2022

The practical application of Singular Value Decomposition through Principal Component Analysis

AN INFORMAL REPORT STYLE

JAYDEN DZIERBICKI

Introduction

This paper demonstrates Principal component analysis (PCA) in MATLAB through the practical application of Singular Value Decomposition (SVD). In linear algebra SVD is a form of matrix decomposition, which essentially breaks matrix A into three elements all with unique properties as summarised by the equation 1.

Equation 1: SVD of Matrix A

$$A = U\sum V^T$$

This suggests that the matrix A can be explained by three matrices which represent the following:

- $\sum$: The diagonal matrix with 'singular' values its diagonal, these are analogous to the eigenvalues of the eigenvector decomposition
- U: A matrix whose column vectors are orthogonal, also referred to as 'left singular vectors'
- V: A matrix whose column vectors are orthogonal, also referred to as 'right singular vectors'

The application of SVD expends into many practical applications, though we will explore equation 1 as it relates to PCA. The application of PCA allows for dimension reduction through the principal components (PC) it produces, with application applicable to only numeric qualitative data. We summarise SVD as it relates to PCA in table 1.

Method

Data was obtained from the JCU MA5801 content page and downloaded from the assessment 4 folder on 27th June 2022, labelled "SoTCombined2010.xlsx". The data was saved locally and imported into MATLAB (*Version 9.10.0.1669831 (R2021a) Update 2*) via the readtable() function, with the MATLAB code supplied in appendix A.

Method: Data exploration

We plotted the relationship of all variables against one another by utilising a double loop in MATLAB, this was achieved by utilising a combination of the figure(), subplot() and scatter() functions; and in addition produced a heat plot by calculating the correlation of all numerical variables.

Method: Remove Missing observations:

The data was then processed to remove any missing observations by calling the rmmissing() function, and qualitative data was removed in preparation for PCA, such as the name of the country.

Method: Data preparation for PCA:

Furthermore, the raw quantitative data was centred and scaled by calling on equation 2, this was achieved by generating a loop in MATLAB (appendix A). The output was then confirmed and compared against similar output by calling on MATLABs inbuilt function normalize(), to confirm the correct implementation of the loop after being made aware of such function.

Equation 2: Used to centre and scale data prior to application of PCA.

$$\mathbf{X}_i = \frac{\widetilde{\mathbf{X}}_i - \boldsymbol{\mu}_i}{\sqrt{\alpha_i}}$$

Method: Perform SVD

We then performed SVD by calling on the svd() function via equation 3.

Equation 3: SVD function used in MATLAB to generate SVD values for PCA.

$$[U, S, V] = svd(X)$$

The function in equation 3 generates three new outputs in MATLABs workspace as summarised in table 1. We then combined variable names onto the vector V for ease of analysis.

Table 1: Summary of output from equation 2

Output	Description
U	Principal directions
S (Equivalent to ∑from equation 1)	Variance of principal components
V	Principal component vectors

Method: Determine number of PC to retain

One of the most used methods used in practical application to determine the number of PC is the elbow method (Zhuang, Wang & Ji, 2022). We determined the initial number of PC by calling on the plot() function and plotting the ratio of each principal component against the sum of weights, and visually inspected the graph to select the cut off at the 'elbow'.

In order to further refine our model, we undertook analysis of residuals by plotting the residuals. This was achieved by generating the reduced matrix as seen in equation 4, with code supplied in appendix A; and analysing the matrix of scores by utilising the find() function in MATLAB by finding scores with a magnitude exceeding 1 in PC3 through to PC13; and also referring to the current literature in our decision making process.

Equation 4: Equation used to obtain a dimensionally reduced representation of the data set, which will be used for analysis of the residuals

$$\hat{\mathbf{X}} = \mathbf{U}\hat{\mathbf{S}}\hat{\mathbf{V}}^T$$

where $\hat{\mathbf{S}}$ and $\hat{\mathbf{V}}$ are the first k columns of $\mathbf{S}$ and $\mathbf{V}$, respectively.

Method: Calculate the Matrix of Scores

We calculate the matrix of scores by equation 5 in MATLAB.

Equation 5: Matrix of scores equation

$$T = U * S$$

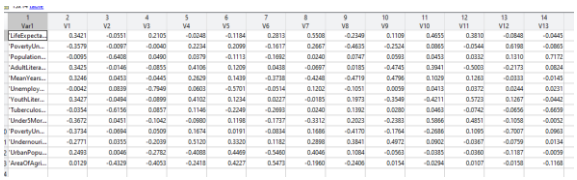
Results & Discussion

Data exploration

As seen in figure 7 (appendix A) we plotted life expectancy on the Y-axis for all numerical variables and observe many variables with obvious linear relationships with life expectancy (variable 2 & 3 for example), as well as non-linear relationships (possibly variable 12), and no clear relationship (variable 8). In addition, we observe many variables have moderate to strong correlations as seen in the heatmap in figure 8 (appendix A). We can see in figure 8 for example that life expectancy is heavily negatively correlated with Mortality age under 5 per 1000 births (-0911); whilst life expectancy has a very weak/almost no correlation with tuberculosis cases (-0.00366). This observation is an interesting observation on face value and is visible in the plot of figure 7 looking at variable 8 as mentioned previously, whilst not explored here a transformation such as a log transformation could have been explored.

Results & Discussion: PCA

We review the various principal components as found in V. Due to these values being constructs of linear combinations of the initial variables the output of V doesn't have any 'real' meaning in relation to the original variables. Though, these values from an absolute numerical point of view are equal to the coefficients of the variable and provide information about which variables which give the largest contribution to the specific components. Values with a high absolute value describe strong correlation; and the sign indicating if that correlation is positive or negative (Taskesen, 2022). As seen in figure 1 we have the output of all the PC from 1 through to 13.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	Var1	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
1	LifeExpecta...	0.3421	-0.0551	0.2105	-0.0248	-0.1184	0.2813	0.5308	-0.2349	0.1109	0.4655	0.3810	-0.0848	-0.0445
2	PovertyUn...	-0.3579	-0.0097	-0.0040	0.2234	0.2099	-0.1617	0.2667	-0.4635	-0.2524	0.0865	-0.0544	0.6198	-0.0865
3	Population...	-0.0095	-0.6408	0.0490	0.0379	-0.1113	-0.1692	0.0240	0.0747	0.0593	0.0453	0.0332	0.1310	0.7172
4	AdultLitera...	0.3425	-0.0146	-0.0855	0.4106	0.1209	0.0438	-0.0697	0.0185	-0.4745	0.3941	-0.5003	-0.2173	0.0824
5	MeanYears...	0.3246	0.0453	-0.0445	0.2829	0.1439	-0.3738	-0.4248	-0.4719	0.4796	0.1029	0.1263	-0.0333	-0.0145
6	Unemploy...	-0.0042	0.0839	-0.7949	0.0603	-0.5701	-0.0514	0.1202	-0.1051	0.0059	0.0413	0.0372	0.0244	0.0231
7	YouthLiter...	0.3427	-0.0494	-0.0899	0.4102	0.1234	0.0227	-0.0185	0.1973	-0.3549	-0.4211	0.5723	0.1267	-0.0442
8	Tuberculos...	-0.0354	-0.6156	0.0857	0.1146	-0.2249	-0.2693	0.0240	0.1392	0.0280	0.0483	-0.0742	-0.0656	-0.6659
9	Under5Mor...	-0.3672	0.0451	-0.1042	-0.0880	0.1198	-0.1737	-0.3312	0.2023	-0.2283	0.5866	0.4851	-0.1058	-0.0052
10	PovertyUn...	-0.3734	-0.0894	0.0509	0.1674	0.0191	-0.0834	0.1686	-0.4170	-0.1764	-0.2686	0.1065	-0.7007	0.0963
11	Undernouri...	-0.2771	0.0355	-0.2039	0.5120	0.3320	0.1182	0.2898	0.3841	0.4972	0.0902	-0.0367	-0.0759	0.0134
12	UrbanPopu...	0.2493	0.0046	-0.2782	-0.4088	0.4469	-0.5460	0.4046	0.1084	-0.0563	-0.0385	-0.0360	-0.1187	-0.0059
13	AreaOfAgri...	0.0129	-0.4329	-0.4053	-0.2418	0.4227	0.5473	-0.1960	-0.2406	0.0134	-0.0294	0.0107	-0.0158	-0.1168

Figure 1: Principal components 1 through to 13 of PCA analysis performed in MATLAB.

For analysis, we will define the coefficient magnitude as either 'strong', 'moderate' or 'weak' (Azid, 2015), in addition we will not report on values not categorised in these categories (Azid, 2015) for the purpose of this paper, and will ideally focus on strong or moderate associations.

- > 0.75 strong
- 0.5 – 0.75 moderate
- 0.3 – 0.49 weak

We provide a quick summary of the first two principal components, and which variables provide the 'most' importance for the corresponding component

- Principal component 1
    - Youth literacy (0.34)
    - Adult literacy (0.34)
    - Life expectancy (0.32)
    - Means year of schooling (0.32)
    - Poverty under 1PerDay (-0.35)
    - Under 5 mortality (-0.36)
    - Poverty under 2PerDay (-0.37)

Whilst these are not strong, it suggests that the first PC is associated weakly with many socioeconomic indictors, both positively and negatively. This suggests that the positive indicators and negative indicators will move in opposite directions, and those indictors with the same sign (+/-) will move in the same direction together. As all the magnitudes are relatively the same/close together this principal component we consider generically based on socio-economic indictors.

- Principal component 2
    - PopulationUnder15 (-0.64)
    - Tuberculosis cases (-0.61)
    - Area of Agriculture (-0.43)

We have 2 moderate and 1 weak variable associated with the second PC, this suggests that the second principal component is associated negatively with a country's population under 15; and negatively with tuberculosis cases, with little connection between other variables. This suggests that these variables will move in the same direction given the negative sign and has little influence from the other variables. As seen in the heatmap in figure 8 we observe that tuberculosis (variable 8) is heavily correlated with population under 15 (variable 3) with a value of 0.976; and then moderately correlated with area of agriculture (variable 13), with poor correlations with the other variables, suggesting the relationship between the heatmap and PC2 for variable 8. This pattern is also observant for PC1 in relation to the heatmap if we look at the correlations of variable 1.

Results & Discussion: Determine number of PC to retain & Matrix of Scores

One of the limitations of the elbow method is it is visual method used to determine the number of PC to retain; and can be slightly subjective. In addition, there are various other methods which when employed can yield different results to the elbow method (Zhuang, Wang & Ji, 2022), such as Kaiser's rule, which sets the cut off for eigenvalues greater than 1 (Braeken & van Assen, 2022). As seen in figure 3 if we employed the elbow method we would retain the first three principal components which explain only 75% of total variation in the dataset, whilst with kaiser's rule would retain

the first 12 as seen in figure 2.

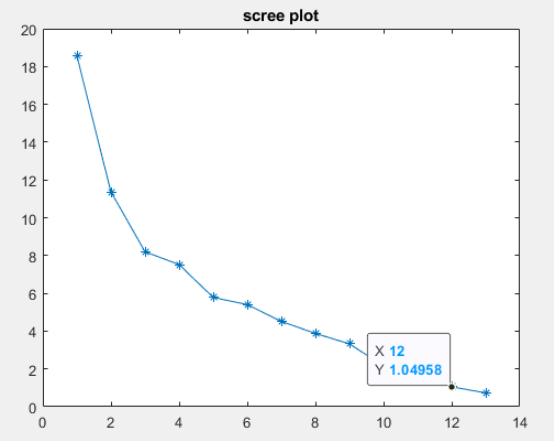


Figure 2: Scree plot, plotting value of eigenvalues, we observe all eigenvalues through to 12 exceed 1 in relation to Kaiser's rule

What we can see visually in figure 3 is that relatively speaking, after the third PC the contribution of PC's becomes smaller in comparison to the first three, relatively speaking.

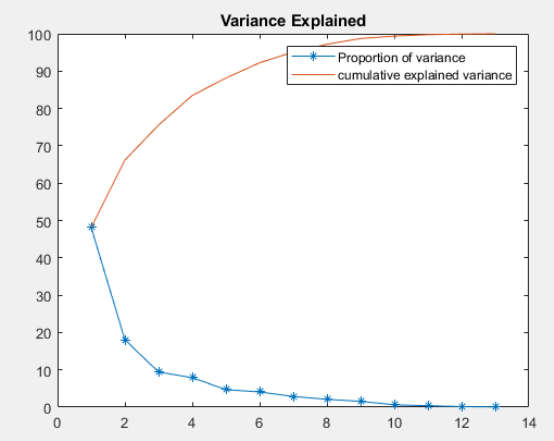


Figure 3: Plot of proportion of variance and cumulative explained variance for each principal component. We observe with the elbow method we would select the first 3 principal components

This was one of the first issues we ran into when determining the number of principal components to retain, two different methods widely used in the literature yielding two very different results. We then decided to do analysis on residuals as seen in figure 4; and we observe an obvious outlier at observation 50 (**country Brazil being the most obvious outlier**) via the residual plot when selecting for the first 3 PC via the elbow method, along with residual values in excess of 1. This suggests that we need to further refine our number of PC required beyond three, and for this reason we decide to refine our model further through further analysis.

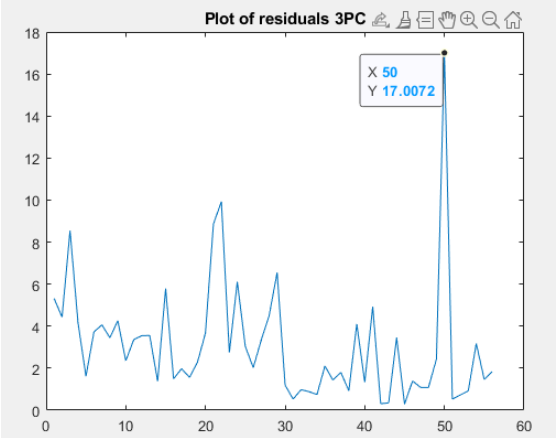


Figure 4: Plot of residuals when selecting first 3 PC. We observe large residuals based on magnitude and evidence of possible outliers, especially observation 50

When reviewing the MATRIX of scores (T) we observed that each column was largely comprised of values with a magnitude less than 1. In addition, this becomes more apparent as we increase the number of principal components as indicated by column number and can be seen in figure 5. This is because the matrix of scores help to determine how much any particular case deviates from the general trend, we know since the first PC accounts for approximately 50% of total variation in the data set (figure 3), then we expect that most cases would have larger scores in this column, which is true as seen in figure 5.

	1	2	3	4	5	6	7
16	-2.5844	0.4758	1.2981	0.7140	0.2805	0.4588	0.3495
17	-2.5552	-0.3059	-0.7117	1.1069	0.1764	0.2399	0.2018
18	-1.5741	0.7353	0.2995	-0.5591	-0.1853	-0.4157	-0.7409
19	-1.8325	0.0221	0.4885	1.0981	0.0737	0.4802	-0.6562
20	-3.1001	0.3984	-1.5679	1.3848	0.5116	-0.7853	0.0605
21	-1.5214	0.6076	-2.9543	-1.5420	-2.4356	0.2501	0.0245
22	-4.4926	0.1309	0.8042	-2.6917	-0.7177	0.9008	-0.5158
23	-1.7566	0.5830	0.1162	-1.2113	-0.6764	-0.0472	0.3706

Figure 5: Random section of matrix of scores (T) demonstrating the reduction of scores for the retention of later principal components with a greater proportion of values less than 1.

If a particular case may be an outlier, then we would expect a large score contribution from

one of the later PC, and as such utilise the find() function on the matrix T for magnitudes between 0.7 and 1, summarising the output below in table 2.

Table 2: Summary matrix of scores and cumulative variance for refinement of analysis

PC	Number > 1 in T matrix	Number > 0.7 in T matrix	% Cumulative variance	Per cent point increase
3*	16	-	75%	-
4*	18	-	83%	8
5*	7	-	88%	5
6*	4	-	92%	4
7*	5	-	95%	2.8
8	2	12	97.2%	2.2
9	3	7	98.7%	1.5
10	**1**	**2**	**99.3%**	**0.6**
11	0	0	99.7%	0.4
12	0	0	99.9%	0.2
13	0	0	100%	0.1

*Values rounded
- Not included in analysis

Based on table 2 we decide to select the first 10 principal components for our analysis. This explains a staggering 99% total variation of our model and reduces the number of observations with a magnitude greater than 1 (***Democratic Republic of the Congo > 1***) & 0.7 (***Democratic Republic of the Congo & Botswana > 0.7***) in the matrix of scores. As seen in table 2 there is no obvious evidence of outliers when selecting for T values with a magnitude greater than 0.7; we plot out the new residuals as seen in figure 6 and observe we have reduced the magnitude of the residuals down relative to figure 4, with no evidence of observation 50 being an outlier and all residuals now having a value less than 1, as most residuals appeared to exceed 1 in figure 5. We do not select the preceding PC due to diminishing returns, and no obvious outliers present when revieing the matrix of scores and selecting the threshold of |0.7|, despite Kaiser's rule which looks at eigenvalues exceeding 1 (Braeken & van Assen, 2022).

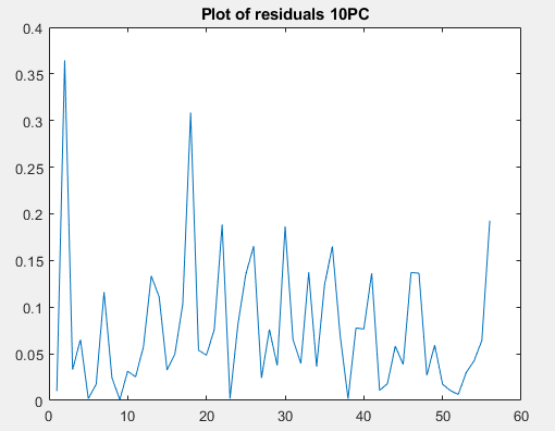


Figure 6: Plot of residuals for first 10 principal components, we see an improvement from previous residual plot due to a reduction in magnitude of residuals, and observation 50 no longer being detected as an outlier.

Conclusion

What we see from the PCA is that the first PC has a negative relationship between the positive socioeconomic indicators, and the negative socioeconomic indictors; as such we can see that an increase in positive socioeconomic indictors results in a decrease in negative socioeconomic indictors. Though, as previously discussed these values are considered weak (Azid, 2015), and additionally it is worth noting that the first PC explains up to 50% of the variation in the data. It was further noted that the second PC has a negative relationship with both a country's population under 15; and number of tuberculosis cases, with little connection between other variables. We did not analysis beyond the second PC as it was not requested in the analysis.

The analysis also raised issues of relaying on visual methodologies such as the widely used elbow method as reported in the literature, which in our example underestimated the number of principal components to retain with both large residuals (figure 4) and evidence of outliers (table 2); and additionally highlighting the use of different methods yielding different results in PC retention. Through our analysis of the residuals and matrix of scores we were able to reduce both the number of outliers (table 2)

and values of the residual plot to below 1 (figure 6), though by retaining 10 PC we have only reduced the dimension of our data by 3. This suggests that a future study would benefit at comparing and contrasting the methods shown here; and other methods available when determining the number of PC to retain, along with the benefit and weakness of such methods and both practical and theoretical implications.

Reference list

Azid, Azman. (2015). Re: How can I interpret PCA results?. Retrieved from: https://www.researchgate.net/post/How-can-I-interpret-PCA-results/566fd9425dbbbdb5df8b4567/citation/download.

Braeken, J., & van Assen, M. (2017). An empirical Kaiser criterion. Psychological Methods, 22(3), 450-466. doi: 10.1037/met0000074

Taskesen, E. (2022). What are PCA loadings and Biplots?. Retrieved 3 August 2022, from https://towardsdatascience.com/what-are-pca-loadings-and-biplots-9a7897f2e559

Zhuang, H., Wang, H., & Ji, Z. (2022). findPC: An R package to automatically select the number of principal components in single-cell analysis. Bioinformatics, 38(10), 2949-2951. doi: 10.1093/bioinformatics/btac235

Appendix A

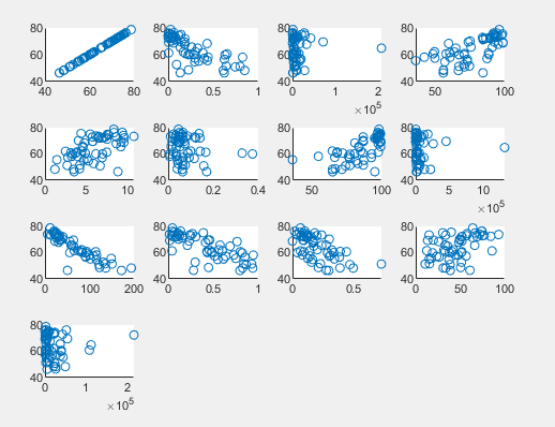


Figure 7: Y-Axis representing life expectancy against all numerical variables in order from variable 1 to variable 13 left to right, top to bottom

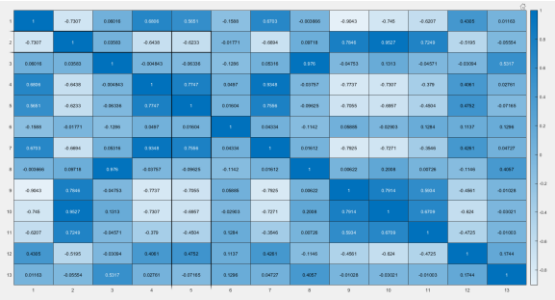


Figure 8: Heatmap of numerical variables, showing multiple correlation between the various variables

```
% This code demosntrates the prac
application of SVD by doing PCA
% Author Jayden Dzierbicki
% Read 9/08/2022

%%
clear all
%%

% Read in data input from csv
dataframe =
readtable("SotTCombined2010.xlsx")
;

%%

% Ensure rows have complete set of
observations
dataframe_2 =
rmmissing(dataframe);


% We observe that the dataframe
contains both qualitative data
(country
```

```
% name) and quantitative data,
this means we must remove column 1
from our
% analysis

%%

% Remove country name for analysis
dataframe_3 = dataframe_2(:,
2:width(dataframe_2)); % Table
version
X_bar = table2array(dataframe_3);
% Name as X-bar as data in in raw
form



%%

% Plot relationships for quick
visual analysis of relationships
(mainly
% for my own practise in matlab).
Allows to look at structure of raw
% numeric data.
for y_axis = 1:width(X_bar)
    figure() % New figure each
loop
    for x_axis = 1:width(X_bar)
        subplot(4,4, x_axis);

scatter(X_bar(:,x_axis),X_bar(:,y_
axis))
    end
end

% We observe some 'strong'
relationships in many varibles,
suggesting high
% +/- correlation




%%

% Prepeare for centring and
scalining matrix of raw data for
PCA and
% initlisise constant values used
in loop
col_mean = mean(X_bar);
col_var = var(X_bar);

%%

% Loop to normalise data for PCA
(center and scale)
X = zeros(length(X_bar),
width(X_bar));
for row = 1:length(X_bar)
    for column = 1:width(X_bar)
        X(row,column) =
(X_bar(row,column) -
col_mean(column)
)/sqrt(col_var(column));
        % Essentially loop through
and replace row/columnwise each
        % observation using
equation
    end
end

% This is an inbuilt function
allowing us to test if the above
loop worked
% results are the same - I learnt
this after developing the above
code
X_inbuilt_test = normalize(X_bar);



%%
% Produce corelation matrix after
normalising above for further
analyis
Cor = X'*X/(length(X)-1);
figure()
heatmap(Cor)
% We observe some strong/negative
correlations in the heatmap



%%

% Perform SVD using our X matrix
as now scaled and centered to
avoid
% influence of extreme values
[U,S,V] = svd(X);

% The code below will allow us to
easily intepret our PCA results as
they
% relate to the varible names. To
do this we look at the magnitude
of the
% varible in each PCA and look at
magnitude
varible_names =
cell2table([dataframe_3.Properties
.VariableNames]');
v_table = array2table(V);
PCA = [varible_names v_table];

%%

% Plot the variance and scree plot
of PCA/aide in determing number of
PC
```

```
S2 = S.^2; % Grab weights for PCA
lam = diag(S2)/sum(diag(S2))*100 %
Ratio of each PC to total sum of
weights
cum_lam=cumsum(lam); % Cumulative
sum

figure()
plot(1:13, lam, "-*" , 1:13,
cum_lam); % Plot to see amount
variance
title("Variance Explained")
legend("Proportion of variance",
"cumulative explained variance")
% We can use various methods to
deterimine number of PC to use in
our
% analyisis
figure()
plot(1:13, diag(S), "-*")
title("scree plot")

%% Scores matrix
% Once we know how many PC to
retain, we set nres = # retain
nres=10;
X_red=
U(:,1:nres)*S(1:nres,1:nres)*V(:,1
:nres)'; % Reduced matrix, based
on nres first PC
residuals = sum(((X-X_red).^2)');
% Calculate residuals based on
number PC retained
residuals_2 = sum((X -
X_red).^2,2);
figure()
plot(1:length(residuals),
residuals); % Plot residuals
title("Plot of residuals 10PC")


%%

% Compute the score matrix
T = U*S; % Hard to eyeball
currently

% All in first ignored PC
[row, col] =
find(abs(T(:,3:13))>1); % Set
equal to 50th observation from
residual plot
row;
col = col+2;

% We observe that if we select the
10th PC we have matrix of scores
all
% less then 0.7 with nothing
standing out
[row, col] =
find(abs(T(:,8:13))>0.7);
row
col = col+7
```