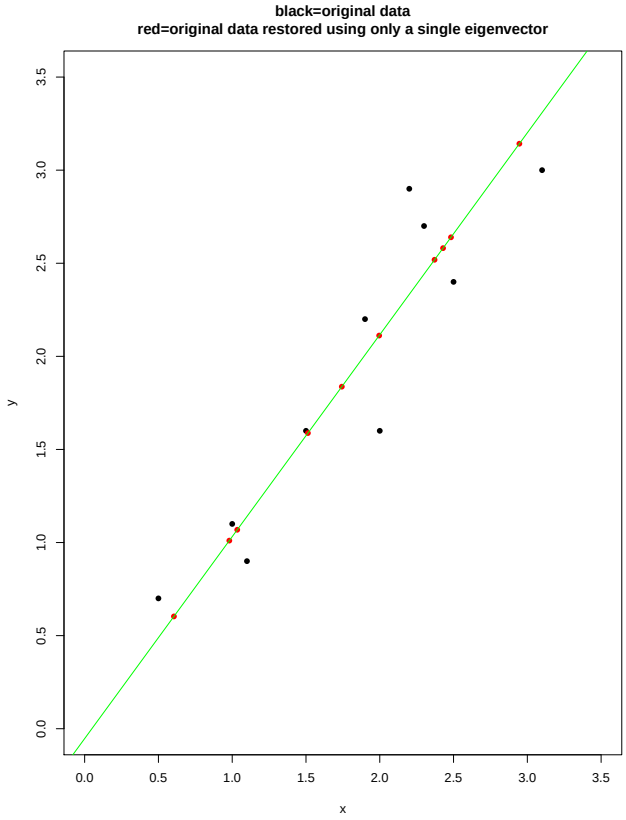
CHALMERS



Principal Components Analysis of Remote Measurement Emission Particles

WEIHUA WANG

Department of Space, Earth and Environment
Chalmers University of Technology
Gothenburg, Sweden 2017

Principal Components Analysis of Remote Measurement Emission Particles

Weihua Wang

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Space, Earth and Environment
Chalmers University of Technology
Gothenburg, Sweden 2017

Principal Components Analysis of Remote Measurement Emission Particles
Weihua Wang

© Weihua Wang, 2017.

Supervisor: Johan Mellqvist, Department of Space, Earth and Environment

Department of Space, Earth and Environment
Research Group of Optical Remote Sensing
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Principal Components Analysis of Remote Measurement Emission Particles
Weihua Wang
Department of Space, Earth and Environment
Chalmers University of Technology

Abstract

This report aims to use the Principal Component Analysis (PCA) to preliminary separate and recognize the light intensity at 500 nm (I500), various trace gas species (e.g. SO_2, NO_2, NH_3 and so on) and the aerosol traces, indicated by the light intensity ratio of 340 nm to 675 nm ($IR_{340/675}$). The data came from field campaign of several days of vehicle-based optical remote sensing measurements in Tianjing, China. The selected trace gas species were measured by spectrometers using the Differential Optical Absorption Spectroscopy (DOAS) and the Solar Occultation Flux (SOF) techniques. Meanwhile, the aerosol traces can be calculated from recordings by a miniature 'Flame' spectrometer. Through PCA, we found that in some cases the Principal Component (PC) has positive correlation with its constituent parts of linear combination.

Keywords: PCA, Aerosol traces, DOAS, SOF.

Acknowledgements

The author would like to sincerely thank Professor Johan Mellqvist and Dr. John Johansson for their significant and kind help in theoretical guidance, simulations, supervision and comments on this project.

Weihua Wang, 2017

List of Abbreviations

AERONET	AErosol RObotic NETwork
API	Application Programming Interface
CARSNET	China Meteorological Administration Aerosol Remote Sensing Network
CCD	Charge-coupled Device
DOAS	Differential Optical Absorption Spectroscopy
EFA	Exploratory Factor Analysis
HCHO	Formaldehyde
IR	Intensity Ratio
LOS	Line Of Sight
MODIS	Moderate-resolution Imaging Spectroradiometers
OD	Optical Depth
PCA	Principal Components Analysis
PCs	Principal Components
PM	Particulate Matter
RMS	root-mean-squared
SD	Standard Deviation
SVD	Singular Value Decomposition
SOF	Solar Occultation Flux
UV	Ultraviolet
VIS	Visible light
VOCs	Volatile Orgaic Compounds

Contents

List of Figures **vii**

List of Tables **ix**

1 Introduction **1**
1.1 Brief description of proposed research questions 2

2 Principal Components Analysis (PCA) **3**
2.1 Background mathematics 3
2.2 Apply PCA on example data set 5
2.3 Retrieving the original data back 9

3 Data Analysis **10**
3.1 PCA processing 10

4 Results and Discussion **12**
4.1 PCA for DOAS data 12
4.2 PCA for SOF data 19
4.3 Conclusion 26
4.4 Future Work 27

References **28**

List of Figures

2.1 Original data set 6
2.2 Mean subtracted data set. The red and blue lines correspond to the two eigenvectors, respectively. 6
2.3 Data set after applying the PCA so that eigenvectors are the axes. 8
2.4 Compared to the original data set, the reconstructed data set only use the main principle component. 8

3.1 Concentration of the sulfur dioxide with (red) and without (blue) despiking. 11

4.1 PCA results of the standard deviations, the rotation coefficients, and so on in terms of the standardized variables of the SO2, NO2, HCHO, I500 and $IR_{340/675}$ on May 8. 13
4.2 Rescaled PC1 and concentrations of the SO2, NO2 and HCHO on May 8. 13
4.3 Rescaled PC2 and concentrations of the SO2, NO2 and HCHO on May 8. 14
4.4 Rescaled PC3 and concentrations of the SO2, NO2 and HCHO on May 8. 14
4.5 Rescaled PC4 and concentrations of the SO2, NO2 and HCHO on May 8. 15
4.6 Rescaled PC5 and concentrations of the SO2, NO2 and HCHO on May 8. 15
4.7 Scatter plot of HCHO and PC2 on May 8. 16
4.8 Scatter plot of NO2 and PC2 on May 8. 17
4.9 Scatter plot of SO2 and PC2 on May 8. 17
4.10 Scatter plot of IR and PC2 on May 8. 18
4.11 Scatter plot of I500 and PC2 on May 8. 18

4.12 PCA results of the rotation coefficients, the standard deviations and so on in terms of the ethylene, ammonia, butane, I500 and $IR_{340/675}$ on May 8. All variables were standardized before PCA. 19
4.13 Rescaled PC1 and concentrations of the C2H4, NH3 and C4H10 on May 8. 20
4.14 Rescaled PC2 and concentrations of the C2H4, NH3 and C4H10 on May 8. 20
4.15 Rescaled PC3 and concentrations of the C2H4, NH3 and C4H10 on May 8. 21
4.16 Rescaled PC4 and concentrations of the C2H4, NH3 and C4H10 on May 8. 21
4.17 Rescaled PC5 and concentrations of the C2H4, NH3 and C4H10 on May 8. 22
4.18 Scatter plot of C2H4 and PC2 on May 8. 22
4.19 Scatter plot of NH3 and PC2 on May 8. 23
4.20 Scatter plot of C4H10 and PC2 on May 8. 23
4.21 Scatter plot of IR and PC2 on May 8. 24
4.22 Scatter plot of I500 and PC2 on May 8. 24
4.23 Scatter plot of IR and PC1 on May 8. 25
4.24 Scatter plot of I500 and PC1 on May 8. 25

List of Tables

2.1 Original data set 6
2.2 Mean adjusted data set 6

Chapter 1

Introduction

Aerosol generally refers to liquid or solid particles suspended in the air. The aerosol, especially those anthropogenic aerosol like haze, can result in adverse impact on human health. In specific, aerosol particles can be inhaled deeply into the lungs, potentially causing serious consequences of respiratory and cardiovascular disease, such as asthma, obesity and metabolic syndrome [1,2].

China is one of the world's major sources of aerosol. With the rapidly expanding economic and industrial developments of the last three decades, the components of aerosol were greatly extended from simple form of biomass burning aerosol associated with agricultural activities or desert dust aerosol from the north and west desert regions, to urban and industrial aerosol emitted from various anthropogenic sources, such as combustion of fossil fuels by vehicles, electricity generation from power plants and other industrial processes [3, 4].

Therefore, the aerosols have brought enough attention to the Chinese government and academic researches. Besides of using data from the MODIS (Moderate Resolution Imaging Spectroradiomer) or the AERONET (AErosol RObotic NETwork) or both, China has established ground-based national networks of CARSNET (China Aerosol Remote Sensing NETwork) and CSHNET (Chinese Sun Hazemeter Network). They are fixed sampling sites located either in the urban, suburban or rural areas. Although expansion in recent years, the existing networks are still inadequate for a comprehensive evaluation of the aerosol properties across the massive area of China.

For principal components analysis (PCA) on air pollution data, studies have been made to interpret the principal components (PCs) in terms of different

air pollutants. In specific, research has been made to quantify the association between daily mortality (cardiovascular diseases, respiratory, pneumonia, etc) and the source-related components, for instance routinely collected concentrations of the air pollutants, in Netherlands [5].

1.1 Brief description of proposed research questions

The flexible, pre-customized mobile measurements instead of the fixed sampling spots are therefore quite necessary to provide supplementary studies. This is the reason why this project chooses the vehicle-based measurements which are rare in Chinese aerosol studies. From the mobile measurement, concentrations of the selected gas species can be recorded by the DOAS (Differential Optical Absorption Spectroscopy) and SOF (Solar Occultation Flux) technique, while the solar light intensity at 500 nm (I500) and the aerosol traces was indicated by the intensity ratio of 340 nm to 675 nm ($IR_{340/675}$) can be gauged by the 'Flame'.

Compared to previous studies where the selected gas species share the same unit of concentration, the input quantities to PCA in this report have different physical dimensions. Therefore, all variables must be standardized before put into the PCA.

The research question we proposed is that the I500, the $IR_{340/675}$ and gas species of the majority from the vehicle-based measurements can be recognized and separated by the PCA. In specific, we would expect the first PC presents information about the light intensity measured by the 'Flame' spectrometer. The second PC shows information about the sulfur dioxide concentration because it displays the largest variance among all the measured gas species on May 8. The third and forth PC would convey information about certain gas species whose variance of concentration ranks the second and the third among all the gas species. The last PC can reflect something about the noise term.

Besides of the introduction chapter, PCA will be described in detail in chapter 2. The data processing procedure before applying the PCA will be expressed in chapter 3. The the results and conclusionare discussed in last chapter.

Chapter 2

Principal Components Analysis (PCA)

In this chapter an example of showing how the PCA is performed will be given step by step according to a reference tutorial [6]. Before the PCA example is given some fundamental mathematics will be briefly reviewed.

PCA is a powerful algorithm for multivariate data analysis. It is very versatile with applications in many disciplines, for example human face recognition in computer vision, image compression, computer networks monitoring and disturbance detection, etc.

Main idea for the PCA is to reduce the dimensionality of a data set consisting of a large number of possibly not all independent variables, while preserving as much as possible of the variation present in the data set. Specifically, variables of the original data set is orthogonal transformed into a new set of linearly uncorrelated variables, i.e. the principal components (PCs), which are constructed such that the first few PCs span most of the variation present in all of the original variables [7].

2.1 Background mathematics

The mean value, denoted as $\bar{X}$, of a data set X is given by

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \tag{2.1}$$

where n refers to the number of elements in the data set.

The standard deviation (SD) of a data set reflects how spread out the data is from the mean value. One definition for a *sample* data set is

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}, \tag{2.2}$$

where s refers to the SD of a sample. Note that dividing by n instead of by (n-1) is applied to calculating the SD of an entire *population*.

Variance, denoted here as s^2, is just a square operation of the SD,

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}. \tag{2.3}$$

The SD and variance only work for one dimensional data set. For data set of two dimensions, covariance is used to find out how much the dimensional data vary from the mean with respect to each other. The mathematical expression for the covariance between data set X and Y is

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{n-1}. \tag{2.4}$$

Covariance is always calculated between two dimensional data. If a data set contains more than two dimensional data, a covariance matrix is used to represent all possible covariance values between all the different dimensions. The mathematical description for the covariance matrix for a n dimensional data set is,

$$C^{n\times n} = \big(c_{i,j}, c_{i,j} = cov(Dim_i, Dim_j)\big). \tag{2.5}$$

where $C^{n\times n}$ is a n by n square matrix, and each entry of $c_{i,j}$ in the matrix is the covariance between the ith dimension, Dim_x, and the jth dimension, Dim_j. Take one example, the covariance matrix of a three dimensional data set of dimensions x, y and z can be denoted by

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}. \tag{2.6}$$

Note that the matrix is symmetrical about the main diagonal. And along the main diagonal, the covariance value is actually one of the dimensions and itself.

Consider two multiplications between a matrix and a vector shown in Equation (2.7)

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 11 \\ 5 \end{pmatrix}. \tag{2.7}$$

Result is not an integer multiply of the original vector. While, in Equation (2.8) the resulting vector is exactly 4 times of the original vector.

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix}. \tag{2.8}$$

The vector in Equation (2.8) represents an arrow pointing from the origin, (0,0), to the point (3,2). The square matrix in Equation (2.8) can be regarded as a transformation matrix because if one multiply this matrix on the left of the vector, the result is just another vector that is transformed from it's original place.

Mathematical expression for an eigenvector follows

$$Av = \lambda v, \tag{2.9}$$

where λ is a scalar known as eigenvalue, A is a square matrix and the non-zero column vector v refers to eigenvector. Note that eigenvector can only be found for square matrices.

2.2 Apply PCA on example data set

This subsection will perform a PCA on an example set of data step by step.

Step 1: Prepare the data

A small data set of 2 dimensions are used here, see Table 2.1 and Figure 2.1, to illustrate how PCA can be performed.

Step 2: Subtract the mean

For PCA to perform correctly, the mean have to be subtracted from each of the original data dimensions. The mean adjusted data set can be seen from Table 2.2.

Table 2.1: Original data set

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2.0	1.6
1.0	1.1
1.5	1.6
1.1	0.9

Table 2.2: Mean adjusted data set

x	y
0.69	0.49
-1.31	-1.21
0.39	0.99
0.09	0.29
1.29	1.09
0.49	0.79
0.19	-0.31
-0.81	-0.81
-0.31	-0.31
-0.71	-1.01

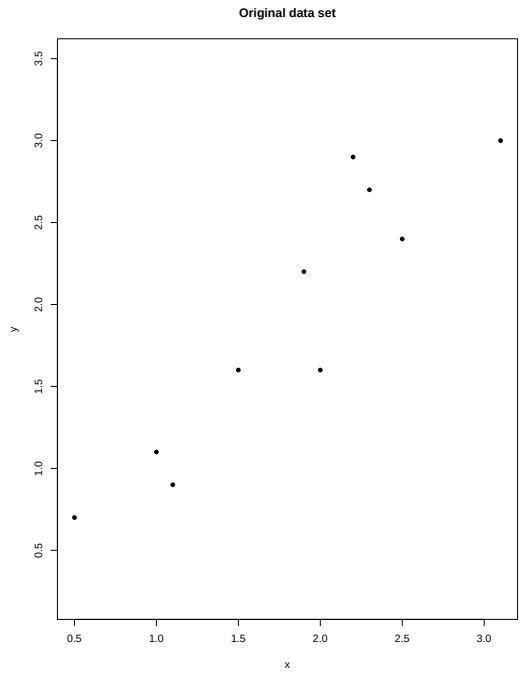


Figure 2.1: Original data set

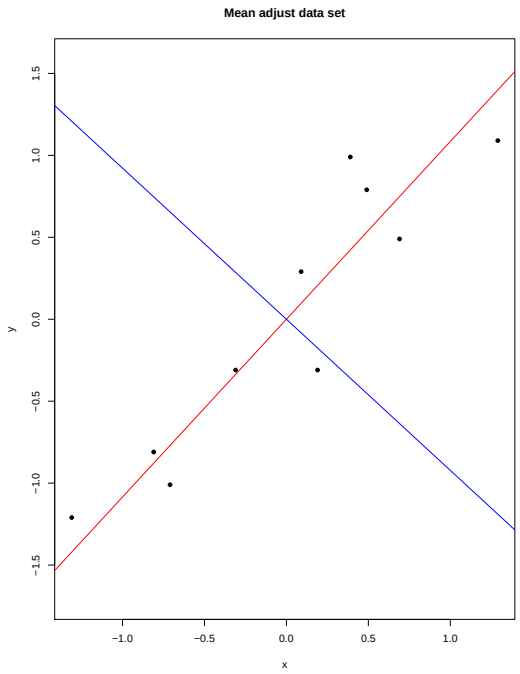


Figure 2.2: Mean subtracted data set. The red and blue lines correspond to the two eigenvectors, respectively.

Step 3: Calculate the covariance matrix

The 2 by 2 dimensional covariance matrix calculated from the mean adjusted data set is

$$cov = \begin{pmatrix} 0.6165556 & 0.6154444 \\ 0.6154444 & 0.7165556 \end{pmatrix}. \tag{2.10}$$

Step 4: Calculate the eigenvectors and eigenvalues

The eigenvectors and eigenvalues calculated from the square covariance are,

$$eigenvalues = \begin{pmatrix} 1.2840277 \\ 0.0490834 \end{pmatrix}. \tag{2.11}$$

$$eigenvectors = \begin{pmatrix} 0.6778734 & -0.7351787 \\ 0.7351787 & 0.6778734 \end{pmatrix}. \tag{2.12}$$

Both eigenvectors are *unit* eigenvectors. In Figure 2.2 the adjusted data has a stronger pattern along the red eigenvector than the blue eigenvector.

Step 5: Choosing components and forming a feature vector

The eigenvectors with the largest eigenvalues are the PCs of the data set. In general, once the eigenvectors are calculated, the next step is to sort them by eigenvalue, highest to lowest. The highest eigenvalue corresponds to the highest order of significance, vice visa. Those components of least significance can be ignored. Although some information are thrown away, the lost is not so much as long as the eigenvalues are small. Finally, the dimensions of the data are reduced. If the original data represents an image, image compression thus can be realized.

Right now a feature vector needs to be constructed. This is done by taking the eigenvectors that one wants to keep from the list of eigenvectors, and forming a matrix with these eigenvectors in the columns.

$$FeatureVector = (eig_1, eig_2, eig_3, ..., eig_n). \tag{2.13}$$

In this example, the less significant eigenvector is omitted

$$FeatureVector = \begin{pmatrix} 0.6778734 \\ 0.7351787 \end{pmatrix}. \tag{2.14}$$

Therefore, the feature vector only has a single column.

Step 6: Deriving the new data set

Once the feature vector is formed, final step in PCA is obtained by matrix multiplication between the transpose of the feature vector and the transposed mean adjusted data set.

$$FinalData = FeatureVector^T \times DataAdjust^T, \tag{2.15}$$

where the eigenvectors in $FeatureVector^T$ are now in rows and the most significant eigenvectors are at the top. The data items in $DataAdjust^T$ are in columns and each row corresponds to a unique dimension. $FinalData$ is the result data set, with each data items in columns and dimensions along rows.

If both eigenvectors are kept for deriving the new data set, the result can be found in Figure 2.3. This plot is basically the original data, rotated so that the eigenvectors replace the previous axes of x and y. This is clear since no information are lost in this transformation.

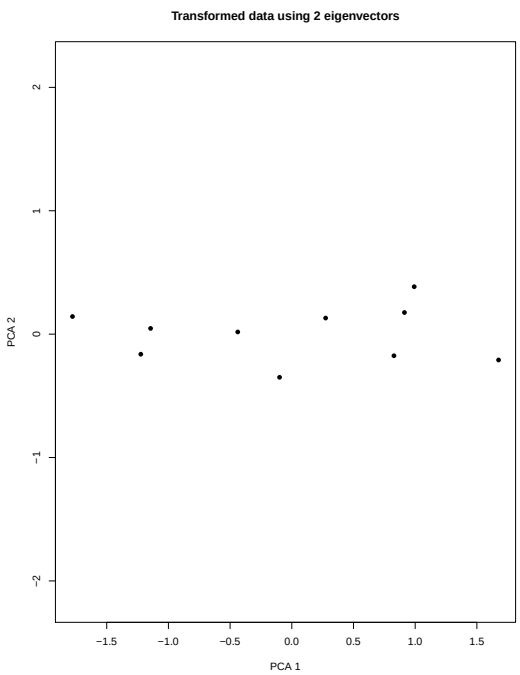


Figure 2.3: Data set after applying the PCA so that eigenvectors are the axes.

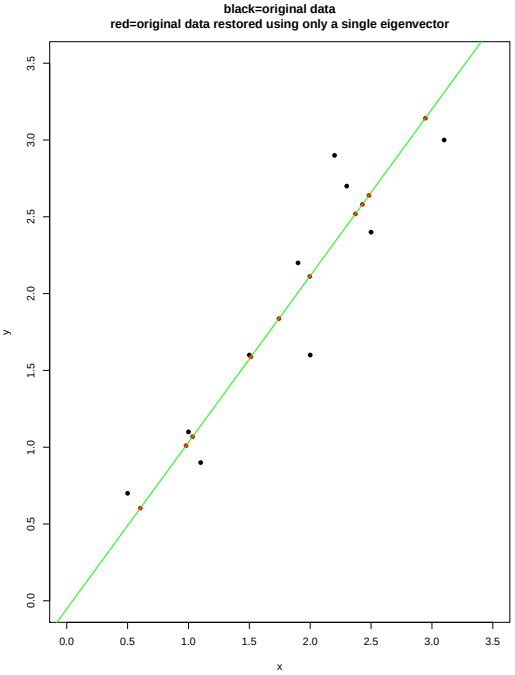


Figure 2.4: Compared to the original data set, the reconstructed data set only use the main principle component.

2.3 Retrieving the original data back

It is understandable that the original data set can be exactly retrieved back only if all the eigenvectors are taken in the final transformation. If, however, the number of eigenvectors are reduced, then the retrieved data must lost some information.

The final transform is recalled in Equation (2.15). To get the mean adjusted data back,

$$DataAdjust^T = FeatureVector^{-1} \times FinalData \tag{2.16}$$

Here, inverse of the feature vector actually equals to the transpose of itself because all the *unit* eigenvectors are taken to form the feature vector. The equation now becomes

$$DataAdjust^T = FeatureVector^T \times FinalData. \tag{2.17}$$

To get the original data back, the mean of that original data have to be added since the mean was subtracted at the very beginning. So, for completeness,

$$OriginalData = (FeatureVector^T \times FinalData) + OriginalMean \tag{2.18}$$

Equation (2.18) also applies to the case when not all the eigenvectors are included to build the feature vector.

Figure 2.4 shows what a reconstructed data would look like when a reduced feature vector is used (i.e. information is lost). Compare to the original data in black dots, the variation along the principle eigenvector (the red line in Figure 2.2) has been kept, while the variation along the other eigenvector (the ignored blue line in Figure 2.2) has gone.

The reproducible R code regarding the above PCA example are attached in Appendix B.

Chapter 3

Data Analysis

The calculations of PCA and rotations were performed in software of **R** with version 3.3.3 (2017-03-07) under **Ubuntu** (12.04) environment.

3.1 PCA processing

PCA can be applied by many functions and packages in R language. The most widely used functions are *prcomp* and *princomp*. Usually, the *prcomp* is the preferred approach for numerical accuracy. Its calculation is done by a singular value decomposition (**SVD**) of the centered and possibly scaled data matrix. While, the *princomp* is calculated by using *eigen* function on the correlation or covariance matrix. The calculation of *princomp* provides compatible output with the result from other commercially software, for example S-PLUS. In this project we use the function *prcomp* from the *stats* package.

Before application of the PCA, all data variables have to be pre-processed. Here, data outliers like the spikes must be detected and removed. Besides, data normalization is strongly suggested because data skewness and the magnitude of the variables can influence or even ruin the output PCs.

The spikes were removed by median filtering with a sliding window. Although a window width of 3 or 5 was good to use, we should take care of the case when the isolated concentration peaks appeared, for example the sulfur dioxide peaks due to the chimney emission on May 8. Because those individual peaks can be wrongly filtered as spikes. Thus, we choose to keep the concentration of the sulfur dioxide without undergoing despiking procedure. Figure 3.1 shows how the sulfur dioxide peaks can be affected from despiking. The

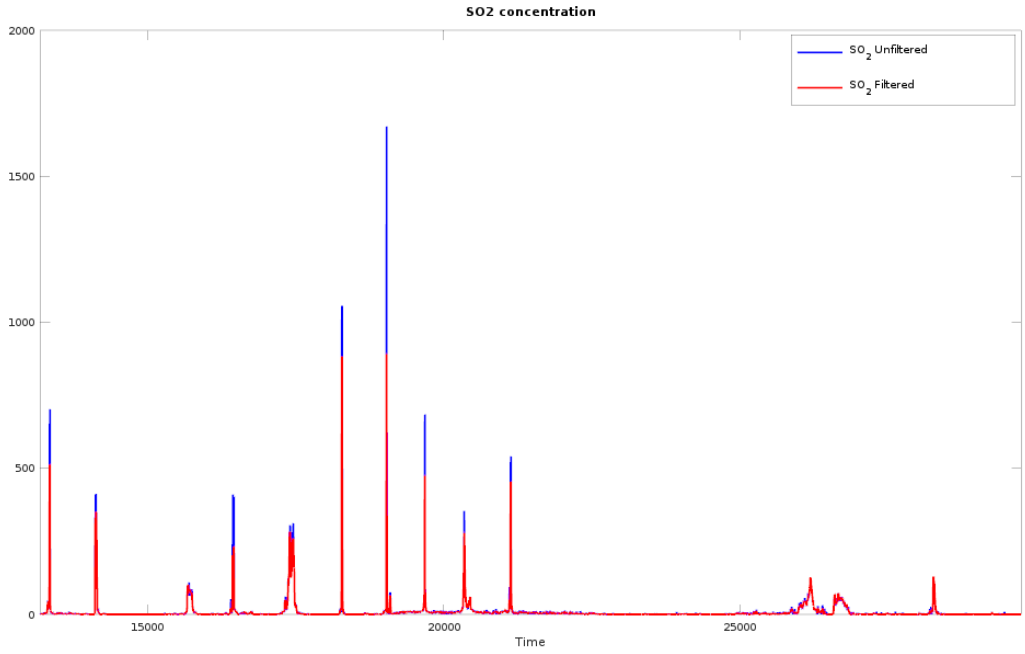


Figure 3.1: Concentration of the sulfur dioxide with (red) and without (blue) despiking.

clearly shortened peaks can actually influence the PCA results and here the unfiltered sulfur dioxide concentration was used for PCA.

There are two ways to normalize the data, either to do it manually without using any built-in function or to standardize the variables automatically by the *prcomp* function. By setting the options of *center* and *scale*. to TRUE in *prcomp* function, all variables will be shifted to become zero centered and scaled to have unit variance. However, since the PCs are constructed such that each one is a linear combination of the standardized data, we choose to normalize the data manually before the PCA takes place. Thus the options *center* and *scale*. are not necessary to be set in *prcomp*.

Chapter 4

Results and Discussion

We apply PCA on standardized variables which include standardized air pollution concentrations, standardized I500 and $IR_{340/675}$. The PCs with the largest variances are most interested, because they are supposed to contain the most information and thus can be used as summary of the original data. How many PCs to be used can be decided from the 'Cumulative Proportion'.

4.1 PCA for DOAS data

Results of applying the PCA to selected gas species of the sulfur dioxide, nitrogen dioxide, formaldehyde as well as selected light intensity of I500 and selected light intensity ratio of $IR_{340/675}$ on May 8 are shown from Figure 4.1 to Figure 4.6. Before PCA, all data sets were spikes filtered except the sulfur dioxide concentration. Note that the time elements approximately before the car-parking instants during the measurements were skipped. Therefore, the time axis did not start from the zero instant.

In Figure 4.1, we can see that the standard deviations for the five PCs are about 1.40679, 1.11129, 0.98266, 0.75580 and 0.49911, respectively. This makes sense because the PCA first seeks the maximum variance from the input standardized data. Then, it removes this variance and seeks the second linear combination which explains the maximum proportion of the remaining variance, and so on. From the 'Cumulative Proportion', we can see that the first four PCs account for 95% of the total proportion, thus the PC5 can actually be neglected if data compression is required. In Figure 4.1, we see the loadings of the standardized variables, for instance for the PC2,

$$PC2 = 0.76776120 \cdot HCHO + 0.44266338 \cdot NO_2 + 0.38202331 \cdot SO_2 + 0.25627761 \cdot IR_{340/675} - 0.05451469 \cdot I_{500}. \quad (4.1)$$

```
Standard deviations:
[1] 1.4067859 1.1112923 0.9826638 0.7558040 0.4991144

Rotation:
                   PC1         PC2         PC3        PC4         PC5
HCHOaconc  -0.10332866  0.76776120 -0.25508278  0.5700417  0.09925278
NO2aconc    0.46986836  0.44266338 -0.21863076 -0.6628430  0.31002022
SO2conc    -0.03899135  0.38202331  0.91588770 -0.1016774 -0.05787314
IR_340T675 -0.58295023  0.25627761 -0.21848621 -0.4459026 -0.58985209
I_500      -0.65360203 -0.05451469  0.02338514 -0.1628622  0.73672353
Importance of components:
                          PC1    PC2    PC3    PC4     PC5
Standard deviation     1.4068 1.1113 0.9827 0.7558 0.49911
Proportion of Variance 0.3958 0.2470 0.1931 0.1143 0.04982
Cumulative Proportion  0.3958 0.6428 0.8359 0.9502 1.00000
```

Figure 4.1: PCA results of the standard deviations, the rotation coefficients, and so on in terms of the standardized variables of the SO2, NO2, HCHO, I500 and $IR_{340/675}$ on May 8.

Since the numbers in rotation matrix is orthogonal, this table can be read from left to right, for example for the standardized concentration of the

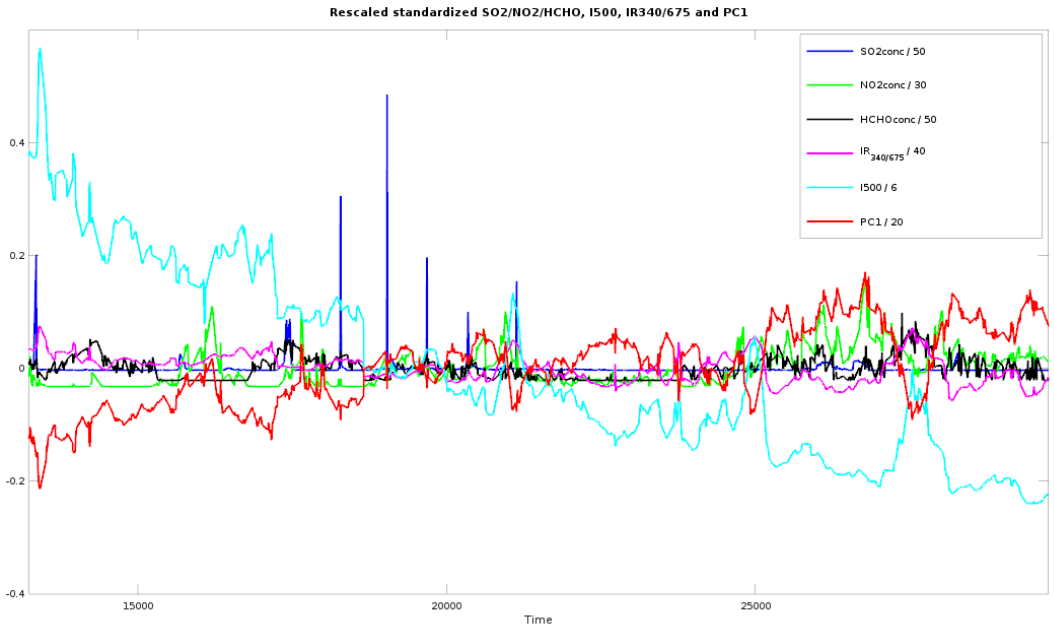


Figure 4.2: Rescaled PC1 and concentrations of the SO2, NO2 and HCHO on May 8.

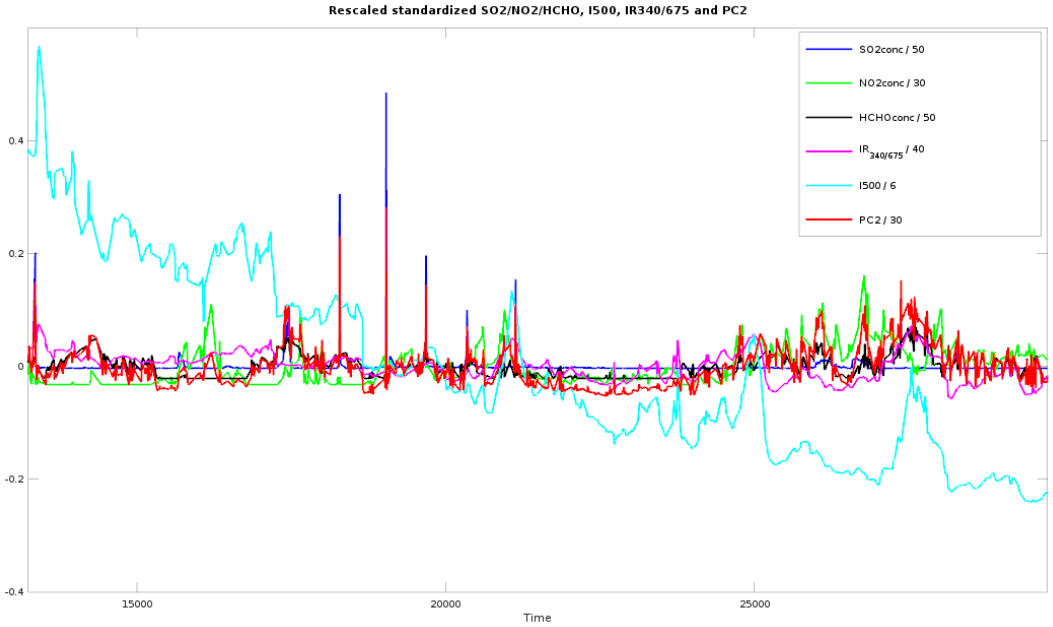


Figure 4.3: Rescaled PC2 and concentrations of the SO2, NO2 and HCHO on May 8.

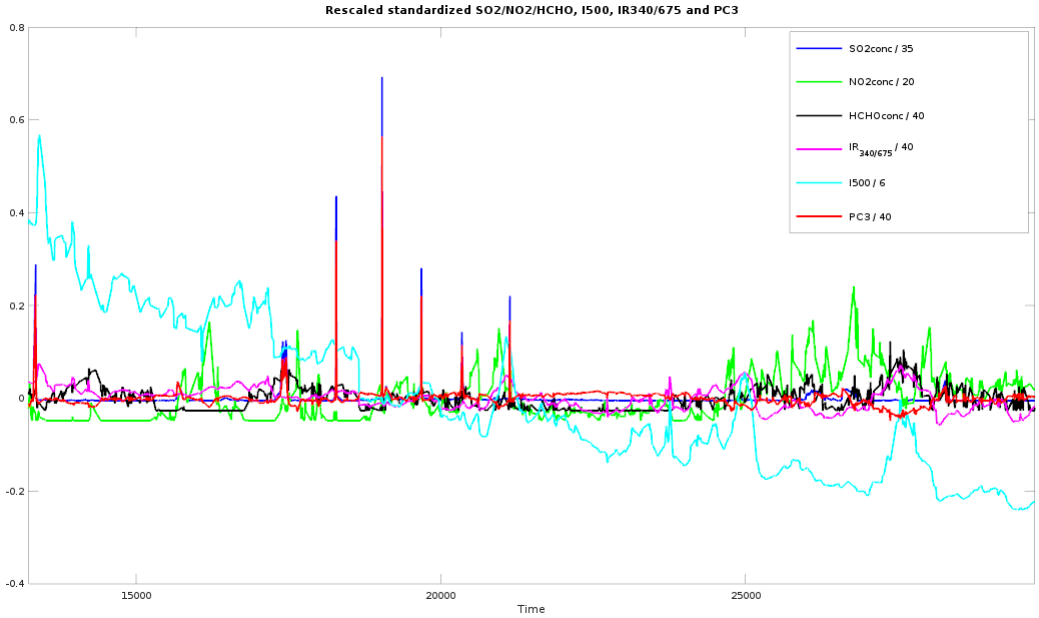


Figure 4.4: Rescaled PC3 and concentrations of the SO2, NO2 and HCHO on May 8.

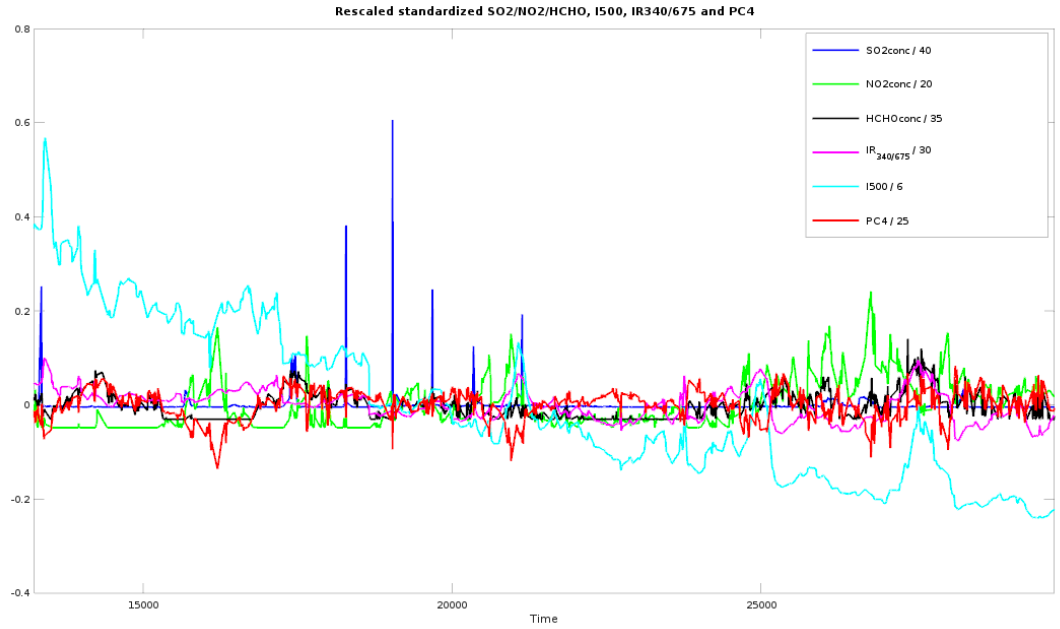


Figure 4.5: Rescaled PC4 and concentrations of the SO2, NO2 and HCHO on May 8.

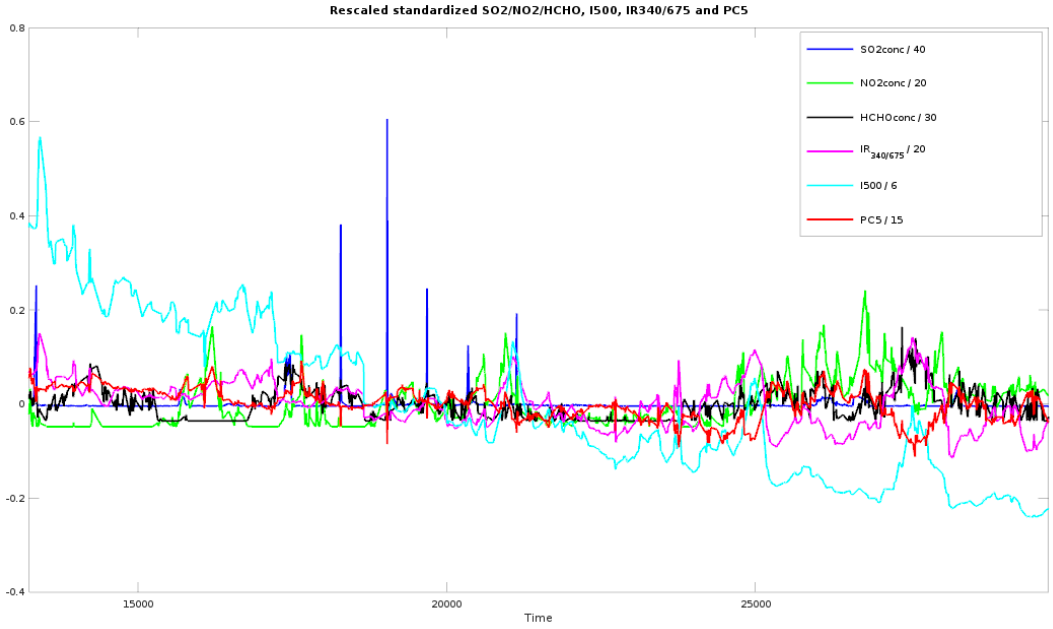


Figure 4.6: Rescaled PC5 and concentrations of the SO2, NO2 and HCHO on May 8.

formaldehyde we have,

$$HCHO = -0.10332866 \cdot PC1 + 0.76776120 \cdot PC2 - 0.25508278 \cdot PC3 + 0.5700417 \cdot PC3 + 0.09925278 \cdot PC5. \quad (4.2)$$

In Equation (4.1), formaldehyde accounts for the largest coefficient with 0.76776120, while loadings from the sulfur dioxide and nitrogen dioxide can not be neglected. Only I500 has a little negative loadings. In general, PC2 follows the variation of the selected gas species as can be seen from Figure 4.3.

In order to explore the correlation between PC2 and each contribution term, the scatter plots were generated from Figure 4.7 to Figure 4.11. Compared to other scatter plots, there is some positive correlation between formaldehyde and PC2.

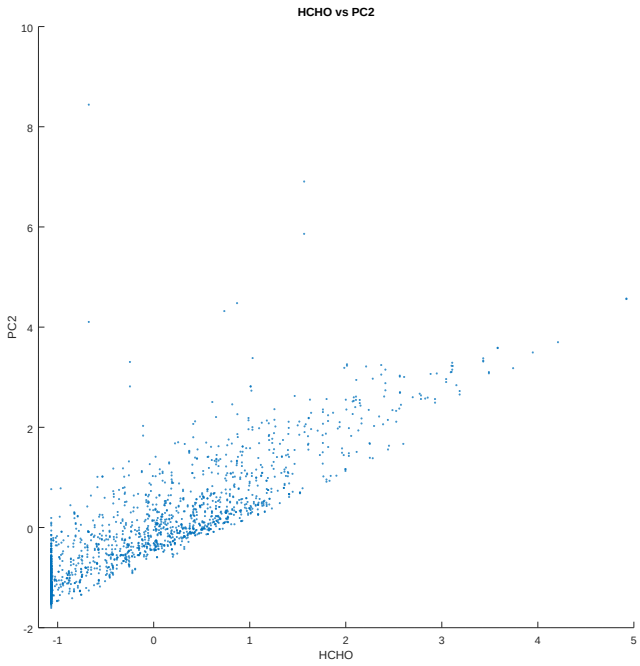


Figure 4.7: Scatter plot of HCHO and PC2 on May 8.

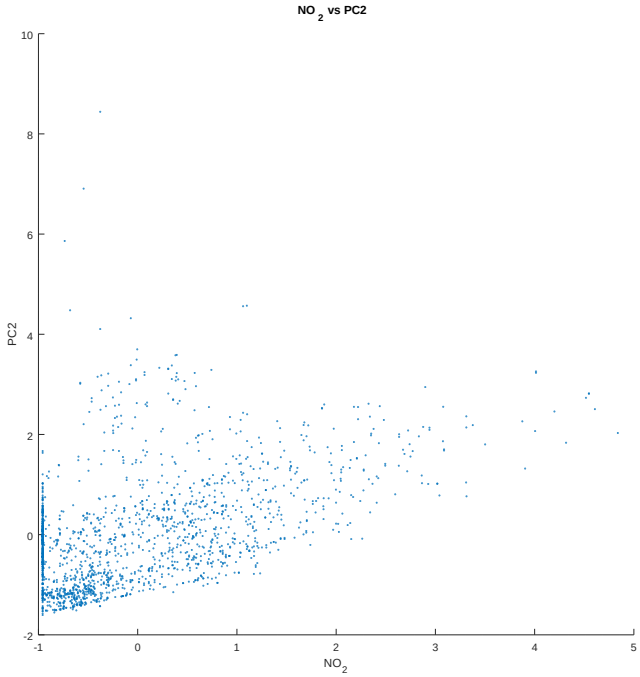


Figure 4.8: Scatter plot of NO2 and PC2 on May 8.

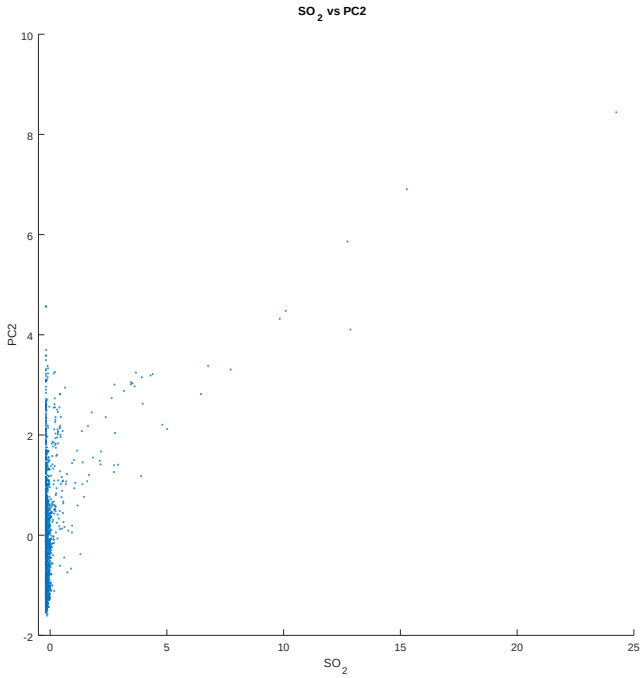


Figure 4.9: Scatter plot of SO2 and PC2 on May 8.

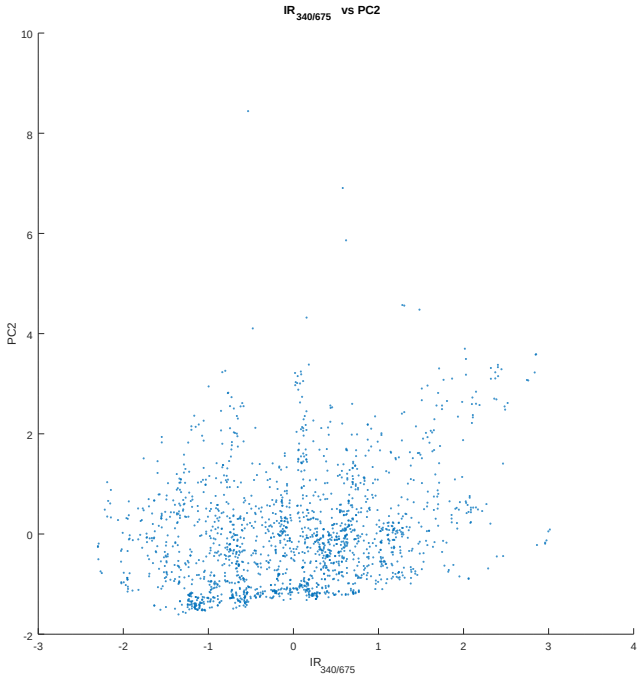


Figure 4.10: Scatter plot of IR and PC2 on May 8.

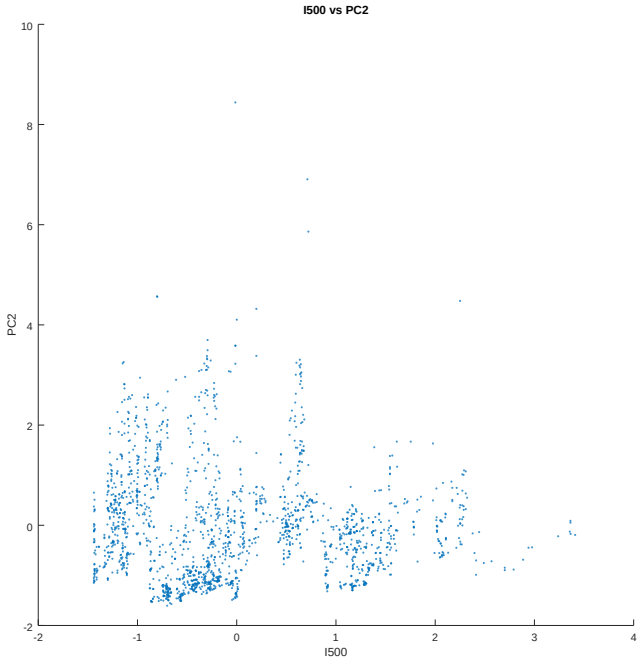


Figure 4.11: Scatter plot of I500 and PC2 on May 8.

4.2 PCA for SOF data

Results of applying the PCA to selected gas species of the ethylene, ammonia, butane as well as I500 and $IR_{340/675}$ on May 8 are shown from Figure 4.12 to Figure 4.17. Before PCA, all data sets were spikes filtered. Because the beginning and ending time for SOF measurements were '13:38:33' and '15:29:11' respectively, the time axis did not start from the zero instant.

As mentioned before, each PC is a linear combination of the standardized constituent terms. Take PC2 as an example, both ammonia and butane account for the most contributions with coefficients of 0.7083138 and 0.6752423, separately. While only small contributions come from the $IR_{340/675}$ and I_{500}. The ethylene, however, has a little negative loadings of -0.1094069. In general, PC2 follows the variation of the selected gas species as can be seen from Figure 4.14.

```
Standard deviations:
[1] 1.4112934 1.1417809 0.9703082 0.8262577 0.2835271

Rotation:
                PC1        PC2         PC3          PC4          PC5
C2H4conc    0.22925955 -0.1094069 -0.96657585  0.0006478073 -0.034653930
NH3conc    -0.05408143  0.7083138 -0.09267546 -0.6973436682 -0.022131857
C4H10conc  -0.15371405  0.6752423 -0.11273979  0.7124788254  0.009134424
IR_340T675  0.67611511  0.1197365  0.17206326  0.0686296704 -0.703004797
I_500       0.68099604  0.1266201  0.12176351  0.0370847744  0.709936690
Importance of components:
                         PC1    PC2    PC3    PC4     PC5
Standard deviation     1.4113 1.1418 0.9703 0.8263 0.28353
Proportion of Variance 0.3983 0.2607 0.1883 0.1365 0.01608
Cumulative Proportion  0.3983 0.6591 0.8474 0.9839 1.00000
```

Figure 4.12: PCA results of the rotation coefficients, the standard deviations and so on in terms of the ethylene, ammonia, butane, I500 and $IR_{340/675}$ on May 8. All variables were standardized before PCA.

From the 'Cumulative Proportion', we can see that the first four PCs account for 98% of the total proportion. We can neglect the PC5 as it contained the noise term. We might infer that PC1 has a strong correlation with $IR_{340/675}$ and I_{500} as can be seen from Figure 4.13. Meanwhile, PC2, PC3 and PC4 can reflect the selected standardized gas species components more or less as can be seen from Figure 4.14, Figure 4.15 and Figure 4.16, although the reflection may include negative contribution.

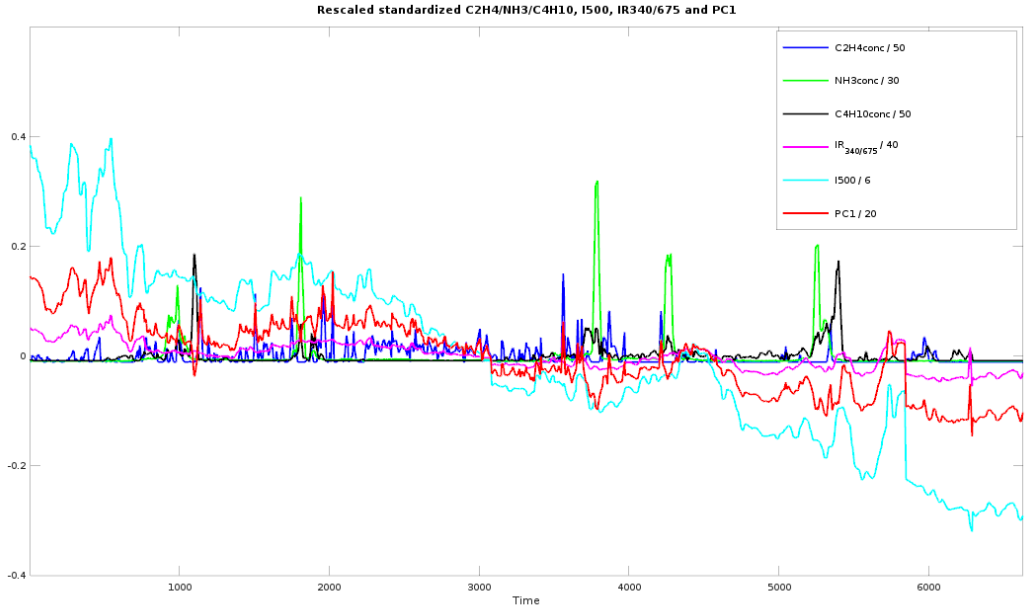


Figure 4.13: Rescaled PC1 and concentrations of the C2H4, NH3 and C4H10 on May 8.

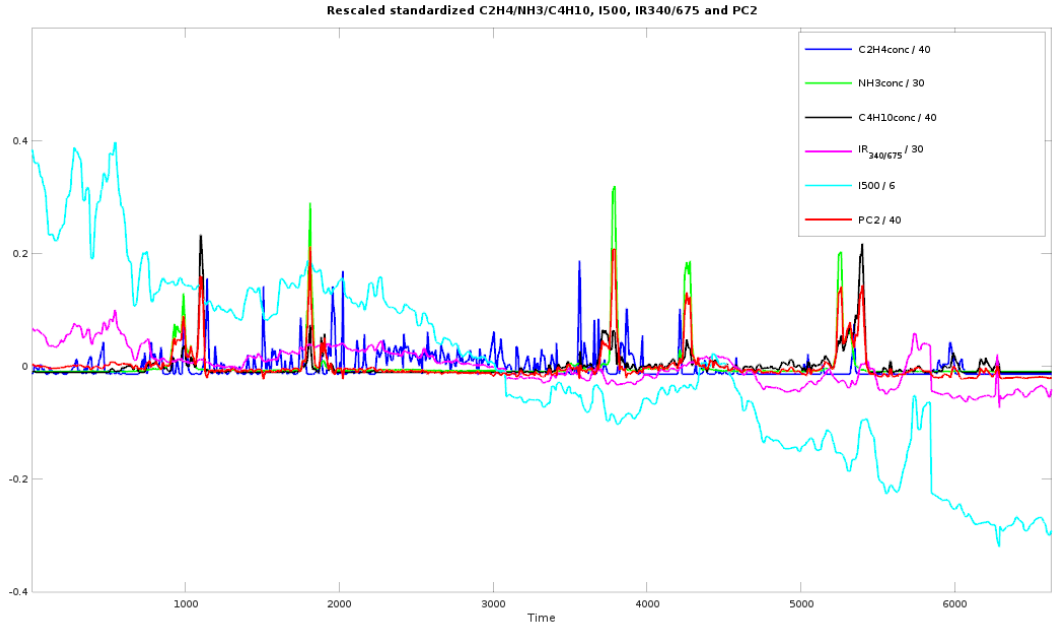


Figure 4.14: Rescaled PC2 and concentrations of the C2H4, NH3 and C4H10 on May 8.

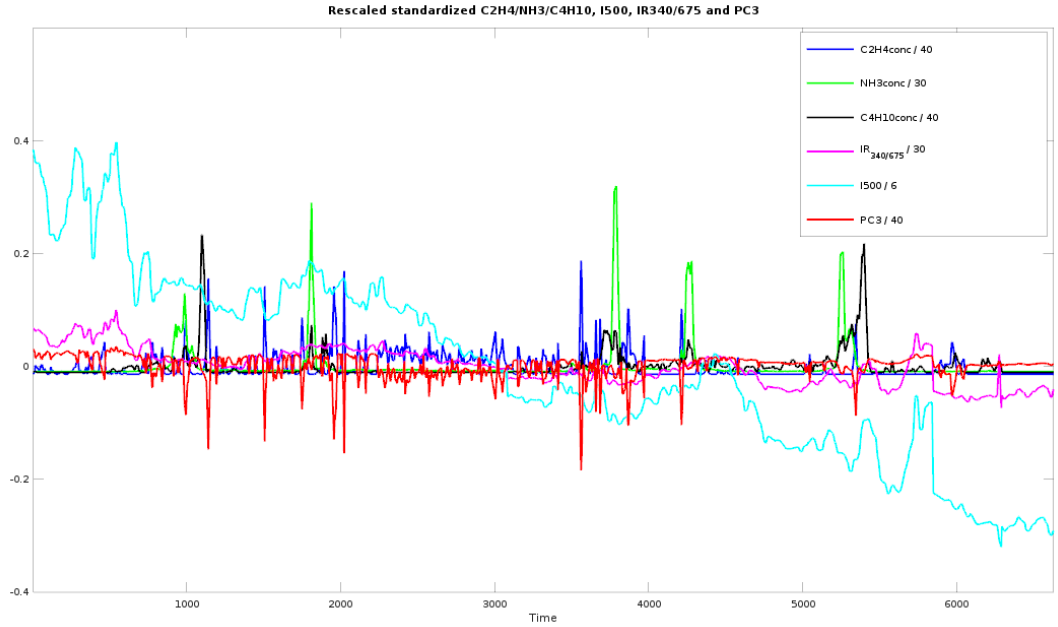


Figure 4.15: Rescaled PC3 and concentrations of the C2H4, NH3 and C4H10 on May 8.

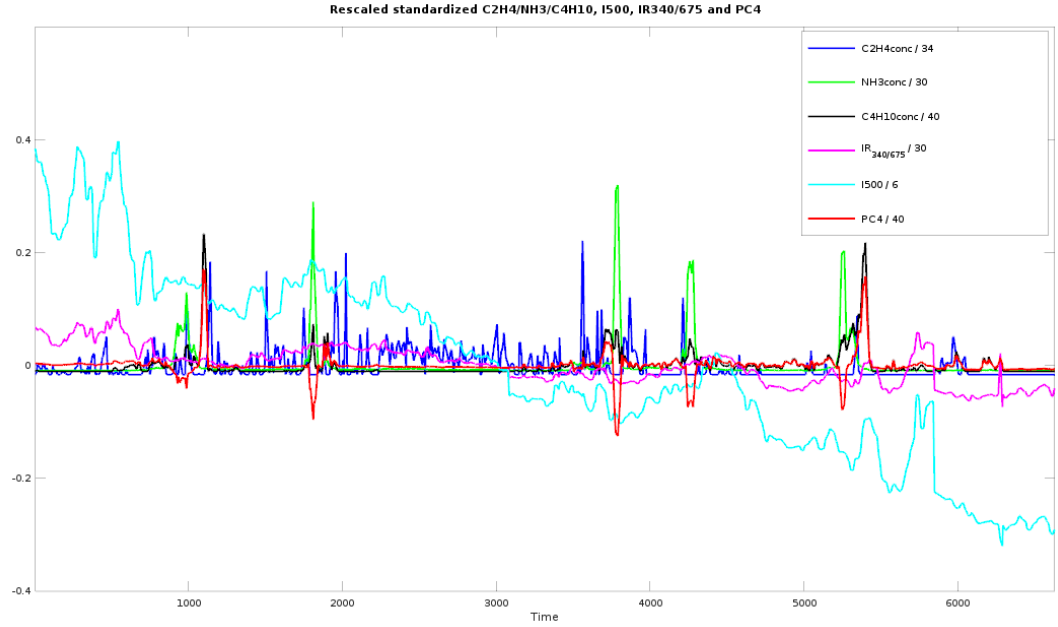


Figure 4.16: Rescaled PC4 and concentrations of the C2H4, NH3 and C4H10 on May 8.

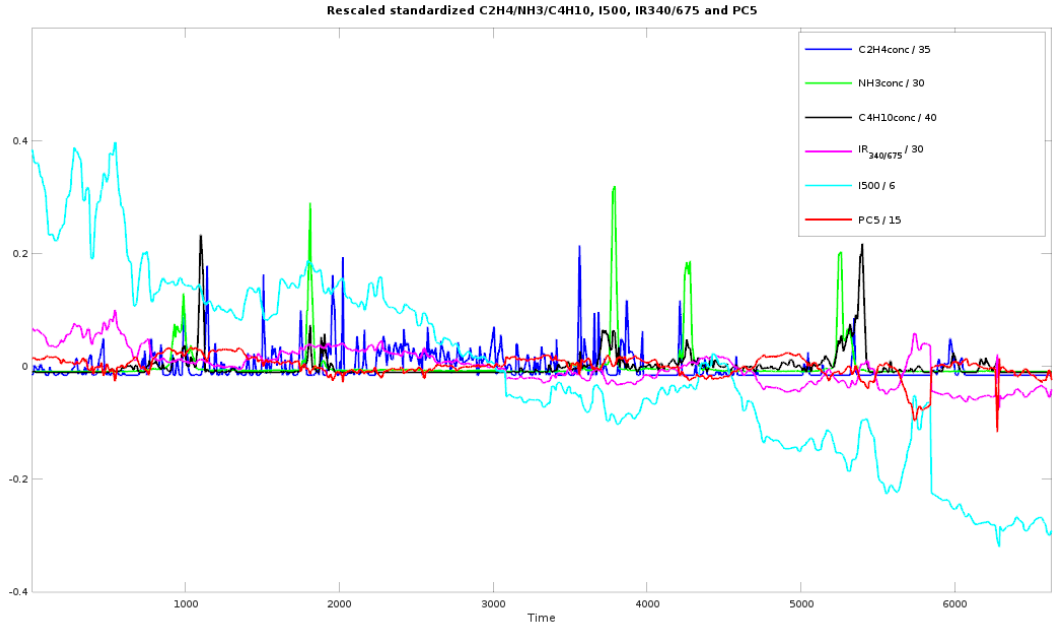


Figure 4.17: Rescaled PC5 and concentrations of the C2H4, NH3 and C4H10 on May 8.

In order to explore the correlation between PC2 and each contribution term, the scatter plots were generated from Figure 4.18to Figure 4.22.

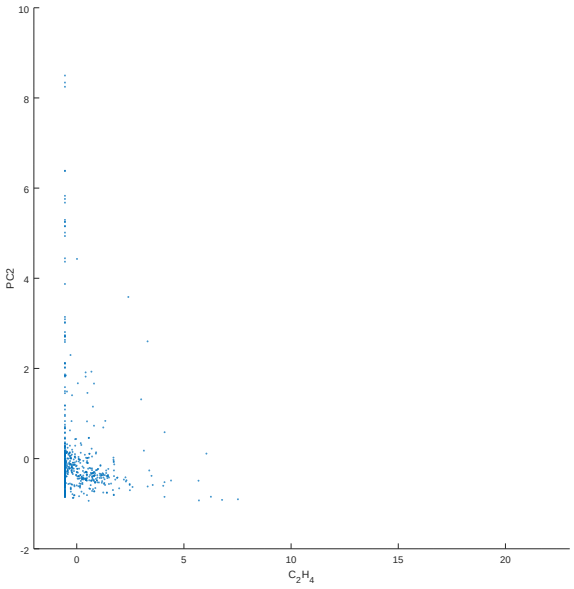


Figure 4.18: Scatter plot of C2H4 and PC2 on May 8.

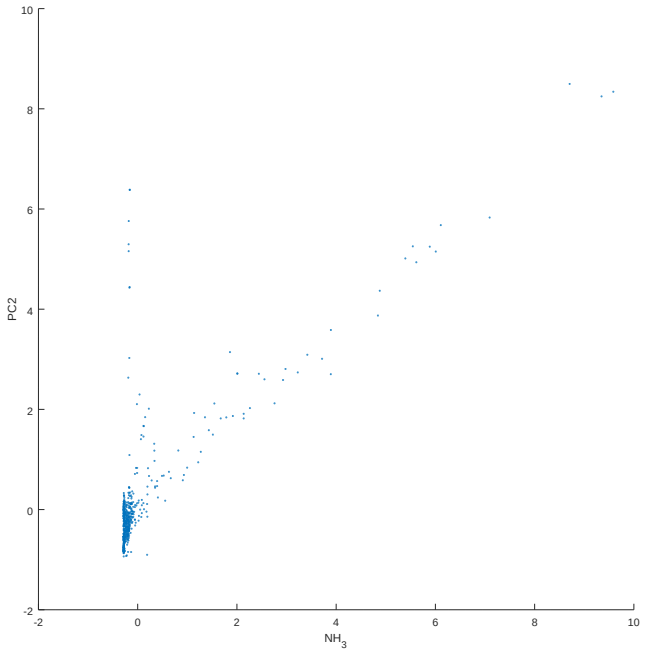


Figure 4.19: Scatter plot of NH3 and PC2 on May 8.

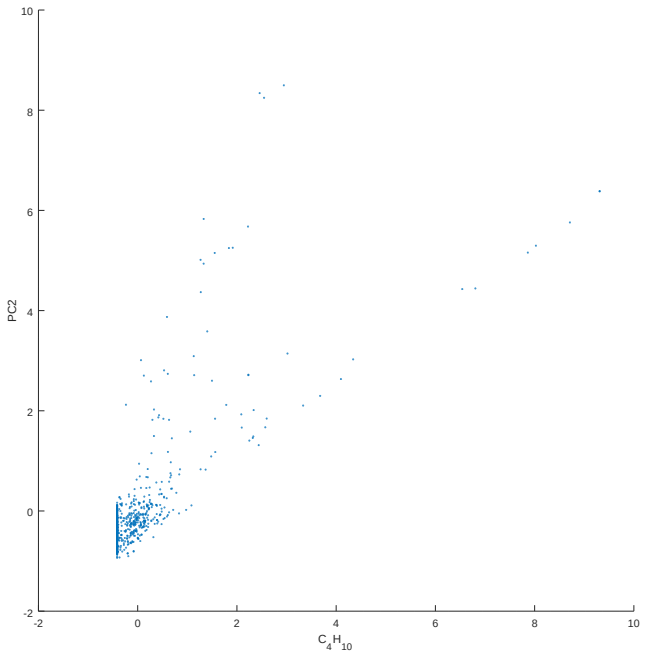


Figure 4.20: Scatter plot of C4H10 and PC2 on May 8.

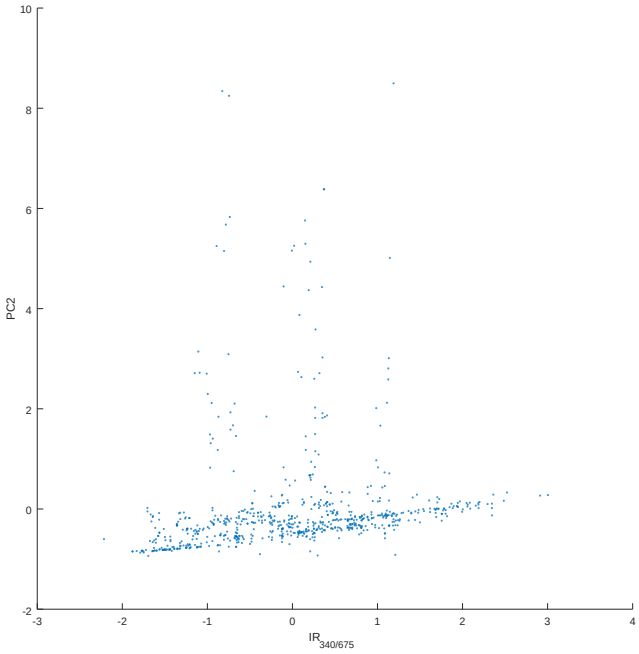


Figure 4.21: Scatter plot of IR and PC2 on May 8.

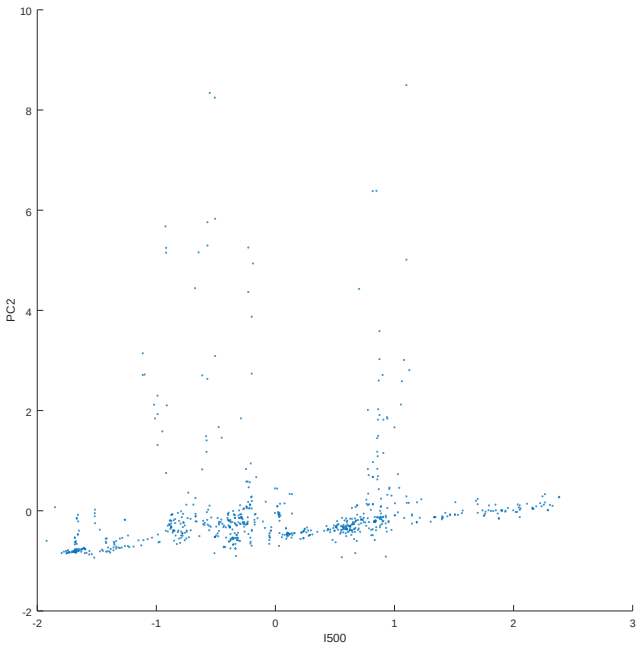


Figure 4.22: Scatter plot of I500 and PC2 on May 8.

The scatter plots among PC1 and I500, $IR_{340/675}$ can be referred in Figure 4.23 and Figure 4.24. Compared to other results, there exist positive correlations.

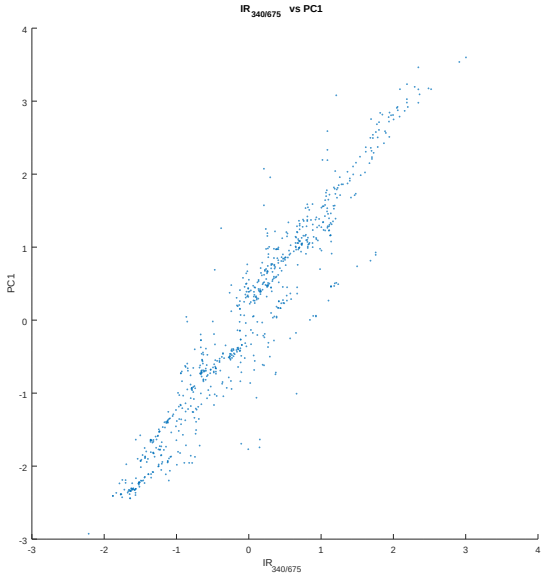


Figure 4.23: Scatter plot of IR and PC1 on May 8.

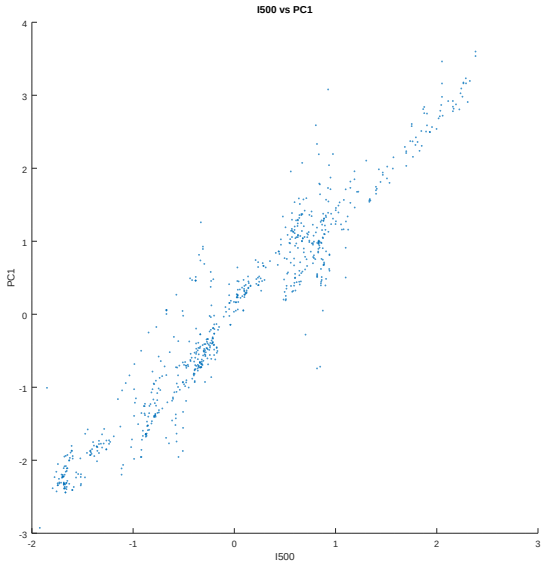


Figure 4.24: Scatter plot of I500 and PC1 on May 8.

4.3 Conclusion

The PCA method has been applied on data of the May 8, hoping to separate various gas species from the light intensity of I500 and the aerosol traces of $IR_{340/675}$.

Results of the PCA with DOAS data analysis show that the PC1 has a strong relationship with the standardized I500, although the correlation is negative as can be seen from Figure 4.7. The PC2 has a strong relationship with a combination of the standardized gas species of sulfur dioxide, nitrogen dioxide and formaldehyde as can be seen from Figure 4.8. The rest PCs either follows several re-scaled standardized concentrations superimposing together or a re-scaled light intensity in minus direction.

Results of the PCA with SOF data analysis show that the PC1 has strong positive correlation with the standardized I500 and $IR_{340/675}$ as can be seen from Figure 4.13. PC2 has strong positive correlation with the standardized concentration of the ammonia and butane as can be seen from Figure 4.14. The PC3 has a strong negative correlation with the standardized ethylene as can be seen from Figure 4.15. Most of the PC4 seems to be a superimposed result by standardized concentration of the butane and ammonia. Last bu not the least, although the rotation loading of the standardized I500 is about 0.709, the standardized $IR_{340/675}$ has an obvious negative correlation with the PC5 as can be seen from Figure 4.17.

In conclude, the PCA does not result in good reorganization and separation of each input component. The reason can be that the measured gas species are not independent with each other. Besides, the light intensity of I500 and intensity ratio of $IR_{340/675}$ show good correlation in most cases which has been done in the report of 'Optical Remote Measurements of Particles in Emission Gas Plumes'. Whether a few interpretable underlying factors or latent variables reside among the data variables requires exploratory factor analysis (EFA).

4.4 Future Work

Future work can focus on applying EFA to the data set thus finding the latent relational structure among the variables. The following up procedure can include study of the basic idea of the EFA, factor analysis to achieve simple explicable structure, validate the structure to ensure model's adequacy and arrived at the names of factor from the variables at last [8].

References

[1] G. Hoek, B. Brunekreef, S. Goldbohm, and P. Fischer, "Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study," *The Lancet*, September 2002.

[2] Q. Sun, X. Hong, and L. E. Wold, "Cardiovascular Effects of Ambient Particulate Air Pollution Exposure," *Circulation – Contemporary Reviews in Cardiovascular Medicine*, June 2010.

[3] T. Eck, B. N. Holben, J. S. Reid, and S. Kinne, "Wavelength dependence of the optical depth of biomass burning, urban, and desert dust aerosols," *Journal of Geophysical Research Atmospheres*, December 1999.

[4] H. Chen, X.-Y. Zhang, X. Xin, P. Goloub, B. Holben, and H. Zhao, "Ground-based aerosol climatology of China: aerosol optical depths from the China Aerosol Remote Sensing Network (CARSNET) 2002-2013," *Atmospheric Chemistry and Physics*, April 2015.

[5] C. Quant, P. Fischer, E. Buringh, C. Ameling, D. Houthuijs, and F. Cassee, "Application of principal component analysis to time series of daily air pollution and mortality," tech. rep., RIVM report 650010035, 2003.

[6] L. Smith, "A tutorial on Principal Components Analysis," tech. rep., University of Otago, February 2002.

[7] I. Jolliffe, *Principal Component Analysis*. Springer, 2nd ed., 2002.

[8] P. Panda, "Exploratory Factor Analysis in R." `https://www.promptcloud.com/blog/exploratory-factor-analysis-in-r/`, February 2017. Retrieved December 19, 2017.