Análisis de Datos Funcionales
Introducción y Aplicaciones

Ana M. Aguilera

Departamento de Estadística e I.O. Universidad de Granada
Instituto de Matemáticas Universidad de Granada (IMAG)

ugr Universidad de Granada

Ciclo de Conferencias RedIUM
IMAG, 19 marzo, 2021

What is functional data (FD)?

© A.M. Aguilera
aaguiler@ugr.es

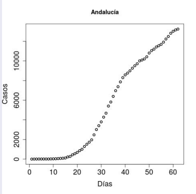
Data that are in the form of functions of a continuous argument

1. Curves
   - Data recorded continuously in time or other continuous argument (meteorology, biosciences, economics, chemometrics, electronic, ...)
2. Surfaces or hypersurfaces
   - Functions of two or more variables as spatio-temporal data depending on time, latitude and longitude (environment, medicine, ...)
3. Multivariate functional data
   - Vector of curves (biomechanics, environment, ...)

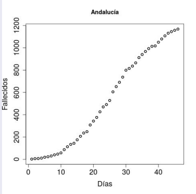
COVID-19 data

© A.M. Aguilera
aaguiler@ugr.es

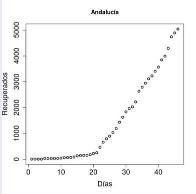
Cases of COVID-19 in Andalucía 20/02/2020 to 27/04/2020 (first wave)



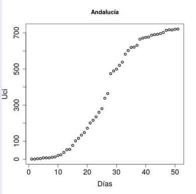
Casos



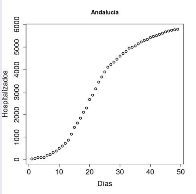
Fallecidos



Recuperados



Uci



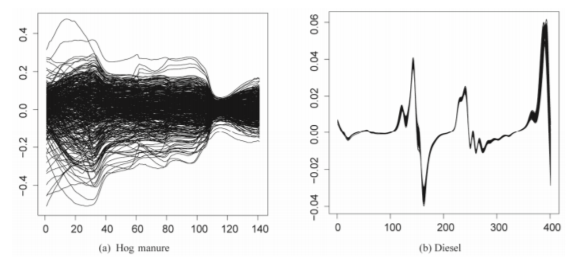
Hospitalizados

Chemometrics

© A.M. Aguilera
aaguiler@ugr.es

Hog manure and diesel

Absorbance spectra for samples of hog manure and diesel

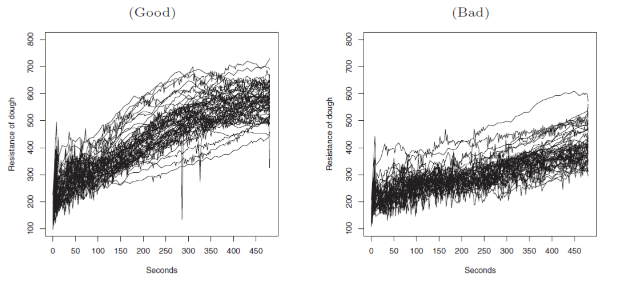


Chemometrics

© A.M. Aguilera
aaguiler@ugr.es

Kneading data

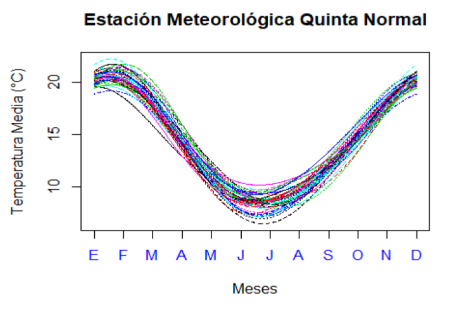
Curves of resistance of dough for 115 different flours registered every two seconds



Meteorological data

Chilean weather data

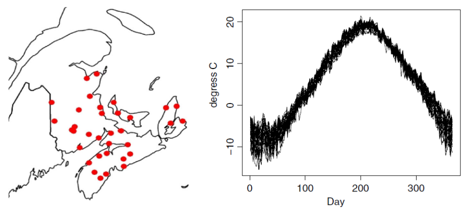
Monthly temperature curves for 46 years (1971-2017) observed at 15 Chilean weather station Quinta-Normal



Meteorological data

Canadian weather data

Averages (over 30 years) of daily temperature curves observed at 35 Canadian Maritime weather stations

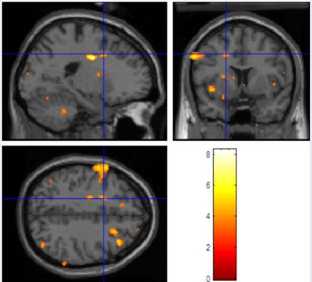


Brain imaging

© A.M. Aguilera
aaguiler@ugr.es

FMRI (functional magnetic resonance imaging)

- A 3-dimensional brain image has thousands of thousands of cubic volumes (voxels) and records the neural activity with high spatial resolution
- The activity level (signal amplitude) is represented by different colors (dark red areas: the most active; light yellow areas: less active; gray areas: very low activity)
- The signal amplitude is recorded over time at each voxel



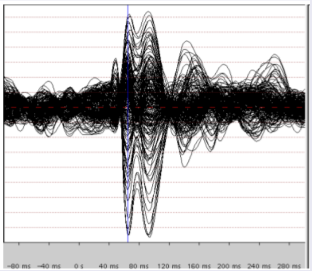
© A.M. Aguilera
aaguiler@ugr.es

Brain imaging

MEG (magneto-encephalography)

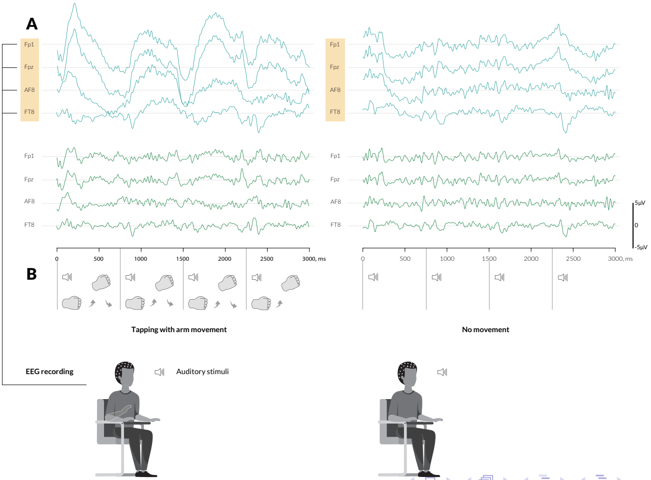
MEG can track the magnetic field accompanying synchronous neuronal activities with temporal resolution of millisecond level

Measures are taken over time by sensors outside the head of a patient



EEG (electroencephalography)

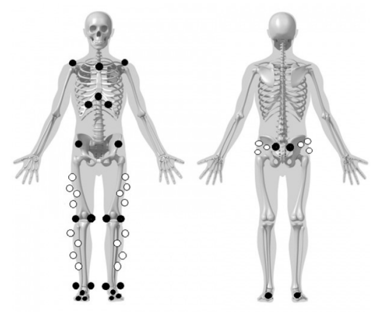
EEG data recorded across different conditions from a single subject with a 64-channels eego mylab system (Art Science and Interaction Lab (ASIL) of Ghent University, Belgium)



Biomechanics

Gait analysis: experimental study developed in the Sport and Health Institute of the University of Granada (IMUDS)

Motion is characterized by the angular position of the different parts of the body along the gait cycle (3D motion capture system (Qualisys AB, Goteborg, Sweden))

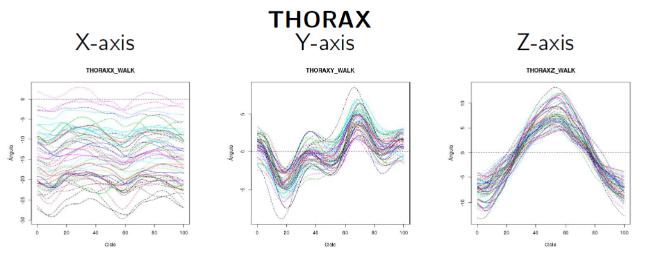


© A.M. Aguilera
aaguiler@ugr.es

Gait analysis

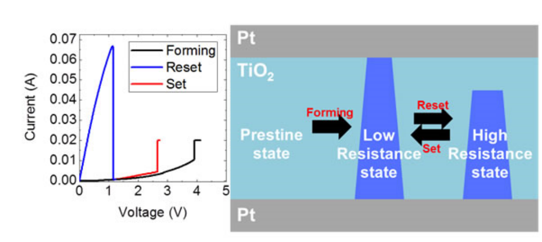
Analyzing kinematics of children using trolleys or backpacks in their daily transportation to school with different loads

The flexion/extension (X-axis), adduction/abduction (Y-axis) and internal/external (Z-axis) rotation of thorax, pelvis, hip, knee and ankle were recorded walking over a platform (101 measurements)



© A.M. Aguilera
aaguiler@ugr.es

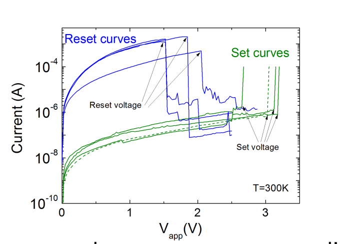
Optimum design of resistive random access memories (RRAMs)
Department of Electronics and Computer Technology (UGR)
Institute of Microelectronics of Barcelona (IMB-CNM, CSIC)



- Two metal plates acting as electrodes with a dielectric in between
- The conductive filaments are formed (set) and destroyed (reset) within the resistive switching device operation
- Sample of current-voltage curves corresponding to the reset-set cycles

Electronics: experimental data

Variability in cycle-to-cycle change in the I-V curves



- Experimental current-voltage curves corresponding to several set-reset cycles. Although the voltage is negative, absolute values are considered for simplicity
- The reset points are determined by the sudden drop of the current (rupture of the conductive filament)
- The set points are characterized by the creation of the conductive filament

Characteristics of Functional Data

© A.M. Aguilera
aaguiler@ugr.es

- Represent the evolution of a smooth process on a continuoum
- The sample curves are observed at discrete time points that could be different among sample units
  - Dense (high frecuency measurements)
  - Sparse (less frequent data)
- May be measured noisily an infrequently, but have enough data to estimate the underlying smooth process
- In many cases, functional data have high resolution and low noise
- The number of variables is much greater than the sample size (high dimensional data)

What is functional data?

© A.M. Aguilera
aaguiler@ugr.es

- A functional random variable X takes values in a functional space H
- The most usual functional data set is given by a sample of curves

  $$\{x_i(t) : i = 1, \ldots, n\}$$

  obtained as realizations of a stochastic process $\{X(t) : t \in T\}$

  $$x(t) \in L_2(T) = \left\{ f : T \longrightarrow \mathbb{R} : \int_T f^2(t)\, dt < \infty \right\}$$

  the usual inner product $\langle f, g \rangle = \int_T f(t)\, g(t)\, dt$, $\forall f, g \in L^2(T)$
- The metric is fundamental for FDA techniques and must be consistent with the data
- H : Hilbert space with scalar product $\langle , \rangle$ and norm $||f||^2 = \langle f, f \rangle$
  Example: $\langle f, g \rangle_\lambda = \langle f, g \rangle + \lambda \left\langle D^2(f), D^2(g) \right\rangle$

What is functional data analysis (FDA)?

© A.M. Aguilera
aaguiler@ugr.es

Statistical techniques for describing and modeling functional data

Goals

- To estimate the true functional form from noisy and discrete observations
- To display the data to highlight relevant characteristics
- To study important sources of pattern and variation among the data
- To model relationship of functional data to other related variables
- To study relationships between derivatives of the functions

Historical references

- Ramsay, J.O. (1982). When the data are curves. Psychometrika 47 (4), 379-396
- Ramsay, J.O. and Silverman, B.W. (1999). Functional Data Analysis. Springer-Verlag.

J.O. Ramsay

Department of Psychology.
McGill University. Canada

B.W. Silverman

Department of Statistics.
Oxford University

Historical references

© A.M. Aguilera
aaguiler@ugr.es

- Basic tool: Functional Principal Component Analysis (FPCA)
  - Karhunen-Loève expansion of a stochastic process
    - M. Loève (1946). Analyse harmonique générale d'une fonction aléatoire. Comptes Rendus de l'Académie des Sciences, 220, 380-382
    - K.O. Karhunen (1946). Zür spektraltheorie stochastischer prozesse. Annales Academiae Scientarum Fennicae, Ser. A, 34
  - J.C. Deville (1974). Méthodes Statistiques et Numériques de l'Analyse Harmonique. Annales de l'INSEE, 15, 3-101
  - J. Dauxois, A. Pousse and Y. Romain (1985). Asymtotic Theory for the Principal Component Analysis of a vector Ramdom Function: Some Applications to Statistical Inference. Journal of Multivariate Analysis, 12, 136-154

From Discrete to Functional Data

© A.M. Aguilera
aaguiler@ugr.es

Represent data recorded at discrete times as a continuous function

- Allow evaluation of the curve at any time point
- Evaluate rates of change
- Reduce noise
- Allow registration onto a common time-scale

Basis-expansion methods

$\{\phi_j(t)\}_{j=1,\ldots,\infty}$ basis of the functional space $H = L^2[T]$

$$x_i\,(t) = \sum_{j=1}^{\infty} a_{ij}\phi_j\,(t)$$

- Polynomials (unstable)
- Fourier or Trigonometric (periodic curves)
- B-splines (smooth curves)
- Wavelets (curves with discontinuities and sharp spikes)

© A.M. Aguilera
aaguiler@ugr.es

Basis-expansion methods

$$x_i = \sum_{j=1}^{K} a_{ij}\phi_j = \phi^T a_i$$

$\phi = (\phi_1, \ldots, \phi_K)^T$; $a_i = (a_{i1}, \ldots, a_{iK})^T$

Problem

Estimating basis coefficients from discrete observations y_{ik} of $\{x_i(t) : i = 1, \ldots, n; t \in T\}$ at time points $\{t_{ik} : k = 1, \ldots, m_i\}$

Basis-expansion methods

© A.M. Aguilera
aaguiler@ugr.es

$$x_i = \sum_{j=1}^{K} a_{ij}\phi_j = \phi^T a_i$$

$\phi = (\phi_1, \ldots, \phi_K)^T$; $a_i = (a_{i1}, \ldots, a_{iK})^T$

Problem

Estimating basis coefficients from discrete observations y_{ik} of $\{x_i(t) : i = 1, \ldots, n; t \in T\}$ at time points $\{t_{ik} : k = 1, \ldots, m_i\}$

Solution

- Noisy data: $y_{ik} = x_i(t_{ik}) + \varepsilon_{ik} \Longrightarrow$ Least squares smoothing by penalizing the roughness of de curve minimizing

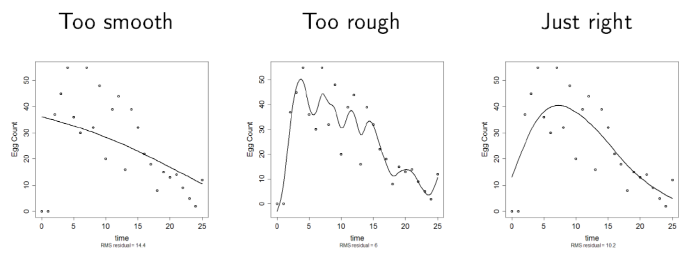
  $$\sum_{k=1}^{m_i} (y_{ik} - x_i(t_{ik}))^2 = (y_i - \Phi_i a_i)^T (y_i - \Phi_i a_i)$$

  $y_i = (y_{i1}, \ldots, y_{im_i})$; $\Phi_i = (\phi_j(t_{ik}))_{m_i \times K}$
- Data without error: $y_{ik} = x_i(t_{ik}) \Longrightarrow$ Interpolation (spline cúbica, ...)

© A.M. Aguilera
aaguiler@ugr.es

What is smoothness?

Eliminate small wiggles and retain the true shape



- Roughness measure: $Lx(t)$ (x is smooth if $Lx(t) = 0$)
- Examples:
  - Curvature: $Lx(t) = D^2x(t) \Rightarrow$ straight lines are smooth
  - Harmonic acceleration: $Lx(t) = w^2Dx(t) + D^3x(t) \Rightarrow$ sinusoid functions $cos(wt)$ and $sin(wt)$ are smooth
- Total size of roughness: $\int_T [Lx(t)]^2 dt$

Penalized smoothing

© A.M. Aguilera
aaguiler@ugr.es

- Noisy data: $y_{ik} = x_i(t_{ik}) + \varepsilon_{ik}$
- Penalizing the roughness of de curve by minimizing

$$\sum_{k=1}^{m_i} (y_{ik} - x_i(t_{ik}))^2 + \lambda \int_T (Lx_i(t))^2 dt = (y_i - \Phi_i a_i)^T (y_i - \Phi_i a_i) + \lambda a_i^T P_d a_i$$

$y_i = (y_{i1}, \ldots, y_{im_i})$; $\Phi_i = (\phi_j(t_{ik}))_{m_i \times K}$
$Lx_i(t)$ measures roughness of x_i, λ a smoothing parameter
$P_d = (\langle L\phi_j, L\phi_k \rangle)$ penalty matrix

Two main problems

- Choosing an appropriate basis
- Selecting the dimension of the basis and the smoothing parameters (leave-one-out cross-validation, generalized cross-validation, ...)

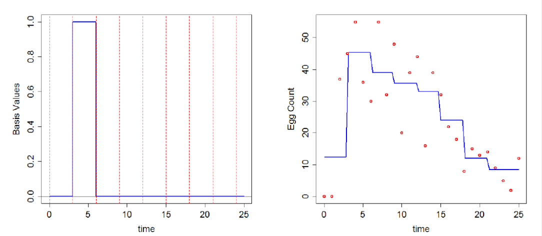
B-spline bases (de Boor, 2001)

© A.M. Aguilera
aaguiler@ugr.es

- Splines are polynomial segments (order m) constrained to be smooth at the joins (knots) (continuous derivatives up to order m-2)
- Basis defined by the order m (order = degree+1) of the polynomial and the location of the knots
  (dimension = order + number interior knots)

Eggs laid from 789 medflies during the first 25 days of their lives

Medfly data with knots every 3 days (Carey et al., 1998)
(Splines of order 1: piecewise constant, discontinuous)



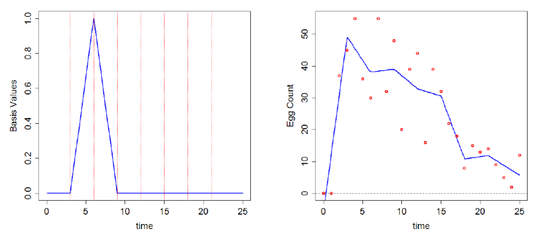
B-spline bases (de Boor, 2001)

© A.M. Aguilera
aaguiler@ugr.es

- Splines are polynomial segments (order m) constrained to be smooth at the joins (knots) (continuous derivatives up to order m-2)
- Basis defined by the order m (order = degree+1) of the polynomial and the location of the knots
  (dimension = order + number interior knots)

Eggs laid from 789 medflies during the first 25 days of their lives

Medfly data with knots every 3 days (Carey et al., 1998)
(Splines of order 2: piecewise linear, continuous)

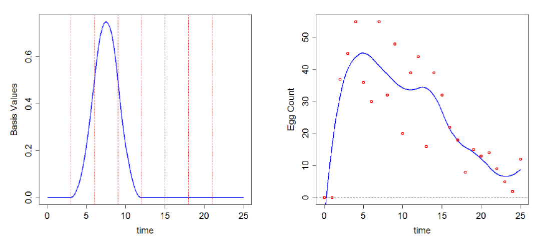


B-spline bases (de Boor, 2001)

- Splines are polynomial segments (order m) constrained to be smooth at the joins (knots) (continuous derivatives up to order m-2)
- Basis defined by the order m (order = degree+1) of the polynomial and the location of the knots
  (dimension = order + number interior knots)

Eggs laid from 789 medflies during the first 25 days of their lives

Medfly data with knots every 3 days (Carey et al., 1998)
(Splines of order 3: piecewise quadratic, continuous derivatives)



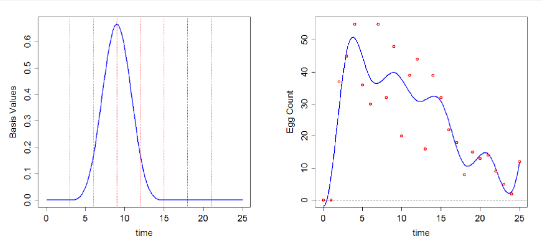
B-spline bases (de Boor, 2001)

© A.M. Aguilera
aaguiler@ugr.es

- Splines are polynomial segments (order m) constrained to be smooth at the joins (knots) (continuous derivatives up to order m-2)
- Basis defined by the order m (order = degree+1) of the polynomial and the location of the knots
  (dimension = order + number interior knots)

Eggs laid from 789 medflies during the first 25 days of their lives

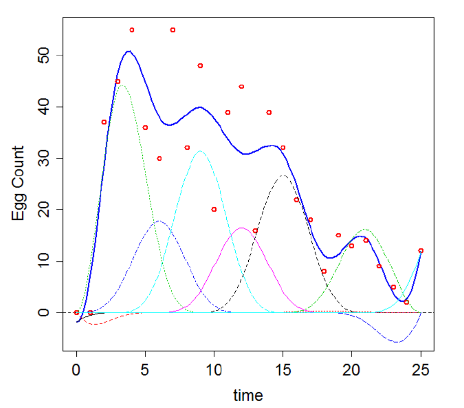
Medfly data with knots every 3 days (Carey et al., 1998)
(Splines of order 4: piecewise cubic, continuous 2nd derivatives)



B-spline smoothing

© A.M. Aguilera
aaguiler@ugr.es

Example of basis expansions with cubic B-splines for egg count of a medfly



B-spline expansions

© A.M. Aguilera
aaguiler@ugr.es

- Regression splines: Least squares smoothing without penalty (De Boor, 1977) ($\lambda = 0$)
- Smoothing splines: Penalized least squares smoothing based on d-order derivatives(O'Sullivan 1986)

$$P_d = \left(\int_T D^d \phi_j(s) \, D^d \phi_k(s)' \, ds \right)$$

- Penalized splines: Penalized least squares smoothing based on d-order differences (P-splines) (Eilers y Marx, 1996)

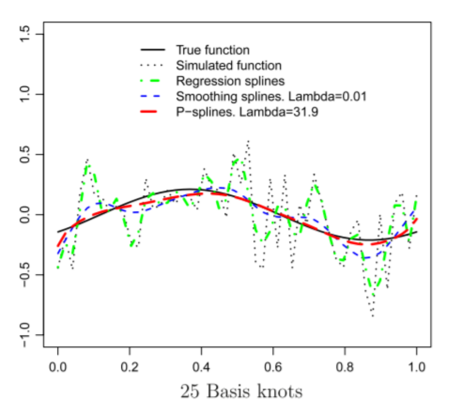
$$P_d = \left(\triangle^d \right)' \triangle^d$$

(Operador de diferencias de orden d $\triangle^d$)

B-spline smoothing

© A.M. Aguilera
aaguiler@ugr.es

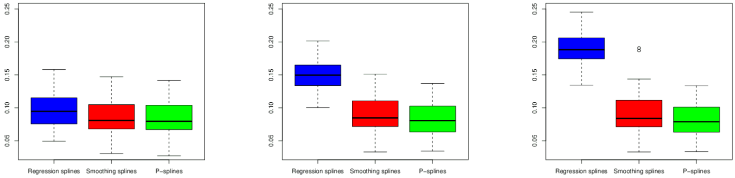
Comparative study (Aguilera et al., 2013)



Regression splines (green), smoothing splines (blue) and P-splines (red) with 25 basic knots

B-spline expansions

Mean square error for 100 curves approximated with regression splines (blue), smoothing splines (red) and P-splines (green) with 5, (left) 15 (center) and 25 (right) knots



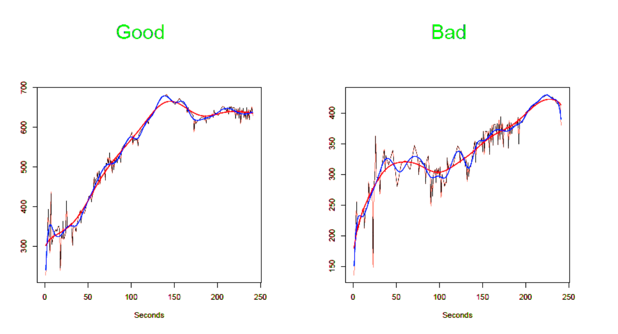
Advantages of P-splines(Aguilera et al., 2013)

- Provide the lowest approximation errors
- Less numerical complexity and computational cost
- The choice and position of knots is not determinant so that it is sufficient to choose a relatively large number of equally spaced basis knots (Ruppert et al., 2003)

Kneading data

© A.M. Aguilera
aaguiler@ugr.es

Curves of resistance of dough during the kneading process

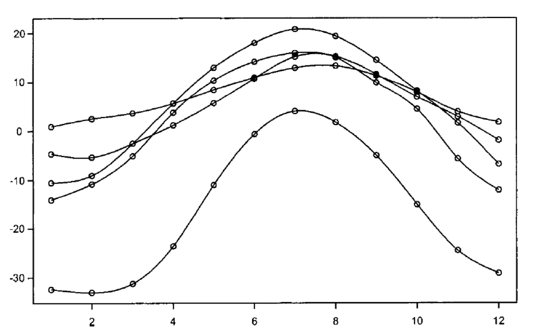


Original sample curve (brown), regression splines (blue), and P-spline approach (red) of a sample curve for good flour (left) and bad flour (right)

Weather curves: B-spline interpolation

© A.M. Aguilera
aaguiler@ugr.es

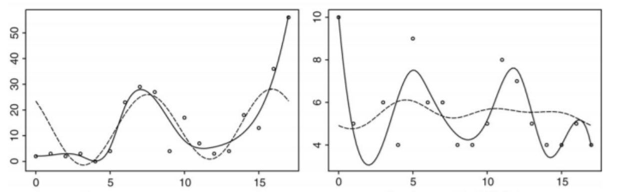
Quasi-natural cubic spline interpolation of curves of monthly temperatures in Celsius degrees used to predict the risk of drought (Escabias et al, 2005)



Fourier basis

Sine and cosine functions of increasing frequency

$$\phi_0(t) = 1, \phi_{2j-1}(t) = \sin\left(\frac{2\pi jt}{T}\right), \phi_{2j}(t) = \cos\left(\frac{2\pi jt}{T}\right) \quad j = 1, \ldots$$

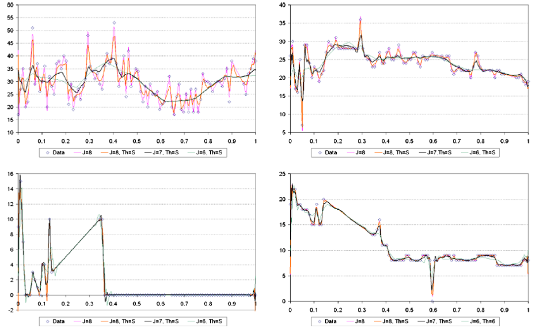


Least squares approximations of stress level curves obtained with Fourier (broken line) and B-spline (solid line) bases and discrete-time observations (points) (Aguilera et al., 2008)

Wavelets bases

Dilations and translations of the mother and father functions ϕ and ψ

$$\phi_{j,k}(t) = 2^{j/2}\phi\left(2^j t - k\right) \quad \text{and} \quad \psi_{j,k}(t) = 2^{j/2}\psi\left(2^j t - k\right)$$



Curves of stress (left) and lupus (right): wavelet projection (Symmlet 4 family, resolution level J=8) (pink), smoothing with soft thresholding J=8 (red), J=7 (black) and J=6 (blue) (Aguilera et al., 2015)

Descriptive analysis of functional data

© A.M. Aguilera
aaguiler@ugr.es

Sample of curves from $X : \{x_i(t) : t \in T, i = 1, \ldots, n\}$

- Mean function

$$\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$$

- Variance function

$$Var_X(t) = \frac{1}{n-1}\sum_{i=1}^{n} (x_i(t) - \bar{x}(t))^2$$

- Covariance function

$$Cov_X(t,s) = \frac{1}{n-1}\sum_{i=1}^{n} (x_i(t) - \bar{x}(t))(x_i(s) - \bar{x}(s))$$

- Correlation function

$$Corr_X(t,s) = \frac{Cov_X(t,s)}{\sqrt{Var_X(t)Var_X(s)}}$$

FCCA Chilean temperatures and precipitacions

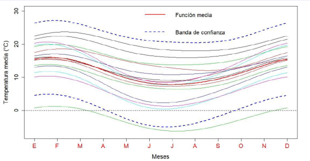
© A.M. Aguilera aaguiler@ugr.es

Chilean map

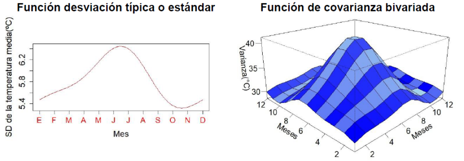
Temperatures for 15 Chilean stations

© A.M. Aguilera
aaguiler@ugr.es

Mean curve and point-wise confidence bands



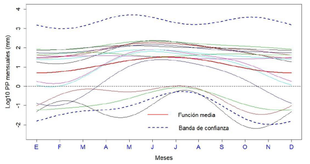
Standard deviation curve and covariance surface



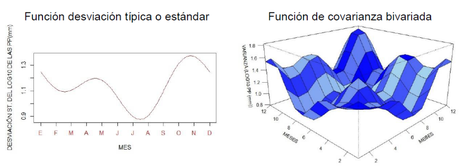
Precipitations ($\log_{10}$) for 15 Chilean stations

© A.M. Aguilera
aaguiler@ugr.es

Mean curve and point-wise confidence bands



Standard deviation curve and covariance surface



Descriptive analysis of functional data

Pairs of curves sampled from $(X, Y) : \{(x_i(t), y_i(t)) : t \in T, i = 1, \ldots, n\}$

- Cross covariance function

$$Cov_{X,Y}(t,s) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i(t) - \bar{x}(t)) \left(y_i(s) - \bar{y}(s)\right)$$

- Cross correlation function

$$Corr_{X,Y}(t,s) = \frac{Cov_{X,Y}(t,s)}{\sqrt{Var_X(t) Var_Y(s)}}$$

Functional PCA

Objective: dimension reduction by uncorrelated linear combinations with maximum variance

Functional PCA

© A.M. Aguilera
aaguiler@ugr.es

Objective: dimension reduction by uncorrelated linear combinations with maximum variance

Multivariate PCA

- Vector of p variables

$$X = (X_1, \ldots, X_p)^T$$

- $z = \sum_j v_j X_j$
- Covariance matrix Σ
- Eigen-decomposition

$$\Sigma = V \Delta V^T = \sum_{j=1}^{p} \lambda_j v_j v_j^T$$

- $v_i^T v_j = \delta_{ij}$

Functional PCA

© A.M. Aguilera
aaguiler@ugr.es

Objective: dimension reduction by uncorrelated linear combinations with maximum variance

Multivariate PCA

- Vector of p variables

$$X = (X_1, \dots, X_p)^T$$

- $z = \sum_j v_j X_j$
- Covariance matrix Σ
- Eigen-decomposition

$$\Sigma = V \Delta V^T = \sum_{j=1}^{p} \lambda_j v_j v_j^T$$

- $v_i^T v_j = \delta_{ij}$

Functional PCA

- Functional var. on $L^2(T)$

$$X = \{X(t) : t \in T\}$$

- $z = \int_T f(t) X(t) dt$
- Covariance surface C(t,s)
- Eigen-decomposition

$$Cov(t,s) = \sum_{j=1}^{\infty} \lambda_j f_j(t) f_j(s)$$

- $\int_T f_i(t) f_j(t) dt = \delta_{ij}$

Functional Principal Component Analysis (FPCA)

© A.M. Aguilera
aaguiler@ugr.es

- Functional variable X with sample functions in $L_2(T)$

$$< f, g >= \int_T f(t)g(t)dt$$

- Principal components: $z_j = \langle X, f_j \rangle = \int_T X(t) f_j(t)\, dt$

$$Max_f\, Var \langle X, f \rangle = \langle C(f), f \rangle$$

$$s.t. \left\{ ||f||^2 = \int_T f^2(t)dt = 1; \langle f_k, f \rangle = \int_T f_k(t) f(t)\, dt = 0 \;\; k \leq j-1 \right\}$$

Functional Principal Component Analysis (FPCA)

- Functional variable X with sample functions in $L_2(T)$

$$< f, g >= \int_T f(t)g(t)dt$$

- Principal components: $z_j = \langle X, f_j \rangle = \int_T X(t) f_j(t) dt$

$$Max_f \, Var \langle X, f \rangle = \langle C(f), f \rangle$$

$$s.t. \left\{ ||f||^2 = \int_T f^2(t)dt = 1; \langle f_k, f \rangle = \int_T f_k(t) f(t) dt = 0 \;\; k \leq j-1 \right\}$$

- Covariance operator: $C(f)(t) = \langle C(t,.), f \rangle = \int_T Cov(t,s) f(s) ds$

- Weight functions (Deville, 1974)

$$C(f_j)(t) = \lambda_j f_j(t) \quad t \in T$$

Functional Principal Component Analysis (FPCA)

© A.M. Aguilera
aaguiler@ugr.es

- Functional variable X with sample functions in $L_2(T)$

$$< f, g >= \int_T f(t)g(t)dt$$

- Principal components: $z_j = \langle X, f_j \rangle = \int_T X(t) f_j(t)\, dt$

$$Max_f \, Var \langle X, f \rangle = \langle C(f), f \rangle$$

$$s.t. \left\{ ||f||^2 = \int_T f^2(t)dt = 1; \langle f_k, f \rangle = \int_T f_k(t) f(t)\, dt = 0 \;\; k \leq j-1 \right\}$$

- Covariance operator: $C(f)(t) = \langle C(t,.), f \rangle = \int_T Cov(t,s) f(s)\, ds$
- Weight functions (Deville, 1974)

$$C(f_j)(t) = \lambda_j f_j(t) \quad t \in T$$

- Principal component decomposition (Karhunen-Loève, 1946)

$$X(t) = \sum_{j=1}^{\infty} z_j f_j(t)$$

Functional Principal Component Analysis (FPCA)

© A.M. Aguilera
aaguiler@ugr.es

- Principal component decomposition (Karhunen-Loève, 1946)

$$X(t) = \sum_{j=1}^{\infty} z_j f_j(t)$$

Functional Principal Component Analysis (FPCA)

© A.M. Aguilera
aaguiler@ugr.es

- Principal component decomposition (Karhunen-Loève, 1946)

$$X(t) = \sum_{j=1}^{\infty} z_j f_j(t)$$

- Principal component reconstruction

$$X^q(t) = \sum_{j=1}^{q} z_j f_j(t)$$

Explained variance $V^q = \left(\sum_{j=1}^{q} \lambda_j\right) \big/ \left(\sum_{i=1}^{\infty} \lambda_i\right)$

Functional Principal Component Analysis (FPCA)

© A.M. Aguilera
aaguiler@ugr.es

- Principal component decomposition (Karhunen-Loève, 1946)

$$X(t) = \sum_{j=1}^{\infty} z_j f_j(t)$$

- Principal component reconstruction

$$X^q(t) = \sum_{j=1}^{q} z_j f_j(t)$$

Explained variance $V^q = \left(\sum_{j=1}^{q} \lambda_j\right) / \left(\sum_{i=1}^{\infty} \lambda_i\right)$

- FPCA of basis expansions $\Leftrightarrow$ Multivariate PCA of matrix $A\Psi^{1/2}$ (Ocaña et al., 2007)
  - $x_i(t) = \sum_{j=1}^{K} a_{ij}\phi_j(t)$
  - A matrix of basic coefficients $A = (a_{ij})_{i=1,\ldots,n;j=1,\ldots,K}$
  - Ψ matrix of inner products $\left(< \phi_j, \phi_k >_{L^2(T)}\right)_{j,k=1,\ldots,K}$

Smoothed FPCA

- Hybrid FPCA (Silverman, 1996)

  Generalized linear combinations with maximum variance restricted to orthonormality of weight functions with the inner product

$$\langle f, g \rangle_\lambda = \langle f, g \rangle + \lambda \langle D^p(f), D^p(g) \rangle$$

Smoothed FPCA

© A.M. Aguilera
aaguiler@ugr.es

- Hybrid FPCA (Silverman, 1996)

  Generalized linear combinations with maximum variance restricted to orthonormality of weight functions with the inner product

$$\langle f, g\rangle_\lambda = \langle f, g\rangle + \lambda \langle D^p(f), D^p(g)\rangle$$

- Principal components $z_j = \langle X, f_j\rangle = \int_T X(t) f_j(t)\, dt$

$$Max_f\, Var\, \langle X, f\rangle = \langle C(f), f\rangle$$

$$s.t. \left\{||f||_\lambda^2 = 1; \langle f_k, f\rangle_\lambda = 0 \;\; \forall k \leq j-1\right\}$$

Smoothed FPCA

- Hybrid FPCA (Silverman, 1996)

  Generalized linear combinations with maximum variance restricted to orthonormality of weight functions with the inner product

$$\langle f, g\rangle_\lambda = \langle f, g\rangle + \lambda \langle D^p(f), D^p(g)\rangle$$

- Principal components $z_j = \langle X, f_j\rangle = \int_T X(t) f_j(t)\, dt$

$$Max_f\, Var\,\langle X, f\rangle = \langle C(f), f\rangle$$

$$s.t. \left\{||f||_\lambda^2 = 1; \langle f_k, f\rangle_\lambda = 0 \;\; \forall k \leq j-1\right\}$$

- FPCA in $L^2(T)$ of smoothed data $S(x_i)$ (Ocaña et al., 1999)
- For basis expansion of data (Aguilera and Aguilera-Morillo, 2013)

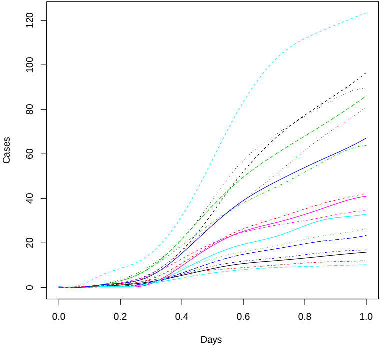
$$S(x_i) = \phi^T((\Psi + \lambda P_d)^{-1}\Psi)^{1/2} a_i$$

FPCA of COVID-19 cases in Spain

© A.M. Aguilera
aaguiler@ugr.es

- Data: daily cumulative informed cases of COVID-19 for 17 autonomous communities (ACs) in Spain from 20/02/2020 to 27/04/2020 (first wave of COVID-19)
- The curves are daily observed starting the day that at least one case is reported. Therefore, the period of observation and the number of observations are different for each AC
- Data homogeneization: the number of cases per 10000 inhabitants is considered and the first observation for each curve is the day that exceeds by first time the maximum of the first reported values
- Data registration: all the curves were registered in the common interval $[0, 1]$ (Ramsay and Silverman, 2005)

FPCA of COVID-19 cases in Spain

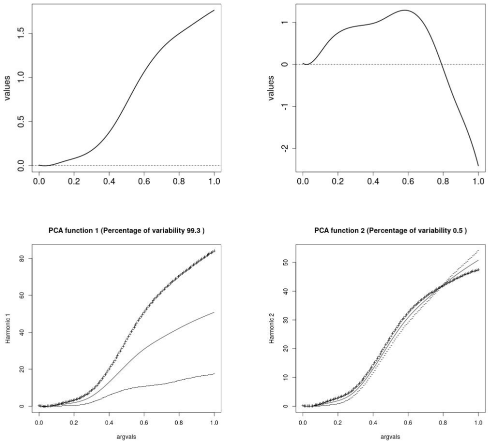


Regression splines (10 B-spline basis functions) for the number of daily cumulative informed cases by COVID-19 per 10000 inhabitants registered in [0,1] for 17 autonomous communities in Spain

FPCA of COVID-19 cases in Spain

© A.M. Aguilera
aaguiler@ugr.es

Weight functions as perturbations of the mean: $\bar{x}(t) \pm 2\sqrt{\lambda_j} f_j(t)$



$f_1(t)$: Quick increase in cases as time passed with the infection curve out of control. The rest of components are difficult to interpret

© A.M. Aguilera
aaguiler@ugr.es

Functional PCA rotation

VARIMAX rotation criterion on FPCA of X(t) (Acal et al., 2021)

- FPCA of X(t) $\leftrightarrow$ PCA of matrix $A\Psi^{1/2}$

$$
\begin{aligned}
Cov(A\Psi^{1/2}) &= V\Lambda V^T \text{ with } \Lambda = Diag(\lambda_j) \\
A\Psi^{1/2} &\approx Z_q V_q^T = (Z_q \Lambda_q^{-1/2})(V_q \Lambda_q^{1/2})^T = \tilde{Z}_q \tilde{V}_q^T \\
x^q &= Z_q f_q \quad f_q^T = \Phi^T \Psi^{-1/2} V_q
\end{aligned}
$$

$A_{n\times K}$ basic coefficients; $\Psi_{K\times K}$ inner products

V_q matrix with first q eigenvectors

$x^q = (x_1^q, \ldots, x_n^q)^T; \quad f_q = (f_1, \ldots, f_q)^T; \quad \Phi = (\Phi_1, \ldots, \Phi_K)^T$

- Rotating loadings of PC scores (eigenvectors) $V_q^R = V_q R \quad (RR^T = I)$

  Orthonormal rotated principal factors: $f_q^{R^T} = \phi^T (\Psi^{-1/2} V_q^R)$

  Correlated PC scores: $Z_q^R = Z_q R$

- Rotating loadings of standardized PC scores $\tilde{V}_q^R = \tilde{V}_q R = V_q \Lambda_q^{1/2} R$

  Non-orthonormal rotated principal factors: $\tilde{f}_q^{R^T} = \phi^T \left(\Psi^{-1/2} \tilde{V}_q^R \right)$

  Uncorrelated PC scores: $\tilde{Z}_q^R = \tilde{Z}_q R$

Varimax FPCA of COVID-19 cases in Spain

© A.M. Aguilera
aaguiler@ugr.es

PC	FPCA	Rotation
1	99.32	44.36
2	0.52	38.14
3	0.12	0.67
4	0.03	14.82

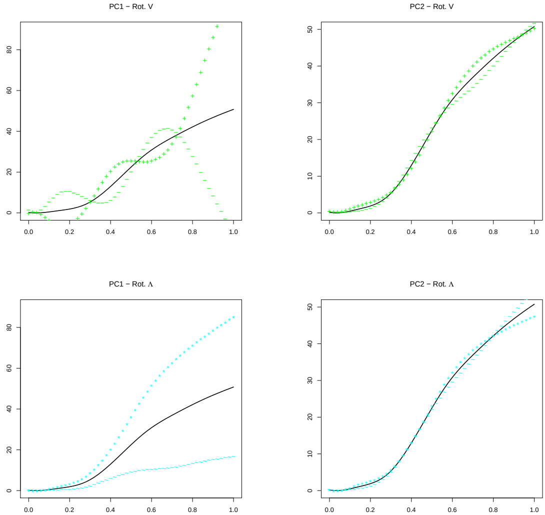
Percentages of variance explained by the first four PCs of COVID-19 curves

Andalucía, AN	Castilla-La Mancha, CM	Madrid, MD
Aragón, AR	Castilla-León, CL	Murcia, MC
Asturias, AS	Cataluña, CT	Navarra, NC
Islas Baleares, IB	Comunidad Valenciana, VC	País Vasco, PV
Canarias, CN	Extremadura, EX	Rioja, RI
Cantabria, CB	Galicia, GA	

Abbreviation of the seventeen Spanish autonomous communities

Varimax FPCA of COVID-19 cases in Spain

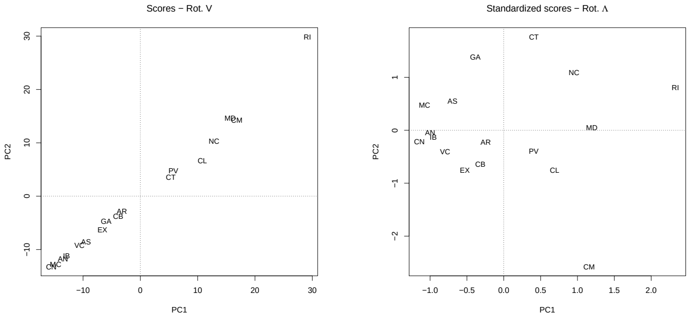
© A.M. Aguilera
aaguiler@ugr.es



Weight functions as perturbations of the mean: $\bar{x}(t) \pm 2\sqrt{\lambda_j} f_j(t)$

Varimax FPCA of COVID-19 cases in Spain

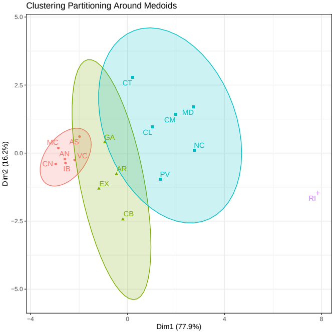
© A.M. Aguilera
aaguiler@ugr.es



PC scores of the seventeen Spanish autonomous communities on the first two rotated principal components of COVID-19 cases

Varimax FPCA of COVID-19 cases in Spain

© A.M. Aguilera
aaguiler@ugr.es



Functional clustering of COVID-19 cases

Functional regression

© A.M. Aguilera
aaguiler@ugr.es

Estimating a response variable Y from related predictors X where at least one variable is functional

- Functional linear models
  - Functional covariate, scalar response
  - Scalar or categorical covariate, functional response
  - Functional covariate, functional response
- Generalized linear models
  - Functional logit models ....
- Generalized additive models
- Mixed effects models

Scalar response models

- Objective: To model dependence of Y on $X(t)$
- Observations: $\{(y_i, x_i(t_j)) : j = 1, \ldots, m; i = 1, \ldots, n\}$

Scalar response models

© A.M. Aguilera
aaguiler@ugr.es

- Objective: To model dependence of Y on $X(t)$
- Observations: $\{(y_i, x_i(t_j)) : j = 1, \ldots, m; i = 1, \ldots, n\}$
- Model formulation:

$$y_i = \alpha + \sum_j \beta_j x_i(t_j) + \epsilon_i \Longrightarrow y_i = \alpha + \int \beta(t) x_i(t) dt + \epsilon_i$$

(multivariate linear model with sums replaced by integrals)

Scalar response models

© A.M. Aguilera
aaguiler@ugr.es

- Objective: To model dependence of Y on $X(t)$
- Observations: $\{(y_i, x_i(t_j)) : j = 1, \ldots, m; i = 1, \ldots, n\}$
- Model formulation:

$$y_i = \alpha + \sum_j \beta_j x_i(t_j) + \epsilon_i \Longrightarrow y_i = \alpha + \int \beta(t) x_i(t) dt + \epsilon_i$$

(multivariate linear model with sums replaced by integrals)

- Least squares estimation (ill-posed problem)

$$\beta = argmin \sum_{i=1}^{n} \left(y_i - \left(\alpha + \int \beta(t) x_i(t) dt \right) \right)^2$$

Model estimation

Solutions

1. Penalized least squares estimation

$$\beta = argmin \sum_{i=1}^{n} \left(y_i - \left(\alpha + \int \beta(t) x_i(t) dt \right) \right)^2 + \lambda \int_T (L\beta(t))^2$$

$\beta(t)$ smooth parameter function $\beta(t) = \sum_j b_j \phi_j(t)$

Model estimation

© A.M. Aguilera
aaguiler@ugr.es

Solutions

1. Penalized least squares estimation

$$\beta = argmin \sum_{i=1}^{n} \left(y_i - \left(\alpha + \int \beta(t) x_i(t) dt \right) \right)^2 + \lambda \int_T (L\beta(t))^2$$

$\beta(t)$ smooth parameter function $\beta(t) = \sum_j b_j \phi_j(t)$

2. Regression on uncorrelated predictor variables Z

$$z_{ij} = < x_i, \beta_j > = \int_T x_i(t) f_j(t) dt \Longrightarrow y_i = \alpha + \sum_j \beta_j z_{ij} + \epsilon_i$$

- Functional principal component regression (PCR)
- Functional partial least squares regression (PLSR)

Functional ANOVA

© A.M. Aguilera
aaguiler@ugr.es

- Objective: To model dependence of $Y(t)$ on a categorical covariate with J groups
- Observations: $\{(y_{ij}(t_k) : j = 1, \ldots, J; n; i = 1, \ldots, n; k = 1, \ldots, K\}$

© A.M. Aguilera
aaguiler@ugr.es

Functional ANOVA

- Objective: To model dependence of $Y(t)$ on a categorical covariate with J groups
- Observations: $\{(y_{ij}(t_k) : j = 1, \ldots, J; n; i = 1, \ldots, n; k = 1, \ldots, K\}$
- Model formulation:

$$y_{ij}(t) = \mu(t) + \alpha_j(t) + \epsilon_{ij}(t) \quad \left(\sum_j \alpha_j(t) \right) = 0$$

Functional ANOVA

© A.M. Aguilera
aaguiler@ugr.es

- Objective: To model dependence of $Y(t)$ on a categorical covariate with J groups
- Observations: $\{(y_{ij}(t_k) : j = 1, \ldots, J; n; i = 1, \ldots, n; k = 1, \ldots, K\}$
- Model formulation:

$$y_{ij}(t) = \mu(t) + \alpha_j(t) + \epsilon_{ij}(t) \quad \left(\sum_j \alpha_j(t)\right) = 0$$

- Estimation: MANOVA for funcional principal components Z
- Testing problem: $H_0 : \alpha_j = \mu_j - \mu = 0$

Biomechanics

© A.M. Aguilera
aaguiler@ugr.es

Experimental study developed in the biomechanics laboratories of the Sport and Health Institute of the University of Granada (IMUDS)

- Objective: Analyze biomechanics of children using school trolleys or backpacks in their daily transportation to school with different loads

© A.M. Aguilera
aaguiler@ugr.es

Biomechanics

Experimental study developed in the biomechanics laboratories of the Sport and Health Institute of the University of Granada (IMUDS)

- Objective: Analyze biomechanics of children using school trolleys or backpacks in their daily transportation to school with different loads
- Data: n=53 participants (25 boys, 28 girls)

  For each join (ankle, hip, knee, pelvis, thorax) and angle direction (X, Y, Z) we have 53×7 curves representing the gait cycle under each one of $J = 7$ conditions

  1. Walk
  2. Backpack with 10% of the body weight
  3. Backpack with 15% of the body weight
  4. Backpack with 20% of the body weight
  5. Trolley with 10% of the body weight
  6. Trolley with 15% of the body weight
  7. Trolley with 20% of the body weight

Biomechanics

© A.M. Aguilera
aaguiler@ugr.es

Experimental study developed in the biomechanics laboratories of the Sport and Health Institute of the University of Granada (IMUDS)

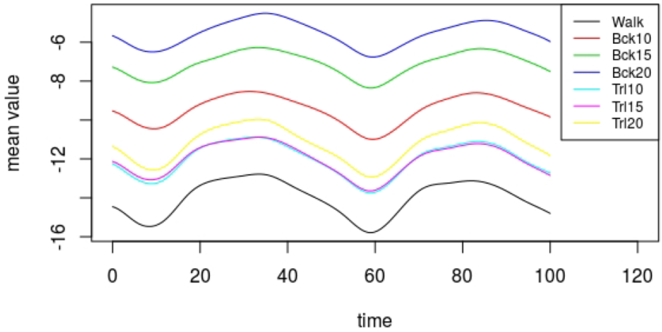
- Objective: Analyze biomechanics of children using school trolleys or backpacks in their daily transportation to school with different loads
- Data: n=53 participants (25 boys, 28 girls)

  For each join (ankle, hip, knee, pelvis, thorax) and angle direction (X, Y, Z) we have 53×7 curves representing the gait cycle under each one of $J = 7$ conditions

  1. Walk
  2. Backpack with 10% of the body weight
  3. Backpack with 15% of the body weight
  4. Backpack with 20% of the body weight
  5. Trolley with 10% of the body weight
  6. Trolley with 15% of the body weight
  7. Trolley with 20% of the body weight
- Functional ANOVA Test: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$

Gait data: thorax

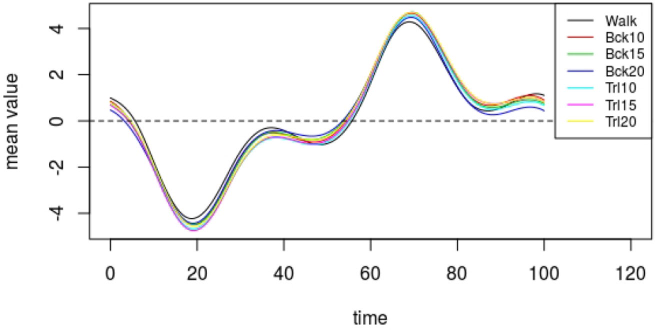
© A.M. Aguilera
aaguiler@ugr.es



Sample mean curves (Axis X)

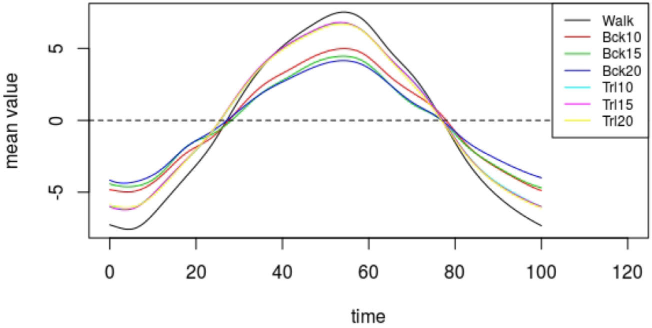
Gait data: thorax

© A.M. Aguilera
aaguiler@ugr.es



Sample mean curves (Axis Y)

Gait data: thorax



Sample mean curves (Axis Z)

© A.M. Aguilera
aaguiler@ugr.es

Functional logit regression

- Objective: Estimating a binary response Y with categories $\{Y_1 = 1, Y_2 = 0\}$ from a functional predictor $\{X(t) : t \in T\}$
- Observations: $\{(y_i, x_i(t_j)) : j = 1, \ldots, m; i = 1, \ldots, n\}$

Functional logit regression

- Objective: Estimating a binary response Y with categories $\{Y_1 = 1, Y_2 = 0\}$ from a functional predictor $\{X(t) : t \in T\}$
- Observations: $\{(y_i, x_i(t_j)) : j = 1, \ldots, m; i = 1, \ldots, n\}$
- Model formulation: $y_i = \pi_i + \varepsilon_i$

$$\pi_i = P[Y = 1/X(t) = x_i(t)] = \frac{\exp\left\{\alpha + \int_T x_i(t)\beta(t)\,dt\right\}}{1 + \exp\left\{\alpha + \int_T x_i(t)\beta(t)\,dt\right\}}$$

- Functional generalized linear model

$$\log\left[\pi_i/(1 - \pi_i)\right] = \alpha + \int_T x_i(t)\beta(t)\,dt$$

Functional logit regression

© A.M. Aguilera
aaguiler@ugr.es

- Objective: Estimating a binary response Y with categories $\{Y_1 = 1, Y_2 = 0\}$ from a functional predictor $\{X(t) : t \in T\}$
- Observations: $\{(y_i, x_i(t_j)) : j = 1, \ldots, m; i = 1, \ldots, n\}$
- Model formulation: $y_i = \pi_i + \varepsilon_i$

$$\pi_i = P[Y = 1/X(t) = x_i(t)] = \frac{\exp\left\{\alpha + \int_T x_i(t)\,\beta(t)\,dt\right\}}{1 + \exp\left\{\alpha + \int_T x_i(t)\,\beta(t)\,dt\right\}}$$

- Functional generalized linear model

$$\log\left[\pi_i/(1-\pi_i)\right] = \alpha + \int_T x_i(t)\,\beta(t)\,dt$$

- Estimation: Logit regression of Y on uncorrelated variables Z (PC (Escabias et al., 2004; 2005), PLS (Escabias et al., 2007))

Kneading data

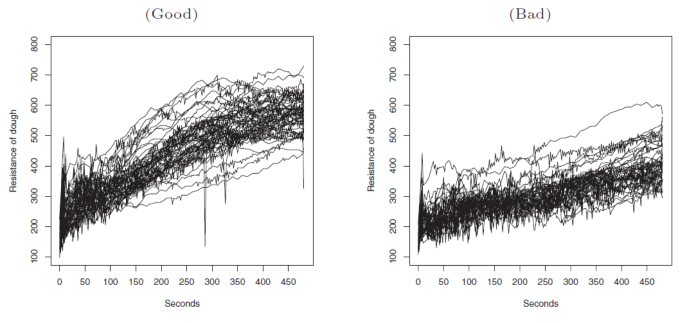
© A.M. Aguilera
aaguiler@ugr.es

- Objective: Predicting the quality of cookies from Danone (good or bad) from the resistance of dough during the kneading process $\{X(t) : t \in [0, 480]\}$ (Aguilera-Morillo and Aguilera, 2015)
- Functional predictor: curves of resistance of dough observed every two seconds $\{X_i(t_j) : t_j \in [0, 480]\}$ $\{t_j = 2 \times j : j = 0, \ldots, 240\}$
- Binary response: $Y = 1$ for good cookies and $Y = 0$ for bad cookies
- Sample of 90 flours divided in a training sample of size 60 and a test sample of size 30
- Solution: FPCA on P-spline smoothing of curves and Logit regression on PCs

Kneading data

© A.M. Aguilera
aaguiler@ugr.es

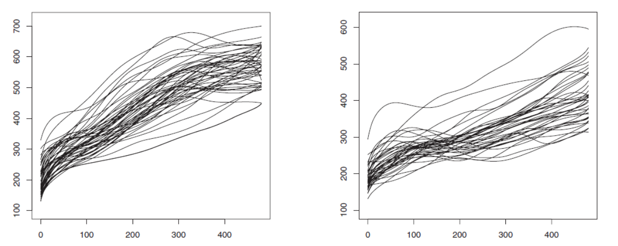
Raw curves of resistance of dough



Kneading data

© A.M. Aguilera
aaguiler@ugr.es

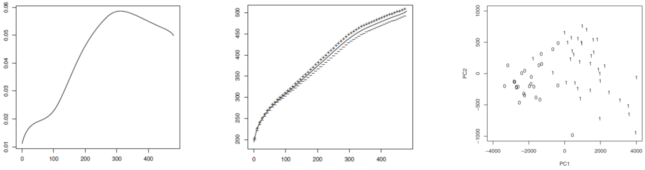
P-spline smoothing of curves of resistance of dough



Kneading data: FPCA

© A.M. Aguilera
aaguiler@ugr.es

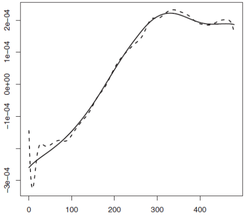
First eigenfunction (left), perturbed mean curve (center) and dispersion graph between the first and second PCs (right)



- The first principal component (93.59% of the total variability) gives negative weights to bad cookies and positive weights to good cookies
- The first PC curve represents the main features of the curve of resistance of a good cookie
- The main mode of variation of the resistance curves is associated to the quality of the cookies and allows to identify the best quality flours

Kneading data: Logit regression on first PC

© A.M. Aguilera
aaguiler@ugr.es



Functional parameter interpretation

- Shape
  A good flour must have less resistance during the early period of the kneading process, and more resistance in the late period
- Odds ratios
  - $\hat{OR}_1^{(\Delta 10)} = \exp\left(10\int_0^{186}\hat{\beta}(t)\,dt\right) = 0.56 \Longrightarrow$ a constant increase of 10 units in resistance during period $(0, 186)$ will halve the odds of producing a good cookie
  - $\hat{OR}_2^{(\Delta 10)} = \exp\left(10\int_{186}^{480}\hat{\beta}(t)\,dt\right) = 1.9 \Longrightarrow$ the same increase in period $(186, 480)$ will double the odds of good cookie

Lupus flare/stress relationship

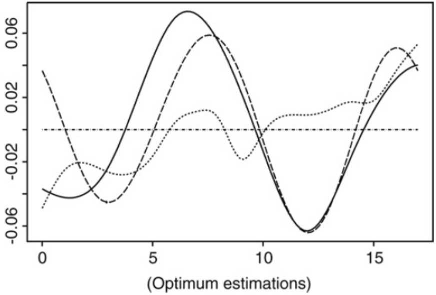
© A.M. Aguilera
aaguiler@ugr.es

- Objective: Predicting the probability of lupus flare from time evolution of stress level (Aguilera et al., 2008)
- Data: 44 SLE patients who were asked to respond to different stress tests over a period of 18 days (several patients did not provide observations on the third and twelfth days)
- Principal component logit regression in terms of only the third PC of stress (less than 10% variability))
- The consequences of high stress levels on a lupus flare have a lag of approximately five days
- The odds of lupus flare are doubled when the stress level increases by five times the third eigenfunction
- CCR 96%

Autoimmune Diseases Section of the Internal Medicine Department of the Virgen de las Nieves hospital (Granada, Spain)
Department of Personality, Testing and Psychological Treatment of the University of Granada

Lupus flare/stress relationship

© A.M. Aguilera
aaguiler@ugr.es



Parameter interpretation

- Shape interpretation: people with a high stress level around the maxima ($t = 6.6$ and $t = 17$) have a higher probability of suffering a lupus flare, while an absolute high level around the minima ($t = 1.2$ and $t = 12$) would reduce this probability
- As a result, the consequences of high stress level on a lupus flare have a lag of approximately five days

Functional regression summary

© A.M. Aguilera
aaguiler@ugr.es

All linear regression methods reduce to

- Turn sums into integrals (ill-posed problem)
- Basis expansion and penalized estimation $\Longrightarrow$ Multicollineality and high dimension
- Perform FPCA or FPLS $\Longrightarrow$ Regression on uncorrelated predictor variables (PC or PLS scores)

Applied in functional versions of

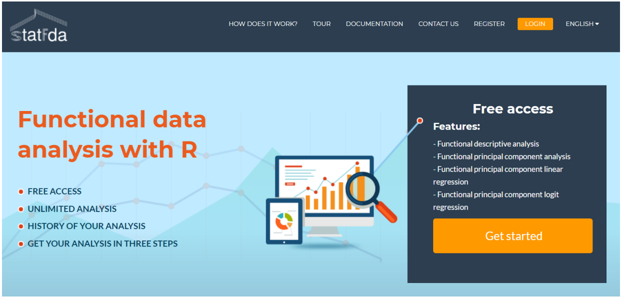
- Generalized linear models
- Generalized additive models
- Mixed models
- Time series of functions
- Cluster Analysis, Linear Discriminant Analysis, ...

FDA packages on R software

- Library fda: functional data models described in the book of Ramsay and Silverman (2005)
- Library fda.usc: fda library plugin with depth measures, functional outliers detection, functional regression, functional data classification, functional ANOVA
- Library fds: functional data sets
- Library logitFD: functional logistic regression
- Library pfica: penalized independent component analysis for univariate functional data
- Library *far*: modelling Hilbertian autoregressive processes autoregresivos hilbertianos
- Library *ftsa*: functional time series analysis
- Library *MFDF*: modelling finance functional data
- Library *refund*: regression with functional data
- Library *nfda*: non-parametric functional data analysis
- Library *rainbow*: graphics for exploratory functional data analysis

StatFda

© A.M. Aguilera
aaguiler@ugr.es



Research project P11-FQM-8068 "Functional Statistical Methods. Development of a Web Interface for their Application" Consejería de Innovación Ciencia y Empresa. Junta de Andalucía

References

- Kokoszka, P. (2017). Introduction to Functional Data Analysis. Chapman and Hall/CRC.
- Zhang, J-T. (2014). Analysis of Variance for functional data. CRC.
- Horvath , L. and Kokoszka, P. (2012). Inference for functional data with applications. Springer-Verlag.
- Shi, J.Q. and Choi, T. (2011). Gaussian process regression analysis for functional data. CRC.
- Ramsay, J.O., Hooker, G. and Graves, S. (2009). Functional Data Analysis with R and MATLAB. Springer-Verlag.
- Ferraty F, Vieu P (2006). Nonparametric Functional Data Analysis. Springer-Verlarg.
- Ramsay, J.O. and Silverman, B.W. (1999, 2005(2nd edition)). Functional Data Analysis. Springer-Verlag.
- Ramsay, J.O. and Silverman, B.W. (2002). Applied Functional Data Analysis: Methods and Case studies. Springer-Verlag.
- Valderrama, M.J., Aguilera, A.M. y Ocaña, F.A. (2000). Predicción Dinámica mediante Análisis de Datos Funcionales. Hespérides-La Muralla.

Acknowledgements

- Research group FQM-307 "Modelling and forecasting with functional data" Consejería de Innovación Ciencia y Empresa. Junta de Andalucía
- Research proyect MTM2017-88708-P "Methodological and applied contributions for stochastic and functional modelling of statistical data" Ministerio de Economía y Competitividad, Gobierno de España
- Department of Statistics and O.R. University of Granada
- Institute of Mathematics. University of Granada (IMAG)
- Working group "Functional Data Analysis". Sociedad de Estadística e I.O. (SEIO)