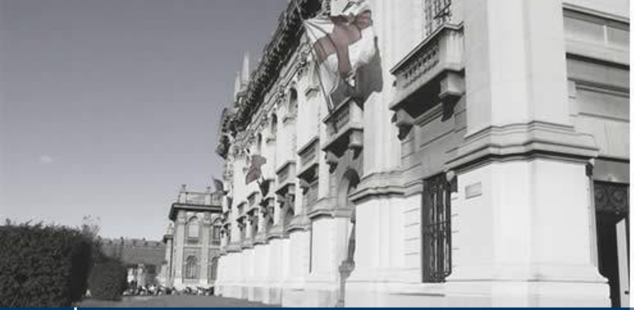


POLITECNICO DI MILANO

January 2014, Geilo, Norway

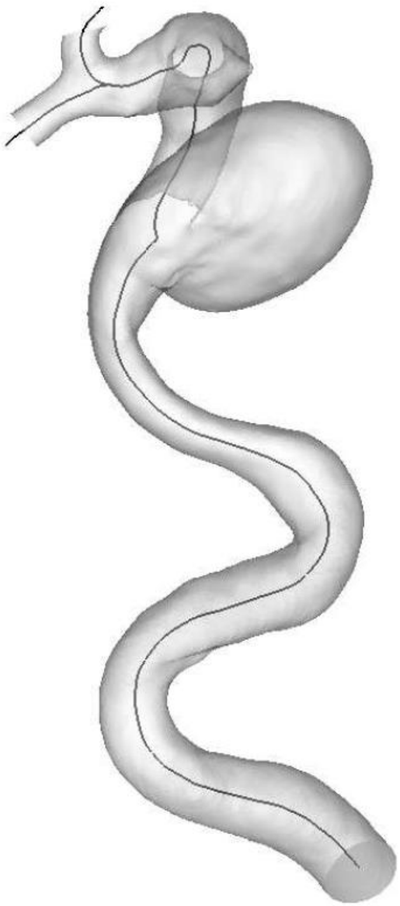
14th Winter School in eScience

Big Data Challenges to Modern Statistics

High-dimensional and complex data: the example of data on functional spaces

Laura M. SANGALLI

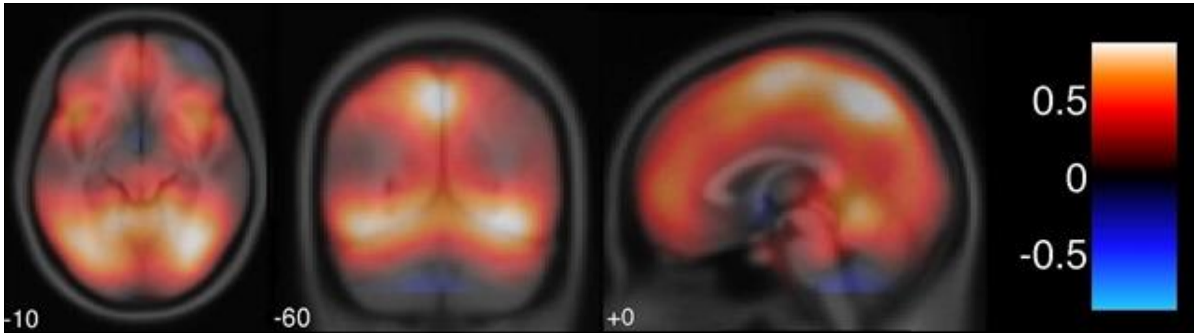
MOX - Dipartimento di Matematica, Politecnico di Milano



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 the internal structures of a body provided by diagnostic medical scanners



Reconstruction of an inner carotid artery with aneurysm, from angiographic images

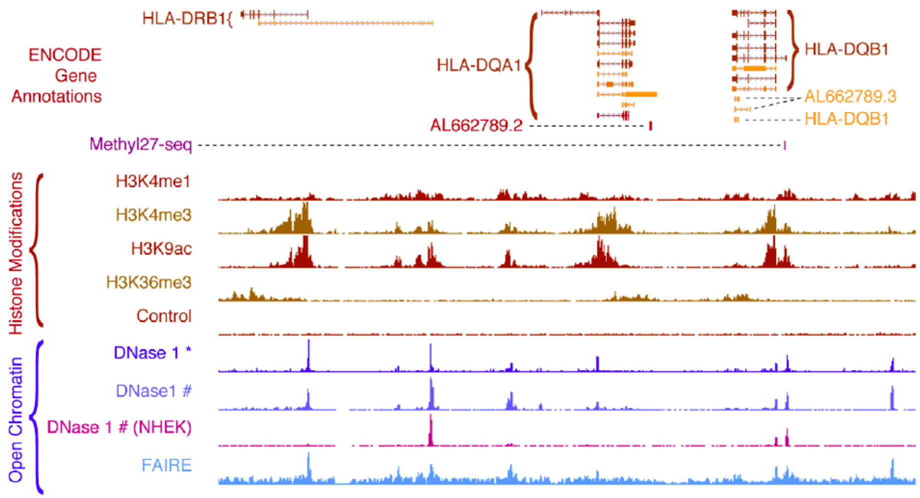
Sangalli, Secchi, Vantini, Veneziani (2009) J. R. Stat. Soc. Ser. C

Magnetic Risonance Imaging of a brain during a reading task

Aston, Turkheimer, Brett (2006) Hum. Brain Map.

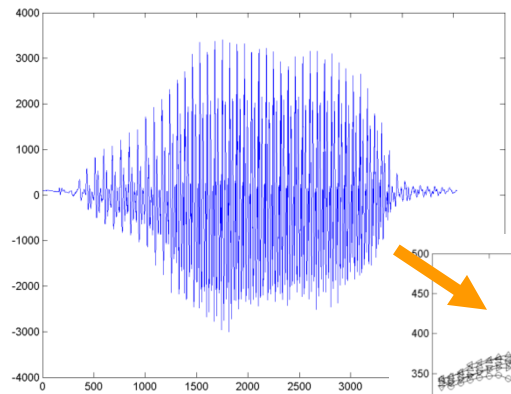
Functional data: where they come from

► measurements of gene expression levels

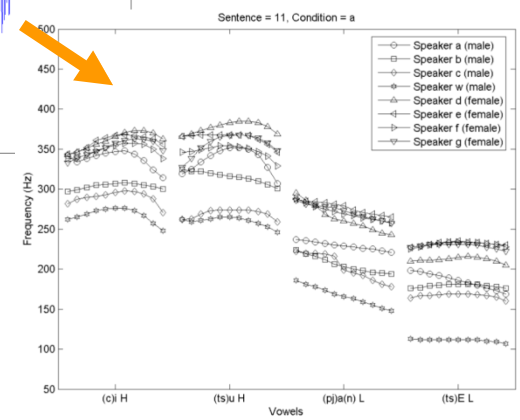


The ENCODE Project Consortium, 2011, PLoS Biology

► large speech databases describing linguistic constructs expressed via spectrum data



Aston, Chiou, Evans (2010)
J. R. Stat. Soc. Ser. C

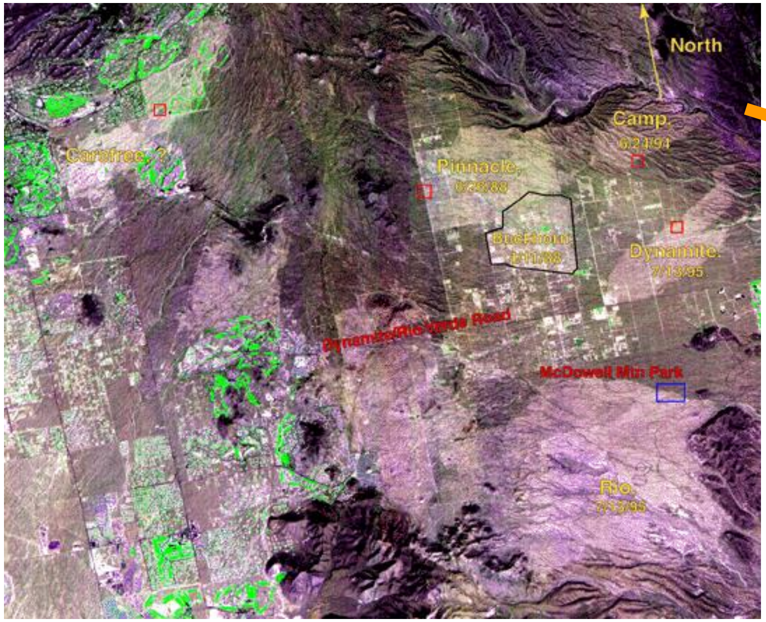


► images of steady or moving objects/individuals recorded by computer vision devices

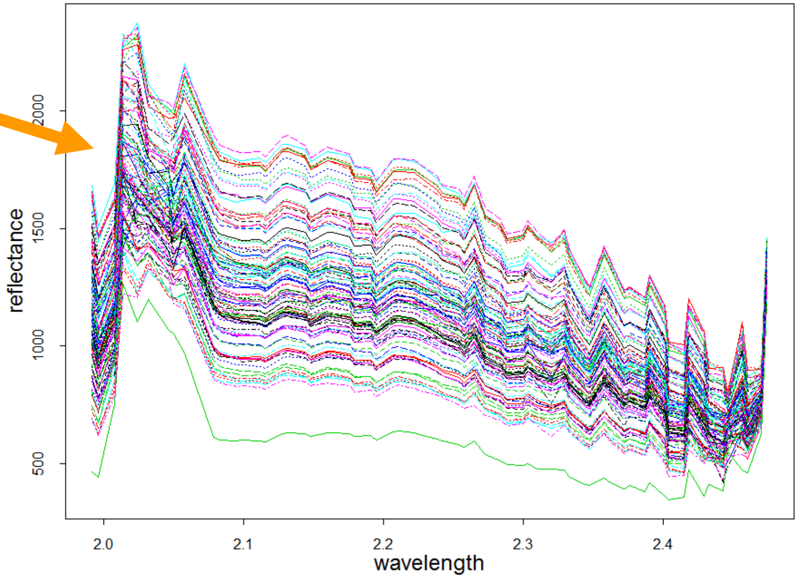
Kaziska, Srivastava (2007)
J. Amer. Statist. Assoc.

Functional data: where they come from

- multi-spectral data from satellite remote sensing



Northwest Scottsdale / Rio Verde area



The analysis of complex and high dimensional data poses new and challenging problems in research

It is fueling one of the most fascinating and fast growing research fields of modern statistics

General references on FDA

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

http://www.functionaldata.org

Software:

- R package fda, available from CRAN; corresponding Matlab code
- R package Refund, available from CRAN
- Matlab code PACE
- R package mgcv

Hal Varian – Google Chief Economist

The New York Times, 2009

“I keep saying the sexy job in the next ten years will be statisticians.

[…] *The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.”*

SIEMENS

Fondazione Politecnico di Milano

€

ANEURISK

A CONJECTURE

The pathogenesis of cerebral aneurysms is conditioned by the geometry of the cerebral vessels through its effects on blood fluid dynamics

MODELLISTICA E CALCOLO SCIENTIFICO – MOX – MODELING AND SCIENTIFIC COMPUTING

EMORY

LaBS – Laboratory of Biological Structure Mechanics – Department of Structural Engineering

MARIO NEGRI – ISTITUTO DI RICERCHE FARMACOLOGICHE

OSPEDALE MAGGIORE POLICLINICO, MANGIAGALLI E REGINA ELENA – FONDAZIONE IRCCS DI NATURA PUBBLICA

Azienda Ospedaliera Ospedale Niguarda Ca'Granda

Statistics

Numerical Analysis

Bio-Engineering

Computer Sciences

Neurosurgery

Neuroradiology

Numerical Analysis

Now at EMORY

Alessandro Veneziani

Principal Investigator

Tiziano Passerini

Marina Piccinelli

Statistics

Piercesare Secchi

Simone Vantini

Laura Sangalli

Valeria Vitelli

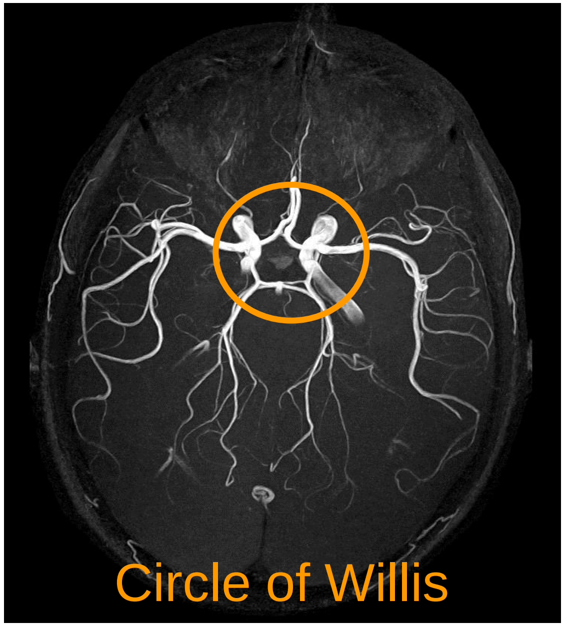
(now at University of Oslo)

Cerebral Aneurysms

- Cerebral aneurysms: deformations of cerebral arteries, mostly placed on vessels belonging to or connected to the Circle of Willis

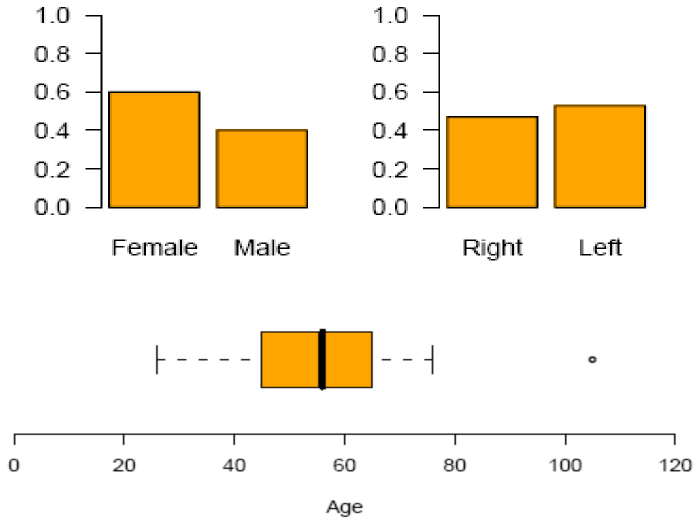
Aneurysms EPIDEMIOLOGY

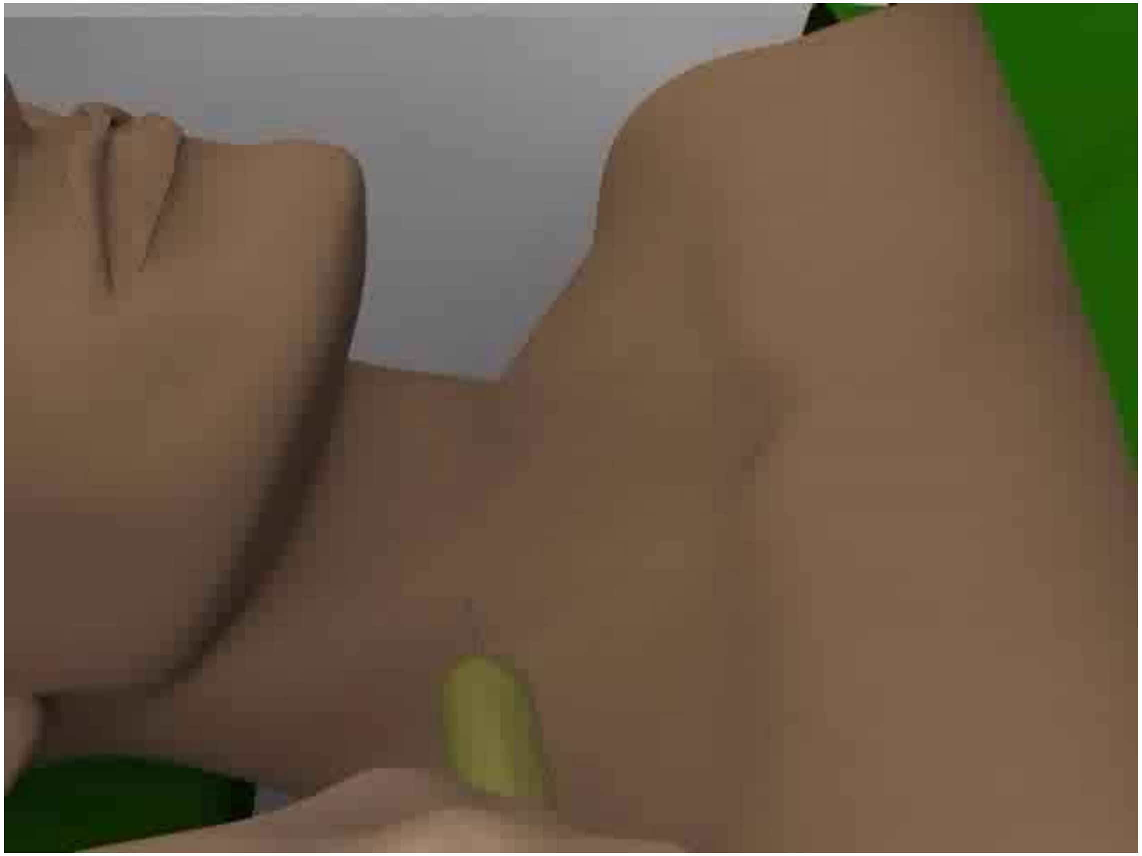
- Incidence rate of cerebral aneurysms:
  1/20 people
- Incidence rate of ruptured cerebral aneurysms per year:
  1/10000 people per year
- Mortality due to a ruptured aneurysm:
  > 50%: Out of 9 patients with a ruptured aneurysm:
  - 3 are expected to die before arriving at the hospital
  - 2 to die after having arrived at the hospital
  - 2 to survive with permanent cerebral damages
  - 2 to survive without permanent cerebral damages



Data: 3D-angiographies

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.

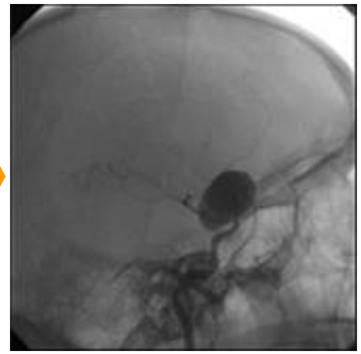
Sequence of X-Rays → 3D-array of gray scaled pixels

From X-rays to Centerlines and Local Maximal Inscribed Sphere Radius

Contrast Fluid Injections

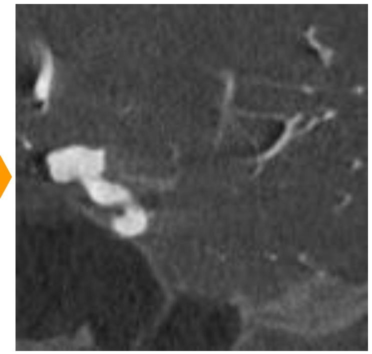
1

X-rays (one direction)



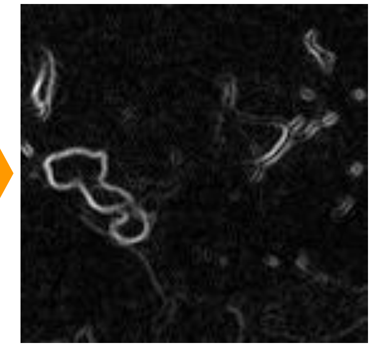
2

3d-array (one slice)



3

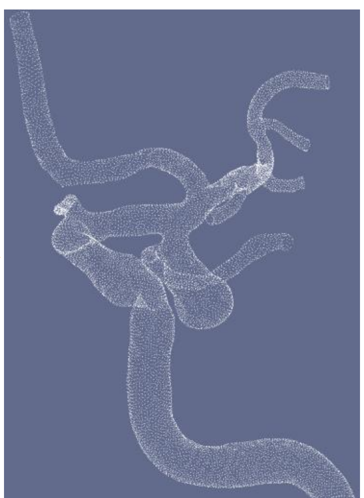
Gradient 3d-array (one slice)



4

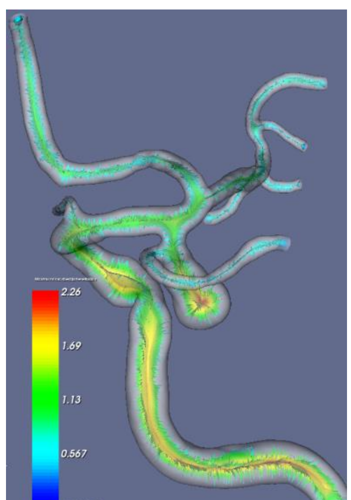
Surface Points

4



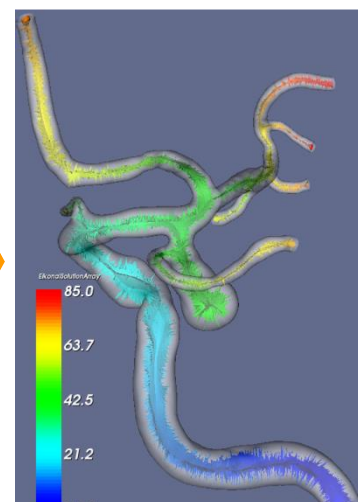
5

Voronoi Diagram



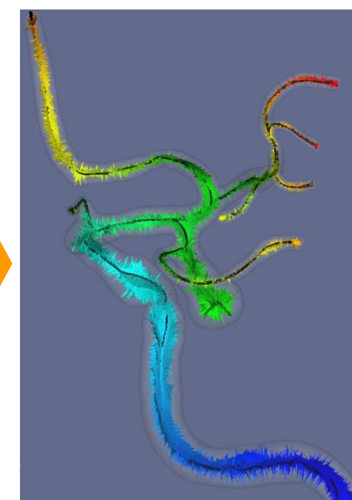
6

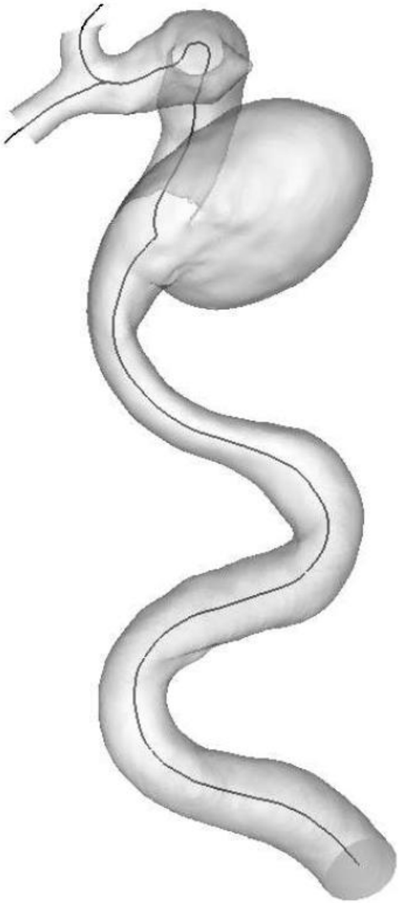
Eikoinal Equation



7

Centerline+MISR





Preprocessing:
Image reconstruction

Focus on Internal Carotid Artery (ICA)

For each patient *i* elicitation of 3-spatial coordinates of ICA centerline

$$\{(x_{ij}, y_{ij}, z_{ij}) : j = 1, \ldots, n_i\}$$

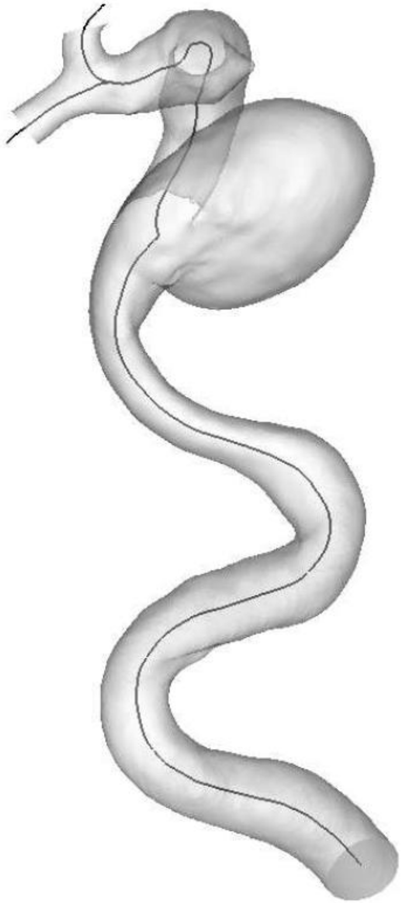
and vessel radius

$$\{R_{ij} : j = 1, \ldots, n_i\}$$

alone a fine grid of points $(350 \leq n_i \leq 1380)$

Two geometric quantities that greatly influence the haemodynamics: vessel **radius** and **curvature** (curvature of its centerline)

→ Choice of data objects, atoms of the analysis



Focus on Internal Carotid Artery (ICA)

For each patient *i* elicitation of 3-spatial coordinates of ICA centerline

$$\{(x_{ij}, y_{ij}, z_{ij}) : j = 1, \ldots, n_i\}$$

and vessel radius

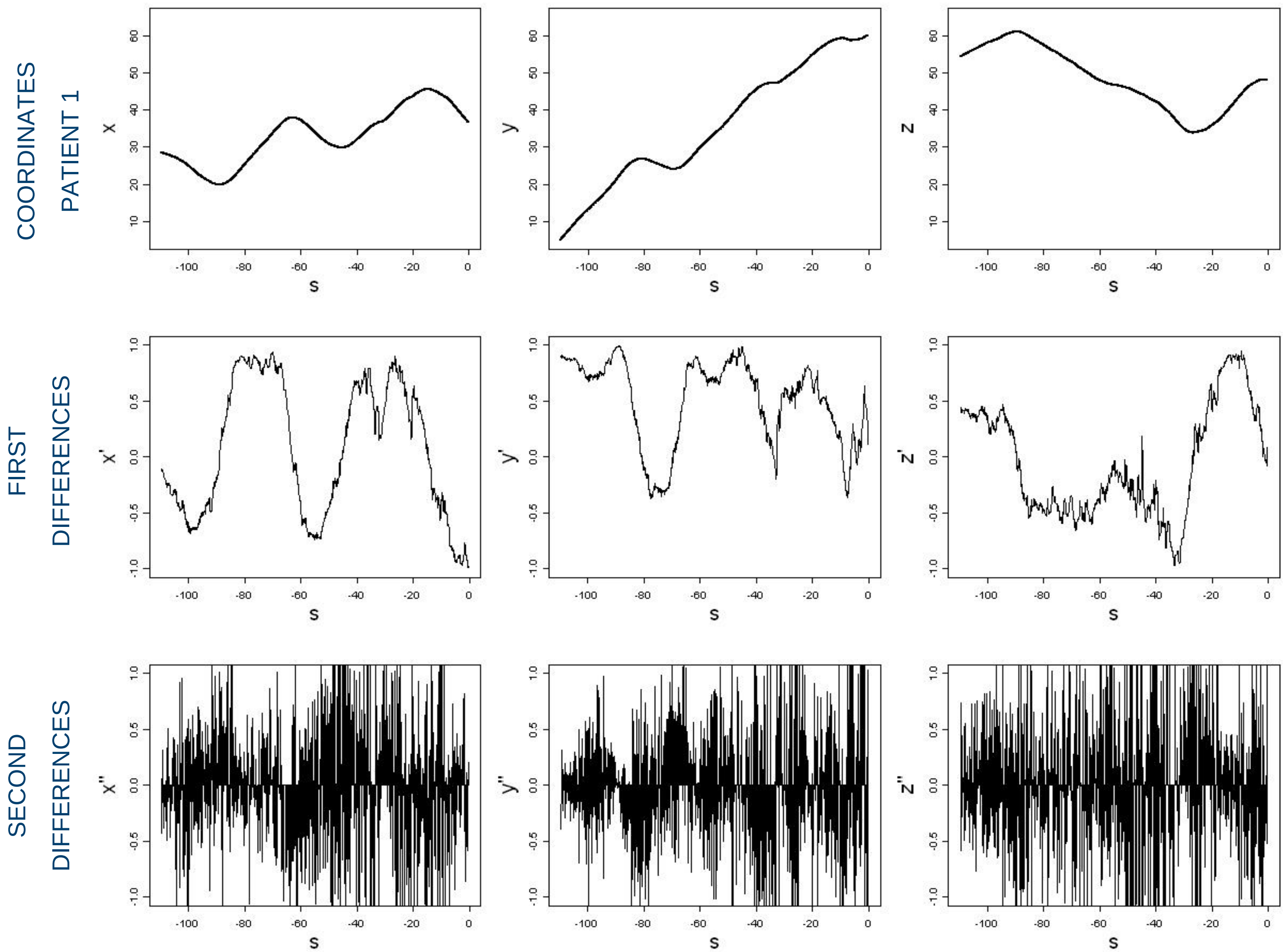
$$\{R_{ij} : j = 1, \ldots, n_i\}$$

alone a fine grid of points $(350 \leq n_i \leq 1380)$

Approximate curvilinear abscissa: $\{s_{ij} : j = 1, \ldots, n_i\}$ $\quad s_{i1} = 0$

$$s_{ij} - s_{ij-1} = -\sqrt{(x_{ij} - x_{ij-1})^2 + (y_{ij} - y_{ij-1})^2 + (z_{ij} - z_{ij-1})^2}, \; j = 2, \ldots, n_i$$

Centerline data



Very high signal-to-noise ratio

Fine grid of observed points

Rough estimates of first and second derivatives by means of central differences

FIRST GOAL:
accurate estimation of centerlines *curvature* functions

Noisy and discrete data $\longrightarrow$ functional representations

Smoothing, regularization, curve fitting

$$z_i = f(s_i) + \epsilon_i \qquad i = 1, \dots, n$$

$\psi_1, \dots, \psi_K$: K basis functions

$$\hat{f}(s) = \sum_{k=1}^{K} \hat{c}_k \psi_k(s)$$

$\rightarrow$ Find $\hat{c}_k, k = 1, \dots, K$ (i.e., find $\hat{f}$) by minimizing

$$\text{SSE} = \sum_{j=1}^{n} (z_i - f(s_i))^2 = \sum_{j=1}^{n} \left(z_i - \sum_{k=1}^{K} c_k \psi_k(s_i) \right)^2$$

$$\Psi = \begin{bmatrix} \psi_1(s_1) & \psi_2(s_1) & \cdots & \psi_K(s_1) \\ \psi_1(s_2) & \psi_2(s_2) & \cdots & \psi_K(s_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(s_n) & \psi_2(s_n) & \cdots & \psi_K(s_n) \end{bmatrix}$$

$$\boldsymbol{z} = (z_1, \dots, z_n)^t$$

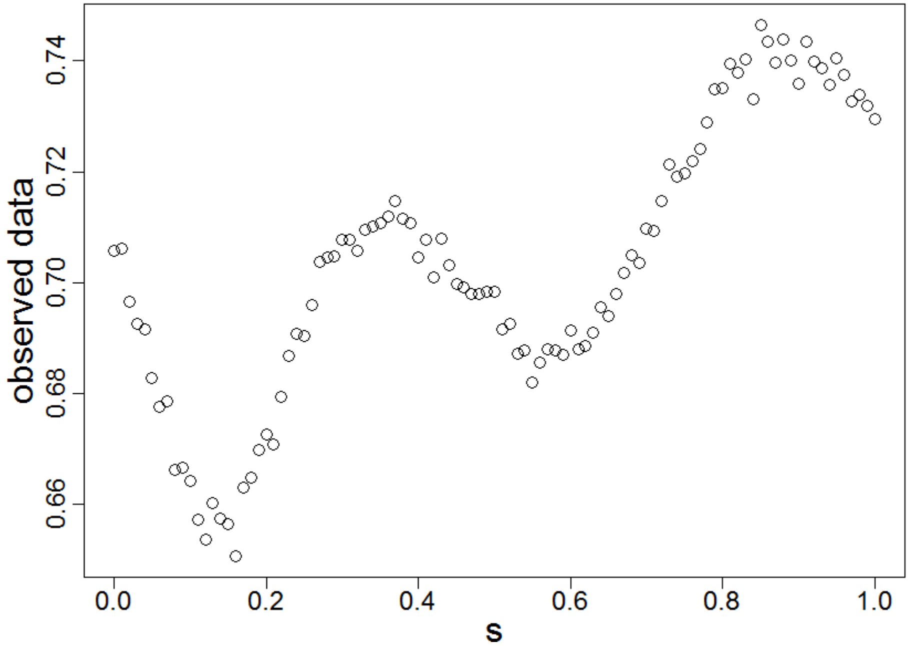
$$\boldsymbol{f} = (f(s_1), \dots, f(s_n))^t$$

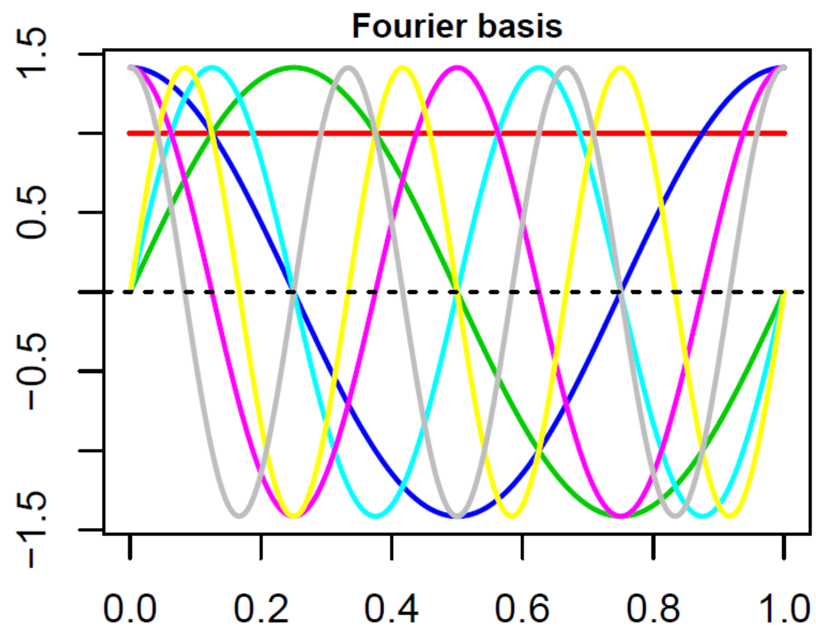
$$\boldsymbol{c} = (c_1, \dots, c_K)^t$$

$$z_i = f(s_i) + \epsilon_i \qquad i = 1, \dots, n$$

$\psi_1, \dots, \psi_K$: K basis functions

$$\hat{f}(s) = \sum_{k=1}^{K} \hat{c}_k \psi_k(s) \qquad \rightarrow \text{Find } \hat{c}_k, k = 1, \dots, K \text{ (i.e., find } \hat{f}) \text{ by minimizing}$$





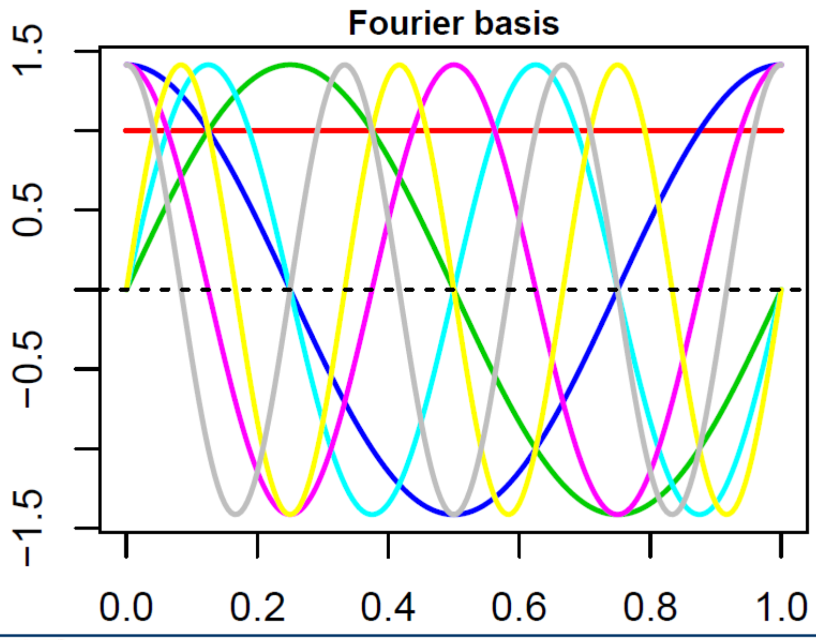
Smoothing, curve fitting

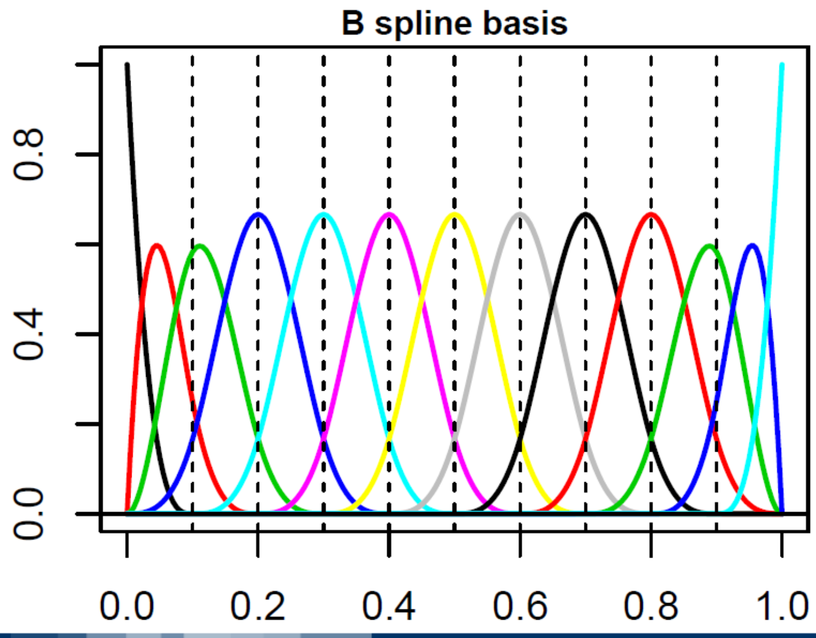
$$\text{SSE} = (\boldsymbol{z} - \Psi \boldsymbol{c})^t (\boldsymbol{z} - \Psi \boldsymbol{c})$$

$$\hat{\boldsymbol{c}} = \left(\Psi^t \Psi\right)^{-1} \Psi^t \boldsymbol{z}$$

$$\hat{\boldsymbol{z}} = \hat{\boldsymbol{f}} = \Psi \hat{\boldsymbol{c}} = \Psi \left(\Psi^t \Psi\right)^{-1} \Psi^t \boldsymbol{z} = S \boldsymbol{z}$$

$$df = K = tr(S) = tr(S^t S) = tr(2S - S^t S)$$





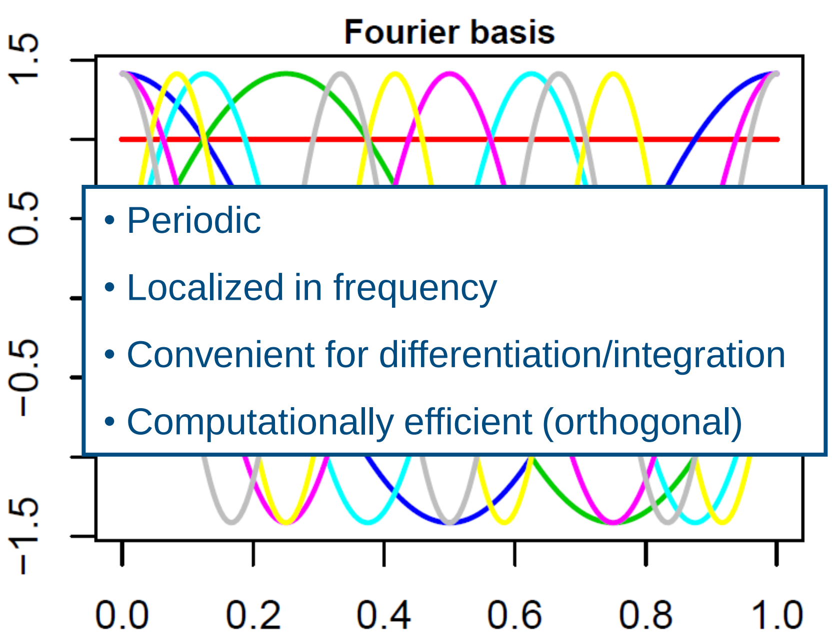
Smoothing, curve fitting

$$\mathrm{SSE} = (\boldsymbol{z} - \Psi \boldsymbol{c})^t (\boldsymbol{z} - \Psi \boldsymbol{c})$$

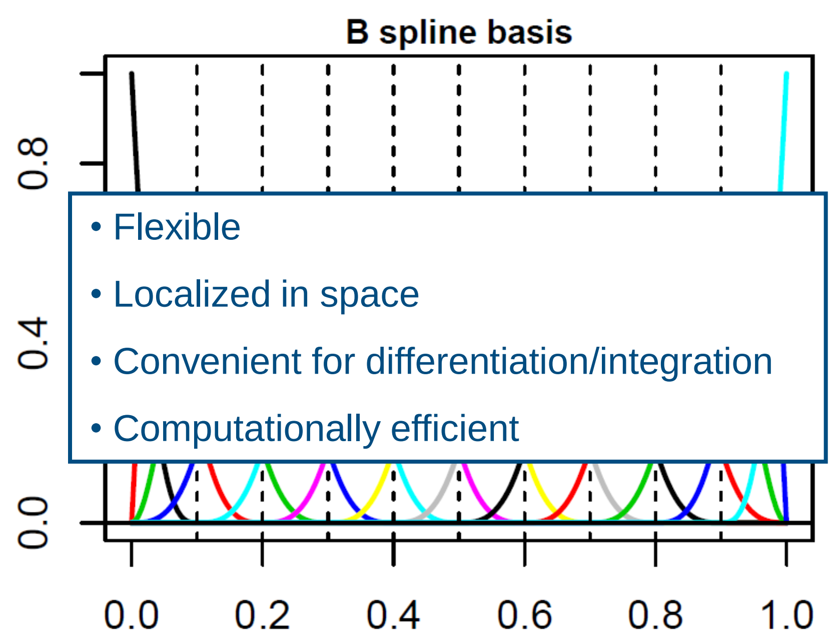
$$\hat{\boldsymbol{c}} = \left(\Psi^t \Psi\right)^{-1} \Psi^t \boldsymbol{z}$$

$$\hat{\boldsymbol{z}} = \hat{\boldsymbol{f}} = \Psi \hat{\boldsymbol{c}} = \Psi \left(\Psi^t \Psi\right)^{-1} \Psi^t \boldsymbol{z} = S \boldsymbol{z}$$

$$df = K = tr(S) = tr(S^t S) = tr(2S - S^t S)$$



- Periodic
- Localized in frequency
- Convenient for differentiation/integration
- Computationally efficient (orthogonal)

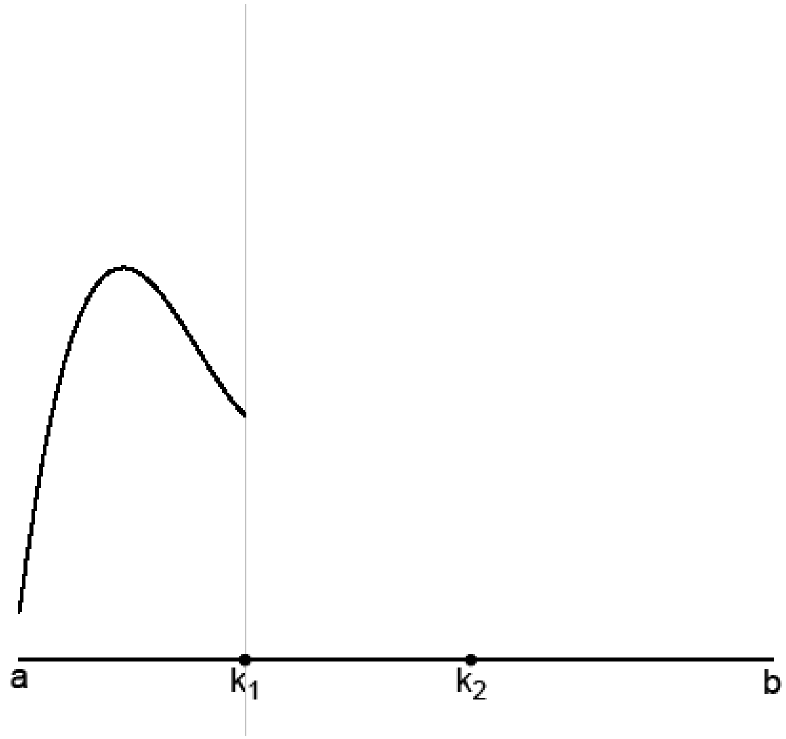


- Flexible
- Localized in space
- Convenient for differentiation/integration
- Computationally efficient

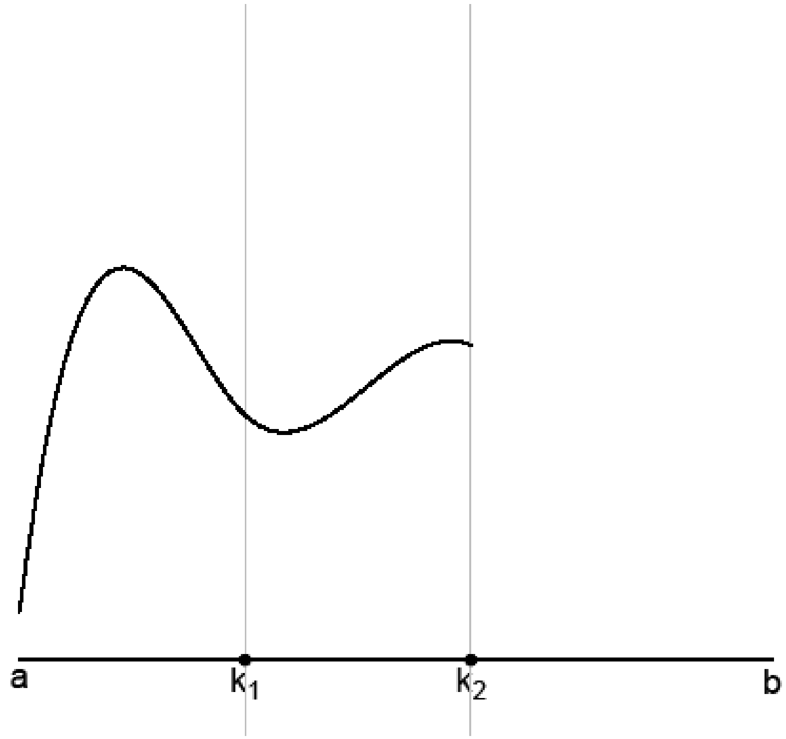
What is a spline

What is a spline

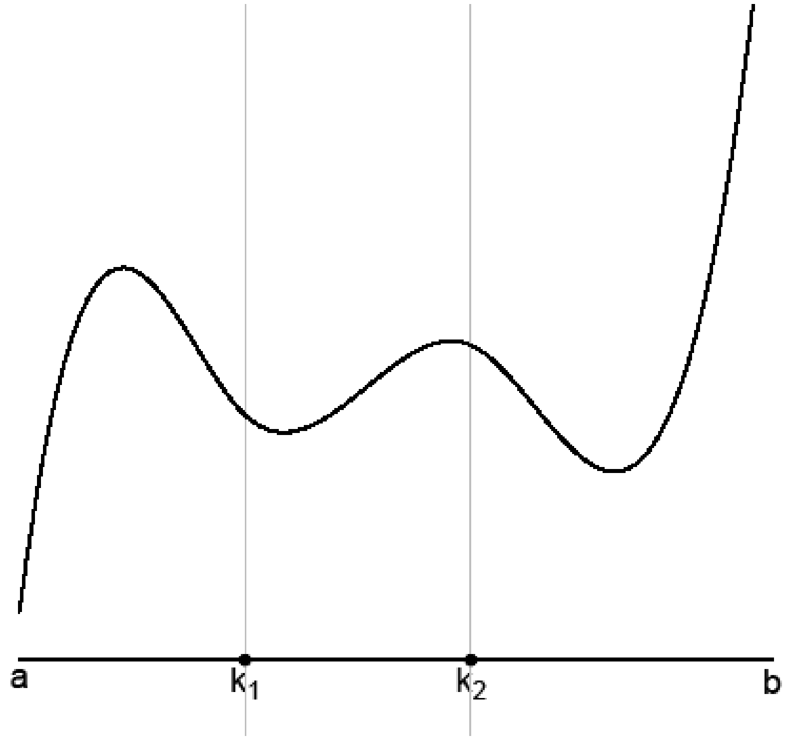
What is a spline



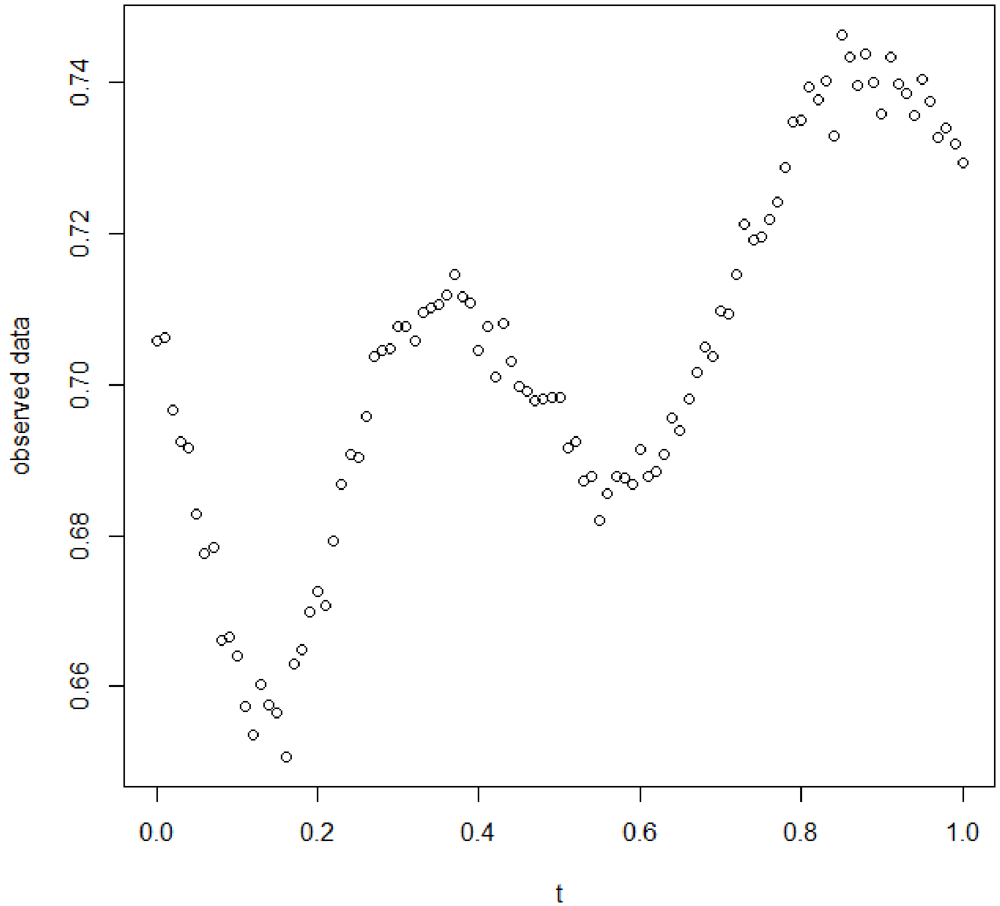
What is a spline



What is a spline

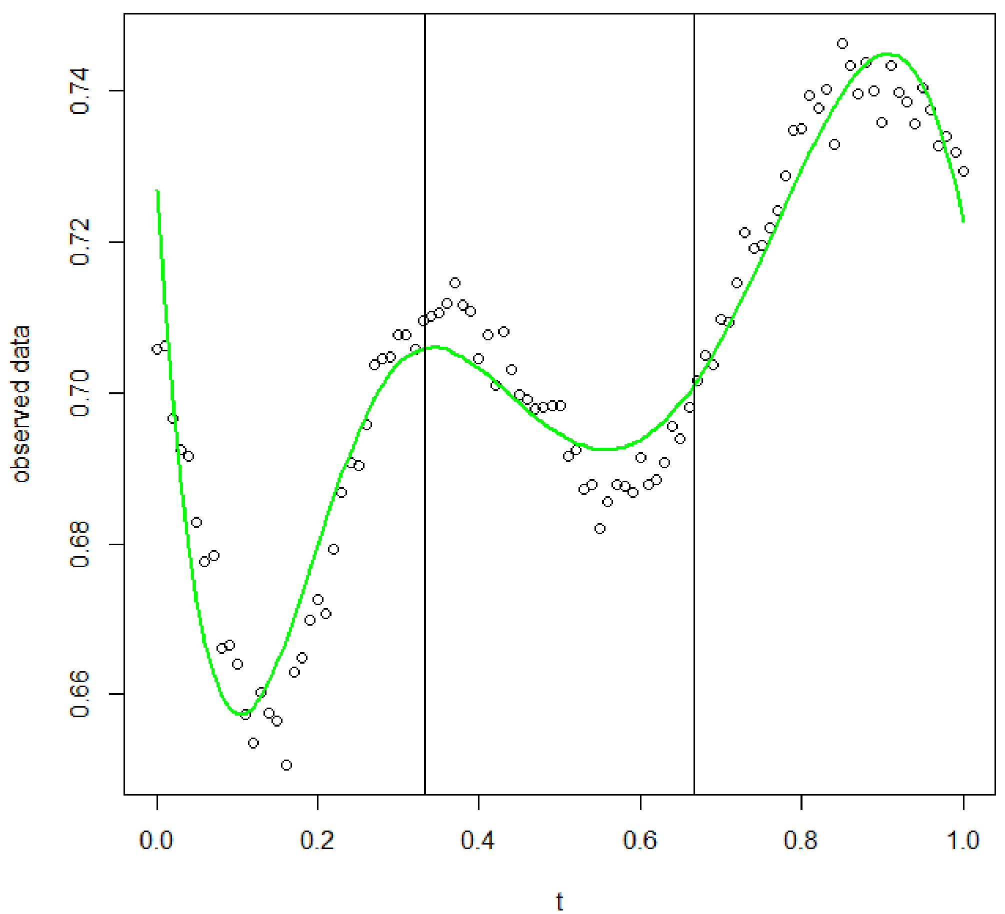


1) Use a functional space with only few dimensions (few basis) $K << n$



Smoothing, curve fitting

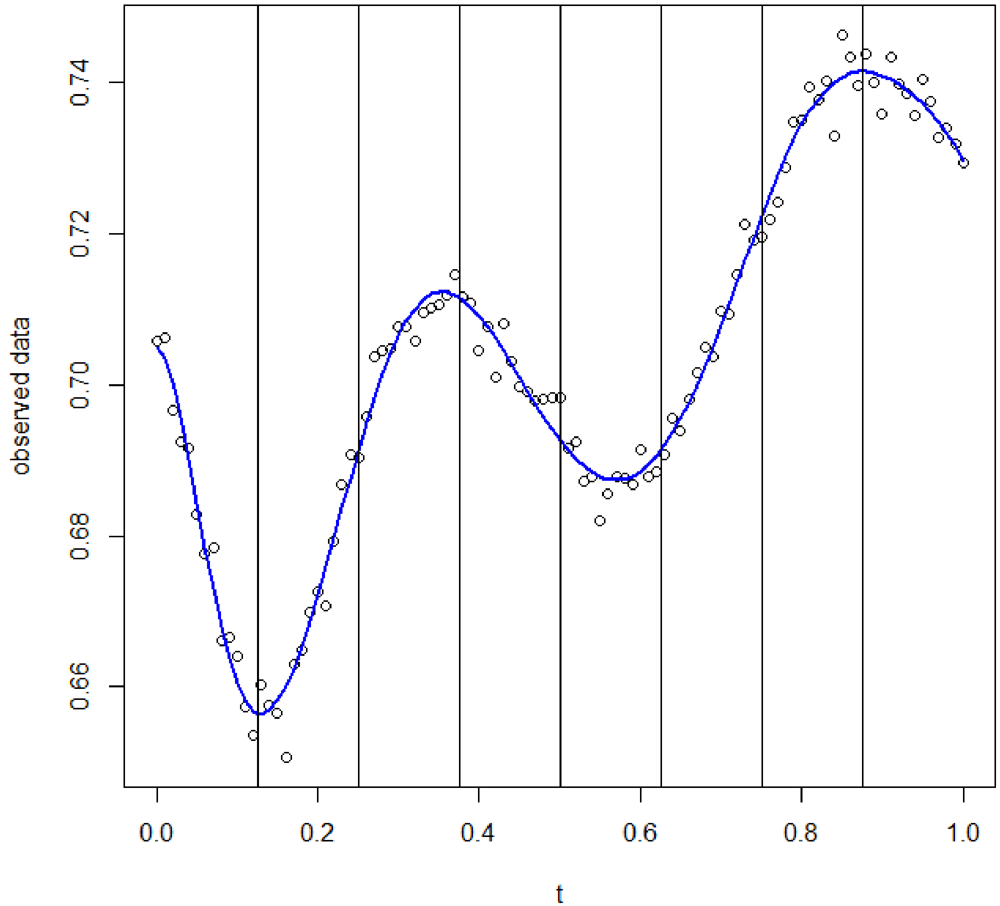
1) Use a functional space with only few dimensions (few basis) K << n



K = 7

(Spline of order 5)

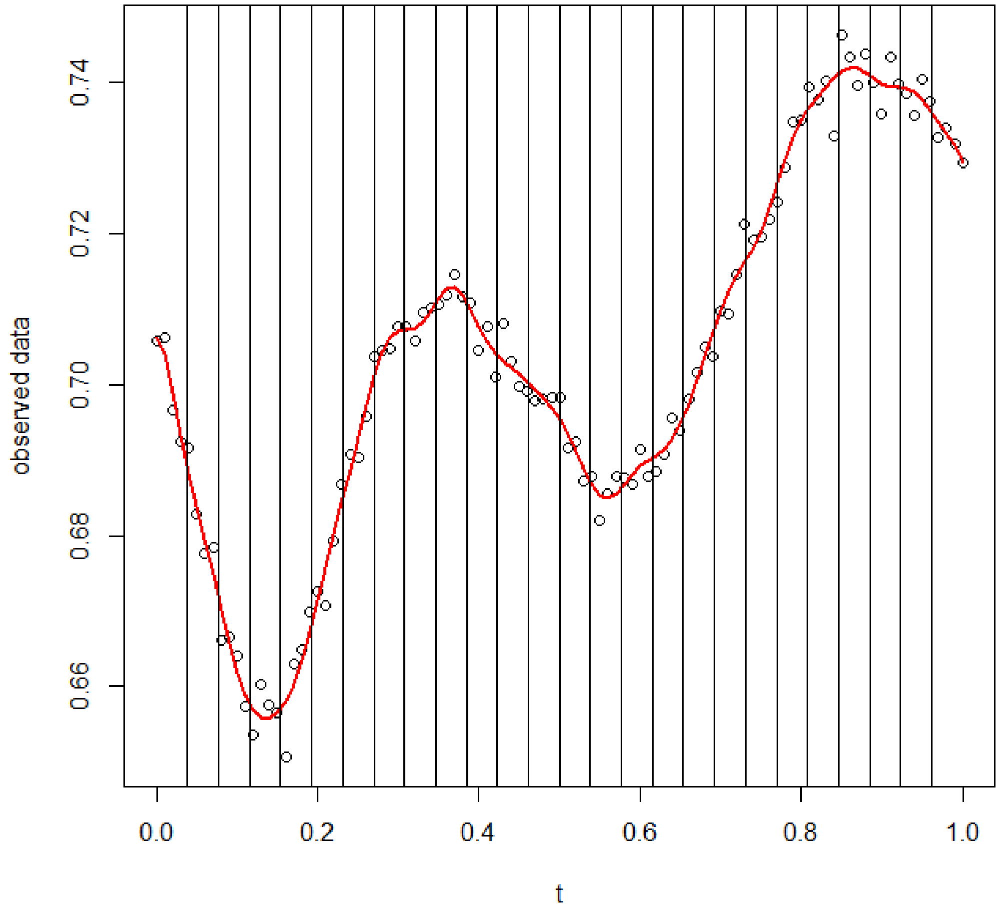
1) Use a functional space with only few dimensions (few basis) K << n



K = 12

(Spline of order 5)

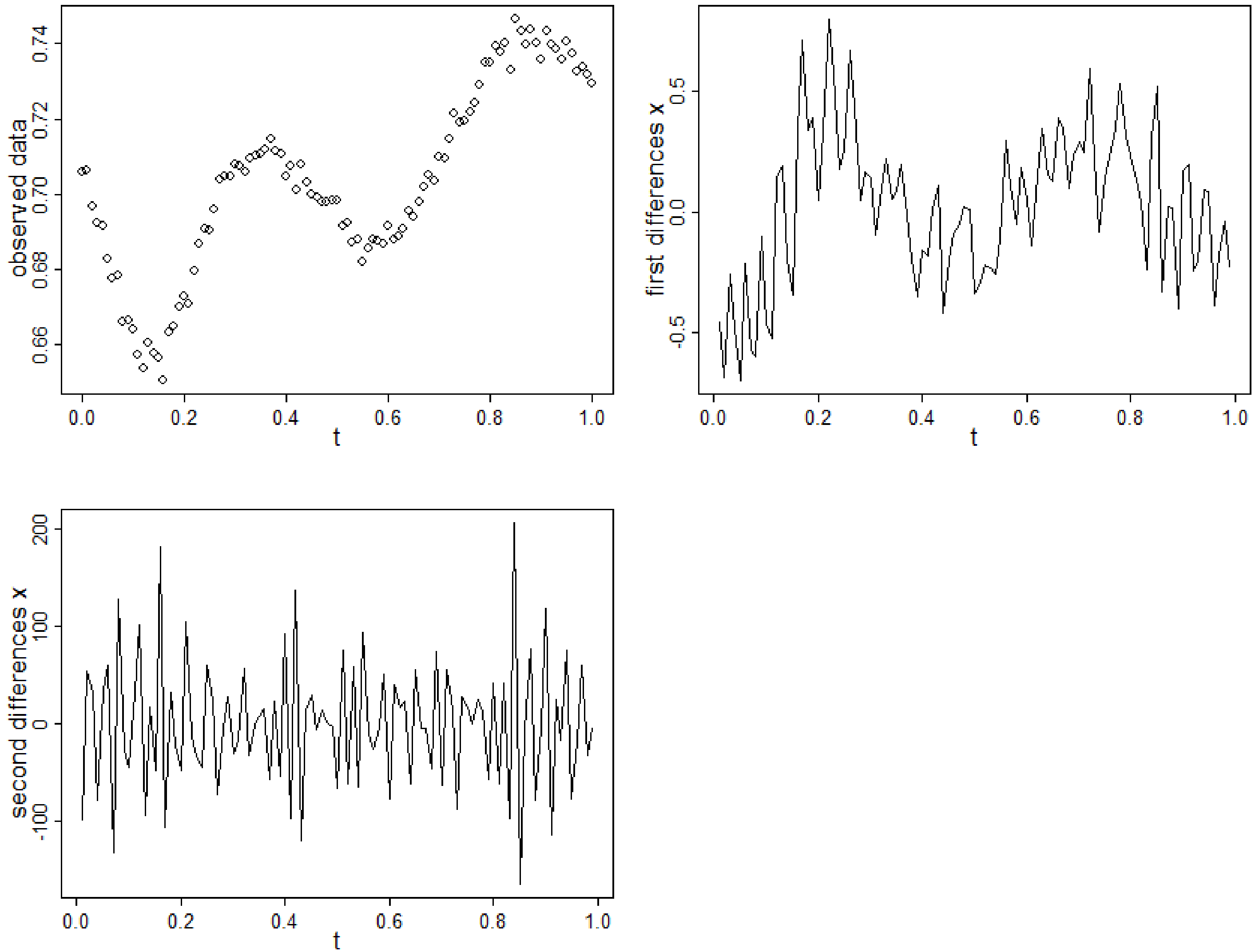
1) Use a functional space with only few dimensions (few basis) K << n



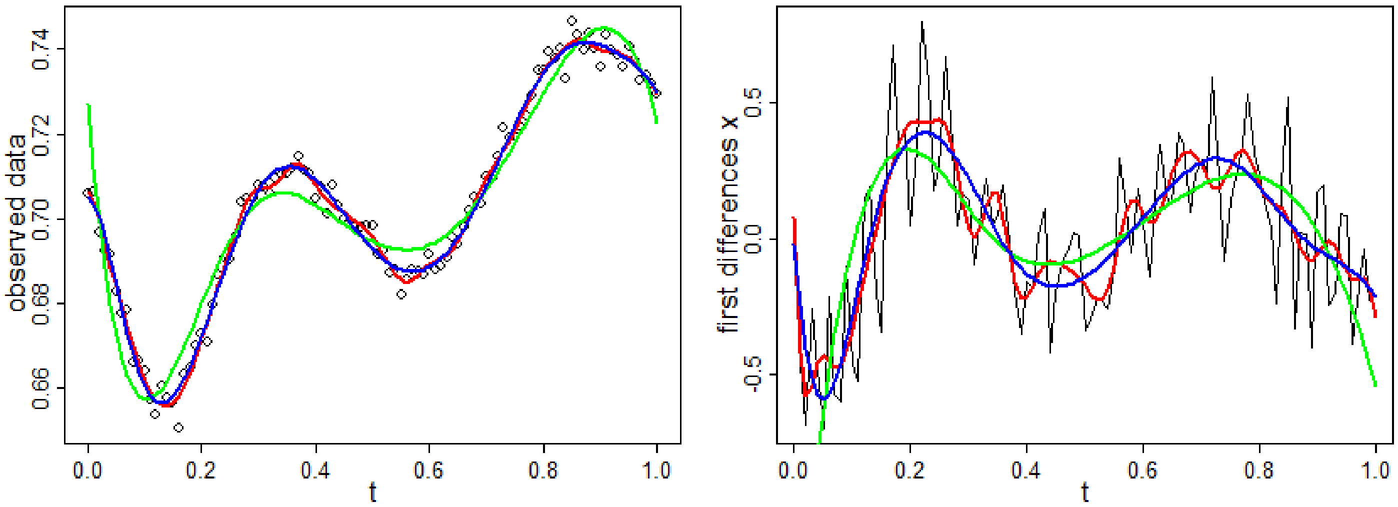
K = 30

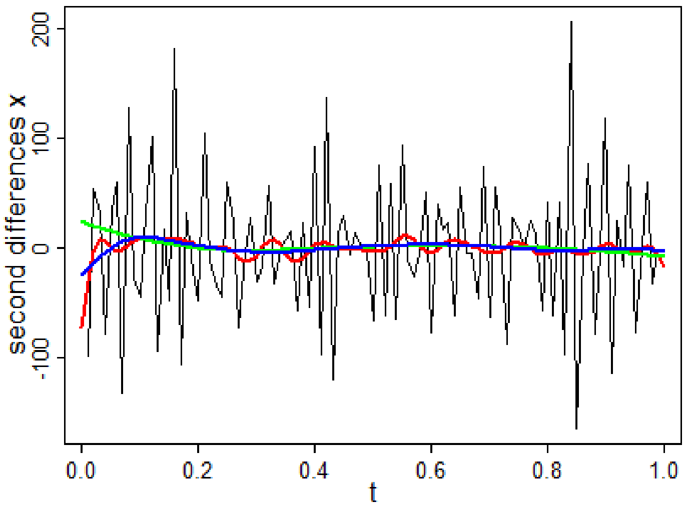
(Spline of order 5)

Smoothing, curve fitting



Smoothing, curve fitting





Choose the number K of basis

K << N

K = 7

K = 12

K = 30

2) Use a rich functional space but with regularization K ~ n

$$\mathrm{SSE}_\lambda = \mathrm{SSE} + \lambda \int \left(f''(s)\right)^2 ds$$

$$\{R_\psi\}_{(k,l)} \quad (k,l)\text{-entry} : \int \psi_k''(s)\psi_l''(s)\, ds$$

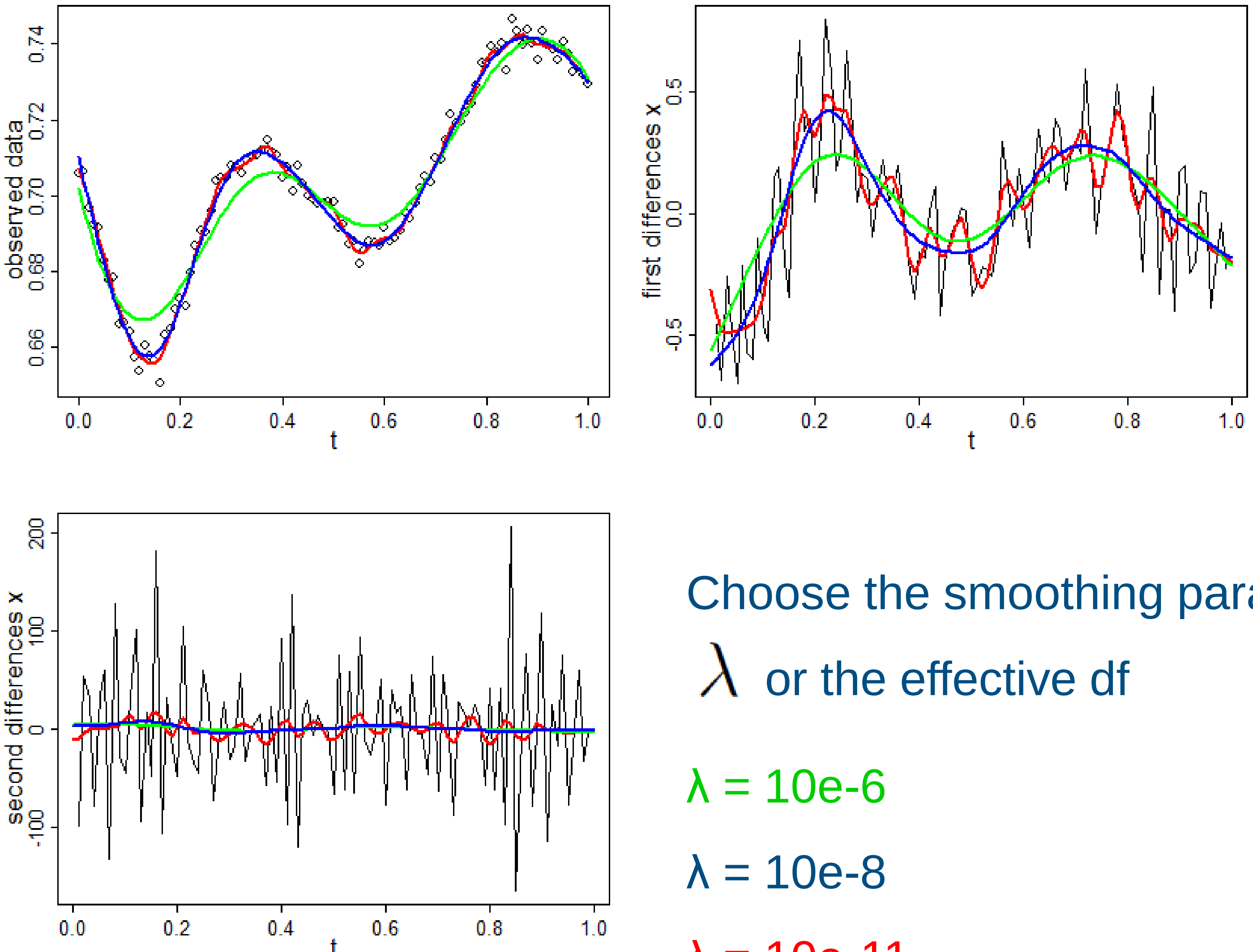
$$\mathrm{SSE}_\lambda = \mathrm{SSE} + \lambda \boldsymbol{c}^t R_\psi \boldsymbol{c}$$

$$\hat{\boldsymbol{c}}_\lambda = \left(\Psi^t \Psi + \lambda R_\psi\right)^{-1} \Psi^t \boldsymbol{z}$$

$$\hat{\boldsymbol{z}} = \hat{\boldsymbol{f}} = \Psi \left(\Psi^t \Psi + \lambda R_\psi\right)^{-1} \Psi^t \boldsymbol{z} = S\boldsymbol{z}$$

$$df = tr(S) < K \qquad \left(\text{or } df = tr(S^t S) \text{ or } df = tr(2S - S^t S)\right)$$

Smoothing, curve fitting



Choose the smoothing parameter λ or the effective df

λ = 10e-6

λ = 10e-8

λ = 10e-11

$$\hat{f}'(s) = \sum_{k=1}^{K} \hat{c}_k \psi'_k(s) \qquad \hat{f}''(s) = \sum_{k=1}^{K} \hat{c}_k \psi''_k(s)$$

Smoothing requires special care when the curve estimate is asked, not only to provide a good smoothing of the data, but also to reflect the features of the curve that are represented by its derivatives

Curve derivatives (or their functions) are very often
- objects of analysis
- helpful for further processing and analysis of the data (curve alignment/clustering)

$$\mathrm{SSE}_\lambda = \mathrm{SSE} + \lambda \int \left(f^{[d]}(s)\right)^2 ds$$

$$\mathrm{SSE}_\lambda = \mathrm{SSE} + \lambda \int \left(L f(s)\right)^2 ds$$

3) Choose basis adaptively to data

K << 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

  Unidimensional curves: see, e.g., Hastie, Tibshirani, Friedman (2009) Springer

  Multidimensional curves: see, e.g., Pigoli, Sangalli (2012) CSDA

- Functional Principal Components Analysis (or other basis constructed from data)

3) Choose basis adaptively to data K << 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

Unidimensional curves: see, e.g., Hastie, Tibshirani, Friedman (20

Multidimensional curves: see, e.g., Pigoli, Sangalli (2012) CSDA

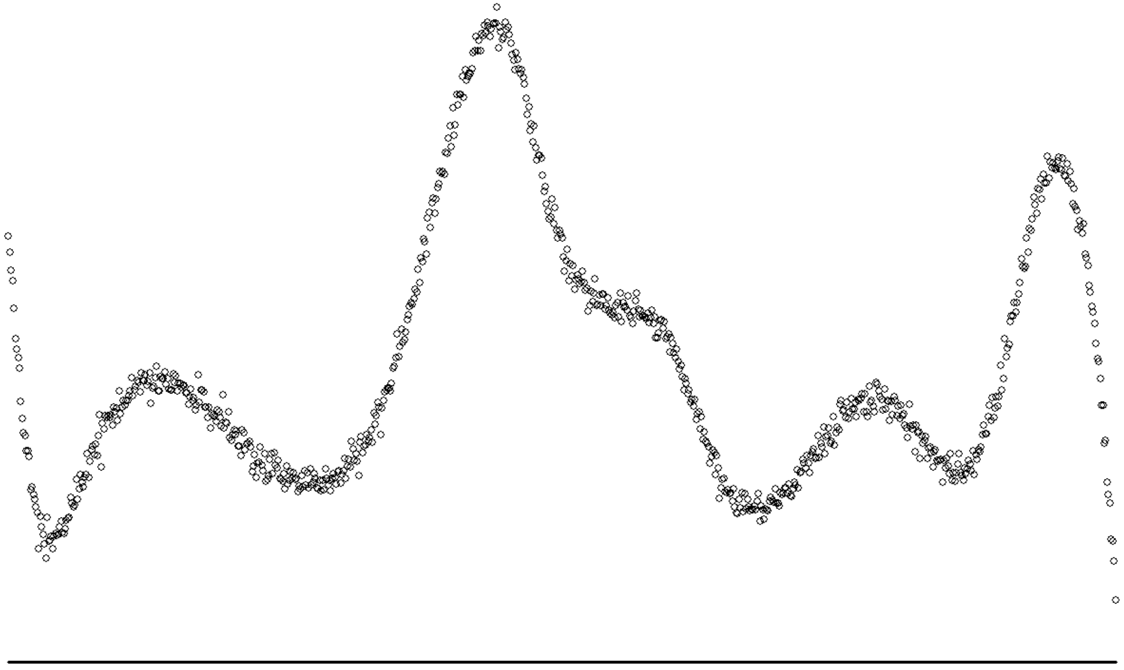
- Good for modeling sharp local features
- Localized in both space and frequency
- An analytical expression may not exist
- Computationally efficient (orthogonal)

- Functional Principal Components Analysis (or other basis constructed from data)

Iterative algorithm for the search of optimal knots in free-knot regression splines

Free knot regression splines (e.g., Zhou-Shen JASA 2001)

K << n
Adaptive basis

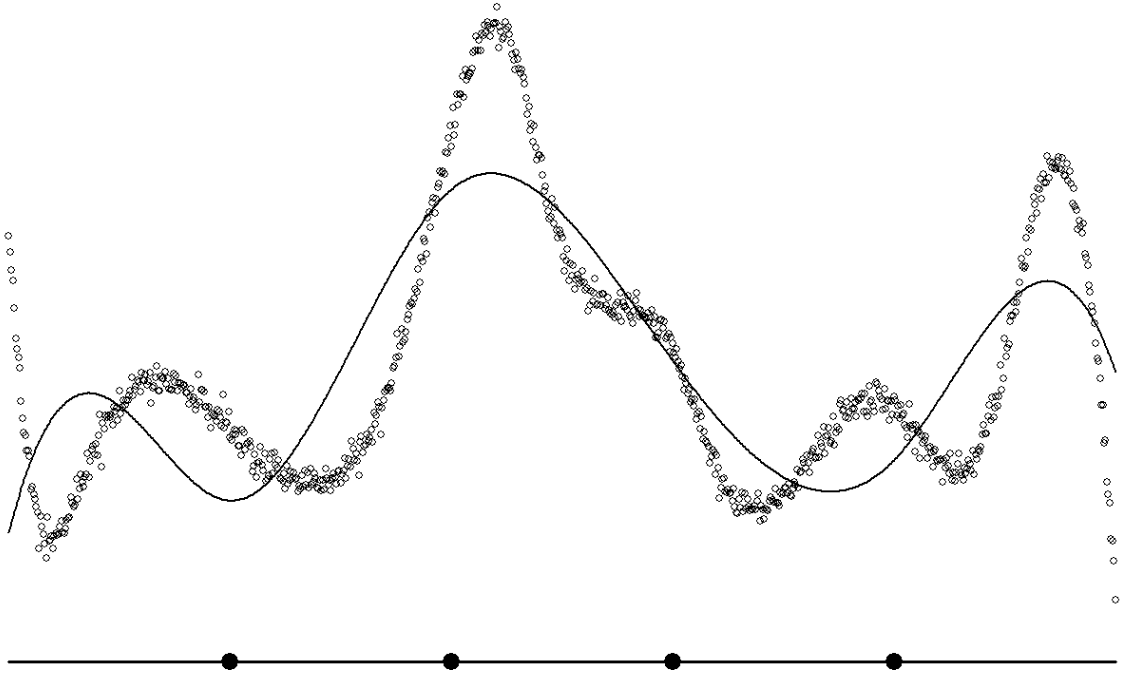


$$\mathrm{SSE} + \lambda K$$

Iterative algorithm for the search of optimal knots in free-knot regression splines

Free knot regression splines (e.g., Zhou-Shen JASA 2001)

K << n
Adaptive basis



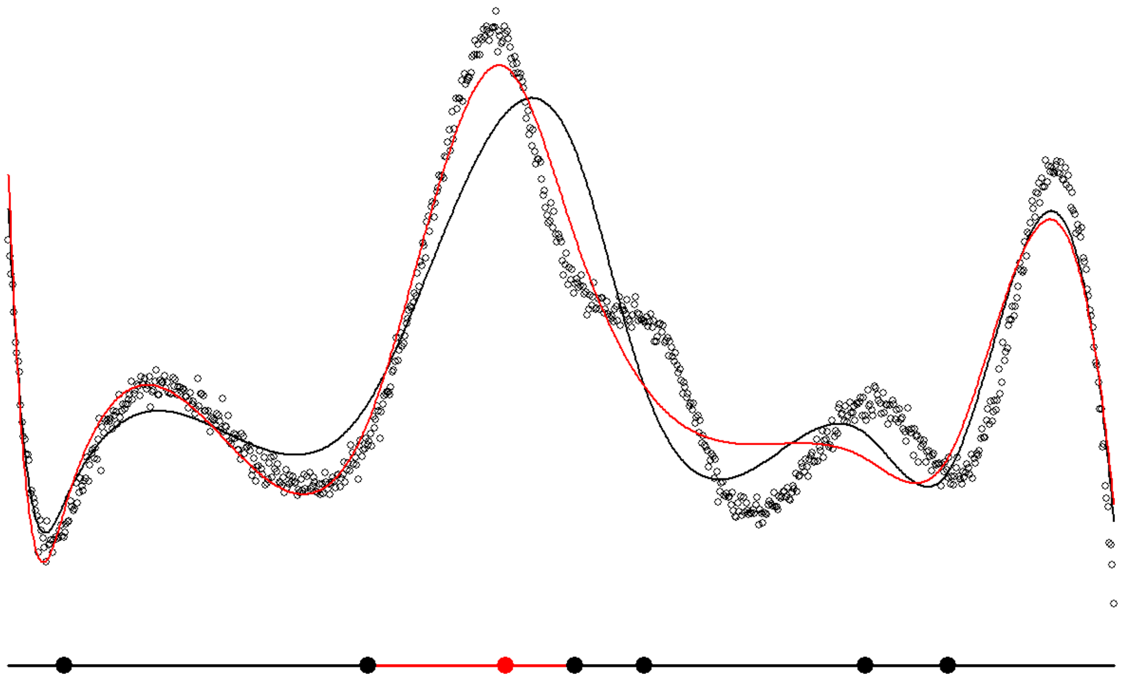
$$\mathrm{SSE} + \lambda K$$

knot initialization

Iterative algorithm for the search of optimal knots in free-knot regression splines

Free knot regression splines (e.g., Zhou-Shen JASA 2001)

K << n
Adaptive basis



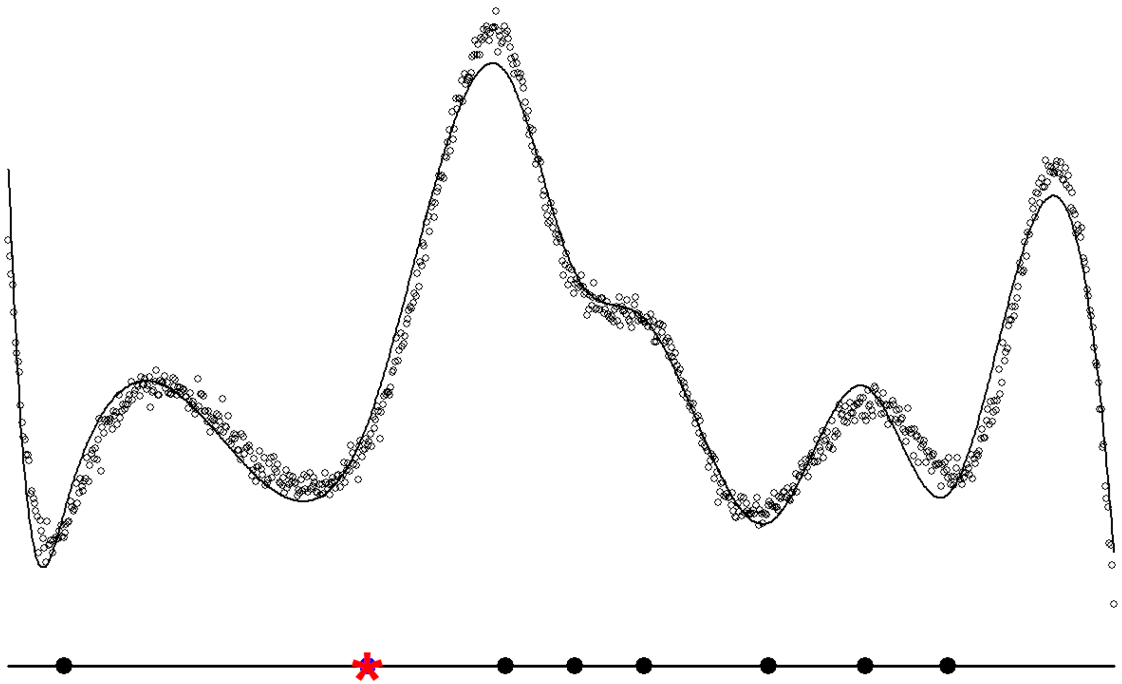
$$\mathrm{SSE} + \lambda K$$

knot addition

Iterative algorithm for the search of optimal knots in free-knot regression splines

Free knot regression splines (e.g., Zhou-Shen JASA 2001)

K << n
Adaptive basis



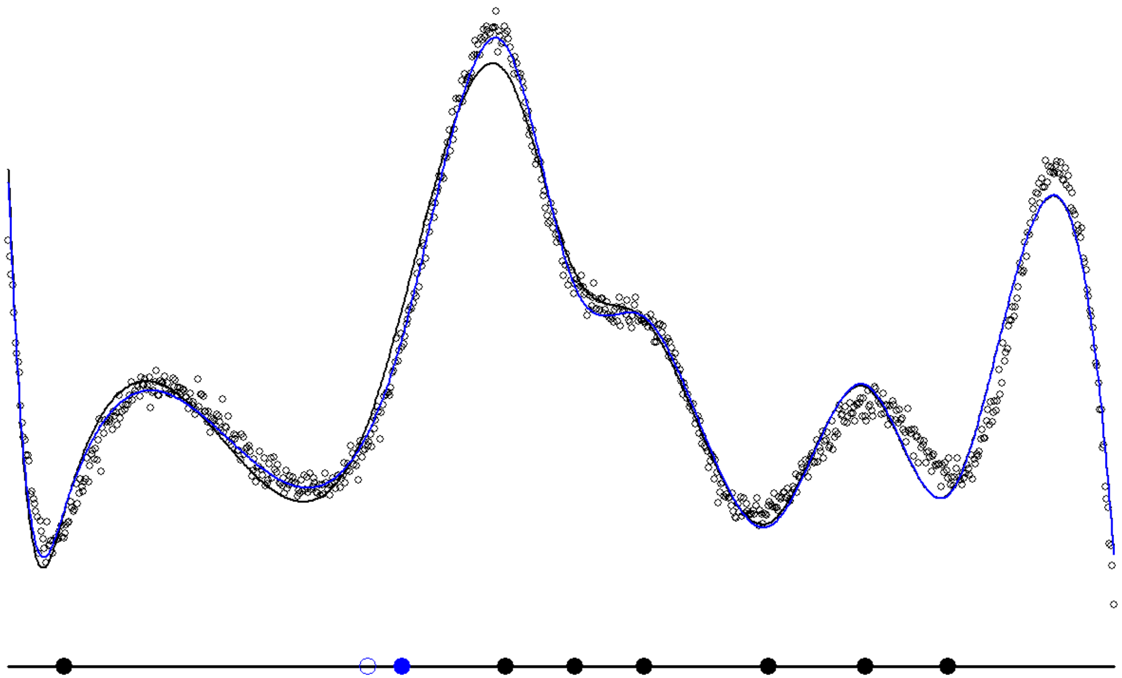
$$\mathrm{SSE} + \lambda K$$

Knot delation/relocation

Iterative algorithm for the search of optimal knots in free-knot regression splines

Free knot regression splines (e.g., Zhou-Shen JASA 2001)

K << n
Adaptive basis



$$\mathrm{SSE} + \lambda K$$

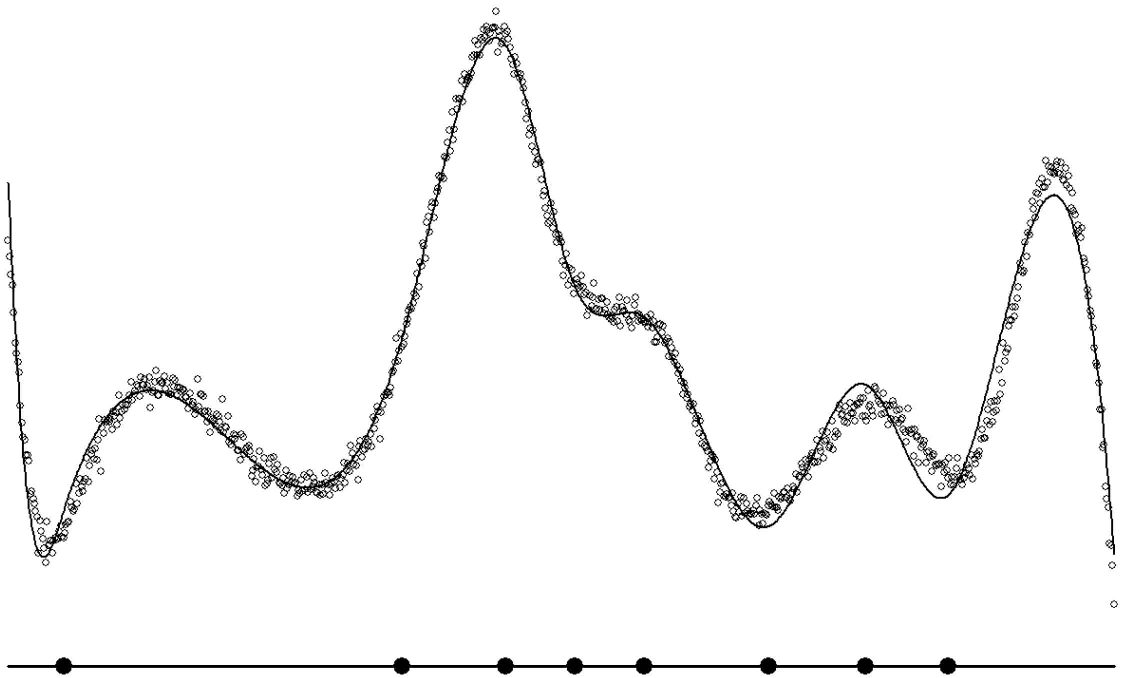
Knot delation/relocation

Iterative algorithm for the search of optimal knots in free-knot regression splines

Free knot regression splines (e.g., Zhou-Shen JASA 2001)

K << n

Adaptive basis

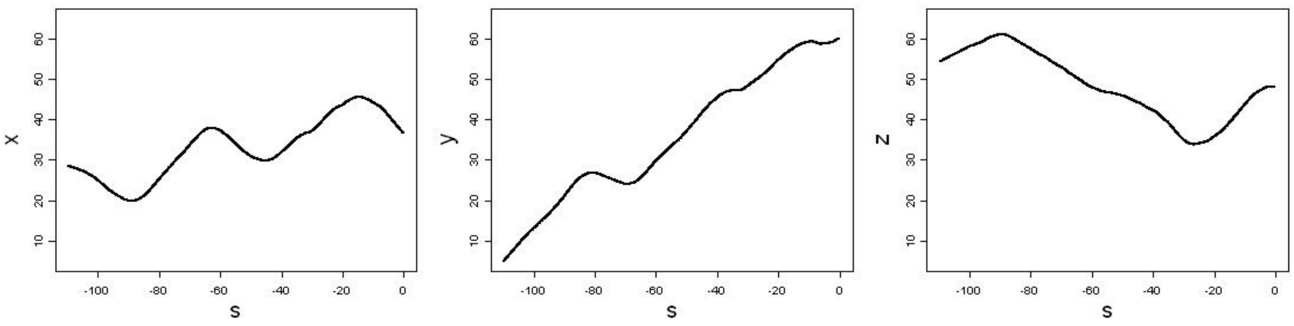


$$\mathrm{SSE} + \lambda K$$

Stopping rule

Centerline data

COORDINATES
PATIENT 1



Very high signal-to-noise ratio

Fine grid of observed points

BACK to AneuRisk data

Preprocessing:
accurate curve estimates

$$\{b_{r,m}^{[\mathbf{k}]}(s) : r = 1, \ldots, m + n_k\}$$

b-spline basis system for the vector space of splines of order m with knot vector $\mathbf{k} = (k_1, \ldots, k_{n_k})$

Functional estimates of the 3-spatial coordinates $\left(\hat{x}(s), \hat{y}(s), \hat{z}(s)\right)$

by **3D free-knot regression splines**

$$\hat{x}(s) = \sum_{r=1}^{m+n_{\hat{k}}} \hat{\lambda}_r^{[x]} b_{r,m}^{[\hat{\mathbf{k}}]}(s) \qquad \hat{y}(s) = \sum_{r=1}^{m+n_{\hat{k}}} \hat{\lambda}_r^{[y]} b_{r,m}^{[\hat{\mathbf{k}}]}(s) \qquad \hat{z}(s) = \sum_{r=1}^{m+n_{\hat{k}}} \hat{\lambda}_r^{[z]} b_{r,m}^{[\hat{\mathbf{k}}]}(s)$$

FIND

$$\hat{n}_k, \hat{\mathbf{k}} = \left(\hat{k}_1(s), \ldots, \hat{k}_{n_k}(s)\right), \hat{\lambda}^{[x]}, \hat{\lambda}^{[y]}, \hat{\lambda}^{[z]}$$

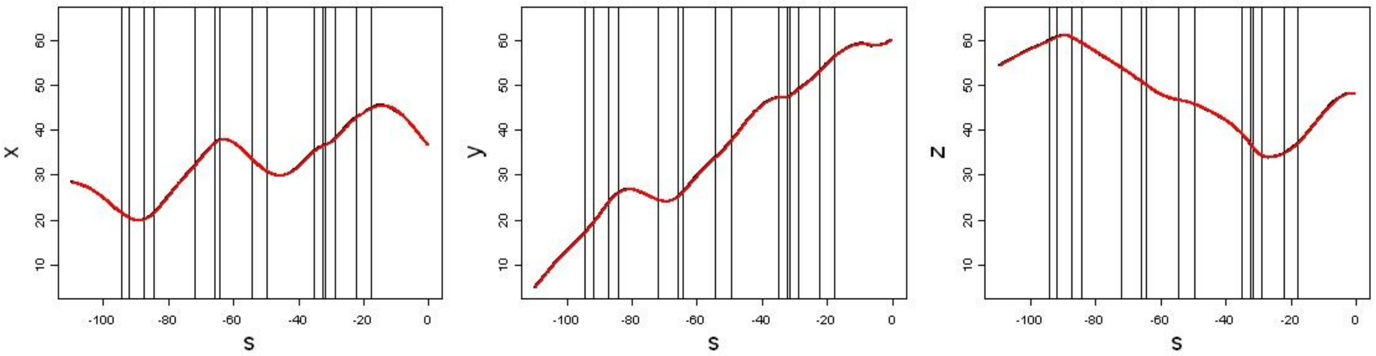
by minimizing

$$\sum_{j=1}^{n} \left(x_j - \sum_{r=1}^{m+n_k} \lambda_r^{[x]} b_{r,m}^{[\mathbf{k}]}(s_j)\right)^2 + \sum_{j=1}^{n} \left(y_j - \sum_{r=1}^{m+n_k} \lambda_r^{[y]} b_{r,m}^{[\mathbf{k}]}(s_j)\right)^2 + \sum_{j=1}^{n} \left(z_j - \sum_{r=1}^{m+n_k} \lambda_r^{[z]} b_{r,m}^{[\mathbf{k}]}(s_j)\right)^2 + \mathcal{C}(m + n_k)$$

FIX $m = 5$ to obtain smooth estimates of the curvature (function of second derivative)

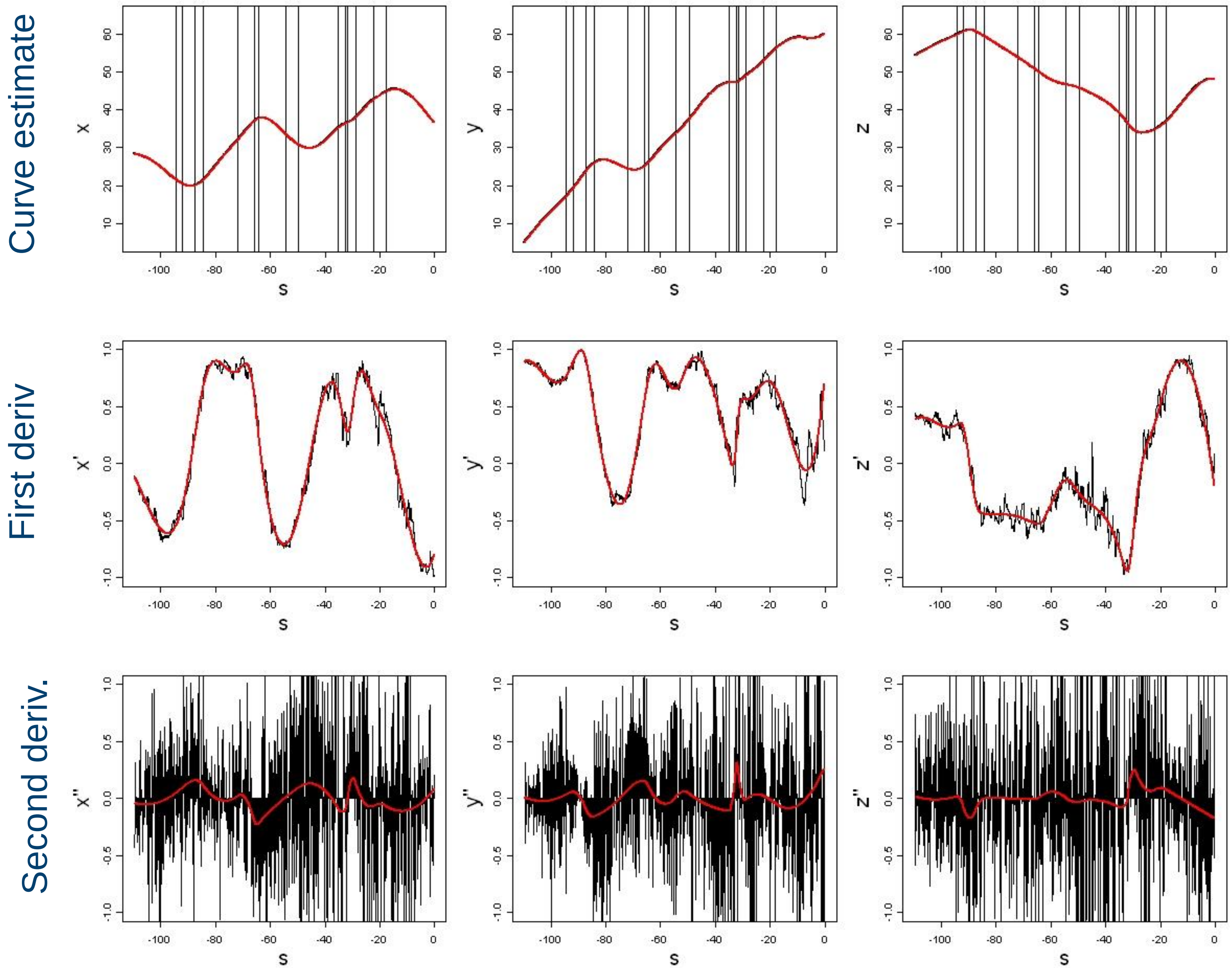
3D free-knot splines

Curve estimate

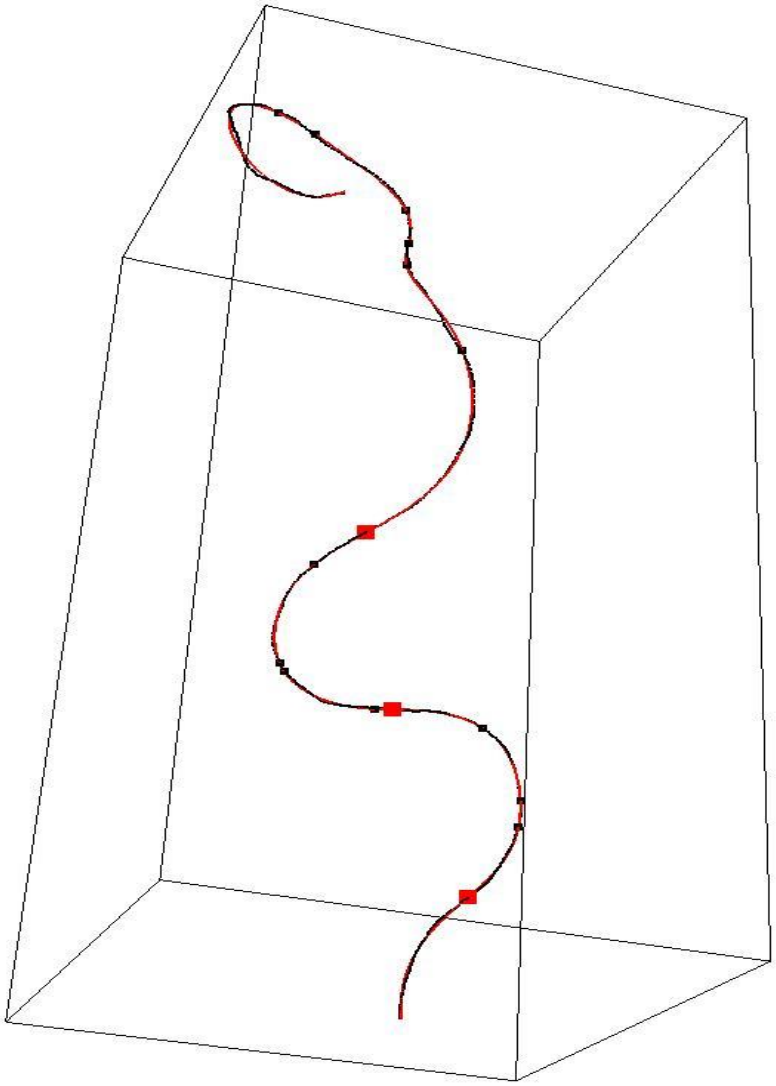


Derivatives of splines are still splines with the same knot vector and coefficients directly computed from the coefficients of the original spline

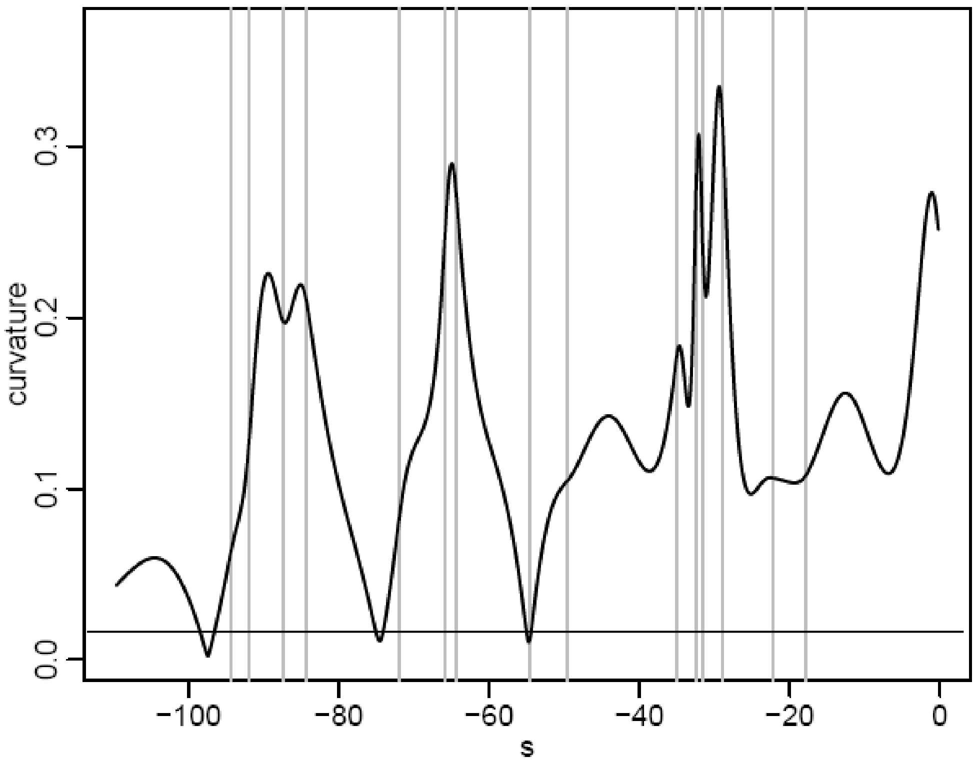
3D free-knot splines



3D free-knot splines

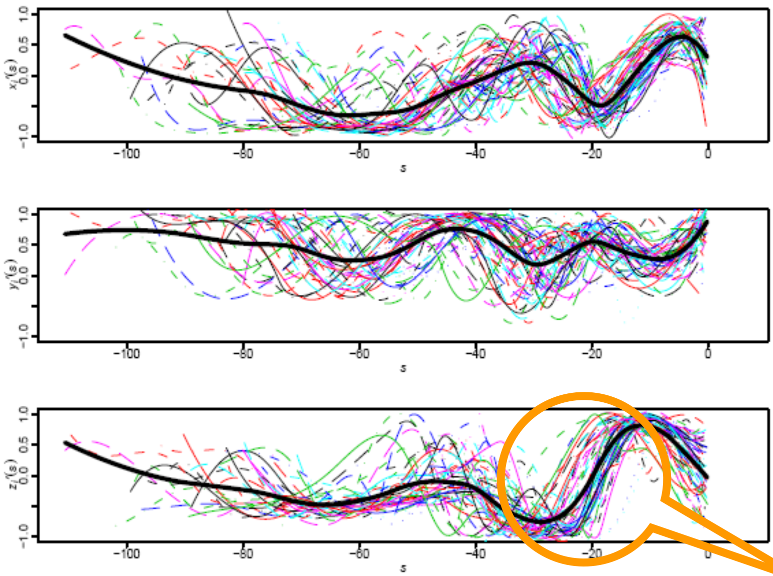


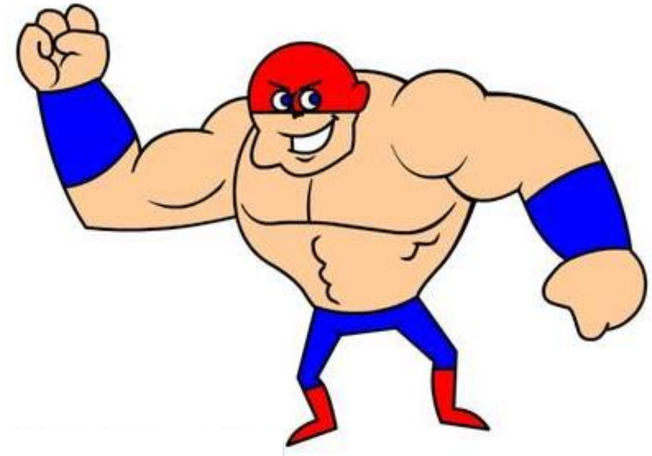
Curvature function



$$C_i(s) = \frac{||(x_i'(s)\ y_i'(s)\ z_i'(s)) \times (x_i''(s)\ y_i''(s)\ z_i''(s))||}{||(x_i'(s)\ y_i'(s)\ z_i'(s))||^3}$$

Centerline first derivatives





Phase Variability
(strongly dependent on dimensions of
body structure and arteries)

Visualise data

► To enable meaningful comparisons across patients we need to decouple between-patients *phase variability* and between-patients *amplitude variability*

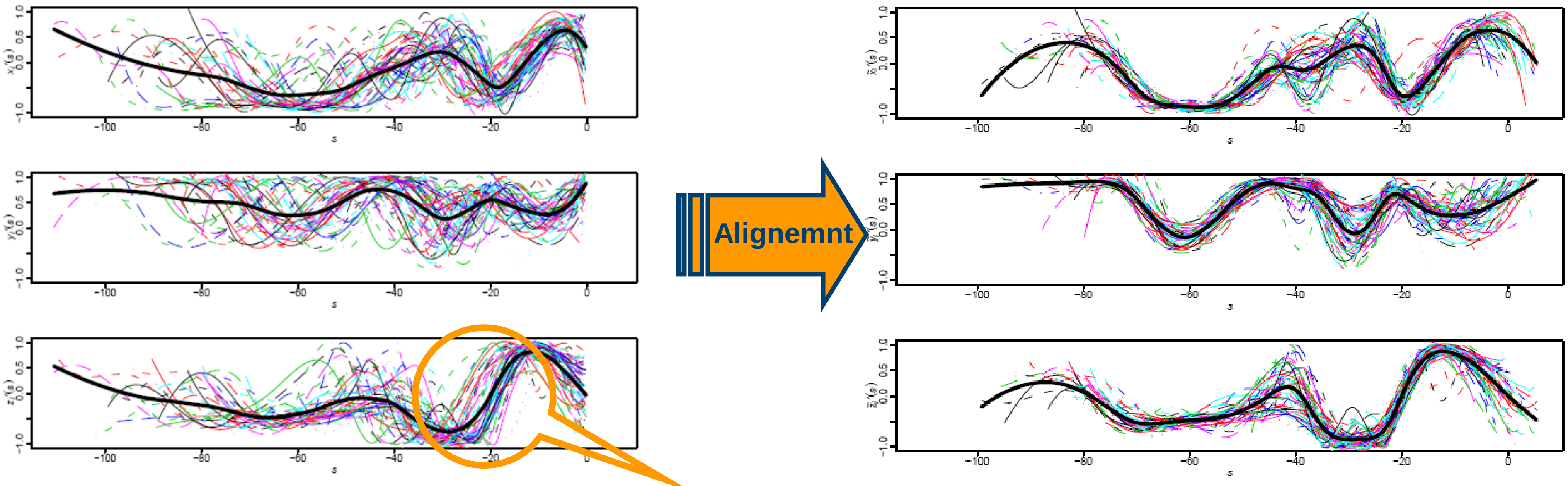
due to *differences in the dimensions* of patients carotids

due to *differences in the morphological shapes* of patients carotids

Curve alignment

Sangalli Secchi Vantini Veneziani, 2009, JASA

Centerline first derivatives



Phase Variability
(strongly dependent on dimensions of
body structure and arteries)

Visualise data

► 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

Decoupling and studying Phase and Amplitude variabilities

Registration, Alignment, Warping

MBI contest

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)

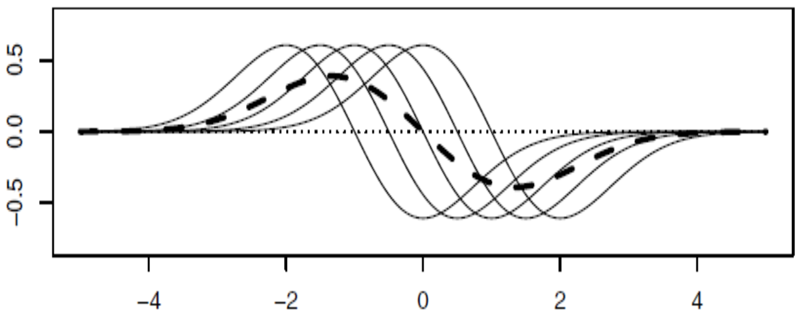
| Description | Schedule | Participants | Titles & Abstracts | Resources | **Apply for Event** | Flyer

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.

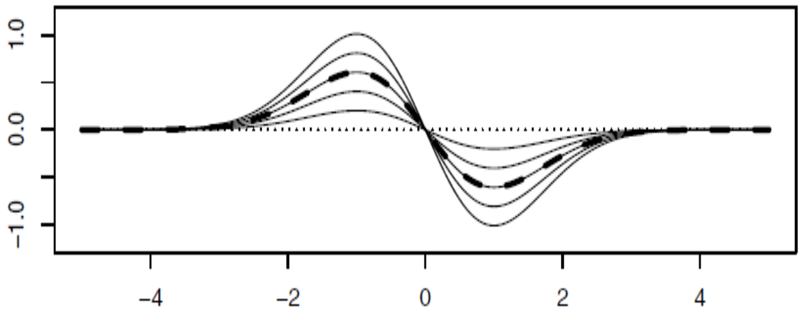
The registration or alignment of features in curves and images by smooth, one-to-one transformations of time or space, respectively, is an emerging hot topic that presents many challenges. From its beginnings with dynamic time warping in the late 50's, followed by the landmark registration methods of 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

→ Forthcoming Special Section of the *Electronic Journal of Statistics*

Phase variability



Amplitude variability



<u>Registration of a set of functions</u>

Find **suitable** warping functions $h_1, \ldots, h_n$ such that $c_1 \circ h_1, \ldots, c_n \circ h_n$ are the most **similar**.

Landmark Approach: known **landmarks** along the curves thar are alinged 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.

$\mathcal{C}$: set of curves $\mathbf{c}(s) : \mathbb{R} \rightarrow \mathbb{R}^d$

Aligning $\mathbf{c}_1(s) \in \mathcal{C}$ to $\mathbf{c}_2(s) \in \mathcal{C}$ means finding a warping function $h(s) : \mathbb{R} \rightarrow \mathbb{R}$

such that the two curves $\mathbf{c}_1(h(s))$ and $\mathbf{c}_2(s)$ are the most similar

$\mathcal{C}$: set of curves $\mathbf{c}(s) : \mathbb{R} \rightarrow \mathbb{R}^d$

Aligning $\mathbf{c}_1$ to $\mathbf{c}_2$ according to (ρ, W) means finding $h^* \in W$

that maximizes $\rho(\mathbf{c}_1 \circ h, \mathbf{c}_2)$ $\quad (\mathbf{c} \circ h)(s) := \mathbf{c}(h(s))$

Similarity index $\rho(\cdot, \cdot) : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}$

Class W of warping functions $h(s) : \mathbb{R} \rightarrow \mathbb{R}$

►The choice of (ρ, W) is *problem-specific*

It defines what is meant by *phase* and *amplitude* variability

Phase and amplitude variabilities

$\mathcal{C}$: set of curves $\mathbf{c}(s) : \mathbb{R} \to \mathbb{R}^d$

Aligning $\mathbf{c}_1$ to $\mathbf{c}_2$ according to (ρ, W) means finding $h^* \in W$

that maximizes $\rho(\mathbf{c}_1 \circ h, \mathbf{c}_2)$ $\qquad (\mathbf{c} \circ h)(s) := \mathbf{c}(h(s))$

Similarity index $\rho(\cdot,\cdot) : \mathcal{C} \times \mathcal{C} \to \mathbb{R}$

AneuRisk data:

two vessel centerlines can be considered similar if they are identical except for shifts and dilations along the three main axes

► The choice of (ρ, W) is *problem-specific*

It defines what is meant by *phase* and *amplitude* variability

Phase and amplitude variabilities

$\mathcal{C}$: set of curves $\mathbf{c}(s): \mathbb{R} \to \mathbb{R}^d$

Aligning $\mathbf{c}_1$ to $\mathbf{c}_2$ according to (ρ, W) means finding $h^* \in W$ that maximizes $\rho(\mathbf{c}_1 \circ h, \mathbf{c}_2)$ $\quad (\mathbf{c} \circ h)(s) := \mathbf{c}(h(s))$

Similarity index $\rho(\cdot,\cdot): \mathcal{C} \times \mathcal{C} \to \mathbb{R}$

$$\rho(\mathbf{c}_1, \mathbf{c}_2) = \frac{1}{d} \sum_{p=1}^{d} \frac{\int c'_{1p}(s) c'_{2p}(s) ds}{\sqrt{\int c'_{1p}(s)^2 ds} \sqrt{\int c'_{2p}(s)^2 ds}}$$

c_{ip}: pth component of $\mathbf{c}_i = (c_{i1}, \ldots, c_{id})$

Class W **of warping functions** $h(s): \mathbb{R} \to \mathbb{R}$

$$W = \{h : h(s) = ms + q \text{ with } m \in \mathbb{R}^+, q \in \mathbb{R}\}$$

$$\rho(\mathbf{c}_1, \mathbf{c}_2) = 1 \quad \Leftrightarrow \quad \text{for } p = 1, \ldots, d,\ \exists \theta_{0p} \in \mathbb{R}, \theta_{1p} \in \mathbb{R}^+ : c_{1p}(s) = \theta_{0p} + \theta_{1p} c_{2p}(s).$$

► The choice of (ρ, W) is *problem-specific*

It defines what is meant by *phase* and *amplitude* variability

Phase and amplitude variabilities

(ρ, W) must satisfy properties that ensure that the aligning problem is well-posed and the corresponding procedure is coherent

- ρ
  - Bounded
  - Reflexive $\rho(\mathbf{c}, \mathbf{c}) = 1$
  - Symmetric $\rho(\mathbf{c}_1, \mathbf{c}_2) = \rho(\mathbf{c}_2, \mathbf{c}_1)$
  - Transitive $[\rho(\mathbf{c}_1, \mathbf{c}_2) = 1 \wedge \rho(\mathbf{c}_2, \mathbf{c}_3) = 1] \Rightarrow \rho(\mathbf{c}_1, \mathbf{c}_3) = 1$
- W
  - Convex vector space
  - Group structure with respect to function composition
- (ρ, W) Properties of coherence

$\rho\,(\mathbf{c}_1, \mathbf{c}_2) = \rho\,(\mathbf{c}_1 \circ h, \mathbf{c}_2 \circ h)\,, \quad \forall\ h \in W$ Invariance, isometry, parallel horbit

(ρ, W) defines on $\mathcal{C}$ a partition in equivalence classes

(the one associated to (ρ, W) in previous slide is the same given by Shape Invariant Models)

shape-analysis

dissimilarity $\mathcal{E}$	class $\mathcal{H}$
$\|\|f_1 - f_2\|\|$	$\mathcal{H}_{shift}$
$\|\|f_1' - f_2'\|\|$	$\mathcal{H}_{shift}$
$\left\|\left\|(f_1 - \bar{f}_1) - (f_2 - \bar{f}_2)\right\|\right\|$	$\mathcal{H}_{shift}$
$\left\|\left\|(f_1' - \bar{f}_1') - (f_2' - \bar{f}_2')\right\|\right\|$	$\mathcal{H}_{shift}$
$\left\|\left\|\frac{f_1}{\|\|f_1\|\|} - \frac{f_2}{\|\|f_2\|\|}\right\|\right\|$	$\mathcal{H}_{affinity}$
$\left\|\left\|\frac{f_1'}{\|\|f_1'\|\|} - \frac{f_2'}{\|\|f_2'\|\|}\right\|\right\|$	$\mathcal{H}_{affinity}$
$\left\|\left\|\text{sign}(f_1')\sqrt{\|f_1'\|} - \text{sign}(f_2')\sqrt{\|f_2'\|}\right\|\right\|$	$\mathcal{H}_{diffeomorphism}$

Curve alignment: iterative procedure

If we had a template (prototype) ICA centerline φ we could align each centerline to this template

The template centerline is unknown and need to be itself estimated from the data

find $\varphi \in \mathcal{C}$ and $\underline{\mathbf{h}} = \{h_1, \ldots, h_N\} \subset W$ 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 $\underline{\mathbf{g}} = \{g_1, \ldots, g_N\} \subset W$

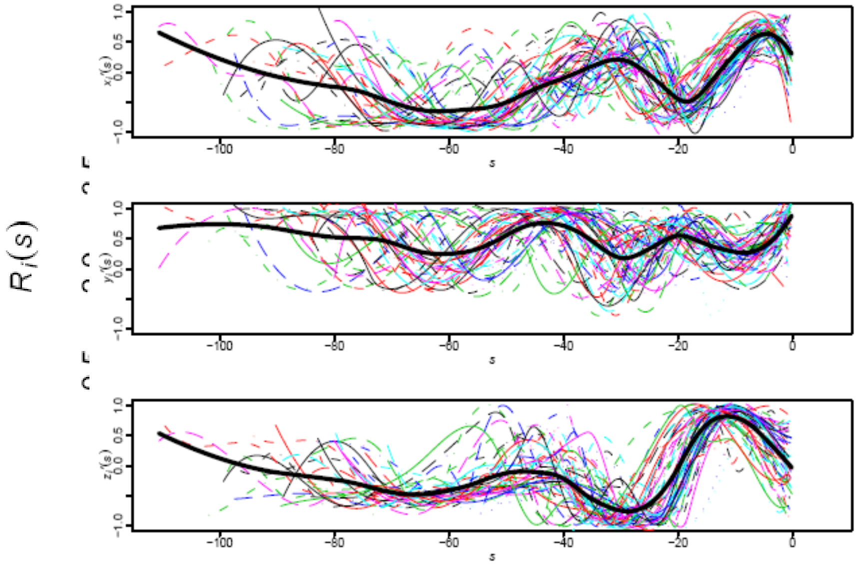
➡ Iterative Procrustes procedure that alternates

- *template estimation step*: the template centerline is estimated from the curves obtained in the previous alignment step

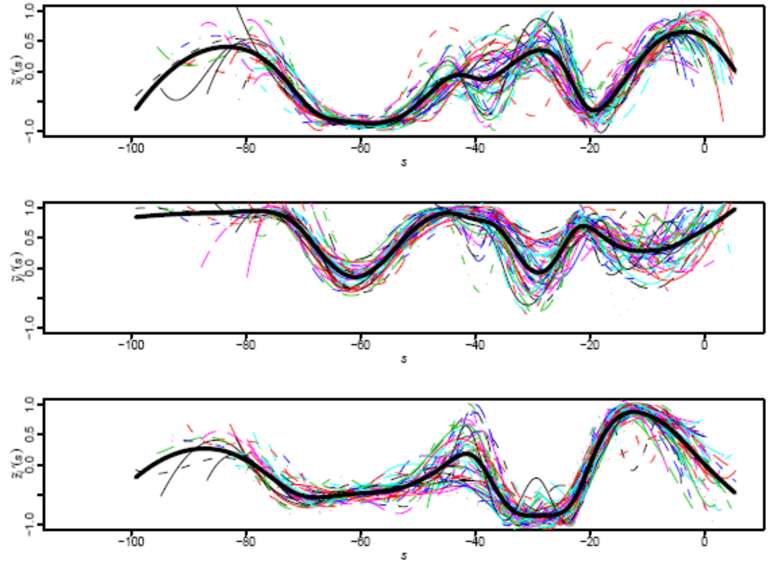
- *aligment step*: the centerlines are aligned to the template centerline estimated in the previos template estimation step

Aneurysm location on aligned ICA radius and curvature profiles

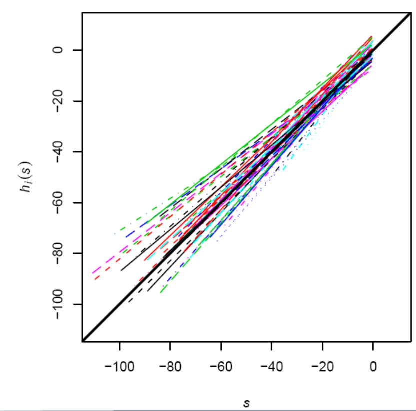
Original centerlines



Aligned centerlines



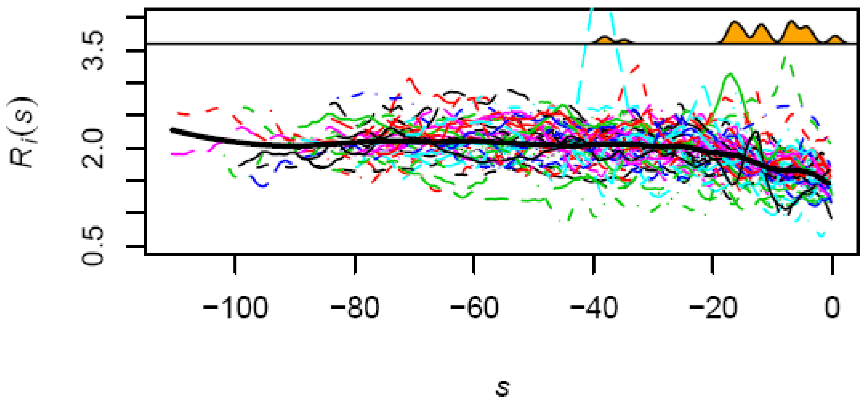
Warping functions (phase variab)

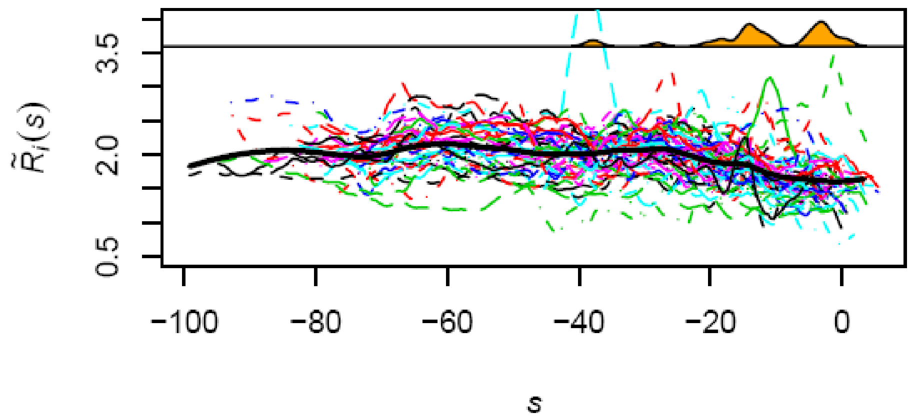


Extract value from data

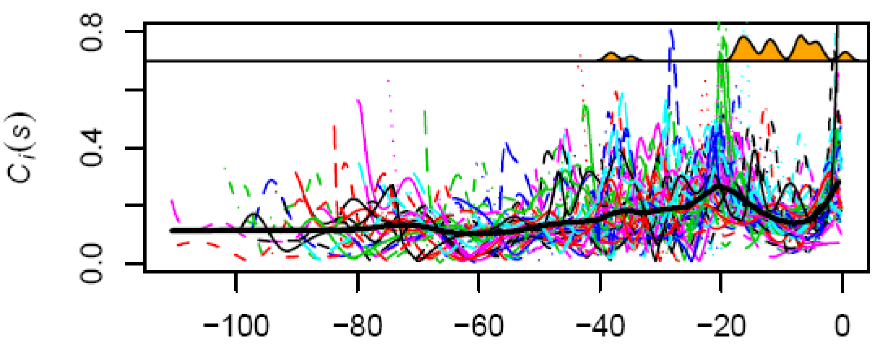
Aneurysm location on aligned ICA radius and curvature profiles

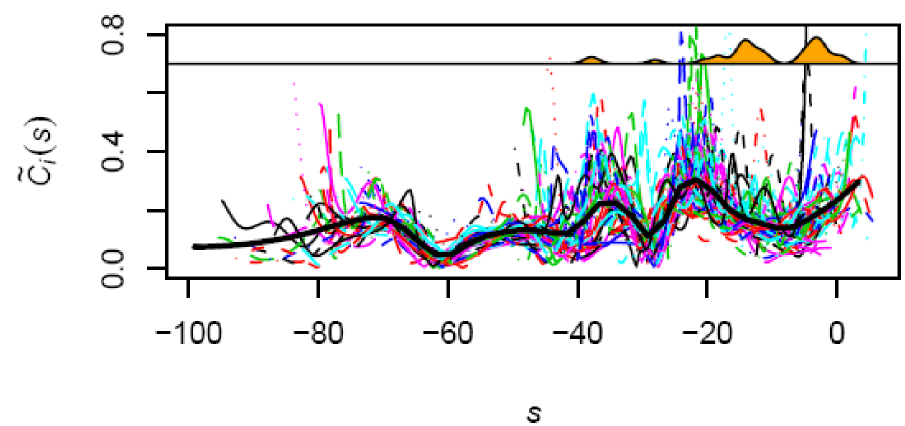
Radius





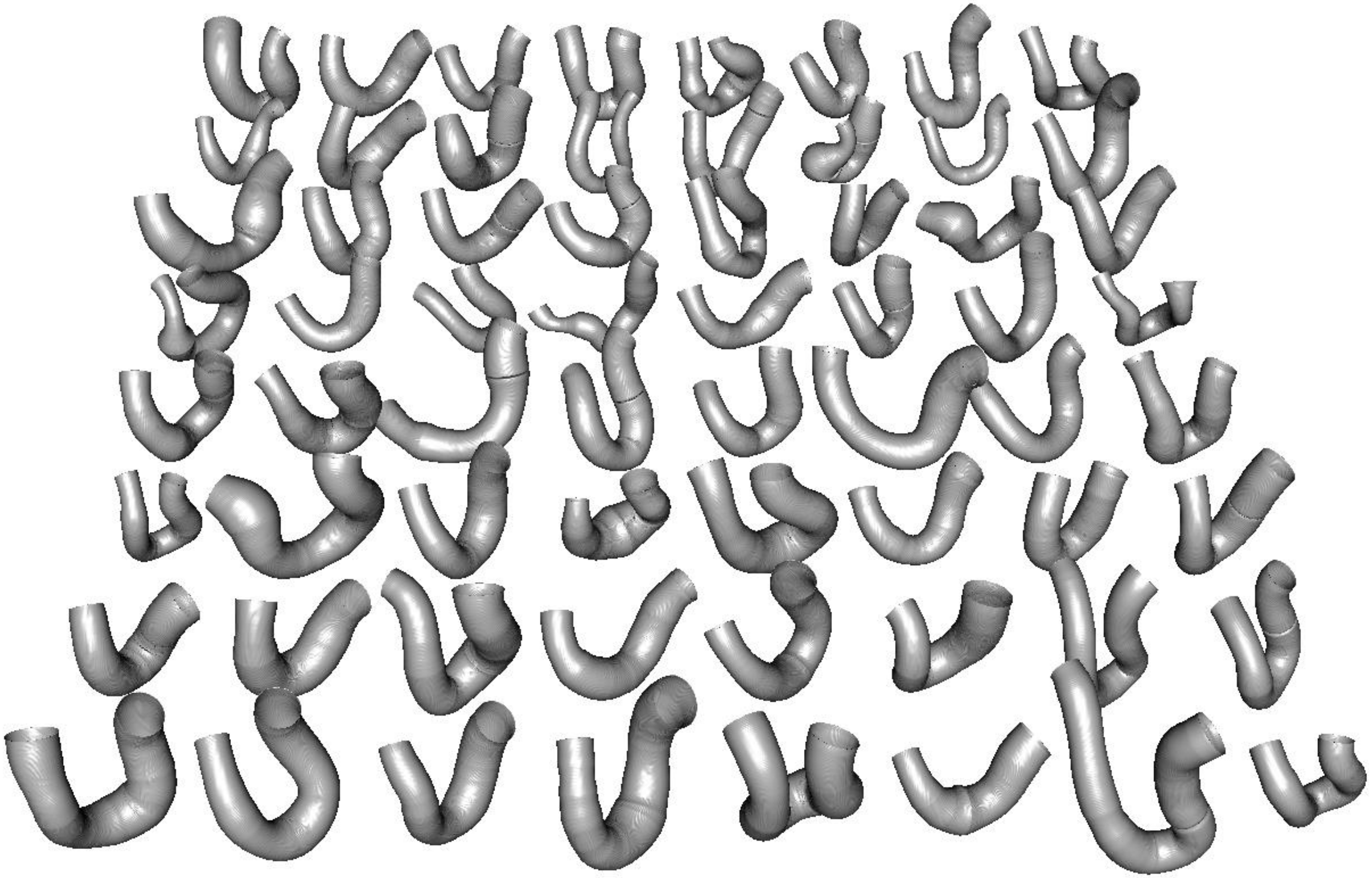
Curvature





Extract value from data

The sample of 65 ICA: each patient is represented by the **REGISTERED** centerline and radius of ICA

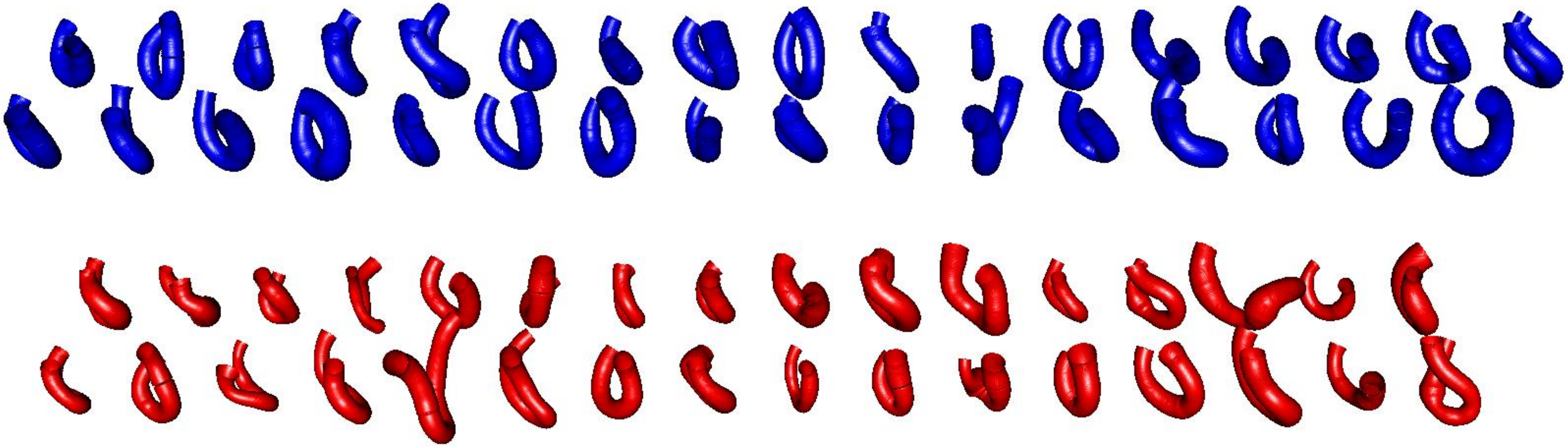


Two groups of patients

We want to use vessel radius and curvature to discriminate:

Aneurysm at or after ICA bifurcation

Upper Group: 33



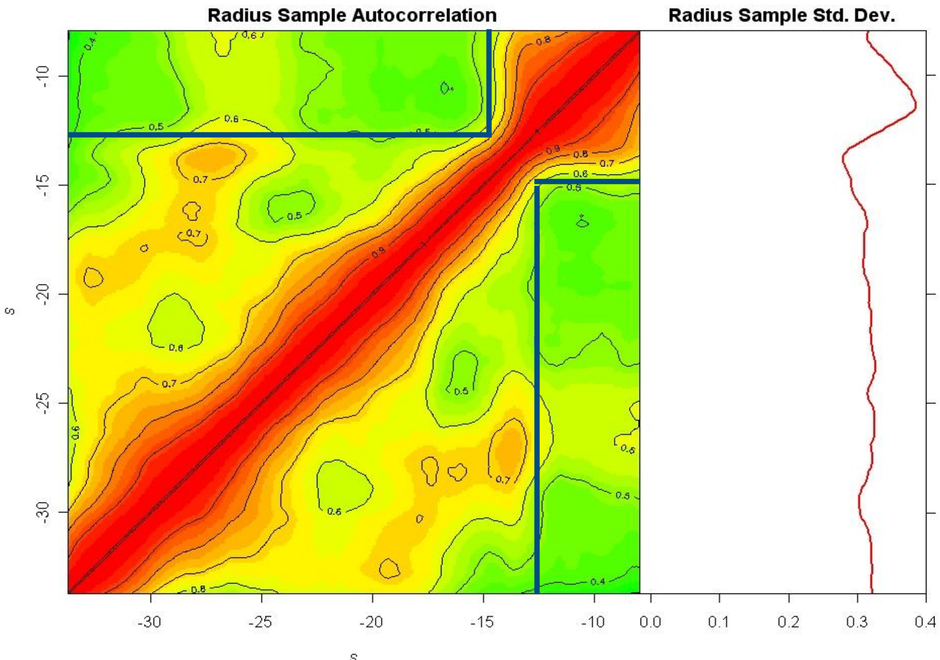
Lower Group: 32

Aneurysm before ICA bifurcation or no aneurysm

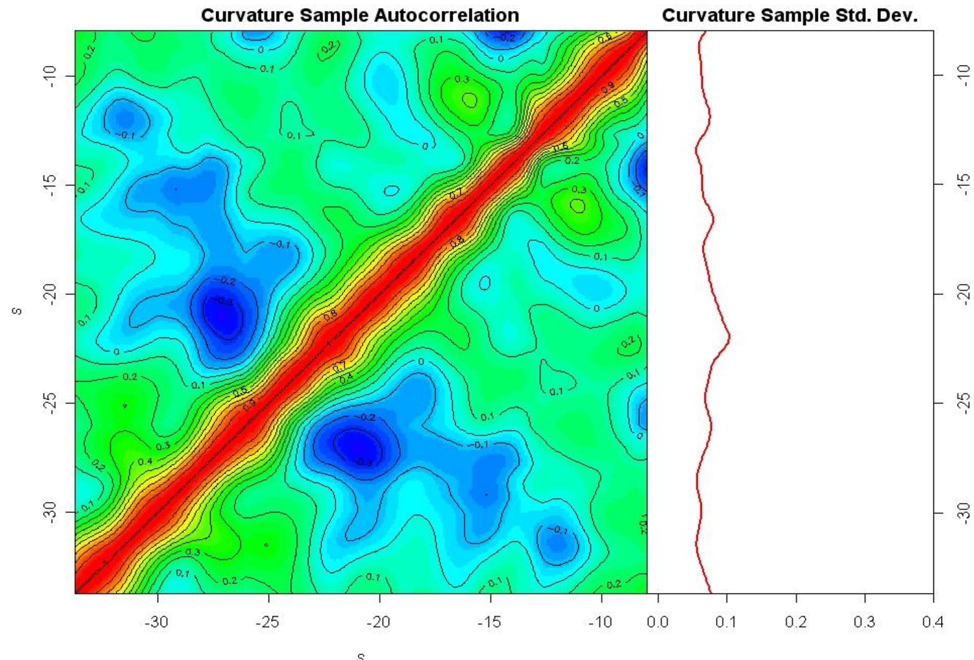
Extract value from data

Covariance structure

Sample Autocorrelation Function and Std. Dev. for Radius Profiles of aligned centerlines



Sample Autocorrelation Function and Std. Dev. for Curvature Profiles of aligned centerlines



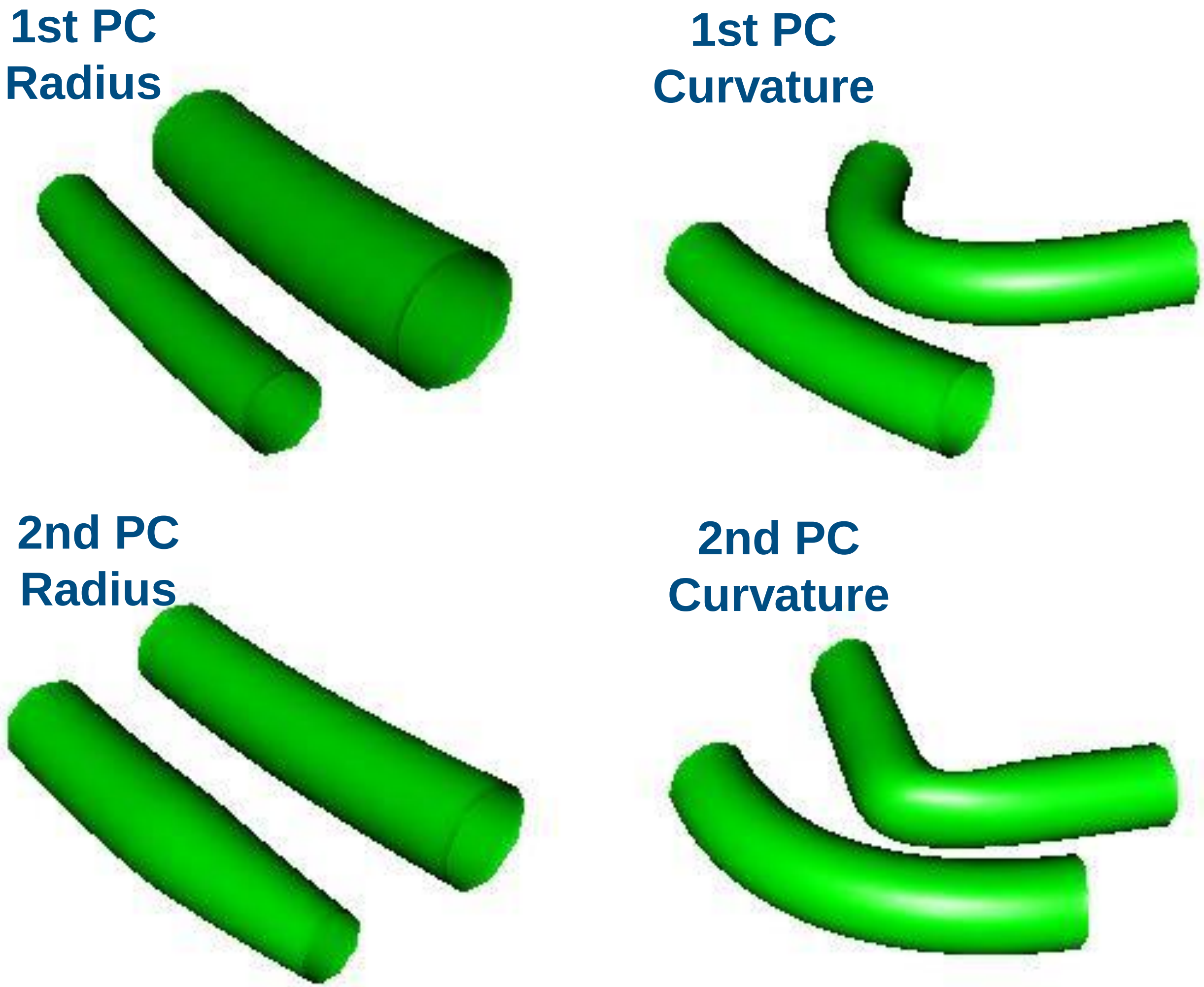
First step: esplore variability by through spectral decomposition of radius and curvature sample autocovariance functions (Functional Principal Component Analysis)

Second step: quadratic discriminant analysis on relevant scores

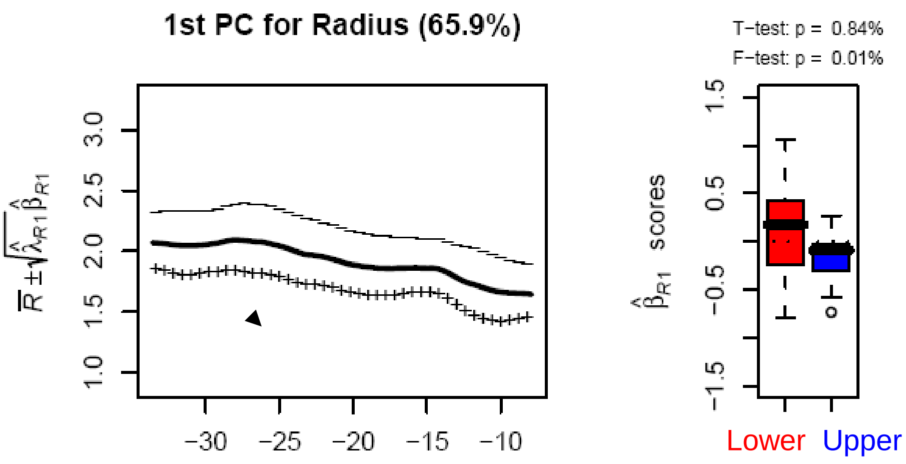
Explore population variability

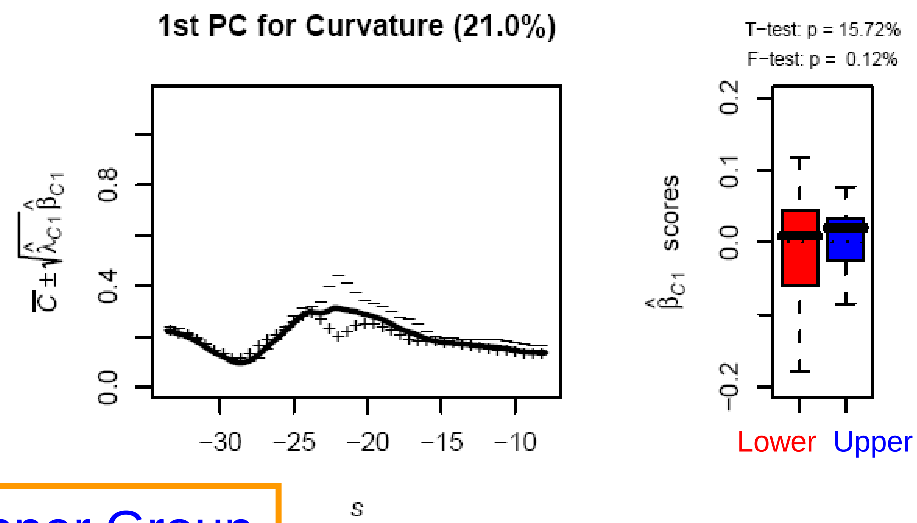
and Supervised Classification, Discrimination

Functional Principal Component Analysis

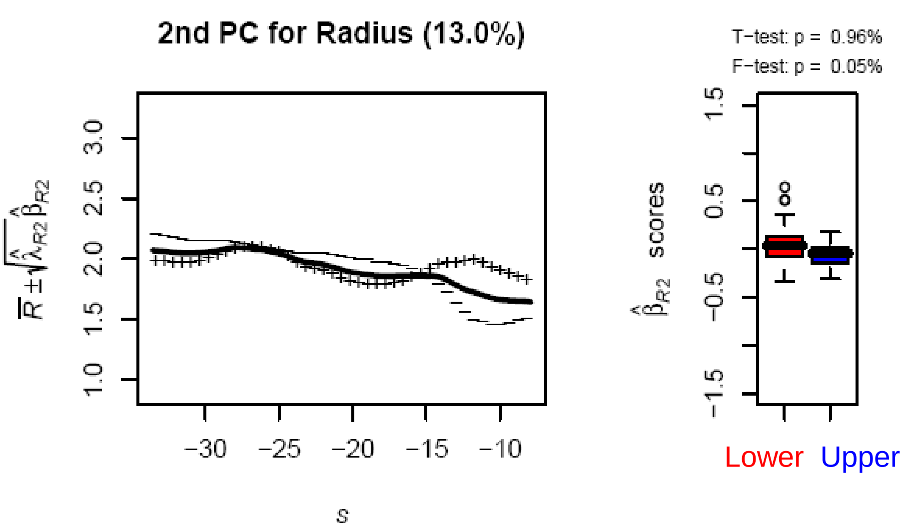


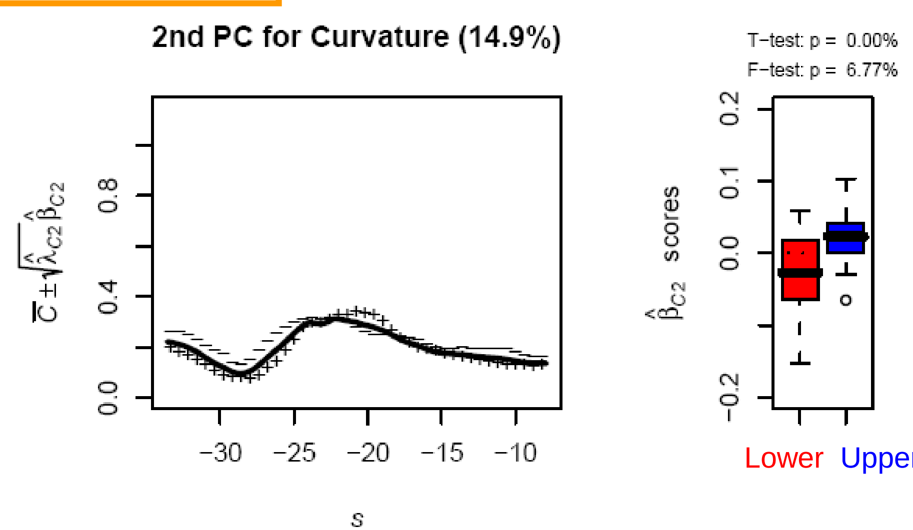
Functional Principal Component Analysis





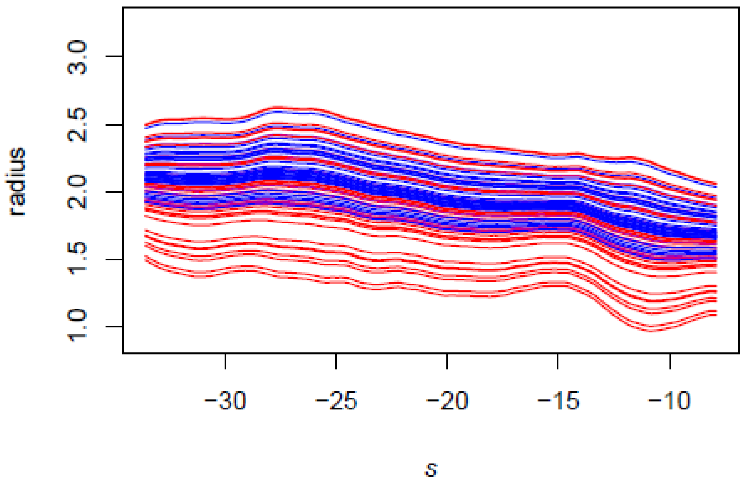
Lower Group - Upper Group



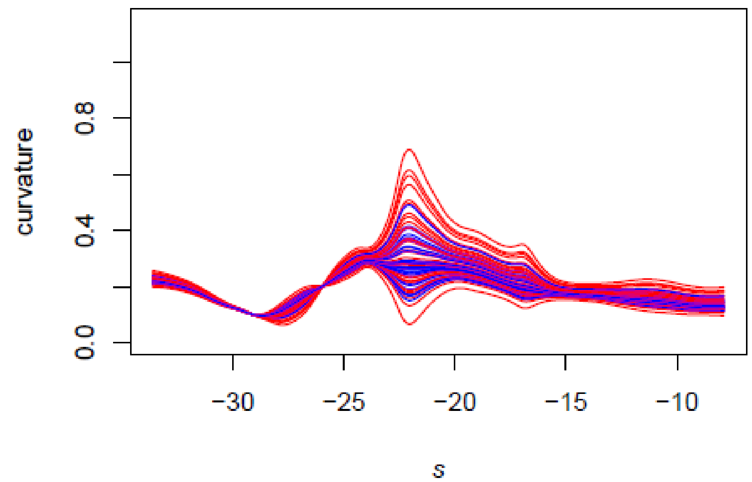


Functional Principal Component Analysis of of Radius and Curvature Functions

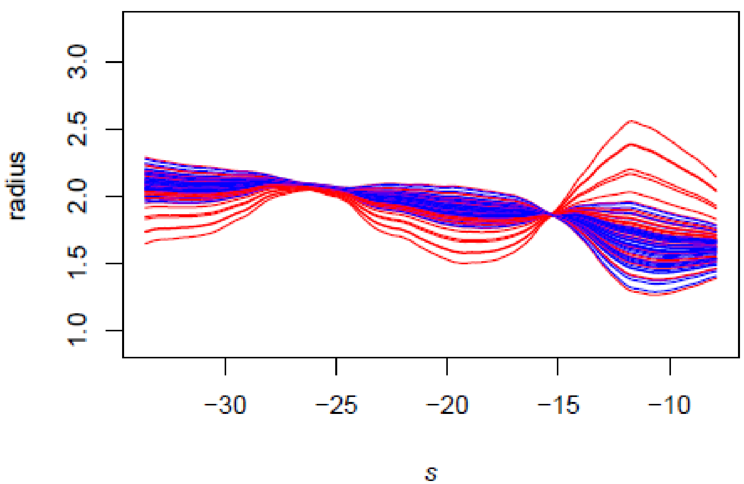
1st PC for Radius (65.9%)



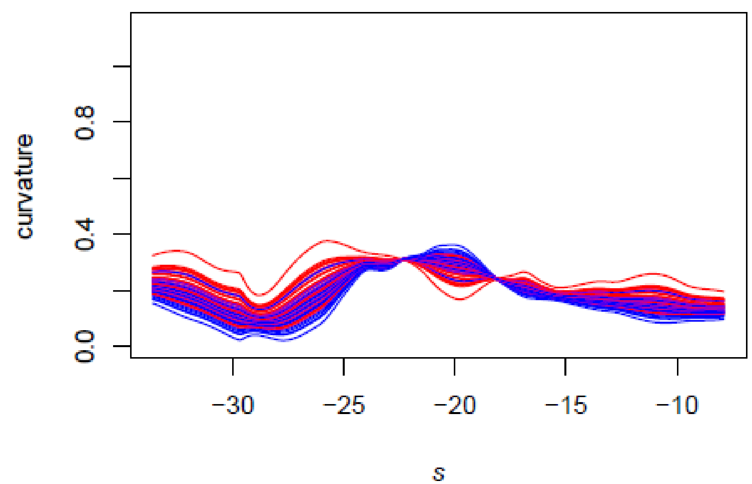
1st PC for Curvature (21.0%)



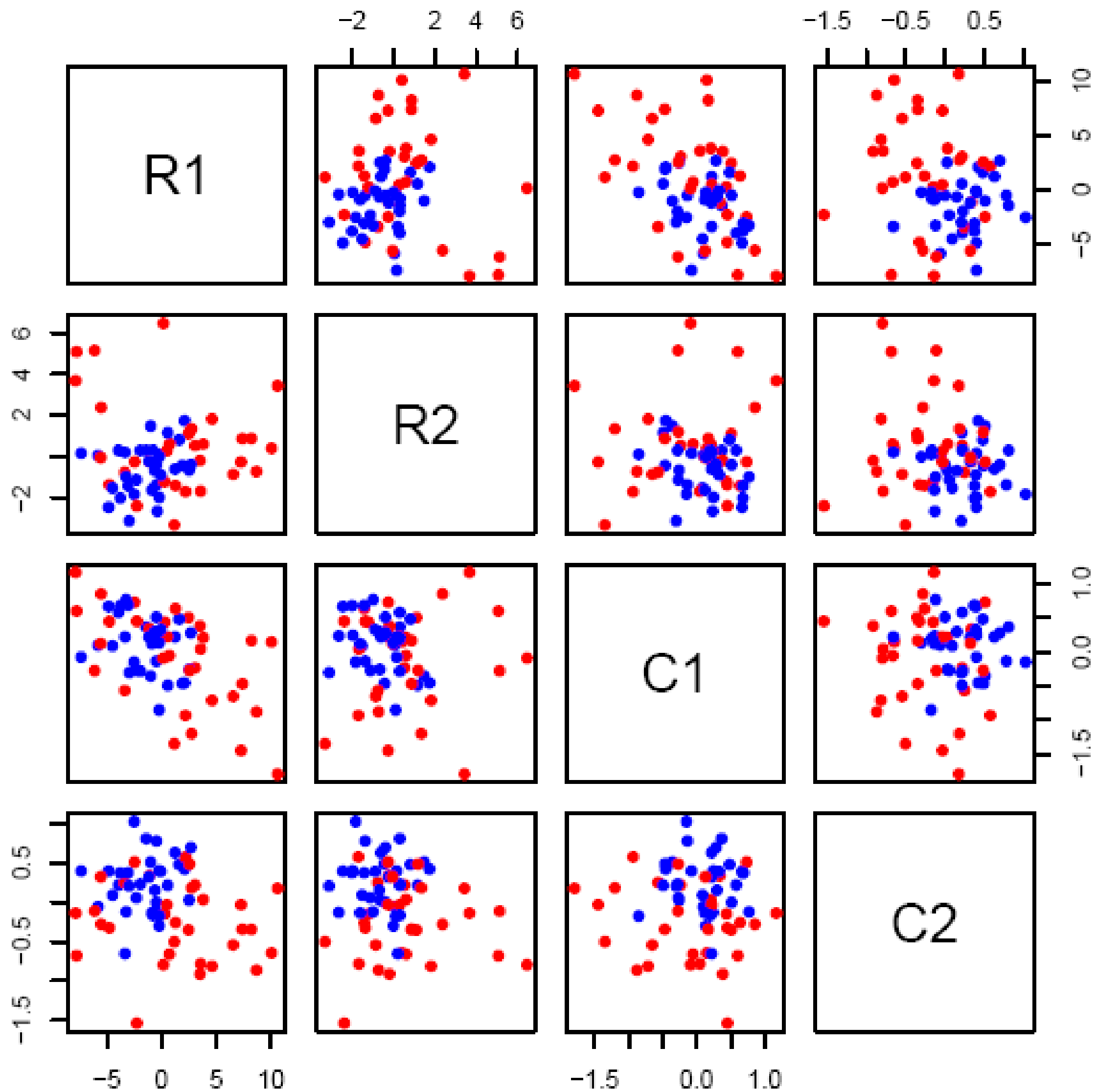
2nd PC for Radius (13.0%)



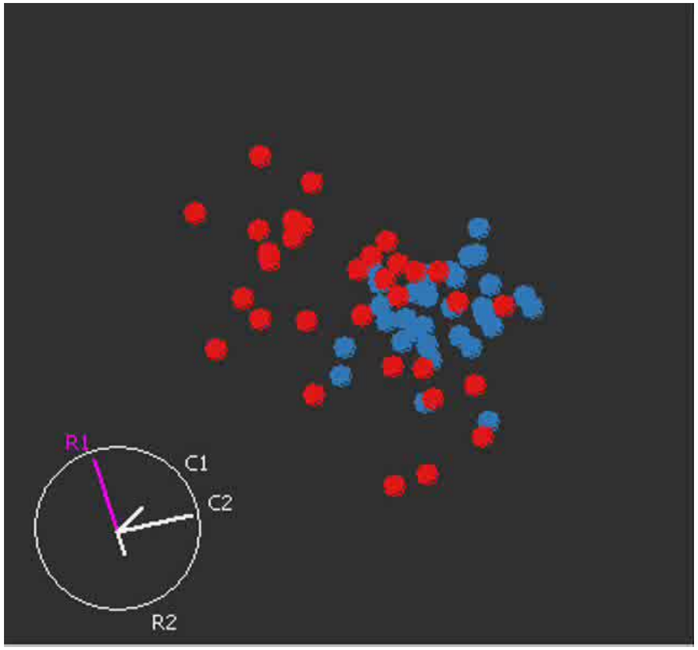
2nd PC for Curvature (14.9%)



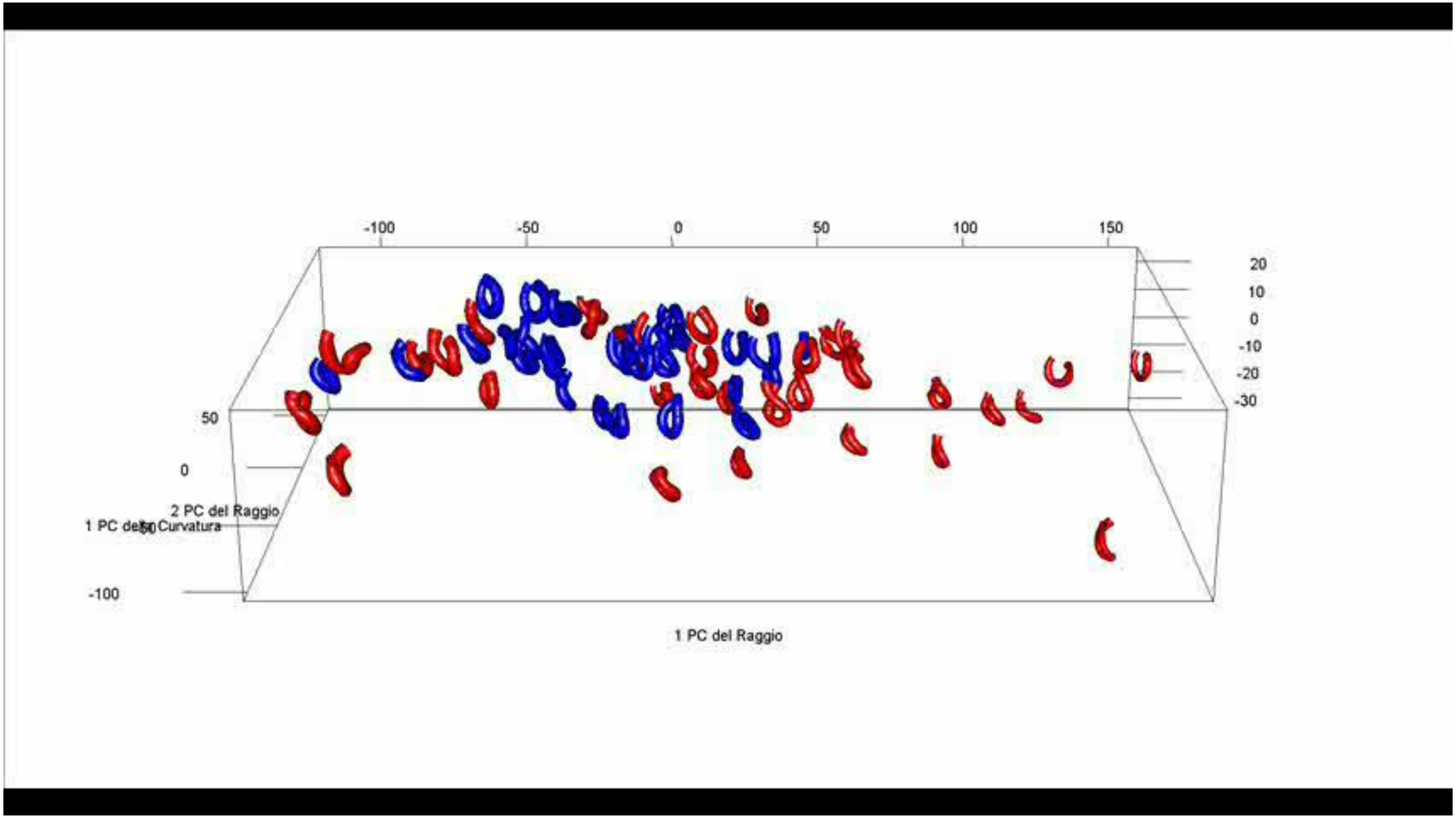
Functional Principal Component Analysis of Radius and Curvature Functions



Functional PCA Scores



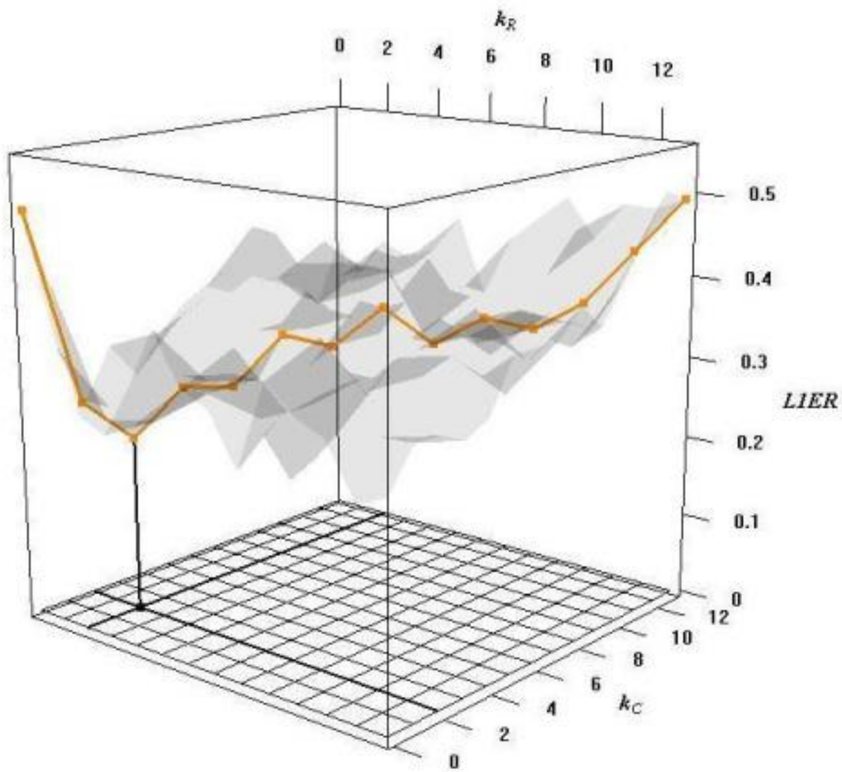
Functional Principal Component Analysis

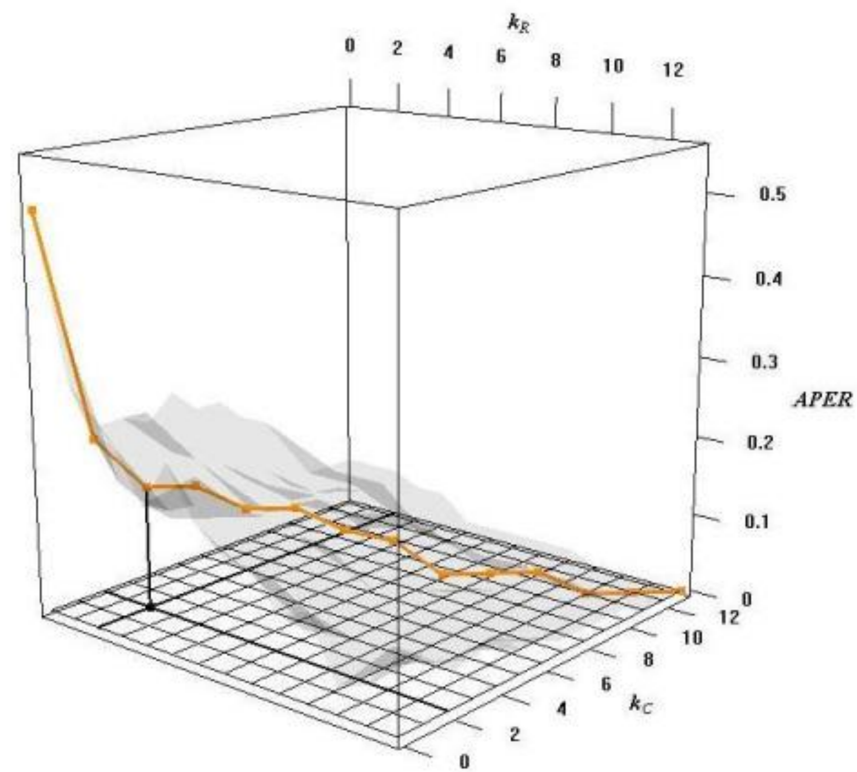


Representation of the 65 ICA in the space generated by:

2 PC of radius and 1 PC of curvature

Quadratic Discriminant Analysis of Functional PCA scores





L1ER = 21.5%

	Lower Predicted	Upper Predicted
Lower	23	9
Upper	5	28

	Lower Predicted	Upper Predicted
Lower	35.4%	13.8%
Upper	7.7%	43.1%

APER = 16.9%

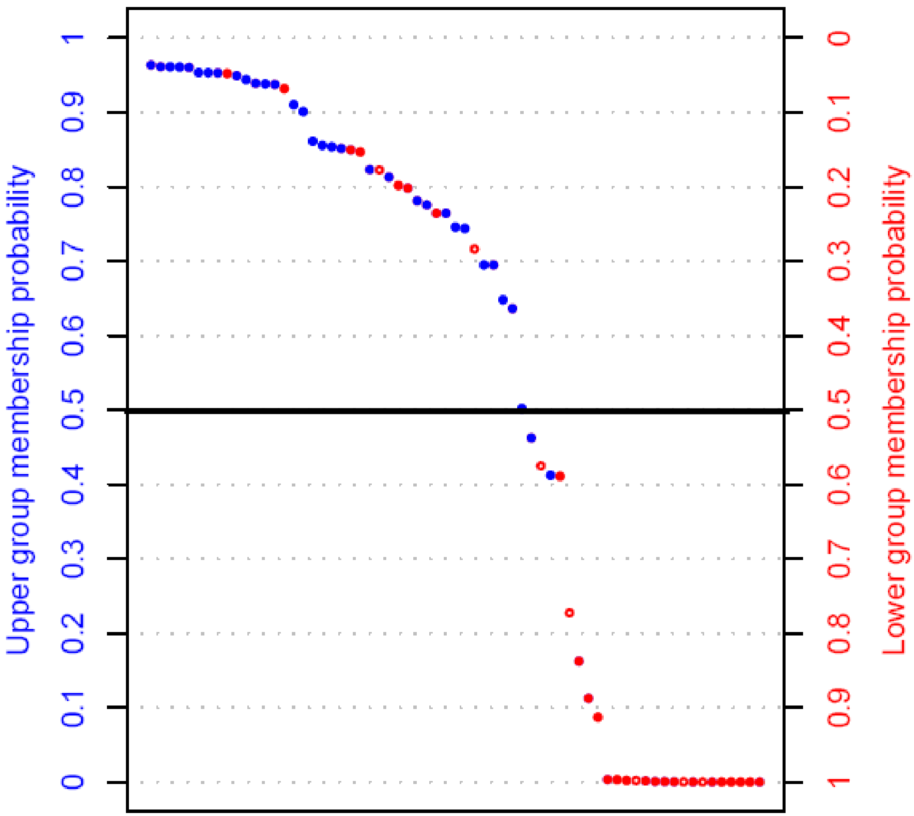
	Lower Predicted	Upper Predicted
Lower	23	9
Upper	2	31

	Lower Predicted	Upper Predicted
Lower	35.4%	13.8%
Upper	3.1%	47.7%

Quadratic Discriminant Analysis of Functional PCA scores

➡ **Upper group** patients are very well characterize by this two geometric features

A quadratic discriminant analysis of scores correctly identifies 31 out of the 33 patients in this group

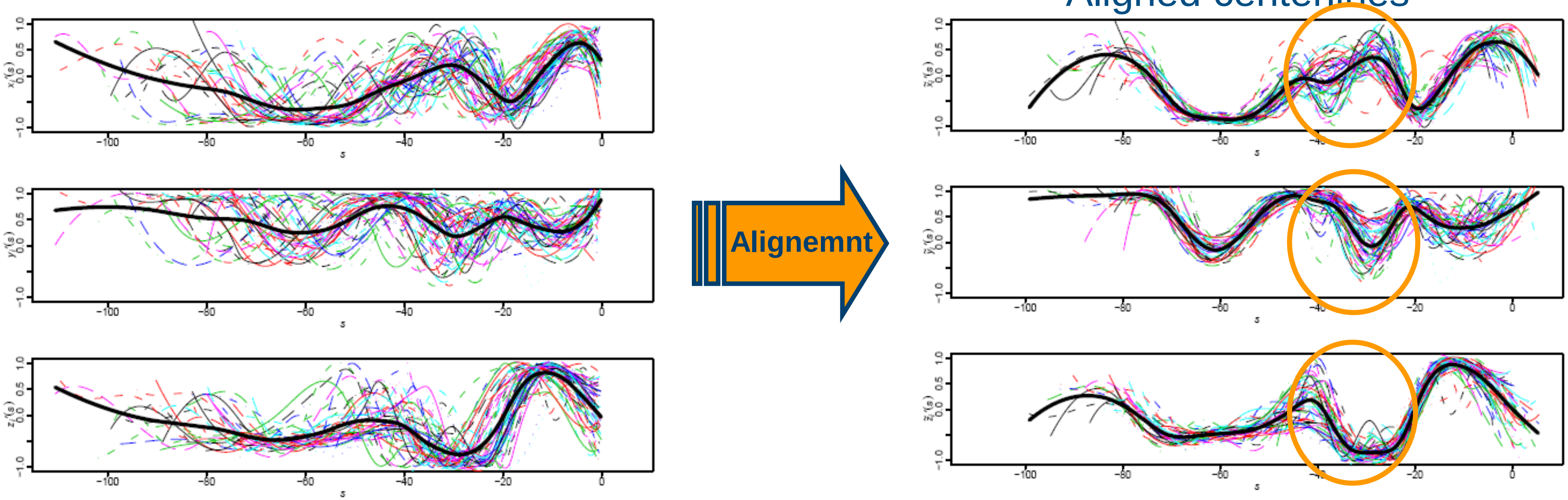


- Large vessels
- Strong tapering
- High within-patient curvature variability
- Lower between-patient variability

Curve clustering and alignment

Sangalli Secchi Vantini Vitelli, 2010, CSDA

Aligned centerlines



GOAL: Identify ICA's with
different morfological shapes

Need to be able to:

jointly align and cluster

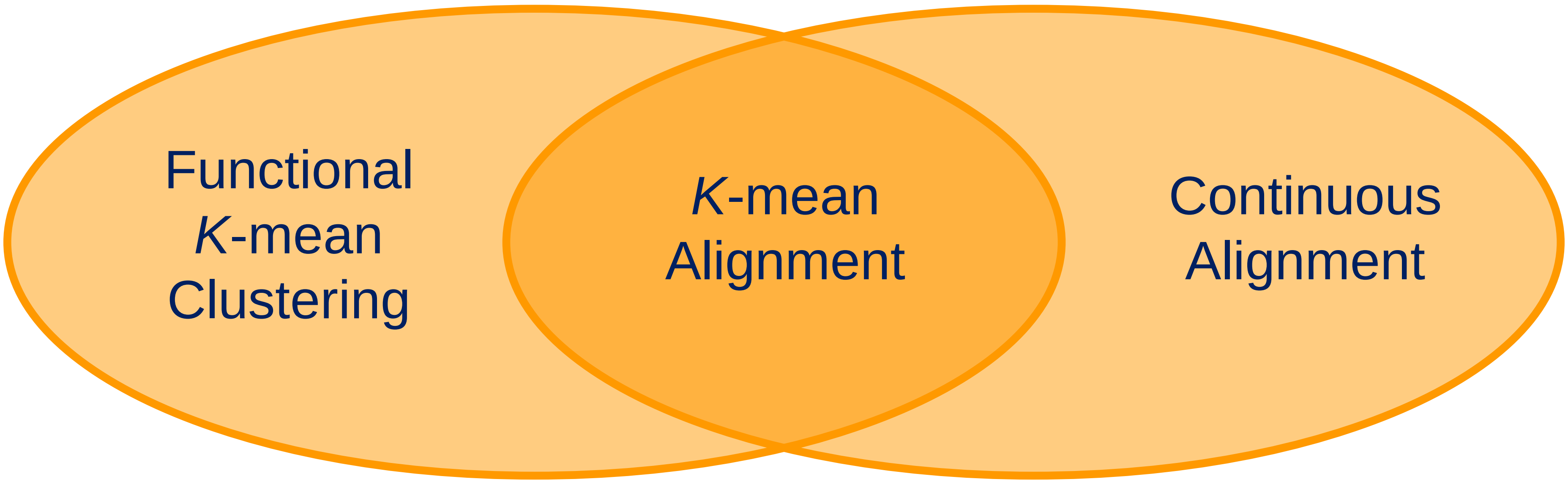
the *N* centerlines

in multiple groups (*k* groups)

having unknwon templates

Usupervised Classification, Clustering

Curve clustering and alignment



→ *K*-mean Clustering
with warping allowed

→ Continuous Alignment
with *K* templates

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

$$\{\mathbf{c}_1, \ldots, \mathbf{c}_N\}$$

with respect to k template curves

$$\underline{\varphi} = \{\varphi_1, \ldots, \varphi_k\}$$

Domain of attraction of φ_j

$$\Delta_j(\underline{\varphi}) = \{\mathbf{c} \in \mathcal{C} : \sup_{h \in W} \rho(\varphi_j, \mathbf{c} \circ h) \geq \sup_{h \in W} \rho(\varphi_r, \mathbf{c} \circ h), \forall\, r \neq j\}, \quad j = 1, \ldots, k$$

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 and alignment

Trivial case: $\underline{\varphi} = \{\varphi_1, \ldots, \varphi_k\}$ known

In order to cluster and align the set of N curves $\{\mathbf{c}_1, \ldots, \mathbf{c}_N\}$ with respect to $\underline{\varphi}$:

for $i = 1, \ldots, N$

- 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: $\underline{\varphi} = \{\varphi_1, \ldots, \varphi_k\}$ unknown

need to be themselves estimated from the data, leading to a complex optimization problem

Given $\{\mathbf{c}_1, \ldots, \mathbf{c}_N\} \subset \mathcal{C}$, find $\underline{\varphi} = \{\varphi_1, \ldots, \varphi_k\} \subset \mathcal{C}$, $\{\lambda_1, \ldots, \lambda_n\} \subset \{1, \ldots, k\}$ and $\underline{\mathbf{h}} = \{h_1, \ldots, h_N\} \subset W$ that maximise $\frac{1}{N} \sum_{i=1}^{N} \rho(\varphi_{\lambda_i}, \mathbf{c}_i \circ h_i)$

An approximate solution to his optimization problem is given by the following iterative procedure

$\underline{\varphi}_{[q-1]} = \{\varphi_{1\,[q-1]}, \ldots, \varphi_{k\,[q-1]}\}$: set of templates after iteration $q-1$

$\{\mathbf{c}_{1\,[q-1]}, \ldots, \mathbf{c}_{N\,[q-1]}\}$: N curves aligned and clustered to $\underline{\varphi}_{[q-1]}$

q-th iteration of the algorithm

Template identification step. For $j = 1, \ldots, k$, the template of the jth cluster $\varphi_{j\,[q]}$ is estimated using all curves assigned to cluster j at iteration $q-1$.

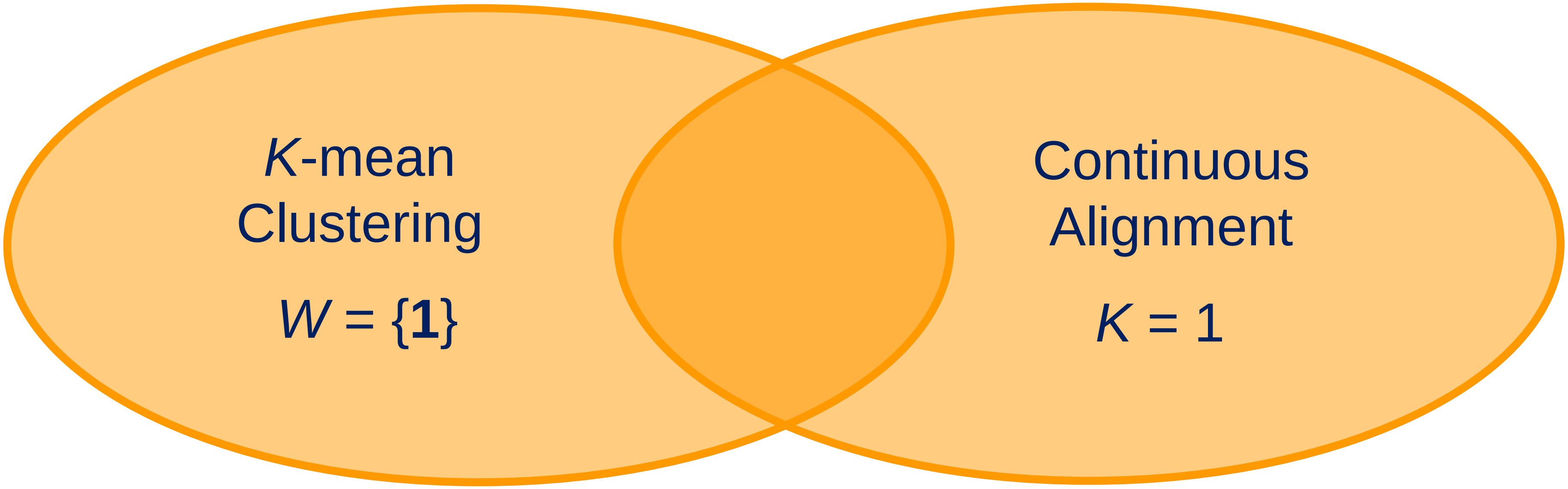
$$\varphi_{j\,[q]} = \arg\max_{\varphi \in \mathcal{C}} \sum_{i:\lambda_i = j} \rho(\varphi, \mathbf{c}_{i\,[q-1]})$$

Assignment and alignment step. The set of curves $\{\mathbf{c}_{1\,[q-1]}, \ldots, \mathbf{c}_{N\,[q-1]}\}$ is clustered and aligned to the set of templates $\underline{\varphi}_{[q]} = \{\varphi_{1\,[q]}, \ldots, \varphi_{k\,[q]}\}$.

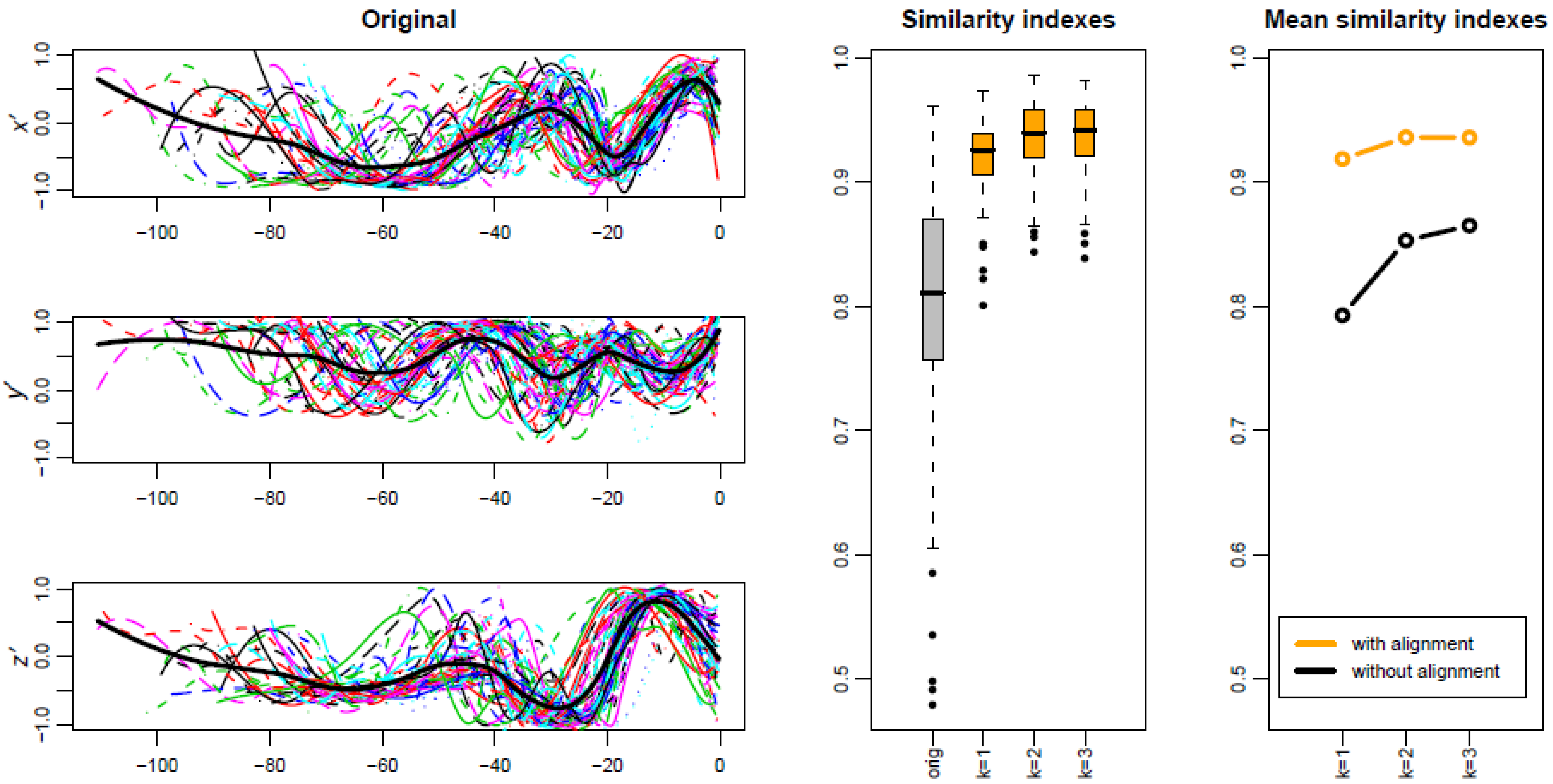
Normalization step. For $j = 1, \ldots, k$, all curves assigned to cluster j are warped along a common warping function, so that the average warping undergone by curves assigned to the same cluster is the identity transformation (thus avoiding the drifting apart of clusters or the global drifting of 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.

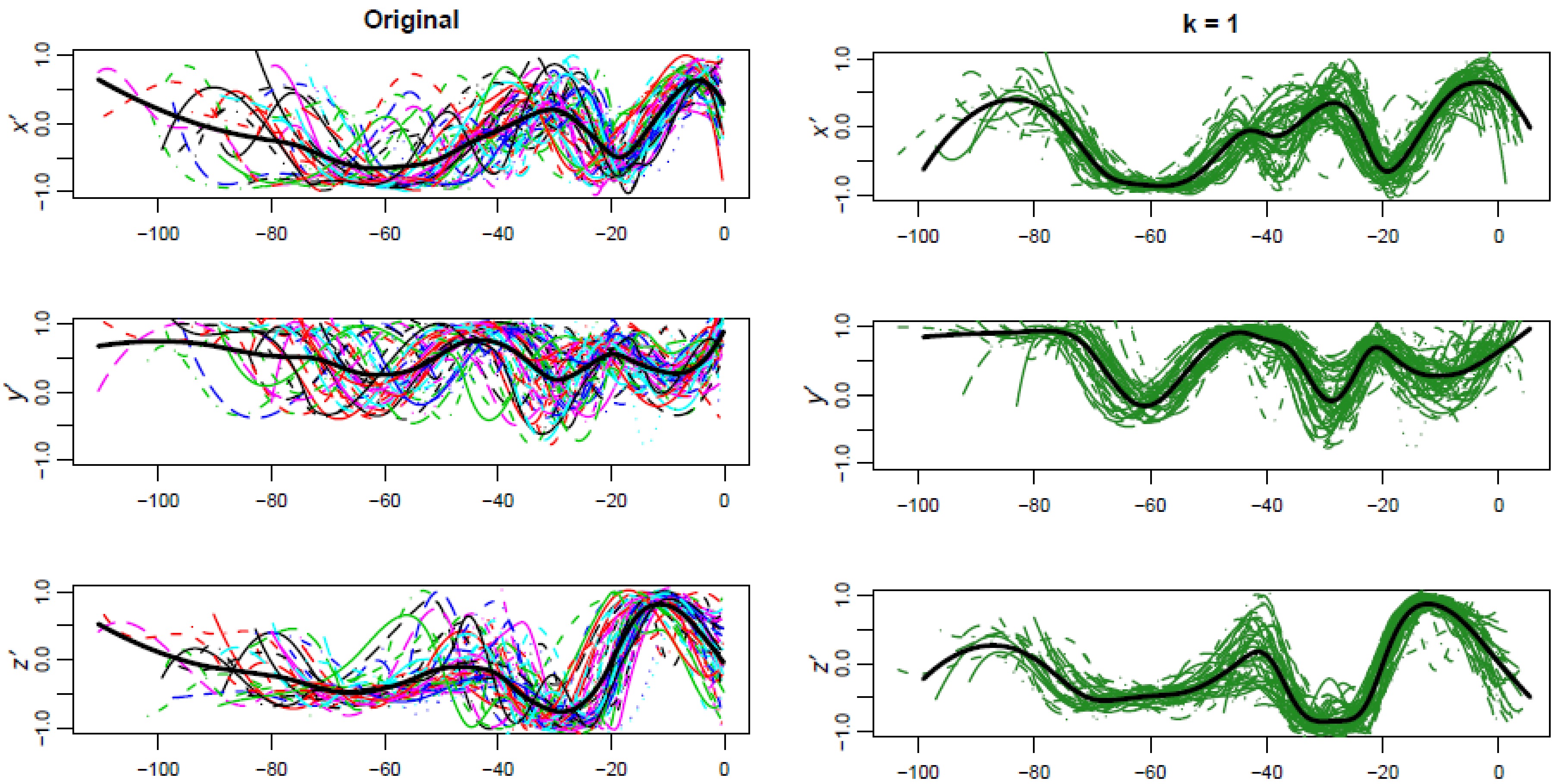
K-mean Alignment:



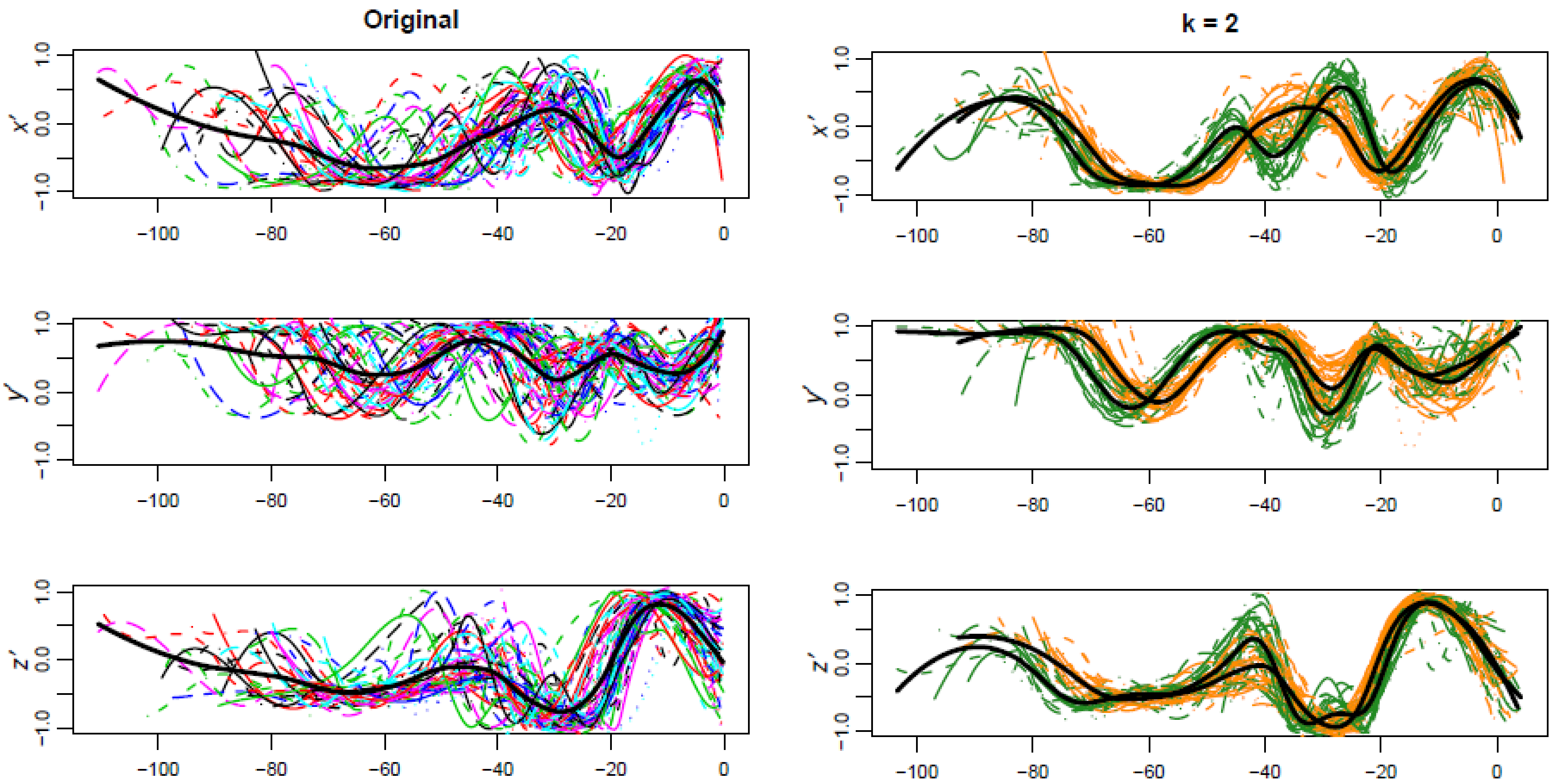
k-mean alignment



One-mean alignment



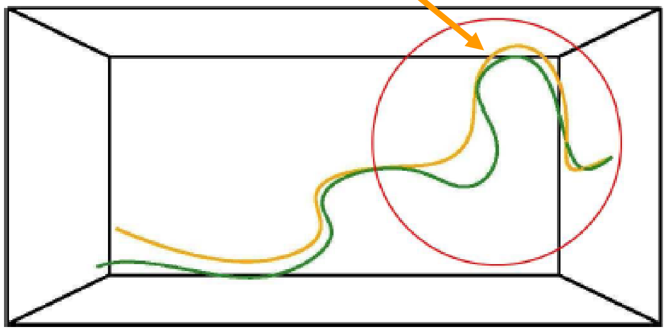
Two-mean alignment

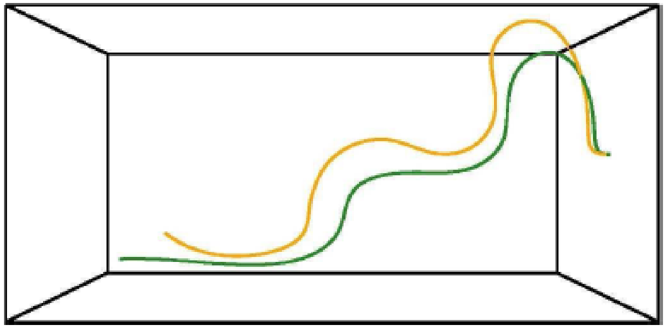


Curve clustering and alignment

Ω shaped ICA's

(1-bend siphon in distal part)





S shaped ICA's

(2-bend siphon in distal part)

The procedure identify two prototype shapes of ICA's that are described in the medical literature

Krayenbuehl et. Al. (1982)

NO interesting insights

Simple clustering without alignment is driven by phase variability and fails to identify different morphological shapes

Curve clustering and alignment

	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%

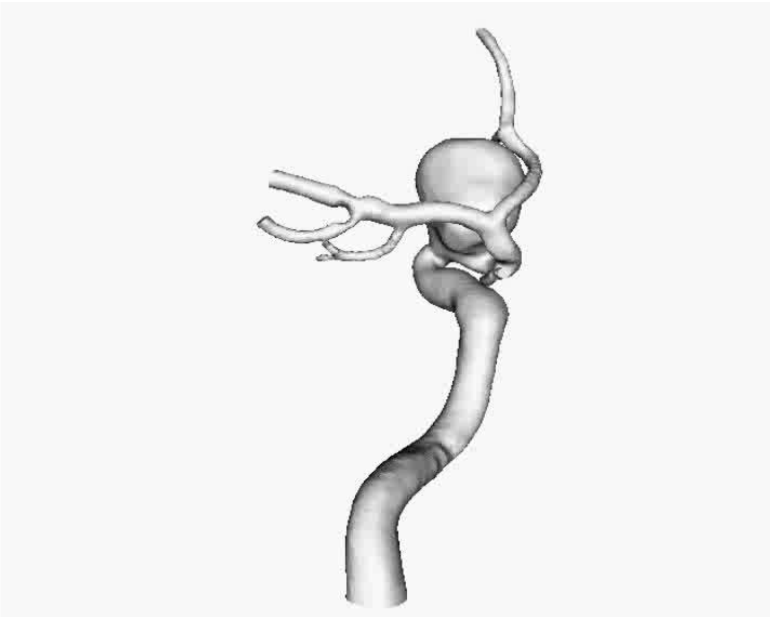
► The ICA syphon acts as a fluid dynamical shock-absorber to steady blood flow in the brain

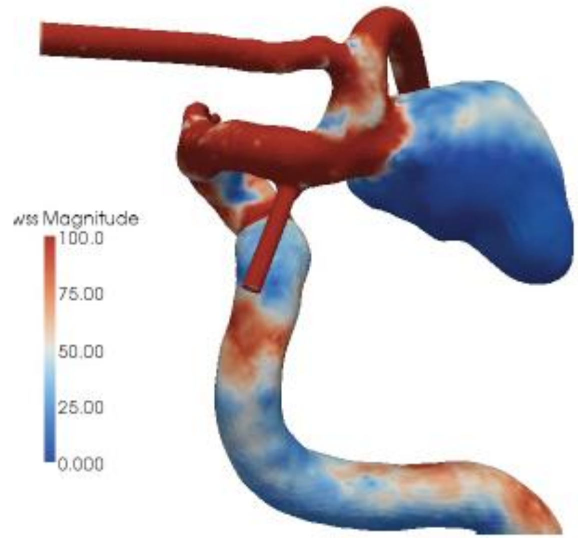
⟶ **S** shaped ICA's seems to be more effective in making the blood-flow steadier with respect to **Ω** shaped ICA's

Blood Flow Numerical Simulations

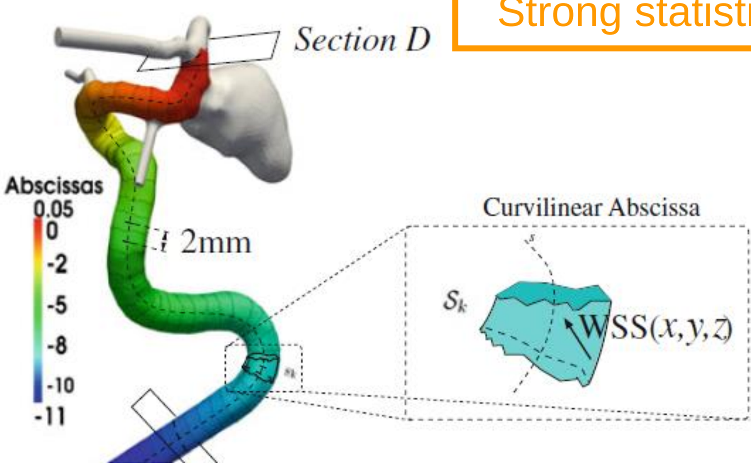
Passerini et al. CVET 2012

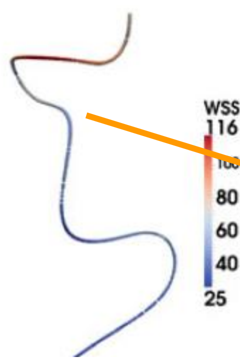
Hemodynamic data obtained by Computational Fluid Dynamics



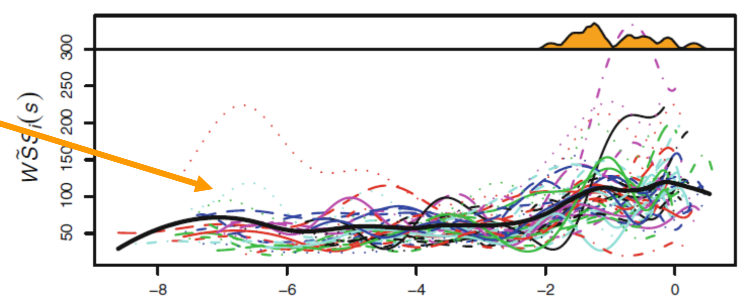


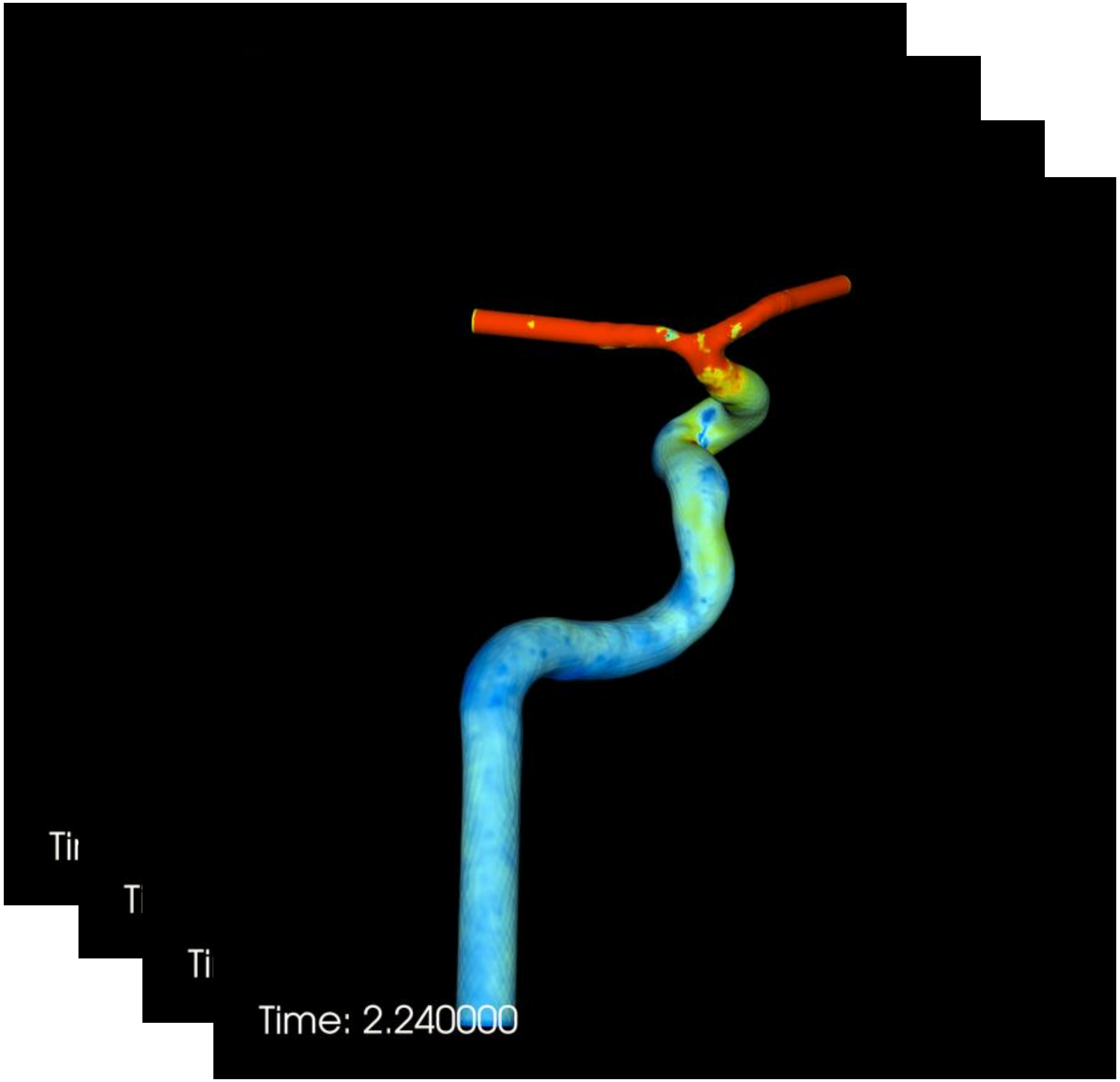
Strong statistical evidence in favor of the conjecture





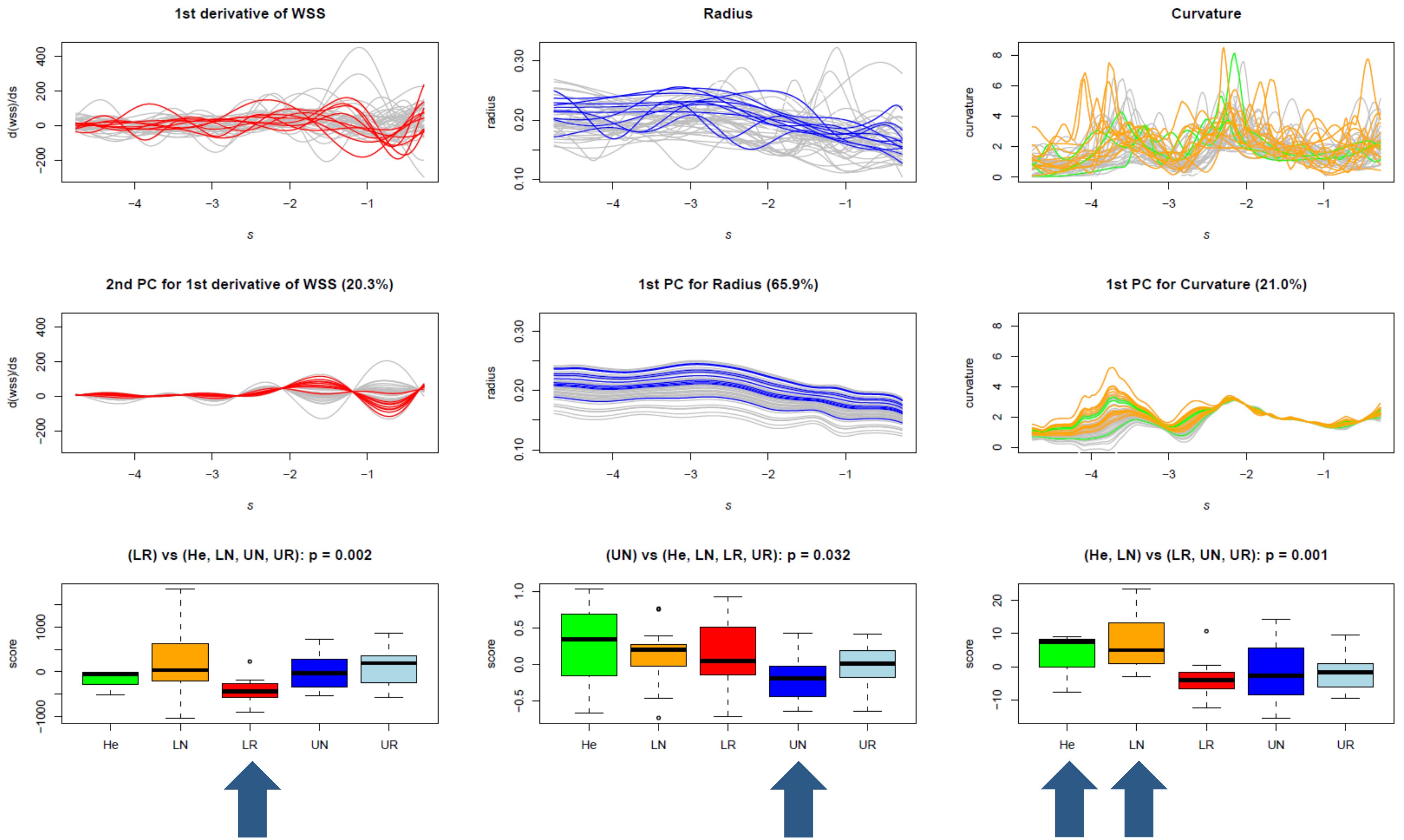
WSS functions





- He: Healthy
- LN: Lower Non-ruptured
- LR: Lower Ruptured
- UN: Upper Non-ruptured
- UR: Upper Ruptured

Joint statistical analysis of geometrical and haemodynamical features



Wall Shear Stress peak

Wide Carotid Artery

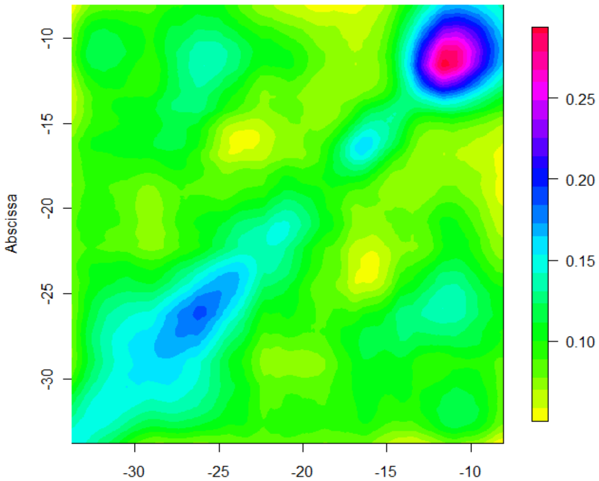
Two siphons

Comparison of covariance structures

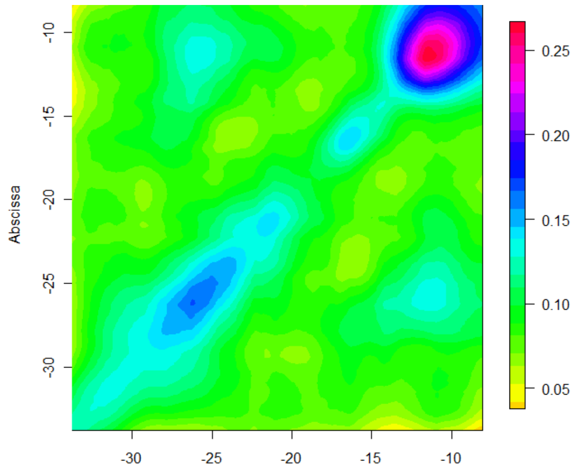
D. Pigoli, J.A.D. Aston, I. Dryden, P. Secchi, 2012

Radius

Aneurysm before bifurc.



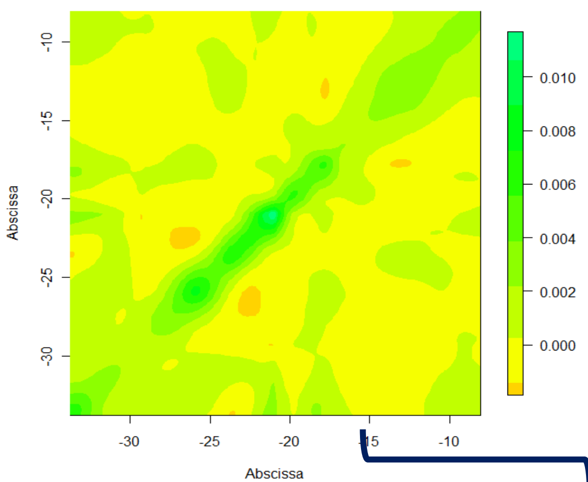
No Aneurysm



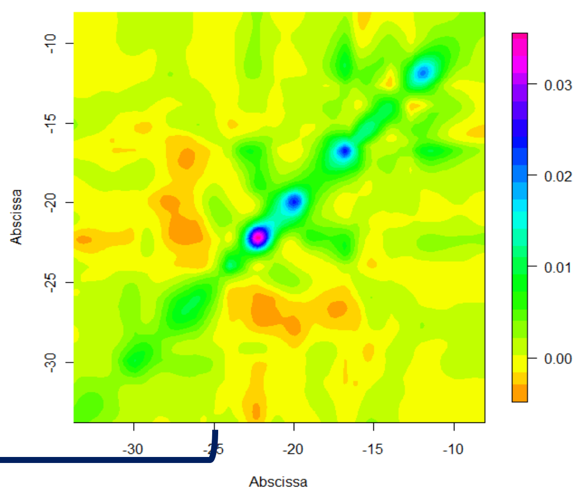
p-value: 0.855

Curvature

Aneurysm before bifurc.



No Aneurysm



Lower Group

p-value: 0.61

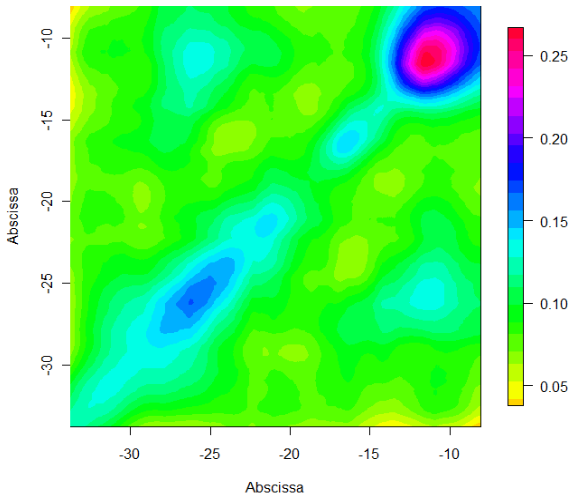
Permutation test based on Square Root distance between covariance operators

Comparison of covariance structures

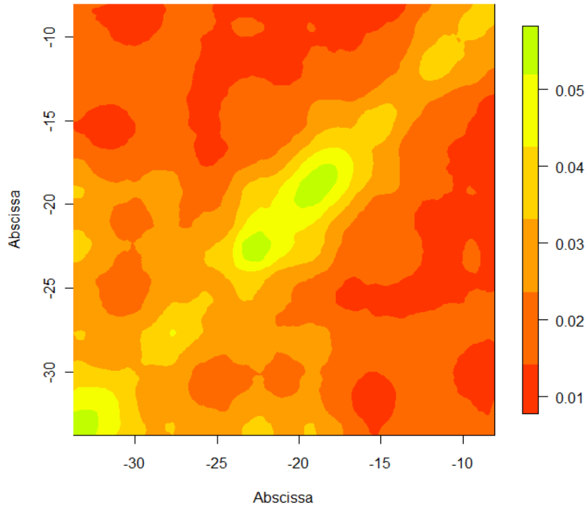
D. Pigoli, J.A.D. Aston, I. Dryden, P. Secchi, 2012

Radius

Lower Group



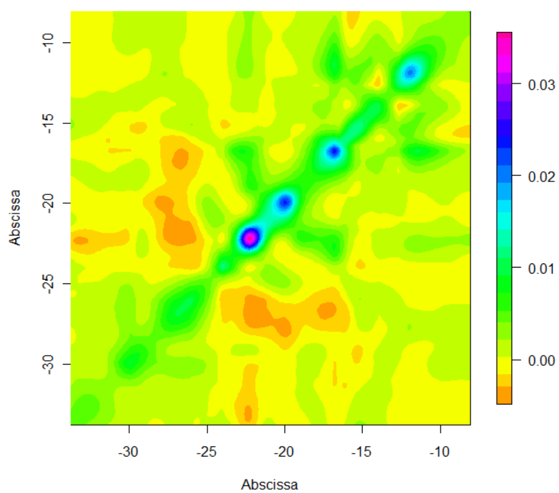
Upper Group



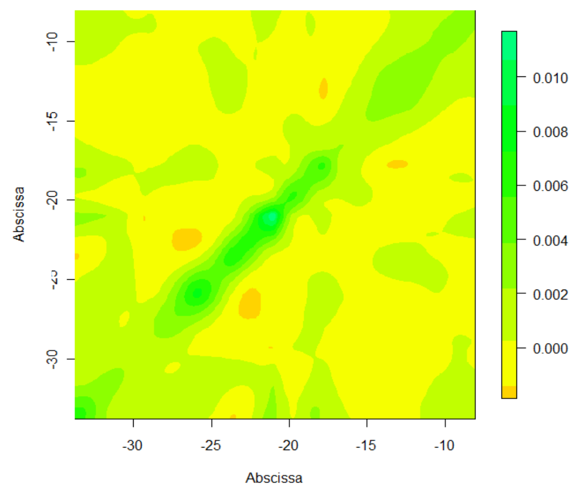
p-value <0.001

Curvature

Lower Group



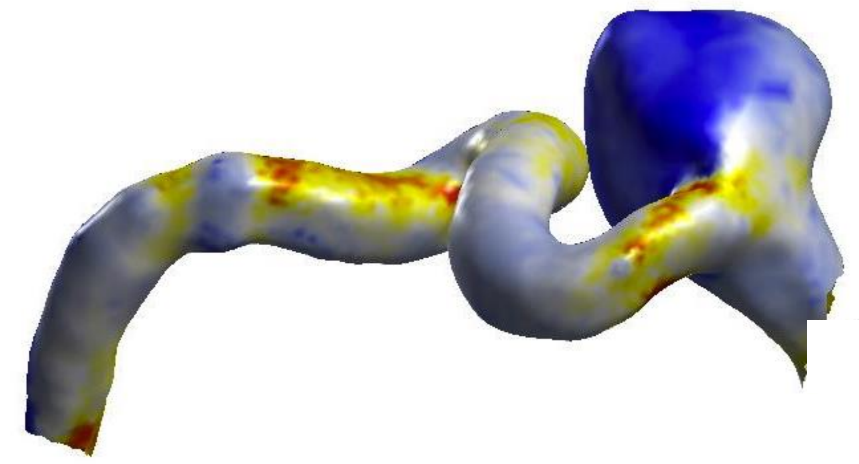
Upper Group

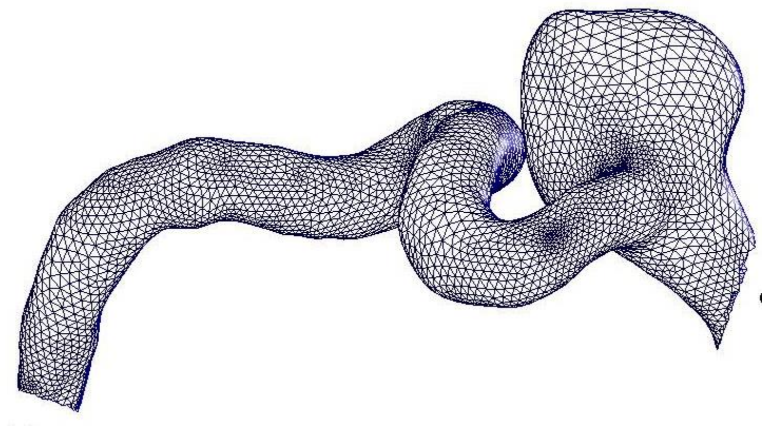


p-value:0.005

Permutation test based on Square Root distance between covariance operators

Spatial regression models over bi-dimensional Riemannian manifolds





on Friday

Main references for the Aneurisk project

http://mox.polimi.it/it/progetti/aneurisk/

- Sangalli, L.M., Secchi, P., Vantini, S. (2013), AneuRisk65: three-dimensional cerebral vascular geometries. TechRep. MOX 44/2013, Dipartimento di Matematica, Politecnico di Milano.

- Sangalli, L.M., Secchi, P., Vantini, S. (2013), Analysis of AneuRisk65 data: K-mean Alignment. TechRep. MOX 45/2013, Dipartimento di Matematica, Politecnico di Milano.

- Antiga, L., Ene-Jordache, B. and Remuzzi, A. (2003), " Computational Geometry for Patient-Specific Reconstruction and Meshing of Blood Vessels from MR and CT angiographies",
IEEE Transactions on Medical Imaging, 22, 674-684.

- 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.

- Piccinelli, M., Veneziani, A., Steinman, D., Remuzzi, A., and Antiga, L. (2009), "A framework for geometric analysis of 852 vascular structures: applications to cerebral aneurysms,"
IEEE Trans. Med. Imaging, 28, 1141–1155.

- 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.

- Passerini, T., Sangalli, L.M., Vantini, S., Piccinelli, M., Bacigaluppi, S., Antiga, L., Boccardi, E., Secchi, P.,Veneziani, A. (2012), "An Integrated CFD-Statistical Investigation of Parent Vasculature of Cerebral Aneurysms", *Cardiovascular Engineering and Technology,* 3, pp. 26-40.

- Pigoli, D., Sangalli, L.M. (2012), “Wavelets in Functional Data Analysis: estimation of multidimensional curves and their derivatives”, *Computational Statistics and Data Analysis*, 56, 1482–1498.

- Vantini, S. (2012) “On the Definition of Phase and Amplitude Variability in Functional Data Analysis”, *Test*, 21, 676-696.

- Helle Sørensen, Jeff Goldsmith, Laura M. Sangalli (2013), “An introduction with medical applications to functional data analysis”. Statistics in Medicine, 32, pp. 5222–5240.

- Pigoli, D., Aston, J.A.D., Dryden, I. and Secchi, P. (2012), “Distance and Inference for Covariance Functions”, Tech. rep. MOX 35/2012, Dipartimento di Matematica, Politecnico di Milano.

- Ettinger, B., Perotto, S., Sangalli, L.M. (2012), “Spatial regression models over two-dimensional manifolds”, Tech. rep. MOX 54/2012, Dipartimento di Matematica, Politecnico di Milano.

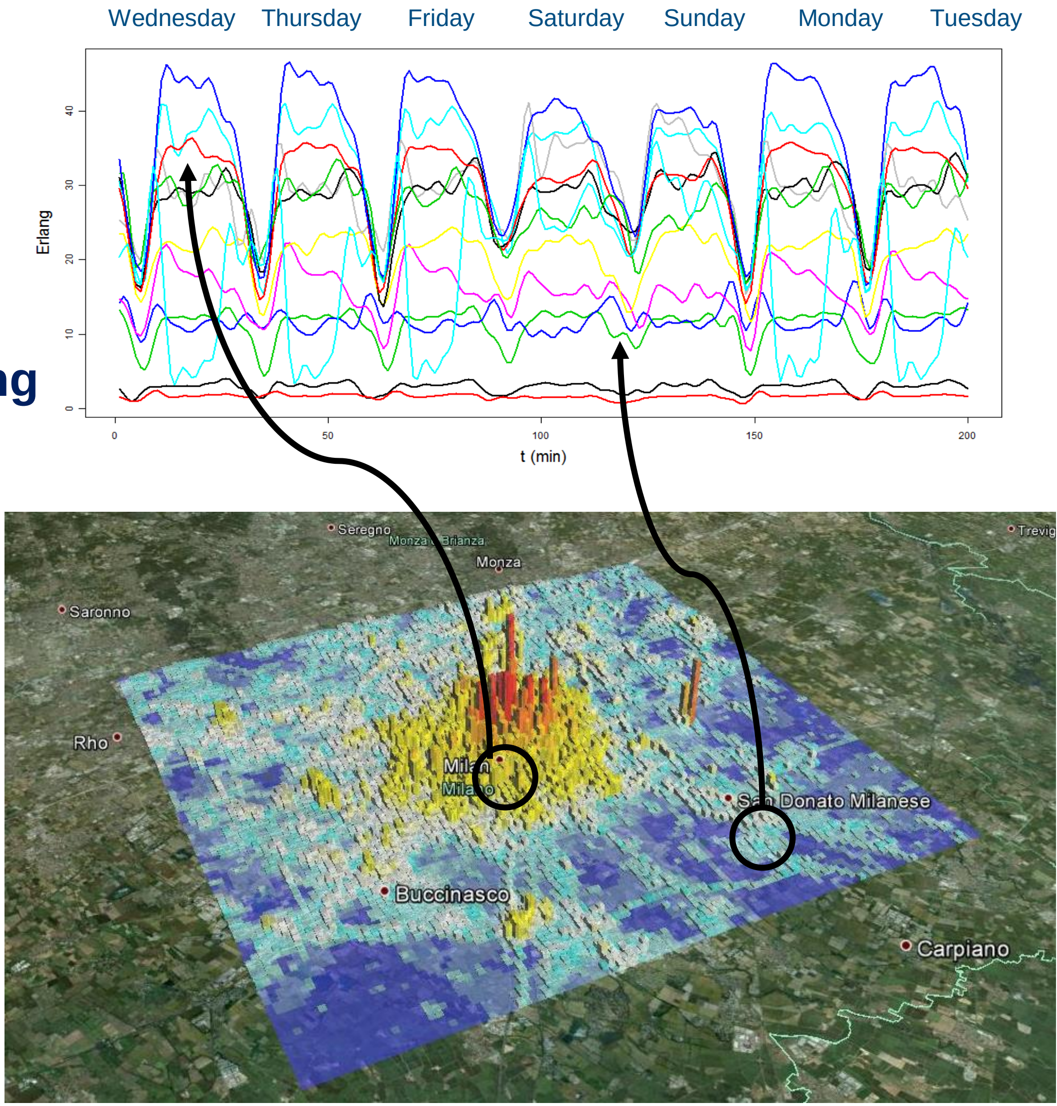
Web page: http://mox.polimi.it/it/progetti/aneurisk/
Web repository: http://ecm2.mathcs.emory.edu/aneurisk

Software: R package for alligning and clustering functional data: fdakma, available from CRAN

Other ongoing projects at MOX

green move

Design and implement a **vehicle-sharing system** in Milan, with electric vehicles



6 Departments of Politecnico di Milano

Fondazione Politecnico di Milano

RegioneLombardia

Data: measures along time of the use of the Telecom mobile phone network across a lattice covering the area of Milan (Italy)

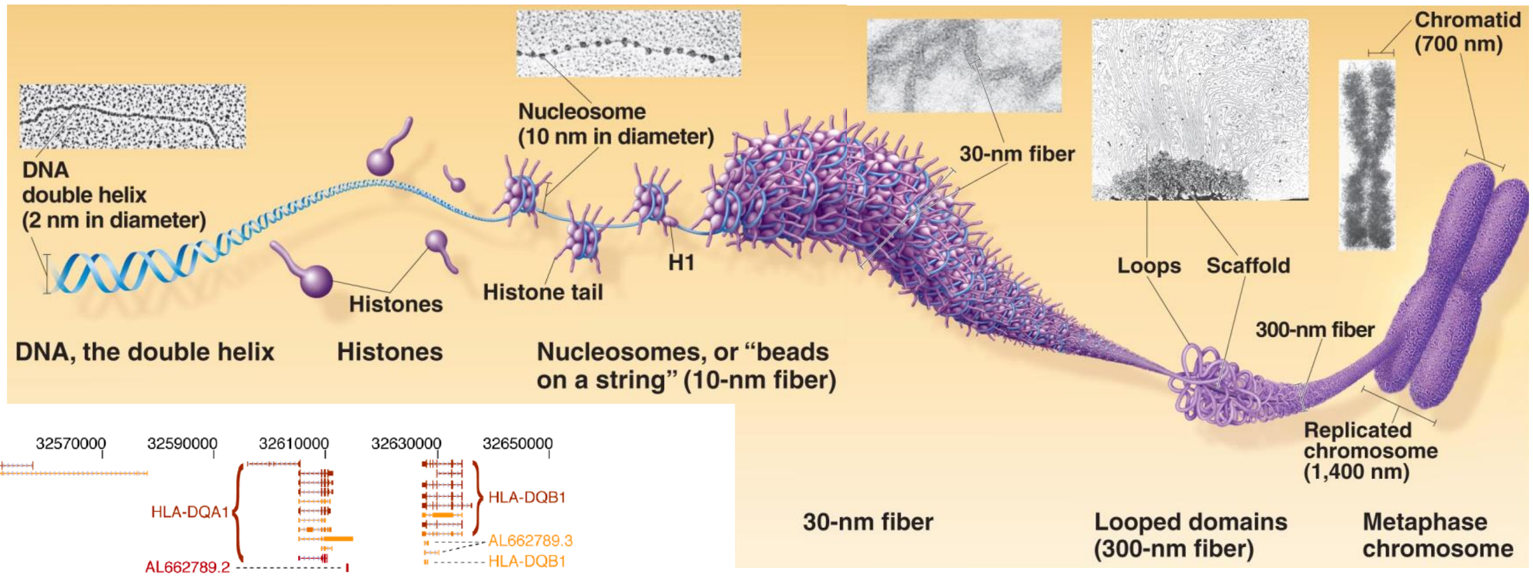
At MOXStat: P. Secchi, S. Vantini, V. Vitelli, P. Zanini

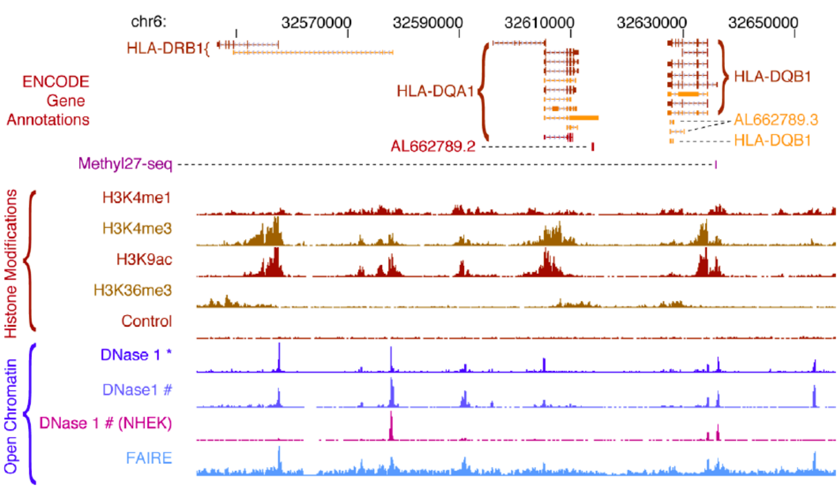
Other ongoing projects at MOX

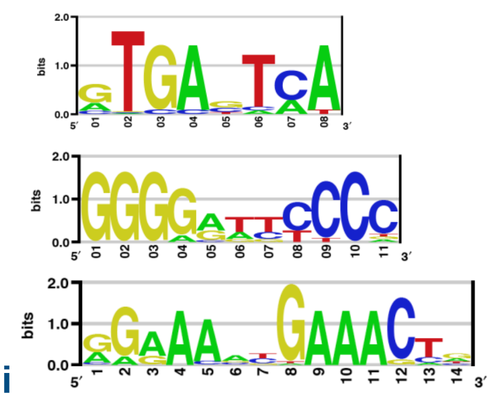
POLITECNICO DI MILANO

IEO
European Institute of Oncology

Epigenoma project







At MOXStat: M. Cremona, A. Parodi, L. Sangalli, P. Secchi , S. Vantini

Other ongoing projects at MOX

The *Project on Chronic Heart Disease* of Regione Lombardia

Utilization of Regional Health Service Databases for evaluating Epidemiology, short- and medium-term outcomes and process indexes for patients hospitalized for chronic heart failure.

Scientific Director: Dr Maria Frigerio (A.O. Niguarda Ca' Granda, Milano)

"The main objective of this project is to give a global picture of data in epidemiological and clinical treatment of Chronic Heart Diseases."

Ministero della Salute

RegioneLombardia

The main goals that the ***Project on Chronic Heart Failures (CHF)*** aims to achieve are:

1. Modelling the combined endpoint of death and hospitalizations process of patients affected by HF
2. Accounting for multiple events, providing a more detailed information on the disease-control process, and a more precise understanding of the prognosis of patients.
3. Identifying specific groups of patients/hospital characterized by different evolution pattern

Statistics role in this context is aimed to

- Summarising information arising from highly complex database
- Exploiting information carried out by administrative source of data (costless & real-time updated)

At MOXStat: A. Paganoni, F. Ieva, N. Tarabelloni

Statistics @ MOX !

- Anna Maria Paganoni
- Piercesare Secchi
- Laura Sangalli
- Simone Vantini
- Andrea Ghiglietti
- Paolo Zanini
- Alessandra Menafoglio
- Alessia Pini
- Marzia Cremona
- Mara Bernardi
- Alice Parodi
- Nicholas Tarabelloni