Dimension reduction

- If the points do not fill the canvas fully, it means that it lives in a lower dimension
- The higher the dimension, the more concentration of points in the centre

Principle component analysis

- A smaller set of variables that contains as much information as the original as possible
- A sequence of linear combinations of variables that have maximal variance, and are mutually uncorrelated

Construction of PC1 :

- Line rotation until it has the greatest variance in the data
- PC1 is a new variable created from a linear combination

$Z_1 = \phi_{11} x_1 + \phi_{21} x_2 + \dots + \phi_{p1} x_p$, with constraint on $\sum_{j=1}^{p} \phi_{j1}^2 = 1$

loading

Loading vector $\phi_1 = [\phi_{11}, \dots, \phi_{p1}]^T$, set the direction in the feature space

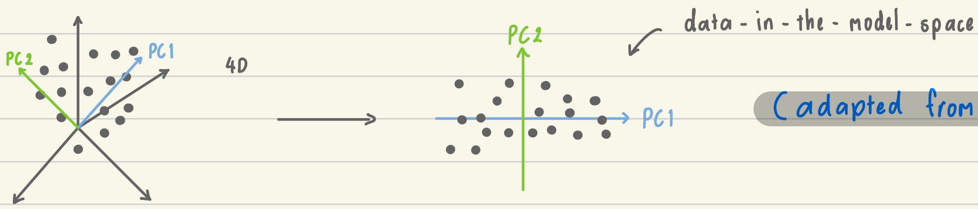
$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \dots + \phi_{p1} x_{ip}$

Construction of PC2 :

- Line orthogonal (perpendicular) to PC1, with next highest variance

$z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \dots + \phi_{p2} x_{ip}$

- There are at most $\min(n-1, p)$ PCs



(adapted from BrendiA github)

Total variance :

- If variables are standardised, TV = # of variables

Proportion of variance explained :

- $PVE_m = \frac{V_m}{TV}$, $CPVE_m = \sum_{m=1}^{k} \frac{V_m}{TV}$
- Elbow rule

- Scaling of variables matter, mean = 0, variance = 1
- Outliers can affect results