PCA with Economic, and Energy Indicator Variables

Math 585 - Final Project

Nate Islip

Draft Due Date: 03/18/2021

Table of Contents:

1 Introduction:

Concerns regarding climate change in the United States and worldwide have grown exponentially due to rising sea levels, melting glaciers, and rising temperatures [1]. These concerns have ultimately led to federal and local mandates regulating greenhouse gas (GHG) production among states and businesses and tax incentive programs incentivizing firms to invest in clean energy projects to reduce emissions [2]. The concern is not only with regards to environmental sustainability but with resource sustainability. Since fossil fuels are finite, it is not a matter of if we run out of fossil fuels but when.

With climate change and its impacts on environmental and economic sustainability clear, many have researched the phenomenon with different countries and economic indicators. Xiucheng Dong et al. propose a sustainability development theory as a function of resource, market, enterprise, and technology and policy. Furthermore, Xiucheng Dong et al. institute the natural industry sustainability index and evaluate it using the Principal component analysis [2]. Their study identifies economic indicators and fuel measurements such as reserve replacement ratio, production, imports, exports, and net profits. At the same time, this study provides critical insight into the sustainability analysis and its impact on policy Thai-Ha Le et al. elaborate on the nexus between renewable and nonrenewable energy sources, a factor not mentioned in Xiucheng Dong et al. and many other papers. According to Thai-Ha le et al., this is mainly due to the lack of data for multiple countries. In their study, fixed effects are dependent on Real GDP and an indicator of emissions (EMI) with explanatory variables governance indicator, real gross fixed capital, total labor force, and renewable energy consumption. Their findings are consistent with other papers and previously made assumptions. This paper seeks to combine the use of macroeconomic indicators, and energy usage data to examine the correlations between these variables and observe whether there is a relationship between energy consumption, in both renewable and nonrenewable energy, and economic well-being.

1.1 Problem Statement:

This paper will examine the energy transition from fossil fuels to renewable energy in the U.S. in-terms of economic growth. Moreover, we will first examine the relationship between the features of the data, and reduce the dimension of the data frame to a smaller number of principal components that explain a percentage of the variation. **Moreover, I intend to see if there is a correlation between energy emissions, and consumption and economic growth.**

The main challenge facing the question of "how will the transition from fossil fuels, to renewable energy impact our economic growth" would be data availability. This is a massive question, which requires a ton of historical, and real-time data. I was only able to capture aggregated measures for economic growth, and energy indicators in renewable energy, and natural gas. This means many features that may be more important than the current features included in this study, were not included. Therefore, given more access to individual gas, and renewable energy types may lead to specific relationships between fuel types, and economic growth indicators such as GDP. The second challenge, which is out of the scope of this paper, would be predicting future energy costs, and production is very difficult due to the complex dynamics represented in our economy. For instance, we cannot control for when pandemics may occur, and how to how severely they might impact demand, and supply.

What I find interesting in the context of this problem is the future implications of finding patterns in the data relating the transition from renewable energy to fossil fuels, and the impact this may have on economic growth. For instance, as we transition to more sustainable energy production methods, many aspects of the economy will shift, as energy (in a general sense) is how the economy functions. For instance, the labor market could shift drastically as the depletion of finite fossil fuels decreases, may lead to fewer jobs, or the implementation of renewable energy could introduce more jobs.

1.2 Data:

This paper's data is predominantly from the United States Energy Information Administration (EIA), Federal Reserve Economic Data (FRED), and World Bank [3, 6]. Data for economic indicators are from FRED and the World Bank, while energy measurements such as emissions, energy imports exports, and prices are from the EIA. There are 131 observations from November 2010 - January 2020. I chose to work with a monthly data due to limitations regarding data availability. Since most measurements pertaining to emissions were not recorded until the 2000's, there would be very few observations.

Below is a table containing 9 variable descriptions, means and standard deviations. Note, RE stands for renewable energy, NG stands for natural gas, GDP is gross domestic product, and NGPL is natural gas plant liquid. Furthermore, renewable energy encompasses all sources of renewable energy such as wind, solar, and hydro power, while natural gas is a single source of fossil fuels, however, natural gas accounts for a majority of fossil fuels as it contains methane, natural gas liquids, carbon dioxide, and water vapor. This fuel type is also one of the main sources of electricity production in the United States.

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Total Consumption (RE)	131	0.839	0.103	0.617	0.766	0.915	1.059
Total Production (RE)	131	846.217	106.486	622.652	769.046	925.649	1,065.092
Total Consumption (NG) Total	131	2,269.020	435.747	1,617.274	1,922.718	2,549.884	3,417.203
NGPL Production	131	145.968	44.846	80.179	106.405	177.056	240.773
Total Imports (NG)	131	249.029	37.129	174.461	220.519	273.173	384.586
Total Exports (NG)	131	210.776	113.027	75.938	127.023	283.751	526.513
Equity Market Volatility (%)	131	0.332	0.252	0.000	0.164	0.458	1.883
GDP Normalized	131	99.583	1.595	90.654	99.748	100.180	100.782
Total Unemployment %	131	11.851	3.572	7	8.6	14.7	23

Figure 2 is a correlation matrix of all the variables in the data set. This step is necessary to identify any initial variables that may be highly correlated with one another, and should be dropped. Any value above 0.8, could be considered highly correlated. The two variables that are highly correlated with one-another are, **total consumption of (RE)**, and **total renewable energy production**. This makes sense since, consumption and production have a symbiotic relationship. As consumption increases, so will production. There is a similar relationship occurring between **natural gas exports** and **natural gas imports**. We can fix this by removing 2 of the 4 variables, one from each fuel type.

The first step in analyzing this data would be to observe the correlation matrix for the 9 variables to see if any are highly correlated. This can be seen in figure

Figure 1: Correlation Matrix for 9 variables, prior to reduction

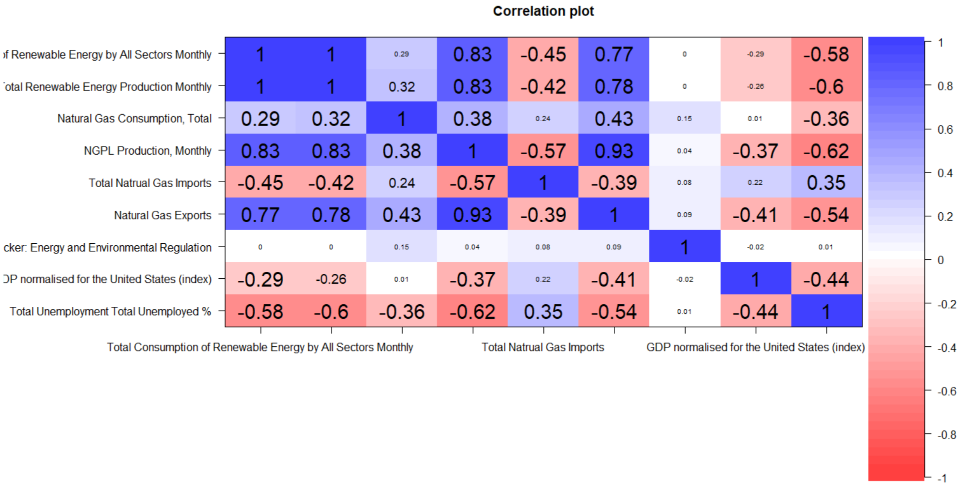
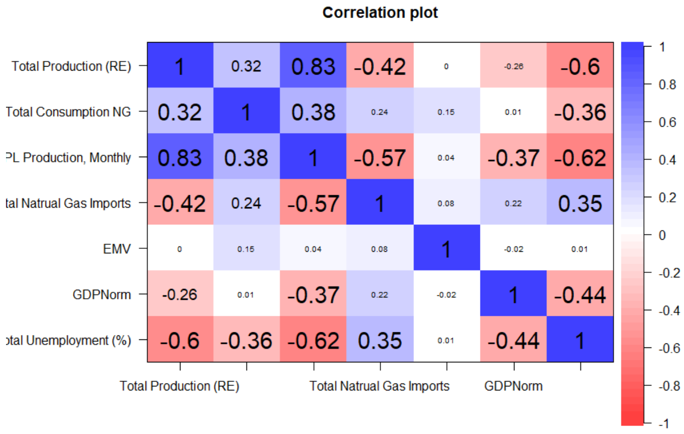


Figure 2: Correlation Matrix for 7 variables, after reduction



2 Methodology:

2.1 Theory:

Before moving forward with computations, I will present the theory used in the textbook **Applied Multivariate Statistical Analysis** [4]. In the context of this problem, we will be taking the data set containing both energy and economic variables and interpreting the results, and reducing the data to fewer components that contribute more variation than if the data includes all p variables. First, consider the following structure of the data set. Suppose the data $\mathbf{x}_1,\ \mathbf{x}_2,\ \ldots,\mathbf{x}_n$ represent n independent drawings from some p-dimensional population with mean vector μ and covariance matrix Σ. These data yield the sample mean vector $\overline{\mathbf{x}}$, the sample covariance matrix $\mathbf{S}$, and the sample correlation matrix $\mathbf{R}$ (8.3.441).

2.1.1 Standardization:

The first step in our analysis is to standardize the variables in the matrix $\mathbf{R}$ (note: $\mathbf{R}$ is the correlation matrix, and since the variables are standardized, the choice of $\mathbf{S}$ or $\hat{\Sigma}$ are irrelevant). Standardizing the variables are extremely important since PCA is projecting data onto directions that maximize the variance, therefore we need to have unit variance. Furthermore, the variables are not of the same units (8.3.449) . This unit variance will be shown in section 2.2.3.

To first perform standardization we construct,

$$z_j = \mathbf{D}^{-1/2}(x_j - \overline{x}) = \begin{bmatrix} \frac{x_{j1}-\overline{x}_1}{\sqrt{s_{11}}} \\ \frac{x_{j2}-\overline{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \frac{x_{jp}-\overline{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \qquad j = 1, 2, \ldots, n \tag{1}$$

Then we are given the $n \times p$ matrix of standardized observations,

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & s_{22} & & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2}\cdots & z_{np} & \end{bmatrix} \tag{2}$$

and the sample covariance matrix $\mathbf{S_z} = \frac{1}{n-1}(\mathbf{Z} - \mathbf{1}\overline{z}')(\mathbf{Z} - \mathbf{1}\overline{\mathbf{z}})$ with sample mean vector $\overline{\mathbf{z}}$. Then,

$$\mathbf{R} = \frac{1}{n-1} \begin{bmatrix} \frac{(n-1)s_{11}}{s_{11}} & \frac{(n-1)s_{12}}{\sqrt{s_{11}\sqrt{s_{22}}}} & \cdots & \frac{(n-1)s_{1p}}{\sqrt{s_{11}\sqrt{s_{pp}}}} \\ \frac{(n-1)s_{12}}{\sqrt{s_{11}\sqrt{s_{22}}}} & \frac{(n-1)s_{22}}{s_{22}} & \cdots & \frac{(n-1)s_{2p}}{\sqrt{s_{22}\sqrt{s_{pp}}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{(n-1)s_{1p}}{\sqrt{s_{11}\sqrt{s_{pp}}}} & \frac{(n-1)s_{2p}}{\sqrt{s_{22}\sqrt{s_{pp}}}} & \vdots & \frac{(n-1)s_{pp}}{s_{pp}} \end{bmatrix} \tag{3}$$

2.1.2 Sample Principal Components:

Note, the sample principal components of the standardized observations (given above) are given by

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}} \qquad i, k = 1, 2, ..., p \tag{4}$$

in addition to sample variance $(\hat{y}_k) = \hat{\lambda}_k$, and sample covariance $(y_i, \hat{}\hat{y_k}) = 0$ with the matrix $\mathbf{R}$ in place of $\mathbf{S}$ [4].

Finally, the standardized observations $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ with correlation matrix $\mathbf{R}$ and eigenvalue-eigen vector pairs $(\hat{\lambda}_1,\ \hat{\mathbf{e}}_1),\ (\hat{\lambda}_2,\ \hat{\mathbf{e}}_2), \ldots,$ with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$ then the ith sample principal component is given by,

$$\hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{z} = \hat{e}_{i1}z_1 + \hat{e}_{i2}z_2 + \ldots + \hat{e}_{ip}z_p \tag{5}$$

Therefore, the sample principal components are denoted as $\hat{y}_1,\ \hat{y}_2,\ \ldots, \hat{y}_p$.

2.1.3 Equal Correlation Structure (future work):

In the textbook, the authors refer to Large Sample inferences and testing for the equal correlation Structure. While I do not compute this in the study, it can be left for future work. This could be used to determine the quality of the principal component approximation. Below is a brief introduction into the process if testing for equicorrelation structure. Using Lawley's procedure (8.5.456),

$$\overline{r}_k = \frac{1}{p-1}\sum_{i=1, i\neq k}^{p} r_{ik} \qquad k = 1, 2, \ldots p \qquad \overline{r} = \frac{2}{p(p-1)}\sum\sum_{i<k} r_{ik} \tag{6}$$

$$\hat{y} = \frac{(p-1)^2[1-(1-\overline{r})^2]}{p-(p-2)(1-\overline{r})^2} \tag{7}$$

The large sample approximate α-level test is to reject H_0 in favor of H_1 if

$$T = \frac{(n-1)}{(1-\overline{r})^2}\left[\sum\sum_{i<k}(r_{ik}-\overline{r})^2 - \hat{y}\sum_{k=1}^{p}(\overline{r}_k-\overline{r})^2\right] > \chi^2_{(p+1)(p-2)/2}(\alpha) \tag{8}$$

using the equations above and where $\chi^2_{(p+1)(p-2)/2}(\alpha)$ is the upper (100α)th percentile of a chi-square distribution with $(p+1)(p-2)/2$ degrees of freedom.

2.1.4 Interpretation of Sample PCs:

After computing the ith principal component, we analyze the Total (standardized) sample variance $= \text{tr}(\mathbf{R}) = p = \hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_p$ and,

$$r_{\hat{y}_i, z_k} = \hat{e}_{ik}\sqrt{\hat{\lambda}_i}$$

for $i,\ k = 1, 2, \ldots, p$. Furthermore, one of the most important aspects of the PCA is the total sample variance explained by the ith sample principal component, and is defined as

$$\text{Proportion (ith sample PC)} = \frac{\hat{\lambda}_i}{p} \qquad i = 1, 2, \ldots, p \tag{9}$$

Moreover, we can calculate the proportion of total population variance due to the kth principal component,

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \qquad k = 1, 2, ..., p \tag{10}$$

These techniques are used in the next section titled computation, and will be used to interpret how much variation each PC contributes.

2.2 Computations:

2.2.1 Eigenvalues, Eigenvectors and Scree Plot:

Below is a list of eigenvalues, and eigenvectors generated by taking the covariance of the standardized variables. The size of the eigenvalues are used to decide which principal components to keep and which ones to drop. For instance, the eigenvalues below are listed below from largest to smallest,

```
        {
        EEs <- Eigen(cor(s.df)) # Eigens for Correlation matrix of Standardized Obs
        EEva <- as.data.frame(EEs[["values"]])
        EEve <- as.data.frame(EEs[["vectors"]])
        }

```

these eigenvalues are used for the scree plot to determine the optimal number of principal components.

$$
\begin{aligned}
\hat{\lambda_1} &= 2.87278423 & \hat{\mathbf{e}}_1' &= \begin{bmatrix}-0.5321 & -0.2452 & -0.5645 & 0.3563 & -0.0147 & 0.1234 & 0.4423\end{bmatrix}\\
\hat{\lambda_2} &= 1.50533628 & \hat{\mathbf{e}}_2' &= \begin{bmatrix}0.04416 & -0.4423 & 0.1104 & -0.3820 & -0.1453 & -0.6585 & 0.4355\end{bmatrix}\\
\hat{\lambda_3} &= 1.22223848 & \hat{\mathbf{e}}_3' &= \begin{bmatrix}-0.0567 & -0.5015 & -0.0977 & -0.3889 & -0.5311 & 0.4588 & -0.3033\end{bmatrix}\\
\hat{\lambda_4} &= 0.89661960 & \hat{\mathbf{e}}_4' &= \begin{bmatrix}-0.0801 & -0.3712 & -0.0350 & -0.3409 & 0.8330 & 0.1880 & -0.0969\end{bmatrix}\\
\hat{\lambda_5} &= 0.33644401 & \hat{\mathbf{e}}_5' &= \begin{bmatrix}-0.6296 & 0.5199 & 0.0568 & -0.5690 & -0.0522 & -0.00054 & 0.0601\end{bmatrix}\\
\hat{\lambda_6} &= 0.13151772 & \hat{\mathbf{e}}_6' &= \begin{bmatrix}0.5538 & 0.2887 & -0.5514 & -0.3665 & 0.0054 & 0.2015 & 0.3619\end{bmatrix}\\
\hat{\lambda_7} &= 0.03505969 & \hat{\mathbf{e}}_7' &= \begin{bmatrix}-0.0464 & 0.0370 & -0.5926 & -0.0399 & 0.0153 & -0.5146 & -0.6154\end{bmatrix}
\end{aligned}
$$

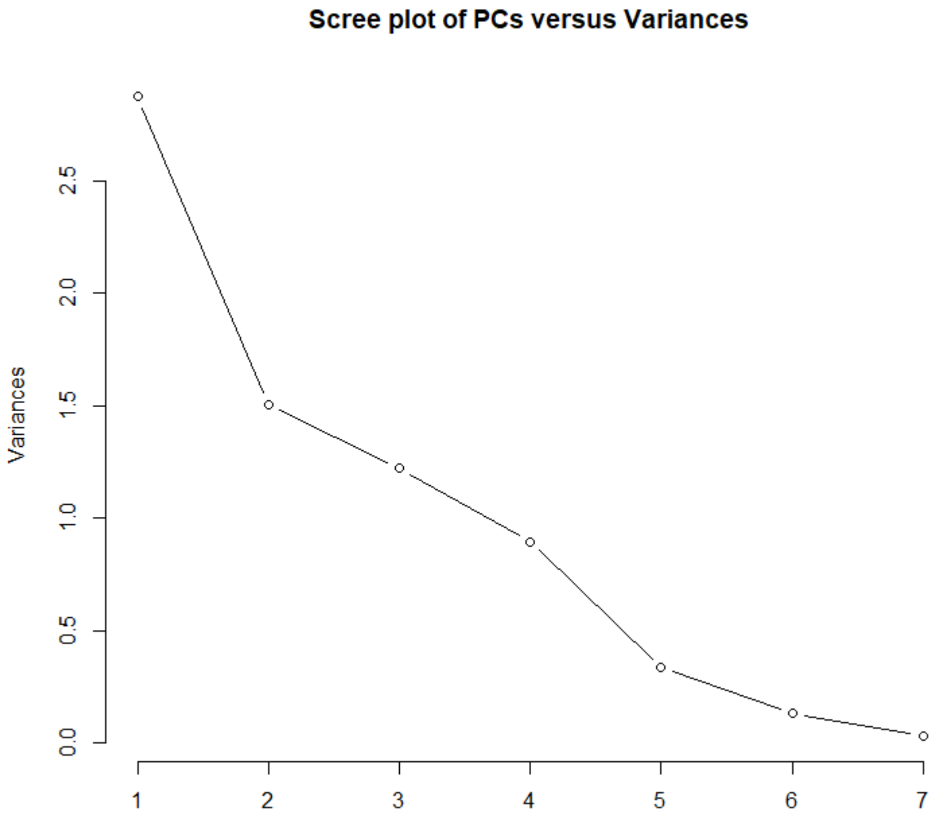
Furthermore, using the standardized variables, we obtain the the 7 sample principal components.

$$
\begin{aligned}
\hat{y}_1 &= \hat{\mathbf{e}}_1\mathbf{z} = -0.5321z_1 + -0.2452z_2 + -0.5645z_3 + 0.3563z_4 + -0.0147z_5 + 0.1234z_6 + 0.4423z_7\\
\hat{y}_2 &= \hat{\mathbf{e}}_2\mathbf{z} = 0.04416z_1 + -0.4423z_2 + 0.1104z_3 + -0.3820z_4 + -0.1453z_5 + -0.6585z_6 + 0.4355z_7\\
\hat{y}_3 &= \hat{\mathbf{e}}_3\mathbf{z} = -0.0567z_1 + -0.5015z_2 + -0.0977z_3 + -0.3889z_4 + -0.5311z_5 + 0.4588z_6 + -0.3033z_7\\
\hat{y}_4 &= \hat{\mathbf{e}}_4\mathbf{z} = -0.0801z_1 + -0.3712z_2 + -0.0350z_3 + -0.3409z_4 + 0.8330z_5 + 0.1880z_6 + -0.0969z_7\\
\hat{y}_5 &= \hat{\mathbf{e}}_5\mathbf{z} = -0.6296z_1 + 0.5199z_2 + 0.0568z_3 + -0.5690z_4 + -0.0522z_5 + -0.00054z_6 + 0.0601z_7\\
\hat{y}_6 &= \hat{\mathbf{e}}_6\mathbf{z} = 0.5538z_1 + 0.2887z - 2 + -0.5514z_3 + -0.3665z_4 + 0.0054z_5 + 0.2015z_6 + 0.3619z_7\\
\hat{y}_7 &= \hat{\mathbf{e}}_7\mathbf{z} = -0.0464z_1 + 0.0370z_2 + -0.5926z_3 + -0.0399z_4 + 0.0153z_5 + -0.5146z_6 + -0.6154z_7
\end{aligned}
$$

```
    {
    plot(pca)
    screeplot(pca, type="line", main = "Scree plot of PCs versus Variances")
    }

```

Figure 3: Scree Plot displaying the level of variation among PC's



2.2.2 PCA Results and Computation:

```
        {
        pca <- prcomp(s.df, scale=T)# PCA Analysis, Scaled for unit variance
        loadings_pca <- data.frame(pca[["rotation"]])
        PCs <- data.frame(PC = paste0("PC", 1:9), # Matrix of components and var
                          VAR = (pca$sdev)^2/sum((pca$sdev)^2))
        }

```

Table 2: Principal components and their proportion of variance

	PC	VAR
1	PC1	0.410
2	PC2	0.215
3	PC3	0.175
4	PC4	0.128
5	PC5	0.048
6	PC6	0.019
7	PC7	0.005

Below, in **table** 3, the principal components are listed from the largest to smallest, with the largest principal component accounting for the most amount of variation, and the second component accounting for the second most variation. Clearly, PC1 accounts for the most variation, 41.04%. PC2 accounts for a cumulative proportion of 62.54% and a proportion of variance of 21.50%. These two components contribute a large majority of the variation, however, PC3, and PC4 are not far from a similar proportion of variance with regards to PC2. According to the cumulative proportion, we could cut the number of components from 7, to either 5 or 6. Since the cumulative proportion reaches 97.62% at PC5, it may be irrelevant to include PC6. This interpretation could also be drawn from the scree plot in **figure** 3. The scree plot shows us a possible elbow at PC5, further supporting our conclusion to only keep 5 out of the 7, principal components.

Table 3: Summary of Principal Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.6949	1.2269	1.1055	0.9469	0.58004	0.36265	0.18724
Proportion of Variance	0.4104	0.2150	0.1746	0.1281	0.04806	0.01879	0.00501
Cumulative Proportion	0.4104	0.6254	0.8001	0.9281	0.97620	0.99499	1.00000

After calculating the proportion of variance, and cumulative proportion it may be helpful to observe the loading's of the PCA model. Loading's describe how much each variable contributes to a particular component. A large loading (positive or negative) indicates that the variable has a strong relationship to a specific principal component. Analyzing the loading's may provide more insight into variables that may or may not contribute to the data sets covariance, and correlation.

Table 4 shows the loading's for each of the principal components. Loading's with significant contributions, determined by their magnitude, have been highlighted for ease of interpretation. PC1 has large, positive loading's for Total Unemployment and Total Natural gas imports. Moreover, PC1 has large, negative loading's for Total renewable energy production and NGPL production. Next, PC2 has a strong positive loading for total unemployment, and strong negative loading's for total natural gas consumption, and total natural gas imports. Furthermore, PC3 has strong positive loading's for GDP normalised, and has strong negative loading's for NG consumption, NG imports, and EMV.

Table 4:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Total Production (RE)	**−0.532**	0.044	−0.057	−0.080	**0.630**	**−0.554**	0.046
Total Consumption NG	−0.245	**−0.442**	**−0.501**	**−0.371**	**−0.520**	−0.289	−0.037
NGPL Production, Monthly	**−0.565**	0.110	−0.098	−0.035	−0.057	**0.551**	**0.593**
Total Natrual Gas Imports	**0.356**	**−0.382**	**−0.389**	**−0.341**	**0.569**	**0.366**	0.040
EMV	−0.015	−0.145	**−0.531**	**0.833**	0.052	−0.005	−0.015
GDPNorm	0.123	**−0.658**	**0.459**	0.188	0.001	−0.202	**0.515**
Total Unemployment (%)	**0.442**	**0.435**	−0.303	−0.097	−0.060	**−0.362**	**0.615**

2.2.3 95% Confidence Ellipse Interpretation:

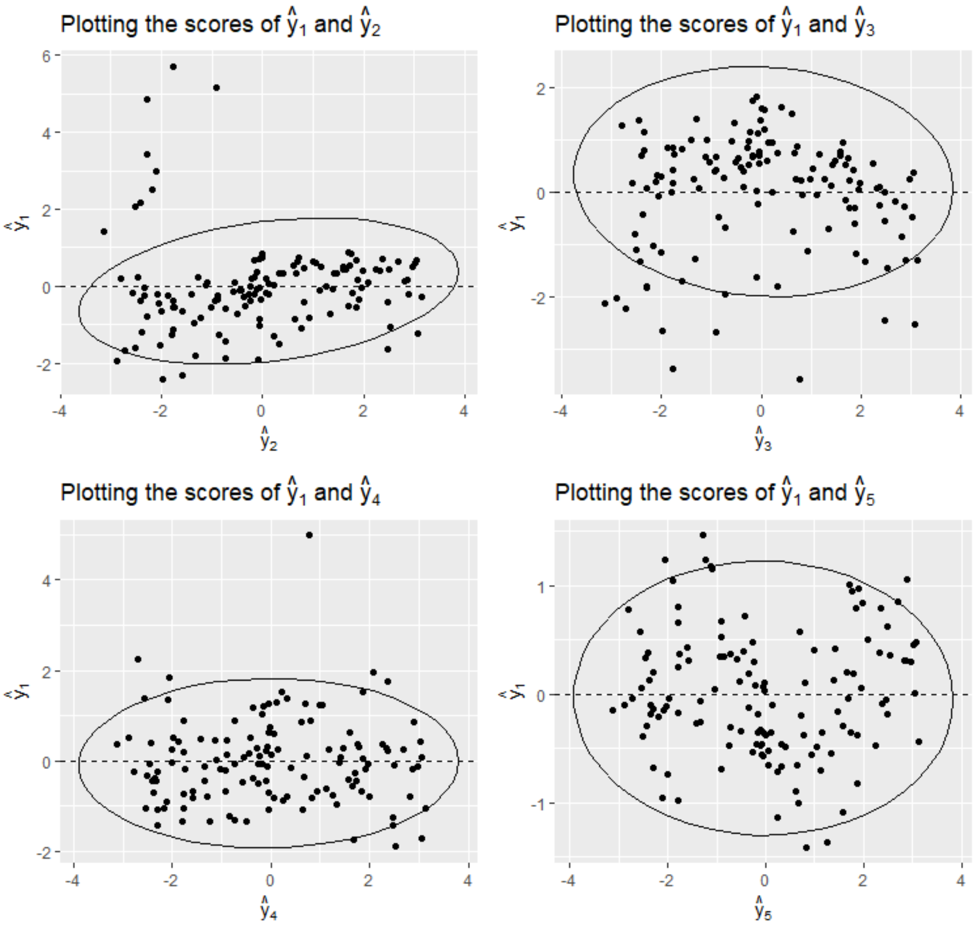
Generating the control ellipse for the principal components may help us with monitoring the quality of the principal components. To check if a process is stable, let us first define stability. If a process is stable over time, the the values of the first two principal components should be stable. To monitor quality, we can construct the ellipse format chart for $(\hat{y}_{j1}, \hat{y}_{j2})$ for $j = 1, 2, \ldots, n$. By the sample variance for $(\hat{y}_k) = \hat{\lambda}_k$,, the two sample components are uncorrelated, therefore the quality ellipse for $n = 131$ large reduces to the collection of possible values $(\hat{y}_1^2, \hat{y}_2)$ such that,

$$\frac{\hat{y}_1^2}{\hat{\lambda}_1} + \frac{\hat{y}_2^2}{\hat{\lambda}_2} \leq \chi_2^2(\alpha) \tag{11}$$

For the purpose of this paper, we have $n = 131$ large and use $\chi_2^2(.05) = 5.99$. The ellipse is centered at $(0, 0)$. This is due to the standardized variables [4]. Below in figure 4 is a matrix of control ellipses generated with all principal components up to PC5. For the purpose of PC1 and PC2 containing the most variation, we will look to interpret figure 4, 1 × 1.

The first figure shows multiple points out of control. In order to deal with the instability of these points, our first step would be to find the score, and its associated eigenvector. Most of these values are dominant in terms of $\hat{y}_2$, however, there are values in $\hat{y}_1$ that cause the instability. Our next step would be to identify where these values are, remove their period and recalculate the eigenvalues. Once removed, the dominance of the eigenvectors should be more stable, and the unstable points no longer outside the control ellipse.

Figure 4: 95% control ellipse for $\hat{y}_1$, $\hat{y}_2$, $\hat{y}_3$, $\hat{y}_4$, $\hat{y}_5$



2.2.4 Biplots:

Below is a 2 × 2 matrix of Bi-plots. Bi-plots are extremely important in Principal component analysis because they provide a lot of information regarding the interpretation of the technique. These plots are broken down as follows: First we plot the scores generated by our PCA output in R, (`pca[['x']]`). These scores can be any combination of principal components, however, we are most interested in PC1 and PC2 since they contain the most variation. Secondly, we plot the loading's, whose columns (listed in figure 5) are eigenvectors, on top of the scores. These loading's are all centered at 0, after standardizing the data [5, 1].

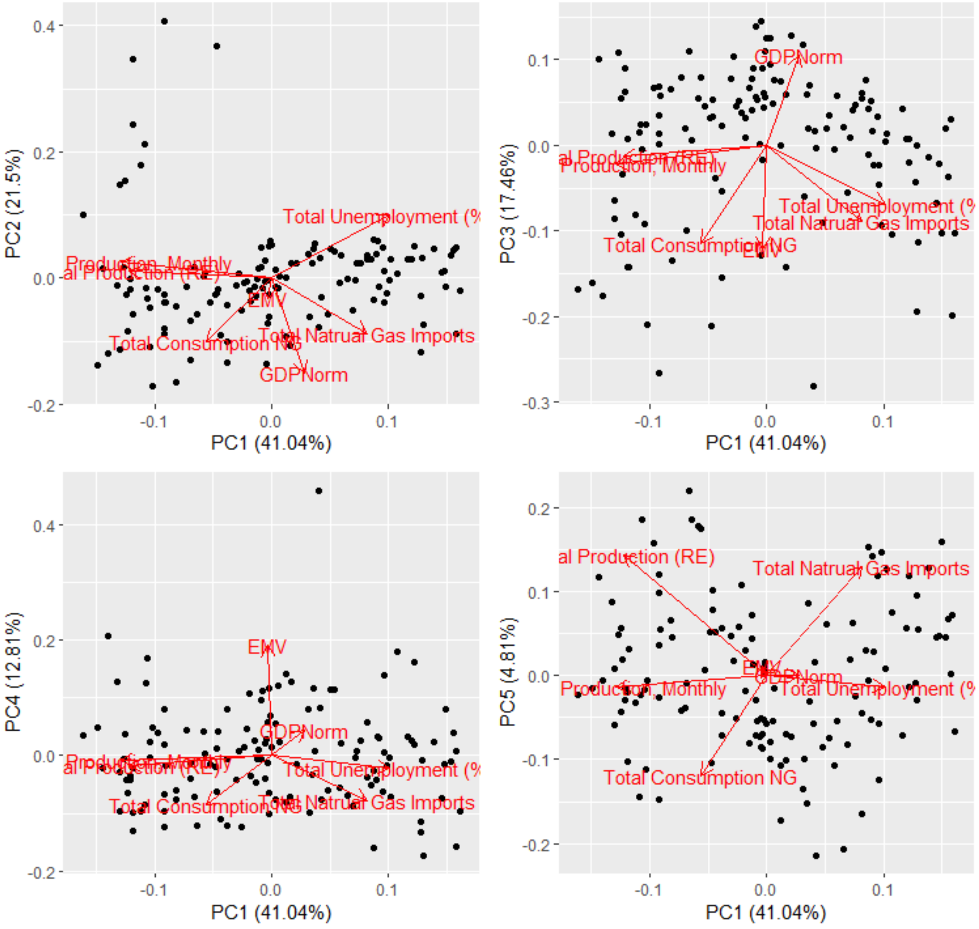
```
        {
        PC12 <- autoplot(pca, loadings = TRUE, loadings.label = T, x = 1, y = 2)
        PC13 <- autoplot(pca, loadings = TRUE, loadings.label = T, x = 1, y = 3)
        PC14 <- autoplot(pca, loadings = TRUE, loadings.label = T, x = 1, y = 4)
        PC15 <- autoplot(pca, loadings = TRUE, loadings.label = T, x = 1, y = 5)
        PC16 <- autoplot(pca, loadings = TRUE, loadings.label = T, x = 2, y = 3)
        PC17 <- autoplot(pca, loadings = TRUE, loadings.label = T, x = 3, y = 4)

        PC12 + PC13 + PC14 + PC15 +
          plot_layout(ncol = 2, nrow = 2)
        }

```

To interpret the biplots, we must define an important relationship commonly seen in biplots. First, the cosine of the angle between a vector and an axis indicates the importance of the contribution of the corresponding variable to the principal component. Vector's representative of the variable's loadings demonstrates this relationship. For instance, the biplot below is a show the relationship of PC1, against PC2, PC3, PC4, and PC5. These are the 5 PCs that demonstrated the most variation in the data matrix. In the first biplot, where PC1 (41.04%) is plotted against PC2 (21.5%), we see a cluster of vectors along zero. Furthermore, total unemployment is facing the direction of PC2, meaning it contributes the most variation to PC2. Alternatively, Total consumption, GDP normalized, and Total natural gas imports contribute an equal amount of variation to PC1. Looking at the second biplot in position 1x2, where PC1 (41.04%) is plotted against PC3 (17.46%), we observe a strong contribution to the variation between normalized GDP and PC3. The other variables, total consumption NG, Total unemployment, and total natural gas, contribute somewhat equal amounts of variation to PC1 (41.04%).

Figure 5: Biplot for different arrangements of Principal Components



3 Conclusion & Discussion:

3.1 Final Discussion:

In this paper, the goal was to find patterns in the data which contained economic variables, and energy related variables regarding production, consumption, exports, pricing and market volatility. The first step in analyzing the data, as mentioned in the textbook, was to observe the correlation matrix for the original data set. This data set showed perfect correlation among specific variables, and no correlation among others. In doing so, I removed these variables and ended with a data set containing 7 variables. After checking the correlation matrix, again, I performed principal component analysis. Using the code provided above, I was able to reduce the number of components from 7, to 4 or 5. The cumulative proportion of PC4 and PC5 contain up to 92.8% and 97.62% of the variation. To decide whether we should choose PC4 or PC5, we can turn to the Scree plot. Observing figure 3, we are looking for an elbow to determine the optimal number of principal components. The plot demonstrates a definitive elbow at PC5, therefore, we will summarize the data in terms of 5 principal components.

After determining the optimal number of principal components, I decided to delve deeper into the results, and examine which variables contribute the most variation to each principal components, mainly PC1, PC2, PC3, PC4, and PC5. This was done by examining both the raw scores of the PCs, and interpretation them in the context of biplots. Examining the raw scores, total consumption of natural gas contributed large negative scores for PC2, PC3, PC4, and PC5. Moreover, Total natural gas imports contributed strong negative correlations to PC1 - PC5. Furthermore, normalised GDP had a strong, negative, loading score for PC2, and PC3. Lastly, renewable energy production contributed a strong negative loading for PC1. These values help us understand if a variable has a strong relation to a particular relationsip. Most of the values for loading's are negatively correlated to their respective principal components.

While the findings were not significant, the results from PCA showed the data could be reduced significantly to 5 principal components. Furthermore, the biplots showed interesting relationships between Labor unemployment, and the consumption of natural gas, GDP and natural gas imports. Almost all loading's were negatively correlated, meaning as one increases the other decreases.

3.2 Future Research and Work:

After determining an optimal number of principal components from this data set, I believe it would be interesting to apply this first step, to factor analysis which is an extension of PCA. Factor analysis would describe the covariance relationships among a range of variables, among many variables in terms of unobservant quantities, factors [4]. Here we would examine whether the data are consistent with a specific structure. Another alternative would be to apply the sample principal components to a multivariate model, to try and predict future energy emissions, based off economic indicators, and energy data. Alternatively, we could use cluster analysis and observe the clusters generated by the principal components. On a more application basis, a few papers use the principal components, generated by observing similar economic, and energy data, can be used to create indexes for natural gas in the United States. I saw this being done with data from China, however, I did not see it applied to the United States.

References

[1] URL: http://strata.uga.edu/8370/lecturenotes/principalComponents.html.

[2] Xiucheng Dong et al. "Sustainability Assessment of the Natural Gas Industry in China Using Principal Component Analysis". In: *Sustainability* 7.5 (2015), pp. 6102–6118. DOI: 10.3390/su7056102.

[3] *Federal Reserve Economic Data: FRED: St. Louis Fed.* URL: https://fred.stlouisfed.org/.

[4] Richard Arnold. Johnson and Dean W. Wichern. *Applied multivariate statistical analysis.* Prentice Hall, 2007.

[5] Linh Ngo. *How to read PCA biplots and scree plots.* Sept. 2020. URL: https://blog.bioturing.com/2018/06/18/how-to-read-pca-biplots-and-scree-plots/.

[6] Philip Schellekens et al. *World Bank Open Data.* Mar. 2021. URL: https://data.worldbank.org/.