



Data Analysis & Visualisation

CSC3062

BEng (CS & SE), MEng (CS & SE), BIT & CIT

Dr Reza Rafiee

Semester 1 2019



Quick review on some technical terms

- Dataset (data set, cohort)
- Variables (or features) & feature space
- Observation (sample)
- Variation
- Dimension
- Pattern



Principal component analysis (PCA)

- A **linear** dimensionality-reduction technique
 - **Transforming** variables (or features) of a large dataset (i.e., multivariate data) into a smaller one that still contains most of the information in the large dataset



Principal component analysis (PCA)

Reducing data by **projecting** (geometrically) into a lower dimensions which called principal components (PCs)



PCA example

Let's look at
a question & analysis for better
understanding of PCA



PCA example (17 dimensions)

Eating in the UK

Assume we have a dataset including 17 features/dimensions (Table 1). This table shows the average consumption of 17 types of food in grams per person per week for every country in the UK.

The table shows some interesting variations across different food types, but overall differences aren't so notable. Let's see if PCA can eliminate dimensions to emphasize how countries differ.

<http://www.sdss.jhu.edu/~szalay/class/2016-oldold/SignalProcPCA.pdf>

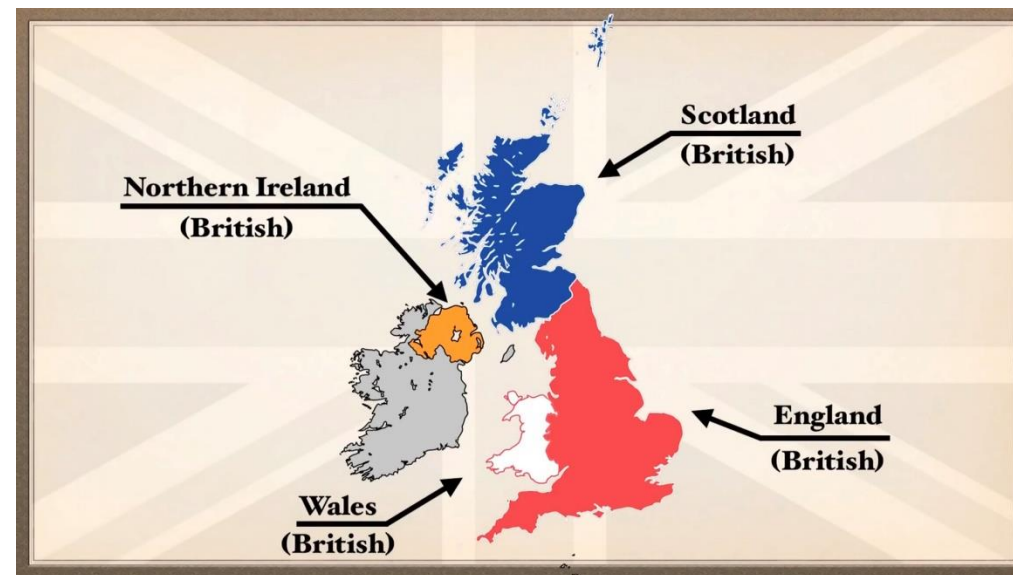
	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

Table 1: UK food consumption in 1997 (g/person/week). Source: DEFRA website



PCA example (17 dimensions)

Eating in the UK






PCA example (17 dimensions)

Eating in the UK

Assume we have a dataset including **17 features/dimensions** (Table 1). This table shows the *average consumption* of 17 types of food in grams per person per week for every country in the UK.

The table shows some interesting **variations across different food types**, but overall differences aren't so notable.

Let's see **if PCA can eliminate dimensions to emphasize how countries differ.**



	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

Table 1: UK food consumption in 1997 (g/person/week). Source: DEFRA website



PCA example (17 dimensions)

Eating in the UK – Question?

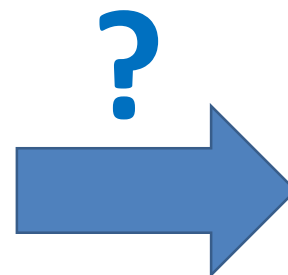
Can PCA reduce the dimension of this dataset (i.e., eliminate dimensions) to highlight how countries differ?



PCA analysis using *prcomp()* package

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

Input_dataset



	England	N Ireland	Scotland	Wales
PC1	-144.993	477.3916	-91.8693	-240.529
PC2	2.532999	58.90186	-286.082	224.6469
PC3	105.7689	-4.8779	-44.4155	-56.4756

Reduced dataset

Summarises of
features



PCA analysis using *prcomp()* package

	England	N Ireland	Scotland	Wales
PC1	-144.993	477.3916	-91.8693	-240.529
PC2	2.532999	58.90186	-286.082	224.6469
PC3	105.7689	-4.8779	-44.4155	-56.4756

Reduced dataset

Question: how many new variables (PCs) will be acceptable when using this transformation (or projection)?



PCA analysis using *prcomp()* package

Input_dataset

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

```
PCA_Model_prcomp <- prcomp(t(Input_dataset), center = T, scale=F)  
# scale =T is appropriate for high-dimensional data
```



PCA analysis using *prcomp()* package

```
PCA_Model_prcomp <- prcomp(t(Input_dataset), center = T, scale=F)
```

```
# scale =T is appropriate for high-dimensional data
```

```
summary(PCA_Model_prcomp)
```

```
# Importance of components:
```

#	PC1	PC2	PC3	PC4
# Standard deviation	324.1502	212.7478	73.87622	3.828e-14
# Proportion of Variance	0.6744	0.2905	0.03503	0.000e+00
# Cumulative Proportion	0.6744	0.9650	1.00000	1.000e+00



PCA analysis using *prcomp()* package

```
PCA_Model_prcomp <- prcomp(t(Input_dataset), center = T, scale=F)
```

```
# scale =T is appropriate for high-dimensional data
```

```
summary(PCA_Model_prcomp)
```

```
# Importance of components:
```

#	PC1	PC2	PC3	PC4
# Standard deviation	324.1502	212.7478	73.87622	3.828e-14
# Proportion of Variance	0.6744	0.2905	0.03503	0.000e+00
# Cumulative Proportion	0.6744	0.9650	1.00000	1.000e+00

```
PCA_Model_prcomp$x # Showing the principle components
```

#	PC1	PC2	PC3	PC4
# England	-144.99315	2.532999	105.768945	-3.765391e-14
# N Ireland	477.39164	58.901862	-4.877895	1.667659e-13
# Scotland	-91.86934	-286.081786	-44.415495	-8.860586e-13
# Wales	-240.52915	224.646925	-56.475555	7.770000e-13



PCA analysis using *prcomp()* package

```
PCA_Model_prcomp <- prcomp(t(Input_dataset), center = T, scale=F)
```

```
# scale =T is appropriate for high-dimensional data
```

```
summary(PCA_Model_prcomp)
```

```
# Importance of components:
```

```
#           PC1    PC2    PC3    PC4
# Standard deviation 324.1502 212.7478 73.87622 3.828e-14
# Proportion of Variance 0.6744 0.2905 0.03503 0.000e+00
# Cumulative Proportion 0.6744 0.9650 1.00000 1.000e+00
```

```
PCA_Model_prcomp$x # Showing the principle components
```

```
#           PC1    PC2    PC3    PC4
# England -144.99315 2.532999 105.768945 -3.765391e-14
# N Ireland 477.39164 58.901862 -4.877895 1.667659e-13
# Scotland -91.86934 -286.081786 -44.415495 -8.860586e-13
# Wales -240.52915 224.646925 -56.475555 7.770000e-13
```



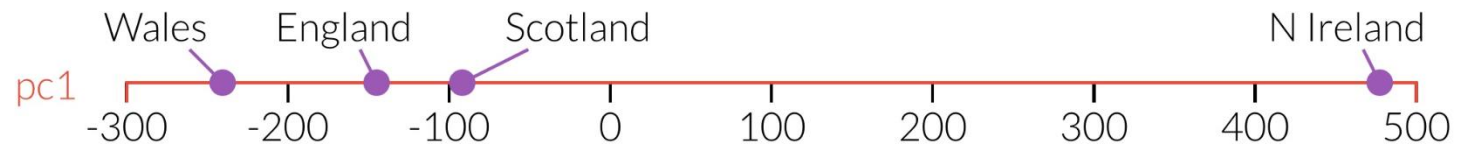
PCA analysis using *prcomp()* package

Eating in the UK

Here is the plot of the data along the first principal component (PCA).

PC1

# England	-144.99315
# N Ireland	477.39164
# Scotland	-91.86934
# Wales	-240.52915

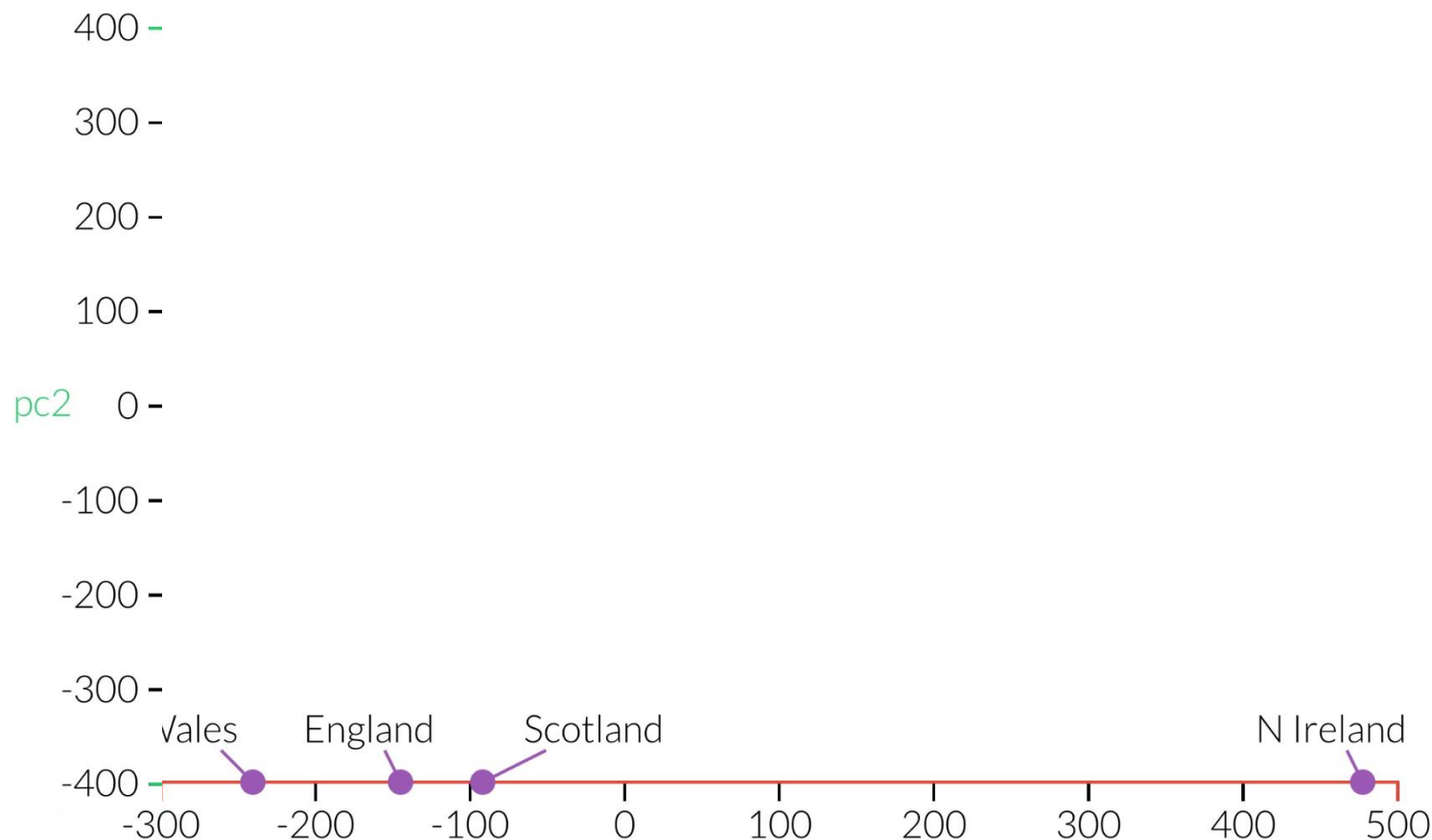




PCA analysis using *prcomp()* package

Eating in the UK

Adding the second principal components (PCA).



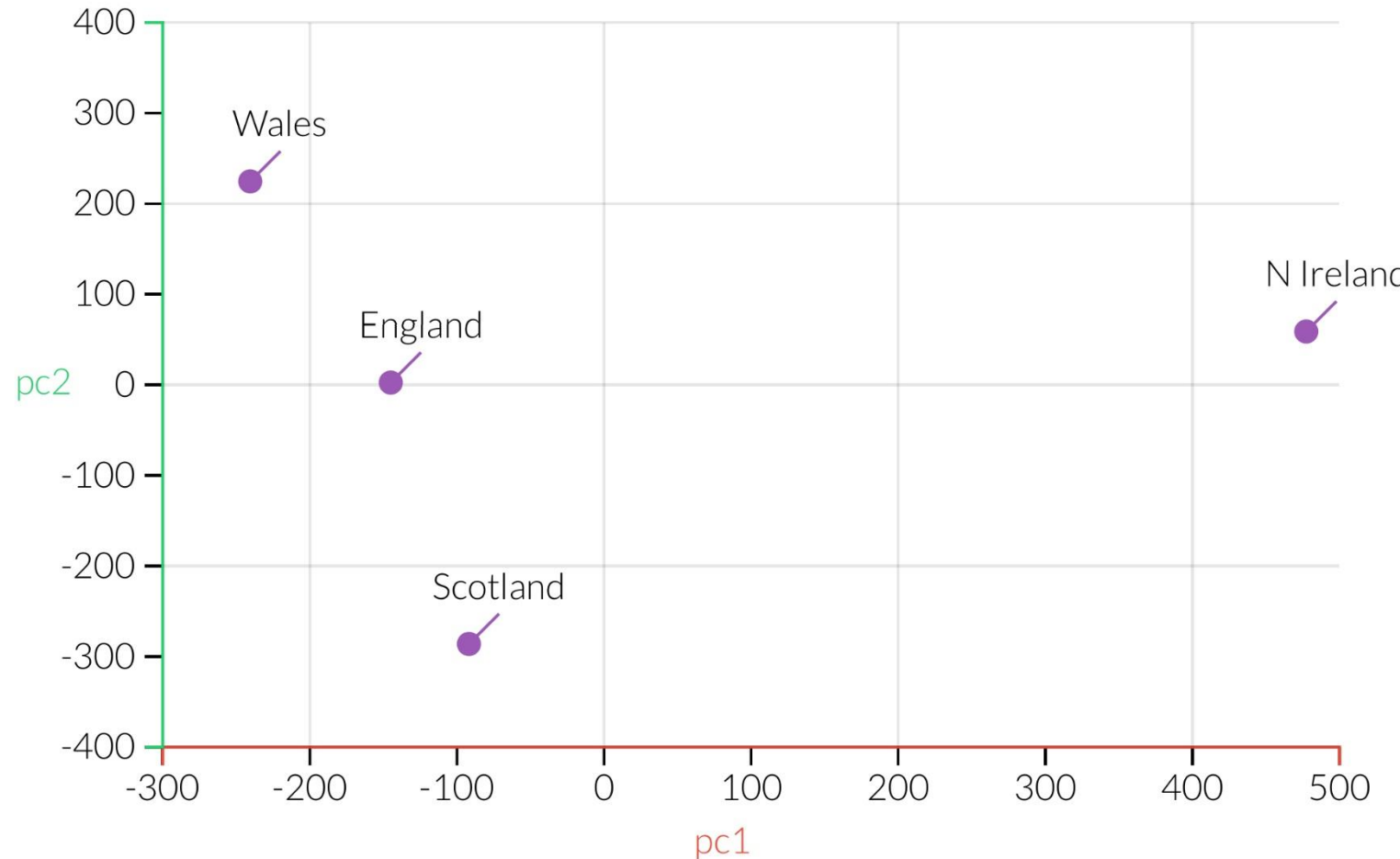


PCA analysis using *prcomp()* package

Eating in the UK

The first and second principal components (PCA).

#	PC1	PC2
# England	-144.99315	2.532999
# N Ireland	477.39164	58.901862
# Scotland	-91.86934	-286.081786
# Wales	-240.52915	224.646925





Interpret the PCs

Eating in the UK

Assume we have a dataset including 17 features/dimensions (Table 1). This table shows the average consumption of 17 types of food in grams per person per week for every country in the UK.

The Northern Irish eat way more grams of fresh potatoes and way fewer of fresh fruits, cheese, fish and alcoholic drinks.

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

Table 1: UK food consumption in 1997 (g/person/week). Source: DEFRA website



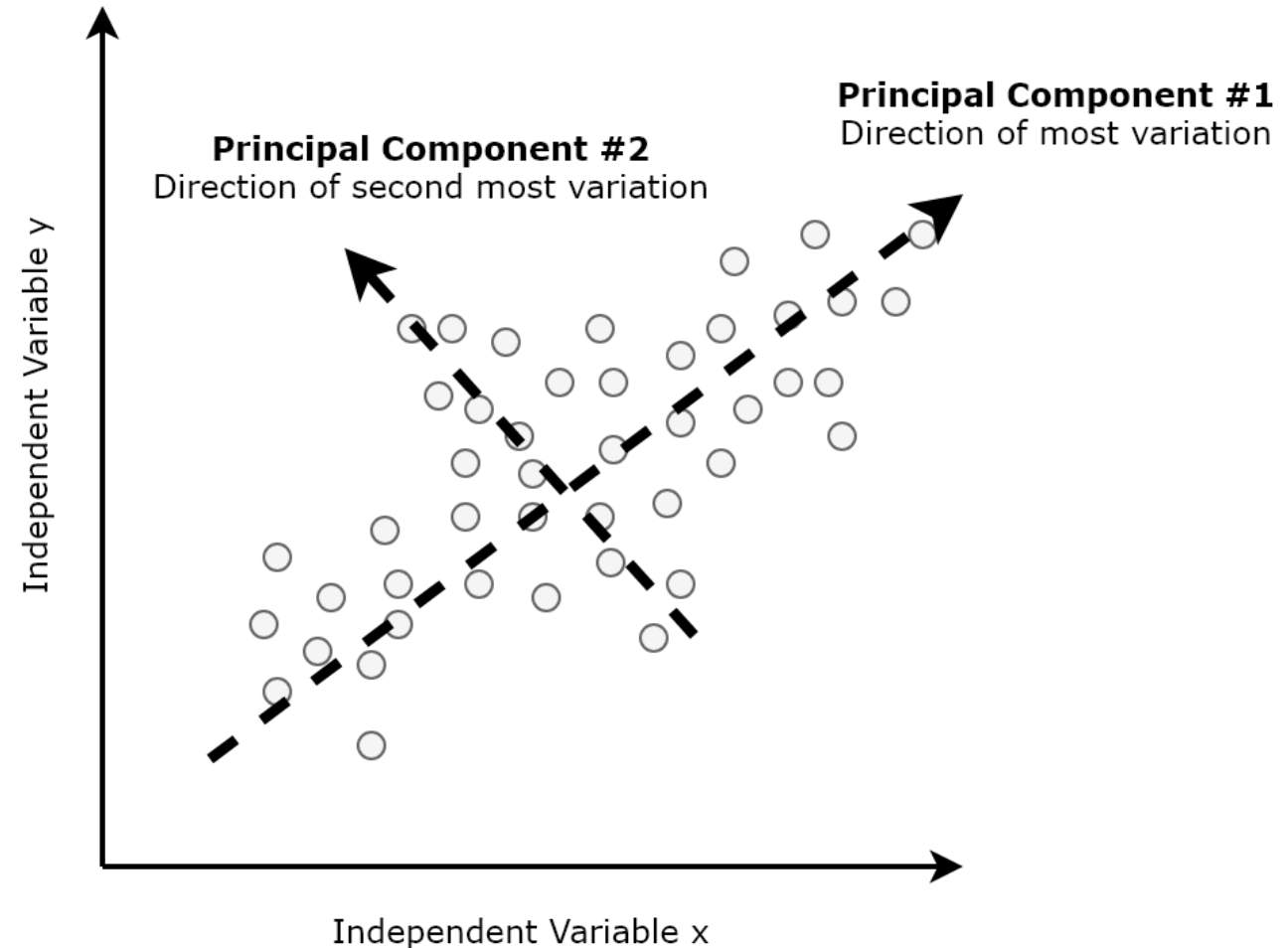
Principal component analysis (PCA)

- A **linear** dimensionality-reduction technique
 - **Transforming** variables (or features) of a large dataset (i.e., multivariate data) into a smaller one that still contains most of the information in the large dataset
- By using PCA, we **reduce the number of features** of a dataset, while preserving as much information as possible.
- Reducing data by **projecting** (geometrically) into a lower dimensions which called principal components (PCs)



Principal component analysis (PCA)

The Principal component directions are directions in the feature space along which the original data are high variable.





Principal component analysis (PCA)

- A technique which is used to emphasise **variation** and **reveal strong patterns** in a dataset. It is often used to make data easy to **explore** and **visualise**.
- It's an **unsupervised learning method** and is similar to clustering.
- It could be considered as a **compression method**.
- Each **feature** could be considered as a different **dimension**.
If you have more than 3 features in your dataset, it could be very difficult to visualise!
- Trade of between accuracy and **simplicity**

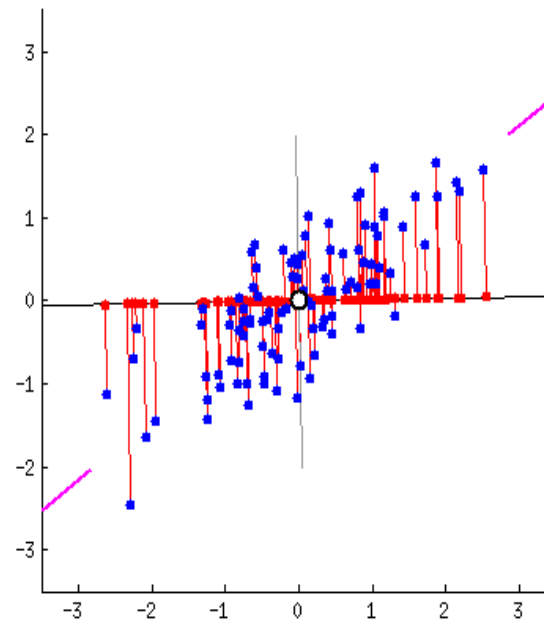


PCA is used **to extract** the important information from a **multivariate** data table and to express this information as a set of few **new variables** called **PCs (principal components)**. These **new variables** correspond to a **linear combination of the originals**. The number of principal components is less than or equal to the number of **original variables**.



PCA

- PCA allows us to **summarise** and to **visualise** the information in a data set containing individuals/observations/samples described by multiple inter-correlated quantitative variables.
- Each variable could be considered as a different dimension. If you have more than 3 variables in your data sets, it could be very difficult to visualize a multi-dimensional hyperspace.





Any Questions?