Notes of

Applied Statistics

From Prof. Piercesare Secchi's lectures
for the MSc in Mathematical Engineering

by Teo Bucci

Politecnico di Milano
A.Y. 2021/2022

LECTURE 2 22/02/2022

$\underline{X} = (X_1, X_2, \dots, X_p)' \in \mathbb{R}^p$ ← Feature space

units: individuals, times, places

$$\begin{array}{c} \\ 1 \\ \\ \\ n \end{array} \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ & & & \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

p: number of features

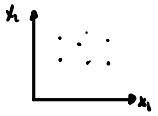
n: sample size

✕ DATA MATRIX
DATA FRAME $n \times p$

↳ ROW PERSPECTIVE

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1' \\ \underline{x}_2' \\ \vdots \\ \underline{x}_n' \end{bmatrix} \quad \underline{x}_i \in \mathbb{R}^p$$

↳ profile of unit i



↳ COLUMN PERSPECTIVE

$\mathbb{X} = [\underline{c}_1 \cdots \underline{c}_p] \quad \underline{c}_i \in \mathbb{R}^n$

↳ Sample of feature i

Y: special variable

GOAL: explain the variability of Y through the variability of $\mathbb{X}$

↳ SUPERVISED PROBLEMS both the features $\underline{X} = (\dots)$ and the target Y are obs. statistical unit

GENERAL PROBLEM OF STAT. LEARNING: supervised setting ? $f: \mathbb{R}^p \to \mathbb{R}$ s.t. $f(\underline{x})$ is the "best" explanation of Y in terms of $\underline{X}$

$$\operatorname*{argmin}_f \mathbb{E}\left[(Y - f(\underline{X}))^2\right]$$

PROP $\operatorname*{argmin}_f \mathbb{E}\left[|Y - f(\underline{x})|^2\right] = \mathbb{E}[Y|\underline{X}]$

proof

$$\mathbb{E}\left[|Y - f(\underline{x})|^2\right] = \mathbb{E}\left[\left[(Y - \mathbb{E}[Y|\underline{x}]) - (f(\underline{x}) - \mathbb{E}[Y|\underline{x}])\right]^2\right] \overset{!}{=}$$

$$= \mathbb{E}\left[(Y - \mathbb{E}[Y|\underline{x}])^2\right] + \mathbb{E}\left[(f(\underline{x}) - \mathbb{E}[Y|\underline{x}])^2\right] - 2\mathbb{E}\left[(Y - \mathbb{E}[Y|\underline{x}])(f(\underline{x}) - \mathbb{E}[Y|\underline{x}])\right]$$

$(*) = 0$

$\mathbb{E}[\mathbb{E}[W|Z]] = \mathbb{E}[W]$

$\mathbb{E}[W]$: does not depend on f

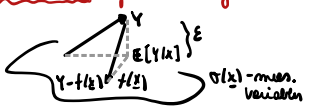
$\mathbb{E}[(f(\underline{x}) - \mathbb{E}[Y|\underline{x}])^2]$: take $f = \mathbb{E}[Y|\underline{x}]$

$$(*) = \mathbb{E}\left[\mathbb{E}\left[(Y - \mathbb{E}[Y|\underline{x}])(f(\underline{x}) - \mathbb{E}[Y|\underline{x}]) \,|\, \underline{X}\right]\right] = 0$$

$\mathbb{E}[Y|\underline{x}] - \mathbb{E}[Y|\underline{x}] = 0$

■

SIGMA-FIELD generated by $\underline{X}$: when you know $\underline{X}$, you know everything that depends on $\underline{X}$ (SIGMA-SPACE)



$Y = \mathbb{E}[Y|\underline{X}] + \varepsilon$

↳ unknown bound on the info carried by $\underline{X}$

↳ BASIC MODEL FOR S.L.

$Y = f(\underline{x}) + \varepsilon$

$f(\underline{x}) = \mathbb{E}[Y|\underline{x}]$

ε r.v. s.t. $\mathbb{E}[\varepsilon] = 0$

ε indep of $\underline{X}$, something I can't explain in term of $\underline{x}$

$\mathrm{Var}(\varepsilon) = \mathbb{E}[\varepsilon^2] - \overbrace{\mathbb{E}^2[\varepsilon]}^{=0} = \mathbb{E}[\varepsilon^2]$

$\mathbb{X} \xrightarrow{\exists} \hat{f}$ estimate of f

↳ known once $\mathbb{X}$ are known

LECTURE 3 24/02/2022

SUPERVISED LEARNING

Basic model

$\mathbb{R} \ni Y = f(\underline{x}) + \varepsilon \quad \underline{X} \in \mathbb{R}^p, \varepsilon \in \mathbb{R}, \mathbb{E}[\varepsilon] = 0$

$f: \mathbb{R}^p \to \mathbb{R} \qquad \varepsilon \perp\!\!\!\perp \underline{X} \qquad \mathbb{E}[Y|\underline{x}] = f(\underline{x})$

GOAL use data to learn f: generate an estimate $\hat{f}$ of f

totally known when $\mathbb{X}$ is known

$\hat{f}: \mathbb{X} \to \hat{f}$

Q: Properties of $\hat{f}$?

Q: Variability of $\hat{f}$?

"Not so large"

$\hat{f} \in \{ f_\theta : \theta \in \mathbb{R}^q \}$ ⇒ Use X to estimate $\hat{\theta}$ so that $\hat{f}(x) = f_{\hat{\theta}}(x)$

↳ We approximate f with a member from this family close enough ($f \approx f_\theta$)

Ex

$f_\theta(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ — LINEAR

$\theta = (\beta_0, \dots, \beta_p) \in \mathbb{R}^{p+1}$

$X \to \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$

↳ Then variables are NOT necessarily the ones received with the dataset

Received: $w_1, \dots, w_K$

Variables: $x_1 = h_1(w_1, \dots, w_K)$ ⋮ $x_K = h_K(w_1, \dots, w_K)$

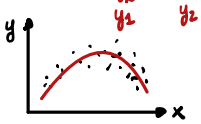
⇒ $f_\theta(w_1, \dots, w_K) = \beta_0 + \beta_1 h_1(w_1 \dots w_K) + \dots + \beta_p h_p(w_1 \dots w_K)$

Ex

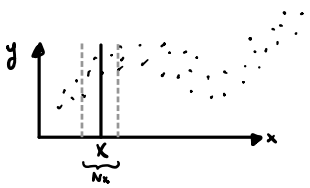
$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 \log x$ — still LINEAR as long as $h_{1 \dots K}$ are known

$y_1 = x$, $y_2 = x^2$, $y_3 = \log x$

↳ this is when PRIOR ENGINEERING / DOMAIN knowledge kicks in! ☺



EX (KNN - K Nearest Neighbors)

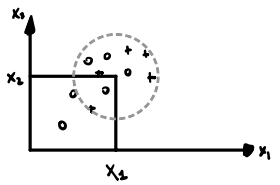


? $\hat{f}$

$N_x(k) = \{ k \text{ closest } x_i \in X \text{ to } x \}$

$\hat{f}(x) = \frac{1}{k} \sum_{x_i \in N_x(k)} Y_i$ $\quad \text{Var}(\hat{f}(x)) = \frac{1}{k} \text{Var}(\varepsilon)$

Ex (KNN for classification)



$\hat{f}(x) = 0$

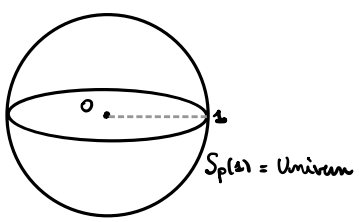
$N_x(k) = \{ k \text{ closest (distance?) } x_i \in X \text{ to } x \}$

$\hat{f}(x) = \begin{cases} 0 & \#\{ Y_i = 0, x_i \in N_x(k) \} > \#\{ Y_i = +, x_i \in N_x(k) \} \\ + & \text{otherwise} \end{cases}$

Curse of dimensionality

↳ There is a "continuity" idea, f needs to be regular enough

↳ If p is large: nothing's around

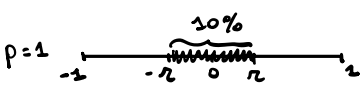


$S_p(r)$ = sphere of radius r in $\mathbb{R}^p$ centered in 0

Friends are uniformly distributed in $S_p(1)$

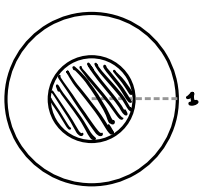
Q: How far do you have to go to reach 10% of your friends?

$p = 1$



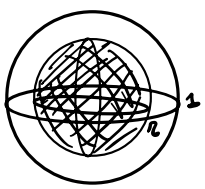
$\frac{\mu(S_1(1))}{\mu(S_1(1))} = 0{,}1 = \frac{2r}{2} = r \Rightarrow r = 0{,}1$

$p = 2$



$0{,}1 = \frac{\mu(S_2(r))}{\mu(S_2(1))} = \frac{\pi r^2}{\pi} = r^2 \Rightarrow r = \sqrt{0{,}1} \simeq 0{,}31$ You have to travel 30% MORE!

$p = 3$



$0{,}1 = \frac{\mu(S_3(r))}{\mu(S_3(1))} = \frac{\frac{4}{3}\pi r^3}{\frac{4}{3}\pi} = r^3 \Rightarrow r = \sqrt[3]{0{,}1} = 0.46$

$p = 100 \Rightarrow r = (0,1)^{1/100} = 0,97$

↳ $p = 1$ $n = 100$) Same information

$p = 50$ $n = (100)^{50}$ but in the real universe there are $\approx 10^{82}$ atoms ☹

↳ Too much!

How to tackle the curse?

↳ Use structured models

↳ reduce dimensionality, PCA, ICA

How to evaluate $\hat{f}$?

Error when predicting Y for a new unit $(Y_0, \underline{X}_0)$ ↳ to be predicted ↳ known

test error — Random — $Y_0 = f(\underline{X}_0) + \varepsilon_0$

$$\mathbb{E}_{\mathcal{X}}\left[\left(Y_0 - \hat{f}(\underline{X}_0)\right)^2\right] = \mathbb{E}_{\mathcal{X}}\left[\left(f(\underline{X}_0) - \hat{f}(\underline{X}_0) + \varepsilon_0\right)^2\right]$$

↳ expectation when $\mathcal{X}$ is known

$\mathbb{E}[\varepsilon_0] = 0$ not random

$$= \mathbb{E}_{\mathcal{X}}\left[\left(f(\underline{X}_0) - \hat{f}(\underline{X}_0)\right)^2 + \varepsilon_0^2 + 2\varepsilon_0\left(f(\underline{X}_0) - \hat{f}(\underline{X}_0)\right)\right]$$

$$= \underbrace{\left(f(\underline{X}_0) - \hat{f}(\underline{X}_0)\right)^2}_{\text{this the part we can reduce}} + \underbrace{\mathrm{Var}(\varepsilon_0)}_{\text{No control on this irreducible}}$$

→ run the model ON the data ("TRAINING ERROR")

↳ poor estimate of the error, will dev. be small

→ use TEST-DATA set known but which you didn't train the model on

↳ is it really a test-data set?

→ Cross-validation

$$\mathbb{E}\left[\mathbb{E}_{\mathcal{X}}\left[\left(Y_0 - \hat{f}(\underline{X}_0)\right)^2\right]\right] = \mathbb{E}\left[\left(f(\underline{X}_0) - \hat{f}(\underline{X}_0)\right)^2\right] + \mathrm{Var}(\varepsilon_0)$$

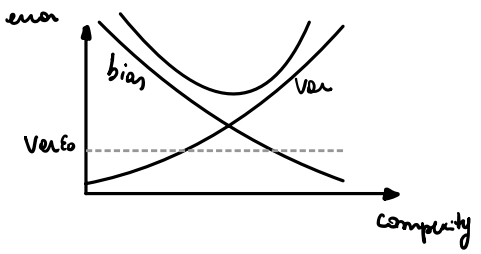
$$= \mathbb{E}\left[\left(f(\underline{X}_0) + \mathbb{E}[\hat{f}(\underline{X}_0)] - \mathbb{E}[\hat{f}(\underline{X}_0)] - \hat{f}(\underline{X}_0)\right)^2\right] + \mathrm{Var}(\varepsilon_0)$$

$$= \left(f(\underline{X}_0) - \mathbb{E}[f(\underline{X}_0)]\right)^2 + \mathbb{E}\left[\left(\mathbb{E}[\hat{f}(\underline{X}_0)] - \hat{f}(\underline{X}_0)\right)^2\right]$$

$$+ 2\,\mathbb{E}\left[\left(f(\underline{X}_0) - \mathbb{E}[\hat{f}(\underline{X}_0)]\right)\left(\mathbb{E}[\hat{f}(\underline{X}_0)] - \hat{f}(\underline{X}_0)\right)\right] + \mathrm{Var}\,\varepsilon_0$$

BIAS-VARIANCE TRADE OFF

$$= \underbrace{\left(f(\underline{X}_0) - \mathbb{E}[\hat{f}(\underline{X}_0)]\right)^2}_{\text{bias}} + \underbrace{\mathrm{Var}(\hat{f}(\underline{X}_0))}_{\text{variance}} + \mathrm{Var}(\varepsilon_0)$$



EX kNN

$$\hat{f}(\underline{x}) = \frac{1}{k}\sum_{\underline{x}_i \in N_k(\underline{x})} Y_i \qquad \mathrm{Var}\left(\hat{f}(\underline{x})\right) = \frac{1}{k}\mathrm{Var}(\varepsilon) \downarrow \quad k \uparrow$$

but bias ↑ k ↑

Read Chap 2 ISS (Intro to Stat Learning)

LECTURE 4 1/3/2022

$S_{12} = \frac{1}{m}\sum_{i=0}^{m}(x_{i2} - \bar{x}_1)(x_{i2} - \bar{x}_2)$ where $\bar{\underline{x}} = \begin{bmatrix}\bar{x}_1 \\ \vdots \\ \bar{x}_p\end{bmatrix}$ vector of sample means

$S = \begin{bmatrix} s_{11} & & \\ & \ddots & \\ & & s_{pp}\end{bmatrix}$ $p \times p$ symmetric matrix of real numbers (sample covariance matrix)

↳ Unit measures is a problem ⇒ Standardize! $r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}\cdot s_{kk}}} \in [-1,1]$

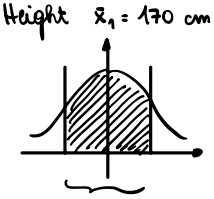
$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ & \ddots & & \\ & & & 1\end{bmatrix}$ correlation matrix

still $p \times p$ Sym. of real numbers

namely $x_j \to \frac{x_j - \bar{x}}{\sqrt{s_{jj}}} = z_j$ $\quad \mathrm{Cov}(z_j, z_k) = \frac{1}{m}\sum_{i=0}^{m}(z_{ij} - \bar{z}_j)(z_{ik} - \bar{z}_k) = r_{jk}$ $j,k = 1 \dots p$

EX

Height $\bar{x}_1 = 170$ cm, $\sqrt{s_{11}} = 3$ cm

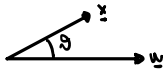


Chebychev: $\mathbb{P}\left[x \in \left[\bar{x}_1 - k\sqrt{s_{11}}, \bar{x}_1 + k\sqrt{s_{11}}\right]\right] \geq 1 - \frac{1}{k^2}$ $\quad k = 2$

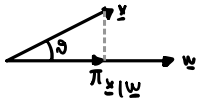
± 2 std dev	75%	Chebychev
± 2 std dev	97%	Gaussian

"Geometry" in $\mathbb{R}^m$

$\underline{x}, \underline{w} \in \mathbb{R}^m$, $\langle \underline{x}, \underline{w} \rangle = \underline{x}'\underline{w}$ $\quad \|\underline{x}\| = \sqrt{\langle \underline{x}, \underline{x} \rangle} = \sqrt{\sum_{i=1}^m x_i^2}$



$$\cos\vartheta = \frac{\langle \underline{x}, \underline{w} \rangle}{\|\underline{x}\| \, \|\underline{w}\|} = \frac{\sum x_i w_i}{\sqrt{\sum x_i^2 \sum w_i^2}}$$



$\pi_{\underline{x}|\underline{w}}$ = operator that projects

$$\pi_{\underline{x}|\underline{w}} = \|\underline{x}\| \cos\vartheta \cdot \frac{\underline{w}}{\|\underline{w}\|} \quad (\underline{w} \text{ unitary vector for to } \underline{w}) = \|\underline{x}\| \frac{\underline{x}'\underline{w}}{\|\underline{x}\| \, \|\underline{w}\|} \cdot \frac{\underline{w}}{\|\underline{w}\|} = \frac{\underline{x}'\underline{w}}{\underline{w}'\underline{w}} \underline{w} = \left(\frac{\underline{w}\underline{w}'}{\underline{w}'\underline{w}} \right) \underline{x}$$

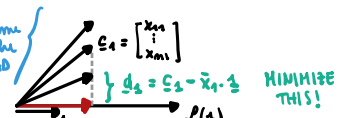
Data

$\mathcal{L}(\underline{1}) = \{ (\overbrace{c, \dots, c}^{n \text{ comp.}}) : c \in \mathbb{R} \}$

$\underline{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$

No variability in this space

They all have the same MEAN but the approx is BAD for many of them



$\underline{c}_1 = \begin{bmatrix} x_{11} \\ \vdots \\ x_{m1} \end{bmatrix}$, $\underline{d}_1 = \underline{c}_1 - \bar{x}_1 \cdot \underline{1}$ MINIMIZE THIS! $\quad \underline{d}_1 \in \mathcal{L}^\perp(\underline{1})$ dim $= m-1$

$\pi_{\underline{c}_1|\underline{1}} = \frac{\underline{1}\underline{1}'}{\underline{1}'\underline{1}} \underline{c}_1 = \frac{1}{m} \sum_{i=1}^m x_{i1} \cdot \underline{1} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_1 \end{pmatrix}$ Best approx on $\mathcal{L}(\underline{1})$

$\|\underline{d}_1\| = \sqrt{\sum d_i^2} = \sqrt{\sum (x_{i1} - \bar{x}_1)^2}$

$\Rightarrow \|\underline{d}_1\|^2 = m S_m$

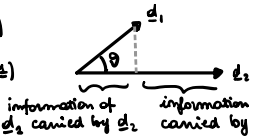
"euros to dollars"

$\vartheta = 0 \Rightarrow \underline{d}_1 \in \mathcal{L}(\underline{d}_2)$ $\quad \underline{d}_1 = k \underline{d}_2$ Same information carried

$\vartheta = \frac{\pi}{2} \Rightarrow$ nothing in $\underline{d}_2$ carries info. about $\underline{d}_1$ and viceversa

Consider $\underline{d}_1 = \underline{c}_1 - \bar{x}_1 \cdot \underline{1} \in \mathcal{L}^\perp(\underline{1})$

$\underline{d}_2 = \underline{c}_2 - \bar{x}_2 \cdot \underline{1} \in \mathcal{L}^\perp(\underline{1})$



$$\cos\vartheta = \frac{\underline{d}_1 \cdot \underline{d}_2}{\|\underline{d}_1\| \, \|\underline{d}_2\|} = \frac{\frac{1}{m} \sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum (x_{i1} - \bar{x}_1)^2 \sum (x_{i2} - \bar{x}_2)^2}} = r_{12} \in [-1, 1]$$ 🙂

Data

$\mathbb{X} = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_n' \end{bmatrix}$ $\quad \underline{x}_i \in \mathbb{R}^p$ $\quad \underline{x}_1, \dots, \underline{x}_n$ iid realizations of $\underline{X}$ random vector of $\mathbb{R}^p$

$\mathbb{P}[\underline{X} \in B] = \nu_{\underline{X}}(B) \quad \forall B \in \mathcal{B}(\mathbb{R}^p)$ ↳ Borel

It might happen $\nu_{\underline{X}}(B) = \int_B f(\underline{x}) \, d\underline{x}$ ↳ density

Ex

$$f(\underline{x}) = \frac{1}{\sqrt{(2\pi)^p |\det \Sigma|}} \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{\mu})' \Sigma^{-1} (\underline{x} - \underline{\mu}) \right\} \qquad \Sigma \in p \times p \text{ pos. def.}, \ \underline{\mu} \in \mathbb{R}^p$$

$\mathbb{E}[\underline{X}] = \underline{\mu}$ $\quad \mathbb{E}[(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})'] = \Sigma = [\sigma_{ij}]$ $\quad i, j = 1, \dots, p$ $\quad \sigma_{ij} = \text{Cov}(X_i, X_j)$

↳ $\underline{c} \in \mathbb{R}^p$ $\quad \mathbb{E}[\underline{c}'\underline{x}] = \mathbb{E}[\sum c_i x_i] = \sum c_i \mathbb{E}[X_i] = \underline{c}'\underline{\mu}$

$(\mathbb{R} \ni) \text{Var}[\underline{c}'\underline{x}] = \underline{c}'\Sigma\underline{c} \ (\in \mathbb{R})$

$\text{Var}(c_1 X_1 + c_2 X_2) = c_1^2 \text{Var} X_1 + c_1^2 \text{Var} X_2 + 2 c_1 c_2 \text{Cov}(X_1, X_2)$

↳ $C \in k \times p$ $\quad \mathbb{E}[C\underline{x}] = C\underline{\mu}$

$\text{Var}(C\underline{x}) = C \Sigma C'$

↳ Methods of moments

$\hat{\underline{\mu}} = \frac{1}{m} \sum_{i=1}^m \underline{x}_i = \bar{\underline{x}}$ $\quad \hat{\Sigma} = \frac{1}{m} \sum_{i=1}^m (\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})' = S_m$

- What are the properties of these two machines?

① $\mathbb{E}[\bar{\underline{x}}] = \underline{\mu}$ (umbiased)

$\text{Cov}(\bar{\underline{x}}) = \frac{1}{m} \Sigma$

② $\mathbb{E}[S_m] = \frac{m-1}{m} \Sigma$ not umbiased $\rightsquigarrow$ $S = \frac{1}{m-1} \sum_{i=1}^m (\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})'$ is umbiased

LECTURE 5 3/3/2022

$\mathbb{X} = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_n' \end{bmatrix} = [\underline{c}_1 \dots \underline{c}_p]$

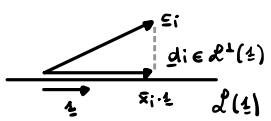
$\underline{x}_i \in \mathbb{R}^p$, $\underline{c}_i \in \mathbb{R}^n$

↳ realizations of $\underline{X}_1 \dots \underline{X}_n$ iid $\sim \underline{\mu}, \Sigma$ $\quad$ estimated by $\bar{\underline{x}} = \frac{1}{n} \sum_{i=1}^n \underline{x}_i$ for $\underline{\mu}$

$S_n = \frac{1}{n} \sum_{i=1}^n (\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})'$ for Σ

Biased ↳ $\frac{1}{n-1}$ umbiased

→ Deviation vector $\underline{d}_i$



$$\underline{d}_i = \underline{c}_i - \bar{x}_i \cdot \underline{1} \overset{!}{=} \underline{c}_i - \frac{\underline{1}\underline{1}'}{\underline{1}'\underline{1}}\underline{c}_i = \left(I - \frac{\underline{1}\underline{1}'}{\underline{1}'\underline{1}}\right)\underline{c}_i$$

Orthogonal projection on $\mathcal{L}^\perp(\underline{1})$

$D = [\underline{d}_1 \dots \underline{d}_p]$ matrix of dev. vectors

$$\overset{!}{=} \left(I - \frac{\underline{1}\underline{1}'}{\underline{1}'\underline{1}}\right)X$$

$$S = \frac{1}{m-1}\begin{bmatrix} \Sigma(x_{i1}-\bar{x}_1)^2 & \Sigma(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)\cdots \\ \vdots & \ddots \end{bmatrix} = \frac{1}{m-1}\begin{bmatrix} \underline{d}_1'\underline{d}_1 & \underline{d}_1'\underline{d}_2 \cdots & \underline{d}_1'\underline{d}_p \\ \vdots & \ddots & \vdots \\ \underline{d}_p'\underline{d}_1 & \cdots & \underline{d}_p'\underline{d}_p \end{bmatrix} = \frac{1}{m-1} D'D = \frac{1}{m-1} \underbrace{X'\left(I - \frac{\underline{1}\underline{1}'}{\underline{1}'\underline{1}}\right)'}_{D'}\underbrace{\left(I - \frac{\underline{1}\underline{1}'}{\underline{1}'\underline{1}}\right)X}_{D}$$

$$\Rightarrow \boxed{S = \frac{1}{m-1} X'\left(I - \frac{\underline{1}\underline{1}'}{\underline{1}'\underline{1}}\right)X}$$

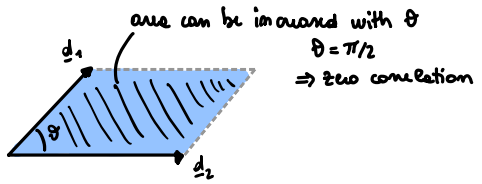
↳ Generalized variance : $\det(S)$ (sample)
$\det(\Sigma)$ (population)

↳ Total variance : $\operatorname{tr}(S)$ (sample)
$\operatorname{tr}(\Sigma)$ (population)

$\begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}$ $\det 1$, $\operatorname{tr} p$ } REC

Ex

$p=2$ $\quad S = \frac{1}{m-1}\begin{bmatrix} \underline{d}_1'\underline{d}_1 & \underline{d}_1'\underline{d}_2 \\ \underline{d}_2'\underline{d}_1 & \underline{d}_2'\underline{d}_2 \end{bmatrix} = \frac{1}{m-1}\begin{bmatrix} \|\underline{d}_1\|^2 & \|\underline{d}_1\|\|\underline{d}_2\|\cos\vartheta \\ \|\underline{d}_1\|\|\underline{d}_2\|\cos\vartheta & \|\underline{d}_2\|^2 \end{bmatrix}$



$$\text{Gen. Var}(S) = \det(S) = \frac{1}{(m-1)^2}\left[\|\underline{d}_1\|^2\|\underline{d}_2\|^2 - \|\underline{d}_1\|^2\|\underline{d}_2\|^2\cos^2\vartheta\right] = \frac{1}{(m-1)^2}\left[\|\underline{d}_1\|^2\|\underline{d}_2\|^2\sin^2\vartheta\right] \propto \text{Area}^2(\text{parallelogram}(\underline{d}_1,\underline{d}_2))$$

PROP $\det(S) = 0 \iff \underline{d}_1 \dots \underline{d}_p$ are linearly dependent

proof ($\Leftarrow$) $\exists\, \underline{c} \neq \underline{0},\ \underline{c} \in \mathbb{R}^p \quad D\underline{c} = \underline{0}$, $\sum \underline{d}_i c_i = 0$

$S = \frac{1}{m-1} D'D \qquad S\underline{c} = \frac{1}{m-1} D'\underbrace{D\underline{c}}_{=\underline{0}} = \underline{0} \Rightarrow$ cols of S are lin. dep $\Rightarrow \det(S) = 0$

($\Rightarrow$) $\exists\, \underline{c} \in \mathbb{R}^p,\ \underline{c} \neq \underline{0}$ s.t. $S\underline{c} = \underline{0} \Rightarrow \frac{1}{m-1} D'D\underline{c} = \underline{0} \Rightarrow \frac{1}{m-1}\underbrace{\underline{c}'D'D\underline{c}}_{\|D\underline{c}\|^2} = 0 \Rightarrow \|D\underline{c}\|^2 = 0 \Rightarrow \underline{d}_1 \dots \underline{d}_p$ lin. dep

Hence if $\det(S) = 0 \quad D\underline{c} = \underline{0} \quad \underline{0} \neq \underline{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix}$ WLOG $c_1 \neq 0 \quad \sum_{i=1}^p c_i \underline{d}_i = \underline{0} \Rightarrow \underline{d}_1 = -\sum_{i=2}^p \frac{c_i}{c_1}\underline{d}_i$

$\Rightarrow x_1 - \bar{x}_1 = -\sum_{i=2}^p \frac{c_i}{c_1}(x_i - \bar{x}_i)$ ■

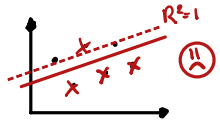
$\det(S) = 0$ never happens ⇝ $\det(S) \approx 0$ is enough
↳ collinearity between variables (TROUBLE!)

X $m \times p$ matrix

Corollary If $p \geq m \Rightarrow \det(S) = 0$
more features than observations

proof $\underline{d}_1 \dots \underline{d}_p \in \mathcal{L}^\perp(\underline{1})$ ← at most $m-1$ systems of linear indep. vectors
$\dim(\mathcal{L}^\perp(\underline{1})) \leq m-1$

if $p \geq m > m-1 \Rightarrow \underline{d}_1 \dots \underline{d}_p$ are lin. dep.



PROP S is positive semi-def

If $\det(S) \neq 0 \Rightarrow S$ is positive def

proof We need to prove: $\forall c \in \mathbb{R}^p \quad c' S c \geq 0$

But $\frac{1}{n-1} c' d' d c = \frac{1}{n-1} \|dc\|^2 \geq 0$

If S is not pos. def $\Rightarrow \exists c \neq 0$ s.t. $c'Sc \overset{\text{equal}}{=} 0 \Rightarrow \frac{1}{n-1} c'd'dc = 0 \Rightarrow \|dc\|^2 \Rightarrow dc = 0 \Rightarrow d_1 \dots d_p$ lin. dep

$\Rightarrow \det(S) = 0$ ↯ ■

REMARK

S symmetric, real-valued $\Rightarrow \exists (\lambda_i, e_i) \in \mathbb{R} \times \mathbb{R}^p$ s.t. $S = \sum_{i=1}^{p} \lambda_i e_i e_i'$

moreover $e_i' e_j = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$

thus $\{e_1 \dots e_p\}$ orthonormal system for $\mathbb{R}^p$

They satisfy $S e_i = \lambda_i e_i \quad i = 1 \dots p$

Also, since it's positive semidef we can order them: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ (convention)

(> if positive def.)

Notation

$P = [e_1 \dots e_p] \Rightarrow \boxed{S = P \Lambda P'}$ $\quad \Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix}$

If $\det(S) \neq 0 \Rightarrow S^{-1} = \sum \frac{1}{\lambda_i} e_i e_i' = P \Lambda^{-1} P$

If $\det(S) = 0 \Rightarrow \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k > \underbrace{\lambda_{k+1} = 0 = \dots = \lambda_p}_{\text{ignore them}} \Rightarrow S^- = \sum_{i=1}^{k} \frac{1}{\lambda_i} e_i e_i'$

If $\det(S) \neq 0$: $\boxed{d_{S^{-1}}(x, y) = \sqrt{(x-y)' S^{-1} (x-y)}}$ Mahalanobis distance

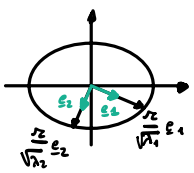
If $S = I$ this is the Euclidean distance

$S_r(\bar{x}) = \{x \in \mathbb{R}^p : (x-\bar{x})'(x-\bar{x}) \leq r^2\}$

$\mathcal{E}_c(\bar{x}) = \{x \in \mathbb{R}^p : (x-\bar{x})' S^{-1} (x-\bar{x}) \leq r^2\}$ ⇝ Ellipse

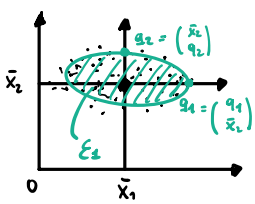
B $p \times p$ pos. def $x \in \mathbb{R}^p : x' B x = 1$

$B = \Sigma \lambda_i e_i e_i'$



Ex

$p=2 \quad S = \begin{bmatrix} s_{11} & 0 \\ 0 & s_{22} \end{bmatrix} \quad s_{11} > s_{22}$



$d_{eucl}(q_1, \bar{x}) = d_{eucl}(q_2, \bar{x})$

↳ very little significative

↳ standardize: $d_{eucl}(\text{Std}(q_1), \text{Std}(\bar{x})) = d_{eucl}\left(\begin{pmatrix} \frac{q_1 - \bar{x}_1}{\sqrt{s_{11}}} \\ 0 \end{pmatrix}, 0\right) = \frac{|q_1 - \bar{x}_1|}{\sqrt{s_{11}}}$

$d_{eucl}(\text{Std}(q_2), \text{Std}(\bar{x})) = d_{eucl}\left(\begin{pmatrix} 0 \\ \frac{q_2 - \bar{x}_2}{\sqrt{s_{22}}} \end{pmatrix}, 0\right) = \frac{|q_2 - \bar{x}_2|}{\sqrt{s_{22}}}$

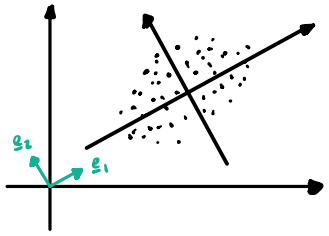
IDEA

The euclidean distance is right after standardization

$$d_{eucl}(\text{Std}(q), \text{Std}(\bar{x})) = d_{eucl}\left(\begin{pmatrix} \frac{q_1 - \bar{x}_1}{\sqrt{s_{11}}} \\ \frac{q_2 - \bar{x}_2}{\sqrt{s_{22}}} \end{pmatrix}, 0\right) = \sqrt{\left(\frac{q_1 - \bar{x}_1}{\sqrt{s_{11}}}\right)^2 + \left(\frac{q_2 - \bar{x}_2}{\sqrt{s_{22}}}\right)^2}$$

$\text{Vol}(\mathcal{E}_c(\bar{x})) = k_p \sqrt[p]{\prod_{i=1}^{p} \pi r_i} = \alpha \sqrt{\det(S)}$ $\quad k_p$: $p=2 \ \pi$, $p=3 \ \frac{3}{3}\pi$

What if... Ex S not diagonal

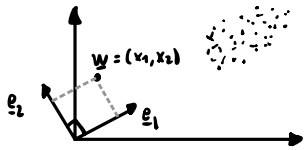


LECTURE 6 8/3/2022

S cov. matrix

$\det(S) > 0$

$S = \sum_{i=1}^{p} \lambda_i e_i e_i' = P \Lambda P'$ $\quad P = [e_1 \dots e_p]$, $\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}$



New system $\tilde{w} = \begin{pmatrix} e_1' w \\ e_2' w \\ \vdots \\ e_p' w \end{pmatrix} = P' w$

$\|w\|^2 = w'w = \tilde{w}'\tilde{w} = w' P' P w$

The data $\mathbb{X}$ becomes: $\tilde{\mathbb{X}} = \begin{bmatrix} (P'x_1)' \\ \vdots \\ (P'x_m)' \end{bmatrix} = \mathbb{X}P$

The covariance becomes $S = \frac{1}{m-1} \mathbb{X}' \left(I - \frac{1 1'}{1'1} \right) \mathbb{X}$

$\tilde{S} = \frac{1}{m-1} \tilde{\mathbb{X}}' \left(I - \frac{1 1'}{1'1} \right) \tilde{\mathbb{X}} \overset{!}{=} \frac{1}{m-1} P' \mathbb{X}' \left(I - \frac{1 1'}{1'1} \right) \mathbb{X} P = P' S P \overset{!}{=} P' P \Lambda P' P = \Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix}$ $\quad S = P \Lambda P'$

PCA: PRINCIPAL COMPONENT ANALYSIS

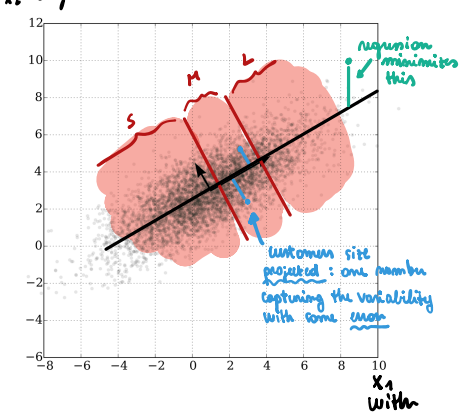
Opt. problem: does $\exists\, a^* \in \mathbb{R}^p$ s.t. $\mathrm{Var}(a^* x) = \sup_{a \in \mathbb{R}^p, \|a\|=1} \mathrm{Var}(a x)$?

Not well posed: if a^* is solution $\Rightarrow$ $\mathrm{Var}(10\, a^* x) = 100\, \mathrm{Var}(a^* x) > \mathrm{Var}(a^* x)$

$\Rightarrow a^*$ not solution

We ask $\|a\| = 1$

↳ a is supposed to be a direction



REMARK: this is not fitting the regression line where we want to predict one variable in term of the other, here we are looking for the subspace which is closest (best approximate) the data.

PROP B pos. semidef. $p \times p$ $\left(B = \sum_{i=1}^{p} \lambda_i e_i' e_i\right)$, then

1) $\sup_{a \in \mathbb{R}^p} \frac{a' B a}{a' a} = \lambda_1$ $\quad \operatorname{argmax} x = e_1$

2) $\sup_{\substack{a \in \mathbb{R}^p \\ a \perp e_1}} \frac{a' B a}{a' a} = \lambda_2$

$\vdots$

p) $\sup_{a \in \mathbb{R}^p} \frac{a' B a}{a' a} = \lambda_p = \inf_{a \in \mathbb{R}^p} \frac{a' B a}{a' a}$ $\quad \operatorname{argmax} x = e_p$

$a \perp e_1 \dots e_{p-1}$

proof $\frac{a' B a}{a' a} = \frac{a' P \Lambda P' a}{a' a} = \frac{y' \Lambda y}{y' y} = \frac{\sum_{i=1}^{p} \lambda_i y_i^2}{\sum y_i^2} \leq \lambda_1 \frac{\sum y_i^2}{\sum y_i^2} = \lambda_1$

but if $a = e_1$: $\frac{e_1' B e_1}{e_1' e_1} = \frac{\lambda_1 e_1' e_1}{e_1' e_1} = \lambda_1$ the sup is reached

step 2) $a \perp e_1$ $\frac{a' B a}{a' a} = \frac{a' P \Lambda P' a}{a' a} = \frac{y' \Lambda y}{y' y} = \frac{\sum_{i=2}^{p} \lambda_i y_i^2}{\sum_{i=2}^{p} y_i^2} \leq \lambda_2 \frac{\sum_{i=2}^{p} y_i^2}{\sum_{i=2}^{p} y_i^2} = \lambda_2$

$$\underline{y} = P'\underline{a} = \begin{bmatrix} \underline{e}_1' \\ \underline{e}_2 \\ \vdots \\ \underline{e}_p' \end{bmatrix} \underline{a} = \begin{bmatrix} 0 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}$$

but if $\underline{a} = \underline{e}_2$ $(\perp \underline{e}_1)$: $\dfrac{\underline{e}_2' B \underline{e}_2}{\underline{e}_2'\underline{e}_2} = \dfrac{\lambda_2 \underline{e}_2'\underline{e}_2}{\underline{e}_2'\underline{e}_2} = \lambda_2$

step p) $\dfrac{\underline{a}' B \underline{a}}{\underline{a}'\underline{a}} = \dfrac{\underline{a}' P \Lambda P' \underline{a}}{\underline{a}'\underline{a}} = \dfrac{\underline{y}' \Lambda \underline{y}}{\underline{y}'\underline{y}} = \dfrac{\sum \lambda_i y_i^2}{\sum y_i^2} \geq \lambda_p \dfrac{\sum y_i^2}{\sum y_i^2} = \lambda_p$ ■

Back to PCA

$$\sup_{\substack{\underline{a} \in \mathbb{R}^p \\ \|\underline{a}\| = 1}} \mathrm{Var}(\underline{a}'\underline{x}) = \sup_{\underline{a} \in \mathbb{R}^p} \frac{\underline{a}'\Sigma\underline{a}}{\underline{a}'\underline{a}} = \lambda_1 \qquad \Sigma = P\Lambda P' = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i'$$

We want $\mathrm{Cov}(\underline{a}'\underline{x}, \underline{e}_1'\underline{x}) \ominus 0$ → $\mathrm{Cov}(\underline{a}'\underline{x}, \underline{e}_1'\underline{x}) = \underline{a}'\Sigma\underline{e}_1 = \lambda_1 \underline{a}'\underline{e}_1 = 0 \Leftrightarrow \underline{a}'\underline{e}_1 = 0 \Leftrightarrow \boxed{\underline{a} \perp \underline{e}_1}$

In general $\mathrm{Cov}(\underline{a}'\underline{x}, \underline{b}'\underline{x}) = \underline{a}'\Sigma\underline{b}$

$$C = \begin{bmatrix} \underline{a}' \\ \underline{b}' \end{bmatrix} \quad \mathrm{Cov}(C\underline{x}) = C\Sigma C' = \begin{bmatrix} \underline{a}' \\ \underline{b}' \end{bmatrix} \Sigma \begin{bmatrix} \underline{a} & \underline{b} \end{bmatrix} = \begin{bmatrix} \underline{a}'\Sigma\underline{a} & \underline{a}'\Sigma\underline{b} \\ \underline{b}'\Sigma\underline{a} & \underline{b}'\Sigma\underline{b} \end{bmatrix}$$

So $\sup_{\substack{\underline{a} \in \mathbb{R}^p \\ \|\underline{a}\| = 1 \\ \underline{a} \perp \underline{e}_1}} \mathrm{Var}(\underline{a}'\underline{x}) = \lambda_2$, In general : $\sup_{\substack{\underline{a} \in \mathbb{R}^p \\ \|\underline{a}\| = 1 \\ \mathrm{Cov}(\underline{a}'\underline{x}, \underline{e}_j'\underline{x}) = 0 \ \forall j = 1 \dots k}} \mathrm{Var}(\underline{a}'\underline{x}) = \lambda_{k+1}$

DEF For $k = 1 \dots p$, $\underline{e}_k$ identifies the direction of the k-th principal component

$Y_k = \underline{e}_k'\underline{x}$ $(Y_k = \underline{e}_k'(\underline{x} - \underline{\mu}))$

$Y_k = e_{1k}x_1 + e_{2k}x_2 + \dots + e_{pk}x_p$ (SCORE: Y_k; LOADINGS: $e_{1k}, e_{2k}, \dots, e_{pk}$)

of it will be $\underline{Y} = P'(\underline{X} - \underline{\mu})$ so that $\mathbb{E}[\underline{Y}] = \underline{0}$

$\mathrm{Cov}(\underline{Y}) = P'\Sigma P = P'P\Lambda P'P = \Lambda$

PROP $\mathrm{Cor}(Y_k, X_i) = e_{ik} \cdot \dfrac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}}$ $\qquad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ & \ddots \\ & & \sigma_{pp} \end{bmatrix}$

proof $\mathrm{Cor}(Y_k, X_i) = \dfrac{\mathrm{Cov}(Y_k, X_i)}{\sqrt{\lambda_k}\sqrt{\sigma_{ii}}} = \dfrac{\lambda_k e_{ik}}{\sqrt{\lambda_k}\sqrt{\sigma_{ii}}} = e_{ik}\dfrac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}}$

$\mathrm{Cov}(Y_k, X_i) = \mathrm{Cov}(\underline{e}_k'\underline{x}, \underbrace{\underline{u}_i'}_{(0 \dots 0 1 0 \dots 0)}\underline{x}) = \underline{e}_k'\Sigma\underline{u}_i = \underline{u}_i'\Sigma\underline{e}_k = \lambda_k \underline{u}_i'\underline{e}_k = \lambda_k e_{ik}$

Is variability between variables VERY different?

- ↳ NO ⇒ PCA
- ↳ YES ⇒ standardize ⇒ PCA

} may get different principal components

LECTURE 7 10/3/2022

REMARK

$\mathrm{GenVar}(\underline{X}) = \mathrm{Det}\,\Sigma = \prod_{i=1}^{p} \lambda_i = \mathrm{Det}\,\Lambda = \mathrm{GenVar}(\underline{Y})$

$\mathrm{TotalVar}(\underline{X}) = \mathrm{tr}\,\Sigma = \mathrm{tr}\,\Lambda = \mathrm{Tot.Var}(\underline{Y})$

INTERPRETATION

$\mathrm{Cor}(Y_i, X_k) = e_{ki}\dfrac{\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$

e_{ki}: tells how much X_k is important in generating Y_i (don't forget that also σ_{kk} plays an important role)

Interpret the loadings only if the variances are of comparable sizes

↳ what does it mean? "ART", depends on the problem and on how good you are...

PCA of std. variables

$\underline{X} \overset{\text{std}}{\rightsquigarrow} V^{-1/2}(\underline{X}-\underline{\mu}) = \underline{Z}$ $\quad V = \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{pp} \end{bmatrix}$

$\underline{Z} = \left(\frac{X_1-\mu_1}{\sqrt{\sigma_{11}}}, \dots, \frac{X_p-\mu_p}{\sqrt{\sigma_{pp}}} \right)'$

$Cov(\underline{Z}) = V^{-1/2} \Sigma V^{-1/2} = \rho = Cor(\underline{X})$

$\hookrightarrow \rho = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i'$ (NOT the same $\lambda_i, \underline{e}_i$ of Σ !)

$\Rightarrow$ PCA of ρ

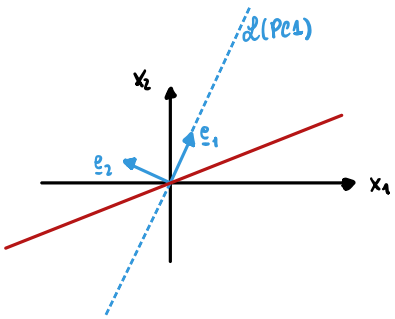
EX

$\underline{X} \in \mathbb{R}^2$ $\quad \Sigma = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix} \Rightarrow \rho = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$ $\quad \underline{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

- PCA on Σ:

  $\lambda_1 = 100.16$ $\quad \underline{e}_1 = (0.04, 0.999)'$ $\quad (\underline{e}_1 \perp \underline{e}_2)$

  $\lambda_2 = 0.84$ $\quad \underline{e}_2 = (0.999, -0.04)'$



- linear space identified by PC1 ($\underline{e}_1$)

  $0.999\, x_1 - 0.04\, x_2 = 0$

  $\hookrightarrow x_2 = \underbrace{\frac{0.999}{0.04}}_{\approx 25} x_1$

- Regression line

  $\left(\frac{Y-\mu_Y}{\sqrt{\sigma_{YY}}} = r \frac{X-\mu_X}{\sqrt{\sigma_{XX}}} \right)$

  $\hookrightarrow \frac{x_2}{10} = 0.4 \frac{x_1}{1} \Rightarrow x_2 = 4 x_1$

- PCA on ρ:

  $\lambda_1 = 1.4$ $\quad \underline{e}_1 = (1/\sqrt{2}, 1/\sqrt{2})'$

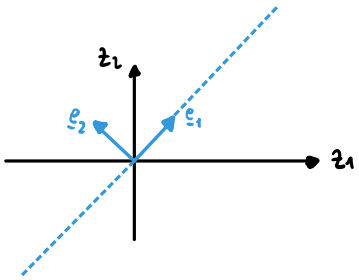
  $\lambda_2 = 0.6$ $\quad \underline{e}_2 = (1/\sqrt{2}, -1/\sqrt{2})'$

- $z_2 = z_1$

  $\frac{x_2}{10} = x_1 \Rightarrow x_2 = 10\, x_1$

- $z_2 = 0.4\, z_1$

  $\frac{x_2}{10} = 0.4\, x_1 \Rightarrow \underbrace{x_2 = 4\, x_1}_{\text{the regression line is SCALE-INVARIANT}}$



In general

$\Sigma = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$ $\quad$ reg line $X_2 = r X_1$ $\quad$ PC1 $X_2 = X_1$

HOWEVER μ, Σ are unknown, but we can estimate them with data ! $\hat{\underline{\mu}} = \bar{\underline{X}}$ $\quad \hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})'$

$\underline{x}_i \mapsto \underline{y}_i = \begin{bmatrix} \underline{e}_1' \underline{x}_i \\ \vdots \\ \underline{e}_p' \underline{x}_i \end{bmatrix}$ $\qquad$ $X = \begin{bmatrix} X_1 & \dots & X_p \\ x_{11} & \dots & x_{1p} \\ x_{21} & & \vdots \\ x_{n1} & & x_{np} \end{bmatrix} \mapsto \left[\begin{array}{ccc|c} Y_1 & Y_2 & \cdots & Y_p \\ \underline{e}_1' \underline{x}_1 & & & \\ \vdots & & & \\ \underline{e}_1' \underline{x}_n & & & \end{array}\right]$

only keep first k components

What is the right k? ART...

OBSERVATION

↳ for $k=1,2,3$ you can PLOT

↳ $\mathrm{Var}(Y_i) = \lambda_i$

↳ λ_1 explains $\frac{\lambda_1}{\sum \lambda_i}$ of the tot. var.

↳ $\lambda_1+\lambda_2$ explain $\frac{\lambda_1+\lambda_2}{\sum \lambda_i}$ of the tot. var.

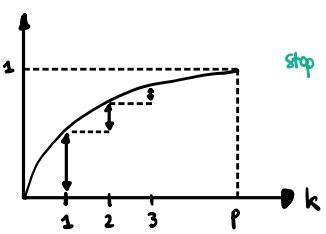
⋮

Stop at 0.80 (80%) rule of thumb

↳ For PCA on ρ

$\bar{\lambda} = \frac{\sum \lambda_i}{p} = \frac{\mathrm{tr}\,\rho}{p} = \frac{p}{p} = 1$ → take the components for which the eigenvalues are $\lambda_i > 1$

↳ $\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$



Stop when adding one dimension doesn't explain a valuable amount of variability

DIFFERENT PERSPECTIVE ON PCA

↳ optimal empirical orthonormal basis

Problem: Find linear space $\mathcal{L}$ of dim. $k<p$ "closest" to the points.

$\mathcal{L} = \mathrm{span}(\eta_1, \eta_2, \dots, \eta_k)$ ← o.n. basis

Find $\eta_1, \dots, \eta_k$ s.t. $\sum_{i=1}^{n} \underbrace{\| (x_i - \bar{x}) - \sum_{j=1}^{k} \eta_j (\eta_j^T (x_i - \bar{x})) \|^2}_{(**)}$ is minimum $= (*)$

$v_i = x_i - \bar{x}$

$$(**) = \| v_i - \sum_{j=1}^{k} \eta_j (\eta_j^T v_i) \|^2 = \left(v_i - \sum_{j=1}^{k} \eta_j (\eta_j^T v_i) \right)^T \left(v_i - \sum_{j=1}^{k} \eta_j (\eta_j^T v_i) \right)$$

$$= v_i^T v_i - 2 \sum_{j=1}^{k} v_i^T \eta_j \eta_j^T v_i + \sum_{j=1}^{k} (v_i^T \eta_j)^2$$

$$= v_i^T v_i - \sum_{j=1}^{k} (v_i^T \eta_j)^2$$

Hence $(*) = \sum_{i=1}^{n} \left(v_i^T v_i - \sum_{j=1}^{k} (v_i^T \eta_j)^2 \right)$ to be min

$\Leftrightarrow \sum_{i=1}^{n} \sum_{j=1}^{k} (v_i^T \eta_j)^2$ to be max

$$\sum_i \sum_j \eta_j^T v_i v_i^T \eta_j = \sum_j \eta_j^T \overbrace{\sum_i v_i v_i^T}^{\text{almost } S} \eta_j = (n-1) \sum_{j=1}^{k} \eta_j^T S \eta_j \text{ to be max}$$

↳ $k=1 \Rightarrow \max_{\substack{\|\eta_1\|=1 \\ \eta_1 \in \mathbb{R}^p}}{}^{(n-1)} \eta_1^T S \eta_1 = \lambda_1 \qquad \eta_1 = e_1 \qquad S = \sum_{i=1}^{p} \lambda_i e_i e_i^T$

↳ by induction: $\eta_1 = e_1, \dots, \eta_k = e_k$

APPROXIMATION ERROR

$$\sum_{i=1}^{n} \| (x_i - \bar{x}) - \sum_{j=1}^{k} e_j e_j^T (x_i - \bar{x}) \|^2 = \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})^T (x_i - \bar{x})}_{(n-1)} - \underbrace{(n-1) \sum_{j=1}^{k} e_j^T S e_j}_{(n-1)\sum_{j=1}^{k} \lambda_j} = (n-1) \sum_{i=1}^{p} \lambda_i - (n-1) \sum_{i=1}^{k} \lambda_i = (n-1) \sum_{i=k+1}^{p} \lambda_i$$

left out from PCA

$$(*) = \mathrm{tr}\left(\sum_{i=1}^{n} (x_i - \bar{x})^T (x_i - \bar{x}) \right) = \sum_{i=1}^{n} \mathrm{tr}\left[(x_i - \bar{x})^T (x_i - \bar{x}) \right] \neq \sum_{i=1}^{n} \mathrm{tr}\left[(x_i - \bar{x})(x_i - \bar{x})^T \right] = \mathrm{tr}\left[\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \right] = (n-1)\,\mathrm{tr}(S) = (n-1) \sum_{i=1}^{p} \lambda_i$$

$\mathrm{tr}(ABC) = \mathrm{tr}(CAB)$

SVD (Singular value Decomposition)

$$X = \underset{n\times m}{U}\ \underset{m\times p}{D}\ \underset{p\times p}{V} \qquad D = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_p \\ & 0 & \end{bmatrix}$$

(unitary, unitary)

$$S = \frac{1}{n-1}X'X = \frac{1}{n-1}V'\underbrace{U'U}_{I}DV = \frac{1}{n-1}V'\overbrace{D'D}^{\Lambda}V = \frac{1}{n-1}P\Lambda P$$

you can reach the same result!

LECTURE 8 11/3/2021

Gaussian distribution

$\mathbb{R}^p \ni \underline{X} \sim N_p(\underline{\mu}, \Sigma)$ if $P[\underline{X}\in B] = \int_B \varphi_{\underline{X}}(\underline{t})\,d\underline{t}$ $\forall B \in \mathbb{R}^p$ Boul

($\underline{\mu} \in \mathbb{R}^p$, Σ $p\times p$ SDP)

$$\varphi_{\underline{X}}(\underline{t}) = \frac{1}{\sqrt{(2\pi)^p \det\Sigma}} \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu})'\Sigma^{-1}(\underline{x}-\underline{\mu})\right\}$$

PROP

- $E[\underline{X}] = \underline{\mu}$
- $Cov(\underline{X}) = \Sigma$

THEOREM (linear comb. is still gaussian)

$\underline{X} \sim N_p(\underline{\mu}, \Sigma) \iff \forall \underline{a}\in\mathbb{R}^p,\ \underline{a}'\underline{X} \sim N_1(\underline{a}'\underline{\mu}, \underline{a}'\Sigma\underline{a})$

proof via characteristic function

COROLLARY

$\underline{X} \sim N_p(\underline{\mu}, \Sigma) \Rightarrow X_i \sim N_1(\mu_i, \sigma_{ii})$ $\forall i = 1,\dots,p$

proof $X_i = \underline{u}_i'\underline{X}$ $\underline{u}_i = (0,\dots,0,\underset{i}{1},0,\dots,0)$

Theorem above $\sim N_1(\underline{u}_i'\underline{\mu}, \underline{u}_i'\Sigma\underline{u}_i)$

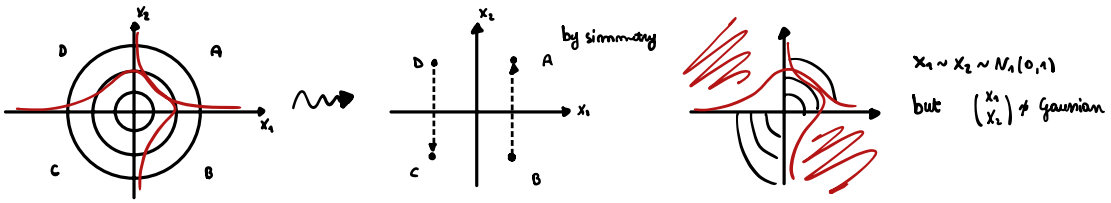
$= N_1(\mu_i, \sigma_{ii})$

OBS ~~$\Leftarrow$~~

$\underline{Z} \sim N_2(\underline{0}, I)$

$X_1 \sim N_1(0,1)$

$X_2 \sim N_1(0,1)$



PROP A $q\times p$ $q<p$

$\underline{X} \sim N_p(\underline{\mu}, \Sigma) \Rightarrow A\underline{X} \sim N_q(A\underline{\mu}, A\Sigma A')$

proof We need to prove $\forall \underline{a}\in\mathbb{R}^q$ $\underline{a}'(A\underline{X}) \sim N_1(\underline{a}'(A\underline{\mu}), \underline{a}'(A\Sigma A)\underline{a})$

$(\underline{a}'A)\underline{X} = (A'\underline{a})'\underline{X} \overset{Thm}{\sim} N_1(\underbrace{(A'\underline{a})'\underline{\mu}}_{\underline{a}'(A\underline{\mu})}, \underbrace{(A'\underline{a})'\Sigma(A'\underline{a})}_{\underline{a}'(A\Sigma A)\underline{a}})$ ∎

PROP $X \sim N_p(\underline{\mu}, \Sigma)$ $\underline{d}\in\mathbb{R}^p \Rightarrow \underline{X}+\underline{d} \sim N_p(\underline{\mu}+\underline{d}, \Sigma)$

proof Exercise.

OBS $t_1,\dots,t_p$ iid $\sim N_1(0,1)$

$\underline{t} = (t_1,\dots,t_p)'$

$\underline{\mu}\in\mathbb{R}^p$, Σ $p\times p$ pos. def

$$\varphi_{\underline{t}}(\underline{t}) = \prod_{i=1}^p \varphi_{t_i}(t_i) = \prod_{i=1}^p \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t_i^2} = \frac{1}{\sqrt{(2\pi)^p}} \exp\left[\overbrace{-\frac{1}{2}\underline{t}'I\underline{t}}^{}\ -\frac{1}{2}\sum_{i=1}^p t_i^2\right] \Rightarrow \underline{Z} \sim N_p(\underline{0}, I)$$

$\underline{X} = \Sigma^{1/2}\underline{Z} + \underline{\mu} \sim N_p(\underline{\mu}, \Sigma)$ If you can generate random gaussian distribution in 1D, you can in any D.

OBS $\mathbb{R}^p \ni \underline{X} = \begin{pmatrix} \underline{X}_1 \in \mathbb{R}^q \\ \underline{X}_2 \in \mathbb{R}^{p-q} \end{pmatrix}$, $\underline{\mu} = \begin{pmatrix} \underline{\mu}_1 \in \mathbb{R}^q \\ \underline{\mu}_2 \in \mathbb{R}^{p-q} \end{pmatrix}$ $\Sigma = \begin{bmatrix} \Sigma_{11}\ q\times q & \Sigma\ q\times(p-q) \\ \Sigma_{21}\ (p-q)\times q & \Sigma_{22}\ (p-q)\times(p-q) \end{bmatrix}$

PROP $\begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix} \sim \underline{X} \sim N_p\left(\begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right) \Rightarrow \underline{X} \sim N_p(\underline{\mu}, \Sigma)$

proof ...

PROP $\begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix} \sim \underline{X} \sim N_p\left(\begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$ If $\Sigma_{12} = 0 \Rightarrow \underline{X}_1 \perp\!\!\!\perp \underline{X}_2$

proof Show that $\varphi_{\underline{X}}(\underline{t}) = \varphi_{\underline{X}_1}(\underline{t}_1) \cdot \varphi_{\underline{X}_2}(\underline{t}_2)$

THEOREM $\underline{X}_1 \mid \underline{X}_2 = \underline{x}_2 \sim N_q\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underline{x}_2 - \underline{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right)$

Ex $p=2$

$\mathbb{R}^2 \ni \underline{X} = \begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_{YY} & \sigma_{YX} \\ \sigma_{XY} & \sigma_{XX} \end{pmatrix} \right)$ $\quad Y \sim N_1(\mu_Y, \sigma_{YY})$

$Y \mid X = x \sim N_1\Big(\underbrace{\mu_Y + \sigma_{YX}\cdot\sigma_{XX}^{-1}(x-\mu_X)}_{\mu_Y + \frac{\sigma_{XY}}{\sqrt{\sigma_{XX}\sigma_{YY}}}\sqrt{\sigma_{YY}}\cdot\frac{x-\mu_X}{\sqrt{\sigma_{XX}}}},\ \underbrace{\sigma_{YY} - \sigma_{XY}\sigma_{XX}^{-1}\sigma_{YX}}_{\sigma_{YY}\left(1 - \frac{\sigma_{XY}^2}{\sigma_{XX}\sigma_{YY}}\right)} \Big)$ (*), ρ^2, ρ

$= N\left(\mu_Y + \rho\frac{\sqrt{\sigma_{YY}}}{\sqrt{\sigma_{XX}}}(x-\mu_X),\ \sigma_{XY}(1-\rho^2) \right)$

When conditioning on X:

- If $\rho \approx 1 \Rightarrow (*) \approx 0$ so I will be pretty sure on Y
- If $\rho \approx 0 \Rightarrow (*) \approx \sigma_{YY}$ my information doesn't change when knowing what's carried by X

↳ I know this even BEFORE observing X

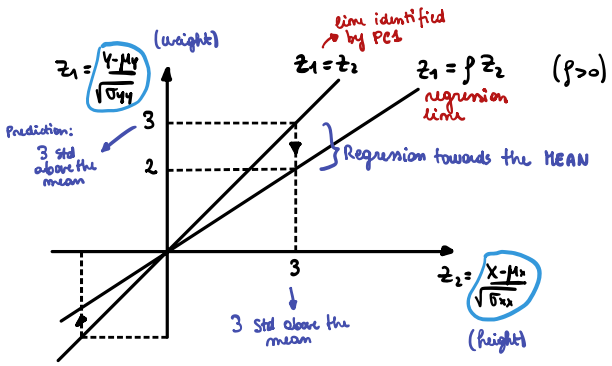
$\mathbb{E}[Y \mid X = x] = \mu_Y + \rho\frac{\sqrt{\sigma_{YY}}}{\sqrt{\sigma_{XX}}}(x-\mu_X)$

↳ regression line: $y = \mu_Y + \rho\frac{\sqrt{\sigma_{YY}}}{\sqrt{\sigma_{XX}}}(x-\mu_X)$

$\Downarrow$

$\frac{y-\mu_Y}{\sqrt{\sigma_{YY}}} = \rho\frac{x-\mu_X}{\sqrt{\sigma_{XX}}}$ (slope)

standardized variables



LECTURE 9 15/3/2022

$\underline{x}_i$ realization of $\underline{X}_i$, $\underline{X}_1, \dots, \underline{X}_m \overset{\text{iid}}{\sim} N_p(\mu, \Sigma)$ (iid? Difficult to trust?) Data: $\mathbb{X} = \begin{bmatrix} \underline{x}_1^t \\ \vdots \\ \underline{x}_m^t \end{bmatrix}$ $\underline{x}_i \in \mathbb{R}^p$

Obvious candidates as estimators:

- for $\underline{\mu}$: $\frac{1}{m}\sum_{i=1}^m \underline{x}_i$
- for Σ: $\frac{1}{m-1}\sum_{i=1}^m (\underline{x}_i - \underline{\bar{x}})(\underline{x}_i - \underline{\bar{x}})$ (Methods of moments)

For gaussian we also have ML

Maximum Likelihood

Suppose you toss a coin 5 times: HTTHT

Model: X_1, X_2, X_3, X_4, X_5 iid $\sim Be(p)$ p?

$\mathbb{P}(\text{What has been observed}) = \mathbb{P}_p(\text{HTTHT}) \overset{\perp\!\!\!\perp}{=} p^2(1-p)^3$ (func. of p) $\Rightarrow L(p) = p^2(1-p)^3 = p^2 - p^5$

A believes $p = 0.4$ $\quad L(0.4) =$ ⬚

B believes $p = 0.5$ $\quad L(0.5) =$ ⬚

In general $L(p \mid x_1 \dots x_m) = p^{\sum x_i}(1-p)^{m-\sum x_i}$

$L'(p) = 2p - 5p^4 = 0$

$\hat{p} = 0 \vee \hat{p} = \sqrt[3]{\frac{2}{5}}$

$\hat{p} = \underset{p}{\operatorname{argmax}}\, L(p) = \underset{p}{\operatorname{argmax}}\, \log L(p)$ (log injective)

$L(\underline{\mu}, \Sigma \mid \underline{X}_1 \dots \underline{X}_m) = \prod_{i=1}^m \frac{1}{\sqrt{(2\pi)^p \det \Sigma}} \exp\left\{ -\frac{1}{2}(\underline{x}_i - \underline{\mu})\Sigma^{-1}(\underline{x}_i - \underline{\mu}) \right\}$

$(\hat{\underline{\mu}}, \hat{\Sigma}) = \underset{\mu \in \mathbb{R}^p,\ \Sigma \text{ pos def } p\times p}{\operatorname{argmax}} L(\mu, \Sigma) = \dots = \left(\bar{\underline{X}}, \frac{m-1}{m} S \right)$

Invariance principle of MLE

Suppose $\vartheta \in \mathbb{R}^k$ is a parameter, suppose $\hat{\vartheta}$ is MLE for ϑ. Consider $h: \mathbb{R}^k \to \mathbb{R}^j$ a mapping. Then $\widehat{h(\vartheta)} = h(\hat{\vartheta})$

Ex $\Sigma = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i'$, $\hat{\Sigma}$ MLE for Σ

$\sum_{i=1}^{p} \tilde{\lambda}_i \tilde{\underline{e}}_i \tilde{\underline{e}}_i'$ $\Rightarrow$ $\tilde{\lambda}_i = \hat{\lambda}_i$ $\tilde{\underline{e}}_i = \hat{\underline{e}}_i$

PROP $\bar{\underline{X}} = \frac{1}{m} \sum_{i=1}^{m} \underline{X}_i \sim N_p(\underline{\mu}, \frac{1}{m}\Sigma)$

proof $\tilde{\underline{X}} = \begin{bmatrix} \underline{X}_1 \\ \vdots \\ \underline{X}_m \end{bmatrix} \in \mathbb{R}^{mp} \sim N_{mp}\left(\begin{pmatrix} \underline{\mu} \\ \vdots \\ \underline{\mu} \end{pmatrix}, \begin{pmatrix} \Sigma & & 0 \\ & \ddots & \\ 0 & & \Sigma \end{pmatrix} \right)$ Let $A = [\overbrace{I_{p\times p} \cdots I_{p\times p}}^{m \text{ times}}] \in \mathbb{R}^{p \times mp}$

$\Rightarrow A\tilde{\underline{X}} = \left(\sum_{i=1}^{m} X_{i1}, \dots, \sum_{i=1}^{m} X_{ip} \right)' \Rightarrow \frac{1}{m} A\tilde{\underline{X}} = \bar{\underline{X}} \sim N_p\left(\frac{1}{m} A \begin{pmatrix} \underline{\mu} \\ \vdots \\ \underline{\mu} \end{pmatrix}, \frac{1}{m^2} A \Sigma A' \right)$ ...

DEF $\underline{z}_1, \dots, \underline{z}_m \overset{iid}{\sim} N_p(\underline{0}, \Sigma)$ Σ pos.def $(\det \Sigma > 0)$ Then $\sum_{i=1}^{m} \underline{z}_i \underline{z}_i' \sim \text{Wishart}(\Sigma, m)$
(p (we don't write it, it's clear from the context))

Properties

1) $A_1 \sim \text{Wishart}(\Sigma, m_1)$, $A_2 \sim \text{Wishart}(\Sigma, m_2)$, $A_1 \perp\!\!\!\perp A_2$ $\Rightarrow A_1 + A_2 \sim \text{Wishart}(\Sigma, m_1 + m_2)$

proof $A_1 = \sum_{i=1}^{m_1} \underline{z}_i \underline{z}_i'$, $A_2 = \sum_{i=1}^{m_2} \tilde{\underline{z}}_i \tilde{\underline{z}}_i'$ $\Rightarrow$ we order them $\underline{z}_1, \dots, \underline{z}_{m_1}, \tilde{\underline{z}}_1, \dots, \tilde{\underline{z}}_{m_2}$

$\underline{W}_1, \dots, \underline{W}_{m_2}, \underline{W}_{m_2+1}, \dots, \underline{W}_{m_1+m_2}$ $\overset{\perp\!\!\!\perp}{\Rightarrow}$ $\underline{W}_1, \dots, \underline{W}_{m_1+m_2} \sim N_p(\underline{0}, \Sigma)$

$\Rightarrow A_1 + A_2 = \sum_{i=1}^{m_1} \underline{z}_i \underline{z}_i' + \sum_{i=1}^{m_2} \tilde{\underline{z}}_i \tilde{\underline{z}}_i' = \sum_{i=1}^{m_1+m_2} \underline{W}_i \underline{W}_i' \sim \text{Wishart}(\Sigma, m_1 + m_2)$

2) $C \in \mathbb{R}^{k\times p}$, $A \sim \text{Wishart}(\Sigma, m)$ $\Rightarrow CAC' \sim \text{Wishart}(C\Sigma C', m)$

proof $A = \sum_{i=1}^{m} \underline{z}_i \underline{z}_i'$ $\Rightarrow CAC^T = \sum_{i=1}^{m} C\underline{z}_i \underline{z}_i' C' = \sum_{i=1}^{m} \underline{W}_i \underline{W}_i'$ where $\underline{W}_i = C\underline{z}_i \sim N_k(\underline{0}, C\Sigma C^T)$

3) $\sigma^2 \in \mathbb{R}$, $\sigma^2 > 0$, $A \sim \text{Wishart}(\Sigma, m)$ $\Rightarrow \sigma^2 A \sim \text{Wishart}(\sigma^2 \Sigma, m)$

proof $\sigma^2 A = \sum_{i=1}^{m} \sigma \underline{z}_i \underline{z}_i' \sigma$ $\underline{W}_i = \sigma \underline{z}_i \sim N_p(\underline{0}, \sigma^2\Sigma)$ use (2)

4) $A \sim \text{Wishart}(\Sigma, m)$ $\Sigma = [\sigma^2]$ $(p = 1)$

$A = \sum_{i=1}^{m} z_i z_i = \sum_{i=1}^{m} z_i^2$ $\Rightarrow \frac{1}{\sigma^2} A = \sum_{i=1}^{m} \frac{z_i}{\sigma} \frac{z_i}{\sigma} = \sum_{i=1}^{m} W_i W_i$ $W_i \sim N(0,1)$ $\Rightarrow \frac{1}{\sigma^2} A \sim \chi^2(m) \Rightarrow$ "$A \sim \sigma^2 \chi^2(m)$"

The Wishart is a multivariate extension of χ^2.

Now, if $A \sim \text{Wishart}(\Sigma, m)$, $\underline{0} \neq \underline{c} \in \mathbb{R}^p$, $\Sigma \in \mathbb{R}^{p\times p}$ $\overset{(2)}{\Rightarrow}$ $\underline{c}' A \underline{c} \sim \text{Wishart}(\overbrace{\underline{c}'\Sigma\underline{c}}^{>0}, m) \overset{(4)}{\sim} \underline{c}'\Sigma\underline{c}\, \chi^2(m)$

$\Rightarrow \underline{c}' A \underline{c} \sim (\underline{c}'\Sigma\underline{c})\, \chi^2(m)$

$\Rightarrow \frac{\underline{c}' A \underline{c}}{\underline{c}'\Sigma\underline{c}} \sim \chi^2(m)$

THEOREM

$\underline{X}_1 \dots \underline{X}_m \overset{iid}{\sim} N_p(\underline{\mu}, \Sigma)$ $\Rightarrow \sum_{j=1}^{m} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})' \sim \text{Wishart}(\Sigma, m-1)$

COROLLARY Thm + (3)

$S = \frac{1}{m-1} \sum_{i=1}^{m} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}}) \sim \text{Wishart}\left(\frac{1}{m-1}\Sigma, m-1 \right)$

$\hat{\Sigma} = \frac{m-1}{m} S \sim \text{Wishart}\left(\frac{1}{m}\Sigma, m-1 \right)$

Summary (THEOREM)

$\underline{X}_1 \dots \underline{X}_m \overset{iid}{\sim} N_p(\underline{\mu}, \Sigma)$:

1) $\bar{\underline{X}} \sim N_p(\underline{\mu}, \frac{1}{m}\Sigma)$
2) $(m-1)S \sim \text{Wishart}(\Sigma, m-1)$
3) $\bar{\underline{X}} \perp\!\!\!\perp S$

LECTURE 10 17/3/2022

THEOREM (LLN) (weak)

$\mathbb{R}^p \ni \underline{X}_1, \dots, \underline{X}_m \overset{iid}{\sim} \underline{\mu}, \Sigma$

$\bar{\underline{X}}_m = \frac{1}{m}\sum_{i=1}^{m} \underline{X}_i \xrightarrow{\mathbb{P}} \underline{\mu}$

$S = \frac{1}{m-1}\sum_{i=1}^{m} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})' \xrightarrow{\mathbb{P}} \Sigma$

THEOREM CLT

$\sqrt{m}(\bar{\underline{X}} - \underline{\mu}) \sim AN_p(\underline{0}, \Sigma)$

$\sqrt{m}(\underline{X}_m - \underline{\mu}) \xrightarrow{\mathcal{L}} N_p(\underline{0}, \Sigma)$

INFERENCE ON $\underline{\mu}$

▶ CASE $m \gg p$

Then CLT yields $\sqrt{m}(\bar{\underline{X}} - \underline{\mu}) \overset{\cdot}{\sim} N_p(\underline{0}, \Sigma) \Rightarrow \sqrt{m}(\bar{\underline{X}} - \underline{\mu})\Sigma^{-1}(\bar{\underline{X}} - \underline{\mu})\sqrt{m} \sim \chi^2(p)$

↳ pivotal ($\mathcal{L}$ doesn't depend on $\Sigma, \underline{\mu}$)

Σ is not known, LLN $S \sim \Sigma$ as $m \to \infty$ thus $\boxed{m(\bar{\underline{X}} - \underline{\mu})S^{-1}(\bar{\underline{X}} - \underline{\mu}) \sim \chi^2(p)}$ pivotal

$\mathbb{P}\left(\underbrace{m(\bar{\underline{X}} - \underline{\mu})S^{-1}(\bar{\underline{X}} - \underline{\mu}) \le \chi^2_{1-\alpha}(p)}_{\text{Ellipse}} \right) = 1-\alpha \longrightarrow \boxed{\mathbb{P}(\bar{\underline{X}} \in \mathcal{E}^{1-\alpha}_{\frac{1}{m}S^{-1}}(\underline{\mu})) = 1-\alpha}$

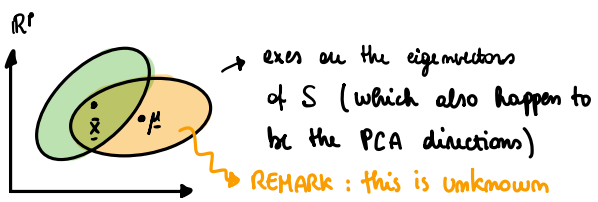
$\Rightarrow \mathbb{P}(d^2_{\frac{1}{m}S^{-1}}(\bar{\underline{X}} - \underline{\mu}) \le \chi^2_{1-\alpha}(p)) = 1-\alpha \qquad \boxed{\mathbb{P}(\underline{\mu} \in \mathcal{E}^{1-\alpha}_{\frac{1}{m}S^{-1}}(\bar{\underline{X}})) = 1-\alpha}$

different centres:

$\mathcal{E}^{1-\alpha}_{\frac{1}{m}S^{-1}}(\underline{\mu}) = \{\underline{y} \in \mathbb{R}^p : m(\underline{y} - \underline{\mu})'S^{-1}(\underline{y} - \underline{\mu}) \le \chi^2_{1-\alpha}(p)\}$

$\mathcal{E}^{1-\alpha}_{\frac{1}{m}S^{-1}}(\bar{\underline{X}}) = \{\underline{\eta} \in \mathbb{R}^p : m(\underline{\eta} - \bar{\underline{X}})'S^{-1}(\underline{\eta} - \bar{\underline{X}}) \le \chi^2_{1-\alpha}(p)\}$

$\bar{\underline{X}} \in \mathcal{E}^{1-\alpha}_{\frac{1}{m}S^{-1}}(\underline{\mu}) \Leftrightarrow \underline{\mu} \in \mathcal{E}^{1-\alpha}_{\frac{1}{m}S^{-1}}(\bar{\underline{X}})$



DEF Confidence region of level $1-\alpha$, $\alpha \in (0,1)$, for $\underline{\mu}$: $\boxed{CR_{1-\alpha}(\underline{\mu}) = \{\underline{\eta} \in \mathbb{R}^p : m(\underline{\eta} - \bar{\underline{X}})'S^{-1}(\underline{\eta} - \bar{\underline{X}}) \le \chi^2_{1-\alpha}(p)\}}$

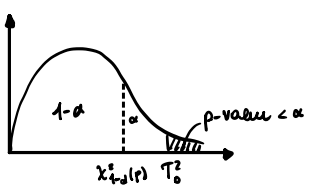
- $(1-\alpha)\%$ of the times this subset will cover the mean $\underline{\mu}$
- Obv. I don't know wether I guessed correctly or not.

We can now build tests and the complement of the confidence region: the rejection region.

TESTING

$H_0: \underline{\mu} = \underline{\mu}_0$ vs $H_1: \underline{\mu} \neq \underline{\mu}_0$

Evidence against H_0: $d_{\frac{1}{m}S^{-1}}(\bar{\underline{X}}, \underline{\mu}_0)$ large



If H_0 is TRUE $\Rightarrow m(\bar{\underline{X}} - \underline{\mu}_0)'S^{-1}(\bar{\underline{X}} - \underline{\mu}_0) \sim \chi^2(p)$

Reject H_0 if $T_0^2 = m(\bar{\underline{X}} - \underline{\mu}_0)'S^{-1}(\bar{\underline{X}} - \underline{\mu}_0) > \chi^2_{1-\alpha}(p)$

↳ Rejection Region $\{\text{data} : T_0^2 > \chi^2_{1-\alpha}(p)\} = \{\text{data} : \text{p-value} < \alpha\}$

▶ CASE n SMALL ⇒ assume gaussianity

$\underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} N_p(\mu, \Sigma)$

$n(\bar{\underline{X}}-\mu)' S^{-1}(\bar{\underline{X}}-\mu) \sim ?$

DEF $W \sim \chi^2(m)$, $Y \sim \chi^2(n)$, $W \perp\!\!\!\perp Y$ ⇒ $\frac{Y/n}{W/m} \sim F(m,n)$ Fisher's distribution

REMARK $t = Z/\sqrt{\frac{W}{m}}$, $Z \sim N(0,1)$, $W \sim \chi^2(m)$, $Z \perp\!\!\!\perp W$ ⇒ $t \sim t(m)$

Then $t^2 = \frac{Z^2}{W/m}$ $Z^2 \sim \chi^2(1)$, $W \sim \chi^2(m)$ ⇒ $t^2 \sim F(1,m)$

REMARK $F(m,n) \xrightarrow[m\to\infty]{\mathcal{L}} \frac{1}{m}\chi^2(m)$

$Y \sim \chi^2(n)$, $W \sim \chi^2(m)$ $\quad \frac{W}{m} \xrightarrow[m\to\infty]{} ?$ $\quad W = \sum_{i=1}^{m} Z_i^2 \overset{Z_i \sim N(0,1)}{\xrightarrow[m\to\infty]{LLN}} \mathbb{E}[Z_1^2] = 1$ $\quad \mathrm{Var}(Z_1) = \mathbb{E}[Z_1^2] - (\mathbb{E}[Z_1])^2 = 1-0 = 1$

HOTELLING'S THEOREM (1931)

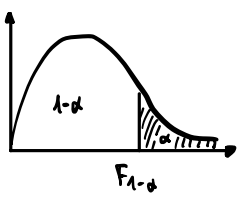
$\underline{X} \sim N_p(\mu, \Sigma)$ $\det \Sigma > 0$
$W \sim \mathrm{Wish}(\frac{1}{m}\Sigma, m)$
$\underline{X} \perp\!\!\!\perp W$
⇒ $\frac{m-p+1}{mp}(\underline{X}-\mu)' W^{-1}(\underline{X}-\mu) \sim F(p, m-p+1)$

COROLLARY $\underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} N_p(\mu, \Sigma)$, $\det \Sigma > 0$ ⇒ $n(\bar{\underline{X}}-\mu)' S^{-1}(\bar{\underline{X}}-\mu) \sim \frac{(n-1)p}{n-p} F(p, n-p)$ pivotal

proof $\sqrt{n}(\bar{\underline{X}}-\mu) \sim N_p(0, \Sigma)$ (exact)
$S \sim \mathrm{Wishart}(\frac{1}{n-1}\Sigma, n-1)$
$\bar{\underline{X}} \perp\!\!\!\perp S$
⇒ (Hotelling $m = n-1$) $\frac{n-1-p+1}{(n-1)p} n(\bar{\underline{X}}-\mu)' S^{-1}(\bar{\underline{X}}-\mu) \sim F(p, n-1-p+1)$

LECTURE 11 18/3/2022

DEF Hotelling's T^2 statistic: $T^2 = n(\bar{\underline{X}}-\mu)' S^{-1}(\bar{\underline{X}}-\mu)$

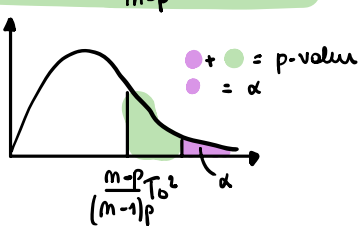


CONFIDENCE REGIONS Fix $\alpha \in (0,1)$ $\quad \mathbb{P}\left(T^2 \le \frac{(n-1)p}{n-p} F_\alpha(p, n-p)\right) = 1-\alpha$

$CR_{1-\alpha}(\mu) = \left\{ \eta \in \mathbb{R}^p : (\eta - \bar{\underline{X}})' S^{-1} (\eta - \bar{\underline{X}}) \le \frac{(n-1)p}{n-p} F_\alpha(p, n-p) \right\}$ ≠ radius from the "n large" case

TESTING
$H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$

If H_0 true ⇒ $T_0^2 = n(\bar{\underline{X}}-\mu_0)' S^{-1}(\bar{\underline{X}}-\mu_0) \sim \frac{(n-1)p}{n-p} F(p, n-p)$
Reject H_0 if $T_0^2 > \frac{(n-1)p}{n-p} F_\alpha(p, n-p)$



REM Don't forget about $\frac{(n-p)}{(n-1)p}$!

▶ The CR identifies all the μ_0 for we can't reject H_0

What happens as $n \to \infty$?
$\frac{(n-1)p}{n-p} F_{1-\alpha}(p, n-p) \xrightarrow{n\to\infty} \chi^2_{1-\alpha}(p)$ as we expected from the previous case
($\frac{(n-1)p}{n-p} \to 1$, $F_{1-\alpha}(p, n-p) \to \frac{1}{p}\chi^2(p)$)

INFERENCE FOR LINEAR COMBINATIONS OF $\underline{\mu}$

$\underline{a} \in \mathbb{R}^p$ $\quad \underline{a}'\underline{\mu} = a_1\mu_1 + \dots + a_p\mu_p$ (my portfolio and not all the market stocks)

Ex $\quad \underline{a} = \underline{u}_i = (0 \dots 0 \underset{i}{1} 0 \dots 0) \Rightarrow \underline{a}'\underline{\mu} = \mu_i$

$\underline{a} = (0 \dots 0 \underset{i}{1} 0 \dots 0 \underset{j}{-1} 0 \dots 0) \Rightarrow \underline{a}'\underline{\mu} = \mu_i - \mu_j$

▶ CASE "ONE-AT-THE-TIME"

▶ $\underline{a}'\bar{\underline{X}}$ (unbiased) estimator for $\underline{a}'\underline{\mu}$

▶ As before: • n small $\Rightarrow$ gaussianity
• n large $\Rightarrow$ CLT to have approximately gaussianity

▶ If $\underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} N_p(\underline{\mu}, \Sigma) \Rightarrow \bar{\underline{X}} \sim N_p(\underline{\mu}, \frac{1}{n}\Sigma) \Rightarrow \underline{a}'\bar{\underline{X}} \sim N_1(\underline{a}'\underline{\mu}, \frac{1}{n}\underline{a}'\Sigma\underline{a}) \Rightarrow$ (standardize) $\boxed{\frac{\underline{a}'\bar{\underline{X}} - \underline{a}'\underline{\mu}}{\sqrt{\underline{a}'\Sigma\underline{a}}}\sqrt{n} \sim N_1(0,1)}$

▶ Σ is usually unknown!
▶ If n large $\Rightarrow$ plug S [In this case you already have a pivotal quantity 😊]

▶ If n is small (assume gaussianity)

We know $(n-1)S \sim \text{Wish}(\Sigma, n-1)$

then $(n-1)\underline{a}'S\underline{a} \sim (\underline{a}'\Sigma\underline{a})\chi^2(n-1) \Rightarrow \frac{(n-1)\underline{a}'S\underline{a}}{\underline{a}'\Sigma\underline{a}} \sim \chi^2(n-1)$ Moreover $\bar{\underline{X}} \perp\!\!\!\perp S$ and so are their transformations, thus

REMARK
$\frac{N(0,1)}{\sqrt{\frac{\chi^2(m)}{m}}} \sim t(m)$ $(\perp\!\!\!\perp)$

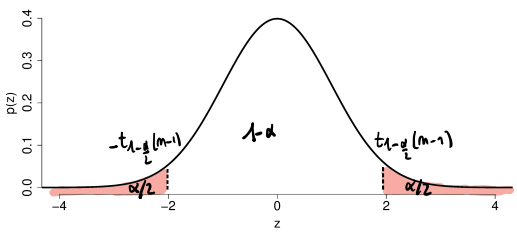
$$\Rightarrow \frac{\frac{\underline{a}'\bar{\underline{X}} - \underline{a}'\underline{\mu}}{\sqrt{\underline{a}'\Sigma\underline{a}}}\sqrt{n}}{\sqrt{\frac{(n-1)\underline{a}'S\underline{a}}{\underline{a}'\Sigma\underline{a}} \cdot \frac{1}{n-1}}} \sim t(n-1) \Rightarrow \boxed{\frac{\underline{a}'\bar{\underline{X}} - \underline{a}'\underline{\mu}}{\sqrt{\underline{a}'S\underline{a}}}\sqrt{n} \sim t(n-1)}$$ pivotal

From this

Random confidence region computed with data

• $\mathbb{P}\left[\frac{|\underline{a}'\bar{\underline{X}} - \underline{a}'\underline{\mu}|}{\sqrt{\underline{a}'S\underline{a}}}\sqrt{n} < t_{1-\frac{\alpha}{2}}(n-1)\right] = 1-\alpha = \mathbb{P}\left[\underline{a}'\underline{\mu} \in \left[\underline{a}'\bar{\underline{X}} \pm t_{1-\frac{\alpha}{2}}(n-1)\sqrt{\frac{\underline{a}'S\underline{a}}{n}}\right]\right] \quad \forall \underline{a} \in \mathbb{R}^p$

• $\boxed{CI_{1-\alpha}(\underline{a}'\underline{\mu}) = \left[\underline{a}'\bar{\underline{X}} \pm t_{1-\frac{\alpha}{2}}(n-1)\sqrt{\frac{\underline{a}'S\underline{a}}{n}}\right]}$



Ex $\quad \underline{a} = \underline{u}_i \qquad CI_{1-\alpha}(\mu_i) = \left[\bar{X}_i \pm t_{1-\frac{\alpha}{2}}(n-1)\sqrt{\frac{S_{ii}}{n}}\right]$

$\underline{a} = (0 \dots 0 \underset{i}{1} 0 \dots 0 \underset{j}{-1} 0 \dots 0) \qquad CI_{1-\alpha}(\mu_i - \mu_j) = \left[\bar{X}_i - \bar{X}_j \pm t_{1-\frac{\alpha}{2}}(n-1)\sqrt{\frac{S_{ii} - 2S_{ij} + S_{jj}}{n}}\right]$

We can do TESTING

$H_0: \underline{a}'\underline{\mu} \overset{(=)}{(\geq)} \delta_0$ (Default pollution)

$H_1: \underline{a}'\underline{\mu} \overset{(\neq)}{(<)} \delta_0$ (We want to prove that it's lower)

$t_0 = \frac{\underline{a}'\bar{\underline{X}} - \delta_0}{\sqrt{\underline{a}'S\underline{a}}}\sqrt{n} \sim t(n-1)$

Reject if $t_0 \overset{(<)}{>} t_{1-\alpha}(n-1)$
$|t_0| > t_{1-\frac{\alpha}{2}}(n-1)$

▶ CASE "SIMULTANEOUS"

In the above we did NOT prove that $\mathbb{P}\left(\underline{a}'\underline{\mu} \in \left[\underline{a}'\bar{\underline{X}} \pm t_{1-\frac{\alpha}{2}}(n-1)\sqrt{\frac{\underline{a}'S\underline{a}}{n}}\right], \forall \underline{a} \in \mathbb{R}^p\right) = 1-\alpha$

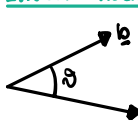
↳ the above would be "no matter what coin I toss I get H ½ of the times"

↳ vs "if I toss all the coins together I get all H with ½ probability"

What should go here to make the statement true?

New problem: $\mathbb{P}\left(\underline{a}'\underline{\mu} \in \left[\underline{a}'\bar{\underline{X}} \pm \boxed{?}\sqrt{\frac{\underline{a}'S\underline{a}}{n}}\right], \forall \underline{a} \in \mathbb{R}^p\right) = 1-\alpha$

LINEAR ALGEBRA RECALL



$\frac{\underline{b}'\underline{d}}{\|\underline{b}\|\|\underline{d}\|} = \cos\vartheta \qquad 0 \leq \frac{(\underline{b}'\underline{d})^2}{\|\underline{b}\|^2\|\underline{d}\|^2} = \cos^2\vartheta \leq 1 \Rightarrow (\underline{b}'\underline{d}) \leq \|\underline{b}\|^2\|\underline{d}\|^2 \quad \forall \underline{b}, \underline{d} \in \mathbb{R}^p$

"=" if $\underline{b} \in \mathcal{L}(\underline{d})$ (C-S inequality)

Let $B \in \mathbb{R}^{p\times p}$ pos. def.

PROP $\forall b, d \in \mathbb{R}^p$, $(b'd)^2 \le (b'Bb)(d'B^{-1}d)$

proof $(b'd)^2 = (\underbrace{b'B^{1/2}}_{w'}\underbrace{B^{-1/2}d}_{z})^2 \overset{CS}{\le} \|w\|^2\|z\|^2 = (b'B^{1/2}B^{1/2}b)(d'B^{-1/2}B^{-1/2}d) = (b'B^2b)(d'B^{-2}d)$ eq. holds if $b'B^{1/2} \in \mathcal{L}(B^{-1/2}d)$, $b \in \mathcal{L}(B^{-1}d)$

PROP $B\in\mathbb{R}^{p\times p}$ pos. def, $d\in\mathbb{R}^p \Rightarrow \max_{x\in\mathbb{R}^p, x\neq 0} \frac{(x'd)^2}{x'Bx} = d'B^{-1}d$

proof CS: $(x'd)^2 \le (x'Bx)(d'B^{-1}d)$

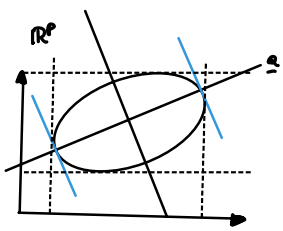
if $x\neq 0 \Rightarrow x'Bx > 0$ (B pos. def) $\Rightarrow \frac{(x'd)^2}{x'Bx} \le d'B^{-1}d \quad \forall x\neq 0$

equality if $x \in \mathcal{L}(B^{-1}d)$

For $a\in\mathbb{R}^p$ $\max_{a\in\mathbb{R}^p, a\neq 0} \frac{(a'(\bar X-\mu))^2}{a'Sa} n = n(\bar X-\mu)S^{-1}(\bar X-\mu) \sim \frac{(n-1)p}{n-p}F(p,n-p)$ (we know it's distribution (Corollary to Hotelling's theorem))

Problem Find $c>0$ s.t. $1-\alpha = \mathbb{P}\left(\frac{|a'\bar X - a'\mu|}{\sqrt{a'Sa}}\sqrt{n} < c, \forall a\in\mathbb{R}^p\right) = \mathbb{P}\left(\frac{(a'(\bar X-\mu))^2}{a'Sa} n < c^2, \forall a\in\mathbb{R}^p\right) = \mathbb{P}\left(\max_{a\in\mathbb{R}^p, a\neq 0}\frac{(a'(\bar X-\mu))^2}{a'Sa} < c^2\right)$

$\Rightarrow \boxed{c^2 = \frac{(n-1)p}{n-p}F_{1-\alpha}(p,n-p)}$ ☺



We can finally conclude: $\mathbb{P}\left[\frac{|a'\bar X - a'\mu|}{\sqrt{a'Sa}}\sqrt{n} < \sqrt{\frac{(n-1)p}{n-p}F_{1-\alpha}(p,n-p)}, \overbrace{\forall a\in\mathbb{R}^p}^{\text{simultaneous}}\right] = 1-\alpha$

which is the same as $\mathbb{P}\left[a'\mu \in \left[a'\bar X \pm \sqrt{\frac{(n-1)p}{n-p}F_{1-\alpha}(p,n-p)}\sqrt{\frac{a'Sa}{n}}\right], \forall a\in\mathbb{R}^p\right] = 1-\alpha$

from which we easily read the simultaneous confidence interval $\boxed{SimCI_{1-\alpha}(a'\mu) = \left[a'\bar X \pm \sqrt{\frac{(n-1)p}{n-p}F_{1-\alpha}(p,n-p)}\sqrt{\frac{a'Sa}{n}}\right]}$

LECTURE 12 22/3/2022

Thus now can take all the possible linear combination we want and we get a confidence interval that is globally correct $(1-\alpha)\%$ of times! Indeed all of the interval are correct with probability $1-\alpha$: all of the interval cover the linear combination $(1-\alpha)\%$ of the time they are used!

So if we have an ellipse of the $CR_{1-\alpha}(\mu)$ then its projection along any direction a gives us the simultaneous confidence interval $SimCI_{1-\alpha}(a^T\mu)$

BONFERRONI STRATEGY FOR SIMULTANEOUS CONFIDENCE INTERVAL

Using the previous method we cover ALL possible linear combinations: we'd like to fix a FINITE number!

Let us fix $a_1,\dots,a_k \in \mathbb{R}^p$: we want the k CI for $a_i'\mu$, $i=1,\dots,k$ with a simultaneous confidence of $1-\alpha$

Consider $CI_{1-\alpha}(a_i'\mu) = \left[a_i'\bar X \pm t_{1-\frac{\alpha}{2}}(n-1)\sqrt{\frac{a_i'Sa_i}{n}}\right]$ (one-at-the-time)

Then

$$\mathbb{P}\left[\bigcap_{i=1}^k \left[a_i'\mu \in CI_{1-\alpha}(a_i'\mu)\right]\right] = 1 - \mathbb{P}\left[\bigcup_{i=1}^k \left[a_i'\mu \notin CI_{1-\alpha}(a_i'\mu)\right]\right] \ge 1 - \sum_{i=1}^k \mathbb{P}\left[a_i'\mu \notin CI_{1-\alpha}(a_i'\mu)\right] = 1 - \sum_{i=1}^k \alpha = 1-k\alpha$$

($\mathbb{P}(A\cup B) = \mathbb{P}(A)+\mathbb{P}(B)-\mathbb{P}(A\cap B) \le \mathbb{P}(A)+\mathbb{P}(B)$ Bonferroni ineq.)

$\Rightarrow$ $\boxed{Bonf\,CI_{1-\alpha}(a_i'\mu) = \left[a_i'\bar X \pm t_{1-\frac{\alpha}{2k}}(n-1)\sqrt{\frac{a_i'Sa_i}{n}}\right]}$ Notice that $Bonf\,CI_{1-\alpha}(a_i'\mu) \xrightarrow{k\to\infty} [-\infty,+\infty]$, good for small k

Simultaneous Testing can be done but it's very conservative (with big data you'd never reject H_0)

H_0: $\begin{cases} H_{01} & a_1'\mu = \delta_1 \\ H_{02} & a_2'\mu = \delta_2 \\ \vdots & \vdots \\ H_{0k} & a_k'\mu = \delta_k \end{cases}$ vs H_1: at least one H_{0i} is false $\Rightarrow$ Reject H_0 if for at least one i $\frac{|a_i'\bar X - \delta_i|}{\sqrt{a_i'Sa_i}}\sqrt{n} > t_{1-\frac{\alpha}{2k}}(n-1)$

Indeed $\mathbb{P}[\text{Reject } H_0 \mid H_0 \text{ true}] = \mathbb{P}\left[\bigcup_{i=1}^{k}\{|t_i| > t_{1-\frac{\alpha}{2k}}(m-1)\} \mid \bigcap_{i=1}^{k} H_{0i} \text{ true}\right] \leq \sum_{i=1}^{k}\mathbb{P}[\dots] = \sum_{i=1}^{k}\frac{\alpha}{k} = \alpha$

So the overall probability of rejecting one H_{0i} when all are true is equal to α

FALSE DISCOVERY RATE (FDR)

(Benjamini & Hochberg 1995)

	citations
2014	21'000
2019	53'000
2022	83'821

} very important strategy for sim. testing!

Let $\mathcal{D}$ be any strategy for testing (eg. Bonferroni)

Decisions taken following $\mathcal{D}$

Truth	Not reject H_{0i}	Reject H_{0i}	
H_{0i}	U	V (false discoveries)	$k_0 = U+V$
H_{1i}	T (missed discoveries)	S (true discoveries)	$k-k_0 = T+S$
	$k-R$	R	

We can observe these

With Bonferroni strategy you reject H_{0i} at level $\frac{\alpha}{k}$

Let I_0 be the set of the true H_{0i}'s $|I_0| = k_0$

$$\mathbb{P}[V \geq 1] = \mathbb{P}\left[\bigcup_{j\in I_0}\{\text{Reject } H_{0j}\}\right] \leq \sum_{j\in I_0}\mathbb{P}[\{\text{Reject } H_{0j}\}] \leq \sum_{j\in I_0}\frac{\alpha}{k} = k_0\frac{\alpha}{k} \leq k\frac{\alpha}{k} = \alpha$$

Family Wise Error Rate (FWER)

We define $Q = \begin{cases} \frac{V}{R}, & R>0 \Leftrightarrow V>0 \\ 0, & R=0 \Leftrightarrow V=0 \end{cases}$ (Random) $\Rightarrow$ FALSE DISCOVERY RATE $\mathbb{E}[Q]$

REMARK $k=k_0$ $S=T=0$ $V=R$ there are no discoveries to be made

$Q = \{0,1\}$

FDR $= \mathbb{E}[Q] = \mathbb{P}[V>0] = \mathbb{P}[V\geq 1] =$ FWER

REMARK $k>k_0$ (*)

if $V=0 \Rightarrow Q=0$

if $V>0 \Rightarrow Q=\frac{V}{R}\leq 1$

moreover $\mathbb{1}[V>0] = \begin{cases} 0 & \text{if } V=0 \\ 1 & \text{if } V>0 \end{cases}$

$\Rightarrow Q \leq \mathbb{1}[V>0]$

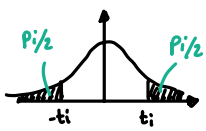
thus if (*) FDR $= \mathbb{E}[Q] \leq \mathbb{E}[\mathbb{1}[V>0]] = \mathbb{P}[V>0] = \mathbb{P}[V\geq 1] =$ FWER

We have a way to control FDR!

Conclusion: Bonferroni is good but not usable, *FDR* is less good but usable!

STRATEGY FOR CONTROLLING FDR

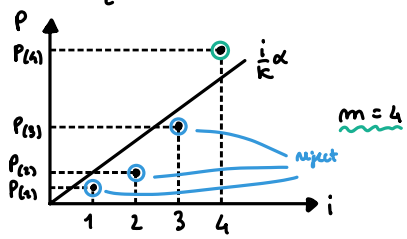
Let p_i be the p-value of the single test H_{0i} vs H_{1i}



$i = 1, \dots, k$

We order the p-values $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(k)}$ ($H_{0(1)}$, $H_{0(2)}$, $H_{0(k)}$)

$m = \max\left\{i \in \{1,\dots,k\} : p_{(i)} \leq \frac{i}{k}\alpha\right\}$ $\quad \alpha \in (0,1)$ $\quad \alpha, k$ fixed



THEOREM (B&H) If $p_1, \dots, p_k$ are independent $\Rightarrow$ The strategy which rejects $H_{0(i)}$ if $i \leq m$ controls FDR at level α (FDR $\leq \alpha$)

THEOREM (BENJAMINI-YEKUTELI 2001)

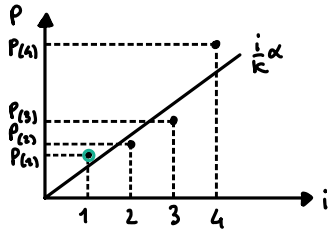
- If $p_1, \dots, p_k$ are positively correlated $\Rightarrow$ B&H 95 strategy controls FDR at level α
- If $p_1, \dots, p_k$ are negatively correlated $\Rightarrow$ diff. procedure: reject $H_{0(i)}$ if $i \leq m^* = \max\left\{j \in \{1,\dots,k\} : p_{(j)} \leq \frac{j}{C_{(k)}k}\alpha\right\}$

where $C_{(k)} = \sum_{j=1}^{k} \frac{1}{j}$

Note: mixed cases are not covered: we need both ALL positively correlated, or ALL negatively correlated p−values!

Efron's modification of B&H strategy

It might happen that $p_{(1)}$ is above the line, B&H would accept all



LECTURE 13 25/3/2022

COMPARISON OF MEAN

e.g. after a treatment

▶ REPEATED MEASURES

e.g. person, family

For each $i=1,\dots,m$ statistical unit, with $p \geq 1$ features, we make $q \geq 2$ measurements

$\underline{X}_{1i} = \begin{pmatrix} \text{height (time 1)} \\ \text{weight (time 1)} \\ \text{salary (time 1)} \end{pmatrix}$ $\xrightarrow{\text{treatment}}$ $\underline{X}_{2i} = \begin{pmatrix} H \text{ (time 2)} \\ W \text{ (time 2)} \\ S \text{ (time 2)} \end{pmatrix}$ ⇖ No dependence between

dependence between

▶ CASE $q=2$ (PAIRS)

$\mathbb{R}^p \ni \underline{\mu}_1 \leftarrow \underline{X}_{11}, \underline{X}_{12}, \dots, \underline{X}_{1m}$ (i=1, i=2, i=m)

$\mathbb{R}^p \ni \underline{\mu}_2 \leftarrow \underline{X}_{21}, \underline{X}_{22}, \dots, \underline{X}_{2m}$ (i=1, i=2, i=m)

$q=2$

REMARK ! "pairs" have to be meaningful, it doesn't make sense to pair a group of French and Italians based on height

The main question is ("did the treatment have an effect ?")

$$\{H_0: \underline{\mu}_1 = \underline{\mu}_2 \text{ vs } H_1: \underline{\mu}_2 \neq \underline{\mu}_1\} \Leftrightarrow \{H_0: \underline{\mu}_1 - \underline{\mu}_2 = \underline{0} \text{ vs } H_1: \underline{\mu}_1 - \underline{\mu}_2 \neq \underline{0}\}$$
$$\underline{\delta} = \underline{0} \qquad \underline{\delta} \neq \underline{0}$$

$\underline{D}_i = \underline{X}_{1i} - \underline{X}_{2i}$ $i = 1,\dots,m$

▶ Assume $\underline{D}_1,\dots,\underline{D}_m$ iid $\sim N_p(\underline{\delta}, \Sigma_D)$

▶ Test statistic $\bar{\underline{D}} = \frac{1}{m}\sum_{i=1}^{m} \underline{D}_i$ $S_D = \frac{1}{m-1}\sum_{i=1}^{m}(\underline{D}_i - \bar{\underline{D}})(\underline{D}_i - \bar{\underline{D}})'$

$\Rightarrow$ $m(\bar{\underline{D}} - \underline{\delta})' S_D^{-1} (\bar{\underline{D}} - \underline{\delta}) \sim \frac{(m-1)p}{m-p} F(p, m-p)$ ← pivotal quantity

- Test $H_0: \underline{\delta} = \underline{\delta}_0$ vs $H_1: \underline{\delta} \neq \underline{\delta}_0$ (e.g. $\underline{\delta}_0 = \underline{0}$)
  $T_0^2 = m(\bar{\underline{D}} - \underline{\delta}_0)' S_D^{-1} (\bar{\underline{D}} - \underline{\delta}_0) \Rightarrow \alpha \in (0,1)$ reject if $T_0^2 > \frac{(m-1)p}{m-p} F_{1-\alpha}(p, m-p)$
- Confidence Region $CR_{1-\alpha}(\underline{\delta}) = CR_{1-\alpha}(\underline{\mu}_1 - \underline{\mu}_2) = \{\underline{\eta} \in \mathbb{R}^p : m(\bar{\underline{D}} - \underline{\eta})' S_D^{-1} (\bar{\underline{D}} - \underline{\eta}) \leq \frac{(m-1)p}{m-p} F_{1-\alpha}(p, m-p)\}$
- Simultaneous Confidence Interval $\text{Sim CI}(\underline{a}'\underline{\delta}) = \left[\underline{a}'\bar{\underline{D}} \pm \sqrt{\frac{(m-1)p}{m-p} F_{1-\alpha}(p, m-p)} \sqrt{\frac{\underline{a}' S_D \underline{a}}{m}}\right]$
- Bonferroni Simultaneous Confidence Interval $\text{Bonf CI}_{1-\alpha}(\mu_{1j} - \mu_{2j}) = \left[\bar{D}_j \pm t_{1-\frac{\alpha}{2p}}(m-1)\sqrt{\frac{S_{jj}}{m}}\right]$

▶ CASE $p=1$ (1 feature, but many repeated measures)

For $i=1,\dots,m$ $p=1$ $\{(\overset{\mathbb{R}}{X_{1i}}, \overset{\mathbb{R}}{X_{2i}}, \dots, \overset{\mathbb{R}}{X_{qi}})' \sim N_q(\underline{\mu}, \Sigma)$ ▶ Assume Gaussianity

($q \geq 2$)

$H_0: \mu_1 = \mu_2 = \dots = \mu_q$ ($\underline{\mu} \in \mathcal{L}(\underline{1})$) vs $H_1: \exists i,j$ s.t. $\mu_i \neq \mu_j$ ($\underline{\mu} \notin \mathcal{L}(\underline{1})$)

Different approach than before

DEF $C \in \mathbb{R}^{(q-1)\times q}$ is called CONTRAST MATRIX if $C = \begin{bmatrix} \underline{c}_1' \\ \vdots \\ \underline{c}_{q-1}' \end{bmatrix}$, $\underline{c}_i \in \mathbb{R}^q$

1) $\underline{c}_1,\dots,\underline{c}_{q-1}$ are lin. indep
2) $\underline{c}_i'\underline{1} = 0$ $\forall i = 1,\dots,q-1$ (i.e. $\underline{c}_i \perp \underline{1}$)

} i.e. $\underline{c}_1,\dots,\underline{c}_{q-1}$ are a basis for $\mathcal{L}^{\perp}(\underline{1})$

$H_0: C\underline{\mu} = \underline{0}$ vs $H_1: C\underline{\mu} \neq \underline{0}$

$\bar{\underline{X}} = \frac{1}{n}\sum_{i=1}^{n} \underline{X}_i \qquad \underline{X}_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{qi} \end{pmatrix} \overset{iid}{\sim} N_{q-1}(\underline{\mu}, \Sigma) \qquad C\bar{\underline{X}} \sim N_{q-1}(C\underline{\mu}, \frac{1}{n} C\Sigma C') \Rightarrow \sqrt{n}(C\bar{\underline{X}} - C\underline{\mu}) \sim N_{q-1}(\underline{0}, C\Sigma C')$

$S = \frac{1}{n-1}\sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})' \Rightarrow (n-1)S \sim \text{Wishart}(\Sigma, n-1) \Rightarrow (n-1)CSC' \sim \text{Wish}(C\Sigma C', n-1)$

$\Rightarrow$ by Hotelling's theorem $n(C\bar{\underline{X}} - C\underline{\mu})'(CSC')^{-1}(C\bar{\underline{X}} - C\underline{\mu}) \sim \frac{(n-1)(q-1)}{n-q+1} F(q-1, n-q+1)$ → pivotal quantity

- Test reject $H_0: C\underline{\mu} = \underline{0}$ at level α if $T_0^2 = n(C\bar{\underline{X}})'(CSC')^{-1}(C\bar{\underline{X}}) > \frac{(n-1)(q-1)}{n-q+1} F_{1-\alpha}(q-1, n-q+1)$

...

⚠ I could choose many different bases

Example

$C_1 = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 1 & \cdots & & & -1 \end{bmatrix}$ $\quad C_2 = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & & 1 & -1 \end{bmatrix}$

comparing with baseline $\quad$ compare with previous

Notice $\exists B \in \mathbb{R}^{(q-1)\times(q-1)}$ s.t. $\det(B) \neq 0$ and $C_1 = BC_2 \Rightarrow$ ↳ maps a basis into another

$T_0^2 = n(BC_2\bar{\underline{X}})'(BC_2SC_2'B')^{-1}(BC_2\bar{\underline{X}})$
$= n(C_2\bar{\underline{X}})'B'(B')^{-1}(C_2SC_2')B^{-1}B(C_2\bar{\underline{X}})$ same!

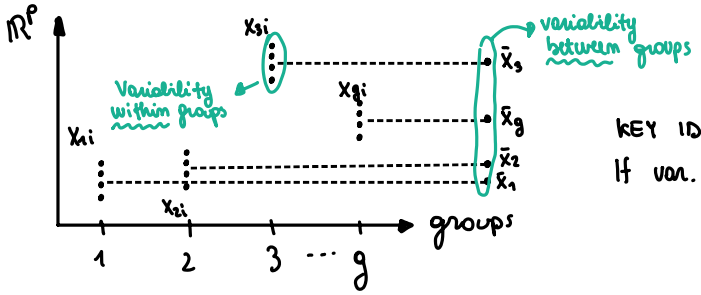
▶ MANOVA (MULTI-VARIATE ANALYSIS OF VARIANCE)

$m_1 \cdots m_g$ observations

g indep. groups $\begin{bmatrix} \underline{X}_{11}, \ldots, \underline{X}_{1m_1} \overset{iid}{\sim} N_p(\underline{\mu}_1, \Sigma) \\ \underline{X}_{21}, \ldots, \underline{X}_{2m_2} \overset{iid}{\sim} N_p(\underline{\mu}_2, \Sigma) \\ \vdots \\ \underline{X}_{g1}, \ldots, \underline{X}_{gm_g} \overset{iid}{\sim} N_p(\underline{\mu}_g, \Sigma) \end{bmatrix}$ / p features

GOAL: make inference on the differences of means among groups: $\underline{\mu}_1, \ldots, \underline{\mu}_g$

Σ have to be equal
- do testing for this
- if not, transform the data until yes



KEY IDEA:
If var. between > var. within $\Rightarrow$ means are different (i.e. treatment had effect)

▶ CASE $g=2$ $p \geq 1$ (2 groups, many features)

▶ Assume gaussianity

$\bar{\underline{X}}_1 \sim N_p(\underline{\mu}_1, \frac{1}{m_1}\Sigma)$, $\bar{\underline{X}}_2 \sim N_p(\underline{\mu}_2, \frac{1}{m_2}\Sigma)$ $\Rightarrow \bar{\underline{X}}_1 - \bar{\underline{X}}_2 \sim N_p(\underline{\mu}_1 - \underline{\mu}_2, (\frac{1}{m_1} + \frac{1}{m_2})\Sigma) \Rightarrow (\frac{1}{m_1} + \frac{1}{m_2})^{-1/2}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2 - (\underline{\mu}_1 - \underline{\mu}_2)) \sim N_p(\underline{0}, \Sigma)$

$S_1 = \frac{1}{m_1 - 1}\sum_{i=1}^{m_1}(\underline{X}_{1i} - \bar{\underline{X}}_1)(\underline{X}_{1i} - \bar{\underline{X}}_1)'$ est. of $\Sigma_1 \Rightarrow (m_1-1)S_1 \sim \text{Wish}(\Sigma, m_1-1)$

$S_2 = \frac{1}{m_2 - 1}\sum_{i=1}^{m_2}(\underline{X}_{2i} - \bar{\underline{X}}_2)(\underline{X}_{2i} - \bar{\underline{X}}_2)'$ est. of $\Sigma_2 \Rightarrow (m_2-1)S_2 \sim \text{Wish}(\Sigma, m_2-1)$

$\Rightarrow (m_1-1)S_1 + (m_2-1)S_2 \sim \text{Wish}(\Sigma, m_1+m_2-2)$

$S_{pooled} = \frac{(m_1-1)S_1 + (m_2-1)S_2}{m_1+m_2-2}$ s.t. $(m_1+m_2-2)S_{pooled} \sim \text{Wish}(\Sigma, m_1+m_2-2)$

$\bar{\underline{X}}_2 \perp S_{pooled} \perp \bar{\underline{X}}_1$ since $S_1 \perp \bar{\underline{X}}_1$, $S_2 \perp \bar{\underline{X}}_2 \Rightarrow (m_1+m_2-2)S_{pooled} \perp (\frac{1}{m_1} + \frac{1}{m_2})^{-1/2}[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) - (\underline{\mu}_1 - \underline{\mu}_2)]$

Hotelling $\Rightarrow$ $(\frac{1}{m_1} + \frac{1}{m_2})^{-1}[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) - (\underline{\mu}_1 - \underline{\mu}_2)]' S_{pooled}^{-1} [(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) - (\underline{\mu}_1 - \underline{\mu}_2)] \sim \frac{(m_1+m_2-2)p}{m_1+m_2-1-p} F(p, m_1+m_2-1-p)$ → pivotal

- Test $H_0: \underline{\mu}_1 - \underline{\mu}_2 = \underline{\delta}_0$ vs $H_1: \underline{\mu}_1 - \underline{\mu}_2 \neq \underline{\delta}_0$ We reject at level $\alpha \in (0,1)$ if

$$T_0^2 = \left(\frac{1}{m_1}+\frac{1}{m_2}\right)^{-1}\left[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) - \underline{\delta}_0\right]' S_{pooled}^{-1}\left[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) - \underline{\delta}_0\right] > \frac{(m_1+m_2-2)p}{m_1+m_2-1-p} F_{1-\alpha}(p, m_1+m_2-1-p)$$

- Confidence Region

$$CR_{1-\alpha}(\mu_1-\mu_2) = \left\{ \underline{\eta} \in \mathbb{R}^p : \left(\frac{1}{m_1}+\frac{1}{m_2}\right)^{-1}\left[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) - \underline{\delta}\right]' S_{pooled}^{-1}\left[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) - \underline{\delta}\right] < \frac{(m_1+m_2-2)p}{m_1+m_2-1-p} F_{1-\alpha}(p, m_1+m_2-1-p) \right\}$$

- Simultaneous Confidence interval $Sim\,CI_{1-\alpha}(\underline{a}'(\mu_1-\mu_2))$ $\forall \underline{a}$ ...
- Bonferroni Simultaneous Confidence interval $Bonf\,CI_{1-\alpha}$ ...

▶ SPECIAL CASE BEHRENS-FISHER PROBLEM

$$\Downarrow \quad \underline{X}_{11}, \dots, \underline{X}_{1m_1} \overset{iid}{\sim} N_p(\underline{\mu}_1, \Sigma_1)$$
$$\underline{X}_{21}, \dots, \underline{X}_{2m_2} \overset{iid}{\sim} N_p(\underline{\mu}_2, \Sigma_2)$$

$H_0: \Sigma_1 = \Sigma_2$ vs $H_1: \Sigma_1 \neq \Sigma_2$ Here we don't want to reject!
- some tests based on Gaussianity (Extended Levene Test)
- Permutation tests (nonparametric)

If m_1, m_2 are large enough we don't need Gaussianity
$\Rightarrow$ by CLT $\bar{\underline{X}}_1 \sim N_p(\underline{\mu}_1, \frac{1}{m_1}\Sigma_1)$ $\Rightarrow$ $\bar{\underline{X}}_1 - \bar{\underline{X}}_2 \sim N_p(\underline{\mu}_1 - \underline{\mu}_2, \frac{1}{m_1}\Sigma_1 + \frac{1}{m_2}\Sigma_2)$
$\bar{\underline{X}}_2 \sim N_p(\underline{\mu}_1, \frac{1}{m_2}\Sigma_2)$

Therefore $\left(\frac{1}{m_1}+\frac{1}{m_2}\right)^{-1}\left[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) - (\underline{\mu}_1 - \underline{\mu}_2)\right]' \left[\frac{1}{m_1}\Sigma_1 + \frac{1}{m_2}\Sigma_2\right]^{-1}\left[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) - (\underline{\mu}_1 - \underline{\mu}_2)\right] \sim \chi^2(p)$

(Σ_1, Σ_2: unknown but $S_1 \to \Sigma_1$, $S_2 \to \Sigma_2$ in $\mathbb{P}$ and thus $\mathcal{L}$)

So our piv. quantity is

$$\left(\frac{1}{m_1}+\frac{1}{m_2}\right)^{-1}\left[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) - (\underline{\mu}_1 - \underline{\mu}_2)\right]' \left[\frac{1}{m_1}S_1 + \frac{1}{m_2}S_2\right]^{-1}\left[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) - (\underline{\mu}_1 - \underline{\mu}_2)\right] \sim \chi^2(p)$$

LECTURE 14 29/3/2022

▶ CASE $g \geq 2$ $p = 1$ (many groups, 1 feature) → ANOVA case, not multivariate
(one-way)
↳ 1 feature

$$\Downarrow \begin{cases} X_{11}, \dots, X_{1m_1} \sim N(\mu_1, \sigma^2) \\ \vdots \\ X_{g1}, \dots, X_{gm_g} \sim N(\mu_g, \sigma^2) \end{cases}$$

Goals:
- $H_0: \mu_1 = \dots = \mu_g$ (e.g. fertiliser has no effect) vs $H_1: \exists i,j : \mu_i \neq \mu_j$ (e.g. has some effect)
- If reject H_0 $\Rightarrow$ estimate the μ_i

New parametrisation of the problem: $\mu_i = \mu + \tau_i$ (μ: overall mean; τ_i: how diff. group i is from the mean)

$H_0: \tau_1 = \dots = \tau_g = 0$ vs $H_1: \exists \tau_i \neq 0$

$g \mapsto g+1$ parameters (μ, τ_i)
$\Rightarrow$ we need one more constraint! (*)

Suppose $m_1 + \dots + m_g = m$, $\bar{X} = \frac{1}{m}\sum_{i=1}^{g}\sum_{j=1}^{m_i} X_{ij}$

$$\mathbb{E}[\bar{X}] = \frac{1}{m}\sum_{i=1}^{g}\sum_{j=1}^{m_i}(\mu+\tau_i) = \frac{1}{m}\sum_{i=1}^{g}(m_i\mu + \tau_i m_i) = \mu\frac{1}{m}\sum_{i=1}^{g} m_i + \frac{1}{m}\sum_{i=1}^{g} m_i\tau_i = \mu + \frac{1}{m}\sum_{i=1}^{g} m_i\tau_i$$

$\bar{X}$ unbiased $\Leftrightarrow$ $\mathbb{E}[\bar{X}] = \mu$ $\Leftrightarrow$ $\sum_{i=1}^{g} m_i\tau_i = 0$ (*)

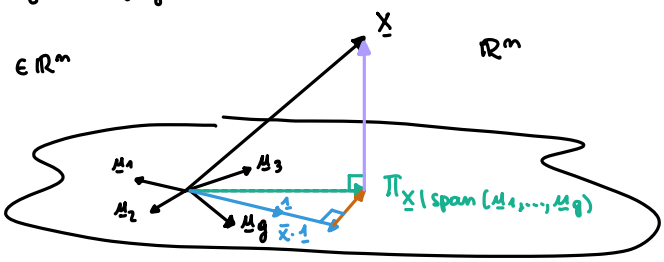
Est. for τ_i? $\mathbb{E}[\bar{X}_i - \bar{X}] = \mathbb{E}[\bar{X}_i] - \mu = \frac{1}{m_i}\sum_{j=1}^{m_i}\underbrace{\mathbb{E}[X_{ij}]}_{\mu+\tau_i} - \mu = \mu + \tau_i - \mu = \tau_i$

($\bar{X}_i$: mean of group i; $\bar{X}$: overall mean; $\sum_{j=1}^{m_i}(\mu+\tau_i) = m_i(\mu+\tau_i)$)

DECOMPOSITION OF VARIANCE

Let $\mathbb{R}^m \ni \underline{X} = (x_{11}, x_{12}, \dots, x_{1m_1}, x_{21}, \dots, x_{2m_2}, \dots, x_{g1}, \dots, x_{gm_g})'$

$$\underline{u}_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \} m_1 \qquad \underline{u}_2 = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \} m_2 \qquad \underline{u}_g = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \} m_g$$



REMARK

1) $\underline{u}_1, \dots, \underline{u}_g$ are lin. indep
2) $\underline{u}_i' \underline{u}_j = m_i \delta_{ij} = m_j \delta_{ij} = \begin{cases} 0 & i \neq j \\ m_i & i = j \end{cases}$
3) $\mathbb{R}^m \ni \underline{1} \in \text{span}(\underline{u}_1, \dots, \underline{u}_g)$

$$\pi_{\underline{X} | \text{span}(\underline{u}_1, \dots, \underline{u}_g)} = \sum_{i=1}^{g} \frac{\underline{u}_i (\underline{u}_i' \underline{X})}{\underline{u}_i' \underline{u}_i} = \sum_{i=1}^{g} \left(\frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij} \right) \cdot \underline{u}_i = \underline{\sum_{i=1}^{g} \bar{x}_i \cdot \underline{u}_i}$$

$$\pi_{\underline{X} | \underline{1}} = \bar{x} \cdot \underline{1} \quad (\bar{x} \in \mathbb{R})$$

$$\pi_{\sum_{i=1}^{g} \bar{x}_i \cdot \underline{u}_i | \underline{1}} = \frac{\underline{1}\underline{1}'}{\underline{1}'\underline{1}} \left(\sum_{i=1}^{g} \bar{x}_i \cdot \underline{u}_i \right) = \sum_{i=1}^{g} \bar{x}_i \cdot \frac{\underline{1}\underline{1}'}{\underline{1}'\underline{1}} \underline{u}_i = \frac{1}{m} \sum_{i=1}^{g} m_i \bar{x}_i \cdot \underline{1} = \left(\frac{1}{m} \sum_{i=1}^{g} \sum_{j=1}^{m_i} x_j \right) \underline{1} = \bar{x} \cdot \underline{1}$$

$$\Rightarrow \underline{X} = \bar{x} \cdot \underline{1} + \sum_{i=1}^{g} (\bar{x}_i - \bar{x}) \underline{u}_i + \sum_{i=1}^{g} \sum_{j=1}^{m_i} (x_{ij} - \bar{x}_i) \underline{u}_i$$

all ORTHOGONAL!

$$\Rightarrow \|\underline{X}\|^2 = \|\bar{x} \cdot \underline{1}\|^2 + \left\| \sum_{i=1}^{g} (\bar{x}_i - \bar{x}) \underline{u}_i \right\|^2 + \left\| \underline{X} - \sum_{i=1}^{g} \bar{x}_i \underline{u}_i \right\|^2$$

$$\underbrace{\sum_{i=1}^{g} \sum_{j=1}^{m_i} x_{ij}^2}_{SS_{OBS}} = \underbrace{m \bar{x}^2}_{SS_{MEAN}} + \underbrace{\sum_{i=1}^{g} (\bar{x}_i - \bar{x})^2 m_i}_{SS_{TREATMENT}} + \underbrace{\sum_{i=1}^{g} \sum_{j=1}^{m_i} (x_{ij} - \bar{x}_i)^2}_{SS_{RESIDUALS}}$$

↳ not captured by the lin. comb. of $\underline{u}_1, \dots, \underline{u}_g$ with $\underline{X}$

Which is equivalent to

$$\underbrace{\sum_{i=1}^{g} \sum_{j=1}^{m_i} (x_{ij} - \bar{x})^2}_{SS_{CENTERED}} = \underbrace{\sum_{i=1}^{g} (\bar{x}_i - \bar{x})^2 m_i}_{SS_{TREATMENT}} + \underbrace{\sum_{i=1}^{g} \sum_{j=1}^{m_i} (x_{ij} - \bar{x}_i)^2}_{SS_{RESIDUALS}}$$

IDEA: Reject H_0 if $\frac{SS_{TREAT}}{SS_{RES}} = (\star)$ is large

If H_0 is true, $(\star) \sim$?

- We know that $\underline{X}$ is Gaussian $\Rightarrow$ each term is Gaussian
- they are also orthogonal $\Rightarrow$ they are independent
- In $(\star)$ we have the quotient of two squared normals $\Rightarrow$ Fisher (χ^2)

Focus on $\sum_{i=1}^{g} \sum_{j=1}^{m_i} (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{g} (m_i - 1) S_i^2 \sim \sigma^2 \chi^2 \left(\sum_{i=1}^{g} (m_i - 1) \right) \sim \sigma^2 \chi^2 (m - g)$

↳ $S_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m} (x_{ij} - \bar{x}_i)$ est. of σ^2 of group i $\sim \sigma^2 \chi^2 (m_i - 1)$

if H_0 is true $\Rightarrow \mu_1 = \mu_2 = \dots = \mu_g = \mu \Rightarrow \frac{1}{m-1} \sum_{i=1}^{g} \sum_{j=1}^{m_i} (x_{ij} - \bar{x})^2 = S^2$ is an estimator of σ^2

$$\Rightarrow (m-1) S^2 = \sum_{i=1}^{g} \sum_{j=1}^{m_i} (x_{ij} - \bar{x})^2 \sim \sigma^2 \chi^2 (m-1)$$

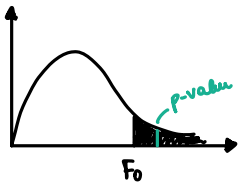
Thus we can conclude:

$$SS_{centred} = SS_{treat} + SS_{res}$$
$$\sim \sigma^2 \chi^2 (m-1) \quad \sim \chi^2 (g-1) \quad \sim \sigma^2 \chi^2 (m-g)$$

$$F_0 = \frac{\sum_{i=1}^{g} m_i (\bar{X}_i - \bar{X})^2 / g-1}{\sum_{i=1}^{g}\sum_{j=1}^{m_i} (X_{ij} - \bar{X}_i)^2 / m-g} \sim F(g-1, m-g)$$

↘ pivotal

Reject H_0 if $F_0 > F_{1-\alpha}(g-1, m-g)$



GENERAL CASE $p \geq 1$, $g \geq 2$

$\mathbb{R}^p \ni \underline{X}_{ij} = \underline{\mu} + \underline{\tau}_i + \underline{\varepsilon}_{ij}$

where

- $\underline{\varepsilon}_{ij} \overset{iid}{\sim} N_p(\underline{0}, \Sigma)$
- $\underline{\mu} \in \mathbb{R}^p$
- $\underline{\tau}_1, \dots, \underline{\tau}_g \in \mathbb{R}^p$ s.t. $\sum_{i=1}^{g} m_i \underline{\tau}_i = \underline{0}$
- $\hat{\underline{\mu}} = \frac{1}{m}\sum_{i=1}^{g}\sum_{j=1}^{m_i} \underline{X}_{ij} = \bar{\underline{X}}$ unbiased
- $\hat{\underline{\tau}}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} \underline{X}_{ij} - \bar{\underline{X}} = \bar{\underline{X}}_i - \bar{\underline{X}}$ unbiased

Model

Goal: $H_0: \underline{\mu}_1 = \dots = \underline{\mu}_g$ vs $H_1: \exists i,j \ \underline{\mu}_i \neq \underline{\mu}_j$

same as: $H_0: \underline{\tau}_1 = \dots = \underline{\tau}_g = \underline{0}$ vs $H_1: \exists \underline{\tau}_i \neq \underline{0}$

REMARK Fix $k \in \{1, \dots, p\}$ (solary) $\Rightarrow$ each $X_{ijk} = \mu_k + \tau_{ik} + \varepsilon_{ijk}$ $\sim N_1(0, \sigma_{kk})$ is an ANOVA model

Why don't we just "press the button" p times? CORRELATION!

Decomposition of covariance

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{m_i} (\underline{X}_{ij} - \bar{\underline{X}})(\underline{X}_{ij} - \bar{\underline{X}})'}_{\text{Cov}_{TOT}} = \underbrace{\sum_{i=1}^{g} (\bar{\underline{X}}_i - \bar{\underline{X}})(\bar{\underline{X}}_i - \bar{\underline{X}})' m_i}_{\text{Cov}_{BETWEEN} = B} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{m_i} (\underline{X}_{ij} - \bar{\underline{X}}_i)(\underline{X}_{ij} - \bar{\underline{X}}_i)'}_{\text{Cov}_{WITHIN} = W}$$

proof use $\underline{X}_{ij} - \bar{\underline{X}} = (\bar{\underline{X}}_i - \bar{\underline{X}}) + (\underline{X}_{ij} - \bar{\underline{X}}_i)$

LECTURE 15 31/3/2022

By analogy with ANOVA we'd like consider B/W but we can't take the ratio of 2 matrices!

- B: $\text{rank}(B) \leq s = \min(g-1, p)$ (If we have 3 groups, $p = 100$, B 100×100 with rank 2)
- $W = \sum_{i=1}^{g} (m_i - 1) S_i$
  $= \frac{1}{m_i - 1}\sum_{j=1}^{m_i} (\underline{X}_{ij} - \bar{\underline{X}}_i)(\underline{X}_{ij} - \bar{\underline{X}}_i)'$ est of Σ
  $\Downarrow$
  $\frac{1}{m-g} W = \frac{1}{m-g}\sum_{i=1}^{g} m_i S_i$ est. of Σ

- Wilks' Λ proposal: test statistic $\boxed{\Lambda_W = \frac{\det(W)}{\det(B+W)}}$ large $\Leftrightarrow$ B is not adding much to W

$\Rightarrow$ Reject H_0 if Λ_W is small

If H_0 is true the distribution of Λ_W is known as long as

- $p \geq 1$, $g = 2, 3$
- $p = 2$, $g \geq 1$
- as $m \to \infty$: $\underbrace{-\left(m - 1 - \frac{p+g}{2}\right)\log(\Lambda_W)}_{\text{reject } H_0 \text{ if } > \chi^2_\alpha(p(g-1))} \sim \chi^2(p(g-1))$ (Bartlett's)

Note that since we have a minus in front of the pivotal statistics, we are taking $-\log$: so we reject for small value of Λ_W, so we reject for values of $-\log \Lambda_W$ which are big.

Rejecting H_0

If H_0 is rejected we need to compare $\underline{\tau}_i, \underline{\tau}_k$ component-wisely using Bonferroni's simultaneous confidence intervals for $\tau_{il} - \tau_{kl}$ $i,k = 1, \dots, g$ to see when there was difference $l = 1, \dots, p$

An estimator is $(\bar{X}_{il} - \bar{X}_{\ell}) - (\bar{X}_{kl} - \bar{X}_{\ell}) = \bar{X}_{il} - \bar{X}_{kl} \sim N(\tau_{il} - \tau_{kl}, \sigma_{ll}(\frac{1}{m_i} + \frac{1}{m_k}))$

$(\bar{X}_{il} - \bar{X}_{\ell})$: est of τ_{il}; $(\bar{X}_{kl} - \bar{X}_{\ell})$: est of τ_{kl}

↑ Not pivotal, estimate of σ_{ll}: $\frac{1}{m-g} W_{ll}$

$pg(g-1)\frac{1}{2}$ SBCI at level $1-\alpha$

$$SBCI_{1-\alpha}(\tau_{il} - \tau_{kl}) = \left[\bar{X}_{il} - \bar{X}_{kl} \pm t_{\frac{\alpha}{2}, \frac{2}{pg(g-1)}(m-g)} \sqrt{\frac{W_{ll}}{m-g}\left(\frac{1}{m_i} + \frac{1}{m_k}\right)}\right]$$

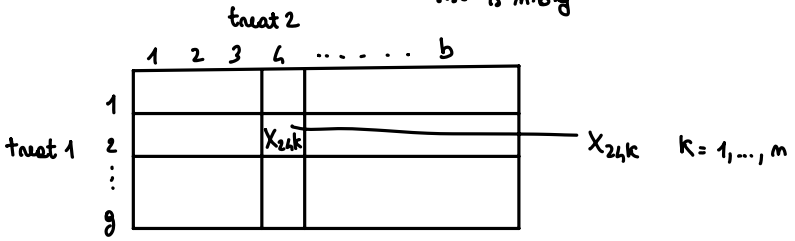
NB If $p=1$ we get SBCI in ANOVA

▶ SHORT EXCURSUS ON TWO-WAY ANOVA

{ treatment 1: g levels $\{1, 2, \dots, g\}$ factor 1
{ treatment 2: b levels $\{1, 2, \dots, b\}$ factor 2

We observe $X_{ijk} \in \mathbb{R}$ $i = 1, \dots, g$; $j = 1, \dots, b$; $k = 1, \dots, m$ (balanced)

↳ total sample size is $m \cdot b \cdot g$



Model

$X_{ijk} = \mu_{ij} + \underbrace{\varepsilon_{ijk}}_{\overset{iid}{\sim} N(0, \sigma^2)}$

$\mu_{ij} = \mu + \tau_i + \beta_j + \gamma_{ij}$ (gb | 1 | g | b | gb)

μ: effect of I treat; τ_i: effect of II treat; γ_{ij}: effect of interaction

Constraints

$0 = \sum_{i=1}^{g} \tau_i$ (1) $\quad 0 = \sum_{j=1}^{b} \beta_j$ (1) $\quad \sum_{j=1}^{b} \gamma_{ij} = 0$ (g) $\quad \sum_{i=1}^{g} \gamma_{ij} = 0$ (b)

(g+b-1)

⇒ in this way we don't overparametrize!

Estimators

$\hat{\mu} = \bar{X} = \frac{1}{mgb} \sum_{ijk} X_{ijk}$

$\bar{X}_{i\cdot} = \frac{1}{bm} \sum_{jk} X_{ijk}$

$\bar{X}_{\cdot j} = \frac{1}{mg} \sum_{ik} X_{ijk}$

$\hat{\tau}_i = \bar{X}_{i\cdot} - \bar{X}$

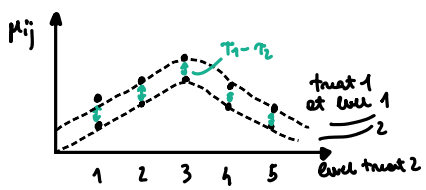
$\hat{\beta}_j = \bar{X}_{\cdot j} - \bar{X}$

$\hat{\gamma}_{ij} = \bar{X}_{ij} - (\bar{X}_{i\cdot} - \bar{X}) - (\bar{X}_{\cdot j} - \bar{X}) - \bar{X} = \bar{X}_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}$

$\bar{X}_{ij} = \frac{1}{m}\sum_{ij} X_{ijk}$: $\not{j}$

REMARK

Suppose $\gamma_{ij} = 0 \ \forall i,j \Rightarrow X_{ijk} = \underbrace{\mu + \tau_i + \beta_j}_{\mu_{ij}} + \varepsilon_{ijk}$

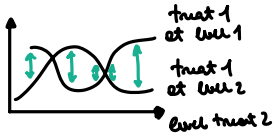


$\mu_{1j} = \mu + \tau_1 + \beta_j$

$\mu_{2j} = \mu + \tau_2 + \beta_j$

Additive model

If there are interactions → TEST FOR INTERACTIONS!



Complete model

then:

$H_0: \gamma_{ij} = 0 \ \forall i,j$	vs	$H_1: \exists \gamma_{ij} \neq 0$	(A)	
$H_0: \tau_i = 0 \ \forall i$	vs	$H_1: \exists \tau_i \neq 0$	(B)	GOAL!
$H_0: \beta_j = 0 \ \forall j$	vs	$H_1: \exists \beta_j \neq 0$	(C)	

Decomposition of variance:

		Dof
$\sum_{i=1}^{g}\sum_{j=1}^{b}\sum_{k=1}^{m}(x_{ijk}-\bar{x})^2$	$SS_{centred}$	$gbm-1$
$= \sum_{i=1}^{g}(\bar{x}_{i\cdot}-\bar{x})^2 bm$	$SS_{treat\,1}$	$g-1$
$+ \sum_{j=1}^{b}(\bar{x}_{\cdot j}-\bar{x})^2 gm$	$SS_{treat\,2}$	$b-1$
$+ \sum_{i=1}^{g}\sum_{j=1}^{b}(\bar{x}_{ij}-\bar{x}_{j\cdot}-\bar{x}_{\cdot i}+\bar{x})^2 m$	$SS_{interaction}$	$(g-1)(b-1)$
$+ \sum_{i=1}^{g}\sum_{j=1}^{b}\sum_{k=1}^{m}(x_{ijk}-\bar{x}_{ij})^2$	$SS_{residuals}$	$gb(m-1)$

if $m=1$ (possible) overfitting! To take into account variability we want $m \geq 2$
As m is greater, the "meter" is more precise

(A) Reject at level α if $\dfrac{\frac{1}{(g-1)(b-1)} SS_{interaction}}{\frac{1}{gb(m-1)} SS_{residuals}} > F_{1-\alpha}\left((g-1)(b-1), gb(m-1)\right)$

If H_0 not rejected (H_0 true)
the model is $X_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk}$ additive! (simpler) ⇒ in this case $SS_{interaction}$ is neglected and becomes part of SS_{res} (also the Dof becomes $gb(m-1) + (g-1)(b-1)$)

(B) Suppose the model is additive
(C similar) Reject at level α if $\dfrac{SS_{treat\,1}/(g-1)}{(SS_{res}+SS_{inter})/(gb(m-1)+(g-1)(b-1))} > F_{1-\alpha}(g-1, gbm-b-g+1)$

Suppose the model is complete
Reject at level α if $\dfrac{SS_{treat\,1}/(g-1)}{(SS_{res})/(gb(m-1))} > F_{1-\alpha}(g-1, gb(m-1))$

LECTURE 16 1/4/2022

CLASSIFIERS

Each unit is represented by $(x_1, .., x_p)^t \in \mathcal{X}$ (e.g. $\mathbb{R}^p$) with features $L \in \{1, ..., g\}$
Goal: learn $\delta: \mathcal{X} \to \{1, ..., g\}$

Supervised classification

Training set

$$\mathbb{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mp} \end{bmatrix} \begin{bmatrix} \ell_1 \\ \vdots \\ \ell_m \end{bmatrix} \quad (L)$$

$x_{ij} \in \mathbb{R}$
$\ell_i \in \{1, ..., g\} \ \forall i$

Learn δ by using: • training set • model (discriminant analysis)

Unsupervised classification

Training set

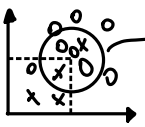
$$\mathbb{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mp} \end{bmatrix} \begin{bmatrix} \ell_1 \\ \vdots \\ \ell_m \end{bmatrix} \quad (L)$$

$x_{ij} \in \mathbb{R}$
$\ell_i \in \{1, ..., g\} \ \forall i$

Estimate $\hat{g}$ and $\hat{\ell}_1, ..., \hat{\ell}_m$ (Cluster analysis)
then ⇒ supervised classification (i.e. find $\delta: \mathcal{X} \to \{1, ..., \hat{g}\}$)

you assume they exist and they are unknown (neither the labels nor how many groups)

Ex KNN



$\mathbb{R}^p$ p large curse of dimensionality!

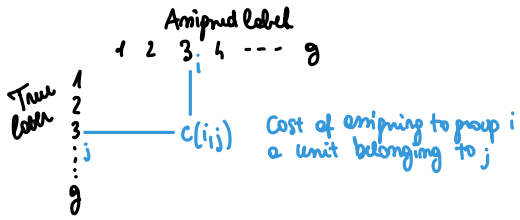
Ex Logistic regression

Supervised classification

Ingredients:

- $\underline{X} \mid L=i \sim f_i(\underline{x})$ density on $\mathbb{R}^p$
  they need to be different
- prior distribution
  $p_i = \mathbb{P}[L=i] \quad i = \{1,\dots,g\}$
  $p_i \geq 0 \quad \sum_{i=1}^{g} p_i = 1$

- Costs of misclassification



$c(i|j) \geq 0$
$c(i|i) = 0$
$c(i|j) \neq c(j|i)$ in general

REMARK $\delta: \mathbb{R}^p \to \{1,2,\dots,g\}$ is a class
$R_i = \{\underline{x} \in \mathbb{R}^p : \delta(\underline{x}) = i\} = \delta^{-1}(i)$
$\{R_1,\dots,R_g\} \leftrightarrow \delta$
↳ partition of $\mathbb{R}^p$

Optimality criterion δ is optimal if it minimizes $\mathbb{E}[\underbrace{\text{cost of misclassification}}_{CM}] = ECM$

Ex $g=2$

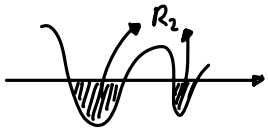
$\underline{x} \in \mathbb{R}^p \quad \delta(\underline{x}) = \begin{cases} 1 & \text{if } \underline{x} \in R_1 \\ 2 & \text{if } \underline{x} \in R_2 \end{cases}$

$$ECM(\delta) = \int_{R_1} c(1|2)\, f_2(\underline{x})\, p_2\, d\underline{x} + \int_{R_2} c(2|1)\, f_1(\underline{x})\, p_1\, d\underline{x}$$

R_1, R_2 → Unknowns

$$= \int_{\mathbb{R}^p} c(1|2) f_2(\underline{x}) p_2\, d\underline{x} - \int_{R_2} c(1|2) f_2(\underline{x}) p_2\, d\underline{x} + \int_{R_2} c(2|1) f_1(\underline{x}) p_1\, d\underline{x}$$

$$= c(1|2)\, p_2 + \int_{R_2} \underbrace{\left[c(2|1) f_1(\underline{x}) p_1 - c(1|2) f_2(\underline{x}) p_2 \right]}_{g(\underline{x})} d\underline{x}$$



$R_2 = \{\underline{x} \in \mathbb{R}^p \quad c(2|1) f_1(\underline{x}) p_1 \leq c(1|2) f_2(\underline{x}) p_2\}$
$R_1 = \{\underline{x} \in \mathbb{R}^p \quad c(2|1) f_1(\underline{x}) p_1 > c(1|2) f_2(\underline{x}) p_2\}$
↳ the equal sign doesn't matter

I will classify to group 2 all units such that if I make a mistake I pay less than what I would pay if I were to classify the same units in the other group and make a mistake

General case: $g \geq 2$

$$ECM(\delta) = \int_{R_1} \sum_{k \neq 1} c(1|k) f_k(\underline{x}) p_k\, d\underline{x} + \cdots \overbrace{\int_{R_i} \sum_{k \neq i} c(i|k) f_k(\underline{x}) p_k\, d\underline{x}}^{\text{general term}} + \cdots + \int_{R_g} [\dots]\, d\underline{x} = \sum_{i=1}^{g} \int_{R_i} \sum_{k \neq i} c(i|k) f_k(\underline{x}) p_k\, d\underline{x}$$

$\hat{\delta} = \underset{\delta}{\operatorname{argmin}}\ ECM(\delta)$ $\quad \hat{\delta} \leftrightarrow \{\hat{R}_1, \dots, \hat{R}_g\}$

$$\hat{R}_1 = \left\{ \underline{x} \in \mathbb{R}^p : \sum_{k \neq 1} c(1|k) f_k(\underline{x}) p_k \leq \sum_{k \neq j} c(j|k) f_k(\underline{x}) p_k, \ \forall j \in \{2, \dots, g\} \right\}$$

$$\hat{R}_i = \left\{ \underline{x} \in \mathbb{R}^p : \sum_{k \neq i} \underbrace{c(i|k)}_{w} f_k(\underline{x}) p_k \leq \sum_{k \neq j} \underbrace{c(j|k)}_{w} f_k(\underline{x}) p_k, \ \forall j \in \{1, \dots, g\} \setminus \{i\} \right\}$$

(w constant)

REMARK $\quad \underline{x} \in R_i \iff \dfrac{\sum_{k \neq i} c(i|k) f_k(\underline{x}) p_k}{\sum_{t=1}^{g} f_t(\underline{x}) p_t} \leq \dfrac{\sum_{k \neq j} c(j|k) f_k(\underline{x}) p_k}{\sum_{t=1}^{g} f_t(\underline{x}) p_t} \quad \forall j \in \{1, \dots, g\} \setminus \{i\} \iff (*)$

doesn't change!

$$\frac{\mathbb{P}(\underline{X}=\underline{x} \mid L=k)\mathbb{P}(L=k)}{\sum_{t=1}^{g} \mathbb{P}(\underline{X}=\underline{x} \mid L=t)\mathbb{P}(L=t)} = \frac{\mathbb{P}(\underline{X}=\underline{x} \mid L=k)\mathbb{P}(L=k)}{\mathbb{P}(\underline{X}=\underline{x})} = \mathbb{P}(L=k \mid \underline{X}=\underline{x})$$

Bayes

$$(*) \iff \sum_{k \neq i} c(i|k)\, \mathbb{P}(L=k \mid \underline{X}=\underline{x}) \leq \sum_{k \neq j} c(j|k)\, \mathbb{P}(L=k \mid \underline{X}=\underline{x}) \quad \forall j \in \{1, \dots, g\} \setminus \{i\}$$

Special cases

1) All costs are equal:

$c(i|j) = w > 0 \quad \forall i \neq j$ (default for a computer)

$c(i|i) = 0$

$$\Rightarrow \quad \underline{x} \in R_i \iff \sum_{k \neq i} \mathbb{P}(L=k \mid \underline{X}=\underline{x}) \leq \sum_{k \neq j} \mathbb{P}(L=k \mid \underline{X}=\underline{x}) \quad \forall j \neq i$$

$$\iff 1 - \mathbb{P}[L=i \mid \underline{X}=\underline{x}] \leq 1 - \mathbb{P}(L=j \mid \underline{X}=\underline{x}) \quad \forall j \neq i$$

$$\iff \mathbb{P}[L=i \mid \underline{X}=\underline{x}] \geq \mathbb{P}(L=j \mid \underline{X}=\underline{x}) \quad \forall j \neq i$$

(BAYES CLASSIFIER)

2) $c(i|j) = w > 0$ $\oplus$ $p_1 = p_2 = \dots = p_g = \frac{1}{g}$ (equally indifferent before looking at the data of which class the unit belongs to)

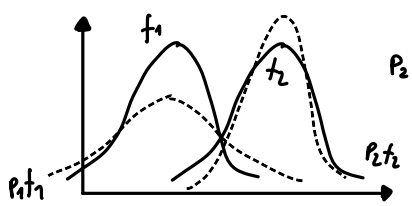
$c(i|i) = 0$

$$\Rightarrow \quad \underline{x} \in R_i \iff \mathbb{P}(L=i \mid \underline{X}=\underline{x}) \geq \mathbb{P}(L=j \mid \underline{X}=\underline{x})$$

↳ Abused in practice!

$$\iff \frac{f_i(\underline{x}) p_i}{\sum f_t(\underline{x}) p_t} \geq \frac{f_j(\underline{x}) p_j}{\sum f_t(\underline{x}) p_t}$$

$$\iff f_i(\underline{x}) \geq f_j(\underline{x}) \quad \forall j \neq i$$



$p_2 > p_1$ → the prior makes you more conservative

LECTURE 17 5/4/2022

3) QDA (Quadratic Discriminant analysis)

Hypothesis:

- $c(i|j) = w > 0 \quad \forall j \neq i$
- $f_i(\underline{x}) = \dfrac{1}{\sqrt{(2\pi)^p |\Sigma_i|}} \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{\mu}_i)' \Sigma^{-1} (\underline{x} - \underline{\mu}_i) \right\} \quad \underline{x} \in \mathbb{R}^p, \ \Sigma_i \text{ pos. def.}, \ \underline{\mu}_i \in \mathbb{R}^p$

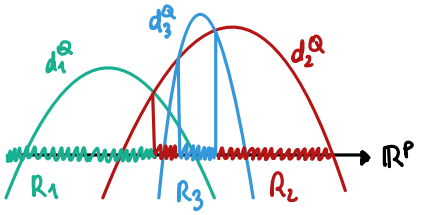
$$\underline{x} \in R_i \iff f_i(\underline{x}) p_i \geq f_j(\underline{x}) p_j \quad \forall j \neq i$$

$$\iff p_i \frac{1}{\sqrt{(2\pi)^p |\Sigma_i|}} \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{\mu}_i)' \Sigma_i^{-1} (\underline{x} - \underline{\mu}_i) \right\} \geq p_j \frac{1}{\sqrt{(2\pi)^p |\Sigma_j|}} \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{\mu}_j)' \Sigma_j^{-1} (\underline{x} - \underline{\mu}_j) \right\} \quad \forall j \neq i$$

$$\Leftrightarrow \quad \log p_i - \tfrac{1}{2}\log|\Sigma_i| - \underbrace{\tfrac{1}{2}(\underline{x}-\underline{\mu}_i)'\Sigma_i^{-1}(\underline{x}-\underline{\mu}_i)}_{d^2_{\Sigma_i}(\underline{x},\underline{\mu}_i)} \geq \underbrace{\log p_j - \tfrac{1}{2}\log|\Sigma_j| - \tfrac{1}{2}(\underline{x}-\underline{\mu}_j)'\Sigma_j^{-1}(\underline{x}-\underline{\mu}_j)}_{= d_j^Q(\underline{x})} \qquad \forall j \neq i$$

MALANOBIS DISTANCE — QUADRATIC DISCRIMINANT FUNCTIONS

$$\Leftrightarrow \quad d_i^Q(\underline{x}) \geq d_j^Q(\underline{x}) \qquad \forall j \neq i$$



4) LDA (Linear Discriminant Analysis)

- $c(i|j) = w > 0 \quad \forall j \neq i$
- $f_i(\underline{x})$ is $N_p(\underline{\mu}_i, \Sigma)$ — SAME FOR ALL $i = 1,\dots,g$ GROUPS

$$\underline{x} \in R_i \Leftrightarrow f_i(\underline{x})p_i \geq f_j(\underline{x})p_j \quad \forall j \neq i$$

$$\Leftrightarrow \log p_i - \cancel{\tfrac{1}{2}\log|\Sigma_i|} - \tfrac{1}{2}(\underline{x}-\underline{\mu}_i)'\Sigma_i^{-1}(\underline{x}-\underline{\mu}_i) \geq \log p_j - \cancel{\tfrac{1}{2}\log|\Sigma_j|} - \tfrac{1}{2}(\underline{x}-\underline{\mu}_j)'\Sigma_j^{-1}(\underline{x}-\underline{\mu}_j) \qquad \forall j \neq i$$

$$\Leftrightarrow \log p_i - \cancel{\tfrac{1}{2}\underline{x}'\Sigma^{-1}\underline{x}} + \underline{x}'\Sigma^{-1}\underline{\mu}_i - \tfrac{1}{2}\underline{\mu}_i'\Sigma^{-1}\underline{\mu}_i \geq \log p_j - \cancel{\tfrac{1}{2}\underline{x}'\Sigma^{-1}\underline{x}} + \underline{x}'\Sigma^{-1}\underline{\mu}_j - \tfrac{1}{2}\underline{\mu}_j'\Sigma^{-1}\underline{\mu}_j \qquad \forall j \neq i$$

(no more quadratic!)

$$\Leftrightarrow \underline{x}'\Sigma^{-1}\underline{\mu}_i + \log p_i - \tfrac{1}{2}\underline{\mu}_i'\Sigma^{-1}\underline{\mu}_i \geq \underbrace{\underline{x}'\Sigma^{-1}\underline{\mu}_j + \log p_j - \tfrac{1}{2}\underline{\mu}_j'\Sigma^{-1}\underline{\mu}_j}_{d_j(\underline{x})} \qquad \forall j \neq i$$

$$\underline{x} \in R_i \Leftrightarrow d_i(\underline{x}) \geq d_j(\underline{x}) \quad \forall j \neq i$$

VERY ROBUST!

IMPORTANT Gaussianity is not much important, instead focus on modifying data to have the same covariance matrix for all groups.

▶ Where is DATA used? To estimate the unknowns $\underline{\mu}_i, \Sigma_i, f_i(\underline{x})$ $i = 1,\dots,g$ (training data)

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_n' \end{bmatrix} \begin{bmatrix} \ell_1 \\ \ell_n \end{bmatrix}$$

WARNING Don't use data to estimate p_i, $i = 1,\dots,g$ unless you know what you're doing.

⇒ ok if sample is "random and large" enough to estimate the frequencies, however things with small frequency will not be many (dov.) and the classifier won't have a lot of information to learn how to classify them!

In LDA and QDA estimate $\underline{\mu}_i$ with $\bar{\underline{x}}_i = \frac{1}{n_i}\sum_{j:\ell_j=i} \underline{x}_j$ (both)

Σ_i with $S_i = \frac{1}{n_i - 1}\sum_{j:\ell_j=i}(\underline{x}_j - \bar{\underline{x}}_i)(\underline{x}_j - \bar{\underline{x}}_i)'$ (QDA)

Σ with $S_{pooled} = \frac{1}{n-g}\sum_{i=1}^{g}(n_i - 1)S_i$ (LDA)

If data in each group are not many with respect to p ⇒ parametrize Σ_i, $i = 1,\dots,g$

EX (Naive Bayes classifier)

Consider QDA and assume: $\Sigma_i = \begin{bmatrix} \sigma_{11}^{(i)} & & 0 \\ & \ddots & \\ 0 & & \sigma_{pp}^{(i)} \end{bmatrix}$ indep components

$$d_i^Q(\underline{x}) = \log p_i - \frac{1}{2}\underbrace{\log|\Sigma_i|}_{\log \sum_{i=1}^{p} \sigma_{ii}} - \frac{1}{2}\underbrace{(\underline{x}-\underline{\mu}_i)'\Sigma_i^{-1}(\underline{x}-\underline{\mu}_i)}_{\sum_{j=1}^{p}\frac{(x_j-\mu_{ij})^2}{\sigma_{jj}}}$$

σ_{jj} is estimated with $\frac{1}{m_i-1}\sum_{j:\ell_j=i}(x_{ij}-\bar{x}_{ij})^2$

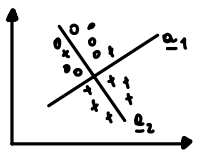
REMARK PCA and then classification is a bad recipe

Fisher's argument for LDA

- Remove gaussianity assumption!

Covariance between groups $B = \frac{1}{g-1}\sum_{i=1}^{g}(\underline{\mu}_i-\bar{\underline{\mu}})(\underline{\mu}_i-\bar{\underline{\mu}})'$ $\quad \bar{\underline{\mu}} = \frac{1}{g}\sum_{i=1}^{g}\underline{\mu}_i$

Covariance within groups



GOAL:
Find $\underline{a} \in \mathbb{R}^p$ that maximizes Variability between groups

Reparametrize the problem $\underline{\mu} = \Sigma^{1/2}\underline{a}$ $\quad (\underline{a} = \Sigma^{-1/2}\underline{\mu}) \Rightarrow \frac{\underline{a}'B\underline{a}}{\underline{a}'\Sigma\underline{a}} = \frac{\underline{\mu}'\Sigma^{-1/2}B\Sigma^{-1/2}\underline{\mu}}{\underline{\mu}'\Sigma^{-1/2}\Sigma\Sigma^{-1/2}\underline{\mu}} = \frac{\underline{\mu}'\Sigma^{-1/2}B\Sigma^{-1/2}\underline{\mu}}{\underline{\mu}'\underline{\mu}}$

$$\operatorname*{argmax}_{\underline{\mu}\in\mathbb{R}^p}\frac{\underline{\mu}'\Sigma^{-1/2}B\Sigma^{-1/2}\underline{\mu}}{\underline{\mu}'\underline{\mu}} = \underline{e}_1 \Rightarrow \underline{a}_1 = \Sigma^{-1/2}\underline{e}_1$$

If $\Sigma^{-1/2}B\Sigma^{-1/2} = \sum_{l=1}^{s}\lambda_l \underline{e}_l\underline{e}_l'$ $\quad s = \operatorname{rank} B = \min\{g-1, p\}$

as in PCA $\quad \underline{a}_1 = \Sigma^{-1/2}\underline{e}_1 \to \underline{a}_1'\underline{X}$ 1st discriminant score (Fisher's score)

$\underline{a}_s = \Sigma^{-1/2}\underline{e}_s \quad \underline{a}_s'\underline{X}$ sth " "

$$A = \begin{bmatrix}\underline{a}_1' \\ \vdots \\ \underline{a}_s'\end{bmatrix} \Rightarrow \operatorname{Cov}(\underline{e}_i\underline{X}, \underline{a}_j\underline{X}) = \underline{a}_i'\Sigma\underline{a}_j = \underline{e}_i'\Sigma^{-1/2}\Sigma\Sigma^{-1/2}\underline{e}_j = \underline{e}_i'\underline{e}_j = \begin{cases}0 & i\neq j\\ 1 & i=j\end{cases} \Rightarrow \operatorname{Cov}(A\underline{X}) = I$$

$\underline{X} \xrightarrow{A}$ $\begin{bmatrix}\underline{a}_1'\underline{X} \\ \underline{a}_2'\underline{X} \\ \vdots \\ \underline{a}_s'\underline{X}\end{bmatrix}$ take k $\quad k = 2, 3$ (for making pictures !!)

- How to build a classifier out of this?

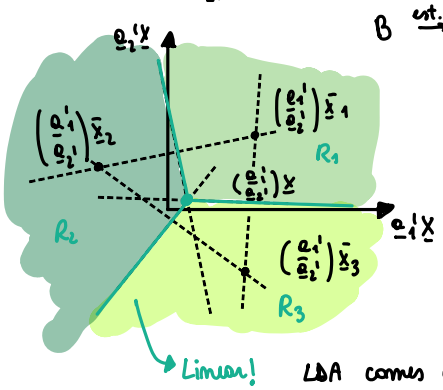
Training data

$$\mathbb{X} = \begin{bmatrix}\underline{x}_1' \\ \vdots \\ \underline{x}_m'\end{bmatrix}\begin{bmatrix}\ell_1 \\ \vdots \\ \ell_m\end{bmatrix}$$

$\underline{\mu}_i \xrightarrow{\text{est.}} \bar{\underline{x}}_i$

$\Sigma \xrightarrow{\text{est.}} S_{\text{pooled}} = \frac{1}{m-g}\sum_{i=1}^{g}(m_i-1)S_i$

$B \xrightarrow{\text{est.}} \hat{B} = \frac{1}{g-1}\sum_{i=1}^{g} m_i(\underline{x}_i-\bar{\underline{x}})(\underline{x}_i-\bar{\underline{x}})' \quad \bar{\underline{x}} = \frac{1}{m}\sum\underline{x}_j$



$$\underline{x} \in R_i \Leftrightarrow \sum_{j=1}^{k}(\underline{a}_j'\underline{x} - \underline{a}_j'\bar{\underline{x}}_i)^2 \leq \sum_{j=1}^{k}(\underline{a}_j'\underline{x} - \underline{a}_j'\bar{\underline{x}}_\ell)^2 \quad \forall j \neq i$$

Linear! LDA comes out even without gaussianity

[Logistic regression useful when distrib. is unknown]

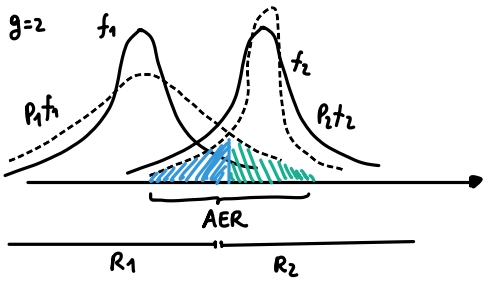
LECTURE 18 7/4/2022

EVALUATING A CLASSIFIER

Let $\delta : \mathbb{R}^p \to \{1 \dots g\}$ be a classifier

$$AER(\delta) = \sum_{k \neq 1} \int_{R_k} f_1(\underline{x}) p_1 \, d\underline{x} + \dots + \sum_{k \neq g} \int_{R_k} f_k(\underline{x}) p_k \, d\underline{x}$$

ACTUAL ERROR RATE → unknown (f...)



NON-PARAMETRIC APPROACH TO ESTIMATE AER

- **Naive approach**

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_m^T \end{bmatrix} \begin{bmatrix} l_1 \\ \vdots \\ l_m \end{bmatrix} \qquad \delta(\underline{x}_i) = \hat{l}_i \qquad \varepsilon_i = \begin{cases} 0 & l_i = \hat{l}_i \\ 1 & l_i \neq \hat{l}_i \end{cases} \Rightarrow \quad APER = \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i$$

APPARENT ERROR RATE (TRAINING ERROR)

CONFUSION MATRIX ($g=2$)

Predicted \ True	1	2
1	n_{11}	n_{12}
2	n_{21}	n_{22}

$$APER = \frac{n_{12} + n_{21}}{n}$$

→ useful to see the weak points of the classifier

- **Use a test set**

$$\tilde{\mathbb{X}} = \begin{bmatrix} \tilde{\underline{x}}_1^T \\ \vdots \\ \tilde{\underline{x}}_{\tilde{m}}^T \end{bmatrix} \begin{bmatrix} l_1 \\ \vdots \\ l_m \end{bmatrix}$$

Apply δ (learnt from $\mathbb{X}$, NOT from $\tilde{\mathbb{X}}$) to $\tilde{\mathbb{X}}$ $\quad \delta(\tilde{\underline{x}}_i) = \hat{l}_i \quad \varepsilon_i = \begin{cases} 0 & \hat{l}_i = \tilde{l}_i \\ 1 & \hat{l}_i \neq \tilde{l}_i \end{cases}$ $\quad \widehat{AER}(\delta) = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i$

In between? CROSS-VALIDATION

- **L10 - Leave one out**

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_m^T \end{bmatrix} \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_m \end{bmatrix}$$

FOR $i = 1, \dots, m$
  $\mathbb{X}_{-i}$ = without datum i
  learn a classifier δ_{-i} from $\mathbb{X}_{-i}$
  test δ_{-i} on i $\quad \varepsilon_i = \begin{cases} 0 & \delta_{-i}(\underline{x}_i) = l_i \\ 1 & \delta_{-i}(\underline{x}_i) \neq l_i \end{cases}$
END

$$\widehat{AER}(\delta) = \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i$$

→ it's a mean of NON-INDEPENDENT measurements

Small bias :)
High variance :(

- **k-fold cross validation**

($k \leq m$, in "practice" $k = 5, 10$ (?))
you take out more data ⇒ K parts

$$ER_i = \frac{1}{m_i} \sum_{j} \varepsilon_j$$

↳ # of j s.t. they belong to part-i

$$\widehat{AER}(\delta) = \frac{1}{m} \sum_{i=1}^{k} m_i \, ER_i$$

OBS $k=m \Rightarrow$ L10

Before splitting $\Rightarrow$ Permute data! To avoid building classifiers of linked variables (depending if the original dataset was ordered)

↳ m! permutations

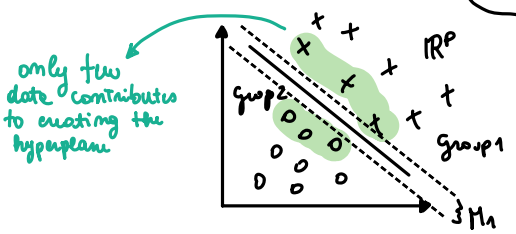
↳ $(k<m)$, run d permutations and get different estimates $\widehat{AER}_1(\delta), \dots, \widehat{AER}_d(\delta)$ → mean and confidence interval 2-test

Short exercises on separating hyperplanes and SVM (Supported Vector Machine) ["Elements of Statistical Learning"]

Idea $(g=2)$

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_m' \end{bmatrix} \begin{bmatrix} l_1 \\ \vdots \\ l_m \end{bmatrix} \quad l_i \in \{1,2\}$$

We assume there exists a hyperplane separating the features in the two groups. This generates a partition of $\mathbb{R}^p$ and thus a δ



When are two A,B subsets of $\mathbb{R}^p$ separated?

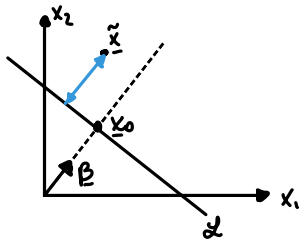
Consider the convex hulls of A and B

1) $CH(A) \neq \emptyset$, $CH(B) \neq \emptyset$
2) $CH(A) \cap CH(B) = \emptyset$
3) ~~either CH(A) or CH(B) is open~~
3) CH(A) and CH(B) are closed and one is compact (V)

$\Rightarrow \exists$ hyperplane (Hahn-Banach)

Hyperplanes

$\mathcal{L}$ hyperplane in $\mathbb{R}^p$: affine subspace of dim $p-1$



Let $\beta \perp \mathcal{L}$, $\|\beta\| = 1$, $\beta \in \mathbb{R}^p$ $\quad \underline{x}_0 \in \text{span}(\beta) \cap \mathcal{L}$ $\quad \beta_0 = \|\underline{x}_0\|$

$\underline{x} \in \mathcal{L} \Leftrightarrow \pi_{\underline{x}|\beta} = \underline{x}_0 \Leftrightarrow \underline{x}'\beta = \beta_0$

$x_1\beta_1 + \dots + x_p\beta_p = \beta_0$

$$y_i = \begin{cases} 1 & x_i \in \text{group 1} \\ -1 & x_i \in \text{group 2} \end{cases}$$

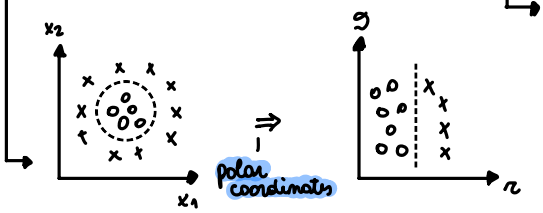
$y_i(\underline{x}_i\beta - \beta_0) \geq 0$

$\underline{x} \notin \mathcal{L}$: $\underline{x}'\beta > \beta_0$ or $\underline{x}'\beta < \beta_0$ $\quad \underline{x}'\beta - \beta_0$ is the distance (with sign) between $\underline{x}$ and $\mathcal{L}$

$M_1 = \min\{ y_i(\underline{x}_i\beta - \beta_0) : i = 1 \dots m \}$ (margin) $\Rightarrow$ Find $\beta \in \mathbb{R}^p$ s.t. $\beta = \underset{\|\beta\|=1, \beta \in \mathbb{R}^p}{\arg\max}\ M_1$ (Sol: see ref.)

What if group 1 and 2 are not separated by a hyperplane?

- Increase the dimension p of the embedding space → the curse of dimensionality plays in our favour
- Solve the "soft" problem allowing some overlapping



$$\begin{cases} \max_{\beta, \beta_0} M \\ \|\beta\| = 1 \\ y_i(\beta'\underline{x} - \beta_0) \geq (1-\varepsilon_i)M \quad \text{for } i = 1, \dots, m \end{cases}$$

↳ $\varepsilon_i \geq 0$, $\sum_{i=1}^{m} \varepsilon_i \leq C$

LECTURE 19 8/4/2022

UNSUPERVISED CLASSIFICATION (CLUSTER ANALYSIS)

- You don't know the labels nor how many they are
- Goal: - Estimate g
  - Estimate $\{l_1, \dots, l_m\}$

$\Rightarrow$ at the end you have a training set $\quad \mathbb{X} = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_m' \end{bmatrix} \begin{bmatrix} \hat{l}_1 \\ \vdots \\ \hat{l}_m \end{bmatrix}$

- Two lines of attack
  - parametric modeling
    - → based on ML (Maximum Likelihood)
    - → EM
  - nonparametric approach based on distances
- Idea: two units belonging to the same cluster are more **similar** than two units belonging to different clusters.

$d : \mathbb{R}^p \times \mathbb{R}^p \to [0, +\infty)$

pseudo-metric:
- $d(\underline{x}, \underline{x}) = 0 \quad \forall \underline{x} \in \mathbb{R}^p$
- $d(\underline{x}, \underline{y}) = 0 \Leftrightarrow \underline{x} = \underline{y}$ (strongly, also $\Rightarrow$)
- $d(\underline{x}, \underline{y}) = d(\underline{y}, \underline{x}) \quad \forall \underline{x}, \underline{y} \in \mathbb{R}^p$
- $d(\underline{x}, \underline{y}) \leq d(\underline{x}, \underline{z}) + d(\underline{z}, \underline{y}) \quad \forall \underline{x}, \underline{y}, \underline{z} \in \mathbb{R}^p$
- $d(\underline{x}, \underline{y}) \leq \max\{d(\underline{x}, \underline{z}) + d(\underline{z}, \underline{y})\}$

distance / metric

ultra-metric

Example of metrics

→ Better to standardize !

- $d(\underline{x}, \underline{y}) = \sqrt{(\underline{x}-\underline{y})'(\underline{x}-\underline{y})}$ Euclidean
- $d(\underline{x}, \underline{y}) = \sqrt{(\underline{x}-\underline{y})'\Sigma^{-1}(\underline{x}-\underline{y})}$ Mahanobis (Σ usually unknown)
- $d(\underline{x}, \underline{y}) = \left(\sum_{i=1}^p |x_i - y_i|^m\right)^{1/m}$ ℓ^m distances → $m=2$ Euclidean, $m=1$ Manhattan
- $d(\underline{x}, \underline{y}) = \sum_{i=1}^p \frac{|x_i - y_i|}{x_i + y_i}$ $\underline{x}, \underline{y} \in (\mathbb{R}^+)^p$ Canberra

$\underline{x}, \underline{y} \in \{0,1\}^p$

$\underline{x}' = (0\ 1\ 1\ 0\ 0\ 0\ 1)$

$d_{Eulid.}(\underline{x}, \underline{y}) = \sqrt{\sum_{i=1}^p (x_i - y_i)^2} = \sqrt{\#\text{ discordances}} = \sqrt{c+b}$ ⇒ we can build a table

$\underline{x}$ \ $\underline{y}$	0	1
0	a	b
1	c	d

You can transform categorical variables

EX $x \in \{\text{blue, green, brown}\}$ ⇒ $x_1 <_0^1$ blue / not blue, $x_2 <_0^1$ green / not green

(1,0) → blue
(0,1) → green
(0,0) → not green nor blue → brown

EX Mix: $d = \lambda d^{quant} + (1-\lambda) d^{cat.}$

Receive $\mathbb{X} = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_n' \end{bmatrix} \xrightarrow{d} D = \begin{bmatrix} 0 & & \\ d_{21} & 0 & \\ d_{31} & d_{32} & 0 \\ \vdots & & \ddots \end{bmatrix}$ (D: distance matrix)

How to measure the distance between SUBSETS ?

Let U, V be finite subsets of $\mathbb{R}^p$

- $d(U,V) = \min\{d(\underline{x}, \underline{y}) : \underline{x} \in U, \underline{y} \in V\}$ SINGLE LINKAGE (d(Italy, Switzerland) = 0)
- $d(U,V) = \max\{d(\underline{x}, \underline{y}) : \underline{x} \in U, \underline{y} \in V\}$ COMPLETE LINKAGE
- $d(U,V) = \frac{1}{|U|+|V|} \sum_{\substack{\underline{x} \in U \\ \underline{x} \notin V}} d(\underline{x}, \underline{y})$ AVERAGE LINKAGE
- $d(U,V) = d(\text{centre of } U, \text{centre of } V)$ → also to be defined! CENTERED LINKAGE

Hierarchical Agglomerative Clustering

- Set a distance and a linkage
- Every unit is a cluster
- Iterate until convergence
  - aggregate the two clusters which are closest

EX $n=5$

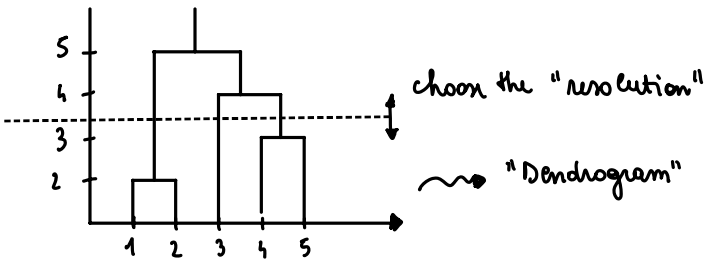
D:

	1	2	3	4	5
1	0				
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

	{1,2}	3	4	5
{1,2}	0			
3	5	0		
4	9	4	0	
5	8	5	3	0

	{1,2}	3	{4,5}
{1,2}	0		
3	5	0	
{4,5}	8	4	0

	{1,2}	{3,4,5}
{1,2}	0	
{3,4,5}	5	0



OBS Jitter the data ⇒ add noise
Careful when using single linkage ⇒ chain effect
CL, AL ⇒ ellipsoidal clusters

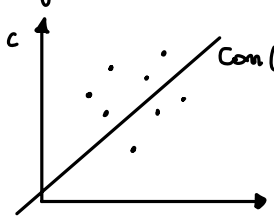
LECTURE 19 21/04/2022

New distance matrix

	1	2	3	4	5
1	0				
2	2	0			
3	5	5	0		
4	5	5	4	0	
5	5	5	4	3	0

cophenetic distances (ultrametric)

For every couple of units $\underline{x}_i, \underline{x}_j$ in $\mathbb{X}$ (training set) you have 2 distances $\begin{cases} d_{ij} & \text{distance in } D \\ c_{ij} & \text{distance in } C \end{cases}$



$Corr(D,C)$ = Cophenetic Correlation Coeff.

Ward agglomerative algorithm

$\mathbb{X} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_m^T \end{bmatrix}$ We want to divide it in $\overbrace{C_1, \dots, C_k}^{k}$ clusters $\quad C_i \cap C_j = \emptyset \ \forall i \neq j, \ \bigcup_{i=1}^{k} C_i = \mathbb{X}$

for $i = 1, \dots, k$

$$\bar{\underline{x}}_j = \frac{1}{\#C_j} \sum_{\underline{x}_i \in C_j} \underline{x}_i$$

$$ESS_j = \sum_{\underline{x}_i \in C_j} \| \underline{x}_i - \bar{\underline{x}}_j \|^2$$

} variability around the barycenter

- When every unit is a cluster ⇒ ESS = 0

iterate:
aggregate two clusters which cause the minimum increment of ESS

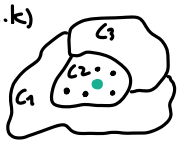
k-Means

$\mathbb{R}^p$, d, $\mathbb{X}$

Let $c_1, \dots, c_k$ be clusters partitioning $\mathbb{X}$

DEF $\bar{x}_j$ CENTROID of c_j $(j = 1 \dots k)$

$$\bar{x}_j = \underset{x \in \mathbb{R}^p}{\arg\min} \sum_{x_i \in c_j} d^2(x_i, x)$$



$\bar{x}_j$ may not belong to $\mathbb{X}$

↳ If d is euclidean this is the baricenter

⇒ "Find $c_1, \dots, c_k$ s.t. $\sum_{j=1}^{k} \sum_{x_i \in c_j} d^2(x_i, \bar{x}_j)$ is minimum" ⇝ Greedy solution with k-means !

- Initialization step
  - either : specify $c_1, \dots, c_k$ randomly (distributing k labels randomly among the n units)
  - or : specify $\bar{x}_1, \dots, \bar{x}_k$ centroids in $\mathbb{R}^p$ randomly
- Iterate until convergence
  1. for $j = 1, \dots, k$ compute c_j so that at the end we have $\bar{x}_1, \dots, \bar{x}_k$] Difficult step ! Minimization ⇒ Way out
  2. for all x_i attribute it to cluster c_j if $d^2(x_i, \bar{x}_j) = \min \{ d^2(x_i, \bar{x}_j),\ j = 1, \dots, k \}$

  Convergence ⇔ step 1 doesn't change

1. modified $\bar{x}_j = \underset{x \in \mathbb{X}}{\arg\min} \sum_{x_i \in c_j} d^2(x_i, x)$

Medoid (K-medoid)

only points of the training set

Label switching problem

Suppose K=3

num 1 \ num 2	1	2	3
1			*
2		*	
3	*		

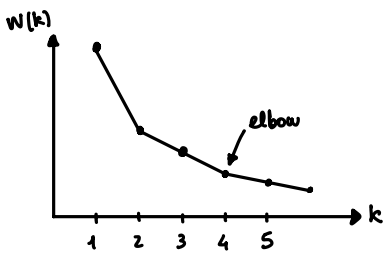
They give you the same cluster, but with different names

(In p=2 easy, pictures! in p=100 permute the matrix...)

How to choose k?

One way to do it

$$W(k) = \sum_{j=1}^{k} \sum_{x_i \in c_j} d^2(x_i, \bar{x}_j)$$



Graphical representation

- Fisher's score
- Multidimensional scaling (MDS)
  - ↳ Raw data ⇒ New representation

$\mathbb{R}^p$, d	$\mathbb{R}^q$ $(q \ll p)$ d_{Euclid}
$D = [d_{ij}]$	$Y = \begin{bmatrix} y_1' \\ \vdots \\ y_m' \end{bmatrix}$ s.t. $d_{Euclid.}(y_i, y_j) \approx d_{ij}$
$X = \begin{bmatrix} x_1' \\ \vdots \\ x_m' \end{bmatrix}$	

- Classical MDS

  Find $y_1, \dots, y_m$ in $\mathbb{R}^q$ s.t. $\sum_{ij} (d_{ij} - \delta_{ij})^2$ is min (no unique sol!)

OBS if d is Euclidean distance in $\mathbb{R}^p$ $\Rightarrow$ PCA span$(e_1, \dots, e_q)$

- Kruskal 'MDS

$$\text{STRESS} = \frac{\sum_{i \neq j} (\theta(d_{ij}) - \delta_{ij})^2}{\sum_{i \neq j} \delta_{ij}^2} \quad \text{minimize w.r.t. } \theta$$

LECTURE 20 26/4/2022

$Y \in \mathbb{R}$ target

$\underline{X} = (X_1, \dots, X_p)' \in \mathbb{R}^p$ features

$Y = f(\underline{X}) + \varepsilon \quad \varepsilon \perp\!\!\!\perp \underline{X}$

↳ How to capture it ? • Use prior knowledge
• Data

Optimal f w.r.t. MSE $\Rightarrow$ $f(\underline{X}) = \mathbb{E}[Y | \underline{X}]$ (σ-field (everything you know when ya know $\underline{X}$))

We know an analytical expression when both $\underline{X}, Y$ are Gaussian

- parametric model
$\hat{f}(\underline{X}) = \beta_0 + \beta_1 X_1^2 + \beta_2 \sin x_2 + \beta_3 x_1 x_3$ PRIOR KNOWLEDGE
- totally data driven LINEAR MODELS

- CART - Random forests
- NN - DL

Good for predictions but it's black-box

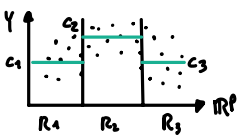
BETTER WITH LOTS OF DATA (the features already explain everything without need of transformations)

CART (Classification and Regression Trees)

Idea: $\hat{f}$ is piecewise constant

$$\hat{f}(\underline{X}) = \sum_{j=1}^{J} c_j \mathbb{1}_{\{\underline{X} \in R_j\}}$$

finite partition



$c_j = \bar{Y}_j$ = mean of Y in R_j

Goal: find J and $\{R_1, \dots, R_J\}$ s.t. $\sum_{j=1}^{J} \sum_{\underline{x}_i \in R_j} \| y_i - \bar{y}_j \|^2$ is min

⚠ WARNING overfitting

CART is a greedy algorithm for solving ↑

Iterate the Step

Step: find s_1 cutoff such that $\sum_{i=1}^{n} (y_i - \bar{y}_i)^2 - \left[\sum_{x_{i1} < s_1} (y_i - \bar{y}_1)^2 + \sum_{x_{i1} > s_1} (y_i - \bar{y}_2)^2 \right]$ is max

(Variability of Y after the cut; Variability of Y; R_1; R_2)

Repeat for $i = 2, \dots, p$

Having $s_1, \dots, s_p$, choose s_j^* that maximizes ↑

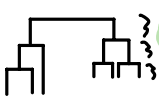
$R_1 = \{ \underline{x}_i : x_{ij} \leq s_j^* \}$ $R_2 = R_1^c$

Stop splitting if the number of units in a region is less than m ($m = 5, 10$, choose wisely)

How to control overfitting ?

▶ prune the tree

↳ penalize the number of leaves J



proportional to the decrease in variability !

$$\min \left\{ \sum_{j=1}^{J} \sum_{\underline{x}_i \in R_j} (y_i - \bar{y}_j)^2 + \alpha J \right\} \quad \alpha > 0$$

PENALIZATION PARAMETER (usually found through cross-validation)

Linear Models for regression

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mp} \end{bmatrix} \begin{matrix} Y \\ \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \end{matrix} \Rightarrow \text{Design matrix } Z = \begin{bmatrix} 1 & z_{11} & \cdots & z_{1n} \\ 1 & z_{21} & \cdots & z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{m1} & \cdots & z_{mn} \end{bmatrix} \begin{matrix} Y \\ \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix} \end{matrix}$$

$z_i = g_i(x_1, \dots, x_p)$ known $\quad i = 1, \dots, n$

$\hat{f}(\underline{z}) = \beta_0 + \beta_1 z_1 + \cdots + \beta_n z_n \qquad \underline{z} = (1\ z_1 \cdots z_n)' \in \mathbb{R}^{n+1}$

LM: $Y = \underline{z}'\underline{\beta} + \varepsilon \qquad \underline{\beta} = (\beta_0\ \beta_1 \cdots \beta_n)' \in \mathbb{R}^{n+1}$

$\varepsilon \perp\!\!\!\perp \underline{z} \quad \mathbb{E}[\varepsilon] = 0 \quad \mathrm{Var}(\varepsilon) = \sigma^2$

Model for the experiment generating the data:

$\underline{Y} = (Y_1 \dots Y_m)' \in \mathbb{R}^m$

$Z \in \mathbb{R}^{m \times (n+1)}$

$\underline{Y} = Z\underline{\beta} + \underline{\varepsilon}$

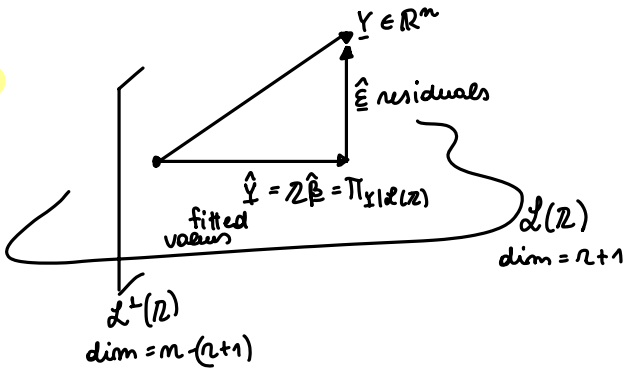
OLS $\quad \underline{\varepsilon} \perp\!\!\!\perp Z \quad \mathbb{E}[\underline{\varepsilon}] = \underline{0} \quad \mathrm{Cov}(\underline{\varepsilon}) = \sigma^2 I$ (no gaussianity assumption here!)

Ordinary Least Squares method

Find $\underline{\beta}$ s.t. $\|\underline{Y} - Z\underline{\beta}\|^2$ is min

PROP If Z is full-rank ($\mathrm{rank}\, Z = n+1$)

$$\Rightarrow \hat{\beta} = \underset{\underline{\beta} \in \mathbb{R}^{n+1}}{\mathrm{argmin}} \|\underline{Y} - Z\underline{\beta}\|^2 = (Z'Z)^{-1} Z'\underline{Y}$$



proof $\quad Z\beta = \beta_0 \underline{1} + \beta_1 \underline{c}_1 + \cdots + \beta_n \underline{c}_n$

$Z = [\underline{1} \quad \underline{c}_1 \quad \cdots \quad \underline{c}_n]$

$(n+1) \times (n+1)$ full rank

$$Z'Z = \sum_{i=1}^{n+1} \lambda_i \underline{e}_i \underline{e}_i' \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n+1} > 0$$

$$\Downarrow$$

$$(Z'Z)^{-1} = \sum_{i=1}^{n+1} \frac{1}{\lambda_i} \underline{e}_i \underline{e}_i'$$

$$\mathcal{L}(Z) \ni \underline{q}_i := \frac{1}{\sqrt{\lambda_i}} Z\underline{e}_i \qquad i = 1, \dots, n+1$$

$$\underline{q}_i' \underline{q}_j = \frac{1}{\sqrt{\lambda_i \lambda_j}} \underline{e}_i' Z'Z \underline{e}_j = \frac{\lambda_i}{\sqrt{\lambda_i \lambda_j}} \underline{e}_i' \underline{e}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \Rightarrow \text{o.n.b. for } \mathcal{L}(Z)$$

$$\Pi_{Y|\mathcal{L}(Z)} = \sum_{i=1}^{n+1} \Pi_{Y|\underline{q}_i} = \sum_{i=1}^{n+1} \underline{q}_i \underline{q}_i' \frac{1}{\underline{q}_i' \underline{q}_i} Y = \sum_{i=1}^{n+1} \frac{1}{\lambda_i} Z\underline{e}_i \underline{e}_i' Z'Y = Z\left(\sum_{i=1}^{n+1} \frac{1}{\lambda_i} \underline{e}_i \underline{e}_i'\right) Z'Y = \underbrace{Z\,\underbrace{(Z'Z)^{-1} Z'}_{H\ \text{Hat matrix}}\, Y}_{\hat{\beta}}$$

REMARK If $\underbrace{\mathrm{rank}\, Z}_{k} < n+1$

We can't invert it, but we can use the generalized Moore-Penrose inverse

LECTURE 21 2/5/2022

Ex (ANOVA)

$$\Downarrow \begin{cases} X_{11}, \dots, X_{1m_1} \overset{iid}{\sim} N_1(\mu_1, \sigma^2) \\ \vdots \\ X_{g1}, \dots, X_{gm_g} \overset{iid}{\sim} N_1(\mu_g, \sigma^2) \end{cases} \qquad Y = (X_{11}\ X_{12} \cdots X_{1m_1} X_{21} \cdots X_{gm_g}) \in \mathbb{R}^{m_1 + \cdots + m_g}$$

Take as design matrix $Z = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix}$ (m_1 rows, m_2 rows, m_g rows; m rows; $g+1$ columns)

$\beta = (\mu, \tau_1, \dots, \tau_g) \in \mathbb{R}^{g+1}$

$Y = Z\beta + \varepsilon$ means $\boxed{X_{ij} = \mu + \tau_i + \varepsilon_{ij}}$ $\begin{matrix} i=1,\dots,g \\ j=1,\dots,m_i \end{matrix}$ if $\varepsilon = (\varepsilon_{11} \cdots \varepsilon_{gm_g}) \in \mathbb{R}^n \sim N_n(0, \sigma^2 I)$

- Notice Z is NOT full-rank
  $\dim(\mathcal{L}(Z)) = g = r < r+1$

- We impose $\sum_{i=1}^{g} n_i \tau_i = 0$ to avoid overparametrization
  $\boxed{\tau_g = -\frac{1}{n_g} \sum_{i=1}^{g} n_i \tau_i}$ consider treatment at (arbitrary!) level g

- Choose $Z = [\underline{z}_0 \; \underline{z}_1 \cdots \underline{z}_g]$ it's full-rank!

$$\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \\ -\frac{n_1}{n_g} \\ \vdots \\ -\frac{n_1}{n_g} \end{bmatrix} \begin{matrix} \} n_1 \\ \\ \\ \\ \\ \} n_g \end{matrix} \quad \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \\ -\frac{n_{g-1}}{n_g} \\ \vdots \\ -\frac{n_{g-1}}{n_g} \end{bmatrix} \begin{matrix} \} n_{g-1} \\ \\ \\ \\ \\ \} n_g \end{matrix}$$

How to estimate $\beta_0, \dots, \beta_r$ and σ^2?

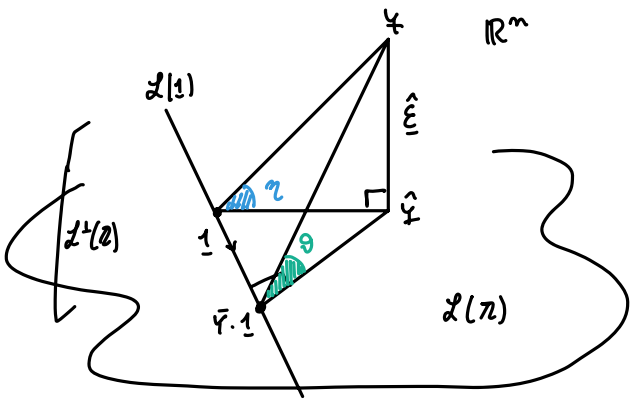
Ordinary Least Squares (OLS)

Assume Z full rank: $\operatorname{rank} Z = r+1 \leq n$

$\hat{Y} = \Pi_{Y|\mathcal{L}(Z)}$ fitted values

$\hat{\varepsilon} = \Pi_{Y|\mathcal{L}^\perp(Z)}$ residuals

$Y = \hat{Y} + \hat{\varepsilon} \Rightarrow \hat{Y} \perp \hat{\varepsilon}$



$\Pi_{Y|\mathcal{L}(\underline{1})} = \bar{Y} \cdot \underline{1}$

$\Pi_{\hat{Y}|\mathcal{L}(\underline{1})} = \frac{\underline{1}\underline{1}'}{\underline{1}'\underline{1}} \hat{Y} = \frac{\underline{1}\underline{1}' H Y}{\underline{1}'\underline{1}} = \frac{\underline{1}\underline{1}' Y}{\underline{1}'\underline{1}} = \bar{Y}\underline{1}$ (H self-adjoint: $(H\underline{1})' = \underline{1}'H' = \underline{1}'$)

$\Rightarrow \|Y - \bar{Y}\underline{1}\|^2 = \|\hat{Y} - \bar{Y}\cdot\underline{1}\|^2 + \|\hat{\varepsilon}\|^2$

$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum \hat{\varepsilon}_i^2$ (centered on the mean)

$\longrightarrow 1 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} + \frac{\sum \hat{\varepsilon}_i^2}{\sum (Y_i - \bar{Y})^2}$ ($\frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = R^2$)

$R^2 = 1 - \frac{\sum \hat{\varepsilon}_i^2}{\sum (Y_i - \bar{Y})^2} = 1 - \sin^2(\vartheta) = \cos^2\vartheta \in [0,1]$ (*)

- $R^2 = 1 \Rightarrow \vartheta = 0 \Rightarrow Y = \hat{Y}$ perfect fit
- $R^2 = 0 \Rightarrow \vartheta = \pi/2 \Rightarrow \hat{Y} = \bar{Y}\cdot\underline{1}$ predict with mean

- "Models through the origin"
  $Z = \begin{bmatrix} \cancel{1} & z_{11} & \cdots \\ \cancel{\vdots} & \vdots & \ddots \end{bmatrix}$ $\beta = (\cancel{\beta_0} \; \beta_1 \cdots \beta_g)$ $\Rightarrow$ $\bar{Y} \overset{?}{=} \bar{\hat{Y}}$ you cannot project on $\mathcal{L}(\underline{1})$ since it's outside $\mathcal{L}(Z)$

  $\beta_0 = 0$

  R^2 cannot be computed, or it can, but (*) doesn't hold anymore, it might be also greater than 1!

  You can only use $\sum Y_i^2 = \sum \hat{Y}_i^2 + \sum \hat{\varepsilon}_i^2$

  $\Rightarrow$ invent $\tilde{R}^2 = \frac{\sum \hat{Y}_i^2}{\sum Y_i^2} = \cos^2\eta \in [0,1]$

- To take into account the model complexity $R^2_{adj} = 1 - \frac{\frac{1}{m-(n+1)}\sum_{i=1}^m \hat{\varepsilon}_i^2}{\frac{1}{m-1}\sum_{i=1}^m (y_i-\bar{y})^2}$

Properties of $\hat{\beta}$ and $\hat{\underline{\varepsilon}}$

Assume full-rank, rank $\mathcal{Z} = n+1 \leq m$

1) $\mathbb{E}[\hat{\beta}] = \beta$ (unbiased)
2) $\mathrm{Cov}(\hat{\beta}) = \sigma^2 (\mathcal{Z}'\mathcal{Z})^{-1}$
3) $\mathbb{E}[\hat{\underline{\varepsilon}}] = \underline{0}$
4) $\mathrm{Cov}(\hat{\underline{\varepsilon}}) = \sigma^2 (I - H)$
5) $\mathbb{E}[\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}] = \mathbb{E}[\sum_{i=1}^m \hat{\varepsilon}_i^2] = \sigma^2 (m-(n+1)) \overset{\text{corollary}}{\Rightarrow} \frac{\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}}{m-(n+1)} = \frac{\sum \hat{\varepsilon}_i^2}{m-(n+1)}$ unbiased for σ^2

proof

1) $\mathbb{E}[\hat{\beta}] = \mathbb{E}[(\mathcal{Z}'\mathcal{Z})^{-1}\mathcal{Z}'\underline{y}] = (\mathcal{Z}'\mathcal{Z})^{-1}\mathcal{Z}'\overbrace{\mathbb{E}[\underline{y}]}^{\mathcal{Z}\beta} = \beta$
2) $\mathrm{Cov}(\hat{\beta}) = \mathrm{Cov}((\mathcal{Z}'\mathcal{Z})^{-1}\mathcal{Z}'\underline{y}) = (\mathcal{Z}'\mathcal{Z})^{-1}\mathcal{Z}'(I\sigma^2)\mathcal{Z}(\mathcal{Z}'\mathcal{Z})^{-1} = \sigma^2(\mathcal{Z}'\mathcal{Z})^{-1}$
3) $\mathbb{E}[\hat{\underline{\varepsilon}}] = \mathbb{E}[(I-H)\underline{y}] = \mathbb{E}[\underline{y}] - \mathbb{E}[H\underline{y}] = \mathcal{Z}\beta - \mathcal{Z}\beta = \underline{0}$
4) $\mathrm{Cov}(\hat{\underline{\varepsilon}}) = \mathrm{Cov}((I-H)\underline{y}) = (I-H)\sigma^2 I (I-H)' = \sigma^2(I-H)(I-H)' = \sigma^2(I-H)^2 = \sigma^2(I-H)$ (the det. must be zero — this will have an impact)
5) $\mathbb{E}[\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}}] = \mathbb{E}[\mathrm{tr}(\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}})] = \mathbb{E}[\mathrm{tr}(\hat{\underline{\varepsilon}}\hat{\underline{\varepsilon}}')] = \mathbb{E}[\mathrm{tr}((I-H)\underline{y}\underline{y}'(I-H)')] = \mathbb{E}[\mathrm{tr}((I-H)\underline{\varepsilon}\underline{\varepsilon}'(I-H)')]$

(1×1 matrix $[\cdot]$; number; $\mathrm{tr}(ABC) = \mathrm{tr}(CAB)$; matrix; $(I-H)(\mathcal{Z}\beta+\underline{\varepsilon}) = (I-H)\underline{\varepsilon}$)

$= \mathrm{tr}((I-H)\,\mathbb{E}[\underline{\varepsilon}\underline{\varepsilon}'](I-H)') = \sigma^2\,\mathrm{tr}((I-H)(I-H)')$

$= \sigma^2\,\mathrm{tr}(I-H)$

$\mathrm{tr}(I) - \mathrm{tr}(H) = m - \mathrm{tr}(\mathcal{Z}(\mathcal{Z}'\mathcal{Z})^{-1}\mathcal{Z}') = m - \mathrm{tr}(\mathcal{Z}'\mathcal{Z}(\mathcal{Z}'\mathcal{Z})^{-1}) = m - \mathrm{tr}(I_{n+1}) = m-(n+1)$

$= \sigma^2(m-(n+1))$

Corollary $S^2 = \frac{\sum \hat{\varepsilon}_i^2}{m-(n+1)}$ unbiased for σ^2

Obs $\mathrm{Cov}(\hat{\beta}) = \sigma^2(\mathcal{Z}'\mathcal{Z})^{-1}$
↳ are correlated, we can control the design matrix (e.g. PCA) Danger zone when regressors are (almost colinear)

Obs Design of Experiments

From now on $\underline{\varepsilon} \sim N_m(\underline{0}, \sigma^2 I)$

Prop Assume $\mathcal{Z}$ full rank and $\underline{y} = \mathcal{Z}\beta + \underline{\varepsilon} \Rightarrow \underline{y} \sim N_m(\mathcal{Z}\beta, \sigma^2 I)$

1) $\hat{\beta}$ and $\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{m}$ are MLE of β and σ^2
2) $\hat{\beta} \sim N_{n+1}(\beta, \sigma^2(\mathcal{Z}'\mathcal{Z})^{-1})$
3) $\hat{\underline{\varepsilon}} \sim N_m(\underline{0}, \sigma^2(I-H))$
4) $\hat{\underline{\varepsilon}} \perp\!\!\!\perp \hat{\beta}$
5) $\hat{\underline{\varepsilon}}'\hat{\underline{\varepsilon}} = \sum_{i=1}^m \hat{\varepsilon}_i^2 \sim \sigma^2 \chi^2(m-(n+1))$

Proof:

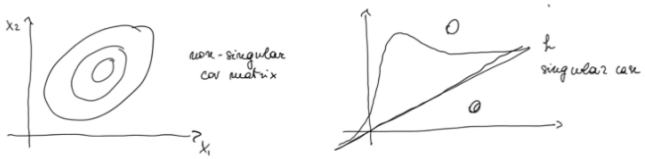
- For 1): Just compute derivative of log likelihood
- For 2),3) and 4) Note that:

$$\begin{bmatrix}\hat{\beta} \\ \hat{\underline{\varepsilon}}\end{bmatrix} = \begin{bmatrix}(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T \\ I - \mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\end{bmatrix}\underline{y} \text{ and } \underline{y} \sim \mathcal{N}_n(\underline{0}, \sigma^2 I)$$

Then we just need to verify: $Cov\begin{bmatrix}\hat{\beta} \\ \hat{\underline{\varepsilon}}\end{bmatrix} = \begin{bmatrix}(\mathbb{Z}^T\mathbb{Z})^{-1} & 0 \\ 0 & I-H\end{bmatrix}$ and so since we have two zero blocks off diagonal than the two vectors are independent!

- For 5): $\underline{\hat{\epsilon}} \sim \mathcal{N}_n(\underline{0}, \sigma^2(I-H))$ but:
  $(I-H)$ is a singular matrix, indeed $det(I-H)=0$ as its rank is $n-(r+1)$ and not n!

  So we have a singular Gaussian distribution: it's a Gaussian distribution defined on a sub-space of Lebesgue measure equal to zero:



  So the error vector $\underline{\epsilon}$ is a non-singular Gaussian, whereas the residual vector $\underline{\hat{\epsilon}}$ is a singular Gaussian constrained on a linear sub-space!

  To prove this we need to compute: $\underline{\hat{\epsilon}}^T(I-H)^{-1}\underline{\hat{\epsilon}}$ but $I-H$ is singular!

  So we can compute the Moore-Penrose Generalised inverse so: $\underline{\hat{\epsilon}}^T(I-H)^{\dagger}\underline{\hat{\epsilon}} \sim \sigma^2\chi^2(n-(r+1))$ and this is the Mahalanobis Distance of a Gaussian from its mean. Then: $\underline{\hat{\epsilon}}^T\underline{\hat{\epsilon}} \sim \sigma^2\chi^2(n-(r+1))$

LECTURE 22 5/5/2022

2) $\Rightarrow \frac{1}{\sigma^2}(\hat{\underline{\beta}}-\underline{\beta})' Z'Z (\hat{\underline{\beta}}-\underline{\beta}) \sim \chi^2(r+1)$
3) $\Rightarrow \frac{1}{\sigma^2}\hat{\underline{\epsilon}}'\hat{\underline{\epsilon}} \sim \chi^2(n-(r+1))$

(2) ⇒ general Penrose inverse; (3) ⇒ ratio is F

PIVOTAL

$$\frac{\frac{1}{\sigma^2}(\hat{\underline{\beta}}-\underline{\beta})' Z'Z (\hat{\underline{\beta}}-\underline{\beta})/(r+1)}{\frac{1}{\sigma^2}\hat{\underline{\epsilon}}'\hat{\underline{\epsilon}}/(n-(r+1))} \sim F(r+1, n-(r+1))$$

S^2

$$CR_{1-\alpha}(\underline{\beta}) = \left\{ \underline{\eta} \in \mathbb{R}^{r+1} : \frac{1}{S^2}(\hat{\underline{\beta}}-\underline{\eta})' Z'Z (\hat{\underline{\beta}}-\underline{\eta}) \leq (r+1) F_{1-\alpha}(r+1, n-(r+1)) \right\}$$

$$\frac{1}{\sigma^2}\hat{\underline{\epsilon}}'\hat{\underline{\epsilon}} \sim \chi^2(n-(r+1)) \rightarrow \frac{(n-(r+1))S^2}{\sigma^2} \chi^2(n-(r+1)) \rightarrow CI_{1-\alpha}(\sigma^2) = \left[\frac{(n-(r+1))S^2}{\chi^2_{\alpha/2}(n-(r+1))}, \frac{(n-(r+1))S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-(r+1))} \right]$$

Linear combinations, $\underline{a} \in \mathbb{R}^{r+1}$

$\underline{a}'\underline{\beta} = a_0\beta_0 + \cdots + \beta_r a_r \sim N_1(\underline{a}'\underline{\beta}, \sigma^2\underline{a}'(Z'Z)^{-1}\underline{a}) \rightarrow \frac{\underline{a}'(\hat{\underline{\beta}}-\underline{\beta})}{\sigma\sqrt{\underline{a}'(Z'Z)\underline{a}}} \sim N(0,1)$ ⫫ $\hat{\underline{\epsilon}}'\hat{\underline{\epsilon}} \sim \sigma^2\chi^2(n-(r+1))$

$\Leftrightarrow \frac{S^2(n-(r+1))}{\sigma^2} \sim \chi^2(n-(r+1))$

$$\frac{\underline{a}'(\hat{\underline{\beta}}-\underline{\beta})}{S\sqrt{\underline{a}'(Z'Z)^{-1}\underline{a}}} \sim t(n-(r+1))$$

PIVOTAL

$$CI_{1-\alpha}(\underline{a}'\underline{\beta}) = \left[\underline{a}'\hat{\underline{\beta}} \pm S\sqrt{\underline{a}'(Z'Z)^{-1}\underline{a}}\; t_{1-\frac{\alpha}{2}}(n-(r+1)) \right]$$

$H_0: \underline{a}'\underline{\beta} = \gamma_0$ vs $\underline{a}'\underline{\beta} \neq \gamma_0$

Reject at $\alpha \in [0,1]$ if $\frac{|\underline{a}'\hat{\underline{\beta}} - \gamma_0|}{S\sqrt{\underline{a}'(Z'Z)^{-1}\underline{a}}} > t_{1-\frac{\alpha}{2}}(n-(r+1))$

Ex (Done in all softwares for linear regression)

$H_0: \beta_i = 0$ vs $\beta_i \neq 0$

Reject if $\frac{|\hat{\beta}_i - 0|}{S\sqrt{\text{diag}_i(Z'Z)^{-1}}} > t_{1-\frac{\alpha}{2}}(n-(r+1))$

Sim. CI for $\underline{a}'\underline{\beta}$

From the Maximum Lemma: $\max_{\underline{a} \in \mathbb{R}^{r+1}} \frac{1}{S^2}\frac{(\underline{a}'(\hat{\underline{\beta}}-\underline{\beta}))^2}{\underline{a}'(Z'Z)^{-1}\underline{a}} = \frac{1}{S^2}(\hat{\underline{\beta}}-\underline{\beta})' Z'Z (\hat{\underline{\beta}}-\underline{\beta}) \sim (r+1)F(r+1, n-(r+1))$

$$\Rightarrow \text{Sim } CI_{1-\alpha}(\underline{a}'\underline{\beta}) = \left\{ \underline{a}'\underline{\beta} \pm \sqrt{\underline{a}'(Z'Z)^{-1}\underline{a}}\sqrt{S^2(r+1)F_{1-\alpha}(r+1, n-(r+1))} \right\}$$

NOT GIVEN BY R ! BUT VERY MEANINGFUL

C $p \times (n+1)$ not random

$H_0: C\beta = 0$ vs $C\beta \neq 0$

$$\rightarrow \begin{bmatrix} c_{11}\beta_0 + c_{12}\beta_1 + \dots + c_{1,n+1}\beta_n \\ c_{21}\beta_0 + \dots + c_{2,n+1}\beta_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

$C\hat{\beta} \sim N_p(C\beta, \sigma^2 C(Z'Z)^{-1}C')$

Under H_0:

$$\frac{\frac{1}{p}(C\hat{\beta})'(\sigma^2 C(Z'Z)^{-1}C')^{-1}(C\hat{\beta})}{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2(m-(n+1))}} \sim F(p, m-(n+1))$$

$\sim \chi^2(p)$ over $\sim \chi^2(m-(n+1))$

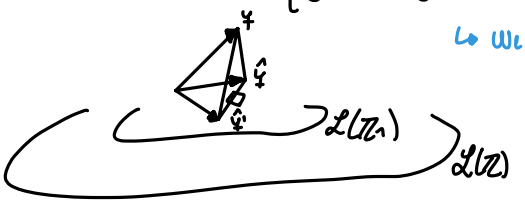
So $\underbrace{\frac{1}{S^2}(C\hat{\beta})'[C(Z'Z)^{-1}C']^{-1}(C\hat{\beta})}_{\text{reject at level } \alpha \text{ if this is } > p F_{1-\alpha}(p, m-(n+1))} \sim p\, F(p, m-(n+1))$

Special case

$Y = \beta_0 + \beta_1 z_1 + \underbrace{\dots + \beta_n z_n}_{\text{can I take out simultaneously the last p?}} + \varepsilon$ $\qquad H_0: \beta_n = \beta_{n-1} = \dots = \beta_{n-(p-1)} = 0$

$\Rightarrow$ Choose $C = \begin{bmatrix} 0 & \dots & 0 & \overbrace{1}^{p} & & \\ \vdots & & \vdots & & \ddots & \\ 0 & \dots & 0 & & & 1 \end{bmatrix} = [0 \; I_p]$ $\quad p \times (n+1)$

↳ We can SHUFFLE to test the subset we want!



$Z = [Z_1 \; Z_2]$ (Z_2: p columns; Z_1: simplified)

Reject if (0) large:

$$\frac{\hat{\varepsilon}_1'\hat{\varepsilon}_1 - \hat{\varepsilon}'\hat{\varepsilon}}{pS^2} \sim F(p, m-(n+1))$$

Very special case

$H_0: \beta_1 = \dots = \beta_n = 0$ (all but β_0)

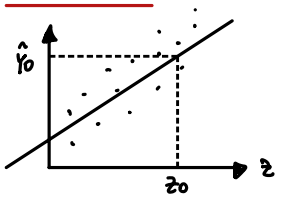
Should I throw everything in the trash bin?

$Z_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ $\qquad \hat{Y}_1 = \bar{Y} \cdot 1$

$\hat{\varepsilon}_1'\hat{\varepsilon}_1 = \sum (Y_i - \bar{Y})^2$

$$\frac{\hat{\varepsilon}_1'\hat{\varepsilon}_1 - \hat{\varepsilon}'\hat{\varepsilon}}{S^2 n} = \frac{\sum (Y_i - \bar{Y})^2 - \sum \hat{\varepsilon}_i^2}{n S^2} = \frac{\sum (\hat{Y}_i - \bar{Y})^2 / n}{\sum \hat{\varepsilon}_i^2 / m-(n+1)} \sim F(n, m-(n+1))$$

PREDICTION



Gauss-Markov theorem given $a \in \mathbb{R}^{n+1}$

The optimal (min. variance) estimator of $a'\beta$ is $a'\hat{\beta}$ (BLUE: Best Linear Unbiased Estimator) $(= a'(Z'Z)^{-1}Z'Y)$

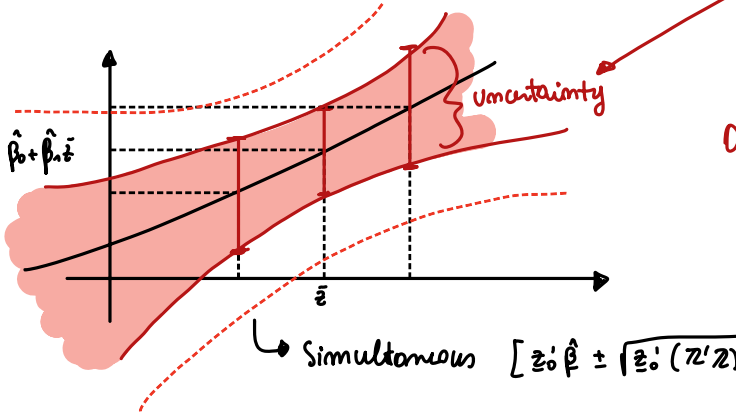
among those which are
- unbiased
- linear functions of Y

$$\frac{z_0'\hat{\beta} - z_0'\beta}{S\sqrt{z_0'(Z'Z)^{-1}z_0}} \sim t(m-(n+1))$$

PIVOTAL

$$\Rightarrow \boxed{CI_{1-\alpha}(z_0'\beta) = \left[z_0'\hat{\beta} \pm S\sqrt{z_0'(Z'Z)^{-1}z_0}\, t_{\alpha/2}(m-(n+1)) \right]} \qquad \alpha \in (0,1)$$

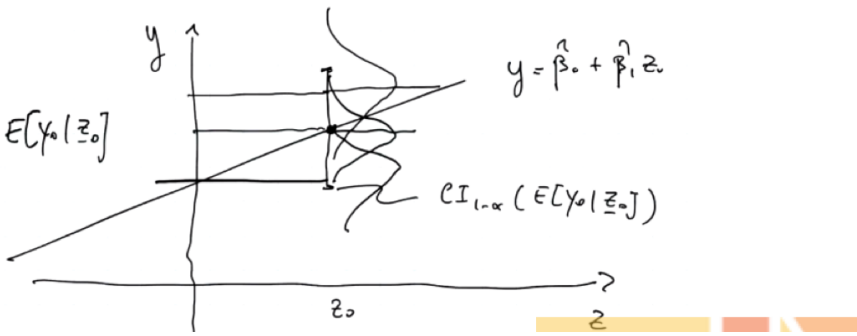
Uncertainty



ONE-AT-THE-TIME ⚠

↳ Simultaneous $\left[\underline{z}_0'\hat{\underline{\beta}} \pm \sqrt{\underline{z}_0'(\mathbb{Z}'\mathbb{Z})^{-1}\underline{z}_0}\sqrt{(r+1)F_{1-\alpha}(r+1, n-(r+1))}\right]$

In the above we have seen the prediction for $\mathbb{E}[y_0|\underline{z}_0]$ Well, what about y_0?



Can we find I such that $\mathbb{P}[y_0 \in I \,|\, \underline{z}_0] = 1-\alpha$?

PREDICTION INTERVAL

Obviously this PI will be larger than the CI for the mean: with the PI we are uncertain about the centre of the distribution and there is an extra variability due to the fact that y is different from its mean!

We know that: $\underline{z}_0^T\hat{\underline{\beta}} \sim \mathcal{N}_1(\underline{z}_0^T\underline{\beta}, \sigma^2\underline{z}_0^T(\mathbb{Z}^T\mathbb{Z})^{-1}\underline{z}_0)$ this is in the training set!

Then y_0 is a new statistical unit: not in the training set! Thus: $y_0 \perp\!\!\!\perp \underline{z}_0^T\hat{\underline{\beta}}$ since $\epsilon_0 \perp\!\!\!\perp \underline{\epsilon}$

Thus: $y_0 - \underline{z}_0^T\hat{\underline{\beta}} \sim \mathcal{N}(0, \sigma^2(1+\underline{z}_0^T(\mathbb{Z}^T\mathbb{Z})^{-1}\underline{z}_0))$ and $\hat{\underline{\epsilon}}^T\hat{\underline{\epsilon}} \sim \sigma^2\chi^2(n-(r+1))$ Moreover: $\hat{\underline{\epsilon}}^T\hat{\underline{\epsilon}} \perp\!\!\!\perp y_0 - \underline{z}_0^T\hat{\underline{\beta}}$ since $\hat{\underline{\epsilon}}^T\hat{\underline{\epsilon}} \perp\!\!\!\perp \hat{\underline{\epsilon}}^T$ and $\hat{\underline{\epsilon}}^T\hat{\underline{\epsilon}} \perp\!\!\!\perp \hat{\underline{\epsilon}}$

Thus $\dfrac{y_0 - \underline{z}_0^T\hat{\underline{\beta}}}{\sqrt{\sigma^2(1+\underline{z}_0^T(\mathbb{Z}^T\mathbb{Z})^{-1}\underline{z}_0}} \sim \mathcal{N}(0,1)$ and $\sqrt{\dfrac{\hat{\underline{\epsilon}}^T\hat{\underline{\epsilon}}}{\sigma^2(n-(r+1))}}$ is a chi-squared. Thus their quotient is a $t-$student that is:

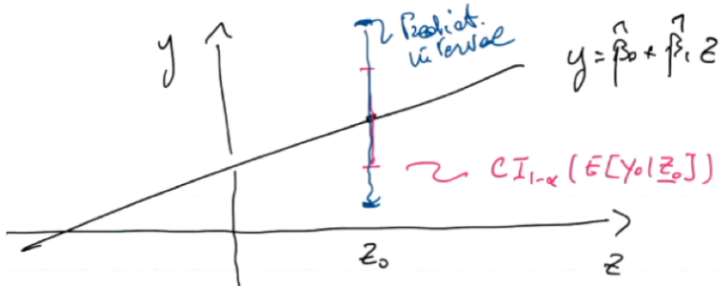
$$\frac{y_0 - \underline{z}_0^T\hat{\underline{\beta}}}{\sqrt{\dfrac{\hat{\underline{\epsilon}}^T\hat{\underline{\epsilon}}}{\sigma^2(n-(r+1))}}} = \frac{y_0 - \underline{z}_0^T\hat{\underline{\beta}}}{S\sqrt{1+\underline{z}_0^T(\mathbb{Z}^T\mathbb{Z})^{-1}\underline{z}_0}} \sim t(n-(r+1))$$

So the prediction interval of probability $1-\alpha$ is:

larger!

$$PI_{1-\alpha}(y_0) = \left[\underline{z}_0^T\hat{\underline{\beta}} \pm S\sqrt{1+\underline{z}_0^T(\mathbb{Z}^T\mathbb{Z})^{-1}\underline{z}_0}\, t_{\alpha/2}(n-(r+1))\right]$$

We see that the prediction interval is larger than the confidence interval:



LECTURE 23 6/5/2022

GLS : Generalized Least Squares

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{n+1}}{\operatorname{argmin}} \|y - X\beta\|^2 \quad \text{(OLS)}$$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{n+1}}{\operatorname{argmin}} (y - X\beta)' W^{-1} (y - X\beta) \quad W \ n \times n \ \text{pos. def}$$

GLS (if $W = I$ we have OLS)

$$(y - X\beta)' W^{-1} (y - X\beta) = (W^{-1/2} y - W^{-1/2} X\beta)' (W^{-1/2} y - W^{-1/2} X\beta)$$
$$= \| \underbrace{W^{-1/2} y}_{\tilde{y}} - \underbrace{W^{-1/2} X}_{\tilde{X}} \beta \|^2$$

$$\Rightarrow \hat{\beta} = (\tilde{X}'\tilde{X})^{-1} \tilde{X}' \tilde{y} = (X' W^{-1} X)^{-1} X' W^{-1} y$$

We relax the assumption of LM

$\text{Cov}(\varepsilon) = \sigma^2 \Sigma \neq \sigma I$ (Σ known) $\Rightarrow$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{n+1}}{\operatorname{argmin}} (y - X\beta)' \Sigma^{-1} (y - X\beta)$$

$$\hat{\beta} = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} y$$

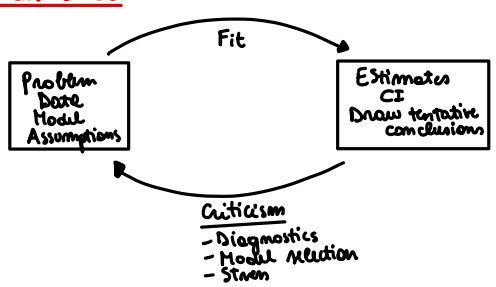
EX

$$\Sigma = \begin{bmatrix} m_1 & & \\ & m_2 & \\ & & \ddots \\ & & & m_n \end{bmatrix} \qquad \text{Var}(Y_i) \propto m_i \qquad Y_i = \sum_{1}^{m_i} T_g$$

"Weighted Least Squares" → You trust less some observations

↳ EX $Y_i = \frac{1}{m_i} \sum_j T_{ij}$ $\quad \text{Var}(Y_i) \propto \frac{1}{m_i}$ $\quad \Sigma = \begin{bmatrix} 1/m_1 & & \\ & \ddots & \\ & & 1/m_n \end{bmatrix}$

DIAGNOSTICS



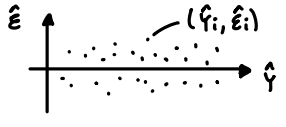
1) Residual Analysis
2) Influencial Cases
3) Collinearity

1) Residual Analysis

Model	Fitted Model	
$y = X\beta + \varepsilon$	$y = X\hat{\beta} + \hat{\varepsilon}$	$\hat{\beta} = (X'X)^{-1}X'y$
$\mathbb{E}[\varepsilon] = 0$	$\mathbb{E}[\hat{\varepsilon}] = 0$	$\hat{\varepsilon} = y - Hy = \pi_{y \mid \mathcal{L}^\perp(X)}$
$\text{Cov}(\varepsilon) = I\sigma^2$	$\text{Cov}(\hat{\varepsilon}) = \sigma^2(I - H)$	
$\varepsilon \sim N_m(0, \sigma^2 I)$		
$\varepsilon \in \mathbb{R}^m$	$\hat{\varepsilon} \in \mathcal{L}^\perp(X) \subseteq \mathbb{R}^m$	

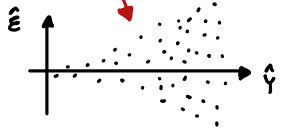
$\dim = m - (n+1)$

Plot the residuals



We don't want PATTERNS !

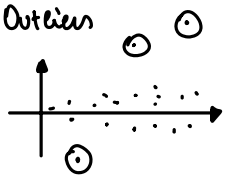
⚠ ETHEROSCHEDASTICITY

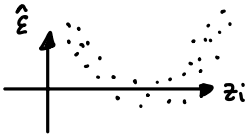


You can't trust all your observations in the same way

⇒ Weighted Least Squares (GLS), Variance Stabilizing Transformations

Outliers





$\Rightarrow$ add z_i^2 to the model
(TRANSFORM z_i to capture this behavior)

$\text{Var}(\varepsilon_i) = \sigma^2$
$\text{Var}(\hat{\varepsilon}_i) = \sigma^2(1-h_{ii})$
The variances are different

$\Rightarrow \dfrac{\hat{\varepsilon}_i}{S\sqrt{1-h_{ii}}}$ STUDENTISED RESIDUALS

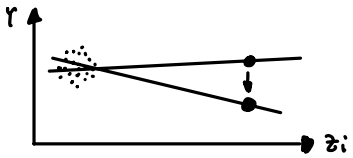
[for carefully designed experiments $h_{ii} \sim 0$ so they are the same]

You can control $H = Z(Z'Z)^{-1}Z'$

2) Influencial Cases

h_{ii} leverages $\in [0,1]$

If $h_{ii} \uparrow 1 \Rightarrow \text{Var}(\hat{\varepsilon}_i) \to 0 \Rightarrow \hat{\varepsilon}_i \to 0$ Fishy, you know this before collecting data!



[meaningful explanation in recording]

Hold out unit $i = 1, \dots, n$ from training set. Does the model change significantly?

Cook's distance $d(\hat{\beta}^{(i)}, \hat{\beta}) = D_i = \dfrac{(\hat{\beta}^{(i)} - \hat{\beta})'(Z'Z)(\hat{\beta}^{(i)} - \hat{\beta})}{S^2(p+1)} = \left(\dfrac{\hat{\varepsilon}_i}{S\sqrt{1-h_{ii}}}\right)^2 \dfrac{h_{ii}}{1-h_{ii}} \cdot \dfrac{1}{p+1}$

3) Collinearity

$\text{Cov}(\hat{\beta}) = \sigma^2(Z'Z)^{-1}$

$\text{Var}(\hat{\beta}_j) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(z_{ij} - \bar{z}_j)^2} \cdot \dfrac{1}{1-R_j^2}$

R_j^2: coefficient of determination when z_j is regressed on $z_1, \dots, \cancel{z_j}, \dots, z_r$

$R_j \uparrow 1 \Rightarrow \text{Var}(\hat{\beta}_j) \uparrow \infty$

Variance Inflation Factor "VIF"

LECTURE 24 10/5/2022

Model Selection

$Y = Z\beta + \varepsilon$

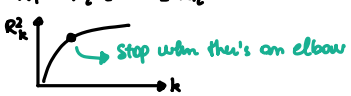
$[\underset{\in \mathbb{R}^n}{1} \; z_1 \dots z_p]$ $\to$ we have 2 possibilities for each model (2^p models)

$"(1+1)^p = \sum_{k=0}^{p} \binom{p}{k} 1^k \cdot 1^{p-k} = \sum_{k=0}^{p} \binom{p}{k}$

- **Brute force / Greedy** $\Rightarrow$ we check them ALL (pretty expensive)

  For $k = 0, \dots, p$
  - fit the $\binom{p}{k}$ models with k regressors
  - Compute R^2 and chose the one with maximum R^2

  $R_1^2 \le R_2^2 \le \dots \le R_p^2$

  Otherwise we could use R_{adj}^2 (not monotonic), AIC, BIC to chose the best k

  ▶ No intelligence, we're just checking them all and choosing according to some criteria (doable only if the models are not to big)

- **Forward / backward selection**

  ↳ 1. select one variable (best model)

  iterate (step k)

  add variable corresponding to maximum increase of $SS_{res}(M_k) - SS_{res}(M_{k+1})$

stop if the F-test fails (you cannot reject H_0)

$$F = \frac{\frac{SS_{res}(M_k) - SS_{res}(M_{k+1})}{SS_{res}(M_{k+1})}}{m-(k+1+1)} \quad \text{is small}$$

Model selection & collinearity

Note: $\underline{y} = \mathbb{Z}\underline{\beta} + \underline{\epsilon}$ is the model for the observed data from OLS we get: $\hat{\underline{\beta}} = (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\underline{y}$ assuming $\mathbb{Z}$ has full rank!

Then the fitted model for the mean of $\underline{y}$ is: $y_0 = \underline{z}_0^T\hat{\underline{\beta}}$

The fitted model always go through the baricentre of the observed data: for instance suppose $\underline{z}_0 = \frac{\mathbb{Z}^T\underline{1}}{\underline{1}^T\underline{1}} = [1 \;\; \bar{z}_1 \;\; \dots \;\; \bar{z}_r]^T$ so it's the vectors of the mean of the regressors!

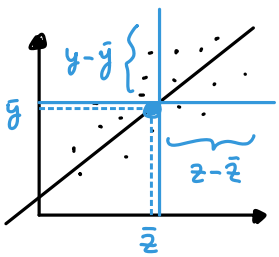
Then: $y_0 = \underline{z}_0^T\hat{\underline{\beta}} = \frac{\underline{1}^T\mathbb{Z}}{n}(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\underline{y} = \frac{1}{n}\underline{1}^T H\underline{y} = \frac{1}{n}(H\underline{1})^T\underline{y} = \frac{1}{n}\underline{1}^T\underline{y} = \bar{y}$ Hence: $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{z}_1 + \cdots + \hat{\beta}_r\bar{z}_r \implies$ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{z}_1 - \cdots - \hat{\beta}_r\bar{z}_r$ So a different representation for the fitted model is:

$$y_0 - \bar{y} = \hat{\beta}_1(z_1 - \bar{z}_1) + \cdots + \hat{\beta}_r(z_r - \bar{z}_r)$$

So instead of working with the original variables, we can first centre the data on the baricentre: $z - \bar{z}$ and $y - \bar{y}$ and then fit the model!

So **centering:** from $\underline{y}$ we consider $\mathbb{R}^n \ni \underline{y}^* = [y_1 - \bar{y}, \dots, y_n - \bar{y}]^T$ and from $\mathbb{Z}$ we consider: $\begin{bmatrix} z_{11} - \bar{z}_1 & \dots & z_{1r} - \bar{z}_r \\ \vdots & & \\ z_{n1} - \bar{z}_1 & \dots & z_{nr} - \bar{z}_r \end{bmatrix} = \mathbb{Z}^*$

which is an $n \times r$ matrix!



⇒ We don't want to lose this property
⇒ $\hat{\beta}_0$ can be recovered accordingly

OLS becomes

$$\begin{cases} \hat{\underline{\beta}}^* = \underset{\underline{\beta} \in \mathbb{R}^r}{\operatorname{argmin}} \; \| \underline{y}^* - \mathbb{Z}^*\underline{\beta} \|^2 \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1^*\bar{z}_1 - \cdots - \hat{\beta}_r^*\bar{z}_r \\ \hat{\beta}_i = \hat{\beta}_i^* \quad i = 1, \dots, r \end{cases}$$

From now on, assume $\mathbb{Z}$ $n \times r$ centred, $\underline{y} \in \mathbb{R}^n$ centred.

1) PCA regression

The problem of collinearity is due to the fact that the regressors aren't orthogonal, thus we can transform them so that we get new orthogonal regressors: we just do PCA of the dependent variables!

$\mathbb{Z} = [\underline{z}_1, \dots, \underline{z}_r]$ with $\underline{z}_i \in \mathbb{R}^n$ (Remember: these are already centred variables!) Then we perform PCA of $\mathbb{Z}$ so that we get: $PC_1, \dots, PC_r$

We can then check for an elbow and reduce the dimensionality: we consider $PC_1, \dots, PC_k$ with $k \leq r$

$\mathbb{Z}^* = [\underline{PC}_1 \cdots \underline{PC}_k]$ fit with this $\mathbb{Z}^*$ $\underline{y} = \mathbb{Z}^*\underline{\beta} + \underline{\epsilon}$, $\underline{\beta} \in \mathbb{R}^k$

$y_0 = \hat{\beta}_1 PC_1 + \cdots + \hat{\beta}_k PC_k$

- $PC_1 = e_{11}z_1 + \cdots + e_{r1}z_r$
- $PC_k = e_{1k}z_1 + \cdots + e_{rk}z_r$

substitute

stituting this into the fitted model:

$$y_0 = \underbrace{z_1(e_{11}\hat{\beta}_1 + \cdots + e_{1k}\hat{\beta}_k)}_{\hat{\gamma}_1} + \underbrace{z_2(e_{21}\hat{\beta}_1 + \cdots + e_{2k}\hat{\beta}_k)}_{\hat{\gamma}_2} + \cdots + z_r\overbrace{(e_{r1}\hat{\beta}_1 + \cdots + e_{rk}\hat{\beta}_k)}^{\hat{\gamma}_r} = z_1\hat{\gamma}_1 + \cdots + z_r\hat{\gamma}_r$$

We solved the problem of collinearity since the principal components are orthogonal, but there is no guarantee that the principal component are good directions for prediction: we might have thrown away the information of $\mathbb{Z}$ correlated with y

Note: there are algorithms that do dimension reduction without loosing too much correlation (e.g: slice-inverse regression).

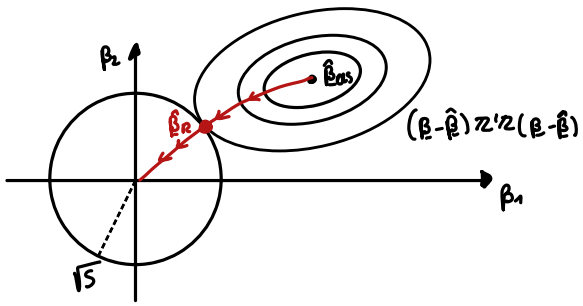
2) Ridge regression (Hoerl & Kennard 70's)

Collinearity $\Rightarrow$ $\mathrm{Var}(\hat{\beta}_i)$ explodes

Idea: constrain $\|\hat{\beta}\|^2$ to be small

$$\text{OLS: } \hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - Z\beta\|^2 = \underset{\beta}{\operatorname{argmin}} \|\underbrace{y-\hat{y}}_{\hat{\varepsilon}} - \underbrace{Z(\beta-\hat{\beta})}_{\in \mathcal{L}(Z)}\|^2 = \underset{\beta}{\operatorname{argmin}} \left[\|\hat{\varepsilon}\|^2 + \|Z(\beta-\hat{\beta})\|^2\right] = \underset{\beta}{\operatorname{argmin}} \|Z(\beta-\hat{\beta})\|^2 = \underset{\beta}{\operatorname{argmin}} (\beta-\hat{\beta})' Z'Z(\beta-\hat{\beta})$$

$$\text{RIDGE: } \begin{cases} \underset{\beta}{\operatorname{argmin}} (\beta-\hat{\beta})' Z'Z (\beta-\hat{\beta}) \\ \|\beta\|^2 \le s \end{cases}$$



OBS

- $\hat{\beta}_R$ is biased
- For suitable choice of s $\quad \overbrace{\mathbb{E}[\|\hat{\beta}_R - \beta\|^2]}^{\text{Variance}} < \mathbb{E}[\|\hat{\beta}_{ols} - \beta\|^2]$
  - ↳ How to choose s?

Lagrangian is $\underset{\beta}{\operatorname{argmin}} [\|Z(\hat{\beta}-\beta)\|^2 + \lambda\|\beta\|^2]$ $\quad \lambda = \lambda(s)$ $\quad \lambda \uparrow s \downarrow$

$$\hat{\beta}_R = (Z'Z + \lambda I)^{-1} Z'y$$

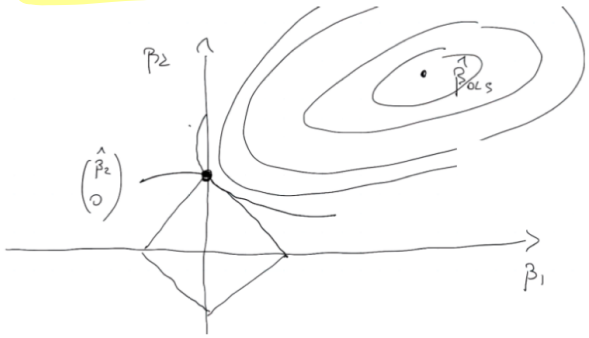
$$\mathbb{E}[\hat{\beta}_R] = (Z'Z + \lambda I)^{-1} Z'Z\beta \quad (\lambda = 0 \text{ OLS})$$

Collinearity ✓

Model selection ✗

3) Lasso regression (Tibshirani)

It's similar to ridge but here we change the constraint so that we have the situation represented below:



So instead of spherical constraints we take a diamond constraint: so that we can hope that the tangent point with the contour point will be on one of the vertices so that one of the $\hat{\beta}_i$ will be exactly zero!

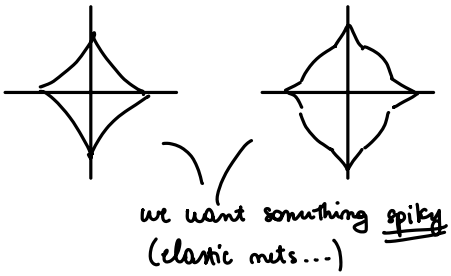
Not only we shrink the $\hat{\beta}$ but we also select the variables: feature selection!

LASSO: $\begin{cases} \underset{\beta}{\operatorname{argmin}} (\beta-\hat{\beta})' X'X (\beta-\hat{\beta}) \\ \|\beta\|_1 \le s \quad (\sum|\beta_i| \le s) \end{cases}$

Collinearity ✓

Model selection ✓

Analytical solution ✗ ⇒ numerical solutions !



we want something spiky (elastic nets...)

LECTURE 25 12/5/2022

Tree-based Methods

- Here we describe *tree-based* methods for regression and classification.
- These involve *stratifying* or *segmenting* the predictor space into a number of simple regions.
- Since the set of splitting rules used to segment the predictor space can be summarized in a tree, these types of approaches are known as *decision-tree* methods.

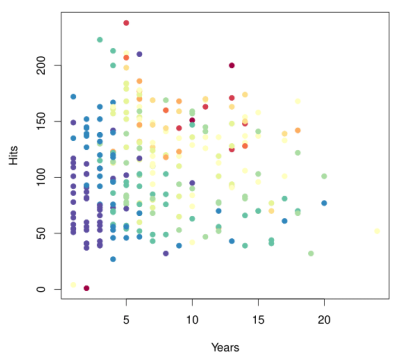
Pros and Cons

- Tree-based methods are simple and useful for interpretation.
- However they typically are not competitive with the best supervised learning approaches in terms of prediction accuracy.
- Hence we also discuss *bagging*, *random forests*, and *boosting*. These methods grow multiple trees which are then combined to yield a single consensus prediction.
- Combining a large number of trees can often result in dramatic improvements in prediction accuracy, at the expense of some loss interpretation.

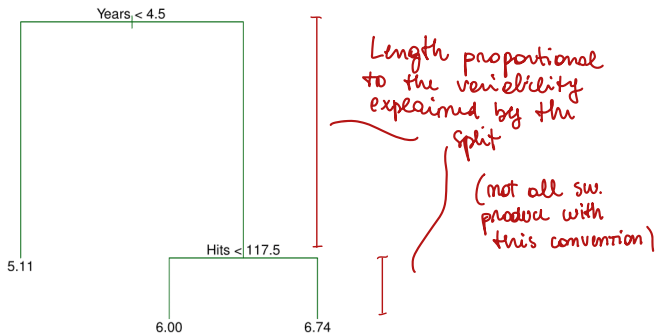
- Decision trees can be applied to both regression and classification problems.
- We first consider regression problems, and then move on to classification.

Baseball salary data: how would you stratify it?

Salary is color-coded from low (blue, green) to high (yellow,red)



Decision tree for these data



Length proportional to the variability explained by the split

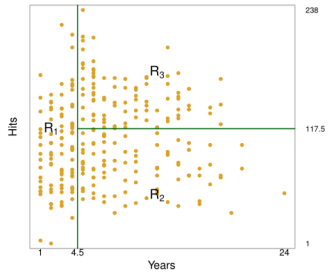
(not all sw. produce with this convention)

Details of previous figure

- For the Hitters data, a regression tree for predicting the log salary of a baseball player, based on the number of years that he has played in the major leagues and the number of hits that he made in the previous year.
- At a given internal node, the label (of the form $X_j < t_k$) indicates the left-hand branch emanating from that split, and the right-hand branch corresponds to $X_j \geq t_k$. For instance, the split at the top of the tree results in two large branches. The left-hand branch corresponds to `Years`<4.5, and the right-hand branch corresponds to `Years`>=4.5.
- The tree has two internal nodes and three terminal nodes, or leaves. The number in each leaf is the mean of the response for the observations that fall there.

Results

- Overall, the tree stratifies or segments the players into three regions of predictor space: R_1 ={X | `Years`< 4.5}, R_2 ={X | `Years`>=4.5, `Hits`<117.5}, and R_3 ={X | `Years`>=4.5, `Hits`>=117.5}.



Terminology for Trees

- In keeping with the *tree* analogy, the regions R_1, R_2, and R_3 are known as *terminal nodes*
- Decision trees are typically drawn *upside down*, in the sense that the leaves are at the bottom of the tree.
- The points along the tree where the predictor space is split are referred to as *internal nodes*
- In the hitters tree, the two internal nodes are indicated by the text `Years`<4.5 and `Hits`<117.5.

Interpretation of Results

- `Years` is the most important factor in determining `Salary`, and players with less experience earn lower salaries than more experienced players.
- Given that a player is less experienced, the number of `Hits` that he made in the previous year seems to play little role in his `Salary`.
- But among players who have been in the major leagues for five or more years, the number of `Hits` made in the previous year does affect `Salary`, and players who made more `Hits` last year tend to have higher salaries.
- Surely an over-simplification, but compared to a regression model, it is easy to display, interpret and explain

Details of the tree-building process

1. We divide the predictor space — that is, the set of possible values for $X_1, X_2, \ldots, X_p$ — into J distinct and non-overlapping regions, $R_1, R_2, \ldots, R_J$.
2. For every observation that falls into the region R_j, we make the same prediction, which is simply the mean of the response values for the training observations in R_j.

More details of the tree-building process

- In theory, the regions could have any shape. However, we choose to divide the predictor space into high-dimensional rectangles, or *boxes*, for simplicity and for ease of interpretation of the resulting predictive model.
- The goal is to find boxes $R_1, \ldots, R_J$ that minimize the RSS, given by

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2,$$

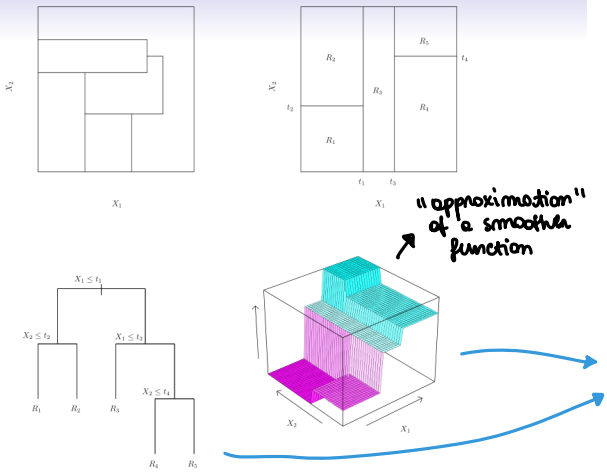
where $\hat{y}_{R_j}$ is the mean response for the training observations within the jth box.

Details— Continued

- We first select the predictor X_j and the cutpoint s such that splitting the predictor space into the regions $\{X|X_j < s\}$ and $\{X|X_j \geq s\}$ leads to the greatest possible reduction in RSS.
- Next, we repeat the process, looking for the best predictor and best cutpoint in order to split the data further so as to minimize the RSS within each of the resulting regions.
- However, this time, instead of splitting the entire predictor space, we split one of the two previously identified regions. We now have three regions.
- Again, we look to split one of these three regions further, so as to minimize the RSS. The process continues until a stopping criterion is reached; for instance, we may continue until no region contains more than five observations.

Predictions

- We predict the response for a given test observation using the mean of the training observations in the region to which that test observation belongs.
- A five-region example of this approach is shown in the next slide.



"approximation" of a smoother function

Almost always there is a bit of interaction but can be useful to understand it

Then tells that the role of X_1 and X_2 is linked

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \beta_3(X_1X_2)$$

or similar

Details of previous figure

Top Left: A partition of two-dimensional feature space that could not result from recursive binary splitting.

Top Right: The output of recursive binary splitting on a two-dimensional example.

Bottom Left: A tree corresponding to the partition in the top right panel.

Bottom Right: A perspective plot of the prediction surface corresponding to that tree.

Pruning a tree

- The process described above may produce good predictions on the training set, but is likely to *overfit* the data, leading to poor test set performance. *Why?*
- A smaller tree with fewer splits (that is, fewer regions $R_1, \ldots, R_J$) might lead to lower variance and better interpretation at the cost of a little bias.
- One possible alternative to the process described above is to grow the tree only so long as the decrease in the RSS due to each split exceeds some (high) threshold.
- This strategy will result in smaller trees, but is too *short-sighted:* a seemingly worthless split early on in the tree might be followed by a very good split — that is, a split that leads to a large reduction in RSS later on.

Pruning a tree— continued

- A better strategy is to grow a very large tree T_0, and then *prune* it back in order to obtain a *subtree*
- *Cost complexity pruning* — also known as *weakest link pruning* — is used to do this
- we consider a sequence of trees indexed by a nonnegative tuning parameter α. For each value of α there corresponds a subtree $T \subset T_0$ such that

$$\sum_{m=1}^{|T|} \sum_{i:\, x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T|$$

is as small as possible. Here $|T|$ indicates the number of terminal nodes of the tree T, R_m is the rectangle (i.e. the subset of predictor space) corresponding to the mth terminal node, and $\hat{y}_{R_m}$ is the mean of the training observations in R_m.

Choosing the best subtree

- The tuning parameter α controls a trade-off between the subtree's complexity and its fit to the training data.
- We select an optimal value $\hat{\alpha}$ using cross-validation.
- We then return to the full data set and obtain the subtree corresponding to $\hat{\alpha}$.

Summary: tree algorithm

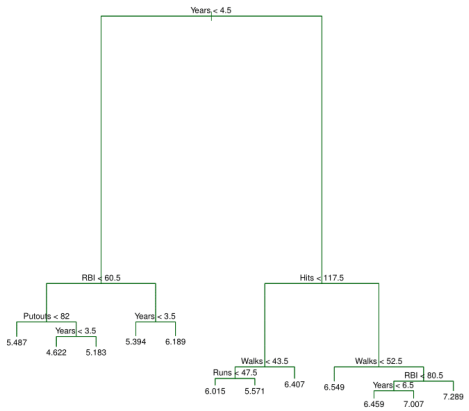
1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.
2. Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α.
3. Use K-fold cross-validation to choose α. For each $k = 1, \ldots, K$:
   - 3.1 Repeat Steps 1 and 2 on the $\frac{K-1}{K}$th fraction of the training data, excluding the kth fold.
   - 3.2 Evaluate the mean squared prediction error on the data in the left-out kth fold, as a function of α.

   Average the results, and pick α to minimize the average error.
4. Return the subtree from Step 2 that corresponds to the chosen value of α.

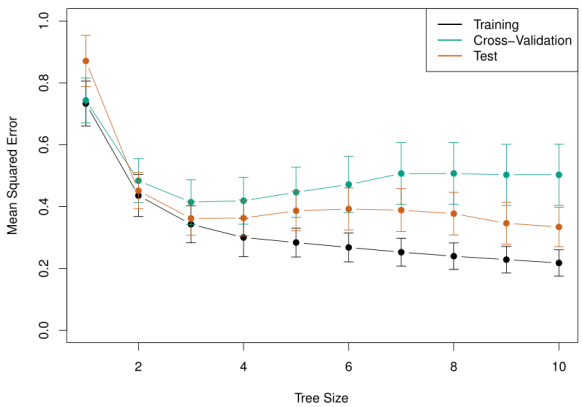
Baseball example continued

- First, we randomly divided the data set in half, yielding 132 observations in the training set and 131 observations in the test set.
- We then built a large regression tree on the training data and varied α in in order to create subtrees with different numbers of terminal nodes.
- Finally, we performed six-fold cross-validation in order to estimate the cross-validated MSE of the trees as a function of α.

Baseball example continued



Baseball example continued



Classification Trees

- Very similar to a regression tree, except that it is used to predict a qualitative response rather than a quantitative one.
- For a classification tree, we predict that each observation belongs to the *most commonly occurring class* of training observations in the region to which it belongs.

Details of classification trees

- Just as in the regression setting, we use recursive binary splitting to grow a classification tree.
- In the classification setting, RSS cannot be used as a criterion for making the binary splits
- A natural alternative to RSS is the *classification error rate*. this is simply the fraction of the training observations in that region that do not belong to the most common class:

$$E = 1 - \max_k(\hat{p}_{mk}).$$

  Here $\hat{p}_{mk}$ represents the proportion of training observations in the mth region that are from the kth class.
- However classification error is not sufficiently sensitive for tree-growing, and in practice two other measures are preferable.

Gini index and Deviance

- The *Gini index* is defined by

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

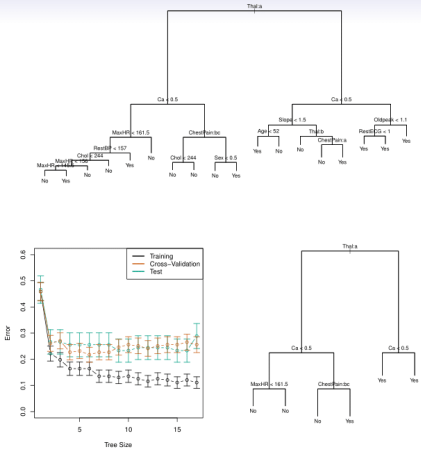
  a measure of total variance across the K classes. The Gini index takes on a small value if all of the $\hat{p}_{mk}$'s are close to zero or one.
- For this reason the Gini index is referred to as a measure of node *purity* — a small value indicates that a node contains predominantly observations from a single class.
- An alternative to the Gini index is *cross-entropy*, given by

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$$

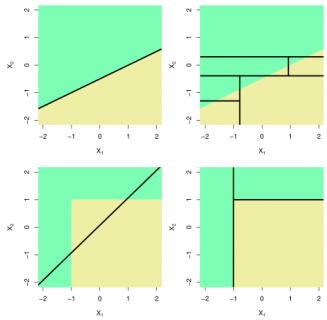
- It turns out that the Gini index and the cross-entropy are very similar numerically.

Example: heart data

- These data contain a binary outcome `HD` for 303 patients who presented with chest pain.
- An outcome value of `Yes` indicates the presence of heart disease based on an angiographic test, while `No` means no heart disease.
- There are 13 predictors including `Age`, `Sex`, `Chol` (a cholesterol measurement), and other heart and lung function measurements.
- Cross-validation yields a tree with six terminal nodes. See next figure.



Trees Versus Linear Models



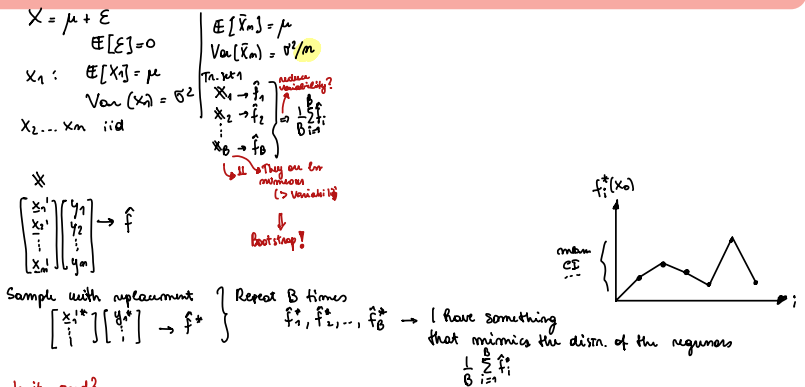
Top Row: True linear boundary; Bottom row: true non-linear boundary.

Left column: linear model; Right column: tree-based model

Advantages and Disadvantages of Trees

- ▲ Trees are very easy to explain to people. In fact, they are even easier to explain than linear regression!
- ▲ Some people believe that decision trees more closely mirror human decision-making than do the regression and classification approaches seen in previous chapters.
- ▲ Trees can be displayed graphically, and are easily interpreted even by a non-expert (especially if they are small).
- ▲ Trees can easily handle qualitative predictors without the need to create dummy variables.
- ▼ Unfortunately, trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches seen in this book.

However, by aggregating many decision trees, the predictive performance of trees can be substantially improved. We introduce these concepts next.



Bagging

- *Bootstrap aggregation*, or *bagging*, is a general-purpose procedure for reducing the variance of a statistical learning method; we introduce it here because it is particularly useful and frequently used in the context of decision trees.
- Recall that given a set of n independent observations $Z_1, \ldots, Z_n$, each with variance σ^2, the variance of the mean $\bar{Z}$ of the observations is given by σ^2/n.
- In other words, *averaging a set of observations reduces variance*. Of course, this is not practical because we generally do not have access to multiple training sets.

Bagging— continued

- Instead, we can bootstrap, by taking repeated samples from the (single) training data set.
- In this approach we generate B different bootstrapped training data sets. We then train our method on the bth bootstrapped training set in order to get $\hat{f}^{*b}(x)$, the prediction at a point x. We then average all the predictions to obtain

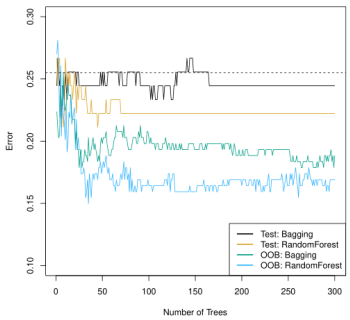
$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x).$$

This is called *bagging*.

Bagging classification trees

- The above prescription applied to regression trees
- For classification trees: for each test observation, we record the class predicted by each of the B trees, and take a *majority vote*: the overall prediction is the most commonly occurring class among the B predictions.

Bagging the heart data



Details of previous figure

Bagging and random forest results for the Heart data.

- The test error (black and orange) is shown as a function of B, the number of bootstrapped training sets used.
- Random forests were applied with $m = \sqrt{p}$.
- The dashed line indicates the test error resulting from a single classification tree.
- The green and blue traces show the OOB error, which in this case is considerably lower

Out-of-Bag Error Estimation

- It turns out that there is a very straightforward way to estimate the test error of a bagged model.
- Recall that the key to bagging is that trees are repeatedly fit to bootstrapped subsets of the observations. One can show that on average, each bagged tree makes use of around two-thirds of the observations.
- The remaining one-third of the observations not used to fit a given bagged tree are referred to as the *out-of-bag* (OOB) observations.
- We can predict the response for the ith observation using each of the trees in which that observation was OOB. This will yield around $B/3$ predictions for the ith observation, which we average.
- This estimate is essentially the LOO cross-validation error for bagging, if B is large.

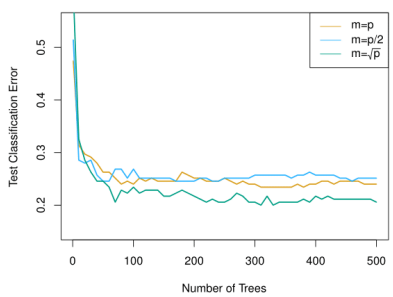
Random Forests

- *Random forests* provide an improvement over bagged trees by way of a small tweak that *decorrelates* the trees. This reduces the variance when we average the trees.
- As in bagging, we build a number of decision trees on bootstrapped training samples.
- But when building these decision trees, each time a split in a tree is considered, *a random selection of m predictors* is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors.
- A fresh selection of m predictors is taken at each split, and typically we choose $m \approx \sqrt{p}$ — that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors (4 out of the 13 for the Heart data).

Example: gene expression data

- We applied random forests to a high-dimensional biological data set consisting of expression measurements of 4,718 genes measured on tissue samples from 349 patients.
- There are around 20,000 genes in humans, and individual genes have different levels of activity, or expression, in particular cells, tissues, and biological conditions.
- Each of the patient samples has a qualitative label with 15 different levels: either normal or one of 14 different types of cancer.
- We use random forests to predict cancer type based on the 500 genes that have the largest variance in the training set.
- We randomly divided the observations into a training and a test set, and applied random forests to the training set for three different values of the number of splitting variables m.

Results: gene expression data



Details of previous figure

- Results from random forests for the fifteen-class gene expression data set with $p = 500$ predictors.
- The test error is displayed as a function of the number of trees. Each colored line corresponds to a different value of m, the number of predictors available for splitting at each interior tree node.
- Random forests ($m < p$) lead to a slight improvement over bagging ($m = p$). A single classification tree has an error rate of 45.7%.

Boosting

- Like bagging, boosting is a general approach that can be applied to many statistical learning methods for regression or classification. We only discuss boosting for decision trees.
- Recall that bagging involves creating multiple copies of the original training data set using the bootstrap, fitting a separate decision tree to each copy, and then combining all of the trees in order to create a single predictive model.
- Notably, each tree is built on a bootstrap data set, independent of the other trees.
- Boosting works in a similar way, except that the trees are grown *sequentially*: each tree is grown using information from previously grown trees.

Boosting algorithm for regression trees

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.
2. For $b = 1, 2, \ldots, B$, repeat:
   - 2.1 Fit a tree $\hat{f}^b$ with d splits ($d + 1$ terminal nodes) to the training data (X, r).
   - 2.2 Update $\hat{f}$ by adding in a shrunken version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x).$$

   - 2.3 Update the residuals,

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i).$$

3. Output the boosted model,

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x).$$

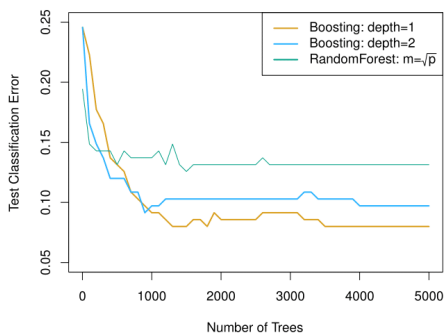
What is the idea behind this procedure?

- Unlike fitting a single large decision tree to the data, which amounts to *fitting the data hard* and potentially overfitting, the boosting approach instead *learns slowly*.
- Given the current model, we fit a decision tree to the residuals from the model. We then add this new decision tree into the fitted function in order to update the residuals.
- Each of these trees can be rather small, with just a few terminal nodes, determined by the parameter d in the algorithm.
- By fitting small trees to the residuals, we slowly improve $\hat{f}$ in areas where it does not perform well. The shrinkage parameter λ slows the process down even further, allowing more and different shaped trees to attack the residuals.

Boosting for classification

- Boosting for classification is similar in spirit to boosting for regression, but is a bit more complex. We will not go into detail here, nor do we in the text book.
- Students can learn about the details in *Elements of Statistical Learning, chapter 10.*
- The R package `gbm` (gradient boosted models) handles a variety of regression and classification problems.

Gene expression data continued



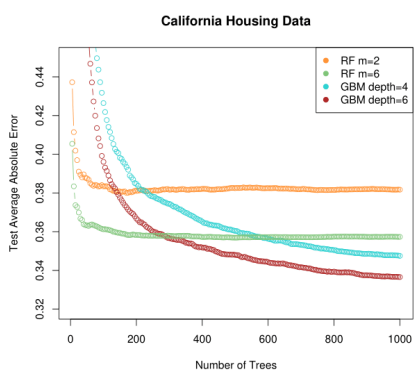
Details of previous figure

- Results from performing boosting and random forests on the fifteen-class gene expression data set in order to predict *cancer* versus *normal*.
- The test error is displayed as a function of the number of trees. For the two boosted models, $\lambda = 0.01$. Depth-1 trees slightly outperform depth-2 trees, and both outperform the random forest, although the standard errors are around 0.02, making none of these differences significant.
- The test error rate for a single tree is 24%.

Tuning parameters for boosting

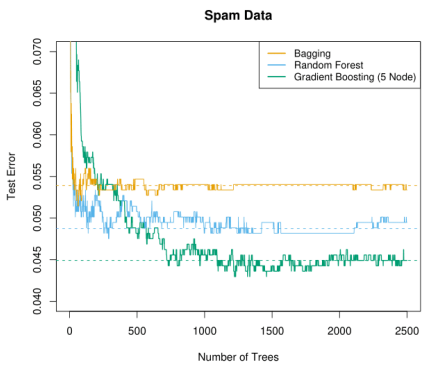
1. The *number of trees* B. Unlike bagging and random forests, boosting can overfit if B is too large, although this overfitting tends to occur slowly if at all. We use cross-validation to select B.
2. The *shrinkage parameter* λ, a small positive number. This controls the rate at which boosting learns. Typical values are 0.01 or 0.001, and the right choice can depend on the problem. Very small λ can require using a very large value of B in order to achieve good performance.
3. The *number of splits* d in each tree, which controls the complexity of the boosted ensemble. Often $d = 1$ works well, in which case each tree is a *stump*, consisting of a single split and resulting in an additive model. More generally d is the *interaction depth*, and controls the interaction order of the boosted model, since d splits can involve at most d variables.

Another regression example



from *Elements of Statistical Learning, chapter 15.*

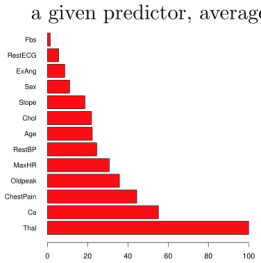
Another classification example



from *Elements of Statistical Learning, chapter 15.*

Variable importance measure

- For bagged/RF regression trees, we record the total amount that the RSS is decreased due to splits over a given predictor, averaged over all B trees. A large value indicates an important predictor.
- Similarly, for bagged/RF classification trees, we add up the total amount that the Gini index is decreased by splits over a given predictor, averaged over all B trees.



Variable importance plot for the `Heart` data

Summary

- Decision trees are simple and interpretable models for regression and classification
- However they are often not competitive with other methods in terms of prediction accuracy
- Bagging, random forests and boosting are good methods for improving the prediction accuracy of trees. They work by growing many trees on the training data and then combining the predictions of the resulting ensemble of trees.
- The latter two methods— random forests and boosting— are among the state-of-the-art methods for supervised learning. However their results can be difficult to interpret.

LECTURE 16/5/2022

MIXED EFFECT MODELS

▶ MOTIVATIONS

General overview

- Regression models are a very common statistical tool.
  A very important assumption of regression models is the independence and homoscedasticity of the observations.
- *How often such assumptions are met in the real world*? Seldom, to be optimistic.
  => a more general framework to properly account for complex structure present into the data is needed.
- Modeling general dependence and heteroscedasticity among observations concerns the design of the variance covariance matrix of the errors.

Ex: It is often the case that statistical units are correlated and/or grouped within a hierarchy:

- Measurements of the same individuals over time
- Patients sharing the same GP, hospital of admission, district, etc.
- Pupils grouped into classes, schools, districts, etc.
- Units belonging to close districts.

General overview

- Observations within groups are more similar (possibly correlated) than observations between groups
  - Independence hypothesis does not hold
  - Ignoring the dependence structure induces biased estimates of parameters
- Linear Mixed Models – LMMs (a.k.a. Random Effects Models or Multilevel/Hierarchical models) are a generalization of traditional regression models (Fixed Effect Models), adding to the linear predictor (a.k.a. fixed component) a random component.
- LMMs allow to disentangle the contribution of different kind of dependencies among observations, focusing on inference related to the presence of groups.

Motivating examples

1. ARMD - Age-Related Macular Degeneration Trial

- The ARMD data arise from a **randomized multi-center clinical** trial comparing an experimental treatment (interferon-a) versus placebo for patients diagnosed with ARMD.
- We focus on the **comparison between placebo and the highest dose** (6 million units daily) **of interferon-α**.
- Patients with macular degeneration progressively lose vision.
- Visual acuity of each of **240 patients** was **assessed at baseline and at four post-randomization timepoints** (4, 12, 24, and 52 weeks).
- Visual acuity was evaluated based on **patient's ability to read lines of letters on standardized vision charts**. The charts display lines of five letters of decreasing size, which the patient must read from top (largest letters) to bottom (smallest letters).
- In our analyses, we will focus on the **visual acuity defined as the total number of letters correctly read.**

=> longitudinal data in the form of up to five visual acuity measurements collected at different, but common to all patients, timepoints.

Long & short format

Motivating examples

2. Multi center trials

- Many trials are multi-center, due to the inadequate number of patients in a single center (rare disease) or to shared studies and protocols.
- Often the analysis ignores the center from which the data were obtained, making the implicit assumption that centers are identical. But they are not: differences usually arise in a) case mix, b) overall success in recruitment and outcome and c) relative benefit (hospital effect).

- Observational multicenter study (N=48 hospitals) on primary lung cancer.
- Investigation of clinical factors associated to 3y death for any cause in **802 pts** undergoing surgery.

- Goals:
  - Risk stratification of the population under study
    => association between mortality and features measured at baseline/entrance
  - Assessment of the grouper effect and scenario analysis

Motivating examples

3. Schools

- Nowadays, Italian (but also international) students are tested by means of standardized tests given at different grades.
- The *school dataset* collects information about the tests of 1000 students enrolled in 50 different primary schools.
- For each student, beside the test result, we observe the gender, the socioeconomic index and the anonymous id of the school in which he/she is enrolled.
- Being enrolled in a school might have a relevant effect on student test scores (e.g. very bad/good teachers) → students are not independent

- **n = 1000 students** within **N = 50 schools**
- Investigation of the association between student-level characteristics and student test scores
- Investigation of the school effect on student test scores

- Goals:
  - Prediction of student test scores
    => association between test score and gender and socioeconomic index of the student
  - Assessment of the school effect and scenario analysis

- LM
- LM 2.0 relaxing homogeneity
- LM 3.0 relaxing independence
- LMM

LM

- LMs are used to quantify the relationship between a dependent variable and a set of covariates with the use of a linear function depending on a (possibly) small number of regression parameters.
- LMs are suitable for analyzing data involving **independent observations with a homogenous variance** (e.g., standard linear regression, ANOVA/ANCOVA models).

- The classical LM for **independent, normally distributed observations** y_j , j = 1,...,n with a constant variance can be specified in a variety of ways. A commonly used specification is:

Model equation at the level of observations

$$y_j = \beta_0 + \beta_1 z_{1j} + \cdots + \beta_p z_{pj} + \varepsilon_j = \boldsymbol{z}_j^t \boldsymbol{\beta} + \varepsilon_j$$

$$\varepsilon_j \sim N(0, \sigma^2) \qquad j \perp j' \quad \text{independent errors}$$

Note that $E[y_j] = \mu_j = \boldsymbol{z}_j^t \boldsymbol{\beta}$

$V[y_j] = V[\varepsilon_j] = \sigma^2$

LM

- LMs are used to quantify the relationship between a dependent variable and a set of covariates with the use of a linear function depending on a (possibly) small number of regression parameters.
- LMs are suitable for analyzing data involving **independent observations with a homogenous variance** (e.g., standard linear regression, ANOVA/ANCOVA models).

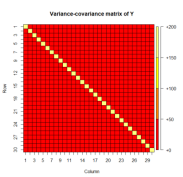
- The classical LM for **independent, normally distributed observations** y_j , j = 1,...,n with a constant variance can be specified in a variety of ways. A commonly used specification is:

Model equation for all data

$$\boldsymbol{y} = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon}$$

Scale parameter

$$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, R) \qquad R = \sigma^2 I_n \in \mathbb{R}^{n \times n}$$



$$\boldsymbol{y} = (y_1, \ldots, y_n)' \qquad \boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$$

$$\mathbf{Z} = \begin{pmatrix} \boldsymbol{z_1} \\ \cdots \\ \boldsymbol{z_j} \\ \cdots \\ \boldsymbol{z_n} \end{pmatrix} = \begin{pmatrix} 1 & z_{11} & \ldots & z_{p1} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & z_{1n} & \ldots & z_{pn} \end{pmatrix} = \begin{pmatrix} \boldsymbol{1'} & \boldsymbol{z_1'} & \ldots & \boldsymbol{z_p'} \end{pmatrix} \in \mathbb{R}^{n \times (p+1)}$$

LM - estimation

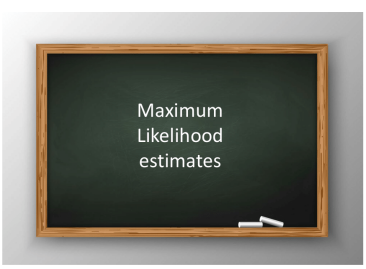
- **Goal**: finding estimates of a set of parameters **β** and σ^2.
- Estimation via least squares (**OLS**)
- However, OLS is less suitable for more complex LMs, including LMMs.

$$\widehat{\boldsymbol{\beta}}_{OLS} = (Z'Z)^{-1} Z'\boldsymbol{y}$$

Note: OLS estimate does not require the normality assumption
Valid under the assumption of uncorrelated residual errors

$$\hat{\sigma}^2_{OLS} = \frac{1}{n-(p+1)} (\boldsymbol{y} - Z\widehat{\boldsymbol{\beta}}_{OLS})'(\boldsymbol{y} - Z\widehat{\boldsymbol{\beta}}_{OLS})$$

Note: OLS estimates are **unbiased**



LM - estimation

Maximum Likelihood (ML) estimation

- The likelihood function for a Normal LM is

$$L_{full}(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = (2\pi\sigma^2)^{-n/2} \prod_{j=1}^{n} exp\left[-\frac{(y_j - \boldsymbol{z}_j'\boldsymbol{\beta})^2}{2\sigma^2}\right]$$

$$l_{full}(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(\boldsymbol{y}_j - \boldsymbol{z}_j'\boldsymbol{\beta})^2$$

- Maximization of l_{full} provides the ML estimates of unknown parameters

$$\widehat{\boldsymbol{\beta}}_{ML} = (Z'Z)^{-1}Z'\boldsymbol{y}$$ Same as OLS

$$\hat{\sigma}^2_{ML} = \frac{1}{n}(\boldsymbol{y} - Z\widehat{\boldsymbol{\beta}}_{ML})'(\boldsymbol{y} - Z\widehat{\boldsymbol{\beta}}_{ML}) = \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{y}_j - \boldsymbol{z}_j'\widehat{\boldsymbol{\beta}}_{ML})^2$$ Biased!

LM - estimation

Restricted Maximum Likelihood (REML) estimation

- To obtain an unbiased estimate for σ^2, an approach that is orthogonal to the estimation of $\boldsymbol{\beta}$ is needed. This is possible considering the likelihood function based on a set of n-(p+1) independent contrast of y.

$$l_{REML}(\sigma^2; \boldsymbol{y}) = -\frac{n-(p+1)}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n} r_j^2 \quad \Rightarrow \quad \hat{\sigma}^2_{REML} = \frac{1}{n-(p+1)}\sum_{j=1}^{n} r_j^2$$

where r_j are the residuals.

- OLS estimators are equivalent to the REML estimates.
  This equivalence holds for classical LM with independent, homoscedastic errors, NOT for more complex formulations.

- The REML objective function does not allow to directly estimate the coefficients $\boldsymbol{\beta}$. The ML formula should be used in this case. Also in this case, the equivalence between ML and REML estimates is true for the classical ML only.

- LM 2.0 relaxing homogenity

LM 2.0

- In the previous case, we formulated the classical LM for independent observations.
  ⇒ key assumptions:
  1. observations are independent and normally distributed with a constant, i.e., homogeneous variance
  2. the expected value of the observations can be expressed as a linear function of covariates.

- We now relax the homoscedasticity assumption and allow for the observations to be heteroscedastic, i.e., to have different variances, while retaining the assumption that the observations are independent and normally distributed.
  ⇒ LMs with heterogeneous variance

- Important concept: *variance function*

- The new setting will ask for suitable estimation methods (e.g WLS, GLS and IRLS estimation).

LM 2.0

- In the classical LM with homogeneous variance, the variance of the dependent variable is $V[y_j] = \sigma^2$
- We now **relax the constant variance assumption** and assume that $V[y_j] = \sigma_j^2$

Therefore:

Model equation at the level of observations

$$y_j = \beta_0 + \beta_1 z_{1j} + \cdots + \beta_p z_{pj} + \varepsilon_j = \boldsymbol{z}_j^t \boldsymbol{\beta} + \varepsilon_j \qquad j = 1, \ldots n$$

$$\varepsilon_j \sim N(0, \sigma_j^2) \qquad j \perp j' \quad \text{independent errors}$$

Note that

$$E[y_j] = \mu_j = \boldsymbol{z}_j^t \boldsymbol{\beta}$$

$$V[y_j] = V[\varepsilon_j] = \sigma_j^2$$

- The model contains n+p+1 > n parameters => the model is **not identifiable!**
- It may become identifiable if we impose **additional constraints on the residual variances**
  1. Assume known variance weights
  2. More general way: to **represent variances** more parsimoniously **as a function of a small set of parameters** => **variance function**.

LM 2.0 – Variance Function

- A more general and flexible way to **introduce variance heterogeneity is by means of a variance function**

$$\lambda(\boldsymbol{\delta}, \mu; \boldsymbol{v})$$

which assumes positive values and is continuous and differentiable with respect to δ.

- Therefore:

$$V[\varepsilon_j] = \sigma^2 \lambda^2(\boldsymbol{\delta}, \mu_j; \boldsymbol{v}_j)$$

where -> $\boldsymbol{v}_j$ is a **vector of (known) covariates** defining the variance function for observation j,
-> $\boldsymbol{\delta}$ contains a **small set of variance parameters**, common to all observations.

Note that:
- because the function λ(·) involves μ_j, it in fact depends on **β**, too
- the parameter σ should be interpreted as a scale parameter only and no more as residual error standard deviation.

LM 2.0 – Variance Function

Therefore:

Model equation at the level of observations

$$y_j = \beta_0 + \beta_1 z_{1j} + \cdots + \beta_p z_{pj} + \varepsilon_j = \boldsymbol{z}_j^t \boldsymbol{\beta} + \varepsilon_j$$

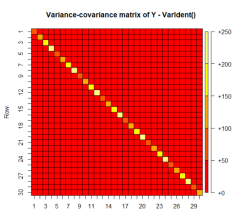
$$\varepsilon_j \sim N(0, \sigma^2 \lambda_j^2)$$

where $\lambda_j^2 = \lambda^2(\boldsymbol{\delta}, \mu_j; \boldsymbol{v}_j)$ and $j \perp j'$ independent errors

Note that

$$E[y_j] = \mu_j = \boldsymbol{z}_j^t \boldsymbol{\beta}$$

$$V[y_j] = V[\varepsilon_j] = \sigma^2 \lambda_j^2$$



Model equation at the level of observations

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, R) \qquad R = \sigma^2 \Lambda\Lambda \quad \text{where} \quad \Lambda = diag(\lambda_1, \ldots, \lambda_n)$$

$$\boldsymbol{R} \in \mathbb{R}^{n \times n}$$

LM 2.0 – Variance Function

Examples of the variance function $\lambda(\cdot)$:

1. Known weights, $\lambda(\cdot) = \lambda(v)$
2. Variance functions depending on δ but not on μ => $\lambda(\cdot) = \lambda(\boldsymbol{\delta};v)$
3. Variance functions depending on δ and μ => $\lambda(\cdot) = \lambda(\boldsymbol{\delta},\mu;v)$
4. Variance functions depending on μ but not on δ => $\lambda(\cdot) = \lambda(\mu;v)$

- We will symbolically refer to groups 2–4 as **<δ>-, <δ,μ>-, and <μ>- group**, respectively.

- **Specification of an LM with heterogeneous variance is very general** and encompasses all four groups of variance functions => the use of a variance function from any of the aforementioned groups does not pose difficulties in terms of the model specification.

  However, in models involving variance functions from groups <δ,μ> or <μ>, the parameters $\boldsymbol{\beta}$ are shared by the mean and variance structures (*mean-variance models)* => require different estimation approaches and inference techniques, as compared to the models involving known weights or variance functions from the <δ>-group.

LM 2.0 – Variance Function

Table 7.1 A summary of the parts of Chap. 7 that contain the information about particular groups of variance functions and the corresponding estimation methods

Group	Arguments $\boldsymbol{\delta}$	Arguments μ_i	Examples	Estimation algorithm	Section
Known weights	–	–	varFixed$(\cdot)$	WLS	7.4.1
<δ>	+	–	Table 7.2	ML/REML	7.4.2
<δ,μ>	+	+	Table 7.3	ML/REML-based GLS	7.8.1.1
<μ>	–	+	Table 7.4	IRLS	7.8.1.2

Table 7.2 Examples of variance functions from the <δ>-group[a]

Function $\lambda(\cdot)$	λ_i	Description
varPower$(\boldsymbol{\delta}; v_i, s_i)$	$\lvert v_i\rvert^{\delta_{s_i}}$	Power of a variance covariate v_i
varExp$(\boldsymbol{\delta}; v_i, s_i)$	$\exp(v_i\delta_{s_i})$	Exponent of a variance covariate
varConstPower$(\boldsymbol{\delta}, v_i, s_i)$	$\delta_{1,s_i} + \lvert v_i\rvert^{\delta_{2,s_i}}$	Constant plus power variance function $\delta_{1,s_i} > 0$
varIdent$(\boldsymbol{\delta}, s_i)$	δ_{s_i}	Different variances per stratum $\delta_1 \equiv 1$, $\delta_s > 0$ for $s \neq 1$

s_j = stratum the j-th obs belongs to

Table 7.3 Examples of variance functions from the <δ,μ>-group[a]

Function $\lambda(\cdot)$	λ_i	Description
varPower$(\boldsymbol{\delta}, \mu_i; s_i)$	$\lvert\mu_i\rvert^{\delta_{s_i}}$	Power of $\lvert\mu_i\rvert$
varExp$(\boldsymbol{\delta}, \mu_i; s_i)$	$\exp(\mu_i\delta_{s_i})$	Exponent of μ_i
varConstPower$(\boldsymbol{\delta}, \mu_i; s_i)$	$\delta_{1,s_i} + \lvert\mu_i\rvert^{\delta_{2,s_i}}$	Constant plus power variance function $\delta_{1,s_i} > 0$

Table 7.4 Examples of variance functions from the <μ>-group[a]

Function $\lambda(\cdot)$	λ_i	Description
varPower$(\mu_i; s_i, \boldsymbol{\delta})$	$\lvert\mu_i\rvert^{\delta_{s_i}}$	Power of $\lvert\mu_i\rvert$, δ_{s_i} known
varExp$(\mu_i; s_i, \boldsymbol{\delta})$	$\exp(\mu_i\delta_{s_i})$	Exponent of μ_i, δ_{s_i} known
varConstPower$(\mu_i; s_i, \boldsymbol{\delta})$	$\delta_{1,s_i} + \lvert\mu_i\rvert^{\delta_{2,s_i}}$	Constant plus power variance function, $\delta_{1,s_i} > 0$, δ_{1,s_i} and δ_{2,s_i} known

LM 2.0 – Estimation

Profiled Likelihood

- The log-likelihood function for the LM 2.0 model is:

$$l_{full}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{j=1}^{n}\log(\lambda_j^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}\lambda_j^{-2}(\boldsymbol{y}_j - \boldsymbol{x}_j'\boldsymbol{\beta})^2$$

  Note that it depends on $\boldsymbol{\delta}$ through λ, then in the case of λ=1 we step back to the ML of the usual LM case.

- **Profiling of a likelihood function** can be done in a variety of ways. Here we follow the profiling approach implemented in the **gls()** function of the **nlme package**.
  => We first profile out the $\boldsymbol{\beta}$ parameters, and then, we profile out σ^2.
- The **advantage** of using the function is that it does not depend on all the initial parameters. Thus, optimization of the function is performed in a **parameter space of a lower dimension**.

Assume that $\boldsymbol{\delta}$ is known. Then,

$$maximize\ l_{full}\ wrt\ \boldsymbol{\beta}\ for\ any\ value\ of\ \boldsymbol{\delta} \Rightarrow \widehat{\boldsymbol{\beta}}(\boldsymbol{\delta}) \Rightarrow l^*_{ML}(\sigma^2, \boldsymbol{\delta})$$

$$\Rightarrow maximize\ l^*_{ML}\ wrt\ \sigma^2\ for\ any\ value\ of\ \boldsymbol{\delta} \Rightarrow \hat{\sigma}^2_{ML}(\boldsymbol{\delta}) \Rightarrow l^*_{ML}(\boldsymbol{\delta})$$

LM 2.0 – Estimation

Profiled Likelihood

- ML estimates of unknown parameters of LM 2.0:

$$\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}}_{ML}) = \left(\sum_{j=1}^{n} \hat{\lambda}_j^{-2} \boldsymbol{x}_j \boldsymbol{x}_j'\right)^{-1} \sum_{j=1}^{n} \hat{\lambda}_j^{-2} \boldsymbol{x}_j y_j \qquad \hat{\sigma}^2_{ML} = \hat{\sigma}^2(\hat{\boldsymbol{\delta}}_{ML}) = \sum_{j=1}^{n} \hat{\lambda}_j^{-2} \hat{r}_j^2 y_j$$

$$\hat{\boldsymbol{\lambda}}_j = \boldsymbol{\lambda}(\hat{\boldsymbol{\delta}}_{ML}; \boldsymbol{v}_j) \quad \hat{\boldsymbol{r}}_j = \boldsymbol{r}_j(\hat{\boldsymbol{\delta}}_{ML})$$

- Biased => REML approach

$$l_{REML}(\sigma^2, \boldsymbol{\delta}) = -\frac{n-(p+1)}{2}\log(\sigma^2) - \frac{1}{2}\sum_{j=1}^{n}\log(\lambda_j^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}\lambda_j^{-2} r_j^2 - \frac{1}{2} log\left[det\left(\sum_{j=1}^{n}\lambda_j^{-2}\boldsymbol{x}_j\boldsymbol{x}_j'\right)\right]$$

$$\Rightarrow \quad \hat{\sigma}^2_{REML}(\hat{\boldsymbol{\delta}}) = \sum_{j=1}^{n} \hat{\lambda}_j^{-2} \hat{r}_j^2 / (n-p-1)$$

- By maximization of l_{REML} with respect to $\boldsymbol{\delta}$, we obtain an estimator $\hat{\boldsymbol{\delta}}_{REML}$ of $\boldsymbol{\delta}$.
  This is used to yield an estimator $\hat{\boldsymbol{\beta}}_{REML}$ of $\boldsymbol{\beta}$.

▶ LM 3.0 relaxing independence

LM 3.0

- The essential assumption for the LMs considered before was that the **observations** collected during the study were **independent of each other**.
- This **assumption is restrictive** in studies which use sampling designs that lead to correlated data.
  - Studies collecting measures over time, i.e., in a longitudinal fashion;
  - Designs involving **hierarchies or grouping** (e.g., cluster-randomization clinical trials; in studies collecting spatially correlated data, etc.)

  Note that for such designs, the distinction between sampling units (e.g., subjects in a longitudinal study) and analysis units (e.g., time-specific measurements) is important.

- We now consider a class of **more general LMs** that allow **relaxing the assumptions of independence**, namely *LMs with fixed effects and correlated residual errors for grouped data*, or simply as LMs for correlated data.
- These models can be viewed as an example of population-averaged models, i.e., models in which the parameters are interpreted as quantifying effects of covariates on the marginal mean value of the dependent variable for the entire population.

- Important concept: *correlation structure*.

LM 3.0

- We now introduce LM with fixed effects and correlated residual errors for grouped data with hierarchical structure.
  - **Single-level of grouping**, with **N groups** (levels of a grouping factor) indexed by *i* (i = 1,...,N)
  - n_i **observations per group** indexed by *j* (j = 1,..., n_i).

- Then, for the group i

$$\boldsymbol{y_i} = Z_i\boldsymbol{\beta} + \boldsymbol{\varepsilon_i} \qquad \boldsymbol{\varepsilon_i} \sim N_{n_i}(\boldsymbol{0}, R_i) \qquad R_i = \sigma^2 \mathrm{R}_i$$

Multiple level of grouping is possible, but requires suitable design of Z matrix and software specifications.

Model equation at the level of group of observations

$$\boldsymbol{y_i} = (y_{i1}, ..., y_{in_i})' \quad \boldsymbol{\varepsilon_i} = (\varepsilon_{i1}, ..., \varepsilon_{in_i})'$$

$$\boldsymbol{Z_i} = \begin{pmatrix} \boldsymbol{z_{i1}} \\ \cdots \\ \boldsymbol{z_{ij}} \\ \cdots \\ \boldsymbol{z_{in_i}} \end{pmatrix} = \begin{pmatrix} 1 & z_{i11} & \cdots & z_{ip1} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & z_{i1n_i} & \cdots & z_{ipn_i} \end{pmatrix} = \begin{pmatrix} \boldsymbol{1}' & \boldsymbol{z}'_{\boldsymbol{i1}} & \cdots & \boldsymbol{z}'_{\boldsymbol{ip}} \end{pmatrix}$$

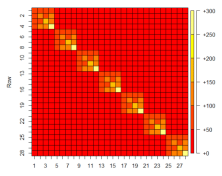
Note that

$$E[y_{ij}] = \mu_{ij} = \boldsymbol{z}_{ij}^t\boldsymbol{\beta} \qquad V[\boldsymbol{y}_i] = V[\boldsymbol{\varepsilon}_i] = \sigma^2 \mathrm{R}_i$$

LM 3.0 – Correlation function

- This **model is not identifiable** in its most general form due to i) the non uniqueness of the R_i representation and ii) the model potentially involves too many unknown parameters.
  => additional constraints

$$R_i = \sigma^2 \mathrm{R}_i = \sigma^2 \Lambda_i C_i \Lambda_i \quad \text{where} \quad \Lambda_i = (\lambda_{i1}, \ldots, \lambda_{in_i}) \quad and \quad C_i = corr\ matrix$$



allows for heteroscedasticity of observations within group *i*

allows for correlation of observations within group *i*

- **Correlation function** (general version): $Corr(\varepsilon_{ij}, \varepsilon'_{ij}) = h[d(t_{ij}, t'_{ij}), \boldsymbol{\varrho}]$ where
  - $\boldsymbol{\varrho}$ is a vector of correlation parameters
  - $d(.)$ is a distance function of vectors of position variables t_{ij} and t'_{ij} corresponding to, ε_{ij} and ε'_{ij}
  - h[.] is a continuous function with respect to $\boldsymbol{\varrho}$, ranging between −1 and 1 and s.t. h(0, $\boldsymbol{\varrho}$) ≡ 1.

LM 3.0 – Correlation function

- By assuming various distances and correlation functions, a variety of correlation structures can be obtained.

- The correlation structures can be classified into **two main groups**:

1. *"Serial" structures*
   => correlation structures which are defined in the context of **time-series** or **longitudinal data**.

Ex: Autocorrelation => $Corr(\varepsilon_{ij}, \varepsilon'_{ij}) = \rho$

2. *"Spatial" structures*
   => correlation structures which are defined in the context of **spatial data**.

Ex: $Corr(\varepsilon_{ij}, \varepsilon'_{ij}) = e^{-s_{ij,ij'}/\varrho}$ => $h(s, \varrho) = e^{-s/\varrho}$
$1 - h(s, \varrho)$ is the semi-variogram

Table 10.1 Examples of serial and spatial correlation structures

Correlation structure	Function $h(\ldots)$	Comment
Serial	(Auto)correlation function	
corCompSymm[a]	$h(k, \boldsymbol{\varrho}) \equiv \varrho$	$k = 1, 2, \ldots; \lvert\varrho\rvert < 1$
corAR1	$h(k, \boldsymbol{\varrho}) \equiv \varrho^k$	$k = 0, 1, \ldots; \lvert\varrho\rvert < 1$
corCAR1	$h(s, \boldsymbol{\varrho}) \equiv \varrho^s$	$s \geq 0; \varrho \geq 0$
corSymm	$h(d(j, j'), \boldsymbol{\varrho}) \equiv \varrho_{jj'}$	$j < j'; \lvert\varrho_{jj'}\rvert < 1$
Spatial	Correlation function	
corExp	$h(s, \boldsymbol{\varrho}) \equiv e^{-s/\varrho}$	$s \geq 0; \varrho > 0$
corGaus	$h(s, \boldsymbol{\varrho}) \equiv e^{-(s/\varrho)^2}$	$s \geq 0; \varrho > 0$
corLin	$h(s, \boldsymbol{\varrho}) \equiv (1 - s/\varrho)\mathrm{I}(s < \varrho)$	$s \geq 0; \varrho > 0$
corRatio	$h(s, \boldsymbol{\varrho}) \equiv 1 - (s/\varrho)^2/\{1 + (s/\varrho)^2\}$	$s \geq 0; \varrho > 0$
corSpher	$h(s, \boldsymbol{\varrho}) \equiv [1 - 1.5(s/\varrho) + 0.5(s/\varrho)^3]\mathrm{I}(s < \varrho)$	$s \geq 0; \varrho > 0$

[a]The names of the structures follow the convention used in the **nlme** package

LM 3.0 - Estimation

- Estimation carried out via WLS or likelihood based methods mentioned before.

Why LMMs

- **Linear mixed effects models (LMMs)** are models in which a **random component** appears in the model function.
- Up to LM 3.0, we modeled the possible dependence among observations acting on the structure of the variance-covariance matrix of the errors.
  This enabled us to **represent a wide range of possible dependence between observations in a flexible way**, in particular the one related to the presence of groups/strata.
  => PB: we have to assume exactly the kind of dependence existing among observations.

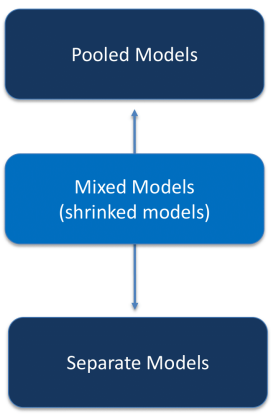
- LMMs enable us to disentangle the effect of the presence of groups from the general kind of dependence occurring among observations.
  => they allow for *partitioning of the global variance* and for *inference on the population of groupers* the actual groups come from.
- More in general, **LMMs represent the most popular and effective approach for the analysis of grouped data**, enabling
  - the *estimation of group level effect*
  - the *modelling of the dependence between observations not deriving by assumptions directly made on the var-cov matrix of the errors, but driven by assumptions on between groups variability*.

LMM – Pooled, separate or to mixed effects estimates?

Back to independent observations averaging over the groups
PB: low number of observations and loosing information

Keep all the observations, Accounting for the presence of groups

Separate model for each group
PB: how to compare/match the estimates



LMM – group-level specification

- For hierarchical data with a single level of grouping (N groups indexed by $i = 1,\dots,N$, with n_i observations per group indexed by $j = 1,\dots, n_i$), we can formulate the classical LMM at a given level of a grouping factor as follows:

Model equation at the level of group of observations

$$\boldsymbol{y}_i = Z_i\boldsymbol{\beta} + W_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i$$

$$\boldsymbol{b}_i \sim N_q(\mathbf{0}, D) \quad \boldsymbol{\varepsilon}_i \sim N_{n_i}(\mathbf{0}, R_i) \quad \text{with} \quad \mathbf{b}_i \perp \boldsymbol{\varepsilon}_i \quad \forall\, i \neq i'$$

$$D = \sigma^2 \mathrm{D} \quad \text{and} \quad R_i = \sigma^2 \mathrm{R}_i \qquad R_i \text{ and } D \text{ positive-definite}$$

$$\boldsymbol{Z_i} = \begin{pmatrix} \boldsymbol{z_{i1}} \\ \dots \\ \boldsymbol{z_{ij}} \\ \dots \\ \boldsymbol{z_{in_i}} \end{pmatrix} = \begin{pmatrix} 1 & z_{i11} & \dots & z_{ip1} \\ \dots & \dots & \dots & \dots \\ 1 & z_{i1n_i} & \dots & z_{ipn_i} \end{pmatrix} = \begin{pmatrix} \mathbf{1}' & \boldsymbol{z'_{i1}} & \dots & \boldsymbol{z'_{ip}} \end{pmatrix} \qquad Z_i \in \mathbb{R}^{n_i \times (p+1)}$$

$$\boldsymbol{b_i} = (b_{i0}, b_{i1}, \dots, b_{iq})' \qquad \boldsymbol{W_i} = \begin{pmatrix} \mathbf{1}' & \boldsymbol{w'_{i1}} & \dots & \boldsymbol{w'_{iq}} \end{pmatrix} \qquad W_i \in \mathbb{R}^{n_i \times (q+1)}$$

- LMMs in their general form are not unique. To make it identifiable we will specify the structure of the matrix R_i in terms of a set of parameters for a variance function and a correlation matrix as in LM 3.0.

LMM – group-level specification

- Generalization to multilevel LMMs is possible, though notationally more complex.
- Let $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N')'$ be the vector of all $n = \sum_{i=1}^{N} n_i$ observed values of the dependent variable.
  Let $\boldsymbol{b} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_N')'$ and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1', \ldots, \boldsymbol{\varepsilon}_N')'$ be the vectors containing all the *N*x*q* random effects and *n* errors, respectively.

Model equation for all data

$$\mathbf{y} = \boldsymbol{Z\beta} + \boldsymbol{Wb} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{b} \sim N_{N(q+1)}(\mathbf{0}, \sigma_b^2 \mathrm{D}) \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma_\varepsilon^2 \mathrm{R}) \quad \text{with} \quad \mathbf{b}_i \perp \boldsymbol{\varepsilon}_i \quad \mathbf{b}_i \perp \mathbf{b}_{i'}$$

where

$$\mathbf{D} = \boldsymbol{I_N} \otimes \boldsymbol{D} = \begin{pmatrix} \mathrm{D} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathrm{D} \end{pmatrix} \text{ and } \mathbf{R} = \begin{pmatrix} \mathrm{R}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathrm{R}_N \end{pmatrix}$$

LMM with *nested structure* for random effects. It is possible to formulate LMMs with non-block-diagonal matrices *W*, D, and *R* (*crossed* random effects).

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z_1} \\ \ldots \\ \boldsymbol{Z_i} \\ \ldots \\ \boldsymbol{Z_N} \end{pmatrix} \in \mathbb{R}^{n \times (p+1)} \qquad \boldsymbol{W_i} = \begin{pmatrix} \boldsymbol{W_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{W_N} \end{pmatrix} \in \mathbb{R}^{n \times N(q+1)}$$

LMM – conditional and unconditional distributions

UNCONDITIONAL DISTRIBUTION of the RANDOM EFFECTS

- The unconditional distribution of the random effects is a **multivariate normal distribution** with zero mean and variance-covariance matrix *D*.

$$D(\sigma^2, \boldsymbol{\theta}_D) = \sigma^2 D(\boldsymbol{\theta}_D)$$

  where $\boldsymbol{\theta}_D$ is a vector of parameters, which represent the (scaled by σ^2) variances and covariances of the elements of **b**.

- Note that the matrix *D* is parameterized using a vector of parameters $\boldsymbol{\theta}_D$.
  - **General case:** any two elements of the vector **b** can be correlated and there are no restrictions imposed on the matrix *D*, except that it is positive-definite and symmetric
    => general structure for *D* with $\boldsymbol{\theta}_D$ containing $q(q+1)/2$ **distinct elements** corresponding to *q* **variances** and $q(q-1)/2$ **covariances** of the random effects included in **b**.
    --> Although *q* is typically small, estimating all the parameters may be difficult if the sample size *n* is limited.
  - **Simplified structure of the matrix *D*:** all elements of the vector $\mathbf{b}_i$ are independent
    => diagonal form for *D* with $\boldsymbol{\theta}_D$ containing *q* **distinct elements** corresponding to *q* **variances**
    Plausibility of the assumption will depend on the data at hand.

LMM – conditional and unconditional distributions

CONDITIONAL DISTRIBUTION of the OBSERVATIONS given the RANDOM EFFECTS

- For the classical LMMs, the conditional distribution $f_{y|b}(\boldsymbol{y}_i|\mathbf{b}_i)$ is multivariate normal, with

$$E[\boldsymbol{y}_i|\mathbf{b}_i] = \boldsymbol{\mu}_i = Z_i\boldsymbol{\beta} + W_i\boldsymbol{b_i} \quad \text{with} \quad \boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})$$
$$V[\boldsymbol{y}_i|\mathbf{b}_i] = V[\boldsymbol{\varepsilon}_i|\mathbf{b}_i] = \sigma^2 \mathrm{R}_i$$

  Thus, conditionally on the (unknown) values of the random effects, the mean value of the dependent-variable is defined by a linear combination of the fixed effects and random effects, respectively.

- In their most general form, LMMs are not identifiable => constraining like in LM 3.0.

$$V[\varepsilon_{ij}|\mathbf{b}_i] = \sigma^2 \mathrm{R}_{ij} = \sigma^2 \lambda^2(\mu_{ij}, \boldsymbol{\delta}; \boldsymbol{v}_{ij})$$

LMM – conditional and unconditional distributions

MARGINAL DISTRIBUTION of the OBSERVATIONS

- The marginal distribution $f_y(\boldsymbol{y}_i)$ of the observations is obtained by "integrating out" the random effects.

$$f_y(\boldsymbol{y}_i) = \int f_{y,b}(\boldsymbol{y}_i, \boldsymbol{b}_i) d\boldsymbol{b}_i = \int f_{y|b}(\boldsymbol{y}_i|\boldsymbol{b}_i) f_b(\boldsymbol{b}_i) d\boldsymbol{b}_i$$

- Given that $f_{y,b}(\boldsymbol{y}_i, \boldsymbol{b}_i)$ and $f_b(\boldsymbol{b}_i)$ are densities of multivariate normal distributions, the marginal distribution of **y** is also multivariate normal and it can be derived analytically.

$$E[\boldsymbol{y}_i] = Z_i\boldsymbol{\beta}$$

$$V[\boldsymbol{y}_i] = \sigma^2[W_i D W_i' + R_i] \; = \sigma^2[W_i D(\boldsymbol{\theta}_D) W_i' + R_i(\boldsymbol{\theta}_R, v_i)]$$

- The **marginal mean value** is defined by a linear combination of the vectors of covariates as for the LMs.
- The **marginal variance-covariance matrix** consists of **two components**.
  - The first one is contributed by the random effects.
  - The second one is related to the residuals.

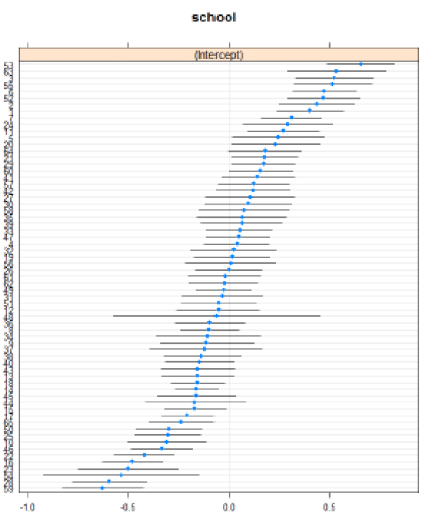
=> *We are partitioning the global variance!*

LMM – Estimation

- **Numerical methods** are always needed in the case of LMMs
- Again, different approaches are possible, leading to possibly different estimates:
  - ML ---> produce variance estimates biased downwards to some degree
  - REML – Restricted (Residual) ML ---> produce unbiased estimates marginalizing wrt nuisance parameters
- Model fitting has 3 distinctive components
  - Estimates of **fixed effects**
  - Estimates of **random effects**
  - Estimates of **variance parameters**
- **WARNING**: variance components are nonegative by definition.
  Nevertheless, methods for estimating them may underestimate variance values, and when real values are close to 0 this may result in unrealistic negative estimates. This may happen when:
  - the # of random effects (q) is small
  - the # of obs per random effect is small

LMM – Inference on parameters

- **Profiled confidence intervals** for the **fixed effects** and for the **standard deviations of random effects and residuals** are provided.
- **Points and intervals estimates for the random effects** can be visualized using a **dotplot**
  --> shows the point and interval estimates for the random effects, ordering them and highlighting which are significantly different from 0.
- To get the **final estimate for the outcome** we add to the contribution of each covariate in the linear predictor the estimate of the random effect the statistical unit belongs to
  --> It may increase or decrease the estimated value for the outcome.
- For each group, we may assess if the corresponding effect is ***positive or negative***, or defining a threshold for labelling them as ***outlier***



LMM – a simple guide to their use

- Compute the **amount of variability the presence of the grouper accounts for**
  => PVRE (Proportion of Variance due to Random Effect)

$$PVRE = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_\varepsilon^2}$$ Random intercept case

- Use the estimate of σ_b^2 for performing **scenario analysis**
  => what if belonging to a $-2\hat{\sigma}_b^2$ or a $+2\hat{\sigma}_b^2$ group, given the same unit conditions?

- We may **cluster the random effects estimates** and use the **medoids** of the groups as representative elements for **quantifying the effect of the «communities of groupers»**.

- We may use the point **estimates of N random effects as a new response** to be modeled through group-specific covariates **OR** we may **introduce group-level covariates** within the same model
  => $(Z \neq W)$

LMMs – Example Multicenter Randomized Trial

- The trial was **randomised double blind comparison of 3 treatments for hypertension** and has been reported by Hall et al. (1991). **Diastolic Blood Pressure (DBP) after treatment** was the primary endpoint.
- One treatment was a **new drug (A)** and the **other two (B and C)** were standard drugs for controlling hypertension (A=Carvedilol, B=Nifedipine, C=Atenodol).
- **29 centres** partecipated in the trial and patients were randomised in order of entry.
- 2 pre treatments and 4 post treatment visits were made as follows:
  - Visit 1 (week 0): Measurements were made to determine wheter pts met eligibility criteria for the trial. Pts who did so received a placebo treatment for 1 week, after which they returned for a second visit
  - Visit 2 (week 1): Measurements were repeated and pts who still satisfied the eligibility criteria were entered into the study and randomized to receive one of the 3 treatments
  - Visit 3-6 (weeks 3,5,7,9): Measurements were repeated at 4 post-treatment visits.
- **311 pts** were assessed to entry into the study. Of these, **288** were suitable and randomised to receiving one of the 3 treatments. **30 pts dropped out** of the study prior to Visit 6.
- Measurements of cardiac function, laboratory values, and adverse events were recordered at each visit.

LMMs – Example Multicenter Randomized Trial

- Focus on a toy-example subsample **N = 3 hospitals**
  n = 9 observations
  3 covariates => p+1 = 4
  Random intercept only => q+1 = 1
- Fit a Linear Mixed Model for dependent (grouped) observations with homoscedastic residuals

$$\mathbf{y} = (176, 194, 156, 150, 150, 160, 150, 160, 160)'.$$

Centre	Treatment	Pre-treatment systolic BP	Post-treatment systolic BP
1	A	178	176
1	A	168	194
1	B	196	156
1	B	170	150
2	A	165	150
2	B	190	160
3	A	175	150
3	A	180	160
3	B	175	160

=> $Z = \begin{pmatrix} 1 & 178 & 1 & 0 \\ 1 & 168 & 1 & 0 \\ 1 & 196 & 0 & 1 \\ 1 & 170 & 0 & 1 \\ 1 & 165 & 1 & 0 \\ 1 & 190 & 0 & 1 \\ 1 & 175 & 1 & 0 \\ 1 & 180 & 1 & 0 \\ 1 & 175 & 0 & 1 \end{pmatrix}$,

EX1: Random intercept only – homoscedastic residuals

(1)

EX: Multicentre Randomized Trial

$N = 3$ hospitals
$m = 9$ observations (group 1 = 4, group 2 = 2, group 3 = 3)

$p+1 = 4$ covariates
$q+1 = 1$ RANDOM INTERCEPT ONLY

> Group Level Formulation

$$\underline{y}_i = Z_i \underline{\beta}^T + W_i b_{i0} + \underline{\varepsilon}_i \longrightarrow \underline{\varepsilon}_i \sim \mathcal{N}_{m_i}(\underline{0}, \sigma^2_\varepsilon R_i)$$

$$b_{i0} \sim \mathcal{N}_1(0, \sigma^2_b)$$

$$\underline{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$$

$$Z_i \in \mathbb{R}^{m_i \times (p+1)} = m_i \times 4 = \begin{cases} i=1 & 4\times 4 \\ i=2 & 2\times 4 \\ i=3 & 3\times 4 \end{cases}$$

$$W_i \in \mathbb{R}^{m_i \times (q+1)} = m_i \times 1$$

$$E[\underline{y}_i] = Z_i \underline{\beta}^T$$

$$V[\underline{y}_i] = W_i \sigma^2_b W_i^T + \sigma^2 R_i$$

$i=1$

$$= \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \sigma^2_b [1111] + \sigma^2_\varepsilon \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

→ Homoschedasticity (but could generalize)

$$= \begin{bmatrix} \sigma^2_b & \sigma^2_b & \sigma^2_b & \sigma^2_b \\ \sigma^2_b & \sigma^2_b & \sigma^2_b & \sigma^2_b \\ \sigma^2_b & \sigma^2_b & \sigma^2_b & \sigma^2_b \\ \sigma^2_b & \sigma^2_b & \sigma^2_b & \sigma^2_b \end{bmatrix} + \begin{bmatrix} \sigma^2_\varepsilon & 0 & 0 & 0 \\ 0 & \sigma^2_\varepsilon & 0 & 0 \\ 0 & 0 & \sigma^2_\varepsilon & 0 \\ 0 & 0 & 0 & \sigma^2_\varepsilon \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2_b + \sigma^2_\varepsilon & \sigma^2_b & \sigma^2_b & \sigma^2_b \\ \sigma^2_b & \sigma^2_b + \sigma^2_\varepsilon & \sigma^2_b & \sigma^2_b \\ \sigma^2_b & \sigma^2_b & \sigma^2_b + \sigma^2_\varepsilon & \sigma^2_b \\ \sigma^2_b & \sigma^2_b & \sigma^2_b & \sigma^2_b + \sigma^2_\varepsilon \end{bmatrix} \in \mathbb{R}^{m_i \times m_i} \quad (*)$$

> All Data Formulation

$$\underline{y} = Z \underline{\beta}^T + W \underline{b}^T + \underline{\varepsilon} \longrightarrow \underline{\varepsilon} \sim \mathcal{N}_9(\underline{0}, \sigma^2 R) \quad R \in \mathbb{R}^{m \times m = 9 \times 9}$$

$$\underline{b} \sim \mathcal{N}_N(\underline{0}, D) \quad D \in \mathbb{R}^{N \times N = 3 \times 3}$$

$$\underline{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$$

$$Z \in \mathbb{R}^{m \times (p+1) = 9 \times 4}$$

$$W \in \mathbb{R}^{m \times N(q+1) = 9 \times 3(1)}$$

$$D = \begin{bmatrix} \sigma^2_b & 0 & 0 \\ 0 & \sigma^2_b & 0 \\ 0 & 0 & \sigma^2_b \end{bmatrix}$$

$$Z = \begin{bmatrix} 1 & z_{11} & z_{21} & z_{31} \\ 1 & & & \\ \vdots & & & \vdots \\ 1 & & & \\ 1 & z_{19} & z_{29} & z_{39} \end{bmatrix}$$

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad \underline{b}_0 = [b_{01}, b_{02}, b_{03}]$$

$$\Rightarrow E[\underline{y}] = Z \underline{\beta}^T$$

$$V[\underline{y}] = W D W^T + \sigma^2_\varepsilon R$$

$$= \begin{bmatrix} (*) & 0 \\ 0 & \end{bmatrix}$$

- observations between groups are uncorrelated (= independent in $\mathcal{N}$ setting)
- observations within groups correlated with dependence induced by random effects

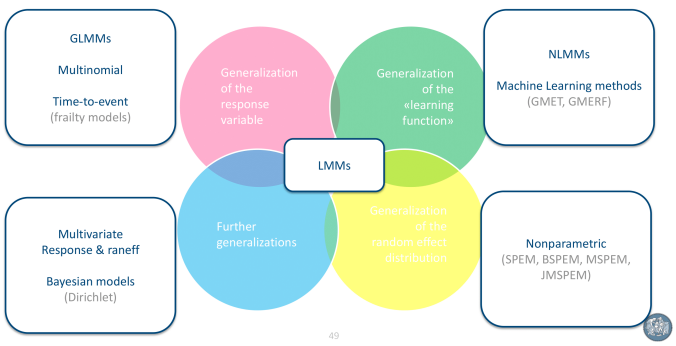
This structure arise from the assumptions on fixed and random effects carried out in the LMM formulation

We might design something similar also working on variance and correlation function of LM 3.0, but then we loose the opportunity of making inference on groups and disentangle the sources.

Take home messages

- **Mind the variance**! Modeling variability is what makes LMs still a powerful tool in the statistical learning landscape.
- The model employing random effects implies a marginal normal distribution which is similar to distributions considered in the context of LMs 3.0, but with the variance-covariance matrix of ***y*** of a very specific parametric form.
- **LMs 3.0 (fixed effects and correlated residual errors) are less restrictive than LMMs.**
  => LMs 3.0 are more flexible than LMMs, but they do not allow making inference about the variability that may be related to different levels of the data hierarchy.
- **Neglecting the correlation structure among the observations leads to a big loss** in the information carried by the data.
- LMMs allow to **extract information at all the levels of the hierarchy and to disentangle the source of variation in the response each level of hierarchy is responsible for** (identifying the important ones!).

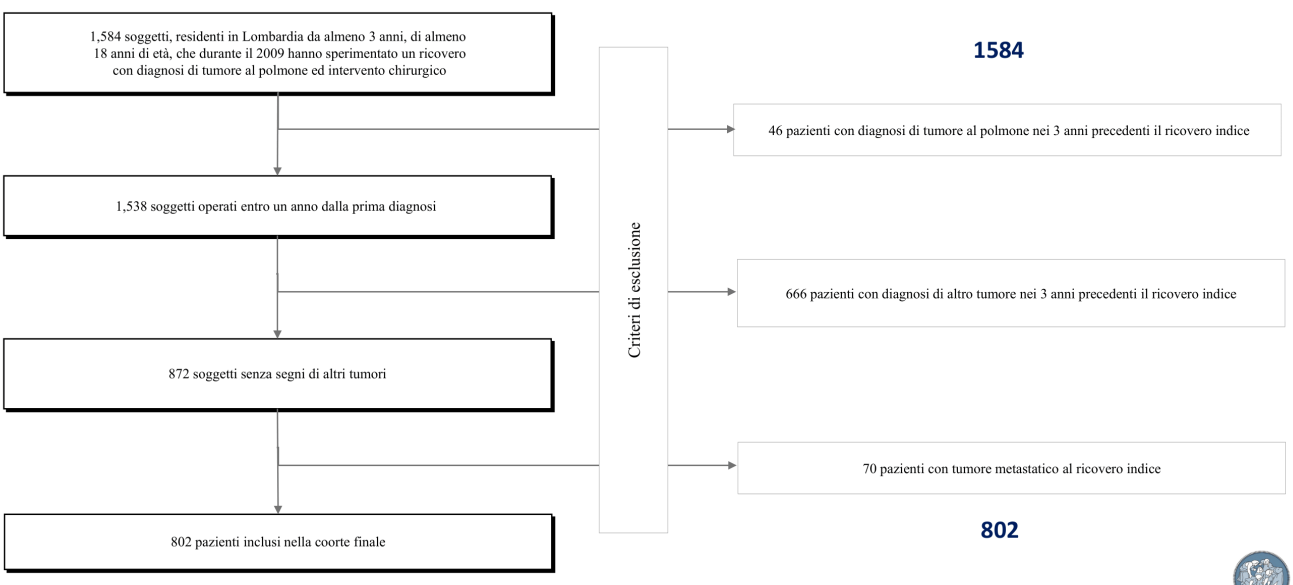
Beyond LMMs



Case study I – Scenario analysis on hospital effect in a multi center trial

- Observational multicenter study (**N=48 hospitals**) on **lung cancer**.
- Investigation of **clinical factors associated to 3y death for any cause** in pts undergoing surgery for lung cancer.
- Goals:
  - **Risk stratification** of the population under study
    (association between mortality and features measured at baseline/entrance)
  - **Assessment of the grouper effect**
  - **Scenario analysis**

Case study I – Cohort selection



Case study I – Hospital volume

48 hospitals, different volume and outcomes

min = 1 -- max = 109

Volume	Numero di ospedali	Numero di pz
<10	27	87
10 – 30	11	206
31 – 50	7	289
>50	3	220

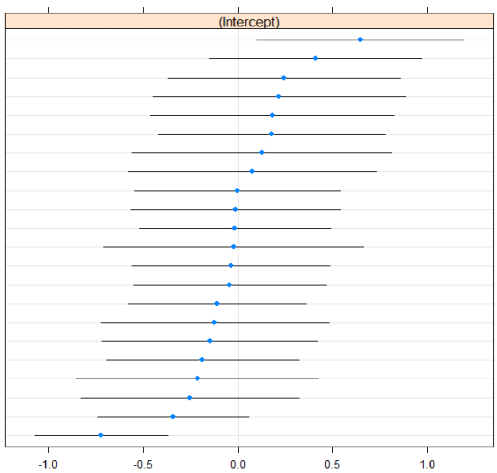
Case study I – GLMM with single level of grouping

$$logit(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 z_{1ij} + \cdots + \beta_p z_{pij} + b_i$$

- b_i random effect $\sim N(0, \sigma_b^2)$
- b_i takes the same value for each observation within the same group and different values in different groups.
- It can be interpreted as ***the effect of being hospitalized in a given structure***.
- Given the patient-specific features, hospital have a significant effect on the log odds of survival probability

% of total variability accounted for by the grouper

VPC = 4.6%



Hospital estimates in the population at hand.

Case study I – GLMM with single level of grouping

$-2\hat{\sigma}$

0 (mean)

$+2\hat{\sigma}$

Man 65 years MCS = 1 NO adjuvant Chemio Low complexity surgery	**0.73**	**0.86**	**0.93**
Man 65 years MCS = 1 NO adjuvantn Chemio High complexity surgery	**0.45**	**0.64**	**0.80**

Case study I – Discussion

- The 3y overall mortality after surgery for lung cancer depends on individual characteristics AND on the hospital the patients are admitted to (protocols? Physicians ability?)

- Each time you have data with grouped structure, using LMMs allows for
  - proper modeling of the **hierarchical structure**
  - better management of **missing data** and **unbalancing between groups**
  - **variance partitioning**
  - **inference** on grouper(s) population => **scenario analysis**

Case study II

Develop an **EM algorithm** for **semi-parametric mixed-effects models** for hierarchical data and apply it to **INVALSI data** as an **unsupervised clustering** tool for Italian schools.

Research questions:

- Are there **differences across the effects of Italian schools** on their student achievements?
- Is it possible to identify **groups of schools** that behave in similar/different ways?
- Do these groups of schools differ in terms of their **characteristics**?

Case study II

Mixed-effects linear model (two-level):

$$\mathbf{y}_i = \beta X_i + \mathbf{b}_i Z_i + \boldsymbol{\epsilon}_i \qquad for \quad i = 1, \ldots, N$$

where
$\mathbf{y}_i$ is the n_i-dimensional vector of **response variable** in the i-th group;
X_i and Z_i are the $(n_i \times (p+1))$ and $(n_i \times (q+1))$ matrices of **fixed and random covariates** respectively;
β is the (p+1)-dimensional vector of **fixed coefficients**;
$\mathbf{b}_i$ is the (q+1)-dimensional vector of **random coefficients** in the i-th group;
$\boldsymbol{\epsilon}_i$ is the (n_i)-dimensional vector of **errors**, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

- Parametric framework $\rightarrow \mathbf{b} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_b)$
- Non-parametric framework $\rightarrow \mathbf{b} \sim$ discrete distribution P^*

Case study II

P^* can be interpreted as the **mixing distribution** that generates the density of the stochastic model in (★).

The ML estimator $\hat{P}^*$ of the random effects distribution P^* can be expressed as a

- **set of points** $(\mathbf{c}_1, \ldots, \mathbf{c}_M)$, where $M \leq N$ and $\mathbf{c}_m \in \mathbb{R}^2$ for $m = 1, \ldots, M$,
- and a **set of weights** $(w_1, \ldots, w_M)$, where $\sum_{m=1}^{M} w_m = 1$ and $w_m \geq 0$ for each $m = 1, \ldots, M$, that represents the proportion of groups belonging to each cluster m.

Case study II

Besides the **estimate of M**, the joint estimation of σ^2, β, $(\mathbf{c}_1, \ldots, \mathbf{c}_M)$ and $(w_1, \ldots, w_M)$ in model (★) is performed through the **maximization of the likelihood**, mixture by the discrete distribution of the random effects:

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) = p(\mathbf{y} | \boldsymbol{\beta}, \sigma^2) = \sum_{m=1}^{M} \frac{w_m}{(2\pi\sigma^2)^{\frac{\sum_{i=1}^{N} n_i}{2}}} exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{n_i} (y_{ij} - c_{0m} - \beta x_{ij} - c_{1m} z_{ij}))^2 \right\},$$

with respect to the fixed coefficient β, the error variance σ^2 and the random effects distribution $(\mathbf{c}_m, w_m)$, for $m = 1, \ldots, M$.
→ **EM algorithm** (SPEM algorithm).

Case study II

The algorithm is inspired by the one proposed by *Azzimonti et. al (2013)*, but with the following improvements:

- the **initialization of the support points** has been reviewed;
- the **maximization** step is made in **closed form**;
- the **covariates are group specific** (they can be different in terms of number of observations and assumed values).

The **support reduction** of the discrete distribution stands on 2 criteria:

- two points $\mathbf{c}_m$ and $\mathbf{c}_k$ closer than a fixed threshold D **collapse to a unique point** $\mathbf{c}_{m,k} = \frac{\mathbf{c}_m + \mathbf{c}_k}{2}$ with weight $w_{mk} = w_m + w_k$;
- we remove points with a weight w **lower than a fixed threshold** $\tilde{w}$, parameterising the remaining weights.

Case study II

The update steps of the SPEM algorithm are the following:

- $w_m^{(up)} = p(\mathbf{b} = \mathbf{c}_m) = \frac{1}{N}\sum_{i=1}^{N} W_{im} \qquad \text{for } m = 1, \ldots, M$
- $(\beta^{(up)}, \sigma^{2(up)}) = \arg\max_{\beta,\sigma^2} \sum_{m=1}^{M}\sum_{i=1}^{N} W_{im} \ln p(\mathbf{y}_i|\beta, \sigma^2, \mathbf{c}_m)$
- $\mathbf{c}_m^{(up)} = \arg\max_{\mathbf{c}} \sum_{i=1}^{N} W_{im} \ln p(\mathbf{y}_i|\beta, \sigma^2, \mathbf{c}) \qquad m = 1, \ldots, M$

where

- $W_{im} = \frac{w_m\, p(\mathbf{y}_i|\beta,\sigma^2,\mathbf{c}_m)}{\sum_{k=1}^{M} w_k\, p(\mathbf{y}_i|\beta,\sigma^2,\mathbf{c}_k)} = p(\mathbf{b}_i = \mathbf{c}_m|\mathbf{y}_i, \beta, \sigma^2)$
- $p(\mathbf{y}_i|\beta, \sigma^2, \mathbf{c}_m) = \frac{1}{(2\pi\sigma^2)^{\frac{n_i}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{ij} - \beta x_{ij} - c_{0m} - c_{1m} z_{ij})^2\right\}$

`SPEM algorithm` → `R package`

Case study II – INVALSI data

- **Students**
  - *Maths **INVALSI test scores** at **grade 8** (MATH8)*,
  - *Maths **INVALSI test scores** at **grade 6** (MATH6)*,
  - socio-economical index (**ESCS**).

- **Schools**
  - Information about **school-body composition**(e.g. percentages of females/immigrants, average ESCS, private/public),
  - information about **school principal** (gender, age, education),
  - **managerial practices** of the school.
- **Size**: 6,572 students within 462 schools.

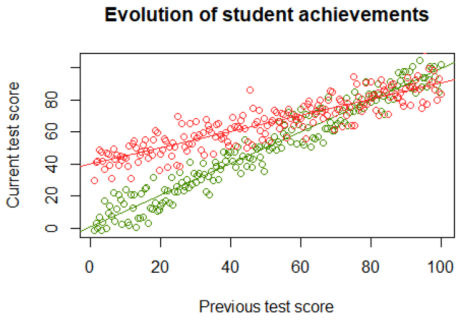
Case study II – INVALSI data

Semi-parametric two-level model for **students** (level 1) nested within **schools** (level 2):

$$\mathbf{y}_i = c_{0m} + c_{1m}\mathbf{z}_i + \beta\mathbf{x}_i + \boldsymbol{\epsilon}_i \qquad i = 1, \ldots, N \qquad m = 1, \ldots, M$$
$$\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2) \qquad \textit{ind.} \qquad (\heartsuit)$$

where
N is the total number of schools,
$\mathbf{y}_i$ is the **maths score at grade 8** of students attending the i-th school,
$\mathbf{z}_i$ is the **maths score at grade 6** of students attending the i-th school,
$\mathbf{x}_i$ is the **socio-economic index** "ESCS" of students attending in the i-th school.

Case study II – INVALSI data



Example of distributions of student achievements within two schools. **Previous score** represents the student score before entering the school, **current score** represents the student score at the end of the period within the school.
How does the relation change across the two schools?

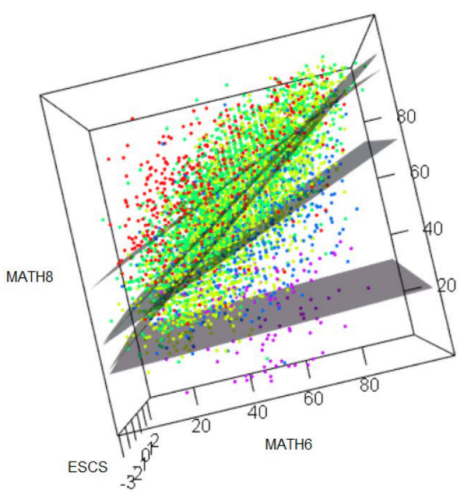
Case study II – INVALSI data

The SPEM algorithm, considering $D = 0.5$ and $\tilde{w} = 0.01$, identifies M=5 distinct clusters of schools:

Cluster	β	c_0	c_1	w
Cluster 1	1.417	46.028	0.454	12.2%
Cluster 2	1.417	22.579	0.707	39.6%
Cluster 3	1.417	30.293	0.648	37.5%
Cluster 4	1.417	31.207	0.393	8.8%
Cluster 5	1.417	25.359	0.027	1.9%

Table: ML estimates of fixed and random coefficients of model (♥) obtained by the SPEM algorithm.

Case study II – INVALSI data



Plot of data with the 5 identified regression planes. Colors represent the 5 clusters.

Case study II – INVALSI data

To explore a posteriori the clusters we apply the following **multinomial logit model**, for each school i, $i = 1, \ldots, N$ and each cluster $m = \{1, 4, 5\}$:

$$ln\left(\frac{P(Y_i = m)}{P(Y_i = C_{ref})}\right) = \beta_{m0} + \sum_{q=1}^{Q} \beta_{mq} v_{iq}.$$

where

- Y_i represents the **cluster of belonging** of school i;
- C_{ref} is the **reference cluster** (union of clusters 2 and 3);
- V is the $N \times q$ matrix of **school level variables**;
- β_m is the vector of coefficients for cluster m.

Case study II – INVALSI data

School level variables	cluster 1	cluster 4	cluster 5
Private School	0.884	−9.187***	−6.147***
Scientific education (yes=1) of the school principal	−0.135	0.171	−6.019***
Central Italy	0.744	0.648	15.691***
Southern Italy	1.201.	1.200.	14.687***

Table: Results of the multinomial logit model. Asterisks denote the significance of the coefficients.

Case study II - discussion

- The SPEM algorithm for hierarchical data can be used as a tool to perform **in-built clustering** in many classification problems
- Contrarily to existing methods, **it does not need to fix a priori the number of mass points** of the random effects discrete distribution
- It represents a **novelty and a value-added** both in the **non-parametric framework** of mixed-effects models and in the **research about school effectiveness**
- When applied to INVALSI data, it identifies **differences across the impacts of Italian schools** on their students achievements

R corner

Useful references to R packages and R functions to implement mixed-effects models

R packages and available codes

R packages employed in the R laboratory session:

1. *nlme*: fixed and mixed-effects regression models with homoscedastic/heteroscedastic and independent/correlated residuals
2. *lme4*: mixed-effects regression models only with homoscedastic and independent residuals (but with a bunch of accessories)
3. *insight*: extract information from a mixed-effects model (e.g., variance decomposition)

R codes relative to **"PoliMi literature"** about mixed-effects models:

1. **GMET()** function: supplementary material of 'Fontana, L., Masci, C., Ieva, F. and Paganoni, A.M. (2021). Performing learning analytics via generalized mixed-effects trees'. Data, 6, 74.'
2. **GMERF()** function: supplementary material of ' Pellagatti M., Masci, C., Ieva, F. and Paganoni, A.M. (2021). Generalized mixed-effects random forest: A flexible approach to predict university student dropout'. Statistical Analysis and Data Mining: The ASA Data Science Journal, 14(3), 241-257.'
3. **SPEM()** function: supplementary material of 'Masci, C., Ieva, F. and Paganoni, A.M. (2018). Semi-parametric mixed-effects models for unsupervised classification of Italian schools'. Journal of the Royal Statistical Society: Series A (Statistics in Society) 182.4, pp.'
4. **MSPEM()** function: supplementary material of 'Masci, C., Ieva, F. and Paganoni, A.M. (2022). Semiparametric multinomial mixed-effects models: a university students profiling tool.' The Annals of Applied Statistics - in press.'

References

References

Journal papers

- **Masci, C., Ieva, F.** and Paganoni, A.M. (2022). Semiparametric multinomial mixed-effects models: a university students profiling tool.' The Annals of Applied Statistics - in press
- Cannistra' M., Masci, C., Ieva, F., Agasisti T. and Paganoni, A.M. (2021). 'Early-predicting dropout of university students: an application of innovative machine learning and multilevel statistical techniques'. Studies in Higher Education, 1-22, DOI:10.1080/03075079.2021.2018415.
- Fontana, L., **Masci, C., Ieva, F.** and Paganoni, A.M. (2021). Performing learning analytics via generalized mixed-effects trees'. Data, 6, 74.
- Masci, C., Ieva, F., Agasisti, T. and Paganoni, A.M. (2021). Evaluating class and school effects on the joint achievements in different subjects: a bivariate semiparametric mixed-effects model'. Computational Statistics, 36, pages 2337–2377.
- Pellagatti M., **Masci, C., Ieva, F.** and Paganoni, A.M. (2021). Generalized mixed-effects random forest: A flexible approach to predict university student dropout'. Statistical Analysis and Data Mining: The ASA Data Science Journal, 14(3), 241-257.
- Gasperoni, F.; **Ieva, F.**; Paganoni, A.M.; Jackson, C.; Sharples, L. (2020) Evaluating the effect of healthcare providers on the clinical path of Heart Failure patients through a novel semi-Markov multi-state model. BMC Health Services Research. 20 (1): 1-11
- Gasperoni, F., **Ieva, F.** Paganoni, A.M., Jackson C., Sharples L.D. (2019) Nonparametric frailty Cox models for hierarchical time-to-event data Biostatistics, 21 (3): 531-544
- **Masci, C., Ieva, F.** and Paganoni, A.M. (2018). Semi-parametric mixed-effects models for unsupervised classification of Italian schools'. Journal of the Royal Statistical Society: Series A (Statistics in Society) 182.4, pp. 1
- **Masci, C.**, Agasisti, T. and Johnes, G. (2018). Student and school performance across countries: A machine learning approach'. European Journal of Operational Research, 269(3), pp. 1072-1085. 313-1342.
- **Masci, C., Ieva, F.**, Agasisti, T. and Paganoni, A.M. (2017). Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements'. Journal of Applied Statistics 44.7, pp. 1296–1317.
- **Masci, C., Ieva, F.**, Agasisti, T. and Paganoni, A. M. (2016). Does class matter more than school? Evidence from a multilevel statistical analysis on Italian junior secondary school students'. Socio-Economic Planning Sciences (54), pp. 47-57.
- Guglielmi, A., **Ieva, F.**, Paganoni, A.M., Ruggeri, F., Soriano, J. (2014). Semiparametric Bayesian modeling for the classification of patients with high observed survival probabilities. Journal of the Royal Statistical Society - Series C, 63 (1): 25-46
- Grieco, N., **Ieva, F.**, Paganoni, A.M. (2012). Performance assessment using mixed effects models: a case study on coronary patient care. IMA Journal of Management Mathematics, 23(2), 117-131
- Guglielmi, A., **Ieva, F.**, Paganoni, A.M., Ruggeri, F. (2012). A Bayesian random-effects model for survival probabilities after Acute Myocardial Infarction. Chilean Journal of Statistics, 3(1): 1-15.
- **Ieva, F.**, Paganoni, A.M. (2010). Multilevel models for clinical registers concerning STEMI patients in a complex urban reality: a statistical analysis of MOMI2 survey. Communications in Applied and Industrial Mathematics, 1(1), 128-147

References

Books

- Gałecki, A., & Burzykowski, T. (2013). Linear mixed-effects model using R. Springer, New York, NY.
- Pinheiro, J., & Bates, D. (2006). Mixed-effects models in S and S-PLUS. Springer science & business media.
- Gelman, A., Hill, J. (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.
- Goldstein, H. (2003) Multilevel Statistical Models. Third edition. London

Codes

- Pinheiro J, Bates D, R Core Team (2022). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-157, https://CRAN.R-project.org/package=nlme.
- Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, **67**(1), 1–48. doi: 10.18637/jss.v067.i01
- Lüdecke D, Waggoner P, Makowski D (2019). "insight: A Unified Interface to Access Information from Model Objects in R." *Journal of Open Source Software*, **4**(38), 1412. doi:10.21105/joss.01412.

LECTURE 26 20/5/2022

SPATIAL STATISTICS

▶ GEOSTATISTICS

- $s_1, \dots, s_m$ spatial locations $\in D$ (fixed)
- $z_{s_1}, \dots, z_{s_m}$ observations collected at those locations
  ↳ correlated
- Aims
  - Model and estimate the spatial dependence
  - Predictions: what happens in s_0?
  - Estimate the mean of z_s

▶ LATTICE DATA  $s_1\ s_2 \dots s_m$ observe everything } Goal: - clustering - modelling the mean

▶ SPATIAL POINT PROCESSES $s_1 \dots s_m$ random locations Goal: study the distr. of $s_1 \dots s_m$

We consider $z_{s_1} \dots z_{s_m}$ scalar

Model $\{z_s, s \in D\}$

$z_s = m_s + \delta_s$

drift $\mathbb{E}[z_s]$ — residual $\mathbb{E}[\delta_s] = 0$

STATIONARITY

▶ STRONG STATIONARITY $(z_{s_1} \dots z_{s_m}) \sim (z_{s_1+h} \dots z_{s_m+h})$ $\forall s_1 \dots s_m \in D$ $\forall h \in \mathbb{R}^d$ $\forall m \geq 1$

▶ 2nd ORDER STATIONARIETY

1) $\mathbb{E}[z_s] = m$ $\forall s \in D$
2) $\text{Cov}(z_{s_1}, z_{s_2}) = C(s_1 - s_2)$

diff. $\to$ (vectors!)

⇐ if the field is gaussian

COVARIOGRAM (COV. FUNCTION)

↳ Properties:

(i) Positive semidefinite $\sum_i \sum_j \lambda_i \lambda_j C(s_i - s_j) \geq 0$ $\forall \lambda_i, \lambda_j \in \mathbb{R}$ $\forall s_1 \dots s_m \in D$

(ii) Symmetric $C(-h) = C(h)$

(iii) Bounded $|C(h)| \leq \underbrace{C(0)}_{\text{variance}}$

VARIOGRAM $2\gamma(s_1 - s_2) = \text{Var}(z_{s_1} - z_{s_2}) = \mathbb{E}[(z_{s_1} - z_{s_2})^2] - (\underbrace{\mathbb{E}[z_{s_1}] - \mathbb{E}[z_{s_2}]}_{=0})^2$

Relation between $2\gamma(\cdot)$ and $C(\cdot)$

$2\gamma(h) = \text{Var}(z_{s_1}) + \text{Var}(z_{s_2}) - 2\,\text{cov}(z_{s_1}, z_{s_2})$

$\overset{!}{=} 2C(0) - 2C(h)$, $h = s_1 - s_2$

Properties

(i) conditional negative def. $\sum_i \sum_j \lambda_i \lambda_j \gamma(s_i - s_j) \leq 0$ $\forall s_i, s_j \in D$ $\forall \lambda_i \in \mathbb{R}$ $\sum_i \lambda_i = 0$

(ii) Sym $\gamma(-h) = \gamma(h)$ $\forall h \in \mathbb{R}^d$

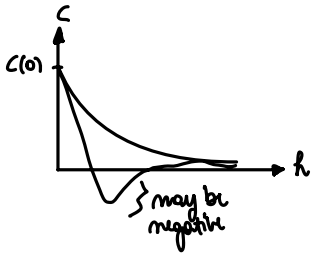
(iii) Non negative $\gamma(h) \geq 0$ $\forall h \in \mathbb{R}^d$

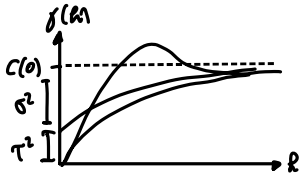
(iv) Zero at origin $\gamma(0) = 0$

(v) sub quadratic growth $\lim_{\|h\| \to \infty} \frac{2\gamma(h)}{\|h\|^2} = 0$

Isotropy: second order stationarity + directional homogeneity

- $\mathbb{E}[z_s] = m$
- $\text{Cov}(z_{s_1}, z_{s_2}) = C(\|h\|) \Rightarrow 2\gamma(\|h\|)$

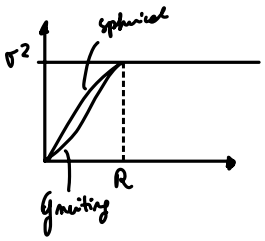




$\tau^2 = \lim_{h \to 0^+} \gamma(h)$ NUGGET

$\sigma^2 + \tau^2 = \lim_{h \to \infty} R(h)$ SILL

R: $\gamma(h) = \sigma^2 + \tau^2 \quad h \geq R$



Valid model: γ s.t. all variogram prop. are satisfied

LECTURE 27 23/5/2022

$C(h) = \underbrace{C(0)}_{\lim_{\|h\| \to \infty} \gamma(\vec{h})} - \gamma(h) \qquad \gamma \rightleftarrows C$

Estimates for γ are more robust.

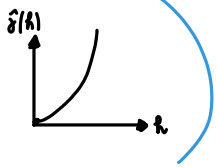
Estimation of the variogram γ

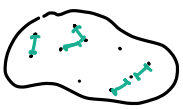
? $2\gamma(h) = \text{Var}(z_{s_1} - z_{s_2}) \quad \|s_1 - s_2\| = h$

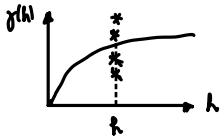
$= \mathbb{E}[(z_{s_1} - z_{s_2})^2] - (\mathbb{E}[z_{s_1}] - \mathbb{E}[z_{s_2}])^2$ → 0 under stationarity

If we're NOT under stationarity, $2\tilde{\gamma} = \mathbb{E}[(z_{s_1} - z_{s_2})^2]$ will estimate $2\gamma(h) + (m_{s_1} - m_{s_2})^2$

Diagnostic! Useful to understand if the process is NOT stationary





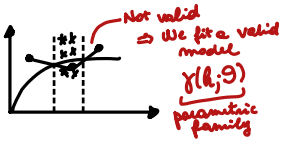


$$2\tilde{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(i,j) \in N(h)} (z_{s_i} - z_{s_j})^2$$

$$N(h) = \{(i,j) : \|s_i - s_j\| \approx h\}$$

↳ Exact h is too strict

Binned Empirical estimator



Spatial predictions (kriging)

$s_1, \dots, s_m \in D$ locations

$z_{s_1}, \dots, z_{s_m}$ observations

z_{s_0}?

In general find $f(z_{s_1} \dots z_{s_m})$ measurable that minimizes $\mathbb{E}[(z_{s_0} - f(\underline{z}))^2]$

⇒ solution $\mathbb{E}[z_{s_0} | \underline{z}]$

properties: 1) Unbiased $\mathbb{E}[\,\cdot\,] = \mathbb{E}[z_{s_0}]$

2) Orthogonal $\mathbb{E}[z_{s_0} | \underline{z}] \perp \underbrace{z_{s_0} - \mathbb{E}[z_{s_0} | \underline{z}]}_{\text{residuals}}$

3) Data interpolation if $s_0 \in \{s_1 \dots s_m\}$, $s_0 = s_i \Rightarrow \mathbb{E}[z_{s_0} | \underline{z}] = z_{s_i}$

USA school: assume it and go on

4) If $\{z_s, s \in D\}$ is Gaussian $\mathbb{E}[z_{s_0} | \underline{z}] = \lambda_0^* + \sum_{i=1}^{m} \lambda_i^* z_{s_i}$

→ We won't look at the BEST function if we don't have Gaussianity, we'll restrict ourselves to a small family

Best Linear Unbiased Predictor

→ Kriging: $z_{s_0} = \lambda_0 + \sum_i \lambda_i z_{s_i}$ ⇒ We look for $\lambda_0, \dots, \lambda_m$ the min $\mathbb{E}[(z_{s_0} - \hat{z}_{s_0})^2]$ s.t. $\mathbb{E}[\underbrace{z_{s_0}^*}_{\text{BLUP}}] = \mathbb{E}[z_{s_0}]$

3 types:

1) Simple Kr : hp $\mathbb{E}[Z_s]$ known everywhere (unless in practice)
2) Ordinary Kr : hp 2nd order stationarity but unknown mean
3) Universal Kr : relax 2nd order stationarity

] $c(h)$ known

▸ Ordinary Kr

Assumptions:

2nd order stationarity & isotropy

$m = \mathbb{E}[Z_s]$

$C(h) = Cov(Z_{s_1}, Z_{s_2}) \quad \|s_1 - s_2\| = h$

Problem:

find $\lambda_0 \dots \lambda_m \in \mathbb{R}$ s.t. $\min \mathbb{E}[(Z_{s_0} - (\lambda_0 + \Sigma_i \lambda_i Z_{s_i}))^2]$

s.t. $\mathbb{E}[(\lambda_0 + \Sigma_i \lambda_i Z_{s_i})] = m \quad \Sigma_i \lambda_i = 1$

Not a convex combination because we don't have $0 \le \lambda_i \le 1$

unb.
$\Rightarrow \lambda_0 + \Sigma_i \lambda_i \mathbb{E}[Z_{s_i}] = m$

$\lambda_0 + \Sigma_i \lambda_i m = m$

I impose uniform unbiasedness : $\begin{cases} \lambda_0 = 0 \\ \Sigma_i \lambda_i = 1 \end{cases}$

Objective functional $\Phi(\underline{\lambda}, \xi) = \mathbb{E}[(Z_{s_0} - \Sigma_i \lambda_i Z_{s_i})^2] + 2\xi(\Sigma_i \lambda_i - 1)$

$= Var(Z_{s_0} - \Sigma_i \lambda_i Z_{s_i}) + 2\xi(\Sigma_i \lambda_i - 1)$

$= \underbrace{Var(Z_{s_0})}_{C(0)} + \Sigma_i \Sigma_j \lambda_i \lambda_j C(\|s_i - s_j\|) - 2\Sigma_i \lambda_i C(\|s_0 - s_i\|) + 2\xi(\Sigma_i \lambda_i - 1)$

$$\begin{cases} \frac{\partial \Phi}{\partial \lambda_i} = 0 = \cancel{2}\Sigma_j \lambda_j C(\|s_i - s_j\|) - \cancel{2} C(\|s_0 - s_i\|) + \cancel{2}\xi & i = 1, \dots, m \\ \frac{\partial \Phi}{\partial \xi} = 0 = 2(\Sigma_i \lambda_i - 1) \end{cases}$$

$(\Sigma)_{ij} = C(\|s_i - s_j\|)$ cov. matrix

$$\rightarrow \boxed{\begin{cases} \Sigma_j \lambda_j C(\|s_i - s_j\|) + \xi = C(\|s_0 - s_i\|) & \text{ok system} \\ \Sigma_i \lambda_i = 1 \end{cases}} \Leftrightarrow \begin{pmatrix} \Sigma & \underline{1} \\ \underline{1}' & 0 \end{pmatrix} \begin{pmatrix} \underline{\lambda} \\ \xi \end{pmatrix} = \begin{pmatrix} \underline{\gamma}_0 \\ 1 \end{pmatrix}$$

$(\underline{\gamma}_0)_i = C(\|s_0 - s_i\|)$ vector of covariances

▸ Universal Kr

Assumptions

$Z_s = m_s + \delta_s \quad s \in D$

$\mathbb{E}[\delta_s] = 0$

$Cov(\delta_{s_1}, \delta_{s_2}) = c(h) \quad h = \|s_1 - s_2\|$

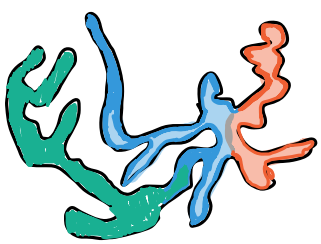
$m_s = \mathbb{E}[Z_s]$ drift

$= \Sigma_{\ell=0}^{L} a_\ell f_\ell(s)$

a_ℓ unknown coeffs. — f_ℓ regressors known everywhere in D

Ex

$(m_s = a_0 + a_1 x + a_2 y \quad (x, y) = s)$



Problem

Find $\lambda_0 \dots \lambda_m \in \mathbb{R}$ that solve $\min \mathbb{E}[(Z_{s_0} - \lambda_0 - \Sigma_i \lambda_i Z_{s_i})^2]$ s.t. $\mathbb{E}[(\lambda_0 + \Sigma_i \lambda_i Z_{s_i})] = m_{s_0} = \sum_{\ell=0}^{L} a_\ell f_\ell(s_0)$

$\lambda_0 - \Sigma_i \lambda_i \mathbb{E}[Z_{s_i}] = \sum_{\ell=0}^{L} a_\ell f_\ell(s_0)$

$\lambda_0 - \Sigma_i \lambda_i \Sigma_\ell a_\ell f_\ell(s_i) = \Sigma_\ell a_\ell f_\ell(s_0)$

$\lambda_0 - \Sigma_\ell a_\ell \Sigma_i \lambda_i f_\ell(s_i) = \Sigma_\ell a_\ell f_\ell(s_0) \rightarrow \begin{cases} \lambda_0 = 0 \\ \Sigma_i \lambda_i f_\ell(s_i) = f_\ell(s_0) \end{cases} \quad \ell = 0 \dots L$

$$\Rightarrow \dots \Rightarrow \begin{pmatrix} \Sigma & F \\ F' & \underline{0} \end{pmatrix} \begin{pmatrix} \underline{\lambda} \\ \underline{\xi} \end{pmatrix} = \begin{pmatrix} \underline{f}_0 \\ \underline{t}_0 \end{pmatrix}$$

LECTURE 28 24/5/2022

To estimate Σ we can estimate $2\gamma(\cdot)$ from the residuals $\delta_{s_1} \dots \delta_{s_n}$

1) Empirical estimate

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} (\delta_{s_i} - \delta_{s_j})^2$$

2) Fit a valid model $\gamma(h;\underline{\theta})$, get $\hat{\underline{\theta}}$

However the δ's are unknown! We first have to estimate them

$$z_{s_i} = \sum_l a_l f_l(s_i) + \delta_{s_i} \quad s_i \in D$$

$$\underline{z} = F\underline{a} + \underline{\delta} \quad Cov(\underline{\delta}) = \Sigma$$

$$\Rightarrow \underset{\underline{a}\in\mathbb{R}^{L+1}}{\operatorname{argmin}} (\underline{z} - F\underline{a})' \Sigma^{-1} (\underline{z} - F\underline{a}) = \hat{\underline{a}}^{GLS} = \dots = (F'\Sigma^{-1}F)^{-1} F'\Sigma^{-1}\underline{z}$$

$\hat{\underline{a}}^{GLS}$ ← BLUE for $\underline{a}$

We need to know Σ

$$\hat{\underline{m}}^{GLS} = F\hat{\underline{a}}^{GLS}$$

$$\hat{\underline{\delta}} = \underline{z} - F\hat{\underline{a}}^{GLS}$$

$\hat{\Sigma}$

Start with $\Sigma = \sigma^2 \mathbb{I}$

LECTURE 29 26/5/2022

- Informally, **functional data** are entities that can be described through a function, e.g., a curve, a surface, an image
- A **functional dataset** consists of a sample of functional observations
- Even though observations are actually discrete and affected by noise, the observed values reflect a **smooth variation of the phenomenon**. One might be interested not only in **point-wise** values, but also in **differential properties** of the data

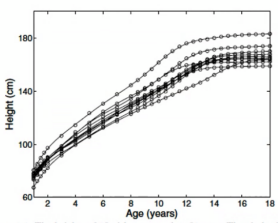


Figure 1.1. The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.

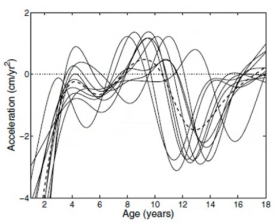


Figure 1.2. The estimated accelerations of height for 10 girls, measured in centimeters per year. The heavy dashed line is the cross-sectional mean, and is a rather poor summary of the curves.

Ramsay Silverman 2005 Springer

Books:

- **Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, Springer, 2nd ed.**
- Ramsay, J.O. and Silverman, B.W. (2002). *Applied Functional Data Analysis*, Springer.
- Ramsay, J.O., Hooker, G. and Graves, S. (2009). *Functional Data Analysis with R and Matlab*, Springer.
- Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*, Springer.
- Horvath, L. and Kokoszka P. (2012). Inference for Functional Data with Applications, Springer.
- Kokoszka P. and Reimherr, M. (2017). Introduction to Functional Data Analysis. Chapman & Hall

Introductory paper:

- Sørensen, H., Goldsmith, J., Sangalli, L.M. (2013), "An introduction with medical applications to functional data analysis". Statistics in Medicine, 32, pp. 5222–5240.

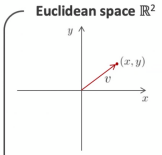
Software: (available from CRAN)

- R package fda (corresponding Matlab code available from http://www.psych.mcgill.ca/misc/fda/)
- R package Refund
- Matlab code PACE
- R package mgcv
- R package fdakma (alignment and clustering)
- R package fdaPDE (functional data over complex multidimensional domains)

The notion of **Hilbert space** generalizes the concept of Euclidean space to spaces of any (even infinite) dimension

- Vectorial structure (linear combinations)
- Distance, angles, projections (measure of dependence, best approximations)

Euclidean space $\mathbb{R}^2$



- Sum: $v_1 + v_2 = (x_1 + x_2, y_1 + y_2)$
- Product by a constant: $c \cdot v = (c \cdot x, c \cdot y)$

Operations (+, ·)

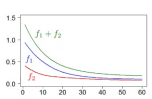
- Norm (lenght of a vector): $\|v\| = (x^2 + y^2)^{1/2}$
- Distance: $\|v_1 - v_2\|^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2$
- Angle: $\vartheta = \arccos \frac{\langle v_1, v_2 \rangle}{\|v_1\| \|v_2\|}$

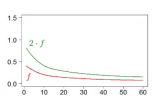
Inner product $\langle v_1, v_2 \rangle = (x_1 \cdot x_2) + (y_1 \cdot y_2)$

L^2: space of real-valued square-integrable functions

- Sum: $(f_1 + f_2)(t) = f_1(t) + f_2(t)$
- Product by a constant: $(c \cdot f)(t) = c \cdot f(t)$

Operations (+, ·)





- Norm: $\|f\|^2 = \int (f(t))^2 dt$
- Distance: $\|f_1 - f_2\|^2 = \int (f_1(t) - f_2(t))^2 dt$
- Angle: $\vartheta = \arccos \frac{\langle f_1, f_2 \rangle}{\|f_1\| \|f_2\|}$

Inner product $\langle f_1, f_2 \rangle = \int (f_1(t) \cdot f_2(t)) dt$

More precisely, L^2 is a quotient space with respect to the equivalence relation: $x \sim y \quad \int [x(t) - y(t)]^2 dt = 0$

Let *H* be a linear space. An inner product on *H* is a bilinear, symmetric, positive definite form

$$\langle \cdot, \cdot \rangle : H \times H \to \mathbb{R}$$

that satisfies

(i) $\langle \lambda x + y, z \rangle = \lambda \langle x, z \rangle + \langle y, z \rangle \quad \forall \lambda \in \mathbb{R}, \ \forall x, y, z \in H$

(ii) $\langle x, y \rangle = \langle y, x \rangle \quad \forall x, y \in H$

(iii) $\langle x, x \rangle \geq 0 \quad \forall x \in H$

(iv) $\langle x, x \rangle = 0 \iff x = 0$

In particular:

- The inner product allows to measure lengths and angles
- It allows to define orthogonality: two vectors in *H* are orthogonal if $\langle x, y \rangle = 0$
- The inner product induces a norm and a metric
- The inner product allows generalizing the Pythagoras' Theorem:

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 \text{ if and only if } \langle x, y \rangle = 0$$

A (real) Hilbert space *H* is an inner product space that is complete, in the norm induced by the inner product.

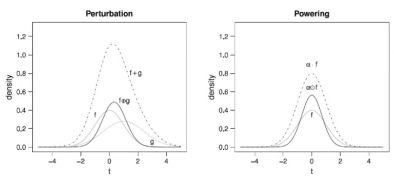
- A Hilbert space is complete in the sense that it contains all the limit points of its Cauchy sequences;
- A Hilbert space is separable if it contains a dense countable subset

- Useful properties:
  - In a Hilbert space one has the notion of orthogonal projection and of best approximations
  - A Hilbert space *H* is separable iff it has an orthonormal basis $\{u_n\}_{n \in \mathbb{N}}$
  - If *H* is separable Hilbert space, $\{u_n\}_{n \in \mathbb{N}}$ is an orthonormal basis and $x \in H$ then

$$x = \sum_{n=1}^{\infty} \langle x, u_n \rangle u_n. \qquad \textbf{Basis expansion}$$

B^2: space of density functions on a closed interval *I*, with log in L^2

- Equivalence relation: f, g are equivalent if they are proportional (*scale invariance*)
- Sum (perturbation): $(f \oplus g)(t) = \dfrac{f(t)g(t)}{\int_I f(s)g(s)\,ds}$
- Product by a constant (powering): $(\alpha \odot f)(t) = \dfrac{f(t)^\alpha}{\int_I f(s)^\alpha\,ds}, \quad t \in I.$
- Inner product: $\langle f, g \rangle_B = \dfrac{1}{2\eta} \int_I \int_I \ln \dfrac{f(t)}{f(s)} \ln \dfrac{g(t)}{g(s)}\,dt\,ds$
- Norm: $\|f\|_B = \left[\dfrac{1}{2\eta} \int_I \int_I \ln^2 \dfrac{f(t)}{f(s)}\,dt\,ds \right]^{1/2}$



Note: the geometry of L^2 doesn't make sense for density functions

- Let *H* be a Hilbert space, whose points are functions defined on a closed interval $T = [t_{min}, t_{max}$ (e.g., range of time during which the data are collected)
- Hereafter, we will always consider functional data in Hilbert spaces

Definition 1:
A **functional random variable** is a random element defined on a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ with values in H
$$X : \Omega \to H$$

Definition 2:
A **functional datum** x is a realization of a functional random variable, i.e., for $\omega \in \Omega$,
$$x = X(\omega) : T = [t_{min}, t_{max}] \to \mathbb{R}$$

Definition 3:
A **functional dataset** is a collection of functional data.

Let $X : \Omega \to H$ be a functional random variable in *H*.

We assume $\mathbb{E}[\|X\|_H^4] < \infty$

Definition 4:
We call Fréchet mean of X the (unique) element μ of *H* that solves
$$\operatorname*{arginf}_{x \in H} \mathbb{E}[\|X - x\|_H^2].$$

- If $H=L^2$ the Fréchet mean coincides a.e. with the point-wise mean
$$\mathbb{E}[X(t)] = \mu(t), \quad t \in T$$
- In any *H*, we can **estimate the mean via the sample estimator**
$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

In $H=L^2$, this is the point-wise sample mean

Let $X : \Omega \to H$ be a **zero-mean** functional random variable in *H*, such that $\mathbb{E}[\|X\|_H^4] < \infty$

Definition 5:
We call covariance operator of X the operator from *H* to *H* defined as

$$Cx = \mathbb{E}[\langle X, x\rangle X], \quad x \in H$$

- If $H=L^2$ the covariance operator can be equivalently defined through a kernel operator

$$[Cx](t) = \int_T c(s,t)x(s)d(s)\,, \quad x \in L^2$$

where the covariance kernel is precisely the point-wise covariance

$$c(s,t) = \mathbb{E}[X(s)X(t)]$$

- In $H=\mathbb{R}^p$, the covariance operator coincides with the linear operator defined by the covariance matrix

Let $X : \Omega \to H$ be a **zero-mean** functional random variable in *H*, such that $\mathbb{E}[\|X\|_H^4] < \infty$

Definition 5:
We call covariance operator of X the operator from *H* to *H* defined as

$$Cx = \mathbb{E}[\langle X, x\rangle X], \quad x \in H$$

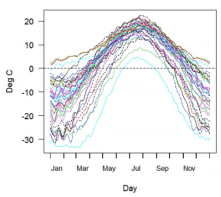
- In any *H*, the covariance operator can be estimated through the sample covariance operator

$$Sx = \frac{1}{N}\sum_{i=1}^{N}\langle X_i, x\rangle X_i, \quad x \in H$$

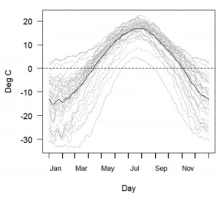
- If $H=L^2$, one can use the alternative definition

$$[Sx](t) = \int_T \widehat{c}(s,t)x(s)d(s)\,, \quad x \in L^2 \qquad \widehat{c}(s,t) = \frac{1}{N}\sum_{i=1}^{N} X(s)X(t)$$

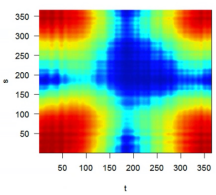
Dataset of Temperatures in Canada (35 observations)



Functional dataset



Sample mean



Sample covariance kernel

$$z_i = f(s_i) + \epsilon_i \qquad i = 1,\ldots,n \qquad \epsilon_1,\ldots,\epsilon_n \text{ i.i.d. } E[\epsilon_i] = 0,\ \mathrm{Var}[\epsilon_i] = \sigma^2$$

Basis expansion

$\psi_1, \ldots, \psi_K$: K basis functions

$$\hat{f}(s) = \sum_{k=1}^{K}\hat{c}_k\psi_k(s) \qquad \to \text{Find } \hat{c}_k, k = 1,\ldots,K \text{ (i.e., find } \hat{f}) \text{ by minimizing}$$

Curve fitting

$$\underline{z} = \Psi \underline{c} + \underline{\epsilon}$$

$$\mathrm{SSE} = \sum_{j=1}^{n}(z_i - f(s_i))^2 = \sum_{j=1}^{n}\left(z_i - \sum_{k=1}^{K} c_k\psi_k(s_i)\right)^2$$

$$\Psi = \begin{bmatrix} \psi_1(s_1) & \psi_2(s_1) & \cdots & \psi_K(s_1) \\ \psi_1(s_2) & \psi_2(s_2) & \cdots & \psi_K(s_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(s_n) & \psi_2(s_n) & \cdots & \psi_K(s_n) \end{bmatrix}$$

$$\boldsymbol{z} = (z_1,\ldots,z_n)^t \qquad \boldsymbol{f} = (f(s_1),\ldots,f(s_n))^t \qquad \boldsymbol{c} = (c_1,\ldots,c_K)^t \qquad \boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)^t$$

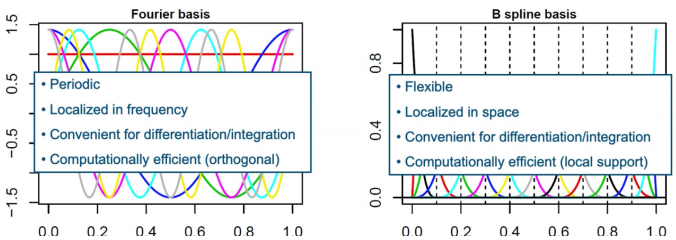
$$\mathrm{SSE} = (\boldsymbol{z} - \Psi\boldsymbol{c})^t(\boldsymbol{z} - \Psi\boldsymbol{c})$$

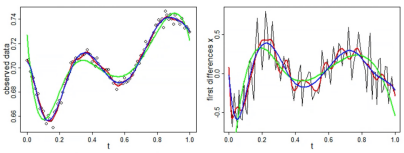
$$\hat{\boldsymbol{c}} = \left(\Psi^t\Psi\right)^{-1}\Psi^t\boldsymbol{z}$$

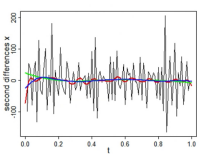
$$\hat{\boldsymbol{z}} = \hat{\boldsymbol{f}} = \Psi\hat{\boldsymbol{c}} = \Psi\left(\Psi^t\Psi\right)^{-1}\Psi^t\boldsymbol{z} = S\boldsymbol{z}$$

$$df = K = tr(S) = tr(S^tS) = tr(2S - S^tS)$$

piecewise polynomial







Choose the number K of basis

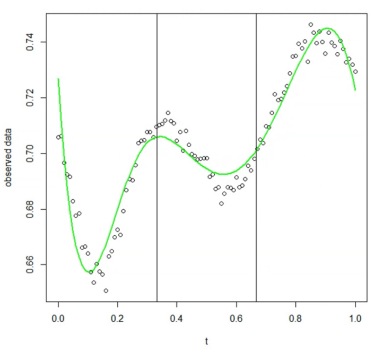
K << N

K = 7

K = 12

K = 30

1) Use a functional space with only few dimensions (few basis) K << n



K = 7
(spline of order 5)

Distributional properties of the estimator:

$$\boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)^t$$

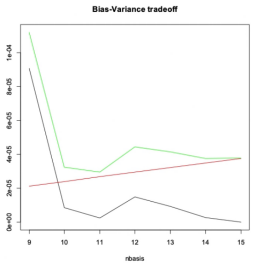
$$\hat{f}(s) = \boldsymbol{\psi}(s)^t \hat{\boldsymbol{c}} = \boldsymbol{\psi}(s)^t (\Psi^t\Psi)^{-1}\Psi^t \boldsymbol{z}$$

$$E[\hat{f}(s)] = \boldsymbol{\psi}(s)^t (\Psi^t\Psi)^{-1}\Psi^t \boldsymbol{f}$$

$$\text{Bias}[\hat{f}(s)] = f(s) - E[\hat{f}(s)]$$

$$\text{Var}[\hat{f}(s)] = E[\{\hat{f}(s) - E[\hat{f}(s)]\}^2] = \sigma^2 \boldsymbol{\psi}(s)^t (\Psi^t\Psi)^{-1}\boldsymbol{\psi}(s)$$

$$\text{MSE}[\hat{f}(s)] = E[\{\hat{f}(s) - f(s)\}^2] = \text{Bias}^2[\hat{f}(s)] + \text{Var}[\hat{f}(s)]$$



▶ Cross-validation

Distributional properties of the estimator:

$$\boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)^t$$

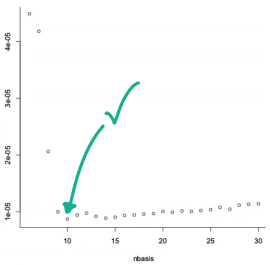
$$\hat{f}(s) = \boldsymbol{\psi}(s)^t \hat{\boldsymbol{c}} = \boldsymbol{\psi}(s)^t (\Psi^t\Psi)^{-1}\Psi^t \boldsymbol{z}$$

$$E[\hat{f}(s)] = \boldsymbol{\psi}(s)^t (\Psi^t\Psi)^{-1}\Psi^t \boldsymbol{f}$$

$$\text{Bias}[\hat{f}(s)] = f(s) - E[\hat{f}(s)]$$

$$\text{Var}[\hat{f}(s)] = E[\{\hat{f}(s) - E[\hat{f}(s)]\}^2] = \sigma^2 \boldsymbol{\psi}(s)^t (\Psi^t\Psi)^{-1}\boldsymbol{\psi}(s)$$

$$\text{MSE}[\hat{f}(s)] = E[\{\hat{f}(s) - f(s)\}^2] = \text{Bias}^2[\hat{f}(s)] + \text{Var}[\hat{f}(s)]$$



Selection of the number of basis K minimizing

$$GCV(K) = \frac{1}{n(1 - tr(S)/n)^2}(\boldsymbol{z} - \hat{\boldsymbol{z}})^t(\boldsymbol{z} - \hat{\boldsymbol{z}})$$
$$= \left(\frac{n}{n-K}\right)\left(\frac{1}{n-K}\right)(\boldsymbol{z} - \hat{\boldsymbol{z}})^t(\boldsymbol{z} - \hat{\boldsymbol{z}})$$

$$\hat{f}'(s) = \sum_{k=1}^{K} \hat{c}_k \psi_k'(s) \qquad \hat{f}''(s) = \sum_{k=1}^{K} \hat{c}_k \psi_k''(s)$$

Smoothing requires special care when the curve estimate is asked, not only to provide a good smoothing of the data, but also to reflect the features of the curve that are represented by its derivatives

Curve derivatives (or their functions) are very often
- objects of analysis
- helpful for further processing and analysis of the data (curve alignment/clustering)

2) Use a rich functional space but with regularization

K ~ n

$$\text{SSE}_\lambda = \text{SSE} + \lambda \int (f''(s))^2 \, ds$$

penalize curvature (potentially a different idea)

$$\{R_\psi\}_{(k,l)} \quad (k,l)\text{-entry} : \int \psi_k''(s)\psi_l''(s)\, ds$$

$$\text{SSE}_\lambda = \text{SSE} + \lambda \boldsymbol{c}^t R_\psi \boldsymbol{c}$$

$$\hat{\boldsymbol{c}}_\lambda = \left(\Psi^t\Psi + \lambda R_\psi\right)^{-1}\Psi^t \boldsymbol{z}$$

$$\hat{\boldsymbol{z}} = \hat{\boldsymbol{f}} = \Psi\left(\Psi^t\Psi + \lambda R_\psi\right)^{-1}\Psi^t \boldsymbol{z} = S\boldsymbol{z}$$

Sub-projection operator

$$df = tr(S) < K \qquad (\text{or } df = tr(S^tS) \text{ or } df = tr(2S - S^tS))$$

Distributional properties of the estimator:

$$\hat{f}(s) = \boldsymbol{\psi}(s)^t \hat{\boldsymbol{c}} = \boldsymbol{\psi}(s)^t \left(\Psi^t\Psi + \lambda R_\psi\right)^{-1}\Psi^t \boldsymbol{z}$$

$$E[\hat{f}(s)] = \boldsymbol{\psi}(s)^t \left(\Psi^t\Psi + \lambda R_\psi\right)^{-1}\Psi^t \boldsymbol{f}$$

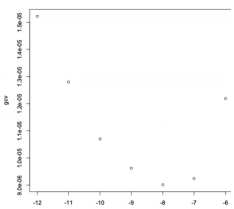
$$\text{Bias}[\hat{f}(s)] = f(s) - E[\hat{f}(s)]$$

$$\text{Var}[\hat{f}(s)] = E[\{\hat{f}(s) - E[\hat{f}(s)]\}^2] =$$
$$= \sigma^2 \boldsymbol{\psi}(s)^t \left(\Psi^t\Psi + \lambda R_\psi\right)^{-1}\left(\Psi^t\Psi\right)\left(\Psi^t\Psi + \lambda R_\psi\right)^{-1}\boldsymbol{\psi}(s)$$

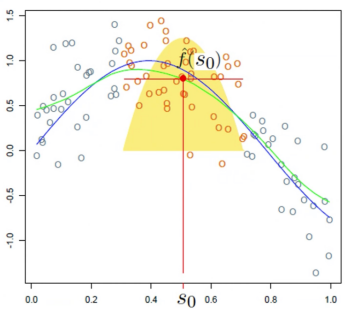
$$\text{MSE}[\hat{f}(s)] = E[\{\hat{f}(s) - f(s)\}^2] = \text{Bias}^2[\hat{f}(s)] + \text{Var}[\hat{f}(s)]$$

Selection of the smoothing parameter λ minimizing

$$GCV(\lambda) = \frac{1}{n(1 - tr(S)/n)^2}(\boldsymbol{z} - \hat{\boldsymbol{z}})^t(\boldsymbol{z} - \hat{\boldsymbol{z}})$$
$$= \left(\frac{n}{n - tr(S)}\right)\left(\frac{1}{n - tr(S)}\right)(\boldsymbol{z} - \hat{\boldsymbol{z}})^t(\boldsymbol{z} - \hat{\boldsymbol{z}})$$



Local polynomial regression (kernel smoother)



Picture taken from Friedman, Tibshirani and Hastie (2013) The Elements of Statistical Learning, Springer

At each abscissa s_0, find $(c_0, \ldots, c_L)$ that minimize

$$\sum_{i=1}^{n} \text{Kern}_h(s_0, s_i)[(z_i - \sum_{l=0}^{L} c_l (s_0 - s_i)^l)]^2 \qquad \text{where } \text{Kern}_h(s_0, s_i) = D\left(\frac{|s_0 - s_i|}{h}\right)$$

Positive functions:

$f(s) = e^{W(s)}$ where $W(s) = \sum_k c_k \psi_k(s)$

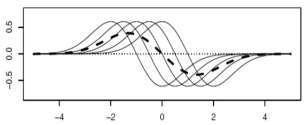
Estimate f by minimizing

$$\sum_{i=1}^{n} \left(z_i - e^{W(s_i)}\right)^2 + \lambda \int W''(s) ds$$

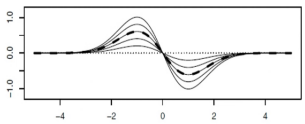
Increasing functions:

$$f(s) = C + \int_{s_0}^{s} exp\{W(t)\} dt \quad \text{where } W(s) = \sum_k c_k \psi_k(s)$$

Phase variability



Amplitude variability



Phase variability: different curves exhibit more or less the same features but that these features occur at different times or space locations for different statistical units.
If not taken properly into account, the misalignment acts as a confounding factor and may blur subsequent analyses.

Registration of a set of functions

Find suitable warping functions $h_1(t), \ldots, h_N(t)$ such that $c_1(h_1(t)), \ldots, c_N(h_N(t))$ are the most similar.

The functions h_i should be increasing; they capture the phase variability. Amplitude variability is the remaining variability in vertical direction among the aligned curves.

In some cases, time or location is merely shifted from curve to curve, for example, because the measurements are started at random time points. For these situations, it is natural to use $h_i(t)=t+delta_i$. In other situations, phase variation is a matter of dilation, in which case $h_i(t)=alpha_i t$ is a natural choice of warping function. In yet other situations, the time or space deformation is more complex.

Landmarks: significant (univocally identifiable) shape-events in a curve, e.g. crossings of zero, peaks, valleys, points of inflection.

$c_1, \ldots, c_N$, where $c_i : [0, T] \to \mathbb{R}^d$

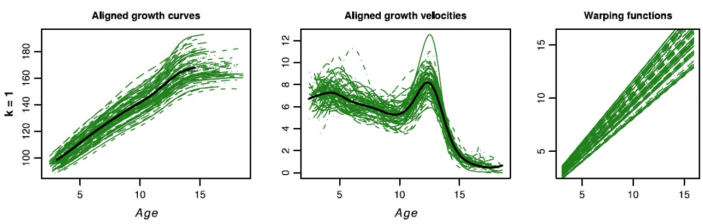
Suppose

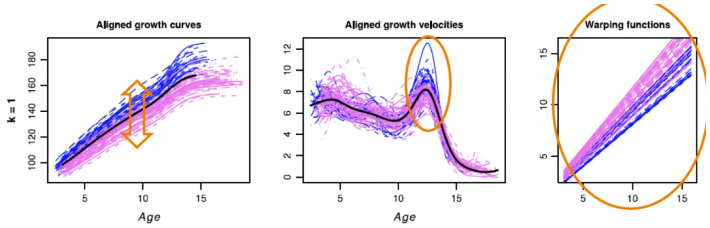
- L landmarks; for the $i-$th curve, located at $t_{i1}, \ldots, t_{iL}$
- a template curve c_0 is available with landmark locations $t_{01}, \ldots, t_{0L}$
  If not, we can define t_{0j} as the average of the t_{ij}'s

Warping function for the $i-th$ curve: any strictly increasing function h_i s.t.

- $h_i(0) = 0$
- $h_i(t_{0j}) = t_{ij}$, for $j = 1, \ldots, L$
- $h_i(T) = T$

▶ $(0,0), (t_{01}, t_{i1}), \ldots, (t_{0L}, t_{iL}), (T,T)$: interpolated by a piece-wise line, a polygon or higher order monotone splines (strictly increasing)





Once the biological clocks are aligned

the height of boys stochastically dominates the one of girls for any registered biological age

boys have a more pronounced growth during puberty (more prominent growth velocity peak)

Neat separation of boys and girls in the phase.

The biological clocks of boys and girls run at different speeds

Goal of **Alignment**:
Decoupling Phase and Amplitude Variability

+

Goal of ***K*-mean** Clustering:
Decoupling Within and Between-cluster (Amplitude) Variability

=

Goal of ***K*-mean Alignment**:
Identifying Phase Variability, Within-cluster Amplitude Variability and Between-cluster Amplitude Variability

LECTURE 30 30/5/2022

p: Principal Component Analysis

Courtesy of P. Secchi

Problem: Given a dataset of *N* zero-mean multivariate observations in $\mathbb{R}^p, X_1, \ldots, X_N$ find the orthonormal directions $\mathbf{a_1}$, ..., $\mathbf{a_p}$ of maximum variability (for the dataset).

Equivalently, for *k= 1,...,p*, find:

$$\begin{aligned} \mathbf{a}_k &= argmax_{\mathbf{a}\in\mathbb{R}^p} Var(\mathbf{a}'\mathbf{X}) \\ \text{subject to:} \quad & \mathbf{a}'\mathbf{a} = 1,\ \mathbf{a}_j'\mathbf{a} = 0 \text{ for } j < k \end{aligned}$$

- We can re-write the problem as

$$\begin{aligned} \mathbf{a}_k &= argmax_{\mathbf{a}\in\mathbb{R}^p} \frac{1}{N}\sum_{i=1}^{N}(\mathbf{a}'\mathbf{X}_i)^2 \\ \text{subject to:} \quad & \mathbf{a}'\mathbf{a} = 1,\ \mathbf{a}_j'\mathbf{a} = 0 \text{ for } j < k \end{aligned}$$

or, equivalently

$$\begin{aligned} \mathbf{a}_k &= argmax_{\mathbf{a}\in\mathbb{R}^p} \frac{1}{N}\sum_{i=1}^{N} < \mathbf{a}, \mathbf{X}_i >^2 \\ \text{subject to:} \quad & ||\mathbf{a}|| = 1,\ < \mathbf{a}_j, \mathbf{a} >= 0 \text{ for } j < k \end{aligned}$$

Note 1. We assume *N>p* and absence of collinearity, i.e. the data matrix is full rank.

Note 2 . If $X_1, \ldots, X_N$ are not zero-mean, they can be centered by subtracting the (sample) mean. For unbisasedness, divide by *N-1* instead of *N*.

Problem: Given a dataset of *N* zero-mean multivariate observations in $\mathbb{R}^p, X_1, \ldots, X_N$ find the orthonormal directions $\mathbf{a_1}$, ..., $\mathbf{a_p}$ of maximum variability, i.e., those solving for k=1,...p,

$$\begin{aligned} \mathbf{a}_k &= argmax_{\mathbf{a}\in\mathbb{R}^p} \frac{1}{N}\sum_{i=1}^{N} < \mathbf{a}, \mathbf{X}_i >^2 \\ \text{subject to:} \quad & ||\mathbf{a}|| = 1,\ < \mathbf{a}_j, \mathbf{a} >= 0 \text{ for } j < k \end{aligned}$$

Solution: Call *S* the sample covariance matrix of $X_1, \ldots, X_N$ Then, the **principal components** are found as the eigenvectors of the matrix *S*; for *k=1,..., p*, they solve the eigen-equation

$$Se_k = \lambda_k e_k$$

The eigenvalue λ_k associated with the eigenvector e_k represents the variability along the direction e_k .

Note. We call *score* u_{ik} the projection of the observation X_i along the direction e_k , i.e.,

$$u_{ik} = \langle X_i, e_k \rangle = X_i' e_k$$

Problem: Given a dataset of *N* zero-mean functional observations in *H*, $X_1, \ldots, X_N$, find the directions of maximum variability (in *H*) of the dataset, i.e., for *k=1,...,N*, find ξ_k maximizing

$$\begin{aligned} & \frac{1}{N}\sum_{i=1}^{N} < \xi, \mathbf{X}_i >_H^2 \\ \text{subject to:} \quad & ||\xi|| = 1,\ < \xi_j, \xi >_H = 0 \text{ for } j < k \end{aligned}$$

- We look for an orthonormal system in *H* maximizing the *variability* of the corresponding projections
- Indeed, $\langle \xi, X_i \rangle_H$ is the projection of X_i «along the direction» ξ (i.e., a «direction» in *H*).
  Note that $\langle \xi, X_i \rangle_H$ is a scalar, hence $\frac{1}{N}\sum_{i=1}^{N}\langle \xi, X_i \rangle_H^2$ is a sample variance in the usual sense.

Note 1. If the data are not zero-mean, they can be centered by subtracting the (sample) mean. *N* should then be replaced by *N-1*.

Note 2. If data are linearly independent and centered on the sample mean, we can find at most *N-1* principal components.

Problem: Given a dataset of *N* zero-mean functional observations in *H*, $X_1, \ldots, X_N$, find the directions of maximum variability (in *H*) of the dataset, i.e., for *k=1,...,N*, find ξ_k maximizing

$$\begin{aligned} & \frac{1}{N}\sum_{i=1}^{N} < \xi, \mathbf{X}_i >_H^2 \\ \text{subject to:} \quad & ||\xi|| = 1,\ < \xi_j, \xi >_H = 0 \text{ for } j < k \end{aligned}$$

- We look for an orthonormal system in *H* maximizing the *variability* of the corresponding projections
- Indeed, $\langle \xi, X_i \rangle_H$ is the projection of X_i «along the direction» ξ (i.e., a «direction» in *H*).
  Note that $\langle \xi, X_i \rangle_H$ is a scalar, hence $\frac{1}{N}\sum_{i=1}^{N}\langle \xi, X_i \rangle_H^2$ is a sample variance in the usual sense.

Note 1. If the data are not zero-mean, they can be centered by subtracting the (sample) mean. *N* should then be replaced by *N-1*.

Note 2. If data are linearly independent and centered on the sample mean, we can find at most *N-1* principal components.