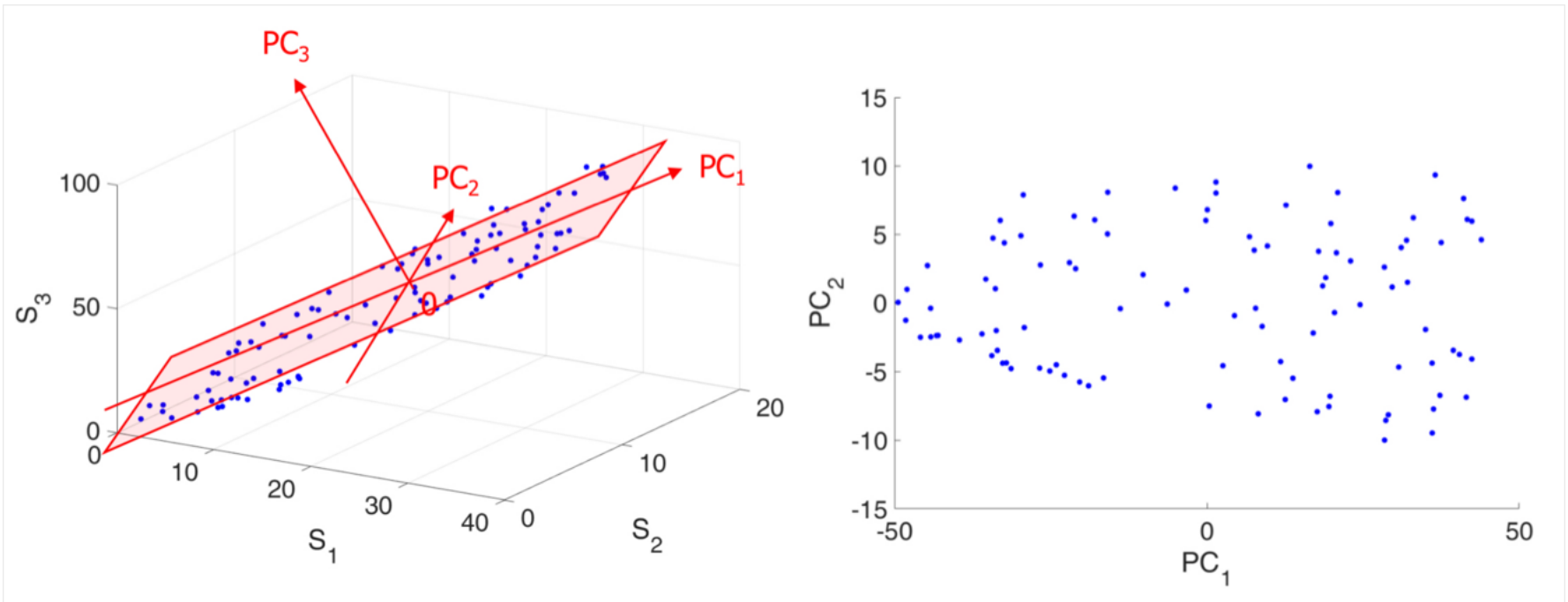
Principal component analysis

- Goal: project high dimensional data onto a lower dimensional subspace while preserving maximal information
- Motivation (the why): visualisation, omitting unimportant variables, saving computational power, speeding up learning algorithms
    - Plotting the first 2 components might reveal new insights, e.g. clusters



1. Subtract average from each dimension
2. Find a direction which maximises the average distance from data points = maximises data variance
3. Repeat previous step in perpendicular direction

- PCA gives components with decreasing order of importance
- PCA is sensitive to rescaling
    - Normalise variables beforehand
- Interpreting components might be tricky, but it is possible (not a black box algorithm)

Input

- Design matrix (feature matrix)

Output

- Variances - show how much scatter each component captures, decreases with each component
    - Used to cut off the number of components to keep
    - Usually we chose to keep a number of components which adds up to a high enough percentage of total variance (usually 80 or 90%)
- Scores - coordinates of data points in the new subspace
    - Used for visualisation
- Coefficients - coordinates of components in the original space, provide interpretation to the components
    - Used for interpretation