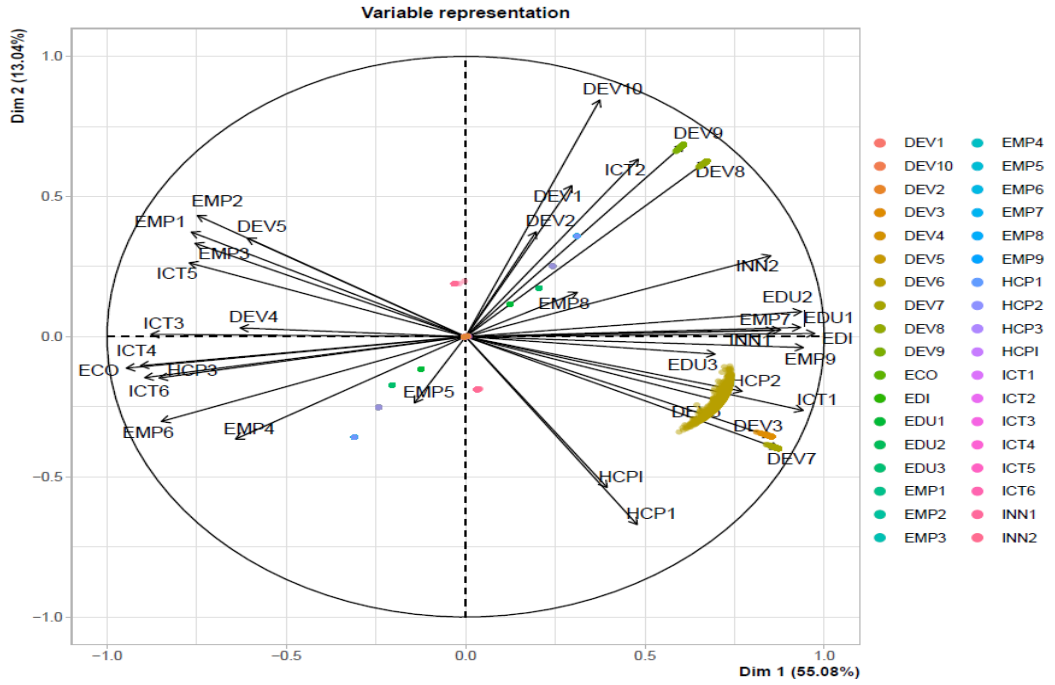

Figure 1. Results of implementing PCA multiple imputation to replace missing values.

3. Methodology

To identify important variables affecting Egypt's transformation to knowledge-based economy, Principal Component Analysis (PCA) is conducted. PCA is an unsupervised learning technique that uses orthogonal transformation to induce a set of observations of probably correlated variables into a set of linearly uncorrelated variables called principal components (Jolliffe, 2002). PCA is very useful to explore large set of variables especially when multicollinearity exists between the variables.

PCA is implemented for dimension reduction purposes. The first principal component captures the largest possible variability in the data. Each successive component has the highest variance possible given that it is orthogonal to the rest of the components. The resulting linear combinations of the variables form uncorrelated orthogonal basis

set. PCA is sensitive to the relative scaling of the original variables. To avoid this problem, PCA is applied on standardized data so no specific variables would dominate the results.

To investigate the existence of multicollinearity between the model variables, we calculated the correlation matrix of the 35 variables at 0.01 significance level. Figure 2 shows the correlation plot. The correlations' colors are scaled according to the direction and magnitude of correlation coefficients. It is important to notice that modeling such a set of variables using multiple regression would reveal biased results due to the high multicollinearity. The PCA is considered as a suitable exploratory method to address multicollinearity.

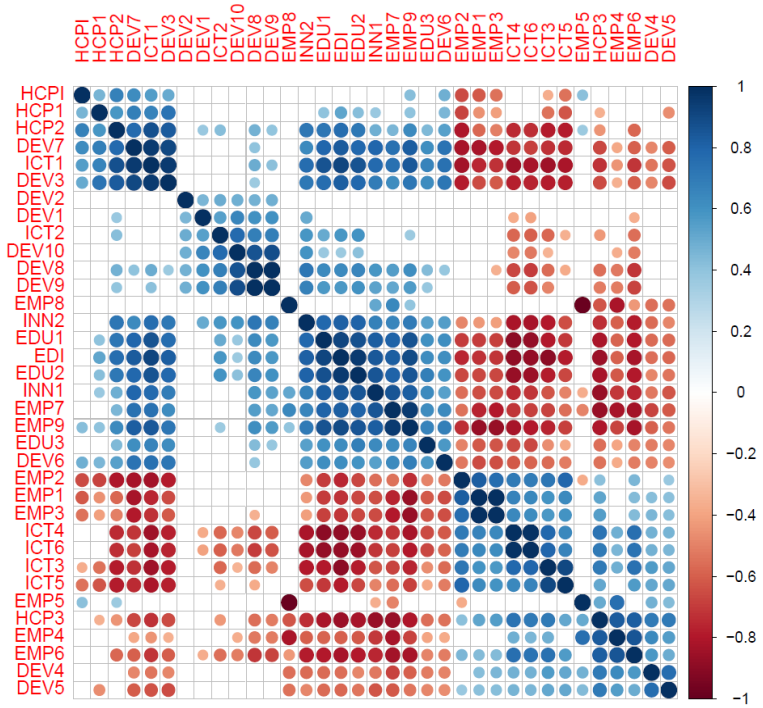

Figure 2. Correlation estimates at 0.01 significance level.

If $x_1, x_2, \dots, x_n$ represent the original variables under study, then PCA aims to project the data to a new direction that explain the most variability of these data such that;

$$\begin{aligned} PC_1 &= \theta_{11}x_1 + \theta_{12}x_2 + \theta_{13}x_3 + \cdots + \theta_{1n}x_n \\ PC_2 &= \theta_{21}x_1 + \theta_{22}x_2 + \theta_{23}x_3 + \cdots + \theta_{2n}x_n \\ &\vdots \\ PC_m &= \theta_{m1}x_1 + \theta_{m2}x_2 + \theta_{m3}x_3 + \cdots + \theta_{mn}x_n \\ &\vdots \\ PC_n &= \theta_{n1}x_1 + \theta_{n2}x_2 + \theta_{n3}x_3 + \cdots + \theta_{nn}x_n \end{aligned} \quad (1)$$

Some data sets could be explained by the first few Principal Components (PCs). Other may require more PCs to be explained though the number of PCs (say m) would be much less than the original number of variables; n. To sum up, PCA measures the association between predictor variables using a correlation matrix. This method aims to understand the directions of the spread of the data using eigenvectors and bringing out the relative importance of these directions using eigenvalues.

To identify the significance of variables under study and their effect on the Egyptian Knowledge Economy, PCA is implemented on the selected 35 variables. For this objective, R is used to run the algorithm and find the results. After signifying the effective variables and their contributions, Principal Component Regression (PCR) is applied to provide a prediction model for the GDP. The results of applying PCA and PCR are presented in the following section.

4. Results and Analyses

PCA provides linear combinations of important variables explaining the most variance in the data. By the end of this research variables representing the knowledge-based economy would be reduced from 35 variables to a limited number of PCs. Figure 3 illustrates the biplot of PC_1 and PC_2. The transparency of each vector depends on the contribution of predictor variable to dimensions or PCs.

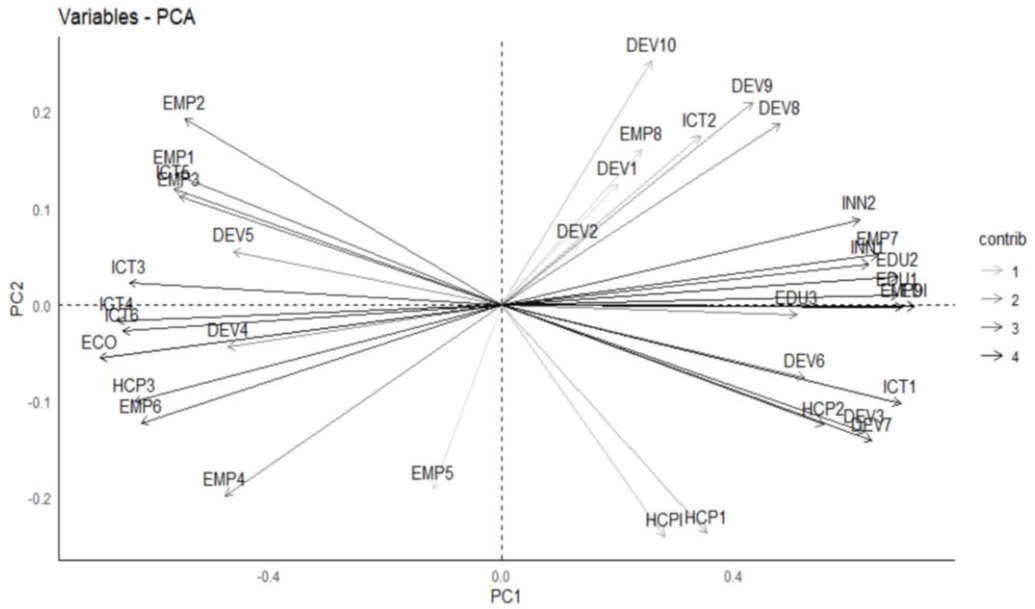

Figure 3. PCA first & second dimensions, vector's color transparency depends on the contribution of variables to each dimension.

Now, it is very important to determine the number of PCs that maximize the variance. For this step, three approaches were followed throughout the literature. First, the percentage of explained variance could be determined in advance of the analysis. For example, some researchers may take the first PCs that explain at least 70% of data variability. Second, a scree plot may provide an indication for the number of PCs that could be utilized (see Figure 4). Notice the elbow

shape in Figure 4. Also note that after PC_1 the percentage of explained variance is rapidly dimensioning.

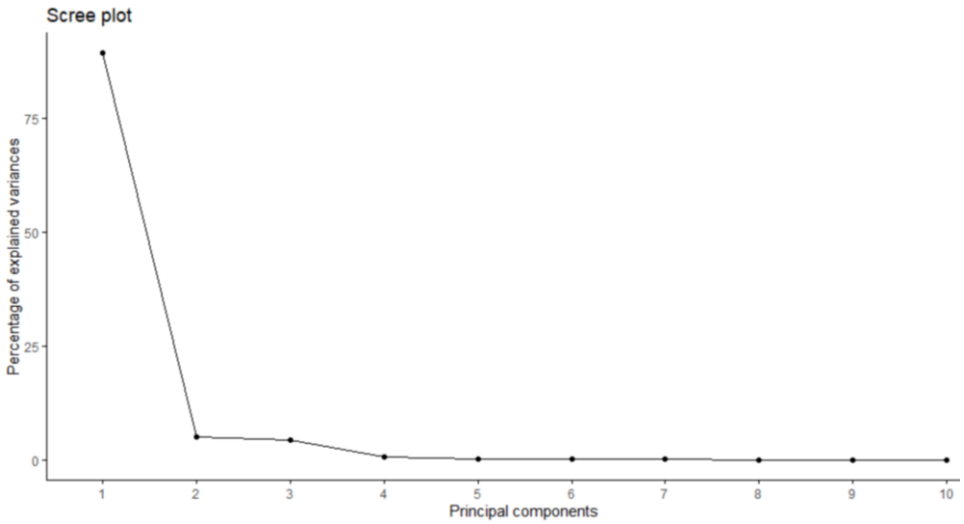

Figure 4. Scree plot depicting percentage of variances explained by each principal component.

Third, the number of PCs is determined according to the eigenvalues. In this case, eigenvalue measures how much variance exists in data in that direction. Any PC with an eigenvalue greater than or equal to one explains more variance than a single observed variable. The first two approaches are subjective. The third approach relies on reported eigenvalues calculated from the data. Following the third approach implies that the first eigenvalue is greater than one (as reported in Table 2). Thereby, the first PC would be considered, and this would explain about 88% of variance.

Table 2. Eigen values for the first five PCs.

	eigenvalue	percentage of variance	cumulative percentage of variance
PC 1	9.77	88.16	88.16
PC 2	0.62	5.57	93.73
PC 3	0.54	4.85	98.57
PC 4	0.07	0.63	99.21
PC 5	0.03	0.27	99.48

Table 3 presents the first eigenvector corresponding to the first eigenvalue. The element θ_{ni} represents the loading or the weight of each variable in the corresponding PC. These loadings represent the importance of each variable to the respective PC. If all variables are equally contributing to the PC, then the weights should be $\frac{1}{\sqrt{35}} = 0.169$. However, we can notice from Table 3 that each variable contributes differently. The variables with higher weights (≥ 0.169) are highlighted with gray. Accordingly, education enrollment, innovation, cellular and computing technology, foreign currency income, employment in services, agriculture and vulnerable activities, survival and life expectancy are considered the most effective features on the knowledge economy.

Table 3. Eigenvector corresponding to the first PC.

	PC_1	
Pillar	Indicator	weight
Education	EDI	0.22
	EDU1	0.21
	EDU2	0.21
	EDU3	0.16
Innovation	INN1	0.20
	INN2	0.19
Information & Communication Technology	ICT1	0.22
	ICT4	-0.21
	ICT3	-0.20
	ICT6	-0.20
	ICT5	-0.18
	ICT2	0.11
Development	DEV3	0.20
	DEV7	0.20
	DEV6	0.16
	DEV4	-0.15
	DEV5	-0.15
	DEV8	0.15
	DEV9	0.13

	DEV10	0.08
	DEV1	0.06
	DEV2	0.04
Employment	EMP9	0.22
	EMP7	0.20
	EMP6	-0.19
	EMP1	-0.18
	EMP3	-0.18
	EMP2	-0.17
	EMP4	-0.15
	EMP8	0.08
	EMP5	-0.04
Human capital	HCP3	-0.20
	HCP2	0.18
	HCP1	0.11
	HCPI	0.09

Now, let us consider the results of running a Principal Component Regression (PCR). This model refers to running regression analysis by regressing the dependent variable 'ECO' on the PCs of the explanatory variables. Again, this method is very useful to reduce dimensions and to address the multicollinearity. The PCs with higher variances will be selected as the model regressors. PCR could lead to efficient prediction of the dependent variable.

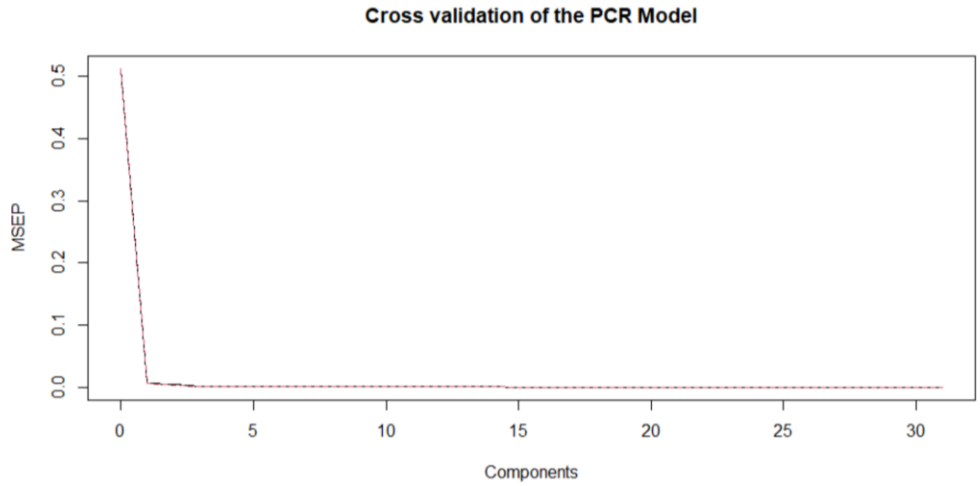

Figure 5. Cross-validation method to identify the best model.

Figure 5 shows the results of cross-validation method followed to determine the best model. The figure depicts that, the minimal Mean Squared Error of Prediction (MSEP) could be reached at the first PC. Therefore, the first principal is sufficient to be used for predicting the GDP. The final PCR model in terms of the original predictors could be displayed in (2) as follows.

$$\begin{aligned} ECO = {} & .05EDI + .05EDU1 + .05EDU2 + .02EDU3 + .05INN1 \\ & + .06INN2 + .04ICT1 + .04ICT2 - .05ICT3 - .05ICT4 \\ & - .03ICT5 - .05ICT6 + .01DEV1 - .02DEV2 \\ & + .05DEV3 - .05DEV4 - .05DEV5 + .02DEV6 \\ & + .03DEV7 + .03DEV8 + .03DEV9 + .03DEV10 \\ & - .01EMP2 - .07EMP4 - .04EMP5 - .06EMP6 \\ & + .04EMP7 + .04EMP8 + .04EMP9 - .03HCPI \\ & + .03HCP1 + .03HCP2 - .07HCP3 \end{aligned}$$

Figure 6 shows graphs the observed GDP values against the ones predicted by the PCR model. From this figure we can conclude that the PCR performs very well in prediction.

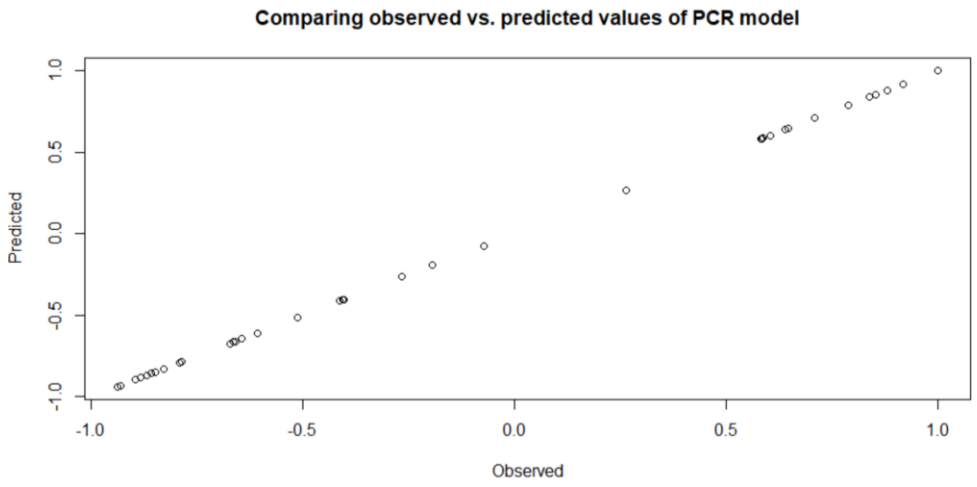

Figure 6. Observed GDP Vs. predicted values by the PCR.

5. Conclusion

Transformation to knowledge-based economy becomes an urgent necessity in the technology era. Identifying important factors affecting this transformation is very crucial. Six pillars are considered in this article; (i) education, (ii) innovation, (iii) information and communication technology, (iv) development, (v) employment, and (vi) human capital. Each aspect contains many variables to investigate their effects on the GDP as indicator for the economy. A total of 35 variables is considered. Thereafter, we followed a dimensionality reduction method, Principal Component Analysis (PCA), to address multicollinearity and highlight important variables. PCA aims to find directions that explain most variance of high-dimensional data. This method would project the data to a new subspace with equal or fewer dimensions than the original data set.

The results reveal that the first Principal Component (PC) could explain about 88% of data variability. Important factors affecting the transformation are education enrollment, innovation, cellular and computing technology, foreign currency income, employment in

services, agriculture and vulnerable activities, survival and life expectancy. We, also, follow the Principal Component Regression (PCR) to regress the GDP on the orthogonal PCs instead of the original collinear predictors. The results yield that the first PC is sufficient to be used for prediction with the minimal Mean Squared Error of Prediction (MSEP).

Another important viewpoint is that some variables that contribute less to the GDP should be given more focus from the government. For example, female labors and industry employment should be provided more support and focus. This also applies on income from services and income from direct foreign investment.

For future research, artificial intelligence models could be applied on the PCA results to predict future performance of the GDP.

References

Amirat, A., & Zaidi, M. (2020). Estimating GDP growth in Saudi Arabia under the government's Vision 2030: A knowledge-based economy approach. Journal of the Knowledge Economy, 11, 1145–1170. doi:https://doi.org/10.1007/s13132-019-00596-2

Aqil, M., Aziz, S. F., Dilshad, M., & Qadeer, S. (2014). Relationship between public education expenditures and economic growth of Pakistan. IOSR Journal of Humanities and Social Science, 19(3), 153–155.

Asongu, S., Tchamyou, V., & Acha-Anyi, P. (2017). Who is Who in Knowledge Economy in Africa? MPRA Paper No. 84043, 1-43.

Barro, R. (2001). Human capital and growth. American Economic Review, 91, 12 - 17.

Ben Hassen, T. (2020). The state of the knowledge-based economy in the Arab world: cases of Qatar and Lebanon. EuroMed Journal of Business, ahead-of-print (ahead-of-print). doi:https://doi.org/10.1108/EMJB-03-2020-0026

Broersma, L., McGuckin, R. H., & Timmer, M. P. (2001). The impact of computers on productivity in the trade sector: Explorations with Dutch microdata. De Economist, 151(1), 53–79.

Burgette, L. F., & Reiter, J. P. (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

Carayannis, E. G., & Campbell, D. F. (2012). Mode 3 knowledge production in quadruple helix innovation systems. New York: Springer.

Hawkins, D. M. (1981). A new test for multivariate normality and homoscedasticity. Technometrics, 23, 105-110, 23, 105-110.

Jamshidian, M., & Jalal, S. (2010). Jamshidian, M. andTests of homoscedasticity, normality, and missing at random for incomplete multivariate data. Psychometrika, 75, 649-674.

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

Khorsheed, M. S. (2015). Saudi Arabia: From oil kingdom to knowledge-based economy. Middle East Policy, 22(3), 147 - 157.

Kurtić, A., & Đonlagić, S. (2012). Determining key factors for knowledge economy development in Bosnia and Hercegovina. Management, Knowledge and Learning International Conference (pp. 413 - 421). Celje, Slovenia: International School for Social and Business Studies.

Maceira, H. M. (2017). Economic Benefits of Gender Equality in the EU. Intereconomics, 52(3), 178–183.

Meyer, D., & Shera, A. (2017). The impact of remittances on economic growth: An econometric model. EconomiA, 18(2), 147-155.

Pal, A., De', R., & Herath, T. (2020). The role of mobile payment technology in sustainable and human-centric development: Evidence from the post-demonetization period in India. Information Systems Frontiers, 22, 607–631.

Širá, E., Vavrek, R., Vozárová, I. K., & Kotuliˇ, R. (2020). Knowledge economy indicators and their impact on the sustainable competitiveness of the EU countries. Sustainability, 4172(12), 1 - 22.

Tchamyou, V. S. (2015). The role of knowledge economy in African. AGDI Working Paper, No. WP/15/049, African Governance and Development Institute, 1-51.

Zhang, D., & Nunamaker, J. F. (2003). Powering E-Learning In the New Millennium: An Overview of E-Learning and Enabling Technology. Information Systems Frontiers, 5(2), 207–218.