Statistics 1

Hadrien Montanelli

1 Random samples

Random sample A *random sample* of size n is a set of random variables $X_1, \dots, X_n$ that are iid.

2 Summary statistics

Sample mean and variance The *sample mean* and the *sample variance* are the random variables defined by

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

3 Maximum likelihood estimation

Likelihood Let $X_1, \dots, X_n$ have joint pmf/pdf $f(\boldsymbol{x};\boldsymbol{\theta})$, which depends on some parameters $\boldsymbol{\theta}$. Given observed values $x_1, \dots, x_n$, the *likelihood* of $\boldsymbol{\theta}$ is the function

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta};\boldsymbol{x}) = f(\boldsymbol{x};\boldsymbol{\theta}).$$

The *log-likelihood* is $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$.

Likelihood (iid) Let $X_1, \dots, X_n$ be a random sample of size n with pmfs/pdfs $f_{X_i}(x_i;\boldsymbol{\theta})$. Then,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{X_i}(x_i;\boldsymbol{\theta}).$$

Maximum likelihood estimator (MLE) The *maximum likelihood estimate* $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ is the $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta})$ for given $\boldsymbol{x}$; $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ is the *maximum likelihood estimator*.

Computing MLEs Either by solving $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = 0$, or by looking at the graph of $L : \boldsymbol{\theta} \mapsto L(\boldsymbol{\theta})$.

4 Parameter estimation

Statistic A *statistic* is any function $T(\boldsymbol{X})$ that does not depend on θ.

Estimator An *estimator* of θ is any statistic $T(\boldsymbol{X})$ that we might use to estimate θ. $T(\boldsymbol{x})$ is the *estimate* of θ obtained via $T(\boldsymbol{X})$ from observed values $\boldsymbol{x}$.

Mean squared error (MSE) The *mean squared error* of an estimator T is defined by

$$\mathrm{MSE}(T) = \mathrm{E}([T-\theta]^2).$$

Bias The *bias* of an estimator T is defined by

$$b(T) = E(T) - \theta.$$

The estimator is *unbiased* if $b(T) = 0$ for all θ. MLEs are often asymptotically unbiased, and have MSEs $\sim 1/n$. For any estimator T, the following relation holds,

$$\mathrm{MSE}(T) = \mathrm{var}(T) + b(T)^2.$$

5 Confidence intervals

Confidence interval (CI) Given two statistics $a(\boldsymbol{X})$ and $b(\boldsymbol{X})$, and $0 < \alpha < 1$, the interval $(a(\boldsymbol{X}), b(\boldsymbol{X}))$ is called a *confidence interval* for θ with confidence level $1-\alpha$ if, for all θ,

$$\mathrm{P}(a(\boldsymbol{X}) < \theta < b(\boldsymbol{X})) = 1 - \alpha.$$

It is also called a $100(1-\alpha)\%$ confidence interval.

CI for normal Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma_0^2)$, where μ is unknown and σ_0^2 is known. Then,

$$\frac{\overline{X} - \mu}{\sigma_0/\sqrt{n}} \sim \mathrm{N}(0,1).$$

Therefore,

$$\mathrm{P}\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma_0/\sqrt{n}} < z_{\alpha/2}\right) = 2\Phi(z_{\alpha/2}) - 1 = 1 - \alpha,$$

where $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. (Note that $\Phi(-x) = 1 - \Phi(x)$.) In other words,

$$\left(\overline{X} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right)$$

is a $100(1-\alpha)\%$ CI for μ. ($z_{\alpha/2} = 1.96$ for $\alpha/2 = 0.05/2$.)

Similarly, one-sided $100(1-\alpha)\%$ CIs are

$$\left(-\infty, \overline{X} + z_\alpha \frac{\sigma_0}{\sqrt{n}}\right) \quad \text{and} \quad \left(\overline{X} - z_\alpha \frac{\sigma_0}{\sqrt{n}}, +\infty\right).$$

Central limit theorem (CLT) Let $X_1, \dots, X_n$ be a random sample of size n of any distribution with mean μ and variance $\sigma^2 < \infty$. Then, for all x,

$$\mathrm{P}\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq x\right) \to \Phi(x) \quad \text{as} \quad n \to \infty.$$

CI using the CLT Let $X_1, \dots, X_n$ be a random sample of size n of any distribution with mean $\mu(\theta)$ and variance $\sigma^2(\theta) < \infty$. Then,

$$\frac{\overline{X} - \mu(\theta)}{\sigma(\theta)/\sqrt{n}} \overset{d}{\approx} \mathrm{N}(0,1),$$

which yields

$$\mathrm{P}\left(-z_{\alpha/2} < \frac{\overline{X} - \mu(\theta)}{\sigma(\theta)/\sqrt{n}} < z_{\alpha/2}\right) \approx 1 - \alpha.$$

At this point, one can either solve the inequality for θ, or estimate $\mu(\theta)$ and/or $\sigma(\theta)$ using the MLE $\hat{\theta}$.

Standard error Let T be an estimator of θ based on $\boldsymbol{X}$. The *standard error* is defined by

$$\mathrm{SE}(T) = \sqrt{\mathrm{var}(T)}.$$

Note that $\mathrm{SE}(T)$ might depend on θ; in that case, the MLE $\hat{\theta}$ might be used to estimate the standard error.

6 Linear regression with intercept

Model For each $1 \le i \le n$,

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2),$$

where σ^2 and $x_1, \ldots, x_n$ are known, and α, β are unknown. This yields pdfs

$$f_{Y_i}(y_i; \alpha, \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right).$$

Log-likelihood

$$\ell(\alpha, \beta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

When the noise is $\overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, the MLE is equivalent to the *least squares estimator* (LSE) obtained by minimizing

$$S(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

(For the LSE, only additive ϵ_i with $\mathrm{E}(\epsilon_i) = 0$ is needed.)

MLE

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}, \quad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\mathrm{cov}(x, y)}{\sigma_x^2} = \frac{\rho_{x,y}\sigma_y}{\sigma_x},$$

with correlation coefficient $\rho_{x,y} = \mathrm{cov}(x, y)/(\sigma_x \sigma_y)$.

Mean and variance of MLE For the computation, it is convenient to define $w_i = x_i - \overline{x}$, so that $\sum_{i=1}^{n} w_i = 0$.

$$\mathrm{E}(\hat{\alpha}) = \alpha, \quad \mathrm{E}(\hat{\beta}) = \beta, \quad \mathrm{var}(\hat{\beta}) = \sigma^2 \Big/ \sum_{i=1}^{n} w_i^2.$$

Confidence interval Since

$$\hat{\beta} \sim \mathrm{N}\left(\beta, \sigma^2 \Big/ \sum_{i=1}^{n} w_i^2\right),$$

a $100(1-\alpha)\%$ CI for β is

$$\left(\hat{\beta} \pm z_{\alpha/2}\sigma \Big/ \sqrt{\sum_{i=1}^{n} w_i^2}\right).$$

7 Linear regression without intercept

Model For each $1 \le i \le n$,

$$Y_i = \beta x_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2),$$

where σ^2 and $x_1, \ldots, x_n$ are known, and β is unknown. This yields pdfs

$$f_{Y_i}(y_i; \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta x_i)^2\right).$$

Log-likelihood

$$\ell(\beta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2.$$

MLE

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}, \quad \mathrm{E}(\hat{\beta}) = \beta, \quad \mathrm{var}(\hat{\beta}) = \sigma^2 \Big/ \sum_{i=1}^{n} x_i^2.$$

Confidence interval Since

$$\hat{\beta} \sim \mathrm{N}\left(\beta, \sigma^2 \Big/ \sum_{i=1}^{n} x_i^2\right)$$

a $100(1-\alpha)\%$ CI for β is

$$\left(\beta \pm z_{\alpha/2}\sigma \Big/ \sqrt{\sum_{i=1}^{n} x_i^2}\right).$$

8 Assessing the fit of a model

Fitted value and residual

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i \quad \text{and} \quad e_i = y_i - \hat{y}_i.$$

Leverage

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}, \quad \text{high when } h_i > 4/n.$$

Mean and variance of residual

$$\mathrm{E}(e_i) = 0, \quad \mathrm{var}(e_i) = \sigma^2(1 - h_i).$$

RSS and RSE and R^2

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2, \quad \mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}}.$$

MSE and R^2

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} e_i^2, \quad \mathrm{R}^2 = 1 - \frac{\mathrm{RSS}}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Studentized residual

$$r_i = \frac{e_i}{\sqrt{\mathrm{var}(e_i)}} = \frac{e_i}{\sigma\sqrt{1 - h_i}}, \quad \text{outliers } |r_i| > 3.$$

Potential problems

- non-linearity (pattern in residual plot);
- varying variance (funnel-type shape in residual plot);
- errors are not independent;
- explanatory variables are measured with error;
- explanatory variables are not linearly independent.

9 Statistics in dimension $d \geq 1$

Covariance and correlation Let $\boldsymbol{X} \in \mathbb{R}^d$ be a random vector. The *covariance matrix* $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ of $\boldsymbol{X}$ has entries

$$\Sigma_{i,j} = \mathrm{cov}(X_i, X_j), \quad 1 \leq i, j \leq d,$$

while the *correlation matrix* $\boldsymbol{\rho} \in \mathbb{R}^{d \times d}$ has elements

$$\rho_{i,j} = \frac{\mathrm{cov}(X_i, X_j)}{\sqrt{\mathrm{var}(X_i)\mathrm{var}(X_j)}}, \quad 1 \leq i, j \leq d.$$

Random sample A *random sample* of size n in dimension $d \geq 1$ is a set of iid random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$.

Sample mean The *sample mean* is the random vector

$$\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i.$$

Sample covariance and correlation The *sample covariance* and *sample correlation* are the random matrices

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T, \quad R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}.$$

Note that $\boldsymbol{R} = \boldsymbol{W}^{-1/2}\boldsymbol{S}\boldsymbol{W}^{-1/2}$ with $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{S})$.

Mean-centred The *mean-centred* version of

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1^T \\ \vdots \\ \boldsymbol{X}_n^T \end{bmatrix} \in \mathbb{R}^{n \times d} \quad \text{is} \quad \begin{bmatrix} \boldsymbol{X}_1^T - \overline{\boldsymbol{X}}^T \\ \vdots \\ \boldsymbol{X}_n^T - \overline{\boldsymbol{X}}^T \end{bmatrix} \in \mathbb{R}^{n \times d}.$$

If $\boldsymbol{X}$ is mean-centred, then

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X}.$$

Properties of covariance Let $\boldsymbol{X} \in \mathbb{R}^d$ be a random vector with covariance matrix $\boldsymbol{\Sigma}$. Then, for any $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^d$,

$$\mathrm{var}(\boldsymbol{\alpha}^T\boldsymbol{X}) = \boldsymbol{\alpha}^T\boldsymbol{\Sigma}\boldsymbol{\alpha},$$
$$\mathrm{cov}(\boldsymbol{\alpha}^T\boldsymbol{X}, \boldsymbol{\beta}^T\boldsymbol{X}) = \boldsymbol{\alpha}^T\boldsymbol{\Sigma}\boldsymbol{\beta}.$$

Linear transformation (MVN) Let $\boldsymbol{X} \in \mathbb{R}^d \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{B} \in \mathbb{R}^{m \times d}$. Then, $\boldsymbol{B}\boldsymbol{X} \sim \mathrm{N}_m(\boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^T)$.

10 MLE in dimension $d \geq 1$

Likelihood (iid) Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be a random sample of size n with pmfs/pdfs $f_{\boldsymbol{X}_i}(\boldsymbol{x}_i; \boldsymbol{\theta})$, which depends on some parameters $\boldsymbol{\theta}$. Then,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n f_{X_i}(\boldsymbol{x}_i; \boldsymbol{\theta}).$$

MLE (MVN) Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be a random sample of size n with pdfs

$$f_{\boldsymbol{X}_i}(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det\boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right).$$

Then,

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i, \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T.$$

11 Linear regression for $d \geq 1$

Model For each $1 \leq i \leq n$,

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_d x_{i,d} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2),$$

where σ^2 and $x_{i,1}, \ldots, x_{i,d}$ are known, and $\beta_0, \ldots, \beta_d$ are unknown. This yields, adding a column of ones in $\boldsymbol{X}$,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n),$$

with $\boldsymbol{Y}, \boldsymbol{\epsilon} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times (d+1)}$, $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$, and pdf

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\beta}) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right).$$

For $\mathrm{N}_d(\boldsymbol{0}, \sigma^2\boldsymbol{I}_d)$ noise, MLE is equivalent to LSE,

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Log-likelihood

$$\ell(\boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

MLE

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

Significance test The t-statistic is used to test the significance of each parameter,

$$t_{\widehat{\beta}_i} = \frac{\hat{\beta}_i - \beta_{i,0}}{\sqrt{\mathrm{var}(\hat{\beta}_i)}} \sim t_{n-d}.$$

12 Logistic regression for $d \geq 1$

Model For each $1 \leq i \leq n$,

$$\mathrm{P}(Y_i = 1) = \frac{1}{1 + e^{-\boldsymbol{\beta}^T\boldsymbol{X}_i}}, \quad \mathrm{P}(Y_i = 0) = 1 - \mathrm{P}(Y_i = 1).$$

This yields pdfs

$$f_{Y_i}(y_i; \boldsymbol{\beta}) = \mathrm{P}(Y_i = 1)^{-y_i}(1 - \mathrm{P}(Y_i = 1))^{1-y_i}.$$

Log-likelihood

$$\ell(\boldsymbol{\beta}) = -\sum_{i=1}^n \log(1 + e^{\boldsymbol{\beta}^T\boldsymbol{X}_i}) + \sum_{i=1}^n y_i\boldsymbol{\beta}^T\boldsymbol{X}_i.$$

13 Principal component analysis

PCA Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$ be a random sample with sample covariance matrix $\boldsymbol{S} \in \mathbb{R}^{d \times d}$. The principal component analysis is the eigenvalue decomposition

$$\boldsymbol{S} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T,$$

where $\boldsymbol{D} \in \mathbb{R}^{d \times d}$ is the matrix of decreasing eigenvalues, and $\boldsymbol{V} \in \mathbb{R}^{d \times d}$ is the orthogonal matrix of eigenvectors.

Loadings and scores matrix The matrix $\boldsymbol{V}$ is called the *loadings matrix*, while the matrix $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{V} \in \mathbb{R}^{n \times d}$ is the *scores matrix*. The rows of $\boldsymbol{Z}$ are called the *principal components* (PCs).

PCA (mean-centred) If $\boldsymbol{X}$ is mean-centred, then the sample covariance matrix of $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{V}$ is $\boldsymbol{D}$.

PCA (correlation matrix) The PCA of $\boldsymbol{R}$ is equivalent to that of $\boldsymbol{S}$ when all the variances S_{ii} are the same.

Biplot A *biplot* is a plot that shows the PC scores together with vectors showing the PC loadings.

Scree plot The *scree plot* is the the visualization of the decreasing sequence of eigenvalues, scaled so that each bar is percentage of the total variance, that is, we plot

$$\frac{100D_i}{\mathrm{tr}(\boldsymbol{D})}, \quad 1 \leq i \leq d.$$

PCA via SVD The singular value decomposition of $\boldsymbol{X}$,

$$\boldsymbol{X} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{Q}^T,$$

where $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{Q} \in \mathbb{R}^{d \times d}$ are orthogonal matrices, and $\boldsymbol{\Lambda} \in \mathbb{R}^{n \times d}$ is a the diagonal matrix, is equivalent to the PCA of $\boldsymbol{S} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$ via

$$\boldsymbol{V} = \boldsymbol{Q}, \quad \boldsymbol{D} = \frac{1}{n-1}\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}, \quad \boldsymbol{Z} = \boldsymbol{X}\boldsymbol{Q} = \boldsymbol{P}\boldsymbol{\Lambda}.$$

Computational cost Note that to get $\boldsymbol{P}$ and $\boldsymbol{\Lambda}$, one can compute the e-value decomposition of $\boldsymbol{X}\boldsymbol{X}^T$ since

$$\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{P}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T)\boldsymbol{P}^T.$$

The computations costs are as follows:

- e-value decomposition of $\boldsymbol{X}^T\boldsymbol{X}$: $\mathcal{O}(d^3)$;
- e-value decomposition of $\boldsymbol{X}\boldsymbol{X}^T$: $\mathcal{O}(n^3)$;
- SVD of $\boldsymbol{X}$: $\mathcal{O}(nd^2)$.

Low-rank approximations Let $\lambda_1, \ldots, \lambda_r$ denote the r largest e-values of $\boldsymbol{S}$ with e-vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r$. Then,

$$\boldsymbol{X} \approx \sum_{i=1}^{r} \boldsymbol{X}\boldsymbol{w}_i\boldsymbol{w}_i^T,$$

is the best rank r-approximation to $\boldsymbol{X}$.

14 Clustering

k-means clustering Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be the matrix of n observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$. The idea is to find, for a given k, the k clusters $C_1, \ldots, C_k$ that minimize

$$\sum_{\ell=1}^{k} \frac{1}{|C_\ell|} \sum_{i,i' \in C_\ell} \|\boldsymbol{X}_i - \boldsymbol{X}_{i'}\|^2.$$

k-means algorithm

- Choose k.
- Randomly assign each observation to one of the clusters $C_1, \ldots, C_k$.
- Iterate the following 2 steps until the cluster assignments stop changing:
  - for each cluster compute the cluster mean,
  $$\boldsymbol{\mu}_\ell = \frac{1}{|C_\ell|} \sum_{i \in C_\ell} \boldsymbol{X}_i,$$
  - re-assign all observations to the cluster whose mean is closest (using Euclidean distance).

Agglomerative clustering It is a type of hierarchical clustering that avoids having to specify the number of clusters in advance.

- Begin with n observations and a measure of the pairwise dissimilarities $d_{i,j}$ for $1 \leq i \neq j \leq n$,
$$\boldsymbol{D}(n) = \begin{bmatrix} d_{2,1} & & & \\ d_{3,1} & d_{3,2} & & \\ \vdots & \vdots & \ddots & \\ d_{n,1} & d_{n,2} & \ldots & d_{n,n-1} \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n-1)}.$$
It is common to use Euclidean distance to measure dissimilarity (but other options exist).
- For $i = n, n-1, \ldots, 2$,
  - find the pair of clusters with the smallest dissimilarity, and fuse these two clusters;
  - compute the new dissimilarity matrix between the new fused cluster and all other $i-1$ remaining clusters, and create an updated matrix of dissimilarities $\boldsymbol{D}(n-1)$.

Linkage methods Computing the dissimilarity matrix requires to compute the distance between two clusters G and H.

- Single Linkage: $d(G, H) = \min_{i \in G, j \in H} d_{i,j}$.
- Complete Linkage: $d(G, H) = \max_{i \in G, j \in H} d_{i,j}$.
- Group average: $d(G, H) = \sum_{i \in G, j \in H} d_{i,j}/(|G||H|)$.

Dendograms The results of an agglomerative clustering of a dataset can be represented as dendrogram, which is a tree-like diagram the allows us to visualize the way in which the observations have been joined into clusters.