**Urban air pollution monitoring to check effect on Environment’s Temperature**

**Arshpreet Kaur Student ID: 40022501**

**INSE 6220 - Advanced Statistical Approaches to Quality**

***Abstract*—**Ease gas Multisensor gadgets can speak to a proficient answer for densifying the inadequate urban air contamination observing cross section. In a past work, we proposed and assessed the adjustment of such a gadget utilizing here and now on-field recorded information for the benzene contamination evaluation. In this work, we exhibit and talk about the outcomes acquired for CO, NO2 and aggregate NOx poisons focus estimation with a similar set up. Traditional air contamination observing station is utilized to give reference information. We demonstrate how a multivariate alignment can be accomplished with the utilization of two weeks in length on-field information recording and neural relapse frameworks. Additionally for these contaminations, no critical execution help was perceptible when longer accounts were utilized. The impact of a suitable element choice for accomplishing ideal exhibitions is likewise talked about contrasting long haul execution consequences of the got adjustments. Advantages and issues of multivariate relationship based adjustment are assessed amid one year long estimation battle.

***Index Terms***— Urban air pollution monitoring, On-field calibration, Multisensor device, Feature selection

**I. INTRODUCTION**

A modest random sample of 25 different Time and day is used to obtain information on the effect on temperature caused by CO, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2). 5 variables are used for each sample age which are as follows:

1. X1 – CO
2. X2- Benzene
3. X3- total nitrogen oxides(NOx)
4. X4- Nitrogen Dioxide (NO2)
5. X5- Temperature

**A) Principal components analysis (PCA**)

Principal components analysis (PCA) is an

Explanatory technique to develop an understanding about data sets. The objective of PCA to reduce the dimensionality of the data set while retaining as much as possible the variation in the data set. Principal components (PCs) are linear transformations of the original set of variables, and are uncorrelated and ordered so that the first few components carry most of the variation in the original data set. The first PC has the geometric interpretation that it is a new coordinate axis that maximizes the variation of the projections of the data points on the new coordinate axis.

The general idea of PCA is as follows: if we have a set of moderately or strongly correlated variables (i.e. the variables share much common information), it may then be possible to construct

new variables that are combinations of these variables that account for much of the original information contained in the data. These linearly uncorrelated variables are called Principle Components. The output of PCA consists of coefficients that define the linear combinations used to obtain the new variables (PC loadings) and the new variables (PCs) themselves. Examining the PC loadings and plotting the PCs can aid in data interpretation, particularly with higher dimensional data.

**PCA Algorithm**

Given a data matrix X, the PCA algorithm consists of four main steps:

1. Compute the centered data matrix by subtracting off‐column means
2. Compute the covariance matrix of the centered data matrix
3. Compute the eigenvectors and eigenvalues using Eigen‐decomposition
4. Compute the transformed data matrix which contains the coordinates of the original data in the new coordinate system defined by the PCs. This matrix will have the correspondence of observations and PC scores.

**Factor Analysis**

Factor analysis is a statistical technique used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. In other words,

it is possible, for example, that variations in three or four observed variables mainly reflect the variations in fewer unobserved variables. Factor analysis searches for such joint variations in response to unobserved latent factors. The observed variables

are modeled as linear combinations of the potential factors, plus "error" terms. The information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset.

There are two types of Factor Analysis; **Exploratory factor analysis (EFA):** It is used to uncover the underlying structure of a relatively large set of variables. A priori assumption is that any variable may be associated with any factor. This is the most common form of factor analysis. There is no prior theory and one uses factor loadings to intuit the factor structure of the data. **Confirmatory factor analysis (CFA):** It seeks to determine if the number of factors and the loadings of measured variables on them conform to what is expected on the basis of pre-established theory.

Indicator variables are selected on the basis of prior theory and factor analysis is used to see if they load as predicted on the expected number of factors. In this project, exploratory factor analysis (EFA) is implemented to analyze wages of male and female workers.Given a data matrix X, the EFA algorithm consists of four main steps:

Step 1: Compute the correlation matrix for all variables and identify variables that are not related to other variables

Step 2: Extract Factors (loadings, scores and specific variance).

X1 = a11F1 + :::: + a1mFm + e1

X2 = a21F1 + :::: + a2mFm + e2

:

Xp = ap1F1 +:::: + apmFm + ep

In matrix form it can be presented as;

Xpx1 = ApxmFmx1 + epx1

Where,

X = V ariables

F = Latent Factors

a = Factors Loadings

e = SpecificErrors

Step 3: Analyze factor loadings, scores and specific variance.

Step 4: Rotate factors (loadings and scores) and analyze the simplified results. After factor extraction it might be difficult to interpret the factors on the basis of their factor loadings. So, rotating factors is better as it make them more meaningful and easier to interpret (each variable is associated with a minimal number of factors), as un-rotated factors are typically not very interpretable (most factors are correlated with many variables).

D. Comparison of PCA and Factor Analysis

Both are dimension-reduction techniques, in the sense that they can be used to replace a large set of observed variables with a smaller set of new variables. The two methods are different in their goals and in their underlying models .In conceptual

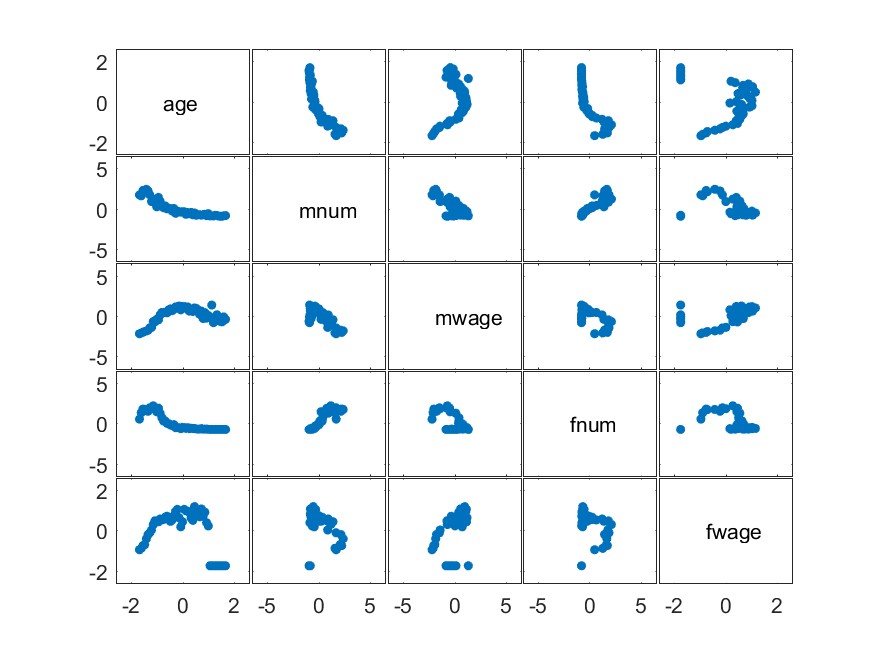
terms, the difference between PCA and Factor Analysis is that PCA analyzes variance and Factor Analysis analyzes covariance. Calculations for both PCA and Factor Analysis involve matrix algebra. Factor analysis takes into account the random error that is inherent in measurement, whereas PCA fails to do so. In respect to the correlation matrices involved in the calculations, that the difference between PCA and EFA in mathematical terms is found in the values that are put in the diagonal of the correlation matrix. In PCA,

1.00s are put in the diagonal meaning that all of the variance in the matrix is to be accounted for (including variance unique to each variable, variance common among variables, and error variance). That would, therefore, by definition, include all of the variance in the variables. In contrast, in EFA, the communalities are put in the diagonal meaning that only the variance shared with other variables is to be accounted for (excluding variance unique to each variable and error variance). That would, therefore, by definition, include only variance that is common among the variables.

**2. PCA IMPLEMENTATION**

We have standardized the data as the variance of the different columns is substantial. That is, PCA is performed on the correlation matrix instead of the covariance. When interpreting correlations it is important to visualize the bivariate relationships between all pairs of variables. This can be achieved by looking at a scatterplot matrix.

From the scatter plot of Movie factors, it can be inferred that bivariate relationship does exist between all the variables.



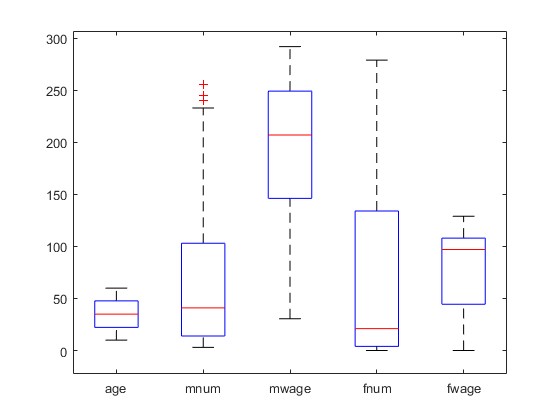
*Figure 1* ‐ *Scatterplot matrix*

From the boxplot we can easily specify that there are

outliers .There are outliers in the box plots of variable

‘mnum’ as we can see in the following graphic. The Fourth variable has the highest impact whereas, age, the first variable, has the lowest impact. Also, the means and variance of variables are not the same. Mwage has the largest mean with largest variance whereas, fnum has the lowest.

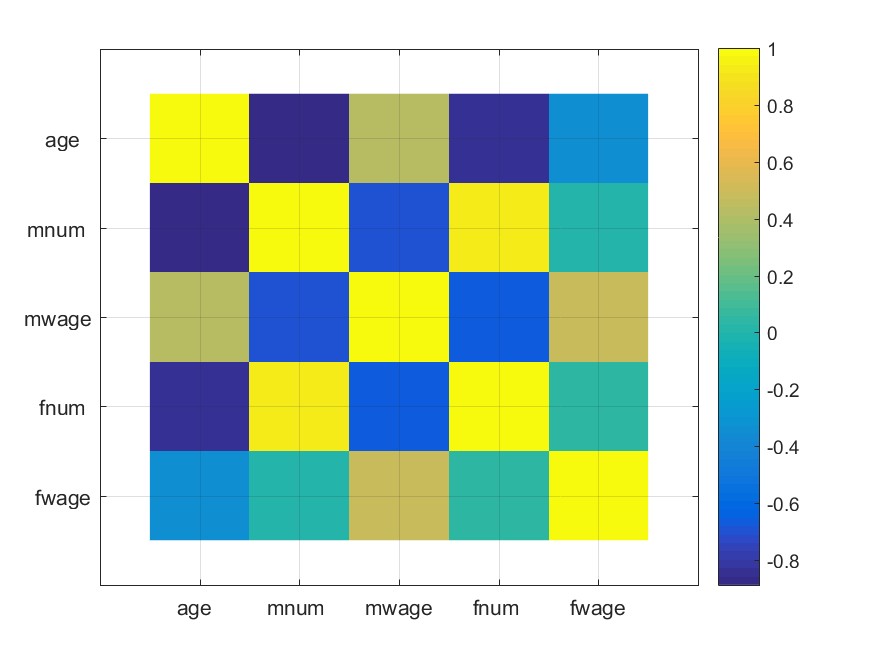
*Figure 2* ‐ *Box Plot*



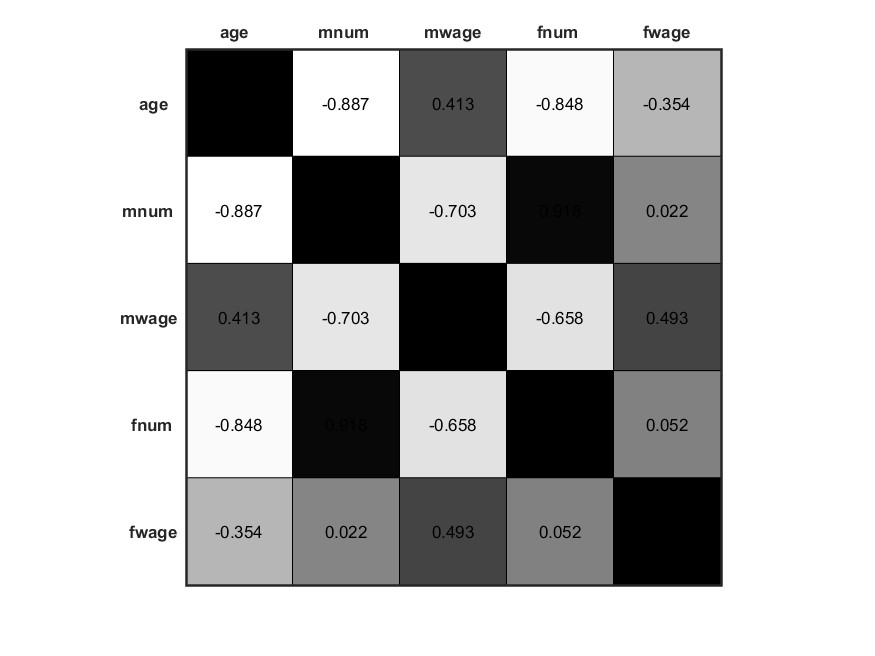
From the correlation matrix plot below, we can see

that the percentage of the variables ‘age’ is

negatively correlated with all except mwage as with it, it is positively correlated. Mnum is negatively correlated with mwage and age. Mwage is negatively correlated with mnum and fnum. Fnum is negatively correlated with mwage also. Whereas Fwage is only negatively correlated with age only. All the correlation is not very strong though. Other interesting relationships between these variables are also evident.



*Figure 3* ‐ *Correlation Matrix of Film Factors*



*Figure 4* ‐ *Correlation Matrix of Film Factor*

The scatter plot of PC2 coefficients vs. PC1

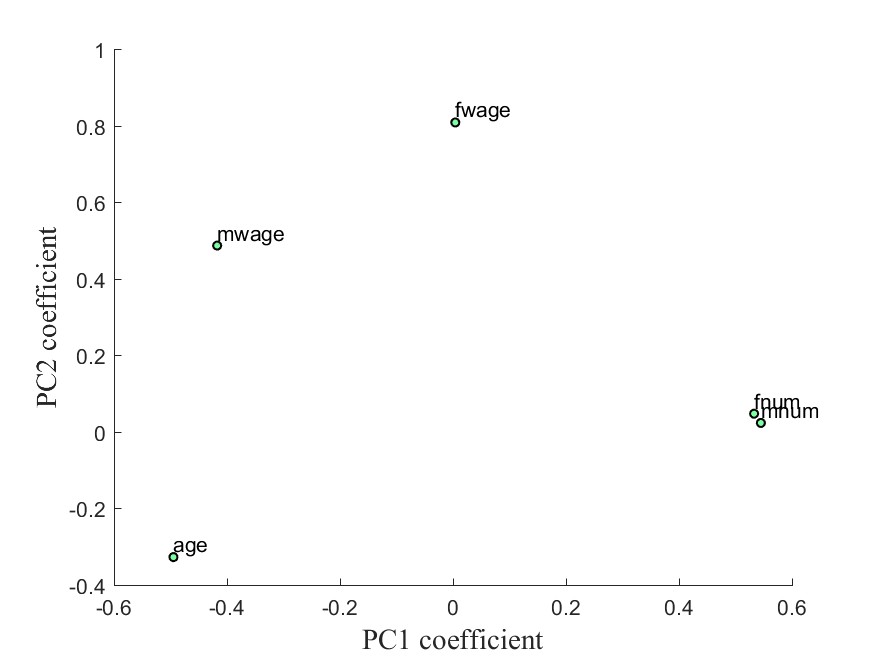
coefficients is shown in Figure 5. This plot helps understand which variables have a similar involvement within PCs.

Variable ‘fwage’ is plotted far away from the rest of the variables and has the maximum magnitude and the Variables ‘fnum’ and ‘mnum’ are plotted close to each other. It shows the strong relation of PC1 and PC2 in terms of ‘fwage’ and weak relation in terms of ‘age’

Also, Variable ‘fnum’ and ‘mnum’ are on the right

and the rest of the variables like ‘age’, ‘fwage’ and

‘mwage’ are on the left side of the plot.



*Figure 5* ‐ *PC2 vs PC1 Coefficients*

The explained variance by the three PCs are: ℓ1 =

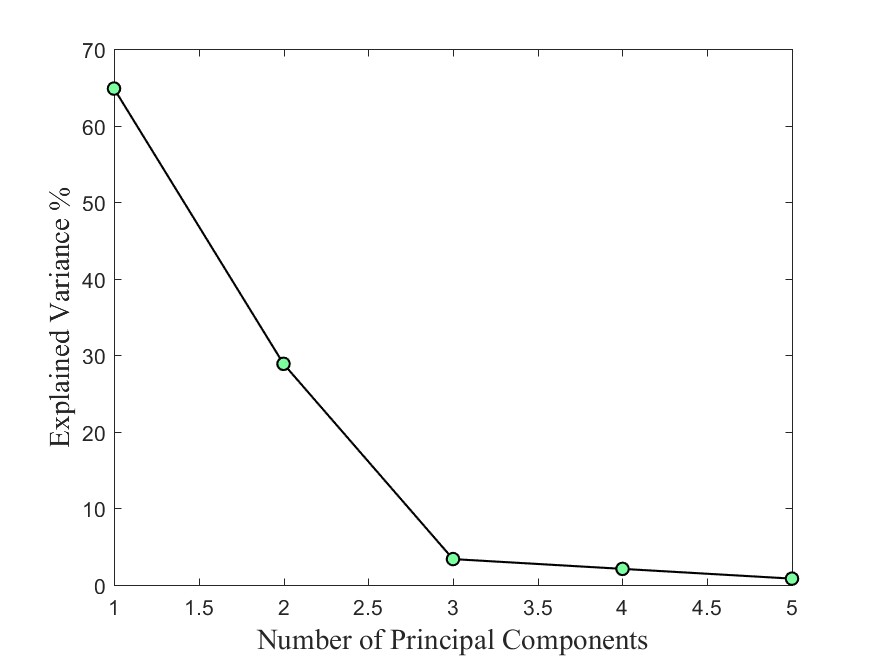
63%, ℓ2 = 28% and ℓ3 = 2%. Notice that PC1, PC2 and PC3 combined account for 93% of the variance in the data.

The Scree and Pareto plots of the explained variance vs. the number of PCs are shown in Figure 6 and 7 respectively. Based on the explained variance by both PC1 and PC2 and also from the scree and Pareto

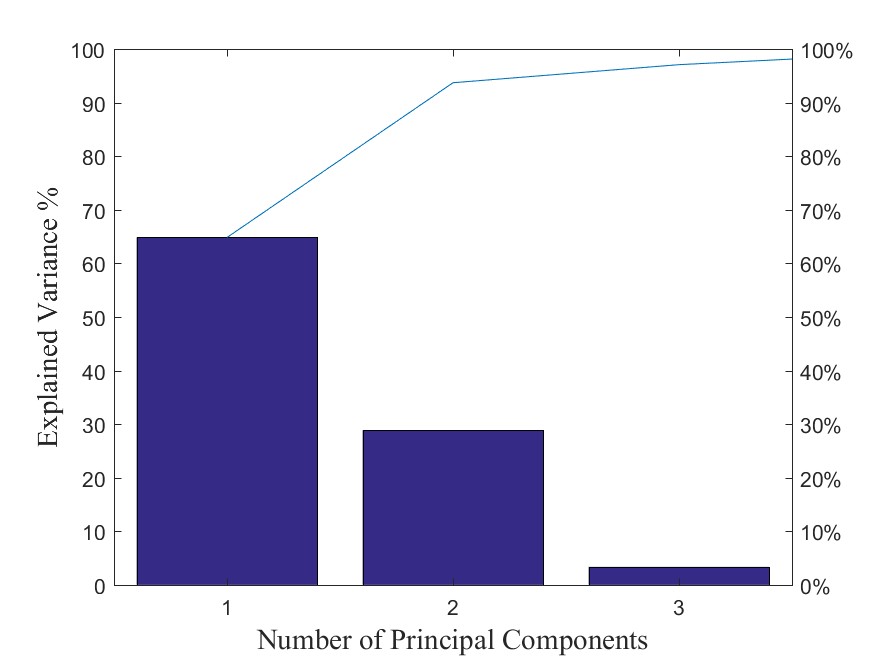
plots, it can be deduced that the lowest‐dimensional

space to represent the diamond data corresponds to d

= 3.



*Figure 6* ‐ *Scree Plot*

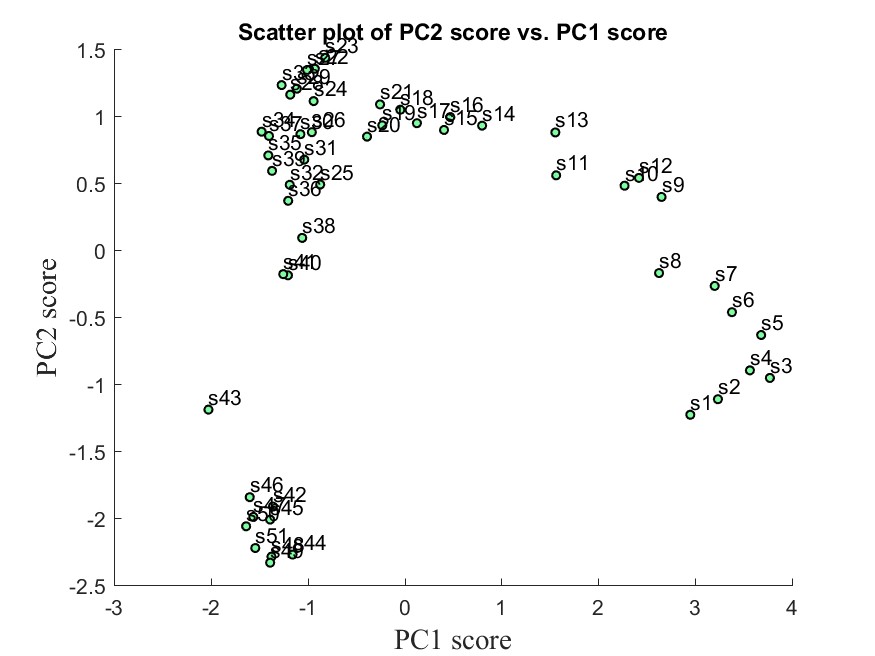


*Figure 7* ‐ *Pareto Plot*

Scatterplot of PC2 Score vs. PC1 Score is shown in

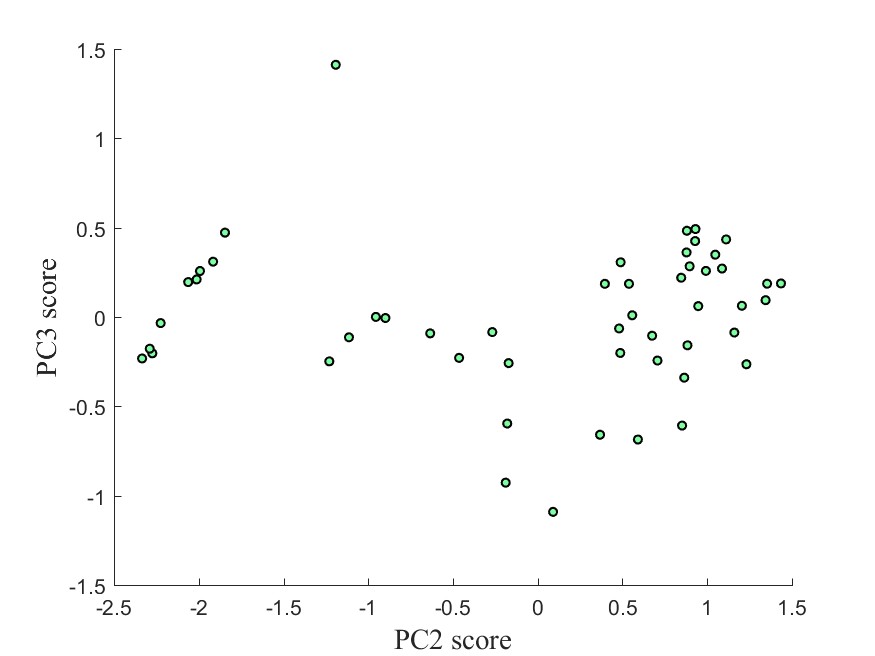
Fig.8. It can be observed that the observation of s43 is exceptionally different from others.

*Figure 8* ‐ *PC1 vs PC2 score*



Scatterplot of PC3 Score Vs. PC2 Score is shown in Fig.5. It can be observed that most of the observations are plotted close to each other except only one.

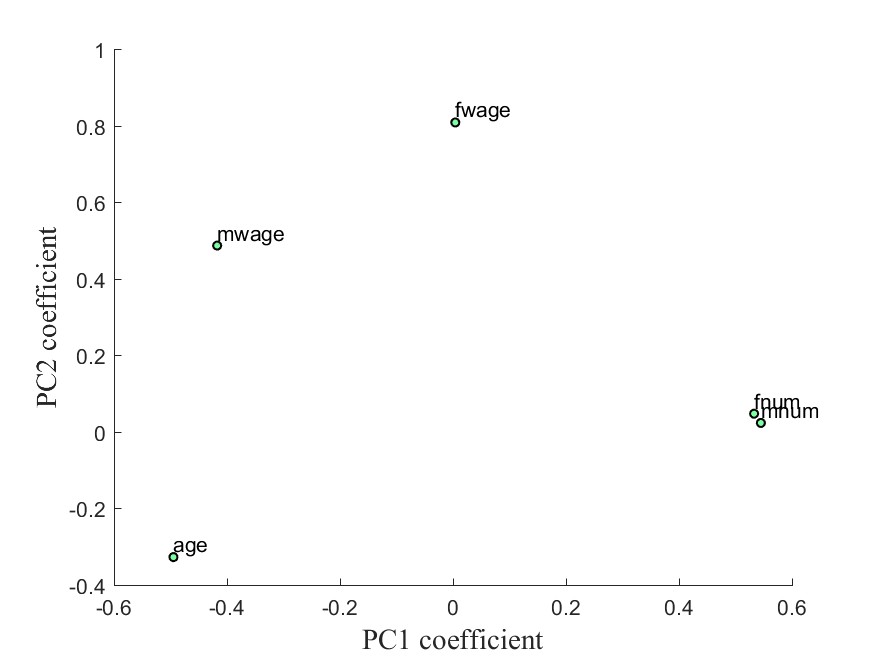
*Figure 9* ‐ *PC2 vs PC3 score-*



The scatterplot of PC2 coefficients vs. PC1

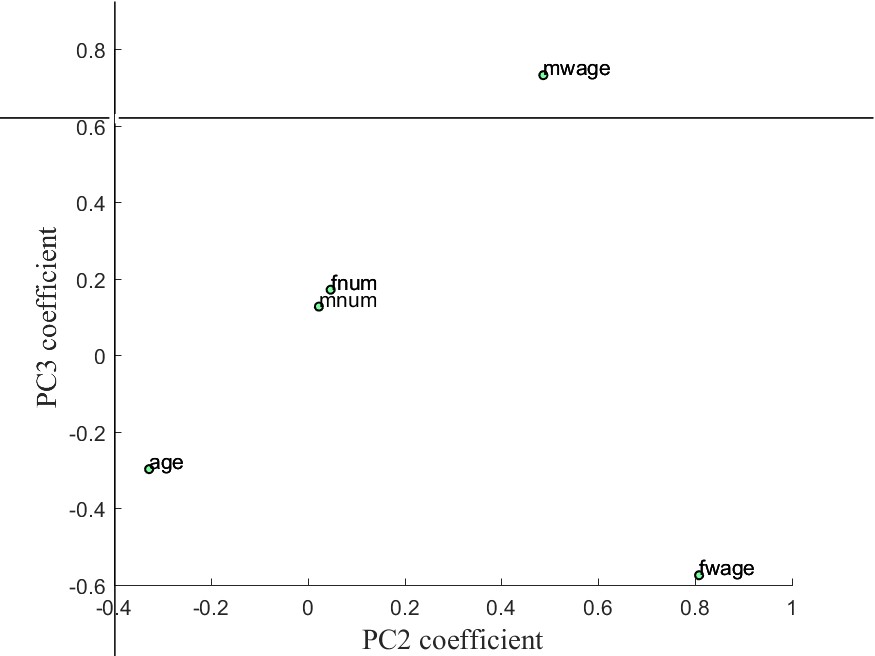
coefficients is shown in Fig. 6. This plot helps understand which variables have a similar involvement within PCs. As it can be observed from figure, that the fwage is plotted far away from other delaying factors and has the maximum magnitude. It

shows the strong relation of PC1 and PC2 in terms of fwage and weak relation in terms of age.



*Figure 10 PC2 coefficient vs PC1 coefficient*

The scatterplot of PC3 coefficients vs. PC2 coefficients is shown in Fig. 7. It can be observed that PC 2 and PC3 have strong relation in terms of mwage and weak relation in terms of fwage.



*Fig 11- PC3 vs PC2 Coefficients*

The biplot is a graphical tool that provides the information on both observations and variables of a data matrix in graphical format. The biplot helps visualize both the principal component coefficients for each variable and the principal component scores for each observation in a single Plot.

The 2D bi‐plot of PC2 vs. PC1 is shown in Figure 12. The axes in the bi‐plot represent the principal

components (columns of Eigen matrix), and the observed variables (rows of Eigen matrix) are represented as vectors.

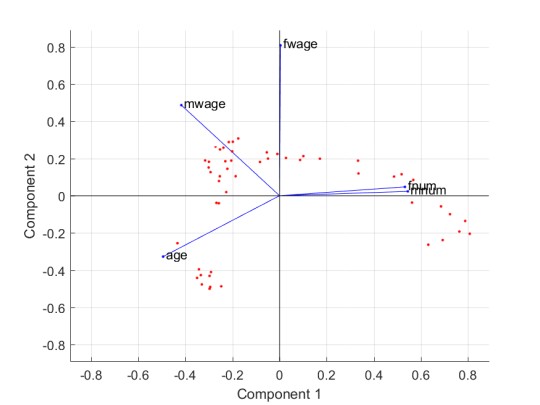
Each observation is represented as a red point in the

bi‐plot. From figure 12, we can see that the first

principal component doesn’t have positive

coefficients for all the variables. As 3 vectors are directed into the right half of the plot and 2 into left half.

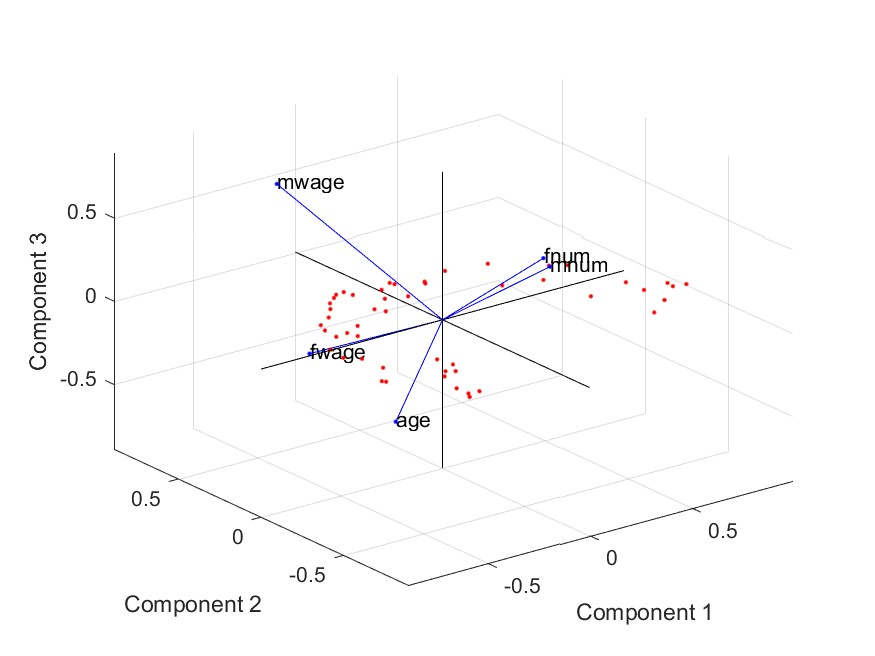
*Figure 12* ‐ *2D Biplot*



The second principal component, represented by the

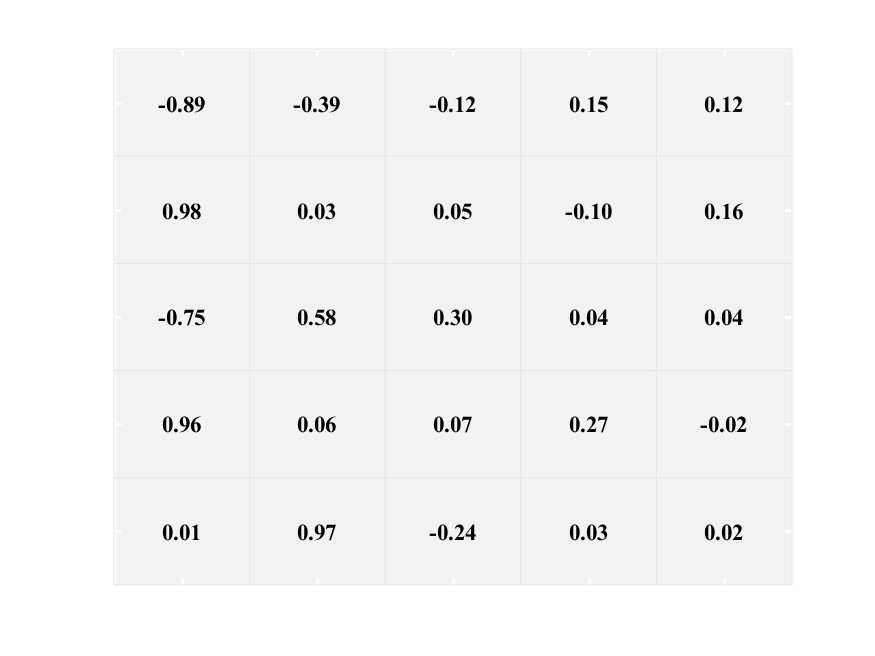
vertical axis, has 4 positive coefficients for the variables **fnum, mnum, mwage** and **fwage** and 1 negative coefficients for the variable **age**. That corresponds to 4 and 1 vectors directed into top and bottom halves of the plot, respectively. The 3D biplot of PC1, PC2 and PC3 is shown in figure 13.

*Figure 13* ‐ *3D Biplot*



From the Component Correlation Matrix, we get the

values for the PCs.



The First PC is given by:

Z1 = -0.89a + 0.98mn – 0.75mw + 0.96fn + 0.01fw

Where-

A=age Mn=mnum Mw= mwage Fn= fnum Fw=

fwage

First PC has contrast between age and mwage together with the mnum and fnum together as evident from the negative coefficients of age and mwage and positive coefficients of mnum and fnum.

Second PC is given by:

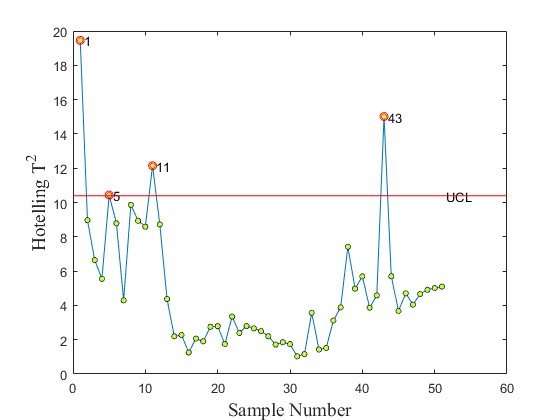
Z2 = -0.39a + 0.03mn + 0.58mw + 0.06fn + 0.97fw

Second PC has contrast between mwage and fwage together with mnum and fnum together as evident from positive coefficients whereas only age is negative coefficients.

The Hotelling and First PC charts are displayed in

Figures 14 and 15. The Hotelling chart indicate that the

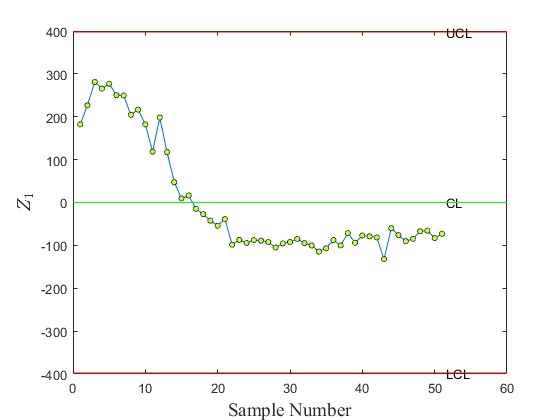
samples 1, 5, 11, 43 are out‐of‐control.



*Figure 14* ‐ *Hotteling Chart*

The First PC chart below shows that all the sample points

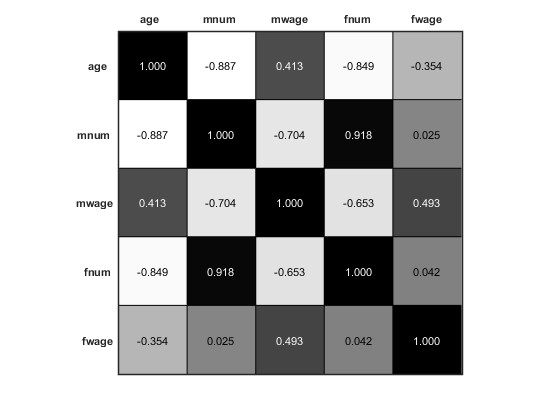
are in control and not outside the control limits.



*Figure 15* ‐ *First PC Charts*

III. FACTOR ANALYSIS IMPLEMENTATION

a. Correlation Matrix



*Figure 16. Correlation Matrix*

With respect to the correlation matrix, two things are important: the variables have to be inter-correlated, but they should not correlate too highly (extreme multicollinearity) as this would cause difficulties in determining the unique contribution of the variables to a factor. Multicollinearity, can be detected via the determinant of the correlation matrix, if the determinant is greater than 0.00001, then there is no multicollinearity.

Corr(X) = 0.0034

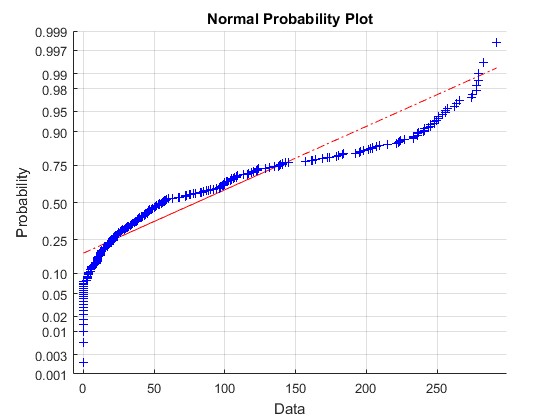
Also the determinant of the correlation matrix, shown above, which is greater than 0.00001, shows there is no multicollinearity, so the data is appropriate for factor analysis.

b. Factor Extraction (Loadings)

Different techniques are used to estimate initial factors, including Principal Components Factor Analysis, Principal Factor Factor Analysis, Iterated Principal Factor Factor Analysis, Maximum Likelihood Factor Analysis, and more, but in this project Maximum Likelihood method is considered. In order to use maximum likelihood method, it is

necessary for the data to be multivariate normal. So, we first check normality of our multivariate data.

*Figure 17. Normal Probability Plot*

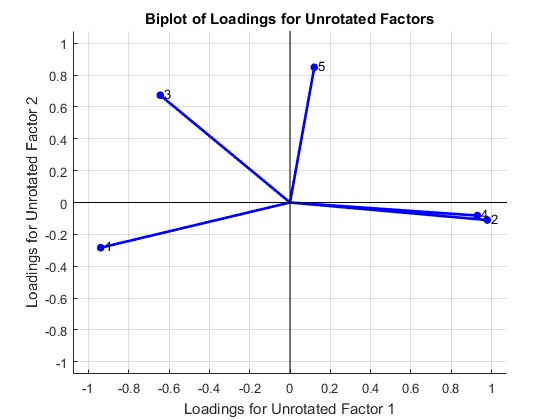


Loadings of Factors=

|  |  |
| --- | --- |
| -0.943 | -0.2676 |
| 0.9762 | -0.1266 |
| -0.6319 | 0.6839 |
| 0.9277 | -0.0977 |
| 0.1348 | 0.8459 |

C. Biplot of Factor Loadings

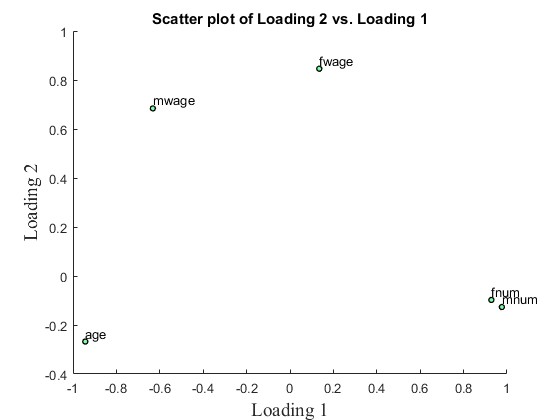
From the biplot of factor loadings, shown in Fig. 14, it can be observed that the second, third, fourth and fifth variables () have a positive loading on the second factor, while second variable (Weather Delay) has only a small loading on the second factor.



*Fig 18. Biplot of loadings for unrotated factors*

D. Scatter Plot of Factor Loadings

Scatter plot of factor loadings, shown in Fig. 15, verifies the result which was concluded from biplot, shown in Fig. 13, that age has only a small loading on the second factor.



*Fig 19. Scatter plot of loading 2 vs loading 1*

E. Specific Variance

A specific variance of 1 would indicate that there is no common factor component in that variable, while a specific variance of 0 would indicate that the variable is entirely determined by common factors.

Estimated Specific Variance =

|  |
| --- |
| 0.0389 |
| 0.0308 |
| 0.1329 |
| 0.1296 |
| 0.2662 |

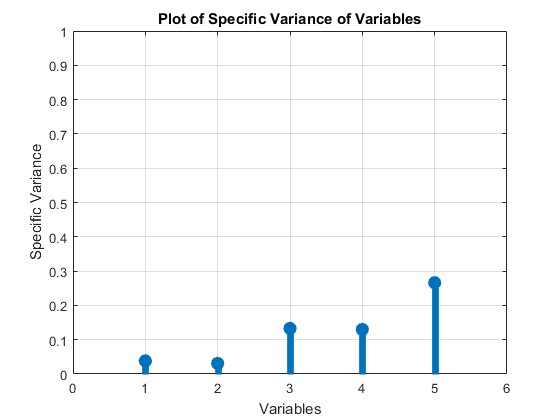
From the specific variance plot, shown in Fig. 16, it can be observed that the least amount of specific variance occurs for the second variable i.e. mnum.

F. Factor Rotation

Rotation statistically manipulates the factors to make them more meaningful. Various methods are used to rotate factors, including Quartimax, Varimax, Promax, and more, out of which some are orthogonal and some are oblique. The choice of rotation is not simple and is critical, so a fairly straightforward way to decide which rotation to take is to carry out the analysis using both types of rotation.

In this project, Varimax (orthogonal) and Promax

(oblique) are implemented and analyzed.



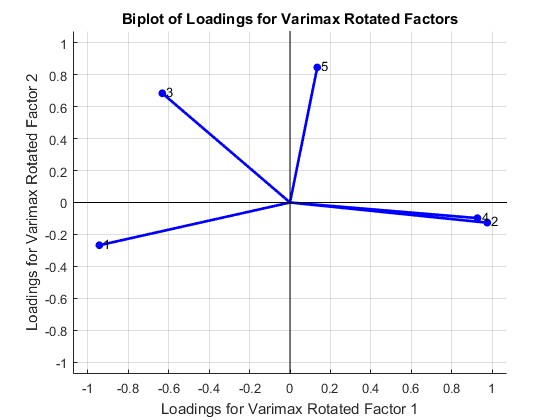
*Figure 20. Plot of Specific Variance*

Varimax, is an orthogonal turn strategy which amplifies

the aggregate of the squared variable loadings over the

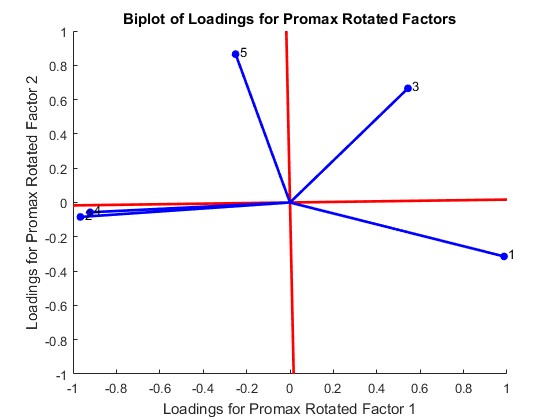
segments. This tends to compel every variable to stack very on as few variables as could be expected under the circumstances. In a perfect world it will make every variable stack on just a single component and accordingly call attention to great summed scores that could be made. Two-figure Varimax pivot is directed

of the two inactive variables removed, appeared in Fig. 21. From the figure, it can be watched that orthogonal turn has not prevailing with regards to giving a basic arrangement of elements.



*Fig 21. Biplot of loadings for varimax rotated factors*

Now Promax oblique rotation will be executed to facilitate streamline. Angled revolutions don't keep the autonomy between the elements. Promax turns the element tomahawks independently, enabling them to have a diagonal point between them.



*Figure 22. Biplot of Loadings for Promax Rotated Factors* From the biplot of loadings for Promax Rotated Factors on oblique axes, shown in Fig. 18, it can be observed that Promax has performed a non-rigid rotation of the axes, and has done a much better job

than varimax at creating a "simple structure". The third variable is in the intermediate position, however fifth variable is close to second factor axis and first, fourth, second variables are close to the first factor axis respectively. This makes an interpretation of these rotated factors more precise. So, it can be concluded that the age is the latent factor.

G. Factor Scores

It is useful to be able to classify an observation based on its factor scores. Biplot of Promax rotated loadings and scores is shown in Fig. 23. From the figure, it can be observed that some observations, points near the left edge of this plot, have the lowest scores for the first factor.

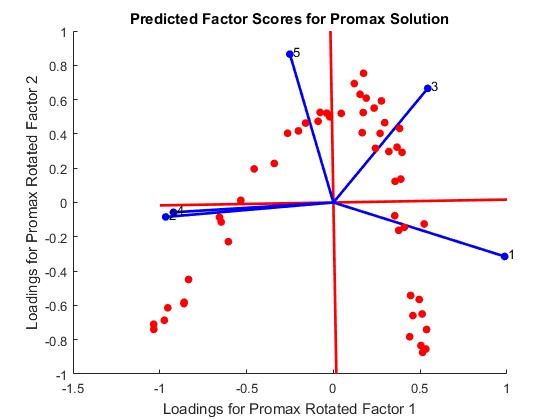
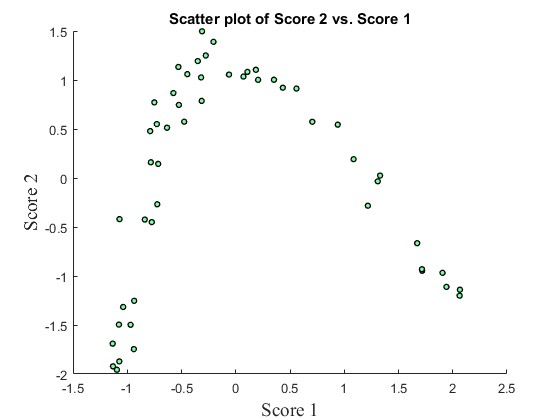


Figure 23. Biplot of Promax Rotated Loadings Vs. Scores



*Fig. 20 Scatter plot of unrotated factor scores*

**H. Final Decision**

Based on the exploratory factor analysis performed on wages of male and female workers, the factor model till age 50 can be accepted.

**IV. CONCLUSIONS**

In this project, through the implementation of two Advanced Statistical Techniques, i.e. Principal Component Analysis and Factor Analysis, on dataset from Wages of Lancashire Cotton Factory Workers in 1833. Based on PCA and FA we can easily interpret the weak attribute which comes out to be age. And it also helped in identifying the latent factors that are enough to predict the result for given project.

**V. REFERENCES**

Factor Analysis, Wikipedia (Dec 02, 2012), Internet:

<http://en.wikipedia.org/wiki/Factor_analysis>

[http://www.mathworks.com/help/wavelet/ref/wmspca](http://www.mathworks.com/help/wavelet/ref/wmspca.html)

[.html](http://www.mathworks.com/help/wavelet/ref/wmspca.html)

[https://vincentarelbundock.github.io/Rdatasets/datasets.](https://vincentarelbundock.github.io/Rdatasets/datasets.html) [html](https://vincentarelbundock.github.io/Rdatasets/datasets.html)

<http://www.mathworks.com/help/stats/factoran.html>

How to Interpret Factor Analysis Results,

eHow (Accessed on Dec 02, 2012). Internet: [http://www.ehow.com/how\_5376446\_interpret-](http://www.ehow.com/how_5376446_interpret-factor-analysisresults.html) [factor-analysisresults.html](http://www.ehow.com/how_5376446_interpret-factor-analysisresults.html)