

Applied Statistics



**Politecnico di Milano – School of Industrial and
Information Engineering**

**Applied Statistics
052498 - 052742**

Academic Year 2024/2025

1 Credits

8 -10 CFU

2 Teachers and tutors

2.1 Course leader

Prof. Piercesare Secchi
MOX - Dipartimento di Matematica
La Nave, III floor
Politecnico di Milano
e.mail: piercesare.secchi@polimi.it
Webex room: <https://politecnicomilano.webex.com/meet/piercesare.secchi>
Office Hours (by appointment): Friday, 16:00 - 18.00.

2.2 Lab teacher

Ing. Guillaume Koechlin
MOX - Dipartimento di Matematica
La Nave, VI floor
Politecnico di Milano
e.mail: guillaume.koechlin@polimi.it
Webex room: <https://politecnicomilano.webex.com/meet/guillaume.koechlin>
Tutoring activities: to be announced

R
we will use Python instead

3 Timetable

Class & Lab:

MON, Room 3.1.5, 10:15-12:15 (Lab&Lecture),
TUE, Room 3.1.4, 11:15-13:15 (Lecture&Lab),
THU, Room 3.1.12, 11:15-13:15 (Lecture&Lab),
FRI, Room 3.1.3, 10:15-12:15 (Lecture&Lab).

4 Web page

The course web page is here:

<https://webeep.polimi.it/course/view.php?id=16791>

Short-term notices, source code for the lab sessions, test results, open problems, etc., will be regularly posted on the course web page.

5 Course Program

The topics covered by the 8 CFU versions of the course are the following:

1. *Exploring a multivariate dataset.* Descriptive statistics and graphical displays. The geometry of a multivariate sample. Sample mean, covariance and correlation. Generalized variance and total variance. The metric induced by the covariance matrix.
2. *Data representation and dimensional reduction.* The analysis of the covariance structure, principal component analysis (PCA).
3. *Classification: discrimination and clustering.* Statistical classification: model, misclassification costs and prior probability. Bayesian supervised classification and the Fisher approach to discriminant analysis. Cross-validation for the evaluation of a classifier. Alternative approaches to classification: CART, support vector machines. Similarity measures. Unsupervised classification; hierarchical and nonhierarchical methods. DBSCAN. K-means and K-medoids. Multidimensional scaling.
4. *Inference about mean vectors.* The multivariate normal distribution, the Wishart distribution, the F distribution. Hotelling T^2 test. Confidence regions and simultaneous comparisons of component means. The Bonferroni method for multiple comparisons. Familywise Error Rate and False Discovery Rate. Comparisons of several multivariate means. ANOVA and MANOVA. Inference for Linear Models. Beyond Ordinary Least Squares: ridge regression, lasso, regularized least squares. Random effects and mixed effects linear models.

In the 10 CFU version of the course, the above topics are complemented with the following two modules of Advances in Statistical Learning:

5. *Introduction to Functional Data Analysis.* Data smoothing, dimensional reduction and representation. Functional principal component analysis. Data registration: phase and amplitude variability. Classification of functional data.
6. *Statistics for spatial data.* Random fields, variogram models and variogram fitting. Spatial prediction and Kriging, Functional data with spatial dependence.

When data
on point depends
on near points

6 Lab sessions and data analysis project

Methods and algorithms will be illustrated in the lab sessions through applications to real data sets; analyses will be performed in R, an open-source package for statistics downloadable at

www.r-project.org

Students shall actively participate in the lab sessions. **All students** – those taking the course for 8CFU and those taking the course for 10CFU – **must** work in a team on a data analysis project developed along the course: each team shall show the project work in progress during routinely scheduled meetings with all other teams.

6.1 Data analysis project

Every student taking the course **must** participate in a data analysis project developed by an **independently** formed team of **3-5 members**. The work in progress of the projects will be collectively discussed during a general meeting to be held **Tuesday, the 8th of April, 2025**. Final analyses and results will be presented in a workshop, which will take place a day **in June 2025 to be collectively decided**.

Data sets available for the team projects will be presented in class **Thursday, the 27th of February**.

Before March 10, each team should send an email to the students

- (1) Xxx Yyyy (xxx.yyyy@mail.polimi.it),
- (2) Www Zzzz (www.zzzz@mail.polimi.it)

containing the following information:

- (a) name and email address of the team leader;
- (b) the title of the project;
- (c) the list of the team members, their names and personal code;
- (d) max 5 lines of abstract with a short description of the data set analyzed and the temporary goals of the project; these could always be updated and modified while the project is under development.

This information will be made public on the course web page as soon as the students (1) and (2) will organize it in a file to be sent to Ing. Guillaume Koechlin.

or we
can choose
our dataset
but have
to prove
why we
choose this
one

pre meeting
o formal team
o choose dataset
o some first
trials analysis
of data

Evaluate also project of
others,

6.2 Open workshop for project presentations

Teams will show the final results and analyses of their projects during an open workshop that will take place after the end of classes (June 2025).

The workshop will consist of a speedy pitch session (2 minutes per team), during which each team very briefly presents its project. This will be followed by a poster session, during which each team will illustrate an A0 poster reporting the results of their analysis to all participants, students, and teachers.

The projects presented at the final workshop will be collectively evaluated **by the course students participating – in presence – to the entire workshop and by the course teachers.**

The teams students (1) and (2) belong to is in charge of the organization of the work-in-progress meeting and of the final workshop (program schedule of the event, chairing the pitch session, presiding over food and drinks for coffee breaks etc.).

7 Exam

The exam consists of two parts:

(can bring everything on exam)

Open book
exam

- (a) A written exam. The written exam will feature multiple-choice questions, that assess both theoretical knowledge and practical skills, along with several data analysis problems to be individually solved with R; two problems for the students following the 8 CFU version of the course, three problems - with extra time - for those registered in the 10 CFU version. For the students taking the course for 10 CFU, the extra problem will be related to the two topics treated in the modules characterizing their additional 2 CFU; working on this problem is **mandatory** to pass the exam of the 10 CFU version of the course. For all students, the use of a personal computer is allowed, as well as that of books, personal notes, etc. (it's an open-book written exam).

In the written exam the student must show the ability to conduct a stylized data analysis, by selecting the appropriate methods and algorithms - among those introduced in the course - for solving the problems, by running the algorithms with R, by identifying the significant results and by reporting them with the precision and property of language which characterize the technical and scientific communication.

- (b) Team project evaluation. Projects will be collectively evaluated by the teachers of the course and by the students participating – in presence – in a final workshop at the end of the course. **The grade of the project expires after the 5th exam session, that is at the beginning of the second semester of the Academic Year 2025/2026.**

During the presentation of their projects, teams must prove their ability to conduct and report a real life statistical data analysis, showing knowledge and understanding of the problem at hand and the nature of the

data, proper judgment for the selection of the appropriate methods and algorithms, sensible interpretations of the generated results - grasping not only their strengths but also their weaknesses - and, finally, communication skills when informing an audience of peers.

To pass the exam, students must pass **both part (a) and part (b)** with a score greater than or equal to 18/30; their final grade is then obtained as the weighted average of the two scores, with weights respectively equal to 0.6 for the written exam and 0.4 for the project evaluation.

8 Course bibliography

required JOHNSON, R.A. and WICHERN, D.W., (2007). *Applied Multivariate Statistical Analysis (sixth edition)*, Prentice Hall, Upper Saddle River

required JAMES G., WITTEN D., HASTIE T. and TIBSHIRANI R. (2013). *An introduction to statistical learning, with application to R*, Springer, New York (<http://www-bcf.usc.edu/~gareth/ISL/>)

required HASTIE T., TIBSHIRANI R. *Statistical Learning MOOC*, <https://www.edx.org/course/statistical-learning>.

suggested RAMSAY, J.O. e SILVERMAN, B.W., (2005). *Functional Data Analysis (second edition)*, Springer Series in Statistics, Springer, New York

additional GALECKI A. e BURZYKOWSKI T. (2013). *Linear Mixed-Effects Models using*. Springer Texts in Statistics, Springer, New York

additional CRESSIE, N. (1993). *Statistics for Spatial Data (Revised Edition)*, John Wiley & Sons

additional HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: data mining, inference and prediction. (Second Edition)*, Springer-Verlag, New York.

additional RAMSAY, J.O. e SILVERMAN, B.W., (2002). *Applied Functional Data Analysis: methods and case studies*, Springer Series in Statistics, Springer, New York

18. 02. 25 Exploring Dataset and Dataframe

- We start from problem, not Data
Questions that I have attack using provided Data.

- Dataframe - statistical units

 $n \times p$
units

n-number of units

p-number of features,
describing unit

	x_1	x_2	...	x_p	
1	x_{11}	x_{12}	-	x_{1p}	$\leftarrow x_1^T$
2	x_{21}	x_{22}	-	x_{2p}	$\leftarrow x_2^T$
3					
:					
n					

x_{ij}

Var j

observing unit i

can be categorical,
and continuous

mode — most frequency object

example: most frequent color of eyes-green

I can give number for categories, but
we still will not be use math on that

1. Questions
2. Analysis

Row perspective

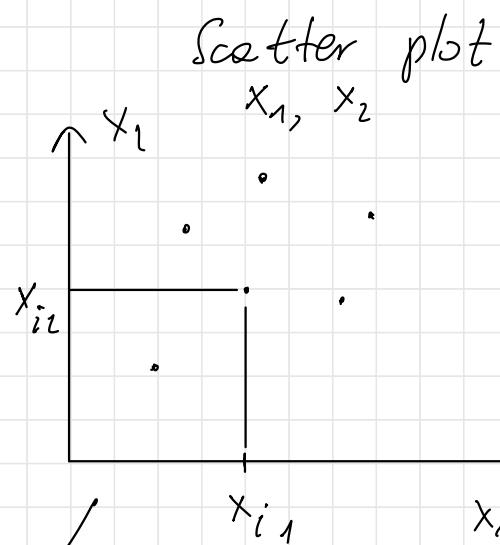
Data: $\underline{x}_i \in \mathbb{R}^P$, $i = 1 \dots n$

$$\underline{x} = \begin{bmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_n \end{bmatrix}$$

Exploring \rightarrow

Basic analysis of

data

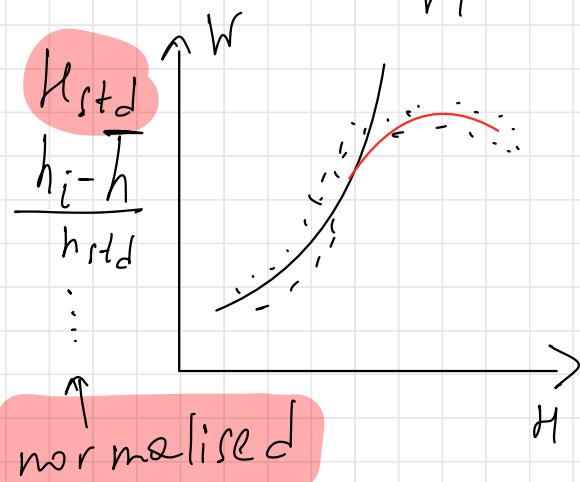


We will plot all combinations

of x_{ij} , $x_{ij} \neq j, i$

Weight Height

w	h_1	h_1^2	h_1^3
w_1	h_1	h_1^2	h_1^3
:	:	:	:
w_n	h_n	h_n^2	h_n^3



normalised

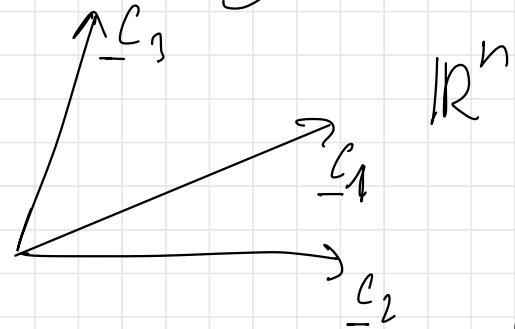
Column perspective

$$\mathbb{X} = [\underline{c}_1 \dots \dots \underline{c}_p]$$

$$\underline{c}_j \in \mathbb{R}^n, j=1 \dots p$$

$$\underline{c}_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \in \mathbb{R}^n$$

$$\underline{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n$$



$$L(\underline{1}) = \text{span}(\underline{1}) = \{ \underline{c} \in \mathbb{R}^n : \beta \cdot \underline{1} = \begin{pmatrix} \beta \\ \vdots \\ \beta \end{pmatrix}, \beta \in \mathbb{R} \}$$

Diagram illustrating the span of $\underline{1}$ as a line through the origin. A horizontal axis is shown with a point d_1 marked. Three vectors $\underline{c}_1, \underline{c}_2, \underline{c}_3$ originate from the origin and lie on this line. A red box highlights the text $\pi \underline{c}_1 | \underline{1} \in L(\underline{1})$.

Can I approximate \underline{c}_1 by my $L(\underline{1})$?

$$\underline{c}_1 - \pi \underline{c}_1 | \underline{1} = d_1 \quad -\text{deviation} \quad \text{of feature one}$$

$$c_1 = \pi_{\underline{L}(1)} \underline{c}_1 + \underline{c}_1^\perp$$

↑ ↗ ↗ ↗
 $\underline{L}(1)$ \perp $\underline{L}^\perp(1)$

$\underline{L}(1)$ dimension: 1 $\underline{L}^\perp(1)$ dimension $n-1$

A short excursion in the geometry
of \mathbb{R}^n $\underline{u}, \underline{v} \in \mathbb{R}^n$

$$\underline{u}, \underline{v} \in \mathbb{R}^n$$

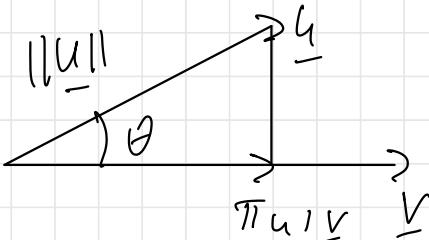
$$\|\underline{u}\| = \sqrt{\underline{u}^T \underline{u}} = \sqrt{\sum_{i=1}^n u_i^2} = \sqrt{\langle \underline{u}, \underline{u} \rangle}$$

$$\underline{u}^T [\dots] \begin{bmatrix} \end{bmatrix}$$

$$\|\underline{v}\|$$

$$\|\underline{u}\|^2 = \langle \underline{u}, \underline{u} \rangle = \underline{u}^T \underline{u}$$

$$\cos \theta = \frac{\langle \underline{u}, \underline{v} \rangle}{\|\underline{u}\| \cdot \|\underline{v}\|} =$$



inner product

$$= \sum_{i=1}^n u_i v_i$$

$$\sqrt{(\sum u_i^2)(\sum v_i^2)}$$

$$\pi_{\underline{u}} \underline{v} = \|\underline{u}\| \cos \theta \cdot \frac{\underline{v}}{\|\underline{v}\|} = \|\underline{u}\| \frac{\langle \underline{u}, \underline{v} \rangle}{\|\underline{u}\| \|\underline{v}\|} \cdot \frac{\underline{v}}{\|\underline{v}\|} =$$

$$= \frac{\langle \underline{u}, \underline{v} \rangle}{\|\underline{v}\|^2} \cdot \underline{v} = \frac{\underline{v}^T \underline{u}}{\underline{v}^T \underline{v}} \cdot \underline{v} = \underbrace{\frac{\underline{v} \underline{v}^T}{\underline{v}^T \underline{v}}} \cdot \underline{u}$$

matrix

projection is

orthogonal projection

idempotentency

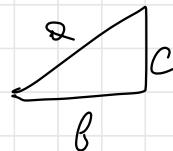
operator

(applying twice, we will get
same)

(symmetric)

(symmetric + idempendancy) \Rightarrow orthogonal

$$\|\underline{u}\|^2 = \|\underline{\pi}_{\underline{u}}\|_V^2 + \|\underline{u} - \underline{\pi}_{\underline{u}}\|_V^2 = \alpha^2 = b^2 + c^2$$



What will be the best approximation of \underline{c}_1
in non variable space

$$\underline{1}^T \underline{1} = n$$

$$\underline{\pi}_{\underline{c}_1, \underline{1}} = \frac{\underline{1} \underline{1}^T}{\underline{1}^T \underline{1}} \underline{c}_1 = \frac{\sum_{i=1}^n x_{i1}}{n} \cdot \underline{1} = \bar{x}_1 \cdot \underline{1}$$

mean of
first column

$$\underline{d}_1 = \underline{c}_1 - \bar{x}_1 \cdot \underline{1} = \begin{pmatrix} x_{11} - \bar{x}_1 \\ x_{21} - \bar{x}_1 \\ \vdots \\ x_{n1} - \bar{x}_1 \end{pmatrix}$$

$$\|\underline{d}_1\|^2 = \underline{d}_1^\top \underline{d}_1 = \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 =$$

$$= (n-1) \underbrace{\frac{1}{n-1} \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}_{\text{sample variance}} = (n-1) s_{11}$$

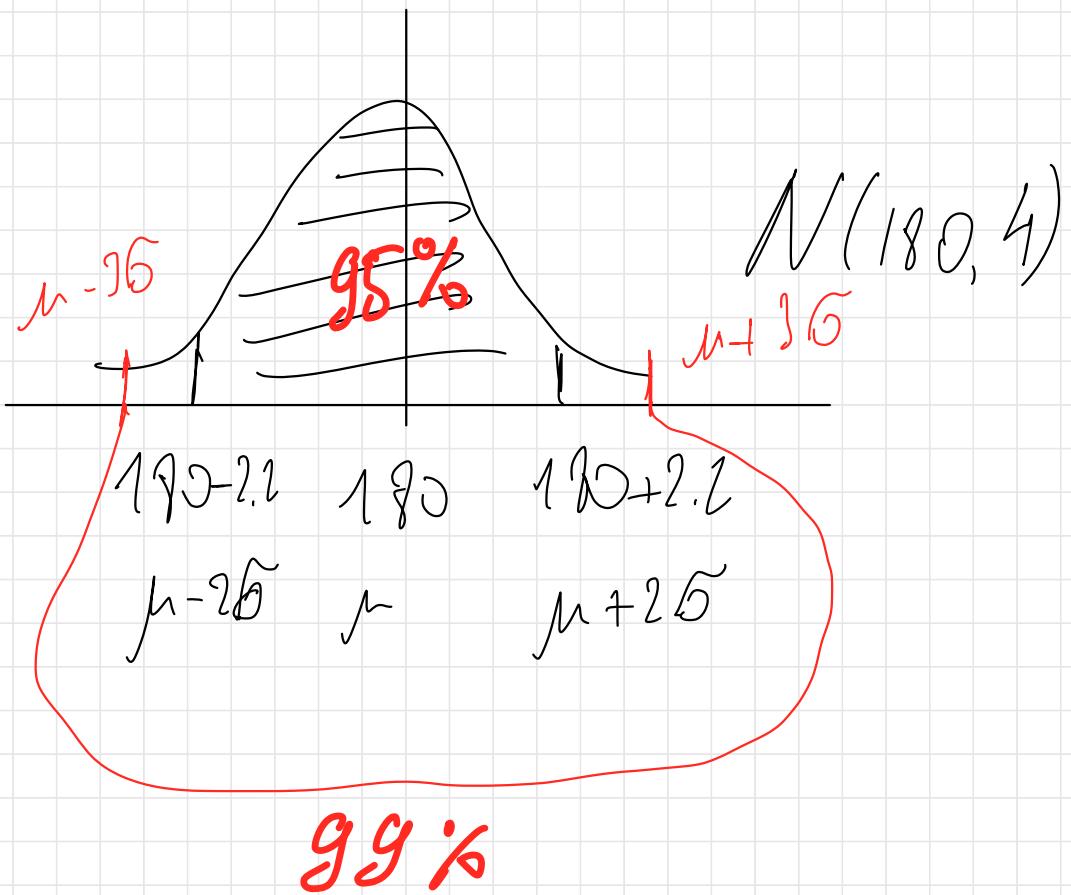
$$s_{11}$$

$$\sqrt{s_{11}} = \text{std dev}$$

X - random variable $\sim \mu, \sigma$ for all distributions

$$P[\mu - k\sigma < X < \mu + k\sigma] \geq 1 - \frac{1}{k^2}$$

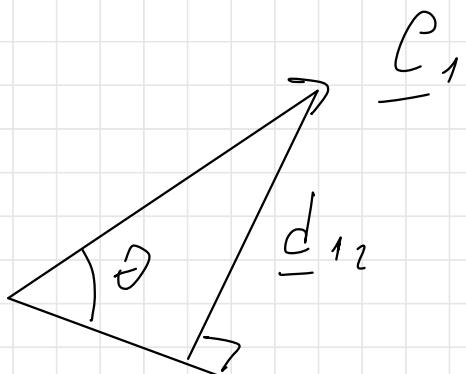
$$\Rightarrow k=2 \Rightarrow P[\mu - 2\sigma < X < \mu + 2\sigma] \geq 1 - \frac{1}{4} = 0.75$$



How many σ -s I am from μ .
 (It's not absolute value kg, meters... ,
 it's deviation)

$$X \rightarrow \frac{X - \mu}{\sigma}$$

flow far we are
 from mean



splitting information
in part

width

weight

rewidth

smth

$$\underline{c}_1 = \pi_{\underline{c}_1 | \underline{c}_L} + \underline{d}_{12}$$

$$\underline{d}_{12} = \underline{c}_1 - \pi_{\underline{c}_1 | \underline{c}_L}$$

$$\pi_{\underline{c}_1 | \underline{c}_L} \rightarrow \underline{c}_2$$

$$\theta = 0 \Rightarrow \xrightarrow{\quad} \xrightarrow{\underline{c}_1} \xrightarrow{\underline{c}_2}$$

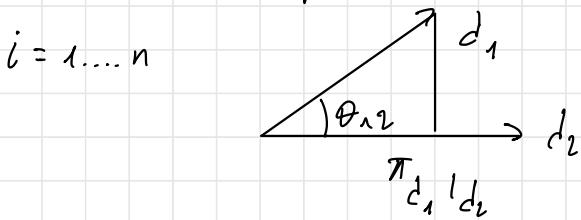
$$\underline{c}_1 = \int \underline{c}_2$$

$$\theta = \frac{\pi}{2} \Rightarrow \underline{c}_1 = ? \underline{c}_2$$

\underline{x}_i

|

$$\mathbb{X} = [\underline{c}_1, \dots, \underline{c}_p] \quad d_i = \underline{c}_i - \pi_{c_i | 1} = \underline{c}_i - \underline{x}_i \cdot 1$$



$$\underline{d}_1 = \pi_{\underline{d}_1 | \underline{d}_2} + (\underline{d}_1 - \pi_{\underline{d}_1 | \underline{d}_2})$$

$$\theta = 0 \Rightarrow \underline{d}_1 = \beta \underline{d}_2$$

$$\underline{c}_1 - \underline{x}_1 \cdot 1 = \beta (\underline{c}_2 - \underline{x}_2 \cdot 1)$$

$$\underline{c}_1 = (\underline{x}_1 - \beta \underline{x}_2) \cdot 1 + \beta \underline{c}_2$$

$$\cos \theta_{12} = \frac{\langle \underline{d}_1, \underline{d}_2 \rangle}{\|\underline{d}_1\| \cdot \|\underline{d}_2\|} = \frac{\underline{d}_1^\top \underline{d}_2}{\sqrt{\underline{d}_1^\top \underline{d}_1 \cdot \underline{d}_2^\top \underline{d}_2}} =$$

$$= \frac{\left(\sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \right)}{\sqrt{\sum_i (x_{i1} - \bar{x}_1)^2 \cdot \sum_i (x_{i2} - \bar{x}_2)^2}} = \frac{s_{12}}{\sqrt{s_{11} \cdot s_{22}}} = \gamma_{12} \in (-1, 1)$$

correlation

$$s_{12} = \frac{1}{n-1} \sum_i (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = \text{covariance } (x_1, x_2)$$

covariance

$$\theta = 0 \iff \gamma_{12} = 1$$

$$\theta = \pi \iff \gamma_{12} = -1$$

$$\theta = \frac{\pi}{2} \iff \gamma_{12} = 0$$

$$0 < \theta < \frac{\pi}{2} \iff \gamma_{12} \in (0, 1)$$

data

\mathbb{X}
 $n \times p$

$$\underline{\bar{X}} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} \in \mathbb{R}^p$$

mean vector

$$S = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ \vdots & \ddots & \ddots & \vdots \\ & & \ddots & S_{pp} \end{bmatrix}$$

$p \times p$

$$R = \begin{bmatrix} 1 & 2_{12} & \dots & 2_{1p} \\ \vdots & \ddots & \ddots & \vdots \\ & & \ddots & 1 \end{bmatrix}$$

$p \times p$

$$\underline{E^x}$$

$$\mathbb{X} \rightarrow \mathbb{X}_{std} = \frac{c_1 - \bar{x}_1}{\sqrt{S_{11}}} \quad \dots \quad , \quad \frac{c_p - \bar{x}_p}{\sqrt{S_{pp}}} \quad \underline{1}$$

compute S for \mathbb{X}_{std}

→ \mathbb{R}

$$\left(\begin{array}{c} \frac{c_{11} - \bar{x}_1}{\sqrt{S_{11}}} \\ \vdots \\ \frac{c_{n1} - \bar{x}_1}{\sqrt{S_{11}}} \\ \vdots \\ \frac{c_{1p} - \bar{x}_p}{\sqrt{S_{pp}}} \\ \vdots \\ \frac{c_{np} - \bar{x}_p}{\sqrt{S_{pp}}} \end{array} \right) = \mathbb{X}_{std}$$

$$S_{ii} = \sum_k^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \cdot \frac{1}{n-1}$$

$$S_{ii} = \sum_k^n (x_{ki} - \bar{x}_i)^2 \cdot \frac{1}{n-1}$$

$$\bar{x}_{1std} = \frac{\bar{c}_1 - \bar{x}_1}{\sqrt{S_{11}}} = 0 \quad \bar{c}_l = \bar{x}_l \quad l \text{ from p}$$

$$\bar{c}_1 = \frac{1}{n} \sum_{i=1}^n c_{i1}$$

$$S_{ij} = \frac{1}{n-1} \sum_{k=1}^n x_{ki} x_{kj}$$

$$S_{11} = \frac{1}{n-1} \sum_{k=1}^n x_{k1}^2$$

$$S_{12} = \frac{1}{n-1} \sum_{k=1}^n x_{k1} x_{k2}$$

$$S_{11std} = \frac{1}{n-1} \sum_{k=1}^n \left(\frac{c_{k1} - \bar{x}_1}{\sqrt{S_{11}}} \right)^2 = \frac{1}{(n-1) S_{11}} \sum_{k=1}^n (c_{k1} - \bar{x}_1)^2 = S_{11}$$

$$= S_{12}$$

$$S_{12std} = \frac{1}{n-1} \sum_{k=1}^n \frac{(c_{k1} - \bar{x}_1)(c_{k2} - \bar{x}_2)}{\sqrt{S_{11}} \sqrt{S_{22}}} = \frac{S_{12}}{\sqrt{S_{11} S_{22}}}$$

$$S_{std} = \begin{bmatrix} 1 & \frac{s_{12}}{\sqrt{s_{11}s_{22}}} & \frac{s_{13}}{\sqrt{s_{11}s_{33}}} & \dots & \frac{s_{1p}}{\sqrt{s_{11}s_{pp}}} \\ & \vdots & \ddots & \vdots & \frac{s_{p-1p}}{\sqrt{s_{p-1,p-1}s_{pp}}} \\ & & & 1 & \end{bmatrix} = R ?$$

✓
yep

20. 02. 25

features

Dataset

$$\mathcal{X} = \begin{bmatrix} 1 & x_1, \dots, x_p \\ \vdots & \vdots \\ n & x_{ij} \end{bmatrix} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \quad - \text{variables}$$

$x_i \in \mathbb{R}^p$

$n \times p$

special variable

(y)

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

target variable

$\mathcal{X} \cup y$ - training set

Supervised problem

both $(x_1 \dots x_p)^T = \underline{x}$ & y are observed
on n statistical units \Rightarrow association
between \underline{x} and y is directly observed
in the training set.

Goal: find $f: \mathbb{R}^p \rightarrow \mathbb{R}$, or $f: \mathbb{R}^p \rightarrow \{a, b, c, d\}$
regression classification

$f(\underline{x})$ "optimal" for exploring the variability of y

Optimality problem

Find $f: \mathbb{R}^p \rightarrow \mathbb{R}$ s.t (such that)

$$E[(f(\underline{x}) - y)^2] \text{ is } \min$$

{ Hilbert space
Banach space }

Unsupervised problem

Target y is hidden. we are not able
to observe targets.

So we try to estimate y .

Note :

$$\begin{aligned} E[(f(\underline{x}) - Y)^2] &= E\left[\left(f(\underline{x}) - E[Y|\underline{x}] + E[Y|\underline{x}] - Y\right)^2\right] \\ &= E\left[\left(f(\underline{x}) - E[Y|\underline{x}]\right)^2\right] + E\left[\left(E[Y|\underline{x}] - Y\right)^2\right] + \\ &\quad \underbrace{+ 2E\left[\left(f(\underline{x}) - E[Y|\underline{x}]\right)\left(E[Y|\underline{x}] - Y\right)\right]}_{\text{Recall } z, w \text{ r.v.s}} \end{aligned}$$

$E[(\cdot)(\cdot)] = E[E[(\cdot)(\cdot)|\underline{x}]]$

$E[z] = E[E[z|w]]$

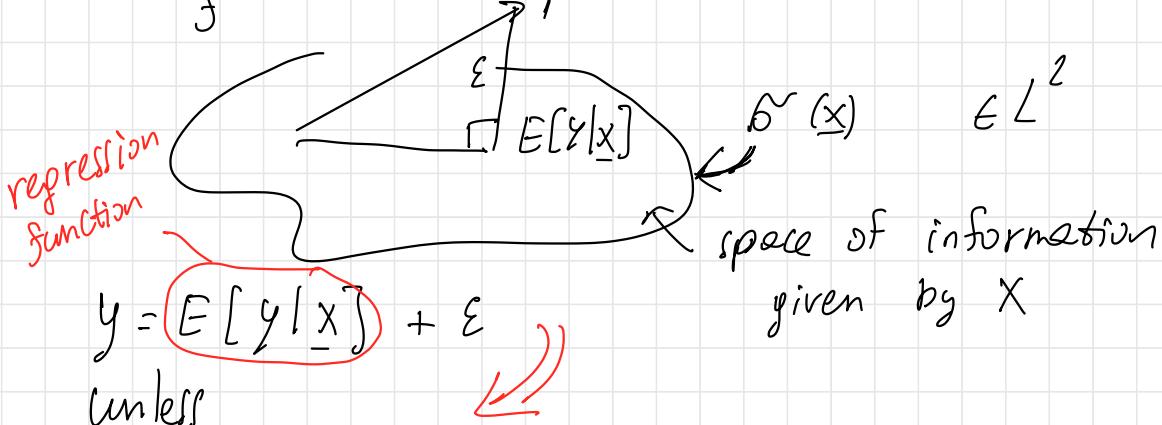
$= E\left[\left(f(\underline{x}) - E[Y|\underline{x}]\right)E\left[E[Y|\underline{x}] - Y|\underline{x}\right]\right]$

0^o

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

" 0

$$\Rightarrow \underset{f}{\operatorname{argmin}} E[(f(\underline{x}) - Y)^2] = E[Y|\underline{x}]$$



$y = f(\underline{x}) + \epsilon$ — General model

$$E[y] = E[E[y|x]] + E[\varepsilon]$$

$$f(x) = E[y|x]$$

$$E[y] = E[y] + E[\varepsilon] \Rightarrow E[\varepsilon] = 0$$

in general we train
 $f(x)$ to be close
to the $E[y|x]$

$$\varepsilon \perp \delta(x)$$

Assume $\text{Var}(\varepsilon) = \sigma^2 < \infty$

$$E[(\varepsilon - E[\varepsilon])^2] \quad || \\ \tilde{\varepsilon} \quad E[\varepsilon^2]$$

Use training set : (\tilde{x}, \tilde{y})
 $n \times p, h \times 1$

to estimate f

by mea sure of

$\hat{f} : \mathbb{R}^p \rightarrow \mathbb{R}$, \hat{f} known once date (\tilde{x}, \tilde{y}) have
been observed

$$\hat{f} : \text{date} \rightarrow \hat{f}$$

estimators

estimate

himbi

ingredients

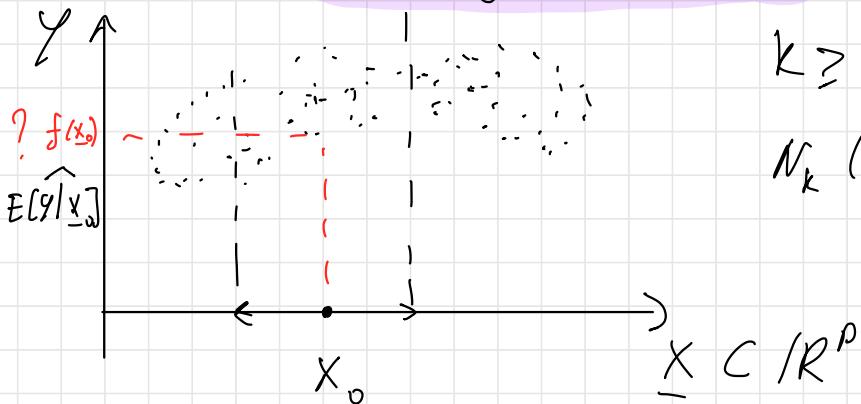
pasta

(augustus ka)

local methods to estimate f

K -nearest neighbour regression

everywhere
ML,
Recommend
System,
Statistic



$$K \geq 1 \quad \forall \underline{x} \in \mathbb{R}^p$$

$$N_k(\underline{x}) = \left\{ \begin{array}{l} k \\ \underline{x}_i \in \mathbb{X} \end{array} \right\}$$

closest to \underline{x}

$$\hat{E}[y|\underline{x}] = \frac{1}{k} \sum_{\underline{x}_i \in N_k(\underline{x})} y_i$$

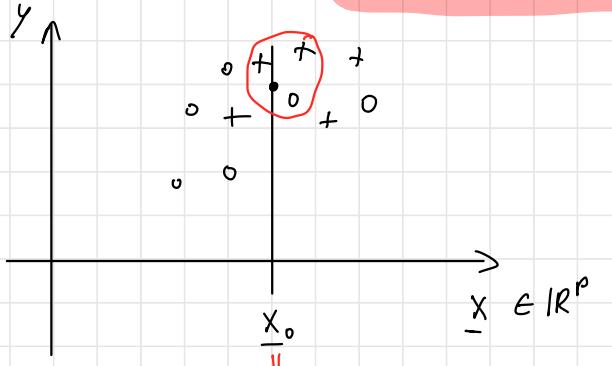
What if we in classification

$$y \in \{-1, +1\}$$

$$f: \mathbb{R}^p \rightarrow \{0, 1\}$$

$$\forall \underline{x} \in \mathbb{R}^p$$

$$\hat{f}(\underline{x}) = \begin{cases} \bullet & \left\{ \begin{array}{l} \bullet \text{ corresponding to} \\ \underline{x}_i \in N_k(\underline{x}) \end{array} \right\} > \\ + & \left\{ \begin{array}{l} + \text{ corresponding to} \\ \underline{x}_i \in N_k(\underline{x}) \end{array} \right\} \\ \text{otherwise} & \end{cases}$$



it's
majority votes

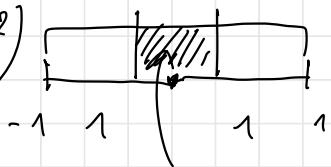
Curse of dimensionality.

imagine
dimension
problem

KNN fine for small p

but worse for large number

$$p = 1 \quad (n = 10^2)$$



uniformly distributed

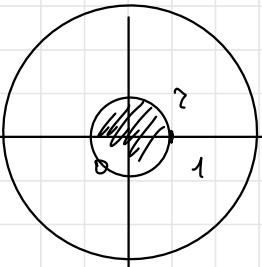
universe
has radius 1

$$10\% \Rightarrow \frac{2^2}{2 \cdot 1} = 0,1$$

to meet

\downarrow
 $Z = 0,1$ 10% of
my friends

$$p = 2 \quad (n = 10^4)$$



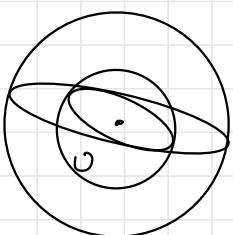
$$\frac{\pi Z^2}{\pi \cdot 1^2} = Z^2 = 0,1$$

$$Z = \sqrt{0,1} = 0,316$$

\Rightarrow now I have to travel 10% of universe

$$p = 3 \Rightarrow \frac{4}{3} \pi R^3$$

$$(n = 10^6)$$



$$Z = 0,464$$

$$\frac{\frac{4}{3} \pi Z^3}{\frac{4}{3} \pi \cdot 1^3} = 0,1 \Rightarrow Z = \sqrt[3]{0,1}$$

T.e. с уменьшением параллельности между точками

10% более вероятны
изменения
вокруг
средней

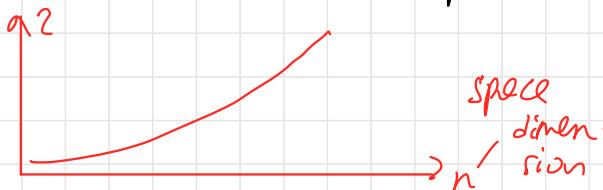
$$P = 100$$

$$Z^{100} = 0.1 \quad Z = \sqrt[100]{0.1} = 0.977$$

Close points similarities disappears

$$(n=10^{200})$$

$$\# \text{ atoms} < 10^{82}$$

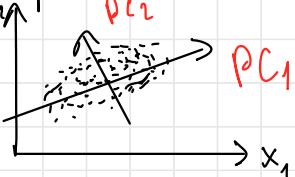


To use local methods we have to reduce dimensionality of the problem.

Reduce the dimensionality of embedding space

(PCA, ICA)

principal component analysis
from ML



find \hat{f} among $\{f_\theta : \theta \in \mathcal{H}\}$

E.g.

$$\hat{f}(\underline{x}) = \underline{\beta} \cdot \underline{x} = \\ = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

\hat{f} estimate of f , when data are known

\underline{x}_0 new $\notin \mathbb{X}$

Generalization error:

Predict \underline{x}_0 : $\hat{f}(\underline{x}_0)$

$$E_{\text{data}} (\gamma_0 - \hat{f}(\underline{x}_0))^2 =$$

$$\gamma_0 = f(\underline{x}_0) + \varepsilon_0$$

$$= E_{\text{data}} \left[(f(\underline{x}_0) + \varepsilon_0 - \hat{f}(\underline{x}_0))^2 \right] = E_{\text{data}} [(f(\underline{x}_0) - \hat{f}(\underline{x}_0))^2] +$$

$$+ E_{\text{data}} [\varepsilon_0^2] + \underbrace{E_{\text{data}} [\varepsilon_0 (f(\underline{x}_0) - \hat{f}(\underline{x}_0))]}_{= 0} =$$

$$= (f(\underline{x}_0) - \hat{f}(\underline{x}_0))^2 + \text{Var}(\varepsilon_0)$$

$$E_{\text{data}} (\gamma_0 - \hat{f}(\underline{x}_0))^2 = (f(\underline{x}_0) - \hat{f}(\underline{x}_0))^2 + \text{Var}(\varepsilon_0)$$

estimate
fixed \hat{f} for today's data
reducible
irreducible

$\text{Err}(\hat{f})$
for all
data

so we can make
it better

estimator

$$E \left[E_{\text{data}} [(\gamma_0 - \hat{f}(\underline{x}_0))^2] \right] =$$

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

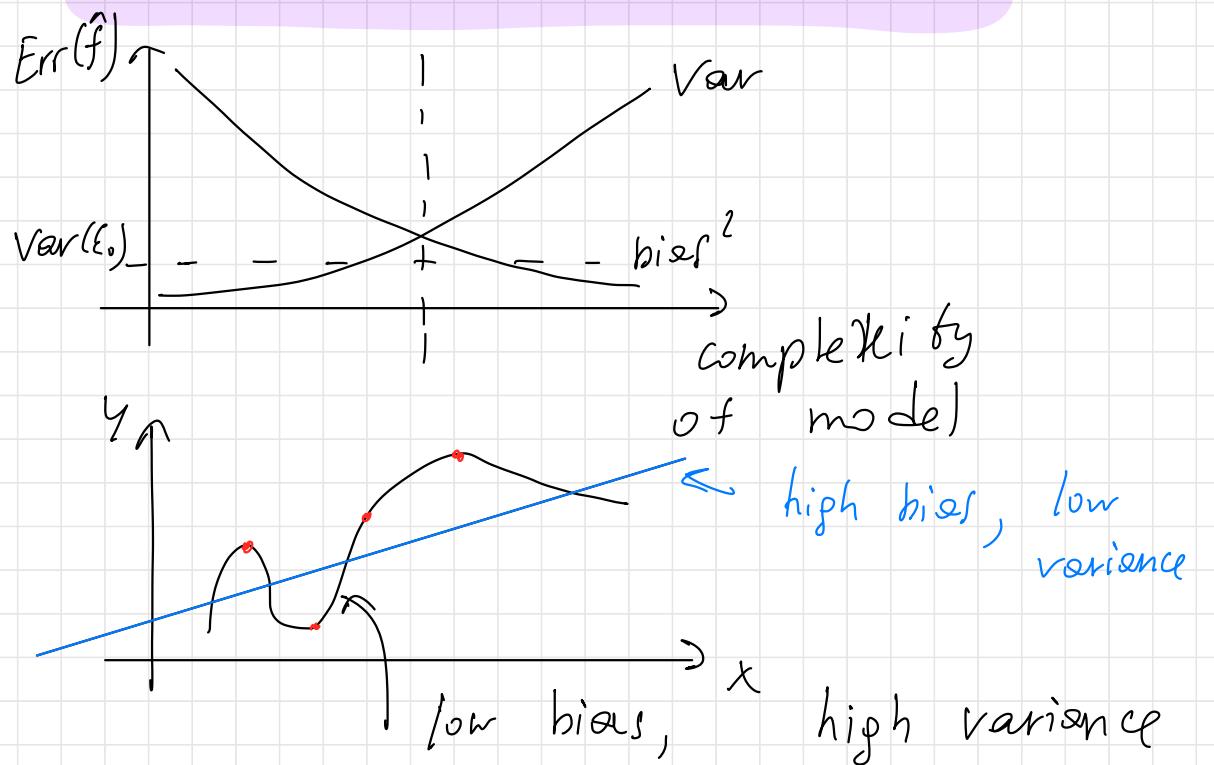
$$= E[(f(\underline{x}_0) - E[\hat{f}(\underline{x}_0)])^2] + E[\hat{f}(\underline{x}_0) - \hat{f}(\underline{x}_0)]^2 + \text{Var}(\varepsilon_0)$$

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

↓ ↑
average of my model variance

of my himhi

Bias - Variance trade-off



Additional ..

- If we apply \ln, x^n and compute correlation, can we say, that now we have not linear correlation between features, but \ln, x^n correlation?
-

About Boxplot

$\vec{x}_n = (x_1, \dots, x_n)$ — sample

Variational series — ordered sample

$$x_{(1)} = x_{\min} \leq x_{(2)} \leq \dots \leq x_{(n)} = x_{\max}$$

$x_{(k)}$ — k-th ordered statistic

r — sample range $= x_{\max} - x_{\min}$

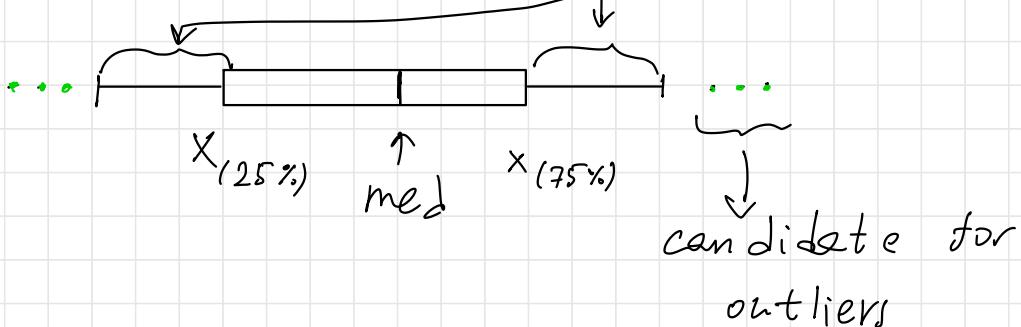
med — median of sample $\Downarrow \begin{cases} x_{(k+1)} & n=2k+1 \\ \frac{x_{(k)} + x_{(k+1)}}{2} & n=2k \end{cases}$

mod — most common element in sample

$$X_{(75\%)} - X_{(25\%)} = \varepsilon \quad \text{Interquartile Range (IQR)}$$

$$1) \quad x_{\min} < x_{(25\%)} - 1.5\varepsilon$$

$$x_{\max} > x_{(75\%)} + 1.5\varepsilon$$



2) If not 1) \Rightarrow



H_0 - main hypothesis

H_1 - alternative hypothesis (class of deviations from the main hypothesis)

Simple hypothesis - a hypothesis that uniquely defines the probability distribution of the model

Composite hypothesis - a hypothesis that does not uniquely define the probability distribution of the model

The main principle of hypothesis testing is: if a low-probability event (assuming H_0 is true) is observed in the experiment, then the null hypothesis does not align with the experiment (consistency).

Otherwise, there is no reason to reject H_0 .

α — significant level, usually set at 0.05, 0.01, 0.1 (determined before the experiment).

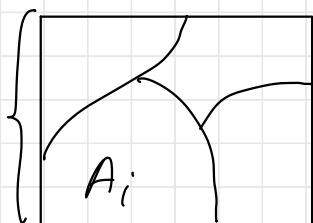
The Pearson's chi-square test for simple H_0

A_1, A_2, \dots, A_m — complete group of events

$$1) A_i \cap A_j = \emptyset \quad i \neq j$$

$$2) A_1 + A_2 + \dots + A_m = \Omega \rightarrow$$

$$3) P(A_i) > 0$$



$$\left. \begin{array}{l} H_0: f \sim F(x) \\ H_1: f \not\sim F(x) \end{array} \right\} \quad \left. \begin{array}{l} f \sim F(x, \theta) \\ H_0: \theta = \delta - \text{simple} \\ H_1: \theta > \delta - \text{composite} \end{array} \right.$$

$\xrightarrow{\quad}$
 X_n - sample A_1, \dots, A_m

$$H_0: P(A_1) = p_1 \quad \dots \quad P(A_m) = p_m$$

$$\widehat{P}(A_1) = \frac{m_1}{n} = \hat{p}_1 \quad \dots \quad \widehat{P}(A_m) = \frac{m_m}{n} = \hat{p}_m$$

- if the observed frequencies significantly differ from the expected frequencies (calculated under H_0), it indicates that the data do not fit the null hypothesis well

$$\Delta = n \sum_{i=1}^n \frac{1}{p_i} (\hat{p}_i - p_i)^2 - \text{measure of the discrepancy between } H_0 \text{ and } \overrightarrow{x_n}$$

$$\Delta = n \sum_{i=1}^m \frac{1}{p_i} \left(\frac{m_i}{n} - p_i \right)^2 = \sum_{i=1}^m \frac{(m_i - np_i)^2}{np_i}$$

Theorem: Under validity of H_0 $\Delta \sim \chi^2_{(m-1)}$

(Proof in Stats3.pdf file page 23) Pearson

$$p\text{-value} = P(\Delta \geq \tilde{\Delta} \mid H_0)$$

(Probability, that deviation will be bigger than our computed deviation $\tilde{\Delta}$)

1) p-value $< \alpha \Rightarrow$ reject H_0 , and as smaller p-value, as reliable, the rejection

RH. (Bepaenning til at der er omgang med kline-kone-kone Δ såsom den samme udvirkning af klinen og konen $\tilde{\Delta}$ nærliggende data, og ikke gavning $\tilde{\Delta}$ mindre betyde i sammenhæng med H_0 relateret)

2) p-value $> \alpha \Rightarrow$ there is no reason to reject H_0 (but it doesn't mean that we accept H_0)

To accept something new we need very
small α

!!! p-value is not probability of main hypothesis. It's degree of confidence in rejecting the hypothesis

24.02.25.

(Week 2)

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

Assumption

$$\underline{x}_1 \leftarrow \underline{x}_1$$

:

$$\underline{x}_n \leftarrow \underline{x}_n$$

data random vectors

i.i.d independent

and identical distribution

$$\underline{x}_1, \dots, \underline{x}_n \text{ iid} \sim \underline{x} \in \mathbb{R}^p$$

$$y_x : \mathbb{B} \rightarrow [0, 1]$$

special case

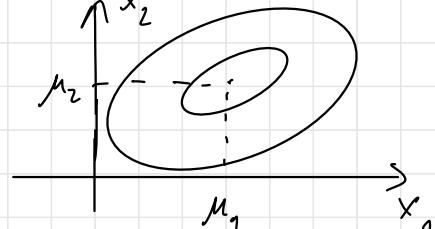
$$y_x(\mathcal{B}) = \int_{\mathcal{B}} f(\underline{x}) d\underline{x}$$

$$y_x(\mathcal{B}) = P[\underline{x} \in \mathcal{B}]$$

f-density of y_x or of \underline{x}

$$\text{E.x. } f(\underline{x}) = \frac{1}{(2\pi)^p \text{Det}(\Sigma)} \exp\left(-\frac{1}{2} (\underline{x} - \underline{\mu})^\top \Sigma^{-1} (\underline{x} - \underline{\mu})\right)$$

$\mu \in \mathbb{R}^p$ Σ $p \times p$ positive defined



We will try to transform data to Gaussian Distribution

Summaries of \underline{D}_x

Column

$$\underline{\mu} = E[\underline{x}] = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \quad \mu_j = E[x_j] \quad j=1\dots p$$

↑ expectation
for each feature

mean of distribution

$$\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}^T$$

$$\Sigma = [\delta_{ij}] \quad p \times p \quad \text{matrix}$$

$$\delta_{ij} = \text{Cov}(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)] \quad i, j = 1 \dots p$$

$$\Sigma = E[(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^T]$$

$$V = \begin{bmatrix} \delta_{11} & \dots & \dots \\ \vdots & \ddots & \delta_{pp} \end{bmatrix}$$

$$\rho = V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}}$$

correlation
matrix

$$V^{-\frac{1}{2}} = \begin{bmatrix} 1/\sqrt{\delta_{11}} & & \\ & \ddots & \\ & & 1/\sqrt{\delta_{pp}} \end{bmatrix}$$

if standardise $\underline{x} \rightarrow \underline{x}_{std}$

$$\text{then } \rho = \text{Cov}(\underline{x}_{std})$$

$$= \begin{pmatrix} \frac{x_1 - \mu_1}{\sqrt{\delta_{11}}} \\ \vdots \\ \frac{x_p - \mu_p}{\sqrt{\delta_{pp}}} \end{pmatrix}$$

Linear comb of components of \underline{x}

$$p=2 \quad \underline{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$c_1 x_1 + c_2 x_2$$

$$E[c_1 x_1 + c_2 x_2] = c_1 E[x_1] + c_2 E[x_2]$$

$c \in \mathbb{R}^2$

$$\underline{c} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

$$= c_1 \mu_1 + c_2 \mu_2$$

$$E[\underline{c}^\top \underline{x}] = \underline{c}^\top \underline{\mu}$$

$$\text{Var}(c_1 x_1 + c_2 x_2) = c_1^2 \text{Var}(x_1) + c_2^2 \text{Var}(x_2) +$$

$$+ 2c_1 c_2 \text{Cov}(x_1, x_2) = c_1^2 \sigma_{11} + c_2^2 \sigma_{22} + 2c_1 c_2 \sigma_{12} =$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \underline{c}^\top \Sigma \underline{c}$$

$$P \quad \underline{x} \in \mathbb{R}^p \quad E[\underline{c}^\top \underline{x}] = \underline{c}^\top \underline{\mu}$$

$$\underline{c} \in \mathbb{R}^p \quad \text{Var}(\underline{c}^\top \underline{x}) = \underline{c}^\top \Sigma \underline{c}$$

G $k \times p$ matrix \underline{x}

$$E[G \underline{x}] = G \underline{\mu}$$

$G \underline{x} \in \mathbb{R}^k$

$$\text{Cov}(G \underline{x}) = G \Sigma G^\top$$

$$k=1 \rightarrow G = c'$$

multiple linear combinations

we had independence between observations but not in features

?

μ and Σ are often unknown (parameters)
→ use data to estimate them

Estimator for μ : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$? result $\in \mathbb{R}^P \rightarrow$

$$\rightarrow \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

for Σ : $S = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top$
~ $n-1$ degrees of freedom

In case μ is known

Estimator for Σ : $S = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^\top$
 n degrees of freedom

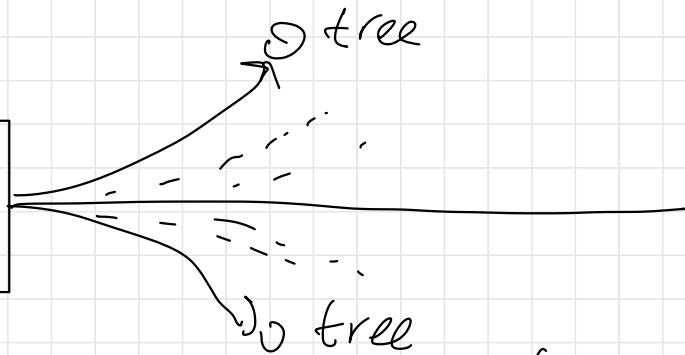
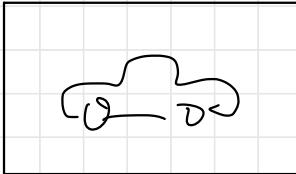
Proposition

① $E[\bar{x}] = \mu$ (unbiased for μ)

$$\text{Cov}(\bar{x}) = \frac{1}{n} \sum$$

(negative definite)

② $E[S] = \Sigma$ (unbiased for Σ)



Proof

$$E[\bar{\underline{x}}] = E\begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{pmatrix} E[\bar{x}_1] \\ \vdots \\ E[\bar{x}_p] \end{pmatrix} =$$

$$= \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} = \underline{\mu}$$

vector of sample means

matrix

$$\text{Cov}(\bar{\underline{x}}) = E[(\bar{\underline{x}} - \underline{\mu})(\bar{\underline{x}} - \underline{\mu})^T] =$$

$$= E \left[\left(\frac{1}{n} \sum_{i=1}^n (\underline{x}_i - \underline{\mu}) \right) \left(\frac{1}{n} \sum_{j=1}^n (\underline{x}_j - \underline{\mu}) \right)^T \right] =$$

$$= \frac{1}{n^2} \sum_i^n \sum_j^n E[(\underline{x}_i - \underline{\mu})(\underline{x}_j - \underline{\mu})^T] =$$

$$E[(\underline{x}_i - \mu)(\underline{x}_j - \mu)^T] = \sum_{\substack{i=j \\ \text{because identical distributed}}} \Omega$$

\nwarrow

$$E[(\underline{x}_i - \mu)] E[(\underline{x}_j - \mu)]$$

\nwarrow because they are independent

Ex. prove that

$$E \left[\sum_{i=1}^n (\underline{x}_i - \bar{\underline{x}}) (\underline{x}_i - \bar{\underline{x}})^T \right] =$$

$$= (n-1) \sum$$

row i of Data frame

$$S = \frac{1}{n-1} \sum_{i=1}^n (\underline{x}_i - \bar{\underline{x}}) (\underline{x}_i - \bar{\underline{x}})^T \quad \begin{matrix} \leftarrow & \text{row perspective} \end{matrix}$$

Column perspective

$$\mathbb{X} = [\underline{c}_1 \dots \underline{c}_p]$$

$$\underline{c}_j = \underline{c}_j - \frac{\underline{1} \underline{1}^T}{\underline{1}^T \underline{1}} \underline{c}_j = \left[\underline{I} - \frac{\underline{1} \underline{1}^T}{\underline{1}^T \underline{1}} \right] \underline{c}_j$$

\mathbb{R}^n

$$\mathbf{d} = \begin{bmatrix} \underline{d}_1 & \dots & \underline{d}_p \end{bmatrix} = \left(\mathbf{I} - \frac{\underline{1}\underline{1}^T}{\underline{1}^T\underline{1}} \right) \mathbf{X}$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \\ & \ddots \end{bmatrix} = \frac{1}{n-1} \begin{bmatrix} \underline{d}_1^T \underline{d}_1 & \underline{d}_1^T \underline{d}_2 \dots \underline{d}_1^T \underline{d}_p \\ \underline{d}_2^T \underline{d}_1 & \underline{d}_2^T \underline{d}_2 \dots \\ & \ddots \end{bmatrix}$$

$$\underline{d}_1^T \underline{d}_1 = \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i1} - \bar{x}_1)$$

$$= \frac{1}{n-1} \mathbf{d}^T \mathbf{d} = \frac{1}{n-1} \mathbf{X}^T \left(\mathbf{I} - \frac{\underline{1}\underline{1}^T}{\underline{1}^T\underline{1}} \right) \left(\mathbf{I} - \frac{\underline{1}\underline{1}^T}{\underline{1}^T\underline{1}} \right) \mathbf{X}$$

$$\Sigma = \frac{1}{n-1} \mathbf{X} \left(\mathbf{I} - \frac{\underline{1}\underline{1}^T}{\underline{1}^T\underline{1}} \right) \mathbf{X}$$

equal

Column persp.

$$\Sigma = \frac{1}{n-1} \sum_{i=1}^n (\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})^T$$

row perspective

Generalized variance	$\text{Det}(S)$	(sample) (provided data)
	$\text{Det}(\Sigma)$	(population) (all data in the world)

Total variance

$\text{tr}(S)$ (sample)
 $\text{tr}(\Sigma)$ (population)

Example (picture)

$$p=2$$

$$S = \frac{1}{n-1} \begin{bmatrix} \underline{d}_1^T \underline{d}_1 & \underline{d}_1^T \underline{d}_2 \\ \underline{d}_2^T \underline{d}_1 & \underline{d}_2^T \underline{d}_2 \end{bmatrix}$$

↔

variability of feature 1
cov of feature 1,2
variability of feature 2

$$\text{Det}(S) = \frac{\|d_1\|^2 \|d_2\|^2 - (\cos \theta_{12})^2 \|d_1\|^2 \|d_2\|^2}{(n-1)^2} =$$

$$= \frac{1}{(n-1)^2} \|d_1\|^2 \|d_2\|^2 (\sin \theta_{12})^2$$

take into account correlation between variables

$$\text{Det}(S) \propto (\|d_1\| \|d_2\| \sin \theta_{12})^2 - \text{shape of}$$

$$\text{tr}(S) = \|d_1\|^2 + \|d_2\|^2 - \text{look only on variability of each variable}$$

For general p

Generalized Variables $\propto \text{Volume}^2 (\underline{d}_1, \dots, \underline{d}_p)$

$$\text{Total variance} = (\|\underline{d}_1\|^2 + \|\underline{d}_2\|^2 + \dots + \|\underline{d}_p\|^2)$$

Proposition

If $\text{Det}(S) > 0 \iff \underline{d}_1, \dots, \underline{d}_p$ are linearly independent

$\underline{d}_1, \dots, \underline{d}_p$ lin dep $\Rightarrow \exists \underline{\lambda} \in \mathbb{R}^p$ s.t. $\underline{\lambda} \neq 0$

$$\lambda_1 \underline{d}_1 + \lambda_2 \underline{d}_2 + \dots + \lambda_p \underline{d}_p = 0$$

$\exists \lambda_1 \neq 0$

$$c_1 - \bar{x}_1 = \frac{-\lambda_1}{\lambda_1} \underline{d}_1 - \frac{\lambda_2}{\lambda_1} \underline{d}_2 - \dots - \frac{\lambda_p}{\lambda_1} \underline{d}_p$$

Prof

$$\exists c \neq 0 \text{ s.t. } c_1 \underline{d}_1 + \dots + c_p \underline{d}_p = 0$$

(\Leftarrow)

$$\underline{d}^\top c = 0$$

$$S = \frac{1}{n-1} \underline{d}^\top \underline{d}^\top = 0$$

$$S_c = \frac{1}{n-1} \underline{d}^\top \underline{d}^\top c = 0 \Rightarrow$$

\Rightarrow cols of S are linear dep \Rightarrow

$$\Rightarrow \text{Rref}(S) = 0$$

$$(\Rightarrow) \text{Rref}(S) = 0 \Rightarrow \exists \subseteq \neq \emptyset \rightarrow \underline{\underline{S}}_{\subseteq} = 0$$

$$\underline{\underline{C}}^T \underline{\underline{S}}_{\subseteq} = 0$$

$$\frac{1}{n-1} \underline{\underline{e}}^T \underline{\underline{d}} \underline{\underline{l}}^T \underline{\underline{d}} \underline{\underline{l}}_{\subseteq} = 0$$

$$\frac{1}{n-1} \|\underline{\underline{d}} \underline{\underline{l}}_{\subseteq}\|^2 = 0 \quad \|\underline{\underline{d}} \underline{\underline{l}}_{\subseteq}\| = 0 \Rightarrow \underline{\underline{d}} \underline{\underline{l}}_{\subseteq} = 0$$

$\Rightarrow \underline{\underline{d}}_1, \dots, \underline{\underline{d}}_p$ - linearly dependent

Prop \times $n \times p$ data

If $p \geq n \Rightarrow \text{Rref}(S) = 0 \Rightarrow$

$\Rightarrow \underline{\underline{d}}_1, \dots, \underline{\underline{d}}_n$ linearly dependent

don't make more features than data

$$\text{Prof} \quad \underline{d}_1, \dots, \underline{d}_p \in \mathbb{R}^n$$

$$\underline{d}_j = \underline{c}_j - \frac{\underline{1} \underline{1}^\top}{\underline{1}^\top \underline{1}} \underline{c}_j \in \mathcal{L}^\perp(\underline{1}) \quad \dim(\mathcal{L}^\perp(\underline{1})) = n-1$$

degree freedom

$$\underline{d}_1, \dots, \underline{d}_p$$

maximum number of lin independent vectors

in $\mathcal{L}^\perp(\underline{1})$ is $n-1 \Rightarrow p > n-1$ i.e $p \geq n \Rightarrow$

$\underline{d}_1, \dots, \underline{d}_p$ are linearly dependent

prove that

$$E \left[\sum_{i=1}^n (\underline{x}_i - \bar{\underline{x}}) (\underline{x}_i - \bar{\underline{x}})^\top \right] =$$

$$= (n-1) \Sigma$$

$$= \sum_{i=1}^n \left(E[(\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})^\top] \right) =$$

$$= \sum_{i=1}^n E[(\underline{x}_i - \mu_i)(\underline{x}_i - \mu_i)^\top] = (n-1) \Sigma$$

25. 02. 25.

Prop S positive semi def \Rightarrow

$$\forall \underline{c} \in \mathbb{R}^n \quad \underline{c}^\top S \underline{c} \geq 0, \underline{c} \neq 0$$

if $\text{Det}(S) > 0$, then S is positive def

Rmk if $\underline{x} \sim \underline{\mu}, \Sigma$

$$\underline{c}^\top \Sigma \underline{c} = \text{Var}(\underline{c}^\top \underline{x}) \geq 0$$

Prof $S = \frac{1}{n-1} \underline{d} \underline{d}^\top \quad \forall \underline{c} \neq 0 \quad \underline{c}^\top S \underline{c} =$

$$= \frac{1}{n-1} \underline{c}^\top \underline{d} \underline{d}^\top \underline{c} = \frac{1}{n-1} \|\underline{d}^\top \underline{c}\|^2 \geq 0$$

Assume $\exists \underline{c} \neq 0$, s.t. $\underline{c}^\top S \underline{c} = 0 \Rightarrow$

$$\Rightarrow \|\underline{d}^\top \underline{c}\| = 0 \Rightarrow \underline{d}^\top \underline{c} = 0 \Rightarrow d_1, \dots, d_p \text{ are lin}$$

dep $\Rightarrow \text{Det}(S) = 0 \Rightarrow$

$\text{Det}(S) \neq 0 \Rightarrow S$ positive def

$p \times p$
 S symmetric coef numbers

Hence, S can be represented as

$$S = \sum_{i=1}^p \lambda_i e_i e_i^T$$

(spectral decomposition)

where $e_i \in \mathbb{R}^p$ $i = 1 \dots p$

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

$e_i \perp e_j$ if $i \neq j$

S-positive semi-def \Rightarrow

$$\lambda_i \geq 0 \quad i = 1 \dots p$$

moreover (λ_i, e_i)

ortho normal basis for \mathbb{R}^p

$$\{e_1, \dots, e_p\}$$

$i = 1 \dots p$ (eigenvalues, eigenvectors)

Gen Var: $\text{Det}(S) = \prod_{i=1}^p \lambda_i$

Hence $\text{Det}(S) > 0 \Rightarrow \lambda_i > 0 \quad i = 1 \dots p$

Total var: $\text{tr}(S) = \sum_{i=1}^p \lambda_i$

Conventions

From now on:

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$$

$$\begin{matrix} | & | & & | \\ e_1 & e_2 & \dots & e_p \end{matrix}$$

{ sorted eigen values
and their vectors }

β $p \times p$ is pos def

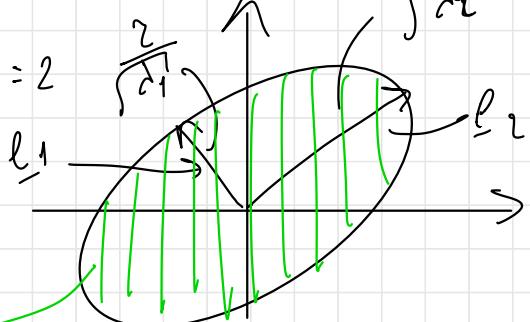
$$Q(\underline{x}) = \underline{x}^T \beta \underline{x} \quad \underline{x} \in \mathbb{R}^p$$

Consider $\left\{ \underline{x} \in \mathbb{R}^p \mid \underline{x}^T \beta \underline{x} \leq 2 \right\} = E_2(0) \rightarrow$

ellipse with center in zero.

$$\beta = \sum_{i=1}^p \lambda_i e_i e_i^T$$

$$p=2$$



$$\text{Vol}(E_2(0)) = k_p \frac{1}{\sqrt{\prod_{i=1}^p \lambda_i}}$$

Mahalanobis dist

A new distance (stat) in \mathbb{R}^p

$$\underline{x}, \underline{y} \in \mathbb{R}^p$$

$$d_{S^{-1}}(\underline{x}, \underline{y}) = \sqrt{(\underline{x} - \underline{y})^T S^{-1} (\underline{x} - \underline{y})}$$

Assuming
- $\det S > 0$

- Note

$$S = \sum d_i e_i e_i^T$$

$$S^{-1} = \sum \frac{1}{d_i} e_i e_i^T$$

$$S^k = \sum (d_i)^k e_i e_i^T$$

$$\sqrt{S} = \sum \sqrt{d_i} e_i e_i^T$$

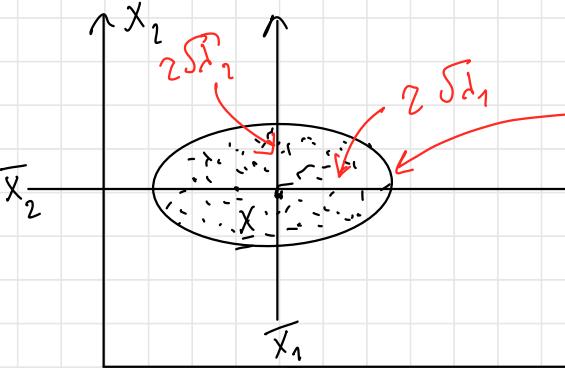
$$= \sqrt{(\underline{x} - \underline{y})^T S^{-1/2} S^{-1/2} (\underline{x} - \underline{y})} = \\ = \sqrt{(S^{-1/2} \underline{x} - S^{-1/2} \underline{y})^T (S^{-1/2} \underline{x} - S^{-1/2} \underline{y})} =$$

$$d_{\text{eucl}} \left(\underline{s}_{\underline{x}}, \underline{s}_{\underline{y}} \right)$$

Example

$$p=2$$

$$\Sigma = \begin{bmatrix} s_{11} & 0 \\ 0 & s_{22} \end{bmatrix} \quad s_{11} > s_{22}$$



$$\mathcal{E}_2(\bar{x}) = \left\{ \underline{x} \in \mathbb{R}^p : (\underline{x} - \bar{\underline{x}})^T \Sigma^{-1} (\underline{x} - \bar{\underline{x}}) \leq 2^2 \right\}$$

$$\underline{x}, \underline{y} \in \mathbb{R}^p$$

$$\text{Vol}(\mathcal{E}_2(\bar{x})) \propto 2^2 \sqrt{d_1 d_2} = 2^2 \sqrt{\det(\Sigma)} = 2^2 \sqrt{\text{GenVar}} X_1 \quad d_{\Sigma^{-1}}(\underline{x}, \underline{y}) = \sqrt{(\underline{x} - \underline{y})^T \Sigma^{-1} (\underline{x} - \underline{y})}$$

$$\underline{d} = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}$$

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

$$= \sqrt{\frac{(x_1 - y_1)^2}{s_{11}}} + \sqrt{\frac{(x_2 - y_2)^2}{s_{22}}} = d_{\text{eucl}}(\underline{x}_{\text{std}}, \underline{y}_{\text{std}}) \text{ normalised}$$

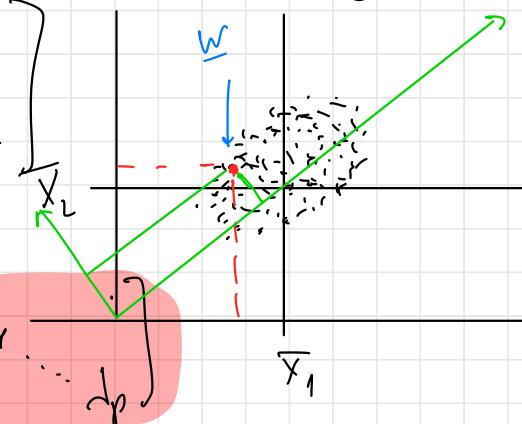
so Mahalanobis distance take into account variance among dimension

$$p=2 \text{ General } \Sigma = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \quad (\text{have correlation})$$

$$\Sigma = \sum_{i=1}^p b_i \underline{e}_i \underline{e}_i^T$$

$$\underline{P} = [\underline{e}_1 \dots \underline{e}_p]$$

$$\Lambda = \begin{bmatrix} 1 & \dots & 1_p \end{bmatrix}$$



$$\Rightarrow S = P \Lambda P^T$$

Old system

$$(x_1, \dots, x_p)$$

$$w$$

$$w = \tilde{w}$$

New system

$$(\underline{e}_1, \dots, \underline{e}_p)$$

$$\tilde{w} = \begin{bmatrix} e_1' & w \\ e_2' & \\ \vdots & \\ e_p' & w \end{bmatrix} = P^T w$$

Old

$$\mathbf{x} = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}$$

$$\tilde{\mathbf{x}} = \mathbf{x}P$$

new

$$\tilde{\mathbf{x}} = \begin{bmatrix} (P' x_1)' \\ \vdots \\ (P' x_n)' \end{bmatrix} = \begin{bmatrix} x_1' P \\ \vdots \\ x_n' P \end{bmatrix} = \mathbf{x}P$$

$$S = \frac{1}{n-1} \mathbf{x}' \left(I - \frac{1 \ 1'}{1' \ 1} \right) \mathbf{x}$$

$$= \left(\frac{1}{n-1} P' \mathbf{x}' \left(I - \frac{1 \ 1'}{1' \ 1} \right) \mathbf{x} \right) P$$

$$\tilde{S} = \frac{1}{n-1} \tilde{\mathbf{x}}' \left(I - \frac{1 \ 1'}{1' \ 1} \right) \tilde{\mathbf{x}} =$$

$$= P' S P = P' P \Lambda P' P = \boxed{P}$$

$$\text{Det}(S) = \prod d_i$$

$$\text{Det}(\tilde{S}) = \prod d_i$$

\Rightarrow I can always find reference system, where correlation disappear
(and our S matrix will be diagonal = Λ)

PCA Principal Component Analysis

$$\mathbb{R}^p \ni \underline{x}, \underline{x} \sim \underline{\mu}, \underline{\Sigma}$$

$$\underline{q} \in \mathbb{R}^p \Rightarrow \text{var}(\underline{q}' \underline{x}) = \underline{q}' \underline{\Sigma} \underline{q}$$

Find \underline{q} s.t. $\text{Var}(\underline{q}' \underline{x})$ is max \Rightarrow

$$\underline{q}^* \text{ sol } 10 \cdot \underline{q}^* \underline{x} \quad \text{Var}(10 \cdot \underline{q}' \cdot \underline{x}) = 100 \text{Var}(\underline{q}' \cdot \underline{x})$$

\Rightarrow additional constraints

$$\underline{q} \text{ s.t. } \|\underline{q}\| = 1 \text{ and } \text{Var}(\underline{q}' \underline{x}) \text{ is max}$$

Find $\underline{\alpha}_1, \underline{\alpha}_2, \dots, \underline{\alpha}_p \subseteq f$

$$\|\underline{\alpha}_i\| = 1 \quad \text{for } i=1 \dots p \quad \text{and}$$

$$\text{Var}(\underline{\alpha}_1^T \underline{x}) = \sup_{\Omega \in \mathbb{R}^p} \text{Var}(\underline{\alpha}^T \underline{x})$$

$$\|\underline{\alpha}\|=1$$

$$\text{Var}(\underline{\alpha}_2^T \underline{x}) = \sup_{\Omega \in \mathbb{R}^p} \text{Var}(\underline{\alpha}^T \underline{x})$$

$$\text{Var}(\underline{\alpha}_1^T \underline{x}, \underline{\alpha}_2^T \underline{x}) = 0$$

$$\|\underline{\alpha}\|=1$$

:

:

$$\text{Var}(\underline{\alpha}_k^T \underline{x}) = \sup_{\Omega \in \mathbb{R}^p} \text{Var}(\underline{\alpha}^T \underline{x})$$

$$\|\underline{\alpha}\|=1$$

$$\text{Var}(\underline{\alpha}_i^T \underline{x}, \underline{\alpha}_j^T \underline{x}) = 0 \quad \text{for } i < j$$

Def. $y_i = \underline{\alpha}_i^T \underline{x}$ i -th principal component

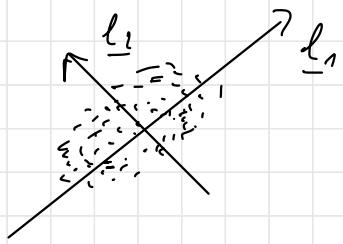
$\underline{\alpha}_i$ i -th loading

y_i i -th score

Theor If $\Sigma = \sum_{i=1}^n \lambda_i \underline{e}_i \underline{e}_i'$

$$1. \quad y_i = e_i' x \quad (\underline{a}_i = \underline{e}_i)$$

$$2. \quad \text{cov}(y_i, y_j) = \begin{cases} \lambda_i & i=j \\ 0 & i \neq j \end{cases}$$



Lemma

B $p \times p$ pos semi def matrix

$$\text{and } B = \sum_{i=1}^p \lambda_i \underline{e}_i \underline{e}_i' = P \Lambda P' \quad \left| \begin{array}{l} \text{so } \underline{e}_i - \text{eigen} \\ \text{value of } B \end{array} \right.$$

Then

$$1. \quad \sup_{\substack{x \neq 0 \\ x \perp l_1}} \frac{\underline{x}' B \underline{x}}{\underline{x}' \underline{x}} = \lambda_1 \quad \text{argmax}_{\substack{x \neq 0 \\ x \perp l_1}} \frac{\underline{x}' B \underline{x}}{\underline{x}' \underline{x}} = \underline{e}_1$$

$$2. \quad \sup_{\substack{x \neq 0 \\ x \perp l_1, l_2}} \frac{\underline{x}' B \underline{x}}{\underline{x}' \underline{x}} = \lambda_2 \quad \text{argmax} \dots = \underline{e}_2$$

$$x \perp \underline{e}_1$$

...

$$p. \quad \sup_{\substack{x \neq 0 \\ x \perp l_1, \dots, x \perp l_{p-1}}} \frac{\underline{x}' B \underline{x}}{\underline{x}' \underline{x}} = \lambda_p = \inf_{\substack{x \neq 0 \\ x \perp l_1, \dots, x \perp l_{p-1}}} \frac{\underline{x}' B \underline{x}}{\underline{x}' \underline{x}} \quad \text{argmax} = \underline{e}_p$$

Prof Lemma

$$\frac{\underline{X}' \underline{B} \underline{X}}{\underline{X}' \underline{X}} = \frac{\underline{X}' \underline{P} \wedge \underline{P}' \underline{X}}{\underline{X}' \underline{X}} = \frac{\underline{X}' \underline{P} \wedge \underline{P}' \underline{X}}{\underline{X}' \underline{P} \underline{P}' \underline{X}} = \begin{cases} Y = P X \\ e_1 \end{cases}$$

$$= \frac{\underline{Y}' \wedge \underline{Y}}{\underline{Y}' \underline{Y}} = \frac{\sum_{i=1}^n \lambda_i Y_i^2}{\sum Y_i^2} \leq$$

$$\leq \frac{\lambda_1 \sum Y_i^2}{\sum Y_i^2} = \lambda_1$$

if we take off $\lambda_i > \lambda_p$

$$\lambda_p \leq \frac{\underline{X}' \underline{B} \underline{X}}{\underline{X}' \underline{X}} \leq \lambda_1, \quad \forall \underline{X} \in \mathbb{R}^p$$

$\underline{B} \underline{e}_1 = \lambda_1 \underline{e}_1$ - by definition
because it eigen value

$$\underline{X}' \underline{P} \underline{P}' \underline{X} = \underline{X}' \underline{X}$$

$$\text{Take } \underline{X} = \underline{e}_1 \Rightarrow \frac{\underline{e}_1' \underline{B} \underline{e}_1}{\underline{e}_1' \underline{e}_1} = \frac{\lambda_1 \underline{e}_1' \underline{e}_1}{\underline{e}_1' \underline{e}_1} = \lambda_1$$

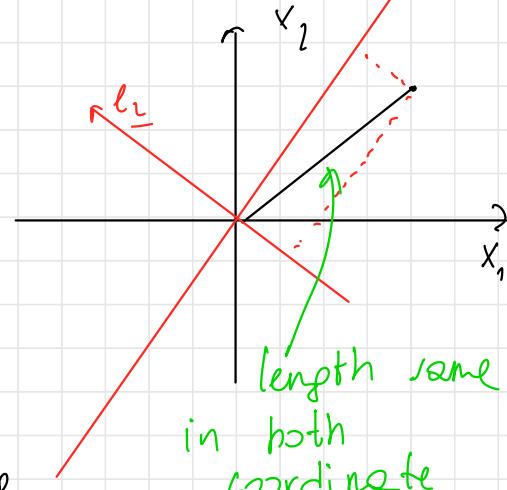
$$\max_{\underline{X} \neq \underline{0}} \frac{\underline{X}' \underline{B} \underline{X}}{\underline{X}' \underline{X}} = \lambda_1$$

argmax $\underline{X}: \underline{e}_1$

$$\text{Take } \underline{X} = \underline{e}_p \Rightarrow \frac{\underline{e}_p' \underline{B} \underline{e}_p}{\underline{e}_p' \underline{e}_p} = \lambda_p$$

$$\max_{\underline{X} \neq \underline{0}} \frac{\underline{X}' \underline{B} \underline{X}}{\underline{X}' \underline{X}} = \lambda_p$$

argmax $\underline{X}: \underline{e}_p$



$$\frac{\underline{x}^T \beta \Delta}{\underline{x}^T \underline{x}} = \frac{\underline{x}^T P \Lambda P^T \underline{x}}{\underline{x}^T P P^T \underline{x}} = \frac{\underline{y}^T \Lambda \underline{y}}{\underline{y}^T \underline{y}} = \frac{\sum_{i=1}^n \lambda_i y_i^2}{\sum y_i^2} \leq \lambda$$

$$\underline{x} \perp \ell_1 \dots \ell_p$$

$$\underline{y} = P \underline{x} = \begin{pmatrix} \ell_1 & \underline{x} \\ \ell_2 & \underline{x} \\ \vdots & \\ \ell_p & \underline{x} \end{pmatrix} = \begin{matrix} 0 \\ \ell_2 \underline{x} \\ 0 \\ \vdots \end{matrix}$$

28.02.25

PCA $\underline{x} \in \mathbb{R}^p$ z. vect $\sim \underline{\mu}, \Sigma$

$$1) \max_{\underline{Q} \in \mathbb{R}^p} \text{Var}(\underline{Q}' \underline{x})$$

$$\|\underline{Q}\| = 1$$

$$\max_{\underline{Q} \in \mathbb{R}^p} \text{Var}(\underline{Q}' \underline{x}) = \max_{\underline{Q} \in \mathbb{R}^p} \frac{\underline{Q}' \Sigma \underline{Q}}{\underline{Q}' \underline{Q}} = \lambda_1$$

$$\|\underline{Q}\| = 1$$

if $\Sigma = \sum_i^p \lambda_i \underline{e}_i \underline{e}_i'$ - (spectral decomposition)

$$\underline{Q}' \underline{Q} = \|\underline{Q}\|^2$$

arg max $\underline{Q}' \underline{e}_i$:

$$y_1 = \underline{e}_1' \underline{x} \quad \text{first-PC} \quad (y_1 = \underline{e}_1' (\underline{x} - \underline{\mu}))$$

usually we have to centre-

lise \underline{x}

$$2) \max_{\underline{Q} \in \mathbb{R}^p} \text{Var}(\underline{Q}' \underline{x})$$

$$\|\underline{Q}\| = 1$$

$$\text{Cor}(\underline{Q}' \underline{x}, \underline{Q}' \underline{x}) = 0 \quad \text{Cor}(\underline{Q}' \underline{x}, \underline{Q}' \underline{x}) = 0$$

$$\max_{\underline{Q} \in \mathbb{R}^p} \frac{\underline{Q}' \Sigma \underline{Q}}{\underline{Q}' \underline{Q}}$$

$$0 = \text{Cov}(\underline{\varphi}' \underline{x}, \underline{e}_1' \underline{x}) = \underline{\varphi}' \sum \underline{e}_1 = \underline{e}_1' \underline{\varphi} \Leftrightarrow$$

$$\underline{\varphi}' \underline{e}_1 = 0 \Leftrightarrow \underline{\varphi} \perp \underline{e}_1$$

$$G \in k \times p \quad \text{Cov}(G \underline{x}) = G \sum G'$$

$$G = \begin{bmatrix} \underline{\varphi}' \\ \underline{e}_1' \end{bmatrix} \quad \text{Cov}(G \underline{x}) = \begin{bmatrix} \underline{\varphi}' \\ \underline{e}_1' \end{bmatrix} \sum \begin{bmatrix} \underline{\varphi} \\ \underline{e}_1 \end{bmatrix} =$$

$$= \begin{bmatrix} \underline{\varphi}' \sum \underline{\varphi} & \underline{\varphi}' \sum \underline{e}_1 \\ \underline{e}_1' \sum \underline{\varphi} & \underline{e}_1' \sum \underline{e}_1 \end{bmatrix}$$

$$\xrightarrow{\quad} = \max_{\underline{\varphi} \in \mathbb{R}^p} \frac{\underline{\varphi}' \sum \underline{\varphi}}{\underline{\varphi}' \underline{\varphi}} = d_2$$

$\underline{\varphi} \perp \underline{e}_1$

$$\text{argmax} : \underline{e}_1 = y_2 = \underline{e}_2' \underline{x} \quad (y_2 = \underline{e}_2' (\underline{x} - \mu))$$

$$k) \max_{\underline{\alpha} \in \mathbb{R}^p} \text{Var}(\underline{\alpha}' \underline{x}) = \max_{\underline{\alpha}} \frac{\underline{\alpha}' \Sigma \underline{\alpha}}{\underline{\alpha}' \underline{\alpha}} = d_k$$

$$\|\underline{\alpha}\|_2 = 1$$

$$\underline{\alpha} = \underline{\ell}_1, \underline{\ell}_2, \dots, \underline{\ell}_{k-1}$$

$$0 = \text{Cov}(\underline{\alpha}' \underline{x}, \underline{\ell}_i' \underline{y}), \quad i = 1 \dots k-1$$

Argmax: $\underline{\ell}_k$

$$y_k = \underline{\ell}_k' \underline{x} \quad (y_k = \underline{\ell}_k' (\underline{x} - \underline{\mu})) \text{ k-th PC}$$

$$\mathbb{R}^D \ni \underline{Y} = P' \underline{x} \quad \text{vector of PCs } (Y = P'(\underline{x} - \underline{\mu}))$$

$$P = [\underline{\ell}_1, \dots, \underline{\ell}_p] \quad \text{Note:}$$

$$\text{i.e. } \Sigma = P \Lambda P' \quad E[\underline{Y}] = P' \underline{\mu}$$

$$\Lambda = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_p \end{bmatrix} \quad (E[\underline{Y}] = P' \underline{0} = \underline{0})$$

$$\text{Cov}(\underline{Y}) = P' \Sigma P = P' P \Lambda P' P = \Lambda = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_p \end{bmatrix}$$

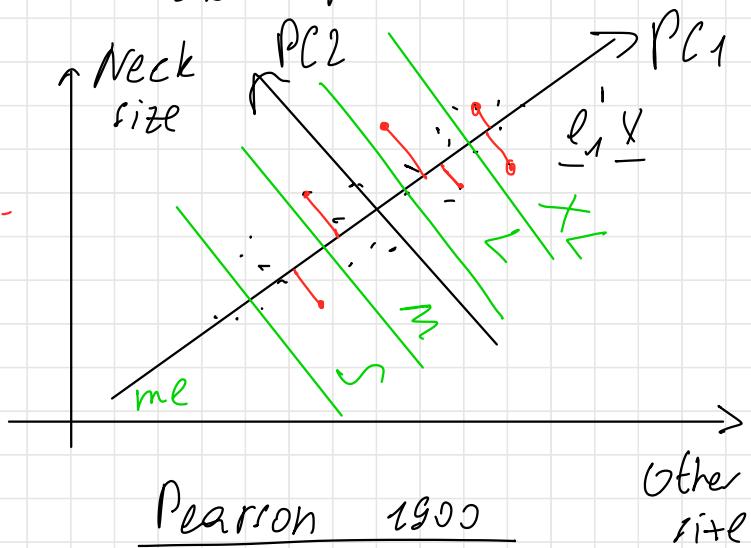
$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix} \quad - \text{may not have effect,}\\ \text{so } y_k \dots y_p \text{ too small}$$

Example

project measurements on

PC with

highest d_i



y_i i -th PC

$$y_i = \sum_{k=1}^n l_{ki} x_k = (\underline{l}_i \underline{x})$$

$$(y_i = \sum_{k=1}^n l_{ki} (x_k - \mu_k))$$

\underline{l}_i - vector of i -th

loadings

y_i - score

loadings - coefficients of the change-of-basis matrix P

Proposition

$$\text{Corr}(y_i, x_k) = \frac{\rho_{ki}}{\sqrt{d_i} \sqrt{\sigma_{kk}}}$$

std of x_k std of y_i

Proof

$$\text{Corr}(y_i, x_k) = \frac{\text{Cov}(y_i, x_k)}{\sqrt{d_i} \cdot \sqrt{\sigma_{kk}}} \quad (\Leftrightarrow)$$

$$\text{Cov}(y_i, x_k) = \text{Cov}(\underline{e}_i' \underline{x}, \underline{u}_k' \underline{x}) = \underline{e}_i' \sum \underline{u}_k =$$

$$= \underline{u}_k' \sum \underline{e}_i = d_i \underline{u}_k' \underline{e}_i = d_i \rho_{ki}$$

why need standartise before PCA

$$\underline{u}_k = \begin{pmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{pmatrix}$$

↑
k

$$(\Leftrightarrow) \frac{\rho_{ki} d_i}{\sqrt{d_i} \sqrt{\sigma_{kk}}} = \rho_{ki} \frac{\sqrt{d_i}}{\sqrt{\sigma_{kk}}}$$

if σ_{kk} not same \approx value for all $k \Rightarrow$

then we can not say that x_k have equal impact to y_i (because corr will be different for each y_i, x_k)

$$\underline{x} \sim \underline{\mu}, \Sigma$$

$$\underline{x}_{\text{std}} = \underline{\zeta} = V^{-\frac{1}{2}} (\underline{x} - \underline{\mu}) \quad V = \begin{bmatrix} \delta_{11} & & \\ & \ddots & \\ & & \delta_{nn} \end{bmatrix}$$

$$\text{Cov}(\underline{\zeta}) = V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}} = \rho = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & p_{ij} & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}$$

correlation

So first standardise and then apply PCA (PCA for ρ , not for Σ)

$$\rho = \sum_{i=1}^p d_i \underline{e}_i \underline{e}_i' \neq \Sigma \quad \text{spect. decomp}$$

~~\neq~~

$$= P' \Lambda P$$

also diff from Σ

$$\underline{y} = P' \underline{\zeta} = P' V^{-\frac{1}{2}} (\underline{x} - \underline{\mu})$$

$$E[\underline{y}] = \underline{0}$$

$$\text{Cov}(\underline{y}) = P' V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}} P = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_p \end{bmatrix}$$

d_i 's eigenvalues of ρ

Example

$$E = \begin{bmatrix} 1 & 1 \\ 1 & 100 \end{bmatrix}$$

$$\underline{\mu} = 0$$

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

but look at
unit of measure

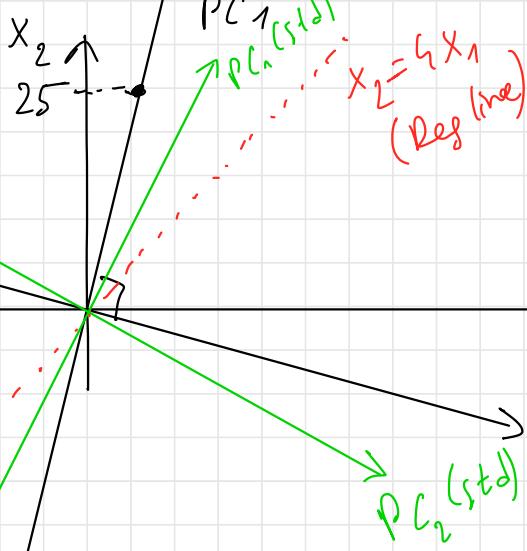
$$\Sigma = d_1 \underline{e}_1 \underline{e}_1^T + d_2 \underline{e}_2 \underline{e}_2^T$$

$$d_1 = 100, 16$$

$$\underline{e}_1 = (0, 0.955)'$$

$$d_2 = 0, 81$$

$$\underline{e}_2 = (0.555, -0, 0.33)'$$



$$PC_1: x_1 0.955 -$$

$$-0.05 x_2 = 0$$

$$x_2 = \frac{0.555}{0.05} x_1$$

$$PC_2: x_1 \cdot 0.05 - 0.555 x_2 = 0$$

Regression line

$$\rho = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

$$\frac{y - \mu_y}{\sqrt{\sigma_y^2}} = \rho \frac{x - \mu_x}{\sqrt{\sigma_x^2}} \quad - \text{Regression line}$$

Reg line: $\frac{x_2}{10} = 0,4 \frac{y_1}{1} \quad x_2 = 4x_1$

Now for standardised

$$\rho = d_1 e_1 e^T + d_2 e_2 e_2^T,$$

$$d_1 = 1 + 0,4 = 1,4$$

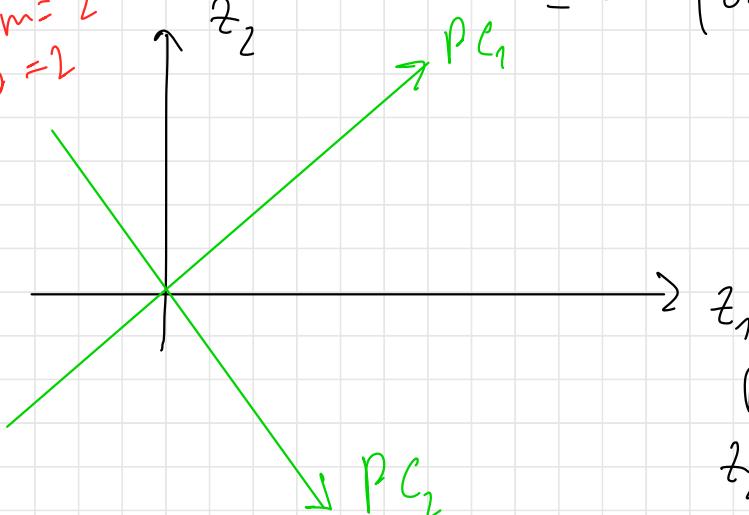
$$e_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)^T$$

$$(d_2 = 1 - 0,4 = 0,6)$$

$$e_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} \right)^T$$

$$\begin{aligned} \text{sum} &= 2 \\ \rho &= 2 \end{aligned}$$

$$z = x_{std} - \bar{x}$$



$\tilde{PC}_1:$

$$z_2 = z_1$$

$$\frac{x_1}{\sqrt{2}} = y_1 \quad x_2 = \sqrt{2}y_1$$

When one of the variance have huge value compared to others \Rightarrow PCA to ρ

PC of Σ

$$\Sigma = \sum_{i=1}^n d_i \underline{e}_i \underline{e}_i' = P \Lambda P' \quad Y = P' X$$

$$\text{Gen Var}(\underline{X}) = \prod_{i=1}^p d_i = \text{Gen Var}(Y)$$

$$\begin{aligned} \text{Total Var}(\underline{X}) &= \text{tr}(\Sigma) = \sum_{i=1}^p d_i = \text{tr}(A) = \\ &= \text{Total Var}(Y) \end{aligned}$$

PC - of P

$$P \Lambda P' = P = \sum_{i=1}^p d_i \underline{e}_i \underline{e}_i' \quad Y = P' Z$$

$$\text{Gen Var}(Z) = \text{Gen Var}(Y)$$

$$\text{Total Var}(Z) = \text{Total Var}(Y)$$

$$Z = V^{-1/2}(\underline{X} - \underline{\mu})$$

$$\text{Corr}(y_i, y_j) = 1$$

Total Var(Z) = tr(P) = p - ?

$$\text{Total Var}(Y) = P$$

$$\sum_{i=1}^n d_i = p \Rightarrow$$

$$J = \frac{\sum d_i}{p} = 1$$

But how to decide which components have to save

$$\text{Cov}(Y) = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & n \end{bmatrix}$$

$$Y = (y_1, y_2, \dots, y_n, y_{n+1}, \dots, y_p)$$

how to define k .

Total Variability explained by

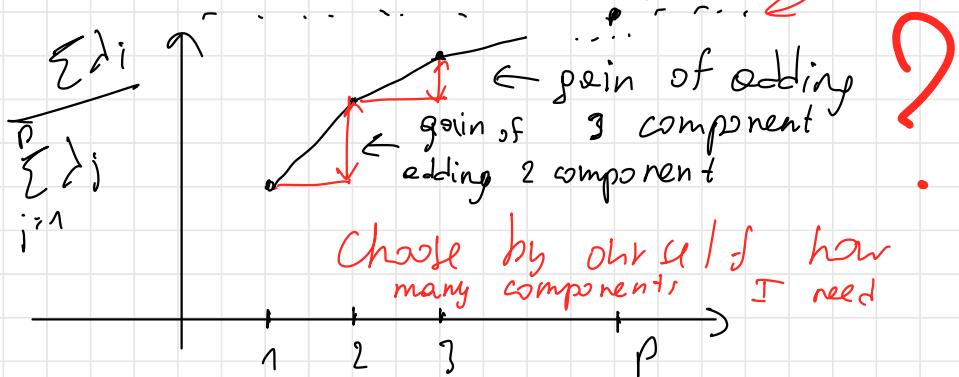
Cum. prop of total variab

$$y_1 \frac{d_1}{\sum d_i}$$

$$\frac{d_1}{\sum d_i}$$

$$y_2 \frac{d_2}{\sum d_i}$$

$$\frac{d_1 + d_2}{\sum d_i} \quad 100\%$$



Data

$$\mathbb{X} = \begin{bmatrix} x_1 & x_2 & \dots & x_p \\ \underline{x}_1 & \underline{x}_2 & \dots & \underline{x}_p \end{bmatrix}$$

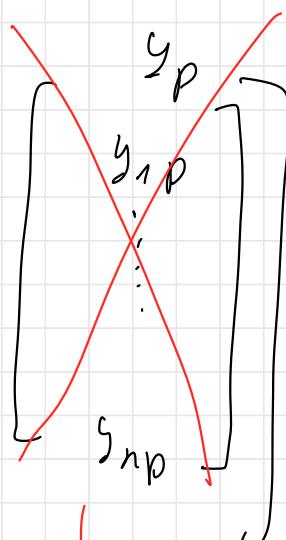
$$S = \sum_{i=1}^p \underline{x}_i \underline{x}_i' = P \Lambda P'$$



$$\rightarrow Y = P'(\underline{x} - \underline{\mu}) \Rightarrow \underline{x}_i \rightarrow P'(\underline{x}_i - \underline{\bar{x}})$$

New Data

$$\tilde{\mathbb{X}} = \begin{bmatrix} y_1 & y_2 & & & y_p \\ y_{11} & y_{12} & \cdots & y_{1k} & y_{1p} \\ y_{21} & y_{22} & & ; & y_{2p} \\ \vdots & & & \vdots & \vdots \\ y_{n1} & y_{n2} & & y_{nk} & y_{np} \end{bmatrix}_{n \times k}^{n \times k} \quad K_{P'}$$



removing from
PCA

Correspondence Analysis

PC's on Contingency tables
for categorical variables

Working
with
frequencies

		Edu				Gender
		1	2	3	4	
M	M	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1..}$
	F	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2..}$
		$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	n

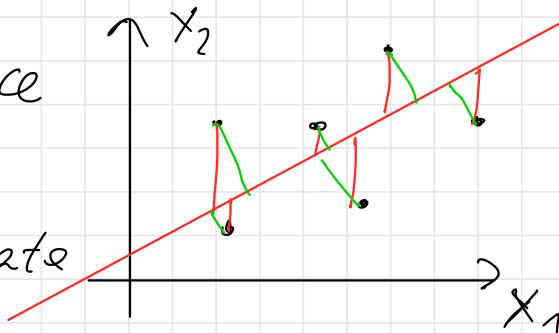
Also look to smallest eigen value
may be it's linearly dependent on
other variables.

Different perspective to PCA (Geometry)

Data: $\underline{x}_1, \dots, \underline{x}_n \in \mathbb{R}^p$ (row perspective)

? Find the linear space
of dimension k

closest to the data



for one -dimension it's not regression
line

But we need minimise projection
(It will be linear space defined by
 k highest principle components)

1.03.25 A different perspective

on PCA.

Let $\mathbb{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ $x_i \in \mathbb{R}^p$

? find linear space L of dimension $k < p$ closest to the data.

L is spanned by y_1, \dots, y_n - orthonormal basis.

$$|y_j| = 1$$

? find y_1, \dots, y_n s.t. $y_i \cdot y_j = 0$ if $i \neq j$

$$\sum_{i=1}^n \|(\underline{x}_i - \bar{\underline{x}}) - \underbrace{\sum_{j=1}^k y_j y_j^\top (\underline{x}_i - \bar{\underline{x}})}_{\text{projection on } L}\|^2 \rightarrow \min$$

projection on L

$$\underline{v}_i = \underline{x}_i - \bar{\underline{x}} \Rightarrow \sum_{i=1}^n \left\| \underline{v}_i - \underbrace{\sum_{j=1}^k y_j y_j^\top \underline{v}_i}_{(*)} \right\|^2 = (*)$$

$$(*) = \left\| \underline{v}_i - \sum_{j=1}^k y_j y_j^\top \underline{v}_i \right\|^2 = (\underline{v}_i - \sum_{j=1}^k y_j y_j^\top \underline{v}_i)^\top (\underline{v}_i - \sum_{j=1}^k y_j y_j^\top \underline{v}_i)$$

$$= \underline{v_i}^T \underline{v_i} - 2 \sum_{j=1}^k \underline{v_i}^T \underline{\eta_j} \underline{\eta_j}^T \underline{v_i} + \sum_{j=1}^k \sum_{t=1}^k \underline{v_i}^T \underline{\eta_j} \underline{\eta_j}^T \underline{\eta_t} \underline{\eta_t}^T \underline{v_i} =$$

$$\underline{\eta_j}^T \underline{\eta_t} = \begin{cases} 0 & j \neq t \\ 1 & j = t \end{cases}$$

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

$$= \underline{v_i}^T \underline{v_i} - \sum_{j=1}^k \underline{v_i}^T \underline{\eta_j} \underline{\eta_j}^T \underline{v_i}$$

$$(\star\star) = \sum_{i=1}^n \left(\underline{v_i}^T \underline{v_i} - \sum_{j=1}^k \underline{v_i}^T \underline{\eta_j} \underline{\eta_j}^T \underline{v_i} \right) =$$

$$= \underbrace{\sum_{i=1}^n \underline{v_i}^T \underline{v_i}}_{\text{does not depend on } \underline{\eta_j}} - \underbrace{\sum_{i=1}^n \sum_{j=1}^k \underline{v_i}^T \underline{\eta_j} \underline{\eta_j}^T \underline{v_i}}_{\rightarrow \min} \rightarrow \min$$

does not depend on $\underline{\eta_j} \Rightarrow \min \sim \underline{\max}$

? find $\underline{\eta_1}, \dots, \underline{\eta_n}$ orthonormal s.t.

$$\sum_{i=1}^n \sum_{j=1}^k \underline{v_i}^T \underline{\eta_j} \underline{\eta_j}^T \underline{v_i} = \sum_{i=1}^n \sum_{j=1}^k \underline{\eta_j}^T \underline{v_i} \underline{v_i}^T \underline{\eta_j} \rightarrow \max =$$

number number (some numbers)

$$= \sum_{j=1}^k \underline{\eta}_j^\top \left(\sum_{i=1}^n \underline{v_i} \underline{v_i}^\top \right) \underline{\eta}_j = \sum_{j=1}^k \underline{\eta}_j^\top \underbrace{\left(\sum_{i=1}^n (\underline{x_i} - \bar{\underline{x}}) / (\underline{x_i} - \bar{\underline{x}}) \right)}_{(h-1) S} \underline{\eta}_j$$

$\underline{v_i} = \underline{x_i} - \bar{\underline{x}}$

$$= (n-1) \sum_{j=1}^k \underline{\eta}_j^\top S \underline{\eta}_j$$

$$\max_{\underline{\eta}_1} (n-1) \underline{\eta}_1^\top S \underline{\eta}_1 = d_1 \quad \leftarrow \text{problem generated by PCA}$$

$$\|\underline{\eta}_1\| = 1$$

$$\arg\max \underline{\eta}_1 \quad \text{if} \quad S = \sum_{i=1}^n d_i \underline{e}_i \underline{e}_i^\top$$

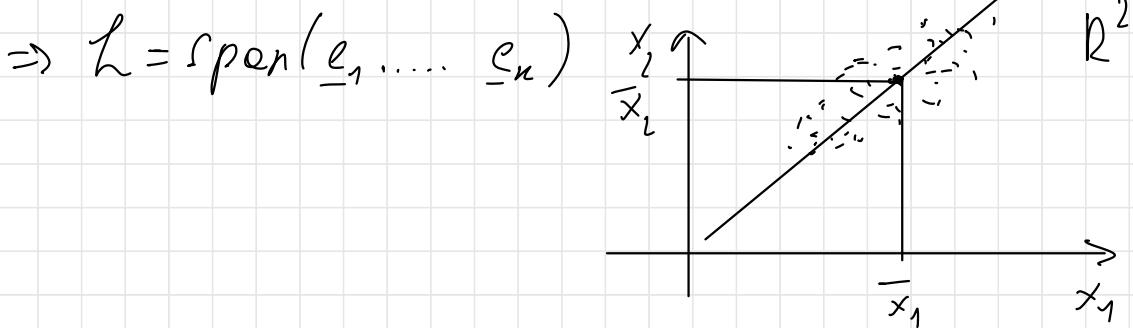
$\underline{\eta}_1$ just try

$$\max_{\underline{\eta}_1} (n-1) \underline{\eta}_1^\top S \underline{\eta}_1 = d_1$$

$$\underline{\eta}_1 \quad \|\underline{\eta}_1\| = 1$$

$$\arg\max: \underline{e}_1 \quad \text{if} \quad S = \sum_{i=1}^n \lambda_i \underline{e}_i \underline{e}_i^\top$$

by on \underline{k} : $\underline{y}_1 = \underline{\ell}_1 \dots \underline{y}_n = \underline{\ell}_n$ $\text{span}(\underline{\ell}_1)$



$$(*) \quad \sum_{i=1}^n \underline{v}_i^\top \underline{v}_i - \sum_{i=1}^n \sum_{j=1}^k \underline{v}_i^\top \underline{y}_j \underline{y}_j^\top \underline{v}_i =$$

$$= \sum_{i=1}^n \|(\underline{x}_i - \bar{\underline{x}}) - \sum_{j=1}^k \underline{v}_i^\top \underline{y}_j \underline{y}_j^\top \underline{v}_i\|^2$$

Approx. error

$$\sum_{i=1}^n \underline{v}_i^\top \underline{v}_i - (n \cdot 1) \sum_{j=1}^k \underline{\ell}_j^\top \underline{\ell}_j = \sum_{j=1}^k \underline{v}_i^\top \underline{v}_i - \sum_{j=1}^k d_j$$

$$\begin{aligned}
 & - (n-1) \sum_{j=1}^k d_j = \quad \text{?} \\
 & = (n-1) \sum_{j=1}^p d_j - (n-1) \sum_{j=p+1}^k d_j = \\
 & = (n-1) \sum_{j=k+1}^p d_j \quad \text{?} \\
 \\
 & \sum_{i=1}^n v_i^\top v_i \in \mathbb{R} \\
 & \operatorname{tr} \left(\sum_{i=1}^n v_i^\top v_i \right) = \\
 & = \operatorname{tr} \left(\sum_{i=1}^n (\underline{x}_i - \bar{\underline{x}})^\top (\underline{x}_i - \bar{\underline{x}}) \right) \\
 \\
 & \operatorname{tr}(ABC) = \operatorname{tr}(CAB) = \\
 & = \operatorname{tr}(B CA) \\
 \\
 & \text{Approx error} \\
 & \text{Also can be} \\
 & \text{chosen to} \\
 & \text{Evaluate number} \\
 & \text{of chosen } k
 \end{aligned}$$

$$\underline{x}_1, \dots, \underline{x}_n \in \mathbb{R}^D$$

closest lin space L of $\dim k$

$$L = \text{span}(\underline{e}_1, \dots, \underline{e}_k)$$

Note that there are an infinite number of basis of $\dim k$ generating L

$$? \quad \underline{w}_1, \dots, \underline{w}_k \quad \text{s.t.} \quad \text{span}(\underline{w}_1, \dots, \underline{w}_k) =$$

$$= L = \text{span}(\underline{e}_1, \dots, \underline{e}_k)$$

ordered according

for instance $\underline{w}_1, \dots, \underline{w}_n$ variability

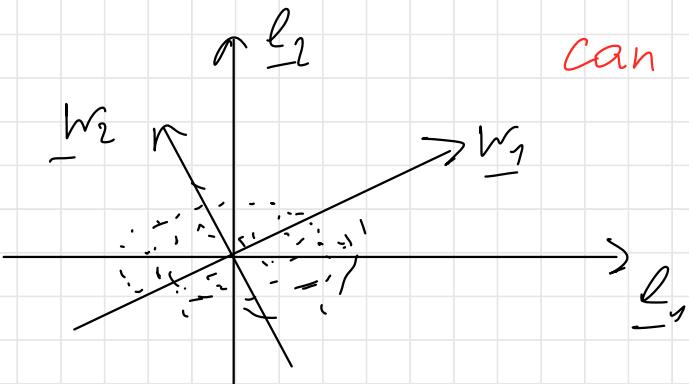
such that $\underline{w} = [\underline{w}_1 \dots \underline{w}_n]$ is sparse

(many zeros)

here we lose order

of variability. $w_j > w_i$ for $j > i$

can not guarantee



VARIMAX ?
for what

RMK \mathbb{X} $n \times p$ data
 SVD of centered data $\left(I - \frac{\mathbb{1}\mathbb{1}^T}{\mathbb{1}^T\mathbb{1}} \right) \mathbb{X} =$ centered data

$$= UDV \quad \left\{ \begin{array}{l} U - n \times n \text{ unitary matrix} \\ U^T U = U U^T = I_n \end{array} \right.$$

$$S = \frac{1}{n-1} \mathbb{X}^T \left(I - \frac{\mathbb{1}\mathbb{1}^T}{\mathbb{1}^T\mathbb{1}} \right) \mathbb{X} \quad \left\{ \begin{array}{l} V - p \times p \text{ matrix} \\ V^T V = V V^T = I_p \end{array} \right.$$

$$= \frac{1}{n-1} V^T D^T U^T UDV \quad \left\{ D = \begin{bmatrix} d_1 & & & \\ & d_2 & & 0 \\ & & \ddots & \\ 0 & \dots & & d_p \\ 0 & & & 0 \end{bmatrix} \quad d_i \geq 0 \right.$$

$$D^T D = \begin{bmatrix} d_1^2 & & \\ & \ddots & \\ & & d_p^2 \end{bmatrix} \quad p \times p \quad d_i = \frac{d_i^2}{n-1}$$

$$\hat{=} \sum_{i=1}^p d_i \underline{v}_i \underline{v}_i^T \quad V = [\underline{v}_1 \dots \underline{v}_p] \quad \underline{v}_i = \underline{e}_i$$

Connection between PCA and SVD of (centered) data frame

Gaussian Dist

$\mathbb{R}^P \ni \underline{x} \sim \mathcal{N}_p(\underline{\mu}, \Sigma)$ where $\underline{\mu} \in \mathbb{R}^P$
 $\Sigma \text{ pos def matrix}$

Rmk If $\underline{x} \sim \mathcal{N}_p(\underline{\mu}, \Sigma) \Rightarrow E[\underline{x}] = \underline{\mu}$ $\text{cov}(\underline{x}) = \Sigma$

$$f(\underline{x}) = \frac{1}{(2\pi)^P \text{Det}(\Sigma)} \exp \left[-\frac{1}{2} (\underline{x} - \underline{\mu})^\top \Sigma^{-1} (\underline{x} - \underline{\mu}) \right]$$

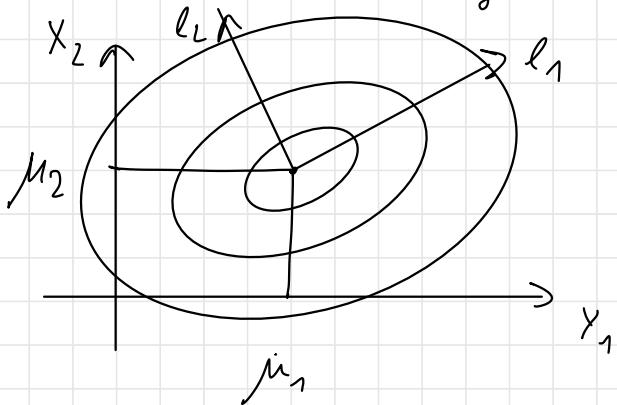
$\underline{x} \in \mathbb{R}^P$

Contours are identified by Mahalanobis distance

$$(\underline{x} - \underline{\mu})^\top \Sigma^{-1} (\underline{x} - \underline{\mu}) = \text{const}$$

↪ ellipse with

$$\text{axis of eigenvectors of } \Sigma = \sum_i^P d_i \underline{e}_i \underline{e}_i^\top$$



Mahalanobis distance

$$f \propto \exp \left[-\frac{1}{2} \underline{d}^\top \Sigma^{-1} (\underline{x} - \underline{\mu}) \right]$$

Ex $p = 1$

$$\mu = \mu \in \mathbb{R}$$

$$\Sigma = [\sigma_{11}]$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2}(x-\mu)\sigma_1^{-1}(x_1)\right]$$

$$= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2} \frac{(x-\mu)^2}{\sigma_1^2}\right] N(\mu, \sigma_1^2)$$

Theo)

$$\underline{x} \sim N_p(\underline{\mu}, \Sigma) \Leftrightarrow$$

$$\underline{Q}' \underline{x} \sim N_1(Q^T \underline{\mu}, Q^T \Sigma Q) \quad \forall \underline{Q} \in \mathbb{R}^{p \times 1}$$

Proof

use the characteristic fct of N_p

Coro $\underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$ and $\underline{x} \sim N_p(\underline{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix})$

(~~X~~)?

$$\Rightarrow \forall i = 1, \dots, p \quad x_i \sim N_1(\mu_i, \sigma_{ii}^2)$$

Proof for $i = 1, \dots, p$ let $\underline{u}_i = (0 \dots 0 \underset{i}{1} 0 \dots 0)^T$

$$\underline{u_i}^T \underline{x} = x_i \xrightarrow{\text{Then}} \mathcal{N}_1(\underline{u_i}^T \underline{x}, \underline{u_i}^T \Sigma \underline{u_i}) =$$

$$= \mathcal{N}_1(\mu_i, \sigma_{ii}^2) \quad (\text{MVR} \Rightarrow \text{each comp Normal.})$$

but each comp Normal $\not\Rightarrow$ MVR

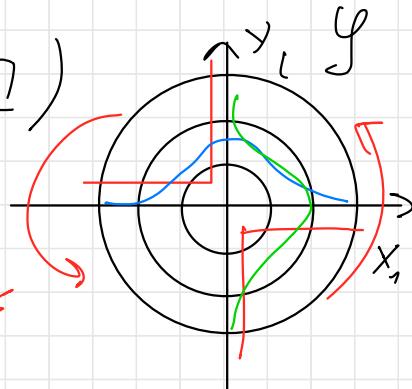
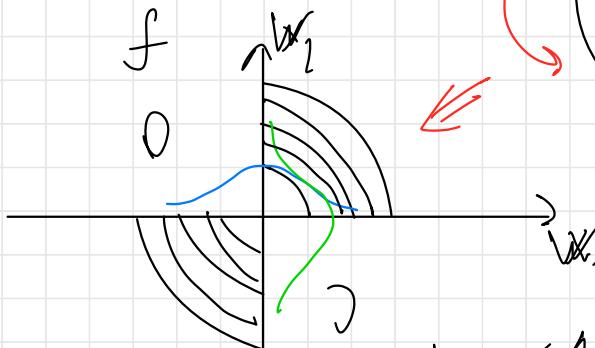
(If one of the component not Gaussian \Rightarrow whole \underline{x} not Gaussian)

Ex $p=2$

$$\underline{x} \sim \mathcal{N}_2(\underline{0}, \Sigma)$$

$$x_1 \sim \mathcal{N}(0, 1)$$

$$x_2 \sim \mathcal{N}(0, 1)$$



but $w \not\sim \mathcal{N}_2$

$$w_1 \sim \mathcal{N}(0, 1)$$

$$w_2 \sim \mathcal{N}(0, 1)$$

we saved marginal but changed
copula.

Look at PCA and try to save Gaussian

Components will be good if highest components will have Gaussian Dist

linear transform of Gaussian dist \rightarrow Gauss dist

Prop $\underline{X} \sim N_p(\mu, \Sigma)$, $A \in \mathbb{R}^{q \times p} \Rightarrow$

$A\underline{X} \sim N_q(A\mu, A\Sigma A^T)$

\mathbb{R}^q

If we need show that something is Gaussian - we have to show that any linear transformation is Gaussian

Don't try linear transformation to get Gaussian from source Dist

(Because it doesn't give you Gaussian)

Proof Need to show: $\forall \underline{Q} \in \mathbb{R}^{q \times q}$

$\underline{Q}'(A\underline{X}) \sim N_1(\underline{Q}'(A\mu), \underline{Q}'(A\Sigma A')\underline{Q})$

$\underline{Q}'(A\underline{X}) = \underline{Q}'AX = (A'\underline{Q})^T \underline{X} \stackrel{\text{Def}}{\sim} \mathbb{R}^p$

$\sim N_1((A'\underline{Q})'\mu, (A'\underline{Q})'\Sigma(A'\underline{Q}))$

$N_1(\underline{Q}'(A\mu), \underline{Q}'(A\Sigma A')\underline{Q}) \stackrel{\text{Def } A\underline{X}}{\sim}$

$$\sim N(\underline{\mu}, \underline{\Sigma})$$

Prop $\underline{x} \sim N_p(\underline{\mu}, \underline{\Sigma}), \underline{d} \in \mathbb{R}^P \Rightarrow$

$$\underline{x} + \underline{d} \sim N_p(\underline{\mu} + \underline{d}, \underline{\Sigma})$$

Proof (exercice) $\sim N_1(\underline{\alpha}^\top \underline{d}, 0)$

$$\forall \underline{\alpha} \in \mathbb{R}^P \quad \underline{\alpha}^\top (\underline{x} + \underline{d}) = \underline{\alpha}^\top \underline{x} + \underline{\alpha}^\top \underline{d} =$$

$$\sim N_1(\underline{\alpha}^\top \underline{\mu}, \underline{\alpha}^\top \underline{\Sigma} \underline{\alpha})$$

$$\Rightarrow N_1(\underline{\alpha}^\top (\underline{\mu} + \underline{d}), \underline{\alpha}^\top \underline{\Sigma} \underline{\alpha})$$

Rmk

$$\underline{z}_1, \underline{z}_2, \dots, \underline{z}_p \quad i.i.d \sim N_p(\underline{0}, \underline{I})$$

$$\begin{aligned} \underline{z} &= \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix} \quad g_p(\underline{z}) = \prod_{i=1}^p g_1(z_i) = \\ &= \prod_{i=1}^p \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z_i^2} = \frac{1}{\sqrt{(2\pi)^p}} e^{-\frac{1}{2} \sum_{i=1}^p z_i^2} \end{aligned}$$

$$= \frac{1}{\sqrt{(2\pi)^p}} e^{-\frac{1}{2} \underline{z}' \underline{I} \underline{z}} \Rightarrow \underline{z} \sim N_p(\underline{0}, \underline{I})$$

Consider $\underline{\mu} \in \mathbb{R}^p$, Σ $p \times p$ posit def

$$\Sigma = \sum_{i=1}^p d_i \underline{e}_i \underline{e}_i'$$

$$d_1 \geq d_2 \geq \dots \geq d_p > 0$$

$$\Sigma^{1/2} = \sum_{i=1}^p \sqrt{d_i} \underline{e}_i \underline{e}_i'$$

$$\underline{x} = \sum \underline{z} + \underline{\mu}$$

$$\sum^{\frac{1}{2}} \underline{z} \sim N_p (\underline{0}, \Sigma^{\frac{1}{2}} \underline{I} \underline{I}^{\frac{1}{2}}) = N(\underline{0}, \Sigma)$$

$$\sum^{\frac{1}{2}} \underline{z} \perp \underline{\mu} \sim N_p (\underline{\mu}, \Sigma)$$

By linear transformation we can get any form of Gaussian Distribution

$$\underline{x} \sim N_p (\underline{\mu}, \Sigma)$$

$$\Sigma^{-\frac{1}{2}} = \sum_{i=1}^p \frac{1}{\sqrt{d_i}} \underline{e}_i \underline{e}_i^\top$$

$$\Sigma^{-\frac{1}{2}} (\underline{x} - \underline{\mu}) \sim N(\underline{0}, \underline{I})$$

Rmk $\underline{x} \sim N_p (\underline{\mu}, \Sigma)$

$$IR \rightarrow w = (\underline{x} - \underline{\mu})' \Sigma^{-1} (\underline{x} - \underline{\mu}) = d \Sigma^{-1} (\underline{x}, \underline{\mu})$$

$$w = (\underline{x} - \underline{\mu})' \Sigma^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} (\underline{x} - \underline{\mu}) = \underline{z}' \underline{z} =$$

$$= \sum_{i=1}^p z_i^2 \quad \underline{z} \sim N_p (\underline{0}, \underline{I})$$

$$z_1, \dots, z_p \text{ i.i.d } \sim N(0, 1)$$

$$z_1 \sim N(0,1) \Rightarrow z_1^2 \sim \chi^2_1(1)$$

$$\sum_{i=1}^p z_i^2 \sim \chi^2(p) \Rightarrow$$

$$W \sim \chi^2(p)$$

$$P\left[\sum_{i=1}^p (\underline{x}_i - \mu_i)^2 \leq \chi^2_{1-\alpha}(p)\right] = 1-\alpha$$

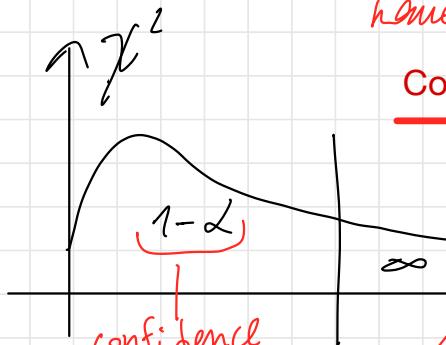
Berzettwerte, 290

same percentage confidence regions
critical values $\chi^2_{1-\alpha}$

$$\mathcal{L} f(0,1)$$

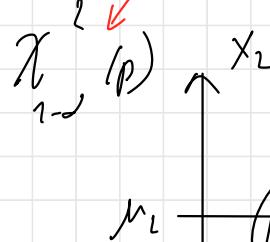
data to be have in this area

$$\sum_{i=1}^p (\underline{x}_i - \mu_i)^2 = z^2$$



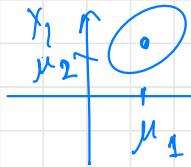
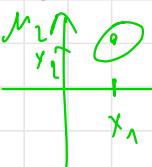
quantile of order
 $\chi^2_{1-\alpha}$

$$\mathcal{L} \chi^2_{1-\alpha}(p)$$



Whellly we have \underline{x}

and we need find μ



mean have
to be in this
area



$$\underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \sim N_p \left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}, \Sigma \right)$$

mixed covariance

$\Sigma = \begin{pmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{pmatrix}$

mixed Covariance

Covariance for X_1

Covariance for X_2

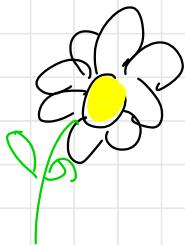
Prop

$$\underline{X} \sim N_p(\underline{\mu}, \Sigma)$$

$$X_1 \sim N_p(\mu_1, \Sigma_{11})$$

prof $X_n = A\underline{X}$, where $A = [I_{q \times q} \quad 0]$
 (A + home)

6.03.25. Lecture



$$\mathbb{R}^p \ni \underline{x} = \begin{pmatrix} \underline{x}_1 \\ \underline{x}_2 \end{pmatrix} \stackrel{q}{\sim} \stackrel{p-q}{\sim}$$

$$\sim N_p \left(\begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} p \times (p-q) \\ \Sigma_{21} / (p-q) \times q & \Sigma_{22} / (p-q) \times (p-q) \end{pmatrix} \right)$$

$$\Rightarrow \underline{x}_1 \sim N_q \left(\underline{\mu}_1, \Sigma_{11} \right)$$

Prop

$$\underline{x} = \begin{pmatrix} \underline{x}_1 \\ \underline{x}_2 \end{pmatrix}, \underline{x}_1 \perp\!\!\!\perp \underline{x}_2 \quad (\Leftrightarrow \Sigma_{12} = \Sigma_{21}^T = [0])$$

$$\left(\begin{pmatrix} \underline{x}_1 \\ \underline{x}_2 \end{pmatrix} \sim N_p \left(\begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \right) \right)$$

! Cov = 0 in Gaussian world means independence

proof show that, $\forall \underline{t} = \begin{pmatrix} \underline{t}_1 \\ \underline{t}_2 \end{pmatrix} \in \mathbb{R}^p$

$$\varphi_{\underline{x}}(\underline{t}) = \varphi_{x_1}(t_1) \cdot \varphi_{x_2}(t_2). \quad (\text{At home})$$

Tes

$$\left(\frac{x_1}{x_2} \right) \sim N_p \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

$$x_1 | x_2 = \underline{x}_2 \sim N_p \left(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (\underline{x}_2 - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \right)$$

Rmk. $x_1 \sim N_p(\mu_1, \Sigma_{11})$ with no information about \underline{x}_2 of \underline{x}_1

$x_1 | x_2 = \underline{x}_1$ how we solve uncertainty

when we lose information of x_2 , which correlates with x_1

$$x_1 | x_2 = \underline{x}_1 \sim N_p \left(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (\underline{x}_1 - \mu_2), \underbrace{\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}}_{\text{decreased uncertainty}} \right)$$

regression function decreased uncertainty
 $E[x_1 | x_2 = \underline{x}_1]$ (doesn't depend on Σ_{11})

in Gaussian world using linear function
 is exact answer (not approximation)

if $\underline{x}_1 \perp\!\!\!\perp \underline{x}_2 \Rightarrow \Sigma_{12} = 0$

$$\underline{x}_1 | \underline{x}_2 = \underline{x}_2 \sim N_2(\mu, \Sigma_{11})$$

Proof. Let

$$A = \begin{pmatrix} I_{q \times q} & -\Sigma_{12} \Sigma_{22}^{-1} \\ 0_{(p-q) \times q} & I_{(p-q) \times (p-q)} \end{pmatrix}$$

$$A \begin{pmatrix} \underline{x}_1 - \underline{\mu}_1 \\ \underline{x}_2 - \underline{\mu}_2 \end{pmatrix} \sim N_p \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, A \Sigma A^T \right)$$

$$A \Sigma A^T = \begin{pmatrix} \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \quad \text{with } \Sigma_{22} \neq 0$$

$$A \begin{pmatrix} \underline{x}_1 - \underline{\mu}_1 \\ \underline{x}_2 - \underline{\mu}_2 \end{pmatrix} = \begin{pmatrix} \underline{x}_1 - \underline{\mu}_1 - \Sigma_{12} \Sigma_{22}^{-1} (\underline{x}_2 - \underline{\mu}_2) \\ \underline{x}_2 - \underline{\mu}_2 \end{pmatrix}$$

$$\text{Hence } \Rightarrow \underline{x}_1 - \underline{\mu}_1 - \Sigma_{12} \Sigma_{22}^{-1} (\underline{x}_2 - \underline{\mu}_2) \perp\!\!\!\perp \underline{x}_2 - \underline{\mu}_2$$

$$\underline{x}_1 - \underline{\mu}_1 - \Sigma_{12} \Sigma_{22}^{-1} (\underline{x}_2 - \underline{\mu}_2) \sim N_p (0, \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})$$

$$\underline{x}_1 - \underline{\mu}_1 - \Sigma_{11} \Sigma_{22}^{-1} (\underline{x}_2 - \underline{\mu}_2) \mid \underline{x}_2 = \underline{x}_2 \sim N_p (0, \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})$$

Thus

$$\underline{x}_1 - \underline{\mu}_1 - \Sigma_{11} \Sigma_{22}^{-1} (\underline{x}_1 - \underline{\mu}_1) \mid \underline{x}_1 = \underline{x}_2 \sim N_p (0, \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})$$

$$\underline{x}_1 \mid \underline{x}_2 = \underline{x}_2 \sim N_p (\underline{\mu} + \Sigma_{12} \Sigma_{22}^{-1} (\underline{x}_2 - \underline{\mu}_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}) \quad \checkmark$$

Ex $p=2$ $\underline{x} = \begin{pmatrix} Y \\ X \end{pmatrix} \leftarrow \begin{matrix} \underline{x}_1 \\ \underline{x}_2 \end{matrix} \sim$

$$\sim \left(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} \right)$$

$$Y \sim N_1(\mu_y, \sigma_{yy}^2)$$

$$Y | X=x \sim N_1\left(\mu_y + \frac{\rho_{xy}}{\sqrt{\sigma_{xx}}} (x - \mu_x), \sigma_{yy}^2 - \frac{\rho_{xy}^2}{\sigma_{xx}} \sigma_{yy}^2\right)$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx} \cdot \sigma_{yy}}}$$

$$\frac{\sigma_{xy}}{\sqrt{\sigma_{xx}} \sqrt{\sigma_{yy}}}$$

$$\begin{aligned} & \sqrt{\sigma_{yy}^2} \\ & \sqrt{\sigma_{xx}} \end{aligned}$$

$$Y | X=x \sim N_1\left(\mu_y + \rho_{xy} \frac{\sqrt{\sigma_{yy}}}{\sqrt{\sigma_{xx}}} (x - \mu_x), \sigma_{yy}^2 (1 - \rho^2)\right)$$

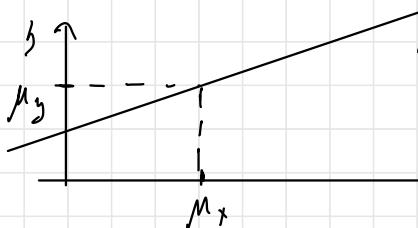
$$E[Y | X=x]$$

?

$$E[Y | X=x] = \mu_y + \rho_{xy} \frac{\sqrt{\sigma_{yy}}}{\sqrt{\sigma_{xx}}} (x - \mu_x) \Rightarrow$$

In Gaussian world, the regression function i.e. the best we can do if we want to make prediction Minimizing MSE is a linear function!

$$y = \mu_y + \rho_{xy} \frac{\sqrt{\sigma_{yy}}}{\sqrt{\sigma_{xx}}} (x - \mu_x)$$



$$y = \beta_0 + \beta_1 x \text{ simple linear regression}$$

$$\beta_1 = \rho_{xy} \frac{\sqrt{\sigma_{yy}}}{\sqrt{\sigma_{xx}}}$$

line

$$\beta_0 = \mu_y - \beta_1 \mu_x$$

standardised value for y is simply standardised value for x multiplied by correlation between x and y

$$\frac{y - \mu_y}{\sqrt{\sigma_{yy}}} = \rho_{xy} \frac{x - \mu_x}{\sqrt{\sigma_{xx}}}$$

so knowing x and correlation with y
we may find y

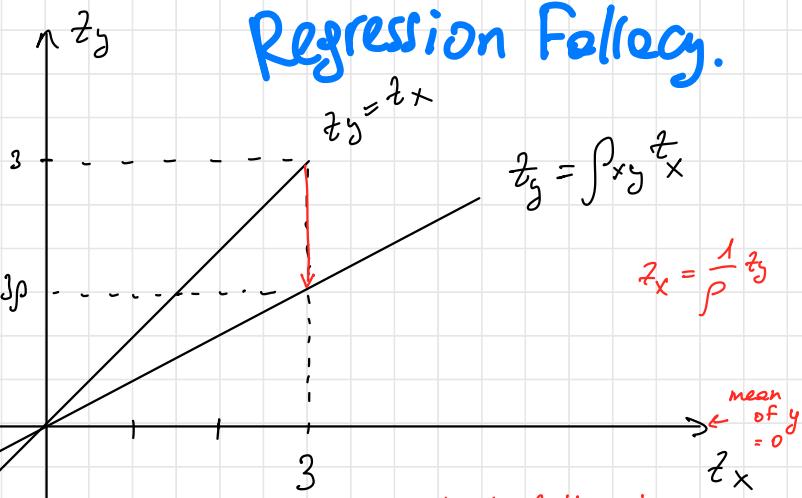
$$\begin{aligned} z_x &= \frac{x - \mu_x}{\sqrt{\sigma_{xx}}} \\ z_y &= \frac{y - \mu_y}{\sqrt{\sigma_{yy}}} \end{aligned}$$

standardised coordinates

x - height of father 3p

y - height of sun

Regression Fallacy.



lets take observation far from μ_x , as example $3\sigma_{xx}$

(but we are in standardised coordinates \Rightarrow just 3) \Rightarrow

height of sun = 3p

$x = \mu_x + 3\sigma_{xx} \Rightarrow y = \mu_y + 3p\sqrt{\sigma_{yy}}$. So father was exceptional, but son is not so exceptional, because $|p| < 1$.

Regression toward mean - Regression effect. BUT, if

I interpret regression effect - it will be **Regression fallacy**.

Example. Trying to predict my grade in calculus 2, using my grade in calculus 1.
For Calc 1 - 30, for Calc 2 predicting -28 (^{Predicted} something not so exceptional)

Try to interpret - if had high grade \Rightarrow will not study ans get less, if had 18 \Rightarrow will get more because will study more - **Regression fallacy**
 But it's all just because it regression to the mean, our outliers try to predict closer to mean value

$$\underline{X} = (x_1, x_2 \dots x_p)^T \sim N_p(\underline{\mu}, \Sigma)$$

$\underline{\mu}, \Sigma$ unknown

Training data: $\mathcal{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}, x_i \in \mathbb{R}^p$

Assumption:

\underline{x} : is realization of \underline{X}_i

$$x_1, \dots, x_n \text{ iid } \sim N_p(\underline{\mu}, \Sigma) \quad (\text{iid } \sim \underline{X})$$

Estimators for $\underline{\mu}$ and Σ :

$$\underline{\bar{x}}_n = \frac{1}{n} \sum_{i=1}^n \underline{x}_i \quad \text{for } \underline{\mu} \quad \leftarrow \text{unbiased}$$

$$\underline{S} = \frac{1}{n-1} \sum_{i=1}^n (\underline{x}_i - \underline{\bar{x}})(\underline{x}_i - \underline{\bar{x}})^T \quad \text{for } \Sigma$$

Maximum Likelihood Estimator (MLE)

i.e. $x_i \sim \text{Bernoulli}(p) \quad x_1, \dots, x_5 \text{ i.i.d. } \sim \text{Bern}(p)$

$$P[X_1=1] = p$$

Data: $X_1=1, X_2=3, X_3=0, X_4=1, X_5=0$

$$P(X_1=1, X_2=3, X_3=0, X_4=1, X_5=0) = p(1-p)(1-p)p(1-p)$$

$$= p^2(1-p)^3$$

$$L(p | \text{data}) = p^2(1-p)^3 \quad L : [0, 1] \rightarrow [0, \infty)$$

$$\ell(p | \text{data}) = \log L(p) = 2\log p + 3\log(1-p)$$

$$\hat{p} = \underset{p \in [0, 1]}{\operatorname{argmax}} L(p) = \frac{1}{5} = \frac{1}{5} \sum_{i=1}^5 x_i$$

$$L(\theta | \text{data}) = P(\text{data} | \theta) \quad \leftarrow \text{estimators}$$

Theorem $\underset{(\mu, \Sigma) \in \mathbb{R}^p, \Sigma \text{ is } p \times p \text{ positive def. matrix}}{\operatorname{argmax}} L(\mu, \Sigma | \underbrace{X_1=x_1, \dots, X_n=x_n}_{\text{i.i.d.}}) = (\hat{\mu}, \hat{\Sigma})$, where

$$L(\mu, \Sigma | \underbrace{X_1=x_1, \dots, X_n=x_n}_{\text{i.i.d.}}) = \prod_{i=1}^n \frac{1}{\sqrt{(2\pi)^p \det \Sigma}} \exp\left(-\frac{1}{2} (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu)\right) \Rightarrow$$

$$\left\{ P(x_1, \dots, x_n | \mu, \Sigma) = P(x_1 | \mu, \Sigma) P(x_2 | \mu, \Sigma) \dots P(x_n | \mu, \Sigma) \right\}$$

\Rightarrow best estimators: $\hat{\mu} = \bar{x}$ sample mean, and $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top = S$ - biased estimator of Σ . $\frac{n}{n-1} S$ - unbiased. So by MLE we get biased estimator of Σ

MLE: we don't know if the estimator is right on average (i.e. unbiased) but it's the best we can do today: we maximise the likelihood of having observed what we have observed!

Invariance property of MLE: suppose

$\underline{\theta} \in \mathbb{R}^K$ is a parameter, for instance: $\underline{\theta} = (\mu, \Sigma)$ and suppose that $\hat{\underline{\theta}} = \hat{\underline{\theta}}(\text{data})$ is the MLE estimator of $\underline{\theta}$ if $h: \mathbb{R}^K \rightarrow \mathbb{R}^J$ is mapping, then $h(\hat{\underline{\theta}}) = \hat{h}(\hat{\underline{\theta}})$

Example. We know that $\hat{\Sigma} = \frac{n-1}{n} \sum$ is the MLE for Σ .

Now if $X_i \stackrel{iid}{\sim} N_p(\mu, \Sigma) \quad \forall i=1, \dots, n$.

How to find estima for σ_d^2 or d_1 , the first eigen value of Σ ?

$$\hat{\Sigma} = \sum_{i=1}^n \hat{\delta}_i \hat{\delta}_i^T \quad \text{from that can find } \hat{d}_1$$

$$\hat{h}(\theta) = h(\hat{\theta})$$

MLE
of $h(\theta)$

[some
property
in another
form]

MLE
depending on
ML approach
more robust
than parametric
models
i.e. more robust
to outliers
and non-
normality

Back to the Gaussian

$\underline{\mu}, \Sigma$

$$L(\underline{\mu}, \Sigma | \text{data}) = \prod_{i=1}^n \frac{1}{\sqrt{(2\pi)^D \text{Det}(\Sigma)}} e^{-\frac{1}{2}(\underline{x}_i - \underline{\mu})^\top \Sigma^{-1} (\underline{x}_i - \underline{\mu})}$$

$$\ell(\underline{\mu}, \Sigma) = \sum_{i=1}^n \log \left(\frac{1}{\sqrt{(2\pi)^D \text{Det}(\Sigma)}} \right) - n \log \sqrt{\text{Det}(\Sigma)} -$$

$$-\frac{1}{2} \sum_{i=1}^n (\underline{x}_i - \underline{\mu})^\top \Sigma^{-1} (\underline{x}_i - \underline{\mu})$$

Motivation diff

$$(\underline{\mu}, \Sigma) = \arg \max \ell(\underline{\mu}, \Sigma)$$

$$\underline{\mu} \in \mathbb{R}^D$$

Σ $p \times p$ pos def matrix

Jhonson
Weier prof

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}_n \text{ unbiased}$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\underline{x}_i - \hat{\mu}) (\underline{x}_i - \hat{\mu})^\top = \frac{n-1}{n} S$$

not $n-1$ (biased)

(asymptotically biased)

MLE's

Distr'n of \bar{x}_n and S

where $\underline{x}_1, \dots, \underline{x}_n$ i.i.d $\sim d_p(\mu, \Sigma)$

Distr' of \bar{x}_n

Prop $\bar{x}_n \sim d_p\left(\mu, \frac{1}{n}\Sigma\right)$

proof $\tilde{\underline{x}} = \begin{pmatrix} \underline{x}_1 \\ \underline{x}_2 \\ \vdots \\ \underline{x}_n \end{pmatrix} \in \mathbb{R}^{np}$

$$\underline{X} \sim N_{np} \left(\begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \\ \vdots \\ \underline{\mu}_n \end{pmatrix}, \begin{pmatrix} \Sigma & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \Sigma \end{pmatrix} \right)$$

$$A = \underbrace{\begin{bmatrix} I_{p \times p} & I_{p \times p} & \cdots & I_{p \times p} \end{bmatrix}}_{n\text{-times}} \quad p \times np$$

$$\frac{1}{n} A \widehat{\underline{X}} = \widehat{\underline{x}}_n \sim N_p \left(\underline{\mu}, \frac{1}{n} \Sigma \right)$$

↑
↑

Prop $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, $A \in \mathbb{R}^{q \times p} \Rightarrow$

$$A \underline{X} \sim N_q(A\underline{\mu}, A \Sigma A^T)$$

\mathbb{R}^q

If we need show that something is Gaussian - we have to show that any linear transformation is Gaussian

10. D3. 25

$$\underline{x}_1, \dots, \underline{x}_n \text{ i.i.d. } \sim N_p(\underline{\mu}, \Sigma)$$

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

$$\underline{\Sigma} = \frac{1}{n-1} \sum_{i=1}^n (\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})^T$$

Def $\underline{z}_1, \dots, \underline{z}_m$ i.i.d. $\sim N_p(\underline{0}, \Sigma)$

$$p \times p \quad \sum_i^m \underline{z}_i \underline{z}_i^T \sim \text{Wishart}(\Sigma, m) \quad \begin{matrix} \text{matrix} \\ 1927 \\ \text{Wishart} \end{matrix}$$

Properties

Note, that \underline{z}_i can be interpreted as co-variance matrices! (distribution of matrix)

$$1. \quad A_1 \sim \text{Wishart}(\Sigma, m_1) \xrightarrow{\underline{H}} \Rightarrow A_2 \sim \text{Wishart}(\Sigma, m_2)$$

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

Proof $A_1 = \sum_{i=1}^{m_1} \underline{z}_{1i} \underline{z}_{1i}^T \sim \underline{z}_{11}, \dots, \underline{z}_{1m_1} \text{ i.i.d. } \sim N_p(\underline{0}, \Sigma)$

$$A_1 = \sum_{j=1}^{m_1} \underline{z}_{2j} \underline{z}_{2j}^T \quad \underline{z}_{2j} \dots \underline{z}_{2m_2} \text{ i.i.d. } \sim N_p(0, \Sigma)$$

Now take them in order

$$\underline{z}_1 \dots \underline{z}_{m_1} \underline{z}_{21} \dots \underline{z}_{2m_2}, \text{ re-name them, maintaining the order} \Rightarrow \underline{w}_1 \dots \underline{w}_{m_1} \underline{w}_{m_1+1} \dots \underline{w}_{m_1+m_2}$$

since every thing is independent $\rightarrow \underline{w}_1 \dots \underline{w}_{m_1+m_2} \sim \text{i.i.d. } N_p(0, \Sigma) \Rightarrow A_1 + A_2 = \sum_{i=1}^{m_1} \underline{z}_{2i} \underline{z}_{2i}^T + \sum_{i=1}^{m_2} \underline{z}_{2i} \underline{z}_{2i}^T = \sum_{i=1}^{m_1+m_2} \underline{w}_i \underline{w}_i^T \sim \text{Wish}(\Sigma, m_1+m_2)$

2. C k x p matrix of const
 $\sim \text{Wish}(\Sigma, m)$

$$\Rightarrow CAC^T \sim \text{Wish}(C\Sigma C^T, m)$$

proof $A = \sum_{i=1}^m \underline{z}_i \underline{z}_i^T \quad \underline{z}_1 \dots \underline{z}_m \sim \text{i.i.d. } N(0, \Sigma)$

$$CAC^T = \sum_i^m C \underline{z}_i \underline{z}_i^T C^T$$

$$C \underline{z}_i \sim d_{\nu_k}(0, C\Sigma C^T) \Rightarrow$$

$$CAC^T \sim \text{Wish}(C\Sigma C^T, m)$$

3. let $\sigma^2 > 0$, $A \sim \text{Wish}(\Sigma, m) \Rightarrow$
 $\sigma^2 A \sim \text{Wish}(\sigma^2 \Sigma, m)$

proof $A = \sum_{i=1}^m z_i z_i^\top$ z_1, \dots, z_m i.i.d. $\sim N_p(0, \Sigma)$

$$\sigma^2 A = \sum_{i=1}^m \sigma^2 z_i z_i^\top \neq$$

$\sigma^2 z_i \sim N_p(0, \sigma^2 \Sigma) \Rightarrow \sigma^2 A \sim \text{Wishart}(\sigma^2 \Sigma, m)$

4. $A \sim \text{Wish}(\Sigma, m)$ $\Sigma = [\sigma^2] \Rightarrow A \sim \sigma^2 \chi^2_m$

by def $A = \sum_{i=1}^m z_i z_i^\top$ $z_i \sim i.i.d N_p(0, \sigma^2)$
 $= \sum_{i=1}^m z_i^2$

The Wishart is a
multivariate extension of
 χ^2

$\frac{1}{\sigma^2} z_1, \dots, \frac{1}{\sigma^2} z_m$ i.i.d. $\sim N_p(0, 1)$

$\frac{1}{\sigma^2} A = \sum_{i=1}^m \frac{z_i^2}{\sigma^2} \sim \chi^2(m)$

$X \sim N(0, 1) \Leftrightarrow \frac{X}{\sigma} \sim N(0, 1)$

look to
that
means
 $\frac{X}{\sigma}$
we multiply and not
variable distribution

so multiplying density we don't change density, we change X .

Remark $\underline{C} \in \mathbb{R}^P$, $A \sim \text{Wish}(\Sigma, m)$
 $P \times P$

$$\underline{?} \quad \underline{C}^T A \underline{C} \sim ?$$

$$\underline{C}^T A \underline{C} \stackrel{\textcircled{1}}{\sim} \text{Wish}(\underline{C}^T \Sigma \underline{C}, m)$$

$$\underline{C}^T \Sigma \underline{C} > 0 \Rightarrow$$

without proof!

$$\underline{C}^T A \underline{C} \sim (\underline{C}^T \Sigma \underline{C}) \chi^2(m) \quad \text{i.e. } \frac{\underline{C}^T A \underline{C}}{\underline{C}^T \Sigma \underline{C}} \sim \chi^2(m)$$

Theo $\underline{x}_1, \dots, \underline{x}_n$ i.i.d $\sim N_p(\mu, \Sigma)$

$$\Rightarrow \sum_i^n (\underline{x}_i - \bar{\underline{x}}) (\underline{x}_i - \bar{\underline{x}})^T \sim \text{Wishart}(\Sigma, n-1)$$

lost one degree of freedom

(corollary
variance)
using theorem
and Property 3

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

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

Rmk . if μ is known \Rightarrow

$$\sum_{i=1}^n \underbrace{(\underline{x}_i - \underline{\mu})}_{\mathcal{T}_i} \underbrace{(\underline{x}_i - \underline{\mu})^\top}_{\mathcal{T}_i^\top} \sim \text{Wish}(\Sigma, n)$$

so if I have μ use this because it have more degree of freedom

Then $\underline{x}_1, \dots, \underline{x}_n$ i.i.d $\sim N_p(\underline{\mu}, \Sigma) \Rightarrow$

$$\bar{x} \perp \mathcal{L}$$

stochastic independent (no dependence)

Summing up

Then $\underline{x}_1, \dots, \underline{x}_n$ i.i.d $\sim N_p(\underline{\mu}, \Sigma)$

1. $\bar{x} \sim N_p(\underline{\mu}, \frac{1}{n} \Sigma)$

2. $(n-1) \mathcal{S} \sim \text{Wish}(\Sigma, n-1)$

3. $\bar{x} \perp \mathcal{L}$ - means that they stochastically independent.

Theorem \bar{X} and S are sufficient statistics: if the data is generated by Gaussian distribution, then all we need to know, to do statistics, is \bar{X} and S .

Note: We can transform data and make it more Gaussian! Some useful transforms:

- Suppose the data has an empirical density with long tail similar to $\chi^2 \Rightarrow$ we can take a logarithm transformation: it just compresses the first part and extend the second part. (Note - that logarithm needs to be centred at suitable point!)

We can also do Box-Cox transformations!

- If x are proportions, so they belong in $[0, 1]$, we can take a sigmoid transformation. Example are: $\log\left(\frac{x}{1-x}\right)$ and $\arctan(x)$

$\underbrace{\text{Long Large numbers (154)}}$ $\xrightarrow{\text{LLN RP}} \underline{x}_1, \underline{x}_2, \dots, \underline{x}_n \dots \sim \overset{\text{i.i.d}}{\mathcal{N}(\mu, \Sigma)}$
 $\underline{\bar{X}_n} \xrightarrow{P} \mu \text{ as } n \rightarrow \infty$ distribution
 (only)

$$\left(P(|\underline{\bar{X}_n} - \mu| > \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty \right)$$

$$\forall \varepsilon > 0$$

$$\text{Also } P(|S - \Sigma| > \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty$$

$$S = \frac{1}{n-1} \sum_{i=1}^n (\underline{x}_i - \underline{\bar{X}})(\underline{x}_i - \underline{\bar{X}})^T$$

CLT

$\mathbb{R} \ni x_1, \dots, x_n \text{ i.i.d } \sim \mu, \sigma^2$

$$\sqrt{n} \frac{\underline{\bar{X}} - \mu}{\sigma} \xrightarrow{n \rightarrow \infty} \mathcal{N}(0, 1)$$

so getting
large sample
with i.i.d $\sim \mu, \sigma^2$
we will get
distribution of
 $\underline{\bar{X}}$

$$\underline{\bar{X}} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

asymptotically normal

$$\text{in } \mathbb{R}^p \quad \sqrt{n}(\underline{\bar{X}} - \mu) \sim \mathcal{N}_p(0, \Sigma)$$

For practical purposes: if n is large $\Rightarrow \underline{\bar{X}_n} \sim \mathcal{N}_p(\mu, \frac{\Sigma}{n})$

Note: If sample is large everything is Gaussian?

NO. If the sample is large, then the sample mean has a distribution which can be approximated by a Gaussian.

Therefore 1 billion coin tosses remain coin tosses: the sample mean of billion Bernoulli has a distribution that is a Gaussian.

We want to use data, which is a partial information about the true population, to infer something about the true population.

Curse of dimensionality: as p gets larger we need larger and larger amounts of data, indeed n must grow exponentially fast with respect to p .

Inference for μ

when n - large
we don't assume
specificity of X

Large sample size (n very large)

P
 $\mathbf{R} \rightarrow \underline{x}_1, \dots, \underline{x}_n \text{ iid } \sim \underline{\mu}, \Sigma \text{ det}(\Sigma) > 0$
assume that it
from CLT $\sqrt{n}(\bar{x} - \mu) \sim N(0, \Sigma)$ Gaussian
(not Asymptotically)

$$(\sqrt{n}(\bar{x} - \mu))^T \Sigma^{-1} (\sqrt{n}(\bar{x} - \mu)) \sim \chi^2(p)$$

$$n(\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu) \sim \chi^2(p) \Rightarrow$$

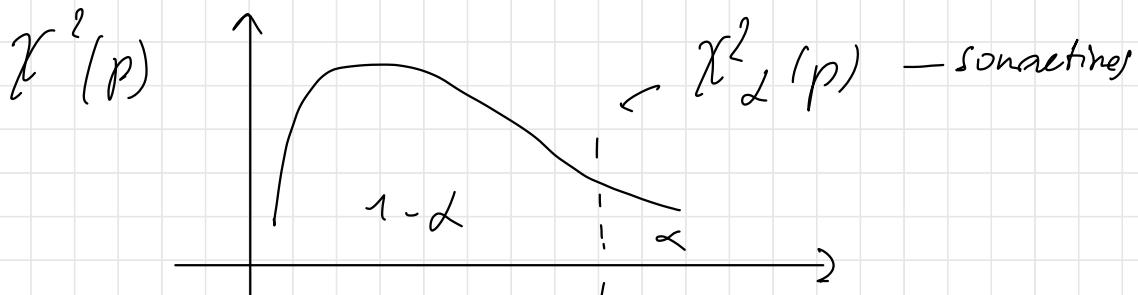
$$\int \rightarrow \sum \quad \begin{cases} \text{(we don't know } \Sigma, \text{ but since)} \\ n \text{ is large } \Rightarrow \text{by LLN} \end{cases}$$

$$n(\bar{x}_n - \mu)^T \Sigma^{-1} (\bar{x}_n - \mu) \sim \chi^2(p)$$

pivotal statistics

It acts as a pivot around which we build up the inferential procedure. Note that for a random variable to be a pivotal statistic we need know its distribution without knowing the value of μ (which unknown, since we want estimate it)

$F_{T \times L}(\cdot, 1)$



$\alpha = 0.05 \Rightarrow p(\mu_n(\mu_0) = 0.05)$
 wrong reject H_0 if χ^2 of distribu-
 $1 - 0.05 = 0.95$ tion lie on this
 line lie on this

$\chi^2_{1-\alpha}(p)$

$1-\alpha$ quantile of χ^2

$$(*) P\left[n(\bar{x}_n - \mu)^T S^{-1} (\bar{x}_n - \mu) \leq \chi^2_{1-\alpha}(p)\right] = 1-\alpha$$

$$P\left[d\left(\frac{1}{n}S\right)^{-1}, (\bar{x}_n, \mu) \leq \chi^2_{1-\alpha}(p)\right] = 1-\alpha$$

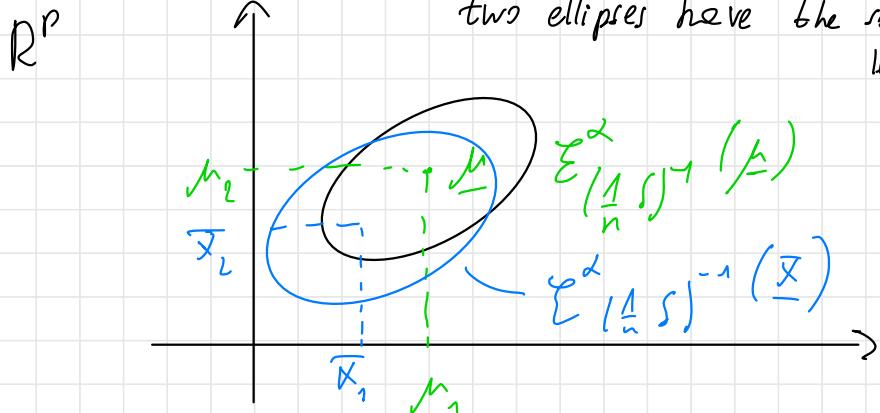
ellipse with
center in μ
and radius
depending on α

? why divide
Minkowski distance

$$\mathcal{E}^\alpha(\mu) = \left\{ \underline{x} \in \mathbb{R}^p : d_{\left(\frac{1}{n}S\right)^{-1}}(\underline{x}, \mu) \leq \chi^2_{1-\alpha}(p) \right\}$$

\approx ellipse

$$\mathcal{E}^\alpha(\bar{x}) = \left\{ \underline{\gamma} \in \mathbb{R}^p : d_{\left(\frac{1}{n}S\right)^{-1}}(\underline{\gamma}, \bar{x}) \leq \chi^2_{1-\alpha}(p) \right\}$$



two ellipses have the same axes, same lengths of semi-axes but a different centre.

$$\underline{x} \in \mathcal{E}_{\left(\frac{1}{n}s\right)^{-1}}(\underline{\mu}) \Leftrightarrow \underline{\mu} \in \mathcal{E}_{\left(\frac{1}{n}s\right)^{-1}}(\underline{x})$$

distance is less than δ \Rightarrow $\delta = 1 - \alpha$

$$P_{\substack{\text{Random} \\ \underline{x}}} [\underline{x} \in \mathcal{E}_{\left(\frac{1}{n}s\right)^{-1}}(\underline{\mu})] = 1 - \alpha$$

\underline{x} is random and the ellipse is given, but we don't know it since we don't know $\underline{\mu}$

$$P_{\substack{\text{Random} \\ \underline{\mu}}} [\underline{\mu} \in \mathcal{E}_{\left(\frac{1}{n}s\right)^{-1}}(\underline{x})] = 1 - \alpha$$

$\underline{\mu}$ is not known but its given while ellipse is random / randomly generated once we have observed the data). Moreover we have that $1/\alpha$ times the random ellipse will cover $\underline{\mu}$

Def Confidence Region for $\underline{\mu}$ at level

$$1 - \alpha, \alpha \in (0, 1)$$

$$CR_{1-\alpha}(\underline{\mu}) = \mathcal{E}_{\left(\frac{1}{n}s\right)^{-1}}(\underline{x}) =$$

$$= \left\{ \underline{\gamma} \in \mathbb{R}^P : n(\underline{\bar{x}} - \underline{\gamma})^\top S^{-1} (\underline{\bar{x}} - \underline{\gamma}) \leq \chi_{1-\alpha}^2(p) \right\}$$

data $\rightarrow \underline{\bar{x}}, S \rightarrow CR_{1-\alpha}(p) \in \mathbb{R}^P$

We have expression for the confidence region around the point estimate for mean: this also quantifies how unclear we are about this value.

notation
The more ellipse is dense shrinked around its center the less is uncertain (nunzugehörig). The uncertainty is given by $\det(S)$ as it represents the volume of ellipse.

Testing

$$H_0: \underline{\mu} = \underline{\mu}_0 \in \mathbb{R}^P \text{ vs } H_1: \underline{\mu} \neq \underline{\mu}_0$$

default to prove
that H_0 is
false

it's always trying

to reject H_0

Assume H_0 is true

This is not
of Shapiro
test

$$T_0^2 = n (\underline{\bar{x}}_n - \underline{\mu}_0)^\top S^{-1} (\underline{\bar{x}}_n - \underline{\mu}_0) \sim \chi^2(p)$$

$Z \in (3, 1)$ How far we are from sample mean

Fix now α level $\alpha \in (0, 1)$ (usually small):
we want to use the data to decide among
two hypothesis

- we can reject H_0 when its true: this is called **Type I Error**. With α we can control the ^{type I error}
- we can reject H_1 when its true: this called **Type II Error**)

Making error α percent of time.

in our case we proof against H_0 when \bar{X} is very far from $\mu_0 \Rightarrow$ we reject H_0
if $T_0^2 > \chi_{\alpha}^2(p)$ (so we only make an error α percent of time)

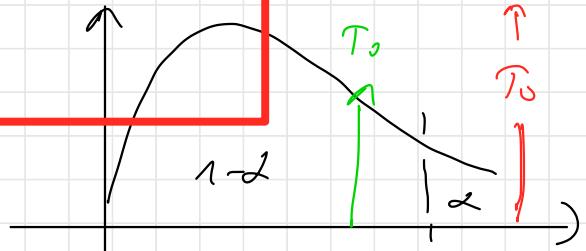
If we see today something that happens 1 time out of 10000 times: either the assumption is wrong or (H_0 is wrong)
we observing miracle

$$d \left(\frac{1}{n} s \right)^{-1} (\bar{X}, \bar{\mu}_0)$$

Reject at level α if

$$T_0^2 > \chi_{1-\alpha}^2(p)$$

not reject Reject



$$\chi_{1-\alpha}^2(p)$$

Reject H_0 - discovering

Rmk 1



$$\chi_{1-\alpha}^2(p)$$

Decision rule

Reject H_0 if p-value is small

Reasoning by contradiction

It's not probability of H_0 //
being true.

Rmk 2

$$T_0^2 > \chi_{1-\alpha}^2(p) \Leftrightarrow \mu_0 \notin \left(\sum_{i=1}^n (\bar{x}_i) \right)$$

Reject

$$\left(CR_{1-\alpha}(\mu) \right)$$

$$p\text{-value} \leq \alpha \Leftrightarrow T_0^2 > \chi_{1-\alpha}^2(p) \Rightarrow$$

- If p-value very small \Rightarrow can reject H_0 .
- If p-value very large \Rightarrow can not reject H_0 .

11.03.25

SnedecorWhat if n is smallRecall

$$P = \frac{Y/n}{W/m}$$

$$? \quad n(\bar{x}_n - \mu_n)^T S^{-1}(\bar{x}_n - \mu_n) \sim ?$$

Assume $\underline{x}_1, \dots, \underline{x}_n$ i.i.d $\sim N(\mu, \Sigma)$

where
 $y \sim \chi^2(n) \rightarrow Y$
 $w \sim \chi^2(m) \rightarrow$

$$F \sim F(n, m)$$

Rmt's

$$1. \quad t = \frac{\bar{z}}{\sqrt{\frac{w}{m}}} \quad \begin{array}{l} \bar{z} \sim N(0, 1) \\ w \sim \chi^2(m) \end{array} \Rightarrow$$

Student $t(m-1)$

$$t^2 = \frac{\bar{z}^2}{\frac{w}{m}} \quad \begin{array}{l} \bar{z}^2 \sim \chi^2(1) \\ w \sim \chi^2(m) \end{array} \Rightarrow t^2 \sim f(1, m)$$

$$2. \quad F(n, m) \xrightarrow[m \rightarrow \infty]{} \frac{1}{n} \chi^2(n)$$

Fisher

proof
 $w \sim \chi^2(m)$

$$w = \sum_{i=1}^n z_i^2 \quad z_i \sim N(0, 1) \quad \text{independent}$$

$$z_1^2 \sim \chi^2(1)$$

Fisher

$$\frac{W}{m} = \frac{1}{m} \sum_{i=1}^m z_i^2 \xrightarrow[m \rightarrow \infty]{LLN} E[z_1^2] = 1$$

Variance of z_i

3. Hotelling's Th (1931)

$$\underline{X} \sim N_p(\underline{\mu}, \underline{\Sigma})$$

$$W \sim \text{Wishart}\left(\frac{1}{m} \underline{\Sigma}, m\right) \Rightarrow \frac{m-p+1}{m-p} (\underline{X} - \underline{\mu})^T W^{-1} (\underline{X} - \underline{\mu}) \sim F(p, m-p+1)$$

$\underline{X} \perp\!\!\!\perp W$

Coro $\underline{x}_1, \dots, \underline{x}_n$ i.i.d $\sim N_p(\underline{\mu}, \underline{\Sigma}) \Rightarrow$

$$n(\underline{\bar{x}}_n - \underline{\mu})^T S^{-1} (\underline{\bar{x}}_n - \underline{\mu}) \sim \frac{(n-1)p}{(n-p)} F(p, n-p)$$

Proof $\underline{\bar{x}}_n \sim N_p(\underline{\mu}, \frac{1}{n} \underline{\Sigma}) \Rightarrow \sqrt{n}(\underline{\bar{x}}_n - \underline{\mu}) \sim N_p(0, \underline{\Sigma})$

$\perp\!\!\!\perp$ to $(n-1)S \sim \text{Wishart}(\underline{\Sigma}, n-1)$

by Hotelling's theorem \Rightarrow

$$n(\underline{\bar{x}}_n - \underline{\mu})^T S^{-1} (\underline{\bar{x}}_n - \underline{\mu}) \sim \frac{(n-1)p}{n-p} \xrightarrow{\text{R.T. Statistic}} F(p, n-p)$$

pivotal

Confidence Region for μ

$$T^2 = n(\bar{X} - \mu)^T S^{-1} (\bar{X} - \mu) -$$

Hotelling's T^2
statistic

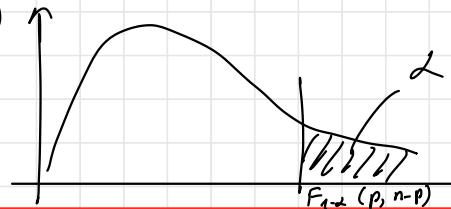
$$\lambda \in (0,1)$$

$$\mathbb{P} \left[n(\bar{X}_n - \mu)^T S^{-1} (\bar{X}_n - \mu) \leq \frac{(n-1)p}{n-p} F_{1-\alpha}(p, n-p) \right] = 1 - \lambda$$

$$d \left(\frac{1}{n} S \right)^{-1} (\bar{X}, \mu)$$

$$CR_{1-\alpha}(\mu) = \left\{ \gamma \in \mathbb{R}^p : n(\bar{X} - \gamma)^T S^{-1} (\bar{X} - \gamma) \leq \frac{(n-1)p}{n-p} F_{1-\alpha}(p, n-p) \right\}$$

different radius from previous



where we had large n

works when we assume gaussianity

yesterday's works even without gaussian

$$\xrightarrow{n \rightarrow \infty} \frac{(n-1)p}{n-p} F_{1-\alpha}(p, n-p) \xrightarrow{n \rightarrow \infty} p \cdot \frac{1}{p} \chi^2_{1-\alpha}(p) = \chi^2_{1-\alpha}(p)$$

- like previously

Test

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

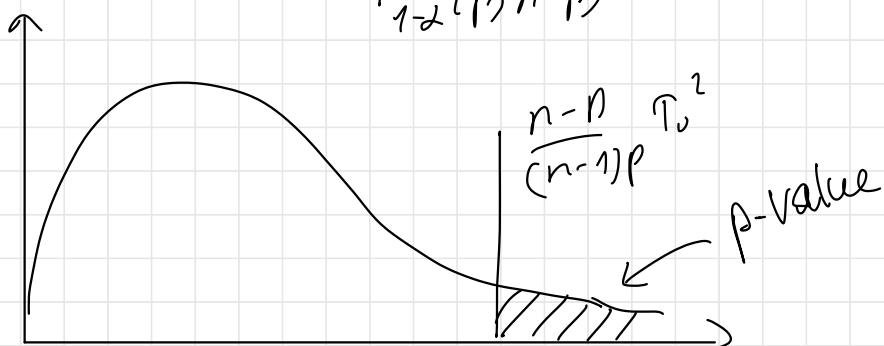
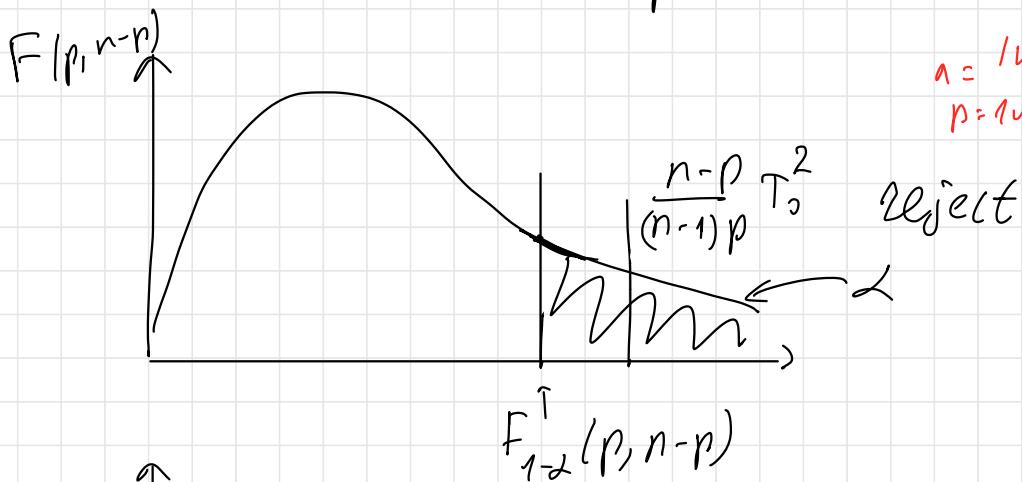
$$T_0^2 = n(\bar{X}_n - \mu_0)^2 \delta^{-1} (\bar{X}_n - \mu_0)$$

If H_0 is true $\Rightarrow T_0^2 \sim \frac{(n-1)p}{n-p} F(p, n-p)$

Given $\lambda \in (0, 1)$

Reject: if $T_0^2 > \frac{(n-1)p}{n-p} F_{1-\lambda}(p, n-p)$

$$\begin{aligned} \alpha &= 1\% \\ p &= 1\% \end{aligned}$$



$$\text{Rmk } CR_{1-\alpha}(\underline{\mu}) = \left\{ \underline{y} \in \mathbb{R}^p : n(\underline{\bar{x}} - \underline{y})^\top S^{-1}(\underline{\bar{x}} - \underline{y}) \leq \right.$$

$$\left. \leq \frac{(n-1)p}{n-p} F_{1-\alpha}(p, n-p) \right\}$$

$$T_0^2 > \frac{(n-1)p}{n-p} F_{1-\alpha}(p, n-p) \Leftrightarrow \underline{\mu}_0 \notin CR_{1-\alpha}(\underline{\mu})$$

$CR_{1-\alpha}(\underline{\mu})$ identifies the $\underline{\mu}$ s for which you cannot reject $H_0: \underline{\mu} = \underline{\mu}_0$ at level α

Ex $\underline{X}_1, \dots, \underline{X}_{10}$ i.i.d $\sim N_2(\underline{\mu}, \Sigma)$

$$n=10, p=2$$

$$\underline{X} = \underline{0}, S = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$F_{0.9}(2, 8) = 9.11$$

$$\alpha = 0.1$$

$$CR_{1-\alpha}(\underline{\mu}) = \left\{ \underline{y} \in \mathbb{R}^p : 10(\underline{\bar{x}} - \underline{y})^\top S^{-1}(\underline{\bar{x}} - \underline{y}) \leq \frac{9.11}{8} \right\}$$

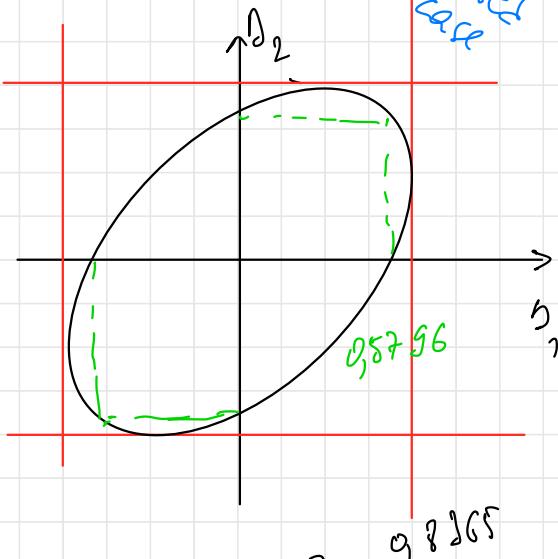
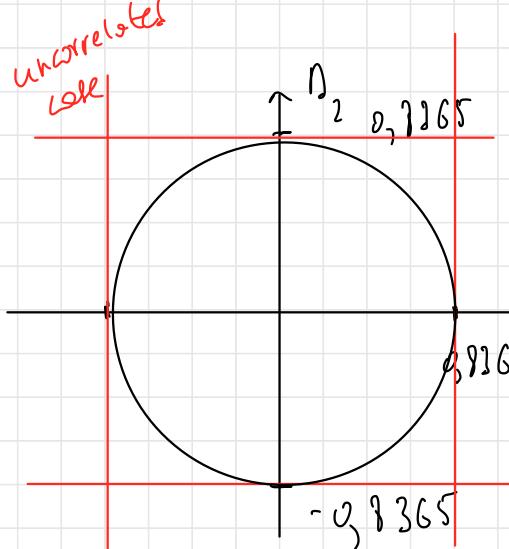
$$\cdot F_{0.9}(2, 8) \} = \left\{ \underline{y} \in \mathbb{R}^p : 10(\underline{\bar{y}}_1^2 + \underline{\bar{y}}_2^2) \leq 6.552 \right\} =$$

This means that the random ellipse produced this circle, but we can not say if $\underline{\mu}$ is within this circle with probability 0.9! We can say that with probability 0.9 the above algorithm produces random ellipses, which includes $\underline{\mu}$

$$= \left\{ \boldsymbol{\gamma} \in \mathbb{R}^2 : \boldsymbol{\gamma}_1^2 + \boldsymbol{\gamma}_2^2 \leq 0.6557 \right\}$$

$$\boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{pmatrix}$$

Correlated case



Assume now $\bar{\mathbf{x}} = \mathbf{0}$ $S = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

$$CR_{1-2}(\mu_1) = \left\{ \boldsymbol{\gamma} \in \mathbb{R}^2 : \boldsymbol{\gamma}^\top S^{-1} \boldsymbol{\gamma} \leq 0.6557 \right\}$$

$$S^{-1} = \begin{pmatrix} \frac{1}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{1}{3} \end{pmatrix}$$

$$CR_{1-2}(\mu_1) = \left\{ \boldsymbol{\gamma} \in \mathbb{R}^2 : \boldsymbol{\gamma}_1^2 - \frac{2}{3}\boldsymbol{\gamma}_1\boldsymbol{\gamma}_2 + \boldsymbol{\gamma}_2^2 \leq \text{quantile } \leq \frac{3}{4}(0.6557) \right\}$$

$$CI_{1-2}(\mu_1) = \left[\bar{x}_1 \pm t(0.95) \sqrt{\frac{s_{11}}{10}} \right] = [-0.5796]$$

$\alpha = 0.1$

t-test for univariate corl

$$CI_{1-2}(\mu_1) = \left[\pm 0.5796 \right]$$

||
0.9

tools of Stat 101 can't \Leftarrow
work in multi-variate setting!

But now we didn't use
the co-variance to compute
these intervals so the two
above don't change if we
change variance!

They are the same for the
two different cases of S
that presented above

We can not say: $IC(\mu) = IC(\mu_1) \times IC(\mu_2)$

$P[\mu_1 \in CI_{0.9}(\mu_1), \mu_2 \in CI_{0.9}(\mu_2)] = 0.9^2 < 0.9$ (if $X_1 \perp X_2$)
cartesian product can not generate a region that has same confidence

Given $\underline{\alpha} \in \mathbb{R}^P$

i.e. each α_j satisfy $CI = 0.9$, no
no intersection $CI(\mu_1) \cap CI(\mu_2)$ gaussian diff.
Since $0.95 \times 0.95 > 0.9$, i.e. can intersect differ.

? $CI_{1-2}(\underline{\alpha}^\top \underline{\mu})$

Ex $\underline{\alpha} = (0 - 0 \ 1 \ 0 - 0) \quad \underline{\alpha} = (\overset{i}{0} \dots \overset{i}{0} \ 1 \overset{j}{0} \dots \overset{j}{0} - 1 \overset{i}{0} \dots \overset{i}{0})$

:

$CI_{1-2}(\mu_i)$

$CI_{1-2}(\mu_i - \mu_j)$

$\bar{\underline{x}} \sim N_p(\underline{\mu}, \frac{1}{n} \Sigma)$

$\underline{\alpha}' \bar{\underline{x}}$ unbiased est for $\underline{\alpha}' \underline{\mu}$

$\bar{\underline{x}} \sim N_1(\underline{\alpha}' \underline{\mu}, \frac{1}{n} \underline{\alpha}' \Sigma \underline{\alpha})$

$$\frac{\underline{\alpha}' \bar{x} - \underline{\alpha}' \mu}{\sqrt{\underline{\alpha}' \Sigma \underline{\alpha}}} \sim \mathcal{N}(0, 1)$$

$(n-1) \underline{\alpha}' \Sigma \underline{\alpha} \sim \text{Wish}(\underline{\alpha}' \Sigma \underline{\alpha}, n-1)$

$$(n-1) \underline{\alpha}' \Sigma \underline{\alpha} \sim \text{Wish}(\underline{\alpha}' \Sigma \underline{\alpha}, n-1) = (\underline{\alpha}' \Sigma \underline{\alpha}) \chi^2_{(n-1)}$$

$$\frac{(n-1) \underline{\alpha}' \Sigma \underline{\alpha}}{\underline{\alpha}' \Sigma \underline{\alpha}} \sim \chi^2_{(n-1)}$$

$$\frac{\underline{\alpha}' \bar{x} - \underline{\alpha}' \mu}{\sqrt{\underline{\alpha}' \Sigma \underline{\alpha}}} \sim \mathcal{N}(0, 1) \quad \text{Student}$$

$$\frac{\sqrt{(n-1) \frac{\underline{\alpha}' \Sigma \underline{\alpha}}{\underline{\alpha}' \Sigma \underline{\alpha}} - \frac{1}{n-1}}}{\sqrt{\frac{1}{n-1} \chi^2_{(n-1)}}} \sim t(n-1) \quad \downarrow$$

$$\frac{\underline{\alpha}' \bar{x} - \underline{\alpha}' \mu}{\sqrt{\underline{\alpha}' \Sigma \underline{\alpha}}} \sim t(n-1) \quad \text{Student}$$

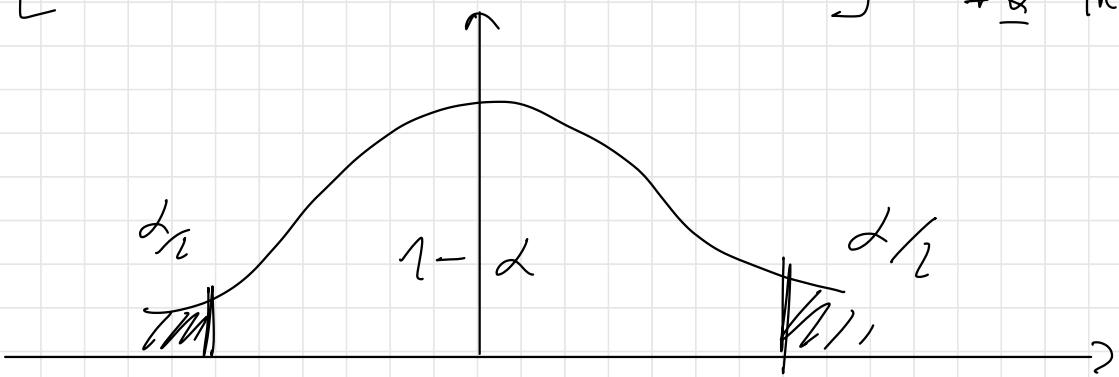
pivotal

$$\lambda \in (0, 1)$$

$$P\left[-t_{1-\frac{\lambda}{2}}(n-1) \leq \frac{\underline{Q}'\bar{x} - \underline{Q}'\mu}{\sqrt{\underline{Q}'\underline{Q}}} S_n \leq t_{1-\frac{\lambda}{2}}(n-1)\right] = 1 - \lambda$$

$$P\left[\underline{Q}'\mu \in \left[\underline{Q}'\bar{x} + t_{1-\frac{\lambda}{2}}(n-1) \sqrt{\frac{\underline{Q}'\underline{Q}}{n}}\right]\right] = 1 - \lambda$$

$\forall \underline{Q} \in \mathbb{R}^p$



$$-t_{1-\frac{\lambda}{2}}(n-1)$$

$$\underline{Q}'\mu$$

$$t_{1-\frac{\lambda}{2}}(n-1)$$

one at
the time

$$CJ_{1-\lambda}(Q'\mu) = \left[\underline{Q}'\bar{x} + t_{1-\frac{\lambda}{2}}(n-1) \sqrt{\frac{\underline{Q}'\underline{Q}}{n}} \right] - V_Q \in \mathbb{R}^n$$

$$Ex \quad \underline{Q} = (0 \ 0 \ 1 \ 0 \ -1) \quad \begin{matrix} | \\ i \end{matrix}$$

$$\underline{Q} = (0 \ 0 \ 1 \ 0 \ -1 \ 0 \ -1 \ 0 \ -1)$$

$$CI_{1-\alpha}(\mu_i) = \left[\bar{x}_i \pm t_{1-\frac{\alpha}{2}}(n-1) \sqrt{\frac{s_{ii}}{n}} \right] \quad i=1\dots D$$

$$CI_{1-\alpha}(\mu_i - \mu_j) = \left[\bar{x}_i - \bar{x}_j \pm t_{1-\frac{\alpha}{2}}(n-1) \sqrt{\frac{s_{ii} - 2s_{ij} + s_{jj}}{n}} \right]$$

Test: $H_0: \underline{\alpha}' \underline{\mu} \leq \underline{d}_0 \quad vs \quad H_1: \underline{\alpha}' \underline{\mu} > \underline{d}_0$

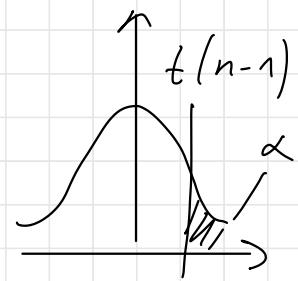
$$\underline{\alpha}' = (0 \dots 0 \ 1 \ 0 \dots 0 \ -1 \ 0 \dots 0)$$

$$H_0: \mu_i - \mu_j \leq d_0 \quad vs \quad H_1: \mu_i - \mu_j > d_0$$

$$t_b = \frac{\underline{\alpha}' \bar{x} - \underline{d}_0}{\sqrt{\underline{\alpha}' \underline{s} \underline{\alpha}}} \quad \text{test stat}$$

$$\mathcal{Z} \in (0, 1)$$

Reject H_0 if $t_b > t_{1-\alpha}(n-1)$

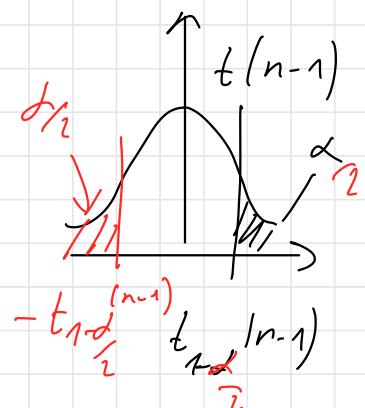


$$t_{1-\alpha}(n-1)$$

$$|t_b| = \frac{(\underline{\alpha}' \bar{x} - \underline{d}_0)}{\sqrt{\underline{\alpha}' S \underline{\alpha}}} \quad \text{for test stat}$$

$\mathcal{L} \in (0, 1)$

Reject H_0 if $|t_b| > t_{1-\frac{\alpha}{2}}(n-1)$



$$\underline{P} \left[\underline{\alpha}' \mu \in \left[\underline{\alpha}' \bar{x} \pm t_{1-\frac{\alpha}{2}}(n-1) \sqrt{\frac{\underline{\alpha}' S \underline{\alpha}}{n}} \right] \right] = 1 - \alpha$$

$\forall \underline{\alpha} \in (\mathbb{R}^p)^p$

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

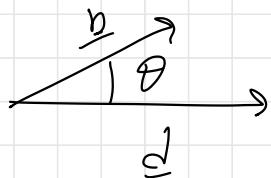
it doesn't mean that all of these CI together cover any combination of μ !

$$\underline{P} \left[\bigcap_{\underline{\alpha}} E_{\underline{\alpha}} \right] = 1 - \alpha \quad \underline{P} [E_{\underline{\alpha}}] = 1 - \alpha$$

$\uparrow \quad P[11111] = 0.5 - \text{no } P[1] = 0.5 \text{ - true}$

means that in all experiment we will have same expected value

A little excursion in ALg



$$\cos \theta = \frac{\underline{b}' \underline{d}}{\|\underline{b}\| \|\underline{d}\|}$$

$$\frac{(\underline{b}' \underline{d})^2}{\|\underline{b}\|^2 \|\underline{d}\|^2} = \cos^2 \theta \leq 1$$

$$(\underline{b}' \underline{d})^2 \leq \|\underline{b}\|^2 \|\underline{d}\|^2 \quad \text{equality if}$$

$$\underline{b} \in \mathcal{L}(\underline{d})$$

$$\left(\sum_i b_i d_i \right)^2 \leq \left(\sum_i b_i^2 \right) \left(\sum_i d_i^2 \right)$$

$$\left(\int f g \right)^2 \leq \left(\int f^2 \right) \left(\int g^2 \right)$$

C-S

Cauchy-Schwarz
Inequality

Prop B pos def $n \times p$

$$\underline{b}, \underline{d} \in \mathbb{R}^p \Rightarrow$$

$$(\underline{b}' \underline{d})^2 \leq (\underline{b}' B \underline{b}) (\underline{d}' B^{-1} \underline{d})$$

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

proof $(\underline{b}^T \underline{d}) = (\underbrace{\underline{b}^T \underline{\beta}^{1/2}}_{(\underline{\beta}^{1/2} \underline{b}^T)} \underbrace{\underline{\beta}^{-1/2} \underline{d}}_T)^2 \leq (\underline{b}^T \underline{\beta} \underline{b})(\underline{d}^T \underline{\beta}^{-1} \underline{d})$

CS

equal if $\underline{\beta}^{1/2} \underline{b} \in \mathcal{L}(\underline{\beta}^{-1/2} \underline{d})$

Cauchy-Schwarz
Inequality

13.03.25

Simultaneous CI's and testing

Max Lemma

β $p \times p$ positive def, $\underline{d} \in \mathbb{R}^p$

$$\max_{\underline{x} \in \mathbb{R}^p} \frac{(\underline{x}' \underline{d})^2}{\underline{x}' \beta \underline{x}} = \underline{d}' \beta^{-1} \underline{d}$$

$$\underline{x} \neq \underline{0}$$

proof $\underline{x} \in \mathbb{R}^p$

$$(\underline{x}' \underline{d})^2 \leq (\underline{x}' \mathbb{I} \underline{x}) (\underline{d}' \beta^{-1} \underline{d}) \quad (C \rightarrow)$$

β pos def $\Rightarrow \underline{x}' \beta \underline{x} > 0$ if $\underline{x} \neq \underline{0}$

hence $\frac{(\underline{x}' \underline{d})^2}{\underline{x}' \beta \underline{x}} \leq \underline{d}' \beta^{-1} \underline{d}$ with equality if $\underline{x} \in \mathcal{L}(\beta^{-1} \underline{d})$

$$\underline{x}_1, \dots, \underline{x}_n \sim$$

i.i.d $N_p(\mu, \Sigma)$

$$\underline{a} \in \mathbb{R}^p$$

$$\frac{\underline{a}' \underline{x} - \underline{a}' \mu}{\sqrt{\underline{a}' \Sigma \underline{a}}} \sqrt{n} \approx \text{pivotel}$$

$$= \frac{\underline{\alpha}^T (\bar{x} - \mu)}{\sqrt{\underline{\alpha}^T \Sigma \underline{\alpha}}} \sqrt{n}$$

by Corollary of
Hotelling Theorem

$$\max_{\underline{\alpha} \in \mathbb{R}^P, \underline{\alpha} \neq 0} n \frac{(\underline{\alpha}^T (\bar{x} - \mu))^2}{\underline{\alpha}^T \Sigma \underline{\alpha}} = n (\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu)$$

$$\sim \frac{(n-1)p}{n-p} F(p, n-p)$$

? $c > 0$ such that $(r, t) \in (0, 1)$

$$P \left[-c \leq \frac{\underline{\alpha}^T (\bar{x} - \mu)}{\sqrt{\underline{\alpha}^T \Sigma \underline{\alpha}}} \sqrt{n} \leq c, \forall \underline{\alpha} \in \mathbb{R}^P, \underline{\alpha} \neq 0 \right]$$

$$\left(\begin{array}{l} \\ \\ \end{array} \right) = 1-\alpha$$

$$P \left[\frac{(\underline{\alpha}^T (\bar{x} - \mu))^2}{\underline{\alpha}^T \Sigma \underline{\alpha}} n \leq c^2, \forall \underline{\alpha} \in \mathbb{R}^P, \underline{\alpha} \neq 0 \right]$$

$$= P \left[\max_{\substack{\underline{\alpha} \in \mathbb{R}^P \\ \underline{\alpha} \neq 0}} \frac{(\underline{\alpha}^T (\bar{x} - \mu))^2}{\underline{\alpha}^T \Sigma \underline{\alpha}} n \leq c^2 \right] =$$

$$= P \left[n(\bar{x} - \mu)^T S^{-1} (\bar{x} - \mu) \leq c^2 \right] \Rightarrow c^2 = \frac{(n-1)p}{n-p} F_{1-\alpha}(p, n-p)$$

$\frac{(n-1)p}{n-p} F(p, n-p)$

hence

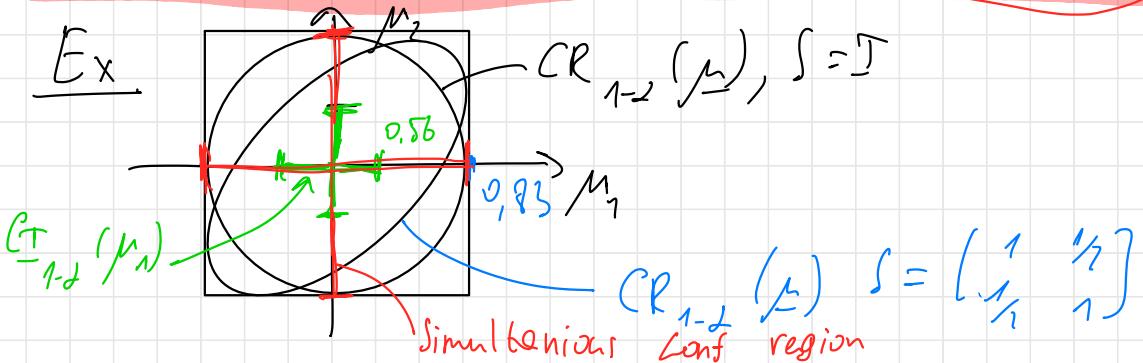
$$P \left[\frac{|\underline{\alpha}' (\bar{x} - \mu)|}{\sqrt{\underline{\alpha}' S \underline{\alpha}}} \sqrt{n} \leq \sqrt{\frac{(n-1)p}{n-p}} F_{1-\alpha}(p, n-p), \forall \underline{\alpha} \in \mathbb{R}^p, \underline{\alpha} \neq \underline{0} \right] = 1 - \alpha$$

Simultaneous confidence interval

The right quantile is
the square root of
an F distribution and
not of a t-student

$$CI_{1-\alpha}(\underline{\alpha}' \mu) = \left[\underline{\alpha}' \bar{x} + \sqrt{\frac{(n-1)p}{n-p}} F_{1-\alpha}(p, n-p) \sqrt{\frac{\underline{\alpha}' S \underline{\alpha}}{n}} \right]$$

$\forall \underline{\alpha} \in \mathbb{R}^p$



Thus now we 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 intervals are correct with probability $1-\alpha$. All of the intervals 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 \mathbf{z} gives us the simultaneous confidence interval $\text{Sim CI}_{1-\alpha}(\mathbf{z}^\top \mu)$

Note: The simultaneous confidence intervals also called Scheffé confidence intervals, and also T^2 confidence intervals!

Ans: Simultaneous confidence intervals are the linear envelope of confidence regions

$$\underline{x}_1 \dots \underline{x}_n \stackrel{i.i.d}{\sim} N(\mu, \Sigma)$$

Аналогично генеральному интервалу для среднего μ .

тут мы используем генеральную формулу для оценки общего ожидания.

$$CI_{1-\alpha} (\underline{\sigma^2} \underline{\mu}) = \left[\underline{\sigma^2} \underline{\bar{x}} \pm t_{2,1} (n-1) \sqrt{\underline{\sigma^2} \underline{S^2} \frac{1}{n}} \right]$$

Т.е. получают один генеральный интервал для общего ожидания.

Тогда получают одновременный интервал для нескольких индивидуальных ожиданий \rightarrow
 \rightarrow simultaneous confidence intervals.

Т.е. если я нахожу одновременные t-интервалы для каждого индивидуального ожидания $\underline{\sigma^2} \underline{\mu}$, то вероятность того, что все эти одновременные ожидания будут одновременно в $1-\alpha$ (т.е. не более α есть хотя бы один из них не входит в

Sim CI - охватывает все возможные индивидуальные ожидания и при этом не входят в $1-\alpha$.

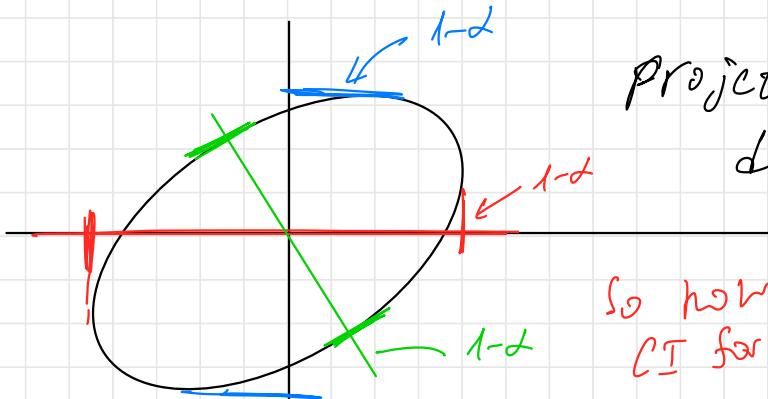
Всем: если генеральная функция $CR_{1-\alpha} (\mu)$, то можно это записать в виде

$\underline{\sigma^2} \underline{\mu}$ даёт Sim CI для $\underline{\sigma^2} \underline{\mu}$

$$Sim CI_{1-\alpha} (\underline{\sigma^2} \underline{\mu}) = \left[\underline{\sigma^2} \underline{\bar{x}} + \sqrt{(n-1) \frac{P}{n-p} F_2(p, n-p)} \sqrt{\underline{\sigma^2} \underline{S^2} \frac{1}{n}} \right]$$

т.е. мы получаем био изображающую структуру.

Confidence as many direction we want, no better direction, the coverage will be $1-\alpha$



projection all directions

so now I have CI for all combinations

Projection of linear combinations of μ

Probability that algorithm producing ... cover value by 95%

Bonferroni method for simultaneous CIs for finite number of linear combination of μ (Because sim CI take all combinations \Rightarrow very large, we want smaller)

Given $\underline{\varphi}_1, \dots, \underline{\varphi}_k \in \mathbb{R}^p$

? Simultaneous CI's for $\underline{\varphi}_1' \underline{\mu}, \dots, \underline{\varphi}_k' \underline{\mu}$

One at the time C_i

$$\underline{\varphi}_i C_{1-\alpha} (\underline{\varphi}_i' \underline{\mu}) = \left[\underline{\varphi}_i' \bar{x} \pm t_{1-\alpha/2}(n-1) \sqrt{\frac{\underline{\varphi}_i' \Sigma \underline{\varphi}_i}{n}} \right]$$

$$P \left[\bigcup_{i=1}^k \left\{ \underline{\varphi}_i' \underline{\mu} \in C_{1-\alpha} (\underline{\varphi}_i' \underline{\mu}) \right\} \right] \text{ Bep 100% ne noumboor } C_i \mu = \alpha$$

$$= 1 - P \left[\bigcup_{i=1}^k \left\{ \underline{\varphi}_i' \underline{\mu} \notin C_{1-\alpha} (\underline{\varphi}_i' \underline{\mu}) \right\} \right] \geq \text{Bep 200% noumboor}$$

$$\overbrace{P(A \cup B) = P(A) + P(B) - P(A \cap B)} \leq P(A) + P(B)$$

(Bonferroni meg)

$$\geq 1 - \sum_{i=1}^k P \left[\underline{\varphi}_i' \underline{\mu} \notin C_{1-\alpha} (\underline{\varphi}_i' \underline{\mu}) \right] = 1 - \alpha k = \frac{\alpha}{k}$$

$$P(\underline{\varphi}_i' \underline{\mu} \in C_{1-\alpha}) < P(\underline{\varphi}_i' \underline{\mu} \in C_{1-\alpha/k}) = 1 - \frac{\alpha}{k}$$

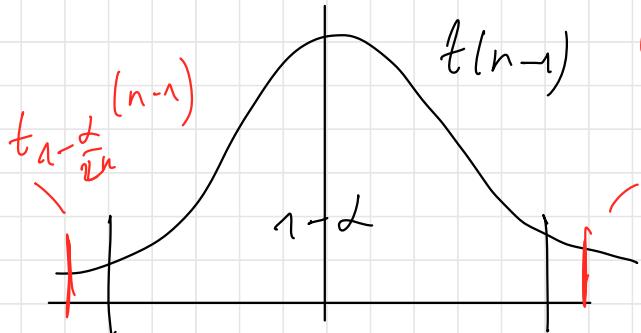
$= 1 - \alpha$ so for each linear combination we

take a confidence interval of level $1 - \alpha/k \Rightarrow$ overall confidence interval is on level $1 - \alpha$ (i.e. ~~beperkt~~ nonparametrisch is ~~significante~~ ~~dele~~ ~~vele~~)

Bonferroni Simultaneous Confidence Intervals

CT's

$$\text{Bonferroni CI}_{1-\alpha}(\bar{x}_i, s_i) = \bar{x}_i + t_{1-\frac{\alpha}{2k}}(n-1) \sqrt{\frac{s_i^2}{k}}$$

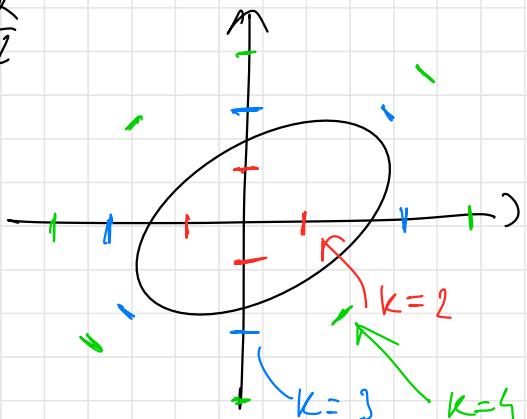


we don't use information about correlation!

increase with k

$$t_{1-\frac{\alpha}{2k}}(n-1)$$

$$t_{1-\frac{\alpha}{2}}(n-1)$$



So if k

will be very large

it can be worse

than standard sum (from only)

If $k = \infty \Rightarrow$ confidence interval is $[-\infty; +\infty]$, so

Bonferroni's Confidence Intervals only work with small finite number of linear combinations!

We have that the Bonferroni's simultaneous confidence interval is larger than the one-at-the-time confidence interval but smaller than the simultaneous confidence Interval.

$$\text{one CI} < \text{Sim Bonferroni} < \text{Sim CI}$$

but when $k \rightarrow \infty$ $\text{Sim BF} > \text{Sim CT}$

Simultaneous Testing with Bonferroni's simultaneous Confidence Interval can be done but since it's very conservative (very strict) it is not used in practice, because with big data we would never come to reject the null hypothesis.

- If we have One-at-the-time Confidence Intervals, then we reject if we are above $t_{\alpha/2} (n-1)$ or symmetrical if we are below $-t_{\alpha/2} (n-1)$
- If we have Bonferroni simultaneous Confidence Intervals with k hypothesis, then we reject if above $t_{\alpha/k} (n-1)$ or below $-t_{\alpha/k} (n-1)$. So the higher k (more hypotheses we have) -the more conservative the procedure gets, the more quantiles are closer to ∞ .

Indeed since we test k hypothesis simultaneously, and we want overall level α , then for each single test we need to test much smaller: α/k

Testing Guilty or
no

$H_0: \bar{\mu}_1 = \bar{\mu}_2 = \dots = \bar{\mu}_k$ vs $H_1: \bar{\mu}_1 = \bar{\mu}_2 = \dots = \bar{\mu}_k$
 & or
 $H_{01}: \bar{\mu}_1 = \bar{\mu}_2 = \dots = \bar{\mu}_{k-1} \neq \bar{\mu}_k$ vs $H_{11}: \bar{\mu}_1 = \bar{\mu}_2 = \dots = \bar{\mu}_{k-1} \neq \bar{\mu}_k$
 & or:
 \vdots \vdots
 $H_{0k}: \bar{\mu}_1 = \bar{\mu}_2 = \dots = \bar{\mu}_k$ vs $H_{1k}: \bar{\mu}_1 = \bar{\mu}_2 = \dots = \bar{\mu}_k$

but at the end, when $k \rightarrow \infty$

Decision rule: probability that loyal made mistake will not be $\alpha\%$, it will be much bigger

Reject H_{0i} $i=1\dots k$

if $\frac{|\bar{\mu}_i - \bar{\mu}_j|}{\sqrt{\frac{\sigma^2}{n}} \sqrt{\frac{\sigma^2}{n}}} > t_{1-\alpha/2}(n-1)$

$$P[\text{reject at least one } H_{0i} \mid \cap H_{0i} \text{ true}]$$

$$= P\left[\bigcup_{i=1}^k \left\{ \frac{|\bar{\mu}_i - \bar{\mu}_j|}{\sqrt{\frac{\sigma^2}{n}} \sqrt{\frac{\sigma^2}{n}}} > t_{1-\alpha/2}(n-1) \right\} \mid \cap H_{0i} \text{ true}\right]$$

$$\leq \sum_{i=1}^k P\left[\underbrace{H_i}_{\text{At least } i-H_0} \text{ true}\right] = \sum_{i=1}^k \frac{\lambda}{k} = \lambda$$

so the overall probability of at least rejecting one null hypothesis when in fact all of them are true is equal to λ

k -thousands and more

Large Scale Hypothesis Testing and False Discovery Rate.

Benjamini & Hochberg (1995)

V - False positive
T - False negative

K simultaneous test

$$k_0 = k - V$$

$$k_1 = k - k_0 = T + S$$

$$U + T = k - R$$

D strategy for testing these k hypothesis

(e.g. Bonferroni's strategy)

Decisions following D

False discoveries

	Do not Reject H_0	Reject H_0	
H_0	T	V	k_0
H_1	T	S	$k - k_0$
missed discoveries			True discoveries
$(k - R)$			(k)

Let $I_0 = \{i \in \{1, \dots, k\} : h_{0i} \text{ is true}\}$
 $K_0 = |I_0|$

$\angle G(0, 1)$

D Bonferroni strategies

$$\begin{aligned} & P[\text{at least one rejected } h_{0i} \mid \bigcap_{i=I_0} I_0] = \\ & = P[V \geq 1] = P\left[\bigcup_{i \in I_0} \{\text{reject } h_{0i} \mid h_{0i} \text{ true}\}\right] \leq \\ & \leq \left(\sum_{i \in I} \frac{\alpha}{k} \right) = k_0 \frac{\alpha}{k} \leq k \frac{\alpha}{k} = \alpha \end{aligned}$$

so when we use Bonferroni's sum we are sure that the probability we reject one or more of true hypothesis is less than α

Family wise Error Rate (FWER)

If D is Bonferroni $\Rightarrow \text{FWER} \leq \alpha$

Now consider $\frac{V}{R}$

$Q = \begin{cases} 0 & \text{if } R = 0 \\ \frac{V}{R} & \text{if } R > 0 \end{cases}$

which part of rejected were rejected wrongly

Def

False Discovery Rate

$$FDR = E[Q]$$

Rmk 1. $k_0 = k$ (nothing to be discovered)

$$\Rightarrow S = 0 \text{ and } V = R$$

$$Q = \begin{cases} 0 & \text{if } V = R = 0 \\ 1 & \text{if } V = R > 0 \end{cases}$$

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

2. $k_0 < k$ if $V = 0 \Rightarrow Q = 0$

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

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

Indicator function

$\left\{ \begin{array}{ll} \mathbb{1}[V > 0] & \rightarrow V > 0 \\ 0 & \text{otherwise} \end{array} \right.$

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

$$\text{FDR} \leq FWER$$

$$= P[V > 0] = P[V \geq 1] = \underline{FWER}$$

Therefore we proved that no matter the number of null hypothesis are true, then $FDR \leq FWER$ and this is exactly why FDR is so appealing: FDR is weaker than FWER, so maybe we can control FDR without being so conservative as Bonferroni is!

Consider a procedure such that: $FDR \leq \alpha$. Then we know that the probability of error (FWER) could be much higher but we don't care.

Therefore if we need to do a lot of multiple testing we control FDR: we are not controlling the probability of the type I error (which will be higher), otherwise we wouldn't have any applicability! Conclusion. Bonferroni is good but not usable, FDR is less good but usable.

A strategy for controlling FDR (B & H 1955)

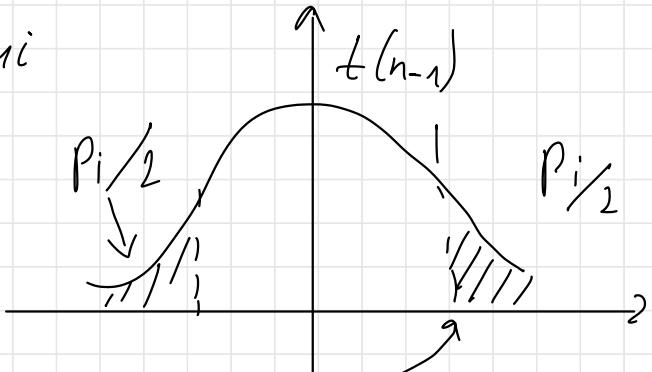
let p_i be the p-value for the test H_{0i} vs H_{1i}

consider $p_1 \dots p_k$

and order

$$p_1 \leq p_2 \dots \leq p_k$$

$$\begin{array}{ccc} | & | \\ h_{0(1)} \text{ vs } h_{1(1)} & h_{0(k)} \text{ vs } h_{1(k)} \end{array}$$



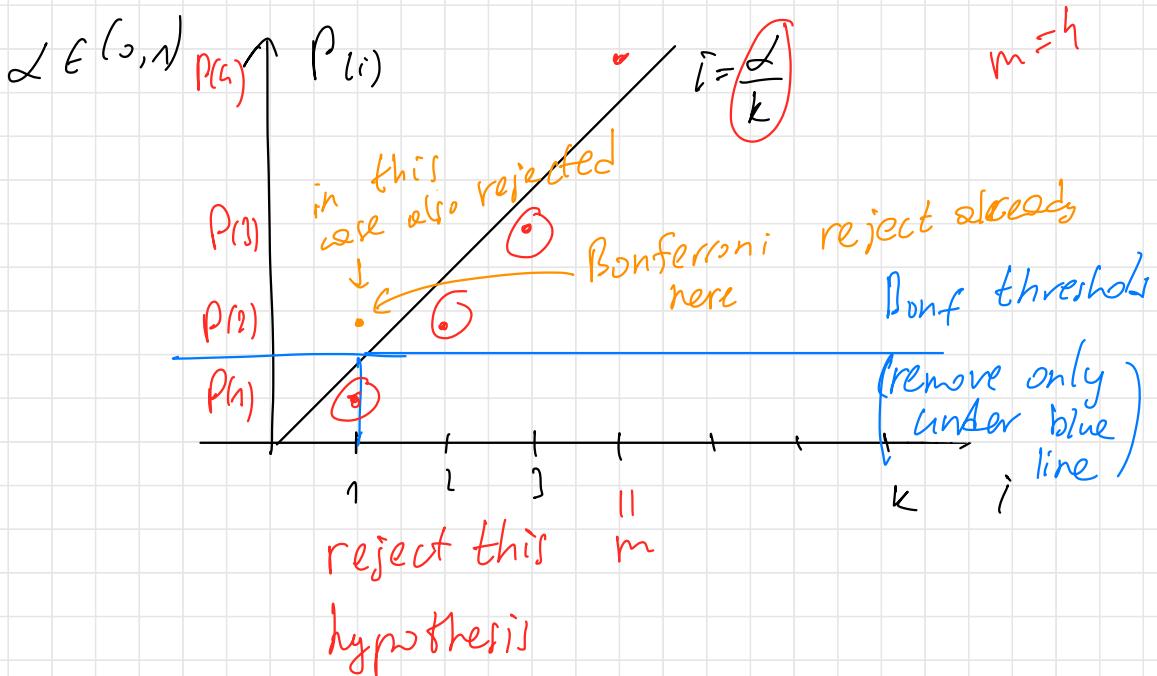
$$t_i = \frac{\bar{Q}_i^T X - \bar{J}_i}{\sqrt{\bar{Q}_i^T S \bar{Q}_i}} \sqrt{n}$$

Theo

If p_1, \dots, p_k are independent
let $m = \max \{ i \in \{1 \dots k\} : p_{(i)} \leq i \cdot \frac{\alpha}{k} \}$

strategy D_{BH} reject $H_{0(i)}$ if $i \leq m \Rightarrow$

D_{BH} controls FDR at level α



Reject everything that before lost
 $P_{(i)}$ lower the line.

Too (Benjamini & Yekutieli , 2001)

1) If P_1, \dots, P_n are positive correlated $\Rightarrow D_{BF}$ controls FDR at level α

2) If P_1, \dots, P_n are negative correlated
 (Any hyp ↑ other decrease) $\Rightarrow D_{BH}^*$

$$\text{let } m^* = \max_k \left\{ i \in \{1, \dots, k\} : P_{(i)} \leq i \cdot \frac{\alpha}{k \cdot C(k)} \right\}$$

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

Strategy: D_{Bh}^* rejects $H_{0(i)}$ if $i \leq m^*$. \Rightarrow

D_{Bh} controls FPR at level α .

Note: mixed cases are not covered: we need both ALL positively correlated, or ALL negatively correlate \downarrow p-value.

Korpe sowi gle untegn, zada nape -
numbers yrobens yberennic, 1-
Kampas by nun yomme oche werst
yndens yberennic $(1 - \frac{\alpha}{k})(1 - \frac{\alpha}{k}) \geq 1 - \alpha$

Wortberichtens reh dawue k-meh deline
y Kampas CI no omfelsnath, zada
hure ohe gath 1- α

In nova $P_{(i)}$ nume gato $i \cdot \frac{\alpha}{k}$ mir mome
un oterpart.

17.03.25 Lecture

Comparing means of
different Gaussian
distributions

in December

Paired data

n units

x_1, \dots, x_p features

$i = 1, \dots, n$

For each unit:

$$\underline{x}_{i1} = \begin{pmatrix} x_{i11} \\ x_{i12} \\ \vdots \\ x_{i1p} \end{pmatrix} \in \mathbb{R}^p$$

might be
dependent

Overall data

$$\underline{x}_{i2} = \begin{pmatrix} x_{i21} \\ \vdots \\ x_{i2p} \end{pmatrix} \in \mathbb{R}^p$$

$$\underline{\mu}_1 \left(\underline{x}_{11} \right), \left(\underline{x}_{21} \right), \dots, \left(\underline{x}_{n1} \right) \quad \text{and} \quad \underline{\mu}_2 \left(\underline{x}_{12} \right), \left(\underline{x}_{22} \right), \dots, \left(\underline{x}_{n2} \right)$$

independent

(but inside can be dependent)

Goal: inference $\underline{\mu}_1 - \underline{\mu}_2$

Note: we need to have paired data:
the two vectors are paired in the
sense that both are observation for
the same statistical unit! As example
of statistical units:

- Same person and we could measure his heart beat, his pressure both before and after treatment.
- Family and we observe the degree of education of father and mother and then the income of father and mother.
- What is bad pairing after having observed the sample. as example.

We take a bunch of Italians and measure their height, same for French group and pair tallest Italian with tallest French ... one's smallest It with smallest French — it is not pairing. We don't observe same statistical unit.

Conclusion: we want to see if treatment has an effect: is there a change?

$$\underline{D}_i = \underline{x}_{i1} - \underline{x}_{i2}$$

$\underline{D}_1, \dots, \underline{D}_n$ - independent i.i.d $\sim N_p(\underline{\delta}, \Sigma)$

$$\underline{d} = \underline{\mu}_1 - \underline{\mu}_2$$

$$\bar{\underline{D}} = \frac{1}{n} \sum_{i=1}^n \underline{D}_i \quad S_{\underline{D}} = \frac{1}{n-1} \sum_{i=1}^n (\underline{D}_i - \bar{\underline{D}})(\underline{D}_i - \bar{\underline{D}})^T$$

$$h(\bar{\underline{D}} - \underline{\delta}) S_0^{-1} (\bar{\underline{D}} - \underline{\delta}) \sim \frac{(n-1)}{n-p} F(p, n-p)$$

pivotel

$$\mathcal{L} \in (0, 1)$$

$$\begin{aligned} CR_{1-\alpha}(\underline{\delta}) &= \left\{ \underline{\gamma} \in \mathbb{R}^p : h(\bar{\underline{D}} - \underline{\gamma})^T S_0^{-1} (\bar{\underline{D}} - \underline{\gamma}) \leq \right. \\ &\leq \left. \frac{(n-1)}{n-p} F_{n-\alpha}(p, n-p) \right\} \end{aligned}$$

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

$$T_0^2 = n (\bar{D} - \underline{f}_0)^T \underline{f}_0^{-1} (\bar{D} - \underline{f}_0)$$

Reject at level $\alpha \in (0, 1)$ if

$$T_0^2 > \frac{(n-1)p}{n-p} F_{1-\alpha}(p, n-p)$$

$$\lim_{n \rightarrow \infty} C_{1-\alpha}^T (\underline{\varphi}) = \left[\bar{D} \pm \sqrt{\frac{(n-1)p}{n-p} F_{1-\alpha}(p, n-p)} \right]$$

$$\left[\bar{D} \pm \sqrt{\frac{\underline{\varphi}' \underline{f}_0^{-1} \underline{\varphi}}{n}} \right] \quad \forall \underline{\varphi} \in \mathbb{R}^p$$

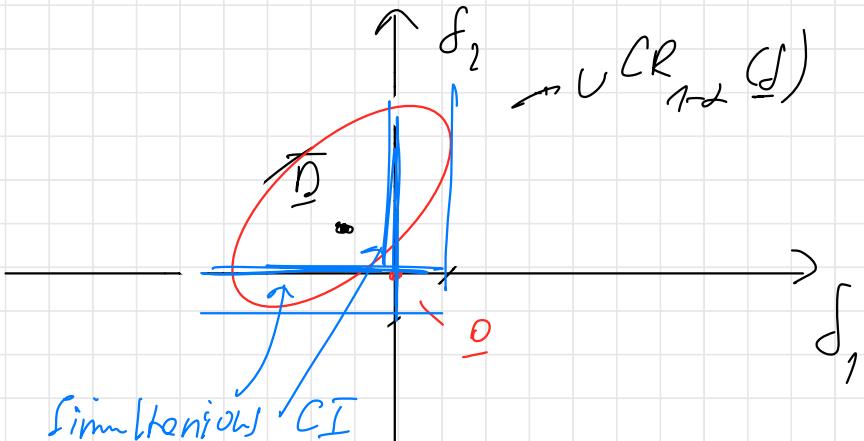
E_x

$$\lim_{\substack{n \rightarrow \infty \\ j=1 \dots p}} C_{1-\alpha} (\underline{f}_j) = \left[\bar{D}_j \pm \sqrt{\frac{(n-1)p}{n-p} F_{1-\alpha}(p, n-p)} \right]$$

$$\left[\frac{\underline{f}_{0,jj}}{n} \right]$$

$$\text{Bonf} \left[C_{1-\alpha} (\underline{f}_j) \right] = \left[\bar{D}_j \pm t_{1-\frac{\alpha}{2p}, n-1} \sqrt{\frac{\underline{f}_{0,jj}}{n}} \right]$$

Rmk



$H_0: \underline{f} = \underline{0}$ vs $H_1: \underline{f} \neq \underline{0}$ \Rightarrow can reject

H_0 , because $\underline{0}$ is not in CR

$\Rightarrow \mu_1 - \mu_2$ (μ_1 is different from μ_2)

If mean of wife and husband different - we still can't say if the difference is in height or weight. We can only say that there are linear combinations where they are different.

(little right move)

$$\bar{D} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

Is the mean diff

different from \underline{D} ?

No?

Repeated Measurements

For unit i : obs Ex.

$x: x_{i1}, x_{i2}, \dots, x_{iq}$ - (weight in different time)
same feature x observed in q

instances

$$\underline{x}_i = (x_{i1}, x_{i2}, \dots, x_{iq}) \in \mathbb{R}^q$$

$$\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n \text{ i.i.d } \sim N_q(\underline{\mu}, \Sigma)$$

$$H_0: \mu_1 = \mu_2 = \dots = \mu_q \quad \text{Vs} \quad H_1: \mu_0 \neq$$

$$\mu_j = E(x_j) \quad j=1 \dots q$$

for all people in month j

$$\underline{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} \quad H_0: \underline{\mu} \in L(\underline{1}) \quad \left(\text{because } \mu_1 = \mu_2 = \dots = \mu_p \right)$$

$$\text{vs } H_1: \underline{\mu} \notin L(\underline{1})$$

Consider $\underline{c}_1, \dots, \underline{c}_{q-1} \in \mathbb{R}^q$ s.t.

1) $\underline{c}_1, \dots, \underline{c}_{q-1}$ are lin independent \Rightarrow

$$2) \underline{c}_i \cdot \underline{1} = 0 \quad i=1, \dots, q-1 \Leftrightarrow \underline{c}_i \perp \underline{1}$$

$$\Rightarrow L^{-1}(\underline{1}) = \text{Span}(\underline{c}_1, \dots, \underline{c}_{q-1})$$

$$C = \begin{bmatrix} \underline{c}_1 \\ \vdots \\ \underline{c}_{q-1} \end{bmatrix} \quad (q-1) \times q$$

contrasts matrix

Ex $C = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{bmatrix}$ What is change
 $q-1$ t_1, t_2, t_3, t_4

$$C = \begin{pmatrix} 1 & -1 & 0 & \hline & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \end{pmatrix}$$

different from baseline

$$H_0: C_{\underline{\mu}} = \underline{0} \quad \text{vs} \quad H_1: C_{\underline{\mu}} \neq \underline{0}$$

Unbiased estimator for $C_{\underline{\mu}}$: $\underline{C\bar{X}}$

$$\underline{C\bar{X}} \sim N_{p-1} (C_{\underline{\mu}}, \frac{1}{n} C \Sigma C^T)$$

$$\sqrt{n} (\underline{C\bar{X}} - C_{\underline{\mu}}) \sim N_{p-1} (\underline{0}, C \Sigma C^T), \quad \text{(independ)} \\ \Downarrow$$

$$(n-1) C \Sigma C^T \sim \text{Wishart}(C \Sigma C^T; n-1) \quad \swarrow$$

\Rightarrow Hotelling's Theorem

$$n (\underline{C\bar{X}} - C_{\underline{\mu}})^T (C \Sigma C^T)^{-1} (\underline{C\bar{X}} - C_{\underline{\mu}}) \sim \\ \sim \frac{(n-1)(p-1)}{n-p+1} F_{1-\alpha} (p-1, n-p+1) \quad \text{pivotel}$$

$$T_0^2 = n (\underline{C\bar{X}})^T (C \Sigma C^T)^{-1} (\underline{C\bar{X}}) \quad \begin{matrix} T_0^2 (n-p+1) \\ (n-1)(p-1) \end{matrix} > F_{1-\alpha}$$

Reject at level $\alpha \in (0, 1)$ if

$$T_0^2 > \frac{(n-1)(p-1)}{n-p+1} F_{1-\alpha} (p-1, n-p+1)$$

What if I will try different C
matrices?

Remark C and \tilde{C} are two
different contrast matrices

$$T_0^2 = n (\tilde{C} \underline{x})^T (\tilde{C} \Sigma \tilde{C}^T)^{-1} (\tilde{C} \underline{x})$$

$$\exists B \quad (q-1) \times (q-1)$$

$$\tilde{C} = BC$$

so contrast matrix is not
given by the problem, is just
a way to define a linear
basis on the space orthogonal
to 1 (all the basis are equivalent)
indeed)

$$T_0^2 = n (BC \underline{x})^T (BC \Sigma C^T B^T)^{-1} (BC \underline{x}) =$$

$$= n (\underline{C} \underline{x})^T \cancel{B^T} \cancel{(B)}^{-1} (\cancel{C} \Sigma \cancel{C}^T)^{-1} \cancel{B}^{-1} \cancel{B} (\underline{C} \underline{x})$$

Generalization x_1, \dots, x_n i.i.d. $\sim N(\mu, \Sigma)$

$$h_0: \mu \in L \quad \text{vs} \quad h_1: \mu \notin L$$

L lin. space of dimension k

$$\mathcal{L}^\perp = \text{span}(\underline{c}_1, \dots, \underline{c}_{p-k})$$

$\underline{c}_1, \dots, \underline{c}_{p-k}$ are basis for \mathcal{L}^\perp

$$C = \begin{bmatrix} \underline{c}_1' \\ \vdots \\ \underline{c}_{p-k}' \end{bmatrix} \Rightarrow h_0: C\mu = 0 \text{ vs } h_1: C\mu \neq 0$$

$$\text{S_n} (C \bar{x} - C\mu) \sim N_{p-k}(0, C\Sigma C^T)$$

$$(n-1) C \bar{x} \sim \text{Wishart}(C\Sigma C^T, n-1) \quad \text{--- II}$$

\Rightarrow Hotelling's Th

$$h(C \bar{x} - C\mu)^T (C\Sigma C^T)^{-1} (C \bar{x} - C\mu) \sim \frac{(n-1)(p-k)}{n-p+k}.$$

$$\cdot F_{n-2}(p-k, n-p+k)$$

Ex.

On unit i : $i = 1 \dots n$

$$\underline{x}_i = \left(\begin{array}{c} (x_{i11}) \\ (x_{i21}) \\ \vdots \\ T \end{array} \right), \left(\begin{array}{c} (x_{i12}) \\ (x_{i22}) \\ \vdots \\ T \end{array} \right), \dots, \left(\begin{array}{c} (x_{i1p}) \\ (x_{i2p}) \\ \vdots \\ T \end{array} \right)$$

feature time
height
weight

unit i

person i
person j
...
person n

$$\left(\begin{array}{c} \mu_{h_1} \\ \vdots \\ \mu_{h_p} \end{array} \right) \leftarrow \left\{ \begin{array}{l} \text{means for } \underline{x}_i \\ \text{Height} \end{array} \right\} \quad \underline{\mu} \in \mathbb{R}^{2p}$$

$$\left(\begin{array}{c} \mu_{w_1} \\ \vdots \\ \mu_{w_p} \end{array} \right) \leftarrow \left\{ \begin{array}{l} \text{means for } \underline{x}_i \\ \text{Weight} \end{array} \right\}$$

$$h_0: \mu_{h_1} = \mu_{h_2} = \dots = \mu_{h_p}$$

&

$$\mu_{w_1} = \mu_{w_2} = \dots = \mu_{w_p}$$

$$\left(\begin{array}{c} 1 \\ 1 \\ \vdots \\ 1 \\ ? \\ \vdots \\ 2 \end{array} \right)$$

$$v_1 h_1: h_0^C$$

↑
something change
over time

orthogonal linear space
2p - 2

MANOVA

We have g independent samples, coming from g different populations: $x_{11}, \dots, x_{1n_1} \stackrel{iid}{\sim} N_p(\mu_1, \Sigma)$ and $x_{21}, \dots, x_{2n_2} \stackrel{iid}{\sim} N_p(\mu_2, \Sigma)$ and so on so forth, until $x_{g1}, \dots, x_{gn_g} \stackrel{iid}{\sim} N_p(\mu_g, \Sigma)$

Our goal is to make inference on the means μ_1, \dots, μ_g this population, even though this procedure is called analysis of variance! We want to see if there is enough variability among the estimator of these means to guarantee that they are different.

We want to compare the variability between with variability within the population.

Suppose we need to decide a certain dose of fertiliser so to optimise crop: what's the best setting for the parameter (dose of fertiliser)? We treat different statistical unit for each group: are there any differences in the means of the output the experiment in generating in the treatment?

It's important to notice that each group has the same co-variance matrix Σ . So that to make the MANOVA troublesome. But how do we know if they are the same since they are unknown? We will estimate them with sample co-variance and then we do some testing on equality of covariance and then decide if we are satisfied enough to say they are the same!

Note if the test says that the covariance matrices are not the same - then we try to transform data until the assumption is satisfied

Note in generality, in which we can have any g and any p the Manova problem is still an open problem, which is being tackled in a non parametric way.

Multivariate Analysis of Variance

MANOVA

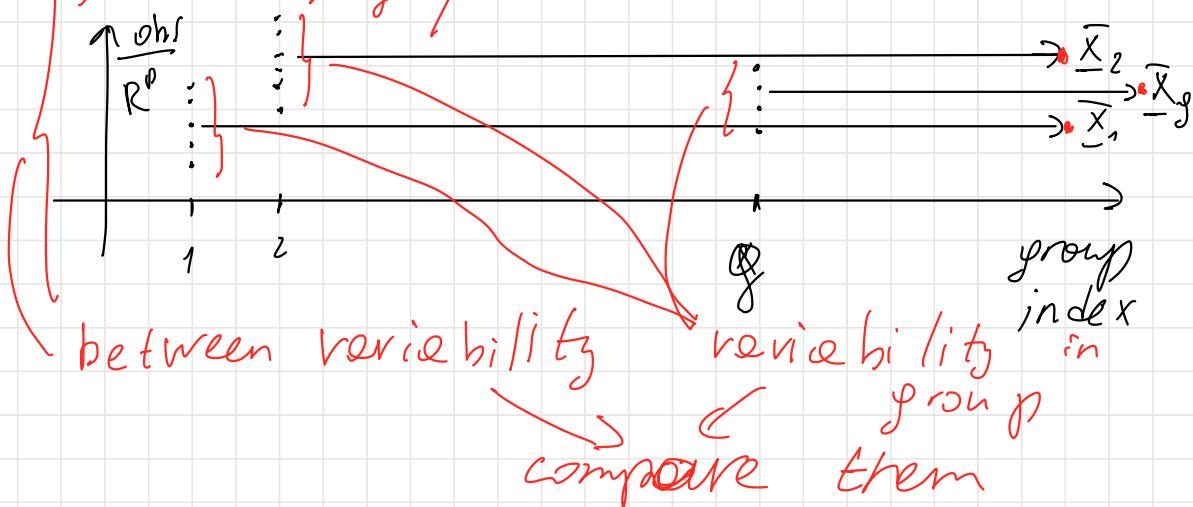
$$\begin{array}{ll} \underline{x}_{11}, \dots, \underline{x}_{1n_1} & i.i.d \sim N_p(\underline{\mu}_1, \underline{\Sigma}) \\ \underline{x}_{21}, \dots, \underline{x}_{2n_2} & i.i.d \sim N_p(\underline{\mu}_2, \underline{\Sigma}) \\ \vdots & \vdots \\ \underline{x}_{g1}, \dots, \underline{x}_{gn_g} & i.i.d \sim N_p(\underline{\mu}_g, \underline{\Sigma}) \end{array}$$

} No pering
independ

} independe nt

Same Σ possibly different means

$\underline{\mu}_1, \dots, \underline{\mu}_g$ / different treatment



(h) ANOVA

compare the variability between groups with within groups

Case 1

$p \geq 1, f = 2$

$\underline{x}_1, \dots, \underline{x}_{n_1}$ i.i.d $N_p(\underline{\mu}_1, \Sigma)$)

$\underline{x}_2, \dots, \underline{x}_{n_2}$ i.i.d $N_p(\underline{\mu}_2, \Sigma)$)

goal: inference on $\underline{\mu}_1 - \underline{\mu}_2$

$$\begin{array}{l} \leftarrow \bar{\underline{x}}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \underline{x}_{1i} \sim N_p\left(\underline{\mu}_1, \frac{1}{n_1} \Sigma\right) \\ \Downarrow \end{array} \Rightarrow$$

$$\bar{\underline{x}}_2 \sim N_p\left(\underline{\mu}_2, \frac{1}{n_2} \Sigma\right)$$

$$\Rightarrow \bar{\underline{x}}_1 - \bar{\underline{x}}_2 \sim N_p\left(\underline{\mu}_1 - \underline{\mu}_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \Sigma\right)$$

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2} \left((\bar{\underline{x}}_1 - \bar{\underline{x}}_2) - (\underline{\mu}_1 - \underline{\mu}_2) \right) \sim N_p(0, \Sigma)$$

$$S_1 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\underline{x}_{i1} - \bar{\underline{x}}_1)(\underline{x}_{i1} - \bar{\underline{x}}_1)^T \text{ est of } \Sigma$$

$$S_2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (\underline{x}_{i2} - \bar{\underline{x}}_2)(\underline{x}_{i2} - \bar{\underline{x}}_2)^T \text{ est of } \Sigma$$

$$\Rightarrow (n_1 - 1) S_1 + (n_2 - 1) S_2 \sim \text{Wish}(\Sigma, n_1 + n_2 - 2)$$

$$S_{\text{pooled}} = \frac{(n_1) S_1 + (n_2 - 1) S_2}{n_1 + n_2 - 2}$$

Weighted
average of
two estimators

$$(n_1 - 1) S_1 \sim \text{Wish}(\Sigma, n_1 - 1) \quad \backslash$$

$$(n_2 - 1) S_2 \sim \text{Wish}(\Sigma, n_2 - 1) \quad /$$

$$(n_1 + n_2 - 2) S_{\text{pool}} \sim \text{Wish}(\Sigma, n_1 + n_2 - 2)$$

$$S_{\text{pool}} \perp \underline{\underline{x}}_1 - \underline{\underline{x}}_2$$

Hotelling's Th: \Rightarrow

$$\left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1} \left((\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2) \right) \left(S_{\text{pooled}} \right)^{-1} \left((\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2) \right) \sim \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F(p, n_1 + n_2 - 1 - p)$$

Test

$$H_0: \mu_1 - \mu_2 = d_0 \quad \text{vs} \quad H_1: \mu_1 - \mu_2 \neq d_0$$

$$T_0^2 = \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1} \left((\bar{x}_1 - \bar{x}_2) - d_0 \right)^T \left(S_{\text{pooled}} \right)^{-1} \left((\bar{x}_1 - \bar{x}_2) - d_0 \right)$$

Reject at level $\alpha \in (0, 1)$

$$\text{if } T_0^2 > \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{1-\alpha}(p, n_1 + n_2 - 1 - p)$$

$$CR_{1-\alpha}(\mu_1 - \mu_2) = \{ \gamma \in \mathbb{R}^p : \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1} \left((\bar{x}_1 - \bar{x}_2) - \gamma \right)^T$$

$$\underbrace{\text{no } x_{12}}_{\text{no } x_{12} \text{ and } x_{22}} \cdot \left(S_{\text{pooled}}^{-1} \left((\bar{x}_1 - \bar{x}_2) - \gamma \right) \right) \leq \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{1-\alpha}(\dots)$$

Try for simultaneous and Bonferroni

18.03.25

$$\underline{x}_{11}, \dots, \underline{x}_{1n_1} \text{ i.i.d. } \sim N_p(\underline{\mu}_1, \underline{\Sigma}_1)$$

$$\underline{x}_{21}, \dots, \underline{x}_{2n_2} \text{ i.i.d. } \sim N_p(\underline{\mu}_2, \underline{\Sigma}_2) \quad \perp\!\!\!\perp$$

⇒ Inference for $\underline{\mu}_1, \underline{\mu}_2$

$$? H_0: \underline{\Sigma}_1 = \underline{\Sigma}_2$$

- Tests generalising Levene tests for $p=1 \Rightarrow$ Anderson

- Permutation test \leftrightarrow Pigoli et al

Large samples n_1, n_2 - large \Rightarrow CLT / linear Theorem

$$\underline{\bar{x}}_1 \sim N_p(\underline{\mu}_1, \frac{1}{n_1} \underline{\Sigma}_1) \quad \perp\!\!\!\perp \Rightarrow$$

$$\underline{\bar{x}}_2 \sim N_p(\underline{\mu}_2, \frac{1}{n_2} \underline{\Sigma}_2)$$

$$\underline{\bar{x}}_1 - \underline{\bar{x}}_2 \sim N_p(\underline{\mu}_1 - \underline{\mu}_2, \frac{1}{n_1} \underline{\Sigma}_1 + \frac{1}{n_2} \underline{\Sigma}_2)$$

$$\left[(\bar{x}_1 - \hat{x}_1) - (\mu_1 - \mu_2) \right]^T \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} \left[(\bar{x}_1 - \hat{x}_1) - (\mu_1 - \mu_2) \right]$$

$$\sim \chi^2(p)$$

$$S_1 \xrightarrow{P} \Sigma_1 \text{ as } n_1 \uparrow \infty$$

\Rightarrow

$$S_2 \xrightarrow{P} \Sigma_2 \text{ as } n_2 \uparrow \infty$$

$$\left[(\bar{x}_1 - \hat{x}_1) - (\mu_1 - \mu_2) \right]^T \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} \left[(\bar{x}_1 - \hat{x}_1) - (\mu_1 - \mu_2) \right]$$

$\sim \chi^2(p)$ pivotal \Rightarrow

$$\begin{cases} CR_{1-\alpha}(\mu_1 - \mu_2) \\ \lim CT_{1-\alpha}(\hat{\alpha} | \underline{x}_1 - \underline{x}_2) \neq \alpha \in \mathbb{R}^P \\ \text{Bonf CT}_{1-\alpha}(\underline{\alpha}' | \underline{\mu} - \underline{\mu}_2) \end{cases}$$

Case $p=1$, $g \geq 2$ (one feature, ≥ 2 samples)

$$\left. \begin{array}{l} X_{1,1}, \dots, X_{1,n_1} \text{ iid } \sim N_1(\mu_1, \sigma^2) \\ X_{2,1}, \dots, X_{2,n_2} \text{ iid } \sim N_2(\mu_2, \sigma^2) \\ \vdots \\ X_{g,1}, \dots, X_{g,n_g} \text{ iid } \sim N_g(\mu_g, \sigma^2) \end{array} \right\} \begin{array}{l} \text{group 1} \\ \text{group 2} \\ \text{independent} \\ \text{group g} \end{array}$$

$$n = n_1 + n_2 + \dots + n_g$$

(what the difference
in treatments,
same vaccine, but different
dose)

ANOVA
one-way

Goal

$$H_0: \mu_1 = \dots = \mu_g \text{ vs } H_1: \mu_i \neq \mu_j$$

(means that at least something changed
and we have new treatment)

If H_0 is rejected

Find the relevant diff $\mu_i - \mu_j$

Model representation

$$x_{ij} = \underbrace{\mu_i}_{\mu_i} + \varepsilon_i + \varepsilon_{ij} \quad \mu \in \mathbb{R}, \quad \varepsilon_i \in \mathbb{R}$$

$$i = 1 \dots g$$

$$\varepsilon_{ij} \text{ i.i.d } \sim N(0, \sigma^2)$$

$$j = 1 \dots n_i$$

to make unbiased

$$\sum \mu_i \varepsilon_i = 0 \quad (\sum \varepsilon_i = 0)$$

$$\underbrace{\mu_1, \mu_2, \dots, \mu_g}_{g \text{ parents}}$$

$$\underbrace{\mu + \varepsilon_1, \mu + \varepsilon_2, \dots, \mu + \varepsilon_g}_{g+1 \text{ parameters}}$$

$n_1 = n_2 = \dots = n_g$ (balanced experiment)
 equal number of mouses

we need constraint on ε

Estimator of μ : take mean over all samples

$$\bar{x} = \frac{1}{n} \sum_{i=1}^g \sum_{j=1}^{n_i} x_{ij}$$

check unbiased $E[\bar{x}] = \frac{1}{n} \sum_{i=1}^g \sum_j E[x_{ij}] =$

$$= \frac{1}{n} \sum_i \sum_j (\mu + \varepsilon_i) = \frac{1}{n} \sum_{i=1}^g n_i (\mu + \varepsilon_i) =$$

$$= \frac{1}{n} \sum_i n_i \mu + \frac{1}{n} \sum_i n_i \varepsilon_i = \mu + \frac{1}{n} \sum_{i=1}^g n_i \varepsilon_i$$

Estimator of $\{2_i\}$, $i=1, \dots, p$.

If $\bar{X}_i = \frac{1}{n} \sum_{j=1}^{n_i} X_{ij}$ mean in group i

$\Rightarrow \bar{X}_i - \bar{X}$ - unbiased for 2_i

$$E[\bar{X}_i - \bar{X}] = \mu + 2_i - \mu = 2_i \text{ (unbiased)}$$

Variance decomposition formula

$$\underline{X} = \begin{pmatrix} X_{11}, & \dots & X_{1n_1}, & X_{21}, & \dots & X_{2n_2}, & \dots & X_{p1}, & \dots & X_{pn_p} \end{pmatrix}^\top$$

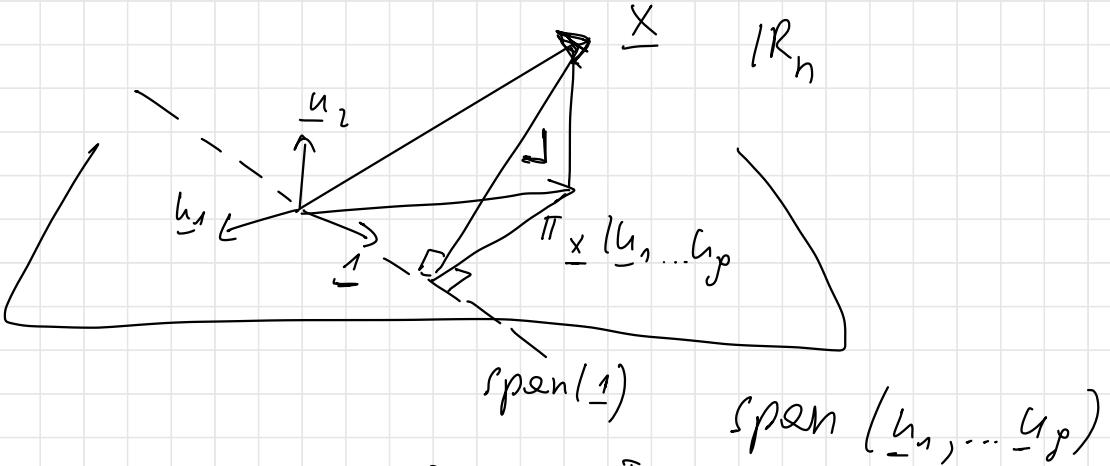
$\in \mathbb{R}^n$

$$\underline{u}_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{n-p} \quad \underline{u}_2 = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}_{n-p} \quad \dots \quad \underline{u}_y = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}_{n-p}$$

1) $\underline{u}_1, \dots, \underline{u}_y$ - are linear independent

2) $\underline{u}_i^\top \underline{u}_j = 0 \quad i, j = 1, \dots, p \quad \underline{u}_i \perp \underline{u}_j$
(orthogonal)

3. $\underline{1} \in \text{Span}(\underline{u}_1, \dots, \underline{u}_g)$ since $\underline{1} = \sum_{i=1}^g \underline{u}_i$



$$\pi_X | \underline{u}_1, \dots, \underline{u}_g = \sum_{i=1}^g \frac{\underline{u}_i \underline{u}_i^T}{\underline{u}_i^T \underline{u}_i} X = \sum_{i=1}^g \frac{1}{h_i} \sum_{j=1}^{n_i} x_{ij} \cdot \underline{u}_i$$

$$= \sum_{i=1}^g \bar{x}_i \underline{u}_i$$

$$\pi_X | \underline{1} = \frac{\underline{1} \underline{1}^T}{\underline{1} \underline{1}} X = \left(\frac{1}{h} \sum_{i=1}^g \sum_{j=1}^{n_i} x_{ij} \right) \underline{1} = \bar{x} \cdot \underline{1}$$

$$\pi_{\sum_{i=1}^g \bar{x}_i \underline{u}_i | \underline{1}} = \frac{\underline{1} \underline{1}^T}{\underline{1} \underline{1}} (\sum_{i=1}^g \bar{x}_i \underline{u}_i) =$$

$$= \frac{1}{\sum_{i=1}^n x_i} \cdot \frac{1}{\sum_{i=1}^n \underline{u}_i} = \left(\frac{1}{n} \sum_{i=1}^n \overline{x}_i \cdot \underline{x}_i \right) \cdot \underline{1} =$$

$$= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n x_{ij} \cdot \underline{1} = \bar{x} \cdot \underline{1} = \bar{x}$$

$$\underline{x} = \bar{x} \cdot \underline{1} + \sum_{i=1}^n (\bar{x}_i - \bar{x}) \underline{u}_i + \left(\underline{x} - \sum_{i=1}^n \bar{x}_i \underline{u}_i \right)$$

(1) (2) (3)

orthogonal (all of them gaussian)

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

$$+ \left\| \underline{x} - \sum_{i=1}^n \bar{x}_i \underline{u}_i \right\|^2$$

\leftarrow S mean \leftarrow S treat

$$\begin{aligned} S_{obs} &= \sum_{i=1}^n \sum_{j=1}^{n_i} x_{ij}^2 = n \bar{x}^2 + \sum_{i=1}^n (\bar{x}_i - \bar{x})^2 n_i + \\ &+ \sum_{i=1}^n \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \end{aligned}$$

S residual

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

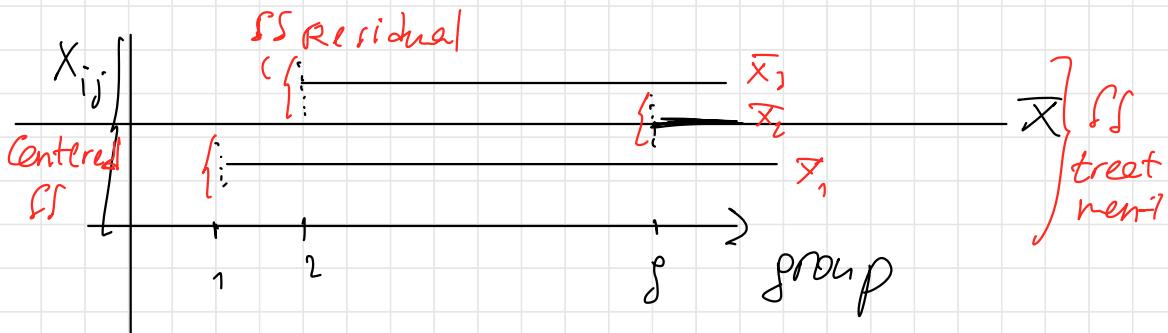
$$+ \|\underline{x} - \sum \bar{x}_i \underline{u_i}\|^2$$

$$\sum_{i=1}^g \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^g (\bar{x}_i - \bar{x})^2 n_i + \sum_{i,j} (x_{ij} - \bar{x}_i)^2$$

Centered \sum_{obs}

\sum_{treat}

\sum_{res}



decompose total variability to
treat and residual variability

$$H_0: \mu_1 = \mu_2 = \dots = \mu_g \quad \text{vs} \quad \mu_1, \dots, \mu_g$$

reject when SS_{treat} large compared
to SS_{res}

So variability among treatment
treatment difference among
means

Reject H_0 if

$$\sum_{i=1}^q (x_i - \bar{x})^2 n_i$$

is large

$$\frac{\sum_{i=1}^q \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{\sum_{i=1}^q n_i}$$

? dist if H_0

is true

$$\underline{x} \sim N_h = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_q \end{pmatrix}, \delta^2 I$$

= (1), (2), (3) are Gaussian, because
obtained from \underline{x} through linear
transformation

- (1) (2) (3) are stochastically independent

Ex \underline{P} is orthogonal projection

$$P \times P \quad (\underline{P} = \underline{P}^T \quad \underline{P}\underline{P} = \underline{P})$$

$$\underline{x} \sim N_p(\underline{\mu}, \Sigma)$$

$$\Rightarrow \underline{P}\underline{x} \perp\!\!\!\perp (\underline{I} - \underline{P})\underline{x}$$

hint: $\begin{pmatrix} \underline{P} \\ \underline{I} - \underline{P} \end{pmatrix} \underline{x} \sim ?$ compute covariance

Note: $\sum_{i=1}^q \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^q (n_i - 1) s_i^2 \sim \sigma^2 \chi^2_{(n-q)}$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \quad \text{est of } \sigma^2$$

$$(n_i - 1) s_i^2 \sim \sigma^2 \chi^2_{(n_i - 1)}$$

- If H_0 is true (same μ_i)

$$\sum_{i=1}^g \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = (n-g) s^2 \sim \sigma^2 \chi^2_{(n-g)}$$

$$s^2 = \frac{1}{n-g} \sum_i \sum_j (x_{ij} - \bar{x})^2 \text{ est of } \sigma^2$$

H_0 :

$$\sum_{i=1}^g \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^g (\bar{x}_i - \bar{x})^2 n_i + \sum_{i=1}^g \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

$$\sim \sigma^2 \chi^2_{(n-g)}$$

$$\sim \sigma^2 \chi^2_{(g-1)} \sim \sigma^2 \chi^2_{(n-g)}$$

when
 H_0 true

F-statistic

$$\sum_{i=1}^g (\bar{x}_i - \bar{x})^2 n_i$$

$$F = \frac{\frac{\sum_{i=1}^g (\bar{x}_i - \bar{x})^2 n_i}{g-1}}{\frac{\sum_{i=1}^g \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{n-g}}$$

$$\sim F(g-1, n-g)$$

Reject H_0 at level $\alpha \in (0, 1)$ if

$$F > F_{1-\alpha}(g-1, n-g)$$

Case 1 $p \geq 1, g \geq 2$

(MANOVA)

$$\text{IR}^P \ni \underline{x}_{ij} = \underbrace{\mu + \underline{\gamma}_i}_{M_i} + \varepsilon_{ij} \quad i=1 \dots g$$
$$\varepsilon_{ij} \text{ i.i.d. } \sim N_p(\underline{0}, \Sigma) \quad j=1 \dots n_i$$

$$\sum_{i=1}^g n_i \underline{\gamma}_i = \underline{0}$$

For component $k = 1, \dots, P$

$$\begin{cases} \underline{x}_{ijk} = \mu_k + \underline{\gamma}_{ik} + \varepsilon_{ijk} \\ \varepsilon_{ijk} \text{ i.i.d. } \sim N_p(\underline{0}, \Sigma_{kk}) \end{cases} \quad \sum_{i=1}^g n_i \underline{\gamma}_{ik} = 0$$

ANOVA for component k

If Σ is diagonal (independent component) \Rightarrow can run p ANOVAs

But ? $\Sigma_{k\ell}$, $k, \ell = 1 \dots p$, $k \neq \ell$
 (dependence between humidity and size important)

Goal $H_0: \underline{\mu}_1 = \dots = \underline{\mu}_p$ vs $H_1: H_0$ c
 equivalent: $\underline{\Sigma}_1 \dots = \underline{\Sigma}_p = 0$ vs $H_1: H_0$ c

Covariance Decomposition

Correlation
tg

$$\sum_{i=1}^p \sum_{j=1}^{n_i} (\underline{x}_{ij} - \bar{x})(\underline{x}_{ij} - \bar{x})^T = B$$

$$\bar{x} = \frac{1}{n} \sum_i \sum_j \underline{x}_{ij}$$

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \quad i = 1 \dots p$$

$$= \sum_i^n (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T + \sum_i \sum_{j=1}^{n_i} (\underline{x}_{ij} - \bar{x}_i)(\underline{x}_{ij} - \bar{x}_i)^T$$

$$W = \sum_{i=1}^g (n_i - 1) S_i$$

$$S_T = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\underline{x}_{ij} - \bar{x}_i)(\underline{x}_{ij} - \bar{x}_i)^T$$

Variability between groups $\xrightarrow{\text{if } \beta \neq 0}$ -?
 Variability within groups $\xrightarrow{\text{if } \beta = 0}$

So when $p \geq 1$ M^AN^OV^A - take into account correlation among features and can indicate difference, which can not be indicated by A^NO^VA.

But if we want see which variable exactly have difference in groups.

If we have g groups \Rightarrow we need run $C_g^2 = \frac{g!}{2!(g-2)!} = \frac{g(g-1)}{2}$ tests, to see which group

have different mean, so to check p variables in g groups I will run $p \cdot \frac{g(g-1)}{2}$ tests \Leftarrow M^AN^OV^A

Multivariate ANOVA MANOVA

20. 02. 25

$$\mathbb{R}^P \ni \underline{x}_{ij} = \underbrace{\underline{\mu} - \underline{\gamma}_i}_{\perp \text{ ij}} + \underline{\varepsilon}_{ij} \quad \begin{array}{l} \underline{\mu} \in \mathbb{R}^P \text{ unknown} \\ \underline{\gamma}_i \in \mathbb{R}^P \quad i=1 \dots g \end{array}$$

$\underline{\varepsilon}_{ij} \text{ i.i.d } \sim N_p(\underline{0}, \Sigma)$

$i = 1 \dots g$

$j = 1 \dots n_i$

$$\sum_{i=1}^g n_i \underline{\gamma}_i = \underline{0}$$

assume
always the
same



variability
between groups

Decomposition of cov

$$\sum_{i=1}^g \sum_{j=1}^{n_i} (\underline{x}_{ij} - \bar{\underline{x}}) (\underline{x}_{ij} - \bar{\underline{x}})^T = \sum_{i=1}^g (\underline{x}_i - \bar{\underline{x}}) (\underline{x}_i - \bar{\underline{x}})^T n_i$$

$$\perp \sum_{i=1}^g \sum_{j=1}^{n_i} (\underline{x}_{ij} - \bar{\underline{x}}_i) (\underline{x}_{ij} - \bar{\underline{x}}_i)^T$$

variability within

if $\bar{\underline{x}}_i = \underline{0} \Rightarrow \bar{\underline{x}} = \underline{0} \quad - ?$

$$H_0: \bar{Z}_1 = \bar{Z}_2 = \dots = \bar{Z}_g = \underline{\bar{Z}} \quad \text{vs} \quad H_1: \bar{Z}_i \neq \underline{\bar{Z}}$$

↑

$$\mu_1 = \dots = \mu_g = \mu \quad \text{vs} \quad \mu_1 = \mu_0^c$$

Test statistics:

$$\Lambda_w = \frac{\text{Det}(W)}{\text{Det}(W + \beta)}$$

Wilks - Lambda Proposed

if Λ_w is large it means that the generalised variance within groups is not so different from the overall generalised variance

this means that the treatment didn't produce much effect: the treatment didn't increase the variability among groups

Reject H_0 if Λ is small
(lot of vari- between methods)
ability

Pillai - Lambda Proposed

$$\Lambda_p = f_2 \left(\beta \cdot (\beta + W)^{-1} \right)$$

Reject H_0 if Λ_p is large

Lowley-Kotelling

$$\Lambda_{LH} = f_2(\beta w^{-1})$$

Reject if H_0 if Λ_{LH} is large

The dist of Λ_w when H_0 is true
is known when

- $p \geq 1$ and $g=1, j$
 - $p=2$ and $g \geq 1$
-

Rmk: $\Lambda_w, \Lambda_p, \Lambda_{LH}$ are all based on

the eigenvalues $\lambda_1, \dots, \lambda_g$ of βw^{-1}

$$S = \min(g-1, p) \quad (\text{more groups, than features})$$

Note: β is $p \times p$ matrix, and we obtain it from vectors $\underline{x}_1 - \underline{x}$ which live in

$(g-1)$ dimensional space: if $g-1 < p$, then
 $\det(B) = 0 \Rightarrow$ There are only
 s eigenvectors where $s = \min(g-1, p) =$
 $= \text{rank}(B)$

Therefore we don't take $\frac{\det(B)}{\det(w)}$
as test statistics because most times
 $\det(B) = 0$.

In the above what does it mean
large or small? we would need to
know the distribution of Λ under H_0
above, but we don't know them:
they are generally unknown, so we
capture them through simulations
(e.g. Monte Carlo)

Bartlett's approximation (when H_0 is true):

$$\approx \left(n - 1 - \frac{p+g}{2} \right) \log(\Lambda_W) \sim \chi^2(p(g-1))$$

Reject H_0 if $\left(n - 1 - \frac{p+g}{2} \right) \log \Lambda_W > \chi^2_{1-\alpha}(p(g-1))$

If you reject H_0 . \Rightarrow Bonferroni CI for

Estimator for $Z_{ie} - Z_{ke}$

$$(\bar{x}_{ie} - \bar{x}_e) - (\bar{x}_{ke} - \bar{x}_e) = k_i \bar{x}_e - k_e \bar{x}_e$$

$$= \bar{x}_{ie} - \bar{x}_{ke} \sim N(Z_{ie} - Z_{ke})$$

$$\frac{\delta_{ie}}{n_i} + \frac{\delta_{ke}}{n_e}$$

(δ_{ee} element ll of Σ)

- $W = \sum_{i=1}^g \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(\bar{x}_{ij} - \bar{x}_i)' =$

(what gave most effect to treatment (which feature))

$$= \sum_{i=1}^g (n_i - 1) s_i \quad s_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_i)}{(x_{ij} - \bar{x}_i)}$$

↑
estimator of
 Σ in group i

$\frac{1}{n-g}$ w pooled est
of Σ

respect to degree of freedom,

because we lost degree of freedom after projection

• Estimator of ϕ_{ee} : $\frac{1}{n-g} w_{ee}$

$$\text{Bonf CI}_{1-\alpha} (\hat{\epsilon}_{ie} - \hat{\epsilon}_{ke}) = [\bar{x}_{ie} - \bar{x}_{ke} \pm$$

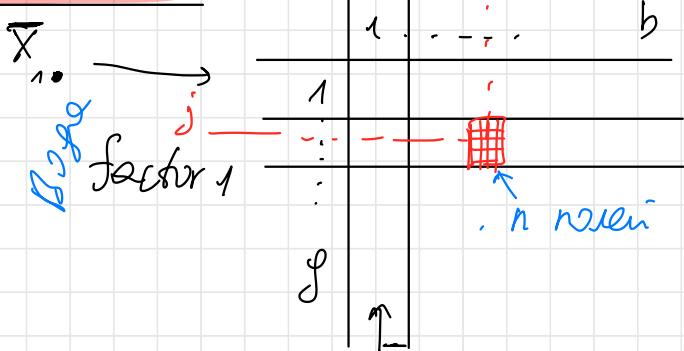
$$\pm t_{1-\frac{\alpha}{2}} \cdot \frac{1}{n-g} \sqrt{\frac{1}{n-g} w_{ee}}$$

$$* = p \cdot \frac{g(g-1)}{2}$$

↑ number of comparisons in $\hat{\epsilon}_{ie} - \hat{\epsilon}_{ke} \in \mathbb{R}^p$

Two-way (M)ANOVA

2 factors



$$X_{ijk} \quad i = 1 \dots g$$

$$j = 1 \dots b$$

$$k = 1 \dots h_{ij} = n \quad (\text{balanced experiment})$$

same sample size in each group (assume)

$$X_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \text{ i.i.d. } \sim N(0, \sigma^2)$$

degrees of freedom: $g \cdot b$

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

α_i - effect of factor 1 at level i

β_j - effect of factor 2 at level j

γ_{ij} - interactions between the two factors

constraints:

$$0 = \sum_{i=1}^g \alpha_i = \sum_{j=1}^b \beta_j = \sum_{i=1}^g \sum_{j=1}^b \gamma_{ij} = \sum_{j=1}^b \sum_{i=1}^g \gamma_{ij}$$

Estimators

$$\bar{X} = \frac{1}{g^b n} \sum_{i=1}^g \sum_{j=1}^b \sum_{k=1}^n x_{ijk} \quad \text{est of } \mu$$

$$\bar{X}_{i \cdot} = \frac{1}{b \cdot n} \sum_{j=1}^b \sum_{k=1}^n x_{1jk} \quad \text{est of } \mu_i$$

$$\bar{X}_{\cdot j} = \frac{1}{g^n} \sum_{i=1}^g \sum_{k=1}^n x_{i1k} \quad \text{est of } \mu_j$$

$$\bar{X}_{i \cdot} - \bar{X} \quad \text{estimator of } \gamma_i$$

$$\bar{X}_{\cdot j} - \bar{X} \quad \text{estimator of } \beta_j$$

$$\bar{x}_{ij} = \frac{1}{n} \sum_{k=1}^n x_{ijk} \quad \text{- est of } f_{ij}$$

$$\bar{x}_{ij} - (\bar{X}_{i \cdot} - \bar{X}) - (\bar{X}_{\cdot j} - \bar{X}) - \bar{X}$$

$$\bar{x}_{ij} - \bar{X}_{i \cdot} - \bar{X}_{\cdot j} + \bar{X} \quad \text{- est of } f_{ij}$$

$$\text{Decomposition of variance} \quad \text{Var} \quad df$$

$$\sum_{l=1}^g \sum_{j=1}^b \sum_{k=1}^n (x_{ijk} - \bar{x})^2 = \text{SS center } g b n - 1$$

$$= \sum_{i=1}^g (\bar{x}_{i\cdot\cdot} - \bar{x})^2 n_b + \text{SS treat}_1 g - 1$$

$$+ \sum_{j=1}^b (\bar{x}_{j\cdot\cdot} - \bar{x})^2 n_g \quad \text{SS treat 2 } b - 1$$

$$+ \sum_{i=1}^g \sum_{j=1}^b \sum_{k=1}^n (x_{ijk} - \bar{x}_{i\cdot\cdot} - \bar{x}_{\cdot j} + \bar{x})^2 n \quad \text{SS inter } (g-1)(b-1)$$

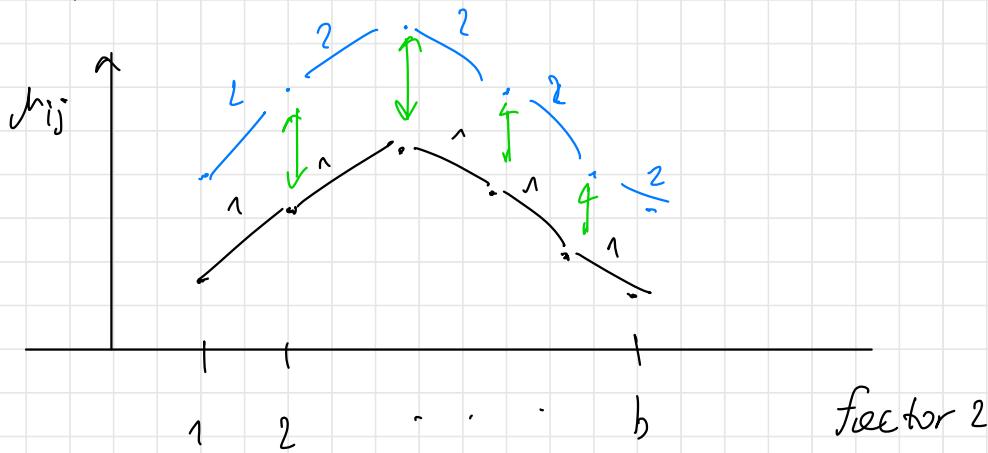
$$+ \underbrace{\sum_{i=1}^g \sum_{j=1}^b \sum_{k=1}^n (x_{ijk} - \bar{x}_{ij\cdot})^2}_{(*)} \quad \text{SS residual } g b (n-1)$$

$$\frac{(*)}{g b (n-1)} \text{ est of } \sigma^2$$

$$\mu_{ij} = \mu + \gamma_i - \beta_j + f_{ij} \quad (\text{complete model})$$

$$\mu_{ij} = \mu - \gamma_i + \beta_j \quad (\text{additive model})$$

Additive model



$$\mu_{ij} = \mu + \gamma_i + \beta_j$$

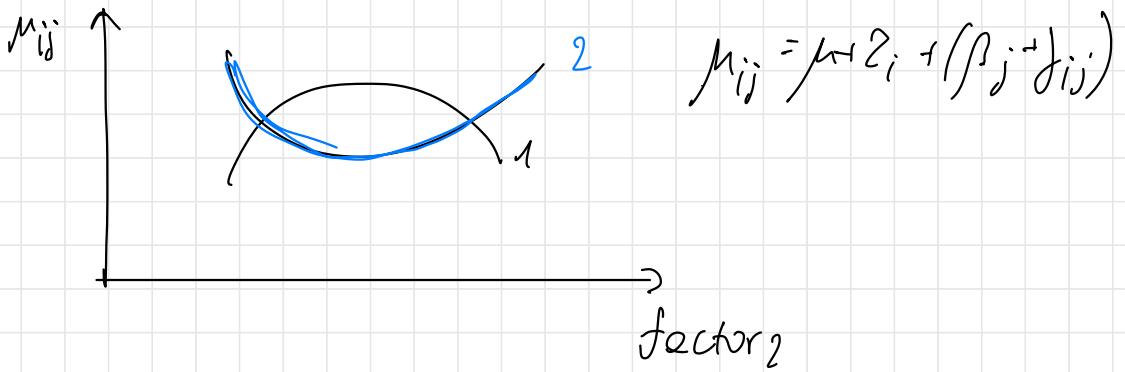
factor 1

$$\mu_{1j} = \mu + \gamma_1 - \beta_j$$

So factor 1
derives & depends on
factor 2

2

$$\begin{aligned}\mu_{2j} &= \mu + \gamma_2 - \beta_j \\ &= \mu_{1j} + (\gamma_2 - \gamma_1)\end{aligned}$$



factor 1 depends on factor 2

We can still put in ANOVA with lot of factors — , but then we will
 —
 —
 ...
 —

lose interactions of factor

Test

$$H_0: \beta_{ij} = 0 \quad \text{vs} \quad H_1: \beta_{ij} \neq 0$$

If H_0 is true

$$\frac{SS_{\text{interact}}}{\frac{(g-1)(n-1)}{SSE_{\text{res}}}} \sim F((g-1)(b-1), gb(n-1))$$

$$h_0 : z_i = 0 \quad \forall i$$

$$\frac{\text{Sd tree}_1}{\frac{g-1}{\text{Sd res}} \sim f(g-1, gb(n-1) + (g-1)(b-1))}$$
$$\frac{gb(n-1) + (g-1)(b-1)}{gb(n-1) + (g-1)(b-1)}$$

Classification

Each unit is represented by (\underline{x}, l)
 $\underline{x} \in \mathbb{R}^p$, $l \in \{1, 2, \dots, p\}$

Supervised classification

Training data classifier Goal - Learn (from data)

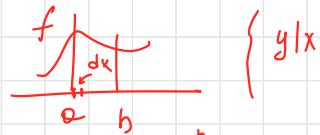
$$\underline{x} \begin{bmatrix} x_1 & \dots & x_p \\ x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_n \end{bmatrix} \begin{array}{c} f: \mathbb{R}^p \rightarrow \{1, 2, \dots, p\} \\ f(\underline{x}) - \text{group membership} \\ (\text{Discriminant Analysis}) \end{array}$$

Unsupervised Classification

l - hidden, (don't know labels)
Goal: estimate $\hat{l}_1, \dots, \hat{l}_n$ (Clustering)

24.02.25

$f(x) = t \Rightarrow$ unit for which class x is attached to group t



Ingredient for building $f: P(x \in [a, b]) = \int_a^b f(x) dx$

① $f: \mathbb{R}^D \rightarrow [0, \infty)$ density of dist of x

when $t = i$

$$f_i(x) dx = P[x \in dx | t=i] \quad i = 1 \dots g$$

different for every group

(e.g. MANOVA) (comparing means of distributions)

② Prior probability:

$$P[t = i] = \underset{g}{\underset{i=1 \dots g}{\text{---}}}$$

$$p_i \geq 0 \text{ and } \sum_{i=1}^g p_i = 1$$

(aren't dependent on features, depend on context)

③ Cost of misclassification

$C(i|j)$ cost of attributing a unit to group i , when in fact it belongs to group j .

$$P(i|j) \geq 0 \quad \leftarrow \text{error}$$

$i, j = 1 \dots g$

$$P(i|i) = 0 \quad P(i|j) \neq P(j|i)$$

(Exemple about disease, predict richer to health person, or predict healthy to sick person (Recall more important)).

$P_{i,j}^c$ - depends on context

so we have to be ready to change

$P_{i,j}^c$ depending on dataset.

(we don't learn it from dataset)

Optimal classifier

Observe that $f: \mathbb{R}^p \rightarrow \{1 \dots g\} \iff$

\iff partition $\{R_1, \dots, R_g\}$ \leftarrow hard classifier
(each feature has label)
partition off \mathbb{R}^p

$$R_i \subseteq \mathbb{R}^p \quad i = 1 \dots g$$

1. $R_i \cap R_j = \emptyset \quad i \neq j$
2. $\bigcup_{i=1}^g R_i = \mathbb{R}^p$

$$R_i = \{ \underline{x} \in \mathbb{R}^p : f(\underline{x}) = i \} = f^{-1}(\{i\}) \quad i=1 \dots g$$

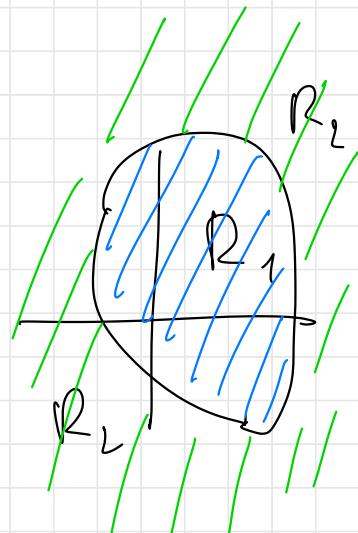
Now we need optimality criterion.

minimize Expected Cost of Misclassification

$$ECM(f)$$

Example :

$$f: \mathbb{R}^p \rightarrow \{1, 2\} \leftrightarrow \{R_1, R_2\}$$



$$ECM(f) = \int_{R_2} C(2|1) f_1(\underline{x}) p_1 d\underline{x} +$$

↑ probability of
feature 1

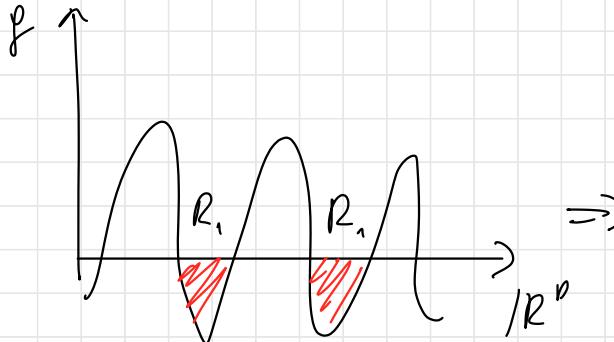
$$+ \int_{R_1} C(1|2) f_2(\underline{x}) p_2 d\underline{x} = \int_{\mathbb{R}^p} C(2|1) f_1(\underline{x}) p_1 d\underline{x} -$$

$$- \int_{R_1} C(2|1) f_1(\underline{x}) p_1 d\underline{x} + \int_{R_1} C(1|2) f_2(\underline{x}) p_2 d\underline{x}$$

$$= C(2|1) p_1 + \int_{R_1} [C(1|2) f_2(\underline{x}) p_2 - C(2|1) f_1(\underline{x}) p_1] d\underline{x}$$

↖ have to be chosen . $d\underline{x}$

And we want to chose R_1 to minimise



integral will
be negative

\Rightarrow min on ECM

$$R_1 = \{x \in \mathbb{R}^D : C(1/2)f_2(x)p_1 \leq C(2/1)f_1(x)p_2\}$$

features payed for putting to group 1 when it

$$R_1 = \{x \in \mathbb{R}^D : C(1/2)f_1(x)p_1 > C(2/1)f_2(x)p_2\}$$

Optimal groups

$$\mathcal{S} \hookrightarrow \{R_1, \dots, R_g\}, \text{ ECM}(\mathcal{S}) = \sum_{k \neq 1} \int C(k|1)f_1(x)p_1 dx$$

$$+ \sum_{k \neq 2} \int_{R_2} C(k|2)f_2(x)p_2 dx + \dots + \sum_{k \neq g} \int_{R_n} C(k|g)f_g(x)p_g dx$$

belongs to 2
but we place to group h

$$= \int_{R_1} \sum_{k \neq 1} C(g|k) f_k(x) p_k dx + \int_{R_2} \sum_{k \neq 2} C(z|k) f_k(x) p_k dx$$

$$+ \dots + \int_{R_g} \sum_{k \neq g} C(g|k) f_k(x) p_k dx$$

$$R_1 = \left\{ \underline{x} \in \mathbb{R}^D \mid \sum_{k \neq 1} C(g|k) f_k(\underline{x}) p_k \leq \sum_{k \neq j} C(j|k) f_k(x) p_k \right\}_{j \neq 1}$$

$$R_2 = \left\{ \underline{x} \in \mathbb{R}^D \mid \sum_{k \neq 2} C(z|k) f_k(x) p_k \leq \sum_{k \neq j} C(j|k) f_k(x) p_k \right\}_{j \neq 2}$$

⋮

expected cost of misclassification
when I put it in group i

$$R_i = \left\{ \underline{x} \in \mathbb{R}^D \mid \sum_{k \neq i} C(i|k) f_k(\underline{x}) p_k \leq \sum_{k \neq j} C(j|k) f_k(x) p_k \right\}_{j \neq i}$$

Optimal classifier

Rmk

1. Costs of misc can be def up to multiplied constant.

Obs

$$f(\underline{x}) = t \quad t \in \{1, \dots, g\}$$

$$\sum_{k \neq t} C(t|k) f_k(\underline{x}) p_k \leq \sum_{k \neq t} C(j|k) f_k(\underline{x}) p_k$$

$$\sum_{j=1}^g f_j(\underline{x}) p_j$$

$$\sum_{j=1}^g f_j(\underline{x}) p_j$$

$$\frac{\sum_{j=1}^g f_j(\underline{x}) p_j d\underline{x}}{\sum_{j=1}^g f_j(\underline{x}) p_j} = \frac{P[\underline{x} \in d\underline{x} | L=k] P[L=k]}{\sum_{i=1}^g P[\underline{x} \in d\underline{x} | L=i] P[L=i]}$$

$P[\underline{x} \in d\underline{x}]$

$$= \frac{P[\underline{x} \in d\underline{x} | L=k] P[L=k]}{P[\underline{x} \in d\underline{x}]} = P[L=k | \underline{x} \in d\underline{x}]$$

Bayes Th

Opt Class $\mathcal{S} \hookrightarrow \{R_1, \dots, R_g\}$

$$R_i = \left\{ \underline{x} \in \mathbb{R}^p : \sum_{k \neq i} C(i|k) P[L=k | \underline{x} = \underline{x}] \leq \right\}$$

$$\leq \sum_{k=j}^C c(j|k) P[L=k | \underline{x} = \underline{x}], j \neq i \}$$

Kopone sumintse u observed specific value
 chazan no no wall for my features
 i, kogo oh k we take some cost for all
 qvaluno, kox chazan wece j

Rmk Assume $c(i|j) = \infty$ if $i \neq j$

$$c(i|i) = 0 \quad i, j = 1 \dots p$$

Opt class $\delta(\underline{x}) = t \Leftrightarrow \sum_{k \neq t} P[L=k | \underline{x} = \underline{x}] \leq$

$$\leq \sum_{k \neq t} P[L=k | \underline{x} = \underline{x}] \text{ for } j \neq t$$

$$1 - P[L=t | \underline{x} = \underline{x}] \leq 1 - P[L=j | \underline{x} = \underline{x}] \text{ if } t$$

$$P[L=t | \underline{x} = \underline{x}] \geq P[L=j | \underline{x} = \underline{x}]$$

$$R_i = \{ \underline{x} \in \mathbb{R}^n : P[L=i | \underline{x} = \underline{x}] \geq P[L=j | \underline{x} = \underline{x}] \}$$

Bayes classifier attribute the labels to the group which maximises the posterior probability

optimal classifier, when missclassification
cost same for every missclassifications.

Assume

- $c(i|j) = C > 0$ if j
- $c(i|i) = 0$
- $P_1 = P_2 = \dots = P_g = \frac{1}{g}$ (uniform distributed prior probabilities)

- we don't specify neither the costs nor the priors but only the density of the features in the different groups
- it's not that we don't specify them: it's just that since we assume they are equal they simplify

\Rightarrow Optimal classifier

$$f(\underline{x}) = t \Leftrightarrow P[l=t | \underline{X}=\underline{x}] \geq P[l=j | \underline{X}=\underline{x}] \quad j \neq t$$

\downarrow (also simplified denominator)

Maximum

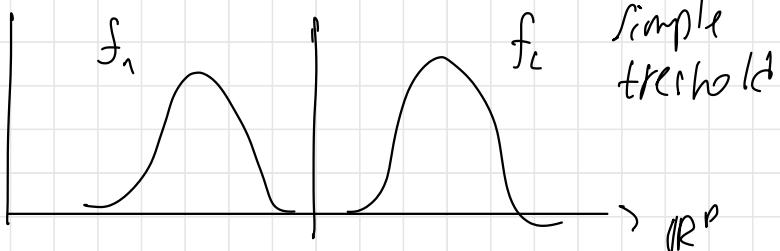
$$f_t(\underline{x}) P_t \geq f_j(\underline{x}) P_j$$

likelihood classifier

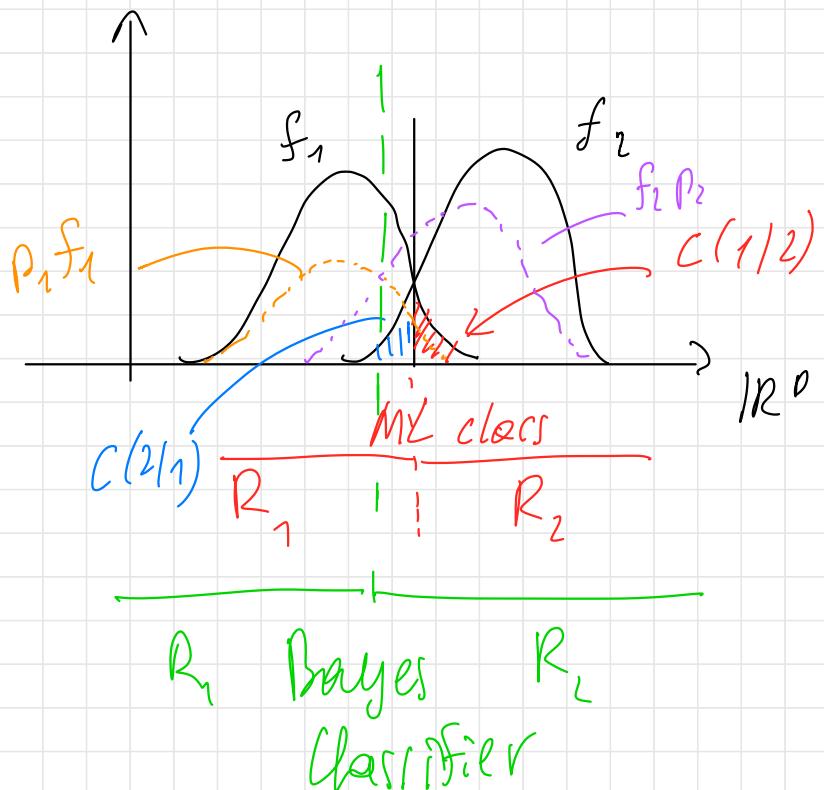
(we make very

strong assumption about data

$$\begin{aligned} \underline{Ex} \\ g=2 \end{aligned}$$



(***)



$$P_2 > P_1$$

Observation

Assume $c(i|j) = c_j > 0 \quad i, j = 1 \dots p$
 $c(i|i) = 0$

Opt classifier $f(x) = t \in \{1 \dots g\} \Leftrightarrow$

$$\Leftrightarrow \sum_{k \neq t} (C_k f_k(x) P_k) \leq \sum_{k \neq t} (C_k f_k(x) P_k)$$

(Can not divide by $C_k P_k$, because it different sums)

$$\text{Def } \pi_k = \frac{C_k p_k}{\sum_{i=1}^k C_i p_i} \geq 0 \quad \sum \pi_k = 1 \Rightarrow$$

$$\sum_{k \neq t} f_k(x) \pi_k \leq \sum_{k \neq j} f_k(x) \pi_k \quad j \neq t$$

→ Bayes Classifier with priors $\pi_1 \dots \pi_g$

Exercise work out the optimal classifier,
when



$$1. \quad C(i|j) = C_i > 0 \quad i \neq j$$

$$C(i|i) = 0 \quad i, j = 1 \dots g$$

$$2. \quad C(i|j) = C_i \cdot h_j \quad i = j \\ > 0 \quad > 0$$

$$C(i|i) = 0$$

$$C(i|j) = C^{x_i} h^{\beta_j} \quad x_i \beta_j \geq 1 \\ C h > 0$$

$$\sum_{k \neq t} c(t|k) f_k(x) p_k \leq \sum_{k \neq j} c(j|k) f_k(x) p_k$$

$$\sum_{j=1}^g f_j(x) p_j$$

$$\sum_{k \neq t} c(t) f_k(x) p_k \leq \sum_{k \neq j} c(j) f_k(x) p_k$$

From the figure (* * *)

In the figure we see the partition we would get from the MLE classifier. It's like saying that if we see a very tall person we suppose its male! What's the error? The integral of f_1 over R_1 plus the integral of f_2 over R_1 (shaded area in the figure)

Suppose we have now a prior: we pick individuals at random in the mechanical engineering department: 80% are males.

So the prior is different because we are considering a specific population with the proportion between male and female different from the proportion in the true whole population!

So we multiply the two previous densities by p_1, p_2 the density are now re-scaled: see again the figure above and in the lastest ~~orange~~ purple lines we can see the Bayes classifier!

Note that the Bayes Classifier has moved the threshold of the MLE classifier to the left, the prior has modified

the density and we are more conservative and make less errors!

So now if we take different cost of misclassification it will have some effect: we weight again differently the two densities.

That's why prior and costs are important: same problem in a different context has different prior and costs but same densities f_1, f_2 .

Example: the prior distribution of how people are dangerous at night is higher than in the morning! We don't move date only the prior: just by doing this we change the way we classify the same individual.

25. 03. 25

Bayes continue...

Special Bayes Classifier

Suppose the prior is Gaussian

$$\underline{x} | L=i \sim N_p(\mu_i, \Sigma_i) \quad i = 1 \dots g$$

Consider Bayes Class

$$f_t(\underline{x}) = t \in \{1, \dots, g\} \Rightarrow P[L=t | \underline{x}] \geq P[L=j | \underline{x}]$$

$$\Leftrightarrow f_t(\underline{x}) p_t \geq f_j(\underline{x}) p_j$$

$$p_t \frac{1}{((2\pi)^p |\Sigma_t|)} \exp\left(-\frac{1}{2} (\underline{x} - \mu_t)^\top \Sigma_t^{-1} (\underline{x} - \mu_t)\right) \geq$$

$$\geq p_j \frac{1}{((2\pi)^p |\Sigma_j|)} \exp\left(-\frac{1}{2} (\underline{x} - \mu_j)^\top \Sigma_j^{-1} (\underline{x} - \mu_j)\right)$$

taking log

$$\log p_t - \frac{1}{2} \log |\Sigma_t| - \frac{1}{2} (\underline{x} - \mu_t)^\top \Sigma_t^{-1} (\underline{x} - \mu_t) \geq$$

$$\Rightarrow \log p_j - \frac{1}{2} \log |\Sigma_j| - \frac{1}{2} (\underline{x} - \underline{\mu}_j)^T \Sigma_j^{-1} (\underline{x} - \underline{\mu}_j)$$

Mahalanobis distance

Closer to $\mu_j \Rightarrow$ higher prob, b/c we belong to this group

Closer $\downarrow \Rightarrow \underbrace{\log p_j - \frac{1}{2} \log |\Sigma_j| - \text{dist} \downarrow}$

higher prob $\Leftrightarrow \uparrow$

$$d_i^Q(\underline{x}) = \log p_i - \frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (\underline{x} - \underline{\mu}_i)^T \Sigma_i^{-1} (\underline{x} - \underline{\mu}_i) \quad i=1\dots g$$

QDA

quadratic discriminant scores

Bayes Class: $\{ \subseteq \{R_1 \dots R_g\} \}$ (QDA):

$$R_i = \{ \underline{x} \in \mathbb{R}^q : d_i^Q(\underline{x}) \geq d_j^Q(\underline{x}), j \neq i \}$$

to whic μ_i \underline{x} is closer

In this case the distribution of the features is Gaussian in each group!

Since it's Bayes Classifier all cost are equal, but as seen above we can have special cost structure!

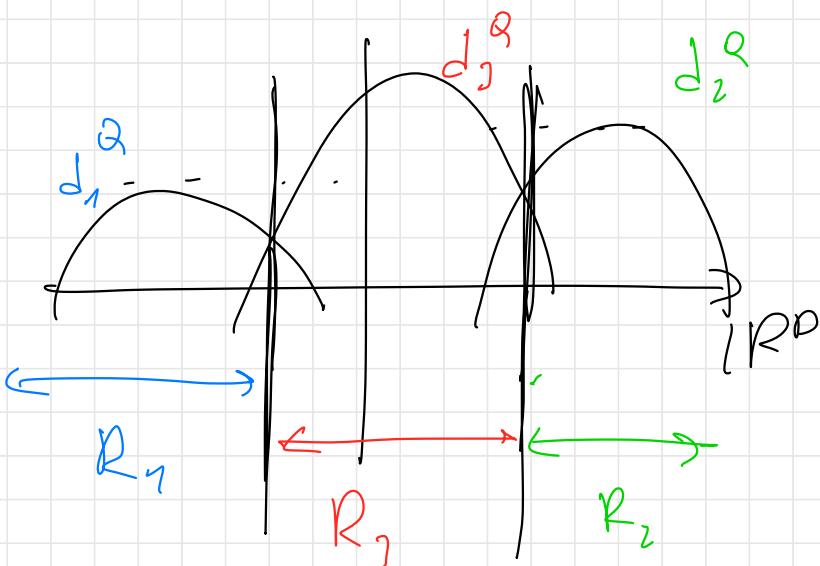
Note: this is very used: we only need to estimate the μ_i and Σ_i for each group

Note: if we are in a space of dimensions more than 3, we consider the boundaries to check out the different partitions!

Note: the fact that QDA is quadratic comes from the term $\frac{1}{2}(\underline{x} - \mu_i)^T \Sigma_i^{-1} (\underline{x} - \mu_i) = \underline{\Sigma}^{-1}(\underline{x}, \mu_i)^2$

QDA says: the closest (with respect to the Mahalanobis Distance) you are to mean of μ_i the more probable your label is from that class!

Example:



So when d_i^Q larger \rightarrow there will be class i .

Assumption $\sum_i = \sum$ for every i ← LDA

$$X | L = i \sim N_p(\mu_i, \Sigma) \quad i = 1 \dots g$$

Bayes classifier

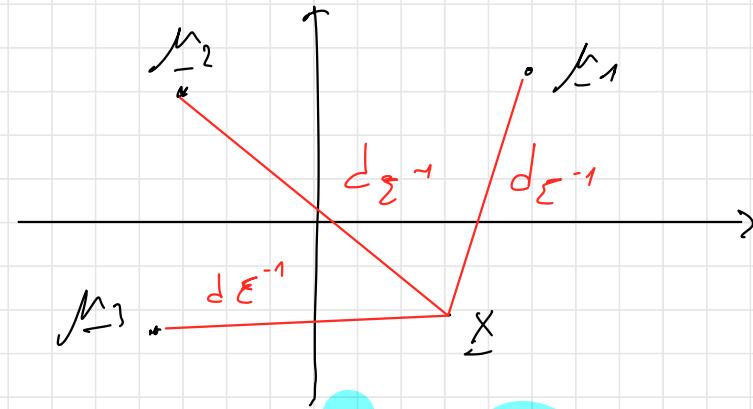
$$f(x) = t \in \{1, \dots, g\} \Leftrightarrow$$

$$\log p_i - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x - \mu_i) \Sigma^{-1} (x - \mu_i) \geq$$

$$\log p_j - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x - \mu_j) \Sigma^{-1} (x - \mu_j)$$

if $p_i = p_j$ (same distribution)

\Rightarrow we will take class closest to $\sqrt{\text{mean}}$

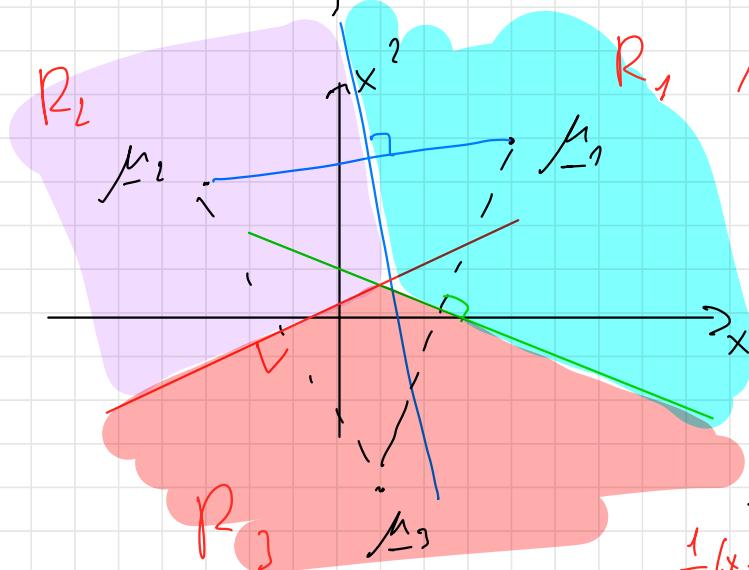


its
mean

R^P

Mahalanobis
distance

(not euclidean)



if $\Sigma = I \Rightarrow$
Euclidean

so if $\Sigma = I$

//

linear borders

$$\frac{1}{2}(\underline{x} - \mu_i)^T \Sigma^{-1} (\underline{x} - \mu_i) = -\frac{1}{2}\underline{x}^T \Sigma^{-1} \underline{x} + \mu_i^T \Sigma^{-1} \underline{x} - \frac{1}{2}\mu_i^T \Sigma^{-1} \mu_i$$

SVM, kernel trick - ?

$$\log p_i - \frac{1}{2} \underline{x}^T \Sigma^{-1} \underline{x} - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \mu_i^T \Sigma^{-1} \underline{x}$$

$$\geq \log p_j - \frac{1}{2} \underline{x}^T \Sigma^{-1} \underline{x} - \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j + \mu_j^T \Sigma^{-1} \underline{x}$$

$$d_i(\underline{x}) = \log p_i - \frac{1}{2} \underline{\mu}_i^T \Sigma^{-1} \underline{\mu}_i + \underline{\mu}_i^T \Sigma^{-1} \underline{x}$$

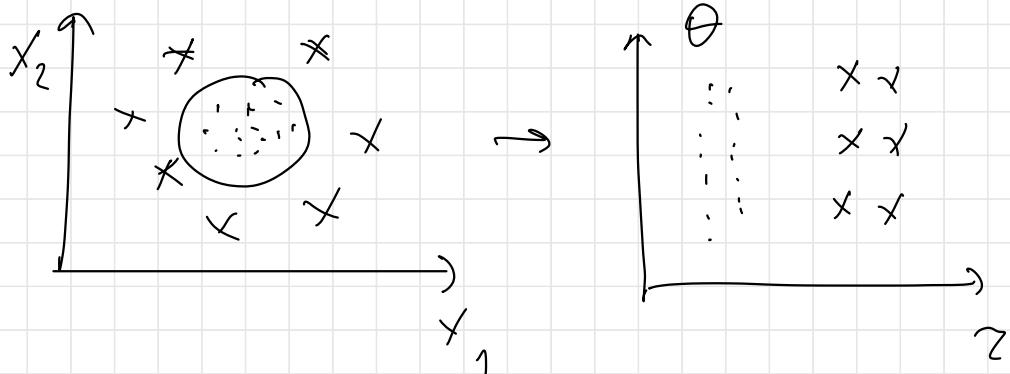
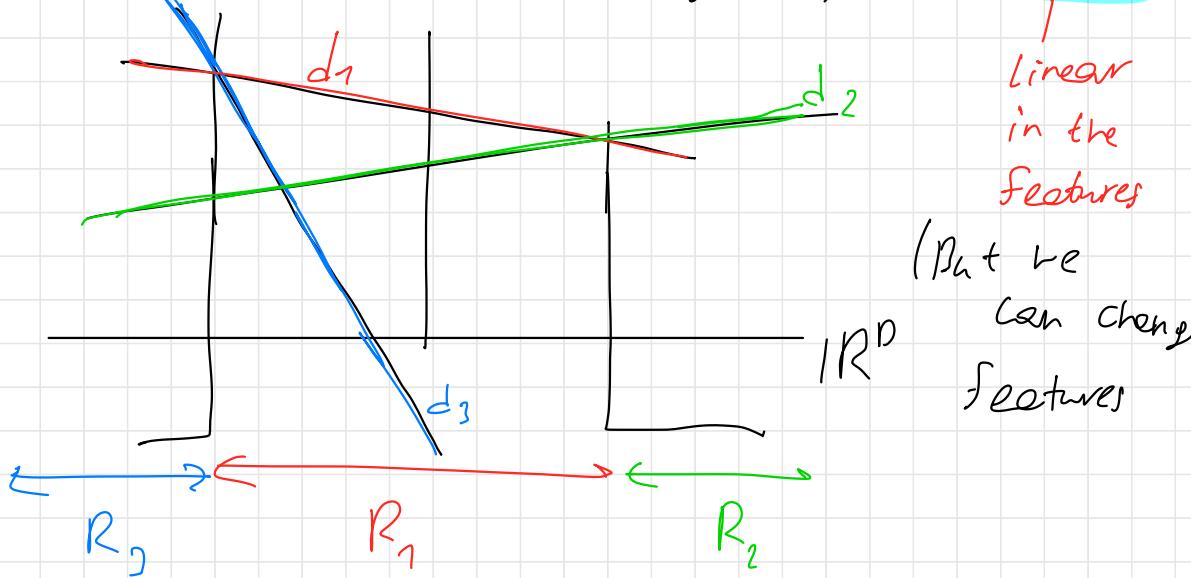
β_0 β_1

linear discriminant score

Bayes Class (LDA) $f(\underline{x}) \leftarrow \{R_1, \dots, R_y\}$

$$R_i = \{ \underline{x} \in \mathbb{R}^D : d_i(\underline{x}) \geq d_j(\underline{x}), j \neq i \}$$

LDA



Conclusion

- QDA Gaussian data, different means and variances in each group
- LDA Gaussian data, different means but equal variances in each group.

Training set

$$\mathcal{X} \left(\begin{array}{c} \underline{x}_1 \\ \vdots \\ \underline{x}_n \end{array} \right) \left(\begin{array}{c} \ell_1 \\ \vdots \\ \ell_n \end{array} \right) \quad \underline{x}_i \in \mathbb{R}^p \\ \ell_i \in \{1, \dots, g\}$$

$h \times p$

use data to "learn"

f_i 's i.e. the dist of \underline{x} in group i

for $i = 1 \dots g$.

prior probability

But don't use training set to define π_i

We may use prior probabilities from Open resources, not from our data.

In QDA: Estimate μ_i with

$$\widehat{\underline{\mu}}_i = \frac{1}{n_i} \sum_{\{j: \ell_j = i\}} \underline{x}_j \quad n_i = |\{j: \ell_j = i\}|$$

$$\Sigma_i \text{ with } S_i = \frac{1}{n_{i-1}} \sum_{\substack{j \\ j: c_j = i}} (\underline{x}_j - \bar{\underline{x}}_i)(\underline{x}_j - \bar{\underline{x}}_i)^T$$

and then plug in the estimators in \hat{d}_i .

In LDA

Estimate Σ with

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

(like weighted average)

- If p large wrt n_i 's \Rightarrow parametrize Σ_i , so if Σ_k is not diagonal \Rightarrow probably it will not be invertible

Noise Bayes \rightarrow Uncorrelated features

$$\Sigma_k = \begin{bmatrix} \tilde{\sigma}_{11}^{(k)} & & & \\ & \ddots & & \\ & & \ddots & \\ 0 & & & \tilde{\sigma}_{pp}^{(k)} \end{bmatrix}$$

component of
features uncorre-
lated

$k = 1 \dots g$

\nwarrow uncorrelated
in class k

$$d_k^Q(\underline{x}) = \log p_k - \frac{1}{2} \sum_{j=1}^P \log \delta_{jj}^{(k)}$$

$$- \frac{1}{2} \sum_{j=1}^P \frac{(x_j - \mu_j)}{\delta_{jj}^{(k)}}$$

$$\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_P \end{bmatrix}$$

Note: it's stable even if $p \gg n$
 Note: we need at least two observations in each group for each component, otherwise we can not estimate variance

Note: doing PCA first and then classification:
this is bad recipe!

Fisher's argument for LDA

FDA

- Robustness of LDA wrt

No Gaussian assumption.

No Gaussian Assumption

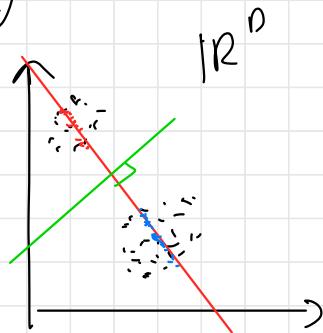
- Dim reduction

$$\underline{x} | L=i \sim \mu_i (\Sigma), i=1 \dots p$$

$$\underline{x} \in \mathbb{R}^p$$

$$E[\underline{x}' \underline{x} | L=i] = \underline{\mu}' \underline{\mu}$$

$$\text{Var}[\underline{x}' \underline{x} | L=i] = \underline{\alpha}' \Sigma \underline{\alpha}$$



So can I find projection, where after

projection Σ will be able separate these groups.

$$(\rho \times 1) \times (\rho \times 1) = \rho \times \rho$$

$$\mathbf{B} = \frac{1}{g-1} \sum_{i=1}^g (\mu_i - \bar{\mu}) (\mu_i - \bar{\mu})'$$

correlation between groups

$$\bar{\mu} = \frac{1}{g} \sum_{i=1}^g \mu_i$$

$$\underset{\underline{\alpha} \in \mathbb{R}^p}{\operatorname{argmax}} \frac{\underline{\alpha}' \mathbf{B} \underline{\alpha}}{\underline{\alpha}' \underline{\alpha}} = \underset{\underline{\alpha} \in \mathbb{R}^p}{\operatorname{argmax}} \frac{1}{g-1} \frac{\sum_{i=1}^g (\underline{\alpha}' \mu_i - \underline{\alpha}' \bar{\mu})^2}{\underline{\alpha}' \underline{\alpha}}$$

covariance within groups

$$\frac{\underline{\alpha}' \mathbf{B} \underline{\alpha}}{\underline{\alpha}' \underline{\alpha}}$$

lets $\sum \frac{1}{2} \underline{\alpha}' \underline{\alpha} = \underline{u}$

$$\Rightarrow \underline{\alpha} = \sum \frac{1}{2} \underline{u}$$

$$\parallel \underline{u}' \left(\sum^{-\frac{1}{2}} \mathbf{B} \sum^{-\frac{1}{2}} \right) \underline{u}$$

maximise

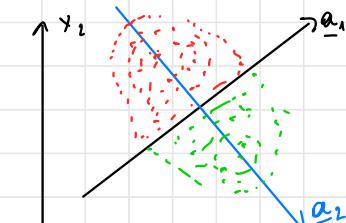
$$\sum^{-\frac{1}{2}} \mathbf{B} \sum^{-\frac{1}{2}} = \sum_{i=1}^g d_i e_i e_i^\top$$

α_1 - doesn't separate group
 α_2 - projection of two groups
 different \rightarrow separable

spec'd
decomp

$$S = \min(g-1, p) \parallel \text{degree of freedom}$$

order of Σ



$$\underset{\underline{h} \in \mathbb{R}^P}{\operatorname{argmax}} \frac{\underline{h}' \sum^{-\frac{1}{2}} \underline{\beta} \sum^{-\frac{1}{2}} \underline{h}}{\underline{h}' \underline{h}} = \underline{\ell}_1$$

↓

$$\underset{\underline{Q} \in \mathbb{R}^P}{\operatorname{argmax}} \frac{\underline{Q}' \underline{\beta} \underline{Q}}{\underline{Q}' \sum \underline{Q}} = \sum^{-\frac{1}{2}} \underline{\ell}_1 \quad Q_1 \text{ first discriminant direct}$$

$\underline{Q}_1^T \underline{X}, \underline{Q}_2^T \underline{X} \dots$ First, second, ... Fisher's Discriminant scores Score

$$\begin{cases} \underline{Q}_1 = \sum^{-\frac{1}{2}} \underline{\ell}_1 & \text{first disc direction} \\ \underline{Q}_2 = \sum^{-\frac{1}{2}} \underline{\ell}_2 & \text{second "} \\ \vdots & \\ \underline{Q}_s = \sum^{-\frac{1}{2}} \underline{\ell}_s & \text{5-th "} \end{cases}$$

So if $g=2$, $\Rightarrow g-1=1 \Rightarrow$ only one

$$A = \begin{bmatrix} \underline{Q}_1' \\ \vdots \\ \underline{Q}_s' \end{bmatrix}$$

$$A \underline{X} = \begin{pmatrix} \underline{Q}_1' \underline{X} \\ \vdots \\ \underline{Q}_s' \underline{X} \end{pmatrix}$$

dimension,
even if P
very large

$$\text{Cov}(\underline{\varrho}_i' \underline{x}, \underline{\varrho}_j' \underline{x}) = \underline{\varrho}_i' \Sigma \underline{\varrho}_j$$

$$= \underline{\varrho}_i' \Sigma^{-\frac{1}{2}} \Sigma \Sigma^{-\frac{1}{2}} \underline{\varrho}_j$$

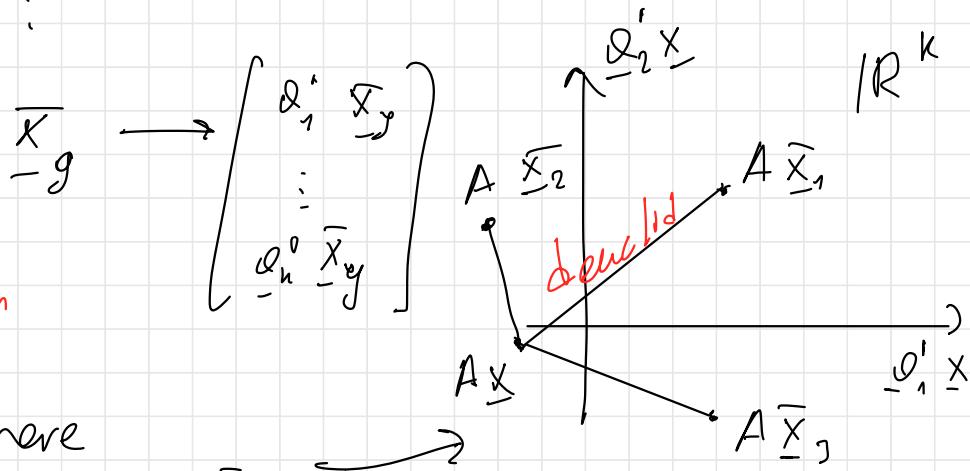
$$= \underline{\varrho}_i' \underline{\varrho}_j \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases}$$

Building a classifier, based on discriminant scores

Estimate μ_1, \dots, μ_p by $\bar{x}_1, \dots, \bar{x}_p$

$$\mathbb{R}^p \ni \bar{x}_1 \rightarrow \begin{pmatrix} \underline{\varrho}_1' \bar{x}_1 \\ \vdots \\ \underline{\varrho}_k' \bar{x}_1 \end{pmatrix}$$

$$k \leq s = \min(p, n)$$



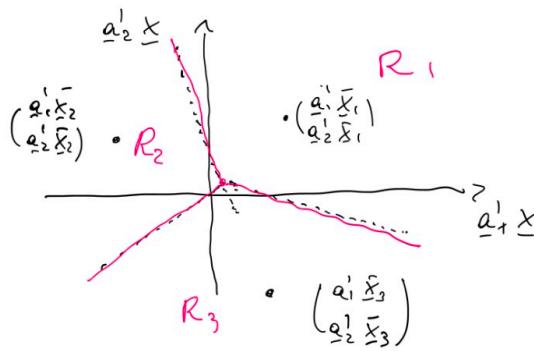
Voronoi
tessellation

here
covariance Identity

$$f(\underline{x}) = t \Leftrightarrow \sum_{j=1}^k (\underline{q}_j' \underline{x} - \underline{q}_j' \bar{\underline{x}}_t)^2 \geq \sum_{j=1}^k (\underline{q}_j' \underline{x} - \underline{q}_j' \bar{\underline{x}}_h)^2 \quad h \neq k$$

This is LDA if $k = s$, $p_1 \dots p_g = \frac{1}{g}$

For example consider the following figure:



we have three group means $(a_1^T \bar{X}_i, a_2^T \bar{X}_i)^T$ where $i = 1, 2, 3$ and we have a point $(a_1^T \underline{x}, a_2^T \underline{x})^T$

We compute the distance to this point between all the three group means above: we attribute this point to the group whose mean is closest to our new point!

Note that in the above figure all the points on one side of the boundary are closest to the group mean which identifies that boundary!

In the above we are building up the **Voronoi Tessellation** of this plane in Fisher's Coordinates: first we project everything on fisher coordinates and then we build up the **Voronoi tessellation** and then that will give us the classifier!

Note that the boundaries between the different regions are all linear: indeed we are minimising the square distance.

Theorem: this classifier is exactly LDA when we take equal priors: indeed we didn't use the prior to build up this classifier: $p_1 = \dots = p_g = \frac{1}{g}$

Note: the above is used with $k = 2, 3$ for graphical reasons: we can plot nice figures! For classification we use LDA, which is overall pretty good!

28.03.25

Evaluating a class

$$f: \mathbb{R}^p \rightarrow \{1 \dots g\}$$

a classifier

$$f \hookrightarrow \{R_1 \dots R_g\}$$

Actual Error Rate

$$AER(f) = \sum_{k=1} \int_{R_k} p_i f_i(x) dx + \sum_{k=2} \int_{R_k} p_i f_i(x) dx$$

$$+ \sum_{k \neq j} \int_{R_k} p_j f_j(y) dx$$

1- AER(f) = Accuracy(f)

Non parametric estimate of APR

Apply f to the training set

$$\hat{X}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{in} \end{pmatrix} \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \quad C_i = \{1 \dots g\} \quad i = 1 \dots n$$
$$x_i \in \mathbb{R}^p$$

for $i = 1 \dots n$

$$f(\underline{x}_i) = \hat{e}_i$$

Compare \hat{e}_i with e_i

Confusion matrix

of units in
training set belong-
ing to group i
and assigned to
group j

		True predictor	Assigned labels		
		1	2	..	g
True labels	1	0			
	2		0	n_{ij}	
:	:				
g					

$$\widehat{AER}(j) = \frac{\text{# mistakes}}{n} = 1 - \frac{\sum_{i=1}^g n_{ii}}{n}$$

A PERtent error Rate

APER (g)

overly optimistic

Ex. KNN with $k=1 \Rightarrow APER = 0$
(just take yourself and it's right prediction)

When $g = 2$ (Binary Classification)

$$f: \mathbb{R}^P \rightarrow \{0, 1\}$$

Confusion matrix

1 - Reject H_0

2 - Not reject H_0

		Assigned		Positive - we predicted ①
		1	0	
True	1	True positive n_{11}	False negative n_{10}	n_{1*}
	0	False positive n_{01}	True negative n_{00}	n_{0*}
		n_{*1}	n_{*0}	n

$$AER = P_1 \int_{R_1} f_1(x) dx +$$

$$P_0 \int_{R_0} f_0(x) dx$$

$$APER(f) = \frac{n_{10} + n_{01}}{n} = \frac{n_{1*}}{n} \cdot \frac{n_{10}}{n_{1*}} + \frac{n_{0*} \cdot n_{01}}{n n_{0*}}$$

trying estimate \rightarrow $\hat{P}_1 \int_{R_1} \hat{f}_1(x) dx + \hat{P}_0 \int_{R_0} \hat{f}_0(x) dx$
actual error rate

we to fully adapting
to the data

Precision:
$$\frac{n_{11}}{n_{1*}}$$

$$n_{11} + n_{01}$$

PPV = 1 - FPR

positive predicted value
 n_m - true discoveries

Recall:
$$\frac{n_{11}}{n_{1*}}$$

sensitivity
 $n_{11} + n_{10}$

how good in computing of diseases

$1 - P[\text{type II error}] = \text{power}$

Specificity =
$$\frac{n_{00}}{n_{0*}}$$

(we want high specificity and high sensitivity)

$1 - P[\text{type I error}] = 1 - \alpha$

F-measure (harmonic mean of Recall and Precision)

$2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$ harmonic mean

Ideal situation

test set

$$\tilde{\mathbf{X}} = \begin{bmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_N \end{bmatrix} \quad \begin{bmatrix} \hat{l}_1 \\ \vdots \\ \hat{l}_N \end{bmatrix}$$

Apply $\delta_{\tilde{\mathbf{X}}}$ to $\tilde{\mathbf{X}}$

$$\delta_{\tilde{\mathbf{X}}} (\hat{x}_i) = \hat{l}_i \quad i = 1 \dots N$$

\Rightarrow Confusion matrix

Compare \hat{l}_i with \tilde{l}_i

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

$$\varepsilon_i = \begin{cases} 1 & \text{if } \hat{l}_i = \tilde{l}_i \\ 0 & \text{else} \end{cases}$$

How to choose test set?

Best - choose randomly

(and save proportion of classes)

Cross Validation

Leave-one-out

(L1O)

n -fold CV

For $i = 1, \dots, n$

(almost unbiased, but
high variability)

$$1. \quad \mathbb{X} = \begin{bmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_i \\ \vdots \\ \underline{x}_n \end{bmatrix} \quad \begin{bmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{bmatrix}$$

Take (\underline{x}_i, e_i) out
of the training
set $\Rightarrow \mathbb{X}_{-i}$

2. Train class on $\mathbb{X}_{-i} \Rightarrow$

$$\delta_{-i} : \mathbb{R}^P \rightarrow \{1, 2, \dots, g\}$$

$$3. \quad \delta_{-i}(\underline{x}_i) = \hat{e}_i$$

4. Compare \hat{e}_i with e_i (for inst compute

$$\varepsilon_i = \begin{cases} 1 & \hat{e}_i = e_i \\ 0 & \text{otherwise} \end{cases}$$

end for all $i = 1 \dots n$

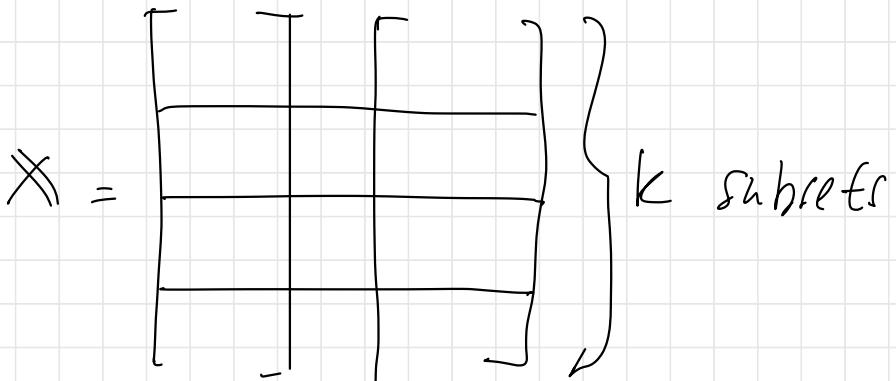
$$\widehat{\text{AER}}(\delta_{\mathbb{X}}) = \frac{\sum \varepsilon_i}{n}$$

k -fold Cross Validation

set $k < n$ ($k = 5, 10$ (practically))
if n is sufficiently large)

split

e



before doing it

- permute the order of data
(swap rows)

for $j = 1 \dots k$.

1. hold out subset j from $\mathbb{X} \Rightarrow \mathbb{X}_{-j}$
2. train classifier (may be any model)

on \mathbb{X}_{-j} $f_{-j} : \mathbb{R}^P \rightarrow \{1 \dots g\}$

3. Apply f_{-j} to subset j

For i : $(\underline{x}_i, \ell_i) \in \text{subset } j$

$$\hat{\ell}_j(\underline{x}_i) = \hat{\ell}_i$$

4. Compare $\hat{\ell}_i = \ell_i$ for $i = (\underline{x}_i, \ell_i)$
pert j $\ell_i \leq 1 \quad \hat{\ell}_i = \ell_i$
end for $0 \quad \text{else}$

\Rightarrow • Confusion matrix (at the end
• $AER(\hat{J}_X) = \frac{1}{n} \sum_{i=1}^n \varepsilon_i$ we tested our
model on each
model

Ex

$$L = \begin{bmatrix} \ell_1 \\ \vdots \\ \ell_{20} \end{bmatrix}$$

ℓ_i - i.i.d $\sim \text{Bin}(0.5)$

$$\underline{X} \in \mathbb{R}^{1000}$$

$$\begin{bmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_{20} \end{bmatrix}$$

\underline{y}_i i.i.d \sim
 $\sim N(\underline{\mu}, \Sigma)$

Select the 50 components of \underline{X}

$(\underline{X} \in \mathbb{R}^{1000})$ that are most correlated with labels $\begin{bmatrix} l_1 \\ \vdots \\ l_{50} \end{bmatrix}$

for example LDA, QDA and their accuracy will be higher. But we over-fit

Cross-validation have to be done for every step.

So in this example we can not build model, but extracting correlated features we will overfit our data.

Repeat k-fold cross-validation

β times by taking β different initial permutation your data.

$\widehat{AER}(f) \dots$

$\widehat{AER}_B(f)$

some classifier
each time make \Rightarrow

k-fold

$$\overline{\widehat{AER}}(f) = \frac{1}{B} \sum_{i=1}^B \widehat{AER}_i(f)$$

$$\text{Var}(\widehat{AER}(f)) = \frac{1}{B-1} \sum_{i=1}^B (\widehat{AER}_i(f) - \overline{\widehat{AER}}(f))^2 \Rightarrow$$

$$C_{1-\alpha}^+ (E[\widehat{AER}(f)]) = \overline{\widehat{AER}}(f) \pm z_{1-\alpha/2} \sqrt{\frac{\text{Var}(\widehat{AER})}{B}}$$

Chapter V Cross

Right understanding very

\Rightarrow same
cross validation

Lecture 1.04.25

SVM (Vapnik) Support Vector Machines

logistic regression - we have to know from previous courses.

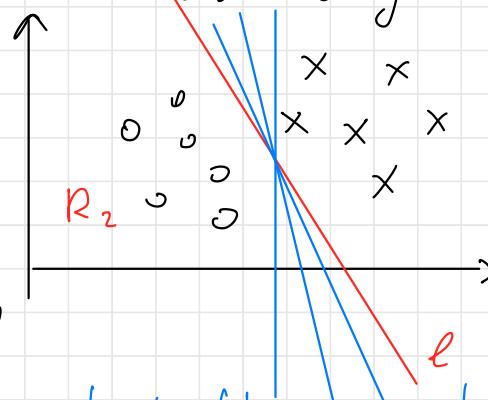
Data

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1 \\ \vdots \\ \vdots \\ \underline{x}_n \end{bmatrix} \begin{bmatrix} \ell_1 \\ \vdots \\ \vdots \\ \ell_n \end{bmatrix} \quad \ell_i \in \{1, 2\}$$
$$\underline{x} \in \mathbb{R}^D$$

ℓ -separating hyperplane

$\Rightarrow \{R_1, R_2\}$ part of \mathbb{R}^D

\Rightarrow classifier $f \leftarrow \{R_1, R_2\}$



Motivating idea:
but there can be
lot of hyper
planes

A, B subsets of \mathbb{R}^n

? \exists separating hyperplane

Let $CH(A)$ be convex hull (~~convex hull~~)
identified by A

$CH(B)$ convex hull identified by B

Theorem \exists separating hyperplane

A and B if:

1. $CH(A) \neq \emptyset, CH(B) \neq \emptyset$

Mann

Banach

2. $CH(A) \cap CH(B) = \emptyset$

theorem

(geometric form)

3. either $CH(A)$ or $CH(B)$ is open

Rmk (2) is equivalent to (3') saying:

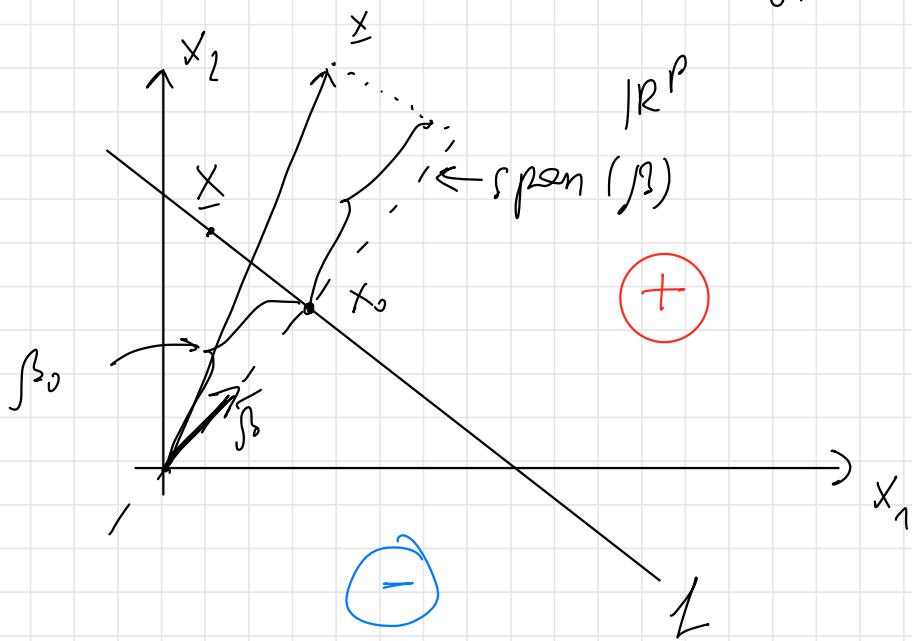
$CH(A)$ and $CH(B)$ are closed and at least one of them is compact

Corollary In case A and B finite \Rightarrow

Check only (1) & (2)

Hyperplanes in \mathbb{R}^p

An hyperplane in \mathbb{R}^p is an affine subspace of \mathbb{R}^p of dimension $p-1$: to identify it is enough to specify the direction or orthogonal to the hyper-plane L



Identified by $\beta \in \mathbb{R}^p$, $\beta \perp L$ and $\|\beta\|=1$

$$x_0 = L \cap \text{span}(\beta)$$

$$x \in L \Leftrightarrow \pi_{x/\beta} = x_0$$

$$\underline{x} \in \mathcal{L} \Leftrightarrow \frac{\beta' \underline{x}}{\|\underline{\beta}\|} = \underline{x}_0 \Leftrightarrow (\beta' \underline{x}) \cdot \underline{\beta}^T = \underline{x}_0$$

let $\|\underline{x}_0\| = \beta_0$

$$\Leftrightarrow \beta' \underline{x} = \beta_0 \Leftrightarrow \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p = \beta_0$$

if $\underline{x} = (x_1, x_2, \dots, x_p)^T$

$\beta' \underline{x} - \beta_0$ distance between \underline{x} and \mathcal{L}
(with sign)

if $\beta' \underline{x} - \beta_0 = 0 \Rightarrow \underline{x} \in \mathcal{L}$

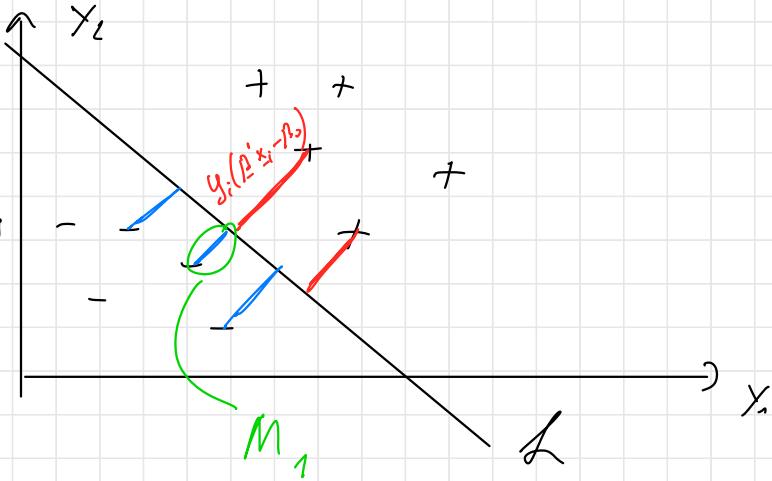
$\beta' \underline{x} - \beta_0 > 0 \Rightarrow \underline{x} \in \text{+}$

$\beta' \underline{x} - \beta_0 < 0 \Rightarrow \underline{x} \in \text{-}$

Deck to training data

* $= \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \begin{bmatrix} \ell_1 \\ \vdots \\ \ell_n \end{bmatrix} e_i^T$

Plus minus



Assume existence of separating hyperplane

(L) for groups 1 & 2.

$$y_i = \begin{cases} +1 & \text{if } \underline{x}_i \in \text{Plus} \\ -1 & \text{if } \underline{x}_i \in \text{Minus} \end{cases}$$

identified by $\underline{\beta}, \beta_0$.

$$y_i (\underline{\beta}' \underline{x}_i - \beta_0) \geq 0 \quad (\text{if } \underline{x}_i \in \oplus \Rightarrow y_i > 0 \Rightarrow \text{res positive})$$

$$\text{if } \underline{x}_i \in \ominus \Rightarrow y_i < 0 \Rightarrow \text{res still positive}$$

$$M_1 = \min \left\{ y_i (\underline{\beta}' \underline{x}_i - \beta_0) : i = 1 \dots n \right\} - \text{Margin}$$

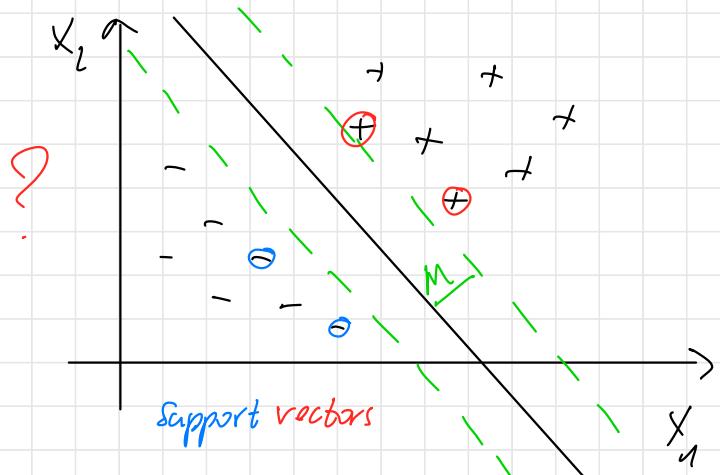
We want to maximise margin

(to find $\underline{\beta}$ which has biggest margin)

Optimisation problem

Find $\underline{\beta}$ and β_0 s.t. $M_1(\underline{\beta}, \beta_0)$ is max

M - margin
(distance)



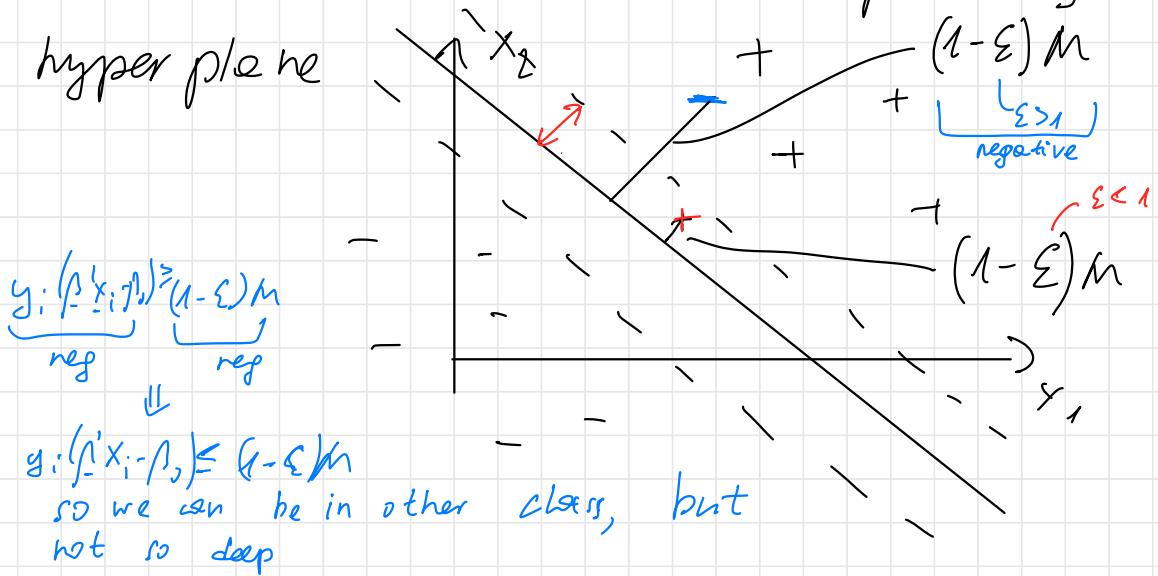
$$\begin{cases} \text{MAX } M(\beta, \beta_0) \\ \beta, \beta_0 \\ \|\beta\| = 1 \\ y_i(\beta^T x_i - \beta_0) \geq M \text{ for } i=1 \dots n \end{cases}$$

\Rightarrow ESL

element of
statistical learning

Hard version
of optimi-
sation problem

What if there is no separating hyperplane



Soft version

$$\left\{ \begin{array}{l} \text{Max } M(\underline{\beta}, \beta_0) \\ \|\underline{\beta}\| = 1 \\ y_i (\underline{\beta}' \underline{x}_i - \beta_0) \geq (1 - \varepsilon_i) M \text{ for } i=1 \dots n \\ \varepsilon_i \geq 0 \text{ and s.t. } \sum_{i=1}^n \varepsilon_i \leq C \end{array} \right. \rightarrow \text{ELS}$$

budget (Regularization)

$C=0 \Rightarrow$ hard problem

Sol'n: $\hat{\underline{\beta}} = \sum_{i=1}^n \frac{\uparrow}{y_i \underline{x}_i}, \hat{\beta}_0 \in \mathbb{R}$

$$\hat{f}(\underline{x}) = \hat{\underline{\beta}}' \underline{x} - \hat{\beta}_0$$

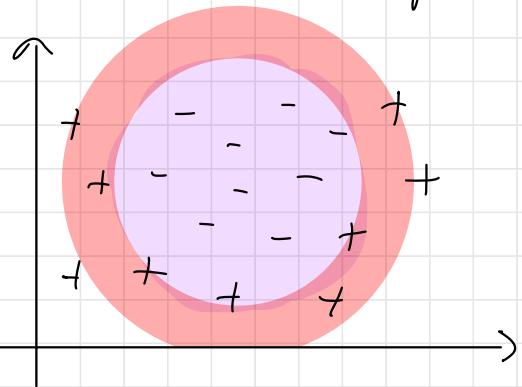
$$C(\underline{x}) = \text{sign}(\hat{f}(\underline{x}))$$

What if there is no obvious separating hyperplane

"Received" variables:

x_1, x_2

features: $z_1 = x_1, z_2 = x_2, z_3 = x_1^2, z_4 = x_2^2, z_5 = x_1 x_2$



$$\tilde{\mathbf{X}} = \begin{pmatrix} z_1 & z_2 & \dots & z_5 \\ x_{11} & x_{11} & x_{11}^2 & x_{11}^3 & x_{11}^4 \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ \mathbf{T} & & & & e_1 \end{pmatrix} \Rightarrow C$$

transformed variables

kernel trick

$$k(\underline{x}) = (h_1(\underline{x}), \dots, h_m(\underline{x}))^\top \in \mathbb{R}^m$$

For $i=1 \dots m$, $h_i: \mathbb{R}^p \rightarrow \mathbb{R}$
 linear independent

Transform received data \mathbf{X} , into $\tilde{\mathbf{X}} =$

$$= \begin{bmatrix} h_1'(\underline{x}) \\ h_2'(\underline{x}) \\ \vdots \\ h_m'(\underline{x}) \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad y_i \in \{-1, +1\} \quad h(\underline{x}) \in \mathbb{R}^m \Rightarrow$$

Separating hyperplane using soft or hard SVM

So we like expand features, to increase distance between features

$$C(\underline{x}) = \text{sign}(\hat{f}(h(\underline{x})))$$

$$\hat{f}(h(\underline{x})) = \hat{\beta}_0 + h(\underline{x})$$

$$\hat{\beta}_0 = \sum_{i=1}^n y_i - \hat{h}(\underline{x}) \Rightarrow \text{data}$$

$$\hat{f}(h(\underline{x})) = \sum_{i=1}^n g_i \hat{h}_i(\underline{x}) - \hat{\beta}_0 \quad \text{what we want classify}$$

\hat{Z}

Only inner product

decide predicted class

Obs

Only the inner product $\hat{h}_i(\underline{x}) \hat{h}(\underline{x})$ enter in the solution

let's consider a fact $K: \mathbb{R}^P \times \mathbb{R}^P \rightarrow [0, \infty)$

$$\text{s.t. } K(\underline{x}, \underline{w}) = \hat{h}(\underline{x}) \hat{h}(\underline{w}) \quad (\text{kernel})$$

$\underline{x}, \underline{w} \in \mathbb{R}^P$

$$\Rightarrow \hat{Z}(\underline{x}) = \sum_{i=1}^n g_i K(\underline{x}_i, \underline{x})$$

Kernel

Original problem $K(\underline{x}, \underline{w}) = \underline{x}^T \underline{w}$

Kernels (popular choices)

- $K(\underline{x}, \underline{w}) = [1 + \underline{x}' \underline{w}]^d$ polynomials of order d (\underline{w})
- $K(\underline{x}, \underline{w}) = \exp[-\gamma \|\underline{x} - \underline{w}\|^2]$ radial basis
- $K(\underline{x}, \underline{w}) = \tanh[k_1 \underline{x}' \underline{w} + k_2]$

4.04.25

Unsupervised Classification - Clustering

$$\tilde{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

$x_i \in \mathbb{R}^p$
 $e_i \in \{1, \dots, k\}$

Much diving idea.

units belonging to the same group (cluster) are more similar than units belonging to different groups.

$$\Rightarrow \tilde{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \begin{bmatrix} \hat{e}_1 \\ \vdots \\ \hat{e}_n \end{bmatrix} \quad x_i \in \mathbb{R}^p \quad \hat{e}_i \in \{1, \dots, \hat{g}\}$$

dissimilarity - distance:

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

with properties:

(1) $d(\underline{x}, \underline{y}) = 0 \Leftrightarrow \underline{x} = \underline{y} \quad \forall \underline{x}, \underline{y} \in \mathbb{R}^n$

(1') $d(\underline{x}, \underline{x}) = 0 \quad \forall \underline{x} \in \mathbb{R}^n \quad (\Leftarrow)$

Ex. $f: \mathbb{R} \rightarrow \mathbb{R}$

$g: \mathbb{R} \rightarrow \mathbb{R}$

$$d(f, g) = \int |f - g|^2 \quad (2) \text{ d. > func}$$

$$d(f, g) = 0 \Rightarrow f = g \quad \begin{array}{l} (1) \Rightarrow \text{not} \\ \text{always} \end{array}$$

(1') but not (2)

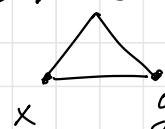
(2) $d(\underline{x}, \underline{y}) = d(\underline{y}, \underline{x}) \quad \forall \underline{x}, \underline{y} \in \mathbb{R}^n$

(symmetry)

(3) $d(\underline{x}, \underline{y}) \leq d(\underline{x}, \underline{z}) + d(\underline{z}, \underline{y})$ triangular inequality

(4) $d(\underline{x}, \underline{y}) \leq \max \{ d(\underline{x}, \underline{z}), d(\underline{z}, \underline{y}) \}$

(ultrametric property)



(4) \Rightarrow (3)

Names: (1), (2), (3) \Rightarrow d is a metric

(1'), (2), (3) \Rightarrow d is pseudometric

(1), (2), (4) $\Rightarrow d$ is ultra-metric

Three steps for clustering

- (1) Define units, unit representation,
space embeddings
 - (2) choosing a d
 - (3) algorithm
- (try to formalize
information to
be clear for
clustering)

Catalog of (trivial) distances

(1) Euclidean (Euclidean to PCA \Rightarrow Cov-^{diag, w/o})

$$d^2(\underline{x}, \underline{y}) = \sum_{i=1}^n (x_i - y_i)^2 \quad \underline{x}, \underline{y} \in \mathbb{R}^p$$

\Rightarrow std variables (or we will have different measurements and spread of values)

(2) $d^2(\underline{x}, \underline{y}) = (\underline{x} - \underline{y})^\top \Sigma^{-1} (\underline{x} - \underline{y})$

Σ covariance matrix X

(with standardization happen, that we assume some covariance structure in every group but this is not necessarily true and we can't even check whether it's true or not.)

euclidean
distance

$$(3) d^p(\underline{x}, \underline{y}) = \sum_{i=1}^p |x_i - y_i|^p \quad d(\underline{x}, \underline{y}) = \sqrt[p]{\sum |x_i - y_i|^p}$$

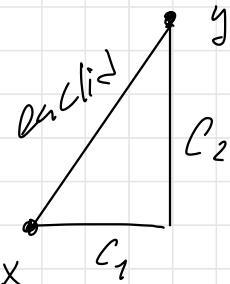
$p \geq 0$

$p=1$ Manhattan distance

(L₁)

$$d_{\text{euclidian}} = c_1^2 + c_2^2$$

$$d_{\text{manht}} = c_1 + c_2$$



(4) Canberra dist

$$\underline{x} \in (\mathbb{R}^+)^p$$

$$x_i \geq 0 \text{ for } i=1 \dots p$$

$$\underline{y} \in (\mathbb{R}^+)^p$$

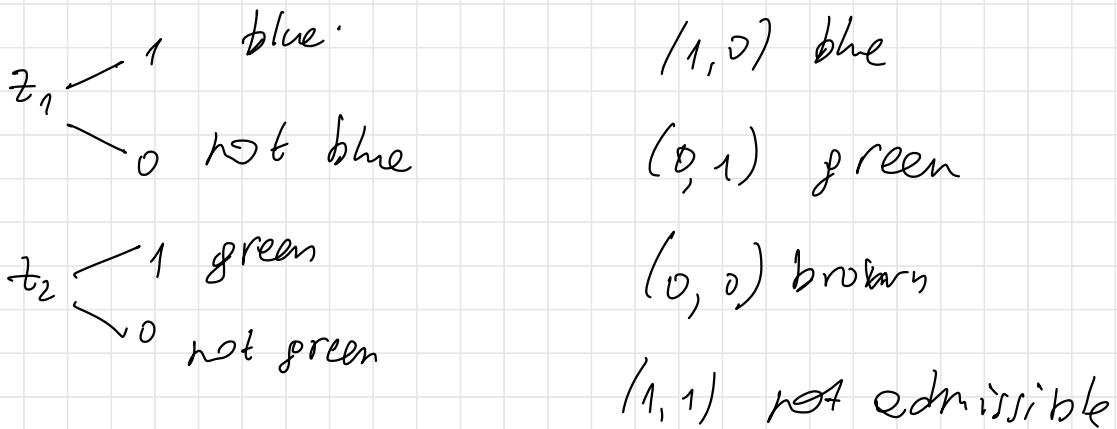
$$d(\underline{x}, \underline{y}) = \sum_{i=1}^p \frac{|x_i - y_i|}{x_i + y_i}$$

sort of normalize with respect to mean

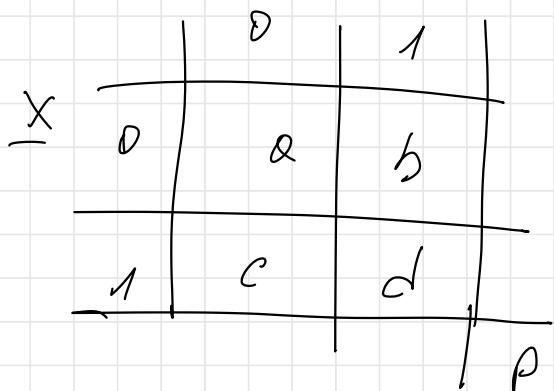
Categorical variables

$$\underline{x}, \underline{y} \in \{0, 1\}^P \quad \underline{x} = (0, 1, 0, 1, 0, 0)$$

Ex $x \in \{\text{blue, green, brown}\}$



$$d_{\text{disc}}^2(\underline{x}, \underline{y}) = \sum_{i=1}^p (x_i - y_i)^2 = \# \text{ of discordances} \quad (\text{mismatches})$$



$$d(\underline{x}, \underline{y}) = 1 - \frac{d}{p}$$

Don't want measure
 distance in (0,0)
 (because both people

without phd not same close as
 two people having phd.

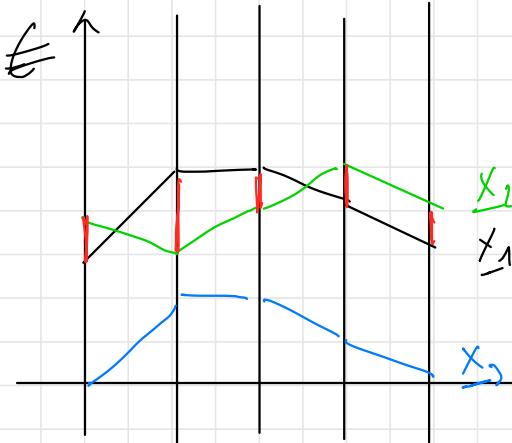
$$\underline{\mathbf{X}} = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_n' \end{bmatrix} = [\underline{c}_1 \dots \underline{c}_p] \quad \underline{c}_i \in \mathbb{R}^p$$

$$d^2(\underline{c}, \underline{b}) = 2(1-p)$$

$\underline{c}, \underline{b} \in \mathbb{R}^n$ if they strongly correlated \Rightarrow they have to be similar

$$p = \text{corr}(\underline{c}, \underline{b})$$

Ex std \underline{c} and \underline{d} $\Rightarrow 2/(1-p) = \text{euclid dist}^2$



Euclidean distance will say, that x_2, x_1 more similar, but x_2 and x_1 have same template (repeated necessary)

$$\underline{x}_1 = k \underline{x}_2$$

and here correlation will define it

$$\mathcal{U} = \{\text{units}\} \quad d: \mathcal{U} \times \mathcal{U} \rightarrow [0, \infty)$$

\underline{u} = sequence of symbols = $(\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \theta)$

$U \vdash$

$\underline{v} =$

"

$= (\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \theta)$

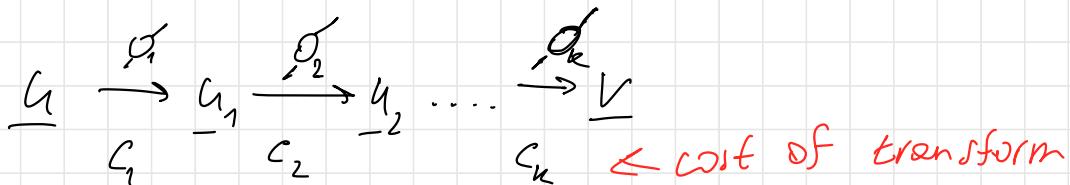
$\downarrow \quad \downarrow$

$k=9$

first position where
two symbols are
different

$$d(\underline{u}, \underline{v}) = \frac{1}{2^k}$$

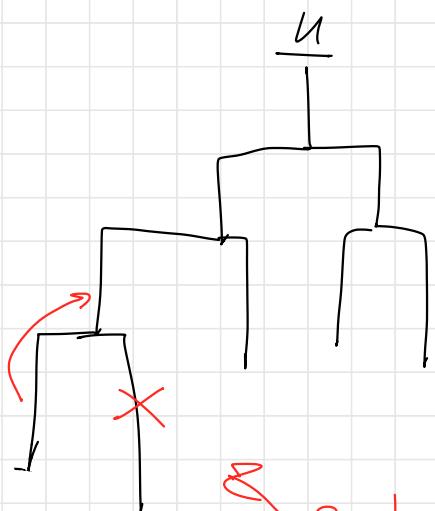
C : set of operations acting on $\underline{u} \in U$



$$d(\underline{u}, \underline{v}) = \min \left(\sum_{i=1}^K c_i : c_i \text{ cost of } O_i, i=1..k \right)$$

edit distances

(like source img $\rightarrow \dots \rightarrow$ flipped img)



and we know what we have to do to make $\underline{U} = \underline{V}$ and this operation (of removing edge) has some cost.

Distance btwn clusters

$$U \subseteq \mathbb{R}^P$$

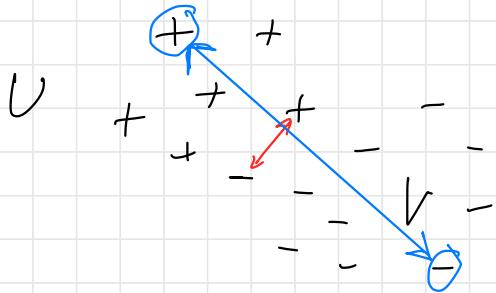
$$V \subseteq \mathbb{R}^P$$

$$d: \mathbb{R}^P \times \mathbb{R}^P \rightarrow [0, \infty)$$

$$? d(U, V) \text{ (linkage)}$$

(1) Single-linkage

$$d(U, V) = \min \{ d(x, y) : x \in U, y \in V \}$$



(2) Complete-linkage

$$d(U, V) = \max \{ d(x, y) : x \in U, y \in V \}$$

(3) Average linkage

$$d(U, V) = \frac{1}{\underbrace{\#U \cdot \#V}_{\text{Number of couples}}} \sum_{\substack{x \in U \\ y \in V}} d(x, y)$$

Number
of couples

(4) \bar{x} centroid of U (being centre)

\bar{y} " " of V

$$\underline{d}(U, V) = d(\bar{x}, \bar{y})$$

$$\bar{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad d_{ij} = d(x_i, x_j) \quad i, j = 1 \dots n$$

$$D = [d_{ij}] \underset{n \times n}{=} \left(\begin{bmatrix} 0 & d_{12} & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n-1, n} & \ddots & 0 \end{bmatrix} \right)$$

dissimilarity matrix

Hierarchical Agglomerative Algorithms

- Initialization: every unit is a cluster
- Repeat until converge
 - Step 1: aggregate the two more similar dist.
 - Step 2: compute the new dissimilarity matrix

Example simple link

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

Iter 1
 $1 \& 2 \Rightarrow \{1, 2\}$
 $d(1, 2) = 2$

$\{1, 2\}$	3	4	5
------------	---	---	---

$$4 \& 5 \Rightarrow \{4, 5\} \quad \underline{\text{Iter 2}}$$

$$d(4, 5) = 3$$

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

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

Iter 3

$$\{4, 5\} \& \Rightarrow \{3, 4, 5\}$$

$$d(\{4, 5\}, \{3\}) = 1$$

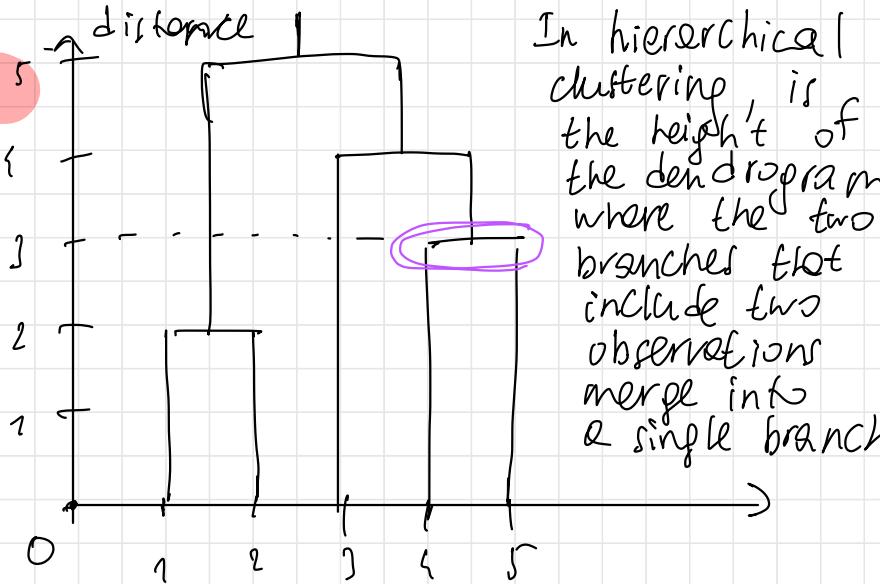
$$\{1, 2\} \& \{3, 4, 5\} \Rightarrow \{1, 2, 3, 4, 5\}$$

Iter 4

$$d(\{1, 2\}, \{3, 4, 5\}) = 5$$

Dendrogram 5

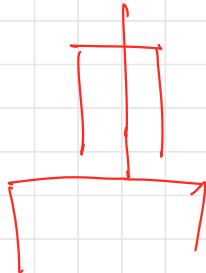
5 merges
with 3



In hierarchical clustering, is the height of the dendrogram where the two branches that include two observations merge into a single branch.

We want make without horizontal line intersection

← not like that



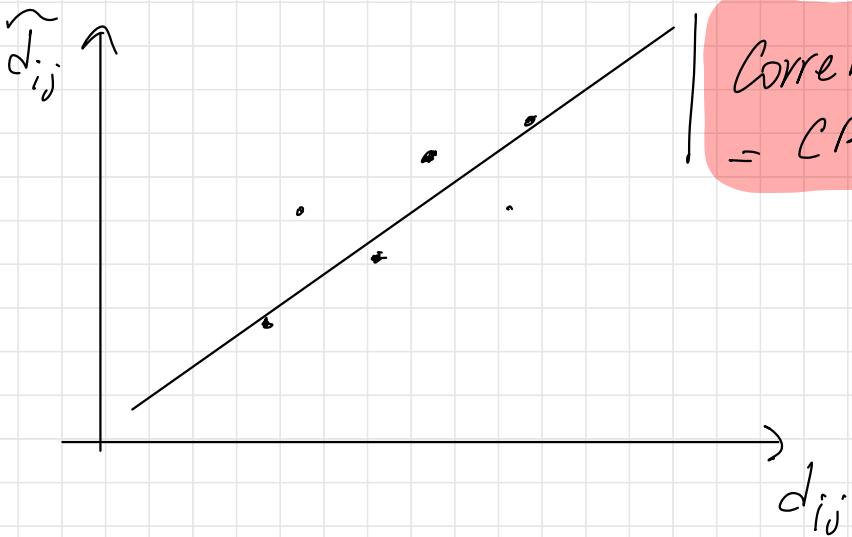
Cophenetic

jittering - adding small error - we will see

is obsr metric	cluster		robust			
	1	2	3	4	5	
Wtde	0					
\tilde{D}	2	0				
cophenet dist matrix	3	5	5	0		
	4	5	5	4	0	
	5	5	5	5	0	

for every couple of units
 $(i, j) \rightarrow d_{ij}, \tilde{d}_{ij}$

distance between each units, being into their account in clusters



Correlation (D, \widehat{D})
 $= CPCC$

as larger CPCC - more captured structure

$\text{Corr} \{ (d_{ij}, \widehat{d}_{ij}) : i, j = 1 \dots n \} = \text{Corr}(D, \widehat{D}) = CPCC$

distance between objects
distance between objects in clusters

- The closer it is to 1 and -1, the better the clustering structure in C is representing a true structure that is in the D matrix!
- Since it's a correlation coefficient it's in between [-1, 1]
- If $CPCC = 0 \Rightarrow$ we couldn't capture any

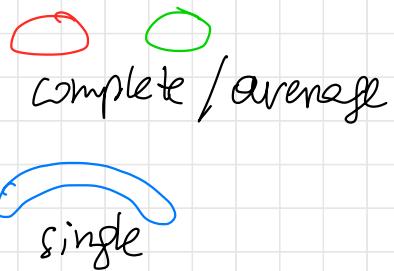
clustering structure by dendrogram

For example we try hierarchical clustering changing linkage, we try them all and see the best **Cophenetic Correlation Coefficient!**

Deciding where to cut the dendrogram is hard: one good idea is to jitter the data and see what happens!

Note: single linkage often generates a chain effect: we create a cluster that is a chain that goes across different clusters! So if clusters are ellipsoid blobs of data we use complete linkage or average linkage!

Note: complete linkage and average linkage have the opposite problem: they generate ellipsoidal clusters!



Lecture 10. 04. 25

Non-hierarchical clustering

k-means

$$d: \mathbb{R}^D \times \mathbb{R}^D \rightarrow [0, \infty)$$

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_n \end{bmatrix} \quad \text{training set.} \quad \underline{x}_i \in \mathbb{R}^D$$

Example Euclidean distance

Fix $k \geq 1 \Rightarrow$ find \underline{c}_1, \dots of \mathbb{X} in

Centroid: (*) $\underline{c}_1, \dots, \underline{c}_k$ k subsets

$$C_i \subseteq \mathbb{X}$$

$$C_1, \dots, C_k$$

$$\underline{\bar{x}}_i = \arg \min_{\underline{x} \in \mathbb{R}^D} \sum d^2(\underline{x}_j, \underline{x})$$

$$1) C_i \cap C_j = \emptyset \text{ if } i \neq j$$

$$\underline{x} \in \mathbb{R}^D \quad \underline{x}_j \in C_i$$

$$2) \cup C_i = \mathbb{X}$$

Ex If d Euclidean distance

$$\underline{\bar{x}}_i = \frac{1}{|C_i|} \sum_{\underline{x}_j \in C_i} \underline{x}_j \quad - \text{just mean}$$

General optimisation problem

Find c_1, \dots, c_k s.t.

$$\sum_{j=1}^k \sum_{x_i \in C_j} d^2(x_i, \bar{x}_j) \text{ is min}$$

k clusters

minimise
sum of
variance
within groups

k-means alg

Initialisation step

- Split \mathbb{X} in k subsets

$c_1, \dots, c_k \rightarrow$ step 1

eqiv

- Choose $\bar{x}_1, \dots, \bar{x}_k$ as initial centroids
of $c_1, \dots, c_k \rightarrow$ step 2

Repeat ✓

Step 1 compute the centroids of
 c_1, \dots, c_k

Step 2 Assign each $\underline{x} \in \mathbb{X}$ to the cluster identified by closest centroid

e.g. assign \underline{x} to C_i if $d(\underline{x}, \bar{\underline{x}}_i) = \min \{d(\underline{x}, \bar{\underline{x}}_1), \dots, d(\underline{x}, \bar{\underline{x}}_n)\}$

End: when centroids do not change

Obs. Solving (*) could be difficult

Easy way out.

$$\bar{\underline{x}}_i = \operatorname{argmin}_{\underline{x} \in \mathbb{X}} \sum_{\underline{x}_j \in C_i} d^2(\underline{x}_j, \underline{x}) \text{ medoid}$$

$\underline{x} \in \mathbb{X}$

$\underline{x}_j \in C_i$

(k-medoid)

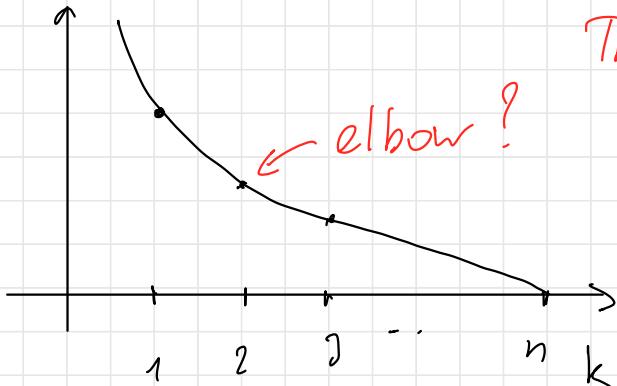
How do we choose k

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

total
within
variability

don't minimise this function

because it will be $k=n$ (number of sample)



The less is k then easier it's to explain what is going on

Note there is no obvious link between PCA and clusters: it may be harmful to do first PCA and then clustering.

The same holds for PCA and classification

Obs Wards' method

$$SS_j = \sum_{x_i \in C_j} \|x_i - \bar{x}_j\|^2$$

Hierarchical Aggl method

(modified linkage)

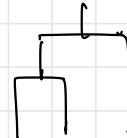
Kon sun crassus
pacchiaro
menygo
vee-
pum

$$SS = SS_1 + SS_2 + \dots + SS_k$$

(if there are k clusters)

Aggregate bottom-up clusters in such a way that the increase of S_S is minimum, (so S_S increases, but by smallest value)

(Kopie Stegmann Klassifizierungsumgebung benutzte passende Werte min)

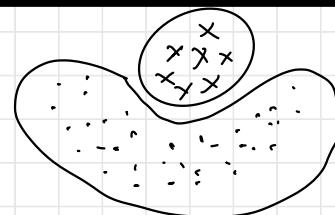


DB-scan

DB = Density Based

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_n \end{bmatrix} \quad \text{training set}$$

$$\underline{x}_i \in \mathbb{R}^P$$



unsupervised



supervised

$$d: \mathbb{R}^P \times \mathbb{R}^P \rightarrow [0, \infty) \text{ dist}$$

Fix $\varepsilon > 0$ For $\underline{x} \in \mathbb{R}^P$ including \underline{x}

$$N_\varepsilon(\underline{x}) = \left\{ \underline{y} \in \mathbb{R}^P : d(\underline{x}, \underline{y}) < \varepsilon \right\}$$



$|N_\varepsilon(\underline{x})|$ = size of $N_\varepsilon(\underline{x})$: number of units in \mathbb{X} belonging to $N_\varepsilon(\underline{x})$

$N_\varepsilon(\underline{x})$

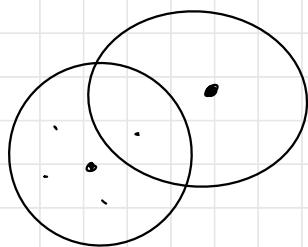
$\min Pts \geq 1$

integer

Point classes

Fix ε , min Pts

$\min Pts = 5$



Core point can
be inside of
region of other core point

- $\underline{x}_i \in \mathbb{X}$ is a

core point

if $|N_\varepsilon(\underline{x}_i)| \geq \text{MinPoint}$

- \underline{x}_i is border point if

\underline{x}_i is not core but
belongs to $N_\varepsilon(\underline{x}_j)$
and \underline{x}_j is core

Density reachable points :

- $\underline{x}_j \in \mathbb{X}$ is directly density reachable
from \underline{x}_i if

1) \underline{x}_i is core

2) $\underline{x}_j \in N_\varepsilon(\underline{x}_i)$

- \underline{x}_j is density reachable from $\underline{x}_i \in \mathbb{X}$

if there is a finite sequence

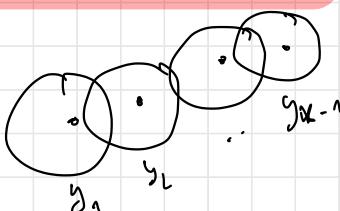
y_1, \dots, y_k s.t.

1) $y_1 = \underline{x}_i$, $y_k = \underline{x}_j$

2) y_{i+1} is directly density reachable from y_i , $i=1, \dots, k-1$

Note: y_1, \dots, y_{k-1}

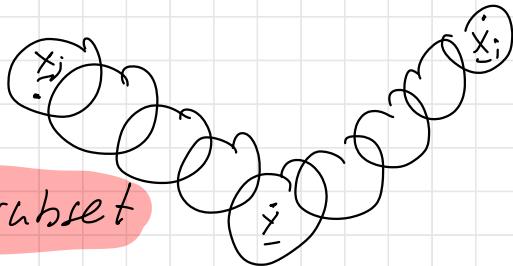
need to be core



- $\underline{x}_i \in \mathbb{X}$, $\underline{x}_j \in \mathbb{X}$ are density connected

if there is $\exists n \underline{x} \in \mathbb{X}$

s.t. both \underline{x}_i and \underline{x}_j are density reachable from \underline{x}



DB-scan identifies subset

C of \mathbb{X} s.t.

1) if $\underline{x}_i \in C$ and \underline{x}_j is density reachable from $\underline{x}_i \Rightarrow \underline{x}_j \in C$

2) \underline{x}_i and $\underline{x}_j \in C$ must be density connected

Problem of high dimension

if p - large \Rightarrow everything isolated

In R it makes choices, so if it assigns same clusters it removes from algorithm
clustering change in borders

Visual clustering

Multi dimensional scaling (MDS)

projection on dimension increasing variability between props.

$$\mathbb{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad x_i \in \mathbb{R}^p \quad d: \mathbb{R}^p \times \mathbb{R}^p \rightarrow [0, \infty)$$

$$q < p \quad \mathbb{X} = \begin{bmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_n \end{bmatrix} \quad y_i \in \mathbb{R}^q$$

reduce dimension
(use embeddings)

$$S: \mathbb{R}^q \times \mathbb{R}^q \rightarrow [0, \infty) \quad \text{euclidean metric}$$

$$d_{ij} = d(\underline{x}_i, \underline{x}_j) \quad f_{ij} = f(\underline{g}_i, \underline{g}_j)$$

$$\underline{x}_i, \underline{x}_j \in \mathbb{R}^P \quad \underline{g}_i, \dots, \underline{g}_j \in \widetilde{\mathcal{X}}$$

and we want

$$d_{ij} \simeq f_{ij}$$

let \underline{x}_i - spatial coordinates of airports in europe, and as distance we take flight time from \underline{x}_i to \underline{x}_j

And now we want to transform
 $\underline{x}_i \rightarrow \underline{g}_i$ where distance between y_i, y_j proportional to $d(\underline{x}_i, \underline{x}_j)$

If it has solution - it has ∞ number of solutions (they will have same distances, but in different position)

Classical MDS find y_1, \dots, y_n c.c.

$$\sum_{i=1}^n \sum_{j=1}^n (d_{ij} - f_{ij})^2 \text{ is minimum}$$

What if d_{ij} also euclidean \Rightarrow
 y_1, \dots, y_n identified by the
scores PC_1, \dots, PC_q

Other possible obj fct

$$\frac{\sum_{i=1}^n \sum_{j=1}^n (\theta(d_{ij}) - f_{ij})^2}{\sum_{i,j} f_{ij}^2} \leftarrow \text{STRESS}$$

$$\sum_{i,j} f_{ij}^2$$

Goal - minimize

$\theta: (0, \infty) \rightarrow [0, \infty)$ stress choosing

y_1, \dots, y_n

and θ

Lecture 14. 04. 25

Supervised Learning - Regression

$y \in \mathbb{R}$ $\underline{x} \in \mathbb{R}^p$
target covariates

Goal: explain the variability of y in terms of \underline{x}

$$\arg\min f: \mathbb{R}^p \rightarrow \mathbb{R} \quad E[(y - f(\underline{x}))^2] = E[y | \underline{x}]$$

Find estimates $\hat{f}: \mathbb{R}^p \rightarrow \mathbb{R}$ of $E[y | \underline{x}]$

Available training set

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_n \end{bmatrix} \quad \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$
$$\underline{x} \quad y$$

data driven approach - CART
(Random Forest)

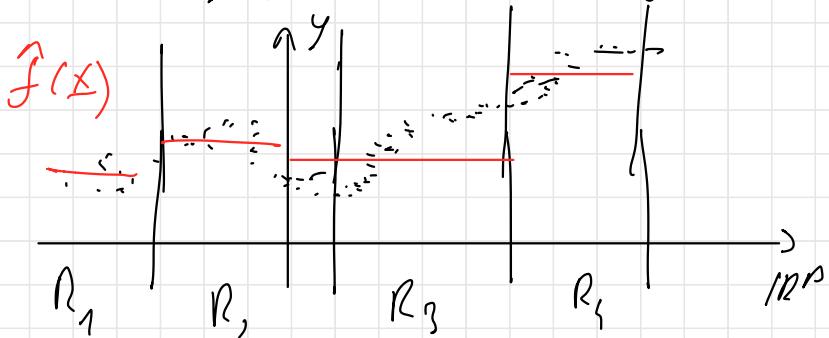
Two approaches

- **data driven**: CART, Random Forest, boosting (this is fine when we don't know much about f .)
- **model based**: linear regression, GLM, LMM (we use some prior knowledge about the problem) (**interpretable**)

CART - Classification And Regression Tree

\Rightarrow Ch on Trees in ISL (red syllabus)

\hat{f} - step wise function. Constant over the sets R_1, \dots, R_g partition of \mathbb{R}^p , such that $R_i \cap X \neq \emptyset$ for $i = 1 \dots g$



$$\bar{y}_i = \frac{1}{|R_i|} \sum_{x_j \in R_i} y_j$$

$$\hat{f}(\underline{x}) = \sum_{i=1}^J \bar{y}_i \underbrace{\Pi_{\{x \in R_i\}}}_{R_i}$$

J - Our groups

General Opt Problem

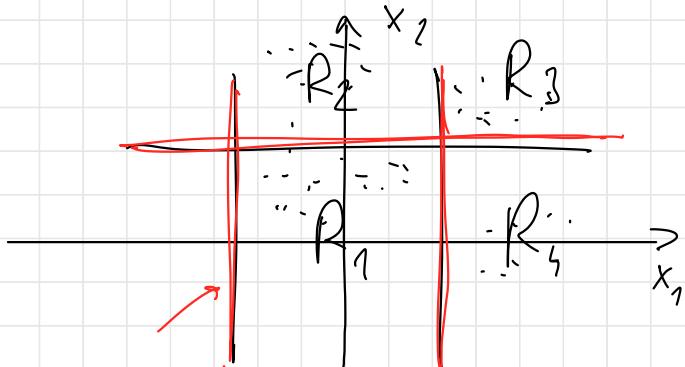
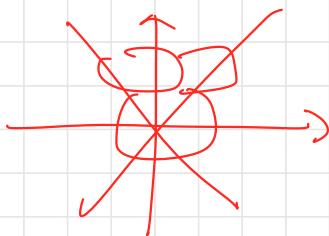
one hot vector of elements $\in \mathbb{R}^J$

$$\sum_{i=1}^J (y_i - \bar{y}_i)^2$$

$$\left\{ \begin{array}{l} \text{argmin} \\ \{R_1, \dots, R_J\} \subset \mathcal{T} \end{array} \right. \quad \begin{array}{l} \text{such that} \\ \sum_{j \in R_i} x_j \in R_i \end{array} \quad \text{⊗}$$

without overfitting the data !

CART - is a greedy algorithm for
"solving" ⊗



borders are linear

Basic step of CART

optimal half-space splitting.

R rectangle in \mathbb{R}^P , $R = (x_1, b_1) \times (x_2, b_2) \times \dots \times (x_p, b_p)$

$$(d, b) = \begin{cases} (2, 5) \\ (2, \infty) \\ (\infty, 5) \end{cases}$$

$$x_1, \dots, x_m \in R^n$$

$$\bar{y} = \frac{1}{|R|} \sum_{x_j \in R} y_j$$

Evaluation function

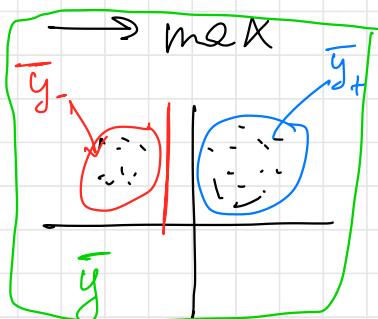
$$\sum_{x_j \in R} (y_j - \bar{y})^2$$

for regression

split R along the direction x_1 (first component of vector feature 1) optimally.

find s_1 s.t.

$$\textcircled{*} \sum_{x_j \in R} (y_j - \bar{y})^2 - \left[\sum_{\substack{x_{j1} \leq s_1 \\ x_{j1} > s_1}} (y_j - \bar{y}_-) + \sum_{x_{j1} > s_1} (y_j - \bar{y}_+)^2 \right]$$



$$\bar{y}_- = \frac{1}{|\{x_j \in R : x_{j1} \leq s_1\}|} \sum_{x_{j1} \leq s_1} y_j$$

$$\bar{y}_+ = \frac{1}{|\{x_j \in R : x_{j1} > s_1\}|} \sum_{x_{j1} > s_1} y_j$$

Solving, we will get s_1^*

repeat for directions $x_2, x_j \dots x_p \Rightarrow$

s_1^*, \dots, s_p^*

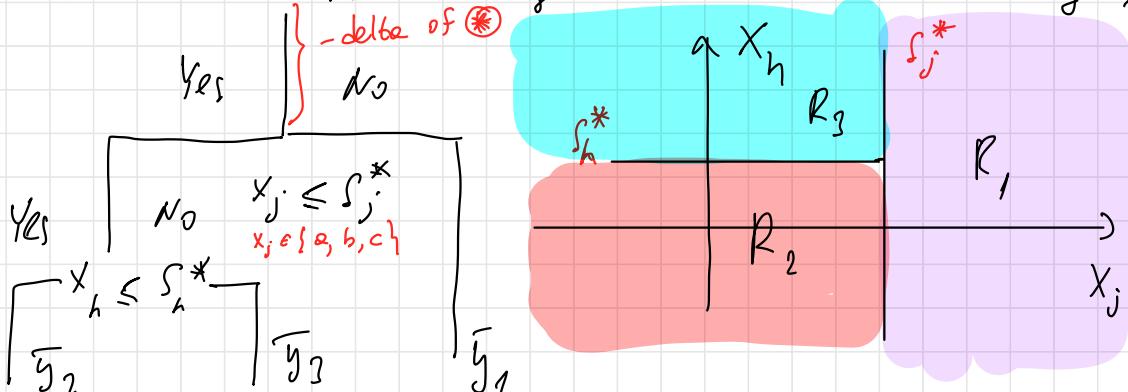
\Rightarrow choose s_k^* which maximise $\textcircled{*}$

We want split rectangle s.t. sum of variability less than variability of source group.

CART step:

start with $R = R^P$

- Proceed iteratively by opt. splitting the rectangles obtained at each iteration
- Stop splitting rect when rectangle has less than or equal to Min Pts (e.g. 5).



We can work with missing data

We can work with categorical variables

Wherever I stay - I can predict

Evaluation function for classification

Labels for units x_1, \dots, x_m - Only elements

Our classes
may be
classes
only
presented
in this rectangle

$$\text{Gini} = \sum_{x_j \in R} \sum_{i=1}^k p_i (1-p_i)$$

in rectangle R

where $1 \dots k$

p_i - frequency

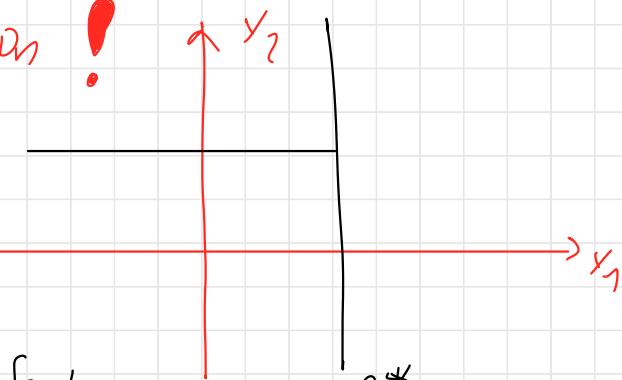
$$\text{Entropy: } - \sum_{x_j \in R} \sum_{i=1}^k p_i \log p_i$$

CART - when try understand model
for linear regression !

we need

$$x_1, x_2, \underbrace{x_1 x_2}_{\text{Our new feature}}$$

to classify $x_1 < \xi_1^*$



Overall cost function: $\sum_{j=1}^J \sum_{x_j \in R_i} (y_j - \bar{y}_i)^2 + \lambda J$

(**) $w(\lambda) = \sum_{i=1}^J \sum_{x_j \in R_i}$

Minimise $w(\lambda)$ s.t.

penalise by λ

Prune the tree bottom

for each created

up to minimize (**)

rectangle

λ - choose from cross validation

Linear model for regression

Training set

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

$$\mathcal{X}: \begin{bmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$? E[y | \underline{x}]$$

first of all we build
design matrix

$$z_i = h_i(\underline{x}_1, \dots, \underline{x}_p)$$

↑ known

$$Y = \begin{bmatrix} 1 & z_1 \\ 1 & z_{11} \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & z_{n1} \end{bmatrix} \quad \begin{bmatrix} z_2 \\ \vdots \\ z_{12} \\ \vdots \\ z_{n2} \end{bmatrix}$$

(what we have to try)

design matrix

function $\left(\begin{array}{c} x_1 \\ \vdots \\ x_n \end{array} \right)$

Model for $E[Y|X] = E[Y|z_1, \dots, z_2] =$

$$= E[Y|x_1, \dots, x_p] = \beta_0 + \beta_1 z_1 + \dots + \beta_2 z_2$$

β_0, \dots, β_2 - 2+1 unknown variables.

$$y = \underbrace{\beta_0 + \beta_1 z_1 + \dots + \beta_2 z_2}_{f(x_1, \dots, x_n)} + \varepsilon \quad \text{Var}(\varepsilon) = \sigma^2 \quad \text{2+2 params}$$

Model for data:

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \mathcal{Z} \text{ design matrix (function of } \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}) \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_2 \end{pmatrix} \in \mathbb{R}^{2+1}$$

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i$$

$$i = 1, \dots, n, \quad E[\varepsilon_i] = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2$$

$$\varepsilon \in \mathbb{R}^n \quad \text{s.t.} \quad E[\varepsilon] = 0, \quad \text{Var}(\varepsilon) = \sigma^2 I$$

so our sample uncorrelated. (independent)

And how we not assume that ε - Gaussian
 ε_i just uncorrelated, we can not say anything about independence

Example. Linear models are very flexible

ANOVA

$$\left\{ \begin{array}{l} X_1, \dots, X_{n_1} \quad i.i.d \sim N(\mu_1, \sigma^2) \\ X_2, \dots, X_{n_2} \quad i.i.d \sim N(\mu_2, \sigma^2) \quad \text{independent} \\ \vdots \\ X_g, \dots, X_{n_g} \quad i.i.d \sim N(\mu_g, \sigma^2) \quad \uparrow \text{same} \end{array} \right.$$

$$y = (x_{11}, x_{12}, \dots, x_{1m_1}, x_{21}, \dots, x_{2m_2}, \dots, x_{g_1}, \dots, x_{gm_g})^T$$

$$x \in \mathbb{R}^{n = n_1 + n_2 + \dots + n_g}$$

$$\mathcal{Z} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \end{bmatrix} n_1 \quad \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix} n_2 \quad \dots \quad \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix} n_g \quad \dots \quad \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix} n_{g+1} \quad \dots \quad \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix} n_{g+1}$$

z_i - dummy variable

$$\beta = \begin{pmatrix} \mu \\ z_1 \\ z_2 \\ \vdots \\ z_g \end{pmatrix} \quad Y = \mathcal{Z}\beta + \varepsilon \quad \varepsilon = (\varepsilon_1, \dots, \varepsilon_{n_1}, \varepsilon_2, \dots, \varepsilon_{n_2}, \dots, \varepsilon_g, \dots, \varepsilon_{n_g})$$

$$\varepsilon_{ij} \stackrel{i.i.d}{\sim} N(0, \sigma^2)$$

$$\varepsilon \sim \mathcal{N}_n(\mu, \sigma^2 I)$$

$$x_{ij} = \underbrace{\mu + z_i}_{\mu_i} + \varepsilon_{ij}$$

but \mathcal{Z} - not full rank \Rightarrow need add constraint

$$\sum_{i=1}^{g+1} n_i z_i = 0$$

$$z_g = - \sum_{i=1}^{g-1} \frac{n_i}{n_g} z_i$$

$$Y = \begin{bmatrix} 1 & \vdots & n_1 & 0 & \vdots \\ \vdots & & \vdots & \vdots & \\ 1 & \vdots & n_2 & 0 & \vdots \\ \vdots & -n_1/n_g & -n_2/n_g & \vdots & \\ 1 & -n_g/n_g & -n_2/n_g & -n_{g-1}/n_g & \\ & & & -n_{g-1}/n_g & \end{bmatrix} \beta = \begin{bmatrix} ? \\ z_1 \\ \vdots \\ z_{g-1} \end{bmatrix}$$

n x g

$$X_{ij} = \mu + \alpha_i + \beta w_{ij} + \varepsilon_{ij}$$

Anne with Covariance

ANCOVA

so how we can expand our Anne Model, adding additional information.

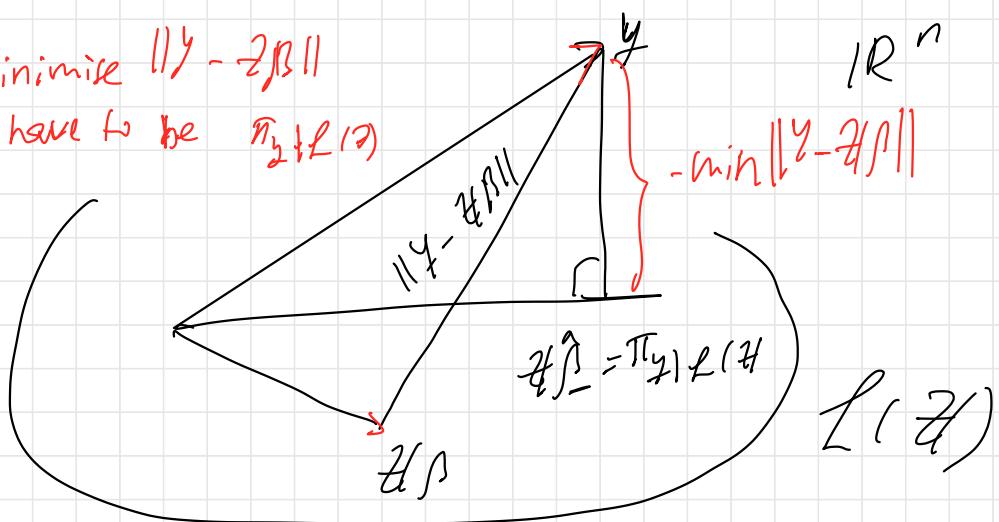
$$Y = X\beta + \varepsilon \quad \text{Estimating } \beta \text{ and } \sigma^2$$

OLS - ordinary Least Squares

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{2+1}}{\operatorname{argmin}} \| Y - X\beta \|^2 = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n (y_i - z_i^\top \beta)^2$$

to minimise $\|y - \mathcal{H}\beta\|$

\mathcal{H} have to be $\mathbb{R}^n + L(\mathcal{H})$



$$\mathcal{H}\beta = [c_0 \dots c_n]\beta = \beta_0 c_0 + \beta_1 c_1 + \dots + \beta_n c_n$$

- And we use euclidean distance, because $\text{cov}(\varepsilon) = \sigma^2 I$ - so it's identity matrix (as example if $\text{cov}(\varepsilon)$ is identity \Rightarrow Mahalanobis distance \Rightarrow Generalised Least squares)

In this picture all vectors from \mathbb{R}^n , but they are not free vary all of them together, for example for ε there are only $n-(r+1)$ degree of freedom

15. 04. 25 Lecture

Linear Models (OLS) (Ordinary least square)

$$Y = \mathcal{Z}\beta + \varepsilon \quad \mathcal{Z} - \text{design matrix}$$

$Y \in \mathbb{R}^n$ obs target

$\mathcal{Z} \in \mathbb{R}^{n \times (2+1)}$ design matrix (known)

$\beta \in \mathbb{R}^{2+1}$ coeff (unknown)

$\varepsilon \in \mathbb{R}^n$: $E[\varepsilon] = 0$, $\text{cov}(\varepsilon) = \sigma^2 I$, σ^2 unknown

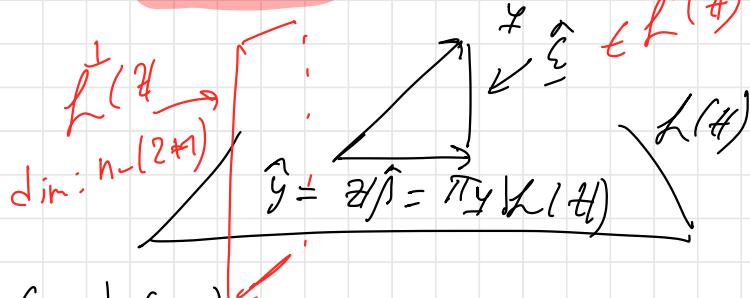
Estimating β and σ^2 (OLS)

Theo. If \mathcal{Z} is full rank ($\text{rank}(\mathcal{Z}) =$

$$= 2+1 \leq n), \text{ then } \hat{\beta} = \underset{\beta \in \mathbb{R}^{2+1}}{\operatorname{argmin}} \|Y - \mathcal{Z}\beta\|^2 =$$

we can choose only $n-(2+1)$ components (other given)

$$= (\mathcal{Z}' \mathcal{Z}^{-1}) \mathcal{Z}' Y$$



proof: $\mathcal{Z}' \mathcal{Z} = (2+1) \times (2+1)$

$$\mathcal{Z}' \mathcal{Z} = \sum_{i=1}^{2+1} d_i \mathcal{Q}_i \mathcal{Q}_i'$$
$$d_1 \geq d_2 \geq \dots \geq d_{2+1} > 0$$

$$(\mathcal{H}' \mathcal{H})^{-1} = \sum_{i=1}^{2+1} \frac{1}{d_i} \underline{e}_i \underline{e}_i'$$

For $i=1 \dots 2+1$ \mathbb{R}^{n+1}

$$\underline{\varphi}_i = \frac{1}{\sqrt{d_i}} \underbrace{\mathcal{H} \underline{e}_i}_{\mathbb{R}^n} \in \mathcal{L}(\mathcal{H})$$

$$\underline{\varphi}_i' \underline{\varphi}_j = \frac{1}{\sqrt{d_i d_j}} \quad \underline{e}_i' \mathcal{H}' \mathcal{H} \underline{e}_j = \frac{1}{d_j} \quad \underline{e}_i' \underline{e}_j = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases}$$

$\Rightarrow \{\underline{\varphi}_1 \dots \underline{\varphi}_{2+1}\}$ orthonormal basis for $\mathcal{L}(\mathcal{H})$

$$\begin{aligned} \pi_y / \mathcal{L}(\mathcal{H}) &= \sum_{i=1}^{2+1} \pi_y | \underline{\varphi}_i = \sum_{i=1}^{2+1} \frac{\underline{\varphi}_i' \underline{\varphi}_i}{\underline{\varphi}_i' \underline{\varphi}_i} y = \\ &= \sum_{i=1}^{2+1} \frac{1}{d_i} \mathcal{H} \underline{e}_i \underline{e}_i' \mathcal{H}' y = \mathcal{H} \left(\sum_{i=1}^{2+1} \frac{1}{d_i} \underline{e}_i \underline{e}_i' \right) \mathcal{H}' y \\ &= \mathcal{H} (\mathcal{H}' \mathcal{H})^{-1} \mathcal{H}' y \end{aligned}$$

$$\pi_y / \mathcal{L}(\mathcal{H}) = \mathcal{H} (\mathcal{H}' \mathcal{H})^{-1} \mathcal{H}' y$$

H hat operator

fitted values

$$\hat{Y} = Z \beta, \quad \hat{\beta} = (Z' Z)^{-1} Z' Y$$

$$\hat{\epsilon} = Y - \hat{Y} \quad - \text{residuals}$$

$$Y = \hat{Y} + \hat{\epsilon}$$

$\hat{Y} \perp \hat{\epsilon}$ ← vector of residuals
(NOT estimator of error)

↖ fitted values

Obs

1. $\text{rank}(Z) = n = 2+1 \Rightarrow Y \in L(Z) \Rightarrow$

$$\hat{Y} = \hat{Y} \Rightarrow \hat{\epsilon} = 0$$

Overfitting (we will have linear space generated by the columns of Z)

2. ~~$\text{rank}(Z) = 2+1 > n$~~ - impossible (more features than observations)

3. $\text{rank}(Z) = k < 2+1 \leq n$

$$Z' Z = \sum_{i=1}^{2+1} d_i e_i e_i' \quad d_1 \geq d_2 \dots d_k > d_{k+1} \geq 0 \dots = d_{2+1}$$

$$(Z' Z)^{-1} = \sum_{i=1}^k \frac{e_i e_i'}{d_i} \quad \text{general inverse of } Z' Z.$$

in that case $\hat{\beta} = (Z' Z)^{-1} Z' Y$

$$\pi_y | \mathcal{L}(y) = \sum_{i=1}^k \pi_y | \varphi_i = \sum_{i=1}^k \frac{\varphi_i \varphi_i'}{\varphi_i' \varphi_i} y =$$

$$= \sum_{i=1}^k \frac{1}{d_i} \underbrace{\varphi_i \varphi_i' \varphi_i'}_{\hat{\Sigma}} y = \varphi \left(\sum_{i=1}^k \frac{1}{d_i} \varphi_i \varphi_i' \right) \varphi' y$$

$$= \varphi \underbrace{\left(\varphi' \varphi \right)^{-1} \varphi'}_{\hat{\Sigma}} y$$

$\hat{\Sigma}$ - we have $n-(2+1)$ degree of freedom
 (variables, that we can choose) so we have
 $n-(2+1)$ variables, expressing variability,
 all other fixed $\hat{\Sigma}$ - not estimator of Σ
 $\hat{\Sigma}$ - if fixed (deterministic)

Coeff. of determinator (R^2)

$$y = \hat{y} + \hat{\Sigma}$$

$$\|y\|^2 = \|\hat{y}\|^2 + \|\hat{\Sigma}\|^2 \quad (\text{I})$$

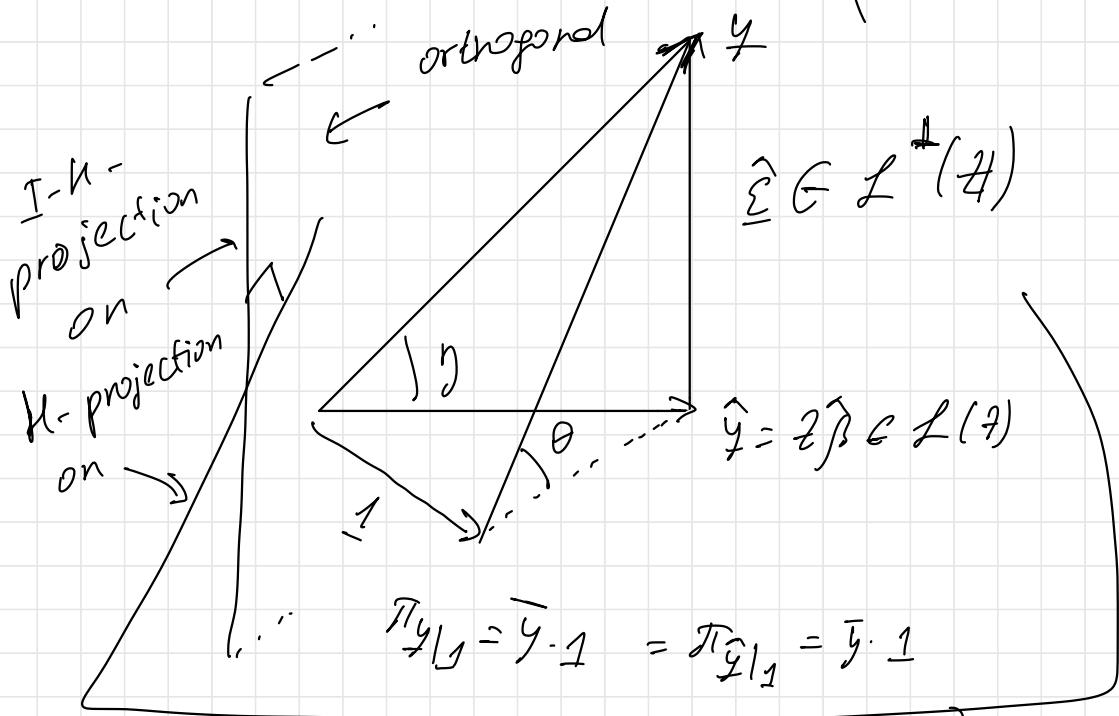
SS_{tot}

SS_{reg}

$SS_{\text{residuals}}$

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 \quad \underline{1} \in L(\mathbb{Z})$$

Recall $\mathcal{X} = \begin{bmatrix} 1 & z_{11} & \dots & z_{12} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & \dots & z_{n2} \end{bmatrix}$ $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_n \end{pmatrix}$



$$\pi_y|_{L(1)} = \bar{y} \cdot 1$$

$$\underbrace{1_H = (H \cdot 1) = (H \cdot 1) = 1'}_{\text{dim: } 2+1 \text{ if }} \quad L(\mathcal{H})$$

$$\pi_{\hat{y}}|_{L(1)} = \frac{1 \cdot 1}{1' \cdot 1} \hat{y}' = \frac{1}{1' \cdot 1} \hat{y}' = \frac{1}{1' \cdot 1} \hat{y} = \bar{y} \cdot 1$$

\mathcal{H} if full rank

$$(\sum_{i=1}^n (y_i - \hat{y}_i)^2)^2 = (\sum_{i=1}^n (y_i - \bar{y})^2) + (\sum_{i=1}^n \hat{e}_i^2)^2 \quad (\text{II})$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{e}_i^2$$

Decomposition of Variance

CSS total

CSS fitted
(reg)

CSS residual

$$1 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} + \frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$\Downarrow R^2$

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \sin^2(\theta)$$

Proportion of total variability explained by the regression model

Obs Also named coefficient of determination

$$1. R^2 \in [0, 1]$$

$$2. R^2=1 \Rightarrow \theta=0 \Rightarrow \hat{y}_i = \bar{y} \quad \text{perfect fit (overfitting)}$$

$$3. R^2=0 \Rightarrow \theta=\frac{\pi}{2} \Rightarrow \hat{y}_i = \bar{y} \cdot 1 \quad \hat{y}_i = \bar{y}$$

$\Leftrightarrow \beta_0 = \text{mean other } \beta_i > 0$

If we don't put first column as design matrix \Rightarrow we can not say that $1 \in L(\mathbf{z})$

Obs If $1 \notin L(\mathbf{z}) \Rightarrow$ (II) doesn't hold! \Rightarrow and we can not use pitagorean theorem. (sometimes we want to do Regression Through the origin in which case we wouldn't use R^2 though)

if $\beta_0 = 0 \Rightarrow$ we run through origin

R^2 may be negative

and $1 \notin L(\mathbf{z})$

But if $\beta_0 = 0 \Rightarrow$ (I) still true

$$\hat{R}^2 = 1 - \frac{\|\hat{\mathbf{z}}\|^2}{\|y\|^2} = 1 - \frac{\sum \hat{z}_i^2}{\sum y_i^2} \Rightarrow \text{compte}$$

$$= 1 - \sin^2 \gamma \quad \text{But how we don't explain}$$

variability

↑
it can be

when $y=0 \Rightarrow x=0$

Negative
 $R^2 - ?$

But we interested
in variability
how our estimator
good compare to
mean

Theo (prop of $\hat{\beta}$ and $\hat{\Sigma}$)

Assume $\text{rank}(Z) = r+1$ (full rank)

1. $E(\hat{\beta}) = \beta$ (unbiased)
2. $\text{cov}(\hat{\beta}) = \sigma^2(Z'Z)^{-1}$ ← Design Of Experiments
3. $E(\hat{\Sigma}) = \Sigma$
4. $\text{cov}(\hat{\Sigma}) = \sigma^2(I - H)$
5. $E[\hat{\Sigma}'\hat{\Sigma}] = E[\sum \hat{\epsilon}_i^2] = \sigma^2(n - (r+1))$

proof

1.
$$E(\hat{\beta}) = E[(Z'Z)^{-1}Z'Y] = (Z'Z)^{-1}Z'Z\beta = \beta$$

$\underbrace{\hspace{10em}}$

$Y = Z\beta + \epsilon \quad E(Y) = Z\beta + E(\epsilon) = Z\beta$

$\underbrace{\hspace{10em}}$
2.
$$\text{cov}(\hat{\beta}) = \text{cov}((Z'Z)^{-1}Z'Y) = (Z'Z)^{-1}Z'\text{cov}(Y)Z = (Z'Z)^{-1}\sigma^2I$$

$$\text{cov}(Y) = \text{cov}(\epsilon) = \sigma^2I \quad (\text{so we can control uncertainty of } \hat{\beta})$$

$$= \sigma^2(Z'Z)^{-1}(Z'Z)(Z'Z)^{-1} = \sigma^2(Z'Z)^{-1}$$

We can control of variability of estimator,

without knowing \hat{y}

$$3. E[\hat{\varepsilon}^2] = E[y - \hat{y}]^2 = E[y^2 - 2y\hat{y} + \hat{y}^2] = E[y^2] - 2E[y\hat{y}] + E[\hat{y}^2] =$$
$$= E[y^2] - 2E[y]\hat{y} = \sigma^2$$

already

$$4. \text{Cov}(\hat{\varepsilon}) = \text{Cov}((I-H)y) = (I-H)\sigma^2 I(I-H)' \text{ in } L(z)$$
$$= \sigma^2 (I-H)(I-H)' = \sigma^2 (I-H)$$
$$y - \hat{y} = y - Hy = (I-H)y$$

\Rightarrow projection
doesn't change

$$E[\hat{\varepsilon}' \hat{\varepsilon}] = E[\text{tr}(\hat{\varepsilon}, \hat{\varepsilon}')] = E[\text{tr}(\hat{\varepsilon}', \hat{\varepsilon})] =$$
$$= E[\text{tr}(\hat{\varepsilon}, \hat{\varepsilon}')] = \text{tr} E[\hat{\varepsilon} \hat{\varepsilon}'] =$$
$$= \text{tr} E[(I-H)y y' (I-H)] = \text{tr} E[(I-H)\underbrace{y y'}_{\varepsilon} (I-H)'] =$$
$$= \text{tr} (I-H) E[\varepsilon \varepsilon'] (I-H)' = \text{tr} (\sigma^2 (I-H)(I-H)') =$$
$$= \text{tr} (\sigma^2 (I-H)) = \sigma^2 (\text{tr} I - \text{tr} H) = \sigma^2 (n - \text{tr}(H)) = \sigma^2 (n - p) =$$
$$\text{tr}(H) = \text{tr}(Z(Z^T Z)^{-1} Z^T) = \text{tr}(Z^T Z (Z^T Z)^{-1}) =$$
$$= \text{tr}(I_{Z+1}) = 2 + 1$$

Def

$$s^2 = \frac{\hat{\Sigma}' \hat{\Sigma}}{n-(2+1)} = \frac{\|\hat{\Sigma}\|^2}{n-(2+1)}$$

Corollary

Gro $E(s^2) = \sigma^2$ i.e. s^2 unbiased for σ^2
 (estimator for σ^2)

From now on:

assume $\underline{\xi} \sim N_n(\underline{0}, \sigma^2 \mathbb{I})$ (*)

$$\Rightarrow \underline{y} = \underline{A} \underline{\beta} + \underline{\xi} \Rightarrow \underline{y} \sim N_n(\underline{A} \underline{\beta}, \sigma^2 \mathbb{I})$$

To

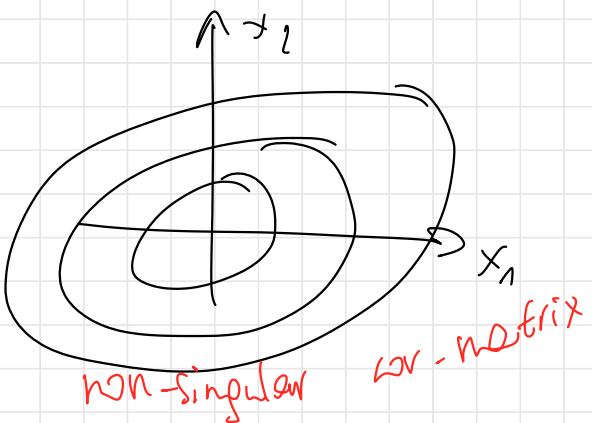
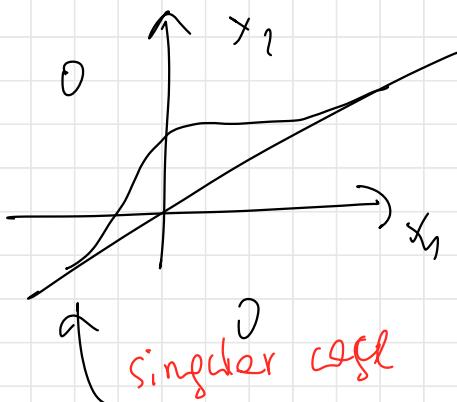
Assume (*) and $\text{rank } (\underline{A}) > 2+1 \leq n$

1. $\hat{\underline{\beta}}$ and $\frac{\hat{\Sigma}' \hat{\Sigma}}{n} = \hat{\sigma}^2$ are ML (maximum likelihood)

estimators of $\underline{\beta}$ and σ^2 resp

2. $\hat{\underline{\beta}} \sim N_{2+1}(\underline{\beta}, \sigma^2 (A' A)^{-1})$

3. $\hat{\Sigma} \sim N_n(\underline{0}, \underbrace{\sigma^2(I-H)}_{\text{singular}})$, $\det(I-H) = 0$
 as its rank is $n-(2+1)$ and not $n!$
 Gaussian distribution



so only among direction we have

$$4. \hat{\Sigma} \parallel \hat{\beta}$$

$$5. \hat{\Sigma} \sim \tilde{\sigma}^2 \chi^2(n-(2+1))$$

proof

1. Write the likelihood and start differentiating...

$$2. \frac{\partial}{\partial \beta} \ln \left(\frac{\hat{\beta}}{\hat{\Sigma}} \right) = \left((\hat{\Sigma}^{-1})^{-1} \hat{\beta} \right)^T \hat{\Sigma} \Rightarrow A \hat{\beta} = \begin{pmatrix} \hat{\beta} \\ \hat{\Sigma} \end{pmatrix}$$

$$\left(\begin{pmatrix} \hat{\beta} \\ \hat{\Sigma} \end{pmatrix} \right) \sim N_{2n+1} \left(A \hat{\beta}, \sigma^2 A A' \right)$$

$$AA' = \begin{pmatrix} (\hat{\Sigma}^{-1})^{-1} & 0 \\ 0 & I-h \end{pmatrix}$$

5. \checkmark

Ey $\underline{x} \sim N_n(\underline{0}, \Sigma)$, $\Sigma = \Sigma_{d_i e_i e_i'}$

$$d_1 \geq d_2 \dots \geq d_k \geq d_{k+1} = 0$$

$$\Rightarrow \underline{x}' \Sigma^{-1} \underline{x} \sim \chi^2_{(n-k)}$$

28. 04. 25

Introduce inference for linear models

Recall: $\mathbf{y} = \mathbf{z}\beta + \varepsilon$ $\mathbf{y} \in \mathbb{R}^n$

\mathbf{Z} $n \times (2+1)$ design matrix

$\beta \in \mathbb{R}^{2+1}$ unknown

$$\varepsilon \sim N_n(0, \sigma^2 I)$$

$$\Rightarrow \hat{\beta} = (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y} \text{ ordinary}$$

$\text{rank}(\mathbf{Z}) = 2+1 \leq n$

$$\hat{\beta} \sim N_{2+1}(\beta, \sigma^2 (\mathbf{Z}' \mathbf{Z})^{-1})$$

$$\hat{\varepsilon} = (\mathbf{I} - \mathbf{H}) \mathbf{y} \quad \mathbf{H} = \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}'$$

11

$$\hat{\varepsilon} \sim N_n(0, \sigma^2 (\mathbf{I} - \mathbf{H}))$$

$$\hat{\varepsilon}' \hat{\varepsilon} = \sum_{i=1}^n \hat{\varepsilon}_i^2 \sim \sigma^2 \chi^2(n - (2+1))$$

$$S^2 = \frac{\hat{\varepsilon}' \hat{\varepsilon}}{n - (2+1)}$$

unbiased for σ^2

we don't know, and we estimate it
 (that why we can not put it here)

$$(\hat{\beta} - \beta)' (\textcircled{6} (Z' Z)^{-1})^{-1} (\hat{\beta} - \beta) \sim \chi^2(n+1)$$

~ Mahalanobis distance

$$\frac{1}{S^2} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \sim \chi^2(n-(2+1))$$

$$\Rightarrow \frac{1}{S^2} (\hat{\beta} - \beta)' (Z' Z) (\hat{\beta} - \beta)$$

$$\frac{2+1}{\frac{1}{S^2}} \sim F(2+1, n-(2+1))$$

(pivotel)

$$\frac{(\hat{\beta} - \beta)' (Z' Z) (\hat{\beta} - \beta)}{S^2} \sim F(2+1, n-(2+1))$$

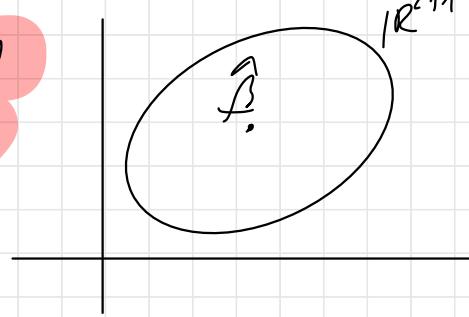
$$\lambda \in (0, 1)$$

← For F-test
checking in par-

tance of β :

$$CR_{1-\alpha}(F) = \left\{ \gamma \in \mathbb{R}^{2+1}: (\hat{y} - \hat{\beta})' Z' Z (\hat{y} - \hat{\beta}) \leq \right.$$

$$\leq (2+1) S^2 \int_{-\infty}^{\infty} (2_{11}, n \cdot (2_{11})) \Big\}$$



linear

Now for different V combinations

Given $\underline{Q} \in \mathbb{R}^{2 \times 1}$, $\mathcal{L} \in \mathbb{C}^{1,1}$

$$? \quad C \sum_{i=1}^n (\underline{Q}' \underline{\beta})$$

$$\underline{Q}' \underline{\beta} \sim N_1(\underline{Q}' \underline{\beta}), \quad D^2 \underline{Q}' (2' 2)^{-1} \underline{Q}$$

$$\frac{\underline{Q}' \underline{\beta} - \underline{Q}' \underline{\beta}}{\sqrt{D^2 \underline{Q}' (2' 2)^{-1} \underline{Q}}} \sim N(0, 1)$$

$$\sim N(0, 1) \quad \text{II} \quad \frac{\underline{\epsilon}' \underline{\epsilon}}{D^2} \sim \chi^2(n-(2+1))$$

$$\frac{\underline{Q}' \underline{\beta} - \underline{Q}' \underline{\beta}}{\sqrt{D^2 \underline{Q}' (2' 2)^{-1} \underline{Q}}} = S^2$$

$$\frac{N(0, 1)}{\sqrt{\chi^2(n-(2+1))}} \sim t(n-(2+1))$$

Student



$$\frac{\underline{Q}^T \hat{\beta} - \underline{Q}' \beta}{\sqrt{s^2 \underline{Q}' (\hat{\Sigma}^T \hat{\Sigma})^{-1} \underline{Q}}} \sim t(n-(p+1))$$

$$C_{1-\alpha}(\underline{Q}' \beta) = \left[\underline{Q}' \hat{\beta} \pm t_{1-\frac{\alpha}{2}}(n-(p+1)) \sqrt{s^2 \underline{Q}' (\hat{\Sigma}^T \hat{\Sigma})^{-1} \underline{Q}} \right]$$

one-at-the-time $C_{1-\alpha}(\hat{\beta}_j) \forall j \in \mathbb{R}^{p+1}$

$$\hat{\beta}_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow i \quad \text{So if } H_0: \beta_i = 0 \Rightarrow \text{may be we can remove this feature cov}(\hat{\beta})$$

Special case:

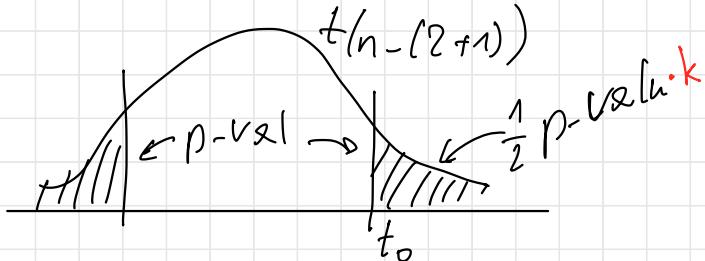
$$C_{1-\alpha}(\beta_j) = (\hat{\beta}_j \pm t_{1-\frac{\alpha}{2}}(n-(p+1)) \sqrt{s^2 \text{diag}_j(\hat{\Sigma}^T \hat{\Sigma})^{-1}})$$

$$j = 0, \dots, p$$

$$H_0: \beta_j = 0 \quad \text{vs} \quad H_1: \beta_j \neq 0$$

Reject at level $\alpha/2$ if $\beta_j \notin C_{1-\alpha}(\beta_j)$

They are not simultaneous



$$t_b = \frac{\hat{\beta}_j - \beta}{\sqrt{\sum \text{diag}(\hat{\Sigma}^{-1})}}$$

in some software ~~*~~~~*~~~~*~~ - means that multiplying by Bonferroni
0.00005 - * 1000 - still will be rejected

less degrees of freedom for model - more precise

from Maximum Lemma:

$$\max_{\underline{Q} \in \mathbb{R}^{2+1}} \frac{(\underline{Q}' \hat{\beta} - \underline{Q}' \beta)^2}{S^2 \underline{Q}' (\hat{\Sigma}^{-1}) \underline{Q}} = \frac{1}{S^2} (\hat{\beta} - \beta)' \hat{\Sigma}^{-1} (\hat{\beta} - \beta) \sim \mathcal{L}(\sigma_1)$$

$$\sim (2+1) F(2+1, n-(2+1))$$

$$\text{Sim CI}_{1-\alpha} (\underline{Q}' \hat{\beta}) = \underline{Q}' \hat{\beta} \pm$$

$$\pm \sqrt{S^2 (2+1) F_{1-\alpha}(2+1, n-(2+1))} \left[\underline{Q}' (\hat{\Sigma}^{-1}) \underline{Q} \right]$$

$$\forall \underline{Q} \in \mathbb{R}^{2+1}$$

So for this CI we assumed that $\underline{\epsilon} \sim N_n(0, \sigma^2 I)$

$$\hat{\Sigma}^2 \sim \delta^2 \chi^2 / (n - (2 + 1))$$

$$\frac{(n - (2 + 1)) \delta^2}{\delta^2} \sim \chi^2 (n - (2 + 1))$$

$$P(\chi^2_{\frac{n-2}{2}} (n - (2 + 1)) \leq \frac{(n - (2 + 1)) \delta^2}{\delta^2}) \leq$$

$$\leq \chi^2_{\frac{n-2}{2}} (n - (2 + 1)) = 1 - \alpha$$

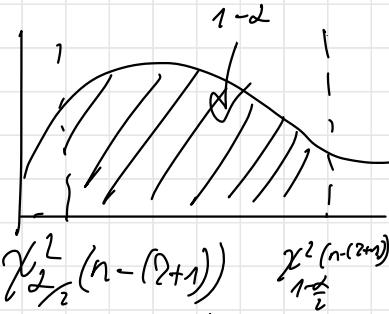
$$CI_{1-\alpha} [\delta^2] = \left[\frac{(n - (2 + 1)) \delta^2}{\chi^2_{\frac{n-2}{2}} (n - (2 + 1))}, \frac{(n - (2 + 1)) \delta^2}{\chi^2_{\frac{n+2}{2}} (n - (2 + 1))} \right]$$

C $p \times (2 + 1)$ matrix of known coefficients

$$H_0: C\beta = 0 \quad \text{vs} \quad C\beta \neq 0$$

$$H_0: C\beta = K_0 \quad \text{vs} \quad C\beta \neq K_0$$

$\text{so here can be any constant}$



F-test for any linear

$$C\hat{\beta} = \begin{pmatrix} C_{11} & C_{12} & \dots & C_{12+1} \\ C_{21} & C_{22} & \dots & C_{22+1} \\ \vdots & \vdots & \ddots & \vdots \\ C_{P_1} & & & C_{P_2+1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \vdots \\ \beta_2 \end{pmatrix} =$$

Combine-
tion

↑ number of linear combinations

$$= \begin{pmatrix} C_{11} \beta_0 + C_{12} \beta_1 + \dots + C_{12+1} \beta_2 \\ \vdots \\ C_{P_1} \beta_0 + \dots + C_{P_2+1} \beta_2 \end{pmatrix}$$

different
linear combi-
nation

Estimator of $C\beta$: $\hat{C}\hat{\beta} \sim_p N_p(C\beta, \sigma^2 C(Z^T Z)^{-1} C)$

$$(\hat{C}\hat{\beta} - C\beta)' [C(Z^T Z)^{-1} C]' (\hat{C}\hat{\beta} - C\beta) \sim \chi^2(p)$$

$$\frac{1}{\sigma^2} \sum \hat{\epsilon}' \hat{\epsilon} \sim \chi^2(n - (2+1)) \quad \text{Matematik}$$

$$\frac{(\hat{C}\hat{\beta} - C\beta)' [C(Z^T Z)^{-1} C]' (\hat{C}\hat{\beta} - C\beta)}{\frac{p}{\sigma^2}} \sim F(p, n - (2+1))$$

(pivotel)

$$\frac{1}{S^2} (C\hat{\beta} - C\beta)^T [C(z'z)^{-1}C]^T (C\hat{\beta} - C\beta) \sim \chi^2(p, n-(p+1)),$$

If H_0 is true, $C\beta = 0$

Reject H_0 at level α if

$$\frac{1}{S^2} (C\hat{\beta})^T [C(z'z)^{-1}C]^T C\hat{\beta} \geq \chi^2_{1-\alpha}(p, n-(p+1))$$

Bonferroni for small number of variables

Special case

$$H_0: \beta_2 = \beta_{2-1} = \dots = \beta_{2-(p-1)} = 0 \quad \text{vs } H_1:$$

$$\exists \beta_i \neq 0, i=2-(p-1), \dots, 2$$

$$Z = \begin{bmatrix} z_1 & z_2 \\ \vdots & \vdots \\ p \text{-variables} \end{bmatrix}$$

Comparing

$$Y = \mathcal{H} \beta + \varepsilon \quad \mathcal{H} \in \mathbb{R}^{n \times (2+1)}$$

Complete model

vs subspace of $L(\mathcal{H})$

$$Y = \mathcal{H}_1 \beta_1 + \varepsilon_1 \quad \mathcal{H}_1 \in \mathbb{R}^{n \times (2-(p-1))}$$

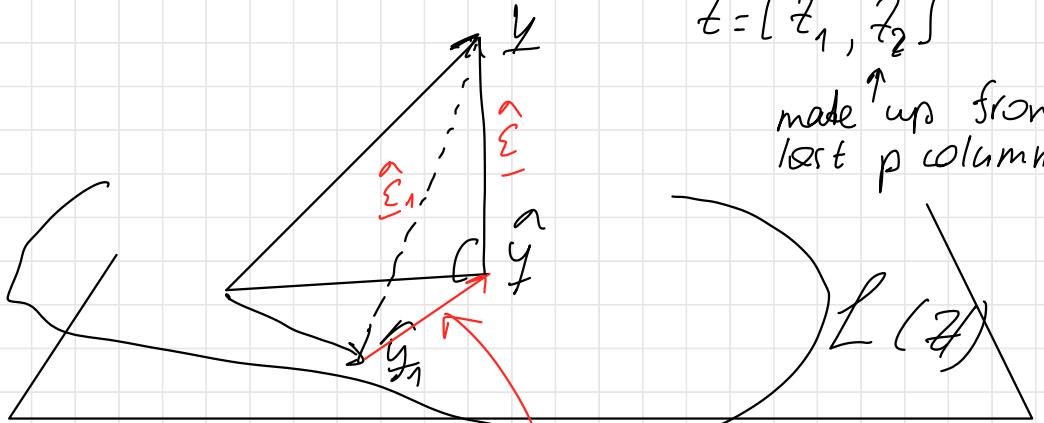
reduced model

$$C = \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ & & & & \ddots & & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & I_p \end{bmatrix}$$

$$(C\beta = 0 \Leftrightarrow \beta_2 = \beta_{2-1} = \dots = \beta_{2-(p-1)} = 0)$$

$$Z = [Z_1, Z_2]$$

made up from first p columns



$$L(\mathcal{H}_1) \subseteq L(\mathcal{H})$$

$$\hat{\Sigma}_1 \hat{\Sigma}_1 - \hat{\Sigma} \hat{\Sigma} =$$

$$= SS_{Res}(\mathcal{H}_1) - SS_{Res}(\mathcal{H})$$

reject if this quantity is large
(if difference small \Rightarrow models very similar and we accept it)

$$\frac{\hat{\epsilon}_1' \hat{\epsilon}_1 - \bar{\hat{\epsilon}}' \bar{\hat{\epsilon}}}{\sigma^2} \sim F(\beta, n-(r+1))$$

$\hat{\epsilon}_1' \hat{\epsilon}_1$

$\bar{\hat{\epsilon}}' \bar{\hat{\epsilon}}$

σ^2

reduction of dimension from $L(2)$ to $L(1)$

$$k(z_1) = \frac{\hat{\epsilon}' \hat{\epsilon}}{n-(r+1)} = \frac{S_{res}(z_1)}{n-(r+1)}$$

Very special case

$$H_0: \beta_1 = \beta_2 = \dots = \beta_r \geq 0 \text{ vs } H_1: \beta_i \neq 0 \quad i=1\dots r$$

$$Z_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad (\text{so we have only intercept})$$

$H_1:$ (in addition to intercept will be at least one more regressor β_i)

$$\hat{\epsilon}_1' \hat{\epsilon}_1 - \bar{\hat{\epsilon}}' \bar{\hat{\epsilon}} = S_{res}(Z_1) - S_{L_{res}}(2) =$$

$$= \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

so in this case our prediction - simple mean

Reject if

$$\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\frac{\sum_{i=1}^n \hat{e}_i^2}{n-(2+1)}} \geq F_{1-\alpha}(c, n-(2+1))$$

Also here

$$p=r$$

so if p-value of this test small \Rightarrow
at least one variable have effect

We are just checking if the variability explained by the model with respect to the variability explained by the residual is big enough to guarantee that the model has some meaning or not

lets go To prediction

$$y = \hat{y}/\beta + \varepsilon \quad \text{fitted model } y = \hat{y}/\hat{\beta}$$

New unit

\underline{z}_0 - observation of $\underline{z}_1, \dots, \underline{z}_r$

the next unit

"Predict $y_0"$

$$y_0 = \underline{z}_0' \underline{\beta} + \varepsilon_0$$

expected value

Estimating:

$$E[\underline{y}_0 | \underline{z}_0] = \underline{z}_0' \hat{\beta} = \beta_0 + \beta_1 z_{01} + \dots + \beta_2 z_{02}$$

obvious estimator: $\underline{z}_0' \underline{\beta}$

$$E[\underline{z}_0' \hat{\beta}] = \underline{z}_0' \underline{\beta} \quad \text{Var}(\underline{z}_0' \hat{\beta}) = \sigma^2 \underline{z}_0' (\underline{z}' \underline{z})^{-1} \underline{z}_0 \Rightarrow$$

$$\Rightarrow \underline{z}_0' \hat{\beta} \sim N(\underline{z}_0' \underline{\beta}, \sigma^2 \underline{z}_0' (\underline{z}' \underline{z})^{-1} \underline{z}_0)$$

Gauss Markov Theorem

$\underline{z}_0' \hat{\beta}$ is the minimum variance estimator

estimator among the estimators of $\underline{z}_0' \underline{\beta}$

s.t. (1) unbiased

(2) linear functions of y

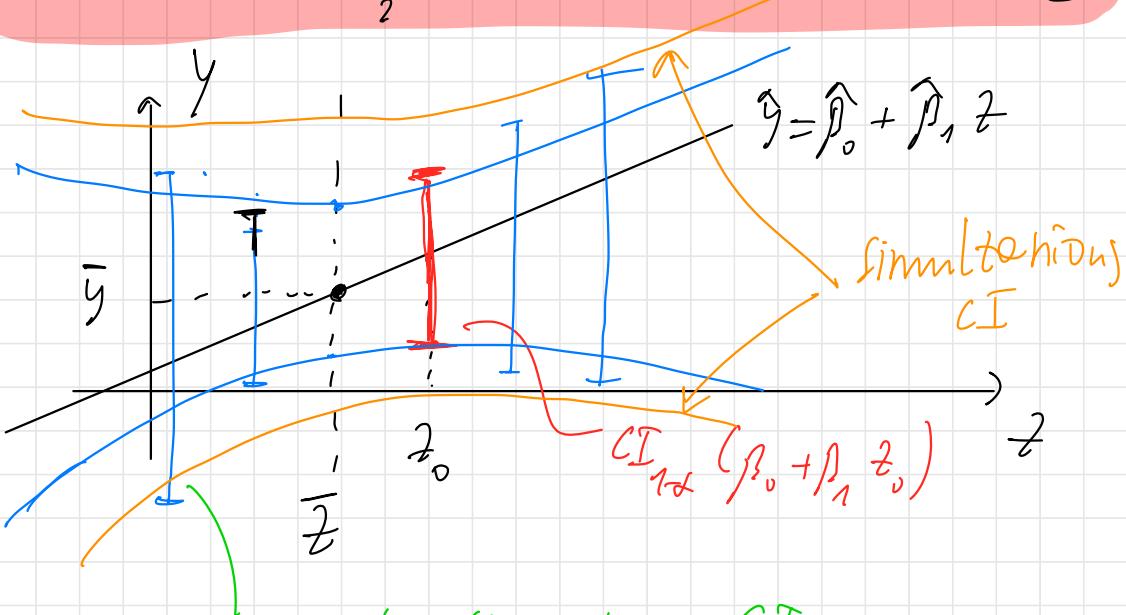
i.e. $\underline{z}_0' \hat{\beta}$ is BLUE

Best Linear Unbiased estimator

Note: $\underline{z}_0' \hat{\beta} = \underline{z}_0' \underbrace{(\underline{z}' \underline{z})^{-1} \underline{z}' y}_{\hat{\beta}}$

$$CI_{1-\alpha} (\hat{y}_0 | \underline{z}_0) = CI_{1-\alpha} [\underline{z}_0' \hat{\beta}] = \text{Interval for mean estimate here we have } CI \text{ for each to}$$

$$= \left[\underline{z}_0' \hat{\beta} \pm t_{1-\frac{\alpha}{2}} \frac{(n-(2+1)) \sqrt{s^2 \underline{z}_0' (\hat{\beta}' \hat{\beta})^{-1} \underline{z}_0}}{2} \right]$$



But all together they are not 95% to be sure about all CI we have to take simultaneous CI

$$\text{Sim } CI_{1-\alpha} (\underline{z}_0' \hat{\beta}) = \left[\underline{z}_0' \hat{\beta} \pm \sqrt{(2+1) F_{1-\alpha} (2+1, n-(2+1))} \right]$$

$$\cdot \sqrt{s^2 \underline{z}_0' (\hat{\beta}' \hat{\beta})^{-1} \underline{z}_0}$$

We can see that the closer we are to the barycenter \bar{z} the smaller the intervals are. This is because the term $\sqrt{\bar{z}^T (\bar{z}^T \bar{z})^{-1} \bar{z}}$ which increases as we move from the mean of the data.

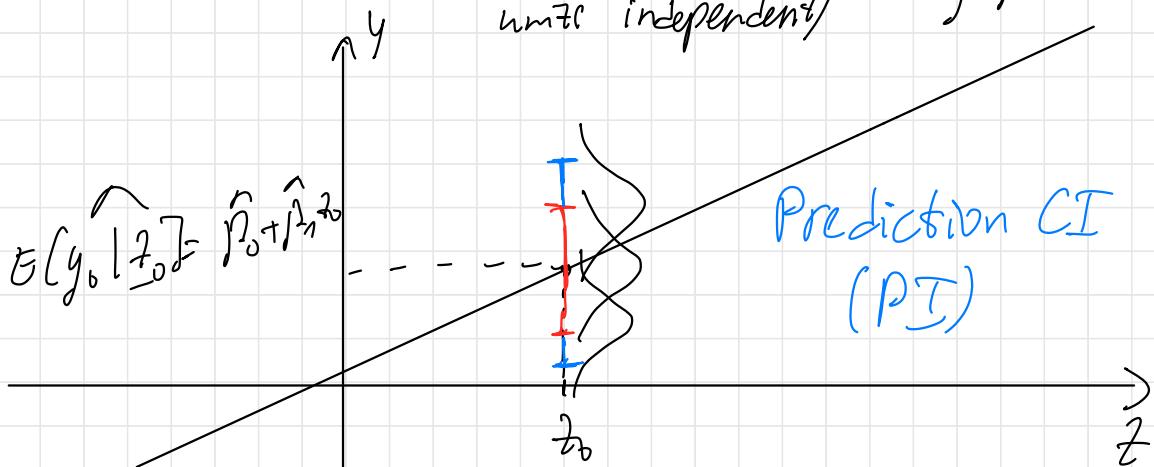
With simultaneous CI we can guarantee that the overall confidence level is $1-\delta$ percent even if make 1 billion prediction! With this new region moreover there will be the true linear model with confidence $1-\delta$.

- Also notice that now we plotted CI for the mean of the distribution of y_0

$$y_0 = \underline{z}' \hat{\beta} + \varepsilon_0$$

$\varepsilon_0 \perp \varepsilon$
(because all
units independent)

$$\hat{y} = \hat{y}_0 + \hat{\beta}' \underline{z}$$



? I_0 interval $P[y_0 \in I_0 | \underline{z}_0] = 1-\alpha$

$$\text{II } y_0 \sim N(\underline{z}' \hat{\beta}, \sigma^2)$$

$$\underline{z}' \hat{\beta} \sim N(\underline{z}' \hat{\beta}, \sigma^2 \underline{z}' (\mathbf{Z}' \mathbf{Z})^{-1} \underline{z})$$

$$y_0 - \underline{z}' \hat{\beta} \sim N\left(0, \sigma^2 \left[1 + \underline{z}' (\mathbf{Z}' \mathbf{Z})^{-1} \underline{z}\right]\right)$$

$$y_0 - \underline{z}' \hat{\beta}$$

$$\frac{1}{\sigma^2 \left[1 + \underline{z}' (\mathbf{Z}' \mathbf{Z})^{-1} \underline{z}\right]} \sim N(0, 1)$$

and $\sum \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\sigma^2 (n-(r+1))}$ is a chi-squared

$$\frac{y_0 - \hat{z}_0^T \hat{\beta}}{\sqrt{\sigma^2 \left[1 + \frac{1}{n} (\hat{z}^T \hat{z})^{-1} \frac{1}{\hat{z}_0} \right]}} \sim t(n-(2+1))$$

$\int \frac{d}{\sigma^2}$

$$P \left[t_{1-\alpha/2} (n-(2+1)) \leq \frac{y_0 - \hat{z}_0^T \hat{\beta}}{\sqrt{\sigma^2 \left[1 + \frac{1}{n} (\hat{z}^T \hat{z})^{-1} \frac{1}{\hat{z}_0} \right]}} \right] <$$

$$\leq t_{1-\alpha/2} (n-(2+1)) \Big] = 1-\alpha$$

PI
 Prediction interval for y_0 : One at the time
 (not simultaneous)

$$\left[\hat{z}_0^T \hat{\beta} \pm t_{1-\alpha/2} (n-(2+1)) \sqrt{\sigma^2 \left[1 + \frac{1}{n} (\hat{z}^T \hat{z})^{-1} \frac{1}{\hat{z}_0} \right]} \right]$$

to take into account variability of mean

Obviously 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

Conclusion:

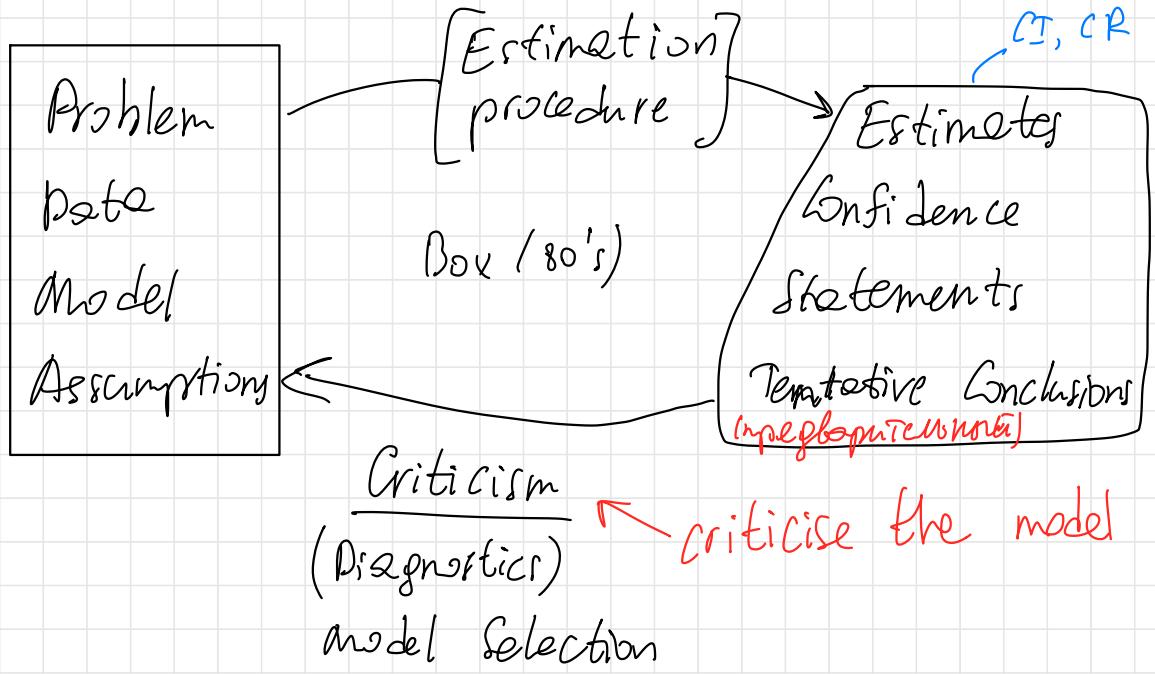
- Confidence interval are used for mean of what we want to predict
- Prediction interval are used for what we predict.

Simultaneous PI:

$$\text{Sim PI}_{1-\alpha} = \left[\hat{\beta}_0^T \hat{\beta} \pm \sqrt{1 + \hat{\beta}_0^T (\hat{\beta}^T \hat{\beta})^{-1} \hat{\beta}_0} \sqrt{(r+1) F_{\alpha}(r+1, n-(r+1))} \right]$$

Lecture 29. 04. 25

Diagnosis for linear models



Model

$$Y = Z\beta + \varepsilon$$

$$\varepsilon \in \mathbb{R}^n$$

$$E[\varepsilon] = 0$$

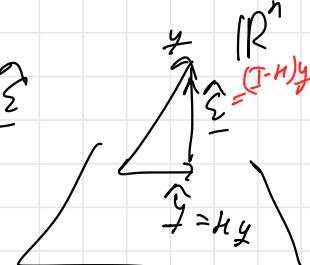
Fitted Model

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

$$\hat{\varepsilon} \in L^1(Z)$$

$$\dim: n - (p+1)$$

$$E[\hat{\varepsilon}] = 0$$



$$L(Z)$$

$$\dim: 2n$$

$$\hat{\varepsilon} \in L^1(Z)$$

$$\dim: n - (p+1)$$

Diagnostics for linear models is based on some basic tools:

- Residual Analysis: we look for outliers, we check for heteroscedasticity (i.e. non constant variance), we check for normality, we check for auto-correlation, and so on so forth
- Influential Cases (Statistical Units)
- Collinearity among the regressors

We know that $E(\hat{\varepsilon}) = 0$ and $\text{Cov}(\hat{\varepsilon}) = \sigma^2(I - H)$ so the residuals are correlated if ε is Gaussian then $\hat{\varepsilon}$ is a singular Gaussian: $\hat{\varepsilon} \sim N_n(0, \sigma^2(I - H))$ so $\hat{\varepsilon}$ are not realisation of ε

$$\text{Var}(\underline{\varepsilon}) = \sigma^2 I$$

not necessarily Gaussian

$$\underline{\varepsilon} \sim N_n(0, \sigma^2 I)$$

homoscedastic

(if all its random variables have the same finite variance - homogeneity of variance)

$$\text{Var}(\hat{\underline{\varepsilon}}) = \sigma^2 (I - H)$$

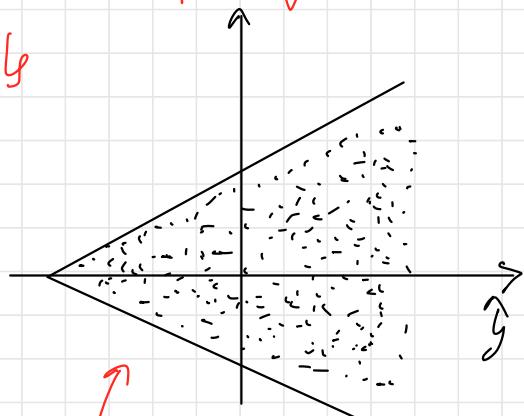
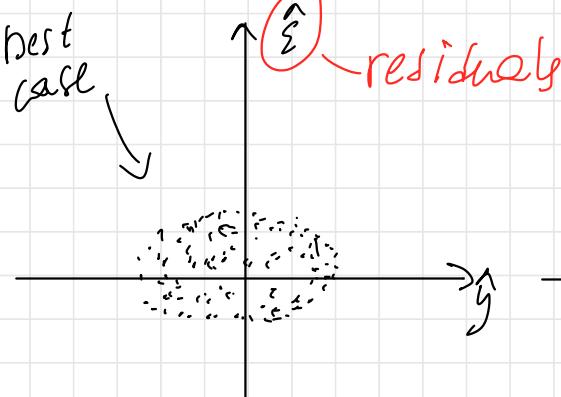
$$\hat{\underline{\varepsilon}} \sim N_n(0, \sigma^2 (I - H))$$

diagnostics going on here.

↙ (homoscedastic)

Residual analysis

random variables have different variance



How to fix it

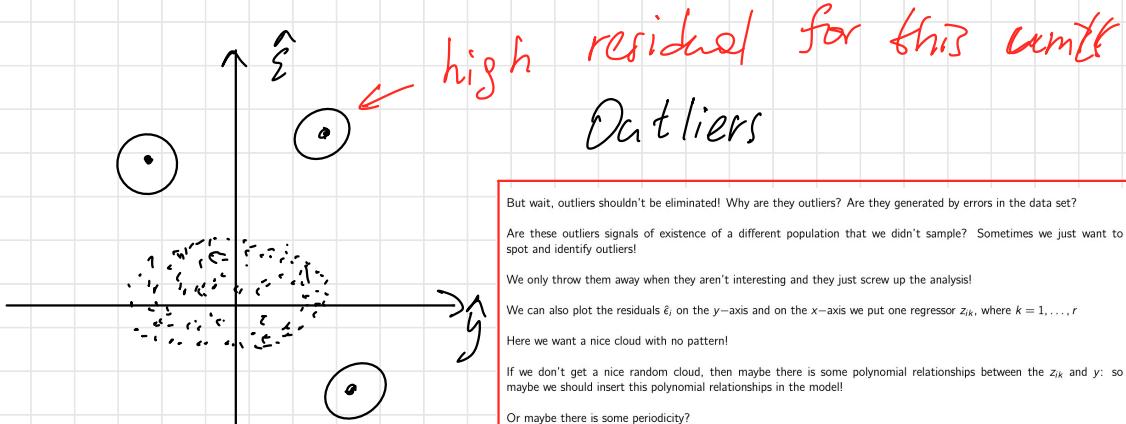
→ transforming features or y

→ Model $\underline{\varepsilon}$ s.t. $\text{cov}(\underline{\varepsilon}) = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix}$

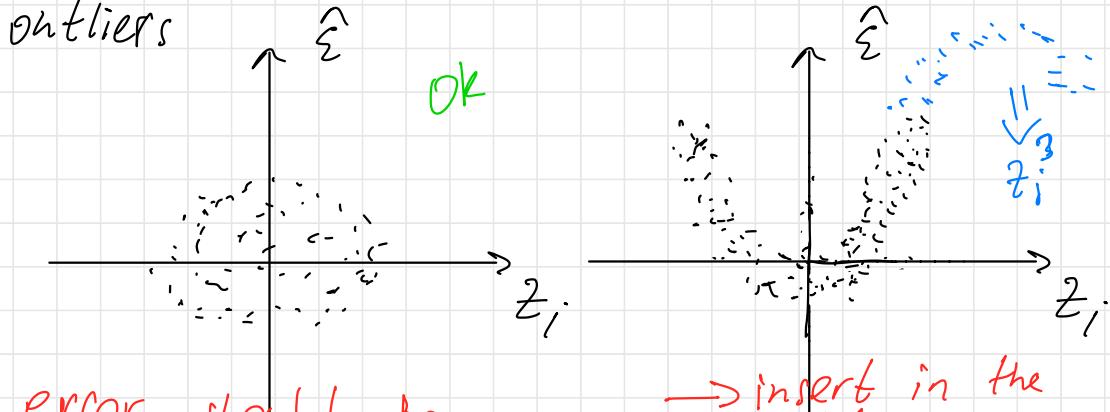
$$(\text{general } \text{cov}(\underline{\varepsilon}) = \sigma^2 \Sigma)$$

weighted sphere

→ LMM (linear mixed models)



Often outliers defined by mistakes,
But sometimes we still want look on
outliers



error should be
independent from
features

→ insert in the
model z_j^2

Gaussianity: We want to check Gaussianity, by making a QQ plot of the residual or of the studentised Residuals.

The residuals will never be Gaussian, we just need to check they are not too far from being Gaussian!

Note: In time series analysis we are worried about auto correlation: how large is the auto correlation? We can use the Durbin-Watson test!

$$\text{Cov}(\underline{\varepsilon}) = \sigma^2 I$$

$$\text{Var}(\varepsilon_i) = \sigma^2$$

$$\text{Cov}(\widehat{\varepsilon}) = \sigma^2 (I - H)$$

$$\text{Var}(\widehat{\varepsilon}_i) = \sigma^2 (1 - h_{ii})$$

h_{ii} = diagonal of $H =$

$$= \text{diag}(H(Z^T Z)^{-1} Z)$$

leverage

Hence consider

$$\frac{\widehat{\varepsilon}_i}{\sqrt{\sigma^2(1-h_{ii})}}$$

standardised residuals \Rightarrow student residuals

repeat residual analysis, using standardised residuals

$$0 \leq h_{ii} \leq 1$$

if $h_{ii} \approx 1 \Rightarrow \text{Var}(\widehat{\varepsilon}_i) \approx 0 \Rightarrow$
 $\Rightarrow \widehat{\varepsilon}_i \approx 0$ since $E[\widehat{\varepsilon}_i] = 0$ by definition

Ex. prove it

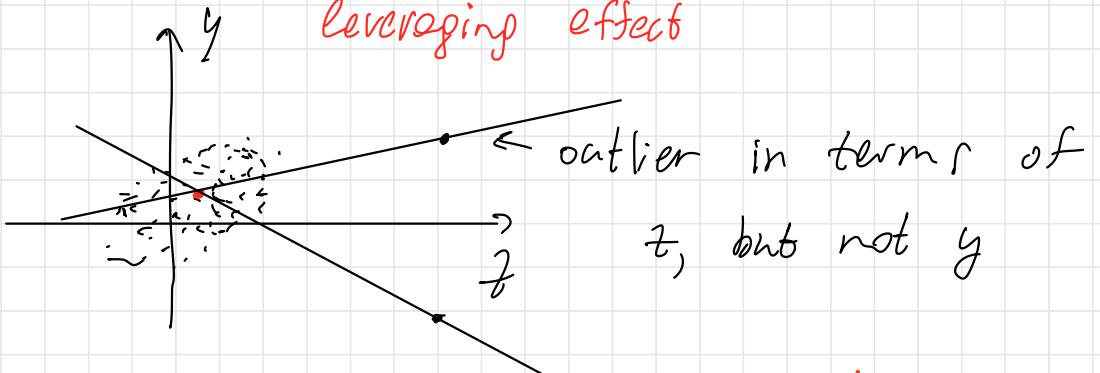
Hint: use the fact that $H = HT$ and $H^T H = H$

$$\text{If } h_{ii} = 1$$

$$\Rightarrow \text{Var}(\widehat{\varepsilon}_i) = 0 \quad E[\widehat{\varepsilon}_i] = 0$$

but h_{ii} only depends on the

design matrix, so even before doing regression we can make sure, by using the design matrix, that an influential point has a small effect



$h_{ii} = 1$ - very bad, because doesn't show anything $E(\hat{\epsilon}_i) = 0$, $\text{Var}(\hat{\epsilon}_i) = 0$

Influential Cases

training set $[Z][Y] : \begin{bmatrix} z_1' \\ \vdots \\ z_n' \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

training set / unit i $[Z_{-i}] [Y_{-i}] = \begin{bmatrix} z_1' \\ \cancel{z_i'} \\ \vdots \\ z_n' \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ \cancel{y_i} \\ y_n \end{bmatrix}$

fit model on $Z_i, Y \Rightarrow \hat{\beta} = (Z^T Z)^{-1} Z^T Y \in \mathbb{R}^{2+1}$

fit model on $Z_{-i}, Y_{-i} \Rightarrow \hat{\beta}_{-i} = (Z_{-i}^T Z_{-i})^{-1} Z_{-i}^T Y_{-i} \in \mathbb{R}^{2+1}$

$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(-i)})^T Z^T Z (\hat{\beta} - \hat{\beta}_{(-i)})}{S^2 (2+1)}$ Cook's distance

if $D_i > 1$ \Rightarrow distance is bigger than we expected

$$D(\hat{\beta}, \hat{\beta}_{(-i)}) = D_i$$

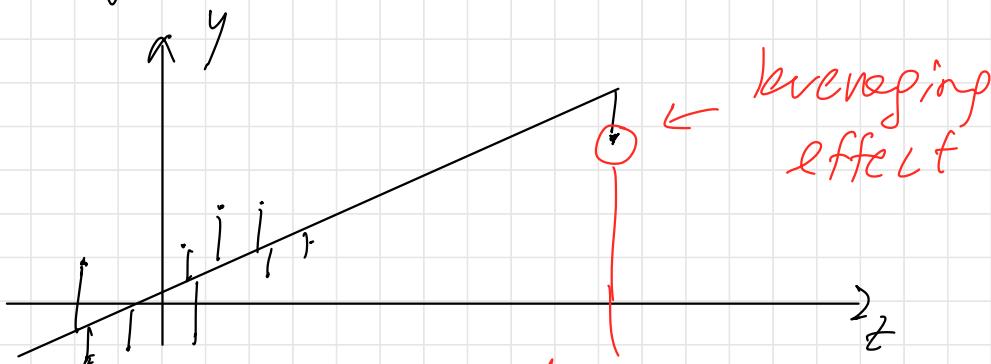
i is influential if D_i is large
if $\hat{\beta}$ outside then they are very distant

$$CR_{1-\alpha}(D) = \left\{ \gamma : D(\gamma, \hat{\beta}) \leq F_{1-\alpha}(2+1, n-(2+1)) \right\}$$

Note:

$$D_i = \left(\frac{\hat{\epsilon}_i}{\sqrt{s^2(1-h_{ii})}} \right)^2 \cdot \frac{h_{ii}}{1-h_{ii}} \cdot \frac{1}{2+1}$$

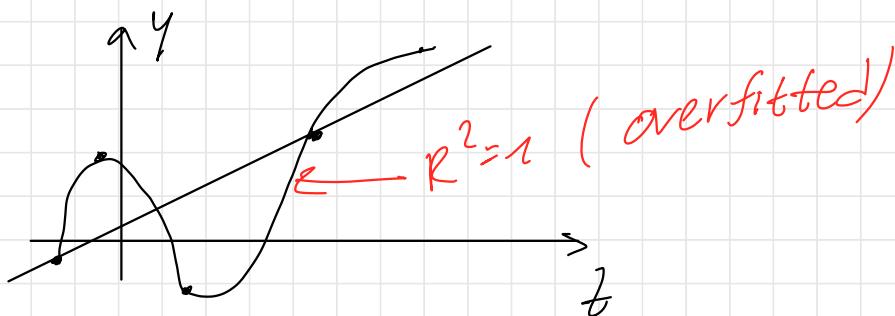
student residuals leverage



this point will influence our model

Assume $\epsilon \sim N(0, \sigma^2 I)$ → Check Gaussianity of $\hat{\epsilon}$
 residuals uncorrelated

→ Check for autocorrelation of the residuals Darwin motion



Collinearity (\longleftrightarrow Model Selection)

columns of design matrix close to be linear dependent. $\Rightarrow (z' z)^{-1}$ - singular
and we can not take $^{-1}$

$$\hat{\beta} = (z' z)^{-1} z' y$$

if variability of \hat{z} exploses \Rightarrow not tractable
numbers

$$\text{cov}(\hat{\beta}) = \sigma^2 (z' z)^{-1}$$

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^n (z_{ij} - \bar{z}_j)^2} \cdot \frac{1}{1 - R_j^2}$$

↑ variability of z_j

R_j^2 - is the (R^2) coefficient of det
when z_j is regressed on $1, z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_n$

$\text{Var}(\hat{\beta}_j) \downarrow$ if $\sum_{i=1}^n (z_{ij} - \bar{z}_j)^2$ (\Rightarrow non collinearity)

$\text{Var}(\hat{\beta}_j) \uparrow$ if $R_j^2 \rightarrow 1$ (\Rightarrow collinearity)

$$\text{VIF} = \frac{1}{1 - R_j^2} \quad \text{Variance Inflation Factor}$$

Here we won't apply PCA, to
make features orthogonal

$\text{VIF} \uparrow \Rightarrow R_j \rightarrow 1 \Rightarrow \text{Var}(\hat{\beta}_j) \uparrow \Rightarrow$ collinearity

$\text{VIF} \downarrow \Rightarrow R_j \rightarrow 0 \Rightarrow \text{Var}(\hat{\beta}_j) \downarrow \Rightarrow$ non collinearity

Model selection

Given: \mathbf{Z}

$$\mathbf{Z} \quad n \times (2+1) \quad 2 \text{ features}$$

as simple model — as more degree of freedom — as less variability of residuals we will have $\text{mean} \rightarrow$ 0 regressor 1

How many models with 1 regressor 2

$$\text{with } 2 \text{ regressors } \binom{2}{2} = \frac{2!}{2!(2-2)!}$$

How many models

$$1) \binom{2}{1} + \binom{2}{2} + \dots + \binom{2}{2} = 2^2$$

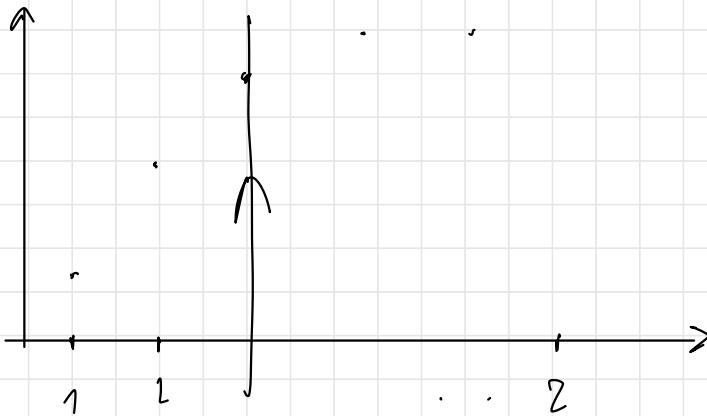
All possible
models
with fixed
 k

Explore them all

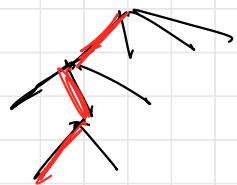
For $k=1, \dots, 2$

1. Fit the $\binom{2}{k}$ models with k regressors
2. Choose the best in term of R^2

$$\text{next } k \Rightarrow R_1^2 \leq R_2^2 \dots \leq R_k^2$$



Use R^2_{adjusted} or $\begin{pmatrix} \text{AIC} \\ \text{BIC} \end{pmatrix}$



Forward selection

- Start with best model with 1 regressor
- Add a regressor and find the best model with 2 regressors
- Run an F test to check if $\beta_i \neq 0$
- Add more regressor until F-test tells to stop ($\beta_i = 0$)

if we keep power $n \Rightarrow$ we have
to keep all previous factors

$$y = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_1 t_2 =$$

$$= \beta_0 + \beta_1 (1 + \beta_1 t_2) t_1 + \dots$$

So if I will use $t_2^3 \Rightarrow$ I need
also save t_2, t_2^2

Lecture 5. 05. 25

Generalisation of OLS (ordinary least squares)

GLS

Model: $y = \beta \mathbf{z} + \varepsilon$

OLS: $\text{Cov}(\varepsilon) = \sigma^2 I$ $\rightarrow \text{Cov}(\varepsilon) = W\sigma^2$

lets assume $W: n \times n$ positive defined matrix

$$\underset{\text{argmin}}{(y - \beta \mathbf{z})' W^{-1} (y - \beta \mathbf{z})} =$$

(OLS: $W = I$) \Rightarrow least square error

$$= (y - \beta \mathbf{z})' W^{-1/2} W^{1/2} (y - \beta \mathbf{z}) =$$

$$= (\underbrace{W^{-1/2} y - \underbrace{W^{-1/2} \beta \mathbf{z}}_y}_y^* \quad \underbrace{W^{1/2} (y - \beta \mathbf{z})}_z^*)' (\underbrace{W^{-1/2} y - \underbrace{W^{-1/2} \beta \mathbf{z}}_y}_y^* \quad \underbrace{W^{1/2} (y - \beta \mathbf{z})}_z^*) =$$

$$= (y^* - \mathbf{z}^* \beta)^* (y^* - \mathbf{z}^* \beta) = \|y^* - \mathbf{z}^* \beta\|^2$$

$$\hat{\beta} = (\mathbf{z}^* \mathbf{z})^{-1} \mathbf{z}^* y^* = (\mathbf{z}^{-1} W^{-1} \mathbf{z})^{-1} \mathbf{z}^* W^{-1} y$$

general solution

$$\text{cov}(y^*) = v^{-1/2} \text{cov}(y) v^{-1/2} = v^{-1/2} \text{cov}(\varepsilon) v^{-1/2}$$

Example

$$y = Z\beta + \varepsilon$$

$$\text{cov}(\varepsilon) = \sigma^2 \Sigma$$

σ^2 unknown
 Σ known

Take $v = \Sigma \Rightarrow \hat{\beta} = (Z'Z)^{-1}Z'\varepsilon^{-1}y$

Note :

$$\text{cov}(y^*) = \text{cov}(\varepsilon^{-1}y) = \sigma^2 \Sigma^{-1/2} \Sigma \varepsilon^{-1/2} = \sigma^2 I$$

(come back to OLS mode)

Particular case

$$\Sigma = \text{diag}(w_1, \dots, w_n), \quad \Sigma = \begin{bmatrix} w_1 & & & 0 \\ & \ddots & & \\ 0 & & \ddots & \\ & & & w_n \end{bmatrix}$$

WLS (weighted least squares)

For instance y_i (average of n_i units)

$$\text{Var}(y_i) = \frac{1}{n_i} \sigma^2 \quad (\text{more trust to sample of many units})$$

$$\Rightarrow w_i = \frac{1}{n_i} \quad v = \text{diag}\left(\frac{1}{n_1}, \dots, \frac{1}{n_n}\right)$$

For instance:

y_i is the sum over n_i units

$$\text{Var}(y_i) = n_i \cdot \sigma^2 \Rightarrow w_i = n_i$$

Collinearity and variable selection

The OLS goes through barycenter of...

Note: the OLS fitted model goes through the barycenter of training set

$$y = \beta' \underline{x} + \varepsilon \quad (\text{no normality assumption})$$

$$\hat{\beta} = (\underline{x}' \underline{x})^{-1} \underline{x}' y \quad \text{OLS estimator of } \beta$$

$$\text{Fitted model: } \hat{y} = \underline{x}' \hat{\beta} \quad y \in \mathbb{R}$$

$$\underline{x} = (1 \ x_1 \ x_2 \ \dots \ x_n)' \in \mathbb{R}^{n+1}$$

$$\text{Take } \underline{x}_0 = (1 \ x_1 \ x_2 \ \dots \ x_n)' =$$

$$= \frac{\underline{x}' \underline{1}}{\underline{1}' \underline{1}}$$

$$y_0 = \hat{\beta}_0 \hat{z} + \hat{\beta}_1 \hat{z}^1 = \frac{\hat{\beta}_0 \hat{z}}{1^T \hat{z}} (\hat{z}^1 \hat{z}^1)^{-1} \hat{z}^1 y =$$

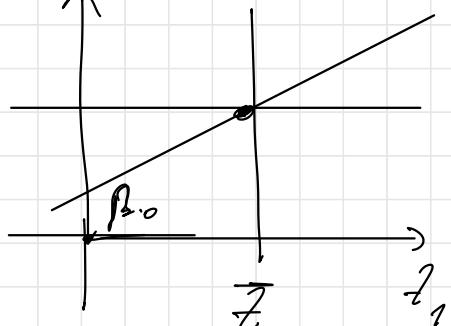
$$= \frac{1^T \hat{z}}{n} y = \frac{(n^{-1})^T y}{n} = \frac{1^T y}{n} = \bar{y} \Rightarrow$$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{z}_1 + \dots + \hat{\beta}_2 \bar{z}_2$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{z}_1 - \dots - \hat{\beta}_2 \bar{z}_2$$

Fitted model

$$y = \hat{\beta}_0 + \hat{\beta}_1 z_1 + \dots + \hat{\beta}_2 z_2$$



$$y - \bar{y} = \hat{\beta}_1 (z_1 - \bar{z}_1) + \dots + \hat{\beta}_2 (z_2 - \bar{z}_2)$$

so everything centered in barycenter

Considering $y \rightarrow y - \bar{y} \cdot 1 = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} = y^*$

$$\mathcal{Z} \rightarrow \begin{pmatrix} z_{11} - \bar{z}_1 & z_{12} - \bar{z}_2 & \dots & z_{12} - \bar{z}_2 \\ \vdots & \vdots & & \vdots \\ z_{n1} - \bar{z}_1 & z_{n2} - \bar{z}_1 & \dots & z_{n2} - \bar{z}_2 \end{pmatrix} \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} = z^*$$

O/LS:

$$\left\{ \begin{array}{l} \text{argmin } \|y^* - \mathcal{H}\beta\|^2 = \hat{\beta}^* = \begin{pmatrix} \hat{\beta}_1^* \\ \vdots \\ \hat{\beta}_n^* \\ \hat{\beta}_2^* \end{pmatrix} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1^* \bar{x}_1 - \dots - \hat{\beta}_2^* \bar{x}_2 \\ \hat{\beta}_i = \hat{\beta}_i^* \quad i=1\dots 2 \end{array} \right\} (*)$$

From now on, assume y and \mathcal{H} to be centered and then recover β by means of $(*)$

Principal Components Regression

$$\mathcal{H} = \begin{bmatrix} z_1 & \dots & z_2 \\ z_{11} & \ddots & z_{12} \\ z_{n1} & \dots & z_{n2} \end{bmatrix} \quad \begin{array}{l} \text{design matrix} \\ n \times 2 \end{array}$$

\Rightarrow PCA on $\mathcal{H} \Rightarrow PC_1 \dots PC_k = PC \dots PC_n$
with $k \leq 2$

$$Z^* = \begin{pmatrix} Z_{11}^* \\ Z_{n1}^* \\ \vdots \\ Z_{1k}^* \\ Z_{nk}^* \end{pmatrix}_{n \times k}$$

$\left. \begin{matrix} PC_1 & \dots & PC_k \end{matrix} \right\}$ $\left. \begin{matrix} Z_{11}^* \\ \vdots \\ Z_{nk}^* \end{matrix} \right\}$ $\left. \begin{matrix} Z_{11}^* \\ \vdots \\ Z_{nk}^* \end{matrix} \right\}$

uncorrelated \Rightarrow collinearity disappears
(by construction)

$$\hat{\beta}^* = (Z^* ' Z^*)^{-1} Z^* ' Y$$

Fitted model $y = \underline{z}^* ' \hat{\beta}^*$ $\underline{z}^* = (PC_1, \dots, PC_k)$

$$= \hat{\beta}_1^* PC_1 + \hat{\beta}_2^* PC_2 + \dots + \hat{\beta}_k^* PC_k$$

$$PC_1 = \ell_{11} Z_1 + \ell_{12} Z_2 + \dots + \ell_{1n} Z_n$$

$$PC_2 = \ell_{21} Z_1 + \ell_{22} Z_2 + \dots + \ell_{2n} Z_n$$

\vdots

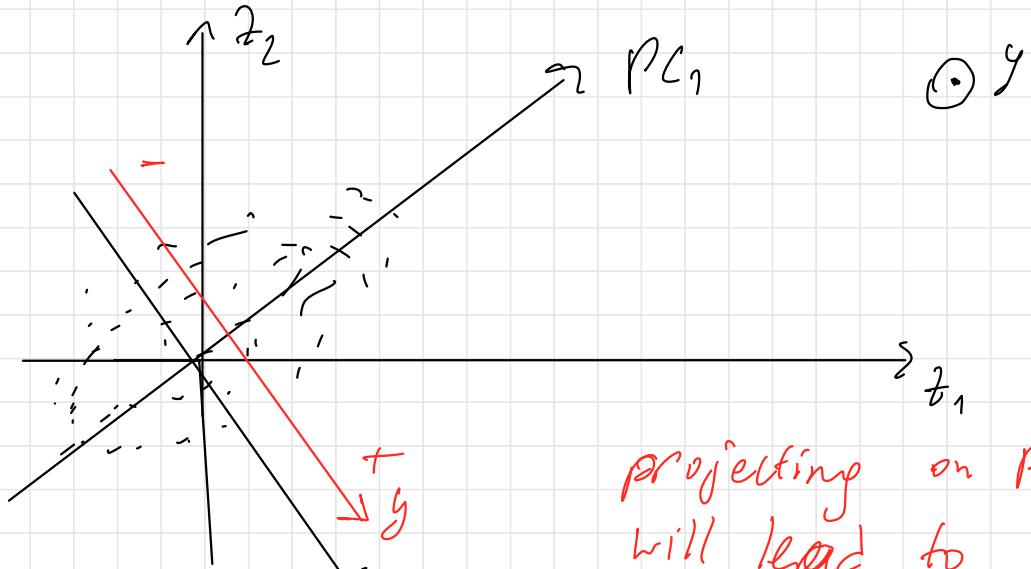
$$PC_k = \ell_{1k} Z_1 + \dots + \ell_{kk} Z_k$$

Fitted model:

$$y = z_1 (\ell_{11} \hat{p}_1^* + \ell_{12} \hat{p}_2^* + \dots + \ell_{1n} \hat{p}_n) +$$
$$+ z_2 (\ell_{21} \hat{p}_1^* + \ell_{22} \hat{p}_2^* + \dots + \ell_{2n} \hat{p}_n) +$$
$$\vdots$$
$$z_p (\quad \cdot \quad - \quad \cdot \quad)$$
$$y = \hat{f}_1 z_1 + \dots + \hat{f}_p z_p \quad \leftarrow \text{all variables will appear but we used } k \text{ PC}$$

Criticism:

- 1) \Rightarrow the model (*) is not sparse / no model selection
- 2) you are not guaranteed that $p_{C1} \dots p_{Ck}$ are the linear combination most correlated with y .



projecting on PC_1
will lead to

losing information
about classes

(correlation with $y = 0$)

Find PC , having correlation with y
(we not try increase variability)

- Canonical Correlation
- Partial least squares

Ridge regression (Hoerl & Kennard 70's)

Collinearity \Rightarrow high variability of

$$\text{Cov}(\hat{\beta}) = (\mathbf{z}' \mathbf{z}) \sigma^2 \quad \hat{\beta} = (\mathbf{z}' \mathbf{z})^{-1} \mathbf{z}' \mathbf{y} = \hat{\mathbf{B}}_{OLS}$$

Ridge Optimization problem:

$$\begin{cases} \underset{\mathbf{f} \in \mathbb{R}^n}{\text{argmin}} \|\mathbf{y} - \mathbf{f}\|_2^2 \\ \|\mathbf{f}\|_2^2 \leq s, \quad s > 0 \end{cases} \quad \mathbf{f} \in \mathcal{L}^\perp(\mathbf{z})$$

$$\|\mathbf{y} - \mathbf{f}\|_2^2 = \|\mathbf{y} - \underbrace{\mathbf{g} + \mathbf{A}(\mathbf{I} - \hat{\mathbf{B}})}_{\in \mathcal{L}(\mathbf{A})}\|_2^2$$

$$\hat{\mathbf{B}}_{OLS} = (\mathbf{A}' \mathbf{A})^{-1} \mathbf{A}' \mathbf{y}$$

$$\hat{\mathbf{y}} = \mathbf{A} \hat{\mathbf{B}} = \mathbf{A} (\mathbf{A}' \mathbf{A})^{-1} \mathbf{A}' \mathbf{y}$$

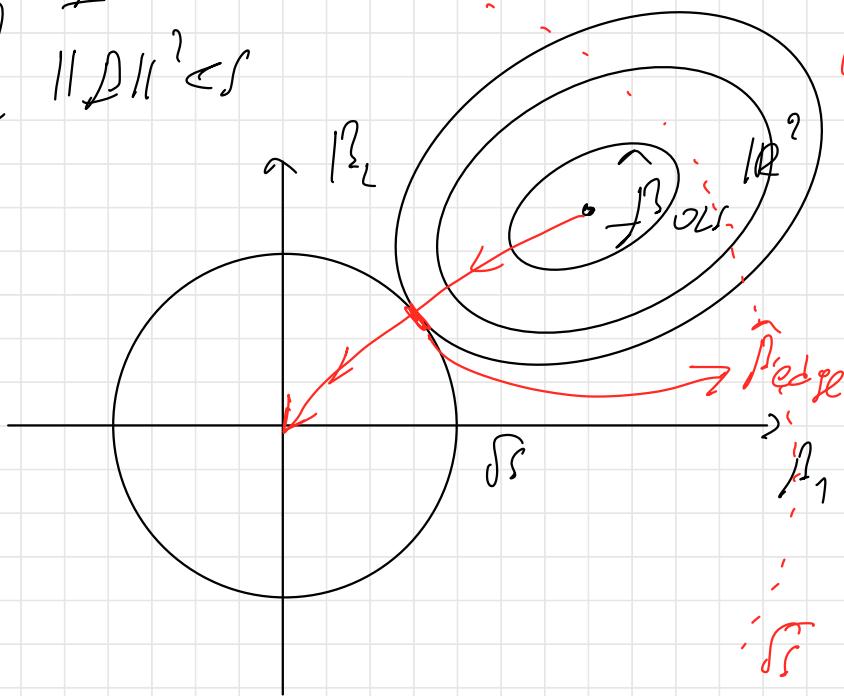
$$\hat{\mathbf{B}} = \|\hat{\mathbf{B}}\|_2^2 + \|\mathbf{A}(\mathbf{B} - \hat{\mathbf{B}})\|_2^2$$

Ridge problem

$$\begin{cases} \underset{\mathbf{B}}{\text{argmin}} \|\mathbf{A}(\mathbf{B} - \hat{\mathbf{B}})\|_2^2 \\ \|\mathbf{B}\|_2^2 \leq s \end{cases}$$

$$\left\{ \begin{array}{l} \underset{\beta}{\operatorname{argmin}} \quad (\beta - \hat{\beta})' \beta + (\beta - \hat{\beta}) \\ \|\beta\|_2^2 \leq s \end{array} \right.$$

ellipses



$\hat{\beta}_{\text{ridge}}$ shrunk estimator of β

To find bridge solve linear model with regularization

Ridge Prob

$$\left\{ \begin{array}{l} \underset{\beta}{\operatorname{argmin}} \quad \|y - z\beta\|^2 \\ \|\beta\|_2^2 \leq s \end{array} \right.$$

To solve consider the

Lagrangian

$$\|y - z\beta\|^2 + \lambda \|\beta\|^2 \quad \text{- optimize } (*)$$

d is a function of s

as larger d -> smaller circle
(stronger regularization)

$$\hat{\beta}_{\text{ridge}} = (\mathbf{Z}' \mathbf{Z} + d \mathbf{I})^{-1} \mathbf{Z}' \mathbf{y}$$

-1

to invert $\begin{cases} \text{solve problem} \\ \text{of collinearity} \\ \text{of variables} \end{cases}$

Criticism:

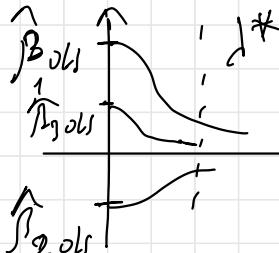
1. $\hat{\beta}_{\text{ridge}}$ is biased $E[\hat{\beta}_{\text{ridge}}] = (\mathbf{Z}' \mathbf{Z} + d \mathbf{I})^{-1} \mathbf{Z}' \mathbf{Z} \beta$

(so to decrease variability, we
lose biased of estimation)

2. H & K proved $\exists d^*$ s.t.

$$E[\|\mathbf{y} - \mathbf{Z}\hat{\beta}_{\text{ridge}}(d)\|^2] \leq E[\|\mathbf{y} - \mathbf{Z}\hat{\beta}_{OLS}\|^2]$$

3. Find the "right" d by cross validation



but with $\hat{\beta}_{\text{ridge}}$ we will never get $\hat{\beta}_j = 0$

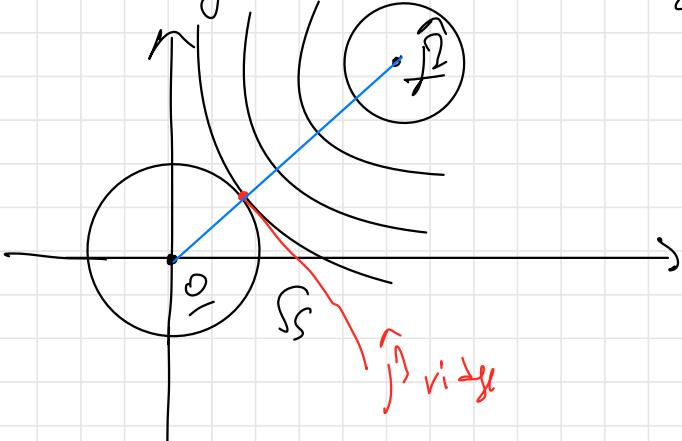
$\hat{\beta}_{DLS}$ - scaling independent
1kg \rightarrow 1000 gram (still
will work,
because
 $\beta_i \cdot 1000$
just be
multiplied)

Obs_i

1. Ridge regression is not scale invariant. \Rightarrow standardize y and z
before fitting

2. If variables are std (not ellipse but circle)

and regressions are orthogonal



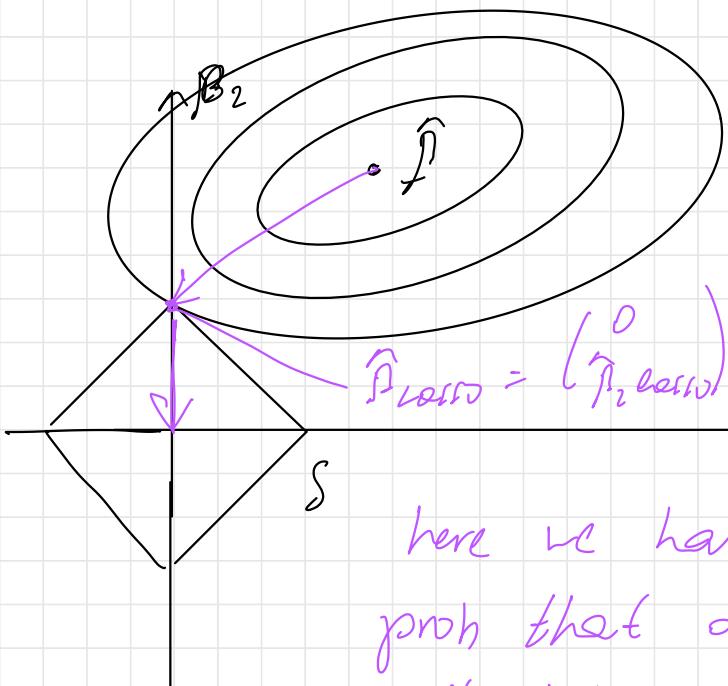
Lasso

(Tibshirani, 96)

instead of $\|\hat{\beta}\|^2 - \text{min}$

$$\begin{cases} \text{argmin } \|y - X\beta\|^2 \\ \|\beta\|_1 \leq s \quad \|\beta\|_1 = \sum |\beta_i| \end{cases}$$

Here we make variable selection



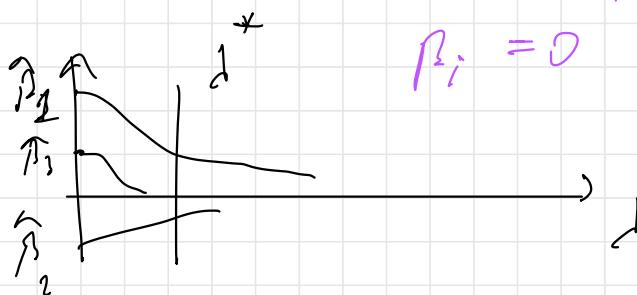
solve the Lagrangian:

$$\text{argmin}_{\beta} \|y - X\beta\|^2 + \lambda \|\beta\|_1$$

Lasso

here we have higher prob that our solution will intersect constraint in point, where one of

$$\beta_i = 0$$



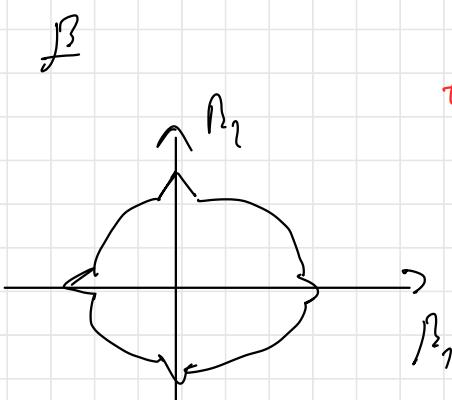
$$\begin{cases} \text{argmin } \|y - f(\beta)\|^2 \\ \|\beta\|_q \leq s \end{cases}$$

$\|f\|_q = \sqrt{\sum |f_i|^q}$ $q \leq 1$

solve model selection problem
(which variables we should use)

Elastic Net

$$\text{argmin}_{\beta} \|y - f(\beta)\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|^2$$



λ_1
takes care of variable selection

λ_2
takes care of collinearity
(to be non collinear)

$$\text{argmin}_{\beta} \|y - f(\beta)\|^2 + \lambda \|\beta\|_1$$

penalise description by远离 ideal solution

g. 05. 25

4 lessons



POLITECNICO
MILANO 1863



repeated late
hierarchically structured
red

Mixed Effect Models

Francesca Ieva

Alessandra Ragni

MOX – Department of Mathematics, Politecnico di Milano, Milan 20133, Italy

Applied Statistics Course

May 2025

Outline

1. Motivations

2. The Linear Mixed Models (LMMs) journey

- The very beginning: classical LM
- LM 2.0: relax the variance homogeneity assumption
- LM 3.0: relax the independence assumption
- LMM

3. Take home messages

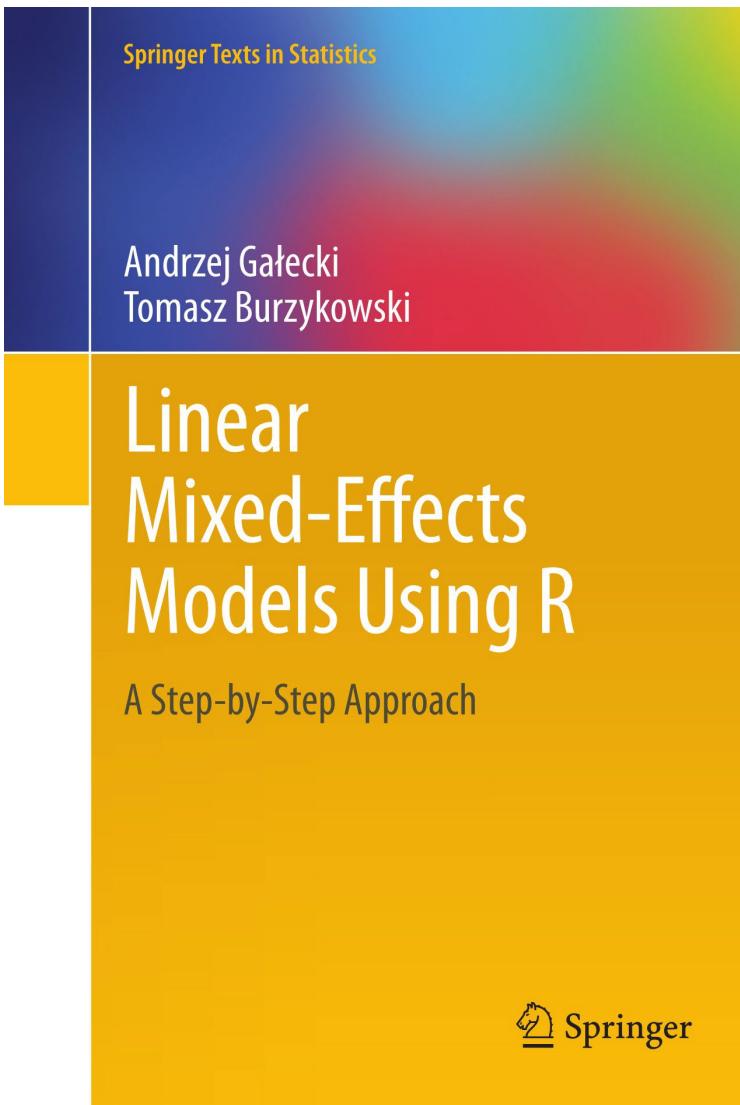
4. Case studies

5. R corner

6. References



Our companion for the journey



has
technical parts
with code



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.

(what about?
features)

two observation
independent

all observation
share same
variance
for the
error

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/administrative neighbours

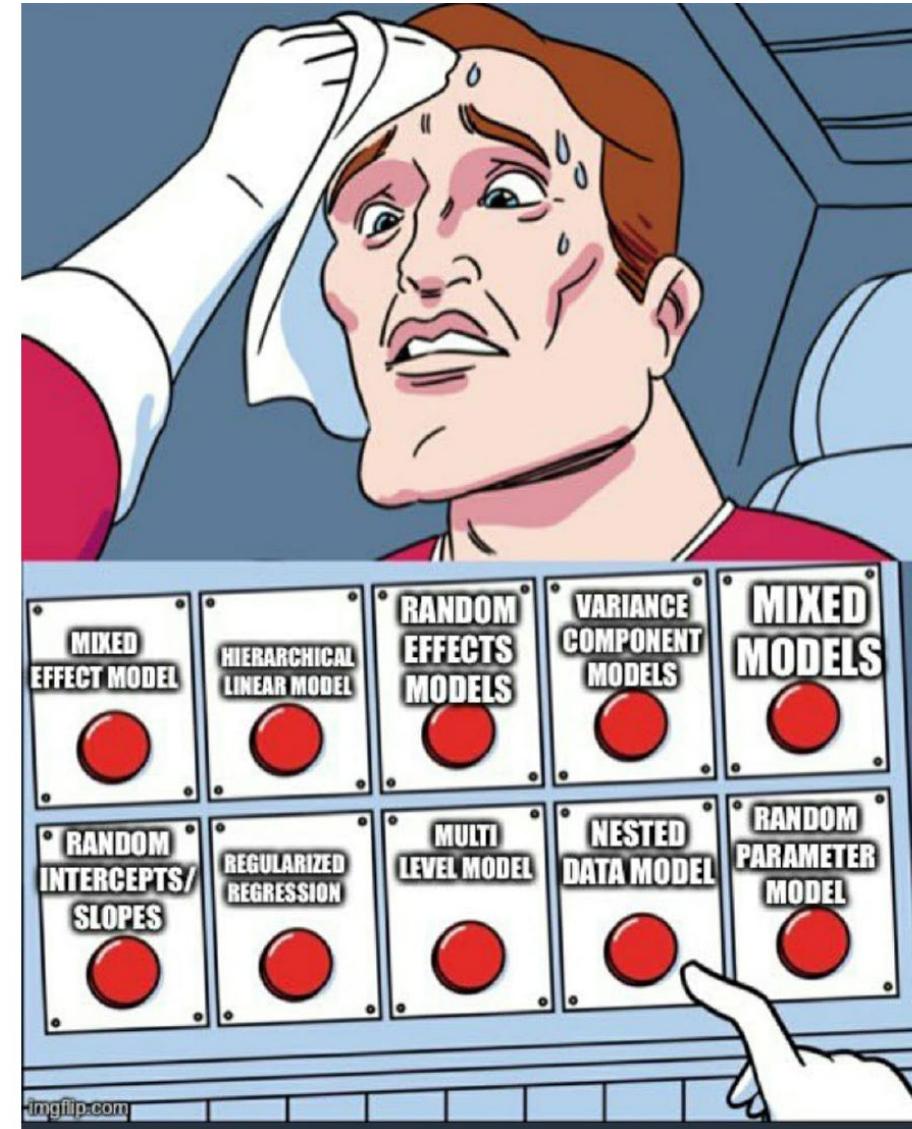
Same groups \Rightarrow have dependence =>
biased inference



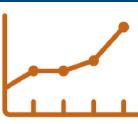
General overview

- 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 and denoise information contents, focusing on inference related to the presence of groups.

more we will be account + residual variability, the better understanding of data



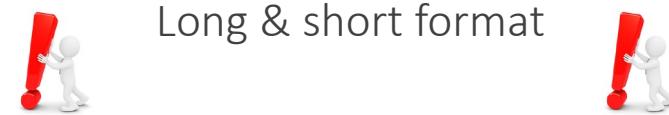
Motivating examples



1. ARMD - Age-Related Macular Degeneration Trial

- The ARMD data arise from a **randomized multi-center clinical trial** comparing an experimental treatment (interferon- α) versus placebo for patients diagnosed with ARMD.
- 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 (defined as the total number of letters correctly read)** 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).
- **Goal:** comparison between placebo and the highest dose (6 million units daily) of interferon- α .

=>*longitudinal data in the form of up to five visual acuity measurements collected at different, but common to all patients, timepoints.*



Motivating examples



2. Multi center trials

how number of patients
affect on the quality of hospital
center *management*

- 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



Motivating examples



4. RIJKZ Data

- Data from the marine benthic data from **9 beaches** along the Dutch coast.
- In each beach, **5 samples were taken at different sites**, and the macro-fauna and abiotic variables were measured.
- We want to **model the species richness at site j on beach i, given the height of site j on beach i compared to mean beach level (NAP_ij) and the exposure on beach i** [an index composed of the following elements: wave action, length of the surf zone, slope, grain size, and the depth of the anaerobic layer] ;
- **n = 45 samples within N = 9 beaches (45=5x9)**

Source: Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R* (Vol. 574, p. 574). New York: Springer.

○ Goals:

- Including the categorical variable factor(Beach) in the model so that the intercept can be estimated differently in each beach, knowing that it would cost 8 regression parameters;
- Test whether the relationship between species richness and NAP is different on each beach, so to include a NAP–Beach interaction term to the model, factor(Beach) + NAP : factor(Beach), so that both intercept and slopes vary across beaches;
- Deal with the fact that we need to estimate 16 “extra parameters” in a **dataset of dimensionality n=45**



The LMM journey

- The very beginning: classical LM
- LM 2.0: relax the variance homogeneity assumption
- LM 3.0: relax the independence assumption
- LMM



The LMM journey



The LMM journey



- 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** $y_j, j = 1, \dots, n$ with a constant variance can be specified in a variety of ways. A commonly used specification is:

Model equation
at the level of
observations

$$y_j = \beta_0 + \beta_1 x_{1j} + \cdots + \beta_p x_{pj} + \varepsilon_j = x_j^t \boldsymbol{\beta} + \varepsilon_j$$

σ^2 for ε

$$\varepsilon_j \sim N(0, \sigma^2)$$

also independence

p+1 PQM

j - index
of units



Note that

$$E[y_j] = \mu_j = x_j^t \boldsymbol{\beta}$$

$$V[y_j] = V[\varepsilon_j] = \sigma^2$$

?



- 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).
 - ↳ all elements on diagonal are equal
- The classical LM for **independent, normally distributed observations** $y_j, j = 1, \dots, n$ with a constant variance can be specified in a variety of ways. A commonly used specification is:

Model equation
for all data

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

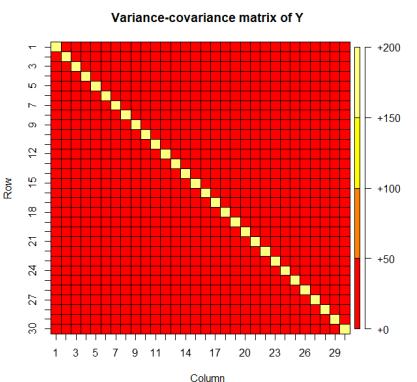
$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, R)$

Scale parameter

$$R = \sigma^2 I_n \in \mathbb{R}^{n \times n}$$

$\mathbf{y} = (y_1, \dots, y_n)'$ $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_j' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{p1} \\ \vdots & \dots & \dots & \vdots \\ 1 & x_{1n} & \dots & x_{pn} \end{pmatrix} = (\mathbf{1} \quad \mathbf{x}_1 \quad \dots \quad \mathbf{x}_p) \in \mathbb{R}^{n \times (p+1)}$$



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.

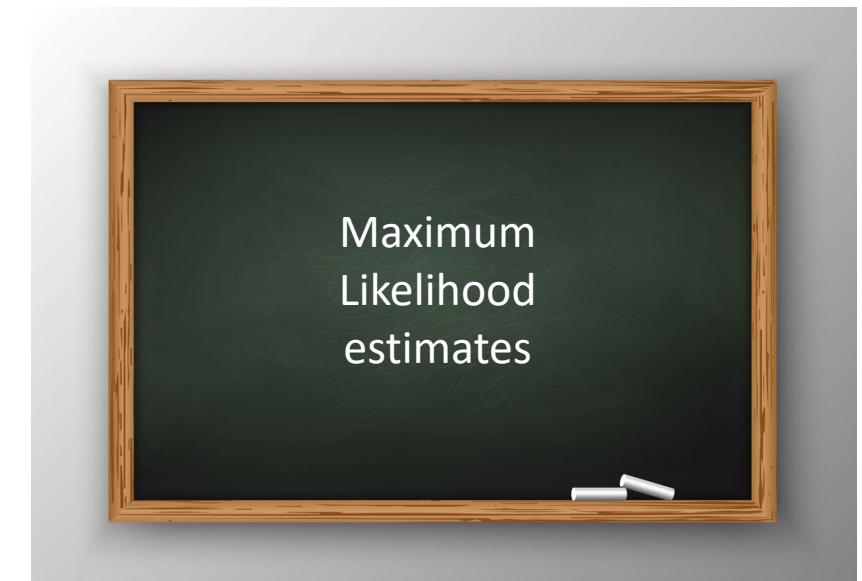
$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y$$

Note: OLS estimate does not require the normality assumption

Valid under the assumption of uncorrelated residual errors

$$\hat{\sigma}_{OLS}^2 = \frac{1}{n-(p+1)} (y - X\hat{\beta}_{OLS})'(y - X\hat{\beta}_{OLS})$$

Note: OLS estimates are unbiased



LM - estimation

Maximum Likelihood (ML) estimation

- The likelihood function for a Normal LM is

$$L_{full}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = (2\pi\sigma^2)^{-n/2} \prod_{j=1}^n \exp\left[-\frac{(y_j - \mathbf{x}'_j \boldsymbol{\beta})^2}{2\sigma^2}\right]$$

$$l_{full}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^n (y_j - \mathbf{x}'_j \boldsymbol{\beta})^2$$

- Maximization of l_{full} provides the ML estimates of unknown parameters

$$\hat{\boldsymbol{\beta}}_{ML} = (X'X)^{-1}X'\mathbf{y} \quad \text{Same as OLS}$$

$$\hat{\sigma}_{ML}^2 = \frac{1}{n} (\mathbf{y} - X\hat{\boldsymbol{\beta}}_{ML})' (\mathbf{y} - X\hat{\boldsymbol{\beta}}_{ML}) = \frac{1}{n} \sum_{j=1}^n (y_j - \mathbf{x}'_j \hat{\boldsymbol{\beta}}_{ML})^2$$

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 β is needed. This is possible considering the likelihood function based on a set of $n-(p+1)$ independent contrast of y .

$$l_{REML}(\sigma^2; \mathbf{y}) = -\frac{n-(p+1)}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^n r_j^2 \Rightarrow \hat{\sigma}_{REML}^2 = \frac{1}{n-(p+1)} \sum_{j=1}^n r_j^2$$

where r_j are the residuals.

Unbiased

- 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 β . 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.

in linear models all these methods some, but when we ignore
some assumptions



The LMM journey



- 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 $V[y_j] = \underline{\sigma_j^2}$

Therefore:

Model equation
at the level of
observations

$$y_j = \beta_0 + \beta_1 x_{1j} + \cdots + \beta_p x_{pj} + \varepsilon_j = x_j^t \boldsymbol{\beta} + \varepsilon_j \quad j = 1, \dots, n$$

$$\varepsilon_j \sim N(0, \sigma_j^2) \quad j \perp j' \text{ independent errors}$$



Note that

$$E[y_j] = \mu_j = x_j^t \boldsymbol{\beta}$$

$$V[y_j] = V[\varepsilon_j] = \sigma_j^2$$

$p+1 + n$
 $\downarrow \downarrow$ different variances

- ❖ 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

$$\lambda(\delta, \mu; v) \quad \leftarrow \text{Variance function}$$

which assumes positive values and is continuous and differentiable with respect to δ .

- Therefore:

$$V[\varepsilon_j] = \sigma^2 \lambda^2(\delta, \mu_j; v_j)$$

where -> v_j is a vector of (known) covariates defining the variance function for observation j,
-> δ contains a small set of variance parameters, common to all observations.

Note that:

- ❖ because the function $\lambda(\cdot)$ 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.

Rappe symmetrische gleiche Varianz für residuals



LM 2.0 – Variance Function

Therefore:

Model equation
at the level of
observations

$$y_j = \beta_0 + \beta_1 x_{1j} + \cdots + \beta_p x_{pj} + \varepsilon_j = x_j^t \boldsymbol{\beta} + \varepsilon_j$$

$$\varepsilon_j \sim N(0, \sigma^2 \lambda_j^2)$$

where $\lambda_j^2 = \lambda^2(\delta, \mu_j; v_j)$ and $j \perp j'$ independent errors

now different
variances

Note that

$$E[y_j] = \mu_j = x_j^t \boldsymbol{\beta}$$

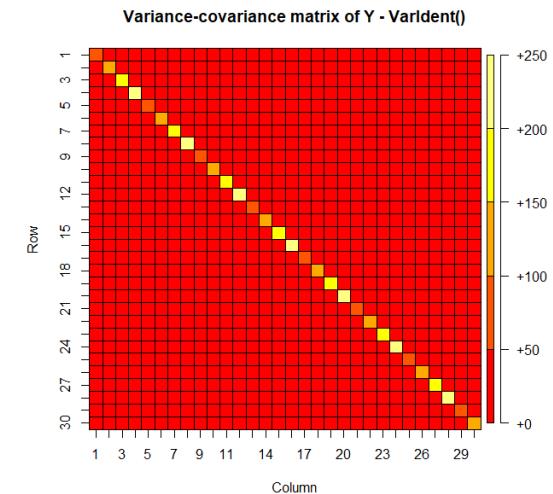
$$V[y_j] = V[\varepsilon_j] = \sigma^2 \lambda_j^2 \quad \longrightarrow$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Model equation
at the level of
observations

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, R) \quad R = \sigma^2 \Lambda \Lambda \quad \text{where } \underline{\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)}$$

$$R \in \mathbb{R}^{n \times n}$$



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 $\mu \Rightarrow \lambda(\cdot) = \lambda(\delta; v)$
3. Variance functions depending on δ and $\mu \Rightarrow \lambda(\cdot) = \lambda(\delta, \mu; v)$
4. Variance functions depending on μ but not on $\delta \Rightarrow \lambda(\cdot) = \lambda(\mu; v)$

- We will symbolically refer to groups 2–4 as $\langle\delta\rangle$ -, $\langle\delta, \mu\rangle$ -, and $\langle\mu\rangle$ - group, respectively.
- Specification of an LM with heterogeneous variance is very general and encompasses all four groups of variance functions \Rightarrow 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 $\langle\delta, \mu\rangle$ or $\langle\mu\rangle$, the parameters β are shared by the mean and variance structures (*mean-variance models*) \Rightarrow require different estimation approaches and inference techniques, as compared to the models involving known weights or variance functions from the $\langle\delta\rangle$ -group.



LM 2.0 – Variance Function

if's up
to us
choose
 δ, μ)
function
for f, r

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		Examples	Estimation algorithm	Section
	δ	μ_i			
Known weights	—	—	varFixed(·)	WLS	7.4.1
$\langle\delta\rangle$	+	—	Table 7.2	ML/REML	7.4.2
$\langle\delta, \mu\rangle$	+	+	Table 7.3	ML/REML-based GLS	7.8.1.1
$\langle\mu\rangle$	—	+	Table 7.4	IRLS	7.8.1.2

Table 7.2 Examples of variance functions from the $\langle\delta\rangle$ -group^a

Function $\lambda(\cdot)$	λ_i	Description
varPower($\delta; v_i, s_i$)	$ v_i ^{\delta_{s_i}}$	Power of a variance covariate v_i
varExp($\delta; v_i, s_i$)	$\exp(v_i \delta_{s_i})$	Exponent of a variance covariate
varConstPower($\delta; v_i, s_i$)	$\delta_{1,s_i} + v_i ^{\delta_{2,s_i}}$	Constant plus power variance function $\delta_{1,s_i} > 0$
varIdent($\delta; s_i$)	δ_{s_i}	Different variances per stratum $\delta_1 \equiv 1, \delta_s > 0$ for $s \neq 1$

so now
not in
parameters
but less,
because
we use function

s_j = stratum the j -th obs belongs to

Table 7.3 Examples of variance functions from the $\langle\delta, \mu\rangle$ -group^a

Function $\lambda(\cdot)$	λ_i	Description
varPower($\delta, \mu_i; s_i$)	$ \mu_i ^{\delta_{s_i}}$	Power of $ \mu_i $
varExp($\delta, \mu_i; s_i$)	$\exp(\mu_i \delta_{s_i})$	Exponent of μ_i
varConstPower($\delta, \mu_i; s_i$)	$\delta_{1,s_i} + \mu_i ^{\delta_{2,s_i}}$	Constant plus power variance function $\delta_{1,s_i} > 0$

Table 7.4 Examples of variance functions from the $\langle\mu\rangle$ -group^a

Function $\lambda(\cdot)$	λ_i	Description
varPower($\mu_i; s_i, \delta$)	$ \mu_i ^{\delta_{s_i}}$	Power of $ \mu_i $, δ_{s_i} known
varExp($\mu_i; s_i, \delta$)	$\exp(\mu_i \delta_{s_i})$	Exponent of μ_i , δ_{s_i} known
varConstPower($\mu_i; s_i, \delta$)	$\delta_{1,s_i} + \mu_i ^{\delta_{2,s_i}}$	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:

$$l_{full}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \sum_{j=1}^n \log(\lambda_j^2) - \frac{1}{2\sigma^2} \sum_{j=1}^n \lambda_j^{-2} (\mathbf{y}_j - \mathbf{x}'_j \boldsymbol{\beta})^2$$

Note that it depends on $\boldsymbol{\delta}$ through λ , then in the case of $\lambda=1$ 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,

$$\text{maximize } l_{full} \text{ wrt } \boldsymbol{\beta} \text{ for any value of } \boldsymbol{\delta} \Rightarrow \hat{\boldsymbol{\beta}}(\boldsymbol{\delta}) \Rightarrow \underline{l}_{ML}^*(\sigma^2, \boldsymbol{\delta})$$

$$\Rightarrow \text{maximize } \underline{l}_{ML}^* \text{ wrt } \sigma^2 \text{ for any value of } \boldsymbol{\delta} \Rightarrow \hat{\sigma}_{ML}^2(\boldsymbol{\delta}) \Rightarrow \underline{l}_{ML}^*(\boldsymbol{\delta})$$



LM 2.0 – Estimation

Profiled Likelihood

- ML estimates of unknown parameters of LM 2.0:

$$\hat{\beta}_{ML} = \hat{\beta}(\hat{\delta}_{ML}) = (\sum_{j=1}^n \hat{\lambda}_j^{-2} \mathbf{x}_j \mathbf{x}'_j)^{-1} \sum_{j=1}^n \hat{\lambda}_j^{-2} \mathbf{x}_j y_j \quad \hat{\sigma}_{ML}^2 = \hat{\sigma}^2(\hat{\delta}_{ML}) = \sum_{j=1}^n \hat{\lambda}_j^{-2} \hat{r}_j^2 / n$$

- Biased => REML approach

$$\hat{\lambda}_j = \lambda(\hat{\delta}_{ML}; v_j) \quad \hat{r}_j = r_j(\hat{\delta}_{ML})$$

$$l_{REML}(\sigma^2, \delta) = -\frac{n - (p + 1)}{2} \log(\sigma^2) - \frac{1}{2} \sum_{j=1}^n \log(\lambda_j^2) - \frac{1}{2\sigma^2} \sum_{j=1}^n \lambda_j^{-2} r_j^2 - \frac{1}{2} \log \left[\det \left(\sum_{j=1}^n \lambda_j^{-2} \mathbf{x}_j \mathbf{x}'_j \right) \right]$$

$$\Rightarrow \hat{\sigma}_{REML}^2(\hat{\delta}) = \sum_{j=1}^n \hat{\lambda}_j^{-2} \hat{r}_j^2 / (n - p - 1)$$

- By maximization of l_{REML} with respect to δ , we obtain an estimator $\hat{\delta}_{REML}$ of δ . This is used to yield an estimator $\hat{\beta}_{REML}$ of β .

I
REML



The LMM journey



- The essential assumption for the LMs considered before was that the **observations** collected during the study were independent of each other. *—unabhängig*
- 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**.

- 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 ($i = 1, \dots, N$)
 - n_i observations per group indexed by j ($j = 1, \dots, n_i$).
- Then, for the group i

Model equation
at the level of group
of observations

$$y_i = X_i \beta + \varepsilon_i$$

date can be grouped

Multiple level of grouping is possible, but requires suitable design of Z matrix and software specifications.

$y_i = (y_{i1}, \dots, y_{in_i})'$ $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})'$

$$X_i = \begin{pmatrix} x_{i1}' \\ \vdots \\ x_{ij}' \\ \vdots \\ x_{in_i}' \end{pmatrix} = \begin{pmatrix} 1 & x_{i11} & \dots & x_{ip1} \\ \vdots & \dots & \dots & \vdots \\ 1 & x_{i1n_i} & \dots & x_{ipn_i} \end{pmatrix} = (\mathbf{1} \quad x_{i1} \quad \dots \quad x_{ip})$$

Note that

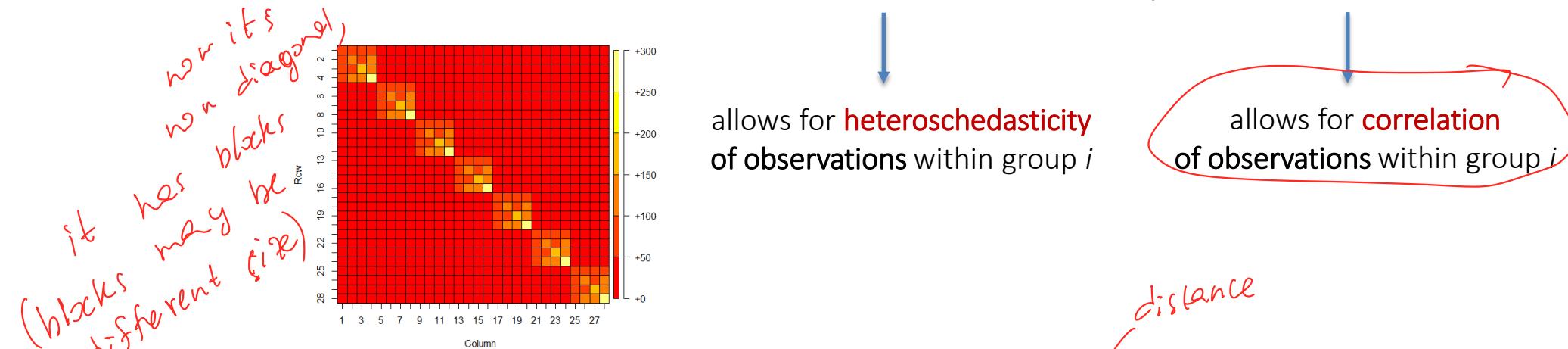
$$E[y_{ij}] = \mu_{ij} = z'_{ij} \beta \quad V[y_i] = V[\varepsilon_i] = \sigma^2 R_i$$



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

$$R_i = \sigma^2 R_i = \sigma^2 \Lambda_i C_i \Lambda_i \quad \text{where} \quad \Lambda_i = (\lambda_{i1}, \dots, \lambda_{in_i}) \quad \text{and} \quad C_i = \text{corr matrix}$$



- Correlation function (general version): $\text{Corr}(\varepsilon_{ij}, \varepsilon'_{ij}) = h[d(t_{ij}, t'_{ij}), \varrho]$ where

- ϱ is a vector of correlation parameters
- $d(\cdot)$ is a distance function of vectors of position variables t_{ij} and t'_{ij} corresponding to, ε_{ij} and ε'_{ij}
- $h[\cdot]$ is a continuous function with respect to ϱ , ranging between -1 and 1 and s.t. $h(0, \varrho) \equiv 1$.



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 => $\text{Corr}(\varepsilon_{ij}, \varepsilon'_{ij}) = \rho$

2. “Spatial” structures

=> correlation structures which are defined in the context of spatial data.

Ex: $\text{Corr}(\varepsilon_{ij}, \varepsilon'_{ij}) = e^{-s_{ij,ij'}/\varrho} \Rightarrow h(s, \varrho) = e^{-s/\varrho}$

$1 - h(s, \varrho)$ is the semi-variogram



Table 10.1 Examples of serial and spatial correlation structures

Correlation structure	Function $h(., .)$	Comment
Serial	(Auto)correlation function	
<i>corCompSymm</i> ^a	$h(k, \varrho) \equiv \varrho$	$k = 1, 2, \dots; \varrho < 1$
<i>corAR1</i>	$h(k, \varrho) \equiv \varrho^k$	$k = 0, 1, \dots; \varrho < 1$
<i>corCAR1</i>	$h(s, \varrho) \equiv \varrho^s$	$s \geq 0; \varrho \geq 0$
<i>corSymm</i>	$h(d(j, j'), \varrho) \equiv \varrho_{j,j'}$	$j < j'; \varrho_{jj'} < 1$
Spatial	Correlation function	
<i>corExp</i>	$h(s, \varrho) \equiv e^{-s/\varrho}$	$s \geq 0; \varrho > 0$
<i>corGaus</i>	$h(s, \varrho) \equiv e^{-(s/\varrho)^2}$	$s \geq 0; \varrho > 0$
<i>corLin</i>	$h(s, \varrho) \equiv (1 - s/\varrho)I(s < \varrho)$	$s \geq 0; \varrho > 0$
<i>corRatio</i>	$h(s, \varrho) \equiv 1 - (s/\varrho)^2 / \{1 + (s/\varrho)^2\}$	$s \geq 0; \varrho > 0$
<i>corSpher</i>	$h(s, \varrho) \equiv [1 - 1.5(s/\varrho) + 0.5(s/\varrho)^3]I(s < \varrho)$	$s \geq 0; \varrho > 0$

^aThe 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.

Still remember
that total number
of persons have
be less now we will have
the number of n
persons of nature, persons
of collinear, persons



Some considerations on LMs 1.0-2.0-3.0

- Relaxing independence and homoscedasticity hypotheses
 - ⇒ More flexible and realistic models
 - ⇒ Increase the number of parameters (identifiability issues => constraints and parametrizations)
- Shaping the variance-covariance matrix of the errors implies that knowledge about the dependency among observations is available
- Inference?

Group structure, want to quantify this effect. We want to say something about new observation.

(like in ANOVA, we assumed some variance and looked only on $f(x)$, now we also look at residuals)

$y = f(x) + \epsilon$ now grouping also effect on ϵ
(variance now depend on smth, as example on μ)



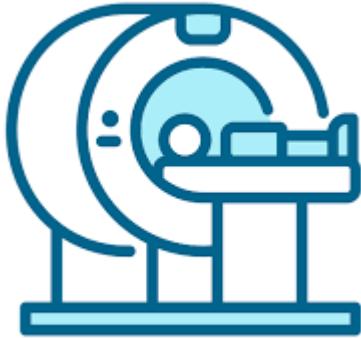
The LMM journey



- 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.
- More in general, LMMs represent the most popular and effective approach for the analysis of dependent 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.
peachy T&Tb
- 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.



Why LMMs

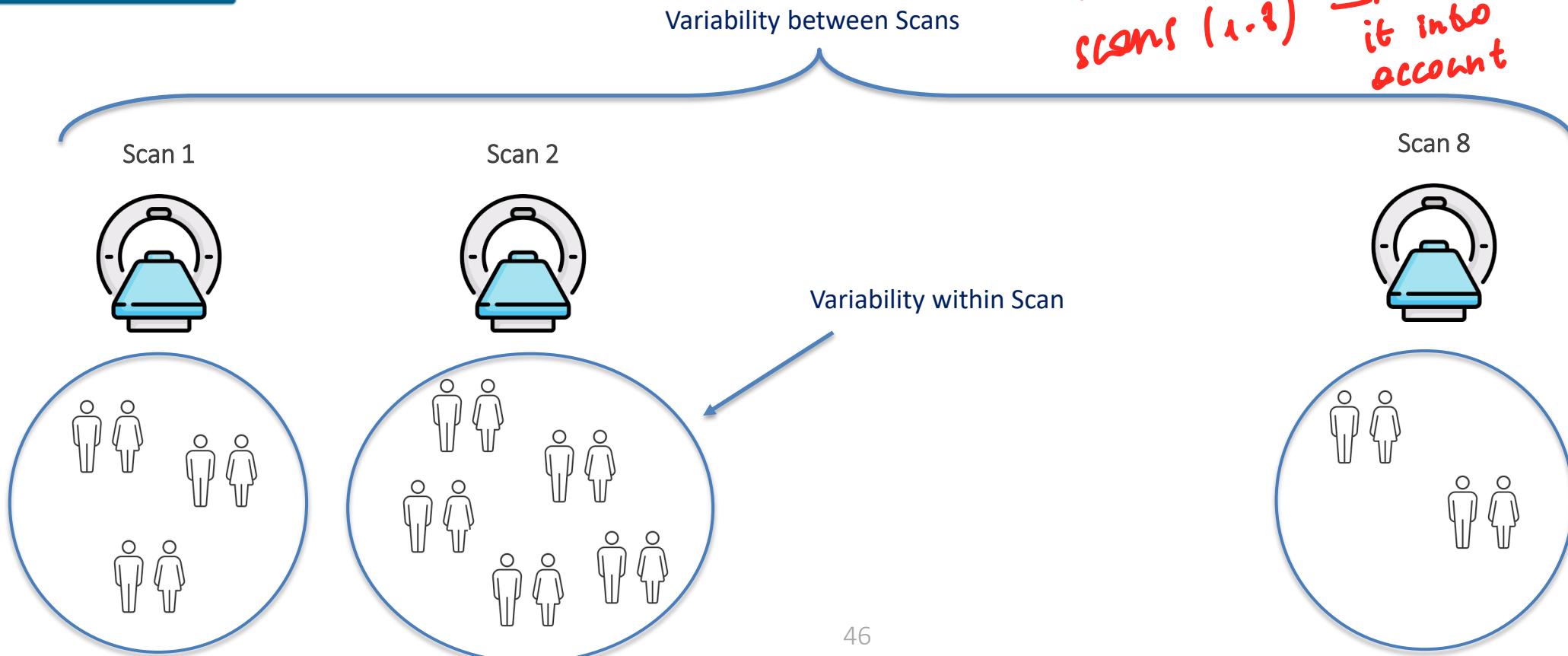


n individuals receive MRI scan in the same hospital

8 different scans (i) are available, people randomly assigned to them

y_{ij} = endpoint measured at individual (j) level

Variability between
scans (i.i) — how to take
it into account



Why LMMs

How to include the information about the Scan effect into the model?

Option 1. => Categorical variable

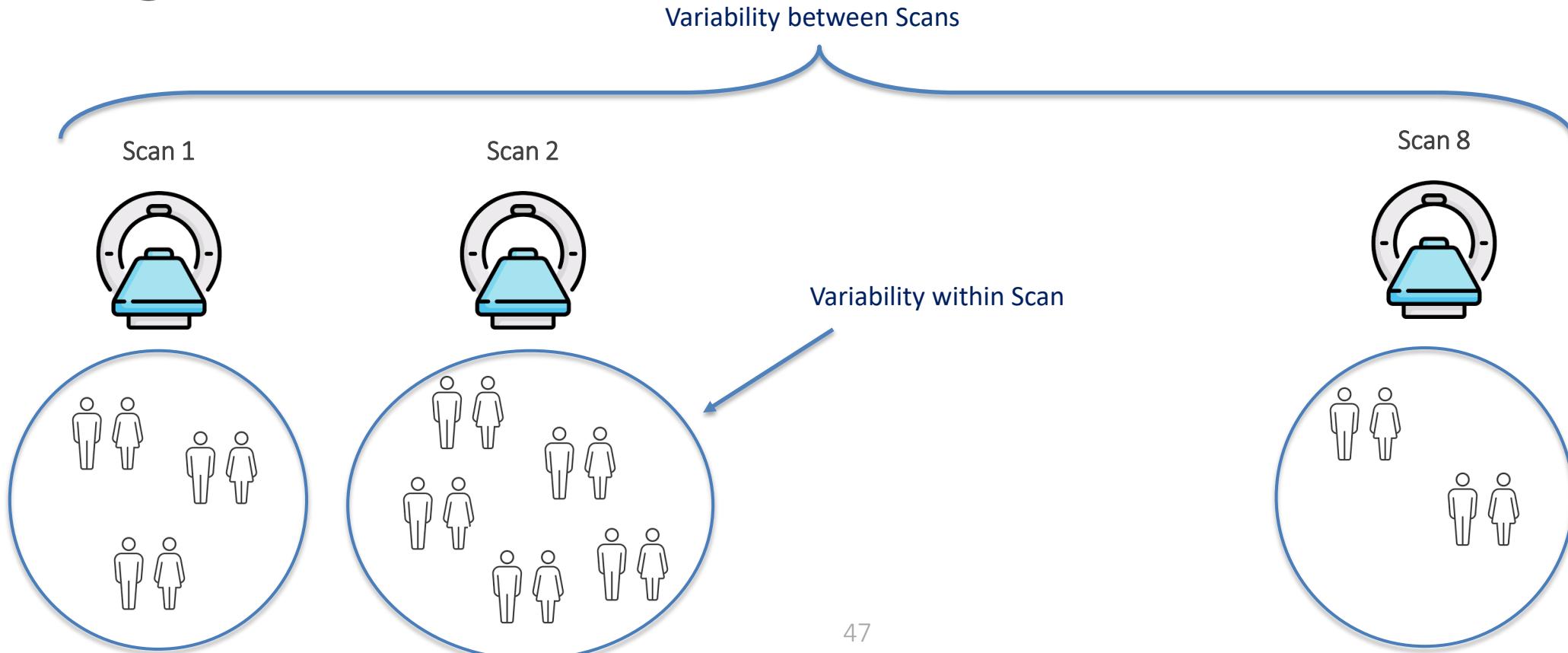


model complexity: if many levels, parameters to be estimated are many



inference: no interest only in the scans which are part of the observed sample

like in ANOVA, with different treatments
 $ANCOVA = ANOVA + Covariance matrix$
but we can not say anything
about new treatment, so
here same problem — we
want say something about
new scan

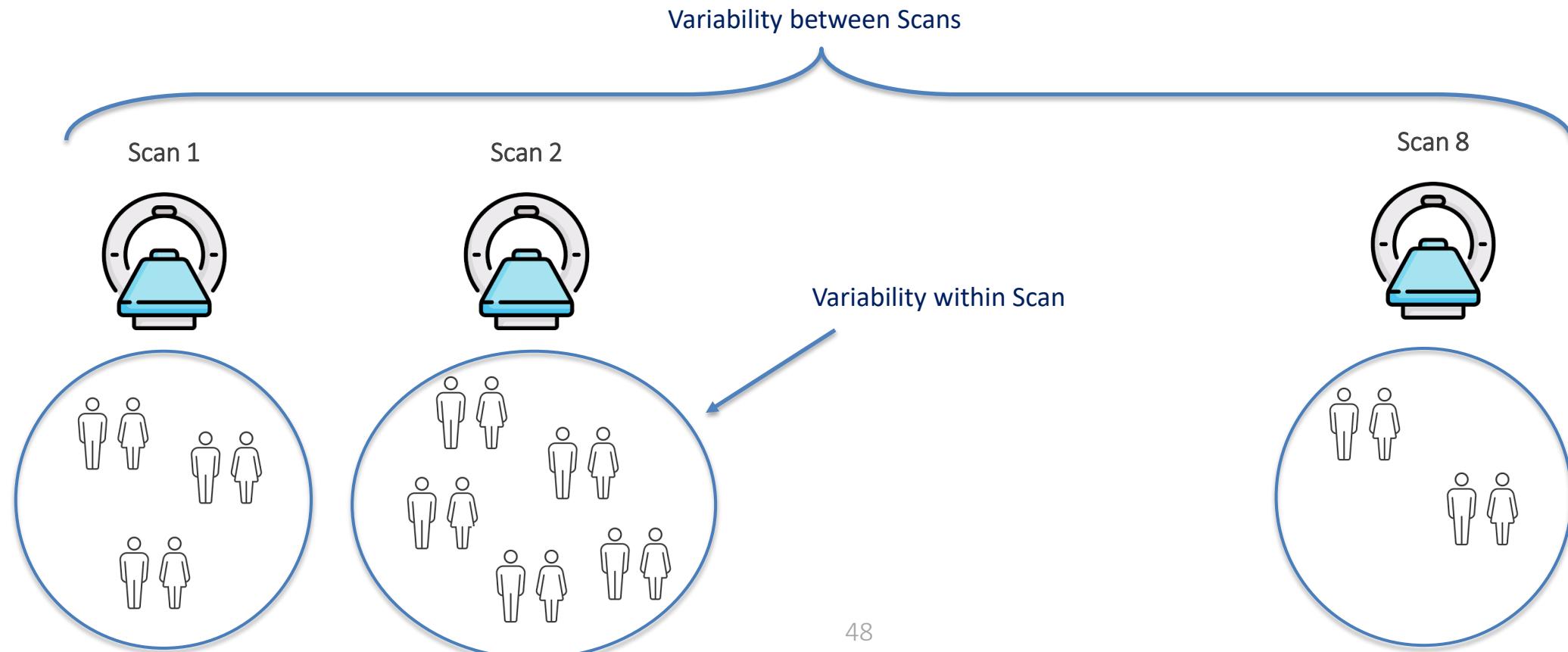


Why LMMs

How to include the information about the Scan effect into the model?

Option 1. => Categorical variable

Option 2. => Random Effect — *what it exactly means*



LMM – Pooled, separate or mixed effects estimates?

Back to independent observations averaging over the groups

PB: low number of observations and loosing information

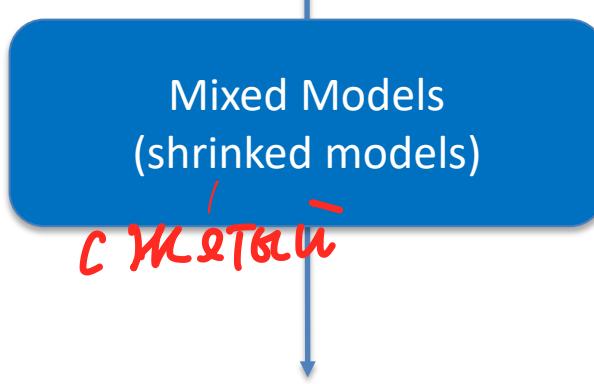


Separate model for each group

PB: how to compare/match the estimates

statistics is about variability
may be $\beta_0 = 22$ $\beta_1 = 4,5$ some
in terms of our variability
in group 1 in group 2

like we have one common scans



Separate Models

n-scans =>
n models
but what if number
of patients in each
group are different -
it may not be enough
to fit each model



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

$$\mathbf{y}_i = X_i \boldsymbol{\beta} + Z_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i \quad (\text{may be } x_i = ?)$$

$\mathbf{b}_i \sim N_q(\mathbf{0}, D)$ $\boldsymbol{\varepsilon}_i \sim N_{n_i}(\mathbf{0}, R_i)$ with $\mathbf{b}_i \perp \boldsymbol{\varepsilon}_i \quad \forall i \neq i'$

$D = \sigma_b^2 D$ and $R_i = \sigma_\varepsilon^2 R_i$ R_i and D positive-definite

$$Z_i = \begin{pmatrix} x'_{i1} \\ \vdots \\ x'_{ij} \\ \vdots \\ x'_{in_i} \end{pmatrix} = \begin{pmatrix} 1 & x_{i11} & \dots & x_{ip1} \\ \dots & \dots & \dots & \dots \\ 1 & x_{i1n_i} & \dots & x_{ipn_i} \end{pmatrix} = (\mathbf{1} \quad \mathbf{x}_{i1} \quad \dots \quad \mathbf{x}_{ip}) \quad X_i \in \mathbb{R}^{n_i \times (p+1)}$$

$\mathbf{b}_i = (b_{i0}, b_{i1}, \dots, b_{iq})'$ $Z_i = (\mathbf{1} \quad \mathbf{z}_{i1} \quad \dots \quad \mathbf{z}_{iq}) \quad Z_i \in \mathbb{R}^{n_i \times (q+1)}$

- 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 $\mathbf{y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_N)'$ be the vector of all $n = \sum_{i=1}^N n_i$ observed values of the dependent variable in the N groups.
Let $\mathbf{b} = (\mathbf{b}'_1, \dots, \mathbf{b}'_N)'$ and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}'_1, \dots, \boldsymbol{\varepsilon}'_N)'$ be the vectors containing all the Nq random effects and n errors, respectively.

Model equation
for all data

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}$$

$$\mathbf{b} \sim N_{N(q+1)}(\mathbf{0}, \sigma_b^2 \mathbf{D}) \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{R}) \quad \text{with} \quad \mathbf{b}_i \perp \boldsymbol{\varepsilon}_i \quad \mathbf{b}_i \perp \mathbf{b}_{i'}$$

where

$$\mathbf{D} = \mathbf{I}_N \otimes \mathbf{D} = \begin{pmatrix} \mathbf{D} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{D} \end{pmatrix} \quad \text{and} \quad \mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{R}_N \end{pmatrix}$$

LMM with *nested structure* for random effects. It is possible to formulate LMMs with non-block-diagonal matrices Z , D , and R (crossed random effects).

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_i \\ \vdots \\ \mathbf{X}_N \end{pmatrix} \in \mathbb{R}^{n \times (p+1)} \quad \mathbf{Z}_i = \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_N \end{pmatrix} \in \mathbb{R}^{n \times N(q+1)}$$



LMM – conditional and unconditional distributions

UNCONDITIONAL DISTRIBUTION of the RANDOM EFFECTS

Symmetrische Parameter

- 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 \mathbf{b} .

- Note that the matrix D is parameterized using a vector of parameters $\boldsymbol{\theta}_D$. *In simplest case, D is diagonal*
 - **General case:** any two elements of the vector \mathbf{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 \mathbf{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 $f_{y|b}(y_i|\mathbf{b}_i)$ is multivariate normal, with

$$E[y_i|\mathbf{b}_i] = \boldsymbol{\mu}_i = X_i \boldsymbol{\beta} + Z_i \mathbf{b}_i \quad \text{with } \boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{in_i})$$

$$V[y_i|\mathbf{b}_i] = V[\boldsymbol{\varepsilon}_i|\mathbf{b}_i] = \sigma^2 R_i$$

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.

$$V[\varepsilon_{ij}|\mathbf{b}_i] = \sigma^2 R_{ij} = \sigma^2 \lambda^2(\mu_{ij}, \boldsymbol{\delta}; \boldsymbol{v}_{ij})$$



LMM – conditional and unconditional distributions

MARGINAL DISTRIBUTION of the OBSERVATIONS

- The marginal distribution $f_y(\mathbf{y}_i)$ of the observations is obtained by “integrating out” the random effects.

$$f_y(\mathbf{y}_i) = \int f_{y,b}(\mathbf{y}_i, \mathbf{b}_i) d\mathbf{b}_i = \int f_{y|b}(\mathbf{y}_i | \mathbf{b}_i) f_b(\mathbf{b}_i) d\mathbf{b}_i$$

- Given that $f_{y,b}(\mathbf{y}_i, \mathbf{b}_i)$ and $f_b(\mathbf{b}_i)$ are densities of multivariate normal distributions, the marginal distribution of \mathbf{y} is also multivariate normal and it can be derived analytically.

$$E[\mathbf{y}_i] = X_i \boldsymbol{\beta}$$

$$V[\mathbf{y}_i] = \sigma^2 [Z_i D Z_i' + R_i] = \sigma^2 [Z_i D(\boldsymbol{\theta}_D) Z_i' + R_i(\boldsymbol{\theta}_R, v_i)]$$

- 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.

residual variability

if $\mathbf{R}_i = \mathbf{I}$ \Rightarrow var of y_i will not

be identical

because $D_i \neq I$

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
or parameters
- Model fitting has 3 distinctive components
 - Estimates of **fixed effects** – *random sampling from normal distribution*
 - Estimates of **random effects**
 - Estimates of **variance parameters**
- **WARNING:** variance components are nonnegative 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

$$y_{ij} = \beta_0 + \beta_1 x_{j1} + \dots + \beta_p x_{jp} + b_j + \varepsilon_{ij} \quad - \text{simplest case}$$

- Profiled confidence intervals for the **fixed effects** and for the standard deviations of random effects and residuals are provided.

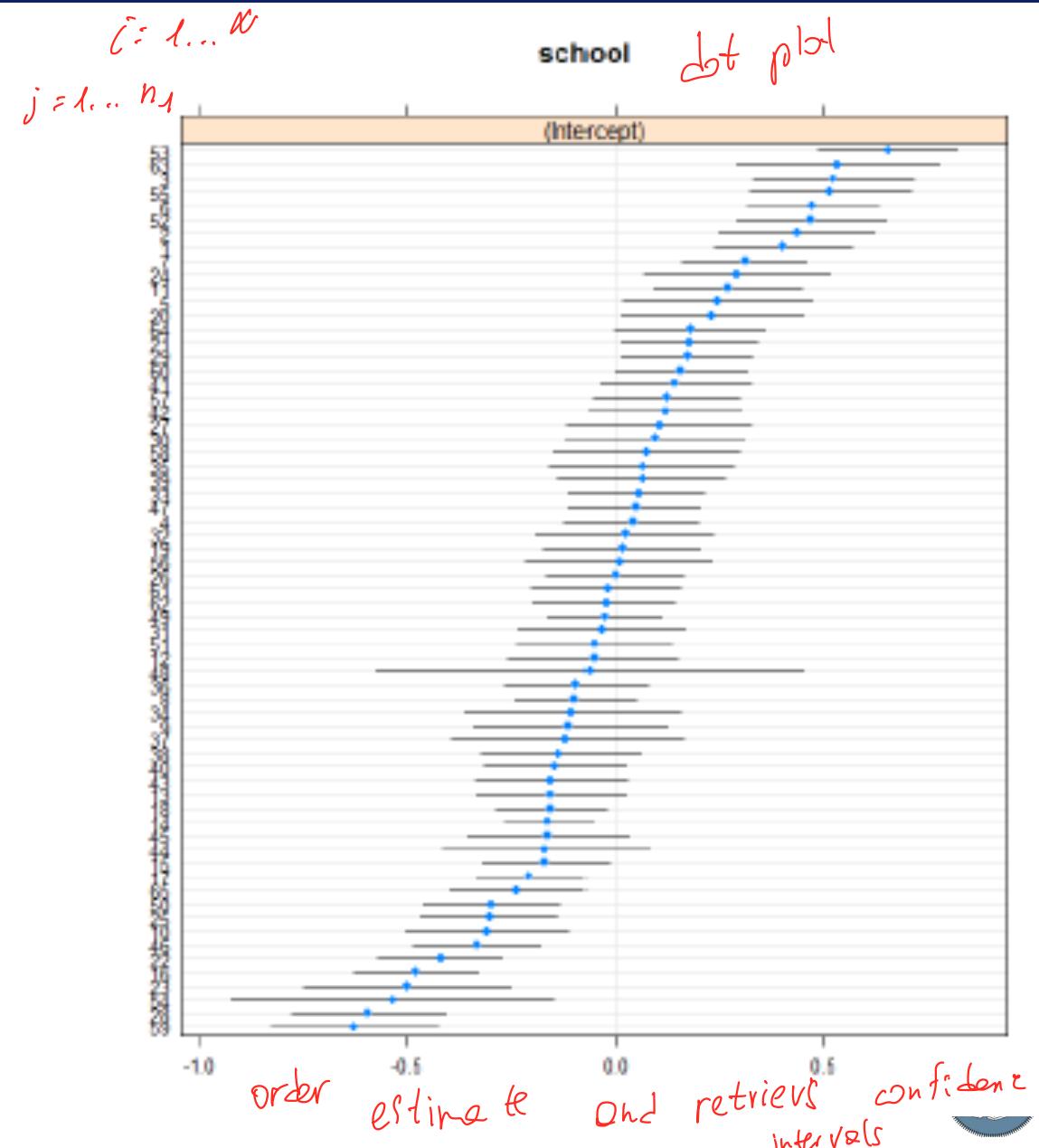
$$\hat{\beta} \quad b \sim \mathcal{N}(0, \Sigma_b)$$

- 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)

$$PVRE = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_{\varepsilon}^2}$$

will not be 20%
5-10%
- of variability explained just by random effect
=> there effect is real

Random intercept case

- Use the estimate of $\hat{\sigma}_b^2$ for performing **scenario analysis**
=> what if belonging to a $-2\hat{\sigma}_b$ or a $+2\hat{\sigma}_b$ 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
=> $(X \neq Z)$

why it help work with new groups?



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** participated 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 whether 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 recorded 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	C	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

=>

$$X = \begin{pmatrix} 1 & 178 & 1 & 0 \\ 1 & 168 & 1 & 0 \\ 1 & 196 & 0 & 0 \\ 1 & 170 & 0 & 1 \\ 1 & 165 & 1 & 0 \\ 1 & 190 & 0 & 1 \\ 1 & 175 & 1 & 0 \\ 1 & 180 & 1 & 0 \\ 1 & 175 & 0 & 1 \end{pmatrix},$$

No One-hot encoding will be linear dependent
 or we can ignore intercept



N = 3 hospitals

m = 9 observations
(group 1=4, group 2=2, group 3=3)

$X = \vec{z}$, $\vec{z} = w$
new rotation

p+1 = 4 covariates
q+1 = 1 RANDOM INTERCEPT ONLY

> Group Level Formulation

$$\underline{y}_i = \underline{Z}_i \vec{\beta} + W_i b_{io} + \varepsilon_i$$

$$\varepsilon_i \sim N_{m_i}(0, \sigma^2 R_i)$$

$$b_{io} \sim N_4(0, \sigma_b^2)$$

$$\vec{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$$

$$\underline{Z}_i \in \mathbb{R}^{m_i \times (p+1)} = m_i \times 4 = \begin{cases} i=1 & 4 \times 4 \\ i=2 & 2 \times 4 \\ i=3 & 3 \times 4 \end{cases}$$

$$V[\underline{y}_i] = W_i \sigma_b^2 W_i^\top + \sigma^2 R_i$$

$$i=1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \sigma_b^2 [1 1 1 1] + \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$W_i \in \mathbb{R}^{m_i \times (q+1)} = m_i \times 1$$

$$= \begin{bmatrix} \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \end{bmatrix} + \begin{bmatrix} \sigma_e^2 & & & \\ & \sigma_e^2 & & \\ & & \sigma_e^2 & \\ & & & \sigma_e^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 \end{bmatrix} \in \mathbb{R}^{m \times m_i} \quad (\star)$$

for only group
i

All Data Formulation

$$\underline{y} = Z \underline{\beta}^T + W \underline{b}^T + \underline{\varepsilon} \rightarrow \underline{\varepsilon} \sim N_9(0, \sigma^2 R) \quad R \in \mathbb{R}^{m \times n = 9 \times 9}$$

$$L \sim N(\mu, D) \quad D \in \mathbb{R}^{N \times N = 3 \times 3}$$

> All Data Formulation

$$y = Z\beta^T + Wb^T + \varepsilon \rightarrow \varepsilon \sim N_g(0, \sigma^2 R) \quad R \in \mathbb{R}^{m \times n} = 9 \times 9$$

$$b \sim N_N(0, D) \quad D \in \mathbb{R}^{N \times N} = 3 \times 3$$

$$\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$$

$$Z \in \mathbb{R}^{m \times (p+1)} = 9 \times 4$$

$$W \in \mathbb{R}^{m \times N(q+1)} = 9 \times 3(4)$$

$$D = \begin{bmatrix} \sigma_b^2 & & \\ & \sigma_b^2 & 0 \\ 0 & & \sigma_b^2 \end{bmatrix}$$

$$Z = \begin{bmatrix} 1 & Z_{141} & Z_{241} & Z_{341} \\ 1 & & & \\ \vdots & & & \end{bmatrix}$$


$\Rightarrow E[y] = Z\beta^T$

$$V[y] = WDW^T + \sigma_e^2 R$$

depends on m
 each block has
 different size

This structure arise from the assumption on fixed and random effects carried out in the LMM formulation

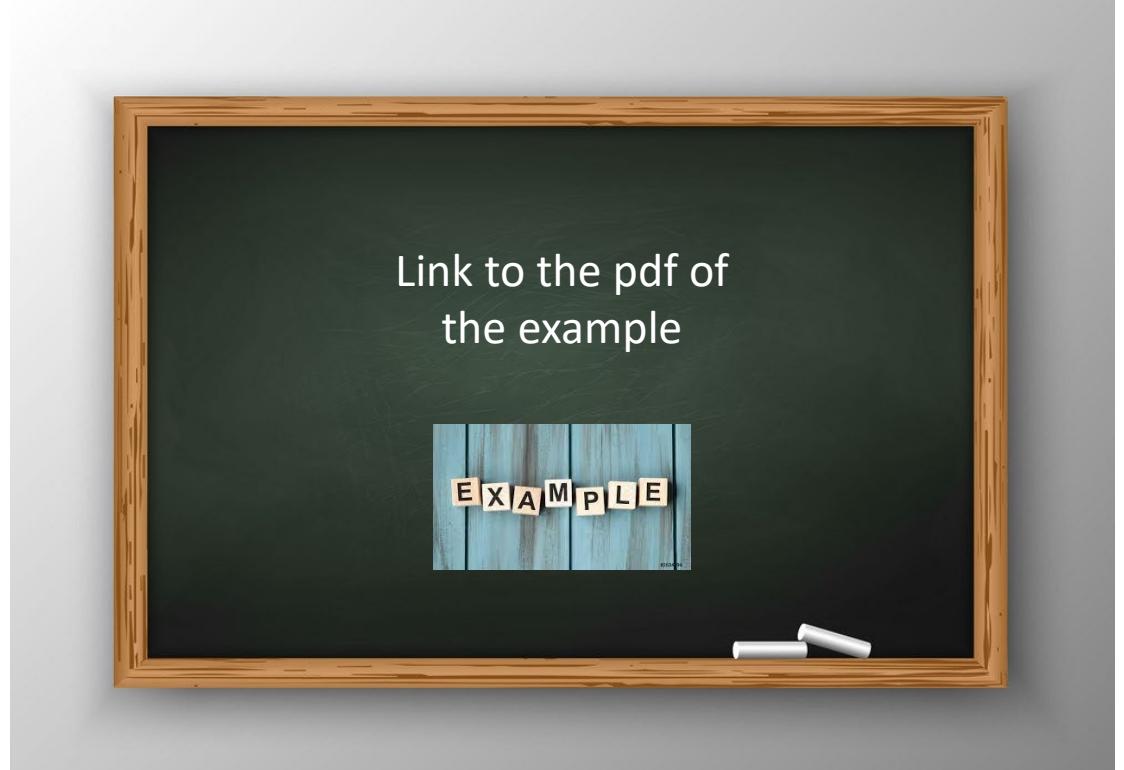
We might design something similar also working on variance and correlation function of LM 3.0, but then we loose the opportunity of making inference on disentangle

observations between groups are uncorrelated (\rightarrow independent in M setting)

LMMs – Example Multicenter Randomized Trial



- EX: Random intercept only – homoscedastic residuals



Take home messages

Beyond LMMs ... the beginning of another journey

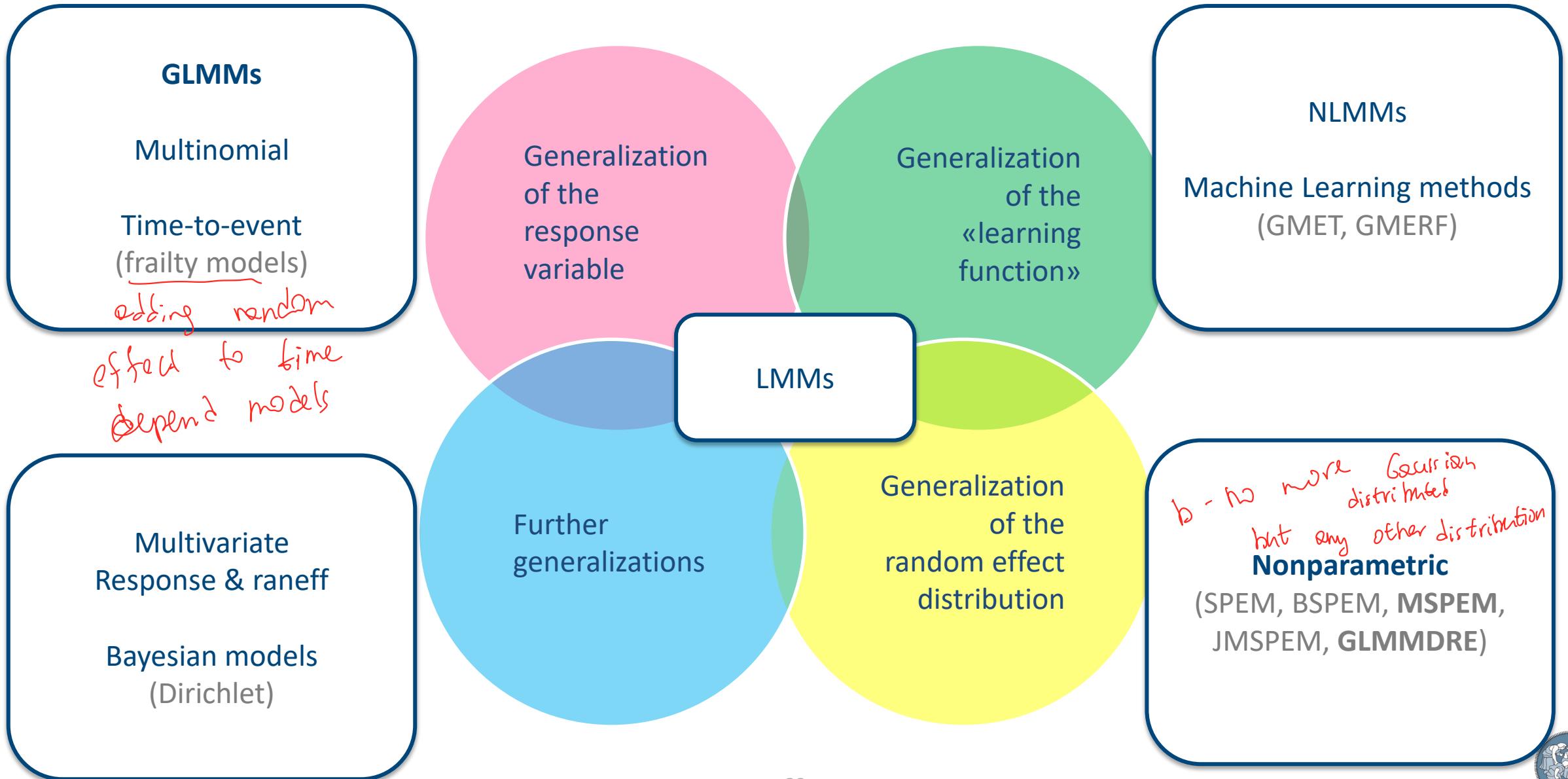


Take home messages

- 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.
A handwritten note in red ink is written over the underlined text. It starts with a red curly brace under "do not allow making inference about the variability that may be related to different levels of the data hierarchy." A red arrow points from this brace to the handwritten note. The handwritten note reads "so we can not ask for new groups".
- 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



GLMM – Generalized Linear Mixed-effect Models

A **Generalized Linear Mixed-effect Model** (GLMM) is an extension of a generalized linear model that includes both fixed and random effects in the linear predictor.

Let us consider observations $j = 1, \dots, n_i$ nested within a group i ($i = 1, \dots, I$). GLMM expresses the n_i -dimensional **response vector** $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})$ for observations in the i -th group as:

$$g(\mathbb{E}[Y_i | \mathbf{b}_i]) = \boldsymbol{\eta}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i$$

where

- i is the group index with I groups ($i = 1, \dots, I$)
- $g(*)$ is a monotonic link function
- $\boldsymbol{\eta}_i$ is the n_i -dimensional linear predictor vector
- $\boldsymbol{\beta}$ is the $(p + 1)$ -dimensional vector of **fixed effects**
- \mathbf{X}_i is the $n_i \times p$ fixed-effects regressor matrix
- $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is the $(q + 1)$ -dimensional vector of **random effects**
- \mathbf{Z}_i is the $n_i \times (q + 1)$ random-effects regressor matrix

- GLMMs handle a **wide range of response distributions** and a wide range of scenarios where observations are grouped rather than completely independent.
- Fixed effects are identified by parameters associated to the entire population, while random ones are identified by group-specific parameters.



Logistic Mixed Effect Models

Mixed effects logistic regression is used to model **binary outcome** variables $Y_{ij} \sim Be(p_{ij})$, in which the log odds of the outcomes are modelled as a linear combination of the predictor variables when data are clustered or there are both **fixed** and **random effects**:

$$\text{logit}(\text{E}[Y_i|\boldsymbol{b}_i]) = \text{logit}(\boldsymbol{p}_i) = \boldsymbol{\eta}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i$$

where

- i is the group index with I groups ($i = 1, \dots, I$)
- $\boldsymbol{y}_i = (y_{i1}, \dots, y_{in_i})$ n_i -dimensional response vector for observations ($j = 1, \dots, n_i$) in group i
- logit link function $g(x) = \text{logit}(x) = \log\left(\frac{x}{1-x}\right)$
- $\boldsymbol{\eta}_i$ is the n_i -dimensional linear predictor vector
- $\boldsymbol{\beta}$ is the $(p + 1)$ -dimensional vector of **fixed effects**
- \boldsymbol{X}_i is the $n_i \times (p + 1)$ fixed-effects regressor matrix
- $\boldsymbol{b}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is the $(q + 1)$ -dimensional vector of **random effects**
- \boldsymbol{Z}_i is the $n_i \times (q + 1)$ random-effects regressor matrix

The random effects \boldsymbol{b}_i are assumed to be **independent** for different groups.

The random effects \boldsymbol{b}_i are defined to have a mean of 0 and therefore **any nonzero mean** for a term in the random effects must be expressed as part of the **fixed-effects** terms.

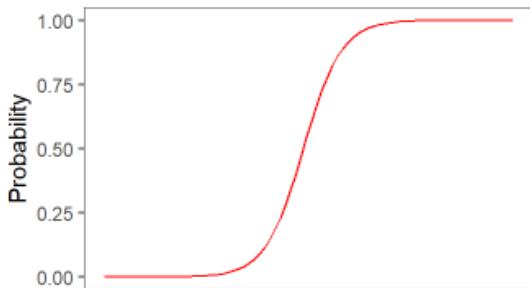
Thus, the columns of \boldsymbol{Z}_i are usually a **subset** of the columns of \boldsymbol{X}_i .



Logistic MEM – Random Intercept and Slope (Single Covariate)

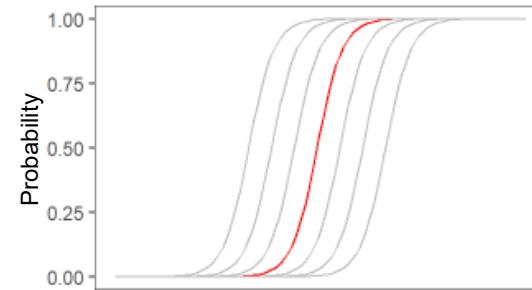
Suppose to have binary response variables $Y_{ij} \sim Be(p_{ij})$, where j is the observation index ($j = 1, \dots, n_i$) within group $i = 1, \dots, I$. Let X_{ij} be the (single) independent covariate.

Logistic regression – Fixed effects model



$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij}$$

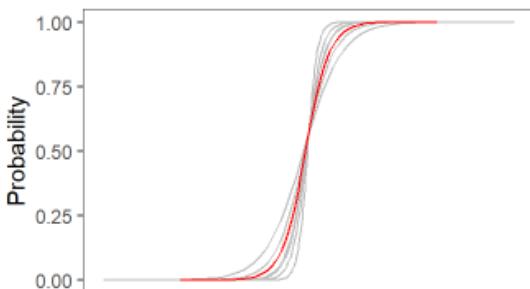
Logistic MEM with random intercept



$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij} + b_{0i}$$

$$b_{0i} \sim \mathcal{N}(0, \sigma_b^2)$$

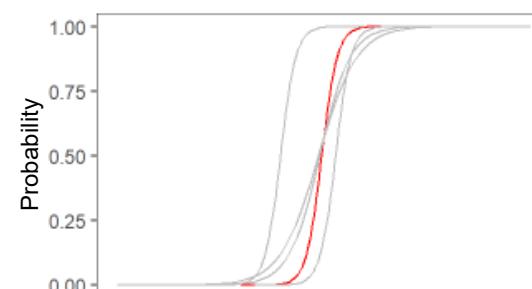
Logistic MEM with random slope



$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij} + b_{1i} x_{ij}$$

$$b_{1i} \sim \mathcal{N}(0, \sigma_b^2)$$

Logistic MEM with random intercept & slope



$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij} + b_{0i} + b_{1i} x_{ij}$$

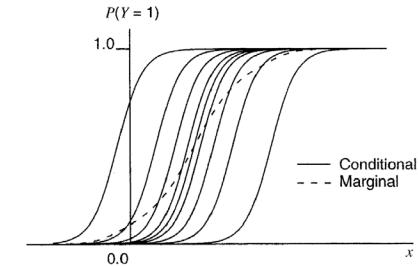
$$b_i = (b_{0i}, b_{1i}) \sim \mathcal{N}(\mathbf{0}, \Sigma)$$



Variance Partition Coefficient (VPC)

Let us consider a logistic mixed-effects model with p predictors x_{kij} ($k = 1, \dots, p$) as fixed effects and only the intercept b_{0i} as random effect:

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{1ij} + \dots + \beta_p x_{pij} + b_{0i} \quad b_{0i} \sim \mathcal{N}(0, \sigma_b^2)$$



Goldstein *et al.* (2002) introduced the **Variance Partition Coefficient** (VPC) to measure the magnitude of the random effects (i.e., VPC is measure of intraclass correlation) in case of **group-specific intercept as random-effects structure**, as follows:

$$VPC = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_{lat}^2} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \pi^2/3}$$

where

- $\hat{\sigma}_b^2$ is the estimated variance of the random intercept
- $\hat{\sigma}_{lat}^2$ is the residual variability that can neither be explained by fixed effects, nor through the group features that are represented by the random intercept. In this case, it is equal to the variance of the standard logistic distribution $\hat{\sigma}_{lat}^2 = \pi^2/3$.



VPC represents the **percentage of unexplained variability** in the response
that is given to the grouping level.



Case studies

- Multicenter observational study on Lung Cancer
- Assessment of pupils learning (INVALSI data)
- RL hospital effect on HF patients survival



Case studies

- Multicenter observational study on Lung Cancer

Rea, F., Ieva, F., Pastorino, U., Apolone, G., Barni, S., Merlini, L., Franchi, M., Corrao, G. (2020)
Surgery-volume and mortality of patients after lung cancer surgery: evidence from an Italian real-world investigation.
[European Journal of Cardio-Thoracic Surgery](#), 58 (1): 70-77



Case study I – Scenario analysis on hospital effect in a multi center trial

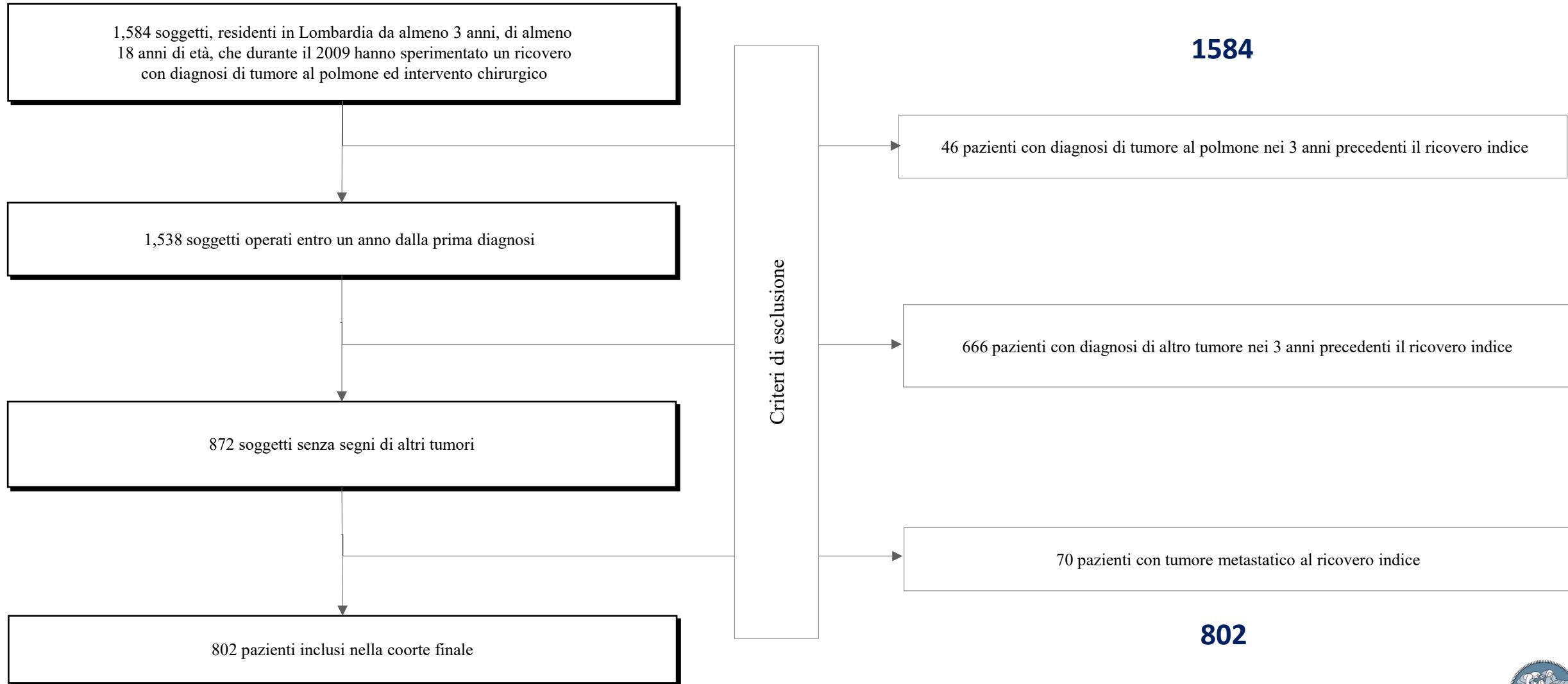


- Observational multicenter study (N=48 hospitals) on **lung cancer**.
- Investigation of **clinical factors** associated to 3y death for any cause in pts undergoing surgery for lung cancer.
- Goals: *population*
 - Risk stratification of the population under study
(association between mortality and features measured at baseline/entrance)
 - Assessment of the grouper effect
 - Scenario analysis

*if hospital seeing
less patients
does surgery with more
practice better*



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 – Scenario analysis on hospital effect



	Volume (# of cases per year)			
	<10 (87 pz)	10-30 (206 pz)	31-50 (289 pz)	>50 (220 pz)
Age [years]				
median (min-max)	69 (34-84)	70 (34-84)	68 (22-83)	68 (41-85)
Gender				
Women (%)	27 (31%)	50 (24%)	86 (30%)	65 (30%)
MCS				
Polymorbid pts (%)	20 (23%)	64 (31%)	62 (21%)	42 (19%)
Main surgery				
Low complexity (3229, 323, others) (Demolizione locale di lesione o tessuto del polmone, Resezione segmentale del polmone, altri)	26 (30%)	66 (32%)	59 (20%)	56 (25%)
Intermediate complexity (324) Lobectomia del polmone	61 (70%)	126 (61%)	205 (71%)	152 (69%)
High complexity (325, 326) Pneumonectomia completa o dissezione radicale delle strutture toraciche	0 (0%)	14 (7%)	25 (9%)	12 (5%)
Adjuvant Chemptherapy	1 (1%)	13 (6%)	15 (5%)	17 (8%)



Case study I – GLMM with single level of grouping

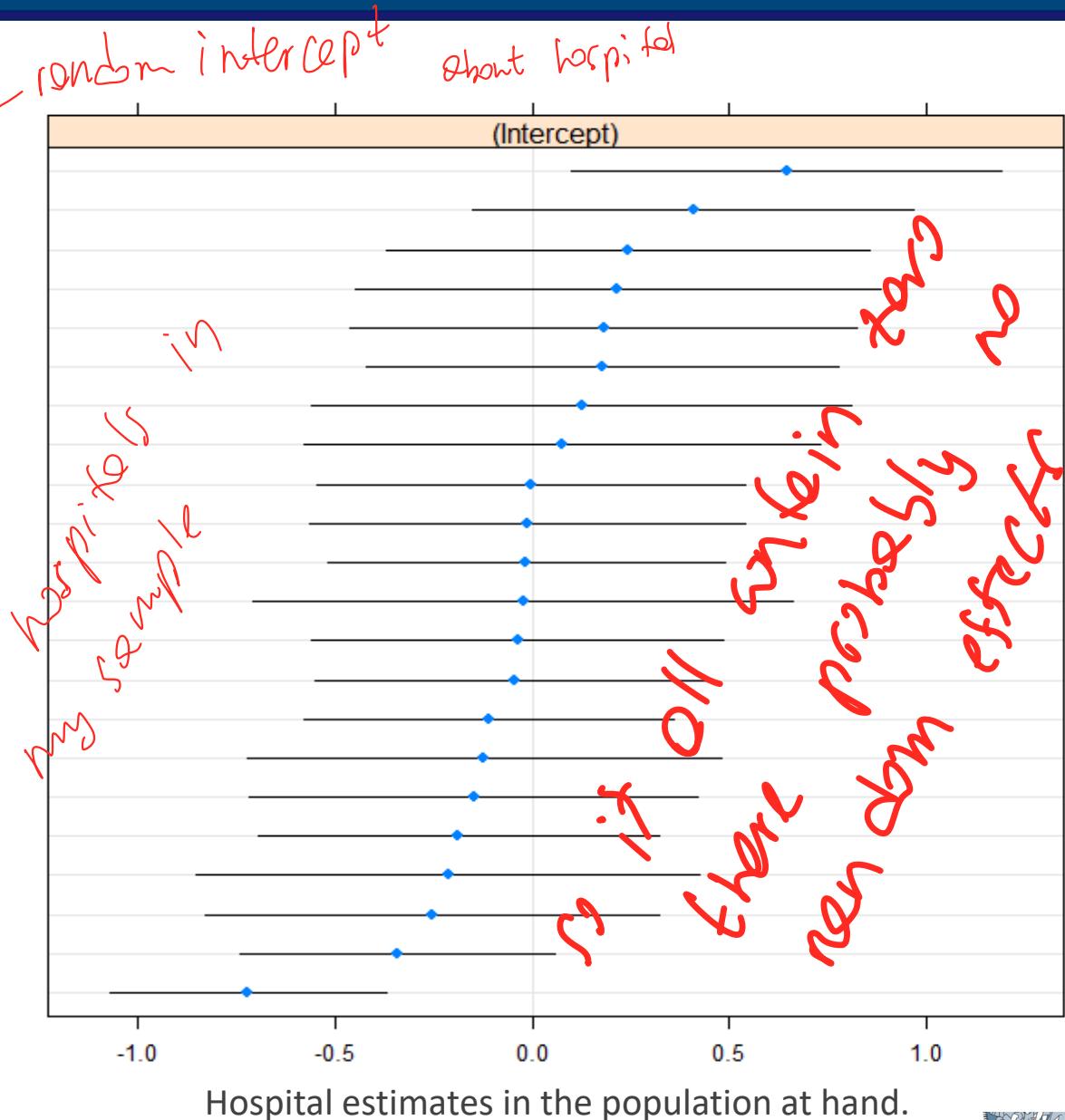


$$\text{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_0 + \beta_1 z_{1ij} + \dots + \beta_p z_{pij} + b_i$$

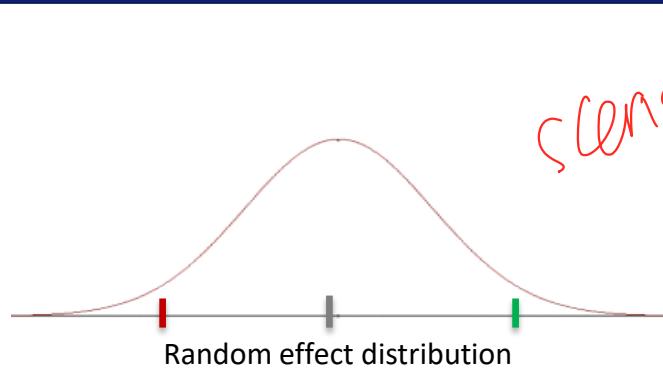
- b_i random effect $\sim N(0, \sigma_b^2)$
- b_i takes the same value for each observation within the same group and different values in different groups.
- It can be interpreted as *the effect of being hospitalized in a given structure*.
- Given the patient-specific features, hospital have a significant effect on the log odds of survival probability

% of total variability accounted for by the grouper

VPC = 4.6%



Case study I – GLMM with single level of grouping



Scenarios analysis



$-2\hat{\sigma}$

we if go worse hospital



0 (mean)

prob of being alive) without effect



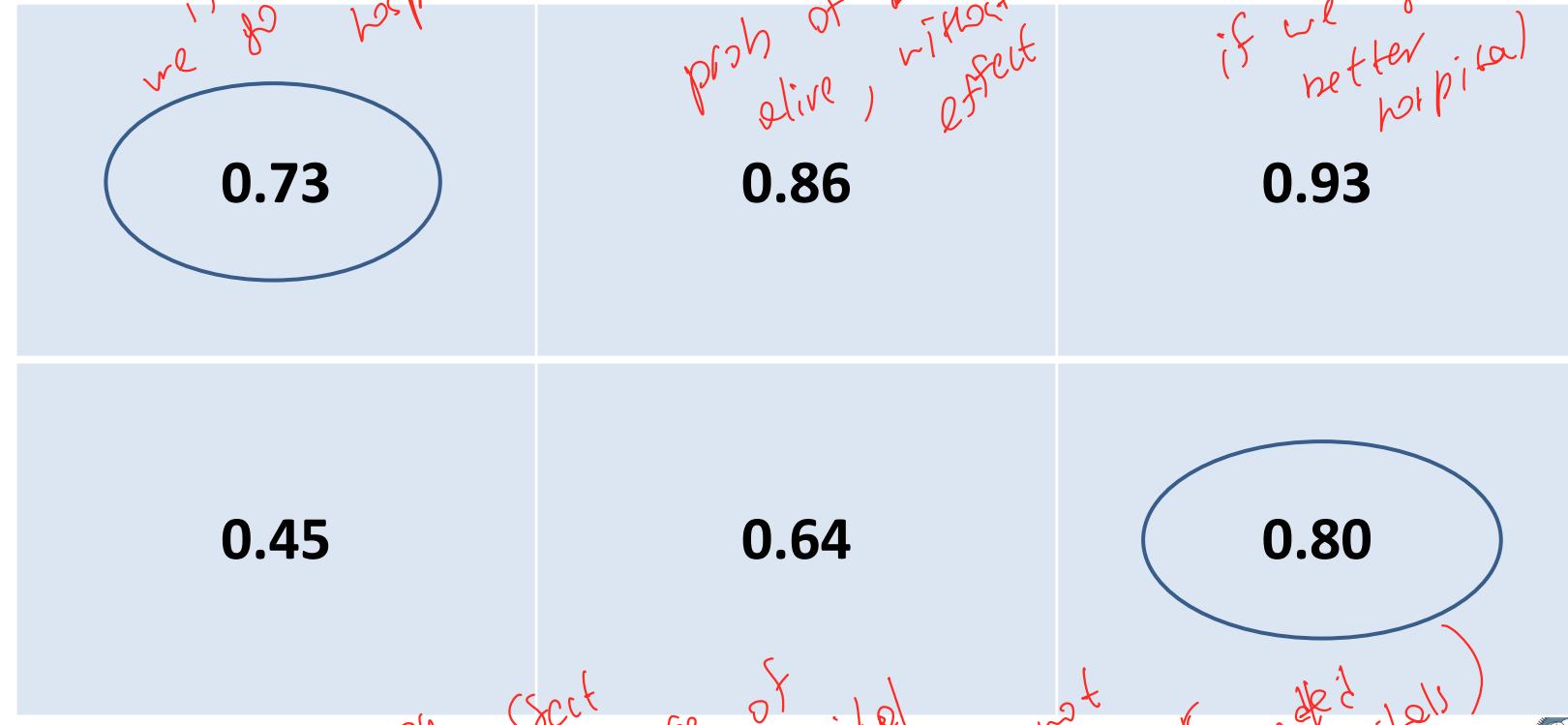
$+2\hat{\sigma}$

if we go better hospital

Man
65 years
MCS = 1
NO adjuvant Chemio
Low complexity surgery



Man
65 years
MCS = 1
NO adjuvant Chemio
High complexity surgery



how can affect choice of hospital (any), not only our preferred hospitals



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**

scenario analysis
new
of
groups



Case studies

- Assessment of pupils learning (INVALSI data)



Case study II – Semiparametric LMM for clustering school effects



➤ INVALSI data: 6572 students within 462 schools

- Student level covariates:
 - Mathematics INVALSI test score at grade 8 (MATH8)
 - Mathematics INVALSI test score at grade 6 (MATH6)
 - Socio-economic index (ESCS)
- School level-covariates:
 - Info about School body composition
 - Info about Scuool principal
 - Info about managerial practices



➤ Aim: 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: Can we measure a significant school effect? Is it possible to cluster these effects? How are these groups characterized?



Case study II – Semiparametric LMM

Mixed-effects linear model:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i$$

- Parametric framework: $\mathbf{b}_i \sim N_{(q+1)}(\mathbf{0}, \sigma_b^2 \mathbf{D})$
- Semiparametric framework: $\mathbf{b}_i \sim P^*$ discrete distribution

Semiparametric mixed-effects linear model:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{c}_m + \boldsymbol{\varepsilon}_i$$

- $m=1, \dots, M$
- $\mathbf{c}_m \in \mathbb{R}^{(q+1)}$ follows a discrete distribution P^* with M (unknown) support points
- Each group i , for $i = 1, \dots, N$, is assigned to a cluster m , that is characterized by random parameters \mathbf{c}_m
→ in-built clustering!



Case study II - Semiparametric LMM

P^* can be interpreted as the mixing distribution that generates the density of the stochastic model.

The ML estimator \hat{P}^* of the random effects distribution P^* can be expressed as:

- a set of M points $(\mathbf{c}_1, \dots, \mathbf{c}_M)$, where $M < N$ and $\mathbf{c}_m \in \mathbb{R}^{(q+1)}$, for $m = 1, \dots, M$,
- and a set of weights (w_1, \dots, w_M) , where $\sum_{m=1}^M w_m = 1$ and $w_m > 0$ for $m = 1, \dots, M$.

w_m represents the proportion of groups belonging to each cluster m.

Besides the estimate of M, the joint estimation of σ^2 , β , $(\mathbf{c}_1, \dots, \mathbf{c}_M)$ and (w_1, \dots, w_M) , is performed through the maximization of the likelihood, mixture by the discrete distribution of the random effects (EM algorithm):

$$L(\beta, \sigma^2 | \mathbf{y}) = p(\mathbf{y} | \beta, \sigma^2) =$$
$$\sum_{m=1}^M \frac{w_m}{(2\pi\sigma^2)^{\frac{\sum_{i=1}^N n_i}{2}}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^N \sum_{j=1}^{n_i} (y_{ij} - c_{0m} - \beta x_{ij} - c_{1m} z_{ij})^2 \right\},$$





Semi-parametric two-level model for **students** (level 1) nested within **schools** (level 2):

$$\begin{aligned} \mathbf{y}_i &= c_{0m} + c_{1m}\mathbf{z}_i + \beta\mathbf{x}_i + \epsilon_i & i = 1, \dots, N & m = 1, \dots, M \\ \epsilon_i &\sim N(\mathbf{0}, \sigma_\epsilon^2) & \text{ind.} & \end{aligned} \quad (\heartsuit)$$

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.





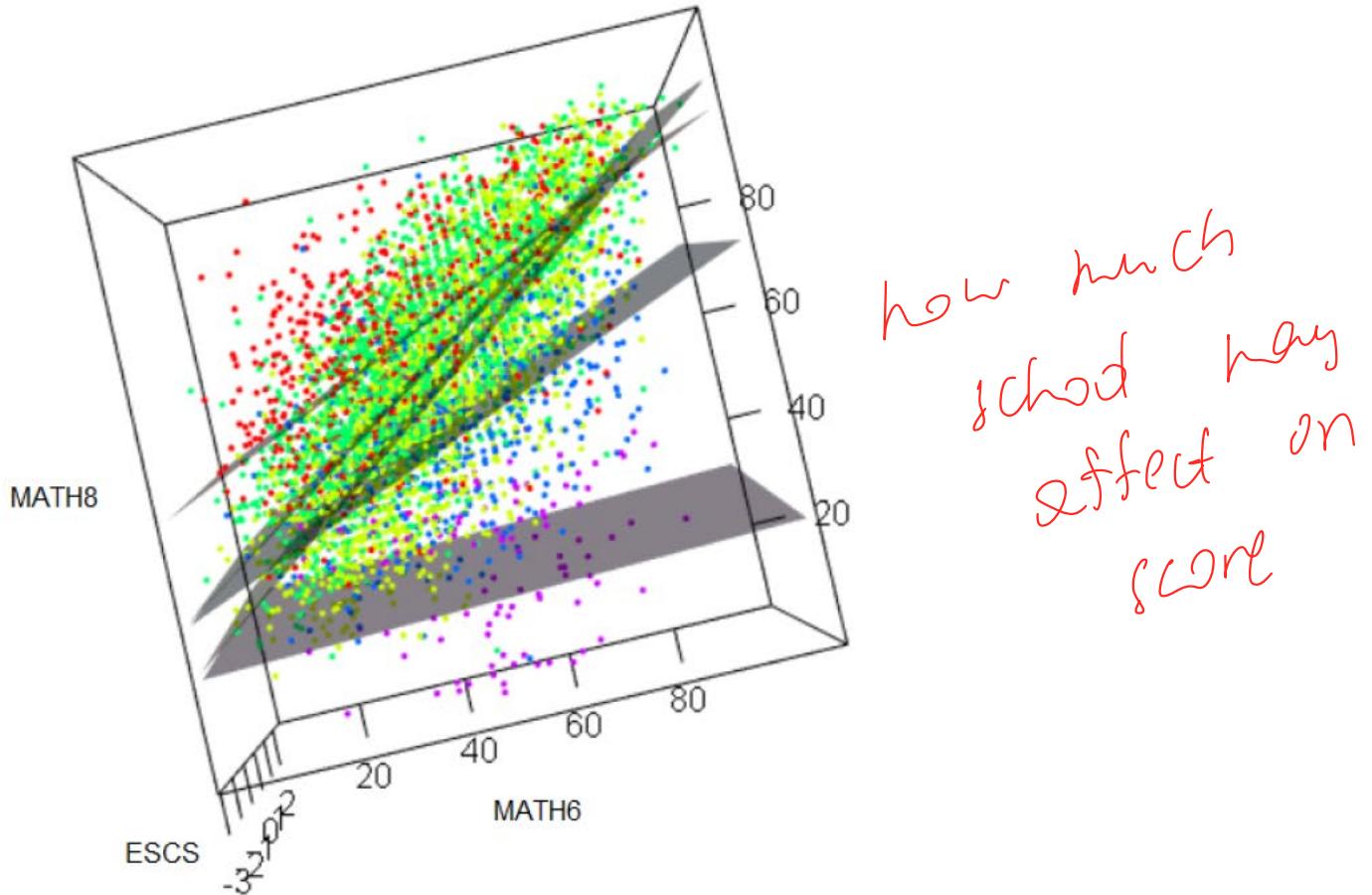
The SPEM algorithm, considering $D = 0.5$ and $\tilde{w} = 0.01$, identifies M=5 distinct clusters of schools:

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.





To explore **a posteriori the clusters** we apply the following **multinomial logit model**, for each school i , $i = 1, \dots, N$ 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 .





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.



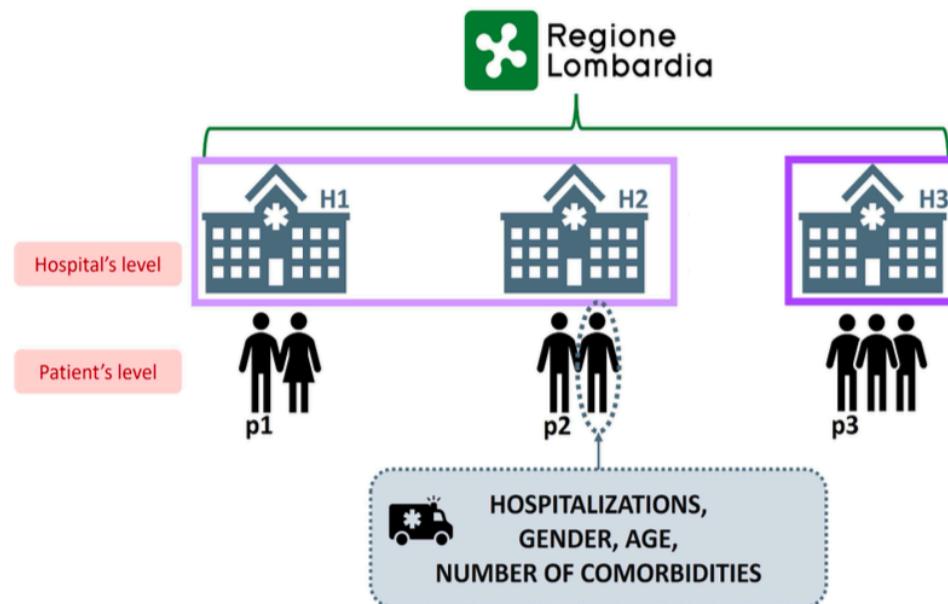
Case studies

- RL hospital effect on HF patients survival



Case study III – Dataset and Motivation

Consider cardiocirculatory pathologies-related **hospitalizations** of **patients** affected by chronic heart failure across **56 hospitals** in Regione Lombardia (Mazzali et al., 2016).



$$\text{logit}(\pi_{ij}) = \text{Hospital effect} + \beta_1 \text{Sex}_{ij} \\ + \beta_2 \text{Comorbidities}_{ij} + \beta_3 \text{Age}_{ij}$$

where π_{ij} is the probability of re-hospitalization in the 1st year for patient i in hospital j .

Aims

- Quantify hospital-specific effects on response
- Reduce dimensionality by **clustering** hospitals
- Enhance **interpretability** → **decision-making**

Ragni, A., Masci, C., Ieva, F., & Paganoni, A. M. (2022). Semi-parametric generalized linear mixed effects models for binary response for the analysis of heart failure hospitalizations. *Proceedings of 51th Scientific Meeting of the Italian Statistical Society*, 2042–2047.



Case study III – The GLMM with Discrete Random Effects model

Let **groups** be indexed by $i = 1, \dots, N$, containing **nested observations** indexed by $j = 1, \dots, n_i$.

Let the **clusters of groups** be indexed by $m = 1, \dots, M$.

Traditional Setting: GLMM

$$g(\mathbb{E}[\mathbf{y}_i | \mathbf{b}_i]) = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i$$

$$\text{s.t. } \mathbf{b}_i \sim \mathcal{N}_Q(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{b}_i})$$

Marginal Likelihood:

$$\mathcal{L}(\boldsymbol{\beta}, (\mathbf{b}_1, \dots, \mathbf{b}_N) | \mathbf{y}) = \prod_{i=1}^N p(\mathbf{y}_i | \boldsymbol{\beta}, \mathbf{b}_i)$$

New Setting: GLMMDRE

$$g(\mathbb{E}[\mathbf{y}_i | (\mathbf{c}_1, \dots, \mathbf{c}_M)]) = \mathbf{X}_i \boldsymbol{\beta} + \sum_{m=1}^M \mathbb{1}_{\{i \in m\}} \mathbf{Z}_i \mathbf{c}_m$$

$$\text{s.t. } p(\mathbf{b}_i = \mathbf{c}_m) = \omega_m, \sum_{m=1}^M \omega_m = 1 \text{ and } \omega_m \geq 0$$

Marginal Likelihood (Aitkin, 1999):

$$\mathcal{L}(\boldsymbol{\beta}, (\mathbf{c}_1, \dots, \mathbf{c}_M) | \mathbf{y}) = \prod_{i=1}^N \sum_{m=1}^M \omega_m p(\mathbf{y}_i | \boldsymbol{\beta}, \mathbf{c}_m)$$



Main steps of the “modified” EM (Dempster et al., 1977)

1. **Initialization** N GLMs are fitted per group to initialize **coefficients**; **weights** are uniformly distributed $\hat{\omega}_m^{(0)} = 1/N$ for $m = 1, \dots, N$;
2. **Support reduction** : Confidence intervals/regions for the two nearest support points are **computed** and **merged if they overlap** → new M is identified;
3. **Parameters update** given number of clusters M & cluster **non-emptiness** check;
4. **Convergence** : When no $1 - \alpha$ confidence regions overlap, and the change in fixed/random effect estimates is negligible.

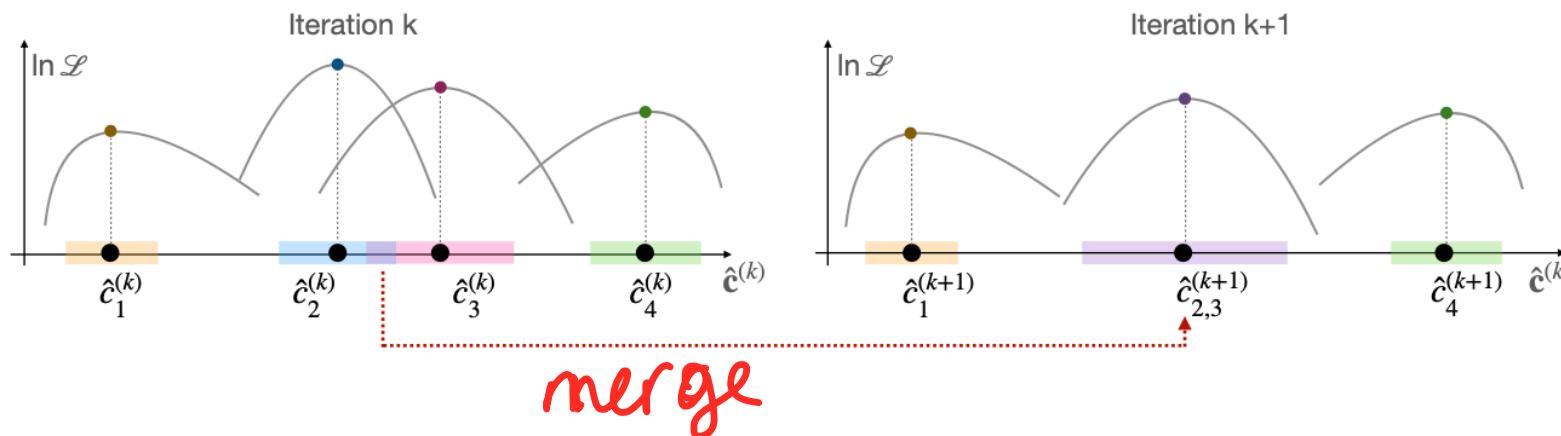
to merge
clusters
to reduce
them



Case Study III – Support reduction procedure

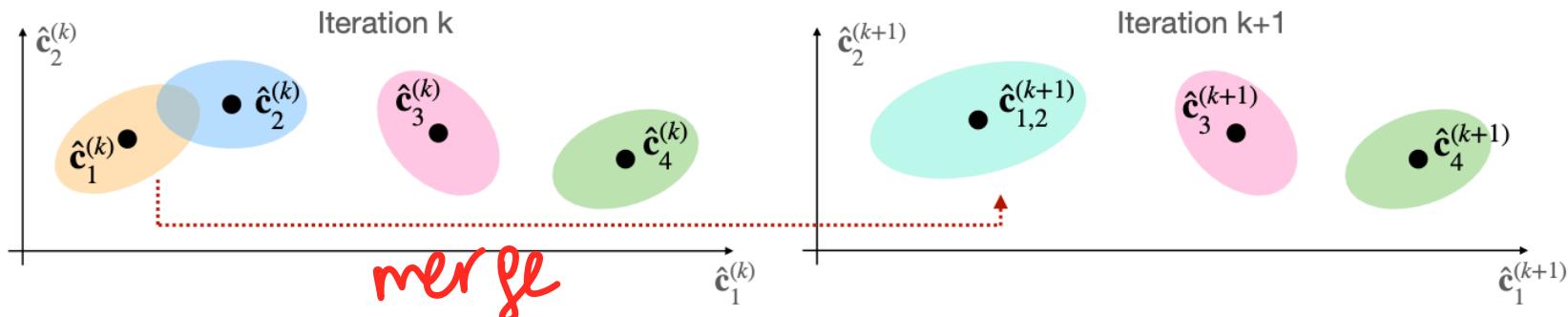
Q=1

$$CI_{1-\alpha}(\hat{c}_m^{(k)}) = \left[\hat{c}_m^{(k)} \pm z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}(\hat{c}_m^{(k)})} \right]$$



$$CR_{1-\alpha}(\hat{c}_m^{(k)}) = \{ \mathbf{c} : (\mathbf{c} - \hat{c}_m^{(k)})' [\text{Var}(\hat{c}_m^{(k)})]^{-1} (\mathbf{c} - \hat{c}_m^{(k)}) \leq \chi^2_{1-\alpha}(Q) \}$$

Q>1



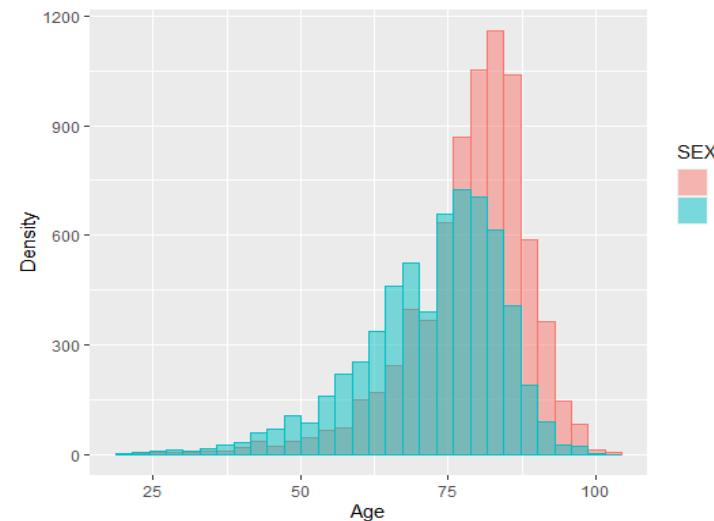
Case Study III – Lombardy Region’s Heart Failure Data

Data collection:

- Data were supplied by **Regione Lombardia (Italy)** [Regione Lombardia. HFData project (2012), Mazzali, Cristina, et al. (2016)]
- In patients experiencing at least one HF event between 2006-2012, we selected hospitalizations for causes related to **cardiocirculatory pathologies**.
- We included in the analysis only
 - patients who didn't leave the study before its end (except for death) and didn't change hospital at different hospitalizations
 - hospitals receiving $100 \leq \text{patients} \leq 400$ and in which at least a patient had a re-hospitalization
- Collected data include 13835 patients, hospitalized into 56 hospitals

Model building:

- For each patient,
 - **Fixed covariates:** *sex, age* and *number of comorbidities* out of the twenty most significant (e.g. dementia, coagulopathy, hypertension, psychosis, arrhythmia,...), recorded at the first hospitalization
 - **Random intercept:** for the (*cluster*) of curing hospital



Case Study III – GLMM vs GLMMDRE

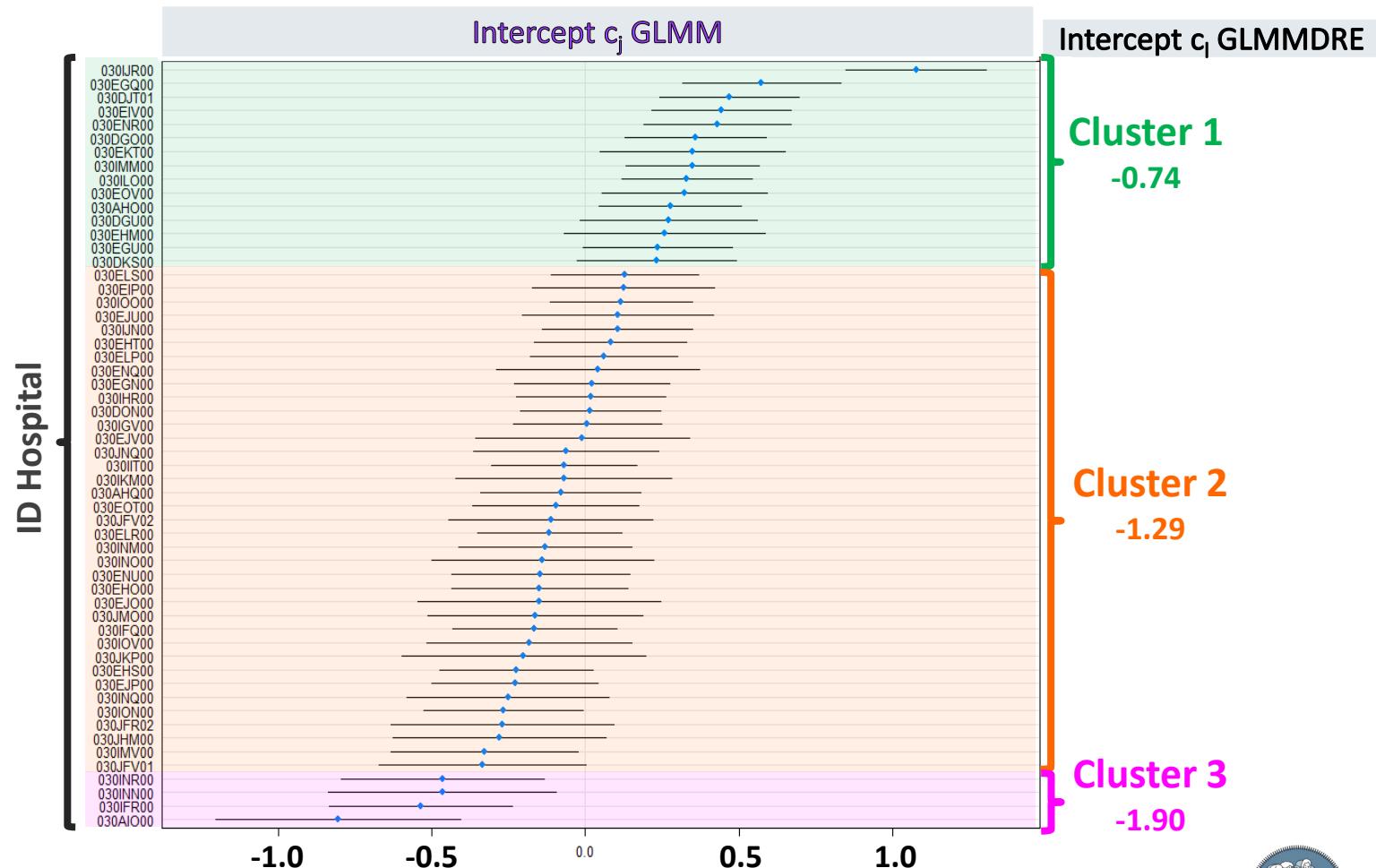
$$\text{logit}(\pi_{ij}) = \text{Intercept} + \beta_{\text{SexF}} \cdot \text{SexF}_{ij} + \beta_{\text{Comorbidities}} \cdot \text{Comorbidities}_{ij} + \beta_{\text{Age}} \cdot \text{Age}_{ij}$$

where π_{ij} is the probability of re-hospitalization in the 1st year for patient i in hospital j

Case study results ($\alpha=0.05$)

	GLMM		GLMMDRE	
	coeff	pval	coeff	pval
Intercept	$c_j - 1.21$	***	c_l	***
β_{SexF}	0.10	*	0.11	***
$\beta_{\text{Comorbidities}}$	0.05	*	0.05	***
β_{Age}	0.08	***	0.08	*

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '' 0.1 '' 1



Case Study III - Conclusions

Simulations results:

- SPGLMM well performs in the individuation of the simulated coefficients
- With appropriate **choice of alpha**, the correct number of simulated clusters is detected

Case study results:

- Clustering of hospitals (given by random intercept) based on patient characteristics and curing hospital
- By varying the **alpha**, different number of clusters respectively with different degrees of dispersion

Relevance of the study:

- Hospitals **evaluation and profiling**, outliers detection for supporting **decision-making strategies**
- The chosen threshold **alpha** (level of confidence) can be more easily interpreted compared to previous methods relying on a pre-defined Euclidean distance



R corner

Useful references to R packages and R functions to implement mixed-effects models



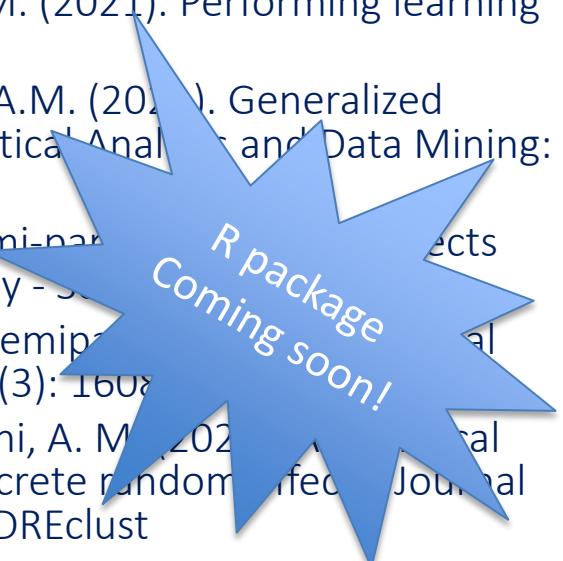
R packages and available codes

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 effects models for unsupervised classification of Italian schools’. Journal of the Royal Statistical Society - Series A, 185(1), 1-26.’
4. **MSPEM()** function: supplementary material of ‘Masci, C., Ieva, F. and Paganoni, A.M. (2022). Semiparametric mixed-effects models: a university students profiling tool.’ The Annals of Applied Statistics, 16(3): 1607-1633.’
5. **algorithm_alpha()** function: supplementary material of ‘Ragni, A., Masci, C., Ieva, F., & Paganoni, A. M. (2022). A statistical significance-based approach for clustering grouped data via generalized linear model with discrete random effects’. Journal of the Royal Statistical Society - Series A. Available at <https://github.com/alessandragagni/glmmDREclust>



References

Journal papers

- Ragni, A., Masci, C., Ieva, F., Paganoni, A.M. (2025) A Statistical Significance-Based Approach for Clustering Grouped Data via Generalized Linear Model with Discrete Random Effects. [Journal of the Royal Statistical Society - Series A](#). doi: 10.1093/jrsssa/qnaf007
- Masci, C., Ieva, F., Paganoni, A.M. (2024) Inferential tools for assessing dependence across response categories in multinomial models with discrete random effects. [Journal of Classification](#). doi: 10.1007/s00357-024-09466-2
- Masci, C., Ieva, F. and Paganoni, A.M. (2022). Semiparametric multinomial mixed-effects models: a university students profiling tool.' The Annals of Applied Statistics, 16(3): 1608-1632 (September 2022) doi: 10.1214/21-AOAS1559
- 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](https://doi.org/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](https://doi.org/10.21105/joss.01412).



Lecture 19.05.25

Ensemble methods

- Supervised problem

- Training set:

$y \in \mathbb{R}$ target

$\underline{x} \in \mathbb{R}^D$ vec of features

$$\mathcal{X} = \begin{bmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_n \end{bmatrix} \quad \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

(\underline{x}_i, y_i) - independent $i = 1 \dots n$

- Model: $y = f(\underline{x}) + \varepsilon$, $E[\varepsilon] = 0$, $\varepsilon \perp \underline{x}$

fitted model: $\hat{f}(\mathcal{X})$

- Recall bias-variance tradeoff

If $\underline{x}_0 \notin \mathcal{X}$, $\underline{x}_0 \notin \begin{bmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_n \end{bmatrix}$

goal predict: $y_0 = f(\underline{x}_0) + \varepsilon_0$

Expected prediction error:

$$E[(y_0 - \hat{f}(\underline{x}_0))^2] = \underbrace{\left[f(\underline{x}_0) - E[\hat{f}(\underline{x}_0)] \right]^2}_{\text{bias}^2} + \underbrace{\text{Var}(\hat{f}(\underline{x}_0))}_{\text{Var r}} + \underbrace{\text{Var}(\varepsilon_0)}_{\text{irreducible}}$$

bias \downarrow \Leftrightarrow var \uparrow low bias, high variance

Example CART - typical example, where we have to work of bias-variance trade-off, minimising variance and not overfit (low bias)

So can we keep low bias and reduce the variance?

one of the solution - increase sample size $n \uparrow$ (but it still have to be i.i.d.)

Paredig matic example

$$\mathbf{X} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad y_i \quad i=1..n \quad \text{i.i.d } \sim f, \sigma^2$$

Model: $y = \mu + \varepsilon$

for $i=1..n$ $E[y_i] = \mu$

$$\hat{f}_{\mathbf{X}} = \bar{y}_n \rightarrow \text{unbiased} \quad \leftarrow \begin{array}{l} \text{bias-var} \\ \text{decomp.} \end{array}$$

$$E[(y_i - \hat{f}_{\mathbf{X}})^2] = \underbrace{(\mu - E[\hat{f}_{\mathbf{X}}])^2}_{0} + \text{Var}(\hat{f}_{\mathbf{X}}) + \text{Var}(\varepsilon_i) = 2\sigma^2$$

$$\hat{f}_{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}_n \Rightarrow E[\bar{y}_n] = \mu$$

$$E[(y_i - \hat{f}_{\mathbf{X}})^2] = \underbrace{\frac{1}{n} \sigma^2}_{0} + \sigma^2 \leq 2\sigma^2$$

$$\downarrow n \rightarrow \infty$$

So idea is taking means of independent observations, which lead to the lower variance

Idea If we had

$$\hat{f}_{\underline{x}_1}, \hat{f}_{\underline{x}_2}, \dots, \hat{f}_{\underline{x}_B} \text{ i.i.d. } \sim \hat{f}_{\underline{x}}$$

$$\Rightarrow \text{my predictor is } \hat{f}_B = \frac{1}{B} \sum_b \hat{f}_{\underline{x}_b} = ?$$

$$\Rightarrow E[\hat{f}_B] = E[\hat{f}_{\underline{x}}]$$

$$\text{Var}[\hat{f}_B] = \frac{1}{B} \text{Var}[\hat{f}_{\underline{x}}]$$

ensemble of predictors - take several models. But how to build this ensemble.

note:

$$E[(y_0 - \hat{f}_B(\underline{x}_0))^2] = [f(\underline{x}_0) - E[\hat{f}_{\underline{x}}(\underline{x}_0)]]^2 +$$

$$+ \frac{1}{B} \text{Var}[\hat{f}_{\underline{x}}(\underline{x}_0)] + \text{Var}(\varepsilon_0) \leq E[(y_0 - \hat{f}_{\underline{x}}(\underline{x}_0))^2]$$

$$\downarrow B \rightarrow \infty$$

How to generate training set:

$$\mathbb{X}_1, \mathbb{X}_2, \dots, \mathbb{X}_B \quad n \times (p+1)$$

$f_{\mathbb{X}_B}$ features \uparrow because of target variable

1) Repeat the n -experiments, generating \mathbb{X} , B times - very costly

- suppose we did this experiments?

$$\underbrace{\mathbb{X}_1 \cup \mathbb{X}_2 \cup \dots \cup \mathbb{X}_B}_{n} \approx \frac{1}{B} \sum_{B=1}^n \mathbb{X}_B$$

fit one model with all collected data.

2) Only one training set \mathbb{X}

\Rightarrow but $\mathbb{X}_1^*, \dots, \mathbb{X}_B^*$ generated by
Bootstrap (Efron, 1975)

\rightarrow resample my training set

For $i = 1 \dots n$, sample a unit from \mathbb{X} :

(x_i^*, y_i^*) with replacement

$$\Rightarrow \hat{X}^* = \begin{bmatrix} \underline{x}_1^* \\ \vdots \\ \vdots \\ \underline{x}_n^* \end{bmatrix} \begin{bmatrix} y_1^* \\ \vdots \\ \vdots \\ y_n^* \end{bmatrix}$$

So we generate by bootstrap, repeated B -times

$$\hat{X}_1^*, \hat{X}_2^*, \dots, \hat{X}_B^* \quad n \times (p+1) \text{ training sets}$$

$$\downarrow$$

$$\hat{f}_{\hat{X}_1^*}, \dots$$

$$\downarrow$$

$$\hat{f}_{\hat{X}_B^*}$$

PROBLEM

\leftarrow independent? \Rightarrow NO

predictor: $\hat{f}_B = \frac{1}{B} \sum_{i=1}^B \hat{f}_{\hat{X}_B^*}$

begging
bootstrap-
aggregation

Breiman (1995)

Are $\hat{X}_1^*, \dots, \hat{X}_B^*$ independent - NO

Fix unit i in $\hat{X} \Rightarrow (\underline{x}_i, y_i)$

$$\begin{aligned} P((\underline{x}_i, y_i) \in \hat{X}^*) &= 1 - P((\underline{x}_i, y_i) \notin \hat{X}^*) \\ &= 1 - \left(1 - \frac{1}{n}\right)^n \rightarrow 1 - e^{-1} \end{aligned}$$

Expected proportion of data shared by X_1 and X_2 is $\approx 1 - \frac{1}{e} \approx 0.63$ for n large

So the problem with bagging is:

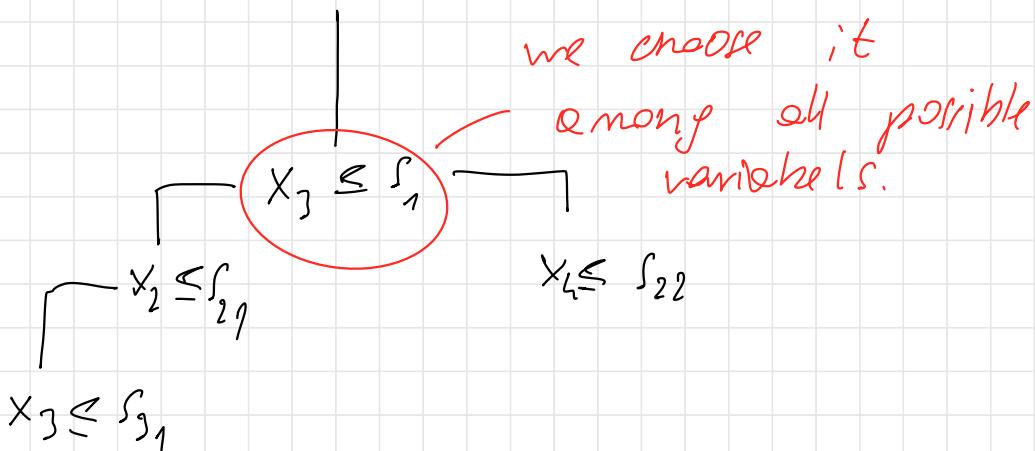
$f_{X_1^*}, \dots, f_{X_n^*}$ are not independent

they are correlated $\Rightarrow \text{Var}(\hat{f}_B) > \frac{1}{B} \text{Var}(\hat{f}_{\bar{X}})$

each tree in each split randomly choose subset of variables from all variables and from them choose optimal

Random Forest (Breiman, 2001) (RF)

We use CART - iteratively split data, using one of feature



- When building a CART, at each split you choose a feature X^* among $(x_1 \dots x_p)$ and a thresh s.t. the decrease in Gini is max.

What if randomly choose these features?

- So when building a tree of a RT at each split you choose X^* (***) among $(X_1^* \dots X_m^*)$ randomly selected from $(x_1 \dots x_p)$ and a thresh... how to choose?

RF:

- $X_1^* \dots X_B^*$ bcof trees
- $f_{X_1^*} \dots f_{X_B^*}$ by (***)
 \uparrow \uparrow (so trees will be less correlated)

Choosing m: if $m=p \Rightarrow$ bagging

- m regression : $m = \frac{1}{3} p$

- m classification : $m = \sqrt{p}$

For regression:

predictor FF: $\hat{f}_B(x) = \frac{1}{B} \sum_{i=1}^B \hat{f}_{x_i^*}(x)$

For classification

predictor RF: majority $\{\hat{f}_{x_1^*}(x), \dots, \hat{f}_{x_B^*}(x)\}$

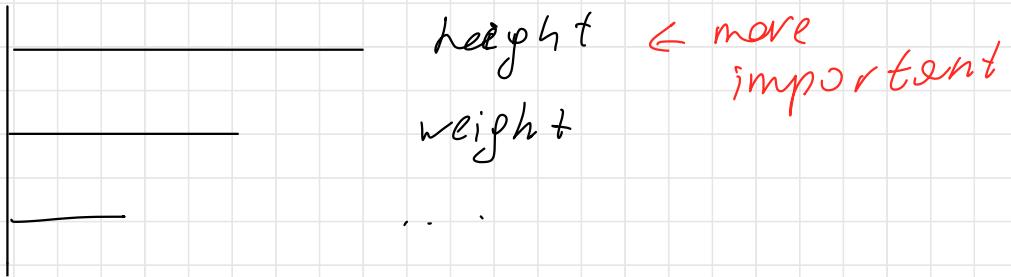
Interpretation

Variable importance

- compute the variability reduction of each variable

At the end sum these reduction for each feature and we will see which variable helped to split tree.

every time x_i is used for generating
a split \Rightarrow log the reduction in
RSS/Gini
 \Rightarrow sum all these reductions \rightarrow index for x_i



Shapley value (Shapley 1951,
cooperative game theory)
(Nobel prize for Economy 2012)

$\{1, \dots, p\}$ players (features, variables)

$$\mathcal{P} = \left\{ S : S \subseteq \{1, \dots, p\} \right\} \text{ set of coalitions}$$

$S \in \mathcal{P}$ S coalition $S \subseteq \{1, 2, \dots, p\}$
(subset of players)

$V: \mathcal{P} \rightarrow \mathbb{R}$ worth function

$v(S)$ is the worth of coalition S

worth of player $i \in \{1 \dots p\}$

If S is a coalition and $i \notin S$
the increase of worth due to adding
 i to S is $v(S \cup \{i\}) - v(S)$

How many coalitions with $|S|$, without
player i

$$\frac{p!}{|S|! (p-|S|-1)!} = p \binom{p-1}{|S|}$$

For $i \in \{1, \dots, p\}$

Shapley value ↴

$$\phi_i = \sum_{S \subseteq \{1, \dots, p\} \setminus \{i\}} \frac{|S|! (p-|S|-1)!}{p!} (v(S \cup \{i\}) - v(S))$$

Properties.

1) Efficiency:

$$\sum_{i=1}^p \phi_i = v(\{1 \dots p\})$$

2) Null player:

$$\text{if } v(s \cup \{i\}) = v(s) \quad \forall s \in \mathcal{P} \Rightarrow \\ \Rightarrow \phi_i = 0$$

3) Symmetry $i, j \in \{1 \dots p\}$

$$\text{and } v(s \cup \{i\}) = v(s \cup \{j\}) \quad \forall s \in \mathcal{P} \Rightarrow \\ \phi_i = \phi_j$$

4) Linearity $v, w: \mathcal{P} \rightarrow \mathbb{R}$ worth let's

$$a, b \in \mathbb{R}$$

$$\phi_i^{av bw} = a \phi_i^v + b \phi_i^w \text{ for } i=1 \dots p$$

Prediction through: \hat{f}

\hat{f} predictor using only features belonging to $s \subseteq \mathcal{S}$

players \hookrightarrow features

V:

Given $\underline{x}_0 \in \mathbb{R}^p$

- \bar{y}

$y_0 = f(\underline{x}_0) + \varepsilon$, specify $V(s) = \hat{f}_{\underline{x}|s}(\underline{x}_0|s)$

Contribution of x_i to predict $\hat{f}_{\underline{x}}(\underline{x}_0)$

$\phi_i(\underline{x}_0) =$

$$\sum_{S \subseteq \{1, \dots, p\} \setminus \{i\}} \frac{p!}{|S|! (p-1-|S|)!} \left(\hat{f}_{\underline{x}|S \cup \{i\}}(\underline{x}_0) - \hat{f}_{\underline{x}|S}(\underline{x}_0) \right)$$

Complex to compute

↓

how to compute faster

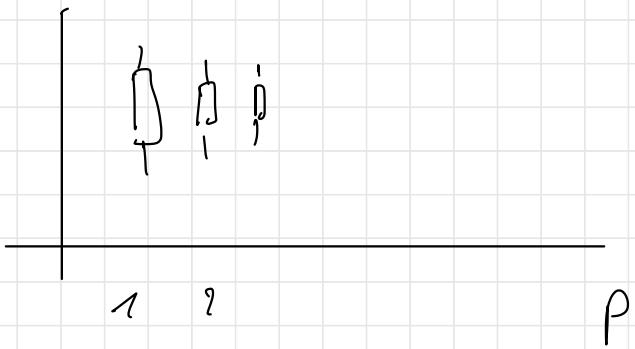
Algorithm

SHAP

Lundberg & Lee (2017)

Tree SHAP Lundberg et al. (2019)
(RF, boosting)

Compute $\phi_i(\underline{x}_k)$, $k = 1, \dots, n$



⇒ C.Mur (2025)

Interpetable Machine Learning a
guide for making black boxe models

Lecture 20.05.25

npjc7pencIDEN-
hse

Every datum is always geo-referenced: time
and location stamp! Spatial statistics:
domain where we look at datum can
be n-dimensional space very abstract

Basic idea: Two datum that are close
in this space are more correlated!

Close with respect what? It also depends
on the notion of distance we consider!
work with Euclidean distance

We have 3 possible types of data:

- Geo-statistics

$d=4$ (latitude, longitude, time)
 $d=1$ dimension of our space

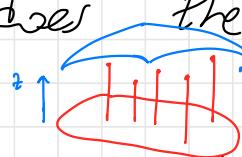
We have a domain $D \subseteq \mathbb{R}^d$ in which
we can identify points $s_i \in D$ and in
each point we have some observation z_{s_i}
(e.g. position)



so we have geographics locations $s_1, \dots, s_n \in G \subset \mathbb{R}^d$ which we assume to be fixed and not random!

At these locations we have observations z_{s_1}, \dots, z_{s_n} random objects (e.g. variables, vectors). \nwarrow partial observation of a random field $\{z_s, s \in G\}$

Our goal is: (characterize the distribution of \rightarrow Estimate the mean \rightarrow Estimate the spatial dependence

- make models: how does the pollution vary in Milan? \uparrow  \rightarrow How to predict this surface
- study spatial dependence and define models for this
- make prediction: we have some new location s_0 can we make predictions in s_0

(we will introduce Kriging Method for this)

- **Lattice (areal) data:** we have a domain D divided in small sub-domains (e.g. grid data or regional data)
 now we have
 region, aggregate
 the region s to z_s

 So we have grid locations $s_1 \dots s_n$ which cover the entire domain (e.g. partition)

and if s_i is a sub-domain, we have an observation z_{s_i}

- $s_1 \dots s_n \subset D$
- z_{s_i} value observed on the area s_i

Now the difference is that s_i is not a point but a sub-region: typically we don't want to make prediction since we have observations all over D , but rather we want to model or cluster!

Note that s_i still fixed

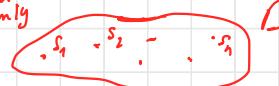
Typical models are Markovian Random Fields!

spatial • Point processes (or patterns): our locations

s_i are now random. So we have a random process realized on some locations $\{s_1, \dots, s_n\}$

z_{s_i} obs at s_i
Examples are murders by a serial killer or earthquake locations!

choose randomly
points



We want to understand if there is a clustering or a bigger presence of locations in some areas with respect to other areas!

The above is the standard point process! But we can also have the so called Marked Point Process in which we also have an observations for each random location!

If also z_1, \dots, z_n with $s_1, \dots, s_n \sim$ Marked sp. point process

If only s_1, \dots, s_n are available \rightarrow std sp point process

So we have the couple (s_i, z_i) for $i = 1, \dots, n$ where s_i is random site and

z_{s_i} is an observation of some feature
in f_i (Poisson model) landslide

Example are: s_i epicenter of earthquake and z_{s_i} is the magnitude!

Geo-statistics $\{z_s, s \in D\}$

Why we should care about spatial dependence? Suppose we have our sampling

sites $s_1, \dots, s_n \in D$ and real random variables z_{s_i} for $i = 1 \dots n$



Stationarity - so the mean and covariance (homogeneity in space)

Second order stationarity $E[z_s] = m$ & $Cov(z_s, z_{s_i}) = C(s_i - s)$ $\forall s, s_i \in D$ (so E and Cov doesn't change)

Note: As a consequence $Cov(z_s) = C(0)$, constant in space

$C: \mathbb{R}^d \rightarrow \mathbb{R}$ is called variogram. So if we know m , C and our data normal \Rightarrow we know everything about our distribution

Properties of C

(i) Positive semi-definite function

$$\sum_i \sum_j d_i d_j C(s_i - s_j) \geq 0 \quad \forall d_i, d_j \in \mathbb{R}$$

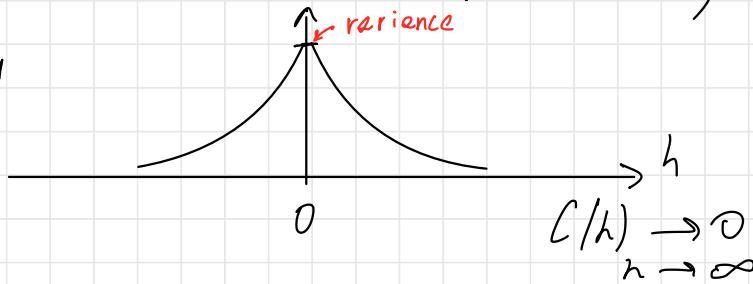
$$s_i, s_j \in D$$

(ii) symmetric function

$$C(-h) = C(h), \forall h \in \mathbb{R}^d$$

(iii) Bounded $|C(h)| \leq C(0), \forall h \in \mathbb{R}^d$

Ex $d=1$



as for away - os
less covariance

Variogram: $2f(s_1 - s_2) = \text{Var}(z_{s_1} - z_{s_2}), s_1, s_2 \in D$

lower variability between z_{s_1}, z_{s_2}

⇒ then more correlated

they are

Note:

$$\begin{aligned} 2f(s_1 - s_2) &= \text{Var}(z_{s_1}) + \text{Var}(z_{s_2}) - 2\text{Cov}(z_{s_1}, z_{s_2}) \\ &= 2C(0) - 2C(s_1 - s_2) \end{aligned}$$

$$\gamma(h) = C(0) - C(h)$$

(Variogram easier to estimate)

Properties: (i) Conditional Negative definite

$$\sum_i \sum_j d_i d_j f(s_i - s_j) \leq 0$$

$\forall d_i, d_j \in \mathbb{R} \text{ s.t. } \sum d_i = 0$
 $\forall s_i, s_j \in D$

↳ Variogram has weaker constraints \Rightarrow
easier to work

(ii) Symmetry: $f(-h) = f(h)$ $\forall h \in \mathbb{R}^d$

(iii) Non negative $f(h) \geq 0$

(iv) Null at 0: $f(0) = 0$

(v) Subquadratic Growth

$$\lim_{\|h\| \rightarrow \infty} \frac{2f(h)}{\|h\|^2} = 0$$

Variogram can not go faster
than $\|h\|^2$ (growth slower than parabola)

Def Isotropy (stronger assumption)

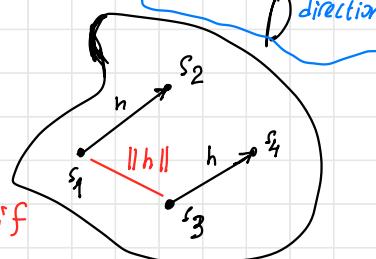
$\{z_s \in D\}$ second order stationary

r.f. is isotropic if:

$$\text{Cov}(z_{s_1}, z_{s_2}) = C/\|s_1 - s_2\|$$

Note: also f will be isotropic if C is so

if we have two couple of points with the same distance then they will have same covariance, even if in different directions



Structural properties of the variogram

- Sill: $\lim_{\|h\| \rightarrow \infty} \gamma(\|h\|) = C(0)$

$$\|h\| \rightarrow \infty$$

- Range:

R s.t. $\gamma(\|h\|) = C(0)$
for $\|h\| \geq R$

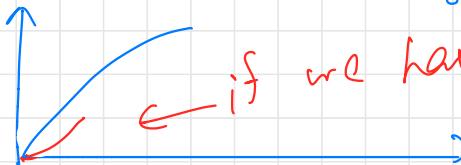
- Nugget $\tau^2 = \lim_{\|h\| \rightarrow 0^+} \gamma(\|h\|)$

if we have
discontinuity \Rightarrow

there is probably
differen (materials)

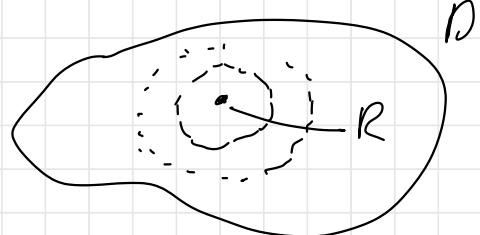
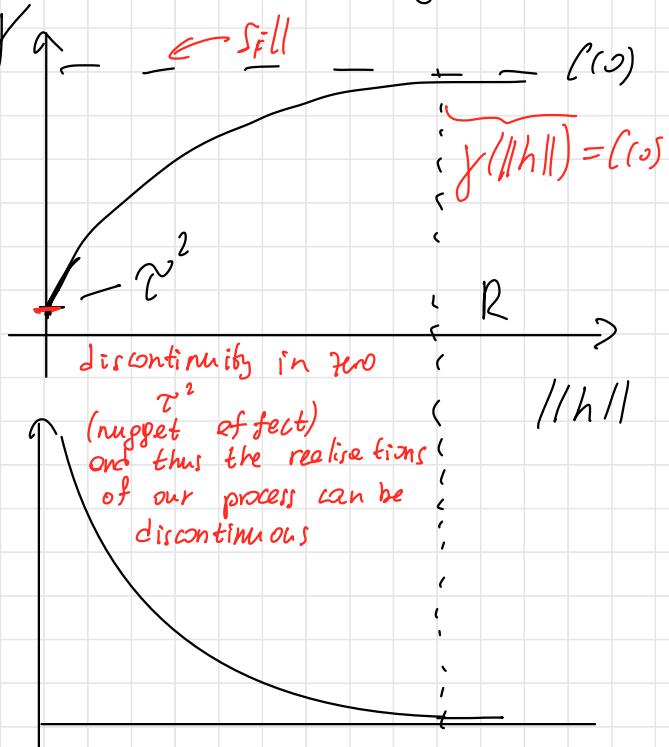
Running this jump of value.

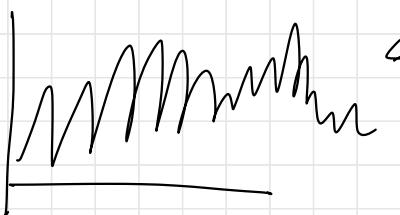
first question - Nugget or not



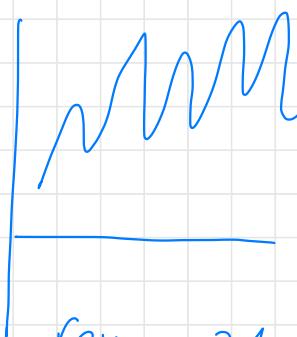
if

we have derivatives here \Rightarrow
there is also in
data





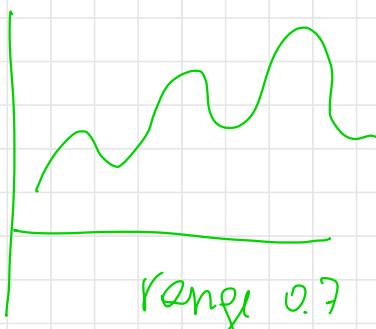
$\leftarrow WRN \Rightarrow$



spherical model
(so correlated with neighbours) in Radius 0.3



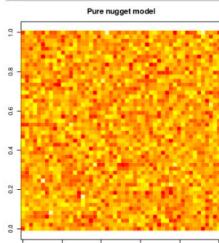
uncorrelated observations
 \Downarrow
go standard models



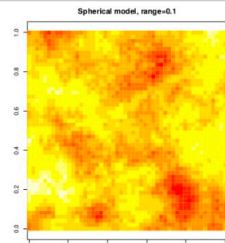
- quadratic
model

The closer to zero the derivative in zeros, the smoother the model

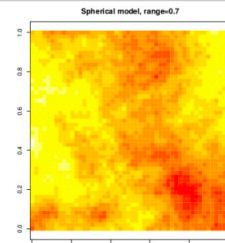
Range



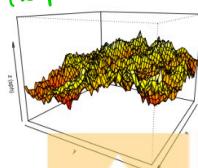
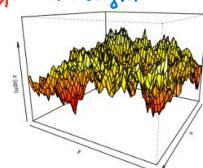
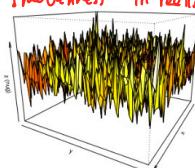
random noise, uncorrelated no smoothness in realization



variogram linear in origin



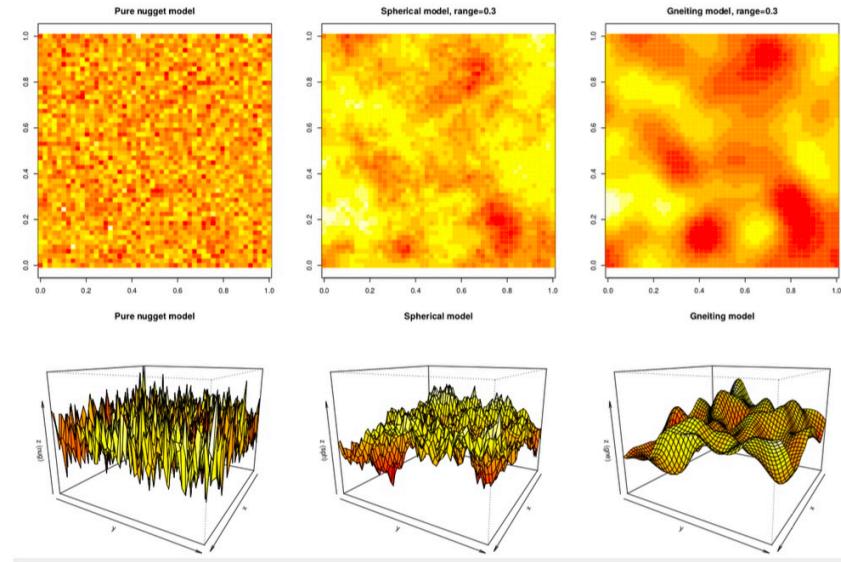
smoother (at point 0.7 distance from the origin, at max, is still correlated)



a point at distance from the origin, at max, is still correlated

\Rightarrow correlation > 0.7 disappear

Consider the following figure in which we have three examples of realisations:



Then:

- The left panel is a white noise
- The central panel is the **spherical model**, with a range of 0.3: this is a model whose vario-gram is linear!
- The third panel is the **Gneiting model** with a range of 0.3: this is a model whose vario-gram is quadratic in zero!

We see that the second and third panel, having the same range, have spots of similar dimension in the contour plot, but we have a clear difference in the regularity (smoothness) of the realisations!

Indeed the closer the derivative (of the vario-gram), in the origin, is to zero, the smoother the realisation of the process!

Suppose \hat{z}_{s_i} comes from a linear model,
 $\therefore \hat{z}_{s_i} = \sum_{\ell=0}^q \alpha_\ell f_\ell(s_i) + \hat{s}_{s_i}$ with $s_i \in D$ where:

- α_ℓ are the regression coefficients
- $f_\ell(s_i)$ are the regressors
- \hat{s}_{s_i} are the residuals (like ε_i in linear model)

If we want estimate coefficients?

$E[\hat{s}_{s_i}] = 0$ but since they are spatially distributed we have that $\text{Cor}(\hat{s}_{s_i}, \hat{s}_{s_j}) \neq 0$
(since we suppose the proximity among different locations induce a spatial dependence in the observations!)

If $\underline{s} = (\hat{s}_{s_1}, \dots, \hat{s}_{s_n})$ vector of residuals, then
 $\text{Cov}(\underline{s}) = \Sigma \neq I$ (not identity $(1 \dots 1)$)

If we use OLS (Ordinary Least Square)
and ignore spatial dependence

$$\hat{\alpha}_{OLS} = (F^T F)^{-1} F^T \underline{z} \quad / F - \text{design matrix}$$

$$\text{Cov}(\hat{\alpha}_{OLS}) = (F^T F)^{-1} F^T [F(F^T F)^{-1}]^T \text{observations}$$

previously data was i.i.d $\Rightarrow \Sigma = \sigma^2 I$

If we use GLS

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{F}^T \boldsymbol{\Sigma}^{-1} \mathbf{F})^{-1} \mathbf{F}^T \boldsymbol{\Sigma}^{-1} \mathbf{z} \Rightarrow \text{Cov}(\hat{\boldsymbol{\beta}}_{GLS}) = (\mathbf{F}^T \boldsymbol{\Sigma}^{-1} \mathbf{F})^{-1}$$

Note: $\text{Cov}(\hat{\boldsymbol{\beta}}_{OLS}) - \text{Cov}(\hat{\boldsymbol{\beta}}_{GLS}) \geq 0$ that is:

positive semi-definite \Rightarrow

$\hat{\boldsymbol{\beta}}_{OLS}$ has higher variance!

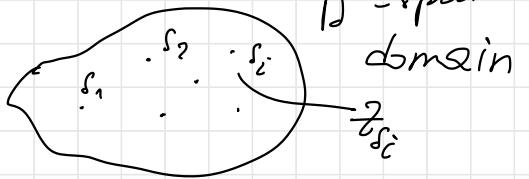
so we need to care about spatial dependence.

Lecture 27. 05.25

- Develop estimators

$$\{z_s, s \in D\}, D \subset \mathbb{R}^d$$

$d=2, 3$



z_{s_1}, \dots, z_{s_N} - observations

Assumptions (*)

- $E[z_s] < \infty \quad \forall s \in D$

- $\text{Var}[z_s] < \infty \quad \forall s \in D$

- Second-order stationarity

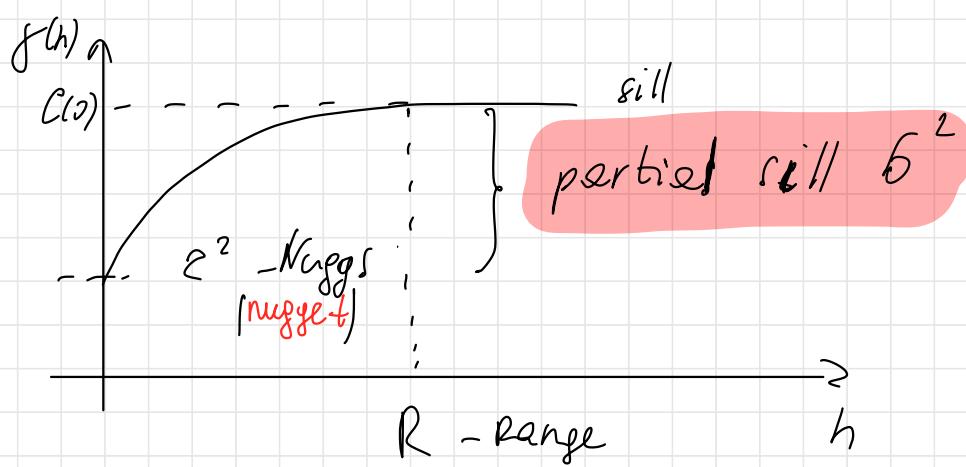
- $- E[z_s] = m \quad \forall s \in D$ *variogram function*

- $- \text{Cor}(z_{s_i}, z_{s_j}) = \underline{C(s_i, s_j)}$ *it's different variogram function*

- + Isotropy: $\text{Cor}(z_{s_1}, z_{s_2}) = \underline{C(||s_1 - s_2||)}$
 ↑ here it function of number

Variogram: $2f(h) = \text{Var}(z_{s_1} - z_{s_2})$ for $s_1, s_2 \in D$ with $h = ||s_1 - s_2||$

(2nd order stationary & isotropy)



Note: Under second order stationarity

$$(f \text{ is isotropic}) \quad 2f(h) = E[(z_{s_1} - z_{s_2})^2] -$$

$$- (E[z_{s_1}] - E[z_{s_2}])^2$$

$\underbrace{\qquad\qquad}_{m}$ $\underbrace{\qquad\qquad}_{m}$
 $\underbrace{\qquad\qquad}_{0}$

partial sill
 c^2

$$\begin{aligned} \text{Var}(x) &= \\ &= E[(x - E[x])^2] = \\ &= E[X^2] - E[X]^2 \end{aligned}$$

Estimating Spatial Dependency

Goal: Estimate $2f(h)$ under assumption (*)

Note $f(h) = C(0) - C(h)$, $C(h) = C(0) - f(h)$

so we will estimate f and $= \lim_{h \rightarrow \infty} f(h) - f(h)$
from that get C

So we have to estimate $E[(z_{s_1} - z_{s_2})^2]$

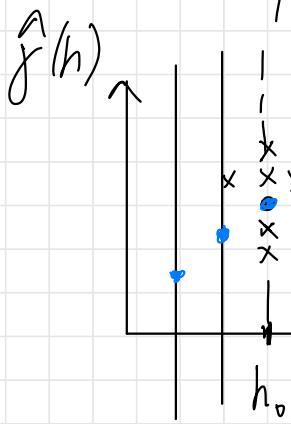
1 Empirical Estimator (first step)

- Fix $h = h_0$

Select all the pairs (s_i, s_j) with $h_0 = \|s_i - s_j\|$

$$N(h_0) = \{(s_i, s_j) \text{ with } h_0 = \|s_i - s_j\|\}$$

$$2\hat{f}(h_0) = \frac{1}{|N(h_0)|} \sum_{(s_i, s_j) \in N(h_0)} (z_{s_i} - z_{s_j})^2$$



Biliné Variogram
but we don't know
guarantee, that
these estimator
will be valid.

2 Fit a valid model to the empirical estimate.

- Suppose we have a parametric valid model $f(h; \theta)$

• Goal: Find $\theta \in \Theta$ which best fits our $\hat{f} = (\hat{f}_1, \hat{f}_2, \dots, \hat{f}_L)'$

with $\hat{f}_i = \hat{f}(h_i)$, $i = 1 \dots L$

(in LRM we had estimators for covariance all of them was maximum likelihood)
we may try see robustness

For instance, Least squares:

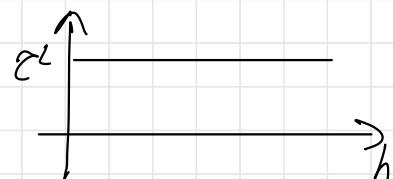
$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^L (\hat{f}_i - f(h; \theta))^2 \quad \begin{matrix} \text{will be OLS} \\ \text{estimators} \end{matrix}$$

$\begin{matrix} \text{will be weighted least} \\ \text{squares} \end{matrix}$

Parametric families

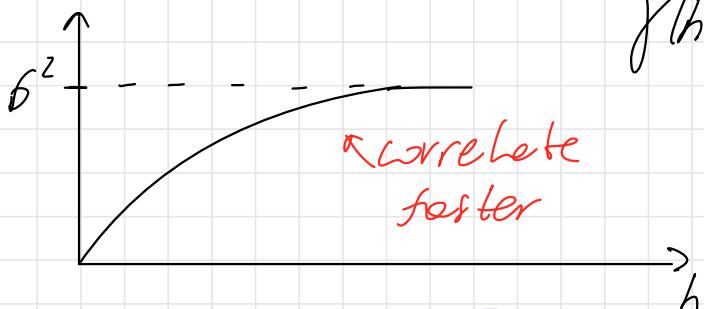
- Pure nugget (constant)

$$f(h) = \begin{cases} 0 & h=0 \\ \varepsilon^2 & h>0 \end{cases}$$



Correspond to random noise, and in random noise it will be best estimator?

- Exponential model

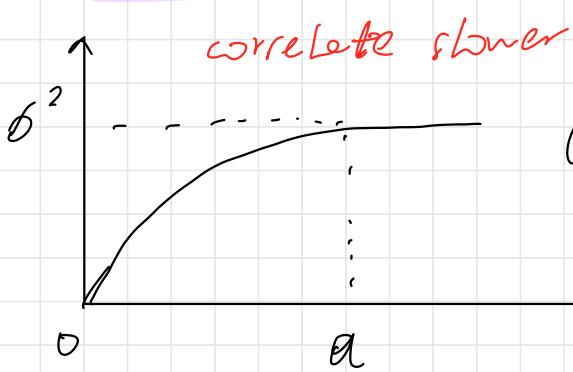


$$f(h) = \begin{cases} \sigma^2(1 - e^{-h/\alpha}), & h > 0 \\ 0, & h = 0 \end{cases}$$

$$\sigma^2 > 0$$

$$\alpha > 0$$

- Spherical model (I just wrote them wrong)



$$f(h) = \begin{cases} \sigma^2 \left(\frac{3}{2} \frac{h}{\alpha} - \frac{1}{2} \left(\frac{h}{\alpha} \right)^3 \right), & h > 0 \\ \sigma^2, & h = 0 \\ 0, & h < 0 \end{cases}$$

- Matern model: ① (smoothness of the line) in some case it will be exponential model

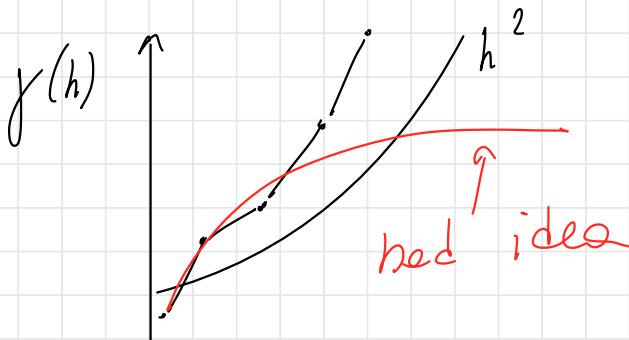
Remark: Properties of valid model

* f_1, f_2 valid, $f_1 + f_2$ - also valid

* $c > 0$, f_2 - valid $\Rightarrow cf_2$ valid

$\gamma_1 \dots \gamma_k$ valid, $c_1 \dots c_n > 0$

$c_1 \gamma_1 + \dots + c_n \gamma_n$ - valid



Recall: (v)
 $\lim_{\|h\| \rightarrow \infty} \frac{\gamma(\|h\|)}{\|h\|^2} = 0$

We are estimating the variogram γ_h (under assumptions (**)) by estimating the

$$\hat{\gamma}(h) = E[(z_{s_1} - z_{s_2})^2] \quad h = \|s_1 - s_2\|$$

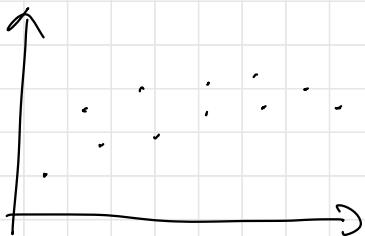
but if we out of stationarity \Rightarrow

$$\hat{\gamma}(h) = E[(z_{s_1} - z_{s_2})^2] - \underbrace{(E[z_{s_1}] - E[z_{s_2}])^2}_{\text{biased part}}$$

When estimating $\hat{\gamma}$ we target

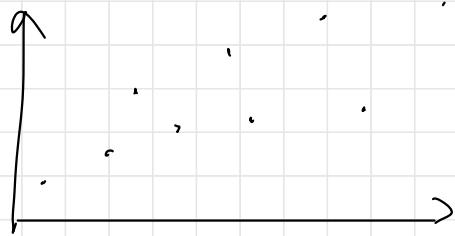
$$E[(z_{s_1} - z_{s_2})^2] = \gamma(h) + (E[z_{s_1}] - E[z_{s_2}])^2 \Rightarrow$$

positive bias if $E[z_{s_1}] \neq E[z_{s_2}]$



stationary

Variogram stable



Not stationary

Variogram unstable

We check directions, to check kriging
because Isotropy assumption - check $\gamma(\|\mathbf{h}\|)$
ignoring direction, but we can say
which directions we want to say

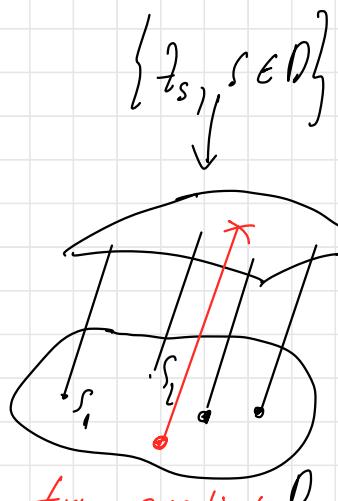
Prediction KRIGING

$$\{z_s, s \in D\} \subset \mathbb{R}^d, d=2, 3$$

z_{s_1}, \dots, z_{s_n} observations

$s_0 \in D$ target location

Goal: Prediction of z_{s_0}



try predict

point estimator \neq prediction

Goal: Find a measurable function $f(\underline{z})$ with $\underline{z} = (z_{1,1}, \dots, z_{n,n})$, which best approximates z_{s_0} in the sense $E[(z_{s_0} - f(\underline{z}))^2]$

Solution: condition expectation

$$f^*(\underline{z}) = E[z_{s_0} | \underline{z}]$$

(i) unbiasedness: $E(E[z_{s_0} | \underline{z}]) = E[z_{s_0}]$

(ii) Orthogonality: $\text{Cor}(E[z_{s_0} | \underline{z}], z_{s_0} - E[z_{s_0} | \underline{z}]) = 0$

(iii) Interpolation: $s_0 \in \{s_1, \dots, s_n\} \Rightarrow E[z_{s_0} | \underline{z}] = z_{s_0}$

our initial date

Under Gaussianity

(iv) Linearity in the date: $E[z_{s_0} | \underline{z}] = d_0 + \sum_{i=1}^n d_i z_{s_i}$ for some $d_0, d_1, \dots, d_n \in \mathbb{R}$

Instead of looking $f^*(\underline{z})$ we will look on properties and functions, having these properties

Kriging: Look for the best linear unbiased Predictor (BLUP)

$$z_s^* = d_0^* + \sum_{i=1}^n d_i^* z_{s_i} \quad \text{where weights } d_0^*, d_1^*, \dots, d_n^*$$

solve: $\min_{d_0, \dots, d_n \in \mathbb{R}} E[(z_{s_0} - (d_0 + \sum_{i=1}^n d_i z_{s_i}))^2]$

s.t. $E[d_0 + \sum_{i=1}^n d_i z_{s_i}] = E[z_{s_0}]$

Assumptions:

- Covariance known: $\text{Cov}(z_{s_1}, z_{s_2}), \forall s_1, s_2 \in D$

Simple Kriging

- Known mean $E[z_s] = m_s, \forall s \in D$

Ordinary Kriging:

- Second order stationarity: $E[z_s] = m$
(isotropy) $\text{Cov}(z_{s_1}, z_{s_2}) = C(\|s_1 - s_2\|)$

Universal Kriging

- Non-stationarity assumptions

$$E(z_s) = n_s, \text{Cov}(z_{s_1}, z_{s_2}) = C(\|s_1 - s_2\|)$$

Ordinary kriging

$$\bullet E[d_0 + \sum_i d_i z_{s_i}] = E[z_{s_0}]$$

$$d_0 + \sum_i d_i m = m$$

Uniform unbiasedness

$$\begin{cases} d_0 = 0 \\ \sum_{i=1}^n d_i = 1 \end{cases}$$

* Objective function

$$\phi(\underline{d}, s) = E[(z_{s_0} - \underbrace{\sum_{i=1}^n d_i z_{s_i}}_{\text{due to unbiasedness}})^2] + 2 \sum_{i=1}^n (d_i - 1)$$

$$(d_1, \dots, d_n) \quad \sum_{i=1}^n d_i z_{s_i} \quad \text{variance}$$

$$= \text{Var}(z_{s_0} - \sum_{i=1}^n d_i z_{s_i}) + 2 \sum_{i=1}^n (d_i - 1) =$$

$$= \text{Var}(z_{s_0}) + \text{Var}(\sum_i d_i z_{s_i}) - 2 \text{Cov}(z_{s_0}, \sum_i d_i z_{s_i})$$

$$= C(d_0) + \sum_{i=1}^n \sum_{j=1}^n d_i d_j C(\|s_i - s_j\|) - \sum_i d_i C(\|s_i - s_i\|)$$

$$+ 2 \sum_i (d_i - 1)$$

$$\frac{\partial \phi}{\partial s_i} = 2 \sum_{j=1}^n \lambda_j C(\|s_i - s_j\|) - 2C(\|s_0 - s_i\|) + 2 \quad i=1..n$$

$$\frac{\partial \phi}{\partial \lambda_i} = 2(\sum_i \lambda_i - 1)$$

$$\begin{cases} \frac{\partial \phi}{\partial \lambda_i} = 0 & i=1..n \\ \frac{\partial \phi}{\partial s} = 0 \end{cases} \rightarrow \begin{cases} \sum_{j=1}^n \lambda_j C(\|s_i - s_j\|) + \sum_i \lambda_i = 1 \end{cases} = C(\|s_0 - s_i\|)$$

Kriging system: coherence with target

$$\begin{pmatrix} \Sigma & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \underline{\lambda} \\ \underline{y} \end{pmatrix} = \begin{pmatrix} \underline{\theta}_0 \\ 1 \end{pmatrix} \quad \begin{matrix} \Sigma_{ij} = C(\|s_i - s_j\|) \\ i, j = 1..n \end{matrix}$$

$$\theta_{0i} = C(\|s_0 - s_i\|)$$

$$i=1..n$$

Lecture 23.05.25

Universal Kriging

$\{z_s, s \in D\}$, $D \subset \mathbb{R}^d$ $d=2, 3$

z_{s_1}, \dots, z_{s_N} obs

s_0 target location in D

Goal: Spatial Prediction of z_{s_0}

Kriging: Look for the BLUP

Best linear
unbiased
predictor

Assumptions:

- No stationarity: $z_s = m_s + f_s$

$m_s = E[z_s]$ $f_s = z_s - m_s$

mean / drift residual (error of model)

$$E[f_s] = 0$$

• Linear model for us

$m_s = \sum_{l=0}^L \alpha_l \underbrace{f_l(s)}_{\text{regressors' (as } z \text{ in linear model)}}$ coefficients (independence in space)

- $\{f_\ell(s)\}$ regressors depending on $s \in D$
 - known everywhere in D (at least I have to know them in f_1, \dots, f_n and s_0)
 - $f_0(s) = 1$
-

Remark: if $L=0 \Rightarrow m_s = m + f_s$
 (stationarity in the mean
 (isotropy))

- f_s : second-order stationarity with known covariogram C

Universal Kr: $z_{s_0}^* = d_0^* + \sum_{i=1}^n d_i^* z_{s_i}$ where
 d_0^*, \dots, d_n^* solve:

$$\min_{d_0, \dots, d_n \in \mathbb{R}} E \left[(z_{s_0} - (d_0 + \sum_i d_i z_{s_i}))^2 \right]$$

$$\text{s.t. } E[d_0 + \sum_i d_i z_{s_i}] = E[z_{s_0}]$$

- Unbiasedness:

$$E[d_0 + \sum_i d_i z_{s_i}] = E[z_{s_0}]$$

for every
possible value
of s_0 under
model (*)

m_{s_i}

$$d_0 + \sum_i d_i \left(\sum_{\ell=0}^L \varrho_\ell f_\ell(s_i) \right) = \sum_{\ell=0}^L \varrho_\ell f_\ell(s_0)$$

$$d_0 + \sum_{\ell=0}^L \varrho_\ell \sum_{i=1}^n d_i f_\ell(s_i) = \sum_{\ell=0}^L \varrho_\ell d_\ell(s_0) \quad \forall \varrho_\ell \in \mathbb{R}$$

After taking derivatives of ϱ_ℓ :

$$\begin{cases} d_0 = 0 \end{cases}$$

$$\sum_i d_i f_\ell(s_i) = f_\ell(s_0) \quad \ell=0 \dots L$$

- Objective function:

$$\begin{aligned} \hat{\phi}(\underline{d}, \underline{s}) &= E \left[\left(z_{s_0} - \sum_{i=1}^n d_i z_{s_i} \right)^2 \right] + \\ &\quad \left(\begin{matrix} d_1 & \dots & d_n \end{matrix} \right)' \left(\begin{matrix} s_0 & \dots & s_L \end{matrix} \right) \\ &\quad + 2 \sum_{\ell=0}^L \sum_{d_\ell} \left(\sum_{i=1}^n d_i f_\ell(s_i) - f_\ell(s_0) \right) \end{aligned}$$

$$= C(\sigma) + \sum_i \sum_j d_i d_j C(||s_i - s_j||) -$$

$$- 2 \sum_i d_i C(||s_0 - s_i||) + 2 \sum_{\ell=2}^L \sum_{\ell} \left(\sum_{i=1}^n d_i f_\ell(s_i) - f_\ell(s_0) \right)$$

UK system

If $\ell \geq 2$, will switch from previous

$$\begin{cases} \sum_j d_j C(||s_i - s_j||) + \sum_{\ell} \sum_{\ell} f_\ell(s_i) = C(||s_i - s_0||) \\ \sum_i d_i f_\ell(s_i) = f_\ell(s_0), \quad \ell = 2 \dots L \end{cases}$$

$i = 1 \dots n$

covariance matrix of observation

Matrix form:

$$\begin{pmatrix} \sum_{\ell=1}^L F \\ F \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_0 \end{pmatrix}$$

F : design matrix of linear model $(*)$

$$F_{il} = f_l(s_i), \quad i = 1 \dots n, \quad \ell = 2 \dots L$$

f_0 design vector at $s_0: f_{0,\ell} = f_\ell(s_0)$

How to estimate our Covariogram?

o Recall. We know how to get $\hat{f}, \hat{c}, \hat{\sigma}^2$ under second order stationarity

- Not do: Using some estimators on z_{s_1}, \dots, z_{s_n}
- Idea: Estimate f, ς, Σ from $\hat{f}_{s_1}, \dots, \hat{f}_{s_n}$
 but we don't know $f_{s_1}, \dots, f_{s_n} \rightarrow$ so
 we can try estimate them, firstly
 estimating m_s

Estimating the diff

$$\text{Model } z_{s_i} = \sum_{\ell=0}^L \underline{\varphi}_\ell f_\ell(s_i) + \varepsilon_{s_i}, \quad i=1 \dots n$$

$$\underline{z} = \underline{F} \underline{\varphi} + \underline{\varepsilon}, \text{ with } \text{cov}(\underline{\varepsilon}) = \Sigma$$

We will use GLS (General Least square)
 (use when Σ is not diagonal)

$$\text{GLS: } \hat{\underline{\varphi}}_L = \underset{\underline{\varphi} \in \mathbb{R}^{L+1}}{\arg \min} (\underline{z} - \underline{F} \underline{\varphi})' \Sigma^{-1} (\underline{z} - \underline{F} \underline{\varphi})$$

$$= \underset{\underline{\varphi} \in \mathbb{R}^{L+1}}{\arg \min} \underline{\varphi}' \Sigma^{-1} \underline{\varphi}$$

$$\hat{\underline{\theta}} = (\underline{F}' \underline{\Sigma}^{-1} \underline{F})^{-1} \underline{F}' \underline{\Sigma}^{-1} \underline{z}$$

more general
 term been
 probably on
 last lectures
 of Linear
 models

Algorithm :

2.- Estimate $\underline{\theta}$ as $\hat{\underline{\theta}} = \hat{\underline{\theta}}^{\text{OLS}}$ knowing $\underline{\Sigma}$

1.- Get $\underline{\Sigma}$ from $\hat{\underline{f}} = \underline{z} - \underline{F}\hat{\underline{\theta}}$

→ Solve LK problem

so to know $\underline{\Sigma}$ we need $\hat{\underline{\theta}}$,

to know $\hat{\underline{\theta}}$ we need $\underline{\Sigma}$,

⇒ Iteratively solve

Q. Initialize : $\hat{\underline{\theta}}^{\text{OLS}} \rightarrow \hat{\underline{f}}^{\text{OLS}} \rightarrow \hat{\underline{\Sigma}}$



 POLITECNICO DI MILANO



Politecnico di Milano
Applied Statistics
May 2025

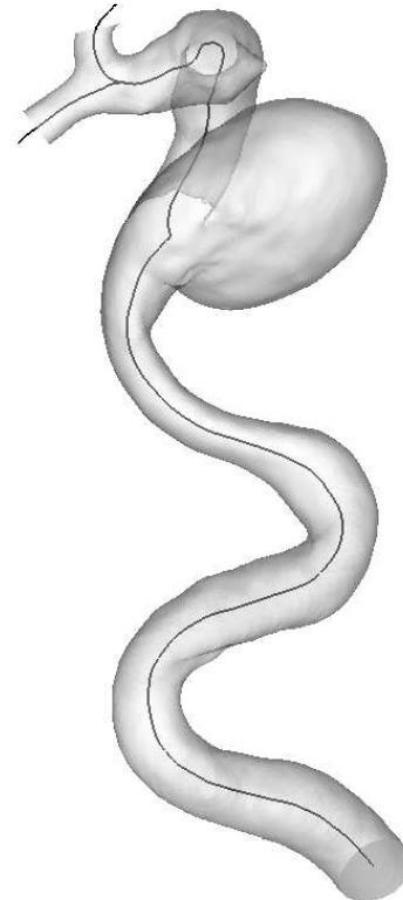


An introduction to functional data analysis

Laura M. SANGALLI

MOX - Dipartimento di Matematica, Politecnico di Milano

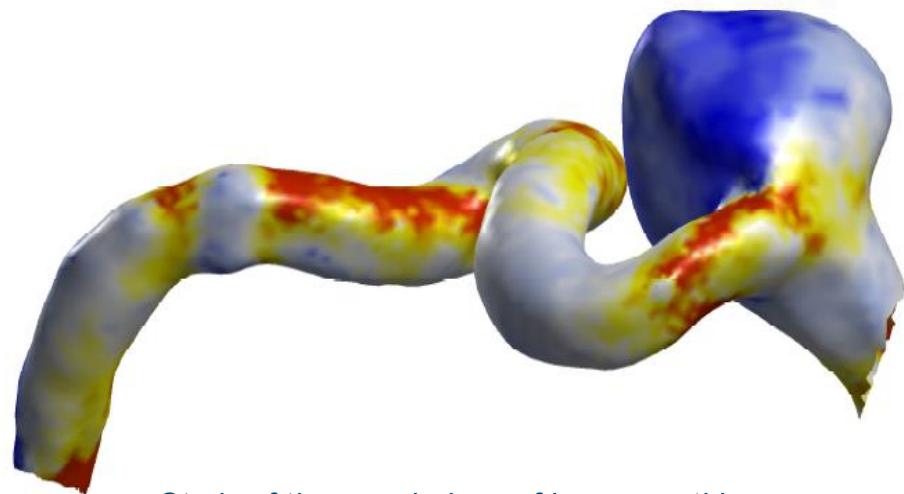
<https://sangalli.faculty.polimi.it/>



Explosive growth in recording **complex** and **high-dimensional** data, e.g., having a **functional nature** (i.e., representable by curves, surfaces, dynamic curves and surfaces), non-euclidean data

2D and 3D images and measures captured in time and space

- ▶ images of internal structures of a body and biological signals recorded by medical scanners



Study of the morphology of inner carotid arteries with aneurysms

AneuRisk project

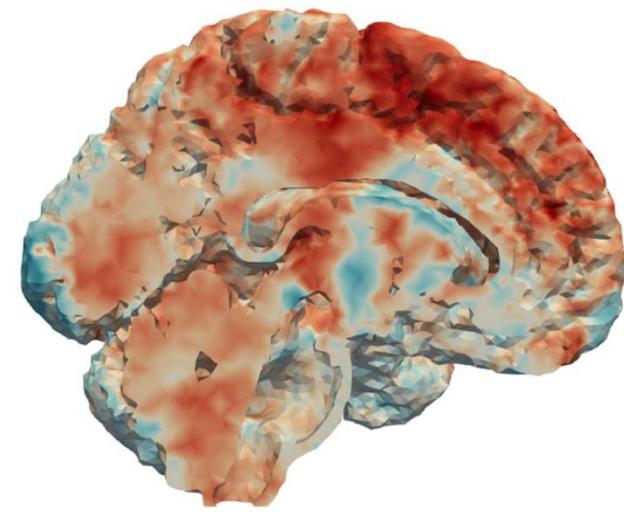
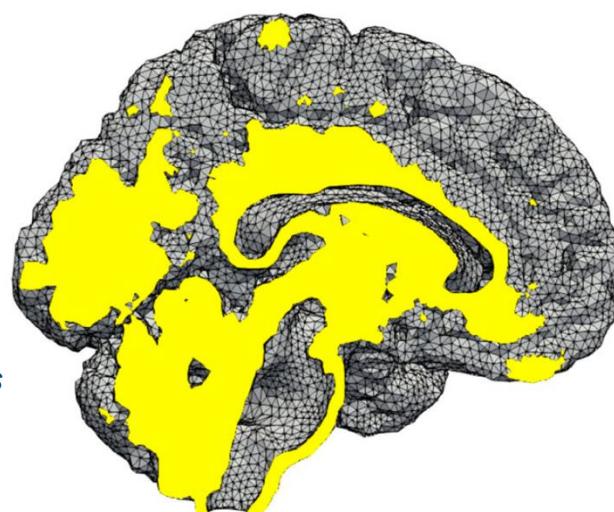
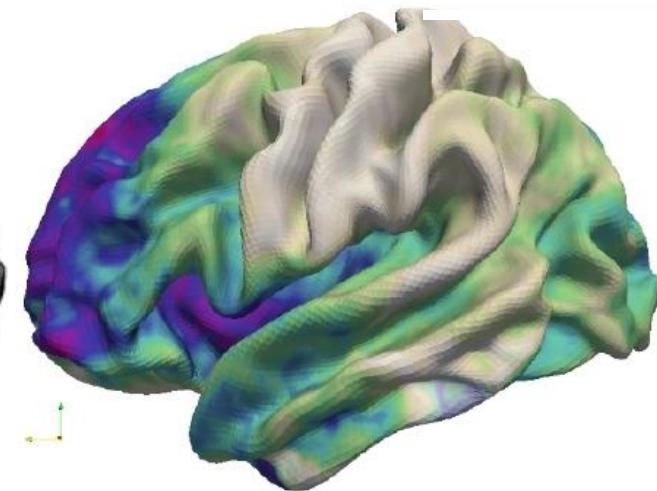
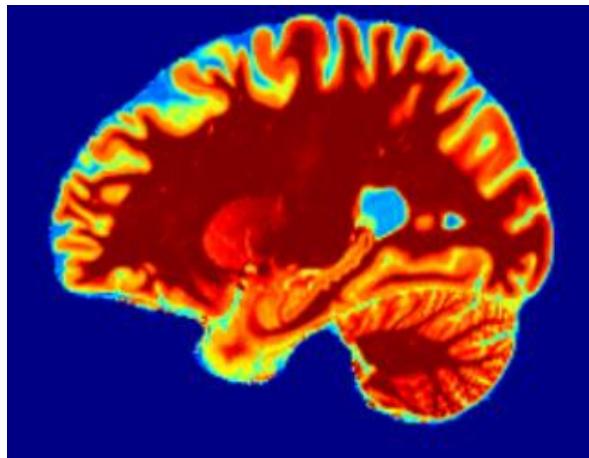
Sangalli, Secchi, Vantini, Veneziani (2009)
JASA and J. R. Stat. Soc. Ser. C

Study of the morphology of inner carotid arteries with aneurysms

AneuRisk project

Ettinger, Perotto, Sangalli (2016) *Biometrika*

- ▶ Study of neuroimaging signals over the cerebral cortex or in the grey matter



Math for Brain project

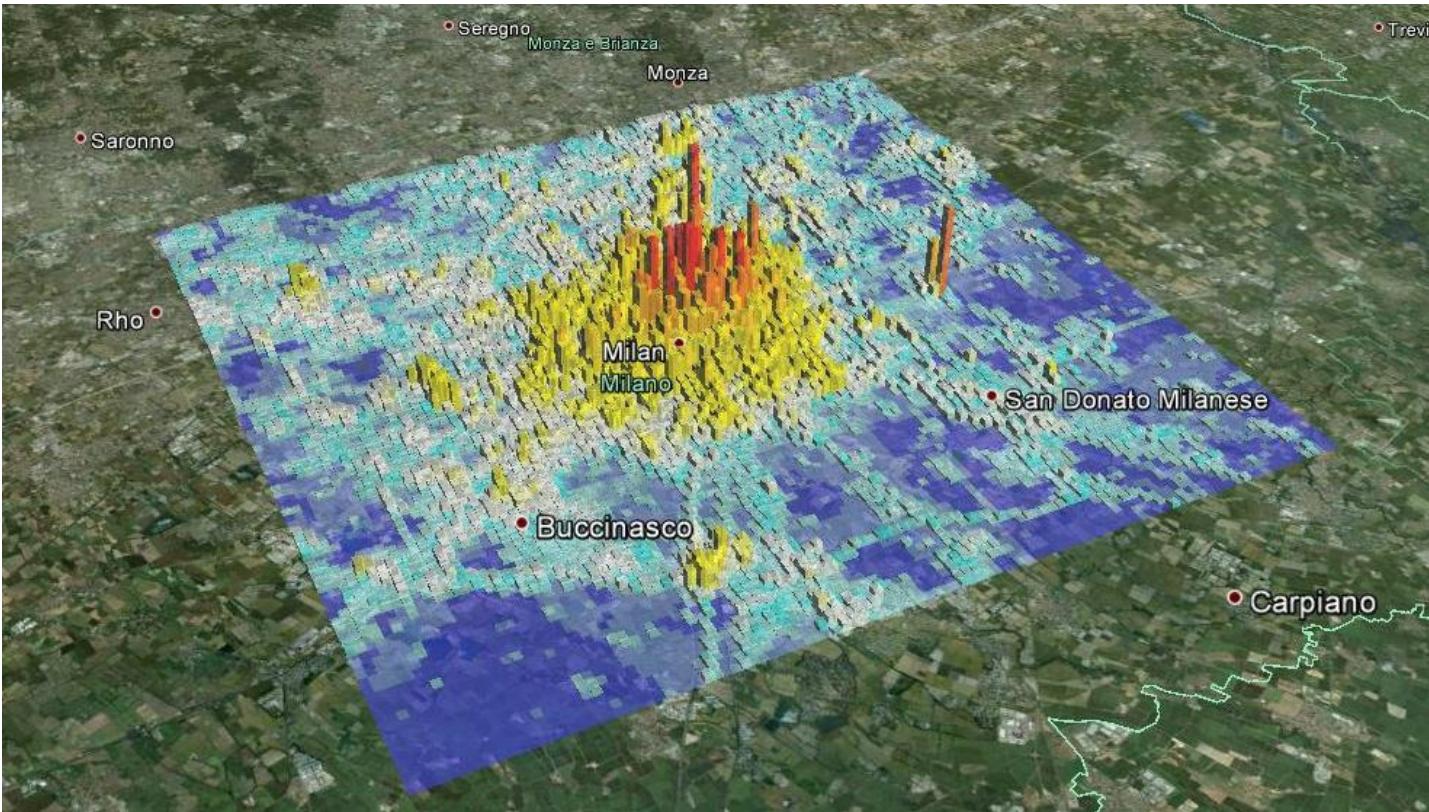
Lila, Aston, Sangalli (2016) AoAS
Arnone, Vicini, Sangalli (2023) Biometrics
Clementi et al. (2023) Plos One

Functional data: where they come from

4

mobile traffic of milan metropolitan

- Study of mobile phone data over the metropolitan area of Milan



RegioneLombardia

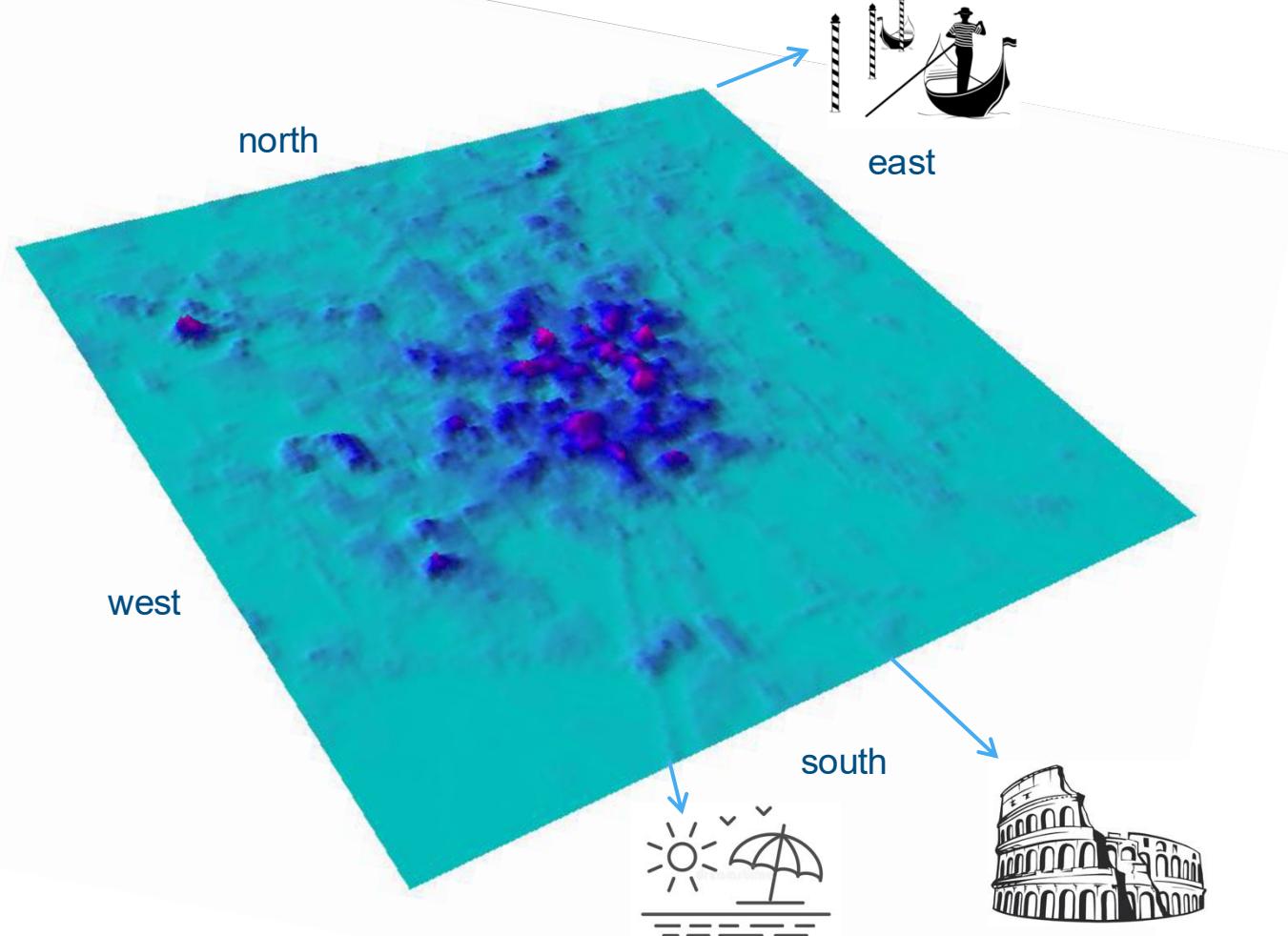
Erlang measures: every 15 mins (18/03/2009 - 24/03/2009) use of Telecom mobile phone network across Milano metropolitan area (13.8M records)

Secchi, Vantini, Vitelli (2015) Statistical Methods & Applications
Zanini, Shen, Truong (2016) The Annals of Applied Statistics

Functional data: where they come from

5

- ▶ Study of mobile phone data over the metropolitan area of Milan

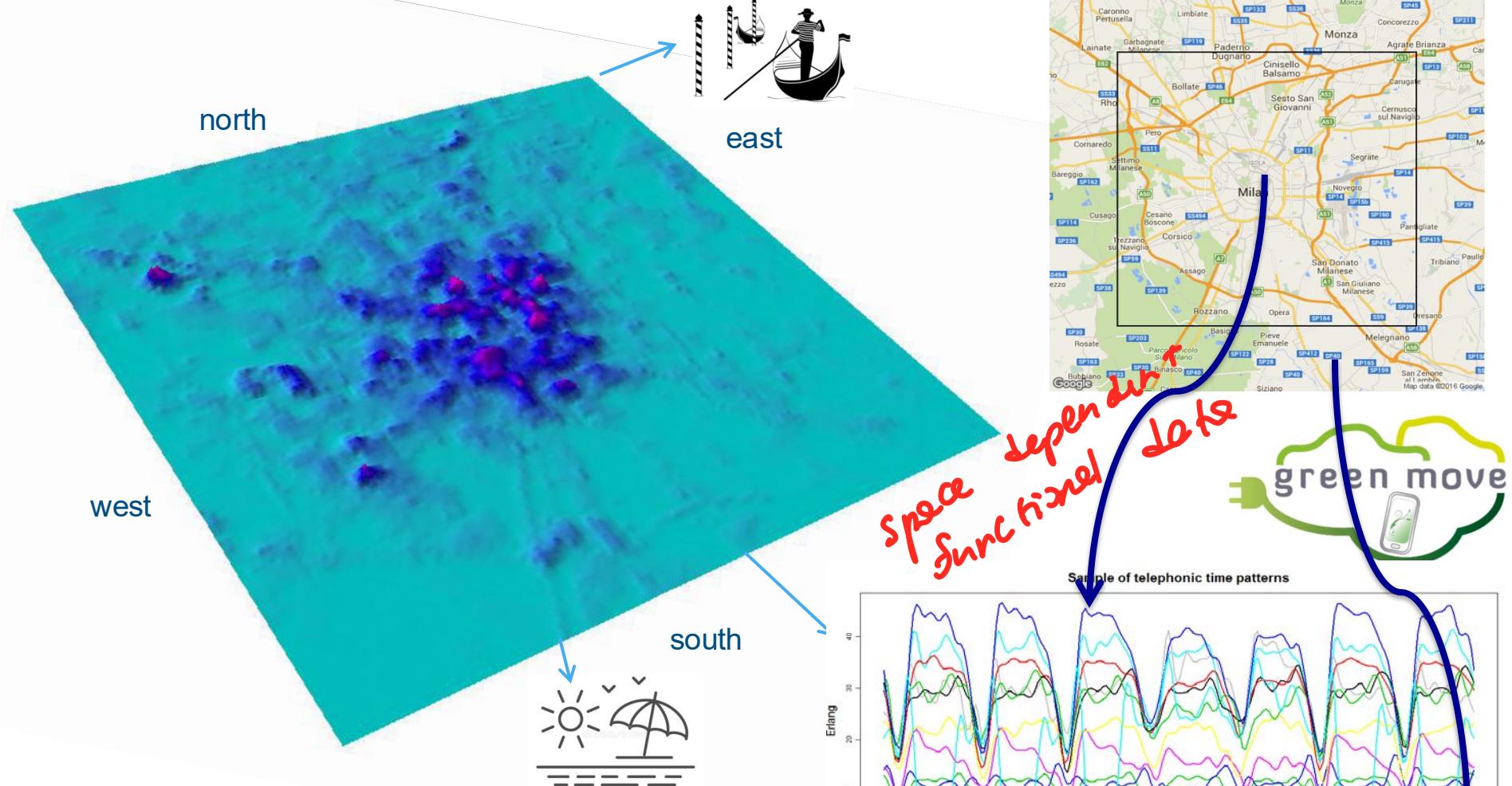


Secchi, Vantini, Vitelli (2015) Statistical Methods & Applications
Zanini, Shen, Truong (2016) The Annals of Applied Statistics

Functional data: where they come from

6

- ▶ Study of mobile phone data over the metropolitan area of Milan

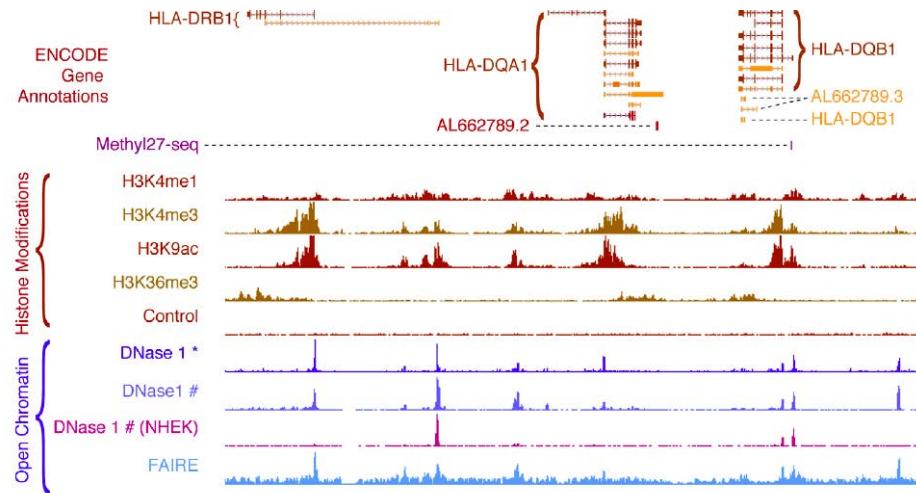


Secchi, Vantini, Vitelli (2015) Statistical Methods & Applications
Zanini, Shen, Truong (2016) The Annals of Applied Statistics

Functional data: where they come from

8

- measurements of gene expression levels via next generation sequencing data



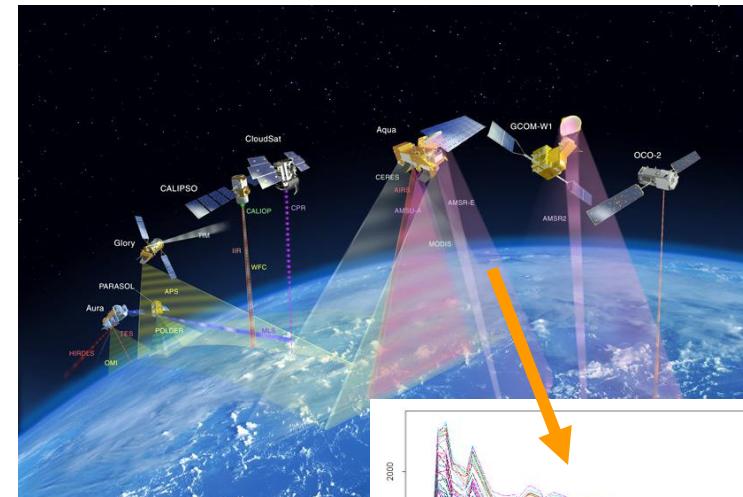
Epigenomic project

Cremona et al. (2015) BMC Bioinformatics

- images of steady or moving objects/individuals recorded by computer vision devices



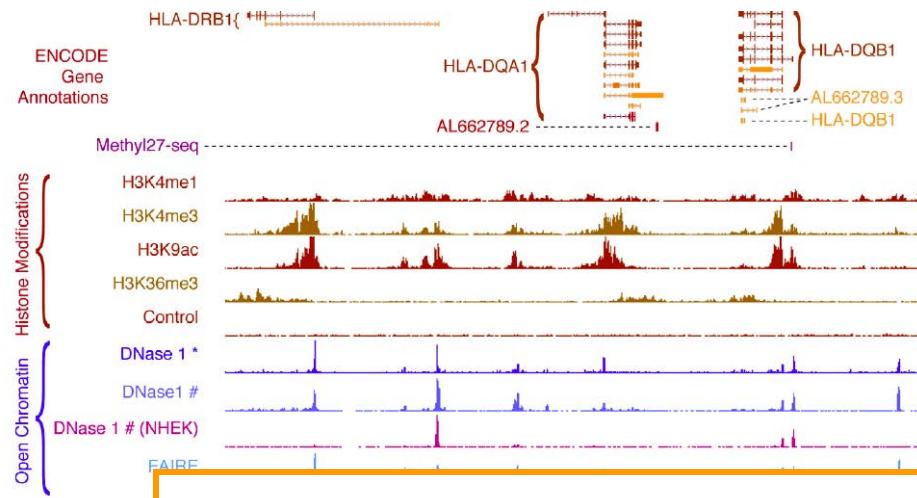
- multi-spectral data from satellite remote sensing



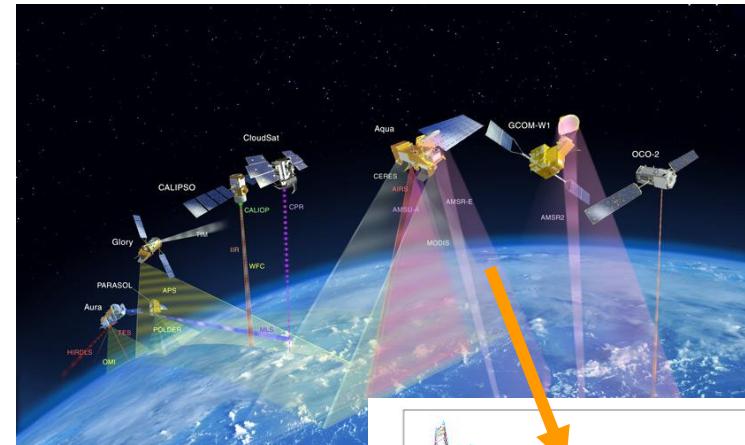
Functional data: where they come from

9

- measurements of gene expression levels via next generation sequencing data

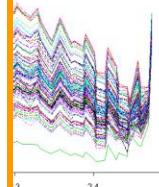
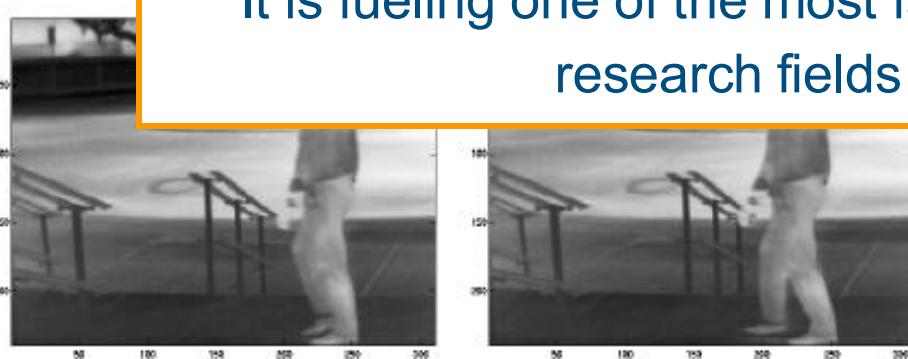


- multi-spectral data from satellite remote sensing



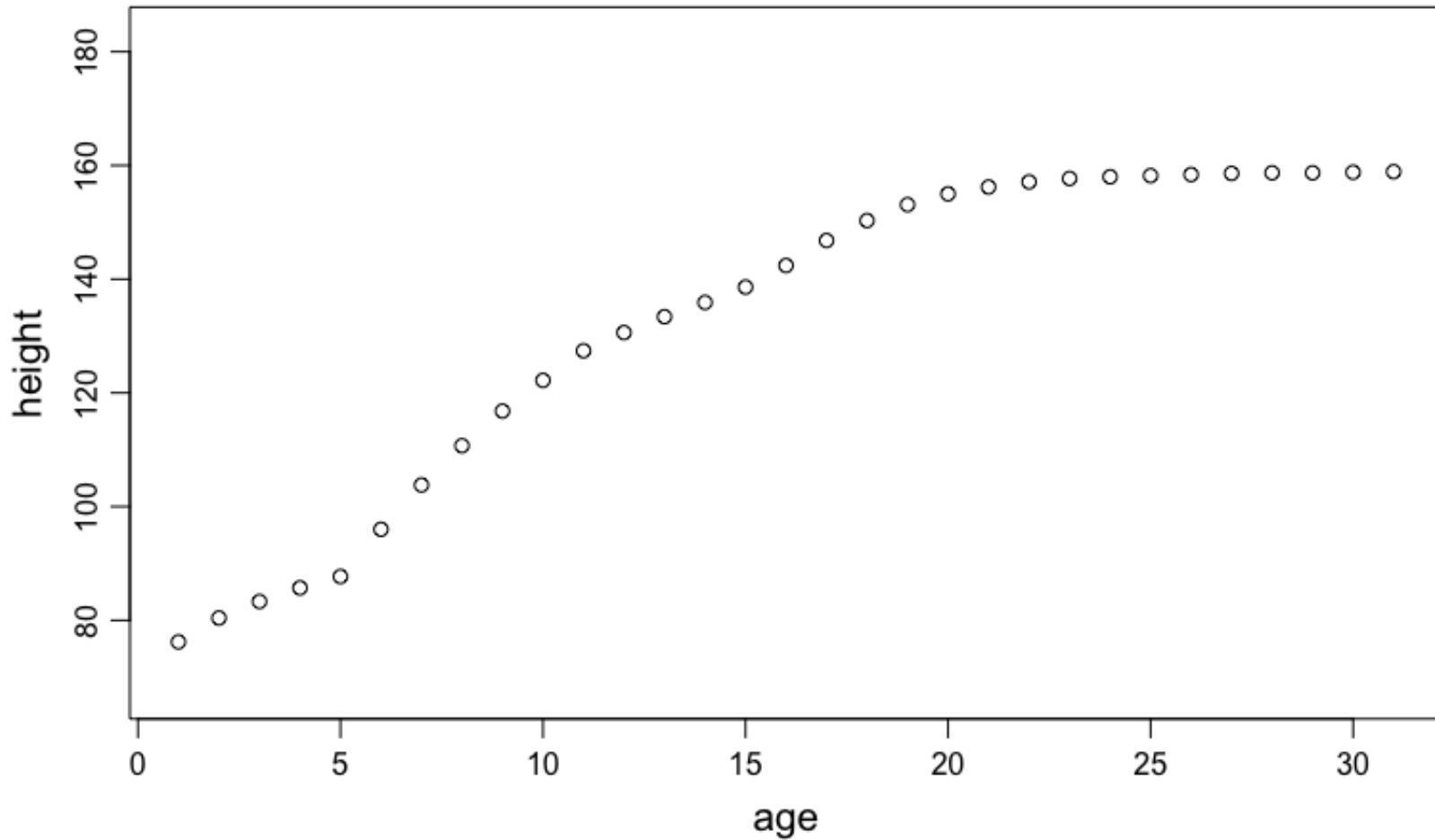
The analysis of complex and high dimensional data poses new and challenging problems in research

It is fueling one of the most fascinating and fastest growing research fields of modern statistics

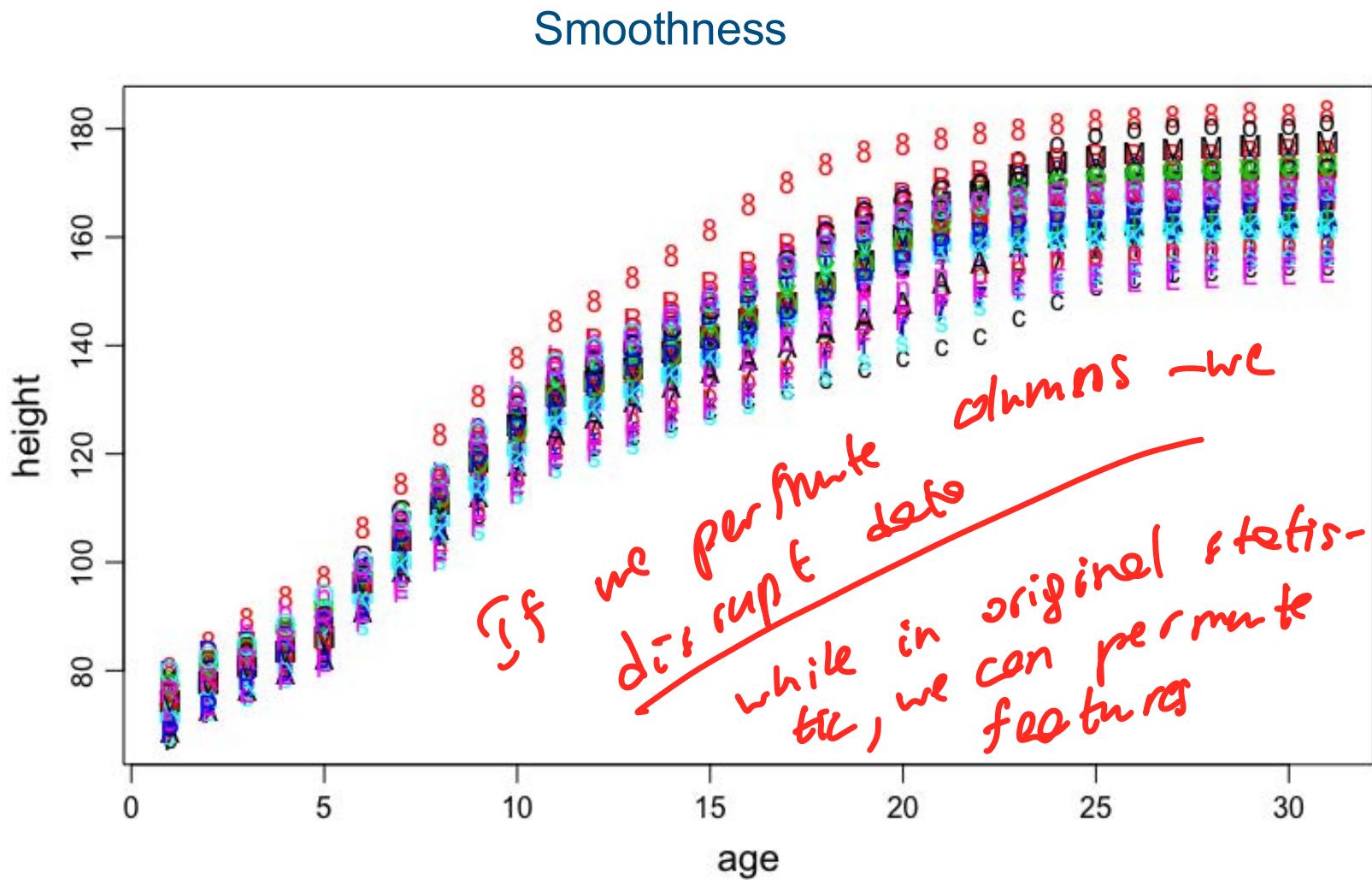


What characterizes functional data?

Smoothness



What characterizes functional data?



What characterizes functional data?

- **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

Molley Sangalli Timothy Tchetgen

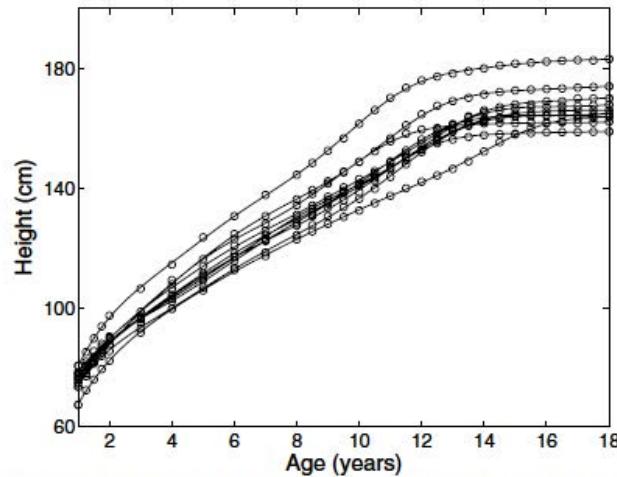


Figure 1.1. The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.

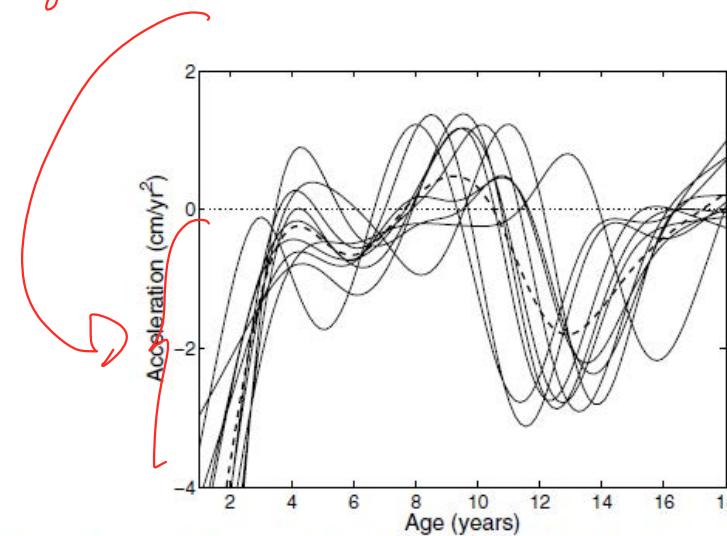


Figure 1.2. The estimated accelerations of height for 10 girls, measured in centimeters per year. The heavy dashed line is the cross-sectional mean, and is a rather poor summary of the curves.

Berkeley Growth Curves as functional data

- Data reflect **smooth** variation of height over time: $h(t)$
- Some interesting features are only visible if **derivatives** are analyzed (e.g., mid-spurt and pubertal growth spurt)
- The grid spacing on the **time axis** is non-uniform
- The function might have been observed on different time points for different individuals
- **Large p small n problems:** classical multivariate methods fail when the number of variables is larger than the sample size (in this case, $p=31$, $n=10$)

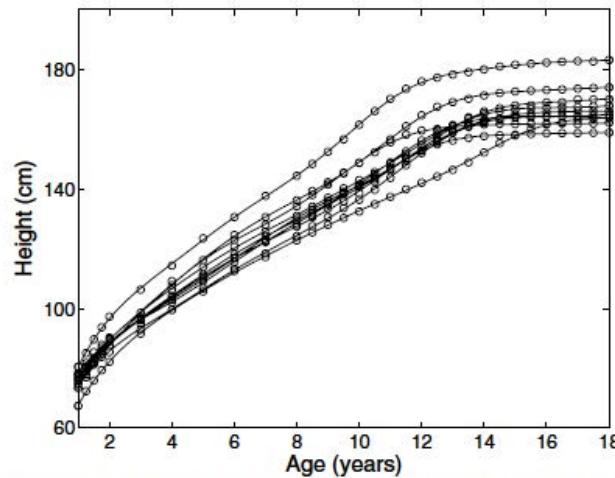


Figure 1.1. The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.

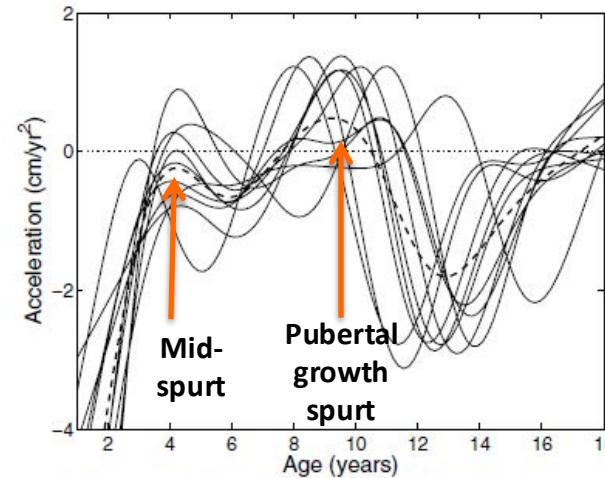


Figure 1.2. The estimated accelerations of height for 10 girls, measured in centimeters per year. The heavy dashed line is the cross-sectional mean, and is a rather poor summary of the curves.

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: CRAN Task View: Functional Data Analysis

- R package fda
- R package Refund
- R package fdapace
- R package fdaCluster (alignment and clustering)
- R package fdaPDE (functional data over complex multidimensional domains)
- Python package scikit-fda

1. Hilbert space model for functional data

- 1.1. Basics notions on Hilbert spaces
- 1.2. Hilbert space embedding for functional data
- 1.3. Formal definition of functional data

2. Smoothing of functional data

- 2.2. Least square smoothing
- 2.3. Smoothing with a differential penalization

3. Data alignment and clustering

- 3.1 Phase and amplitude variability
- 3.2 Landmark and continuous registration
- 3.3 Decoupling phase and amplitude variability
- 3.4 K-mean alignment

4. Dimensionality reduction

- 4.1. Functional Principal Components in Hilbert spaces
- 4.2. Examples in L₂

5. Functional data over complex multidimensional domains



POLITECNICO DI MILANO



Politecnico di Milano
Applied Statistics
May 2025



An introduction to functional data analysis

Part 1 - Hilbert space model for functional data

Laura M. SANGALLI

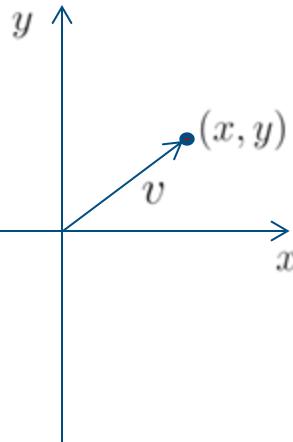
MOX - Dipartimento di Matematica, Politecnico di Milano

<https://sangalli.faculty.polimi.it/>

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)

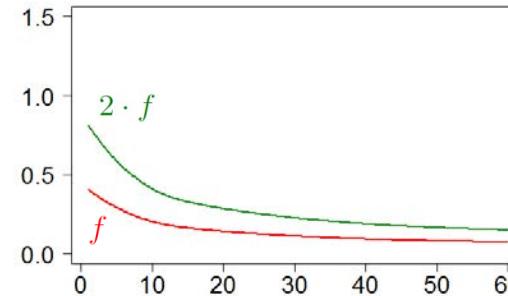
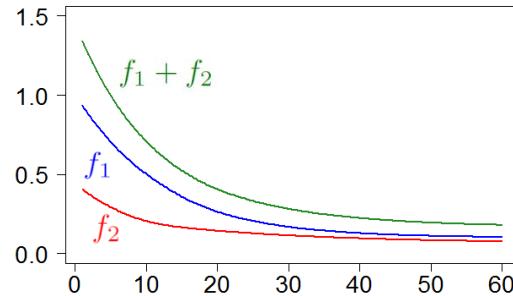
Euclidean space \mathbb{R}^2



- Sum: $v_1 + v_2 = (x_1 + x_2, y_1 + y_2)$
 - Product by a constant: $c \cdot v = (c \cdot x, c \cdot y)$
 - Norm (length 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: $\vartheta = \arccos \frac{\langle v_1, v_2 \rangle}{\|v_1\| \|v_2\|}$
- Operations (+, ·) 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 $(+, \cdot)$



- 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: $x = y$ if $\int [x(t) - y(t)]^2 dt = 0$



Basics notions on Hilbert spaces: a reminder

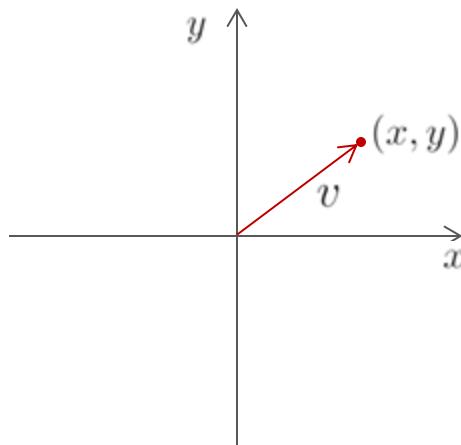
A Hilbert Space approach to the analysis of Functional Data

Courtesy of P. Secchi

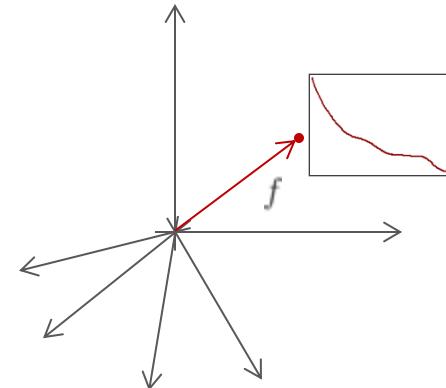
Embedding functional data in un appropriate Hilbert space enable us

- to understand functional data as **points of a space of functions**
- to uplift many methods of **multivariate statistics to functional data**, through the notions of inner product and norm

Multivariate statistics
(Euclidean space)



Functional Data Analysis
(Hilbert space)





Basics notions on Hilbert spaces: a reminder

A Hilbert Space approach to the analysis of Functional Data

Courtesy of P. Secchi

Let H be a linear space. An inner product on H is a bilinear, symmetric, positive definite form

$$\langle \cdot, \cdot \rangle : H \times H \rightarrow \mathbb{R}$$

that satisfies

- (i) $\langle \lambda x + y, z \rangle = \lambda \langle x, z \rangle + \langle y, z \rangle \quad \forall \lambda \in \mathbb{R}, \quad \forall x, y, z \in H$
- (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:

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 \text{ if and only if } \langle x, y \rangle = 0$$



Basics notions on Hilbert spaces: a reminder

A Hilbert Space approach to the analysis of Functional Data

Courtesy of P. Secchi

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

$$x = \sum_{n=1}^{\infty} \langle x, u_n \rangle u_n. \quad \text{Basis expansion}$$

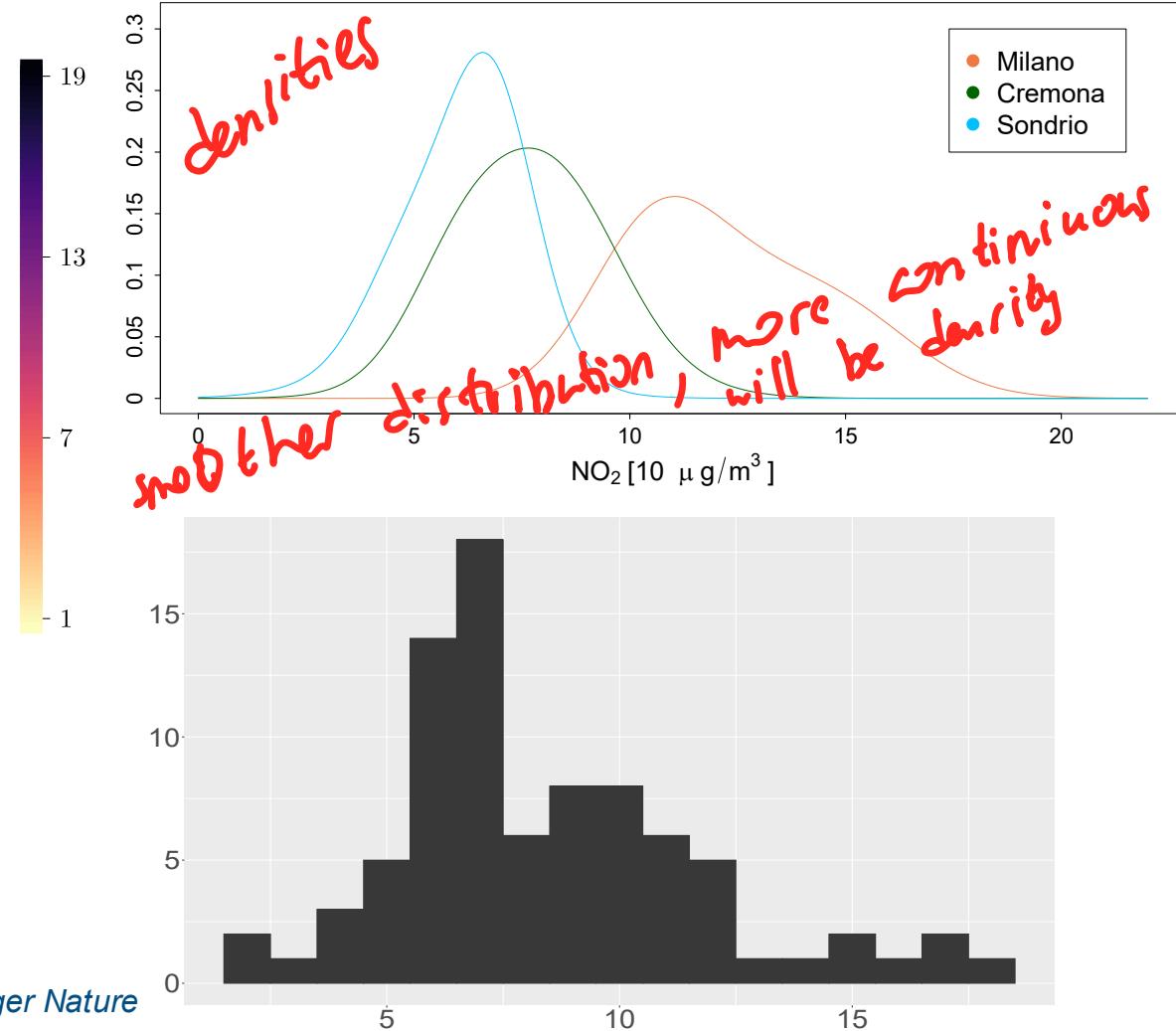
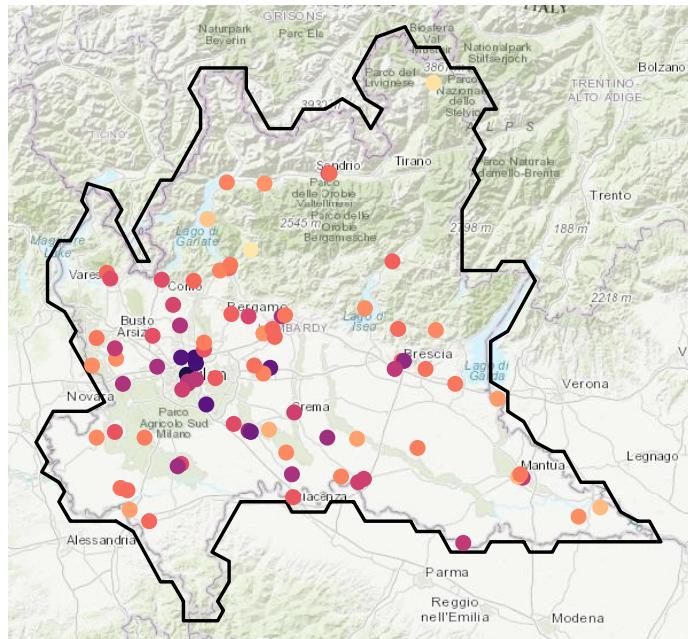
depending on b_k , we can choose different spaces

Constrained functional data densities

22



Funded by the
European Union
NextGenerationEU



ARPA LOMBARDIA
Agenzia Regionale per la Protezione dell'Ambiente

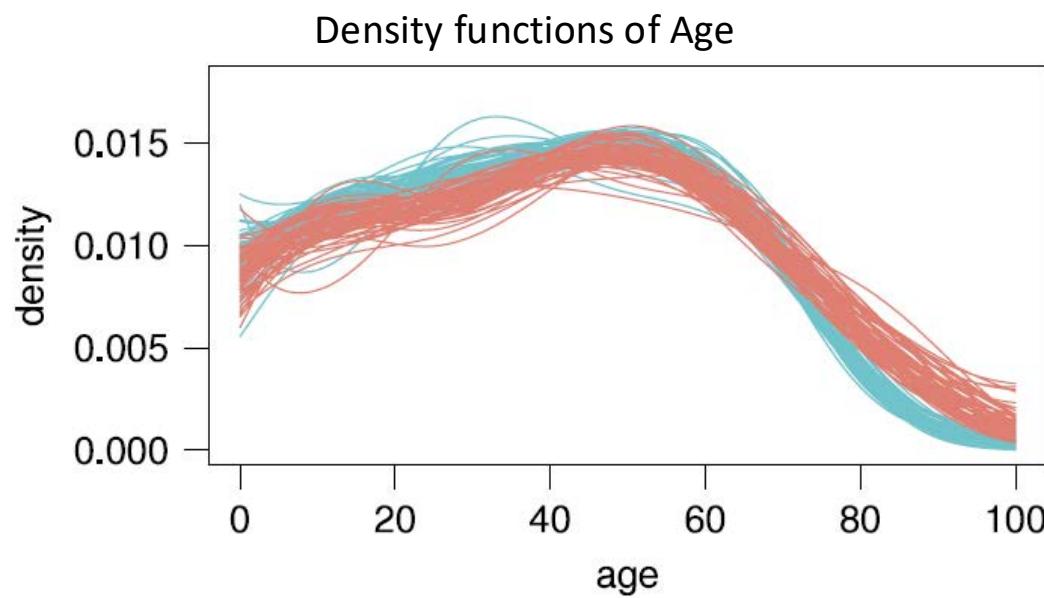
Study of air pollution in Lombardy

GRINS project & CoEnv project

De Sanctis, Di Battista, et al. (2025) EES Springer Nature



DATA SCIENCE
AND STATISTICS

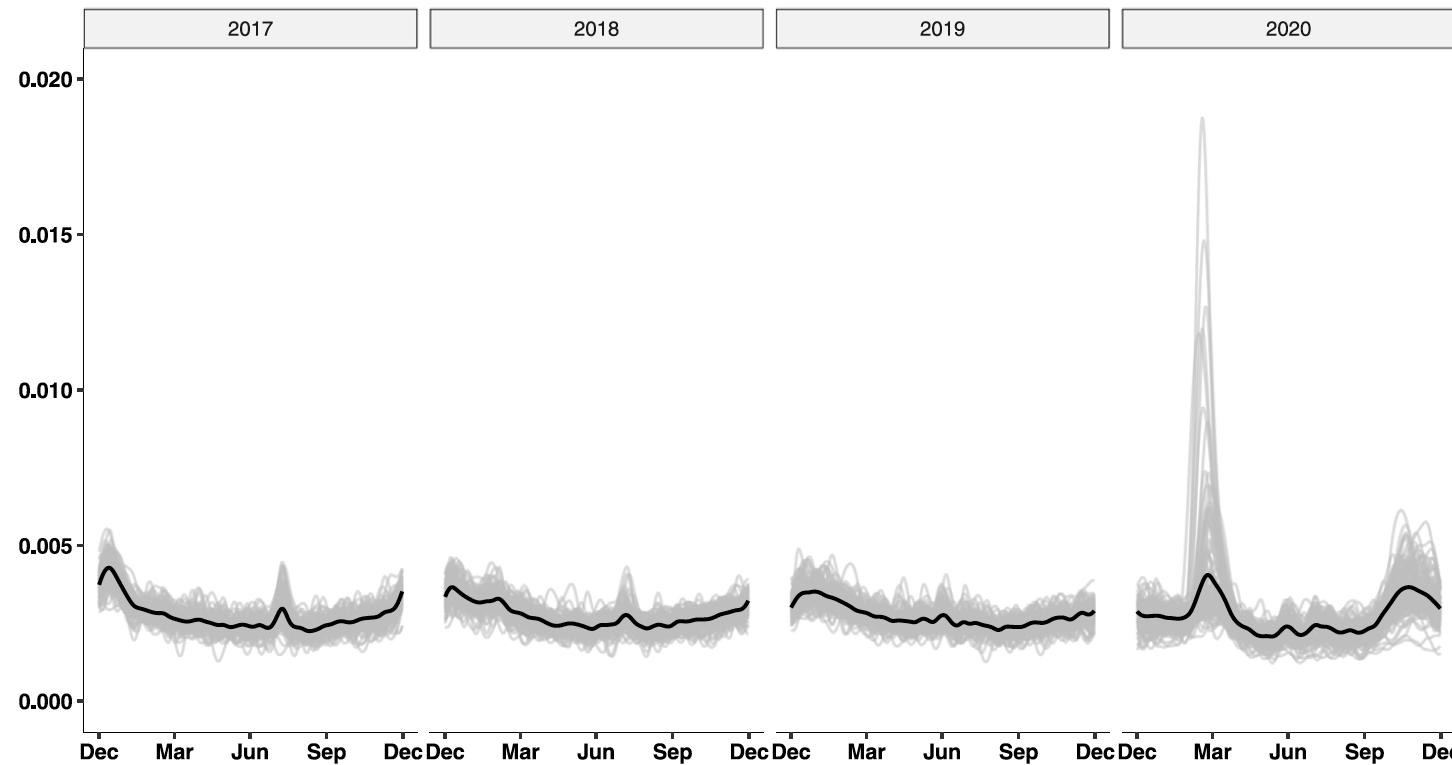


Hron et al. 2015, CSDA

che p'noct'

Smooth estimates of the mortality densities, 70+ years

Provinces



Scimone, Menafoglio, Sangalli, Secchi (2022) *Spatial Statistics*



Basics notions on Hilbert spaces: a reminder

A Hilbert Space approach to the analysis of Functional Data

Courtesy of P. Secchi

B^2 : space of density functions on a closed interval I , with \log in L^2

'Bors Space'

- Equivalence relation: f, g are equivalent if they are proportional (scale invariance)

- Sum (perturbation): $(f \oplus g)(t) = \frac{f(t)g(t)}{\int_I f(s)g(s) ds}$

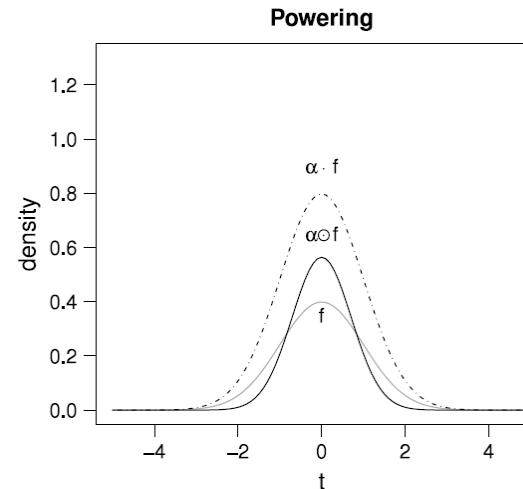
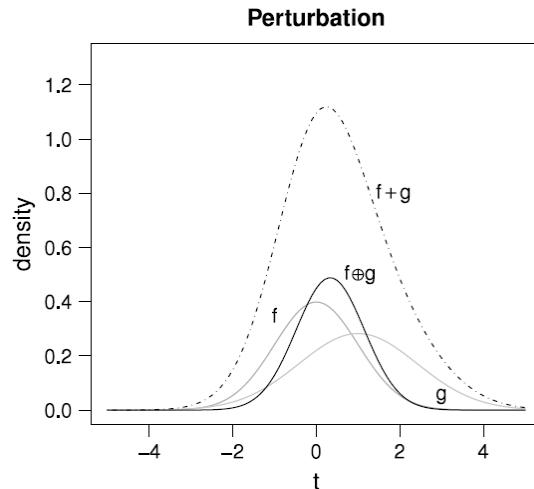
- Product by a constant (powering): $(\alpha \odot f)(t) = \frac{f(t)^\alpha}{\int_I f(s)^\alpha ds}, \quad t \in I$.

- Inner product: $\langle f, g \rangle_B = \frac{1}{2\eta} \int_I \int_I \ln \frac{f(t)}{f(s)} \ln \frac{g(t)}{g(s)} dt ds$

- Norm: $\|f\|_B = \left[\frac{1}{2\eta} \int_I \int_I \ln^2 \frac{f(t)}{f(s)} dt ds \right]^{1/2}$

functions
constrained to

their densities



Note: the geometry of L^2 doesn't make sense for density functions

Basics notions on Hilbert spaces: a reminder

A Hilbert Space approach to the analysis of Functional Data

Courtesy of P. Secchi

B²: space of density functions on a closed interval I, with log in L²

- Equivalence relation: f, g are equivalent if they are proportional (*scale invariance*)
- Sum (perturbation): $(f \oplus g)(t) = \frac{f(t)g(t)}{\int_I f(s)g(s) ds}$.
- Product by a constant (powering): $(\alpha \odot f)(t) = \frac{f(t)^\alpha}{\int_I f(s)^\alpha ds}, \quad t \in I.$
- Inner product: $\langle f, g \rangle_{\mathcal{B}} = \frac{1}{2\eta} \int_I \int_I \ln \frac{f(t)}{f(s)} \ln \frac{g(t)}{g(s)} dt ds$
- Norm: $\|f\|_{\mathcal{B}} = \left[\frac{1}{2\eta} \int_I \int_I \ln^2 \frac{f(t)}{f(s)} dt ds \right]^{1/2}$
- B^2 is isomorphic to L^2 (in fact, all the Hilbert spaces are isomorphic). An isometric isomorphism is provided, e.g., by the **centred log-ratio transformation**

$$\text{clr}(f)(t) = f_c(t) = \ln f(t) - \frac{1}{\eta} \int_I \ln f(s) ds.$$

Exercise: prove that

$$\text{clr}(f \oplus g)(t) = f_c(t) + g_c(t), \quad \text{clr}(\alpha \odot f)(t) = \alpha \cdot f_c(t), \quad \langle f, g \rangle_{\mathcal{B}} = \langle f_c, g_c \rangle_2 = \int_I f_c(t)g_c(t) dt.$$



- First step in fda:
choose appropriate embedding for the data
- **Separable Hilbert spaces are a convenient choice** (projections, best approximations)
- **Note:** Not all the interesting spaces are Hilbert: e.g., the space of continuous functions is not a Hilbert space. Other interesting spaces: Riemannian manifolds (OODA)
- Examples of Hilbert spaces for FDA:
 - L^2 , space of square integrable functions: OK for most data analyses (especially if data are unconstrained)
 - H^2 , Sobolev space of functions that L^2 and whose derivative up to the second order are also in L^2
 - B^2 , space of functional compositions: useful for density functions



Formal definition of functional data

Functional random variables and functional data

Courtesy of P. Secchi

- 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

$$X : \Omega \rightarrow H$$

Definition 2:

A **functional datum** x is a realization of a functional random variable, i.e., for $\omega \in \Omega$,

$$x = X(\omega) : T = [t_{min}, t_{max}] \rightarrow \mathbb{R}$$

Definition 3:

A **functional dataset** is a collection of functional data.



Formal definition of functional data

Functional random variables and functional data

Courtesy of P. Secchi

Let $X : \Omega \rightarrow H$ 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

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

in approx private
hilbert space

In $H=L^2$, this is the point-wise sample mean

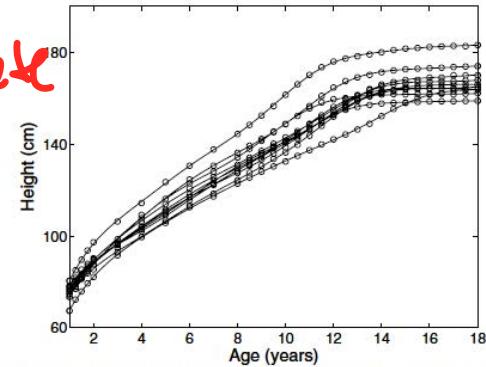


Figure 1.1. The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement.



Formal definition of functional data

Functional random variables and functional data

Courtesy of P. Secchi

Let $X : \Omega \rightarrow H$ 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

$$Cx = \mathbb{E}[\langle X, x \rangle X], \quad x \in H$$

projection

- If $H=L^2$ the covariance operator can be equivalently defined through a kernel operator

$$[Cx](t) = \int_T c(s, t)x(s)d(s), \quad x \in L^2$$

covariance kernel

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



Formal definition of functional data

Functional random variables and functional data

Courtesy of P. Secchi

Let $X : \Omega \rightarrow H$ 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

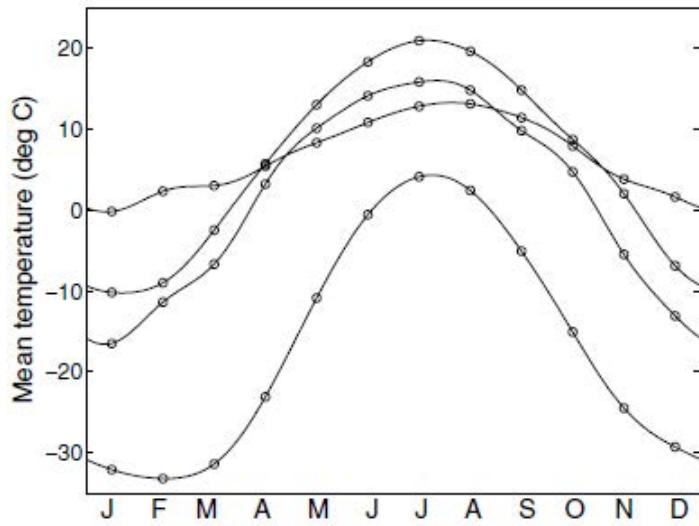
$$Sx = \frac{1}{N} \sum_{i=1}^N \langle X_i, x \rangle X_i, \quad x \in H$$

- If $H=L^2$, one can use the alternative definition

$$[Sx](t) = \int_T \widehat{c}(s, t)x(s)d(s), \quad x \in L^2 \qquad \widehat{c}(s, t) = \frac{1}{N} \sum_{i=1}^N X(s)X(t)$$

Another simple example of functional data

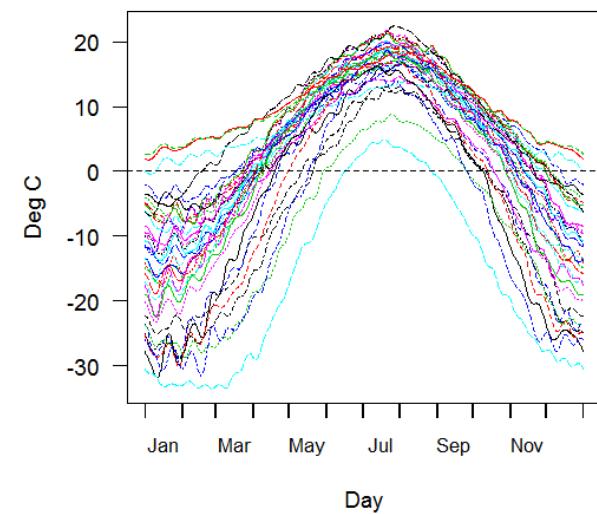
Example: Temperature curves in four locations in Canada (periodic data)



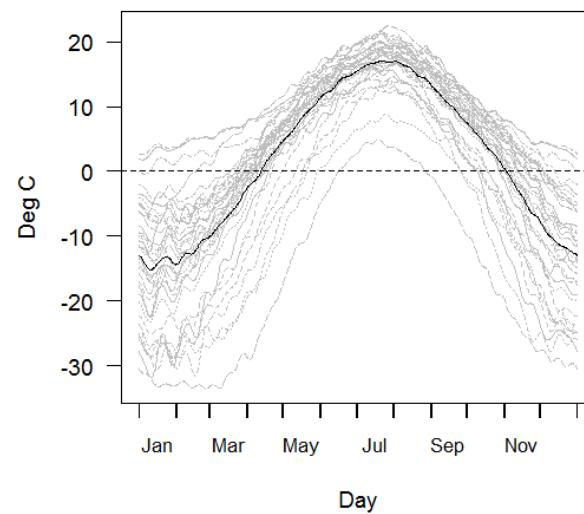
Another simple example of functional data

Courtesy of P. Secchi

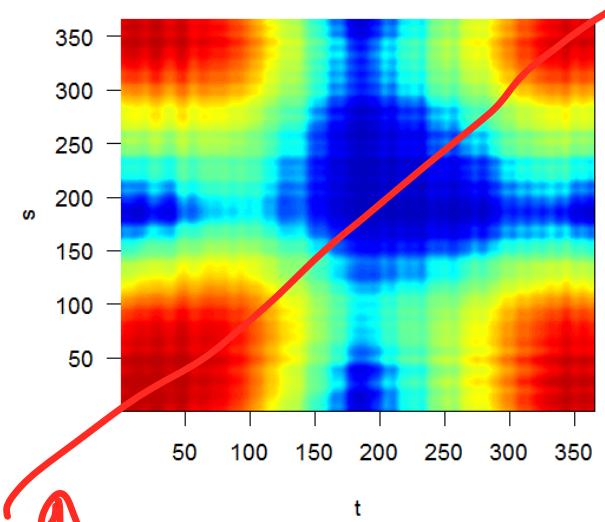
Example: Temperature curves in four locations in Canada (periodic data)



Functional dataset



Sample mean



Sample covariance kernel

main diagonal



POLITECNICO DI MILANO



Politecnico di Milano
Applied Statistics
May 2025



An introduction to functional data analysis

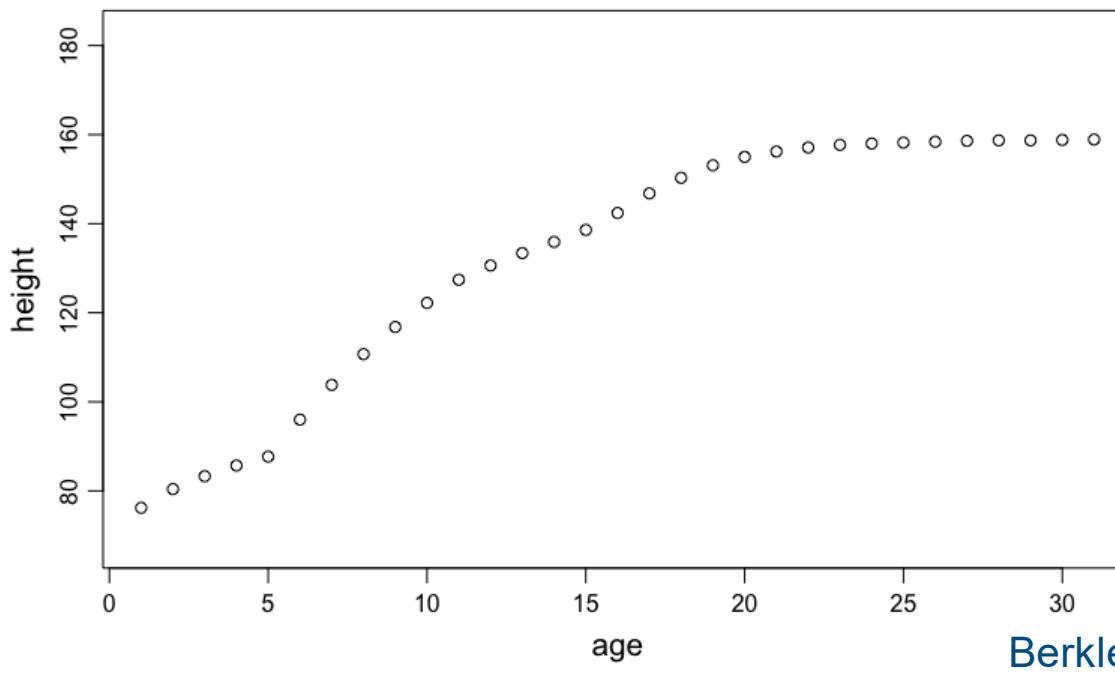
Part 2 - Smoothing functional data

Laura M. SANGALLI

MOX - Dipartimento di Matematica, Politecnico di Milano

<https://sangalli.faculty.polimi.it/>

Smoothing



Noisy and discrete data → functional representations

Smoothing - curve fitting

Chapters 3, 4, 5, 6 of Ramsay and Silverman (2005), *Functional Data Analysis*, Springer

Smoothing

Lessons for the master course in Applied Statistics

May 2025

Laura Sangalli

MOX laboratory – Department of Mathematics - Politecnico di Milano

Email: laura.sangalli@polimi.it

Web: <https://sangalli.faculty.polimi.it>

Office: Building 14th, La Nave, 7th floor



POLITECNICO
MILANO 1863

SMOOTHING

- $Y \in \mathcal{F}$: functional variable

Y-random function

Smoothing model / nonparametric regression model

$$Y = f(t) + \epsilon$$



POLITECNICO MILANO 1863

SMOOTHING

- $Y \in \mathcal{F}$: functional variable

Smoothing model / nonparametric regression model

$$Y_i = f(t_i) + \epsilon_i$$

Observations fixed

Sample: $\{(Y_1, t_1), \dots, (Y_n, t_n)\}$



POLITECNICO MILANO 1863

SMOOTHING

- $Y \in \mathcal{F}$: functional variable

Smoothing model / nonparametric regression model

$$Y_i = f(t_i) + \epsilon_i$$

Sample: $\{(Y_1, t_1), \dots, (Y_n, t_n)\}$

Work conditionally on t (consider covariates fixed)



POLITECNICO MILANO 1863

SMOOTHING

- $Y \in \mathcal{F}$: functional variable

homoscedasticity

Smoothing model / nonparametric regression model ✓

$$Y_i = f(t_i) + \epsilon_i \quad \epsilon_i \sim \text{indep} \quad \mathbb{E}[\epsilon_i] = 0, \text{Var}[\epsilon_i] = \sigma^2 < 0$$

Sample: $\{(Y_1, t_1), \dots, (Y_n, t_n)\}$

Work conditionally on t (consider covariates fixed)



POLITECNICO MILANO 1863

SMOOTHING

- $Y \in \mathcal{F}$: functional variable

Smoothing model / nonparametric regression model

$$Y_i = f(t_i) + \epsilon_i \quad \epsilon_i \sim \text{indep} \quad \mathbb{E}[\epsilon_i] = 0, \text{Var}[\epsilon_i] = \sigma^2 < 0$$

Sample: $\{(Y_1, t_1), \dots, (Y_n, t_n)\}$

Work conditionally on t (consider covariates fixed)

Estimation problem:

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \text{RSS}(f) = \operatorname{argmin}_{f \in \mathcal{F}} \left\{ \sum_{i=1}^n (Y_i - f(t_i))^2 \right\}$$



POLITECNICO MILANO 1863

SMOOTHING

- $Y \in \mathcal{F}$: functional variable

Smoothing model / nonparametric regression model

$$Y_i = f(t_i) + \epsilon_i \quad \epsilon_i \sim \text{indep} \quad \mathbb{E}[\epsilon_i] = 0, \text{Var}[\epsilon_i] = \sigma^2 < 0$$

Sample: $\{(Y_1, t_1), \dots, (Y_n, t_n)\}$

Work conditionally on t (consider covariates fixed)

Estimation problem:

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \text{RSS}(f) = \operatorname{argmin}_{f \in \mathcal{F}} \left\{ \sum_{i=1}^n (Y_i - f(t_i))^2 \right\}$$

*we can find
f in F
minimising*

*REVERSE PROBLEM
ILL POSED!*



POLITECNICO MILANO 1863

SMOOTHING

- $Y \in \mathcal{F}$: functional variable

Smoothing model / nonparametric regression model

$$Y_i = f(t_i) + \epsilon_i \quad \epsilon_i \sim \text{indep} \quad \mathbb{E}[\epsilon_i] = 0, \text{Var}[\epsilon_i] = \sigma^2 < 0$$

Sample: $\{(Y_1, t_1), \dots, (Y_n, t_n)\}$

Work conditionally on t (consider covariates fixed)

Estimation problem:

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}_K} \text{RSS}(f) = \operatorname{argmin}_{f \in \mathcal{F}_K} \left\{ \sum_{i=1}^n (Y_i - f(t_i))^2 \right\}$$

1st approach: restrict search to \mathcal{F}_K , with $\dim(\mathcal{F}_K) = K \ll n$

reduce dimension to $\tilde{\mathcal{F}}_K$ (structured models)

POLITECNICO MILANO 1863

like in linear models
we decided, that our
model is linear



SMOOTHING

- $Y \in \mathcal{F}$: functional variable

if functional data
 $p \rightarrow \infty$ (many features)

Smoothing model / nonparametric regression model

$$Y_i = f(t_i) + \epsilon_i \quad \epsilon_i \sim \text{indep} \quad \mathbb{E}[\epsilon_i] = 0, \text{Var}[\epsilon_i] = \sigma^2 < 0$$

Sample: $\{(Y_1, t_1), \dots, (Y_n, t_n)\}$

Work conditionally on t (consider covariates fixed)

Estimation problem:

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \text{RSS}(f) = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (Y_i - f(t_i))^2 + P(f) \right\}$$

2nd approach: do not restrict \mathcal{F} but add a roughness penalty
(some kind of regularisation)



POLITECNICO MILANO 1863

SMOOTHING

Estimation functional

1st approach: \mathcal{F} approximated by \mathcal{F}_K , with $\dim(\mathcal{F}_K) = K \ll n$

$$\hat{f} = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \text{RSS}(f) = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - f(t_i))^2$$



POLITECNICO MILANO 1863

SMOOTHING

Estimation functional

1st approach: \mathcal{F} approximated by \mathcal{F}_K , with $\dim(\mathcal{F}_K) = K \ll n$

$$\hat{f} = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \text{RSS}(f) = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - f(t_i))^2$$

- ψ_1, \dots, ψ_K : K basis functions s.t. $\mathcal{F}_K = \text{span}(\psi_1, \dots, \psi_K)$



POLITECNICO MILANO 1863

SMOOTHING

Estimation functional

1st approach: \mathcal{F} approximated by \mathcal{F}_K , with $\dim(\mathcal{F}_K) = K \ll n$

$$\hat{f} = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \text{RSS}(f) = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - f(t_i))^2$$

- ψ_1, \dots, ψ_K : K basis functions s.t. $\mathcal{F}_K = \text{span}(\psi_1, \dots, \psi_K)$
- Smoothing model

$$Y_i = f(t_i) + \epsilon_i$$



POLITECNICO MILANO 1863

SMOOTHING

Estimation functional

1st approach: \mathcal{F} approximated by \mathcal{F}_K , with $\dim(\mathcal{F}_K) = K \ll n$

$$\hat{f} = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \text{RSS}(f) = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - f(t_i))^2$$

► ψ_1, \dots, ψ_K : K basis functions s.t. $\mathcal{F}_K = \text{span}(\psi_1, \dots, \psi_K)$

Smoothing model

approximate our model
with basis functions

$$Y_i = f(t_i) + \epsilon_i \quad \rightsquigarrow \quad Y_i = \sum_{j=1}^K c_j \psi_j(t_i) + \epsilon_i$$



POLITECNICO MILANO 1863

SMOOTHING

Estimation functional

1st approach: \mathcal{F} approximated by \mathcal{F}_K , with $\dim(\mathcal{F}_K) = K \ll n$

$$\begin{aligned}\hat{f} &= \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \text{RSS}(f) = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - f(t_i))^2 \quad \text{so now it is linear regression}\\ &= \underset{\mathbf{c} \in \mathbb{R}^K}{\operatorname{argmin}} \text{RSS}(\mathbf{c}) = \underset{\mathbf{c} \in \mathbb{R}^K}{\operatorname{argmin}} \sum_{i=1}^n \left(Y_i - \sum_{j=1}^K c_j \psi_j(t_i) \right)^2\end{aligned}$$

► ψ_1, \dots, ψ_K : K basis functions s.t. $\mathcal{F}_K = \text{span}(\psi_1, \dots, \psi_K)$

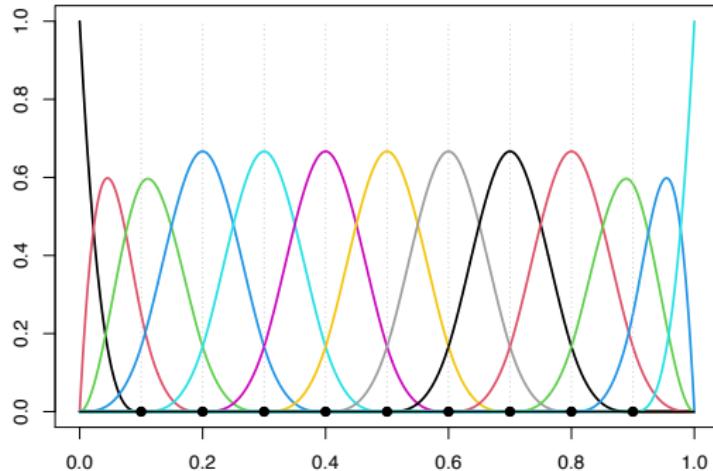
Smoothing model

$$Y_i = f(t_i) + \epsilon_i \quad \rightsquigarrow \quad Y_i = \sum_{j=1}^K c_j \psi_j(t_i) + \epsilon_i$$



POLITECNICO MILANO 1863

B-SPLINE BASIS

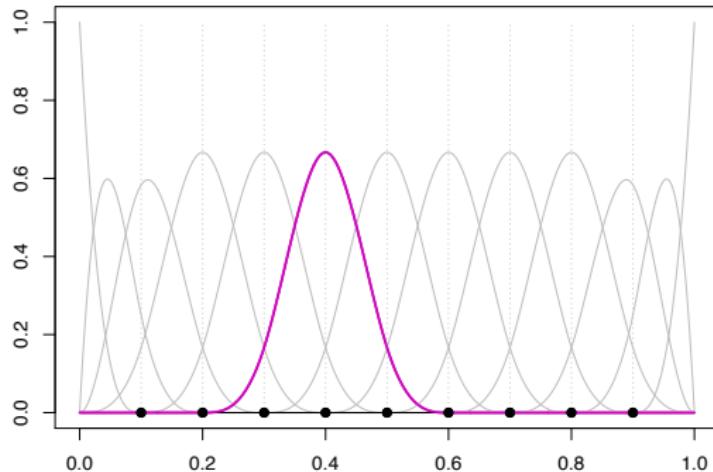


Cubic B-spline basis



POLITECNICO MILANO 1863

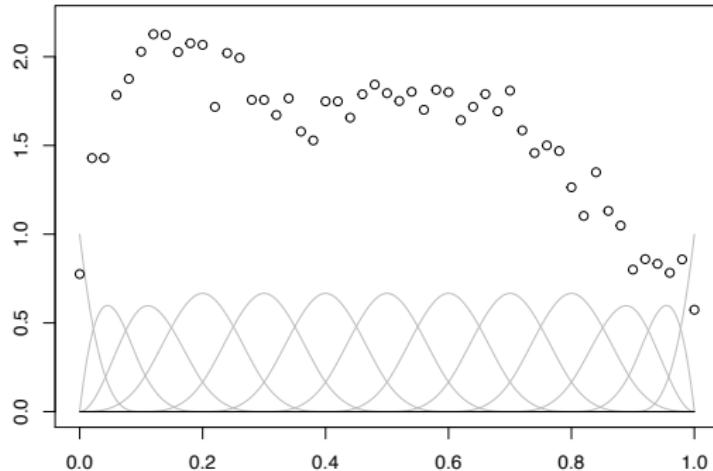
B-SPLINE BASIS



Computationally efficient: local support, band limited structure of key matrices involved, etc



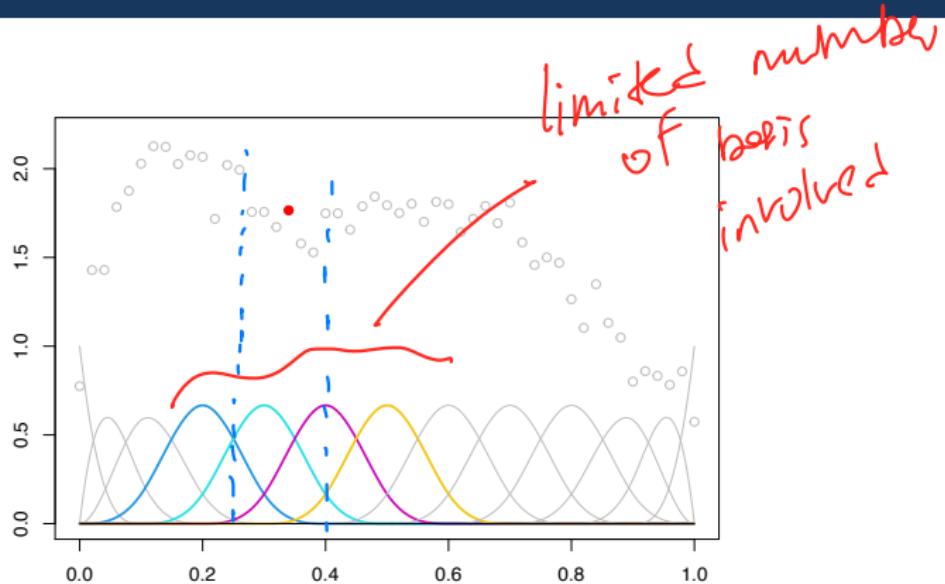
B-SPLINE BASIS



Computationally efficient: local support, band limited structure of key matrices involved, etc



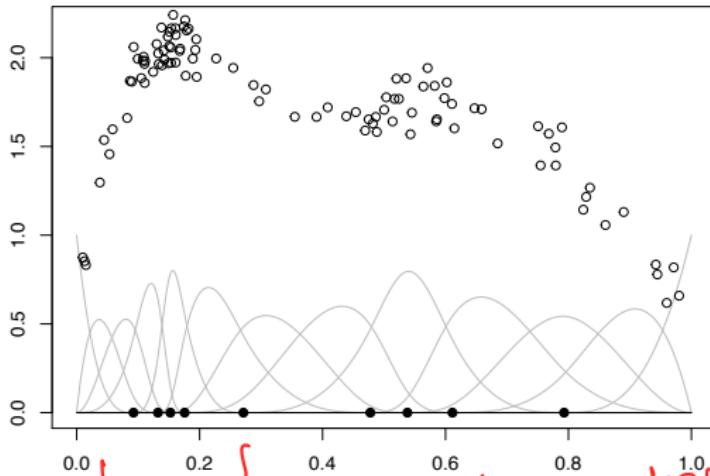
B-SPLINE BASIS



Computationally efficient: local support, band limited structure of key matrices involved, etc



B-SPLINE BASIS



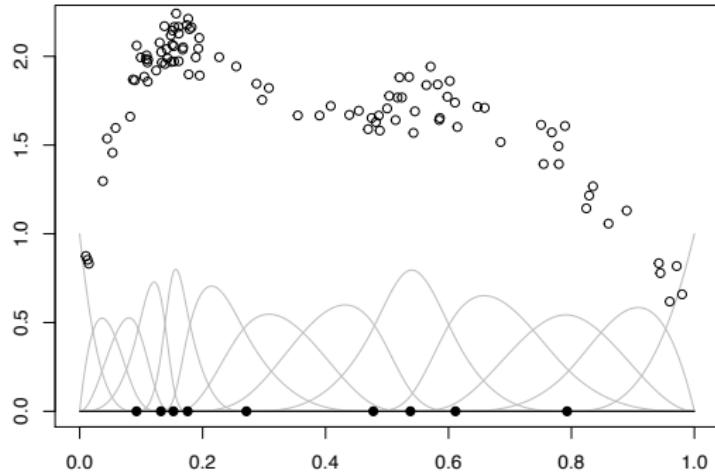
Knots can be placed along the percentiles of t_1, \dots, t_n

use B-spline function ↗

where
we have
more
observations



B-SPLINE BASIS



Knots replication permits discontinuity in derivatives or function itself



REGRESSION SPLINES

Matrix formulation

$$\Psi = \begin{bmatrix} \psi_1(t_1) & \psi_2(t_1) & \dots & \psi_K(t_1) \\ \psi_1(t_2) & \psi_2(t_2) & \dots & \psi_K(t_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(t_n) & \psi_2(t_n) & \dots & \psi_K(t_n) \end{bmatrix}$$

our new design matrix

$$\psi = (\psi_1, \dots, \psi_K)^\top$$

basis functions

$$Y = (Y_1, \dots, Y_n)^\top \quad f = (f(t_1), \dots, f(t_n))^\top$$

$$c = (c_1, \dots, c_K)^\top \quad \epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$$

Nonparametric regression model

$$Y = f + \epsilon = \Psi c + \epsilon \quad \mathbb{E}[\epsilon] = \mathbf{0} \quad \text{Var}[\epsilon] = \sigma^2 I_n$$

$$\hat{f} = \psi^\top \hat{c}$$

estimator



POLITECNICO MILANO 1863

REGRESSION SPLINES

Expression of the estimator

we don't exactly want estimate \mathbb{E} (there for it's not parametric)

$$\hat{\mathbf{c}} = \underset{\mathbf{c} \in \mathbb{R}^K}{\operatorname{argmin}} \text{RSS}(\mathbf{c}) = \underset{\mathbf{c} \in \mathbb{R}^K}{\operatorname{argmin}} \left\{ (\mathbf{Y} - \Psi \mathbf{c})^\top (\mathbf{Y} - \Psi \mathbf{c}) \right\}$$

$$\hat{\mathbf{c}} = (\Psi^\top \Psi)^{-1} \Psi^\top \mathbf{Y}$$

$$\hat{\mathbf{Y}} = \hat{\mathbf{f}} = \Psi \hat{\mathbf{c}} = \Psi (\Psi^\top \Psi)^{-1} \Psi^\top \mathbf{Y} = S \mathbf{Y}$$

$$S = \Psi (\Psi^\top \Psi)^{-1} \Psi^\top : \text{projection matrix (properties: } S^\top S = S)$$

S identical?

$$df = K = \text{tr}(S) = \text{tr}(S^\top S) = \text{tr}(2S - S^\top S)$$

dimension we projecting
columns

$$\hat{\sigma}^2 = \frac{1}{n-K} (\mathbf{Y} - \hat{\mathbf{Y}})^\top (\mathbf{Y} - \hat{\mathbf{Y}}) = \frac{1}{n-\text{tr}(S)} (\mathbf{Y} - \hat{\mathbf{Y}})^\top (\mathbf{Y} - \hat{\mathbf{Y}})$$

number of basis (already include constant, in L^2 was p_m)
 \downarrow
($p+1$)



REGRESSION SPLINES

Properties of the estimator

$$\hat{f}(t) = \psi(t)^\top \hat{\mathbf{c}} = \psi(t)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \mathbf{Y}$$

fixed on time t

$$E[\mathbf{y}] = \Psi \mathbf{c}$$

$$\mathbb{E}[\hat{f}(t)] = \mathbb{E}[\psi(t)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \mathbf{Y}] = \psi(t)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \Psi \mathbf{c} = \psi(t)^\top \mathbf{c}$$

$$\text{Bias}[\hat{f}(t)] = f(t) - \mathbb{E}[\hat{f}(t)] = f(t) - \psi(t)^\top \mathbf{c}$$

Source of bias: discretization

$$\text{Var}[\hat{f}(t)] = \mathbb{E}[\{\hat{f}(t) - \mathbb{E}[\hat{f}(t)]\}^2] = \sigma^2 \psi(t)^\top (\Psi^\top \Psi)^{-1} \psi(t)$$

$$\text{Var}[\hat{f}(t)] = \mathbb{E}[\{\hat{f}(t) - \mathbb{E}[\hat{f}(t)]\}^2] = \hat{\sigma}^2 \psi(t)^\top (\Psi^\top \Psi)^{-1} \psi(t)$$



REGRESSION SPLINES

Bias-Variance trade-off

Number of bases K controls Bias-Variance trade-off.

K can be selected by AIC, C_p Mallows, cross-validation,

Generalized Cross Validation:

$$\begin{aligned} GCV(K) &= \frac{n}{(n-K)} \frac{1}{(n-K)} (\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}}) \\ &= \frac{n}{(n - \text{tr}(S))^2} (\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}}) \end{aligned}$$

minim *leave-one-out CV*
compute once using self date
without making
k-models

Hopefully the chosen value of K is close to that minimizing

$$\text{MSE}[\hat{f}(t)] = \mathbb{E}[\{\hat{f}(t) - f(t)\}^2] = \text{Bias}^2[\hat{f}(t)] + \text{Var}[\hat{f}(t)]$$



REGRESSION SPLINES

Asymptotic properties

Under regularity conditions, as $n \rightarrow \infty$
and $K(n) \rightarrow \infty$ with appropriate rates

$$\text{Bias}[\hat{f}(t)] \rightarrow 0 \quad \text{and} \quad \text{Var}[\hat{f}(t)] \rightarrow 0$$

$$\text{MSE}[\hat{f}(t)] \rightarrow 0$$

$$\hat{f}(t) \approx \text{Gaussian}$$

Limiting Gaussian distribution justifies Wald type inference on $f(t)$:

- C.I. of approx level $(1 - \alpha)$ on $f(t)$: $\hat{f}(t) \pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\hat{f}(t)]}$
- Test on $H_0 : f(t) = f_0(t)$ vs $H_1 : f(t) \neq f_0(t)$ of approx level α

$$\text{Reject } H_0 \text{ if } |\hat{f}(t) - f_0(t)| > z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}[\hat{f}(t)]}$$

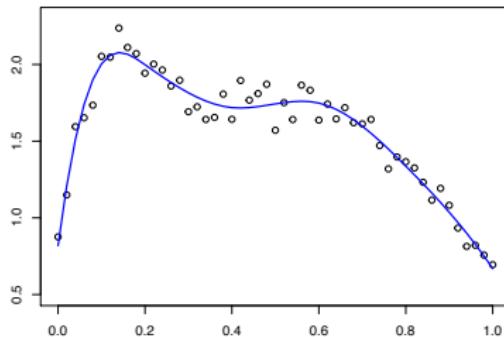


REGRESSION SPLINES

Inference

Caveats:

- This inference is only pointwise!

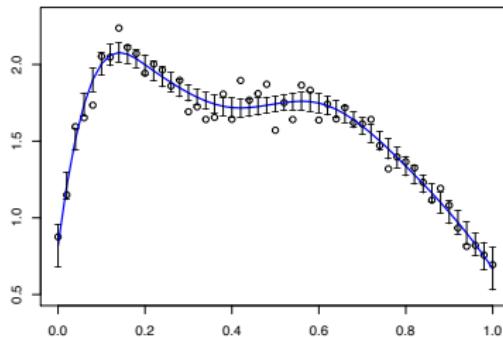


REGRESSION SPLINES

Inference

Caveats:

- This inference is only pointwise!



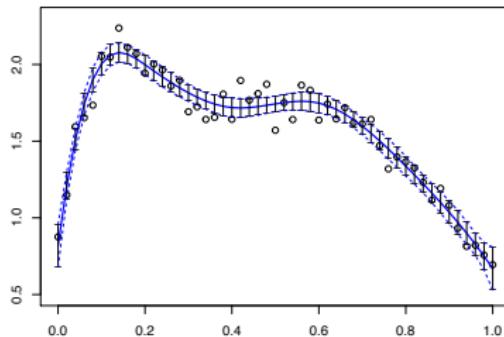
C.I.s are computed one at a time, not simultaneous!

REGRESSION SPLINES

Inference

Caveats:

- This inference is only pointwise!



C.I.s are computed one at a time, not simultaneous!

Even if you join the various C.I.s by bands, do not forget these are C.I.s for $f(t)$ in a specific t .



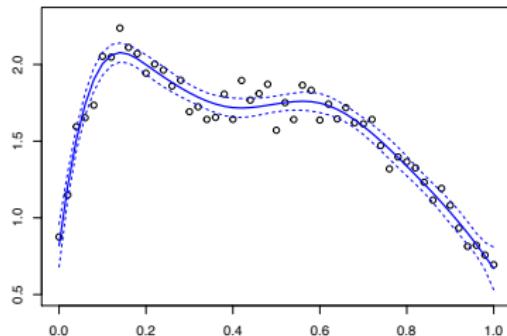
POLITECNICO MILANO 1863

REGRESSION SPLINES

Inference

Caveats:

- This inference is only pointwise!



C.I.s are computed one at a time, not simultaneous!

Even if you join the various C.I.s by bands, do not forget these are C.I.s for $f(t)$ in a specific t .

You cannot interpret the smooth dashed bands as delimiters of a region that includes the true overall f with a given confidence.



REGRESSION SPLINES

Inference

Caveats:

- ▶ This inference is only pointwise!



POLITECNICO MILANO 1863

REGRESSION SPLINES

Inference

Caveats:

- ▶ This inference is only pointwise! *(not simultaneous CI)*
- ▶ This inference does not account for the uncertainty in the selection of K



POLITECNICO MILANO 1863

REGRESSION SPLINES

Inference

Caveats:

- ▶ This inference is only pointwise!
- ▶ This inference does not account for the uncertainty in the selection of K
- ▶ Wald type inference is underconservative in nonparametric regression:



POLITECNICO MILANO 1863

REGRESSION SPLINES

Inference

Caveats:

- ▶ This inference is only pointwise!
- ▶ This inference does not account for the uncertainty in the selection of K
- ▶ Wald type inference is underconservative in nonparametric regression:
 - IC coverage is $< (1 - \alpha)$, i.e., true confidence < nominal



POLITECNICO MILANO 1863

REGRESSION SPLINES

Inference

Caveats:

- ▶ This inference is only pointwise!
- ▶ This inference does not account for the uncertainty in the selection of K
- ▶ Wald type inference is underconservative in nonparametric regression:
 - IC coverage is $< (1 - \alpha)$, i.e., true confidence < nominal
 - $\mathbb{P}(\text{type I error}) > \alpha$



POLITECNICO MILANO 1863

REGRESSION SPLINES

Inference

Caveats:

- ▶ This inference is only pointwise!
- ▶ This inference does not account for the uncertainty in the selection of K
- ▶ Wald type inference is underconservative in nonparametric regression:
 - IC coverage is $< (1 - \alpha)$, i.e., true confidence $<$ nominal
 - $\mathbb{P}(\text{type I error}) > \alpha$

Various alternative approaches, including bootstrap, as well as undersmoothing or oversmoothing approaches (see, e.g., review in Hall and Horowitz 2013)



POLITECNICO MILANO 1863

SMOOTHING SPLINES

Estimation functional

2nd approach: Do not restrict \mathcal{F} but
minimize RSS + roughness penalty

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \text{RSS}_\lambda(f) = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (Y_i - f(t_i))^2 + \lambda \mathcal{P}(f) \right\} \quad \lambda > 0$$

so we work in subset of \mathcal{F} where we restrict \mathcal{F}



POLITECNICO MILANO 1863

SMOOTHING SPLINES

Estimation functional

2nd approach: Do not restrict \mathcal{F} but
minimize RSS + roughness penalty

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \text{RSS}_\lambda(f) = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (Y_i - f(t_i))^2 + \lambda \mathcal{P}(f) \right\} \quad \lambda > 0$$



POLITECNICO MILANO 1863

SMOOTHING SPLINES

Estimation functional

2nd approach: Do not restrict \mathcal{F} but
minimize RSS + roughness penalty

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \text{RSS}_\lambda(f) = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (Y_i - f(t_i))^2 + \lambda \mathcal{P}(f) \right\} \quad \lambda > 0$$

we will always use spline minimizer functions of L^2
and their
first second
derivative

THM: If (a, b) s.t. $a < t_{[1]} < t_{[2]} < \dots < t_{[n]} < b$, $\mathcal{F} = H^2(a, b)$, and
 $\mathcal{P}(f) = \int_a^b (f''(t))^2 dx$ then \hat{f} is a natural cubic splines over (a, b)
with knots at $t_{[1]}, t_{[2]}, \dots, t_{[n]}$.

h^h - penalize h -th derivative



SMOOTHING SPLINES

Estimation functional

2nd approach: Do not restrict \mathcal{F} but
minimize RSS + roughness penalty

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \text{RSS}_\lambda(f) = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (Y_i - f(t_i))^2 + \lambda \mathcal{P}(f) \right\} \quad \lambda > 0$$
$$f = \sum_{j=1}^K c_j \psi_j$$

THM: If (a, b) s.t. $a < t_{[1]} < t_{[2]} < \dots < t_{[n]} < b$, $\mathcal{F} = H^2(a, b)$, and $\mathcal{P}(f) = \int_a^b (f''(t))^2 dx$ then \hat{f} is a natural cubic splines over (a, b) with knots at $t_{[1]}, t_{[2]}, \dots, t_{[n]}$.



SMOOTHING SPLINES

Estimation functional

2nd approach: Do not restrict \mathcal{F} but
minimize RSS + roughness penalty

$$\begin{aligned}\hat{f} &= \underset{f \in \mathcal{F}}{\operatorname{argmin}} \text{RSS}_\lambda(f) = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (Y_i - f(t_i))^2 + \lambda \mathcal{P}(f) \right\} \quad \lambda > 0 \\ &= \underset{\mathbf{c} \in \mathbb{R}^K}{\operatorname{argmin}} \left\{ (\mathbf{Y} - \Psi \mathbf{c})^\top (\mathbf{Y} - \Psi \mathbf{c}) + \lambda \mathbf{c}^\top P \mathbf{c} \right\} \quad P \in \mathbb{R}^K \times \mathbb{R}^K\end{aligned}$$

THM: If (a, b) s.t. $a < t_{[1]} < t_{[2]} < \dots < t_{[n]} < b$, $\mathcal{F} = H^2(a, b)$, and $\mathcal{P}(f) = \int_a^b (f''(t))^2 dx$ then \hat{f} is a natural cubic splines over (a, b) with knots at $t_{[1]}, t_{[2]}, \dots, t_{[n]}$.



SMOOTHING SPLINES

Estimation functional

2nd approach: Do not restrict \mathcal{F} but
minimize RSS + roughness penalty

$$\begin{aligned}\hat{f} &= \underset{f \in \mathcal{F}}{\operatorname{argmin}} \text{RSS}_\lambda(f) = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (Y_i - f(t_i))^2 + \lambda \mathcal{P}(f) \right\} \quad \lambda > 0 \\ &= \underset{\mathbf{c} \in \mathbb{R}^K}{\operatorname{argmin}} \left\{ (\mathbf{Y} - \Psi \mathbf{c})^\top (\mathbf{Y} - \Psi \mathbf{c}) + \lambda \mathbf{c}^\top P \mathbf{c} \right\} \quad P \in \mathbb{R}^K \times \mathbb{R}^K\end{aligned}$$

THM: If (a, b) s.t. $a < t_{[1]} < t_{[2]} < \dots < t_{[n]} < b$, $\mathcal{F} = H^2(a, b)$, and $\mathcal{P}(f) = \int_a^b (f''(t))^2 dx$ then \hat{f} is a natural cubic splines over (a, b) with knots at $t_{[1]}, t_{[2]}, \dots, t_{[n]}$.

Ex: For $\mathcal{P}(f) = \int (f'')^2$ then (j, ℓ) -entry of P is $\int_a^b \psi_j''(t) \psi_\ell''(t) dx$



SMOOTHING SPLINES

Estimation functional

2nd approach: $\hat{f} \in \mathcal{F}_K$, with $\dim(\mathcal{F}_K) = K \approx n$

minimize RSS + roughness penalty

like
Lagrangean

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \text{RSS}_\lambda(f) = \operatorname{argmin}_{f \in \mathcal{F}_K} \left\{ \sum_{i=1}^n (Y_i - f(t_i))^2 + \lambda \mathcal{P}(f) \right\} \quad \lambda > 0$$

$$= \operatorname{argmin}_{\mathbf{c} \in \mathbb{R}^K} \left\{ (\mathbf{Y} - \Psi \mathbf{c})^\top (\mathbf{Y} - \Psi \mathbf{c}) + \lambda \mathbf{c}^\top P \mathbf{c} \right\} \quad P \in \mathbb{R}^K \times \mathbb{R}^K$$

THM: If (a, b) s.t. $a < t_{[1]} < t_{[2]} < \dots < t_{[n]} < b$, $\mathcal{F} = H^2(a, b)$, and $\mathcal{P}(f) = \int_a^b (f''(t))^2 dx$ then \hat{f} is a natural cubic splines over (a, b) with knots at $t_{[1]}, t_{[2]}, \dots, t_{[n]}$.

Ex: For $\mathcal{P}(f) = \int (f'')^2$ then (j, ℓ) -entry of P is $\int_a^b \psi_j''(t) \psi_\ell''(t) dx$



POLITECNICO MILANO 1863

SMOOTHING SPLINES

For large datasets, it may be convenient to employ a mixed approach, where one considers a roughness penalty, but also reduces the dimensionality of the data estimation problem by considering a basis of dimension K , where K is still large but somehow smaller than n .



SMOOTHING SPLINES

Expression of the estimator

like in Ridge regression

$$\hat{\mathbf{c}} = (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y}$$



POLITECNICO MILANO 1863

SMOOTHING SPLINES

Expression of the estimator

$$\hat{\mathbf{c}} = (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y}$$

each column of Ψ is
basis vector

$$\hat{\mathbf{Y}} = \hat{\mathbf{f}} = \Psi \hat{\mathbf{c}} = \Psi (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y} = S \mathbf{Y}$$



SMOOTHING SPLINES

Expression of the estimator

$$\hat{\mathbf{c}} = (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y}$$

$$\hat{\mathbf{Y}} = \hat{\mathbf{f}} = \Psi \hat{\mathbf{c}} = \Psi (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y} = S \mathbf{Y}$$

$$S = \Psi (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top : \text{sub-projection operator } (S^\top S \neq S)$$

because we add penalty,
so now it's not real
projection matrix



SMOOTHING SPLINES

Expression of the estimator

$$\hat{\mathbf{c}} = (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y}$$

$$\hat{\mathbf{Y}} = \hat{\mathbf{f}} = \Psi \hat{\mathbf{c}} = \Psi (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y} = S \mathbf{Y}$$

$S = \Psi (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top$: sub-projection operator ($S^\top S \neq S$)

$df = \text{tr}(S) < K$ (or $df = \text{tr}(S^\top S)$ or $df = \text{tr}(2S - S^\top S)$)

used to measure degree of freedom of estimator, larger λ - more regularization, more reduce complexity of estimator



SMOOTHING SPLINES

Expression of the estimator

$$\hat{\mathbf{c}} = (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y}$$

$$\hat{\mathbf{Y}} = \hat{\mathbf{f}} = \Psi \hat{\mathbf{c}} = \Psi (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y} = S \mathbf{Y}$$

$S = \Psi (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top$: sub-projection operator ($S^\top S \neq S$)
Buyale S -Identical matrix

$$df = \text{tr}(S) < K \quad (\text{or } df = \text{tr}(S^\top S) \text{ or } df = \text{tr}(2S - S^\top S))$$

$$\hat{\sigma}^2 = \frac{1}{n - \text{tr}(S)} (\mathbf{Y} - \hat{\mathbf{Y}})^\top (\mathbf{Y} - \hat{\mathbf{Y}})$$



SMOOTHING SPLINES

Properties of the estimator

$$\hat{f}(t) = \psi(t)^\top \hat{\mathbf{c}} = \psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y}$$

$$\mathbb{E}[\hat{f}(t)] = \mathbb{E}[\psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y}] = \psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \Psi \mathbf{c}$$

$$\text{Bias}[\hat{f}(t)] = f(t) - \mathbb{E}[\hat{f}(t)] = f(t) - \psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \Psi \mathbf{c}$$

Sources of bias: discretization and penalization

(unless true f is s.t. $\mathcal{P}(f) = 0$, the penalty induces a bias)

we can let $d \rightarrow \infty$ and $n \rightarrow \infty$ problem well posed
bias disappear

bigger $\lambda \rightarrow$ less variance
 $\lambda \downarrow \rightarrow \downarrow \text{bias} \uparrow \text{variance}$



SMOOTHING SPLINES

Properties of the estimator

$$\hat{f}(t) = \psi(t)^\top \hat{\mathbf{c}} = \psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y}$$

$$\mathbb{E}[\hat{f}(t)] = \mathbb{E}[\psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y}] = \psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \Psi \mathbf{c}$$

$$\text{Bias}[\hat{f}(t)] = f(t) - \mathbb{E}[\hat{f}(t)] = f(t) - \psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \Psi \mathbf{c}$$

Sources of bias: discretization and penalization

(unless true f is s.t. $\mathcal{P}(f) = 0$, the penalty induces a bias)

$$\text{Var}[\hat{f}(t)] = \sigma^2 \psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \Psi (\Psi^\top \Psi + \lambda P)^{-1} \psi(t)$$



POLITECNICO MILANO 1863

SMOOTHING SPLINES

Properties of the estimator

$$\hat{f}(t) = \psi(t)^\top \hat{\mathbf{c}} = \psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y}$$

$$\mathbb{E}[\hat{f}(t)] = \mathbb{E}[\psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \mathbf{Y}] = \psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \Psi \mathbf{c}$$

$$\text{Bias}[\hat{f}(t)] = f(t) - \mathbb{E}[\hat{f}(t)] = f(t) - \psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \Psi \mathbf{c}$$

Sources of bias: discretization and penalization

(unless true f is s.t. $\mathcal{P}(f) = 0$, the penalty induces a bias)

$$\text{Var}[\hat{f}(t)] = \sigma^2 \psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \Psi (\Psi^\top \Psi + \lambda P)^{-1} \psi(t)$$

$$\widehat{\text{Var}[\hat{f}(t)]} = \hat{\sigma}^2 \psi(t)^\top (\Psi^\top \Psi + \lambda P)^{-1} \Psi^\top \Psi (\Psi^\top \Psi + \lambda P)^{-1} \psi(t)$$



SMOOTHING SPLINES

Bias-Variance trade-off

Smoothness parameter λ controls Bias-Variance trade-off.

Selection of smoothness parameter λ : AIC, C_p , Mallows, cross-validation, Generalized Cross Validation:

$$GCV(\lambda) = \frac{n}{(n - \text{tr}(S))^2} (\mathbf{Y} - \hat{\mathbf{Y}})^\top (\mathbf{Y} - \hat{\mathbf{Y}})$$

(Mimics leave-one-out CV)



POLITECNICO MILANO 1863

SMOOTHING SPLINES

Bias-Variance trade-off

Smoothness parameter λ controls Bias-Variance trade-off.

Selection of smoothness parameter λ : AIC, C_p Mallows, cross-validation, Generalized Cross Validation:

$$GCV(\lambda) = \frac{n}{(n - tr(S))^2} (\mathbf{Y} - \hat{\mathbf{Y}})^\top (\mathbf{Y} - \hat{\mathbf{Y}})$$

Hopefully the chosen λ is close to that minimizing $MSE[\hat{f}]$.



SMOOTHING SPLINES

Under regularity conditions, as $n \rightarrow \infty$,
and $K(n) \rightarrow \infty$ and $\lambda(n) \rightarrow 0$ with appropriate rates

$$\text{Bias}[\hat{f}(t)] \rightarrow 0 \quad \text{and} \quad \text{Var}[\hat{f}(t)] \rightarrow 0$$

$$\text{MSE}[\hat{f}(t)] \rightarrow 0$$

$$\hat{f}(t) \approx \text{Gaussian}$$

Limiting Gaussian distrib justifies Wald type inference on $f(t)$.



REFERENCES

- Green, Peter J. and Bernard W. Silverman (1994). *Nonparametric regression and generalized linear models*. Vol. 58. Monographs on Statistics and Applied Probability. A roughness penalty approach. Chapman & Hall, London, pp. xii+182. ISBN: 0-412-30040-0. DOI: 10.1007/978-1-4899-4473-3. URL: <http://dx.doi.org/10.1007/978-1-4899-4473-3>.
- Hall, Peter and Joel Horowitz (2013). "A simple bootstrap method for constructing nonparametric confidence bands for functions". In: *Ann. Statist.* 41.4, pp. 1892–1921. ISSN: 0090-5364. DOI: 10.1214/13-AOS1137. URL: <https://doi.org/10.1214/13-AOS1137>.
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani ((2021) ©2021). *An introduction to statistical learning—with applications in R*. Second. Springer Texts in Statistics. Springer, New York, pp. xv+607. ISBN: 978-1-0716-1417-4; 978-1-0716-1418-1. DOI: 10.1007/978-1-0716-1418-1. URL: <https://doi.org/10.1007/978-1-0716-1418-1>.
- Wood, Simon N (2017). *Generalized additive models: an introduction with R*. CRC press.

$$\hat{f}'(s) = \sum_{k=1}^K \hat{c}_k \psi'_k(s) \quad \hat{f}''(s) = \sum_{k=1}^K \hat{c}_k \psi''_k(s)$$

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)

$$SSE_\lambda = SSE + \lambda \int (f^{[d]}(s))^2 ds$$

$$SSE_\lambda = SSE + \lambda \int (Lf(s))^2 ds$$

differential operators

(so we penalise severe differentials)
gradient

Comments on
penalization in other
contexts



POLITECNICO DI MILANO



Politecnico di Milano
Applied Statistics
May 2025



An introduction to functional data analysis Part 2 – Smoothing: additional material

Laura M. SANGALLI

MOX - Dipartimento di Matematica, Politecnico di Milano

<https://sangalli.faculty.polimi.it/>

Two alternative approaches to smoothing

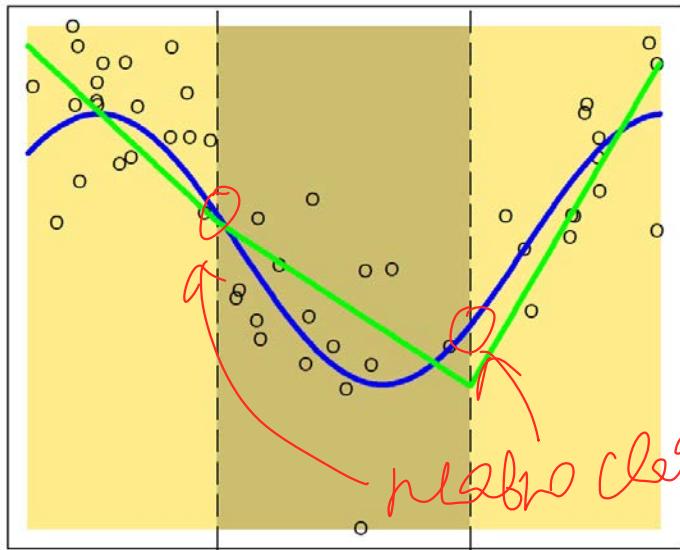
Two main ideas:



- 1) Keep the windows fixed but require appropriate continuity at knots!

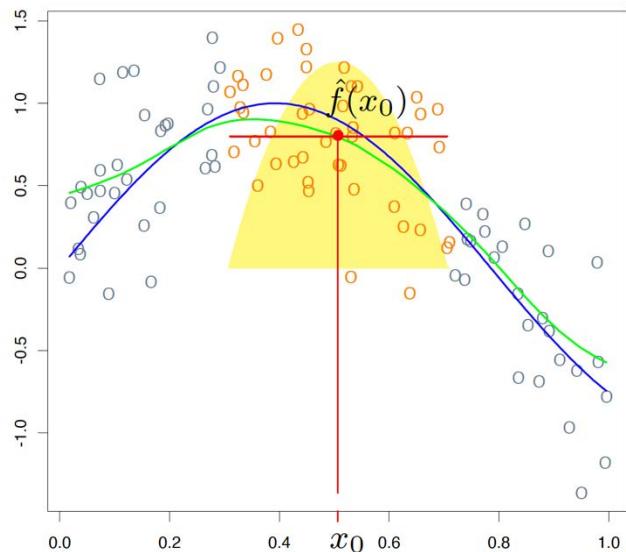
neoprobabilità

split in sub domains
each have polynomial, and →
then polynomial link with
continuity

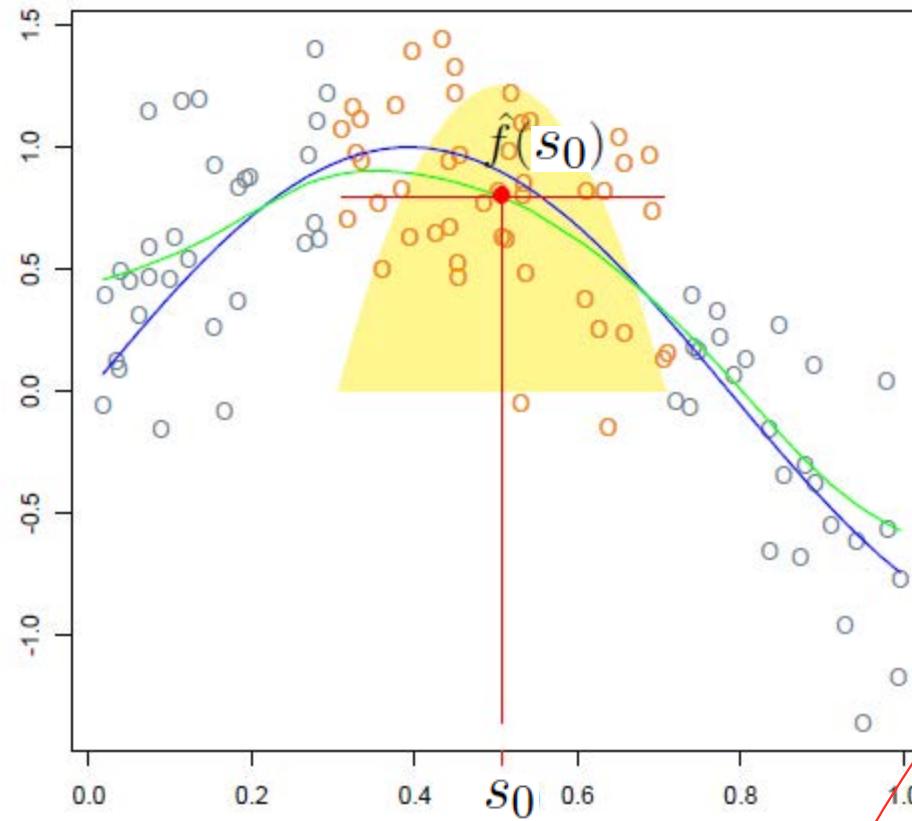


- 2) Fit in moving windows!

*fit in every point
 where I want to
 estimate*



Local polynomial regression (kernel smoother)



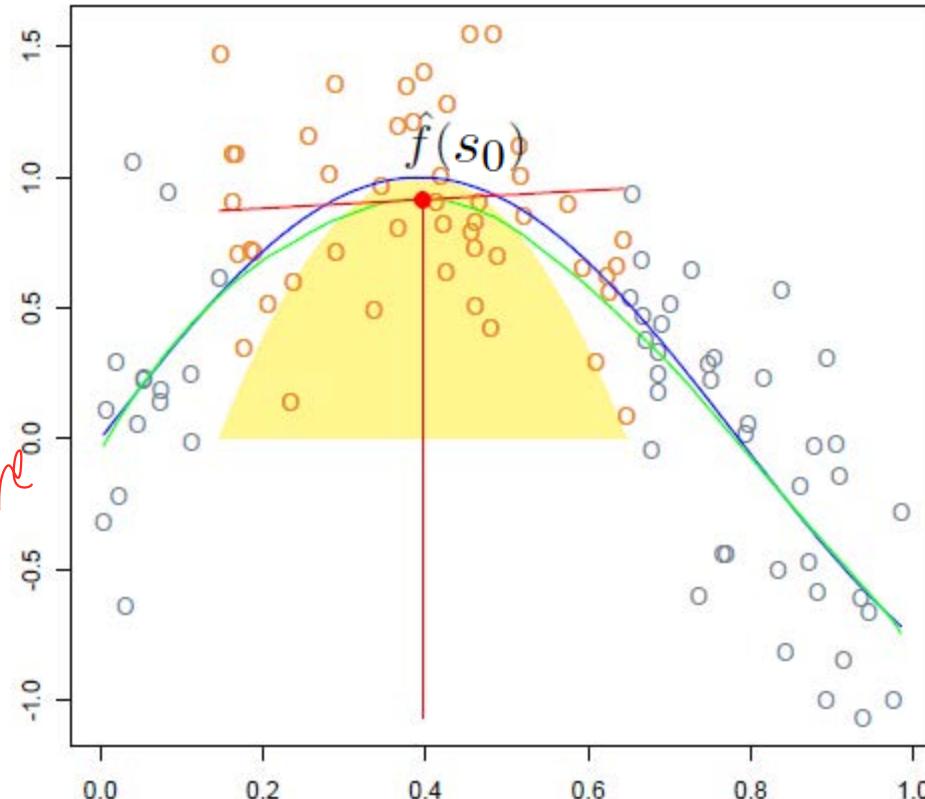
Kernel smoothing
(weighted Least Squares)
where weight come
from kernel

Picture taken from
Friedman, Tibshirani and
Hastie (2013) The Elements of
Statistical Learning, Springer

At each abscissa s_0 , find (c_0, \dots, c_L) that minimize

$$\sum_{i=1}^n \text{Kern}_h(s_0, s_i) [(z_i - \sum_{l=0}^L c_l (s_0 - s_i)^l)^2] \quad \text{where } \text{Kern}_h(s_0, s_i) = D\left(\frac{|s_0 - s_i|}{h}\right)$$

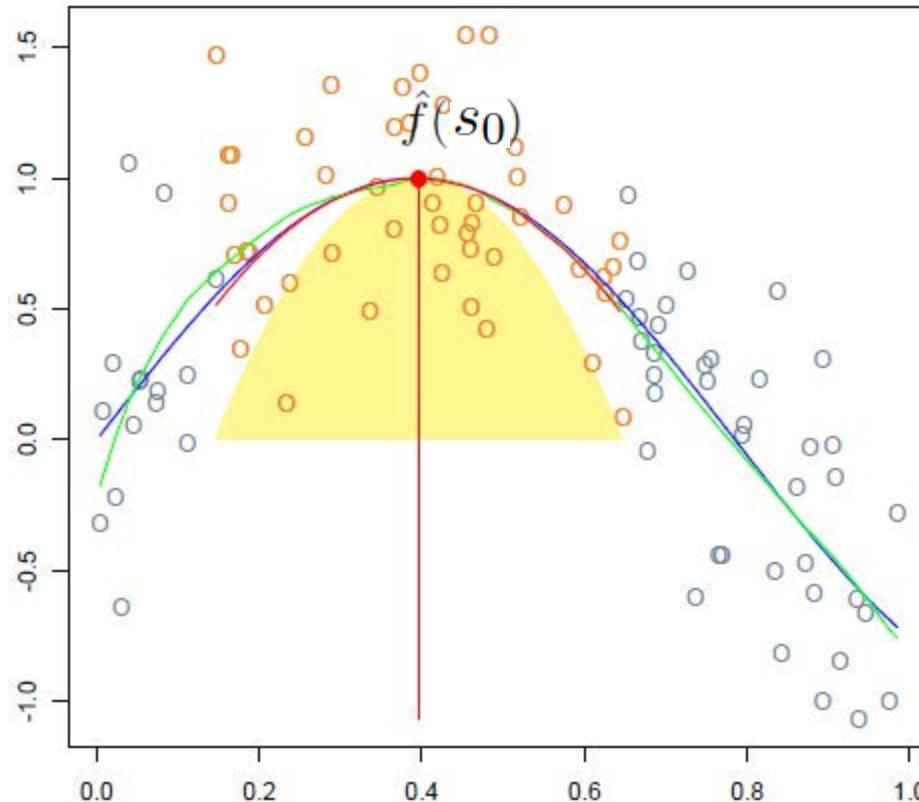
weights
in least squares



At each abscissa s_0 , find (c_0, \dots, c_L) that minimize

$$\sum_{i=1}^n \text{Kern}_h(s_0, s_i) [(z_i - \sum_{l=0}^L c_l (s_0 - s_i)^l)^2] \quad \text{where } \text{Kern}_h(s_0, s_i) = D\left(\frac{|s_0 - s_i|}{h}\right)$$

Picture taken from
Friedman, Tibshirani and
Hastie (2013) The Elements of
Statistical Learning, Springer



Picture taken from
Friedman, Tibshirani and
Hastie (2013) The Elements of
Statistical Learning, Springer

At each abscissa s_0 , find (c_0, \dots, c_L) that minimize

$$\sum_{i=1}^n \text{Kern}_h(s_0, s_i)[(z_i - \sum_{l=0}^L c_l(s_0 - s_i)^l)]^2 \quad \text{where } \text{Kern}_h(s_0, s_i) = D\left(\frac{|s_0 - s_i|}{h}\right)$$

Positive functions:

$$f(s) = e^{W(s)} \quad \text{where } W(s) = \sum_k c_k \psi_k(s)$$

Positive functions:

$$f(s) = e^{W(s)} \quad \text{where } W(s) = \sum_k c_k \psi_k(s)$$

Estimate f by minimizing

$$\sum_{i=1}^n (z_i - e^{W(s_i)})^2 + \lambda \int W''(s) ds$$

Increasing functions:

$$f(s) = C + \int_{s_0}^s \exp\{W(t)\} dt \quad \text{where } W(s) = \sum_k c_k \psi_k(s)$$

Densities (B^2 space): —splines, which have to be density.

Machalová, J., Hron, K., Monti, G., 2015. Preprocessing of centred logratio transformed density functions using smoothing splines. *J. Appl. Stat.* 43.



3) Choose basis adaptively to data

$K \ll n$

Some possibilities:

- Free-knot regression splines

Unidimensional curves: see, e.g., Zhou, Shen (2001) JASA

Multidimensional curves: see, e.g., Sangalli, Secchi, Vantini, Veneziani (2009) JRSSC

- Wavelets

Fourier basis

Unidimensional curves: see, e.g., Hastie, Tibshirani, Friedman (2009)

Multidimensional curves: see, e.g., Pigoli, Sangalli (2012) CSDA

- Good for modeling sharp local features
- Localized in both space and frequency
- An analytical expression may not exist
- Computationally efficient (orthogonal)

- Functional Principal Components Analysis (or other basis constructed from data)



POLITECNICO DI MILANO



Politecnico di Milano
Applied Statistics
May 2025



An introduction to functional data analysis

Part 3 - Alignement *(βlapAb nu BAue)*

Laura M. SANGALLI

MOX - Dipartimento di Matematica, Politecnico di Milano

<https://sangalli.faculty.polimi.it/>

Decoupling and studying Phase and Amplitude variabilities

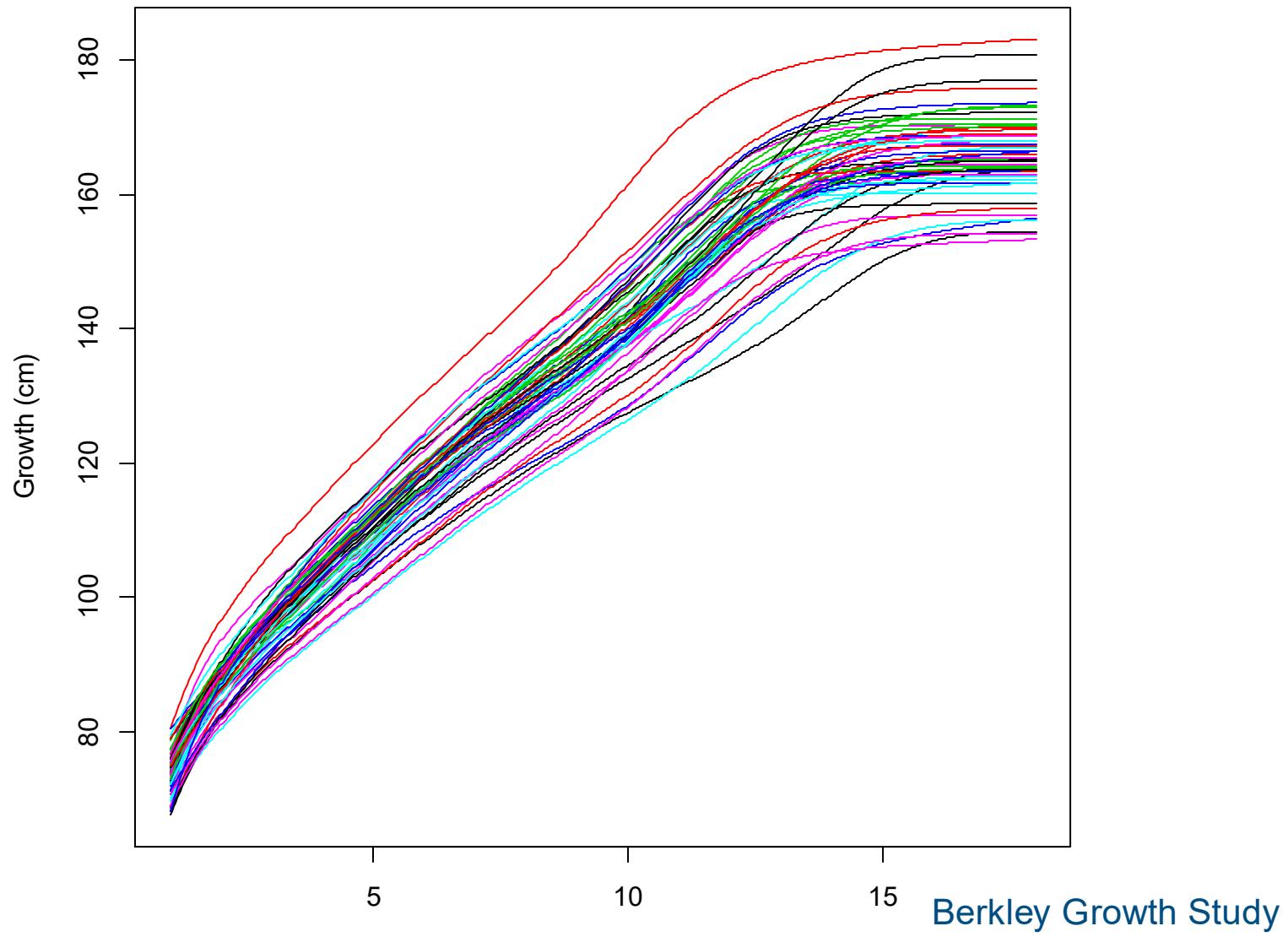
Registration, Alignment, Warping

Chapter 7, Ramsay and Silverman 2005, Springer

Review article:

Marron, J.S., Ramsay, J.O., Sangalli, L.M., Srivastava, A. (2015), Functional Data Analysis of Amplitude and Phase Variation, *Statistical Science*, 30 (4), 468-484.

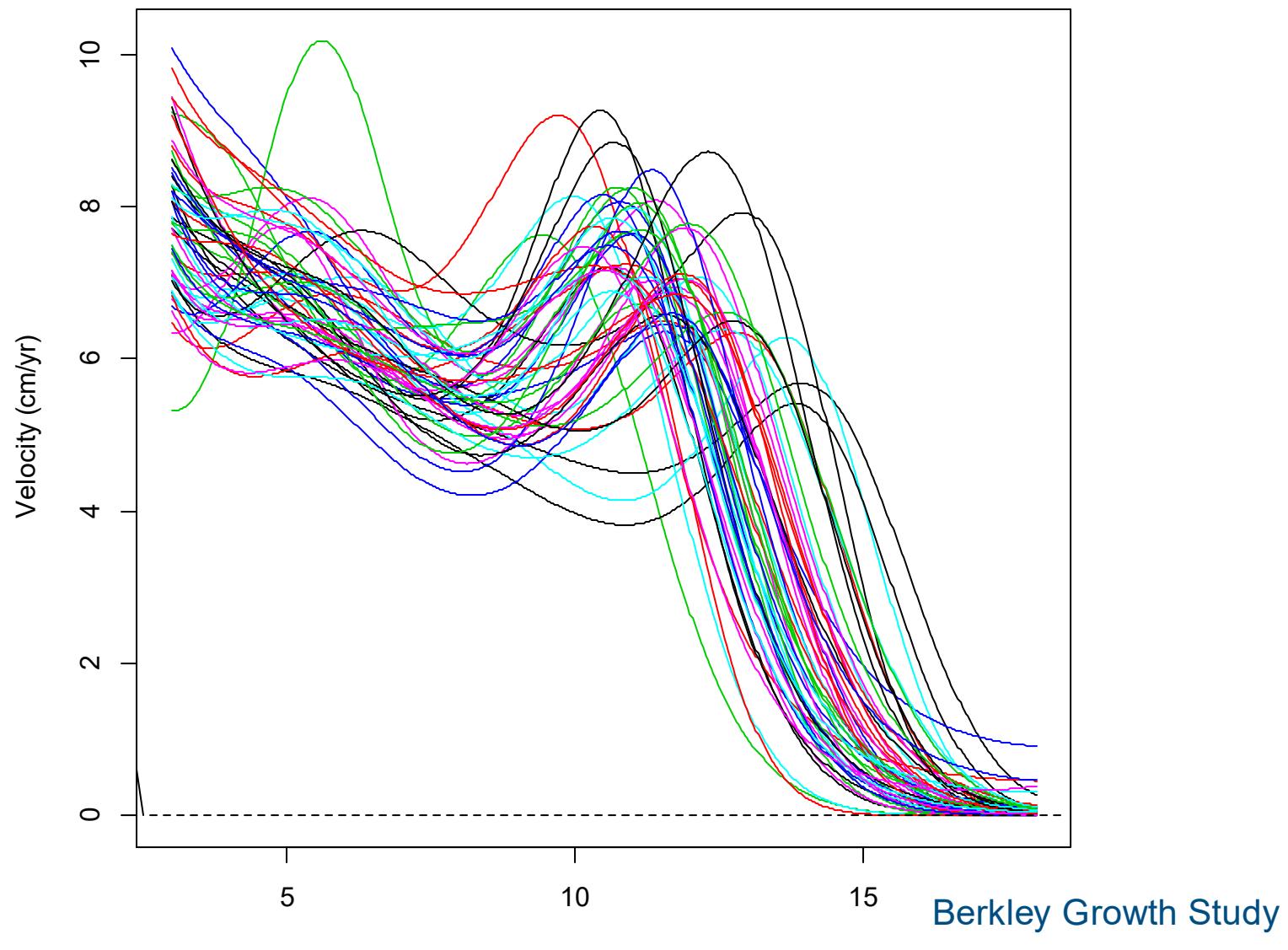




Phase and Amplitude variability

74

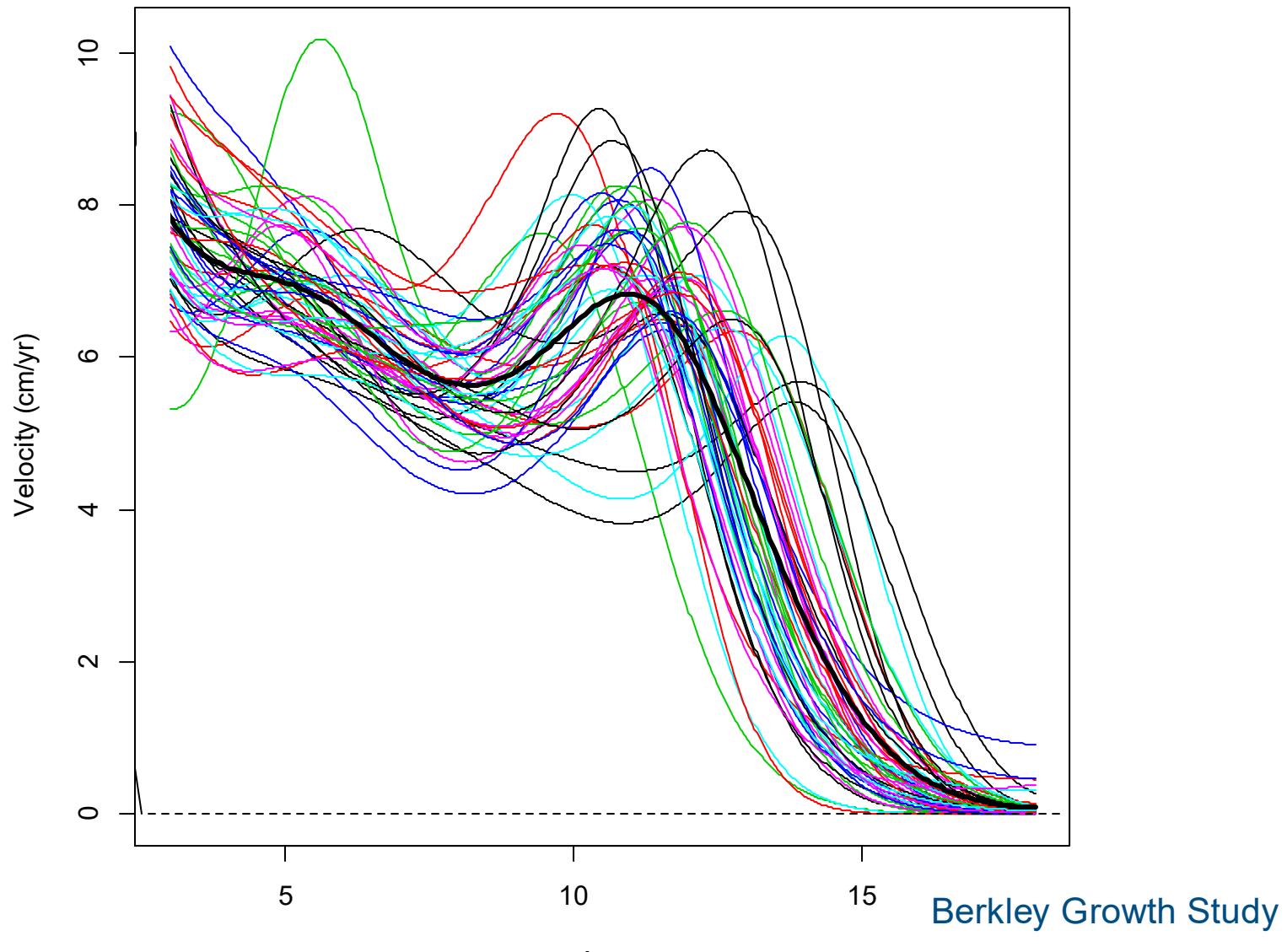
Ramsay Silverman 2005 Springer



Phase and Amplitude variability

75

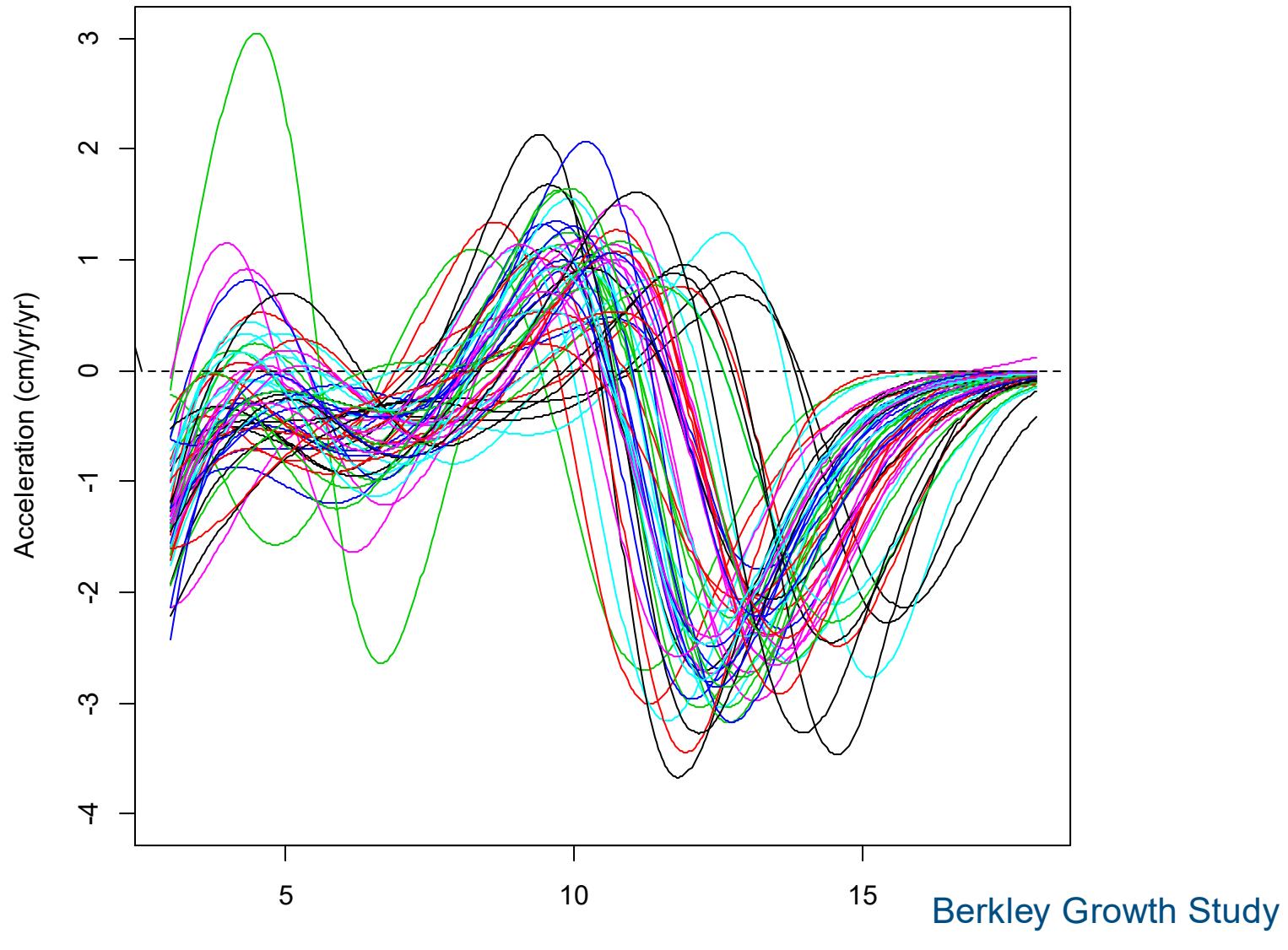
Ramsay Silverman 2005 Springer



Phase and Amplitude variability

76

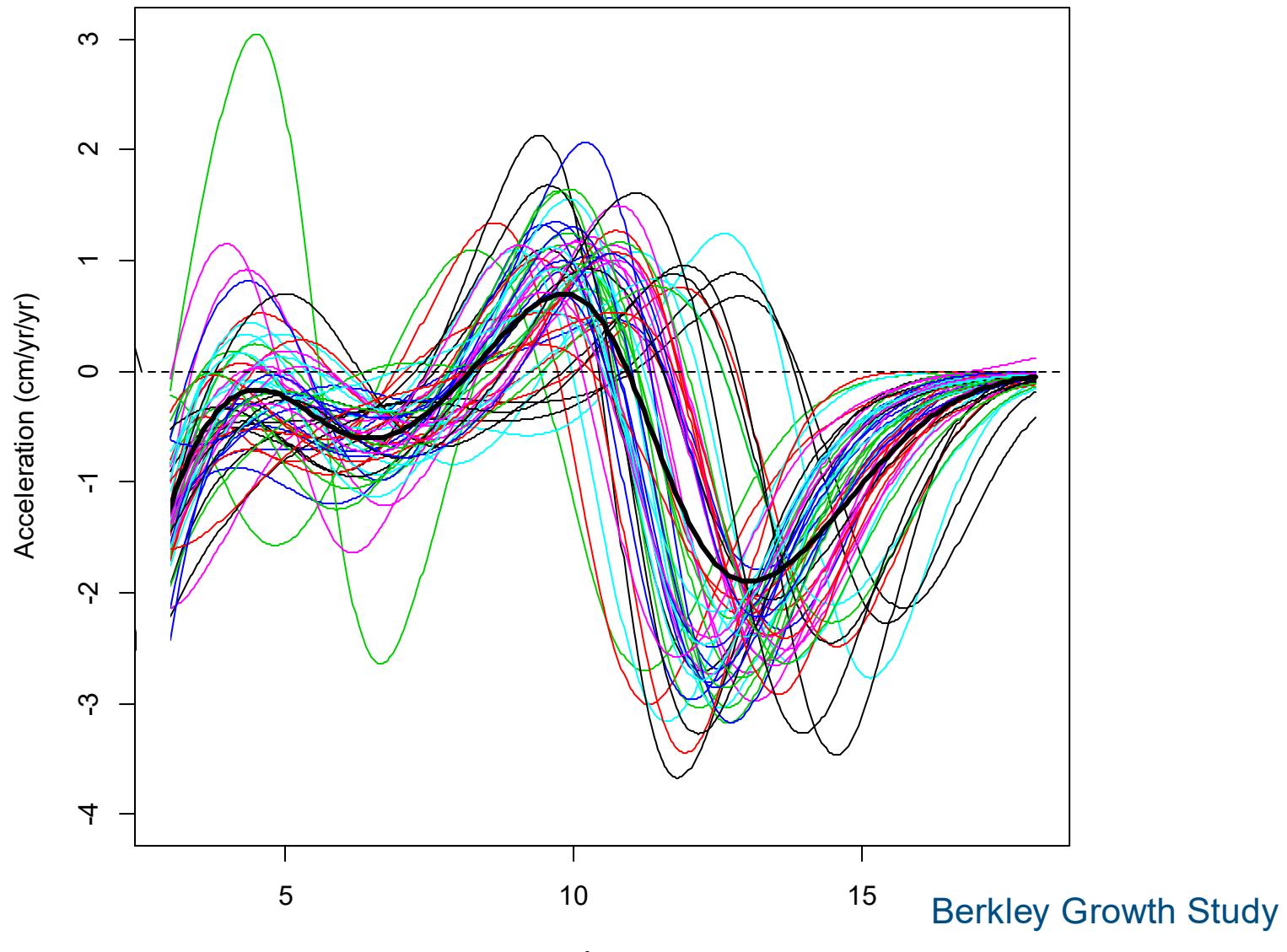
Ramsay Silverman 2005 Springer



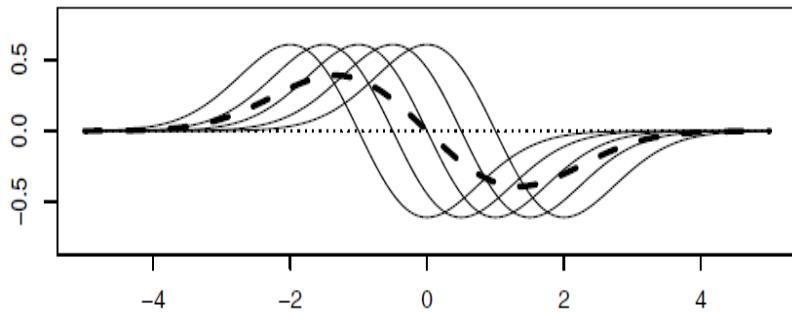
Phase and Amplitude variability

77

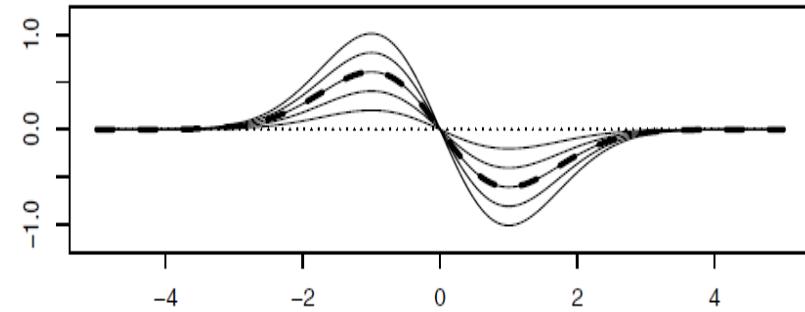
Ramsay Silverman 2005 Springer



Phase variability



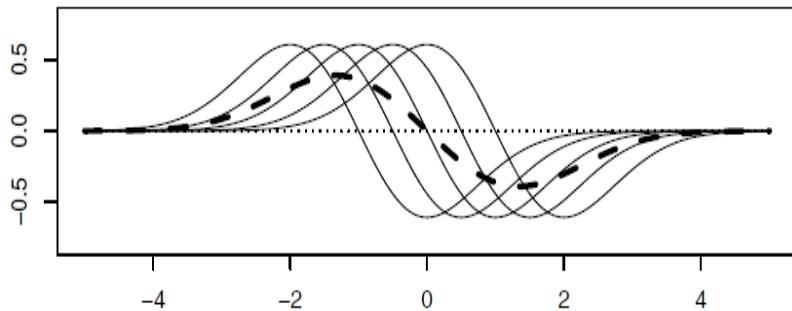
Amplitude variability



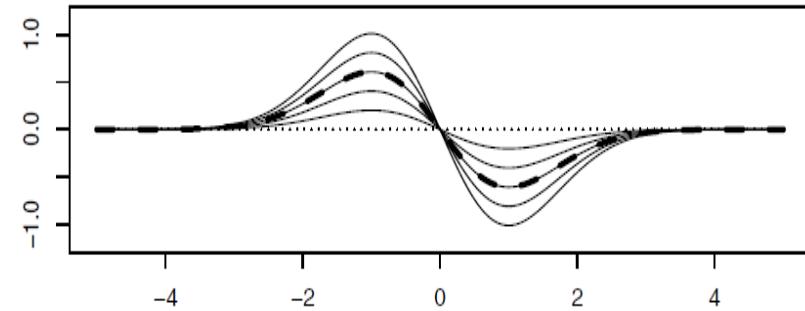
Phase variability: different curves exhibit more or less the same features but that these features occur at different times or space locations for different statistical units.

If not taken properly into account, the misalignment acts as a confounding factor and may blur subsequent analyses.

Phase variability



Amplitude variability



Registration/alignement/warping of a set of functions

Find suitable warping functions $h_1(t), \dots, h_N(t)$ such that $c_1(h_1(t)) \dots, c_N(h_N(t))$ 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 $h_i(t)=t+\delta t_i$
- 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.

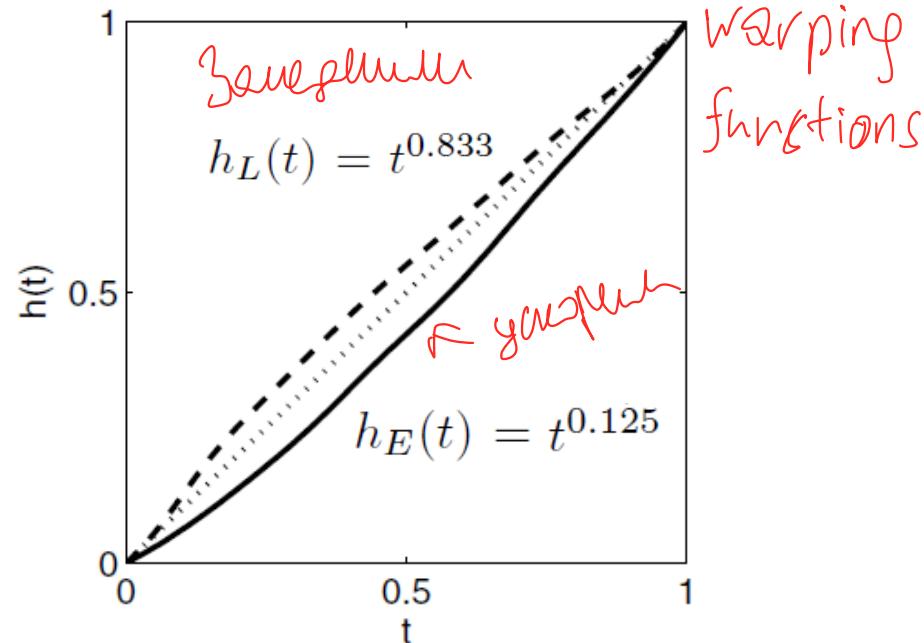
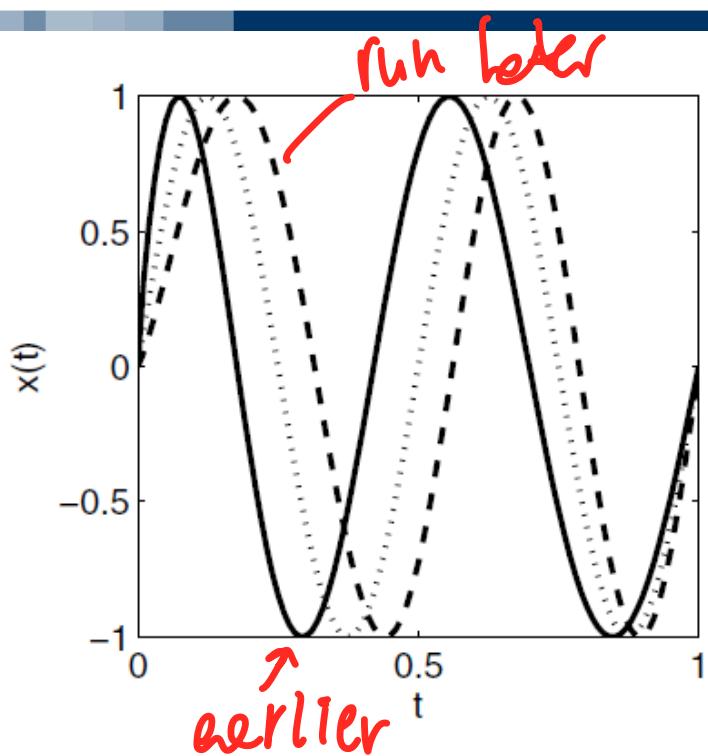
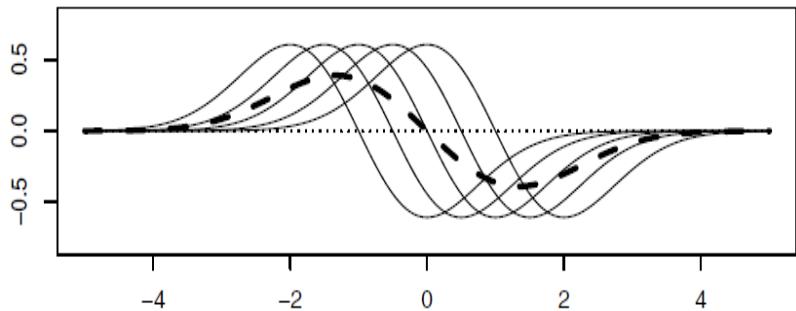


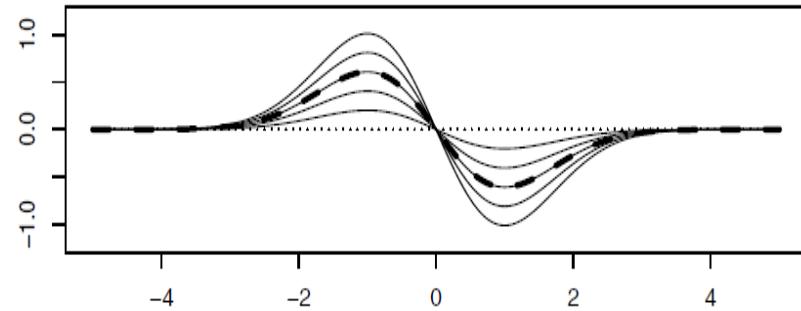
Figure 7.9. The left panel shows the target function, $x_0(t) = \sin(4\pi t)$, as a dotted line; an early function, $x_E(t) = \sin(4\pi t^{0.8})$, as a solid line; and a late function, $x_L(t) = \sin(4\pi t^{1.2})$, as a dashed line. The corresponding warping functions that register the early and late curves to the target are shown in the right panel.

ges gesher

Phase variability



Amplitude variability



Registration of a set of functions

Find suitable warping functions h_1, \dots, h_N such that $c_1 \circ h_1, \dots, c_n \circ h_N$ are the most similar.

→ **Landmark Approach:** known **landmarks** along the curves that are aligned so that landmarks occurs at the same abscissa points.

→ **Continuous Approach:** define a measure of similarity/dissimilarity between curves, that are aligned in order to maximize/minimize their similarity/dissimilarity.

Landmarks: significant (univocally identifiable) shape-events in a curve, e.g. crossings of zero, peaks, valleys, points of inflection.

c_1, \dots, c_N , where $c_i : [0, T] \rightarrow \mathbb{R}^d$

Suppose

- L landmarks; for the i -th curve, located at t_{i1}, \dots, t_{iL}
- a template curve c_0 is available with landmark locations t_{01}, \dots, 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.

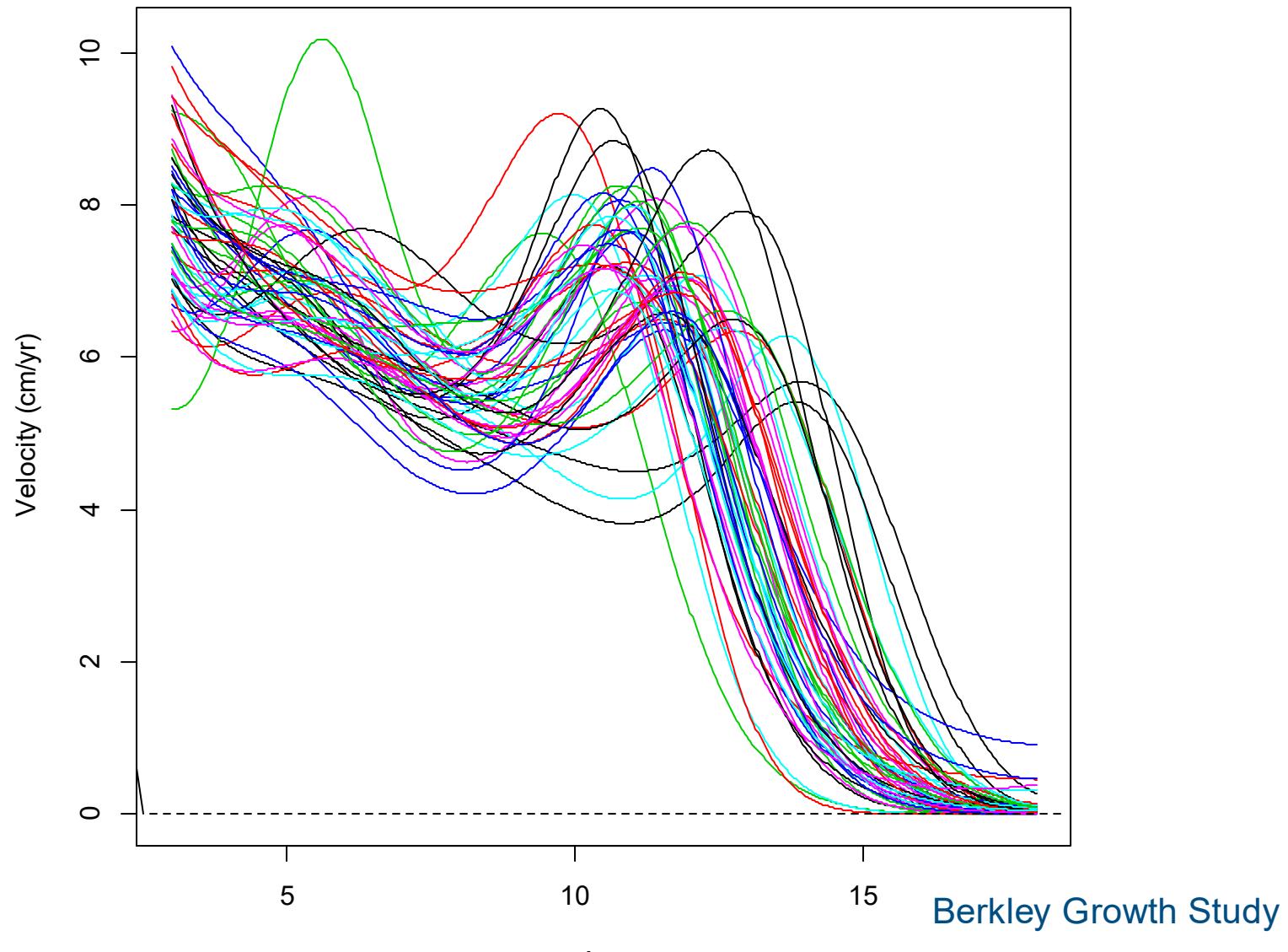
- $h_i(0) = 0$
 - $h_i(t_{0j}) = t_{ij}$, for $j = 1, \dots, L$
 - $h_i(T) = T$
- $(0, 0), (t_{01}, t_{i1}), \dots, (t_{0L}, t_{iL}), (T, T)$: interpolated by a piece-wise line, a polygon or higher order monotone splines (strictly increasing)

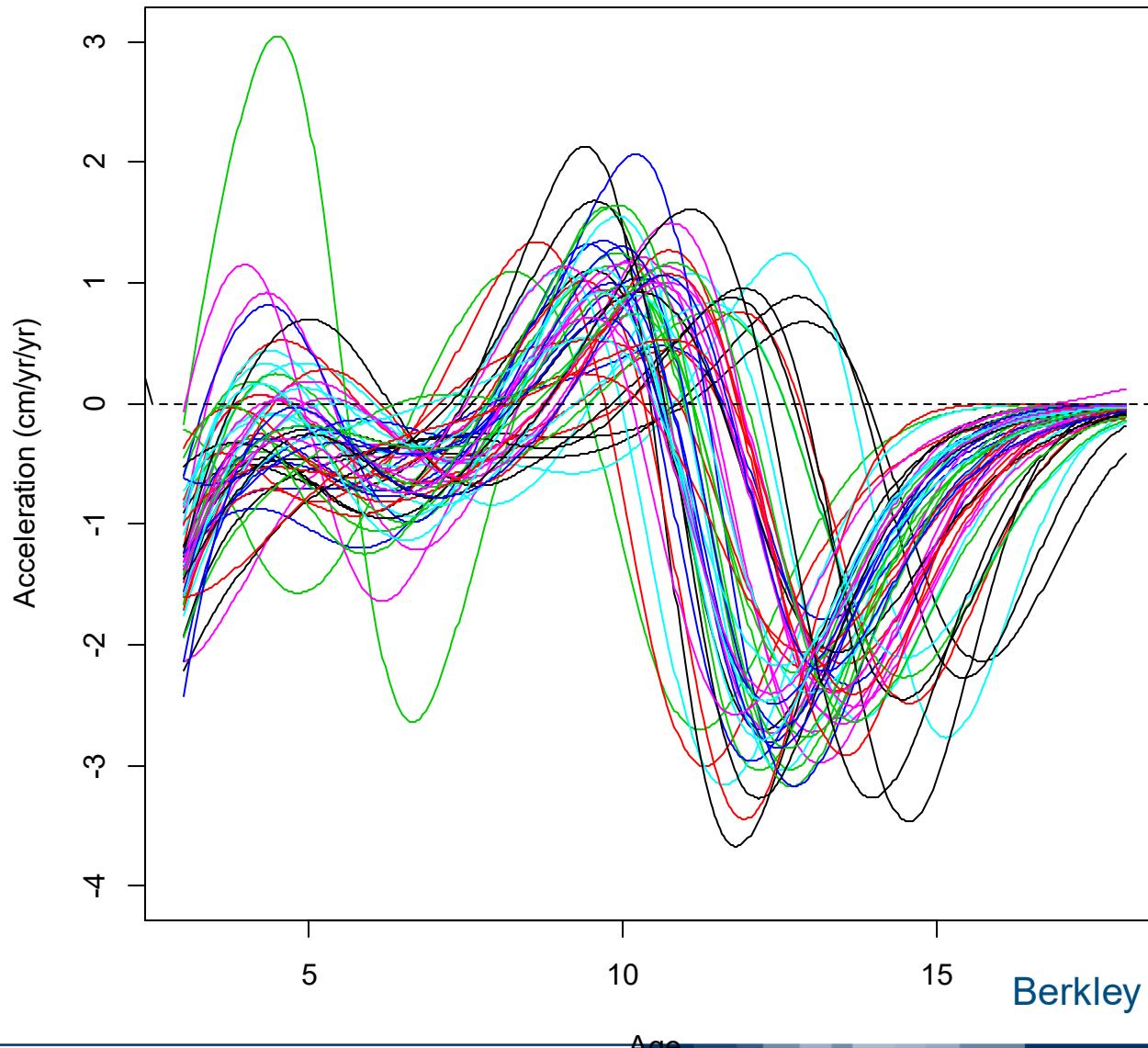
A model for warping functions

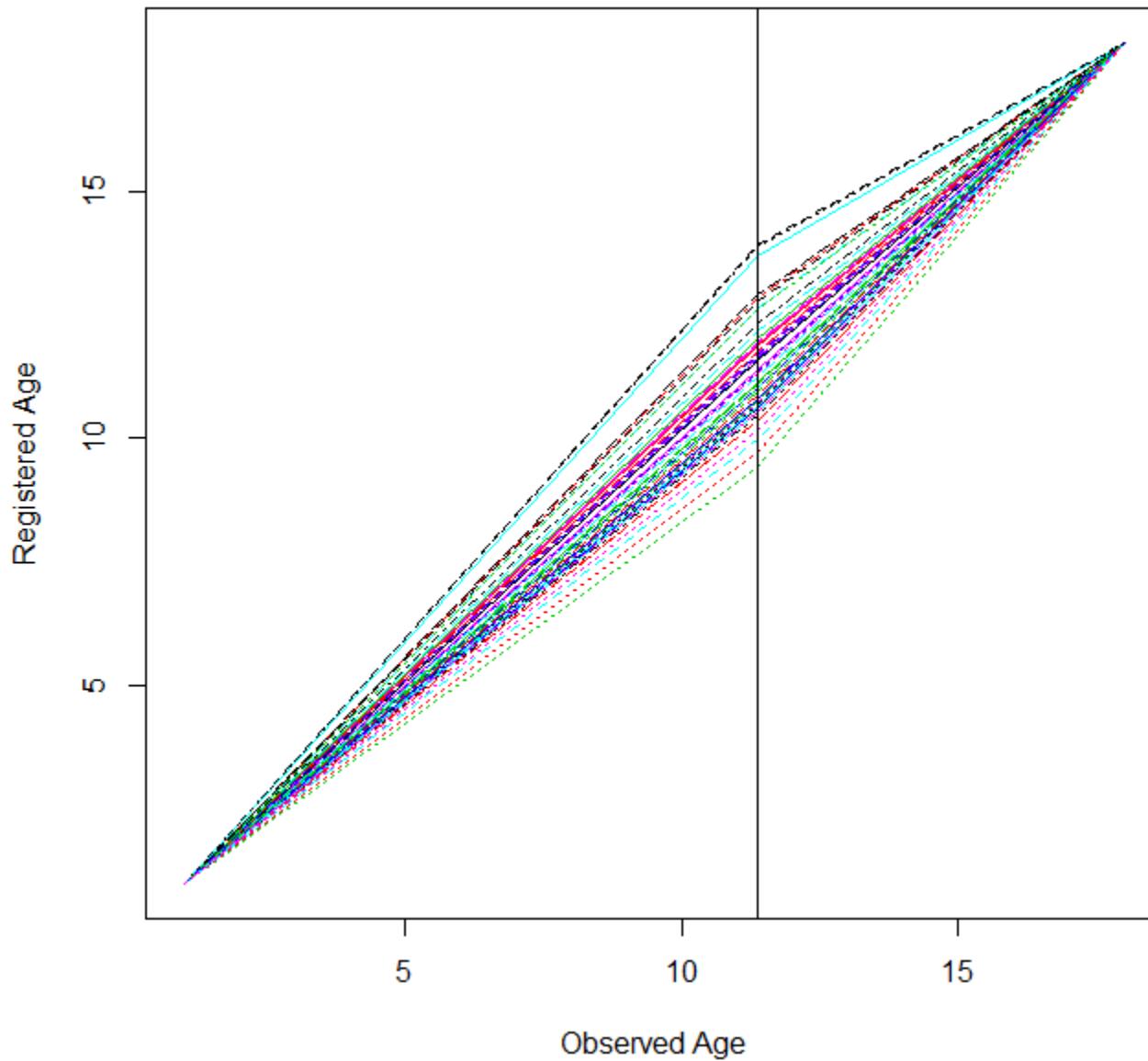
$$h_i(t) = C_{0i} + C_{1i} \int_0^t \exp\{W_i(u)\} du$$

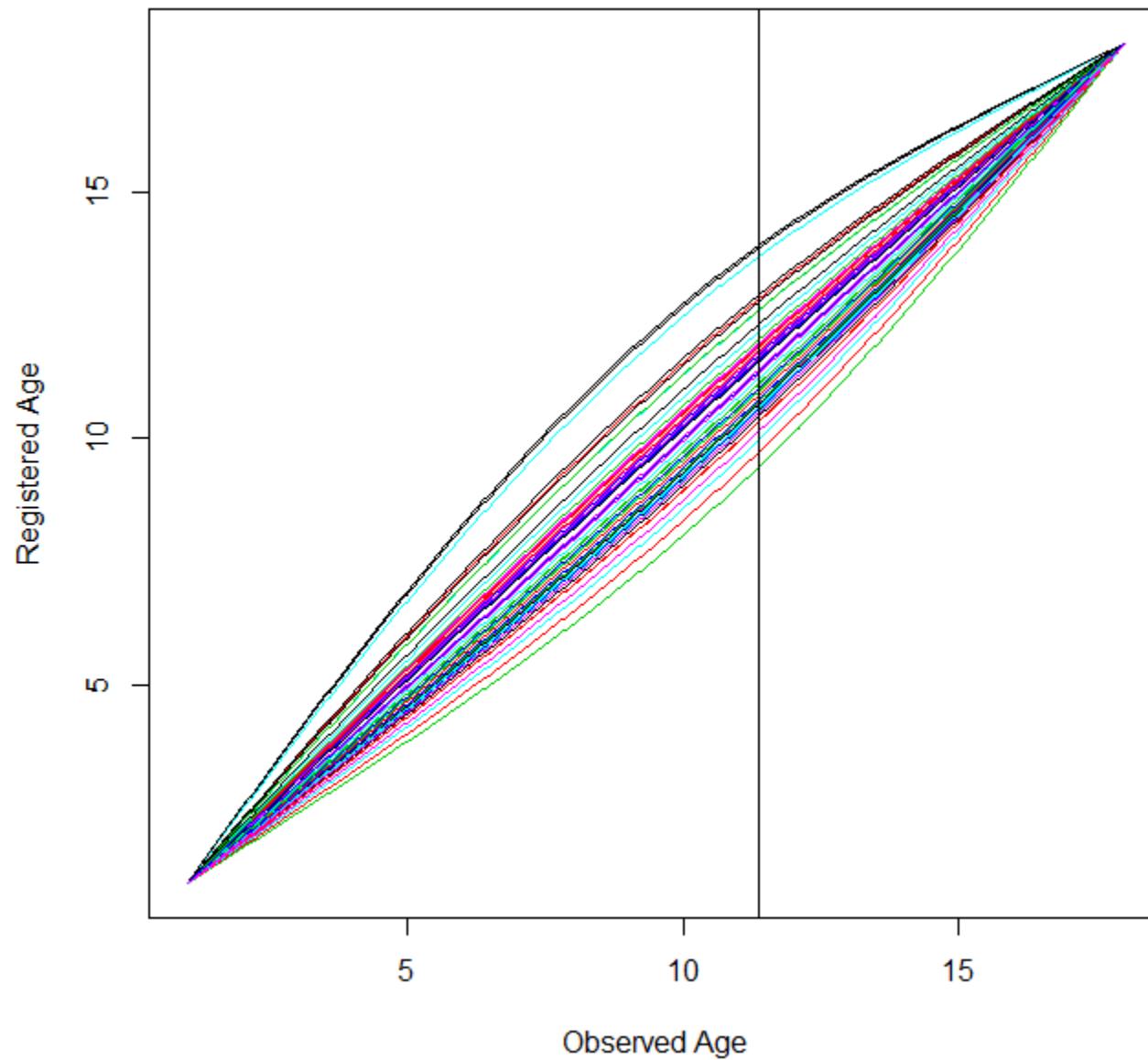
- C_{0i}, C_{1i} fixed by the requirement that $h_i(0) = 0$ and $h_i(T) = T$
- if shift registration is a possibility, C_{0i} can be allowed to pick phase shift
- Clock time corresponds to $W(u) = 0$.
- If $W_i(u)$ is positive, then $h_i(t) > t$, and warped time is growing faster than clock time, and this is what we want if our observed process is running late
- If $W_i(u)$ is negative, then $h_i(t) < t$, and clock time is being slowed down for a process that is running ahead of some target

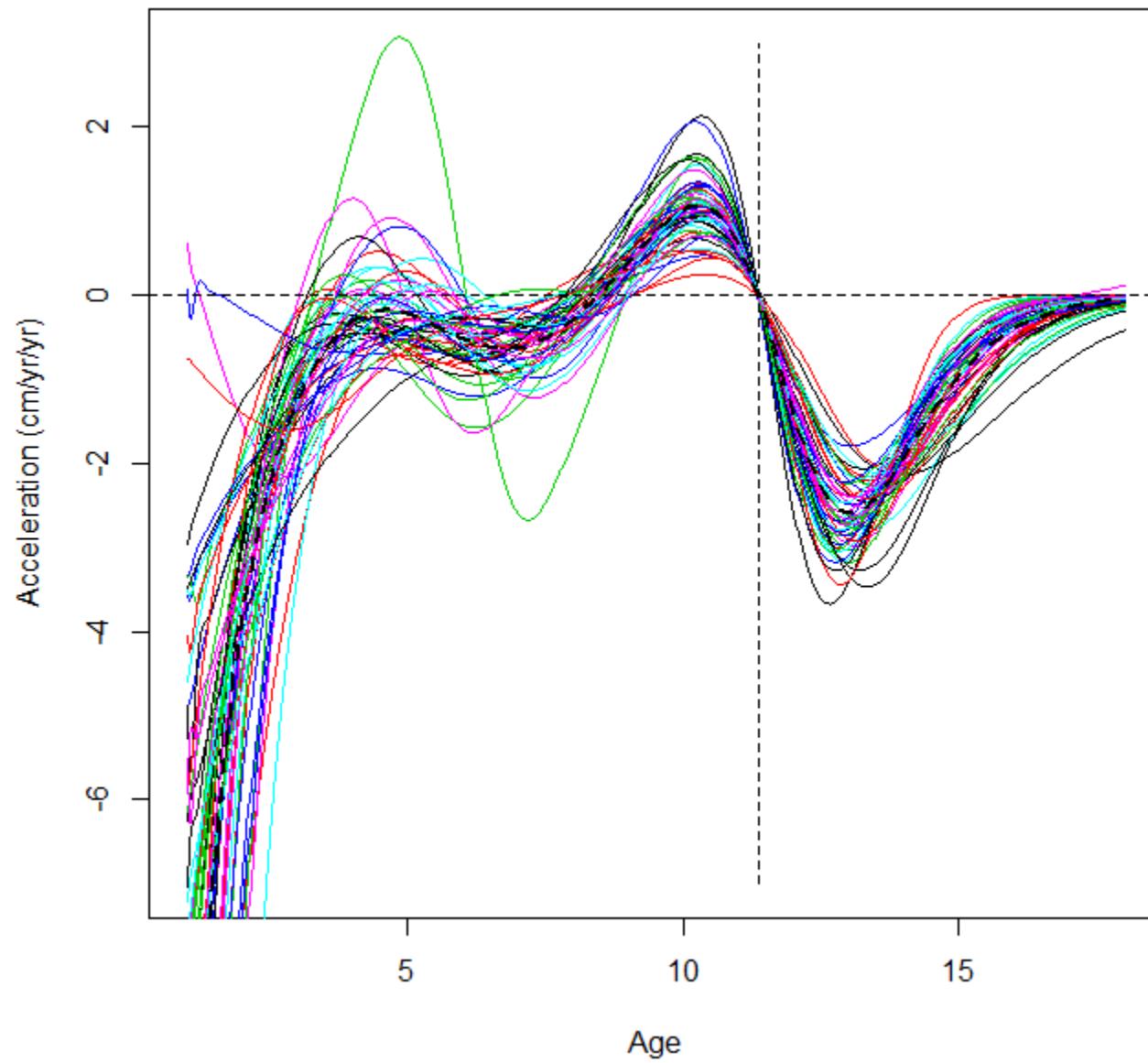


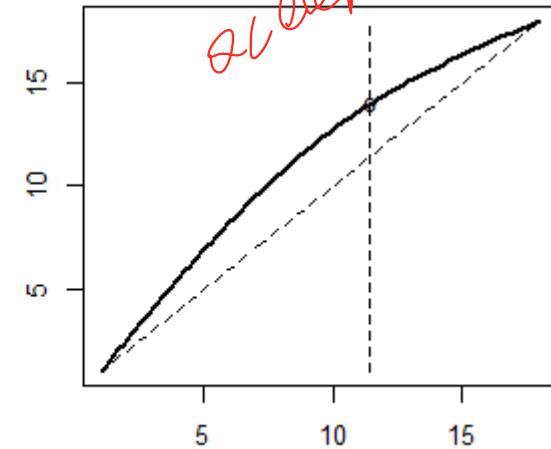
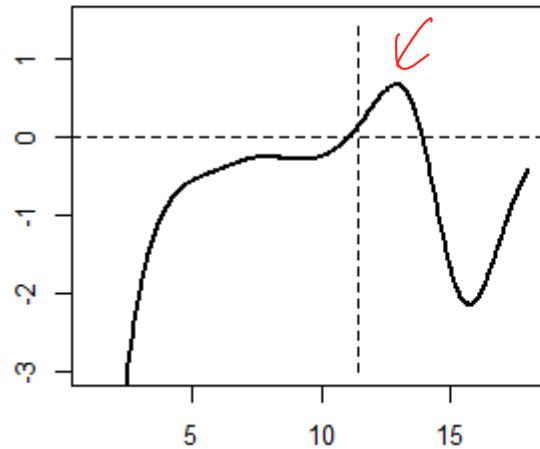
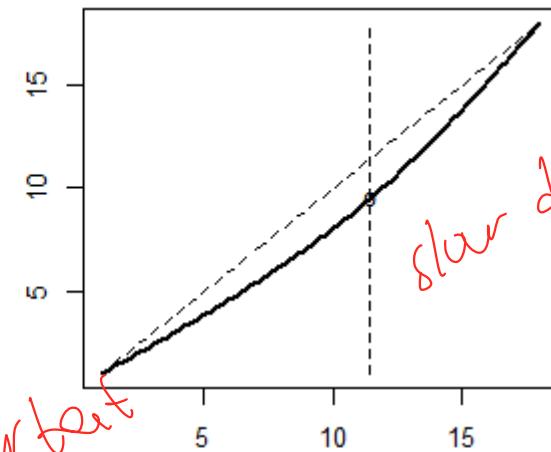
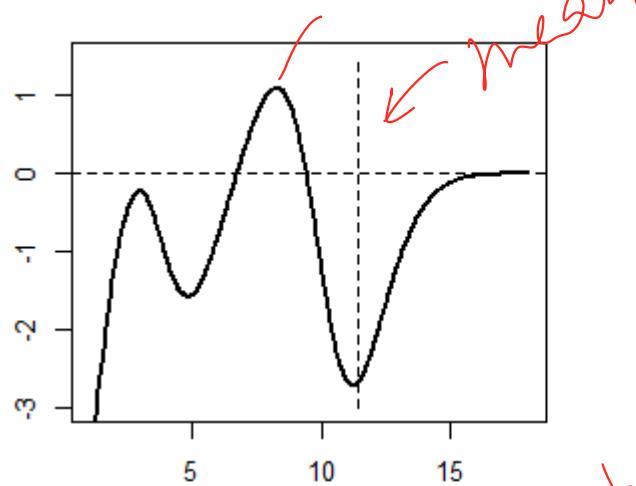












Multidimensional wavelet estimates



RegioneLombardia



92

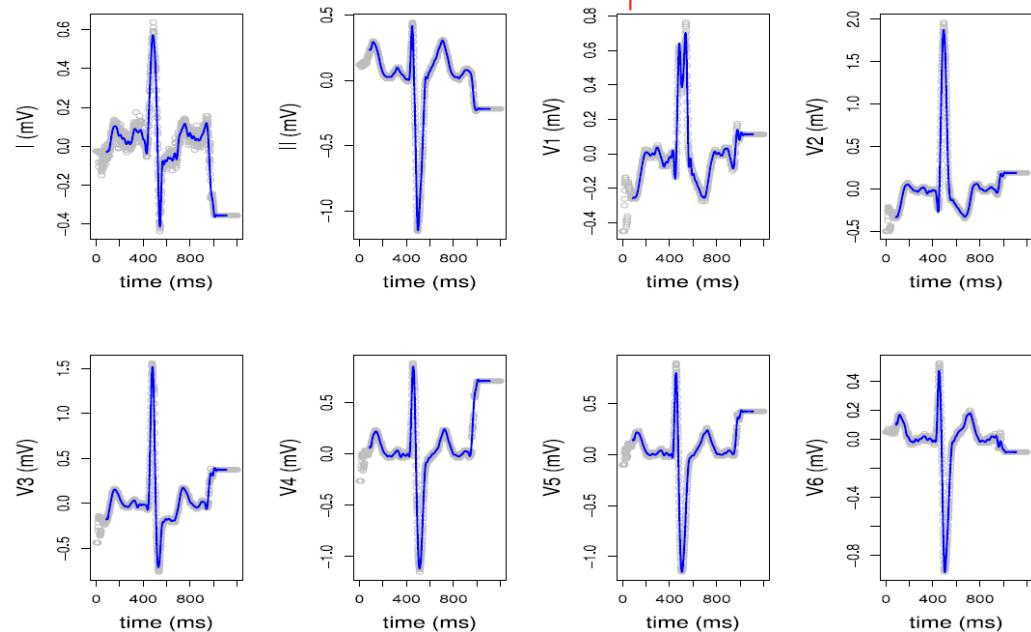
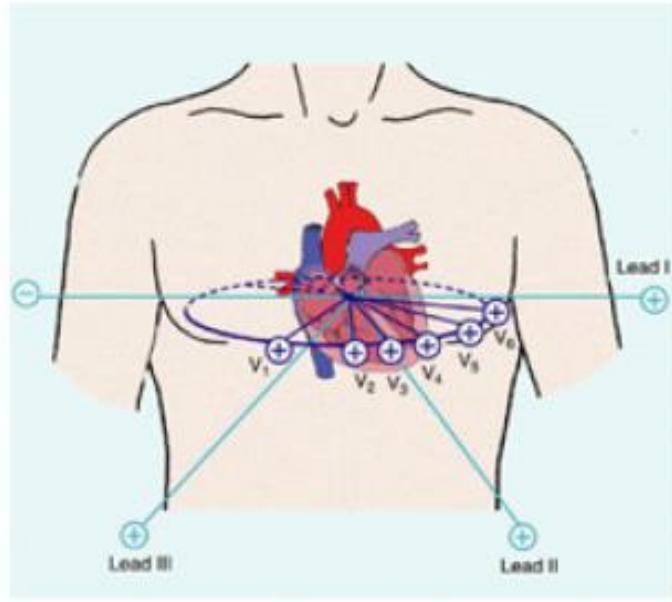
Pigoli and Sangalli 2012 CSDA

Ieva et al. 2013 JRSSC

Case study:

Electro Cardio Gram (ECG) records alignment

measure heart along 8 dimensions
(direction)



Multi-leads ECG: eight-dimensional functional data, whose eight coordinates measure different projections of the heart dynamics in different directions

Multidimensional wavelet estimates

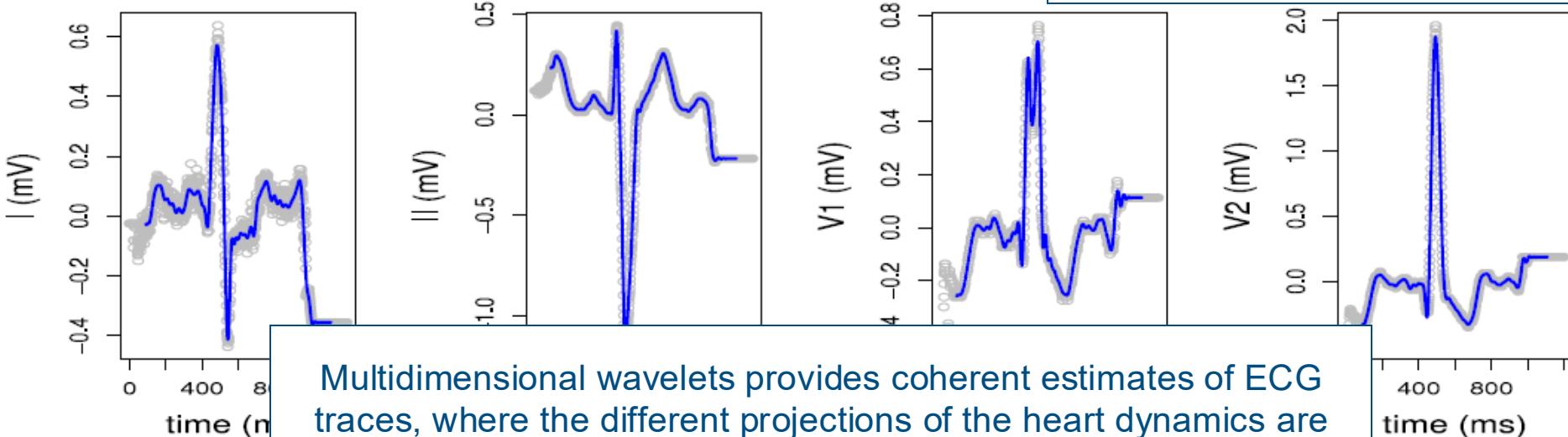


RegioneLombardia

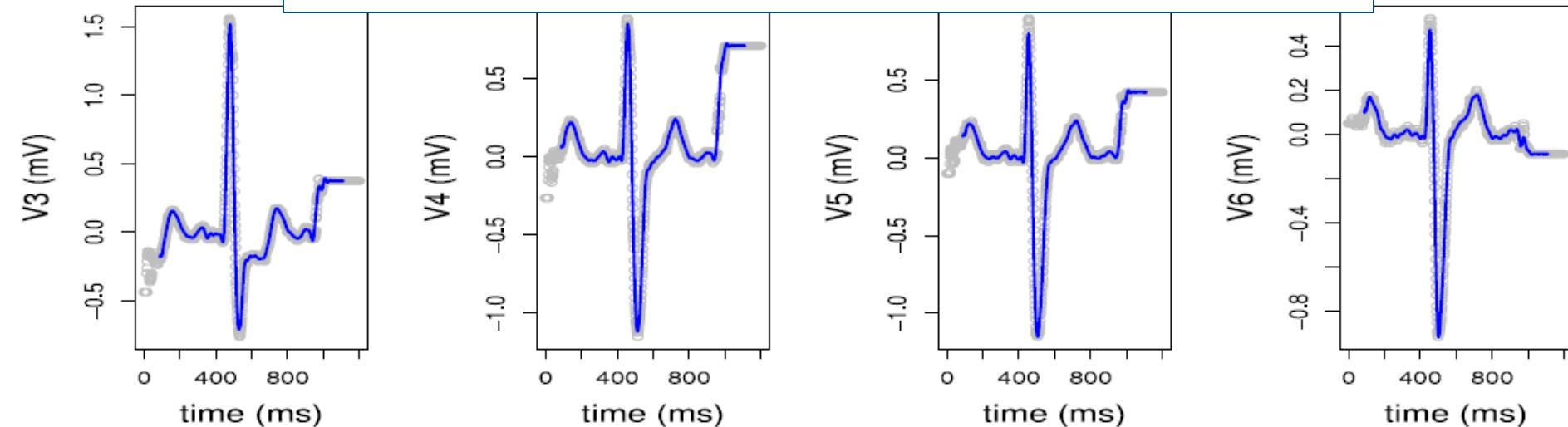


93

Pigoli and Sangalli 2012 CSDA



Multidimensional wavelets provides coherent estimates of ECG traces, where the different projections of the heart dynamics are among them consistent. The estimates also account for the correlation of the components of error on the different projections





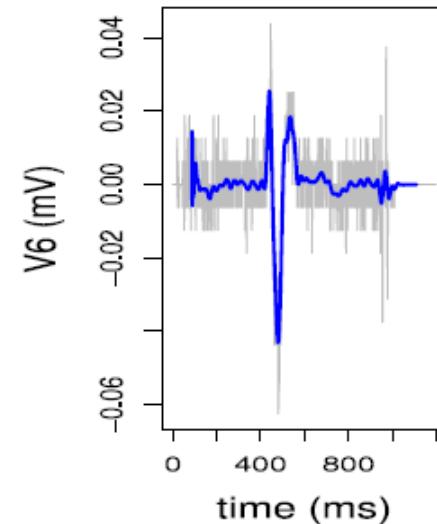
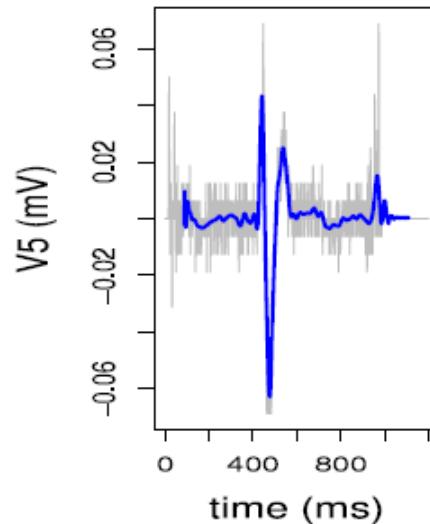
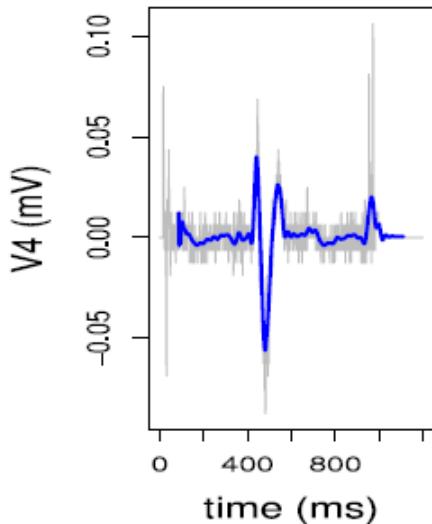
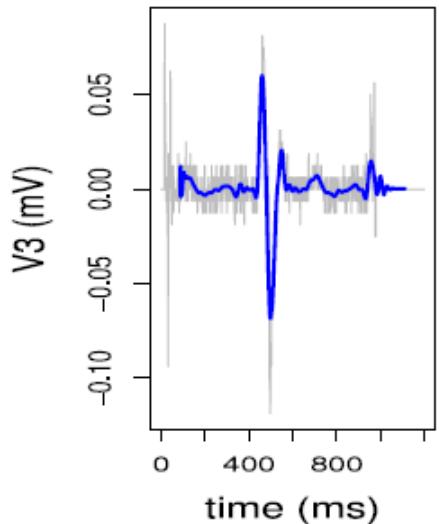
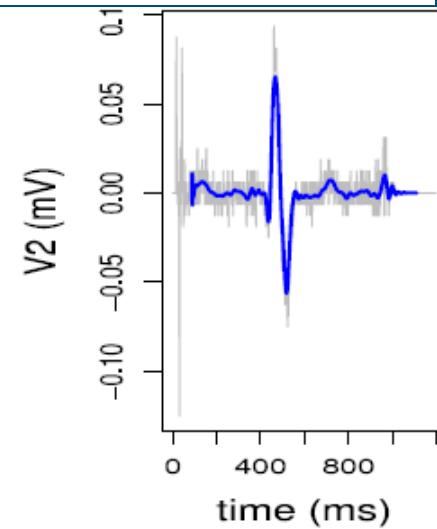
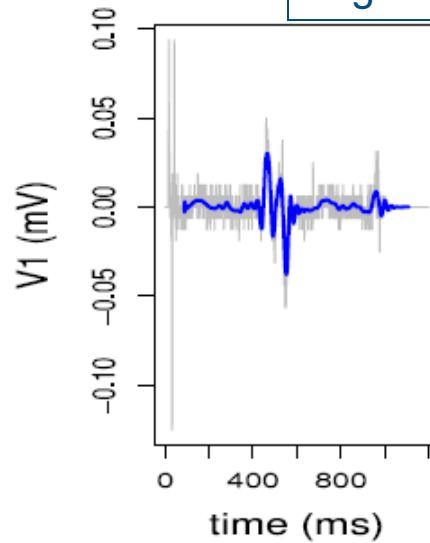
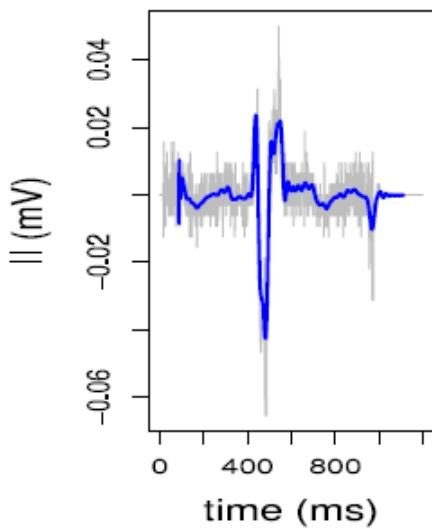
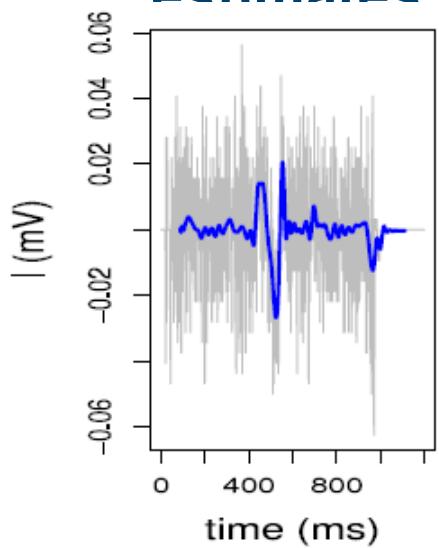
Derivatives of multidimensional wavelet estimates



94

RegioneLombardia

Pigoli and Sangalli 2012 CSDA



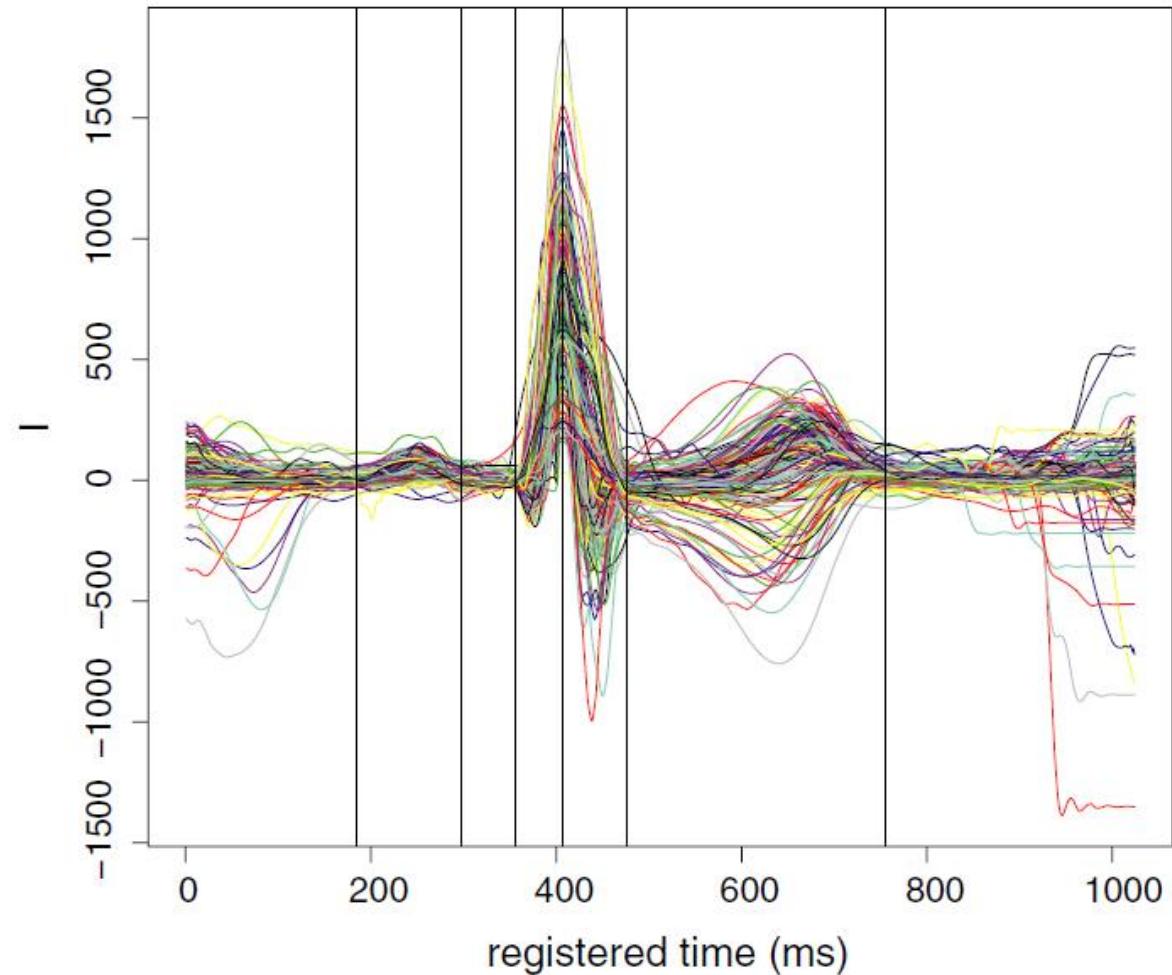
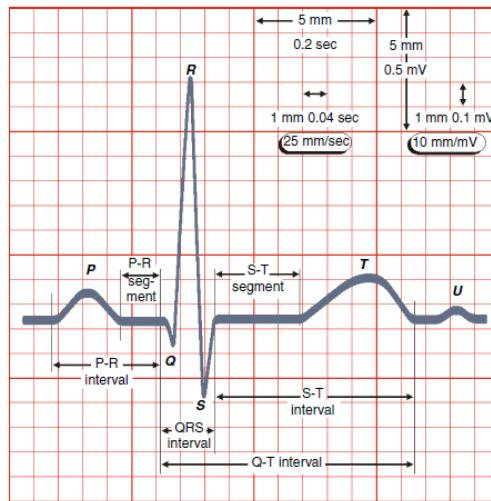


95

RegioneLombardia

leva et al. 2013 JRSSC

The 6 landmarks identify the *P*-wave (*P*, onset, *P*, offset), the *QRS*-complex (*QRS*, onset, *QRS*, offset), the *T*-wave (*T*, offset), the *R*-peak identified on lead I (*I*, peak).



Semi-automatic diagnostic procedure, based on the ECG morphology, that is able to classify physiological and pathological traces

- Landmark-based registration may require significant user input and can be sensitive to the accuracy of the landmark identification.
- In some applications it is not possible to identify well-defined features that can be taken as landmarks

Alternative strategy: Continuous registration

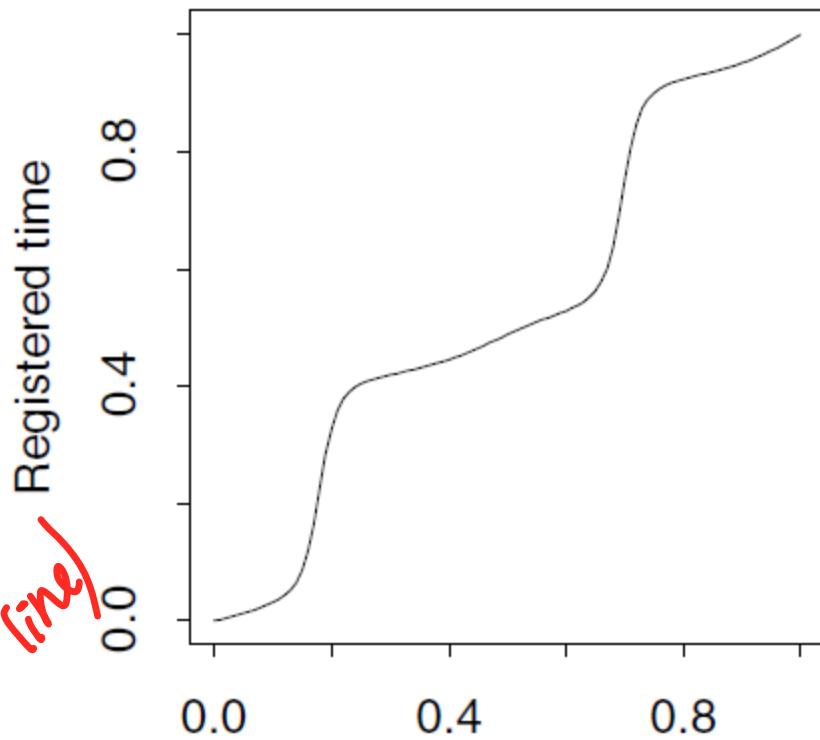
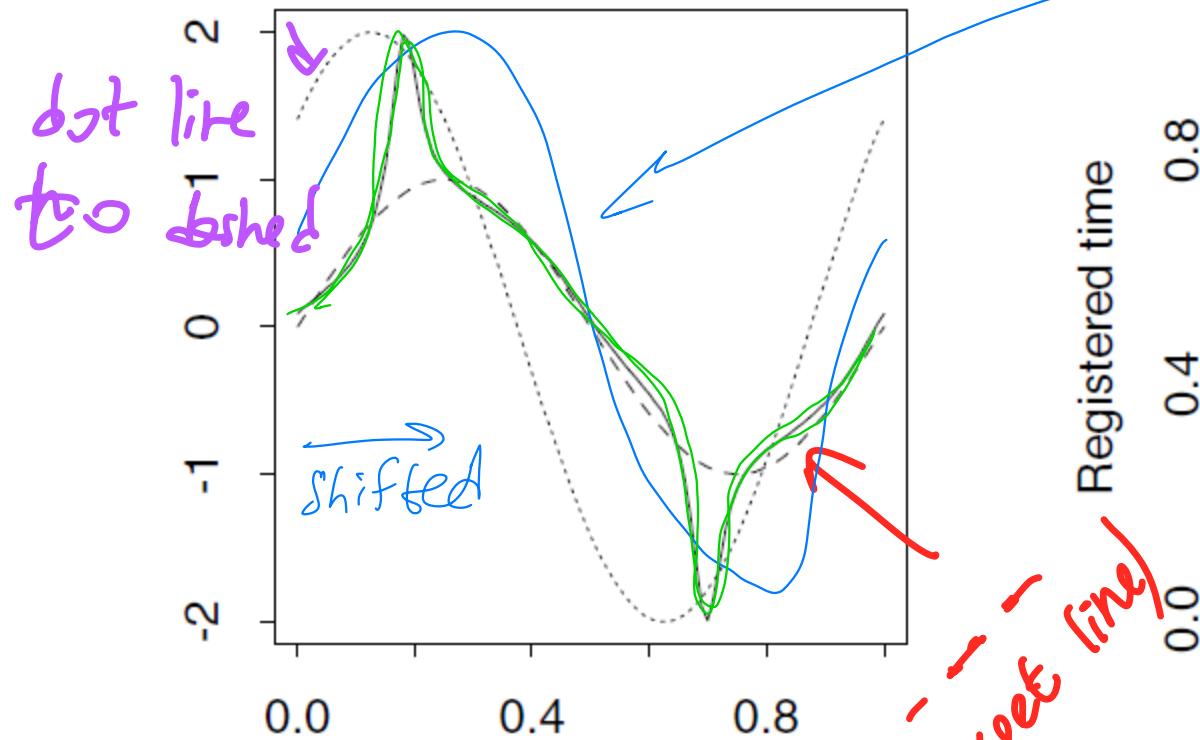
Main idea:

- definition of a suitable distance (or closeness) measure between curves, which measures dissimilarity (or similarity) between curves.
- the curves are thus aligned by warping their time or space abscissa parameters choosing the optimal warping function in some class of admissible warping functions in order to minimize the final distance among the curves or, equivalently, maximize their final similarity.



$\hat{h} = \operatorname{argmin}_h \int C_0(t) - G(h(t))^2 dt$

we try to fit



$h \in W^{1,1}[0,1]$

The problem of decoupling **amplitude** and **phase** variability is not univocally defined.

Different measures of distance/similarity can be considered, and different classes of admissible warping functions (e.g., simple translations or dilations, increasing linear transformations, more complex increasing transformations) leading to different registration results.

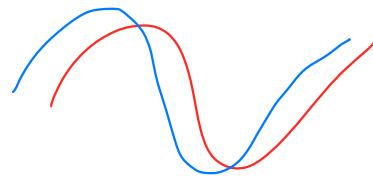
The choice of the couple formed by
dissimilarity/similarity measure & admissible warping functions
defines the distinction between phase variability and amplitude variability in the specific problem under analysis.

This choice must thus be **problem specific**.



(ρ, W) must satisfy properties that ensure that the aligning problem is well-posed and the corresponding procedure is coherent

- ▶ ρ
 - Bounded
 - Reflexive
 - Symmetric
 - Transitive
 - ▶ W
 - Convex vector space
 - Group structure with respect to function composition
 - ▶ (ρ, W)
 - Properties of coherence *shifting shouldn't change*
 - $\rho(\mathbf{c}_1, \mathbf{c}_2) = \rho(\mathbf{c}_1 \circ h, \mathbf{c}_2 \circ h), \quad \forall h \in W$ *W-invariance of similarity index*
 $(\text{Isometry of the group, parallel orbits})$
- $\rho(\mathbf{c}_1 \circ h_1, \mathbf{c}_2 \circ h_2) = \rho(\mathbf{c}_1 \circ h_1 \circ h_2^{-1}, \mathbf{c}_2) = \rho(\mathbf{c}_1, \mathbf{c}_2 \circ h_2 \circ h_1^{-1})$



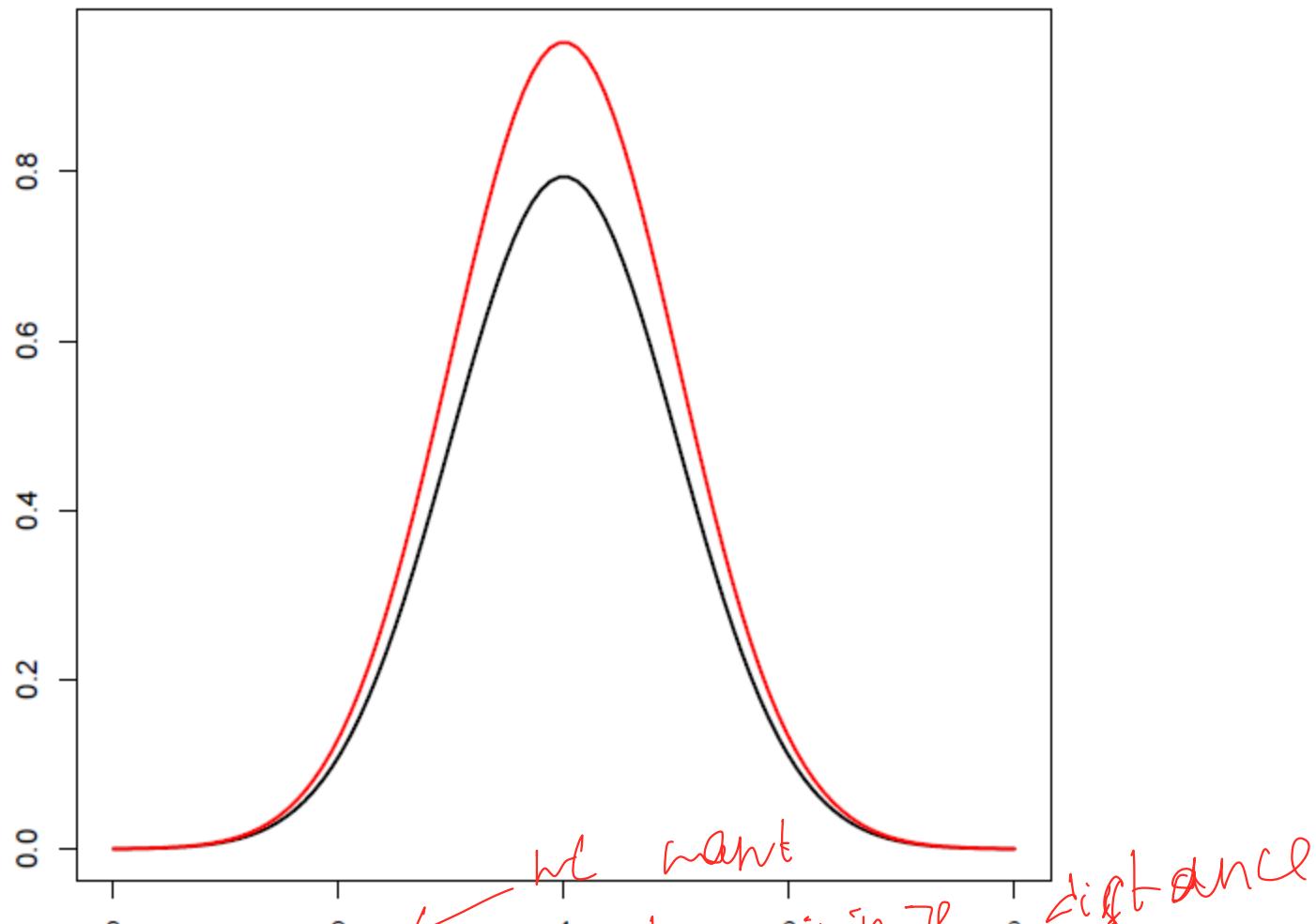
(ρ, W) defines on the considered set of functions \mathcal{C} a partition in equivalence classes

dissimilarity d	warpings W
$\ c_1 - c_2\ $	W_{shift}
$\ c'_1 - c'_2\ $	W_{shift}
$\ (c_1 - \bar{c}_1) - (c_2 - \bar{c}_2)\ $	W_{shift}
$\ (c'_1 - \bar{c}'_1) - (c'_2 - \bar{c}'_2)\ $	W_{shift}
$\left\ \frac{c_1}{\ c_1\ } - \frac{c_2}{\ c_2\ } \right\ $	$W_{affinity}$
$\left\ \frac{c'_1}{\ c'_1\ } - \frac{c'_2}{\ c'_2\ } \right\ $	$W_{affinity}$
$\left\ \text{sign}(c'_1)\sqrt{ c'_1 } - \text{sign}(c'_2)\sqrt{ c'_2 } \right\ $	$W_{diffeomorphism}$



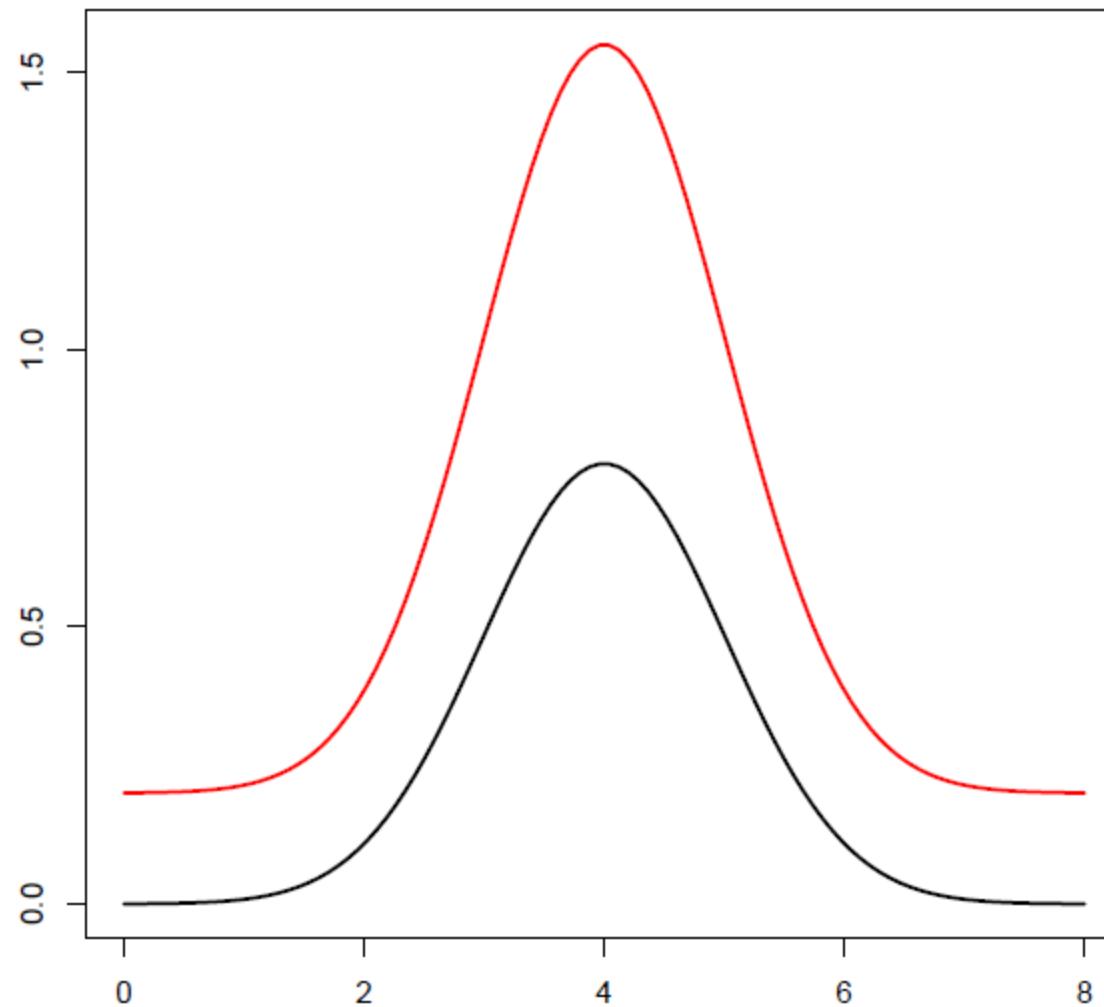


(correlation)
maximum availability
↓



$$\frac{\langle c_0, c_1 \rangle}{\|c_0\| \|c_1\|} = 1$$

$$\left\| \frac{c_0}{\|c_0\|} - \frac{c_1}{\|c_1\|} \right\| = 0 \quad \text{normalized L}_2 \text{ distance}$$



$$\frac{\langle c'_0, c'_1 \rangle}{\|c'_0\| \|c'_1\|} = 1$$

$$\left\| \frac{c'_0}{\|c'_0\|} - \frac{c'_1}{\|c'_1\|} \right\| = 0$$

If a template (prototype) curve φ is known, then it is enough to align each cuve to this template

If the template is unknown then it must be estimated from the data, leading to a complex optimization problem

find $\varphi \in \mathcal{C}$ and $\underline{h} = \{h_1, \dots, h_N\} \subset W$ such that

$$\frac{1}{N} \sum_{i=1}^N \rho(\varphi, \mathbf{c}_i \circ h_i) \geq \frac{1}{N} \sum_{i=1}^N \rho(\psi, \mathbf{c}_i \circ g_i)$$

for any other $\psi \in \mathcal{C}$ and $\underline{g} = \{g_1, \dots, g_N\} \subset W$

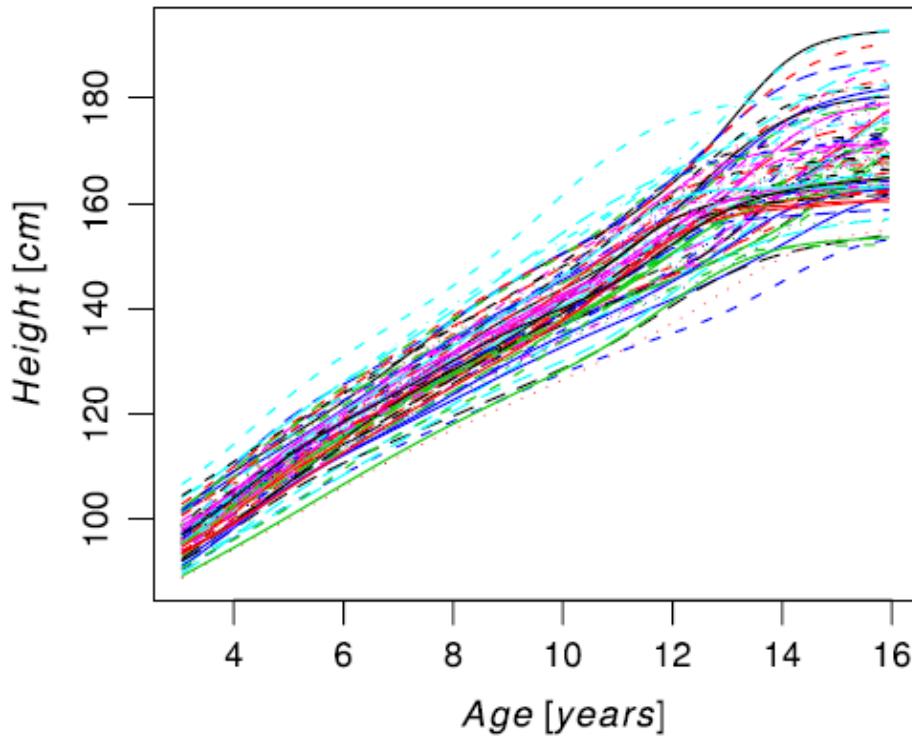
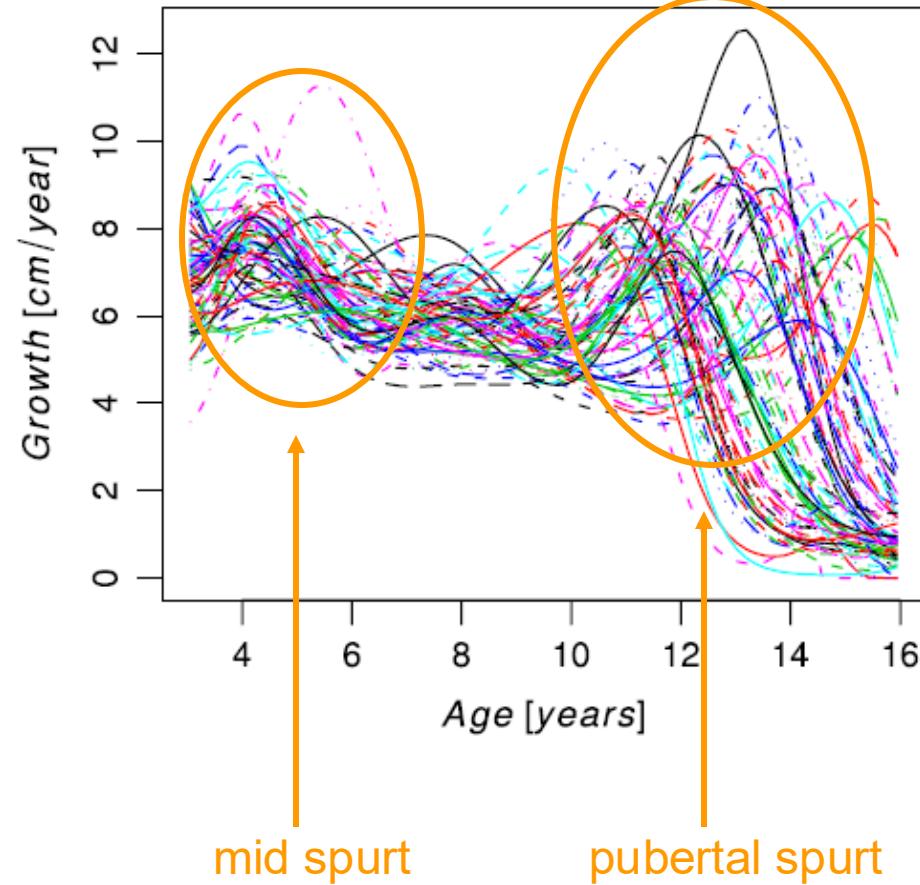
→ Iterative Procrustes procedure that alternates

- **template estimation step:** the template curve is estimated from the curves obtained in the previous alignment step
- **alignment step:** the curves are aligned to the template centerline estimated in the previous template estimation step

Both phase variation and/or the amplitude variation may be associated to the phenomenon (for instance, the pathology) under study!!!

It is thus necessary to study both types of variations to see how they relate to the problem being investigated.



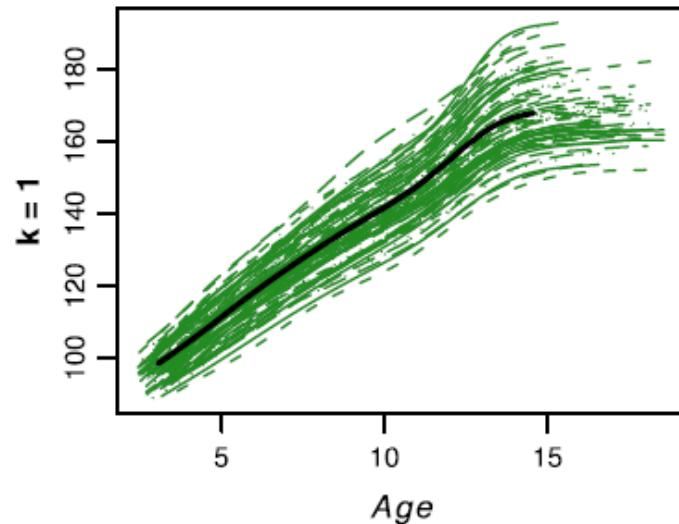
Growth curves**Growth velocities**

93 children, 39 boys and 54 girls

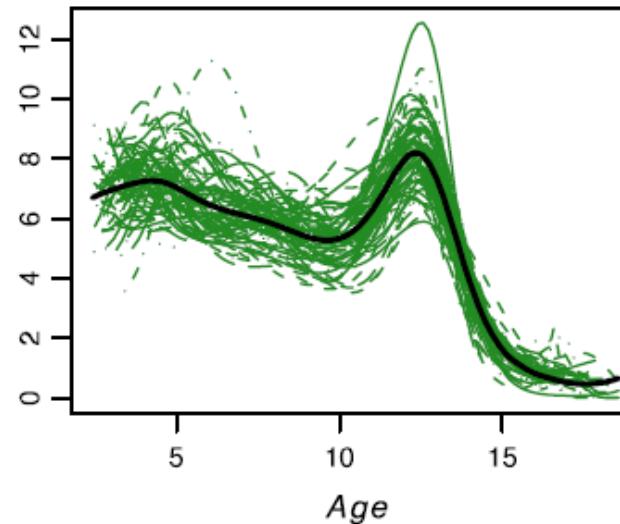
Curves estimated by monotonic cubic regression splines, implemented using the R package *fda*

Does the analysis point out some differences in the growth of boys and girls?

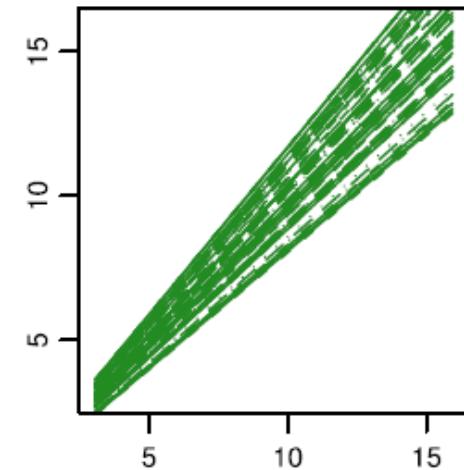
Aligned growth curves



Aligned growth velocities



Warping functions



Results of continuous alignment using the following similarity index and class of warping functions (ρ, W) :

$$\rho(c_i, c_j) = \frac{\int_{S_{ij}} c'_i(s)c'_j(s)ds}{\sqrt{\int_{S_{ij}} c'_i(s)^2 ds} \sqrt{\int_{S_{ij}} c'_j(s)^2 ds}}$$

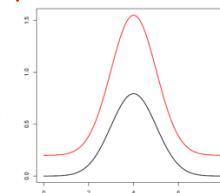
$$\rho(c_i, c_j) = 1 \Leftrightarrow \exists a \in \mathbb{R}, b \in \mathbb{R}^+ : c_i(t) = a + b c_j(t)$$

modify biological starting point

$$W = \{h : h(t) = mt + q \text{ with } m \in \mathbb{R}^+, q \in \mathbb{R}\}$$

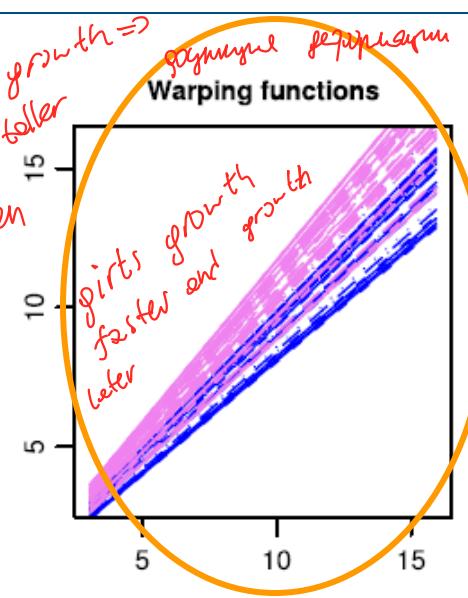
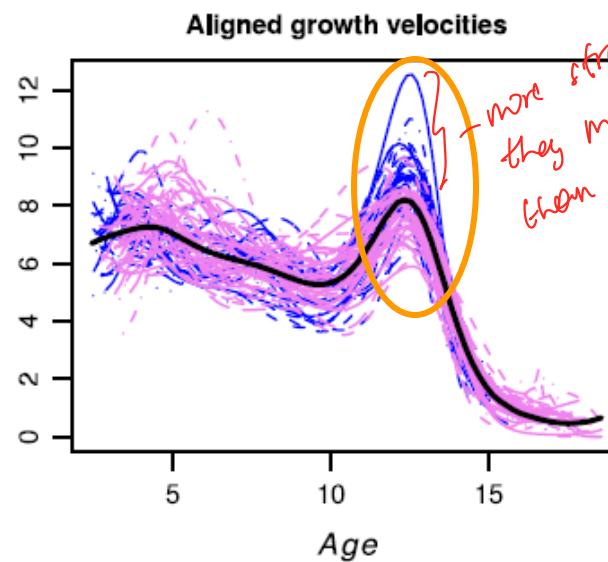
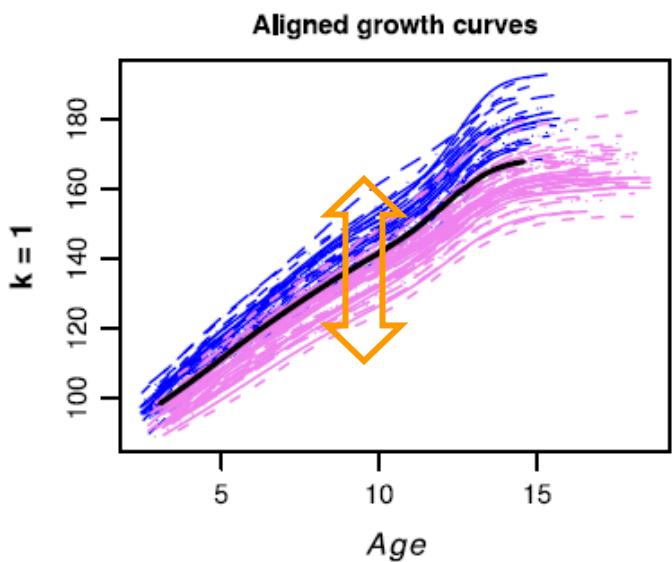
the focus is on growth patterns, rather than on the absolute heights of the children or on their more or less pronounced growths

perfectly aligned



constant modifications of the running speeds of the children biological clocks

Sangalli Secchi Vantini Vitelli 2010 CSDA



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

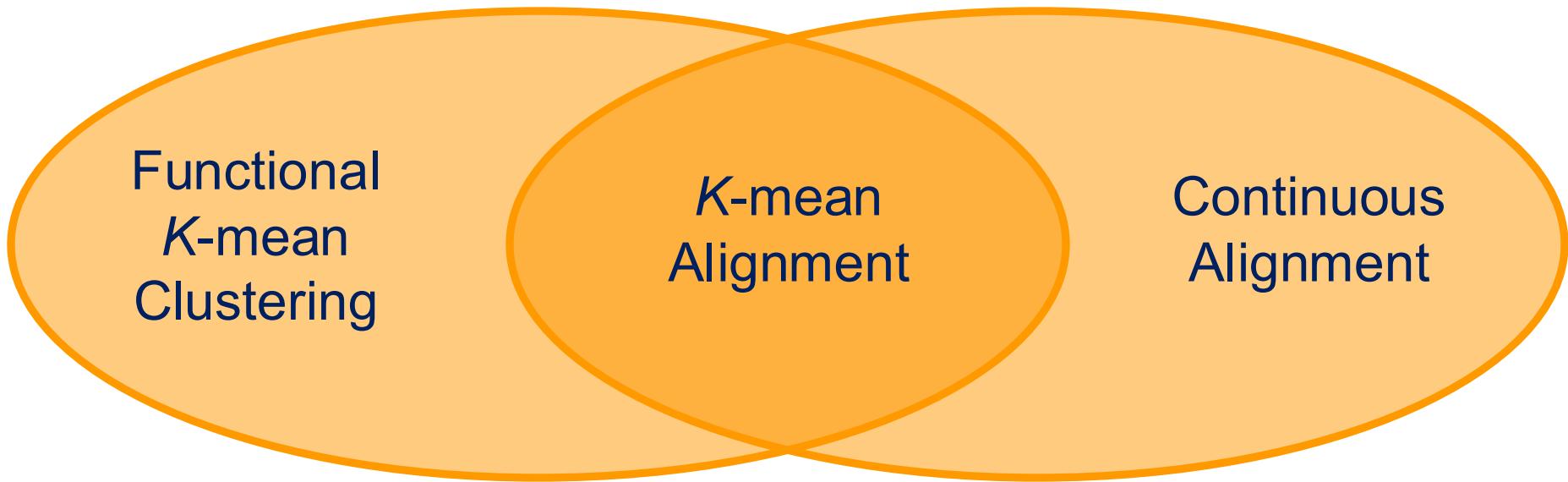


Curve clustering (functional k-mean clustering of curves)

- Heckman, N.E., Zamar, R.H. (2000), "Comparing the shapes of regression functions", Biometrika 87, 135-144.
- Tarpey, T., Kinateder, K.K.J. (2003), "Clustering functional data", J. Classification 20, 93-114.
- Shimizu, N., Mizuta, M. (2007), "Functional clustering and functional principal points", In: Lecture Notes in Artificial Intelligence, vol. 4693. Springer-Verlag, Berlin Heidelberg, 501-508.
- Cuesta-Albertos, J.A., Fraiman, R. (2007), "Impartial trimmed k-means for functional data". Comput. Statist. Data Anal. 51, 4864-4877.

Simultaneous clustering and alignment of curves

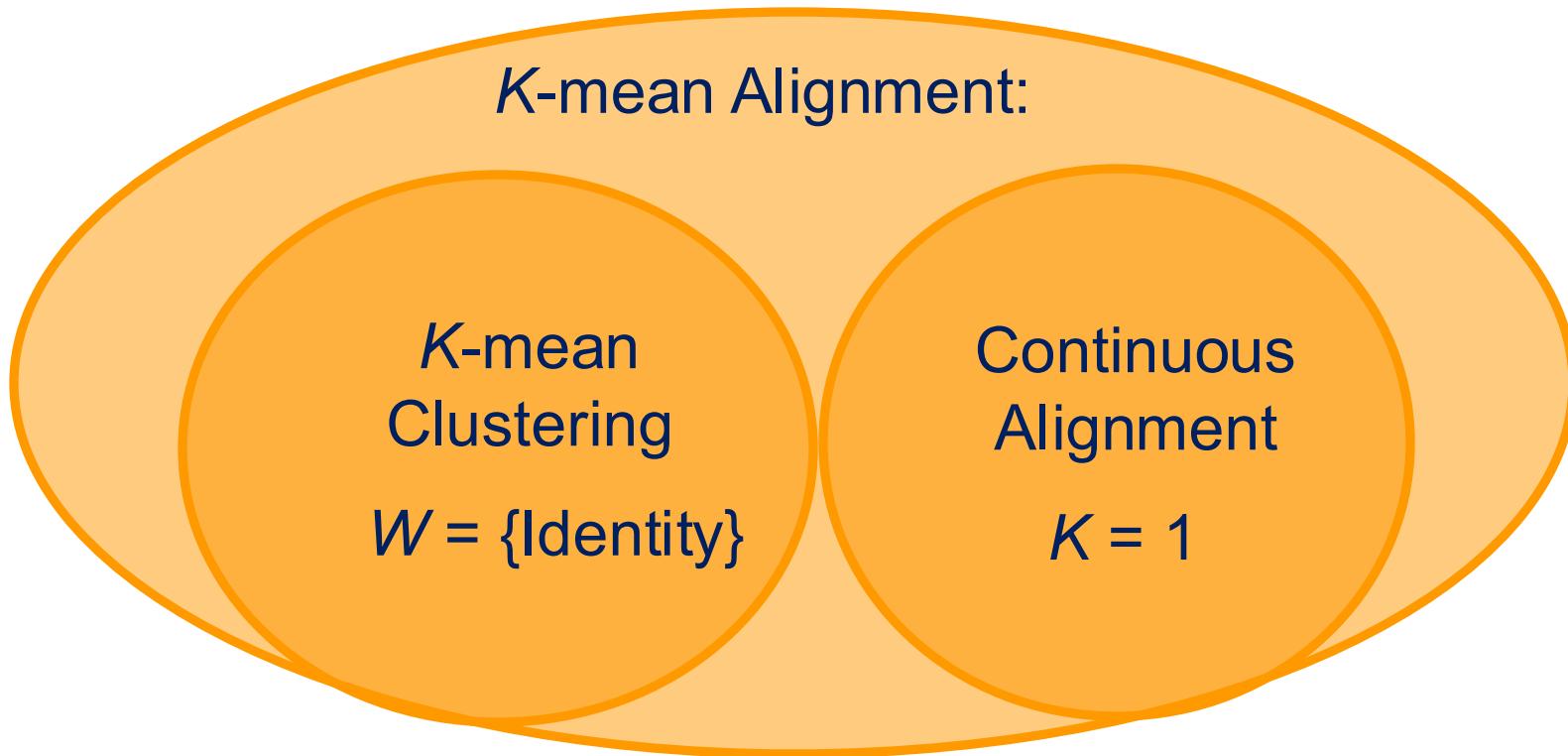
- Tang, R., Müller, H.-G. (2009), "Time-synchronized clustering of gene expression trajectories". Biostatistics 10, 32-45.
- Boudaoud, S., Rix, H., Meste, O. (2010), "Core Shape modelling of a set of curves". Comput. Statist. Data Anal. 54, 308-325.
- Liu, X., Yang, M.C.K. (2009), "Simultaneous curve registration and clustering for functional data". Comput. Statist. Data Anal. 53, 1361-1376.
- Sangalli, L.M., Secchi, P., Vantini, S., Vitelli, V. (2010), "K-means alignment for curve clustering", Comput. Statist. Data Anal., 54, 1219-1233.
- Sangalli, L.M., Secchi, P., Vantini, S. (2014), "Analysis of AneuRisk65 data: K-mean Alignment", Electronic Journal of Statistics, 8 (2), 1891-1904.



→ *K*-mean Clustering
with warping allowed

→ Continuous Alignment
with *K* templates

Code for *K*-mean alignment: R package `fdaCluster` (formerly `fdakma`), available from CRAN



Code for K-mean alignment: R package `fdaCluster` (formerly `fdakma`), available from CRAN

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**

(disclosing clustering in the phase)

Aligning and clustering a set of N curves

$$\{\mathbf{c}_1, \dots, \mathbf{c}_N\}$$

with respect to k template curves

$$\underline{\varphi} = \{\varphi_1, \dots, \varphi_k\}$$

Domain of attraction of φ_j

$$\Delta_j(\underline{\varphi}) = \{\mathbf{c} \in \mathcal{C} : \sup_{h \in W} \rho(\varphi_j, \mathbf{c} \circ h) \geq \sup_{h \in W} \rho(\varphi_r, \mathbf{c} \circ h), \forall r \neq j\}, \quad j = 1, \dots, k$$

Labelling function

$\lambda(\underline{\varphi}, \mathbf{c})$: indicates a cluster the curve \mathbf{c} should be assigned to

$\lambda(\underline{\varphi}, \mathbf{c}) = j$: the similarity index obtained by aligning \mathbf{c} to φ_j is at least as large as the similarity index obtained by aligning \mathbf{c} to any other template φ_r , with $r \neq j$

$\varphi_{\lambda(\underline{\varphi}, \mathbf{c})}$: indicates a template the curve \mathbf{c} can be best aligned to

Curve clustering when curves are misaligned

Trivial case: $\underline{\varphi} = \{\varphi_1, \dots, \varphi_k\}$ known

In order to cluster and align the set of N curves $\{\mathbf{c}_1, \dots, \mathbf{c}_N\}$ with respect to $\underline{\varphi}$:

for $i = 1, \dots, N$

- assign \mathbf{c}_i to the cluster $\lambda(\underline{\varphi}, \mathbf{c}_i)$
- align it to the corresponding template $\varphi_{\lambda(\underline{\varphi}, \mathbf{c}_i)}$

Non-trivial case: $\underline{\varphi} = \{\varphi_1, \dots, \varphi_k\}$ unknown

need to be themselves estimated from the data, leading to a complex optimization problem

Given $\{\mathbf{c}_1, \dots, \mathbf{c}_N\} \subset \mathcal{C}$, find $\underline{\varphi} = \{\varphi_1, \dots, \varphi_k\} \subset \mathcal{C}$,
 $\{\lambda_1, \dots, \lambda_N\} \subset \{1, \dots, k\}$ and $\underline{h} = \{h_1, \dots, h_N\} \subset W$
 that maximise $\frac{1}{N} \sum_{i=1}^N \rho(\varphi_{\lambda_i}, \mathbf{c}_i \circ h_i)$

An approximate solution to
 this optimization problem is
 given by the following
 iterative procedure

$\underline{\varphi}^{[q-1]} = \{\varphi_1^{[q-1]}, \dots, \varphi_k^{[q-1]}\}$: set of templates after iteration $q-1$

$\{\mathbf{c}_1^{[q-1]}, \dots, \mathbf{c}_{N^{[q-1]}}\}$: N curves aligned and clustered to $\underline{\varphi}^{[q-1]}$

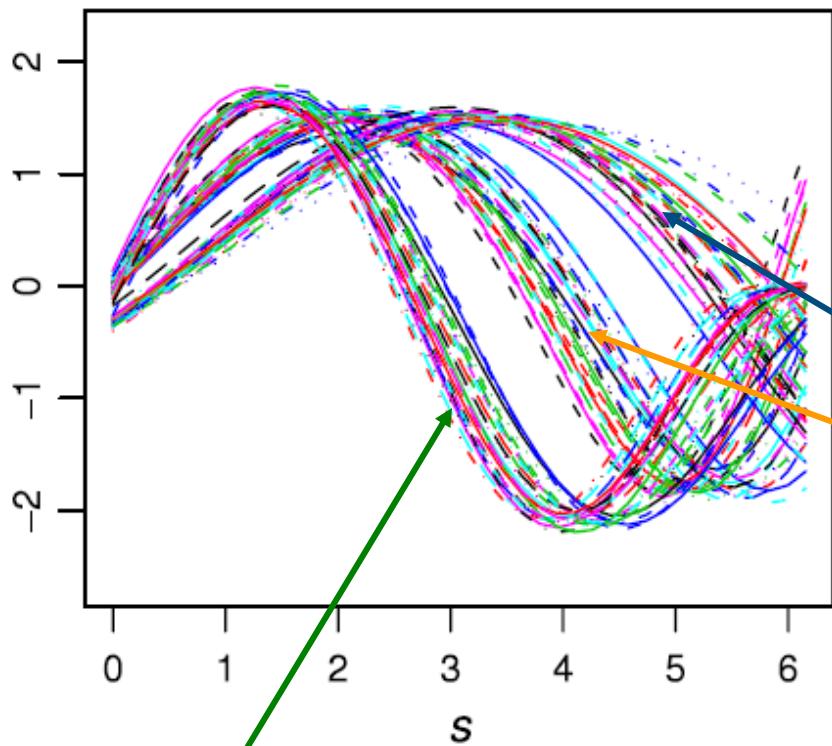
Template identification step. For $j = 1, \dots, k$, the template of the j th cluster $\varphi_j^{[q]}$ is estimated using all curves assigned to cluster j at iteration $q-1$.

$$\varphi_j^{[q]} = \arg \max_{\varphi \in \mathcal{C}} \sum_{i: \lambda_i=j} \rho(\varphi, \mathbf{c}_i^{[q-1]})$$

Assignment and alignment step. The set of curves $\{\mathbf{c}_1^{[q-1]}, \dots, \mathbf{c}_{N^{[q-1]}}\}$ is clustered and aligned to the set of templates $\underline{\varphi}^{[q]} = \{\varphi_1^{[q]}, \dots, \varphi_k^{[q]}\}$.

Normalization step. For $j = 1, \dots, k$, all curves assigned to cluster j are warped along a common warping function, so that the average warping undergone by curves assigned to the same cluster is the identity transformation (thus avoiding the drifting apart of clusters or the global drifting of the overall set of curves).

The algorithm is stopped when, in the assignment and alignment step, the increments of the similarity indexes are all lower than a fixed threshold.



2 AMONG THEM OFFERS
probably have generated these data?

ONE has associated a
further CLUSTERING IN
THE PHASE

$$1 * \sin(s) + 1 * \sin\left(\frac{s^2}{2\pi}\right)$$

$$2 * \sin(s) - 1 * \sin\left(\frac{s^2}{2\pi}\right) + (1 + \varepsilon_{4i})s + (1 + \varepsilon_{2i}) * \sin\left(\frac{(\varepsilon_{3i} + (1 + \varepsilon_{4i})s)^2}{2\pi}\right)$$

$$\rho(c_i, c_j) = \frac{\int_{S_{ij}} c'_i(s)c'_j(s)ds}{\sqrt{\int_{S_{ij}} c'_i(s)^2 ds} \sqrt{\int_{S_{ij}} c'_j(s)^2 ds}}$$

$$\rho(c_i, c_j) = 1 \Leftrightarrow \exists a \in \mathbb{R}, b \in \mathbb{R}^+ : c_i(t) = a + b c_j(t)$$

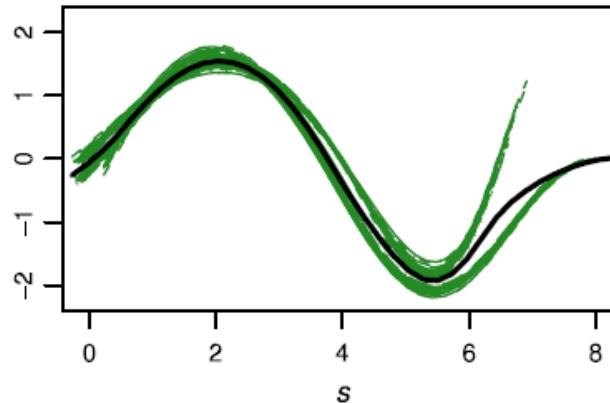
$$W = \{h : h(t) = mt + q \text{ with } m \in \mathbb{R}^+, q \in \mathbb{R}\}$$



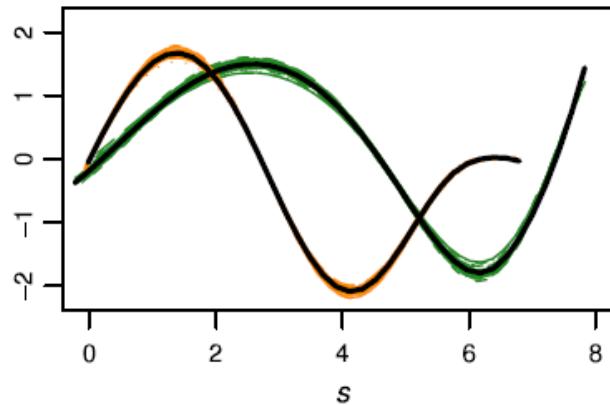


Aligned and clustered curves

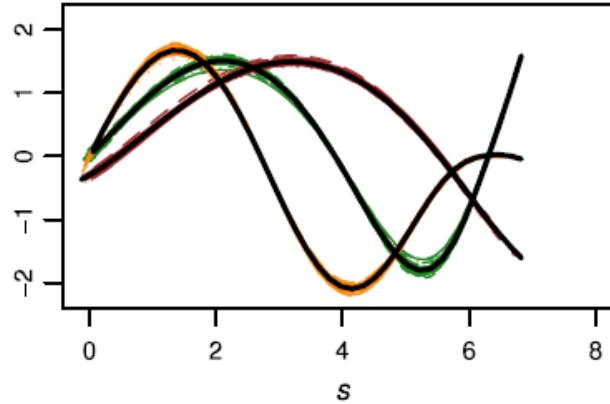
$K = 1$



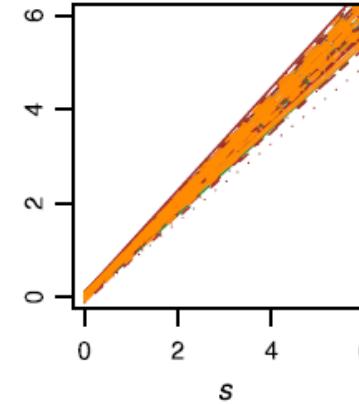
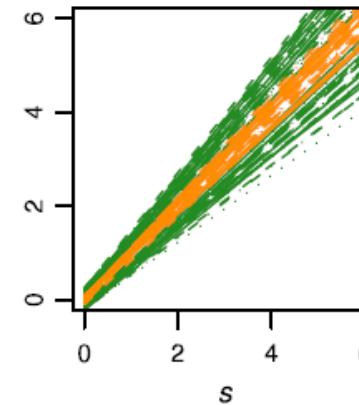
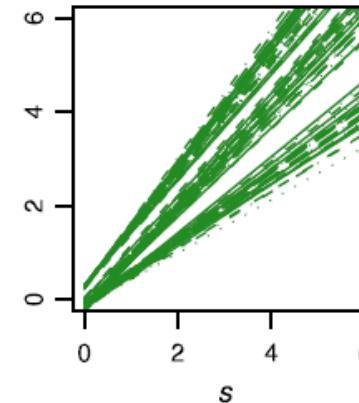
$K = 2$

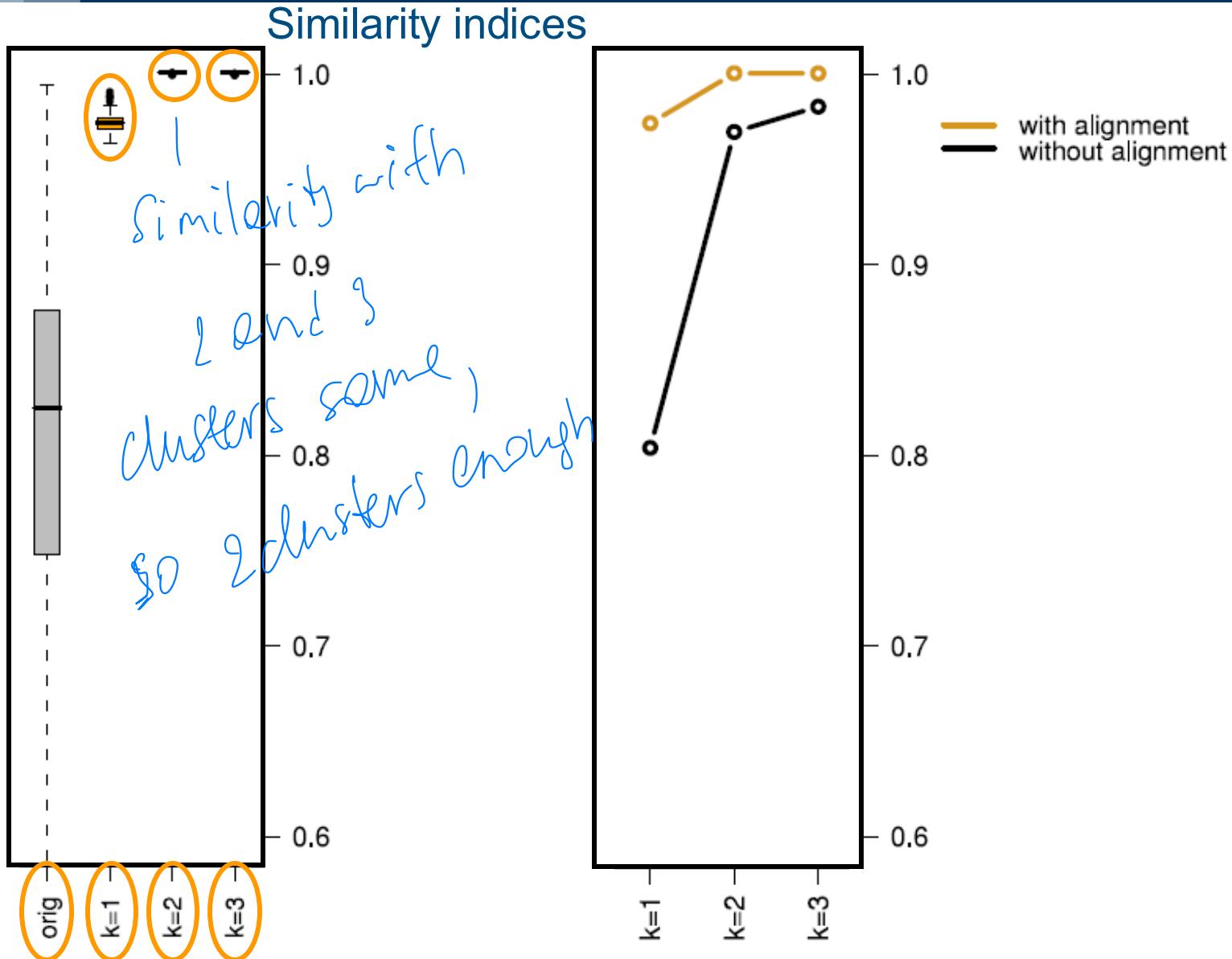


$K = 3$

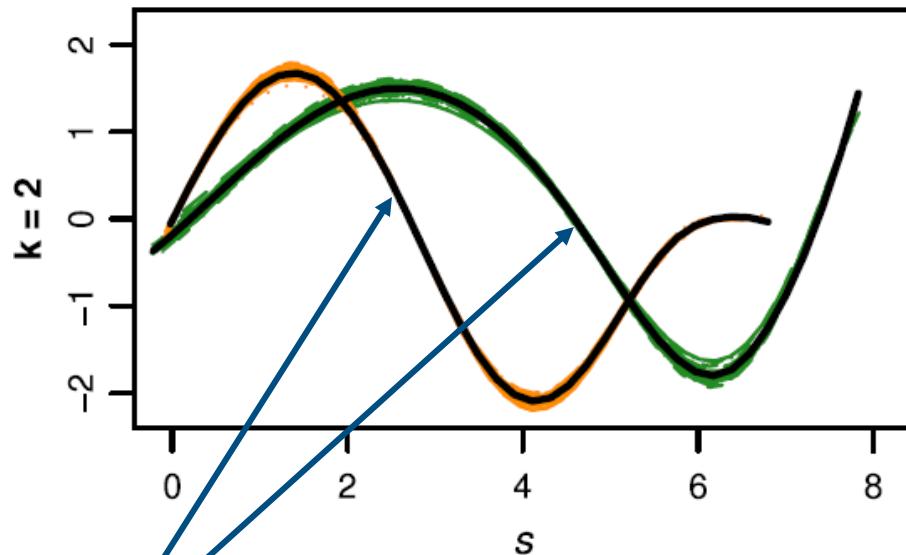


Warping functions¹²⁰



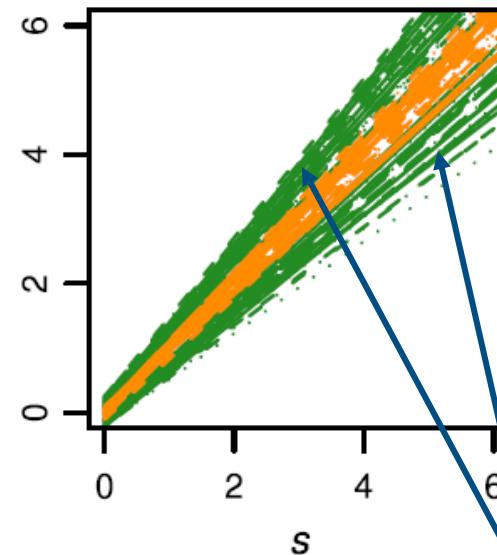


Aligned and clustered curves



k-mean alignment is able to efficiently detect true amplitude clusters and also to disclose clustering structures in the phase

Warping functions





Mathematical Biosciences Institute

[Home](#) [About](#) [News](#) [Events](#) [People](#) [Visitors](#) [Postdoctoral](#) [Committees](#) [Institute Partners](#) [Education](#) [Publications](#)

Calendar
Apply for Workshop
Workshops
Visiting Lecturer Program
Colloquia/Seminars
Summer Programs
Public Lectures
Visitor Info
Annual Programs
Propose MBI Programs

CTW: Statistics of Time Warpings and Phase Variations (November 13-16, 2012)

Organizers: J. S. Marron (UNC), J. O. Ramsay (McGill), L. Sangalli (Politecnico di Milano), A. Srivastava (Florida State)

[Description](#) [Schedule](#) [Participants](#) [Titles & Abstracts](#) [Resources](#) [Apply for Event](#)
[Flyer](#)

Background: A common feature of functional measurements of data over time, space and other continua, is that salient features in the resulting curves and surfaces vary in position from one recording to the next. For example, the growth patterns of children vary in the timing of puberty, human movements in activities like handwriting and golf swings speed up and slow down from one instance to another, seasonal events like hurricanes arrive early some years and late in others, and traffic jams vary in location over city streets from one day to another. At the same time, each of the events can also vary in intensity. We refer to positional variation as phase variation, and intensity variation as amplitude variation. It is now evident that many processes unfold over a system time that not only does not unroll at the same rate as physical clock time, but also tends to vary in a significant way from one realization of a functional event to another.

The registration or alignment of features in curves and images by smooth, one-to-one transformations of time or space, respectively, is an emerging hot topic that presents many challenges. From its beginnings with dynamic time warping in the late 50's, followed by the landmark registration methods of Fred Bookstein, the registration of brain images to a fixed atlas, and its widespread application in functional data analysis, statisticians have realized that nonlinear phase

The Electronic Journal of Statistics

**Special Section on Statistics of Time
Warpings and Phase Variation,
Volume 8 , Number 2, 2014**



POLITECNICO DI MILANO



Politecnico di Milano
Applied Statistics
May 2025



An introduction to functional data analysis Part 3 – Alignment – Case Study: AneuRisk

Laura M. SANGALLI

MOX - Dipartimento di Matematica, Politecnico di Milano

<https://sangalli.faculty.polimi.it/>



AneuRisk data

Available from: <https://mox.polimi.it/research-areas/statistics/research/high-dimensional-complex-data-functional-spatial-and-object-data/aneurisk/>

Main references:

- Sangalli, L.M., Secchi, P., Vantini, S. (2014a), “AneuRisk65: three-dimensional cerebral vascular geometries”, *Electronic Journal of Statistics*, 8, 2, 1879–1890.
- Sangalli, L.M., Secchi, P., Vantini, S. (2014b), “Analysis of AneuRisk65 data: Kmean Alignment”, *Electronic Journal of Statistics*, 8, 2, 1891–1904.
- Sangalli, L.M., Secchi, P., Vantini, S., Vitelli, V. (2010), “K-mean alignment for curve clustering”, *Computational Statistics and Data Analysis*, 54, pp. 1219-1233.
- Sangalli, L.M., Secchi, P., Vantini, S., Veneziani, A. (2009), “Efficient estimation of 3-dimensional curves and their derivatives by free-knot regression splines, applied to the analysis of inner carotid artery centerlines”, *Journal of the Royal Statistical Society Ser C*, 58, 285-306.
- Sangalli, L.M., Secchi, P., Vantini, S., Veneziani, A. (2009), “A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery”, *Journal of the American Statistical Association*, 104, 37-48.

Code for K-mean alignment: R package fdakma, available from CRAN

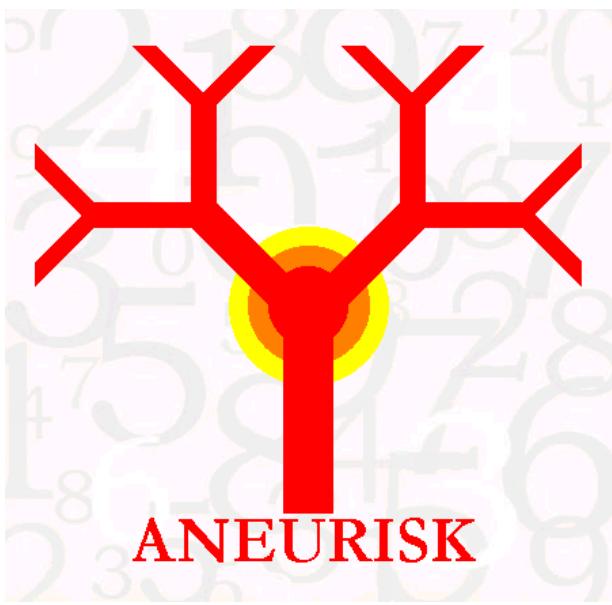
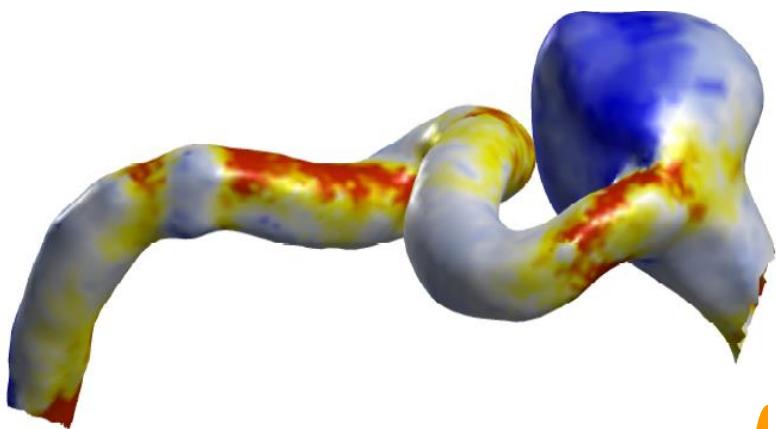




The ANEURISK Project

<https://statistics.mox.polimi.it/aneurisk/>

126



A CONJECTURE

The pathogenesis of cerebral aneurysms is conditioned by the geometry of the cerebral vessels through its effects on blood fluid dynamics



Statistics

Numerical Analysis

Bio-Engineering

Computer Sciences

Neurosurgery

Neuroradiology



Numerical Analysis



Alessandro Veneziani

Principal Investigator



Tiziano Passerini



Marina Piccinelli



Statistics



Piercesare Secchi



Simone Vantini



Laura Sangalli



Valeria Vitelli

(now at University of Oslo)



Edoardo Boccardi

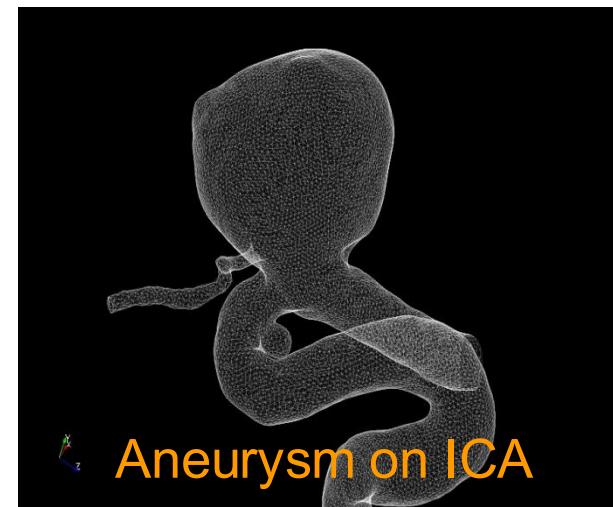
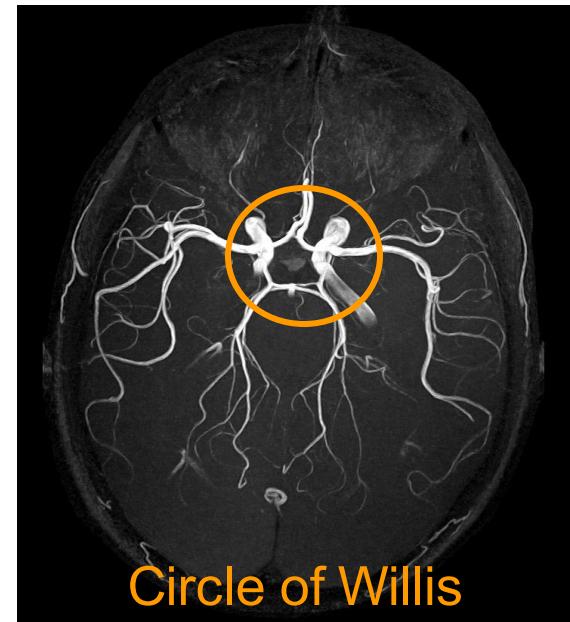
Medicine



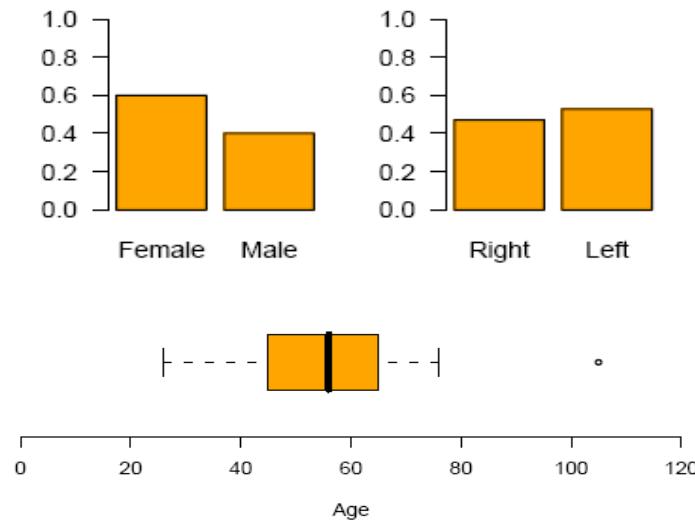
- Cerebral aneurysms: deformations of cerebral arteries, mostly placed on vessels belonging to or connected to the Circle of Willis (Selvity system of blood) coming to brain

Aneurysms EPIDEMIOLOGY

- Incidence rate of cerebral aneurysms:
1/20 people
- Incidence rate of ruptured cerebral aneurysms per year:
1/10000 people per year
- Mortality due to a ruptured aneurysm:
> 50%: Out of 9 patients with a ruptured aneurysm:
 - 3 are expected to die before arriving at the hospital
 - 2 to die after having arrived at the hospital
 - 2 to survive with permanent cerebral damages
 - 2 to survive without permanent cerebral damages



Observational Study conducted at Ospedale Ca' Granda Niguarda – Milano relative to 65 patients hospitalized from September 2002 to October 2005.

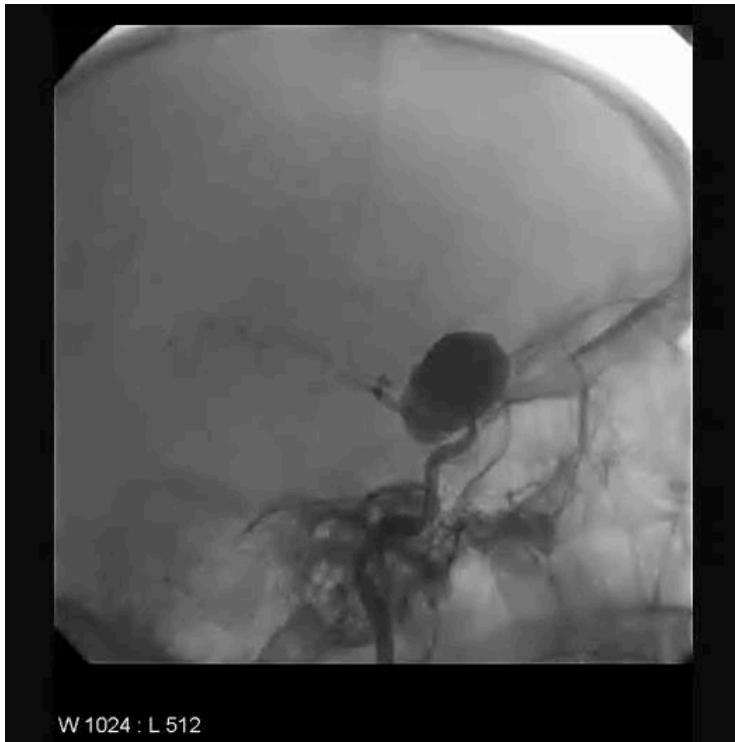


Upper group	Lower group	
Aneurym at or after ICA biforc	Aneurysm before ICA biforc	No aneurysms
33	25	7

Data: 3D-angiographies

Observational Study conducted at Ospedale Ca' Granda Niguarda – Milano relative to 65 patients hospitalized from September 2002 to October 2005.

Sequence of X-Rays



3D-array of
gray scaled pixels



From X-rays to Centerlines and Local Maximal Inscribed Sphere Radius

131

Contrast Fluid Injections



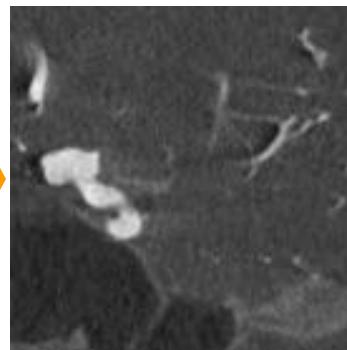
1

X-rays (one direction)



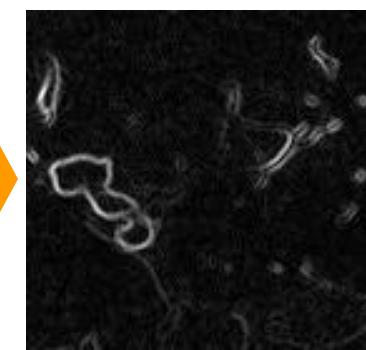
2

3d-array (one slice)



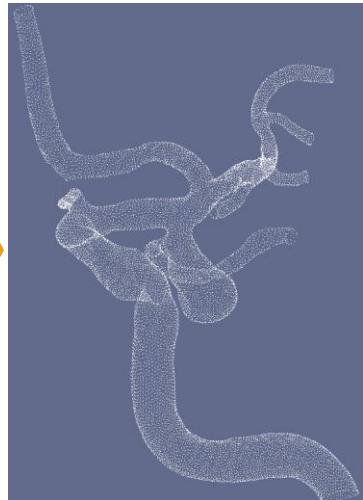
3

Gradient 3d-array (one slice)



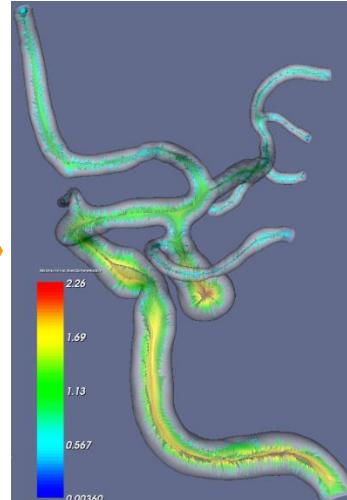
4

Surface Points



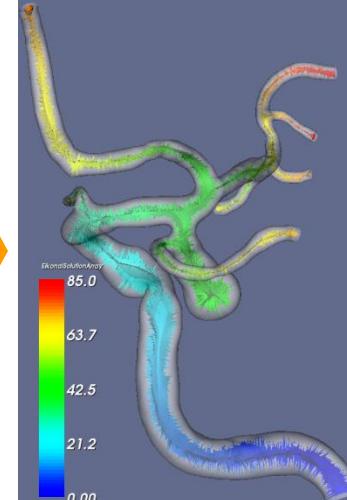
4

Voronoi Diagram



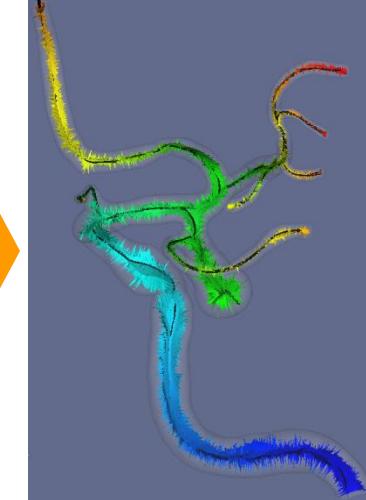
5

Eikonal Equation

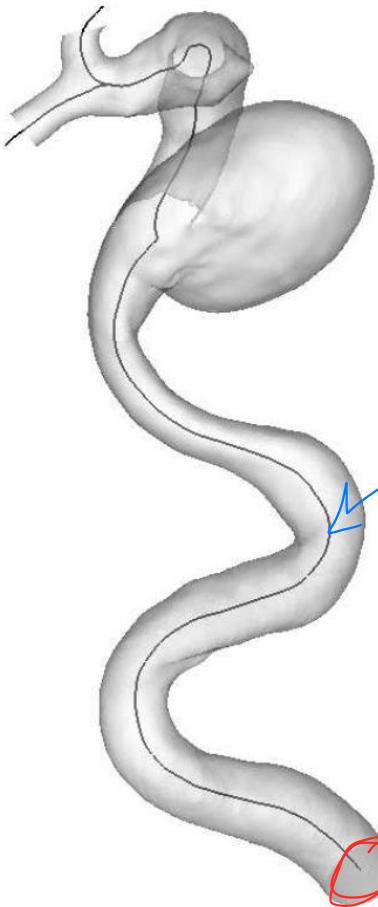


6

Centerline+MISR



7



Focus on Internal Carotid Artery (ICA)

For each patient i elicitation of 3-spatial coordinates of ICA centerline

$$\{(x_{ij}, y_{ij}, z_{ij}) : j = 1, \dots, n_i\}$$

and vessel radius

$$\{R_{ij} : j = 1, \dots, n_i\}$$

alone a fine grid of points

$$(350 \leq n_i \leq 1380)$$

Approximate curvilinear
Preprocessing:

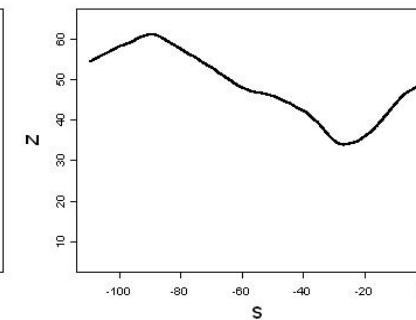
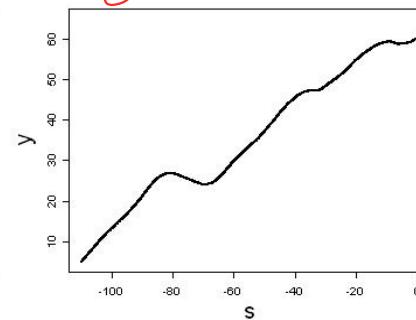
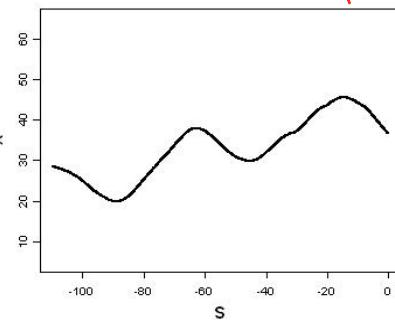
Image reconstruction

$$s_{ij} - s_{ij-1} = -\sqrt{(x_{ij} - x_{ij-1})^2 + (y_{ij} - y_{ij-1})^2 + (z_{ij} - z_{ij-1})^2}, \quad j = 2, \dots, n_i$$

Two geometric quantities that greatly influence the haemodynamics: vessel radius and curvature (curvature of its centerline)

s depending on $x/y/z$

COORDINATES
PATIENT 1



Very high signal-to-noise ratio
Fine grid of observed points

Preprocessing:
accurate curve estimates

b-spline basis system for the vector space

$\{b_{r,m}^{[k]}(s) : r = 1, \dots, m + n_k\}$ of splines of order m
with knot vector $\mathbf{k} = (k_1, \dots, k_{n_k})$

Previously we had knot points and basis function,
now in 3-d dimension we have vectors

Functional estimates of the 3-spatial coordinates
by 3D free-knot regression splines

$$(\hat{x}(s), \hat{y}(s), \hat{z}(s))$$

$$\hat{x}(s) = \sum_{r=1}^{m+n_k} \hat{\lambda}_r^{[x]} b_{r,m}^{[\hat{\mathbf{k}}]}(s) \quad \hat{y}(s) = \sum_{r=1}^{m+n_k} \hat{\lambda}_r^{[y]} b_{r,m}^{[\hat{\mathbf{k}}]}(s) \quad \hat{z}(s) = \sum_{r=1}^{m+n_k} \hat{\lambda}_r^{[z]} b_{r,m}^{[\hat{\mathbf{k}}]}(s)$$

FIND

$$\hat{n}_k, \hat{\mathbf{k}} = (\hat{k}_1(s), \dots, \hat{k}_{n_k}(s)), \hat{\lambda}^{[x]}, \hat{\lambda}^{[y]}, \hat{\lambda}^{[z]}$$

by minimizing

*optimal number
of knots* *coefficient of basis
functions*

$$\sum_{j=1}^n \left(x_j - \sum_{r=1}^{m+n_k} \lambda_r^{[x]} b_{r,m}^{[\mathbf{k}]}(s_j) \right)^2 + \sum_{j=1}^n \left(y_j - \sum_{r=1}^{m+n_k} \lambda_r^{[y]} b_{r,m}^{[\mathbf{k}]}(s_j) \right)^2 + \sum_{j=1}^n \left(z_j - \sum_{r=1}^{m+n_k} \lambda_r^{[z]} b_{r,m}^{[\mathbf{k}]}(s_j) \right)^2 + \mathcal{C}(m+n_k)$$

FIX

$$m=5$$

to obtain smooth estimates of the curvature (function of second derivative)

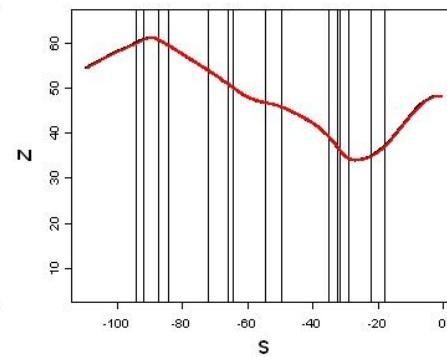
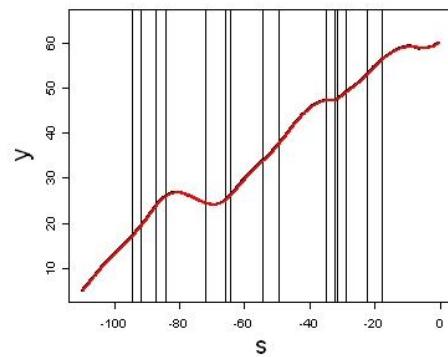
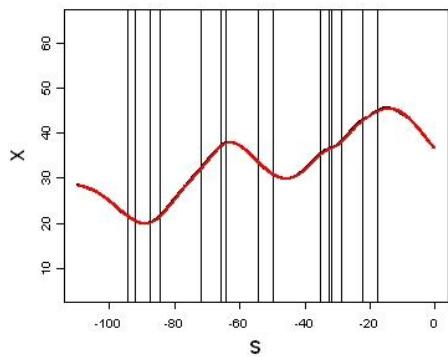


Accurate curve estimation by 3D free-knot splines

136

Sangalli Secchi Vantini Veneziani, 2009, JRSS-C

Curve estimate



Derivatives of splines are still splines with the same knot vector and coefficients directly computed from the coefficients of the original spline

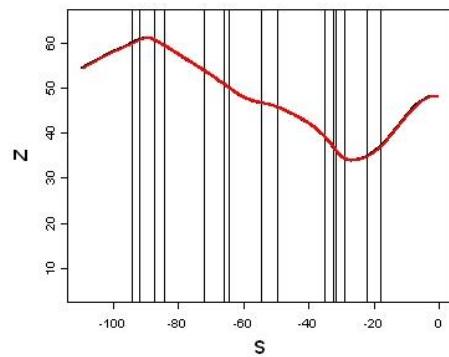
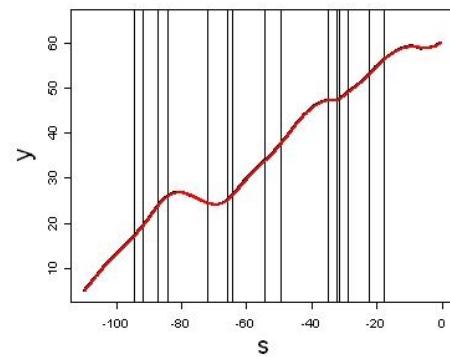
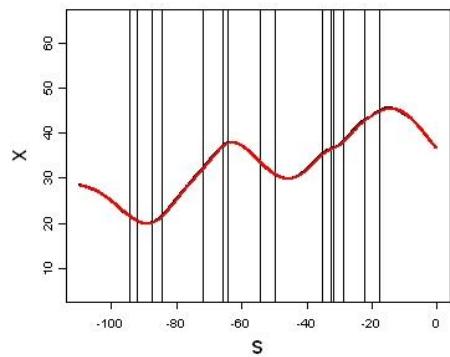


Accurate curve estimation by 3D free-knot splines

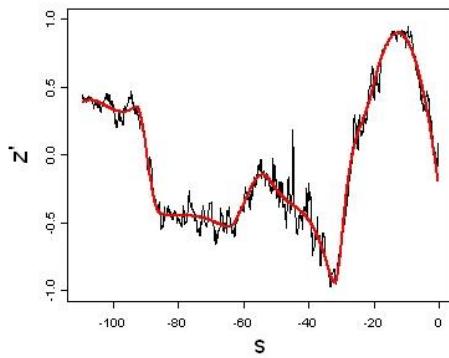
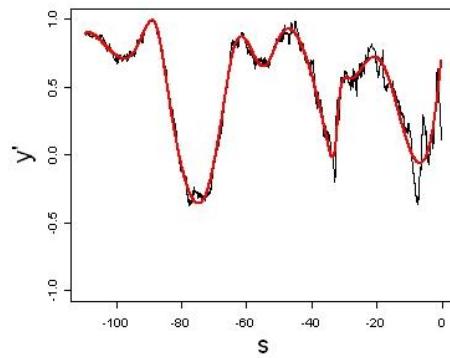
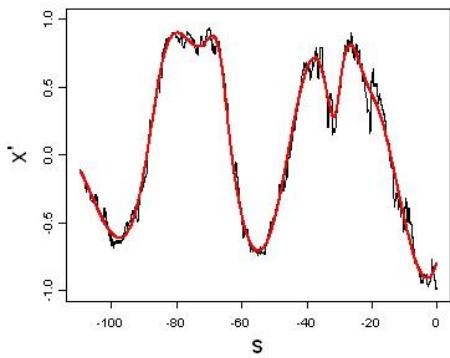
137

Sangalli Secchi Vantini Veneziani, 2009, JRSS-C

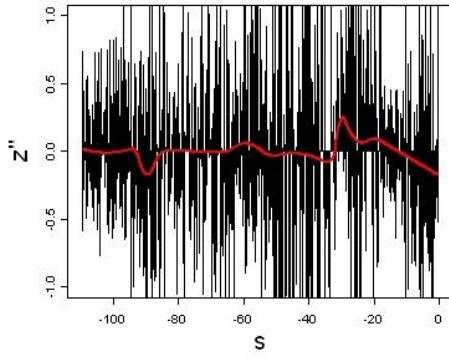
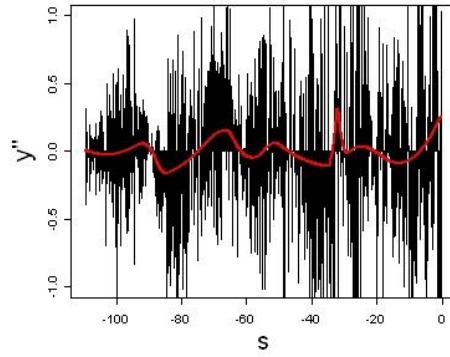
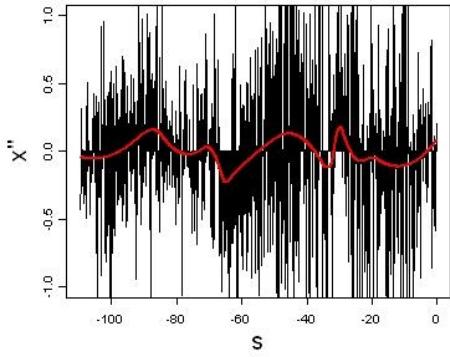
Curve estimate



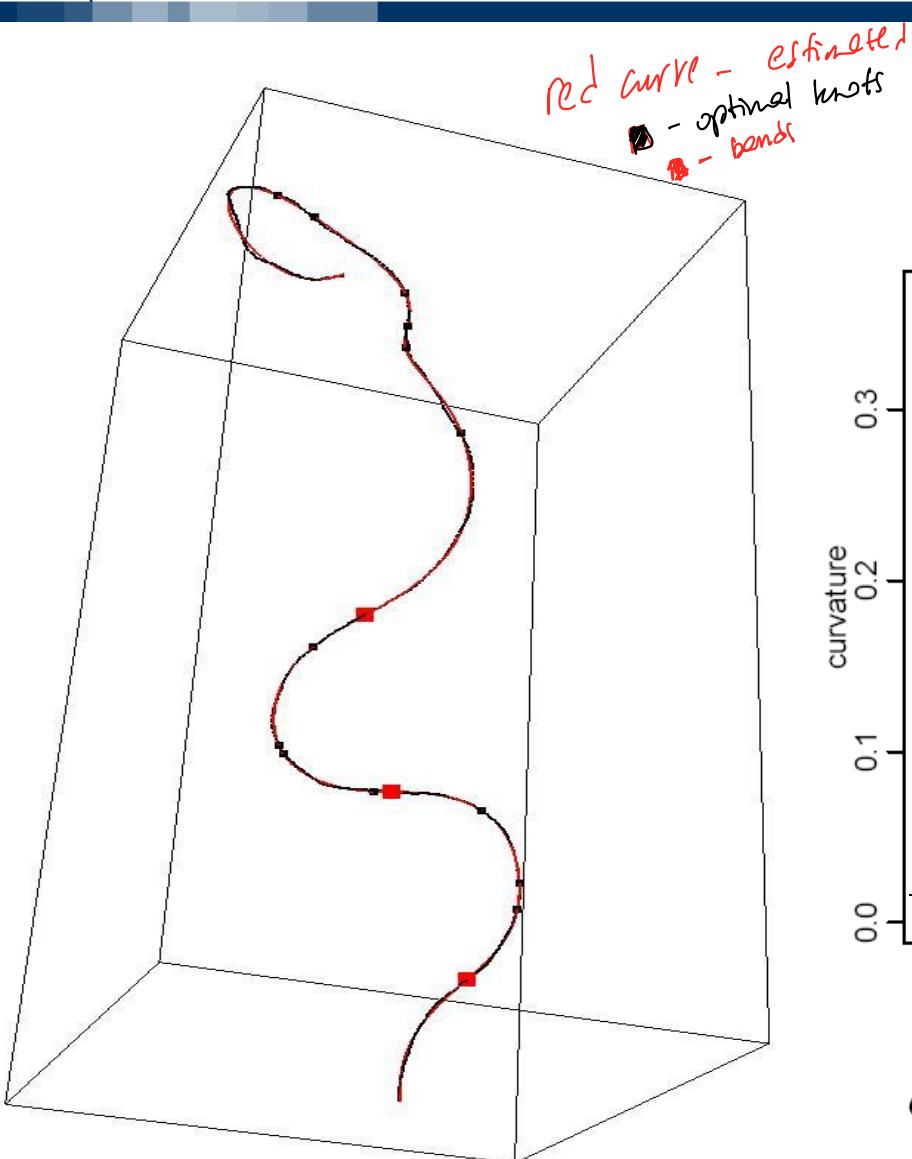
First deriv



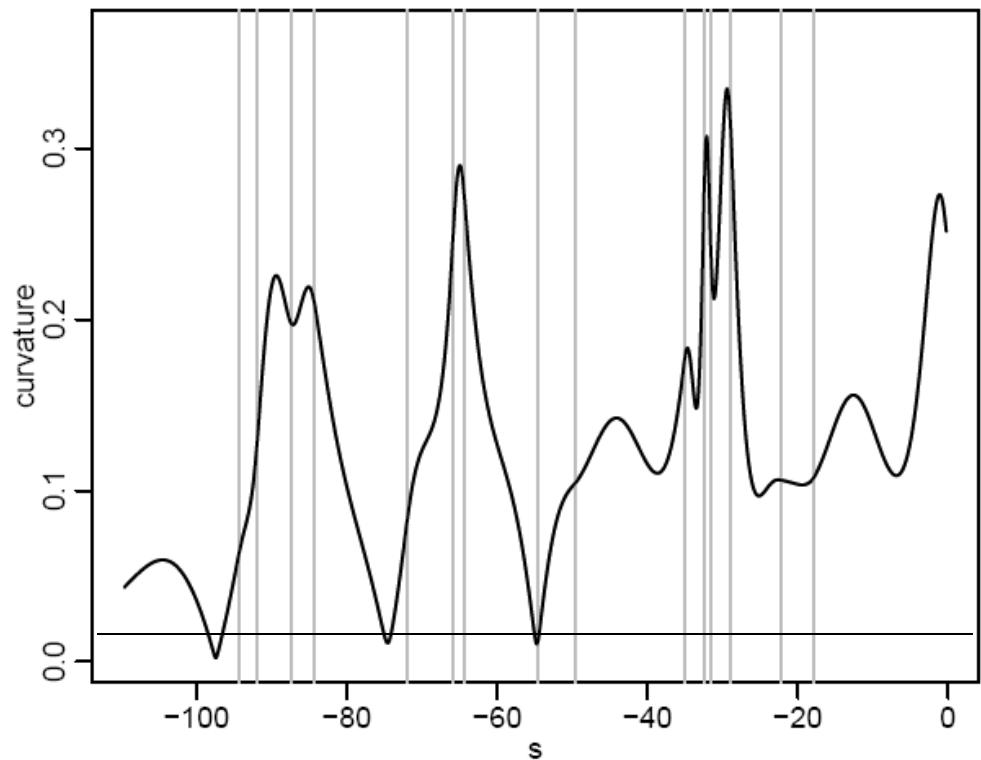
Second deriv.



Accurate curve estimation by 3D free-knot splines

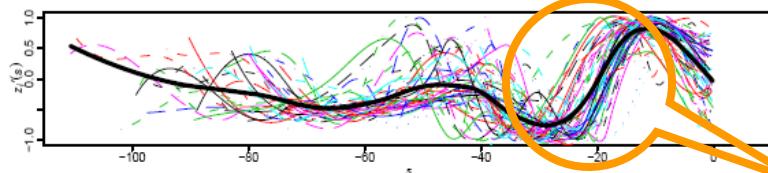
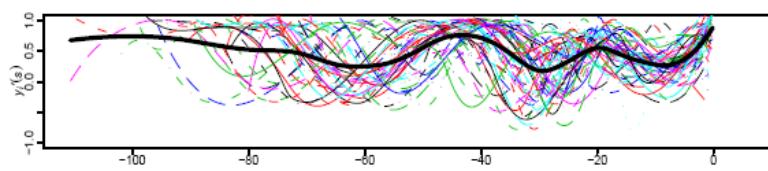
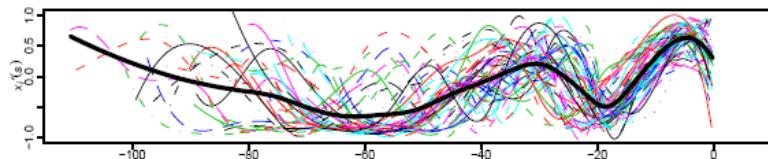


Curvature function



$$C_i(s) = \frac{\|(x'_i(s) \ y'_i(s) \ z'_i(s)) \times (x''_i(s) \ y''_i(s) \ z''_i(s))\|}{\|(x'_i(s) \ y'_i(s) \ z'_i(s))\|^3}$$

Centerline first derivatives



x' (first derivatives)

y'

z'



Visualise data

Phase Variability
(strongly dependent on dimensions of body structure and arteries)

(fit can be women/man)

- To enable meaningful comparisons across patients we need to decouple between-patients *phase variability* and between-patients *amplitude variability*

due to *differences in the dimensions* of patients carotids

due to *differences in the morphological shapes* of patients carotids

\mathcal{C} : set of curves $\mathbf{c}(s) : \mathbb{R} \rightarrow \mathbb{R}^d$

how we made alignment
correlation between first derivatives

Similarity index

$$\rho(\mathbf{c}_1, \mathbf{c}_2) = \frac{1}{d} \sum_{p=1}^d \frac{\int c'_{1p}(s)c'_{2p}(s)ds}{\sqrt{\int c'_{1p}(s)^2 ds} \sqrt{\int c'_{2p}(s)^2 ds}}$$

c_{ip} : pth component of $\mathbf{c}_i = (c_{i1}, \dots, c_{id})$

Class W of warping functions

$$W = \{h : h(s) = ms + q \text{ with } m \in \mathbb{R}^+, q \in \mathbb{R}\}$$

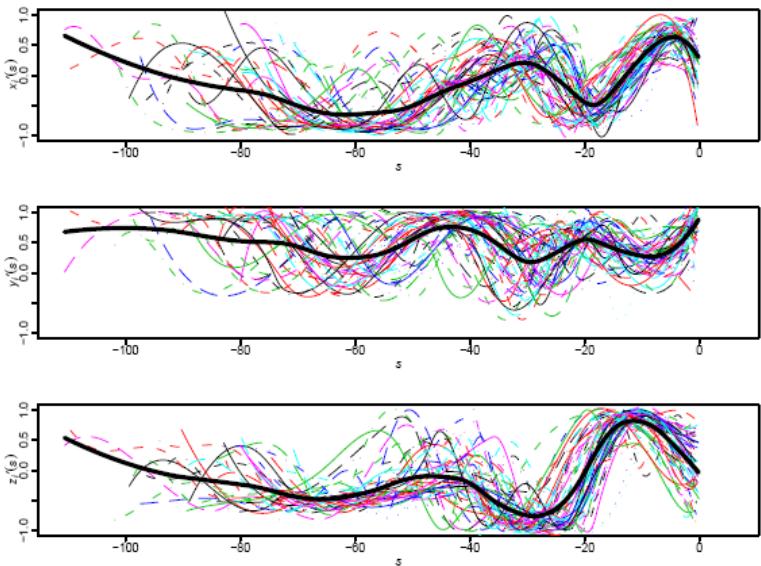
$$\rho(\mathbf{c}_1, \mathbf{c}_2) = 1 \Leftrightarrow \text{for } p = 1, \dots, d, \exists \theta_{0p} \in \mathbb{R}, \theta_{1p} \in \mathbb{R}^+ : \\ c_{1p}(s) = \theta_{0p} + \theta_{1p} c_{2p}(s).$$

Aneurysm location on aligned ICA radius and curvature profiles

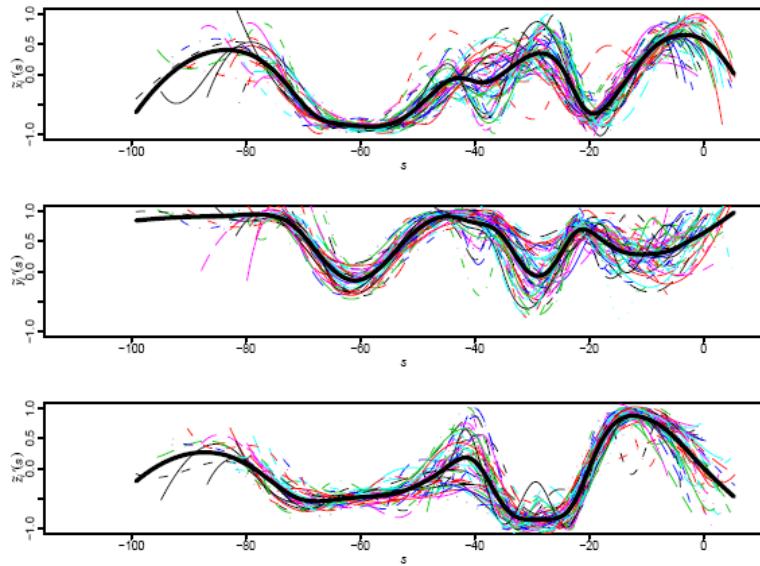
141

Sangalli Secchi Vantini Veneziani, 2009, JASA

Original centerlines

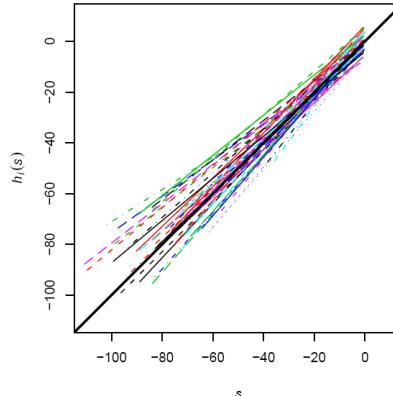


Aligned centerlines



car
ve
align
rotin

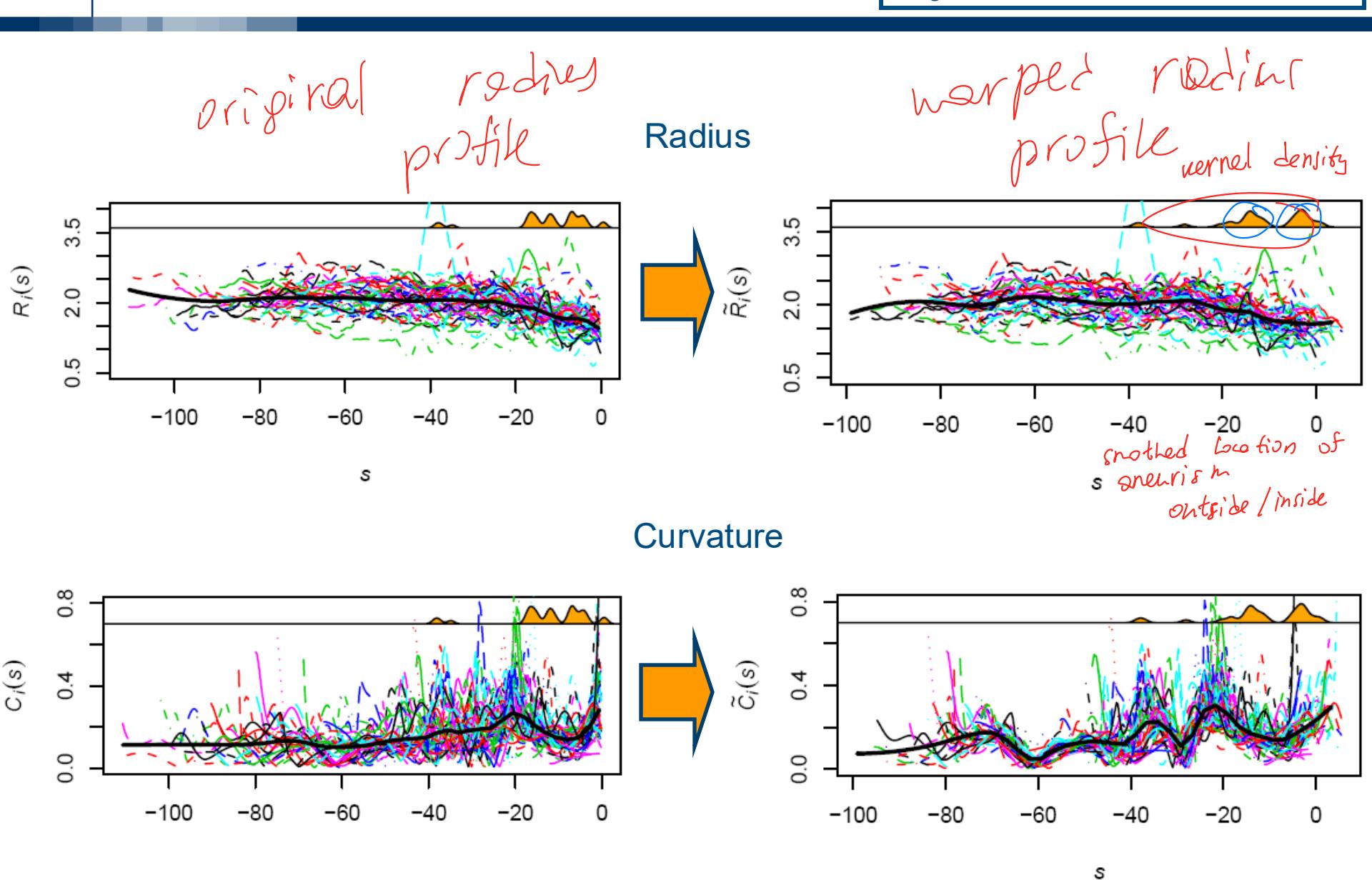
Warping functions (phase variab)

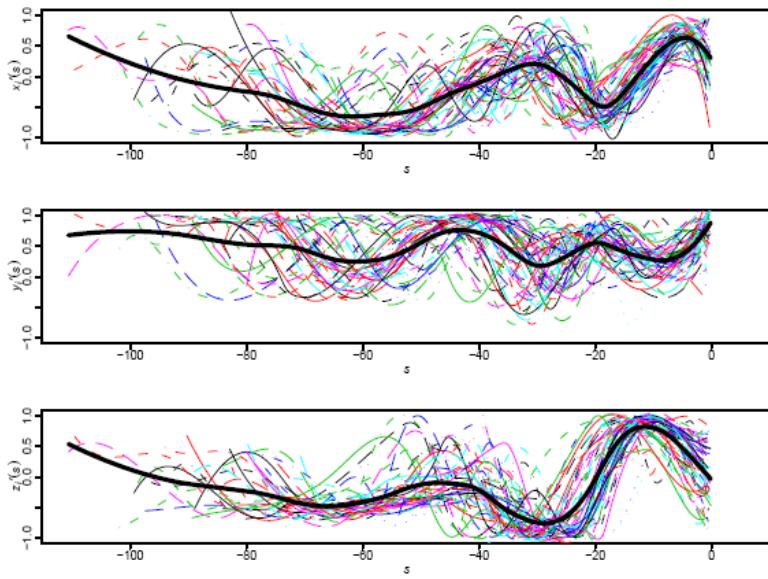


Aneurysm location on aligned ICA radius and curvature profiles

142

Sangalli Secchi Vantini Veneziani, 2009, JASA

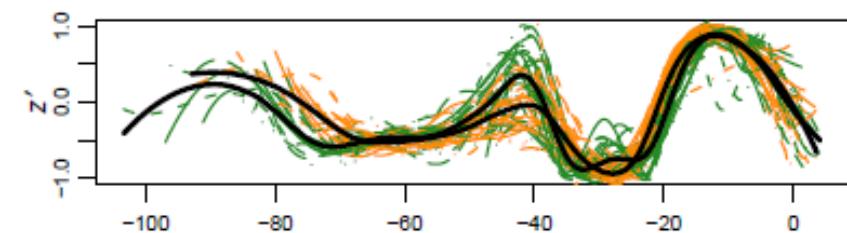
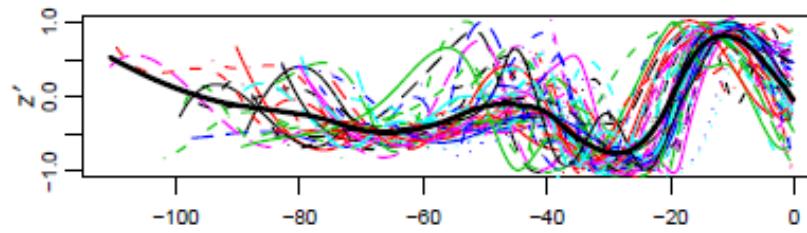
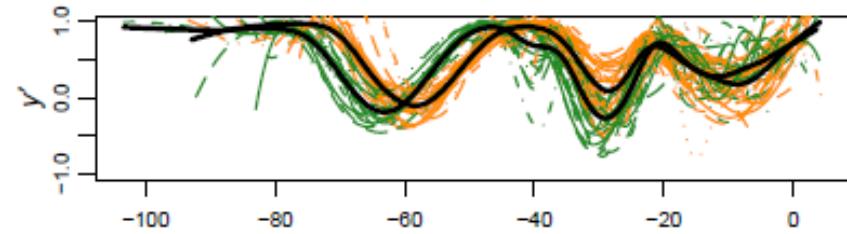
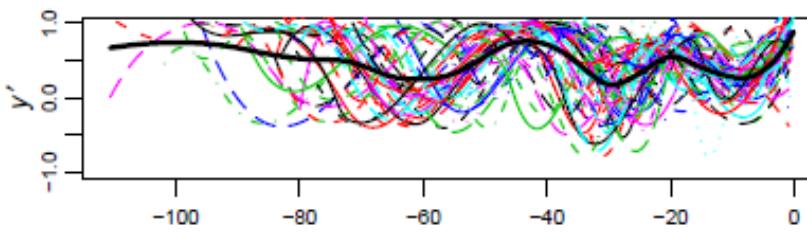
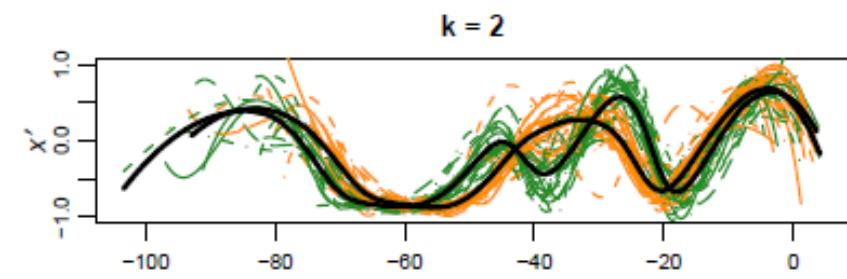
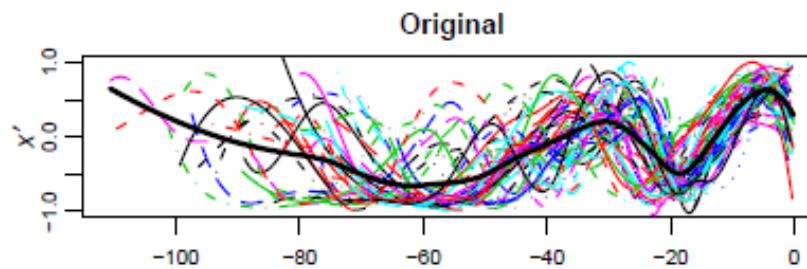


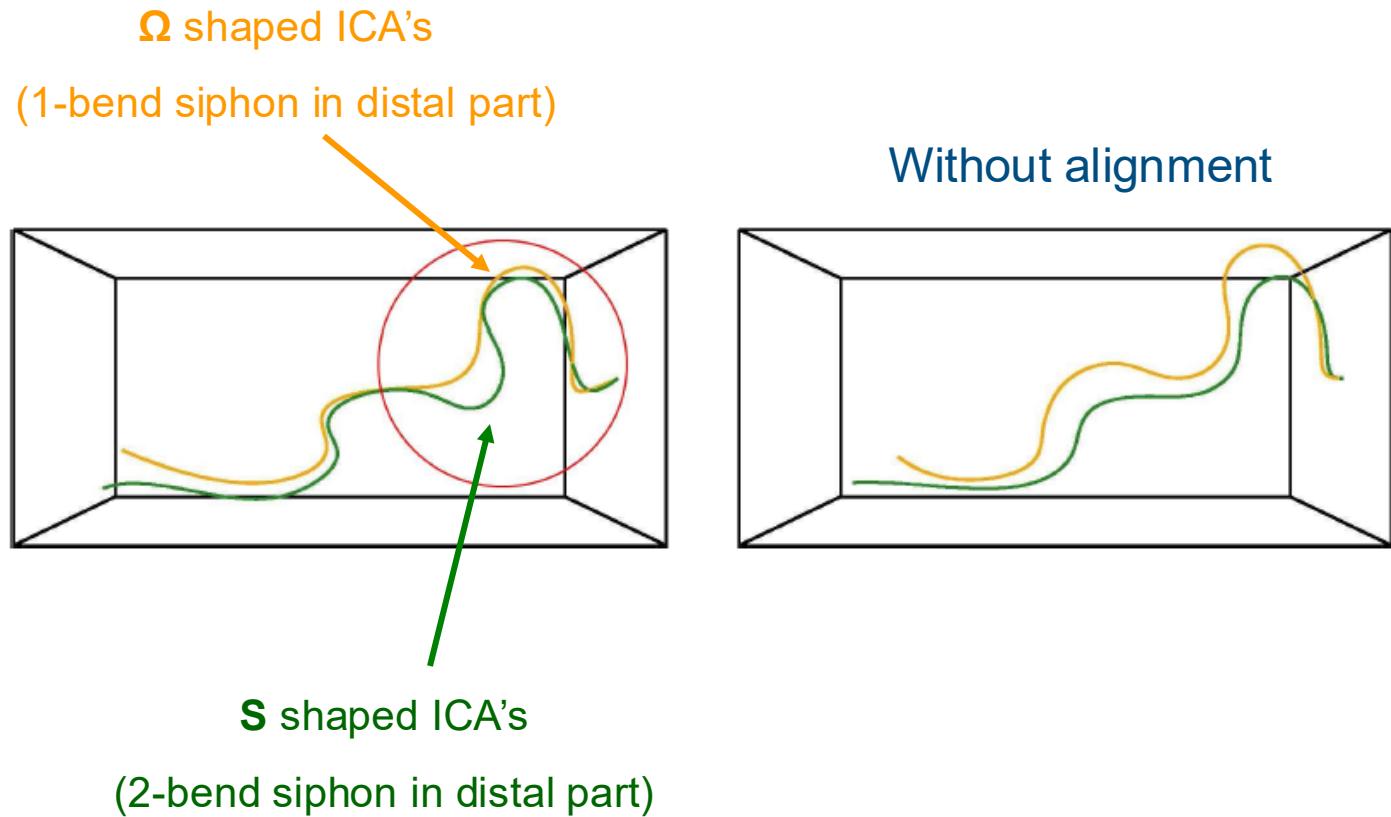


GOAL: Identify ICA's with different morphological shapes

Need to be able to:
jointly align and cluster
the N centerlines
in multiple groups (k groups)
having unknown templates

lines been aligned and clustered





The procedure identify two prototype shapes of ICA's that are described in the medical literature
Krayenbuehl et. Al. (1982)

Simple clustering without alignment is driven by phase variability and fails to identify different morphological shapes

	Aneurysm at or after ICA biforc. (33)	Aneurysm before ICA biforc. (25)	No aneurysms (7)
S shaped ICA's	30%	52%	100%
Ω shaped ICA's	70%	48%	0%

Certain of
box loop
of under-
standing
data

- ▶ The ICA siphon acts as a flow energy dissipator to steady blood flow in the brain

→ better energy dissipator

- S shaped ICA's seems to be more effective in making the blood-flow steadier with respect to Ω shaped ICA's



An introduction to functional data analysis

Part 4 – Functional Principal Component Analysis

Laura M. SANGALLI

MOX - Dipartimento di Matematica, Politecnico di Milano

<https://sangalli.faculty.polimi.it/>

Chapter 8 of Ramsay and Silverman (2005), *Functional Data Analysis*, Springer



Recap: Principal Component Analysis

151

Courtesy of P. Secchi

Problem: Given a dataset of N zero-mean multivariate observations in $\mathbb{R}^p, X_1, \dots, X_N$ find the orthonormal directions a_1, \dots, a_p of maximum variability (for the dataset).

Equivalently, for $k = 1, \dots, p$, find:

$$a_k = \operatorname{argmax}_{a \in \mathbb{R}^p} \operatorname{Var}(a' X)$$

$$\text{subject to: } a'a = 1, a'_j a = 0 \text{ for } j < k$$

- We can re-write the problem as

$$a_k = \operatorname{argmax}_{a \in \mathbb{R}^p} \frac{1}{N} \sum_{i=1}^N (a' X_i)^2$$

$$\text{subject to: } a'a = 1, a'_j a = 0 \text{ for } j < k$$

or, equivalently

$$a_k = \operatorname{argmax}_{a \in \mathbb{R}^p} \frac{1}{N} \sum_{i=1}^N \langle a, X_i \rangle^2$$

$$\text{subject to: } \|a\| = 1, \langle a_j, a \rangle = 0 \text{ for } j < k$$

Note 1. We assume $N > p$ and absence of collinearity, i.e. the data matrix is full rank.

Note 2. If X_1, \dots, X_N are not zero-mean, they can be centered by subtracting the (sample) mean. For unbiasedness, divide by $N-1$ instead of N .

if mean! = 0 $\Rightarrow \frac{1}{N-1}$



Problem: Given a dataset of N zero-mean multivariate observations in \mathbb{R}^p , X_1, \dots, X_N find the orthonormal directions a_1, \dots, a_p of maximum variability, i.e., those solving for $k=1, \dots, p$,

$$a_k = \underset{\mathbf{a} \in \mathbb{R}^p}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^N \langle \mathbf{a}, X_i \rangle^2$$

subject to: $\|\mathbf{a}\| = 1, \langle a_j, \mathbf{a} \rangle = 0 \text{ for } j < k$

Solution: Call S the sample covariance matrix of X_1, \dots, X_N . Then, the principal components are found as the eigenvectors of the matrix S ; for $k=1, \dots, p$, 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.,

$$u_{ik} = \langle X_i, e_k \rangle = X'_i e_k$$

u_{ik} represent the variability captured from data along direction e_k

Problem: Given a dataset of N zero-mean functional observations in H , X_1, \dots, X_N , find the directions of maximum variability (in H) of the dataset, i.e., for $k=1, \dots, N$, find ξ_k maximizing

$$\frac{1}{N} \sum_{i=1}^N \langle \xi, X_i \rangle_H^2$$

subject to: $\|\xi\| = 1, \langle \xi_j, \xi \rangle_H = 0 \text{ for } j < k$

\sim uncorrelated (orthogonal)

- 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.

find closest linear space to data. If we have N data, we can find $N-1$ principal components



Problem: Given a dataset of N zero-mean functional observations in H , X_1, \dots, X_N , find the directions of maximum variability (in H) of the dataset, i.e., for $k=1, \dots, N$, find ξ_k maximizing

$$\frac{1}{N} \sum_{i=1}^N \langle \xi, X_i \rangle_H^2$$

subject to: $\|\xi\| = 1, \langle \xi_j, \xi \rangle_H = 0 \text{ for } j < k$

- As in multivariate principal component analysis, **functional principal components** are related with the eigen-decomposition of the functional counterpart of the (sample) covariance matrix
- Recall that the **sample covariance operator** is defined as

$$Sx = \frac{1}{N} \sum_{i=1}^N \langle X_i, x \rangle X_i, \quad x \in H$$

In L^2 it is equivalently defined as

$$[Sx](t) = \int_T \widehat{c}(s, t)x(s)d(\varepsilon)], \quad x \in L^2 \quad \text{with} \quad \widehat{c}(s, t) = \frac{1}{N} \sum_{i=1}^N X(s)X(t)$$

Note. If data are centered on the sample mean, divide by $N-1$ for unbiasedness.



Problem: Given a dataset of N zero-mean functional observations in H , X_1, \dots, X_N , find the directions of maximum variability (in H) of the dataset, i.e., for $k=1, \dots, N$, find ξ_k maximizing

$$\frac{1}{N} \sum_{i=1}^N \langle \xi, X_i \rangle_H^2$$

subject to: $\|\xi\| = 1, \langle \xi_j, \xi \rangle_H = 0 \text{ for } j < k$

Solution: Let S be the sample covariance operator of X_1, \dots, X_N . Then, the **functional principal components** ξ_1, \dots, ξ_N are found as the eigenfunctions of the operator S , i.e., they solve the eigen-equations

$$S\xi_k = \lambda_k \xi_k$$

The eigenvalue λ_k associated with the eigenvector ξ_k represents the variability along the direction ξ_k

We call functional score u_{ik} the projection of the observation X_i along the direction ξ_k , i.e.,

$$u_{ik} = \langle X_i, \xi_k \rangle$$

Note. If data are centered on the sample mean, we can find at most $N-1$ principal components





- !!! We only have discrete and noisy realizations of X_1, \dots, X_N !!!

for the i -th statistical unit:

$\{x_{i1}, \dots, x_{in}\}$: x_{ij} is the value observed for the i -th statistical unit at s_j

→ for each statistical unit we obtain a functional representation by smoothing

PRE-SMOOTHING APPROACH

- Dimensional reduction: look for an elbow in the cumulative percentage of total variance explained by the first p functional principal components

$$CPV(p) = \frac{\sum_{k=1}^p \hat{\lambda}_k}{\sum_{k=1}^N \hat{\lambda}_k}.$$

- Useful plots: boxplots of scores along the first p directions, to investigate the possible presence (and influence) of outliers, clustering structures, etc

- Interpretation of the loadings:

- Plot directly loadings (*only for expert users*)

- Plot mean +/- eigenfunction multiplied by a proper constant, e.g., std. along the component, which corresponds to sqrt of the eigenvalue: $\bar{X} \pm \sqrt{\lambda_k} \xi_k$

- Plotting the projection of the dataset along each component: $\bar{X} + u_{ik} \xi_k$ or along the first p components: $\bar{X} + \sum_{k=1}^p u_{ik} \xi_k$



Example

Dataset of Canadian temperatures

Ramsay Silverman 2005 Springer

