CORE Skills

Delivering Data Science
In Resources & Energy

Data Analysis III: Multivariate analysis & model building

DAY 6

15-Day Data Science Springboard

Dr Jeremy Mitchell & Dr Ying Yap, Data Mettle

19 January 2022

Program partners

_dm

THE UNIVERSITY OF WESTERN AUSTRALIA

Curtin University

A CORE partnership

Program Timeline

DAY 4, 5 & 6: Data Analysis

Leader Engagement	15 Day Professional Program		
2-hour Leading Data Scientists Leader Support	Preparatory		Introduction to Data Projects
2-hours	Day 1	Day 2	Day 3
January 2022	Wed 24 Nov 2021	Thurs 25 Nov 2021	Wed 1 Dec 2021
Enabling your people's data science upskilling & project delivery in 15 day program	Introduction to the Program Tools	Introduction to the Program Tools:	Zero to Data Science in a day

Data Analysis

Day 4	Day 5	Day 6
Wed 8 Dec 2021	Wed 15 Dec 2021	Wed 19 Jan 2022
Getting to know the program tools - data munging and exploratory data analysis	Simple predictions - regression and statistical model building	Multivariate analysis and model building

Data & Communication Sandbox	Data Fusion and Machine Learning		Data Fusion Sandbox	Special Data Types - Time-series & Networks	Special Data Types - Knowledge Discovery from Natural Language Processing & Text Mining	Special Data	Capstone Project Development & Presentation	Capstone Propeller
Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15
Wed 2 Feb 2022	Wed 9 Feb 2022	Wed 16 Feb 2022	Wed 23 Feb 2022	Wed 2 Mar 2022	Wed 9 Mar 2022	Wed 16 Mar 2022	Wed 23 Mar 2022	Wed 22 June 2022
Effective data storytelling: Communicating results to non-technical audiences	Pros and cons of commonly used statistical and machine learning techniques I	Pros and cons of commonly used statistical and machine learning techniques II	Consolidate approaches covered and test on datasets	The 4th dimension and predictions	Finding needles in wordstacks	Spatial analytics and predictions	Capstone Project pitches to leadership	Project Review

© CORE Hub 2021

Q&A, Issues & Announcements

Before we Get Started

- Resources & Tasks on Github.

- We'll continue talking about **projects** this afternoon.

- Make notes about any ideas, perspectives or issues you encounter throughout the day.
- If you have aspects you'd like to go over throughout the day, feel free to post them to the general channel and we'll try to address the straightforward ones as we break.
- We'll come together to discuss before we close out this afternoon.

© CORE Hub 2021

Schedule

DAY 6

AWST	AEST	Agenda	Educator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	**Robust Regression (DAY 5 continued)**	Ying
08:15	10:15	**More Regression**	Ying
09:15	11:15	*Morning Tea*	
09:30	11:30	**Dimensionality Reduction**	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	Capstone Project Update/Share	Tamryn/All
12:15	14:15	**Clustering and High-D Regression**	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Partial Least Squares** (high-level overview)	Jeremy
13:50	15:50	Apply to your own Problem	
14:45	16:45	**Closeout** – Reflections, Takeaways	Jeremy
14:55	16:55	**Menti** Feedback	Tamryn
17:00	17:00	**Close**	

© CORE Hub 2021

Aims & Learning Outcomes

DAY 6

Aims

- Perform multivariate Exploratory Data Analysis (EDA).
- Analyse central tendency, dispersion and correlations.
- Practice unsupervised learning: principal component analysis (PCA) and clustering.

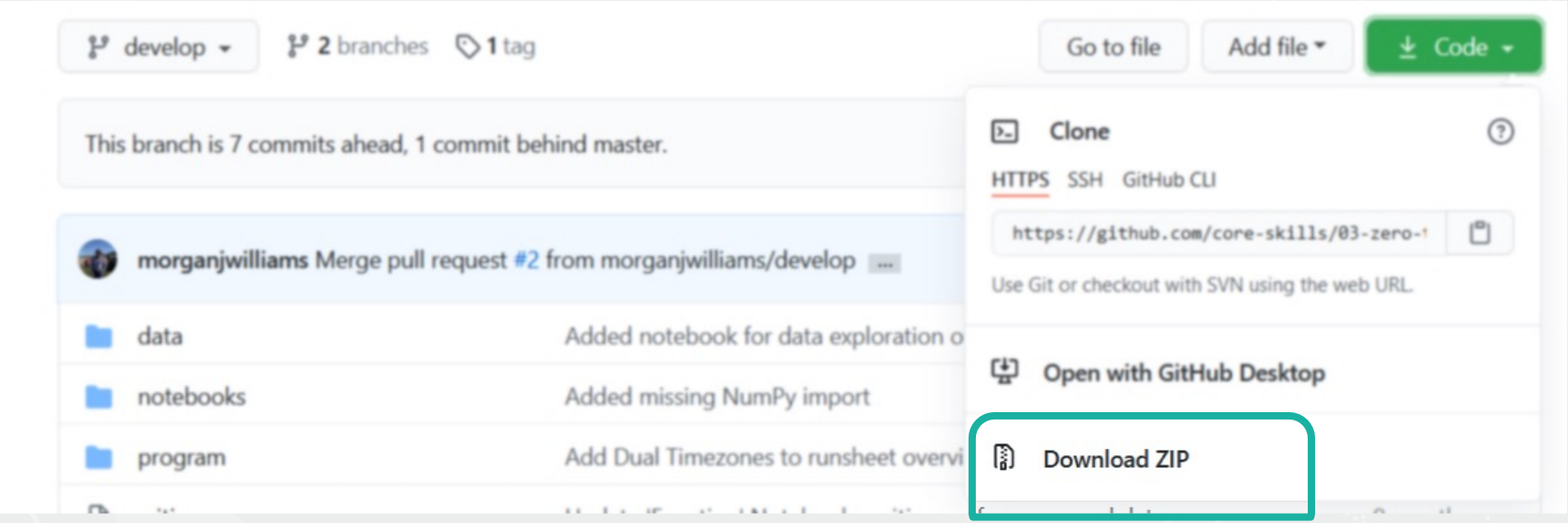
Learning Outcomes

- Understand the differences between univariate and multivariate data and apply multivariate techniques and EDA.
- Understand the basis of dimensionality reduction.
- Apply k-means clustering.
- Critique whether a correlation is spurious.
- Identify if the process used in the activities is reproducible.
- Understand the impact of missing data and recall practical solutions to work with incomplete data sets.

© CORE Hub 2021

GitHub Content for Today

[github.com / core-skills / 06-multivariate-analysis]



© CORE Hub 2021

GitHub – Program Notes

github.com / core-skills / 06-multivariate-analysis / program / 00_overview.md

Overview

Overview | Data Culture | From Here to There | Data Projects | Data Exploration | Closeout

Aim

Provide an overview of a 'typical' data science workflow.

Learning Outcomes

1. To appreciate what data science is
2. To appreciate the fields data science spans
3. Understand the stages of a data science project and define a mental model of it
4. Analyse the opportunity and potential value of data science in your organisation

Schedule

AWST	AEST	Agenda
07:30 - 07:45	09:30 - 09:45	Q&A, Issues & Announcements
07:45 - 09:15	09:45 - 11:15	**Creating a Data Culture**
09:15 - 09:30	11:15 - 11:30	*Morning Tea*
09:30 - 11:00	11:30 - 13:00	**Getting From Here to There**

© CORE Hub 2021

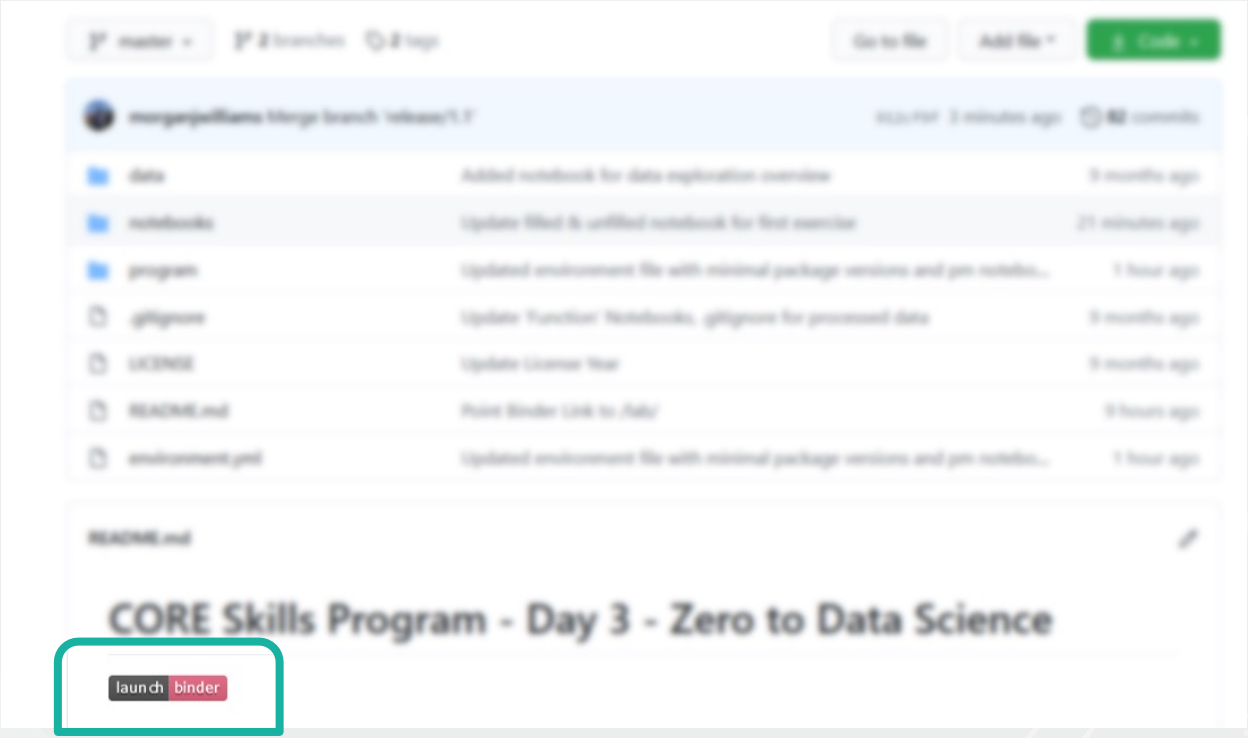
Environment

- Open an Anaconda Prompt
- Navigate to where you have the unzipped repository material

```
conda env create -f environment.yml # make new env

conda activate core06 # activate this env (Windows)

# make this available to Jupyter as a "kernel"
python -m ipykernel install --user --name=core06

jupyter lab # launch Jupyter lab
```


© CORE Hub 2021

Binder Backup



© CORE Hub 2021

More Regression (continued from Day 5)

Schedule

DAY 6

AWST	AEST	Agenda	Educator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	**Robust Regression (DAY 5 continued)**	Ying
08:15	10:15	**More Regression**	Ying
09:15	11:15	*Morning Tea*	
09:30	11:30	**Dimensionality Reduction**	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	Capstone Project Update/Share	Tamryn/All
12:15	14:15	**Clustering and High-D Regression**	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Partial Least Squares** (high-level overview)	Jeremy
13:50	15:50	Apply to your own Problem	
14:45	16:45	**Closeout** – Reflections, Takeaways	Jeremy
14:55	16:55	**Menti** Feedback	Tamryn
17:00	17:00	**Close**	

© CORE Hub 2021

More Regression

Missing data

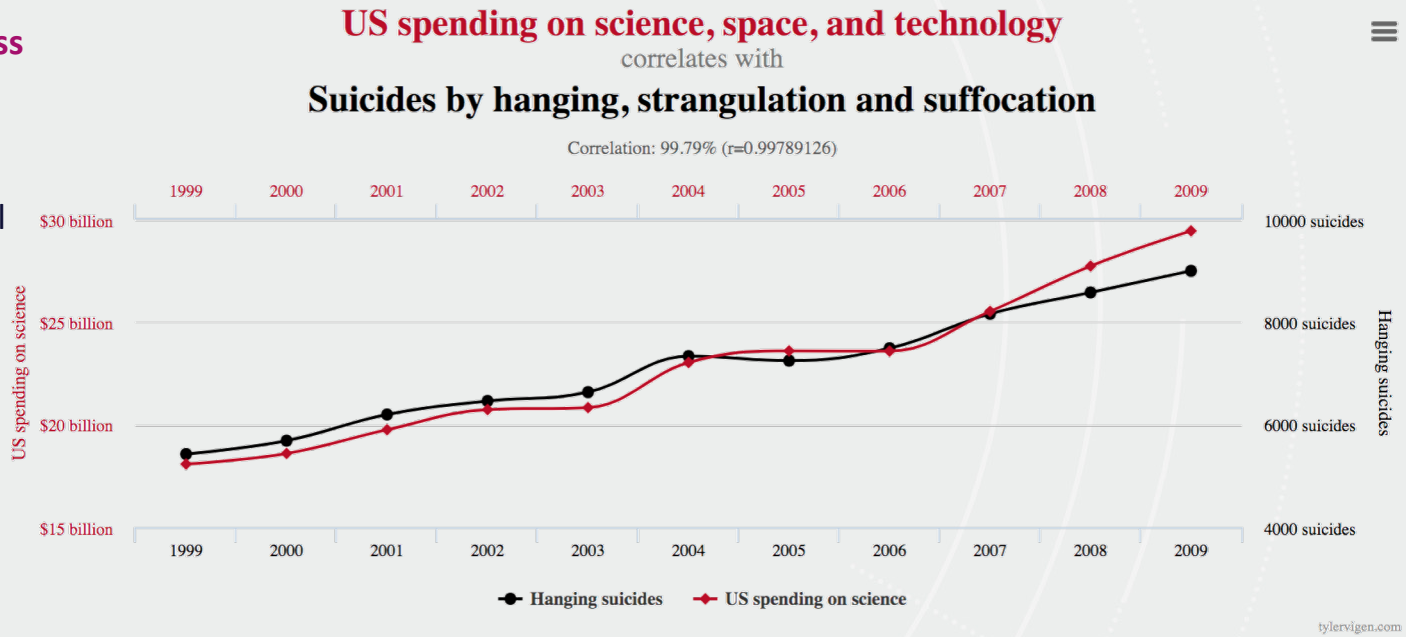
- In some cases, we have missing data that will actually bias our regression results. If possible, we would like to try and account for this or replace the missing data.
- Replacing missing data is called imputation.
- There are multiple ways to do this ranging from interpolation, substitution, and regression
- Each method has drawbacks and the technique used should be appropriate for the type of data

© CORE Hub 2021

Spurious correlation

To regress or not to regress

- We see correlation
- We can build a model to express this relationship
- But should we?



Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

Spurious correlation from tylervigen.com

© CORE Hub 2021

Refresher from last time

Download today's repository

Open am1_more_regression.ipynb and work your way through Exercise 1

- Use the next 20 minutes to complete the exercise
- Then we will move onto multivariate data

© CORE Hub 2021

Multivariate Regression

Multivariate regression

- Multivariate regression is the natural extension of linear regression to include more than one explanatory variable

- The linear regression model is

$$Y = aX + b + \varepsilon$$

- The multivariate regression model for N explanatory variables $X_1, X_2, \cdots, X_N$ is

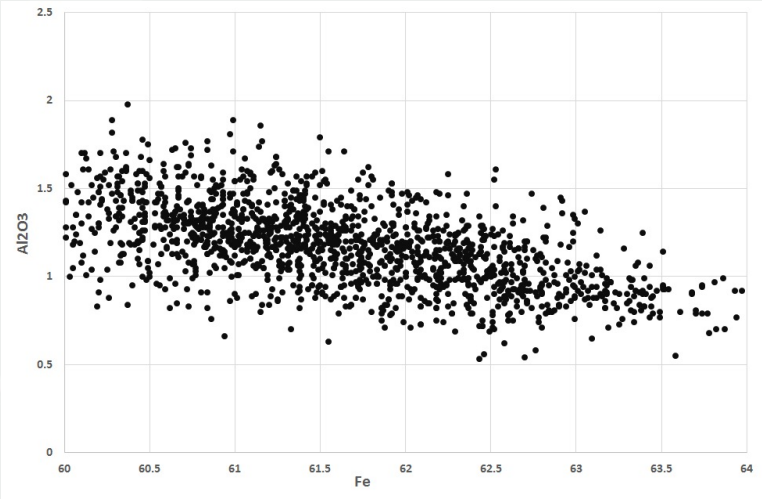
$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_N X_N + \beta_0 + \varepsilon$$

© CORE Hub 2021

Linear vs. multivariate regression

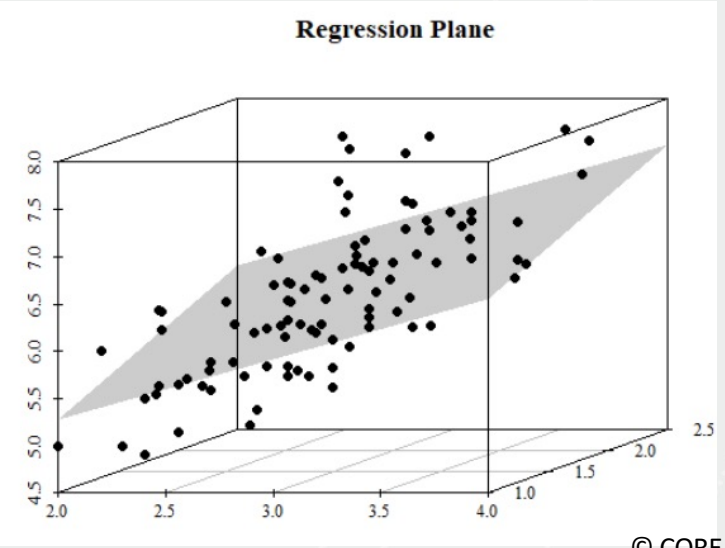
Linear Regression

- The linear regression model fits a line through the points



Multivariate Regression

- The multivariate regression model fits a plane through the points



© CORE Hub 2021

Multivariate regression

Notation

- The multivariate regression model is usually expressed using matrix notation

$$Y = X\beta + \varepsilon$$

- Y is now a vector of dimension $Mx1$ which contains the M samples of the dependent variable
- X is now a matriz of dimension $Mx(N+1)$ which contains the M samples of the N explanatory variables
- β is now a vector of dimension $(N+1)x1$ which contains the parameters of the model
- ε is now a vector of dimension $Mx1$

© CORE Hub 2021

Multivariate regression

- For instance, an Excel table with samples for two explanatory attributes and one dependent variable

(X_1)	(X_2)	(Y)
10	30	24
15	25	27
10	40	29
20	18	31
25	22	25
18	31	33
12	26	26
14	34	28
16	29	31
22	37	39
24	20	33
17	25	30
13	27	25
30	23	42
24	33	40

$$\mathbf{Y} = \begin{pmatrix} 24 \\ 27 \\ 29 \\ 31 \\ 25 \\ 33 \\ 26 \\ 28 \\ 31 \\ 39 \\ 33 \\ 30 \\ 25 \\ 42 \\ 40 \end{pmatrix}, \; X = \begin{pmatrix} 1 & 10 & 30 \\ 1 & 15 & 25 \\ 1 & 10 & 40 \\ 1 & 20 & 18 \\ 1 & 25 & 22 \\ 1 & 18 & 31 \\ 1 & 12 & 26 \\ 1 & 14 & 34 \\ 1 & 16 & 29 \\ 1 & 22 & 37 \\ 1 & 24 & 20 \\ 1 & 17 & 25 \\ 1 & 13 & 27 \\ 1 & 30 & 23 \\ 1 & 24 & 33 \end{pmatrix}, \; \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

15x1 15x(2+1) (2+1)x1

Care should be exercised when entering the data into the program to run the multivariate regression model

© CORE Hub 2021

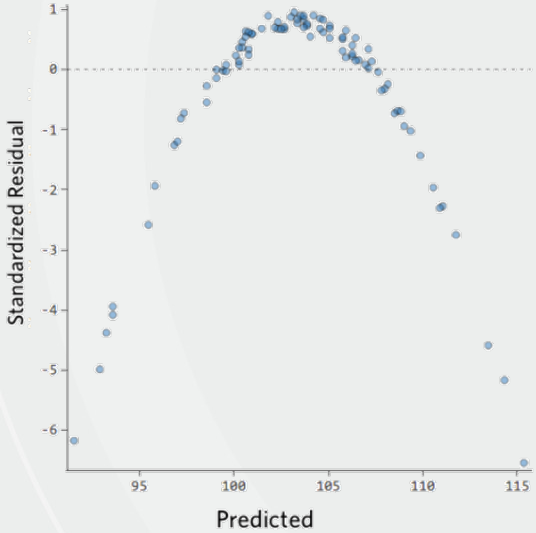
Multivariate regression

- Similar to the linear regression model, the assumptions are made on the error component
- These assumptions are similar but in a multivariate setting
- The error has multivariate Gaussian distribution with mean zero and constant variance
  - Homoscedasticity and no bias
- The error is independent of all explanatory variables and the dependent variable
  - Scatterplot of the error with all explanatory variables should not show any clear pattern

© CORE Hub 2021

Multivariate regression

- Some patterns are indicative that a transformation in one explanatory variable is needed



Worth considering a power transformation in the multivariate regression model, e.g.

$$Y = \beta_1 X_1^2 + \beta_2 X_2 + \cdots + \beta_N X_N + \beta_0 + \varepsilon$$

© CORE Hub 2021

Validating a Multivariate Regression Model

Validating a multivariate regression model

- Validating a multivariate regression model is more involved than validating a regression model
- There are a number of measures or approaches that can be used to carry out the validation
  - Residual analysis
  - Coefficient of multiple determination
  - Adjusted R^2
  - Statistical tests

© CORE Hub 2021

Validating a multivariate regression model

- The coefficient of multiple determination R^2 is the generalisation of the corresponding coefficient of determination in linear regression
- R^2 is the proportion of the variance in the dependent variable that is explained by the model
- $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum \varepsilon_i^2}{\sum (y_i - \bar{y})^2}$
- An important concept in multivariate regression is the analysis of variance (ANOVA)
- $SST = SSE + SSR$ where $SSR = \sum (\hat{y}_i - \bar{y})^2$
- $0 \leq R^2 \leq 1$ and large R^2 values do not necessarily indicate the multivariate regression is a good model

© CORE Hub 2021

Validating a multivariate regression model

- R^2 can be artificially increased by increasing the number of explanatory variables
- An alternative measure is the adjusted R^2, R_a^2
- R_a^2 takes into account the complexity – number of explanatory variables – of the multivariate regression model
- R_a^2 increases only if new explanatory variables improve the model
- A significant difference between R^2 and R_a^2 is indicative that the selection of the number of explanatory variables should be revised

© CORE Hub 2021

Validating a multivariate regression model

- Statistical tests can be used to assess the importance or impact of one explanatory variable in the model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_N X_N + \beta_0 + \varepsilon$$

- Statistical tests are used to determined if one could consider $\beta_i = 0$ or $\beta_i \neq 0$

$$H_0\colon \beta_i = 0 \quad vs \quad H_1\colon \beta_i \neq 0$$

- The test compares two models, one with the explanatory variable X_i and the other without it.

© CORE Hub 2021

Validating a multivariate regression model

- The F-test is used to test the overall significance of the multivariate regression model
- The F-test compares the multivariate regression model with a model with no explanatory variables

$$H_0\colon \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_N = 0$$

$$vs.$$

$$H_1\colon \beta_i \neq 0 \; for \; some \; i \geq 1$$

- A decision can be made by analysing the statistical results

© CORE Hub 2021

Analysing the Statistical Results

Analysing the statistical results

- Most programs provide a set of statistical results to diagnose the multivariate regression model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.933
Model:                            OLS   Adj. R-squared:                  0.928
Method:                 Least Squares   F-statistic:                     211.8
Date:                Thu, 04 Oct 2018   Prob (F-statistic):           6.30e-27
Time:                        14:15:39   Log-Likelihood:                -34.438
No. Observations:                  50   AIC:                             76.88
Df Residuals:                      46   BIC:                             84.52
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.4687      0.026     17.751      0.000       0.416       0.522
x2             0.4836      0.104      4.659      0.000       0.275       0.693
x3            -0.0174      0.002     -7.507      0.000      -0.022      -0.013
const          5.2058      0.171     30.405      0.000       4.861       5.550
```

Number of explanatory variables

coefficients

$$Y = 0.4687X_1 + 0.4836X_2 - 0.0174X_3 + 5.2058$$

© CORE Hub 2021

Analysing the statistical results

- The model is considered significant if the absolute value of the F-statistic is greater than the corresponding tabulated value at a confidence level α
- For this example, all explanatory variables seem to contribute to the response variable

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.933
Model:                            OLS   Adj. R-squared:                  0.928
Method:                 Least Squares   F-statistic:                     211.8
Date:                Thu, 04 Oct 2018   Prob (F-statistic):           6.30e-27
Time:                        14:15:39   Log-Likelihood:                -34.438
No. Observations:                  50   AIC:                             76.88
Df Residuals:                      46   BIC:                             84.52
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.4687      0.026     17.751      0.000       0.416       0.522
x2             0.4836      0.104      4.659      0.000       0.275       0.693
x3            -0.0174      0.002     -7.507      0.000      -0.022      -0.013
const          5.2058      0.171     30.405      0.000       4.861       5.550
```


© CORE Hub 2021

Analysing the statistical results

- Significance of each explanatory variable is obtained using the t and $P > |t|$ values
- The null hypothesis is rejected if the absolute value of t is greater the corresponding tabulated value at confidence level α

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.933
Model:                            OLS   Adj. R-squared:                  0.928
Method:                 Least Squares   F-statistic:                     211.8
Date:                Thu, 04 Oct 2018   Prob (F-statistic):           6.30e-27
Time:                        14:15:39   Log-Likelihood:                -34.438
No. Observations:                  50   AIC:                             76.88
Df Residuals:                      46   BIC:                             84.52
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.4687      0.026     17.751      0.000       0.416       0.522
x2             0.4836      0.104      4.659      0.000       0.275       0.693
x3            -0.0174      0.002     -7.507      0.000      -0.022      -0.013
const          5.2058      0.171     30.405      0.000       4.861       5.550
```


© CORE Hub 2021

Analysing the statistical results

- AIC and BIC stand for Akaike's and Bayesian information criteria respectively
- The smaller the AIC and the BIC values the better the model
- Useful when comparing two different models

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.933
Model:                            OLS   Adj. R-squared:                  0.928
Method:                 Least Squares   F-statistic:                     211.8
Date:                Thu, 04 Oct 2018   Prob (F-statistic):           6.30e-27
Time:                        14:15:39   Log-Likelihood:                -34.438
No. Observations:                  50   AIC:                             76.88
Df Residuals:                      46   BIC:                             84.52
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.4687      0.026     17.751      0.000       0.416       0.522
x2             0.4836      0.104      4.659      0.000       0.275       0.693
x3            -0.0174      0.002     -7.507      0.000      -0.022      -0.013
const          5.2058      0.171     30.405      0.000       4.861       5.550
```


© CORE Hub 2021

Exercise: Basis Regression using templates

Spend the rest of the time working through Exercise 2 in am1_more_regression.ipynb

© CORE Hub 2021

Schedule

DAY 6

AWST	AEST	Agenda	Educator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	**Robust Regression (DAY 5 continued)**	Ying
08:15	10:15	**More Regression**	Ying
09:15	11:15	*Morning Tea*	
09:30	11:30	**Dimensionality Reduction**	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	Capstone Project Update/Share	Tamryn/All
12:15	14:15	**Clustering and High-D Regression**	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Partial Least Squares** (high-level overview)	Jeremy
13:50	15:50	Apply to your own Problem	
14:45	16:45	**Closeout** – Reflections, Takeaways	Jeremy
14:55	16:55	**Menti** Feedback	Tamryn
17:00	17:00	**Close**	

© CORE Hub 2021

Dimensionality Reduction

Variable selection as dimension reduction

Selecting the best set of variables to build a model leads to many benefits:

- Better predictions
- Better model interpretability and explanability
- Lower computational cost of training
- Less impact from the curse of dimensionality

© CORE Hub 2021

Variable selection as dimension reduction

What could be an easy way to select the most useful variables?

© CORE Hub 2021

Dimensionality reduction

scikit-learn.org/stable/tutorial/machine_learning_map/index.html

© CORE Hub 2021

An Introduction to Principal Component Analysis

- Principal component analysis (PCA) is a mathematical technique that transforms a set of N possibly correlated variables into a set of N uncorrelated variables
- The uncorrelated variables are called principal components

$$X_1, X_2, \cdots, X_N \quad \Longleftrightarrow \quad Y_1, Y_2, \cdots, Y_N$$

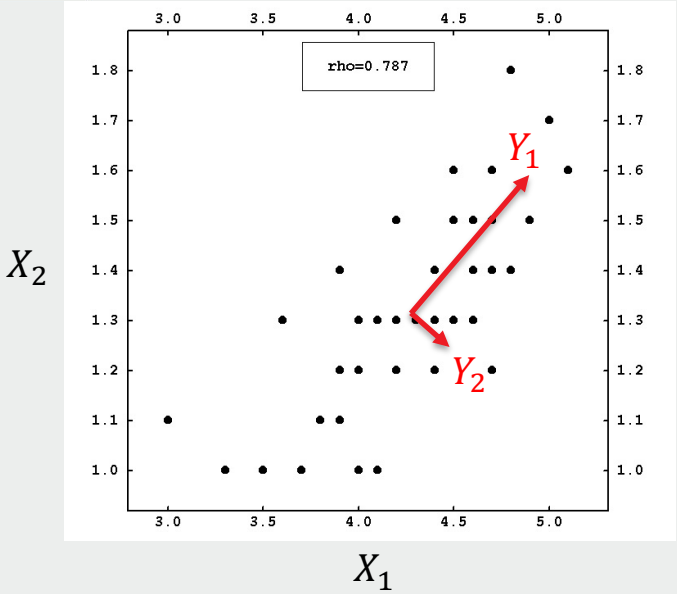
$$\downarrow \qquad\qquad\qquad \downarrow$$

Correlated variables Principal components

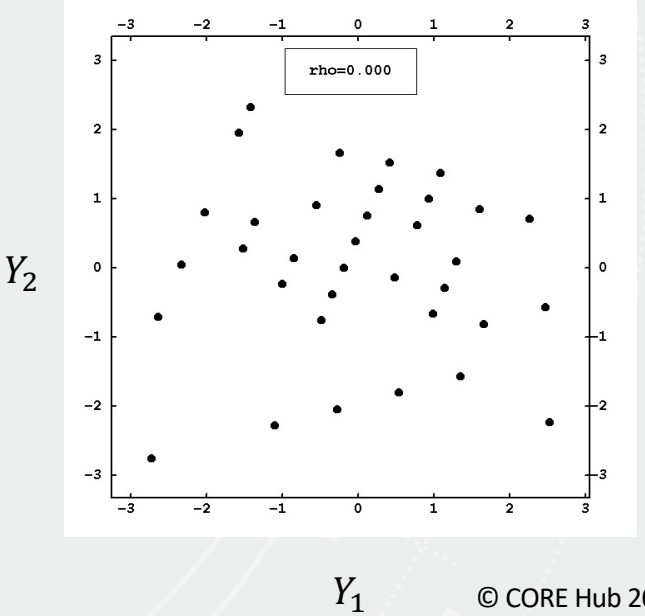
- The input variables should be <u>linearly correlated</u>

© CORE Hub 2021

An Introduction to Principal Component Analysis



Note that PCA variables have different values compared to the input variables

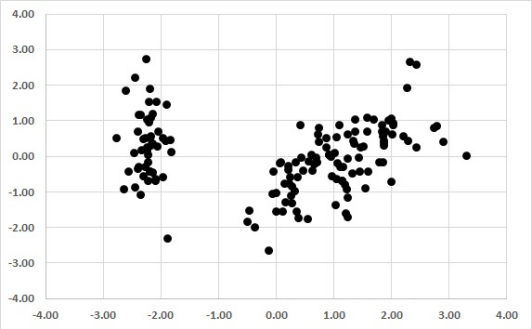


© CORE Hub 2021

An Introduction to Principal Component Analysis

- Excel table with input and PCA variables

Petal.length	Petal.width	Sepal.Length	Sepal.width	F1	F2	F3	F4
1.40	0.20	5.10	3.50	-2.26	0.51	-0.12	-0.02
1.40	0.20	4.90	3.00	-2.09	-0.66	-0.23	-0.10
1.30	0.20	4.70	3.20	-2.37	-0.32	0.05	-0.03
1.50	0.20	4.60	3.10	-2.30	-0.58	0.10	0.07
1.40	0.20	5.00	3.60	-2.39	0.67	0.02	0.04
1.70	0.40	5.40	3.90	-2.07	1.52	0.03	0.00
1.40	0.30	4.60	3.40	-2.45	0.07	0.34	0.04
1.50	0.20	5.00	3.40	-2.23	0.25	-0.08	0.03
1.40	0.20	4.40	2.90	-2.34	-1.10	0.15	0.03
1.50	0.10	4.90	3.10	-2.19	-0.45	-0.25	0.04
1.50	0.20	5.40	3.70	-2.16	1.07	-0.26	-0.02
1.60	0.20	4.80	3.40	-2.33	0.16	0.10	0.13
1.40	0.10	4.80	3.00	-2.22	-0.71	-0.22	0.00
1.10	0.10	4.30	3.00	-2.64	-0.94	0.19	0.02
1.20	0.20	5.80	4.00	-2.19	1.89	-0.47	-0.19
1.5	0.4	5.7	4.4	-2.25	2.72	0.03	-0.05
1.3	0.4	5.4	3.9	-2.2	1.51	0	-0.19
1.4	0.3	5.1	3.5	-2.19	0.51	-0.04	-0.09
1.7	0.3	5.7	3.8	-1.89	1.43	-0.37	-0.06
1.5	0.3	5.1	3.8	-2.34	1.16	0.14	0.04
1.7	0.2	5.4	3.4	-1.91	0.43	-0.42	-0.01
1.5	0.4	5.1	3.7	-2.2	0.95	0.16	-0.06
1	0.2	4.6	3.6	-2.77	0.49	0.34	-0.02
1.7	0.5	5.1	3.3	-1.82	0.11	0.04	-0.15
1.9	0.2	4.8	3.4	-2.23	0.16	0.12	0.27



© CORE Hub 2021

An Introduction to Principal Component Analysis

- Interpretation of PCA variables is usually carried out by analysing the correlation with the input variables

Variables	Y_1	Y_2
X_1	0.99	0.14
X_2	0.99	-0.14

- The first principal component accounts for most of the information in both variables X_1 and X_2
  - Instead of working with two variables, one could work with only one, the first principal component
  - Dimension reduction

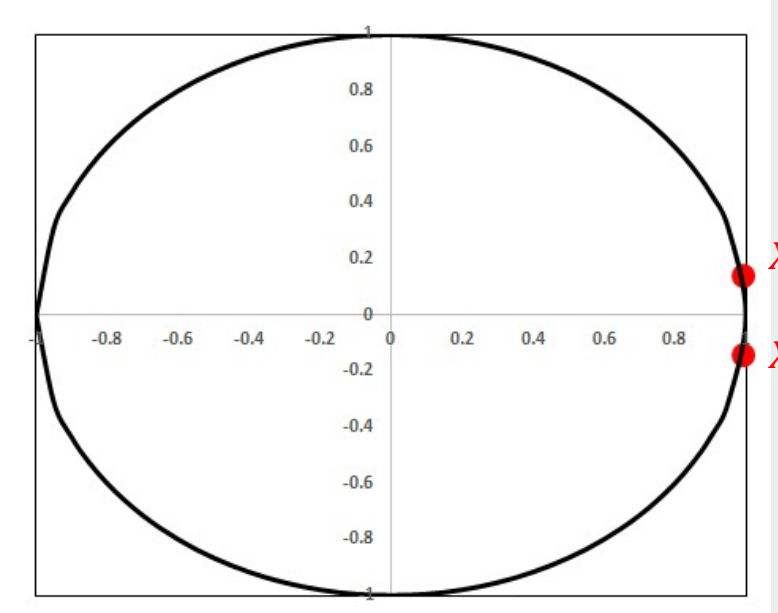
© CORE Hub 2021

An Introduction to Principal Component Analysis

- In practice, the circle of correlation is commonly used when several variable are being analysed

Variables	Y_1	Y_2
X_1	0.99	0.14
X_2	0.99	-0.14

Factor loadings

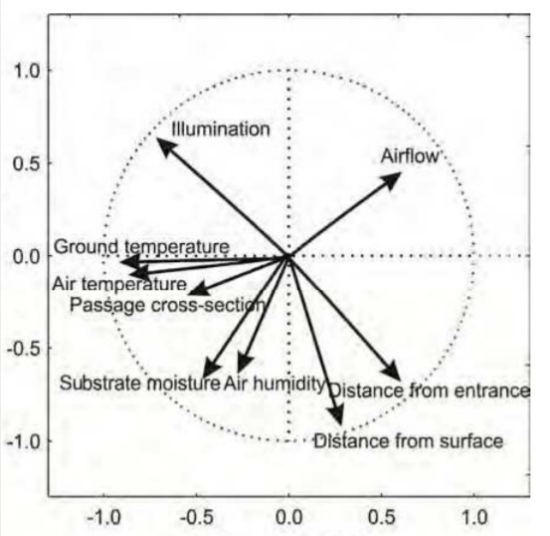


© CORE Hub 2021

An Introduction to Principal Component Analysis

- Example: Environmental data with nine input variables

Parameter \ PC	PC1	PC2
Air temperature	-0.87	-0.10
arc-sin relative air humidity	-0.29	-0.60
Ground temperature	-0.85	-0.06
arc-sin substrate moisture	-0.49	-0.72
Airflow	0.62	0.45
Distance from entrance	0.60	-0.67
Distance from surface	0.26	-0.82
log passage cross-section	-0.55	-0.20
log illumination	-0.69	0.63
$F_{2,99}$; p	19.85; <0.001	9.19; <0.001



- What can we conclude from the circle of correlation?

© CORE Hub 2021

PCA Theory – A Gentle Introduction

- PCA requires to compute the variance-covariance matrix of the N input variables
  - When the variables have different units, for instance, ppm and %, it is better to standardise the variables before computing the variance-covariance matrix
  - This is equivalent to compute the correlation matrix
- A mathematical approach called spectral decomposition is applied to the correlation matrix
- On output one has a set of vectors called eigenvectors and values called eigenvalues
- Each eigenvector is associated to an eigenvalue

© CORE Hub 2021

PCA Theory – A Gentle Introduction

- The principal components correspond to the eigenvectors
- The variance of each principal component is the corresponding eigenvalue
- Principal components are usually sorted in descending order by their corresponding eigenvalues or variance
- First principal component has the highest variance
- Total variance is the sum of the variance of each principal component

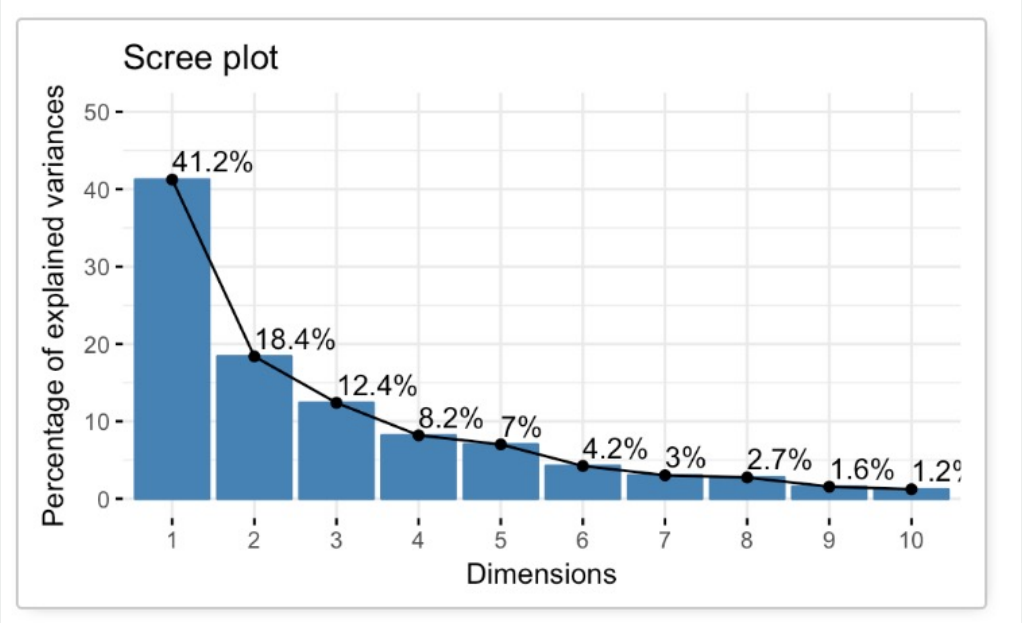
$$\sum var(X_i) = var(Y_1) + var(Y_2) + \cdots + var(Y_N)$$

- Key property for dimensionality reduction

© CORE Hub 2021

PCA Theory – A Gentle Introduction

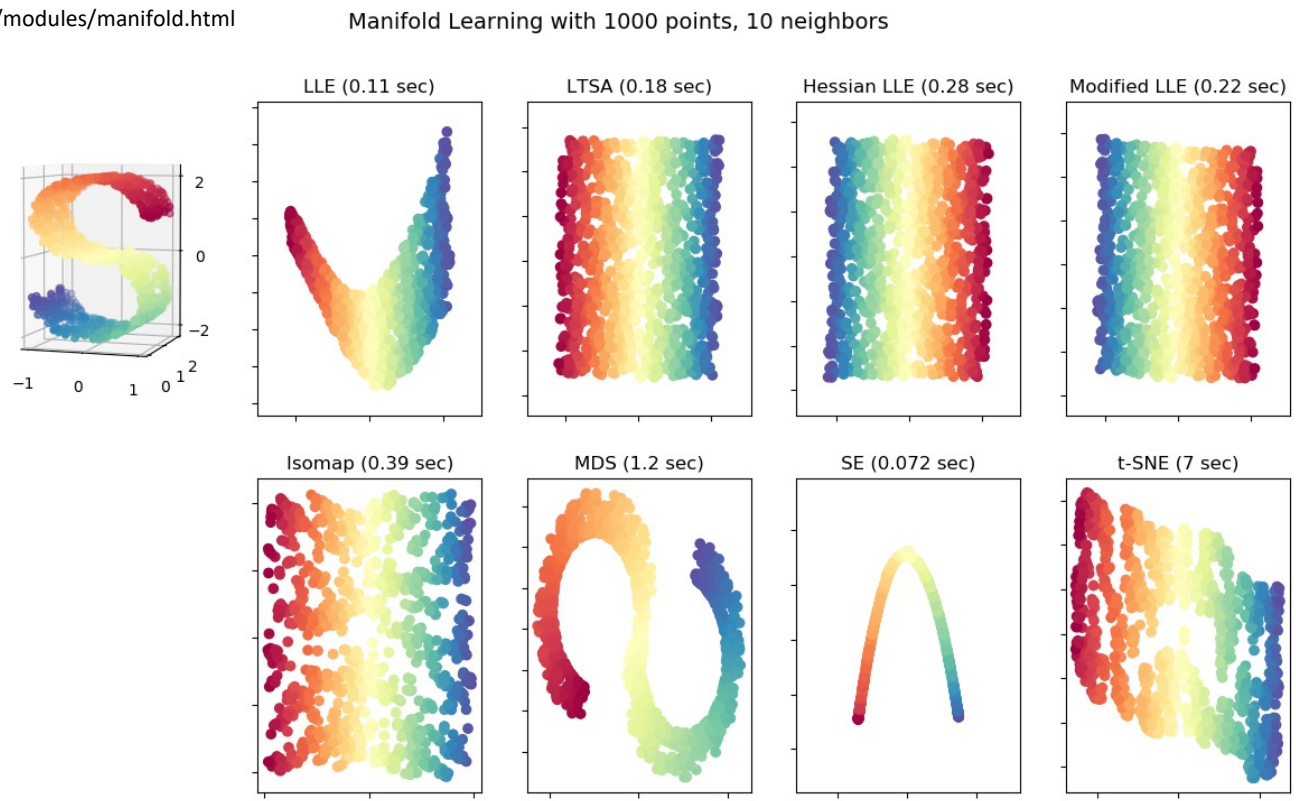
- How to choose the number of principal components to retain?
- The scree plot shows the fraction of the total variance in the data that is explained by each principal component



© CORE Hub 2021

A short overview of manifold learning

Source: scikit-learn.org/stable/modules/manifold.html

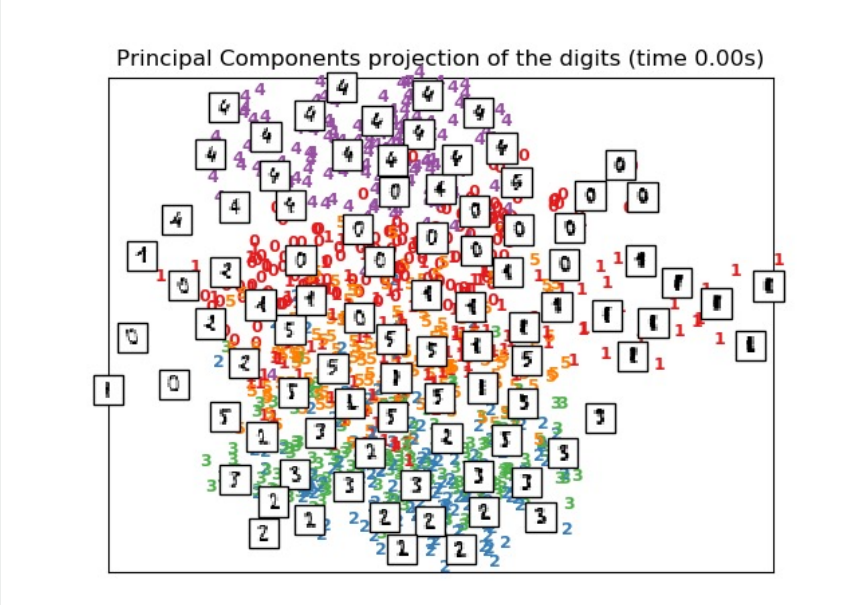


© CORE Hub 2021

About t-distributed Stochastic Neighbor Embedding (t-SNE)

Source: scikit-learn.org/stable/auto_examples/manifold/plot_lle_digits.html#
sphx-glr-auto-examples-manifold-plot-lle-digits-py

t-SNE has become incredibly popular for its impressive ability to manage thousands of dimensions

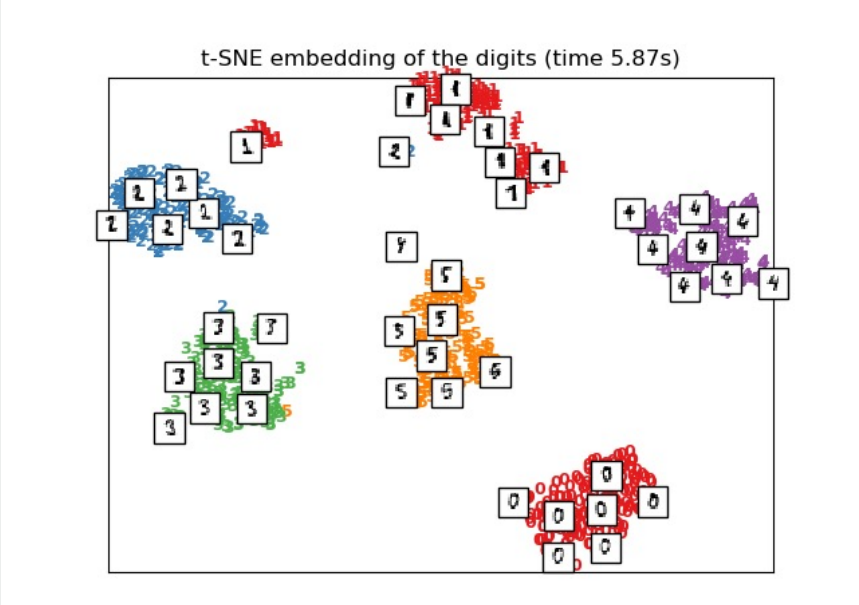


© CORE Hub 2021

About t-distributed Stochastic Neighbor Embedding (t-SNE)

Source: scikit-learn.org/stable/auto_examples/manifold/plot_lle_digits.html#
sphx-glr-auto-examples-manifold-plot-lle-digits-py

t-SNE has become incredibly popular for its impressive ability to manage thousands of dimensions



© CORE Hub 2021

Exercise: The Octane dataset

Open **am2-dimreduction.ipynb**

and go through the notebook

© CORE Hub 2021

Schedule

DAY 6

AWST	AEST	Agenda	Educator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	**Robust Regression (DAY 5 continued)**	Ying
08:15	10:15	**More Regression**	Ying
09:15	11:15	*Morning Tea*	
09:30	11:30	**Dimensionality Reduction**	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	Capstone Project Update/Share	Tamryn/All
12:15	14:15	**Clustering and High-D Regression**	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Partial Least Squares** (high-level overview)	Jeremy
13:50	15:50	Apply to your own Problem	
14:45	16:45	**Closeout** – Reflections, Takeaways	Jeremy
14:55	16:55	**Menti** Feedback	Tamryn
17:00	17:00	**Close**	

© CORE Hub 2021

Clustering

Clustering

- The aim of clustering is to group a set of observations
- Observations in the same cluster should be more similar to each other than observations in other clusters
- Observations usually correspond to high dimensional data and therefore manual definition of groups or clusters is not an easy task
- There are many applications of clustering in practice
    - Petroleum
    - Mining
    - Medicine
    - Medical Imaging
    - Banking
    - Internet

© CORE Hub 2021

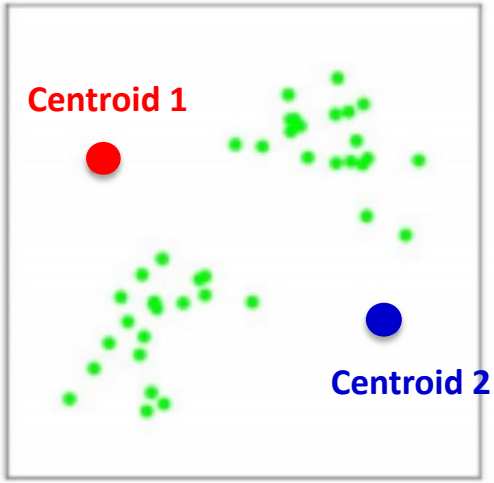
Clustering: K-means

- There are a substantial number of techniques developed to group data into clusters
- The simplest one is called K-means clustering
- K refers to the number of clusters one wants to identify
  - It should be specified in advance but there are newer approaches where this is not required
- K points called centroids are randomly initialised and iteratively move to the center of each cluster
- The process is stopped when there are no further changes in the centroid values

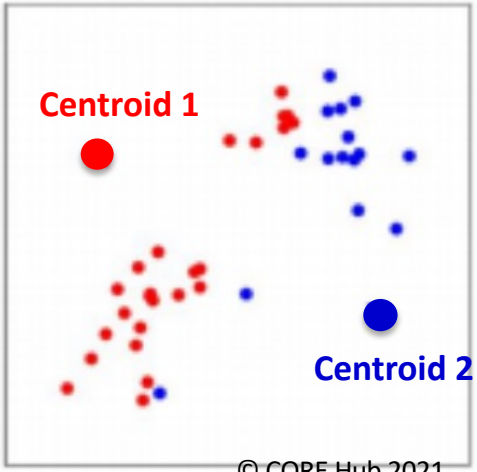
© CORE Hub 2021

Clustering: K-means

- Dataset with two clusters. Centroids are randomly initialised



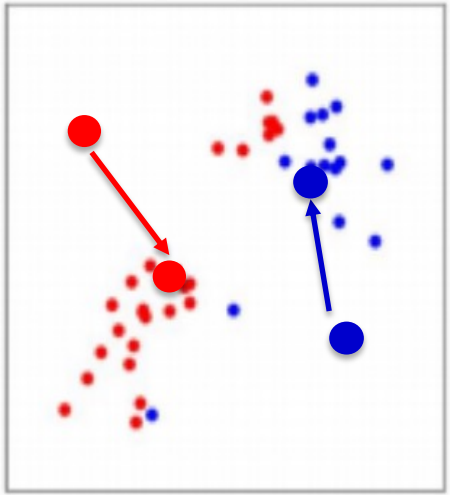
- Cluster 1 is assigned to the closest data points to centroid 1
- Cluster 2 is assigned to the closest data points to centroid 2

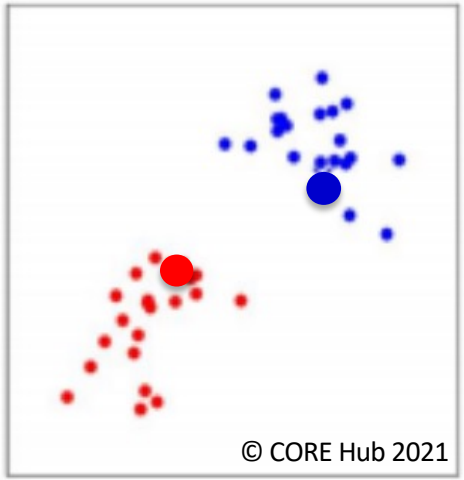


© CORE Hub 2021

Clustering: K-means

- Calculate the centroid of points in cluster 1 and cluster 2
- Move centroid 1 and centroid 2 to these locations
- Cluster 1 is assigned to the closest data points to centroid 1
- Cluster 2 is assigned to the closest data points to centroid 2



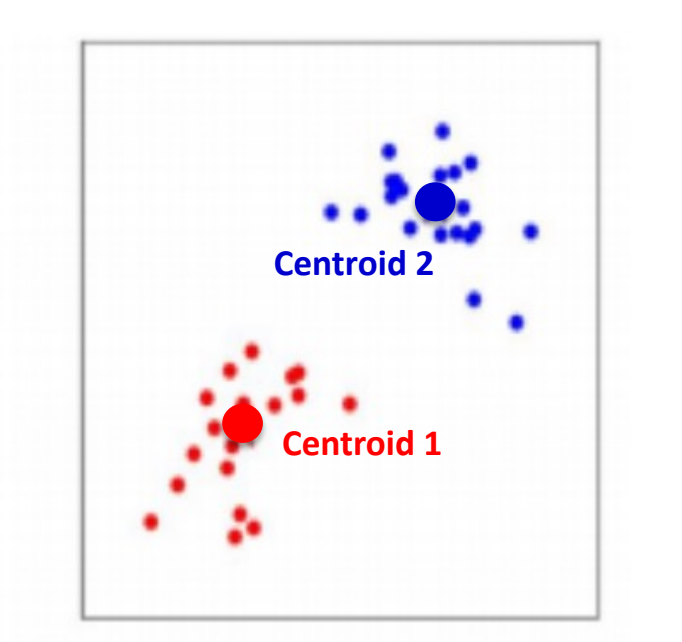


© CORE Hub 2021

Clustering: K-means

- Repeat until there are no more changes in the centroid locations

- The centroids obtained are representative of the cluster of points and are used for further analysis



© CORE Hub 2021

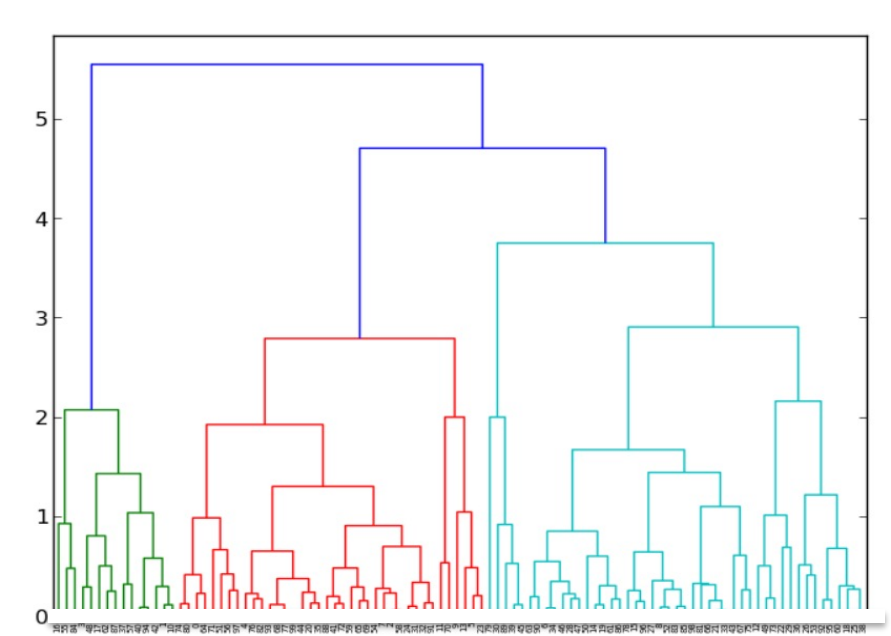
Clustering: Hierarchical

- Hierarchical clustering works using a measure of similarity between observations
- The choice of the similarity measure depends on the problem
- Data values start all in one cluster and splits are carried out recursively using different linkage criteria
- The process is represented using a dendogram
- The clusters are determined visually by analising the dendogram and using expert judment

© CORE Hub 2021

Clustering: Hierarchical

- Example of a dendogram showing three clusters



© CORE Hub 2021

Clustering: Spectral

- Spectral clustering is a more advanced technique for clustering
- It is called 'spectral' because uses a spectral decomposition of a similarity matrix
- The similarity between two observations X and Y is defined by the exponential function
- $S(X,Y) = exp\left(-\frac{\|X-Y\|^2}{2\sigma^2}\right)$
- The scaling parameter σ^2 controls how fast the similarity between the two points falls off with their distance

© CORE Hub 2021

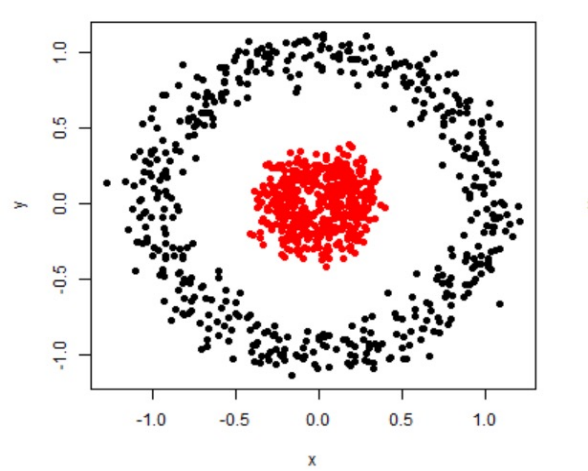
Clustering: Spectral

- The key idea of spectral clustering is to cluster the raw data using the eigenvectors of the similarity matrix
- The first k eigenvectors are used with k the number of clusters
- K-means approach is used with the eigenvectors in a slightly different way to the traditional k-means approach
  - Why not to use K-means on the raw data?
  - The eigenvectors tend to form tight clusters that are not evident in the raw data

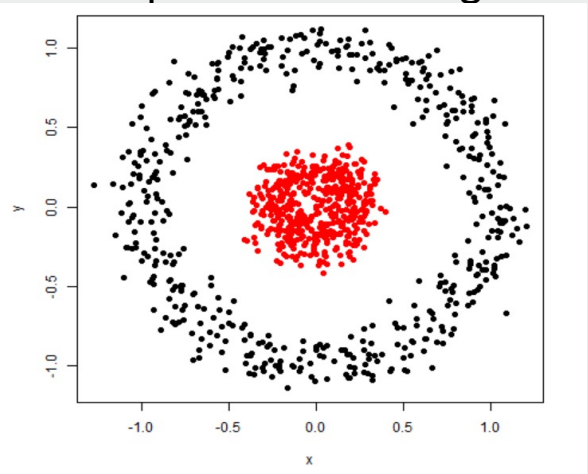
© CORE Hub 2021

Clustering: Comparison

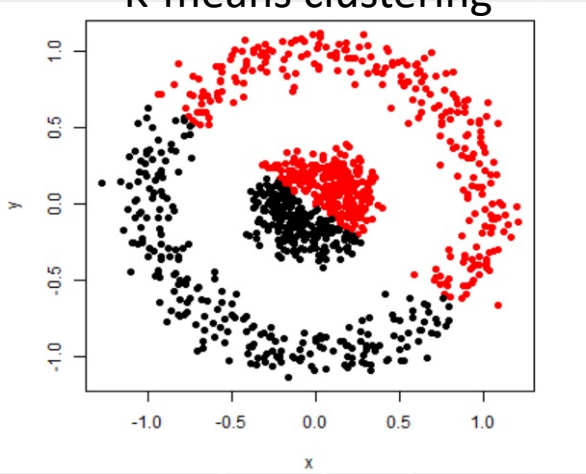
Raw data



Spectral clustering



K-means clustering



© CORE Hub 2021

Open pm1_clustering.ipynb

and work through the exercises

© CORE Hub 2021

Schedule

DAY 6

AWST	AEST	Agenda	Educator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	**Robust Regression (DAY 5 continued)**	Ying
08:15	10:15	**More Regression**	Ying
09:15	11:15	*Morning Tea*	
09:30	11:30	**Dimensionality Reduction**	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	Capstone Project Update/Share	Tamryn/All
12:15	14:15	**Clustering and High-D Regression**	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Partial Least Squares** (high-level overview)	Jeremy
13:50	15:50	Apply to your own Problem	
14:45	16:45	**Closeout** – Reflections, Takeaways	Jeremy
14:55	16:55	**Menti** Feedback	Tamryn
17:00	17:00	**Close**	

© CORE Hub 2021

Partial Least Squares

A short introduction to PLS regression

- PLS regression = Partial Least Squares regression
- Its principle is similar to PCA regression, i.e., help with ill-conditioned multivariate problems
- While PCA regression looks for uncorrelated components as a linear combination of the input variables only, PLS regression also tries to maximize the covariance between input and output variables
- Especially useful when there are more variables than observations or multicollinearity between the input variables

© CORE Hub 2021

Exercise: The Octane dataset

Open **pm2-PLS.ipynb**

and go through the notebook

© CORE Hub 2021

Schedule

DAY 6

AWST	AEST	Agenda	Educator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	**Robust Regression (DAY 5 continued)**	Ying
08:15	10:15	**More Regression**	Ying
09:15	11:15	*Morning Tea*	
09:30	11:30	**Dimensionality Reduction**	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	Capstone Project Update/Share	Tamryn/All
12:15	14:15	**Clustering and High-D Regression**	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Partial Least Squares** (high-level overview)	Jeremy
13:50	15:50	Apply to your own Problem	
14:45	16:45	**Closeout** – Reflections, Takeaways	Jeremy
14:55	16:55	**Menti** Feedback	Tamryn
17:00	17:00	**Close**	

© CORE Hub 2021

Apply to your own Problem

If you don’t have a dataset

- UCI Machine Learning Repository:
  archive.ics.uci.edu/ml/index.php

- Kaggle datasets:
  www.kaggle.com/datasets

© CORE Hub 2021

Schedule

DAY 6

AWST	AEST	Agenda	Educator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	**Robust Regression (DAY 5 continued)**	Ying
08:15	10:15	**More Regression**	Ying
09:15	11:15	*Morning Tea*	
09:30	11:30	**Dimensionality Reduction**	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	Capstone Project Update/Share	Tamryn/All
12:15	14:15	**Clustering and High-D Regression**	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Partial Least Squares** (high-level overview)	Jeremy
13:50	15:50	Apply to your own Problem	
14:45	16:45	**Closeout** – Reflections, Takeaways	Jeremy
14:55	16:55	**Menti** Feedback	Tamryn
17:00	17:00	**Close**	

© CORE Hub 2021

Takeaways & Closeout

High-level Takeaways From Today

- Understand the differences between univariate and multivariate data and apply multivariate techniques and EDA.
- Understand the basis of dimensionality reduction.
- Execute measures of correlations.
- Critique whether a correlation is spurious.
- Identify if the process used in the activities is reproducible.
- Understand the impact of missing data and recall practical solutions to work with incomplete data sets.

© CORE Hub 2021

High-level Takeaways From Today

Your thoughts?

© CORE Hub 2021

CORE Skills

A CORE partnership

COREHUB.COM.AU/SKILLS