Danmarks Tekniske Universitet

DTU

Project 1 - South African Heart Disease

02450 Introduction Machine Learning and Data Mining

GROUP 240

Anders Krenk - s204749
Bence Gattyan - s204773

Section	1	2	3	4	4.1	4.2	5	6.1	6.2	6.3	6.4	6.5
Anders Krenk	0	10	25	100	100	70	60	0	0	0	0	0
Bence Gattyan	100	90	75	0	0	30	40	0	0	0	0	0

Table 1: Individual contributions for each section

October 4, 2022

1 Introduction

Our objective on this report is to process data with feature extraction and visualization. This will give us better insight on our data set which is a necessary step before applying a machine learning algorithm to the problem in later projects. Contributions are indicated for each section in Table 1

2 Description of our data set

The data set we chose is a study on heart-disease in South Africa found on Elements of Statistical Learning. [1] The 1983 article that the book used can be found on journals.co.za [2] The data set we use is taken from a larger data set in the original article.

The data was recorded as a coronary risk factor screening in three rular communities of Afrikaans-speaking whites in south-western Cape. The study considered reversible risk factors such as inactivity, obesity and irreversible risk factors such as chest pain, family history etc. They found that these risk factors appeared singly or in combination in the great majority of the study population.

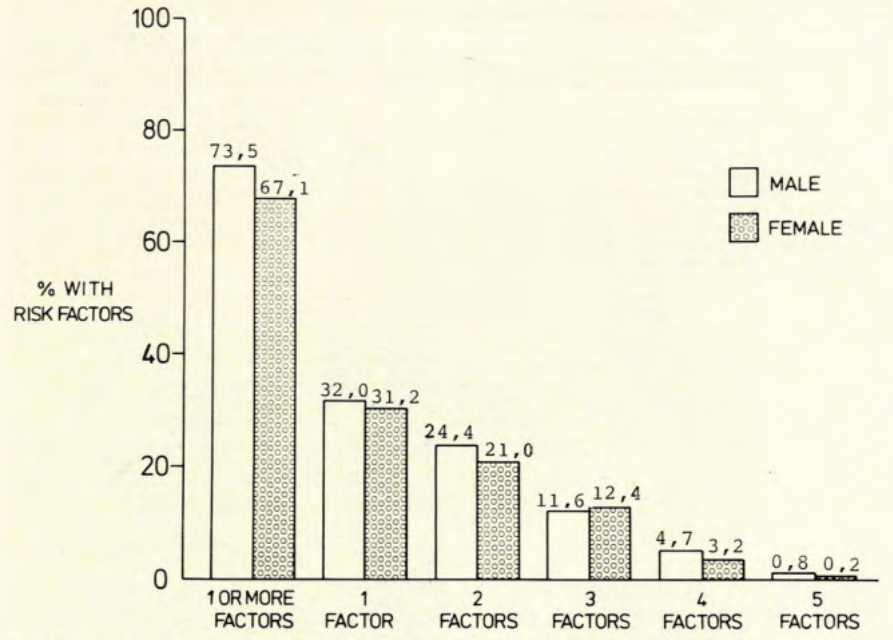

Figure 1: Combination of the three major reversible risk factors, II family history of IHD and a history of chest pain and/or an ECG suggestive of IHD in males and females aged 15-64 years. (Note that the major risk factors considered are hypercholesterolaemia (>= 6.5 mmol/l), hypertension (>= 160/95 mmHg) and cigarette smoking (>= 10/d). Subjects with at least one first-degree relative with fatal or non-fatal IDH were considered to have a positive family history) (Figure caption from the original article)

From this data set we want to learn how the recorded attributes affect the chance of having coronary heart disease and predict it for new observations. It would also help to reduce the number of attributes needed to predict response and sort out the ones that highly correlate so that for future measurements less questions would be needed to record from patients.

For this we will need to reformat the data, mainly standardizing it and converting famhist from Present/Absent to 1/0.

```
sbp             systolic blood pressure
tobacco         cumulative tobacco (kg)
ldl             low densiity lipoprotein cholesterol
adiposity
famhist         family history of heart disease (Present, Absent)
typea           type-A behavior
obesity
alcohol         current alcohol consumption
age             age at onset
chd             response, coronary heart disease

row.names,sbp,tobacco,ldl,adiposity,famhist,typea,obesity,alcohol,age,chd
1,160,12.00, 5.73,23.11,Present,49,25.30, 97.20,52,1
2,144, 0.01, 4.41,28.61,Absent,55,28.87,  2.06,63,1
3,118, 0.08, 3.48,32.28,Present,52,29.14,  3.81,46,0
4,170, 7.50, 6.41,38.03,Present,51,31.99, 24.26,58,1
5,134,13.60, 3.50,27.78,Present,60,25.99, 57.34,49,1
6,132, 6.20, 6.47,36.21,Present,62,30.77, 14.14,45,0
7,142, 4.05, 3.38,16.20,Absent,59,20.81,  2.62,38,0
```

Figure 2: original data

3 The attributes of the data

In Figure 2 you can see the original data that we work with, there were some holes to be filled in about the descriptions of the attributes but after some research and reading the original article we have the following table:

Abbreviation	Attribute	Type	Continuous/Discrete
sbp	systolic blood pressure	Ratio	Continuous
tobacco	cumulative tobacco (kg)	Ratio	Continuous
ldl	low density lipoprotein cholesterol	Ratio	Continuous
adiposity	(BMI related)	Interval	Continuous
famhist	family history of heart disease	Binary	Discrete
typea	type-A behavior (Bortner)	Interval	Continuous
obesity	(BMI related)	Interval	Continuous
alcohol	current alcohol consumption	Ratio	Continuous
age	age at onset	Ratio	Discrete
chd	response, coronary heart disease	Nominal	Discrete

Table 2: Names and corresponding operations/conditions in the FSMD diagram

As previously stated we had to convert famhis but aside from row 262 there was no missing values. Below you can see that the attributes have different scales and means, but even after standardizing they still had different shapes hinting at different distributions.

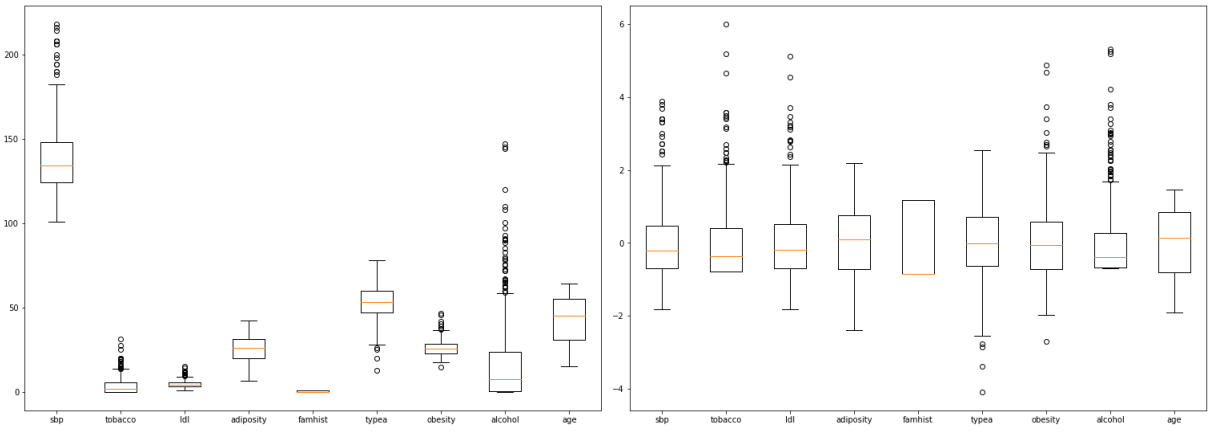

Figure 3: Boxplot for all attributes (original on left, standardized on right)

Now we investigate the individual distributions for each attributes by making histograms.

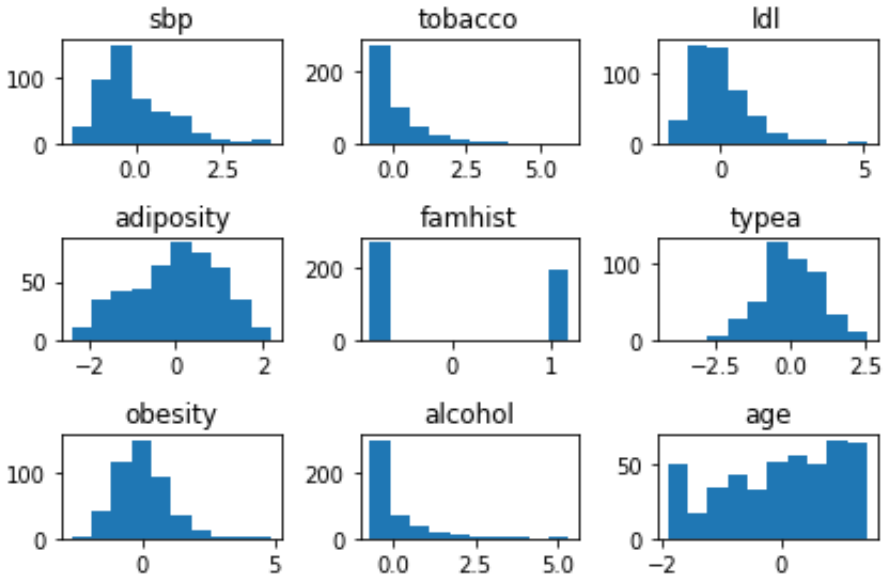

Figure 4: Histogram for all attributes

Tobacco and alcohol usage seems to be exponentially distributed while obesity, sbp, ldl and typea seem normally distributed with skewing to one side. Famhist is split because its binary type, adiposity and age don't seem to have a very obvious distribution with age maybe having a very noisy universal distribution.

And as for the correlation of the attributes, plotting some of attributes against each other reveal correlation between them:

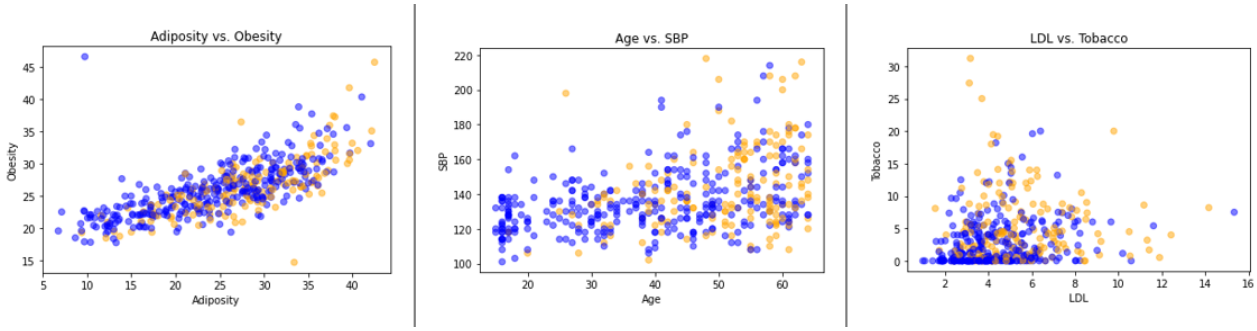

Figure 5: Plotting attributes against each other

The three graphs all show a clustering of negative chd responses toward the right side of the graph, where the more extreme cases of the given attributes reside. The "Adiposity vs. Obesity" graph shows a positive linear relation, and as for the other two the "point clouds" have apparent relation as well.

4 Data visualization and PCA

In this section we investigate how data can be filtered and visualized using principal component analysis. The data in the African Heart Disease data set contains 9 attributes that we will utilize to achieve our primary machine learning aim, classifying if a patient has a negative or positive coronary heart disease (chd) response (a discrete class label).

4.1 Data Format and Feature Transformation

We present the dataset in the standard X,y data format, where the attributes sbp, tobacco, ldl, adiposity, famhist, typea, obersity, alcohol and age (refer to attribute table in section 3 for attribute descriptions) make up the X matrix, and the y vector consists of chd responses for each patient. Before any computation can be done, however, the "famhist" column (which represents a patient's presence of heart disease in the family) requires a feature transformation. Specifically a binarization, as the famhist attribute is binary and can therefore be represented with a "0" and "1", rather than "Absent" and "Present", respectively. This is easily done with a string replacement.

4.2 Principal Component Analysis

With the new cleaned dataset, we can start performing computations. As the data for every attribute varies significantly in value, a standardization of the dataset is relevant for the PCA to provide inferrable results. We can either do this by subtracting every column in the X matrix by their corresponding column means, or by doing the aforementioned AND dividing by the standard deviation. In Python we produce results for the variance explained, standardizing the dataset with both standardization methods:

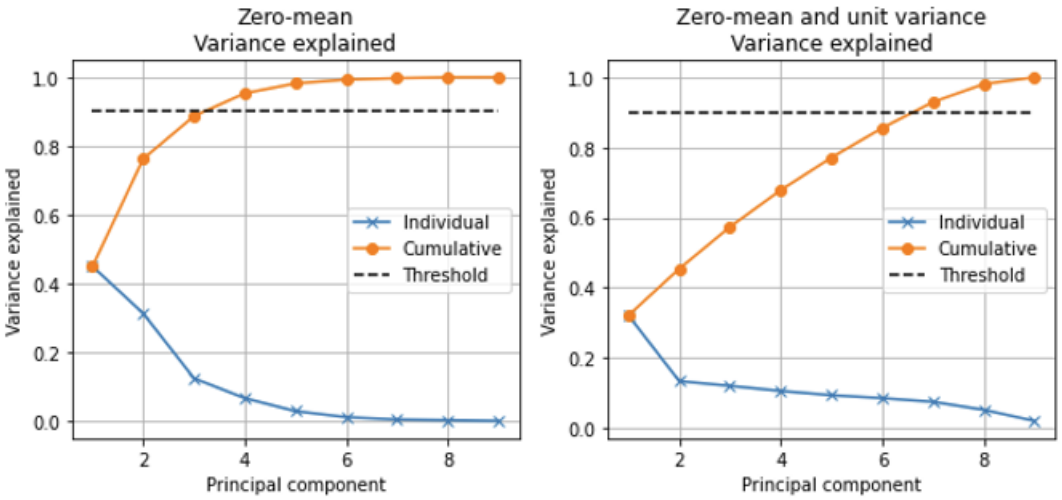

Figure 6: Variance explained with both standardization methods

As seen on the figure, the zero-mean standardization reveals that only 3 principal components will explain 90 percent of the variance, whereas the other standardization method requires 7 principal components to surpass the 90 percent threshold.

Now we plot the principal directions for two PCA components for both methods.

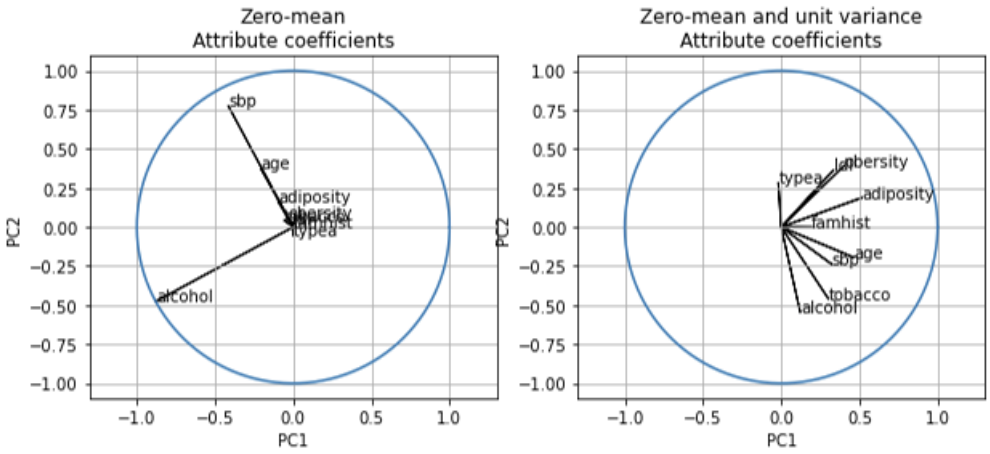

Figure 7: 2D PCA directions

We can see how the attributes get translated to the 2D PCA where the zero mean method has a lot of them overlapping in the same direction while the unit variance one is more spread out. We can see in the next figure how this affects the spread.

We plot the two-dimensional PCA graphs for the both methods:

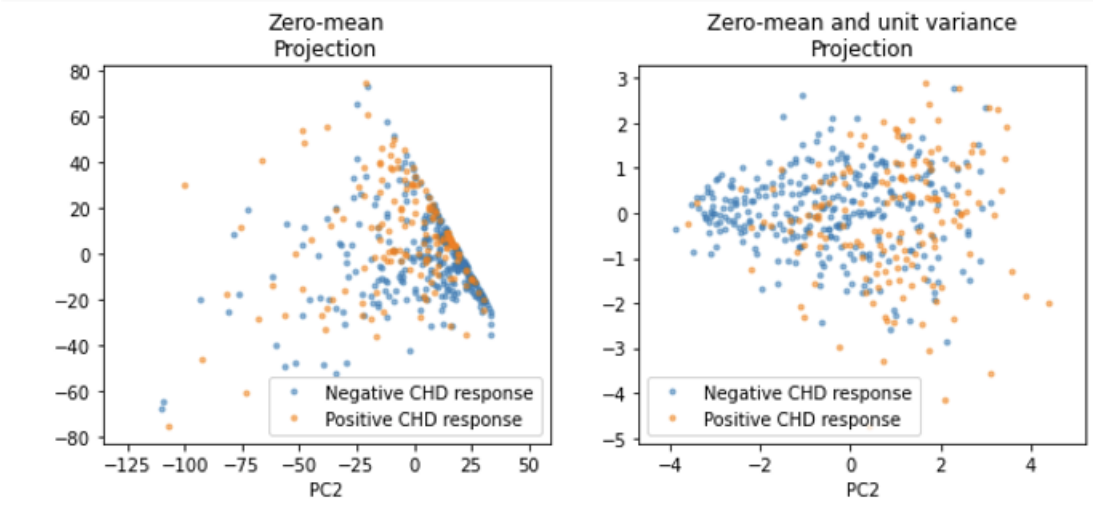

Figure 8: 2D PCA

Initially, the zero-mean projection of the first two principal seems messy. The second standardization projection seems better at representing the patients into two clusters, which is odd as the standardization method used only explains roughly 45 percent (refer to figure

4) of the variance compared to nearly 80 percent variance explained with the zero-mean method. A third principal component is added to visualize the projection more thoroughly and explain more variance:

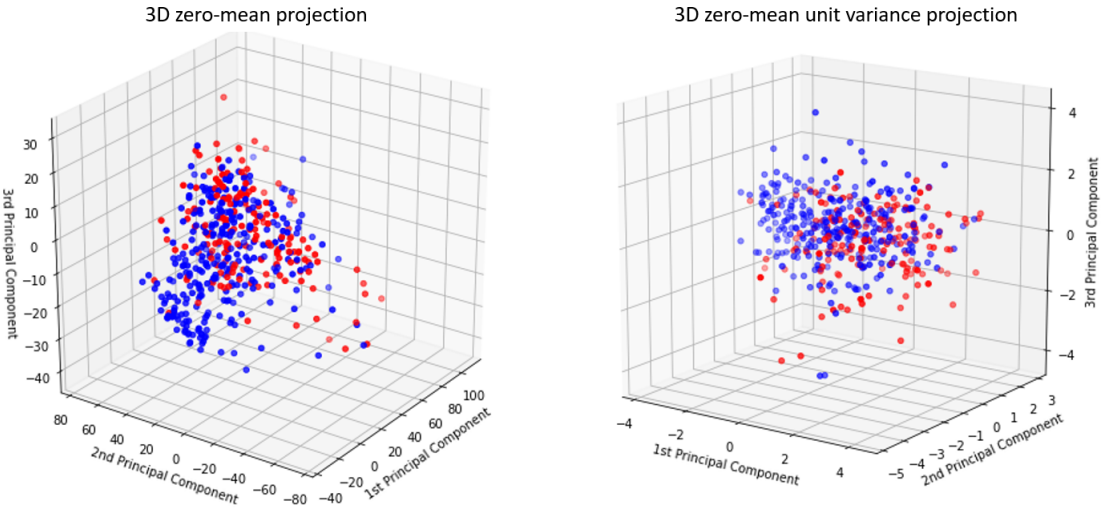

Figure 9: 3D PCA

The second standardization seems the better fit for our classification. In the following section, we will discuss what we have learned about the data.

5 What we learned

The primary machine learning model aim is to classify whether a patient has coronary heart disease or not. The zero-mean unit variance standardization produced better results for the principal directions, as each attribute contributes significantly more than with the zero-mean standardization, where mainly sbp and alcohol contribute. Although each PCA component in the zero-mean unit variance standardization contributes somewhat equal amounts to the variance explained, as opposed to the zero-mean standardization (which almost reaches the 90 percent threshold), the former produces a more inferrable 2D and 3D PCA graph. Looking at the 3D PCA results, two clusters seem to form, indicating that a successful classification of a patient having chd is possible, given that we can separate them better with methods we have yet to learn.

6 Our attempts at the exam questions

6.1 Question 1

Question 1. Spring 2019 question 1: The main dataset used in this exam is the Urban Traffic dataset[3] described in table 1. We will consider the type of an attribute as the highest level it obtains in the type-hierarchy (nominal, ordinal, interval and ratio). Which of the following statements are true about the types of the attributes in the Urban Traffic dataset?

No.	Attribute description	Abbrev.
x_1	30-minute interval (coded)	Time of day
x_2	Number of broken trucks	Broken Truck
x_3	Number of accident victims	Accident victim
x_4	Number of immobile busses	Immobilized bus
x_5	Number of trolleybus network defects	Defects
x_6	Number of broken traffic lights	Traffic lights
x_7	Number of run over accidents	Running over
y	Level of congestion/slowdown (low to high)	Congestion level

Classifying each attribute's type in the type-hierarchy will enable us to provide an answer to the problem:

Attribute	x1	x2	x3	x4	x5	x6	x7	y
Type	Interval	Ratio	Ratio	Ratio	Ratio	Ratio	Ratio	Ordinal

From our classification of the attributes, D is the right answer.

6.2 Question 2

Question 2. Spring 2019 question 2: Consider again the Urban Traffic dataset from table 1 and in particular the 14 and 18'th observation

$$x_{14} = \begin{bmatrix} 26 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad x_{18} = \begin{bmatrix} 19 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Which of the following statements about the p-norm distance $d_p(\cdot,\cdot)$ is correct?

A $d_{p=\infty}(x_{14}, x_{18}) = 7.0$

B $d_{p=3}(x_{14}, x_{18}) = 3.688$

C $d_{p=1}(x_{14}, x_{18}) = 1.286$

D $d_{p=4}(x_{14}, x_{18}) = 4.311$

E Don't know.

```
In [183]: 26-19
Out[183]: 7

In [184]: 2-0
Out[184]: 2

In [185]: 7+2
Out[185]: 9

In [186]: math.sqrt(7**2+2**2)
Out[186]: 7.280109889280518

In [187]: (7**2+2**2)**(1/2)
Out[187]: 7.280109889280518

In [188]: (7**3+2**3)**(1/3)
Out[188]: 7.054004063162272

In [189]: (7**4+2**4)**(1/4)
Out[189]: 7.0116277797172

In [190]: (7**7+2**7)**(1/7)
Out[190]: 7.00015541565479
```

The correct answer is A

6.3 Question 3

Question 3. Spring 2019 question 3: A Principal Component Analysis (PCA) is carried out on the Urban Traffic dataset in table 1 based on the attributes x_1, x_2, x_3, x_4, x_5.

The data is standardized by (i) substracting the mean and (ii) dividing each column by its standard deviation to obtain the standardized data matrix $\tilde{\boldsymbol{X}}$. A singular value decomposition is then carried out on the standardized data matrix to obtain the decomposition $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T = \tilde{\boldsymbol{X}}$

$$\boldsymbol{V} = \begin{bmatrix} 0.49 & -0.5 & 0.08 & -0.49 & 0.52 \\ 0.58 & 0.23 & -0.01 & 0.71 & 0.33 \\ 0.56 & 0.23 & 0.43 & -0.25 & -0.62 \\ 0.31 & 0.09 & -0.9 & -0.19 & -0.24 \\ -0.06 & 0.8 & 0.03 & -0.41 & 0.43 \end{bmatrix}$$

$$\boldsymbol{S} = \begin{bmatrix} 13.9 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 12.47 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 11.48 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 10.03 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 9.45 \end{bmatrix}$$

The sum of the first two singular values is less than 50 to the correct answer is C.

6.4 Question 4

Question 4. Spring 2019 question 4: Consider again the PCA analysis for the Urban Traffic dataset, in particular the SVD decomposition of $\tilde{\boldsymbol{X}}$ in eq. (2). Which one of the following statements is true?

$$\boldsymbol{V} = \begin{bmatrix} 0.49 & -0.5 & 0.08 & -0.49 & 0.52 \\ 0.58 & 0.23 & -0.01 & 0.71 & 0.33 \\ 0.56 & 0.23 & 0.43 & -0.25 & -0.62 \\ 0.31 & 0.09 & -0.9 & -0.19 & -0.24 \\ -0.06 & 0.8 & 0.03 & -0.41 & 0.43 \end{bmatrix}$$

Looking at the principal direction matrix V, each row corresponds to an attribute. We find the correct solution by eliminating the solution scenarios that don't make sense. If we take a look at answer C:

"An observation with a low value of Time of day, a high value of Broken Truck, a low value of Accident victim, and a low value of Defects will typically have a negative value of the projection onto principal component number 4." As we're investigating the sign of the projection onto the fourth principal component, we look at the fourth column. The values of the mentioned attributes are listed in order:

-0,49, 0.71, -0.25, -0,41

The description of the data given in the solution C corresponds to the values shown in V, and since the negative values outweigh the the positive it is reasonable to believe that the observation will typically lead to a negative value of the projection onto principl component 4. The correct answer is C.

6.5 Question 5

Question 5. Spring 2019 question 14: Suppose s_1 and s_2 are two text documents containing the text:

$$s_1 = \begin{Bmatrix} \text{the bag of words representation} \\ \text{becomes less parsimoneous} \end{Bmatrix}$$
$$s_2 = \{\text{if we do not stem the words} \quad \}$$

The documents are encoded using a bag-of-words encoding assuming a total vocabulary size of $M = 20000$. No stopwords lists or stemming is applied to the dataset. What is the Jaccard similarity between documents s_1 and s_2?

The Jaccard similarity between two sets of data is simply the number of overlapping items divided by the number of total unique items. In this case, the items that are unique are:

The, bag, of, words, representations, becomes, less, parsimonious, if, we , do, not, stem

Which amounts to 13. The items that overlap are:

the, words

Which is 2. The Jaccard is therefore 2 / 13 = 0.15384615384. The correct answer is A.

References

[1] J. F. Trevor Hastie, Robert Tibshirani, *The Elements of Statistical Learning - South African Heart Disease*. February 2009.

[2] J. Rossouw, *Coronary risk factor screening in three rural communities - the CORIS baseline study*.