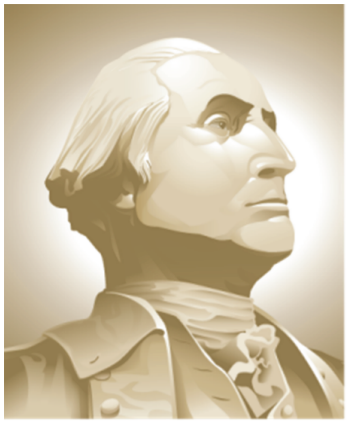
EMSE 6765: DATA ANALYSIS
For Engineers and Scientists

Session 11: Principal Component Analysis (PCA), Introduction, How it works, Mechanics

Version: 11/11/2021



THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Lecture Notes by: J. René van Dorp[1]

www.seas.gwu.edu/~dorpjr

---

[1] Department of Engineering Management and Systems Egineering, School of Engineering and Applied Science, The George Washington University, 800 22nd Street, N.W., Suite 2800, Washington D.C. 20052. E-mail: dorpjr@gwu.edu.

- Researchers deal with dozens or **even hundreds of variables** in their analyses. With that many variables it is **difficult to comprehend the patterns of association or statistical dependence**. This is especially true nowadays in the era "Big Data".

- This is complicated by the fact that there is often **substantial redundancy among the dimensions in the data**, as a result of **high levels of correlation and multicolinearity** within the data.

- **Principal component analysis** is a method for **re-expressing** multivariate data. It allows the researcher to **re-orient the data so that a first few dimensions account for as much of the available information/variation as possible within the original data set**. The researcher must decide on **the number of dimensions to use**, trading of simplicity for completeness.

- **Each new dimension/principal component is uncorrelated with the other principal components** and **hence captures independent pieces of the information/variation puzzle represented in the larger data set**. **Providing meaningful interpretation** to these **principal components/dimensions** is **one of the challenges in PCA analysis.**

- **Example application:** Creating indices from survey data.

QUESTION	RESPONSE
C1	I prefer complex to simple problems
C2	I like to have repsonsibility of handling a situation that requires a lot of thinking
C3	Thinking is not my idea of fun
C4	I would rather do something requiring little thought than something that is sure to challenge my thinking abilities
C5	I try to antiticipate and avoid situations where there is a little chance that I will have to think in depth about something
C6	I find satisfaction in deliberating hard for long hours
C7	I only think hard as I have to
C8	I prefer to think about small daily projects to long-term ones
C9	I like little tasks that require little thought once I have learned them
C10	The idea of relying on thought to make my way to the top appeals me
C11	I really enjoy a task that involves coming up with new solutions to problems
C12	Learning new ways to think doesn't excite me much
C13	I prefer my life to be filled with puzzles that I must solve
C14	The notion of thinking abstractly appeals to me
C15	I prefer tasks that are intellectual, difficult, and important to ones that do no require much thought
C16	I feel relief rather than satisfaction after completing a taks that required a lot of mental effort
C17	It's enough for me that something gets the job done; I don't care how or why it works.
C18	I usually end up deliberating about issues even when they do not affect me personally

- 18 questions to **elicit a persons "need for cognition"**
- Answers are provided on **a scale from 1 to 5** (A Likert Scale).
- **The colored questions are in reverse order** such that a **higher score on the question reflects a lower need for cognition.**

First thought: Create index by adding and subtracting the reverse coded responses

	1	1	-1	-1	-1	1	-1	-1	-1	1	1	-1	1	1	1	-1	-1	1	
Respondent	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	Additive Score
1	1	3	4	4	3	4	3	4	4	3	4	3	4	4	3	2	2	2	-1
2	2	4	1	1	2	1	4	2	2	4	4	3	2	2	4	2	1	4	9
3	4	4	1	1	1	3	1	3	2	4	5	1	5	4	5	2	2	4	24
4	3	1	4	5	4	3	3	3	5	1	1	2	3	5	1	5	5	3	-15
5	4	4	2	2	1	4	2	2	1	4	5	1	4	2	4	1	1	2	20
6	5	5	1	1	1	5	1	3	1	5	5	1	5	5	5	1	2	4	32
7	3	3	1	2	2	1	4	3	2	4	4	2	3	2	3	2	1	2	6
8	3	4	1	1	1	3	2	3	3	4	5	1	3	5	4	2	1	2	18
9	4	4	1	2	2	4	4	3	1	5	5	2	4	4	4	4	1	4	18
10	4	4	2	2	1	3	5	2	2	4	5	1	3	4	4	2	1	5	18
11	2	3	3	3	2	3	4	2	2	4	4	2	4	3	3	3	4	2	3
12	2	3	2	3	3	2	3	3	4	4	5	5	4	2	4	4	2	4	1
13	1	5	1	1	2	5	1	4	4	5	5	1	4	5	4	1	4	4	19
14	3	3	3	2	1	3	2	2	2	3	4	2	3	3	3	2	1	3	11
15	2	5	1	1	1	4	1	1	1	5	5	1	4	4	2	5	1	2	20
16	2	2	3	4	4	3	1	3	5	4	4	3	1	2	4	1	3	3	-2
17	4	4	1	2	3	4	2	3	2	5	5	1	4	4	4	3	2	4	19
18	5	5	1	1	1	1	1	1	1	5	5	1	5	5	5	1	1	4	31
19	5	5	1	1	1	1	4	2	1	5	5	1	5	4	5	1	2	2	23
20	4	5	1	1	1	4	1	1	2	5	5	1	4	4	4	2	1	1	25
21	4	5	1	3	1	4	3	2	4	3	4	3	2	5	5	1	1	5	18
22	2	5	5	2	2	4	5	2	2	4	4	1	5	4	4	5	1	1	8
23	4	3	1	1	1	4	2	2	2	5	5	1	4	5	5	5	2	3	21
24	4	4	2	2	2	3	1	2	1	4	4	3	3	5	4	3	2	4	17
25	1	4	1	2	2	2	4	1	1	5	4	2	1	5	3	2	1	4	13
26	2	4	1	2	2	1	4	2	2	3	4	3	4	4	3	2	2	2	7
27	1	1	1	1	1	4	4	4	2	5	5	1	1	5	3	1	1	1	10

Principal Component Analysis: C1-C18

Variable	PC1	PC2
C1	0.257	-0.133
C2	0.306	-0.049
C3	-0.248	-0.109
C4	-0.273	-0.272
C5	-0.295	-0.104
C6	0.191	-0.161
C7	-0.214	-0.192
C8	-0.218	-0.092
C9	-0.223	-0.285
C10	0.261	-0.144
C11	0.286	-0.139
C12	-0.261	0.078
C13	0.233	-0.349
C14	0.239	-0.412
C15	0.148	-0.280
C16	-0.159	-0.369
C17	-0.227	-0.411
C18	0.096	-0.099

- Note that the reverse coded questions score negatively on the first principal component.

- Note that the absolute values differ from 0.096 to 0.306 indicating not every questions contributed the same to the "need for cognition"

- First principal component accounted for 32.1 % of the variance in the data and had an eigenvalue of 5.7506

Principal Component Analysis: C1-C18

Eigenanalysis of the Correlation Matrix

Eigenvalue	5.7506	1.4249	1.1730	0.9944	0.9527	0.8980	0.8837	0.8159
Proportion	0.319	0.079	0.065	0.055	0.053	0.050	0.049	0.045
Cumulative	0.319	0.399	0.464	0.519	0.572	0.622	0.671	0.716

Eigenvalue	0.7313	0.6709	0.6456	0.5692	0.5566	0.5142	0.4375	0.4003
Proportion	0.041	0.037	0.036	0.032	0.031	0.029	0.024	0.022
Cumulative	0.757	0.794	0.830	0.862	0.893	0.921	0.945	0.968

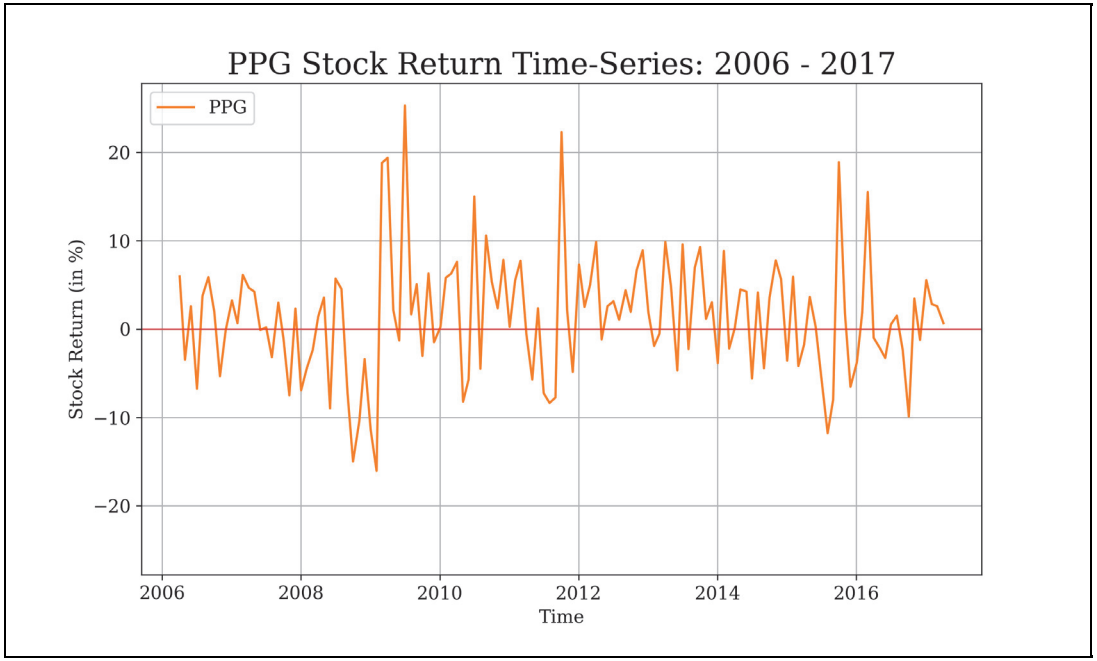
Eigenvalue	0.3051	0.2763
Proportion	0.017	0.015
Cumulative	0.985	1.000

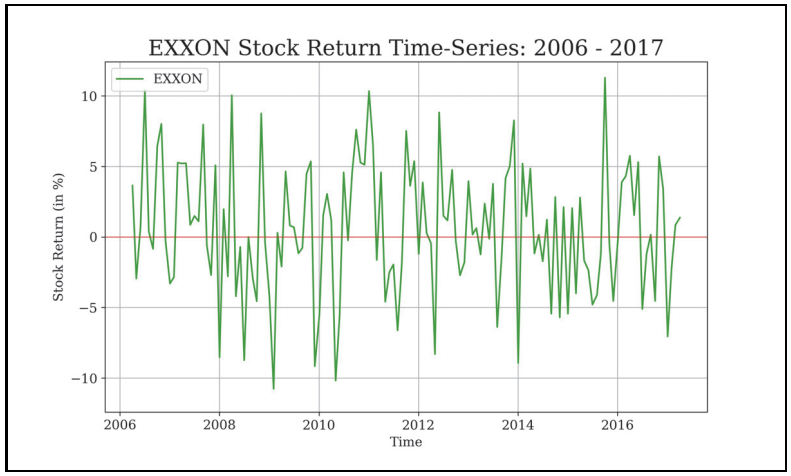
- When performing a **Principal Component Analysis** on the **Correlation Matrix** you *typically* look only for those **Principal Components with an eigenvalue greater than one** (although there are some exceptions).

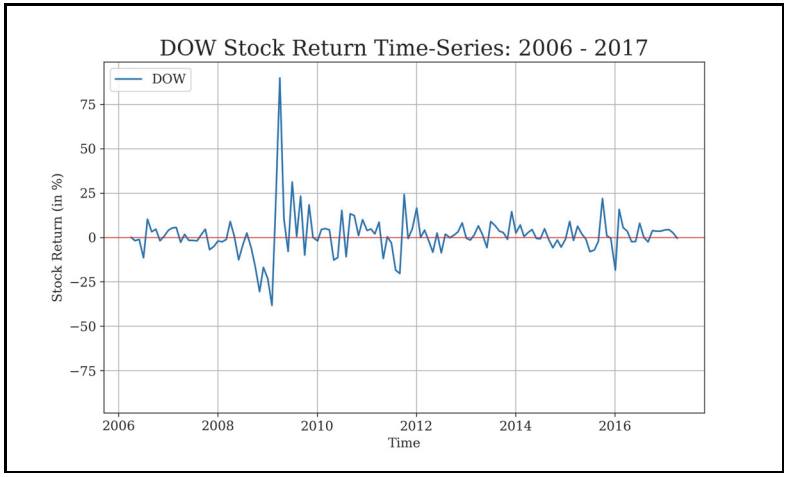
- An initial observation of the above eigen analysis suggests that **three out of the 18 principal components** describe about **46% of the variation in the 18 dimensional data (i.e. question response data)** and the **rest of them can be attributed as measurement error** (certainly if their eigenvalues are much less than one).

- **The monthly rates of return for three stocks** (PPG Industrues, EXON, DOW Chemicals) listed on the New York Stock Exchange were determined for the period March 1976-April 2017. **The monthly rates of return are defined as:**

$$\frac{Current\ Closing\ Price\ -\ Previous\ Closing\ Price}{Previous\ Closing\ Price}$$







- The observations in 133 succesive months appear to be distributed with weak auto-correlations:

	AUTO Correlation		
	PPG	EXXON	DOW
One-Step	0.06	0.01	0.16
Two-Step	-0.11	0.00	-0.08
Three Step	0.07	-0.14	0.05

	Correlation Matrix		
133	PPG	EXXON	DOW
PPG	1.00	0.35	0.71
EXXON	0.35	1.00	0.24
DOW	0.71	0.24	1.00

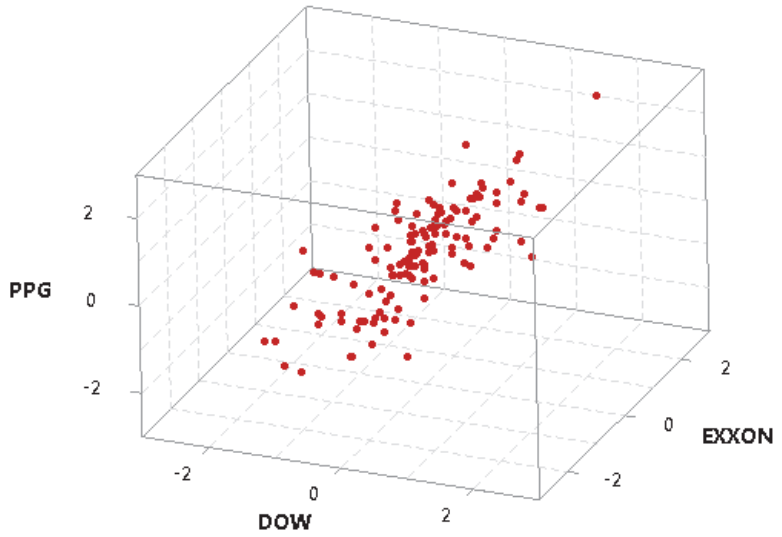
- The rates of return across stocks are **positively correlated**, as one would expect **since stocks tend to move together** in response **to general economic conditions.**

- Lets open the spreadsheet "Stock_3_Raw_Data.xlsx" and perform a PCA analysis of the correlation matrix of these rate of returns.

PPG Industries							
Date	Open	High	Low	Close	Volume	Adj Close	RETURN
4/3/2017	82.019997	83.550003	81.330002	83.129997	11630700	83.129997	0.01365681
3/1/2017	81.699997	84.25	80.309998	82.010002	14445600	82.010002	0.008485022
2/1/2017	84	84.160004	80.760002	81.32	13325600	81.32	-0.021771385
1/3/2017	90.940002	91.339996	83.129997	83.889999	13578400	83.129852	-0.070573889
EXXON							
Date	Open	High	Low	Close	Volume	Adj Close	RETURN
4/3/2017	82.019997	83.550003	81.330002	83.129997	11630700	83.129997	0.01365681
3/1/2017	81.699997	84.25	80.309998	82.010002	14445600	82.010002	0.008485022
2/1/2017	84	84.160004	80.760002	81.32	13325600	81.32	-0.021771385
1/3/2017	90.940002	91.339996	83.129997	83.889999	13578400	83.129852	-0.070573889
DOW CHEMICALS							
Date	Open	High	Low	Close	Volume	Adj Close	RETURN
4/3/2017	63.57	64.43	62.299999	63.27	5217100	63.27	-0.004249307
3/1/2017	62.869999	65.419998	61.970001	63.540001	6845100	63.540001	0.027832886
2/1/2017	59.610001	64.360001	59.330002	62.259998	6155700	61.819389	0.04410527
1/3/2017	57.200001	61.73	56.52	59.630001	6632700	59.208004	0.042118129

Stock data downloaded from:
https://finance.yahoo.com

The intuition of PCA is **best illustrated visually in an example with 3D data**.

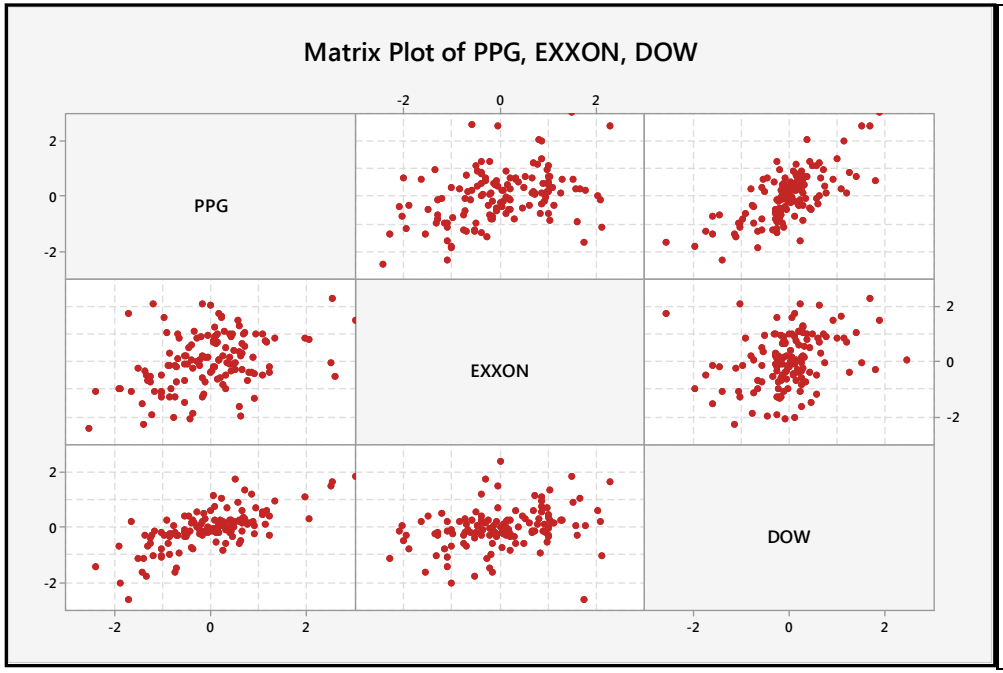


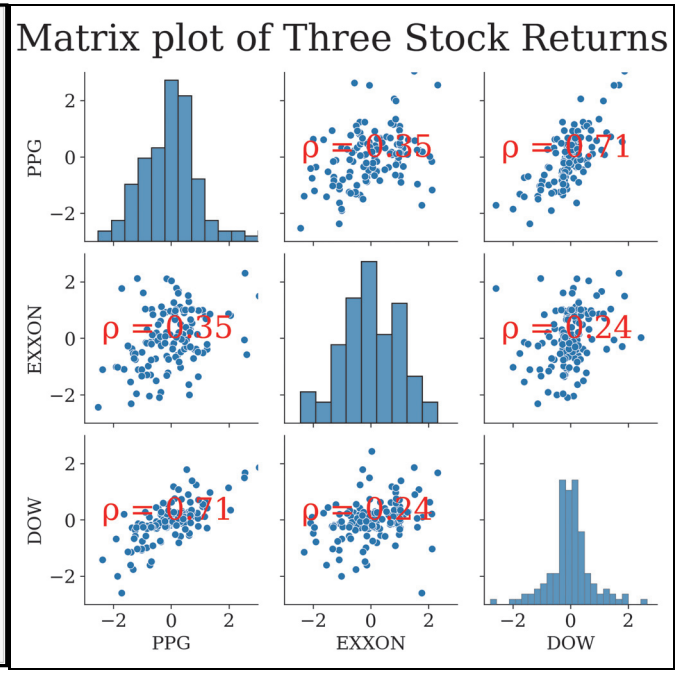
	Correlation Matrix		
133	PPG	EXXON	DOW
PPG	1.00	0.35	0.71
EXXON	0.35	1.00	0.24
DOW	0.71	0.24	1.00
VAR	1.00000	1.00000	1.00000
MEAN	0.00000	0.00000	0.00000

- Note that we are using standardized data since the variance of each variable equals 1. **The sum of the variances is therefore equal to three.**

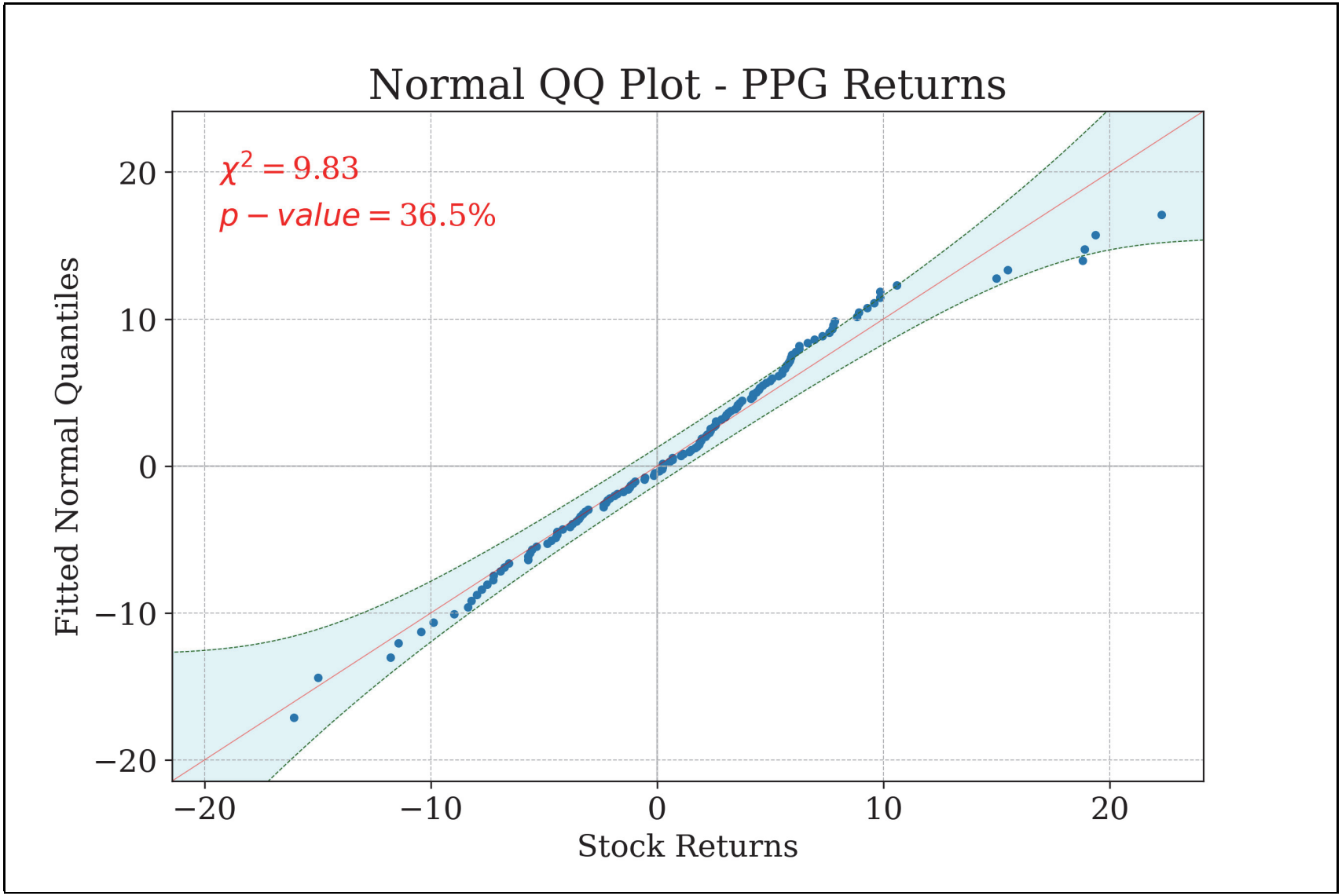
- Correlation is largest between PPG and DOW (both chemical companies and is smallest between DOW and EXXON).

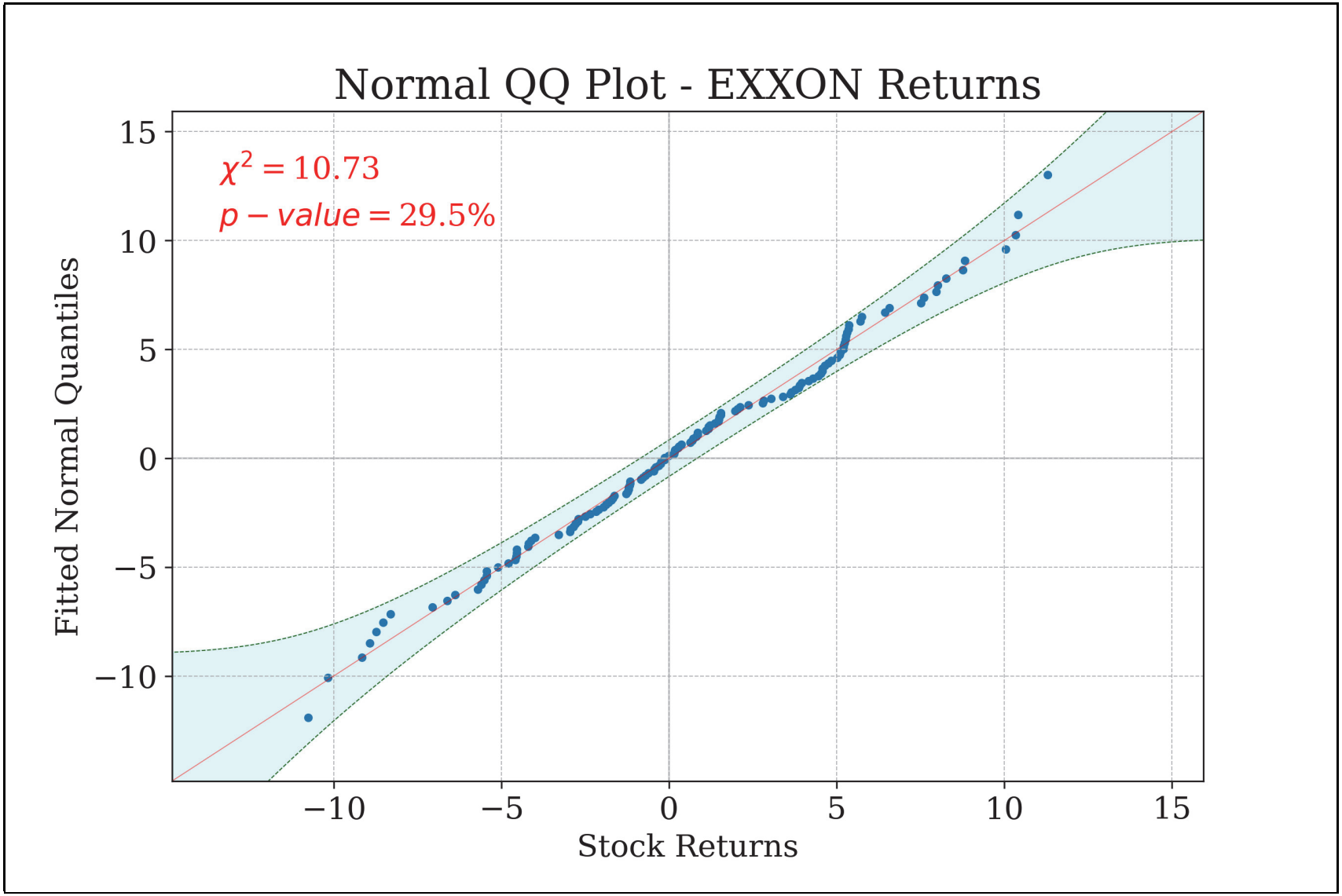
Matrix Plot generated by MINITAB (left) and Python (right)

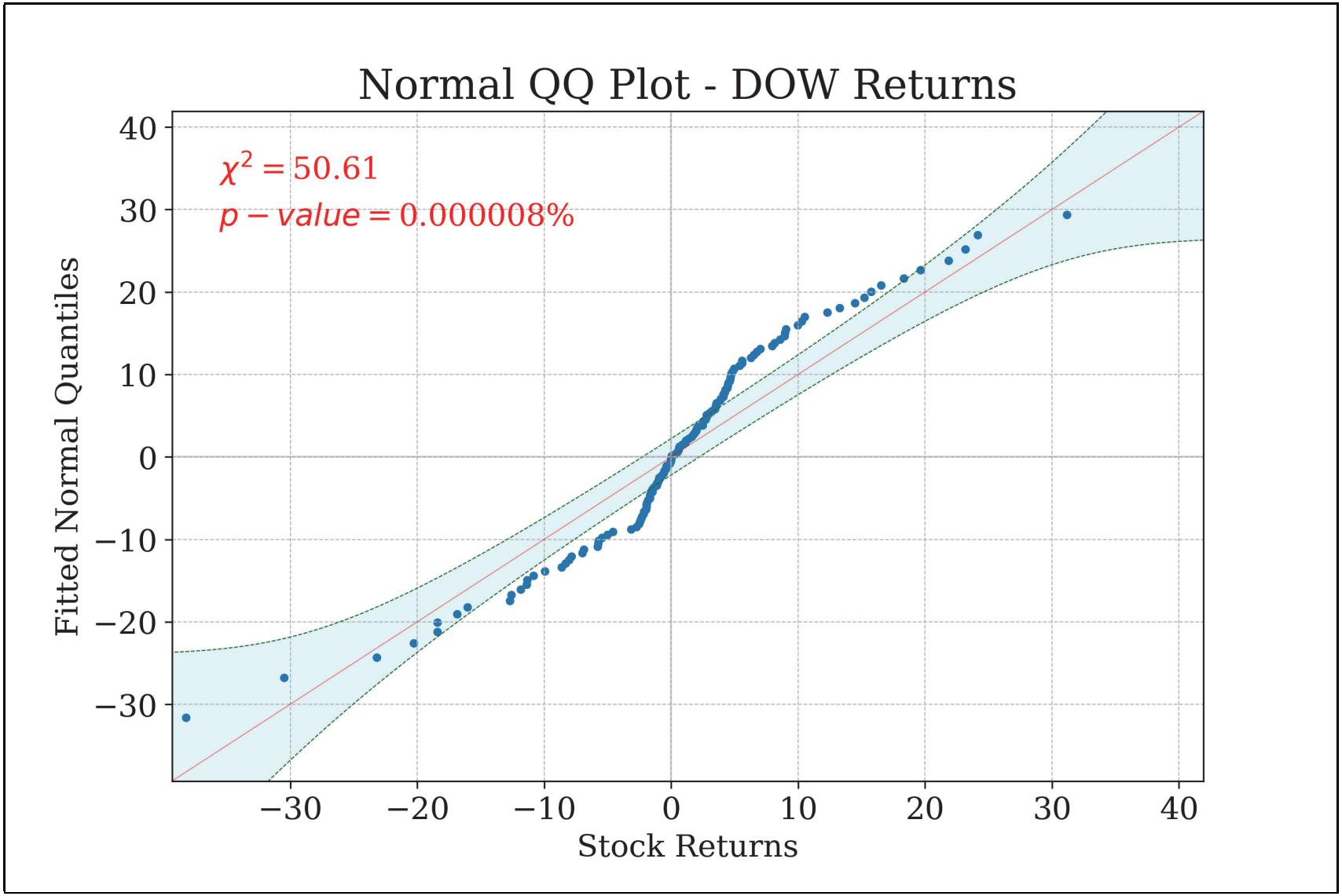




Noting this pattern of positive correlation throughout, one is tempted to ask the following question: **Would it be possible to use a single dimension to capture and convey most (if not all) of the information/variation contained in the standardized stock price data of *PPG*, *EXXON* and *DOW*?**

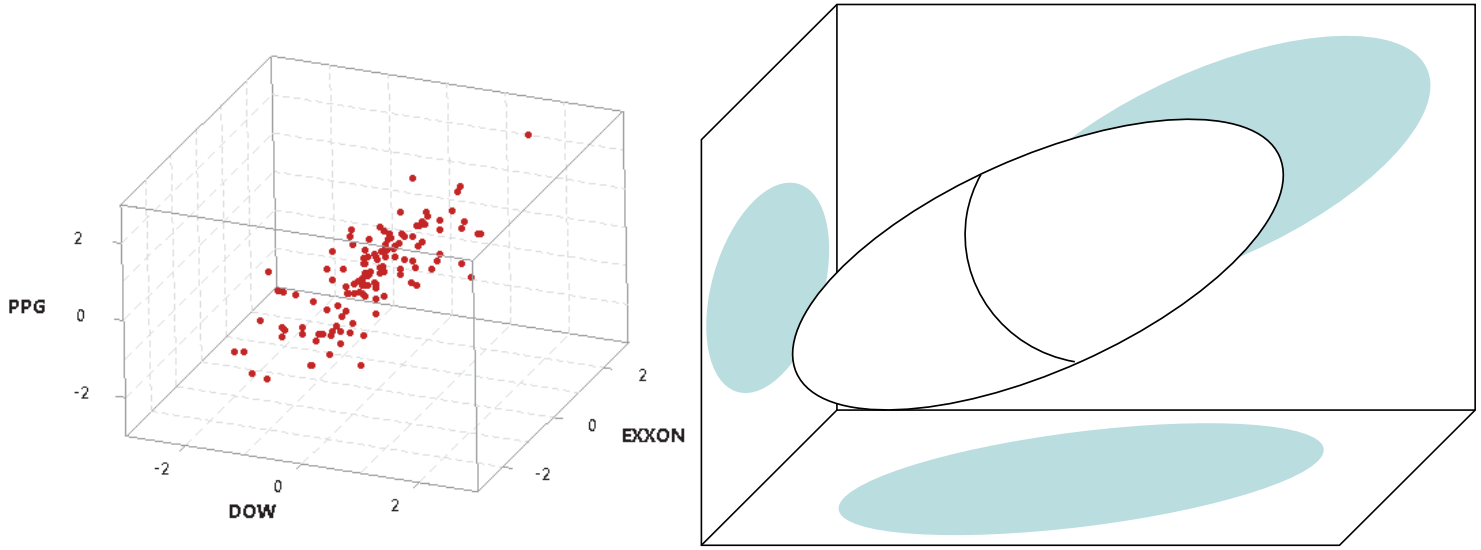

Normal QQ Plot - PPG Returns
$\chi^2 = 9.83$
$p - value = 36.5\%$
Fitted Normal Quantiles
Stock Returns
20
10
0
−10
−20
−20
−10
0
10
20


Normal QQ Plot - EXXON Returns
$\chi^2 = 10.73$
$p - value = 29.5\%$
Fitted Normal Quantiles
Stock Returns
15
10
5
0
−5
−10
−10
−5
0
5
10
15


Normal QQ Plot - DOW Returns
$\chi^2 = 50.61$
$p - value = 0.000008\%$
Fitted Normal Quantiles
Stock Returns
40
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30
40

Principal component analysis: Finding **a smaller number of variables/ dimensions** that accounts for **a large explanation of the original information.**

Stylized view of Scatter plot

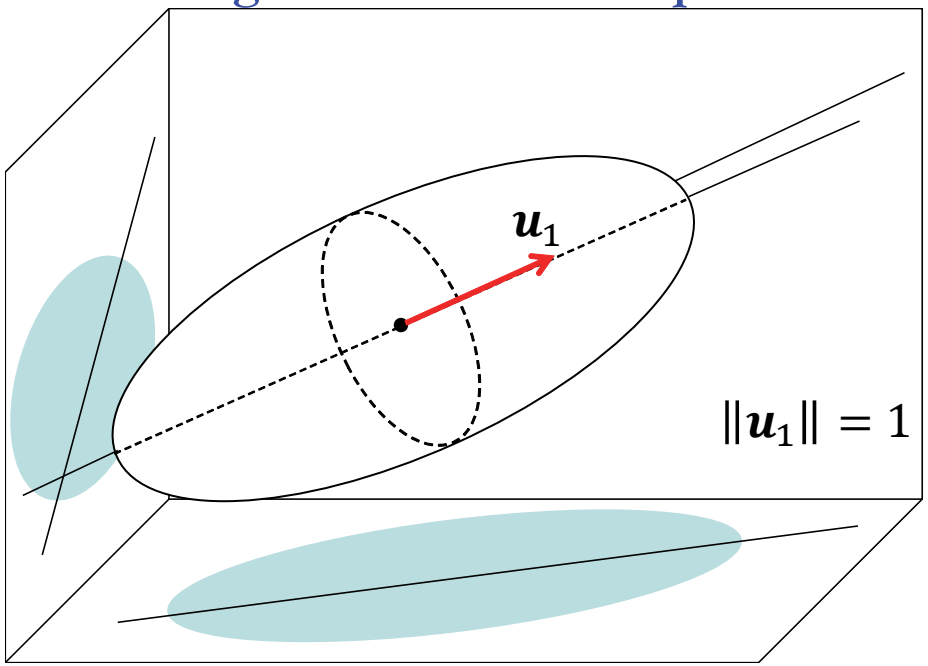


QUESTION:

What linear combination of $PPG, DOW, EXXON$ accounts for **the largest variance?** The answer to this question identifies the first **Principal Component.**

- Recall that **a vector-vector multiplication $X^T u_1$** is **a projection of the vector $X^T = (X_1, X_2, X_3) = (PPG,\ EXXON,\ DOW)$** onto **the line spanned by the vector u_1, where by convention one chooses $\|u_1\| = 1$.**

Stylized view of First Principal Component: Longest Axis of the Ellipsoid.



$$Z_1 = X^T \boldsymbol{u_1} = \sum_{i=1}^{3} u_{i1} X_i,\ \|\boldsymbol{u_1}\| = 1,\ Z_1 = 0.656X_1 + 0.420X_2 + 0.627X_3,$$

$Var(Z_1) = 1.907$ accounting for **$1.907/3 \approx 63.6\%$ of original variance.**

- One can think of $Var(Z_1) = 1.907$ as accounting for $\approx 63.6\%$ of **the combined original total summed variance of 3**. **The greater this value, the more "information" from the original data** is retained **in this single principal component.**

- Let A be an $(n \times m)$ matrix with $n > m$ (more rows than columns) and with columns that are linearly independent, i.e.

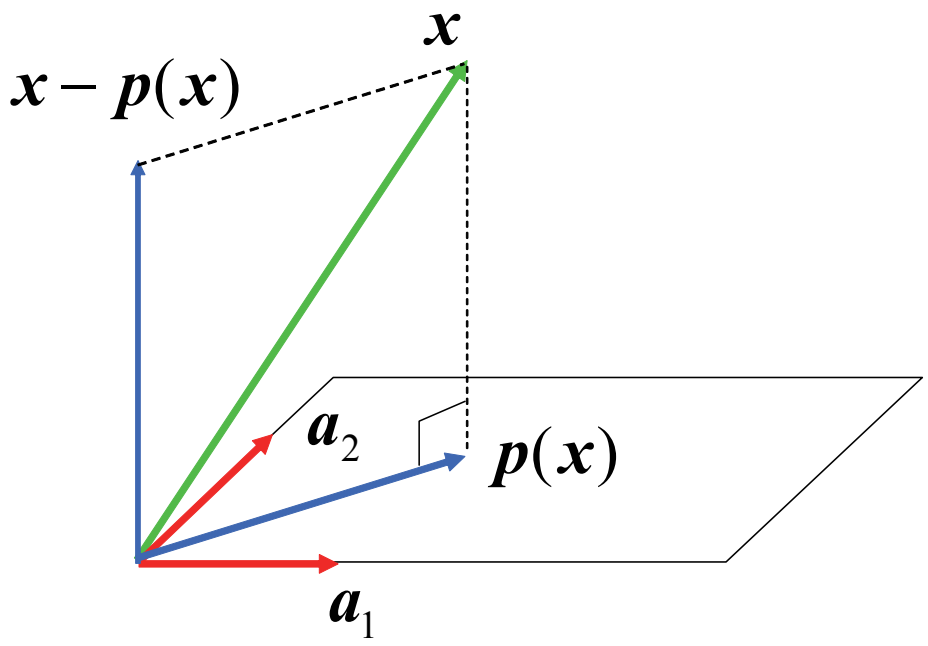
$$\sum_{i=1}^{m} \lambda_i \boldsymbol{a}_i = \mathbf{0} \Leftrightarrow \lambda_i = 0,\ i = 1, \ldots m$$

- $A = (\ \boldsymbol{a}_1 \quad \boldsymbol{a}_2\ )$, Column Space of $A \equiv$ All points $\boldsymbol{y}$ such that $\boldsymbol{y} = \lambda_1 \boldsymbol{a}_1 + \lambda_2 \boldsymbol{a}_2$ for some constants λ_1, λ_2.

- The projection of a vector $\underline{\boldsymbol{x}}$ onto the column space of A, equals:

$$p(\underline{\boldsymbol{x}}) = A(A^T A)^{-1} A^T \underline{\boldsymbol{x}}$$

- Recall that in regression we obtained: $\widehat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$. Hence, **the regression fit $\widehat{\boldsymbol{y}}$ is the projection of the dependent variable $\boldsymbol{y}$** onto **the columns space of the data matrix $\boldsymbol{X}$**.



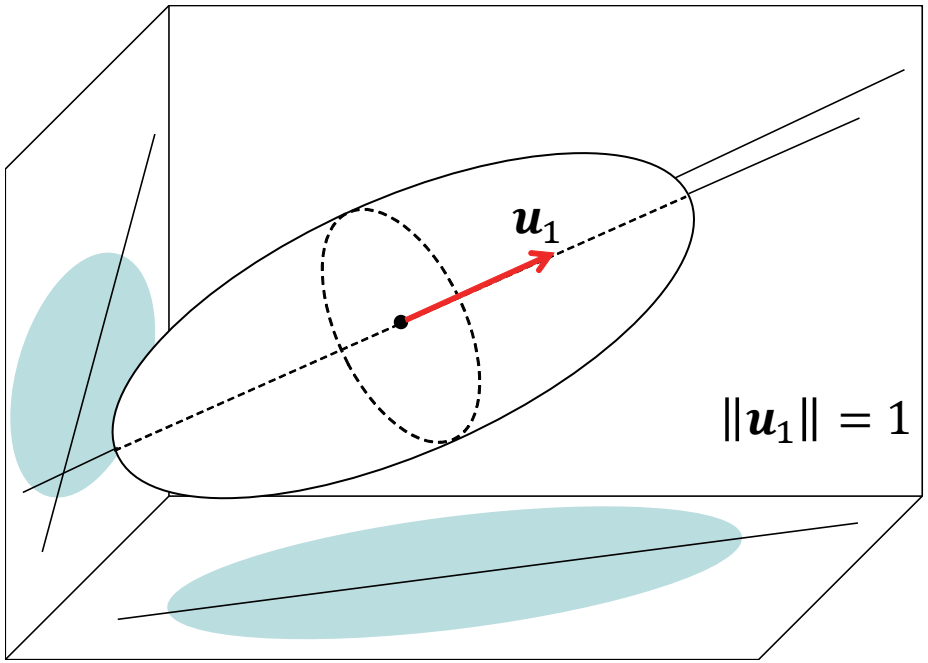
$$p(\underline{\boldsymbol{x}}) = A(A^T A)^{-1} A^T \underline{\boldsymbol{x}}, \quad \underline{\boldsymbol{x}} - p(\underline{\boldsymbol{x}}) = (I - A(A^T A)^{-1} A^T)\underline{\boldsymbol{x}}$$

- The projection of a vector $\underline{\boldsymbol{x}}$ onto **the orthogonal complement of the column space of $\boldsymbol{A}$**, equals:

$$\underline{\boldsymbol{x}} - p(\underline{\boldsymbol{x}}) = (I - A(A^T A)^{-1} A^T)\underline{\boldsymbol{x}}$$

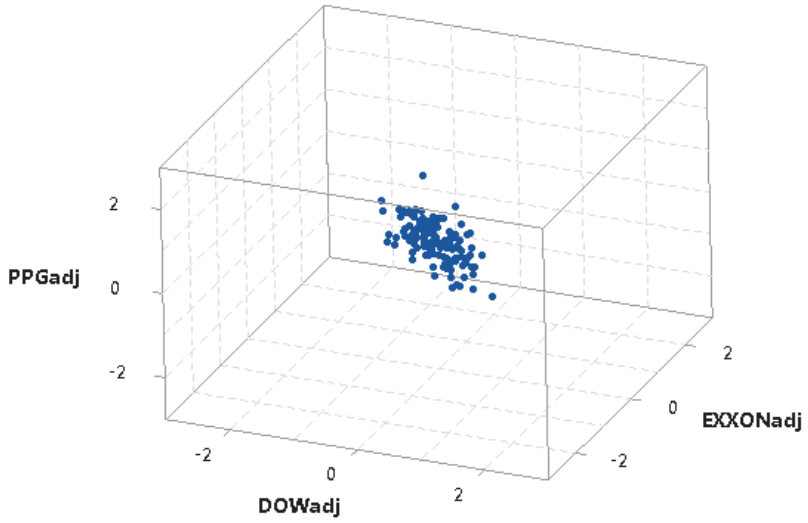
- In regression analysis, the elements of vector $\boldsymbol{y} - \widehat{\boldsymbol{y}}$ are the residuals.

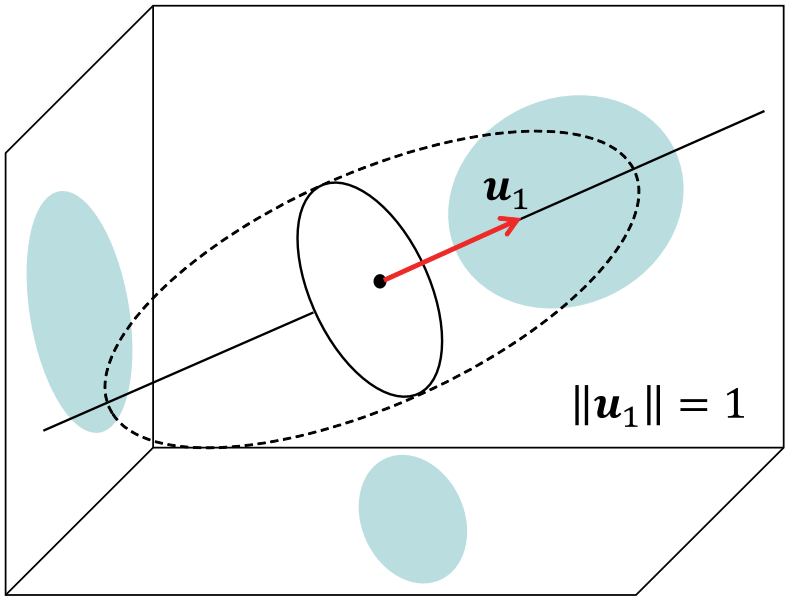
- We can now calculate the projection $p(\boldsymbol{x})$ of each point $\boldsymbol{x}$ onto the line spanned by the vector $\underline{u}_1$ (**constituting the first principal component**). **The difference vector $\boldsymbol{x} - \boldsymbol{p(x)}$ is the projection of $\boldsymbol{x}$ onto the plane perpendicular to the vector $\underline{\boldsymbol{u}}_1$.**



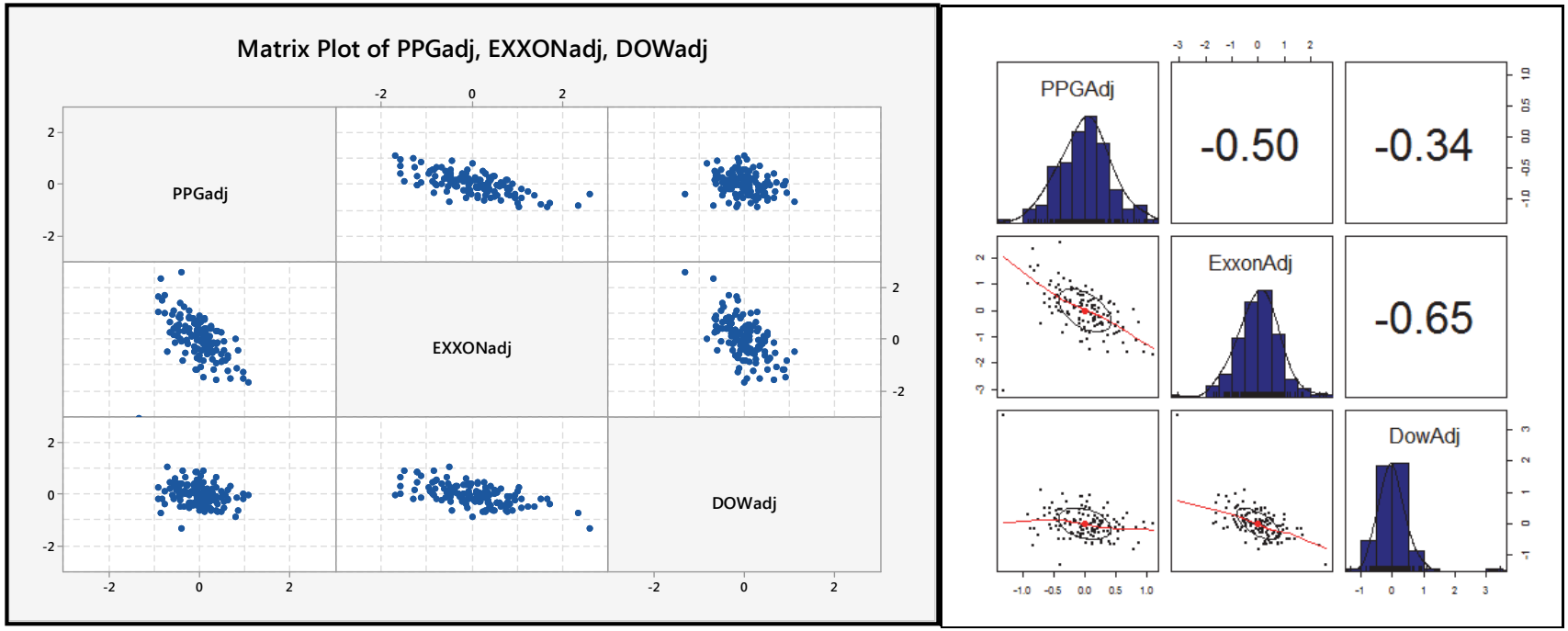
- By considering now the points $\underline{\boldsymbol{x}} - p(\underline{\boldsymbol{x}})$ **we have removed/filtered from the original data** the info contained in **the first principal component.**

- Denoting the coordinates of the projected points $\underline{\boldsymbol{x}} - p(\underline{\boldsymbol{x}})$, with PPG_{adj}, $EXXON_{adj}$ and DOW_{adj} in a **3D scatter plot** we have:





Notice a much smaller spread in the adjusted scatter plot compared to the original scatter plot.



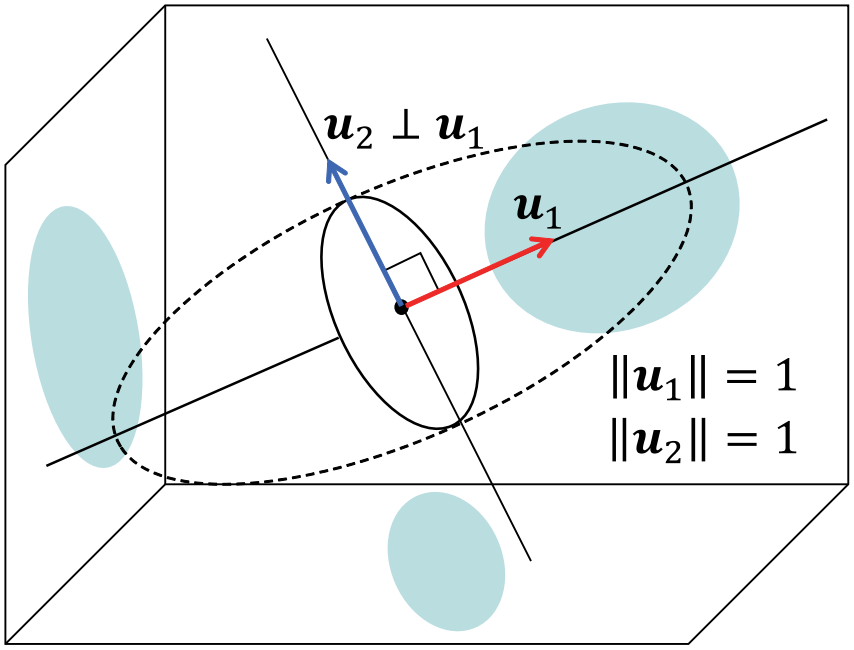
- Observe now that **after removing the information of the first principal component Z_1**, the remaining association suggests **a strong negative correlation** between $(DOW, EXXON)$ and $(PPG, EXXON)$, but also a lesser one between and (PPG, DOW).

- **This is not obvious from the original correlation matrix** and illustrates **why PCA** helps in understanding **more the association** among the variables.

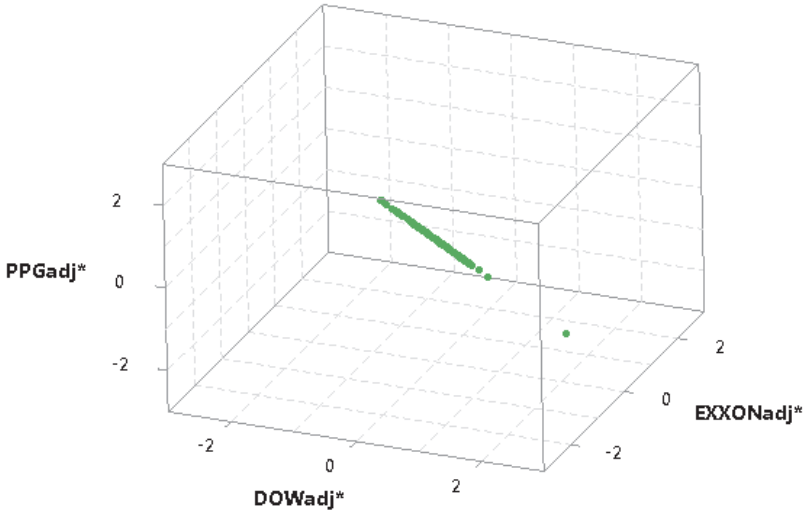
- **To find the second principal component** we are searching for the linear combination of $(X_{1adj}, X_{2adj}, X_{3adj})$ **that has now the largest variance**.

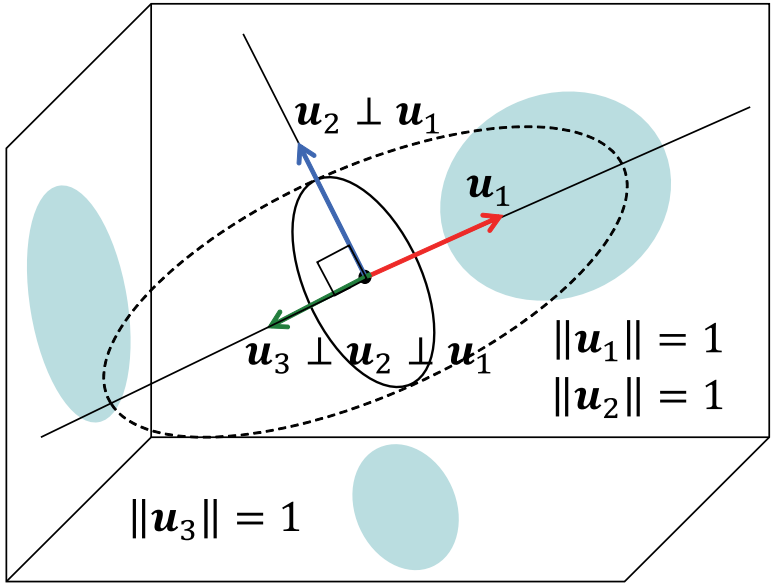
$$Z_2 = X^T \boldsymbol{u_2} = \sum_{i=1}^{3} u_{i2} X_i,\ \|\boldsymbol{u}_2\| = 1, Z_2 = -0.200X_1 + 0.898X_2 - 0.392X_3,$$

$$Var(Z_2) = 0.817 \text{ accounting for } \frac{0.817}{3} \approx 27.2\%,\ \boldsymbol{u}_2 \perp \boldsymbol{u}_1 \text{ (perpendicular)}$$



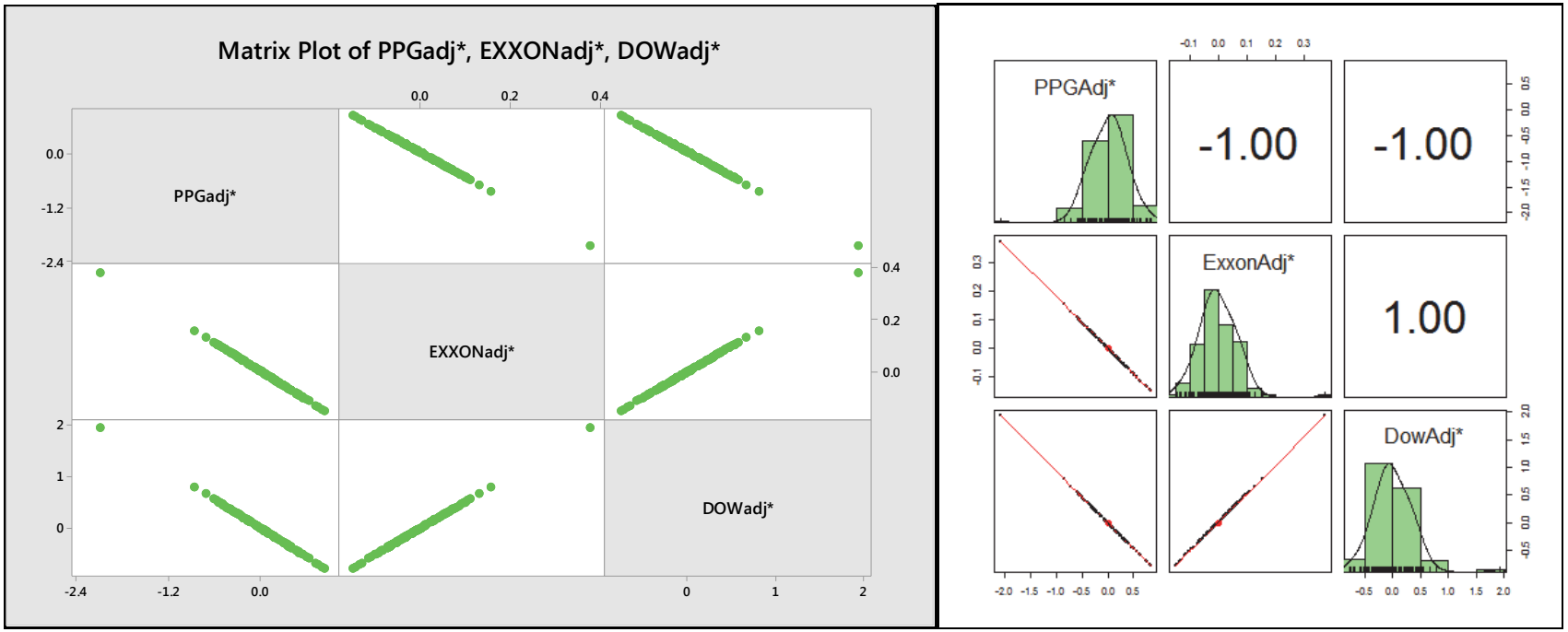
- We can now calculate the projection $p(\underline{\boldsymbol{x}})$ of each point $\underline{\boldsymbol{x}}$ onto the plane spanned by the vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. **The difference vector $\underline{\boldsymbol{x}} - \boldsymbol{p}(\underline{\boldsymbol{x}})$ is the projection of $\underline{\boldsymbol{x}}$ onto the line perpendicular to the vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$.**





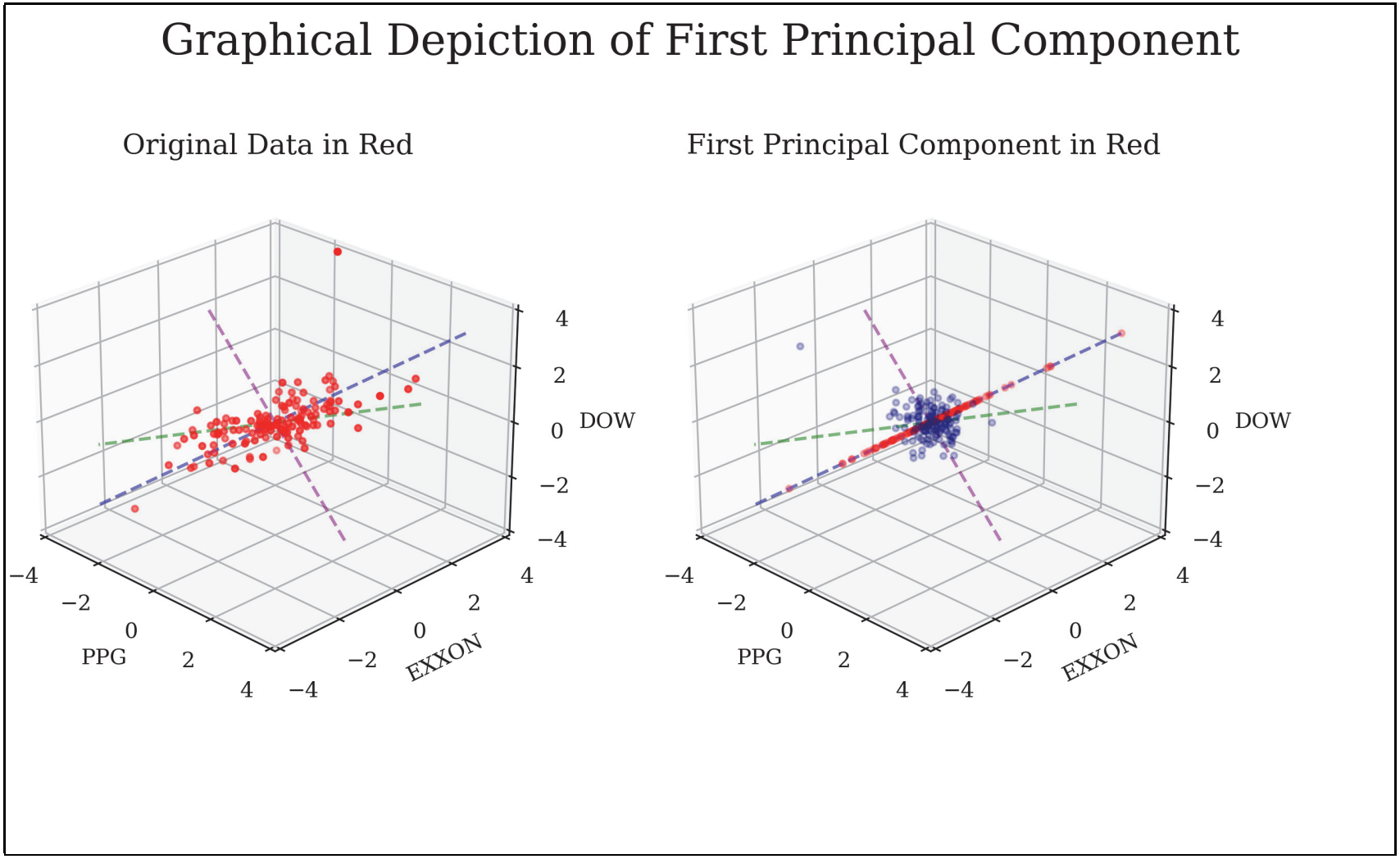
$$Z_3 = X^T \boldsymbol{u_3} = \sum_{i=1}^{3} u_{i3} X_i,\ \|\boldsymbol{u}_3\| = 1,\ Z_3 = -0.727X_1 + 0.131X_2 + 0.673X_3,$$

$$Var(Z_3) = 0.276 \text{ accounting for } \frac{0.276}{3} \approx 9.2\%,\ \boldsymbol{u}_3 \perp \boldsymbol{u}_2 \perp \boldsymbol{u}_1$$



- Observe now that **after removing the information of the first two principal component Z_1 and Z_2**, the remaining association suggests a negative linear relationship between $(PPG, EXXON)$ and (PPG, DOW), but also positive linear relationship between $(EXXON, DOW)$.

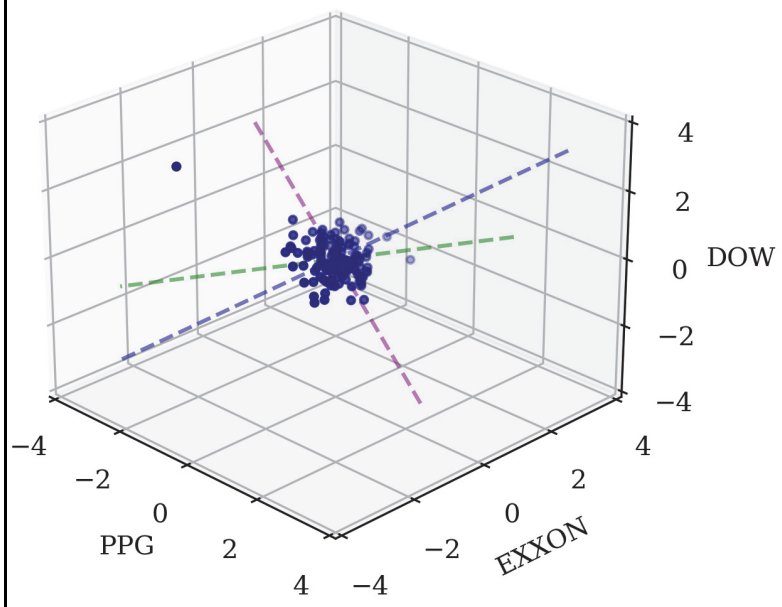
- **This is not obvious from the original correlation matrix** and illustrates **why PCA** helps in understanding **more the association** among the variables.



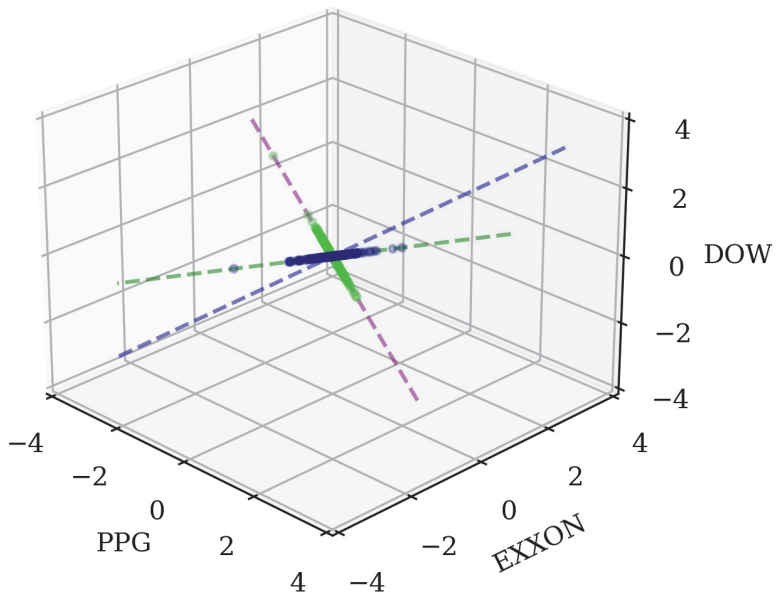
Analysis in "Stocks_3_PCA_FIRST.py"

Graphical Depiction of Second and Third Principal Component

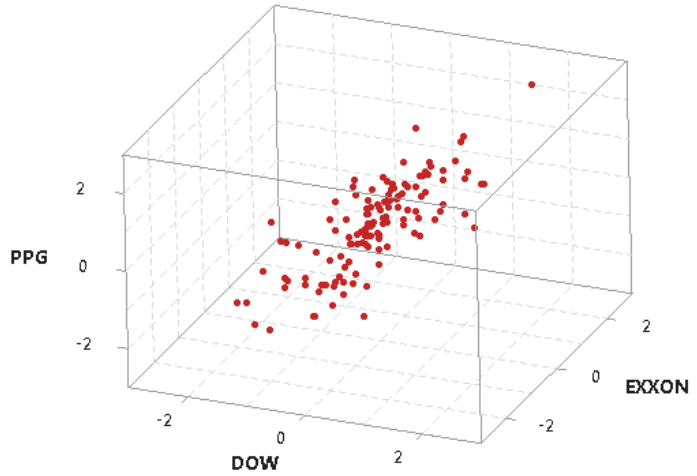
Orthogonal Complement data in blue

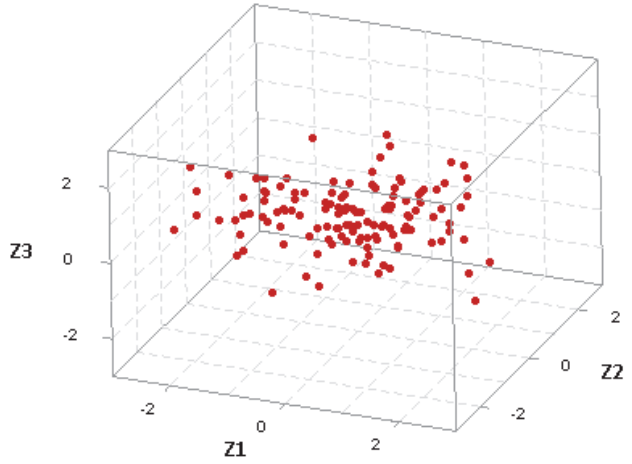


2nd PC in blue, 3rd PC in green

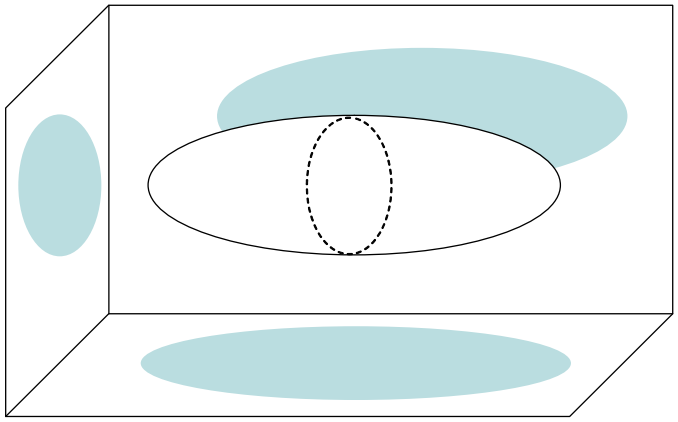


Analysis in "Stocks_3_PCA_Second and Third.py"





Stylized three-dimensional view of shape of distribution of Z_1, Z_2 and Z_3



Correlation Matrix

	Z1	*Z2*	*Z3*
Z1	1.907	0.000	0.000
Z2	0.000	0.817	0.000
Z3	0.000	0.000	0.276

- The new variables Z_1, Z_2, Z_3 are **mutually uncorrelated**, and **each new variables is chosen to account for the maximum amount of variation not already accounted for by the previously chosen variables.**

- Although it takes Z_1, Z_2, Z_3 to explain **all of the original information**, in our example Z_1 and Z_2 together account for a total variance of $1.907 + 0.817 = (= 2.724$ accounting for about $2.724/3.0 \approx 90.8\%$ of the total variance).

- **While this dimension reduction is not a big deal here**, it is easy to imagine the benefits of being able to reduce 20 variables down to say five dimensions.

- **It is usually instructive to examine the relationship** between Z_1, Z_2 and Z_3 and the original variables X_1, X_2, X_3. **One way to do this is look at their correlations**.

- The **correlations** between (X_1, Z_1), (X_2, Z_1), (X_3, Z_1) are called the **principal component loadings** of Z_1 onto the original variables. If all these correlations are high, then Z_1 reflects **a component of shared variance underlying all three original variables.**

Principal Component Loadings

(Correlation between PC's and Original Variables)

	Z1	*Z2*	*Z3*
PPG	0.906	-0.181	-0.382
EXXON	0.580	0.812	0.069
DOW	0.866	-0.354	0.354
Mean	0.000	0.000	0.000
Variance	1.907	0.817	0.276
% Explanation	63.6%	27.2%	9.2%
Cum% Explan.	63.6%	90.8%	100.0%

Total Variance in three standardized variables equals three.

9.2% = 0.276/3

82.1% = $(0.906)^2$

Percent Explanation of Variance in individual variables by PC's

	Z1	*Z2*	*Z3*
PPG	82.1%	3.3%	14.6%
EXXON	33.6%	65.9%	0.5%
DOW	74.9%	12.5%	12.5%

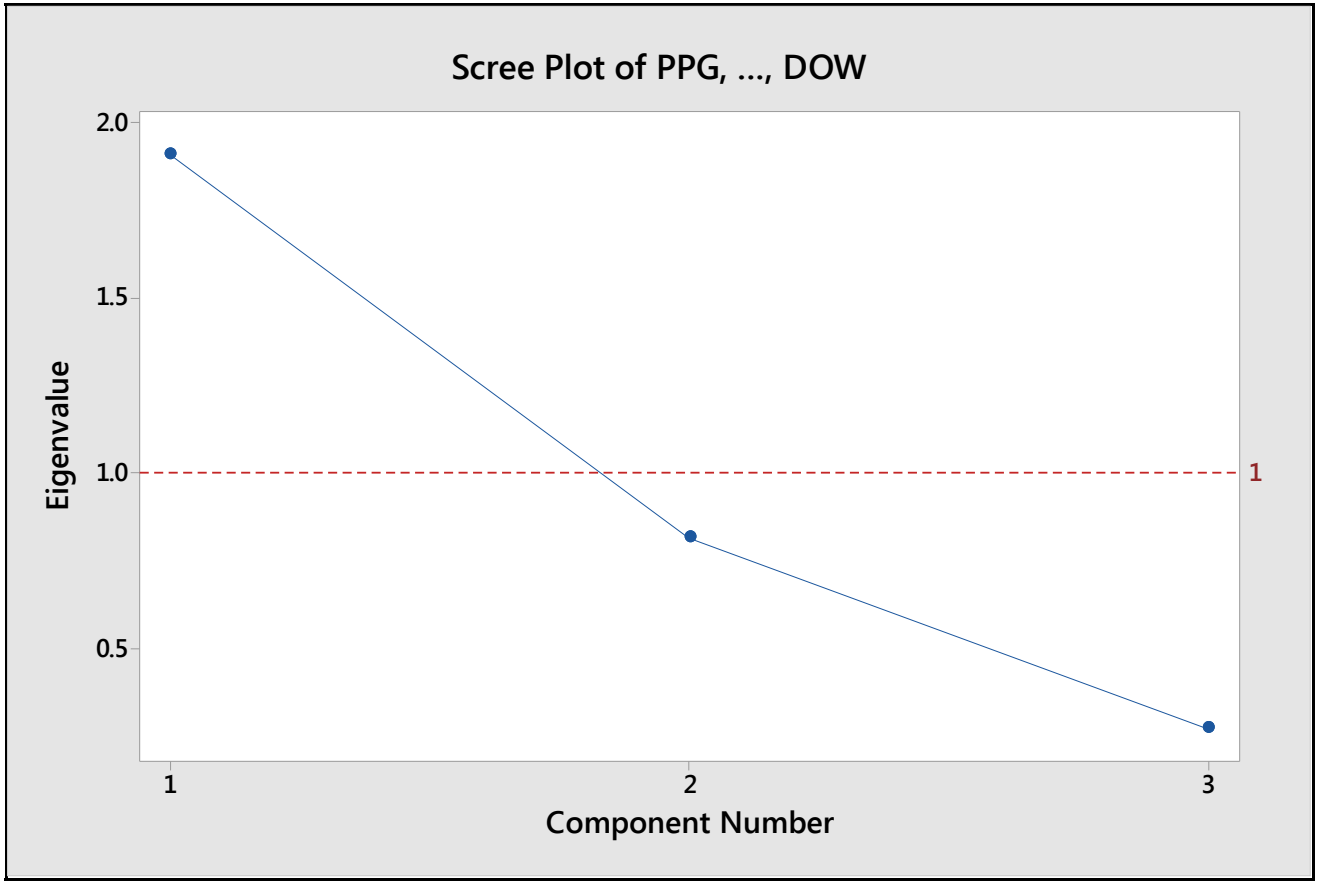
We can add percent explanations because Z1, Z2 and Z3 are uncorrelated.

Cumulative Percent Explanation of Variance in individual variables

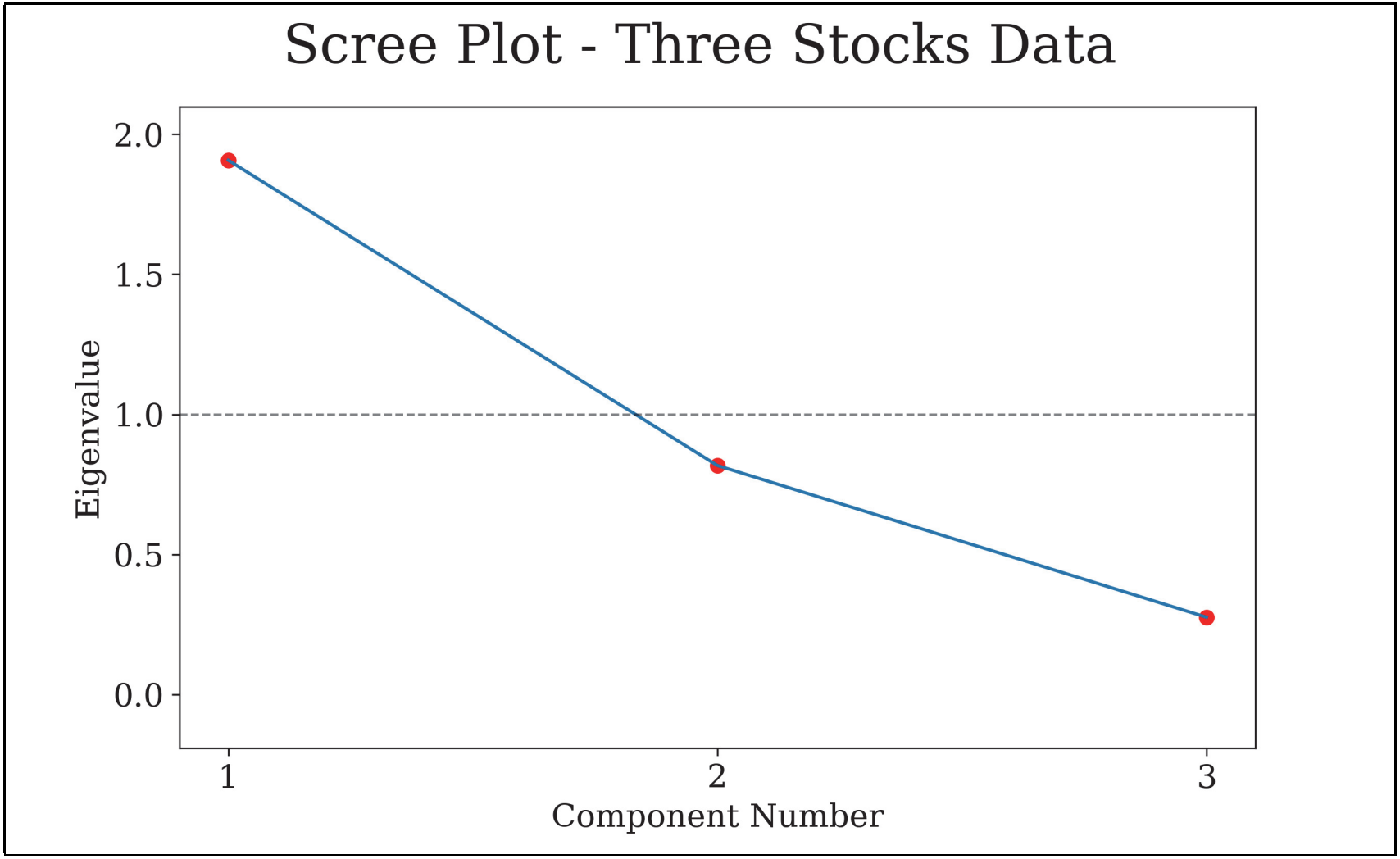
	Z1	*Z2*	*Z3*
PPG	82.1%	85.4%	100.0%
EXXON	33.6%	99.5%	100.0%
DOW	74.9%	87.5%	100.0%

Row sums to 100%

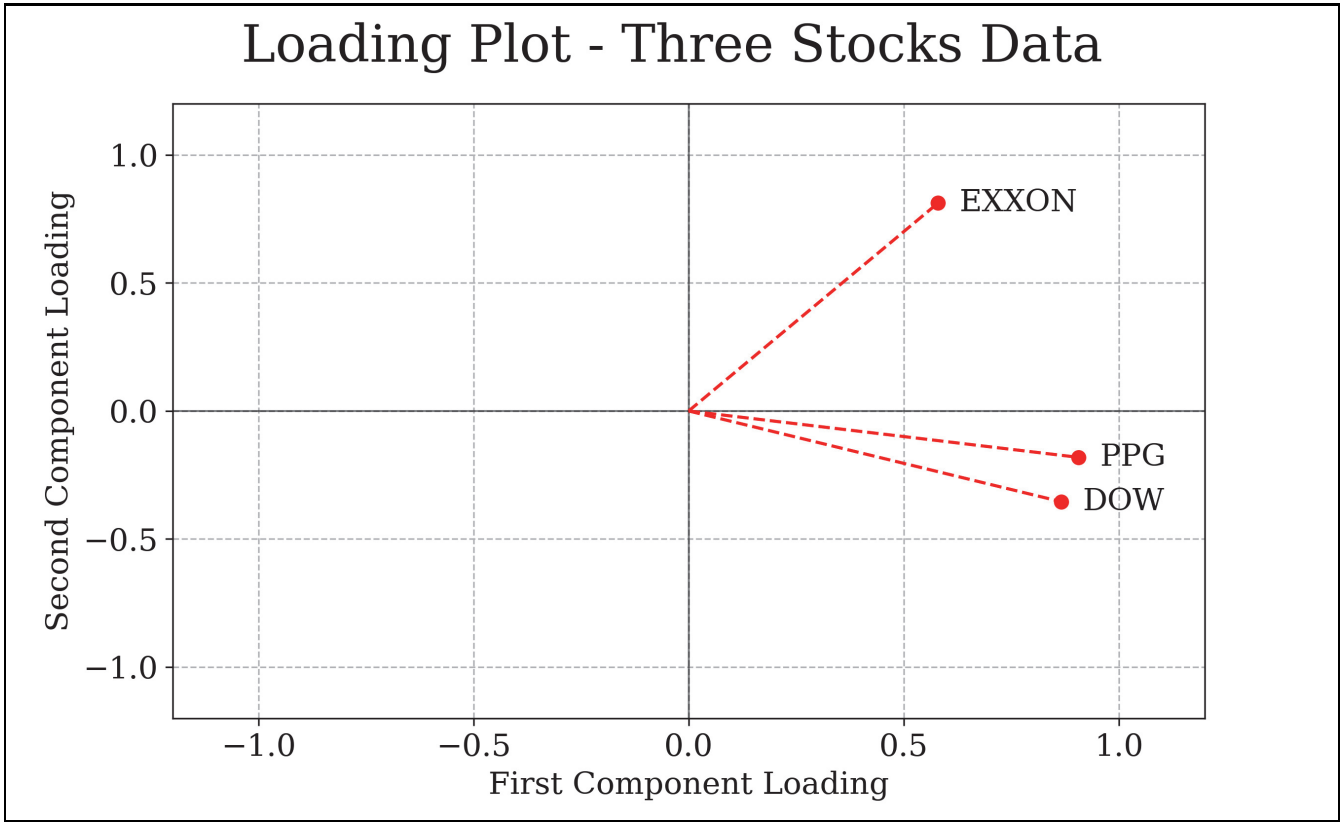
Total Variance in individual standardized variable equals one.



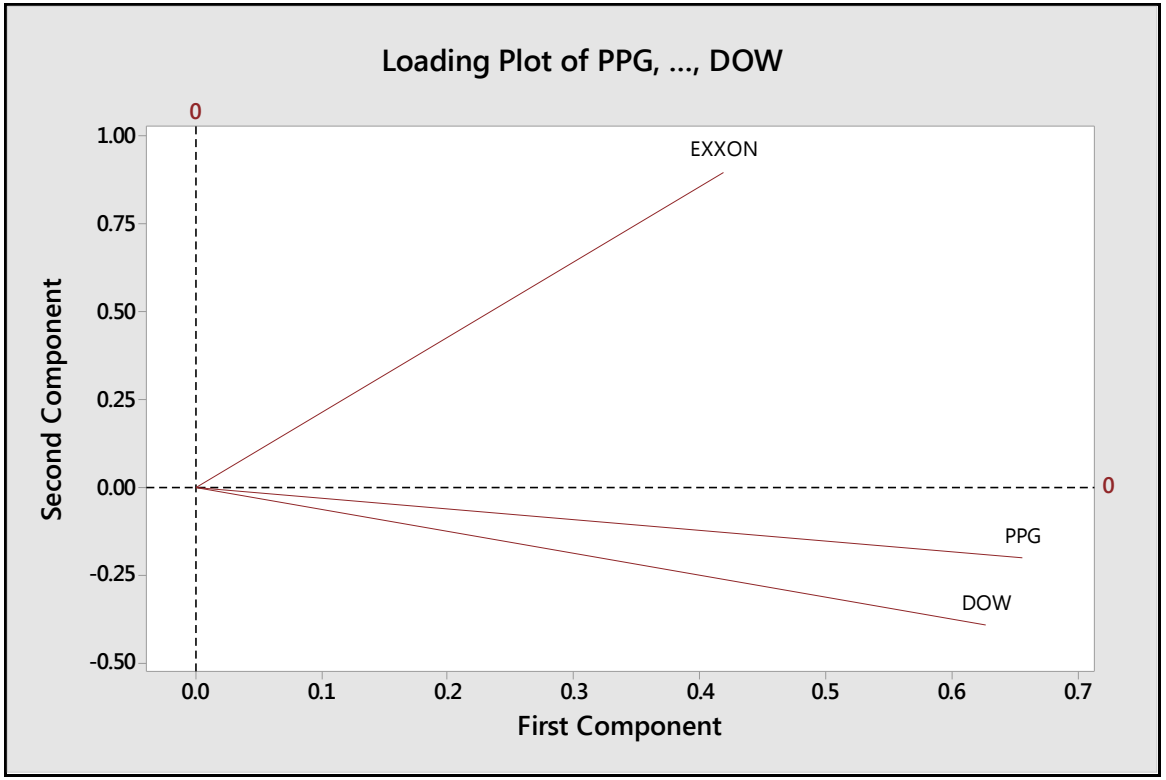
- **Scree Plot:** You **typically only retain** the principle components with **a variance (i.e. eigenvalue) larger than one**, or one less than the elbow of **the scree plot**.



- **Scree Plot:** You typically retain the principle components with a variance (i.e. eigenvalue) larger than one, or one less than the elbow of **the scree plot**.



- **Loading Plot (in Python):** The first principle component is **a market component**. The second principle component is **a classification component**. For example, if the stock of another company has **a negative correlation** with the Z_2 scores that company behaves like a **"chemical company"** otherwise it behaves like an **"oil company"**.



- **Loading Plot (in Minitab):** The interpretation is the same, but Minitab plots as "Loadings" the **elements of the coefficient vectors $\underline{u}_1$ and $\underline{u}_2$** and **not the correlation of the re-expressed Z-Scores with the original X-data.** As such, the Minitab loading plot **does not take the variances of the Z-Scores into account.**

- Start with a standardized $(n \times p)$ data matrix $\boldsymbol{X} = (\boldsymbol{x_1} \quad \boldsymbol{x_2} \quad \ldots \quad \boldsymbol{x_p})$
- Find the first principal component $\boldsymbol{z}$ as a linear combination of the original variables such that it has maximum variance.

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{u},\ \boldsymbol{u}^T = (u_1 \quad u_2 \quad \ldots \quad u_p)$$

- We have $E[\boldsymbol{z}] = \underline{0}$, because the data matrix $\boldsymbol{X}$ is **standardized** and thus

$$Var(\boldsymbol{z}) = \frac{1}{n-1}\boldsymbol{z}^T\boldsymbol{z} = \frac{1}{n-1}\boldsymbol{u}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{u}$$

- But:

$$\frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X} \text{ is a } (p \times p) \text{ matrix}$$

and is **the sample correlation matrix $\boldsymbol{R}$** of the data matrix $\boldsymbol{X}$, **because $\boldsymbol{X}$ is a standardized data matrix**.

- For any square-matrix A, a vector $\boldsymbol{x}$ that satisfies $A\boldsymbol{x} = \lambda\boldsymbol{x}$, (i.e. the matrix product $A\boldsymbol{x}$ is a scalar product of the vector $\boldsymbol{x}$ itself) is called **an eigen-vector of the matrix A. The scalar λ is called the eigen-value.**

- **A square $(p \times p)$ positive definite matrix A** (recall that the variance covariance matrix and the correlation matrix are both positive definite, i.e. $\boldsymbol{x}^T\boldsymbol{\Sigma}\boldsymbol{x} > 0$, for any vector $\boldsymbol{x} \neq \boldsymbol{0}$), has **$p$ eigenvectors that are mutually orthogonal with strictly positive eigenvalues $\lambda_i > 0$.**

- Note that:

$$A\boldsymbol{x} = \lambda\boldsymbol{x} \Leftrightarrow A(-\boldsymbol{x}) = \lambda(-\boldsymbol{x})$$

- Hence, if $\boldsymbol{x}$ is an eigen-vector so is $-\boldsymbol{x}$ and with the same eigen-value as $\boldsymbol{x}$.

- **Conclusion: Eigen-vectors are unique up to their sign**. Hence, when using software programs to determine eigenvectors (or principal components) **the answers may differ in sign only.** This **should be taken into account when interpreting the principal components.**

- It can now be shown that **the coefficients for the p principal components of a standardized $(n \times p)$ data matrix $\boldsymbol{X}$** are the **p eigenvectors of the sample correlation matrix $\boldsymbol{R}$**, where:

$$\boldsymbol{R} = \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X}$$

Homework 1: Verify that the following vectors are the eigenvectors of correlation matrix $\boldsymbol{R}$ in our example using matrix-vector multiplication in Excel.

u1	u2	u3
0.656	-0.200	-0.727
0.420	0.898	0.131
0.627	-0.392	0.673

Homework 2: Verify that the following eigenvalues are associated with the eigenvectors above.

Eigenvalues	1.907	0.817	0.276

- **It is interesting to note that the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ are exactly equal to the variances for the three principal components**. Could this be a coincidence? The answer is of course: No!

- Recall that: $\boldsymbol{Z_i} = X\boldsymbol{u_i}$, $\boldsymbol{R}\boldsymbol{u_i} = \lambda_i \boldsymbol{u_i}$, $\|\boldsymbol{u}_i\| = 1 \Leftrightarrow \boldsymbol{u_i^T u_i} = 1$
  Hence:

$$Var(\boldsymbol{Z_i}) = \frac{1}{n-1}\boldsymbol{u_i^T X^T X u_i} = \boldsymbol{u_i^T R u_i} = \boldsymbol{u_i^T}\lambda_i \boldsymbol{u_i} = \lambda_i \boldsymbol{u_i^T u_i} = \lambda_i$$

- **Because the principal components are orthogonal,** they are uncorrelated, it therefor follows that:

$$Var(\sum_{i=1}^{p} \boldsymbol{Z}_i) = \sum_{i=1}^{p} Var(\boldsymbol{Z}_i) = \sum_{i=1}^{p} \lambda_i = p$$

- **When using the correlation matrix $\boldsymbol{R}$** to derive the principal components (and not the covariance matrix), it may be shown that **sum above totals to $\boldsymbol{p}$.**

$\boldsymbol{Z} = (Z_1, Z_2, Z_3)$: The principal component scores, $Z = \boldsymbol{XU}$,

$\boldsymbol{U} = (\boldsymbol{u}_1 \quad \boldsymbol{u}_2 \quad \boldsymbol{u}_3)$: The principal coefficients or eigenvectors.

- **The principal component loadings** $F = Corr(\boldsymbol{X}, \boldsymbol{Z})$ are obtained by multiplying **each eigen vector $\boldsymbol{u}_i$ by the scalar $\sqrt{\lambda_i}$, i.e. the standard deviation of the i-th principal component.**

- **Denoting $\boldsymbol{f}_{ij}$ as the ij-th element of the matrix** $F = Corr(\boldsymbol{X}, \boldsymbol{Z})$ (the correlation between X_i and the principal component Z_j), **the amount of variance in variable $\boldsymbol{X}_i$ accounted for by the principal components $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_c$ equals:**

$$\sum_{j=1}^{c} f_{ij}^2$$

- When $c = p$, (i.e. the number of principal components is equal to the number of variables), it may be shown that:

$$\sum_{j=1}^{p} f_{ij}^2 = 1$$

- **The monthly rates of return for five stocks** (PPG, DOW, PRAXAIR, EXXON, BP) listed on the New York Stock Exchange were determined for the period April 2017-March 2006. **The weekly rates of return are defined as**

$$\frac{Current\ Friday\ Closing\ Price\ -\ Previous\ Friday\ Closing\ Price}{Previous\ Friday\ Closing\ Price}$$

- The observations in 133 succesive weeks appear to be distributed **without autocorrelation**, but **the rates of return accross stocks are positively correlated**, since, as one expects stocks tend to **move together in repsonse to general economic conditions.**

- Lets open the spreadsheet "Stock_5_Raw_Data.xlsx" and perform a PCA analysis using **only the correlation matrix in Minitab** of these rate of returns.

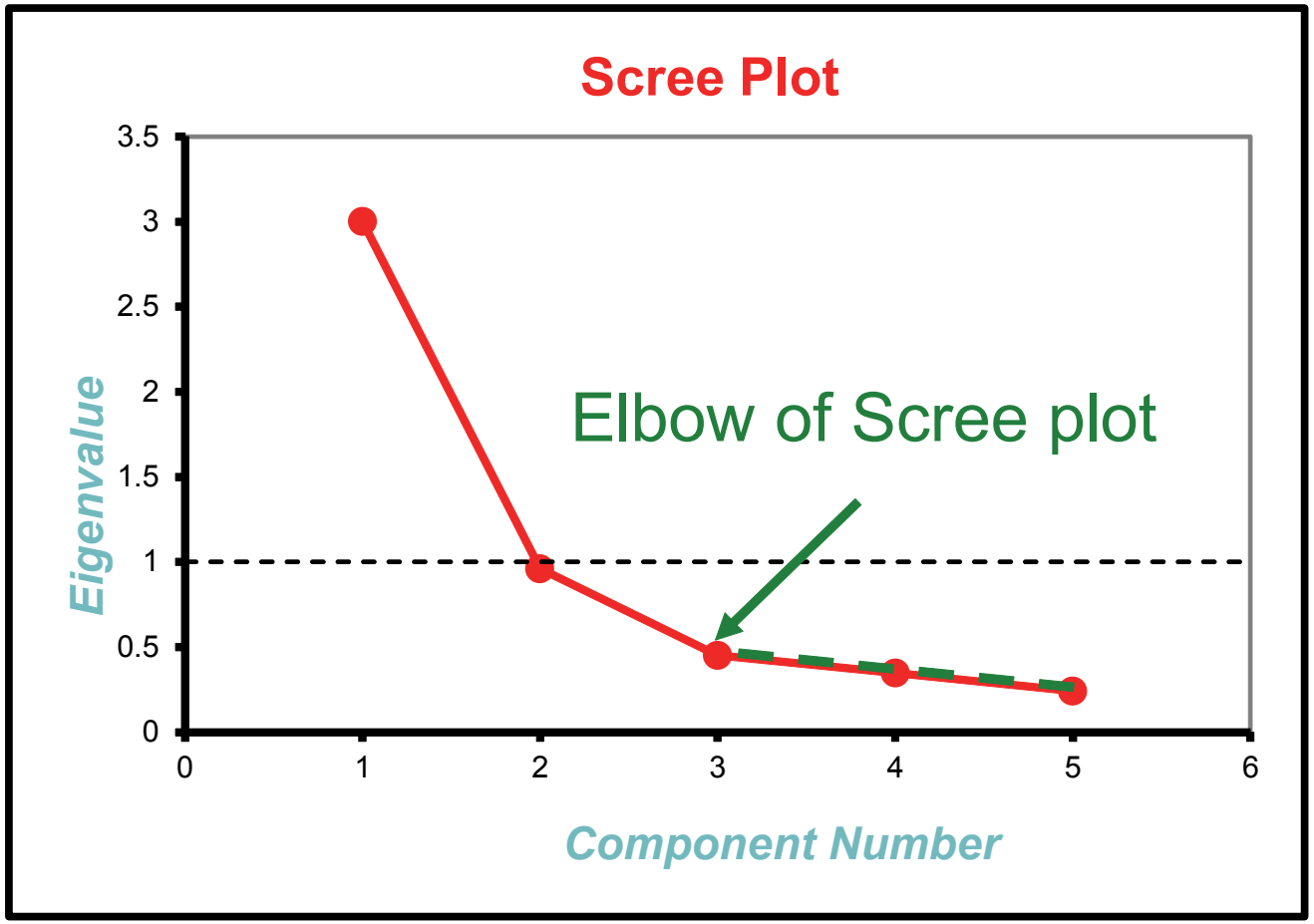
Correlation Matrix:

133	PPG	DOW	PRAXAIR	EXXON	BP
PPG	1.000	0.714	0.688	0.352	0.460
DOW	0.714	1.000	0.542	0.239	0.415
PRAXAIR	0.688	0.542	1.000	0.446	0.494
EXXON	0.352	0.239	0.446	1.000	0.610
BP	0.460	0.415	0.494	0.610	1.000

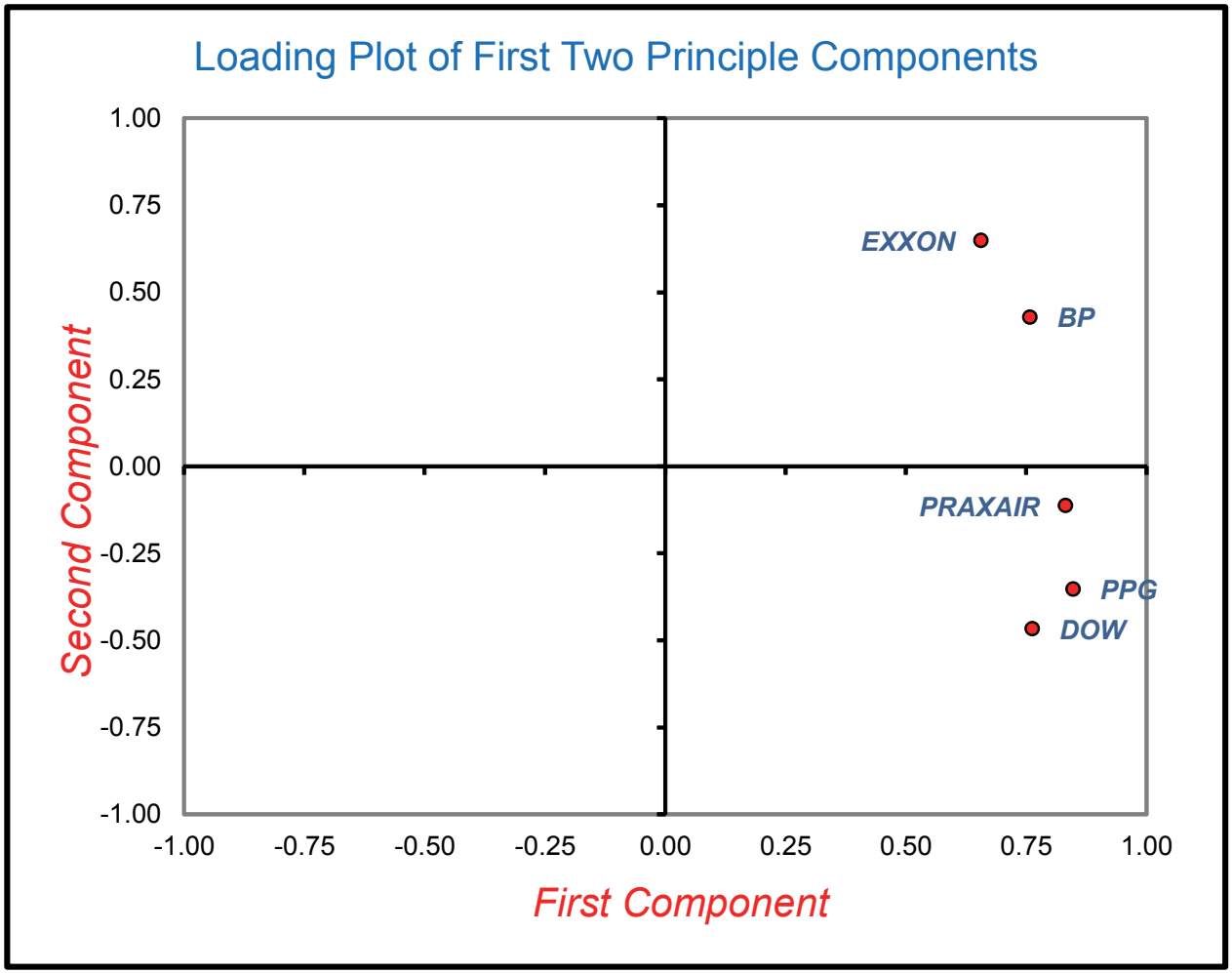
	Z1	Z2	Z3	Z4	Z5	Total
Eigenvalues	3.001	0.9604	0.4508	0.3468	0.2409	5.00
Percentage	60.0%	19.2%	9.0%	6.9%	4.8%	
Cumulative	60.0%	79.2%	88.2%	95.2%	100.0%	

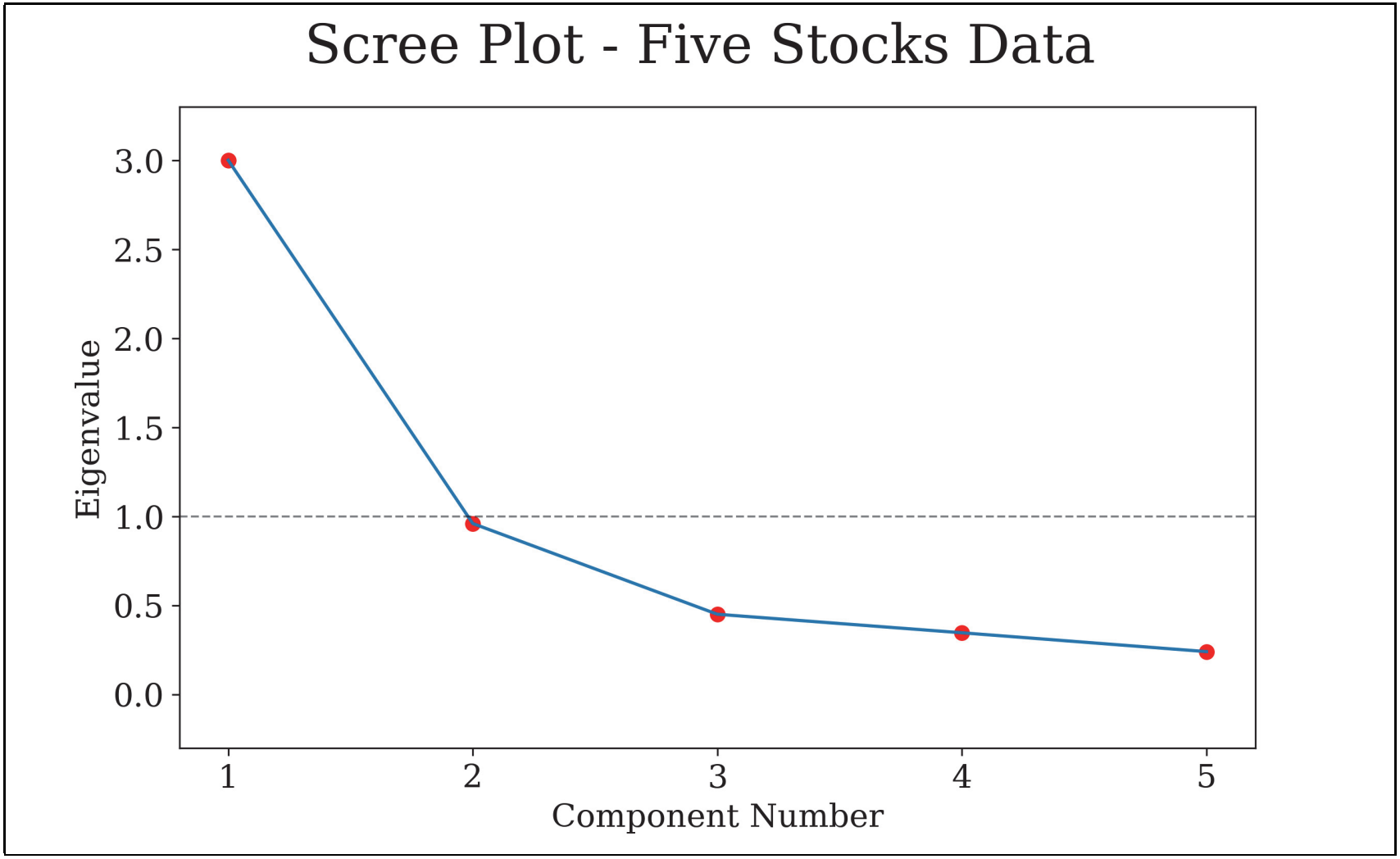
	Loadings	Z1	Z2	Z3	Z4	Z5
X1	PPG	0.848	-0.354	0.066	-0.073	0.382
X2	DOW	0.763	-0.467	-0.323	-0.193	-0.240
X3	PRAXAIR	0.832	-0.113	0.456	0.227	-0.187
X4	EXXON	0.657	0.649	0.113	-0.366	-0.025
X5	BP	0.759	0.428	-0.348	0.344	0.042

Scree Plot in MS Excel

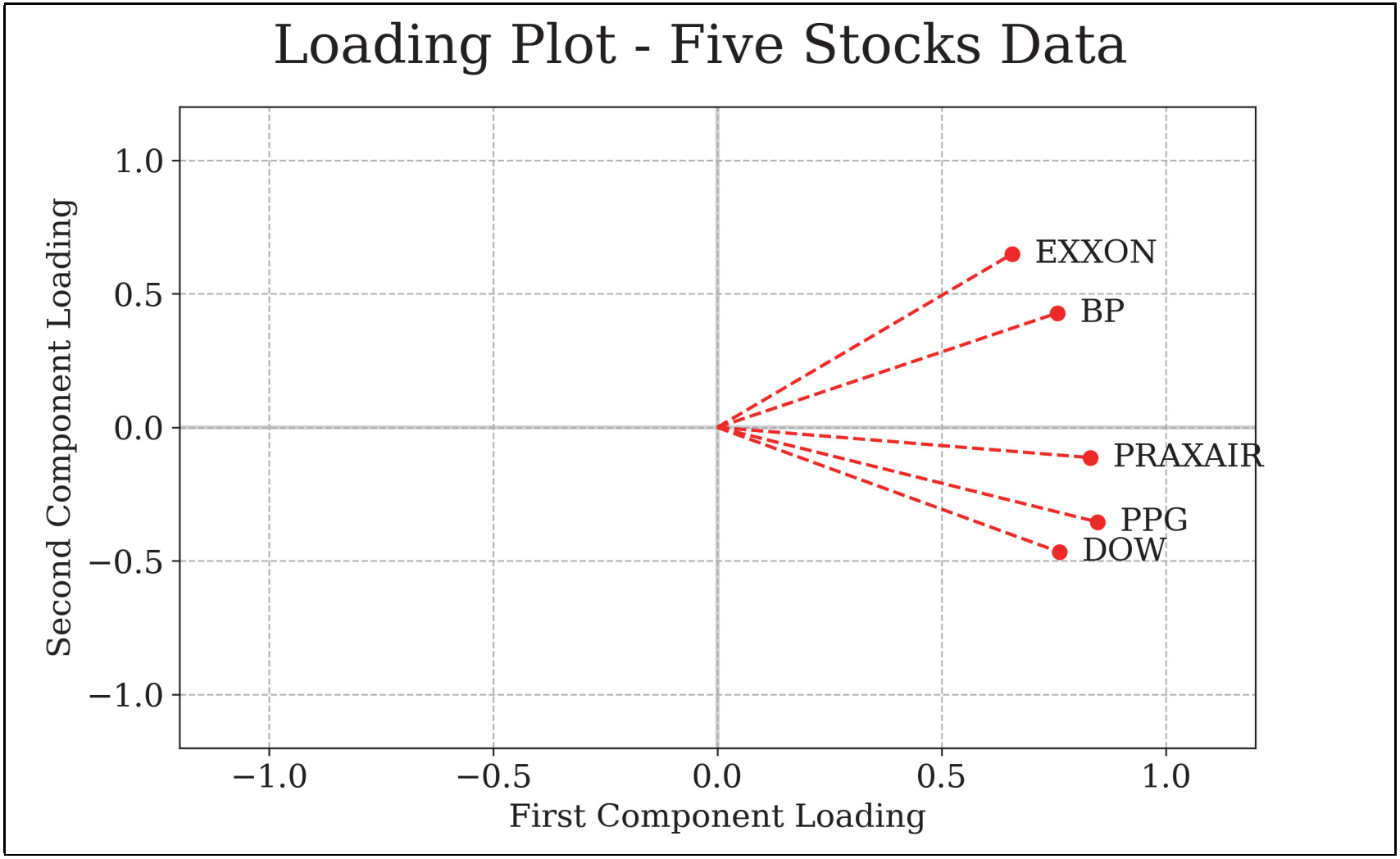


Loading-Plot in Excel





Analysis in "Stocks_5_PCA.py"



Analysis in "Stocks_5_PCA.py"

Conclusion:

From the Scree above it follows that the dependence amongst the groups of five stock can be explained by two components. The first one being a general market component (note that the sign of the loadings on the first component is the same amongst all stocks) and the second one being an industry component. The two sub-groups that could be named are **"Chemical" (*DOW, PRAXAIR* and *PPG*)** and **"Oil" (*EXXON* and *BP*).**

Note that the sign of the loadings for the second component is opposite for those in the group Chemical than those in the group Oil. **Hence, from a portfolio diversification point of view it may make sense to have both stocks of the group Chemical and Oil in ones portfolio as this could reduce port folio volatility**. This behavior **is not clear from the correlation matrix itself since the correlations are over shadowed by the dependency on the first principal component (the general marker component).**

One word of caution is in order: The variance of the second component is rather small (the eigenvalue is only 0.9604). Hence, the **Kaiser's rule would not retain** this particular component. That said, it does allow us to classify two industry sub-groups. Note that **the scree plot does suggest to retain two components.**