Outline

The data matrix

The data set is a $n \times p$ matrix $\mathbf{X} = (x_{ij})$ with values in $\mathbb{R}$:

- each row $\mathbf{x}_i$ represents an individual/observation
- each col $\mathbf{x}^j$ represents a variable/attribute

$$\mathbf{X} = \begin{array}{c} \\ \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_i \\ \vdots \\ \mathbf{x}_n \end{array} \begin{array}{c} \begin{array}{cccccc} \mathbf{x}^1 & \mathbf{x}^2 & \dots & \mathbf{x}^j & \dots & \mathbf{x}^p \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix} \end{array}$$

Cloud of observation in $\mathbb{R}^p$

Individuals can be represented in the variable space $\mathbb{R}^p$ as a point cloud

[Example in $\mathbb{R}^3$]{cloud_centering}{width="60%"}

Center of Inertia

(or barycentrum, or empirical mean)

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = \begin{pmatrix} \sum_{i=1}^{n} x_{i1}/n \\ \sum_{i=1}^{n} x_{i2}/n \\ \vdots \\ \sum_{i=1}^{n} x_{ip}/n \end{pmatrix}$$

We center the cloud $\mathbf{X}$ around $\mathbf{x}$ denote this by $\mathbf{X}^c$

$$\mathbf{X}^c = \begin{pmatrix} x_{11}-\bar{x}_1 & \dots & x_{1j}-\bar{x}_j & \dots & x_{1p}-\bar{x}_p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1}-\bar{x}_1 & \dots & x_{ij}-\bar{x}_j & \dots & x_{ip}-\bar{x}_p \end{pmatrix}$$

Inertia and Variance

Total Inertia:

distance of the individuals to the center of the cloud

$$I_T = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij}-\bar{x}_j)^2 = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i-\bar{\mathbf{x}}\|^2 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{dist}^2(\mathbf{x}_i,\bar{\mathbf{x}})$$

Proportional to the total variance

Let $\hat{\boldsymbol{\Sigma}}$ be the empirical variance-covariance matrix

$$I_T = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2 = \sum_{j=1}^{p}\frac{1}{n}\|\mathbf{x}^j-\bar{x}_j\|^2 = \sum_{j=1}^{p}\mathbb{V}(\mathbf{x}^j) = \mathrm{trace}\Big(\hat{\boldsymbol{\Sigma}}\Big)$$

$\rightsquigarrow$ Good representation has large inertia (much variability)

$\rightsquigarrow$ Large dispertion $\sim$ Large distances between points