Linköping University | Department of Computer and Information Science

Master's thesis, 30 ECTS | Statistics and Machine Learning

2021 | LIU-IDA/LITH-EX-A--2021/001--SE

Quantifying nitrogen oxides and ammonia via frequency modulation in gas sensors

– **DRAFT**

Kvantifiering av kväveoxider och ammoniak via frekvensmodulering i gassensorer

Marcos Freitas Mourão dos Santos

Supervisor : Annika Tillander
Examiner : José M. Peña

External supervisor : Mike Andersson

Linköpings universitet
SE–581 83 Linköping
+46 13 28 10 00 , www.liu.se

Abstract

The abstract resides in file `Abstract.tex`. Here you should write a short summary of your work.

Acknowledgments

Thank you for reading my draft! :)

Contents

List of Figures

List of Tables

List of acronyms and abbreviations

AC Alternating Current. 2

CV Cross Validation. 17

GBCO Gate Bias Cycled Operation. 2

Hz Hertz. 5, 7

mA miliamperes. 4

MSE Mean Squared Error. 17

NIPALS Nonlinear Iterative Partial Least Squares. 13, 15

OLS Ordinary Least Squares. v, 8, 11, 14, 15, 18

PC Principal Component. 12, 13, 14, 17

PCA Principal Components Analysis. 11, 12, 13, 14

PCR Principal Components Regression. v, 11, 14, 18

PLS Partial Least Squares. 14, 15, 17

PLSR Partial Least Squares Regression. v, 2, 11, 14, 15, 18

ppm parts per million. 4, 7

RMSE Root Mean Squared Error. 17

RSS Residual Sum of Squares. 11, 16

SAS Sensor and Actuator Systems. 4

SCR Selective Catalytic Reduction. 1, 2

SiC-FET Silicon Carbide Field Effect Transistor. 2, 4

TCO Temperature Cycled Operation. 2, 7

1 Introduction

1.1 Motivation

Nitric Oxide (NO) and Nitrogen Dioxide (NO_2), commonly referred together as NO_x, are hazardous gases to the environment and to humans. Its main sources are combustion processes in transportation, and industrial processes such as (but not limited to) auto mobiles, trucks, boats, industrial boilers, turbines, etc. (USEPA 2019).

NO_x exposure to humans can cause respiratory illnesses such bronchitis, emphysema and can worsen heart disease (Boningari and Smirniotis 2016). Environmentally, NO_x are deemed precursors of adverse phenomena such as smog, acid rain, and the depletion of ozone (O_3) (Alberto Bernabeo, Webster, and Onofri n.d.). It is of high interest, therefore, to reduce NO_x emissions.

One well studied and successful method of reducing emissions is Selective Catalytic Reduction (SCR), which consists in the reduction of NO_x by ammonia (NH_3) into nitrogen gas (N_2) and water (H_2O) (Forzatti 2001), both harmless components. The process is based in the following reactions (Forzatti 2001):

- $4\,NH_3 + 4\,NO + O_2 \longrightarrow 4\,N_2 + 6\,H_2O$
- $2\,NH_3 + NO + NO_2 \longrightarrow 2\,N_2 + 3\,H_2O$
- $8\,NH_3 + 6\,NO_2 \longrightarrow 7\,N_2 + 12\,H_2O$

One key element in these reactions, however, is the amount of ammonia dosed into the SCR systems. Ammonia itself is hazardous to humans, causing skin and respiratory irritation, among other illnesses (ASTDR 2004). More importantly, ammonia is one of the main sources of nitrogen pollution and it has direct negative impact on biodiversity via nitrogen deposition in soil and water (Guthrie, Giles, Dunkerley, Tabaqchali, Harshfield, Ioppolo, and Manville 2018).

Hence it is also desired to keep ammonia emissions to a minimum. Too much ammonia in the SCR catalyst will guarantee NO_x reduction at the expense of undesired ammonia emissions. Concurrently, too little ammonia will impede SCR to occur properly, beating the purpose of the catalyst and as a consequence, undesired NO_x emissions.

To monitor gasses concentrations, chemical sensors are deployed, one of which is the Silicon Carbide Field Effect Transistor (SiC-FET). The identification and quantification of gasses is normally achieved through multiple sensor in so called sensor arrays. Ideally each sensor in the array needs to have different responses to different compounds (Bastuck 2019). The deployment of multiple sensors, on the other hand, proves itself cumbersome due to the increased chances of failure, and decalibration of the system should one or multiple sensors be replaced (Bastuck 2019).

One solution to this problem is the cycled operation of one single sensor, referred as virtual multi-sensor (Bastuck 2019). By cycling the working point parameters of the sensor, different substances react differently in the sensor surface, which in turn produces different responses. Temperature Cycled Operation (TCO), Gate Bias Cycled Operation (GBCO), and the combination of the two have been proven to increase selectivity of SiC-FET sensors (Bastuck 2019).

TCO, in contrast with a constant temperature evaluation, produces unique transient sensor responses, i.e. each gas mixture yields a slightly different sensor output. This unique gas signature increases selectivity (Bur, Bastuck, Lloyd Spetz, Andersson, and Schütze 2014). Additionally, the high temperatures reached in these cycles help in the cleansing of the sensor surface, preparing it for the new mixtures to come.

Frequency modulation tries to achieve the same goal: avoid steady state responses in exchange of unique signatures that could help identify/quantify the gasses at hand. It consists on operating the sensor in Alternating Current (AC). One then can regulate the frequency of this operation and create cycles of different frequencies, similar to what is done in TCO. This is equivalent to GBCO, but with more frequency changes and achieving overall higher frequencies.

The main question is: given these set of unique sensor responses, how one can quantify the gasses that produced them? Multivariate regression techniques have been shown to be successful: Partial Least Squares Regression (PLSR) has been used in chemometrics extensively and it has been proven to be good at this task (Bastuck 2019) (Wold, Sjöström, and Eriksson 2001). Other multivariate regression methods, naturally, can also be used.

1.2 Aim

The aim of this thesis is to investigate if frequency modulation can be used to simultaneously quantify the concentrations of NO_x and Ammonia in a particular gas mixture.

1.3 Research questions

1. Can frequency modulation be used to simultaneously quantify NO_x and Ammonia concentrations?
2. Which method yields best predictions of gas concentrations?

2 Data

2.1 Data acquisition

The data was acquired at the Sensor and Actuator Systems (SAS) laboratory at Linköping University. The experiment — as shown on Figure 2.1 — consisted of exposing different gas combinations to two SiC-FET sensors under a particular frequency cycle and recording its response, measured in miliamperes (mA). The is then used to extract secondary features, namely average and slope values from certain regions of the frequency cycle.

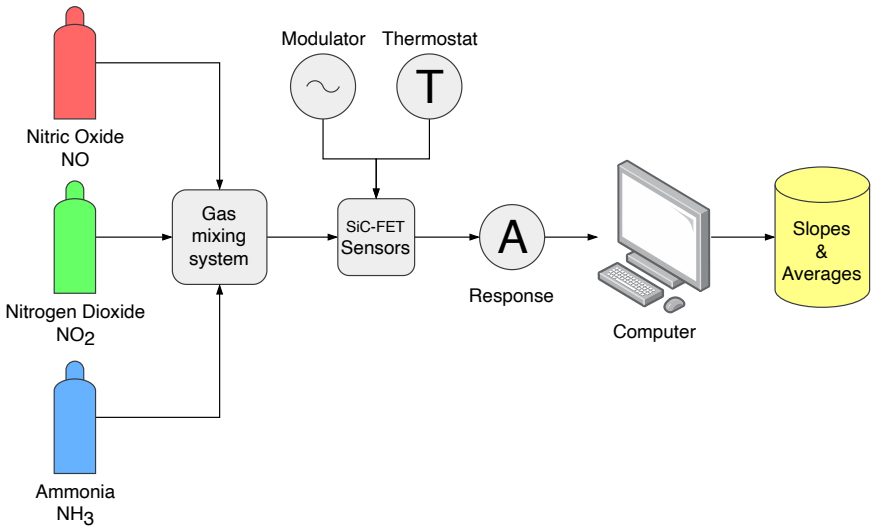

Figure 2.1: Schema of the data acquisition process.

In more detail, NO, NO_2 and NH_3 had five possible concentration values each: 5, 10, 20, 40, and 80 parts per million (ppm). The experiment was designed to encompass all possible combinations of these gases, amounting to 125 different gas mixtures. Each feature was sub-

mitted to the same frequency cycle four times. The cycle consists of 16 unique frequencies: 0.05, 0.1, 0.25, 0.5, 1, 2, 5, 10, 25, 50, 100, 200, 500, 1000, 2500 and 5000 Hertz (Hz). A typical raw sensor response for frequency modulation experiments is shown on Figure 2.2.

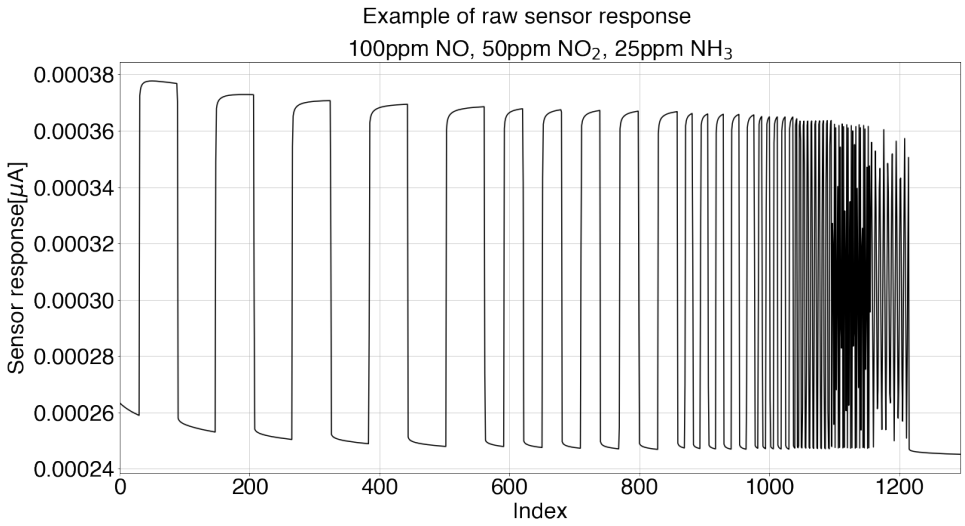

Figure 2.2: An example of raw sensor response

Throughout one cycle, several slope and average features were extracted. The sample rate for feature extraction was set at 4 Hz, i.e. in a cycle of 60 seconds, a total of $60\text{s} \times 4\frac{1}{\text{s}} = 240$ pairs of slopes and averages are recorded, which totals to 480 features per cycle. In other words, during one experiment – 4 cycles of 60 seconds – a total of $480 \times 4 = 1920$ features are extracted.

One way to visualize this process is shown in Figures 2.3. Note that the y-axis is in log-scale due to the different orders of magnitude of frequencies. Moreover, Figure 2.4 gives more insight into feature measurement and Table 2.1 summarizes the data acquisition details.

Table 2.1: Data acquisition details

Parameter	**Value**
Factors (gases)	3
Levels (concentrations)	5
Frequencies	16
Features per cycle	480
Number of cycles	4
Data points per mixture	1920
Number of mixtures	125
Features per experiment	240.000
Number of experiments	3
Total features	720.000

For specific timestamps and measurement durations, the reader is referred to Appendix A.

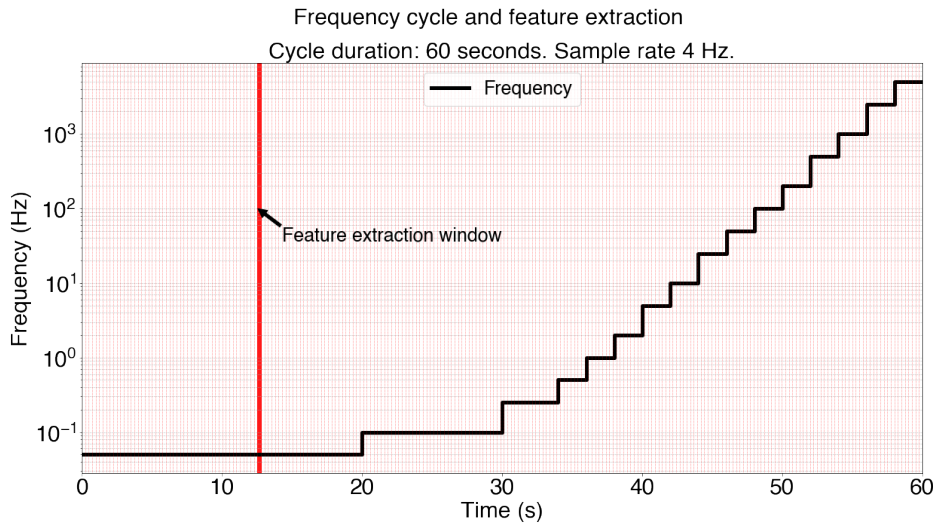

Figure 2.3: Feature measurements times per cycle. The width of the red line indicates the duration of one of the feature measurement windows as an example.

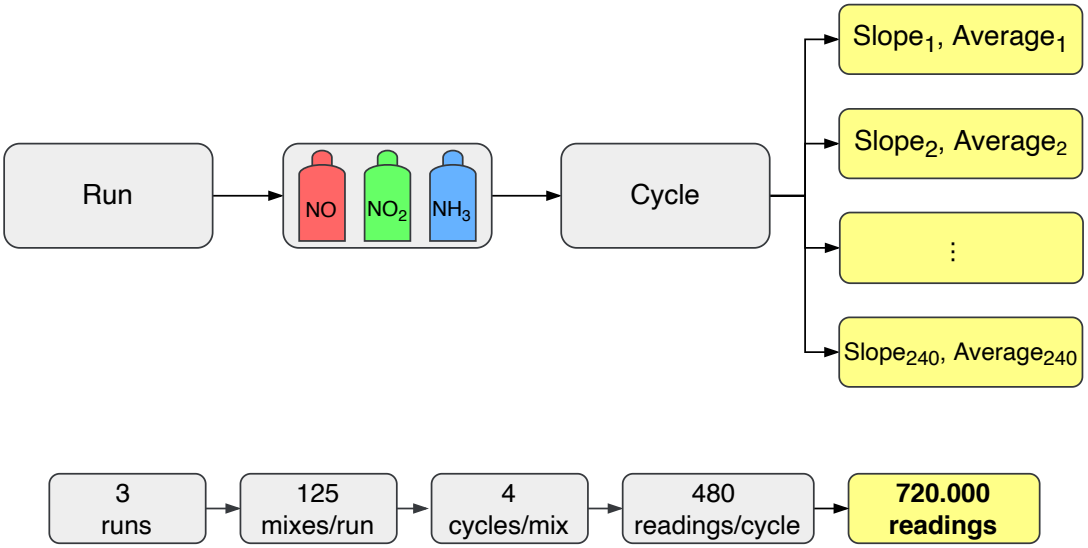

Figure 2.4: A visualization of the feature measurement process.

2.2 Raw data

The experiments were run between 26th and 29th March, 2021. The experiment data was exported as an excel file containing twelve columns, as specified in Table 2.2

Table 2.2: Raw data column details

Name	Description	Unit
Exposure nr	A particular mix of NO, NO_2 and NH_3. Ranges from 1 to 375	-
Cycle nr	The cycle number. Ranges from 1 to 4.	-
Sample nr	Extracted feature index. Ranges from 1 to 240	-
NO	Nitric Oxide concentration	ppm
NO2	Nitrogen Dioxide concentration	ppm
NH3	Ammonia concentration	ppm
Freq	Frequency	Hz
Slope sensor 1	Slope	μA/s
Slope sensor 2	Slope	μA/s
Average sensor 1	Average	μA
Average sensor 2	Average	μA
Sensor temperature	Temperature	degrees Celsius (°C)

Table 2.3: Sample of raw data.

Index	Exposure nr	Cycle nr	Sample nr	NO [ppm]	NO2 [ppm]	NH3 [ppm]	Freq [Hz]	Slope sensor 1 [uA/s]	Slope sensor 2 [uA/s]	Average sensor 1 [uA/s]	Average sensor 2 [uA/s]	Sensor temperature [C]
0	1	1	1	10	5	20	0.05	-18.855169	-22.588416	32.926184	27.961554	274.994683
1	1	1	2	10	5	20	0.05	-28.289268	-28.185027	25.853867	20.915297	274.980487
2	1	1	3	10	5	20	0.05	-0.390916	-0.482129	25.756138	20.794765	274.985895
3	1	1	4	10	5	20	0.05	-0.234549	-0.156366	25.697501	20.755673	275.020372
4	1	1	5	10	5	20	0.05	-0.143336	-0.247580	25.661667	20.693778	275.014964
							⋮					
100000	105	1	161	5	5	40	5.0	-38.366212	-48.495271	30.241896	24.821197	275.021724
100001	105	1	162	5	5	40	5.0	6.619507	8.521964	31.896773	26.951688	274.999415
100002	105	1	163	5	5	40	5.0	-1.941549	6.580416	31.411386	28.596792	275.011584
100003	105	1	164	5	5	40	5.0	27.401023	22.012900	38.261641	34.100017	275.009894
100004	105	1	165	5	5	40	5.0	-27.016623	-28.439121	31.507486	26.990236	275.014400
							⋮					
359995	375	4	236	20	80	5	5000.0	-0.136821	-0.158538	34.129879	30.345597	275.002007
359996	375	4	237	20	80	5	5000.0	0.010859	0.010859	34.132593	30.348312	274.986797
359997	375	4	238	20	80	5	5000.0	-0.043435	0.030405	34.121734	30.355913	274.979811
359998	375	4	239	20	80	5	5000.0	-0.117275	-0.026061	34.092416	30.349398	274.984543
359999	375	4	240	20	80	5	5000.0	0.073840	0.039092	34.110876	30.359171	274.998063

2.3 Pre-processing

The features (slopes and averages) from the same target (a particular exposure) in the raw data file in Table 2.3 are spread across multiple rows, which is not suitable for analysis. Opposed to TCO, the experiments were conducted at constant temperature and therefore the temperature column is discarded. The data was, subsequently, modified to have the desired format: each row containing the predictors for one particular combination of gases. Additionally, the data from each sensor was split into two datasets.

The naming convention for the features is shown in Figure 2.5. First, the frequency in which the measurement was taken followed by the sensor number. After that, the feature name itself is followed by its index, i.e. where in the frequency cycle the measurement was made. This convention allows for easy identification of key information of the cycle and measurement.

The pre-processing results in the format shown in Figure 2.6. Recalling that there are 125 possible mixtures of gases, it is important to note that there are repeated exposures in the data set and those are treated as individual observations. Since each unique gas mixture was exposed 4 times during a cycle, and the experiment was repeated 3 times, this yields a total of $4 \times 3 \times 125 = 1500$ exposures.

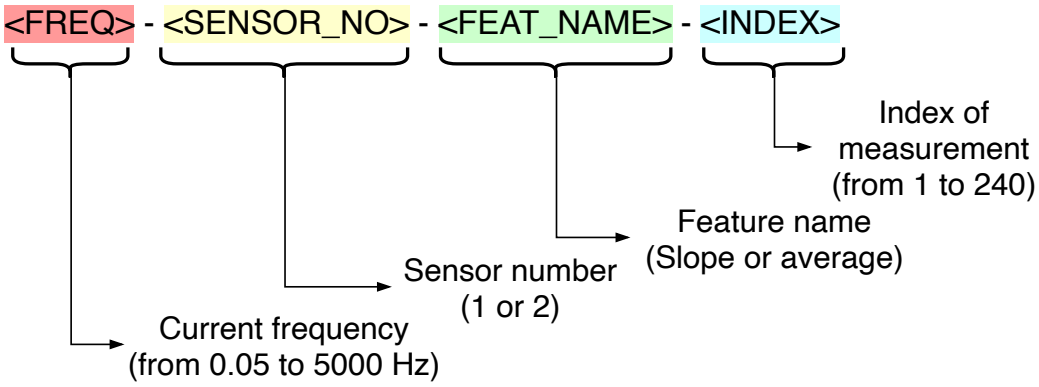

Figure 2.5: Feature naming convention.

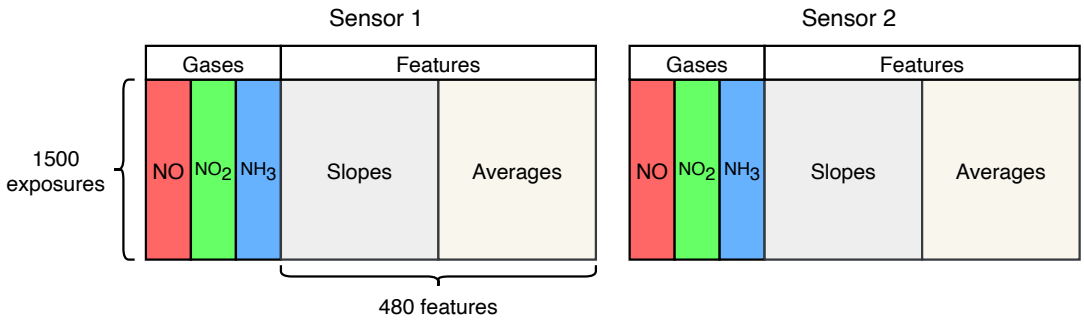

Figure 2.6: Pre-processed data structure.

Finally, the majority of models proposed are scale variant, with the exception of OLS. The data, therefore, was centered and scaled to have unit variance and zero mean. A snippet of the final data set is shown in Table 2.4

Table 2.4: Sample of pre-processed data.

Index	exposure	NO	NO2	NH3	0.05-1-slope-0	0.05-1-slope-1	...	5000.0-1-slope-239	0.05-1-avg-0	0.05-1-avg-1	0.05-1-avg-2	...	5000.0-1-avg-238	5000.0-1-avg-239
0	1.0	10.0	5.0	20.0	2.033884	-2.024181	...	1.256538	2.619349	1.757883	1.728610	...	2.276652	2.301890
1	1.0	10.0	5.0	20.0	-0.661288	0.869532	...	-0.846043	0.567699	2.477401	2.449619	...	2.522155	2.496103
2	1.0	10.0	5.0	20.0	0.232116	0.038766	...	1.546549	1.389922	2.716136	2.746701	...	2.637892	2.669321
3	1.0	10.0	5.0	20.0	-0.854716	1.071274	...	1.546549	0.592878	2.848766	2.866869	...	2.722064	2.753305
4	2.0	20.0	40.0	40.0	1.876790	-1.808299	...	-1.063551	2.963172	2.760346	2.746701	...	2.213523	2.182912
							⋮							
700	176.0	40.0	20.0	40.0	1.735014	-1.850717	...	-2.151093	1.590775	0.071822	0.072959	...	0.098694	0.046562
701	176.0	40.0	20.0	40.0	-0.466435	0.439147	...	-1.208557	-0.393152	-0.046993	-0.033022	...	0.084665	0.055310
702	176.0	40.0	20.0	40.0	-0.509181	0.495187	...	-0.725205	-0.423540	-0.015217	-0.020505	...	0.078820	0.061143
703	176.0	40.0	20.0	40.0	-0.494932	0.467598	...	-2.296098	-0.409070	-0.031795	-0.043871	...	0.070052	0.014485
704	177.0	80.0	40.0	40.0	1.763512	-1.868305	...	2.005733	1.608140	0.076796	0.059607	...	0.074144	0.122381
							⋮							
1495	374.0	80.0	80.0	40.0	-0.429032	0.485703	...	-0.725205	-1.122473	-1.365005	-1.369615	...	-1.476030	-1.490233
1496	375.0	20.0	80.0	5.0	1.083129	-1.205831	...	-1.426065	0.366557	-1.232375	-1.238876	...	-1.479537	-1.510646
1497	375.0	20.0	80.0	5.0	-0.635640	0.696067	...	0.507342	-1.303357	-1.373294	-1.352369	...	-1.461417	-1.445908
1498	375.0	20.0	80.0	5.0	-0.016883	0.086357	...	0.120661	-0.768521	-1.329084	-1.358488	...	-1.481875	-1.475653
1499	375.0	20.0	80.0	5.0	-0.588619	0.638821	...	0.676515	-1.247789	-1.358926	-1.369615	...	-1.486551	-1.466904

3 Theory

The quantification of gases based on the sensor response can be viewed as a multivariate multiple regression problem where the predictors, i.e. features derived from the sensor signal, are used to predict multiple responses, i.e. the concentrations of pertinent gases. This chapter exposes the theory behind some of these models.

The models here listed were chosen as a natural progression from a statisticians point of view: starting with simple models and progressively increasing complexity as insights from the data and the problem are gathered.

3.1 Notation

In favor of consistency and clarity, the notation used throughout this work is presented here. Bold capital letters, e.g. $\mathbf{A}$, are matrices while bold lower case letters are row vectors, e.g. $\mathbf{a}$. Scalars, on the other hand, are denoted as standard lower case letters, e.g. a_{i1}. Transposes and inverses are denoted, respectively with $\cdot^{\top}$ and $\cdot^{-1}$. An example is shown below.

$$\mathbf{A} = [\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_n}]^{\top}$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{a_1} \\ \mathbf{a_2} \\ \vdots \\ \mathbf{a_n} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix}$$

3.2 Ordinary Least Squares Regression

A simple, first approach would be to tackle the problem with a Ordinary Least Squares (OLS) regression model. As Friedman, Hastie, Tibshirani, et al. 2001 explains, each output $\mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_K}]^\top$ has its own linear model. Now, given a set of n observations $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}]^\top$ and each observation having p features, e.g. $\mathbf{x}_i = [x_{i1}, x_{i2}, ...x_{ip}]$, $i = 1, 2, ...n$, the concatenation of all linear models can be written in matrix form as in Equation 3.1.

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{3.1}$$

Where:

- $\mathbf{B}$: $[p+1 \times K]$ matrix of regression coefficients (with the $+1$ referring to the intercept term);
- $\mathbf{E}$: $[N \times K]$ matrix of residuals.

The objective is then to find the coefficients $\hat{\mathbf{B}}$ which minimizes the Residual Sum of Squares (RSS), which is summarized by Equation 3.3 (Friedman, Hastie, Tibshirani, et al. 2001):

$$\hat{\mathbf{B}}^{\text{OLS}} = \arg\min_{\mathbf{B}} \text{RSS}(\mathbf{B}) \tag{3.2}$$

In turn, the RSS, as the name suggests, is defined as the difference between real and predicted values, squared, which in matrix form is written as (Friedman, Hastie, Tibshirani, et al. 2001):

$$\text{RSS}(\mathbf{B}) = \text{Tr}[(\mathbf{Y} - \mathbf{XB})^\top(\mathbf{Y} - \mathbf{XB})] \tag{3.3}$$

Finally, solving for $\hat{\mathbf{B}}$ yields (Friedman, Hastie, Tibshirani, et al. 2001):

$$\hat{\mathbf{B}}^{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \tag{3.4}$$

For the problem at hand, in addition to the high number of features, it is often the case that sensor data points are acquired in quick succession, which in turn leads to highly correlated features (Bastuck 2019), which can result in high variance in a least squares model (Friedman, Hastie, Tibshirani, et al. 2001). It is natural, therefore, to progress towards methods that incorporate dimensionality reduction such as Principal Components Regression (PCR) and Partial Least Squares Regression (PLSR) or shrinkage such as Ridge Regression.

3.3 Principal Component Analysis

One way to define Principal Components Analysis (PCA) is to view it as a orthogonal projection of the data into a principal space of lower dimension such that the variance of this projection is maximized (Bishop 2006).

Just as before, consider the collection of n observations $\mathbf{X}$ with covariance matrix $\mathbf{\Sigma}$. Additionally, consider a matrix $\mathbf{P} = [\mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p_n}]^\top$ where $\mathbf{p_i}$ is a row vector of coefficients referring to the i-th linear combination (Johnson and Wichern 2013):

$$t_i = \mathbf{x_i}\mathbf{p_i}^\top \qquad i = 1, 2, ..., n \tag{3.5}$$

The variance and covariance of these new variables t_i can be written as follows:

$$\text{Var}(t_i) = \mathbf{p_i}^\top \mathbf{\Sigma} \mathbf{p_i} \qquad i = 1, 2, ..., n \tag{3.6}$$

$$\text{Cov}(t_i, t_k) = \mathbf{p_i}^\top \mathbf{\Sigma} \mathbf{p_k} \qquad i, k = 1, 2, ..., n \tag{3.7}$$

The first Principal Component (PC) is then the linear combination with maximum variance, i.e. the linear combinations that maximizes $\text{Var}(t_1)$, with the constraint that the coefficient vector $\mathbf{p_1}$ has unit length. In summary, the first PC is computed as (Johnson and Wichern 2013):

$$\begin{aligned} &t_1 = \mathbf{x_1}\mathbf{p_1}^\top \\ &\quad \text{that maximizes } \text{Var}(\mathbf{x_1}\mathbf{p_1}^\top) \\ &\quad \text{subject to } \mathbf{p_1}^\top\mathbf{p_1} = 1 \end{aligned} \tag{3.8}$$

The second PC, similarly to the first, is the linear combination with maximum variance, but with an added extra constraint: this new linear combination must be orthogonal to the previous one, i.e. they must be linearly independent:

$$\begin{aligned} &t_2 = \mathbf{x_2}\mathbf{p_2}^\top \\ &\quad \text{that maximizes } \text{Var}(\mathbf{x_2}\mathbf{p_2}^\top) \\ &\quad \text{subject to } \mathbf{p_2}^\top\mathbf{p_2} = 1 \\ &\quad \text{and } \text{Cov}(t_1, t_2) = 0 \end{aligned} \tag{3.9}$$

The k-th PC is then:

$$\begin{aligned} &t_k = \mathbf{x_k}\mathbf{p_k}^\top \\ &\quad \text{that maximizes } \text{Var}(\mathbf{x_k}\mathbf{p_k}^\top) \\ &\quad \text{subject to } \mathbf{p_k}^\top\mathbf{p_k} = 1 \\ &\quad \text{and } \text{Cov}(t_j, t_k) = 0 \quad \text{for } k > j \end{aligned} \tag{3.10}$$

In summary, the objective of PCA is find a matrix $\mathbf{P}$ such that the linear transformation

$$\mathbf{T} = \mathbf{X}\mathbf{P}^\top \tag{3.11}$$

yields new variables that are uncorrelated and arranged in decreasing order of variance.

It can be shown that these desired linear combinations can be written in terms of the eigenvalues (λ) and eigenvectors ($\mathbf{e}$) of $\mathbf{\Sigma}$, the covariance matrix of $\mathbf{X}$ (Johnson and Wichern 2013). The elements of eigenvectors are called loadings, while the new features $\mathbf{T}$ are . In short, for the k-th PC:

$$\begin{aligned} t_k &= \mathbf{X_k e_k^\top} \\ \text{Var}(t_k) &= \mathbf{e_k^\top \Sigma e_k} = \lambda_k \\ \text{Cov}(t_j, t_k) &= \mathbf{e_k^\top \Sigma e_j} = 0 \quad \text{for } k \neq j \end{aligned} \tag{3.12}$$

There are several ways of computing PCs. Many of which involving finding aforementioned eigenvalues and eigenvectors. These calculations can be computationally expensive, depending on the desired number of extracted PCs (Bishop 2006). One option is the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm, also called Power Method. It has two clear advantages: "it can handle missing data and computes the components sequentially" (Dunn 2021).

The NIPALS algorithm to compute the first k-th PCs, $\text{PC}_i, \ i = 1, 2, ..., k$ us displayed below as Algorith 1 (Dunn 2021) (Ng 2013) (Wright 2017). Since it computes the loadings and scores sequentially, it is possible to stop it as early as desired. The "truncated" loadings and scores that project $\mathbf{X}$ into the principal subspace of k PCs is defined in Equation 3.13 :

$$\mathbf{T_{|k}} = \mathbf{X P_{|k}^\top} \tag{3.13}$$

Algorithm 1: Nonlinear Iterative Partial Least Squares (NIPALS) for PCA

Result: Matrices of loadings $\mathbf{P_{|k}}$ and scores $\mathbf{T_{|k}}$ of the k-th first Principal Components

Initialize $\mathbf{T_{|k}}$ and $\mathbf{P_{|k}}$
i = 1
$\mathbf{X_1} \coloneqq \mathbf{X}$
while $i < k$ **do**
  repeat
    Choose $\mathbf{t_i}$ as any column of $\mathbf{X_i}$
    Compute loadings $\mathbf{p_i} = (\mathbf{t_i^\top t_i})^{-1} \mathbf{t_i^\top X_i}$
    Scale $\mathbf{p_i} = \frac{\mathbf{p_i}}{\sqrt{\mathbf{p_i^\top p_i}}}$
    Compute scores $\mathbf{t_i} = (\mathbf{p_i^\top p_i})^{-1} \mathbf{p_i^\top X_i}$
  until $\mathbf{t_i}$ *converges*
  Append $\mathbf{t_i}$ to $\mathbf{T_{|k}}$
  Append $\mathbf{p_i}$ to $\mathbf{P_{|k}}$
  Deflate: $\mathbf{X_{i+1}} = \mathbf{X_i} - \mathbf{t_i p_i^\top}$
  $i \mathrel{+}= 1$
end
return $\mathbf{T_{|k}}, \mathbf{P_{|k}}$

In words, the main idea of the algorithm can be summarized as choosing an arbitrary column of $\mathbf{X}$ as the scores vector $\mathbf{t_i}$, shown in line 6. Next, the computation of the i-th loadings vector $\mathbf{p_i}$ by regressing every column of $\mathbf{X}$ via OLS onto the scores $\mathbf{t_i}$. $\mathbf{p_i}$ is then scaled to have unit length in Line 8, which in turn is used to compute the i-th scores vector $\mathbf{t_i}$ by regressing every column of $\mathbf{X}$ via OLS onto the loadings $\mathbf{p_i}$, shown in Line 9. This procedure is repeated until change in $\mathbf{t_i}$ between iterations is small enough. Once convergence is achieved, scores $\mathbf{t_i}$ and loadings $\mathbf{p_i}$ are stored as the i-th column of matrices $\mathbf{T}$ and $\mathbf{P}$ of Equation 3.11, respectively. Finally, the variability explained by $\mathbf{t_i}$ and $\mathbf{p_i}$ from $\mathbf{X}$ is subtracted in a procedure called deflation.

3.4 Principal Component Regression

With the inner workings of PCA explained in the previous section, PCR can be simply reduced to a Least Squares regression on the first k-th PCs, i.e. performing linear regression on $\mathbf{T_{|k}}$ instead of $\mathbf{X}$:

$$\mathbf{Y} = \mathbf{T_{|k}}\mathbf{B} + \mathbf{E} \tag{3.14}$$

And the regression coefficients are found analogously to Equation 3.4:

$$\hat{\mathbf{B}}^{\text{PCR}} = (\mathbf{T}_{|\mathbf{k}}^{\top}\mathbf{T_{|k}})^{-1}\mathbf{T}_{|\mathbf{k}}^{\top}\mathbf{Y} \tag{3.15}$$

Although useful, PCR has a potential flaw: while the new found projection of $\mathbf{X}$ is guaranteed to best explain the variance of predictors, this cannot be said about the responses $\mathbf{Y}$ (James, Witten, Hastie, and Tibshirani 2013). PLSR, on the other hand, solves this issue by supervising the identification of PCs (James, Witten, Hastie, and Tibshirani 2013).

3.5 Partial Least Squares Regression

PLSR, much like PCR, also aims to reduce dimensionality via linear combinations of the inputs. This technique, however, also takes into account the response variables $\mathbf{Y}$. One key advantage of PLSR is that it seeks axes with most variance (like PCR) and high correlation with response variables (Friedman, Hastie, Tibshirani, et al. 2001).

The main idea can be described as finding linear combinations for the data matrix $\mathbf{X}$ and response matrix $\mathbf{Y}$ as follows (Ng 2013), similarly to what was done in Section 3.3. Here, the matrices $\mathbf{W}$ and $\mathbf{U}$ are score matrices, i.e. the transformed PLS variables, and $\mathbf{L}$ and $\mathbf{Q}$ are loading matrices, i.e. the weights of this transformation (projection).

$$\mathbf{W} = \mathbf{X}\mathbf{L}^{\top} \tag{3.16}$$

$$\mathbf{U} = \mathbf{Y}\mathbf{Q}^{\top} \tag{3.17}$$

Instead of simply running NIPALS on $\mathbf{X}$ and $\mathbf{Y}$ separately. PLSR uses information from $\mathbf{Y}$ to decompose $\mathbf{X}$ and *vice-versa* (Ng 2013). Algorithm 2 is an adaptation of Algorithm 1 to incorporate this intended behavior.

Algorithm 2: NIPALS for Partial Least Squares Regression (PLSR)

Result: Matrices of loadings $\mathbf{L_{|k}}, \mathbf{Q_{|k}}$ and scores $\mathbf{W_{|k}}, \mathbf{U_{|k}}$ of the k-th first Partial Least Squares directions

1 Initialize $\mathbf{L_{|k}}, \mathbf{Q_{|k}}$ and $\mathbf{W_{|k}}, \mathbf{U_{|k}}$
2 i = 1
3 $\mathbf{X_1} \coloneqq \mathbf{X}$
4 $\mathbf{Y_1} \coloneqq \mathbf{Y}$
5 **while** $i < k$ **do**
6 **repeat**
7 Choose $\mathbf{u_i}$ as any column of $\mathbf{Y_i}$
8 Compute loadings of $\mathbf{X_i}$ based on score of $\mathbf{Y_i}$: $\ell_\mathbf{i} = (\mathbf{u_i^\top u_i})^{-1}\mathbf{u_i^\top X_i}$
9 Scale $\ell_\mathbf{i} = \frac{\ell_\mathbf{i}}{\sqrt{\ell_\mathbf{i}^\top \ell_\mathbf{i}}}$
10 Compute score of $\mathbf{X_i}$: $\mathbf{w_i} = (\ell_\mathbf{i}^\top \ell_\mathbf{i})^{-1}\ell_\mathbf{i}^\top \mathbf{X_i}$
11 Compute loadings of $\mathbf{Y_i}$ based on score of $\mathbf{X_i}$: $\mathbf{q_i} = (\mathbf{w_i^\top w_i})^{-1}\mathbf{w_i^\top Y_i}$
12 Scale $\mathbf{q_i} = \frac{\mathbf{q_i}}{\sqrt{\mathbf{q_i^\top q_i}}}$
13 Compute score of $\mathbf{Y_i}$: $\mathbf{u_i} = (\mathbf{q_i^\top q_i})^{-1}\mathbf{q_i^\top Y_i}$
14 **until** $\mathbf{u_i}$ *converges*
15 Append $\mathbf{w_i}$ to $\mathbf{W_{|k}}$
16 Append $\ell_\mathbf{i}$ to $\mathbf{L_{|k}}$
17 Append $\mathbf{u_i}$ to $\mathbf{U_{|k}}$
18 Append $\mathbf{q_i}$ to $\mathbf{Q_{|k}}$
19 Deflate $\mathbf{X_i}$: $\mathbf{X_{i+1}} = \mathbf{X_i} - \mathbf{w_i}\ell_\mathbf{i}^\top$
20 Deflate $\mathbf{Y_i}$: $\mathbf{Y_{i+1}} = \mathbf{Y_i} - \mathbf{u_i q_i^\top}$
21 $i \mathrel{+}= 1$
22 **end**
23 **return** $\mathbf{W_{|k}}, \mathbf{L_{|k}}, \mathbf{U_{|k}}, \mathbf{Q_{|k}}$

As with Algorithm 1, Algorithm 2 can be summarized as choosing a column of $\mathbf{Y_i}$ as the initial response score vector $\mathbf{u_i}$. After that, the i-th loadings vector $\mathbf{w_i}$ of $\mathbf{X}$ is computed in Line 8 by regressing every column of $\mathbf{X}$ via OLS onto scores vector of $\mathbf{Y}$, $\mathbf{u_i}$. Similarly to before, the data loadings vector $\mathbf{w_i}$ is scaled to have unit length, which in turn is used to compute the i-th data scores vector $\mathbf{w_i}$ by regressing every column of $\mathbf{X_i}$ via OLS onto the column $\ell_\mathbf{i}$ in Line 10. Now, the i-th response loadings vector $\mathbf{q_i}$ of $\mathbf{Y_i}$ by regressing every column of $\mathbf{Y}$ via OLS onto scores vector of $\mathbf{X}$, $\mathbf{w_i}$, shown in Line 11; This loadings vector is also scaled to have unit length.

Following in Line 13, the i-th response scores vector $\mathbf{u_i}$ is computed by regressing every column of $\mathbf{Y_i}$ via OLS onto the column $\mathbf{q_i}$. This procedure is repeated until change in $\mathbf{u_i}$ between iterations is small enough. In that case the results $\mathbf{w_i}$ and $\ell_\mathbf{i}$ are stored as the i-th

column of matrices $\mathbf{W}$ and $\mathbf{L}$ of Equation 3.16 and $\mathbf{u_i}$ and $\mathbf{q_i}$ are stored as the i-th column of matrices $\mathbf{U}$ and $\mathbf{Q}$ of Equation 3.17. Finally, the variability explained by $\mathbf{w_i}, \ell_\mathbf{i}$ and $\mathbf{u_i}, \mathbf{q_i}$ from $\mathbf{X_i}$ and $\mathbf{Y_i}$, respectively, are removed.

After finding the k partial least squares directions from Algorithm 2 above, the problem, as in Section 3.4, reduces to performing Least Squares Regression using the newfound transformations.

$$\mathbf{Y} = \mathbf{W}_{|\mathbf{k}}\mathbf{B} + \mathbf{E} \tag{3.18}$$

Which in turn, analogously to Equations 3.4 and 3.15, yields the coefficients:

$$\hat{\mathbf{B}}^{\text{PLSR}} = (\mathbf{W}_{|\mathbf{k}}^{\top}\mathbf{W}_{|\mathbf{k}})^{-\mathbf{1}}\mathbf{W}_{|\mathbf{k}}^{\top}\mathbf{Y} \tag{3.19}$$

3.6 Ridge Regression

Ridge regression is also a viable alternative to reduce the problem of highly correlated features (Friedman, Hastie, Tibshirani, et al. 2001). Instead of fitting a least squares model on a subset of predictors or a transformation of them, Ridge allows the use of all features with a continuous shrinkage of its coefficients, which results in less variance (Friedman, Hastie, Tibshirani, et al. 2001).

For the multi-output case, there are two options: use the same penalization parameter λ for all variables $\mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_K}]^{\top}$ or apply different parameters $\boldsymbol{\lambda} = [\lambda_\mathbf{1}, \lambda_\mathbf{2}, ..., \lambda_\mathbf{K}]^{\top}$. In this work, the latter is preferred over the former, as it allows a more fine tuned control of the regression models for each studied gas.

Analogous to Section 3.2, the goal is to minimize the RSS, but now with the penalization term taken into account. Equation 3.20 below shows this objective function in matrix form.

$$\text{RSS}^{\text{Ridge}}(\mathbf{B}, \boldsymbol{\lambda}) = \text{Tr}[(\mathbf{Y} - \mathbf{XB})^{\top}(\mathbf{Y} - \mathbf{XB})] + \text{Tr}[\mathbf{B}^{\top}\mathbf{B} + \boldsymbol{\lambda}\mathbf{I}] \tag{3.20}$$

$$\hat{\mathbf{B}}^{\text{Ridge}} = \underset{\mathbf{B}}{\arg\min}\ \text{RSS}^{\text{Ridge}}(\mathbf{B}) \tag{3.21}$$

The coefficients that minimize the RSS is shown in Equation 3.22 below.

$$\hat{\mathbf{B}}^{\text{Ridge}} = (\mathbf{X}^{\top}\mathbf{X} + \boldsymbol{\lambda}\mathbf{I})^{-\mathbf{1}}\mathbf{X}^{\top}\mathbf{Y} \tag{3.22}$$

The choice of hyper-parameters $\boldsymbol{\lambda} \geq 0$ controls how much shrinkage is applied to the coefficients: larger $\boldsymbol{\lambda}$ implies more penalization to complex models. Although the coefficients are shrunk towards zero, they never reach zero, which makes Ridge regularization unsuitable for feature selection (Friedman, Hastie, Tibshirani, et al. 2001).

3.7 Cross Validation

There are several choices to make for the aforementioned models: How many PCs/PLS components to use? How much penalization to impose in Ridge regression?

A first answer to this would be to split the data into training and validation sets. After fitting models to the former, the latter is used to measure the prediction error via some scoring function. In that sense, it is important to distinguish test error rate from training error rate. The first, also called generalization error, is the score of the fit on an independent, previously unseen test sample. The second, on the other hand, is the average score over the training sample (Friedman, Hastie, Tibshirani, et al. 2001).

Scoring functions measure how much the data deviates from the fit and can be used as a qualitative tool for model selection and comparison. Once this is done, the choice of the model that yields minimum error is trivial. Two examples of wildly used score functions are Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). For the multi-output case of m responses and n observations, they are defined respectively as :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m}(y_{ij} - \hat{y}_{ij})\right)^2 \tag{3.23}$$

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{3.24}$$

This approach, however, is sensitive to the choice of these sets. Additionally, reserving part of the data just for validation might be detrimental for the model fitting process, specially if the number of observations is low (James, Witten, Hastie, and Tibshirani 2013).

One tool that can help alleviating these problems is Cross Validation (CV). More specifically, K-fold CV: it involves equally dividing the training data into K sets. For each subset, the desired model is trained using $k-1$ folds and the prediction error is computed on the remaining fold (Friedman, Hastie, Tibshirani, et al. 2001). As for the final evaluation, it is performed in the held-out test set.

4 Methods

This chapter describes the application of theory shown in Chapter 3 applied to the pre-processed data described in Chapter 2. For further details regarding coding and implementation, the reader is referred to this work's repository.

In general, the use of Scikit-Learn's `pipeline` class alongside linear models made analysis straightforward. Additionally, Scikit-Learn's `GridSearchCV` allowed for a faster evaluation of different hyperparameters such as number of components and shrinkage factor.

4.1 Ordinary Least Squares

4.2 Principal Components Regression

4.3 Partial Least Squares Regression

4.4 Ridge Regression

5 Results

6 Discussion

6.1 Results

6.2 Method

6.3 The work in a wider context

7 Conclusion

Bibliography

Alberto Bernabeo, R., K. Webster, and M. Onofri (n.d.). “Health and Environmental Impacts of Nox: An Ultra- Low Level of Nox (Oxides of Nitrogen) Achievable with A New Technology.” In: *Global Journal of Engineering Sciences* 3 (), pp. 2–7. DOI: 10.33552/gjes.2019.02.000540.

ASTDR (2004). “Sheet for ammonia published by the Agency for Toxic Substance and Disease Registry (ASTDR).” In: 2672, pp. 1–18. URL: https://www.atsdr.cdc.gov/MHMI/mmg126.pdf%5C%0Ahttps://www.atsdr.cdc.gov/mmg/mmg.asp?id=7&tid=2#bookmark02.

Bastuck, Manuel (Jan. 2019). “Improving the performance of gas sensor systems with advanced data evaluation, operation, and calibration methods.” PhD thesis, p. 267.

Bishop, Christopher M (2006). *Pattern recognition and machine learning*. springer.

Boningari, Thirupathi and Panagiotis G. Smirniotis (2016). “Impact of nitrogen oxides on the environment and human health: Mn-based materials for the NOx abatement.” In: *Current Opinion in Chemical Engineering* 13.x, pp. 133–141. ISSN: 22113398. DOI: 10.1016/j.coche.2016.09.004. URL: http://dx.doi.org/10.1016/j.coche.2016.09.004.

Bur, Christian, Manuel Bastuck, Anita Lloyd Spetz, Mike Andersson, and Andreas Schütze (2014). “Selectivity enhancement of SiC-FET gas sensors by combining temperature and gate bias cycled operation using multivariate statistics.” In: *Sensors and Actuators B: Chemical* 193, pp. 931–940. ISSN: 0925-4005. DOI: https://doi.org/10.1016/j.snb.2013.12.030. URL: https://www.sciencedirect.com/science/article/pii/S0925400513015037.

Dunn, Kevin (2021). *Process Improvement Using Data*. McMaster University. ISBN: 9781292037578. URL: https://learnche.org/pid/.

Forzatti, Pio (2001). “Present status and perspectives in de-NOx SCR catalysis.” In: *Applied Catalysis A: General* 222.1. Celebration Issue, pp. 221–236. ISSN: 0926-860X. DOI: https://doi.org/10.1016/S0926-860X(01)00832-8. URL: https://www.sciencedirect.com/science/article/pii/S0926860X01008328.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani, et al. (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.

Guthrie, Susan, Sarah Giles, Fay Dunkerley, Hadeel Tabaqchali, Amelia Harshfield, Becky Ioppolo, and Catriona Manville (2018). *Impact of ammonia emissions from agriculture on biodiversity: An evidence synthesis*. Santa Monica, CA: RAND Corporation. DOI: 10.7249/RR2695.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Johnson, R.A. and D.W. Wichern (2013). *Applied Multivariate Statistical Analysis: Pearson New International Edition*. Pearson Education Limited. ISBN: 9781292037578. URL: https://books.google.se/books?id=xCipBwAAQBAJ.

Ng, Kee Siong (2013). "A simple explanation of partial least squares." In: *The Australian National University, Canberra*.

USEPA (2019). *Nitrogen Oxides Control Regulations*. https://www3.epa.gov/region1/airquality/nox.html. Accessed 2021-02-09.

Wold, Svante, Michael Sjöström, and Lennart Eriksson (2001). "PLS-regression: a basic tool of chemometrics." In: *Chemometrics and Intelligent Laboratory Systems* 58.2. PLS Methods, pp. 109–130. ISSN: 0169-7439. DOI: https://doi.org/10.1016/S0169-7439(01)00155-1. URL: https://www.sciencedirect.com/science/article/pii/S0169743901001551.

Wright, Kevin (2017). *The NIPALS algorithm*. https://cran.r-project.org/web/packages/nipals/vignettes/nipals_algorithm.html. Accessed: 2021-03-12.

A Data acquisition time stamps

Table A.1: Data acquisition timestamps.

Frequency (Hz)	Duration (s)	Feature	Start time (s)	End time (s)
0.05	20	Slope	0,0	0,4
		Average	9,6	10,0
		Slope	10,0	10,4
		Average	19,6	20,0
0.1	10	Slope	20,0	20,4
		Average	24,6	25,0
		Slope	25,0	25,4
		Average	29,6	30,0
0.25	4	Slope	30,0	30,4
		Average	31,6	32,0
		Slope	32,0	32,4
		Average	33,6	34,0
0.5	2	Slope	34,0	34,4
		Average	34,6	35,0
		Slope	35,0	35,4
		Average	35,6	36,0
1.0	2	Slope	36,0	36,4
		Average	36,6	37,0
		Slope	37,0	37,4
		Average	37,6	38,0
2.0	2	Slope	38,0	38,4
		Average	38,6	39,0
		Slope	39,0	39,4
		Average	39,6	40,0
5.0	2	Slope	40,0	40,4
		Average	40,6	41,0
		Slope	41,0	41,4
		Average	41,6	42,0
10.0	2	Slope	42,0	42,4
		Average	42,6	43,0
		Slope	43,0	43,4
		Average	43,6	44,0
25.0	2	Slope	44,0	44,4
		Average	44,6	45,0
		Slope	45,0	45,4
		Average	45,6	46,0
50.0	2	Slope	46,0	46,4
		Average	46,6	47,0
		Slope	47,0	47,4
		Average	47,6	48,0
100.0	2	Slope	48,0	48,4
		Average	48,6	49,0
		Slope	49,0	49,4
		Average	49,6	50,0
200.0	2	Slope	50,0	50,4
		Average	50,6	51,0
		Slope	51,0	51,4
		Average	51,6	52,0
500.0	2	Slope	52,0	52,4
		Average	52,6	53,0
		Slope	53,0	53,4
		Average	53,6	54,0
1000.0	2	Slope	54,0	54,4
		Average	54,6	55,0
		Slope	55,0	55,4
		Average	55,6	56,0
2500.0	2	Slope	56,0	56,4
		Average	56,6	57,0
		Slope	57,0	57,4
		Average	57,6	58,0
5000.0	2	Slope	58,0	58,4
		Average	58,6	59,0
		Slope	59,0	59,4
		Average	59,6	60,0