UNIVERSITY OF EDINBURGH
Business School

Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

Dr Xuefei Lu
The University of Edinburgh Business School

Dimension Reduction

Principal component analysis (PCA)

Principal component analysis (PCA)

PCA is a multivariate technique with the aim of reducing the dimensionality of a dataset while accounting for as much of the original variation (variance) as possible present in the dataset.

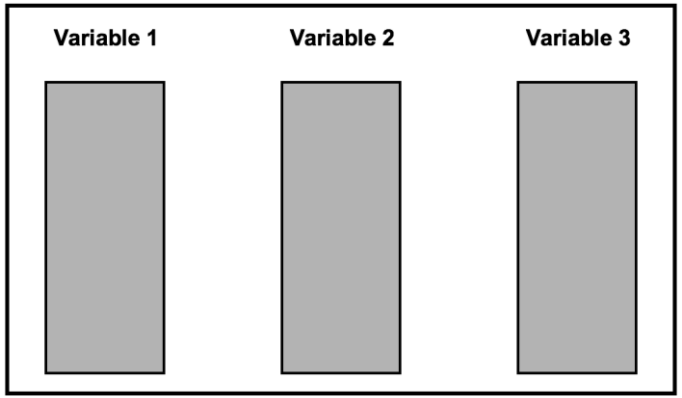
This aim is achieved by transforming the original variables to a new set of variables, the principal components (PCs), that are:

- linear combinations (i.e., weighted sums) of the original predictors
- uncorrelated among themselves
- ordered so that the first few of them account for most of the variation

Potentially, the result of a PCA would be the creation of a small number of new variables that can be used as surrogates for the large number of original variables, and consequently provide a simpler basis for undertaking further analyses like linear regression (but not only!).

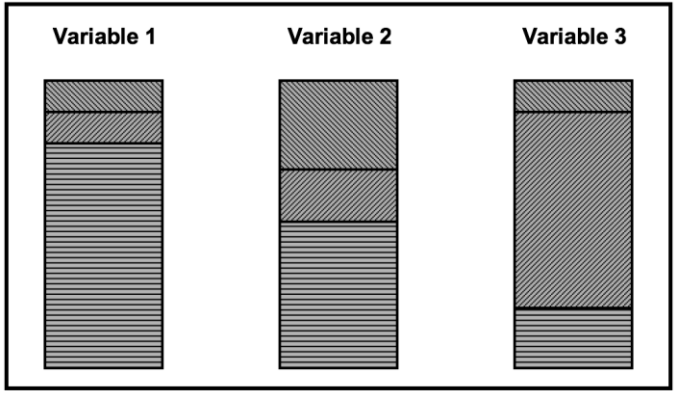
PCA– Heuristic interpretation

For illustration, consider an example with 3 predictors. We represent schematically this situation as follows:



Total original information = 100%

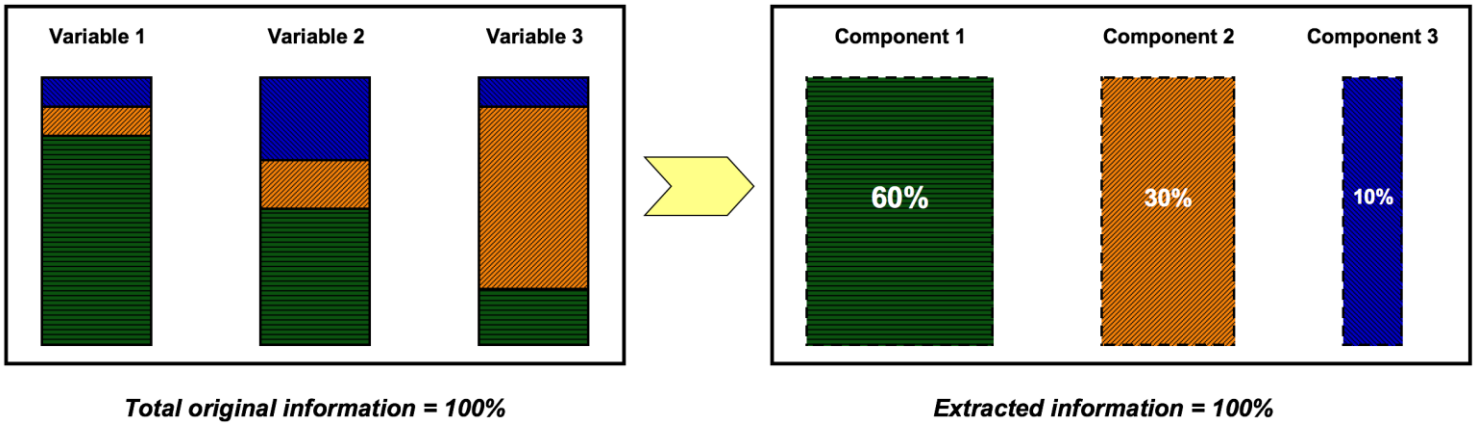
We assume that these variables do not provide "separate" unrelated pieces of information, but they are instead correlated each other. We represent the information shared by the original variables as follows:



Total original information = 100%

PCA– Heuristic interpretation

PCA allows to "aggregate" in new variables (the components) the common information shared by the original variables:

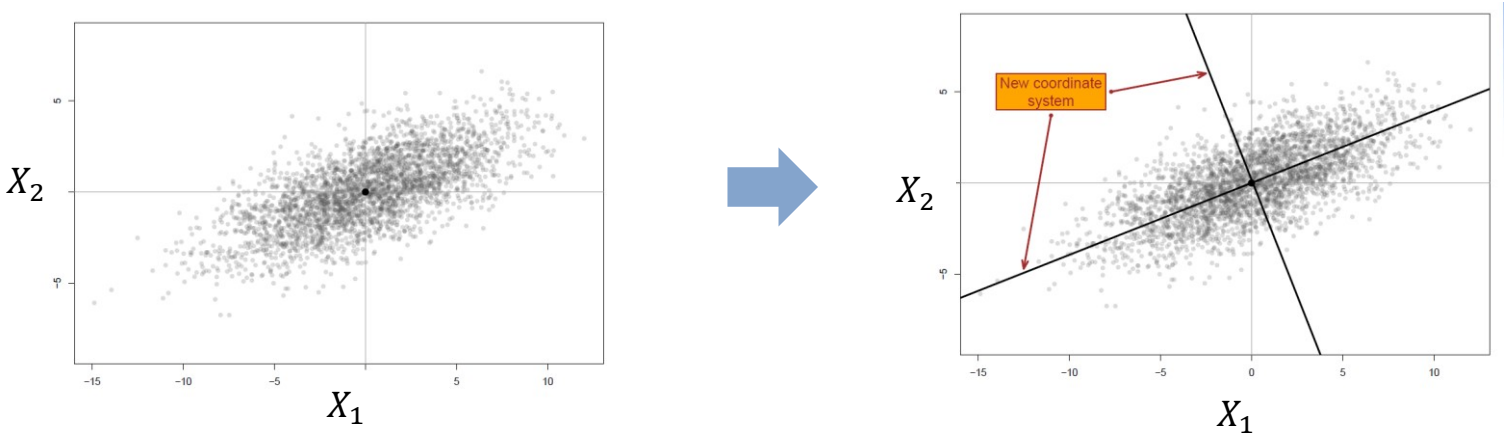


The components are obtained by linearly combining the original variables and are ordered according to the information they convey.

PCA– Geometric intuition

Geometrically, PCA amounts to changing the coordinate system with respect to which the original variables are measured such that, according to the new system, the new variables (the PCs) are no longer correlated.

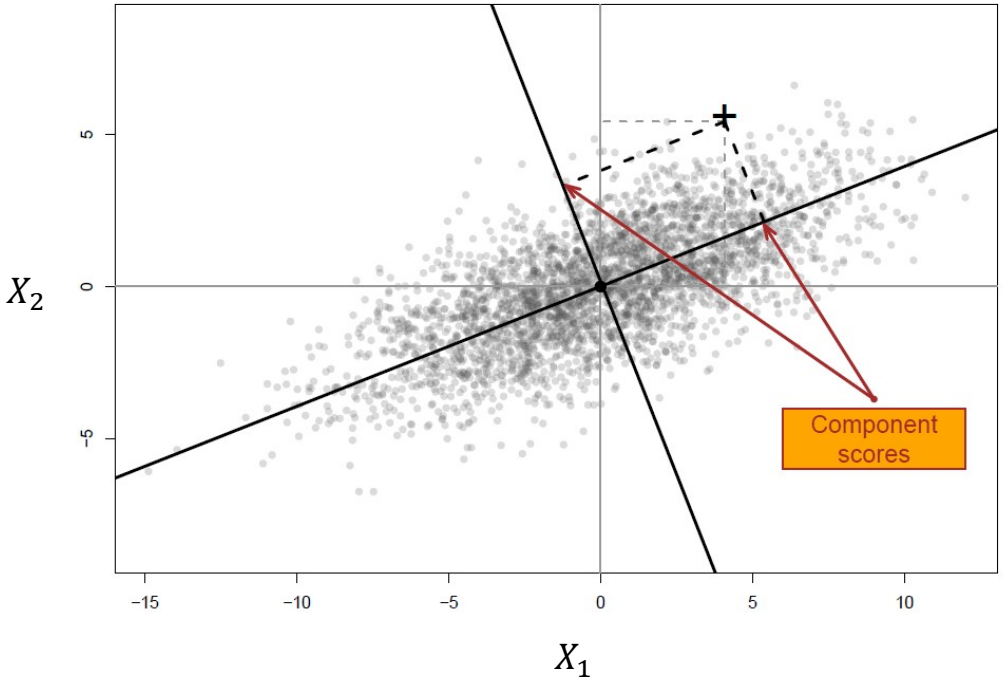
A further hypothetical example with only two variables X_1 and X_2.



PCA– Geometric intuition

Consider now a specific point in the diagram.

The new coordinates are called **the component scores**.



Finding the principal components

More technically, given a set of variables $X_1, \dots, X_p$, the **first** principal component, PC_1, is obtained as

$$PC_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

that is as the linear combination of the original variables with largest variance.

Since the variance of PC_1 could be increased without limit simply by increasing the coefficients $a_{11}, \dots a_{1p}$, a common choice for is to require that the sum of squares of the coefficients

takes value one, that is $\sum_{j=1}^{p} a_{1j}^2 = 1$.

The second principal component, PC_2, is defined as the linear combination

$$PC_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

that has the second greatest variance and satisfying: 1) $\sum_{j=1}^{p} a_{2j}^2 = 1$; 2) being **uncorrelated** with the first component

The third, fourth, and so on are found in a similar way.

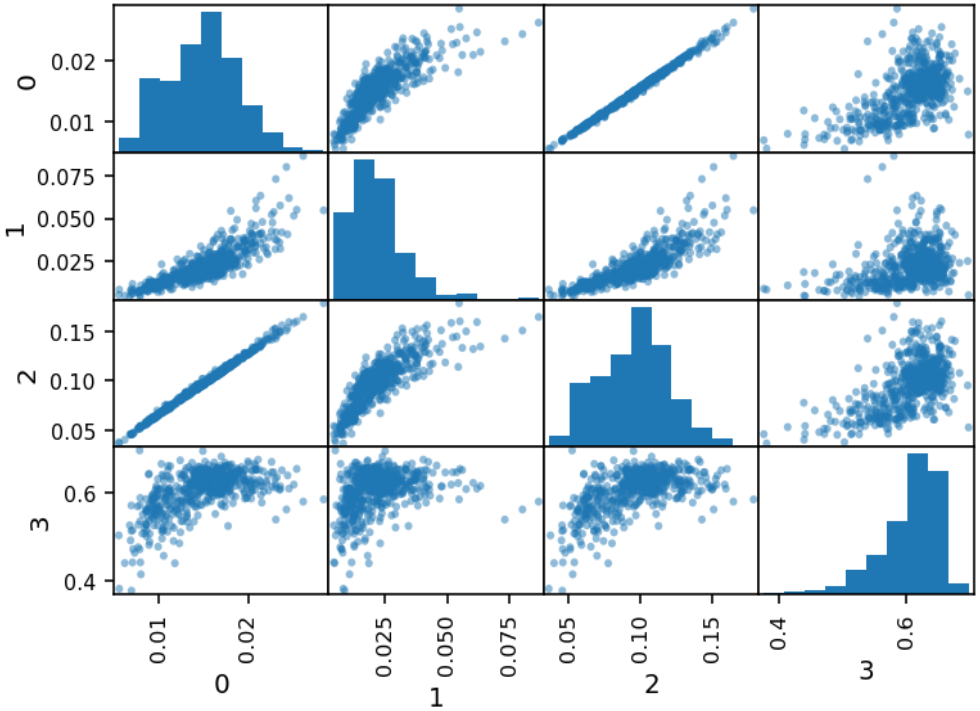
The coefficients $a_{l1}, \dots a_{lp}$ for the l-th component are called the **loadings** (component weights/component coefficients) of the l-th component on the different variables -- importance of a variable on a principle component.

PCA – notes

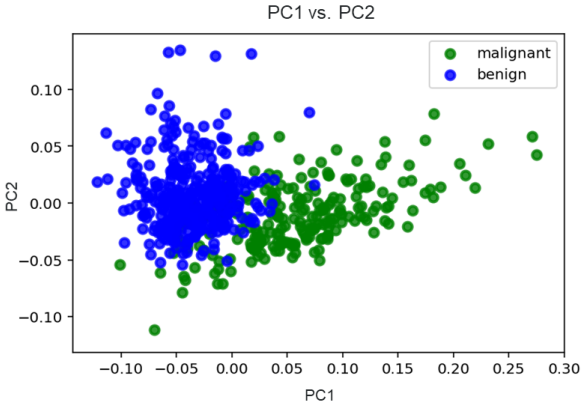
- A very important point with PCA is that it is **not scale-invariant**. This means that rescaling the variables may lead to a very different solution.
- Additionally, if there are large differences between the variances (i.e., the scales) of the original variables, then those whose variances are largest will tend to dominate the early components.
- Therefore, PCA should be always performed on "standardized" variables.

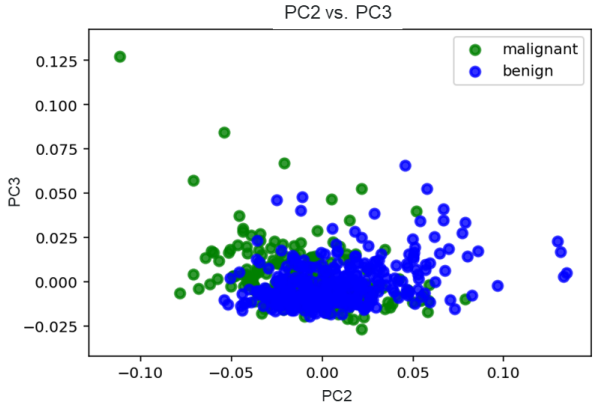
- Note that PCA is unsupervised -- It does not take into account the dependent variable. Thus, it does not necessarily create principal components that are suitable for prediction. Also the principal components might not necessarily explain a particularly interpretable result.

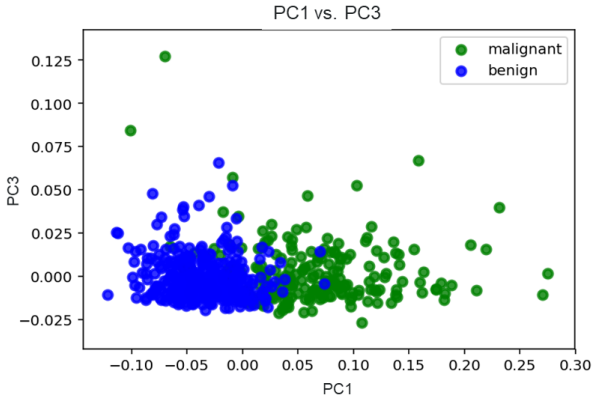
PCA – the breast cancer example



PCA – the breast cancer example







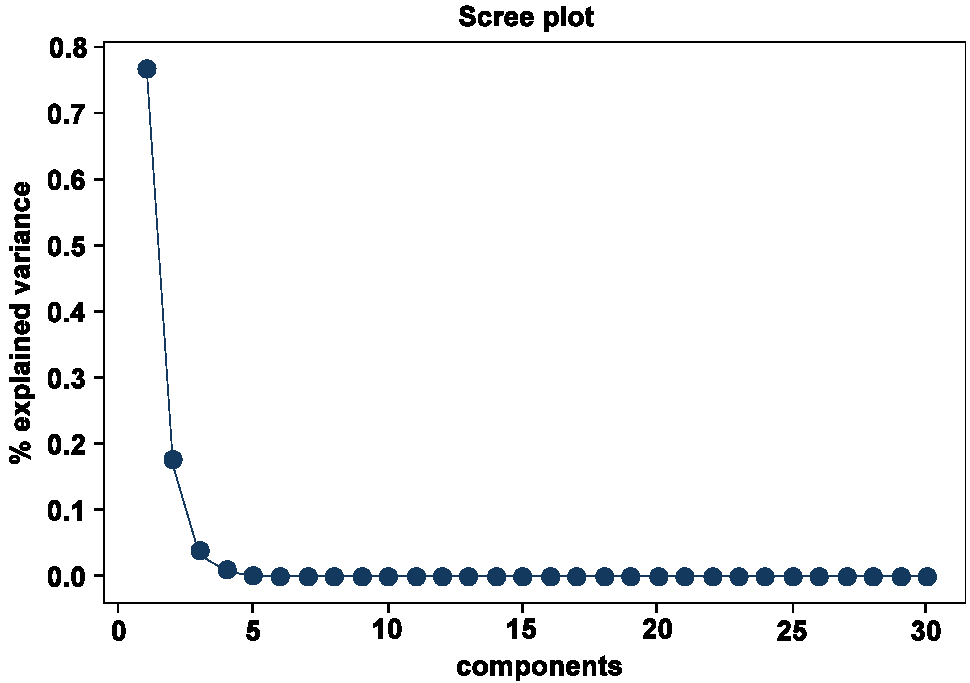
PCA – the breast cancer example

	PC1	PC2	PC3
0	-0.04817	0.078424	-0.00534
1	-0.09319	**0.258465**	0.053075
2	**-0.29956**	**0.488159**	-0.02267
3	**-0.69655**	**-0.4251**	0.019013
4	-0.00051	0.001367	0.000281
5	-0.00026	0.00115	0.000566
…	…	…	…
21	-0.11025	0.353206	-0.00474
22	**-0.25337**	**0.596564**	-0.05951
23	**0.575504**	**0.118657**	-0.12563
24	-0.00064	0.002052	-6.41E-05
25	-0.00024	0.002789	-0.00148

PCA – notes

- The variation in the original p variables is only completely accounted for by all p principal components.
- How many components to retain?
  - Retain just enough components to explain at least 70-80% of the total variation in the original variables
  - Exclude those PCs whose variances are less than one. Note that, due to standardization, the average variance of the original variables is equal to one. Hence, this method suggests to retain those components that account for more variance than the average for the original variables
  - Some authors suggest the examination of the so called **scree plot**, a plot of the component variances versus the component. The number of components selected is the value corresponding to an "elbow" in the curve, that is, to a substantial change in the slope

PCA – scree plot



the breast cancer example

PCA – notes

- By repeating the analysis, it may be very likely the case that you will find opposite signs for some of the loadings and for some component scores.
- This is not a mistake and, most importantly, it does not affect the final interpretation of the results.
- The components provide a new coordinate system for the data. However, components only provide the direction of the new coordinates (the axes themselves), but not their orientation (whether positive or negative).
- If you find an opposite sign for some of the loadings, the corresponding component will also have a reversed sign.

- Study the following file

7 - Coding_for_PCA.ipynb

- Then try the following exercise and answer the questions in the next slides

  8 - Activity_6_performing_your_own_PCA.ipynb + absent.csv

Activity: Performing your own PCA

Quiz

Based on 8 - Activity_6_performing_your_own_PCA.ipynb + absent.csv

Q1. You have performed PCA and can now answer what variables, or better, what linear combinations of variables are explaining the variance in the dataset well. What are the 5 most important variables in the first principle component?

- A. Reason for absence, Transportation expense, Distance from Residence to Work, Service time, Height
- B. Reason for absence, time, Age, Weight, Height
- C. Reason for absence, Transportation expense, Distance from Residence to Work, Service time, Weight
- D. Transportation expense, Distance from Residence to Work, Service time, Age, Weight

Quiz

Based on 8 - Activity_6_performing_your_own_PCA.ipynb + absent.csv

Q2. What are the 5 most important variables in the second principle component?

- A. Reason for absence,Month of absence,Distance from Residence to Work,Work load Average/day,Height
- B. Reason for absence,Month of absence,Distance from Residence to Work,Work load Average/day,Absenteeism time in hours
- C. Distance from Residence to Work,Work load Average/day,Weight,Height,Absenteeism time in hours
- D. Reason for absence,Month of absence,Distance from Residence to Work,Work load Average/day,Weight

Quiz

Based on 8 - Activity_6_performing_your_own_PCA.ipynb + absent.csv

Q3. Based on the two previous questions, which two variables do you think might be worthwhile for inclusion in the final model?

- A. Reason for absence,Distance from Residence to Work
- B. Transportation expense, height
- C. Transportation expense,Work load Average/day
- D. Work load Average/day,Weight

UNIVERSITY OF EDINBURGH
Business School

Let's discuss performing your own PCA
You will now share and discuss the insights you obtained from performing the PCA.

Consider and discuss the following four questions with your peers:

- Did you find it easy to construct and interpret the PCA?
- What variables have you found to be important?
- What can the loadings tell you?
- What about the interaction between variables – were the correlated variables split out over components?

Comment on posts that had similar findings to you, as well as those that differed greatly. Discuss the possible reasons for this and your opinions about your peers' insights.

Quiz

Q1. What metrics are suitable for regression, rather than classification?

- A. RMSE, recall
- B. Recall and precision
- C. RMSE, MAPE
- D. MAPE, recall

Quiz

Q2. What metric is more suitable to evaluate a skewed dataset?

- A. recall
- B. precision
- C. F-score
- D. AUC

Quiz

Q3. Principle component analysis is what type of feature analysis technique?

- A. Filter method
- B. Wrapper method
- C. Embedded method
- D. Feature creation method

Quiz

Q4. Mutual information and Chi-squared values can be used to rank features. Therefore, they are:

- A. Filter method
- B. Wrapper method
- C. Embedded method
- D. Feature creation method

We introduce a dataset and pick some subsets to illustrate that, given the same model, using different data for evaluation can paint quite a different picture of the same model.

In this activity, it's up to you to find good and representative subsets to evaluate the model. Please try the following file:

X - Activity_7_evaluating_your_model.ipynb

Activity: Evaluating your model (optional)

UNIVERSITY OF EDINBURGH
Business School