FEATURE SELECTION BY FEATURE PERMUTATION

- Let's suppose we already have a model. Which are the most important features for **that model**?
  - Note: although attributes are going to be evaluated individually, it's not a filter method because it uses a pre-existing model to sort the features. It's now wrapper either, because attributes are evaluated individually.
- A feature a_k is just a column in the data table.
- First, error $\hat{e}$ of the model is estimated with validation data (different the ones used for training).
- Second, values of feature a_k are randomly ordered and a new model error is computed: $\hat{e}_{ak}$
- If is important, error is going to increase: $\hat{e}_{ak}$- $\hat{e}$ is going to be large. And the other way arround, if it is not important, $\hat{e}_{ak}$- $\hat{e} \sim 0$
- Therefore, features a_k can be sorted by $\hat{e}_{ak}$- $\hat{e}$ (from largest to smallest) and the top features are selected.

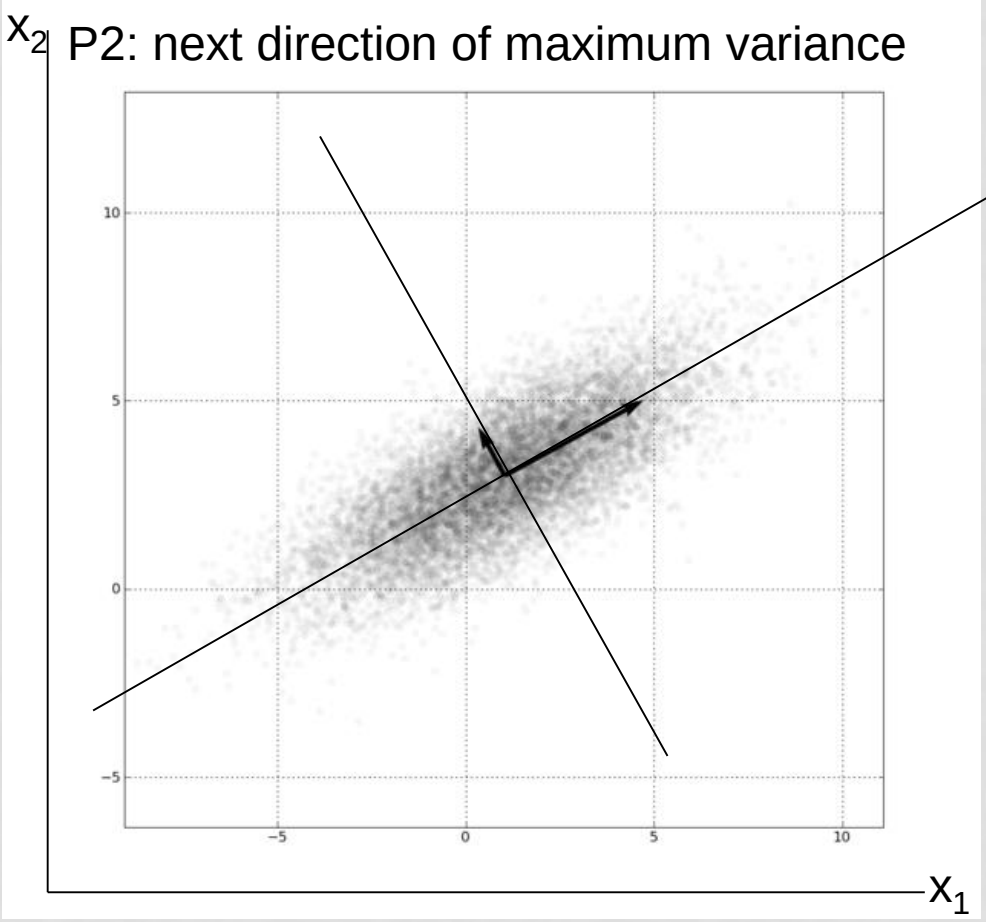
TRANSFORMATION (+ SELECTION) OF ATTRIBUTES

- Principal Component Analysis (PCA)

TRANSFORMATION WITH PRINCIPAL COMPONENT ANALYSIS (PCA)

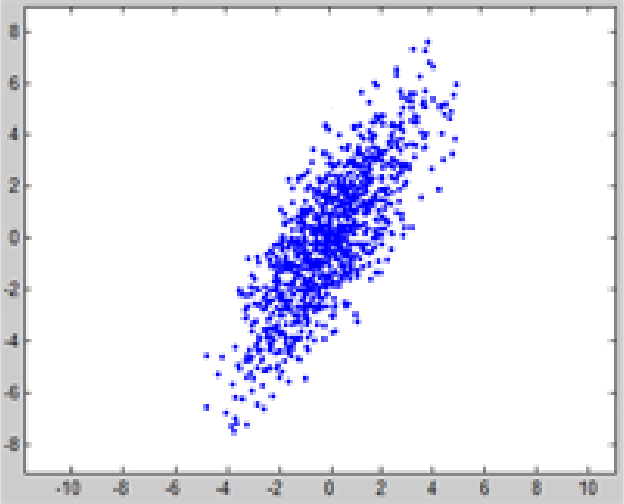
- This method constructs new attributes, as a linear combination of the original input attributes
- The new attributes are sorted by the variance of the new attributed (explained variance)
- Dimensionality can be reduced by choosing the attributes with more variance

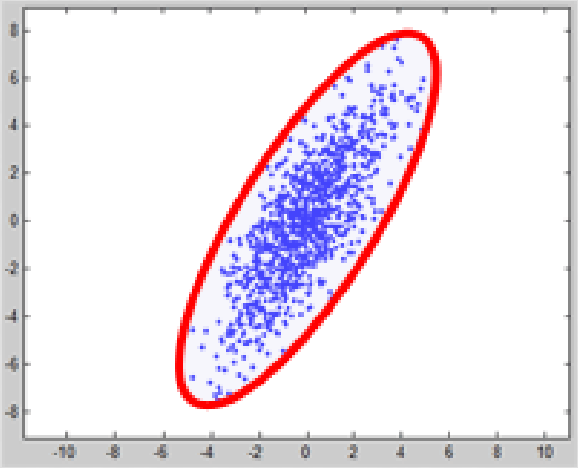
PCA

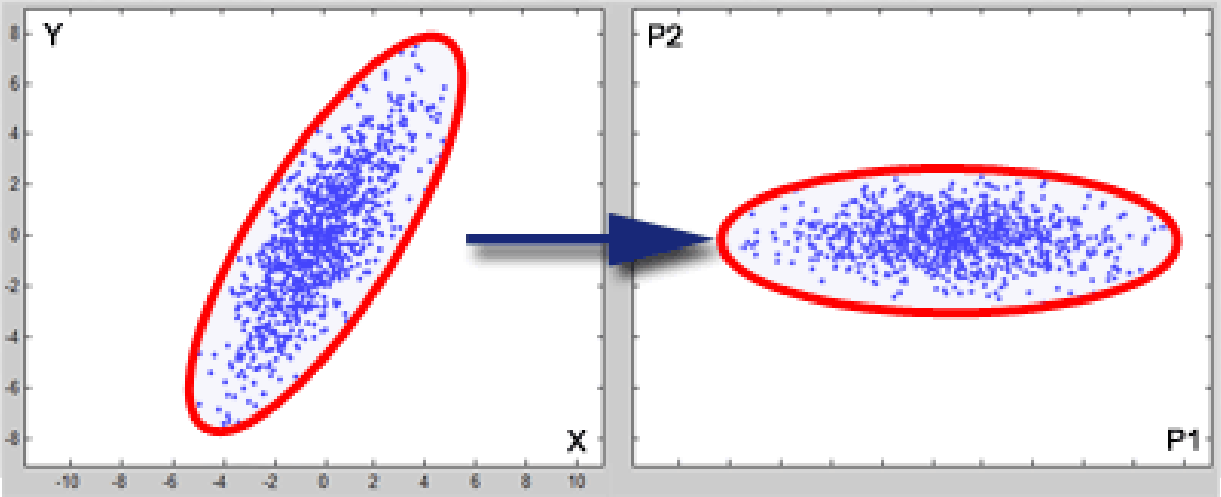


Two new attributes: P1 and P2

PCA TRANSFORMATION







- Linear transformations
- It removes redundancy from attributes (correlation)

$P_1 = k_{11}*x_1+k_{12}*x_2$

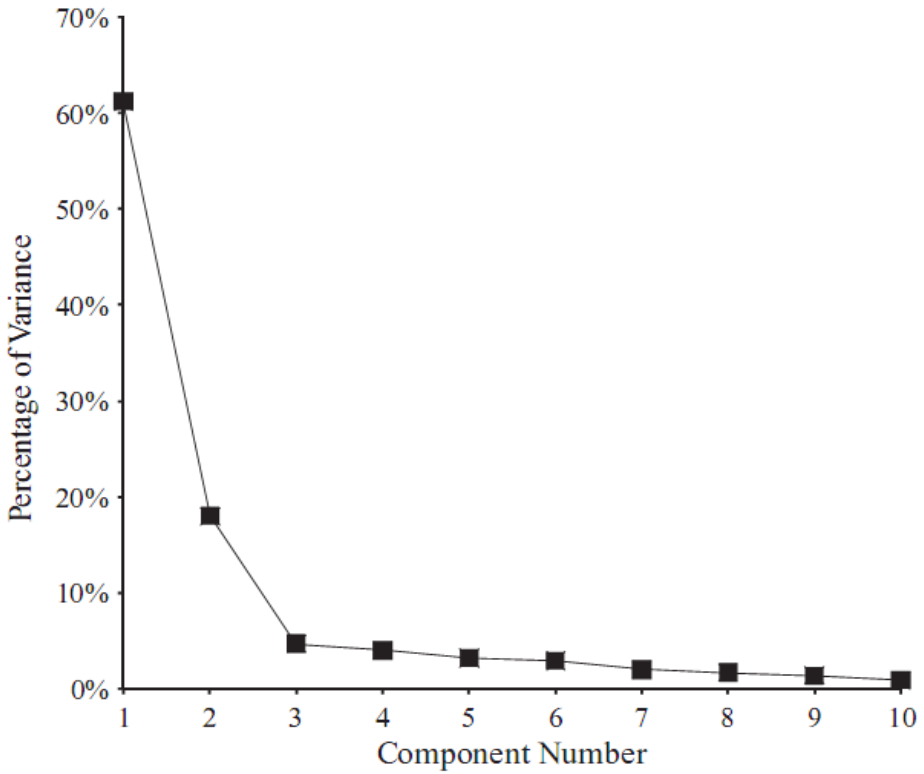
$P_2 = k_{21}*x_1+k_{22}*x_2$

$P = \mathbf{X}*k$

PCA: TRANSFORMATION AND SELECTION

Axis	Variance	Cumulative
1	61.2%	61.2%
2	18.0%	79.2%
3	4.7%	83.9%
4	4.0%	87.9%
5	3.2%	91.1%
6	2.9%	94.0%
7	2.0%	96.0%
8	1.7%	97.7%
9	1.4%	99.1%
10	0.9%	100.0%

(a)

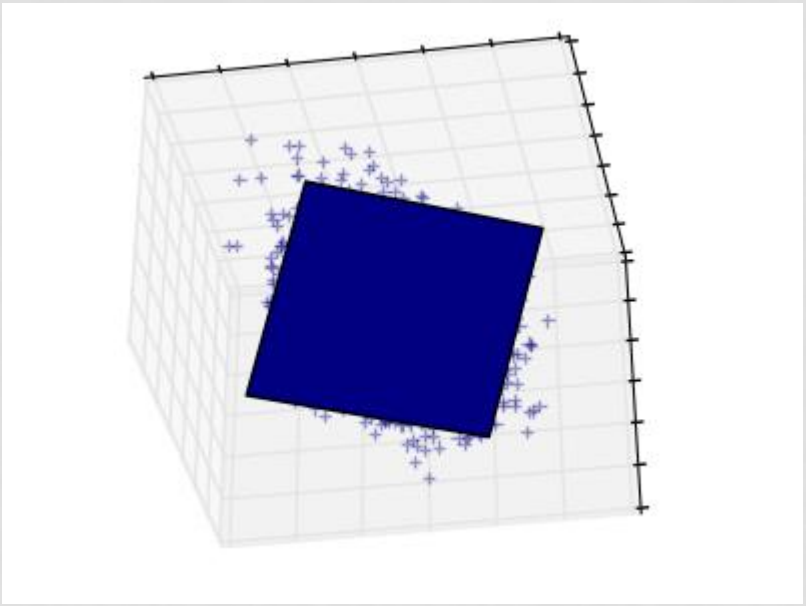


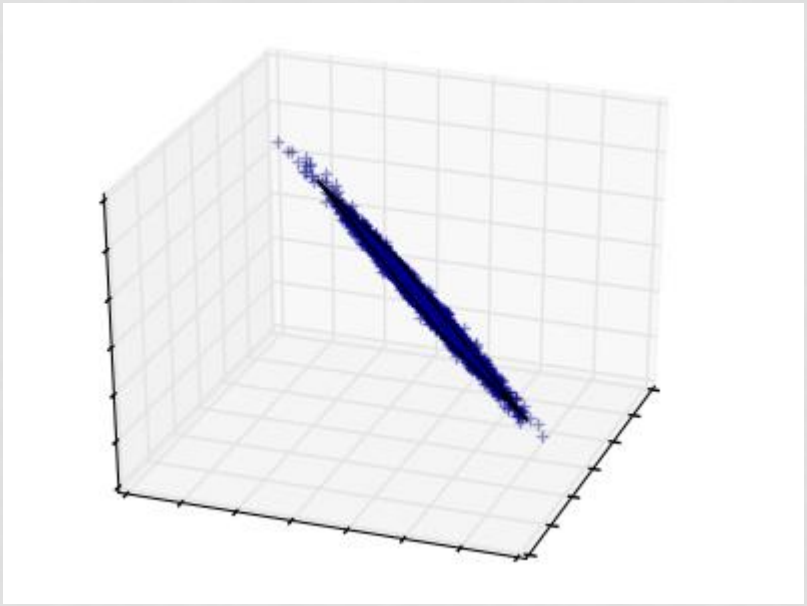
(b)

- Typically, a threshold is set so that the explained variance is larger than 95% (7 in this case)
- If only a few attributes explain most of the variance, the rest can be removed (e.g. imagine two dimensional data embedded in 20 dimensions)

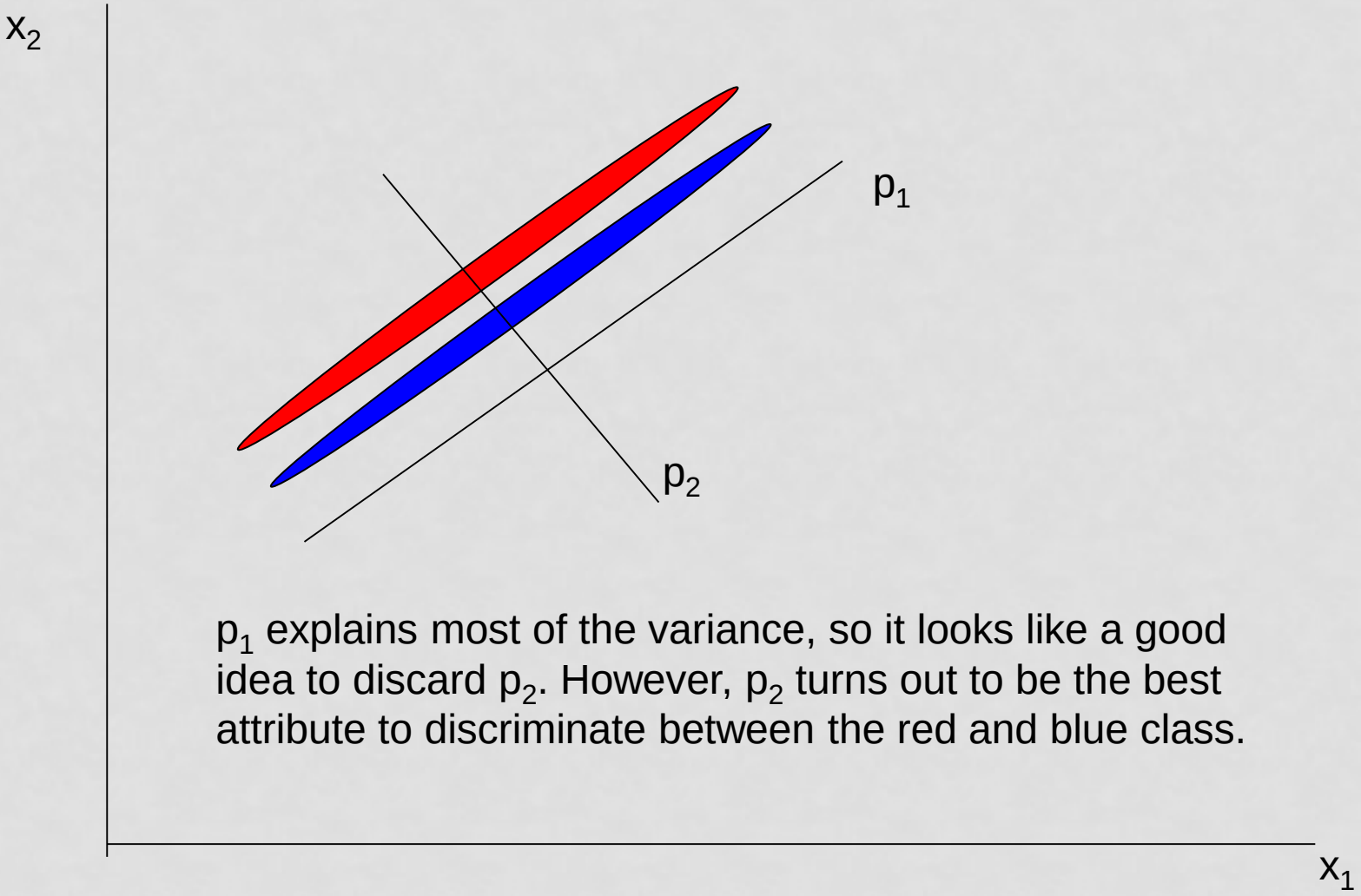
PCA AND ACTUAL DIMENSION OF DATA

- A two dimensional dataset embedded in three dimensions





BEWARE, PCA IS NOT SUPERVISED



p_1 explains most of the variance, so it looks like a good idea to discard p_2. However, p_2 turns out to be the best attribute to discriminate between the red and blue class.

ADVANTAGES / DISADVANTAGES OF PCA

- Advantage: it may find out the actual dimensionality of data
  - E.g.: let's imagine instances in 2D with an ellipsoid shape, but embedded in 20 dimensions. PCA will easily identify that only 2 dimensions are required.
- Advantage: decorrelates attributes (removes redundancy between attributes)
- Disadvantage: PCA is **not supervised**, so there is guarantee that it will find out the attributes that best discriminate between the classes.
- Disadvantage: Slow if lots of attributes.