Functional Data Analysis: Techniques and Applications

R. Todd Ogden and Jeff Goldsmith

March 17, 2014

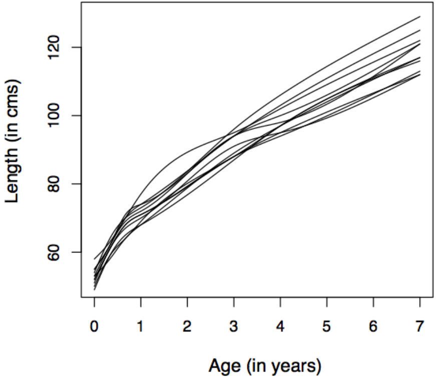
FDNY
FUNCTIONAL DATA NEW YORK
THE DEPARTMENT OF BIOSTATISTICS
Columbia University MAILMAN SCHOOL OF PUBLIC HEALTH

Outline

- Examples, definitions, notation
- Display
- Smoothing
- Functional principal components analysis
- Regression with functional predictors and/or responses

What is functional data?

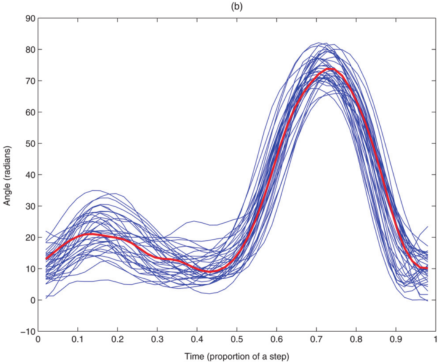
Some examples...



Child height as a function of age.

What is functional data?

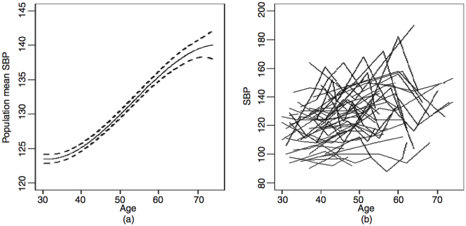
Some examples...



Knee angle as children go through a gait cycle.

What is functional data?

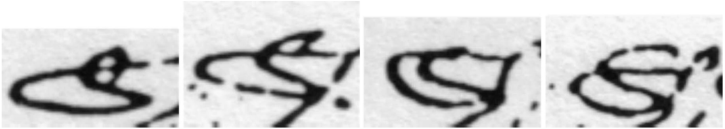
Some examples...



Systolic blood pressure at various ages for 150 subjects.

What is functional data?

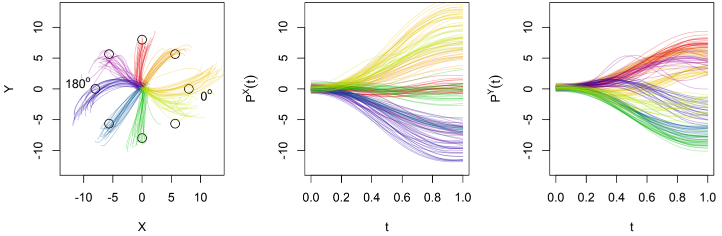
Some examples...



Examples of the S in Shakespeare's signature

What is functional data?

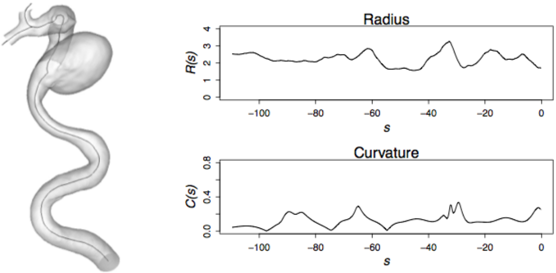
Some examples...



Reaching motions made by a stroke patient

What is functional data?

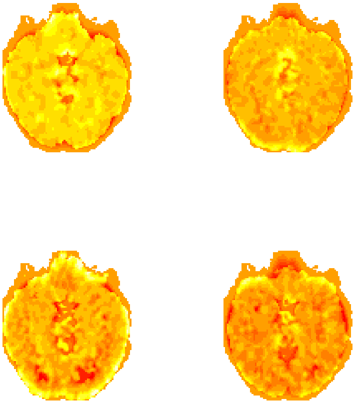
Some examples...



Curvature and radius of the carotid artery.

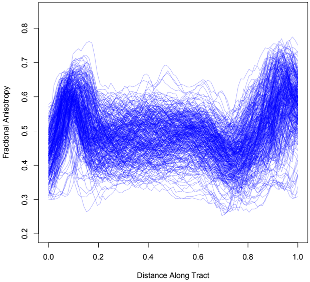
What is functional data?

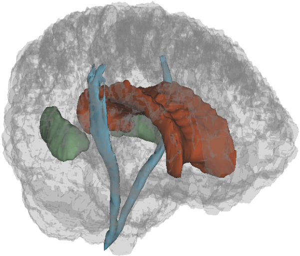
Some examples...



Brain images.

Recurring example: DTI





Tract profiles from diffusion tensor imaging

What is functional data?

Something like a definition:

> *"Observations on subjects that you can imagine as $X_i(s_i)$, where s_i is continuous"*

Functional notation is conceptual; observations are made on a finite discrete grid.

Some characteristics of functional data

The following are sometimes associated with functional data:

- High dimensional
- Temporal and/or spatial structure
- Interpretability across subject domains

Discretization of functional data

- Conceptually, we regard functional data as being defined on a continuum, e.g., $X_i(t)$, $0 \leq t \leq 1$.
- In practice, functional data are observed at a finite number of points.

Discretization of functional data

Dense functional data: Often, this is a fine regular grid, i.e., $\boldsymbol{x}_i = \left(X_i\left(\frac{1}{N}\right), X_i\left(\frac{2}{N}\right), \ldots, X_i(1)\right)$: spectral data, imaging data, accelerometry, ...

Sparse functional data: In other situations, the points at which observations are taken are irregular, often random: CD4 count, blood pressure, etc.

- In such cases, some kind of *interpolation* is necessary.

Functional data are technically multivariate data!

Why not just apply multivariate techniques (MANOVA, clustering, multiple regression, etc.)?

Functional data are technically multivariate data!

Why not just apply multivariate techniques (MANOVA, clustering, multiple regression, etc.)?

- Any technique for functional data should take into account the *structure* of the data — results from multivariate data analyses are generally permutation-invariant, but results from functional data analyses should *not* be!
- Methodological developments in FDA are often extensions of corresponding multivariate techniques.

Functional data are often observed with measurement error

- $X_i(t)$ is smooth (and continuously defined) but we observe

$$\boldsymbol{x}_i = \left(X_i \left(\frac{1}{N} \right) + \epsilon_1, X_i \left(\frac{2}{N} \right) + \epsilon_2, \ldots, X_i (1) + \epsilon_n \right)$$

- It is common to smooth the data before any analysis (topic we'll revisit soon)
- In other situations, accounting for measurement error is built in to the analysis procedure.

Comparison across observations

In order for functional data to be comparable across observations (e.g., across subjects), they must be observed on the same domain, i.e., t must be the same for $X_1(t)$ and $X_2(t)$.

In many cases, this is straightforward:

- Spectral data

Problematic for some other situations:

- Growth curves (for adolescents, “growth spurts” may not line up)
- Brain imaging data (structure is somewhat different from subject to subject)

In such cases it is often possible to *register* the data, e.g., using *landmarks* or by *warping*.

Summary measures for functional data

Suppose we have functional data $\{X_i(t),\ t \in \mathcal{T},\ i = 1, \ldots, n\}$.
Mean: $\mu(t) = EX_i(t)$.

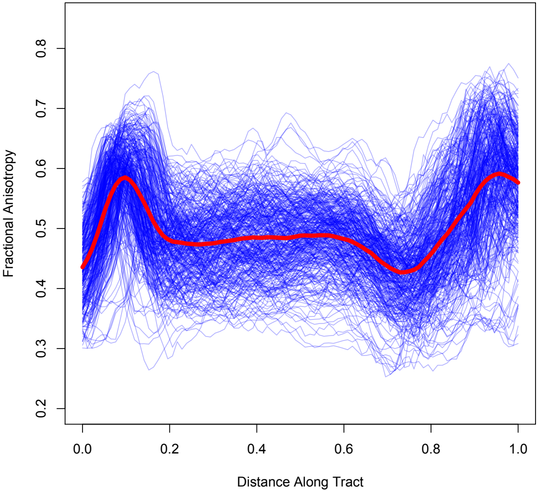
- The mean is itself functional
- Typically, we assume that the mean is smooth
- "Raw" estimator is sample mean: $\bar{X}(t) = \frac{1}{n} \sum X_i(t)$
- A typical estimator of μ would be a smoothed version of $\bar{X}(t)$ (more on this later).

Summary measures for functional data

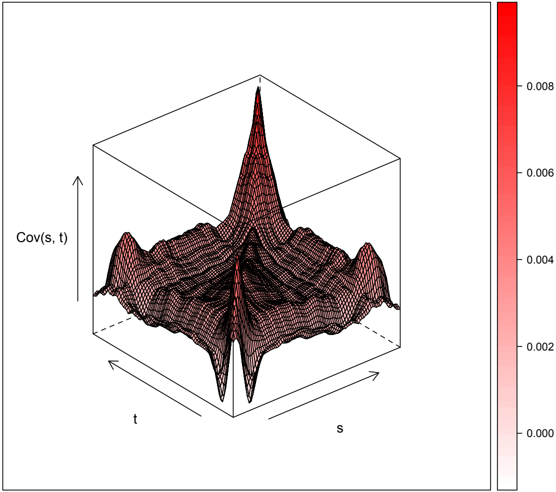
Suppose we have functional data $\{X_i(t),\, t \in \mathcal{T},\, i = 1, \ldots, n\}$.
Variance:
$\Sigma(s,t) = \mathrm{Cov}(X(s), X(t)) = E\left[(X(s) - \mu(s))(X(t) - \mu(t))\right]$

- This is a (two-dimensional) *surface*.
- "Raw" estimator is sample covariance:
  $\hat{\Sigma}(s,t) = \mathrm{Cov}(X_i(s), X_i(t))$
- Would need to smooth this as well.

Summary measures for functional data



Summary measures for functional data



Beyond iid functional data

Although the iid case is quite common, other situations are possible:

- Multilevel functional data:
  - $\{X_{ij}(t),\, t \in \mathcal{T},\, i = 1, \ldots, n,\, j = 1, \ldots, J_i\}$
  - Example: repeated motions in gesture data
- Longitudinal functional data:
  - $\{X_{ij}(t, v_j),\, t \in \mathcal{T},\, i = 1, \ldots, n,\, j = 1, \ldots, J_i\}$
  - Example: DTI data (multiple clinical visits)

Common problems in functional data analysis

Some issues arise regularly in FDA

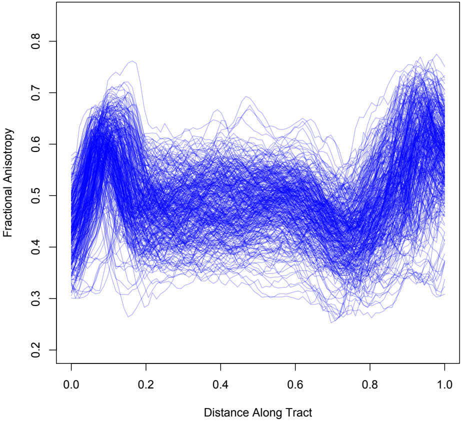
- Data display and summarization
- Smoothing and interpolation
- Patterns in variability: principal component analysis
- Regression (with functional predictors, outcomes, or both)

Data display

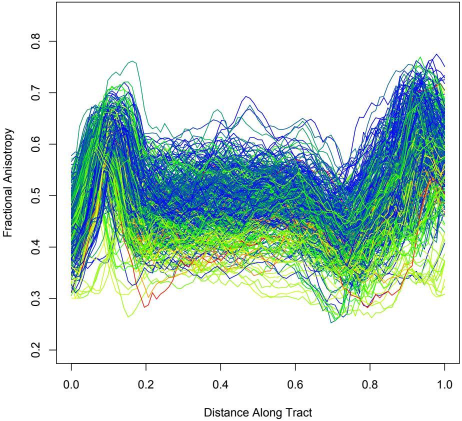
Lots of tools for displaying data

- Spaghetti plots
- Rainbow plots
- 3D rainbow plots
- Examples for all using DTI data follow; R code is available online

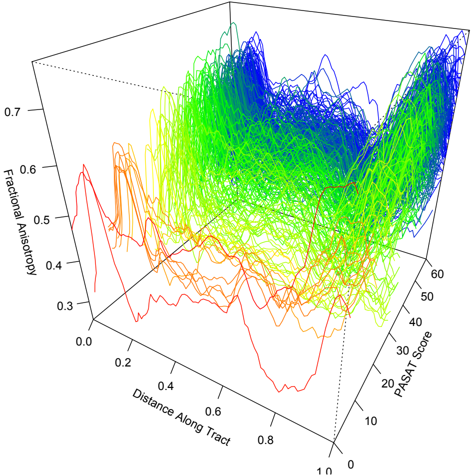
Spaghetti plot



2D rainbow plot



3D rainbow plot



Smoothing

Why do we need smoothing?

- Data are often observed with error
- There's a need to interpolate to a common grid

Smoothing

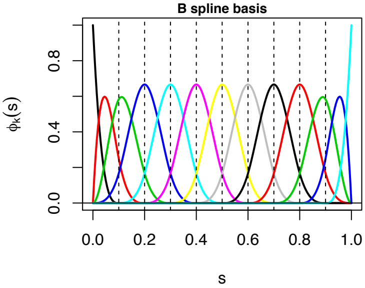
Why do we need smoothing?

- Data are often observed with error
- There's a need to interpolate to a common grid

How are we going to do smoothing?

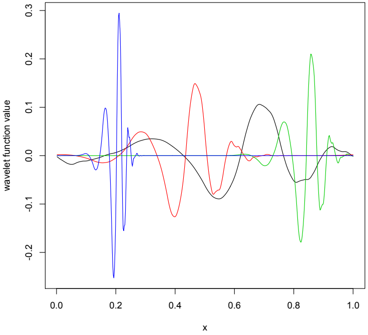
- Use a known set of basis functions
- Regress observed data onto known basis

Some common basis functions: Splines



- Continuous
- Easily defined derivatives
- Good for smooth data

Some common basis functions: Wavelets



- Formed from a single "mother wavelet" function: $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$
- Orthonormal basis
- Particularly good when there are jumps, spikes, peaks, etc.
- Wavelet representation is *sparse*

Minimize sum of squares

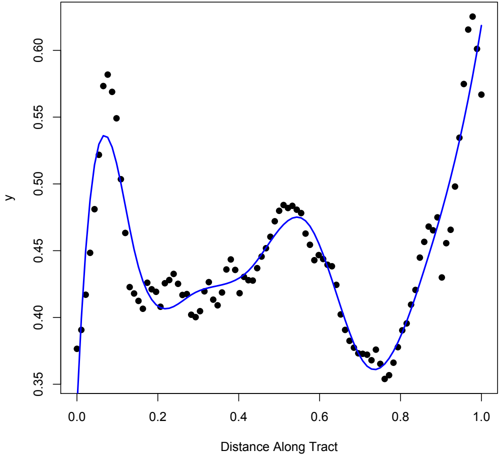
Suppose we want to smooth a curve $Y_i(t)$ observed with error. We can use

$$\hat{Y}_i(t) = \sum_{k=1}^{K} \hat{c}_{ik}\psi_k(t).$$

We only need to estimate the subject-specific scores $\hat{c}_{ik}$; minimize SSE_i with respect to c_{ik}, where

$$SSE_i = \sum \left(Y_i(t_i) - \sum_{k=1}^{K} c_{ik}\psi_k(t_i) \right)^2$$

Example



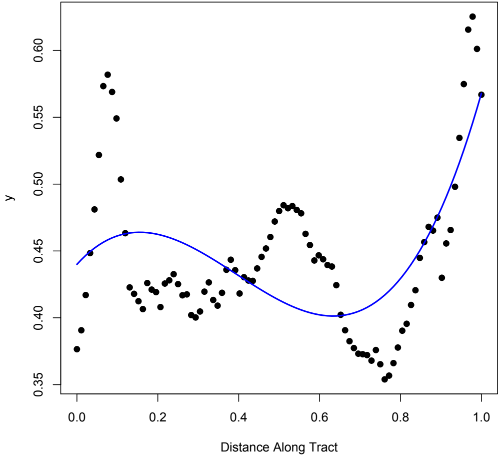
Tuning

For any curve, many possible smooths are available

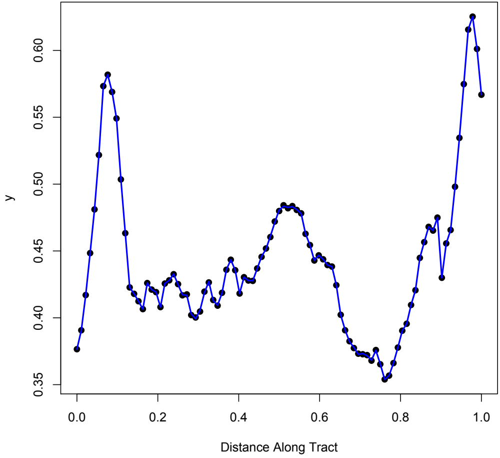
- Depends on the spline basis
- Depends on the number of basis functions
- Depends on the estimation procedure

"Tuning" is the process of adjusting the smoother to the data at hand. This is often implicit.

Example



Example



Penalization

Rather than choosing a smoother "by hand", we could use a lot of basis functions but *explicitly* penalize "wiggliness"

Leads to a penalized SSE:

$$SSE_i = \sum (Y_i(t) - \Psi(t)c_i)^2 + \lambda \mathrm{Pen}(\Psi(t)c_i)$$

- Common penalties are on the derivatives (enforcing smoothness)
- Need to choose *tuning parameter* λ

Data-driven basis

- Previous bases don't depend on the data; only the loadings do.
- FPCA gives a "data-driven" basis: it is constructed from the observed data.
- Looks pretty similar mathematically:

$$\hat{Y}_i(t) = \sum_{k=1}^{K} \hat{c}_{ik} \psi_k(t).$$

- Difference is that the ψ aren't pre-specified.

Data-driven basis

So where do the basis functions ψ come from?

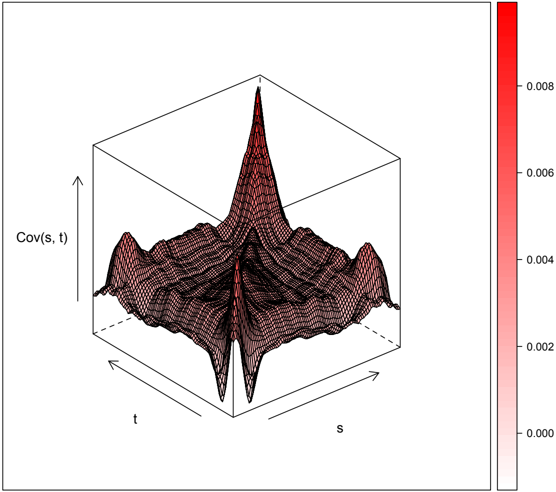
- Construct covariance matrix Σ
- (Remove main diagonal, smooth)
- Spectral decomposition of Σ produces basis functions ψ

Data-driven basis

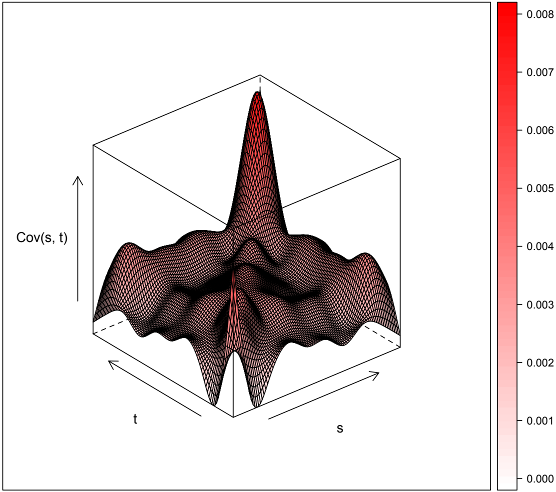
Some properties of FPCA

- The ψ are orthonormal (non-overlapping)
- Also the most parsimonious basis expansion for a given data set
- Basis functions are often interpretable - describe the major directions of variability in the observed data

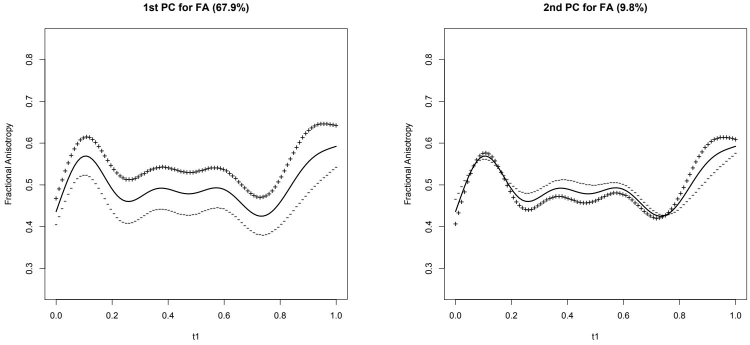
Example



Example



Example



1st PC for FA (67.9%)
2nd PC for FA (9.8%)
Fractional Anisotropy
t1
0.3
0.4
0.5
0.6
0.7
0.8
0.0
0.2
0.4
0.6
0.8
1.0

Data-driven vs Pre-specified

- Data-driven bases are the most parsimonious for a given dataset, but may not transfer to new data
- Data-driven often work better for sparse data (borrowing strength to derive basis functions)
- Pre-specified often have better analytical properties (easily computed derivatives, known forms)

Regression modeling with functional data

- Scalar on function regression
- Function on scalar regression
- Function on function regression

Scalar on function regression: Example scenarios

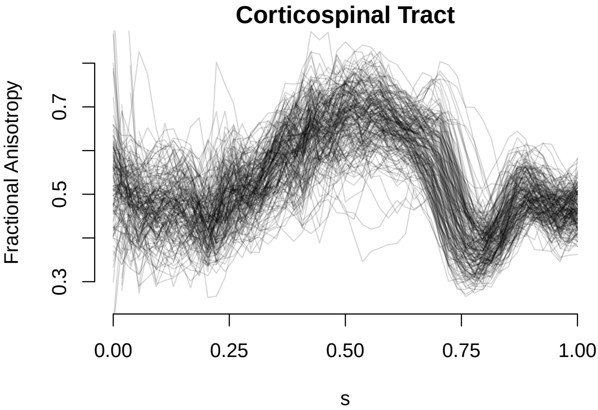
$X =$ temperature (over time) for the year
$Y =$ total rainfall for one year

$X =$ NIR spectrum
$Y =$ water content of a sample

$X =$ brain image
$Y =$ clinical outcome

Example data: DTI

$x_i(s) =$ fractional anisotropy along the corticospinal tract
$Y_i =$ measure of cognitive function



Linear scalar-on-function regression model

Given data $(\{x_1(s), s \in \mathcal{S}\}, Y_1), \ldots, \{x_n(s), s \in \mathcal{S}\}, Y_n)$, the scalar-on-function regression model is:

$$Y_i = \alpha + \int x_i(s)\beta(s)\,ds + \epsilon_i,\ i = 1, \ldots, n$$

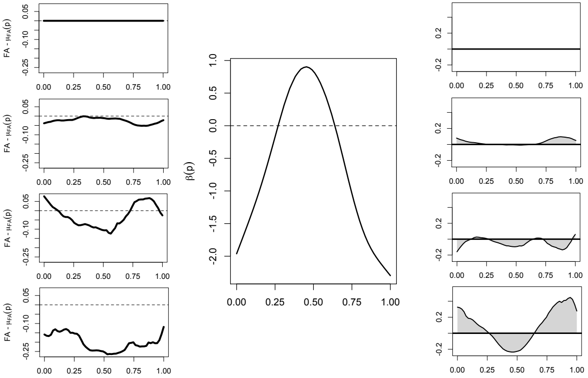
Interpretation of "coefficient function" β:

- Where $\beta(s) > 0$, larger values of $x_i(s)$ lead to higher predicted Y.
- Where $\beta(s) < 0$, larger values of $x_i(s)$ lead to lower predicted Y.
- Where $\beta(s) = 0$, $x_i(s)$ has no effect on Y.

Coefficient Interpretation

$$X_i(s) \quad \longrightarrow \quad \beta(s) \quad \longrightarrow \quad X_i(s)\beta(s) \quad \longrightarrow \quad \int X_i(s)\beta(s)\,ds$$

Observed Tract Profile | Coefficient Function | Profile x Coefficient (Area Under Curve Shaded) | Functional Contribution



0

1.6

-2.6

5.2

Scalar-on-function regression: The need for regularization

But the function $x_i(s)$ is only observed at N points!

- $\boldsymbol{x}_i = (x_i(1/N), x_i(2/N), \ldots, x_i(1))^T$
- $\boldsymbol{\beta} = (\beta(1/N), \beta(2/N), \ldots, \beta(1))^T$

The model becomes

$$\begin{aligned} Y_i &= \alpha + \int x_i(s)\beta(s)\, ds + \epsilon_i \\ &\approx \alpha + (1/N)\boldsymbol{x}^T\boldsymbol{\beta} + \epsilon_i \end{aligned}$$

If we're not thinking "functionally", this is like doing regression with n observations and N predictors!

To get reasonable fits, we must regularize in some way.

Basis functions

Possible basis functions: splines, orthogonal polynomials, principal components, wavelets, etc.

Let

$$\begin{aligned} x_i(s) &= \sum_{k=1}^{K} c_{ik}\psi_k(s) \\ \beta(s) &= \sum_{k=1}^{K} \theta_k\psi_k(s) \end{aligned}$$

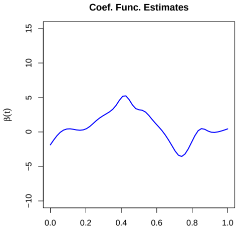
This is now a K-dimensional regression problem.

Scalar-on-function regression: Basis function representation

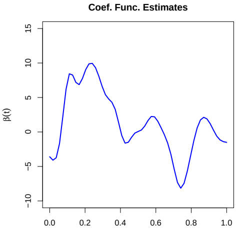
$$
\begin{aligned}
Y_i &= \alpha + \int x_i(s)\beta(s)\,ds + \epsilon_i \\
&= \alpha + \int \left(\sum_{\ell=1}^{K} c_{i\ell}\psi_\ell(s)\right)\left(\sum_{k=1}^{K} \theta_k\psi_k(s)\right) ds + \epsilon_i \\
&= \alpha + \sum_{k=1}^{K}\left[\sum_{\ell=1}^{K} c_{i\ell}\left(\int \psi_\ell(s)\psi_k(s)\,ds\right)\right]\theta_k + \epsilon_i \\
&= \sum_{k=1}^{K} z_k\theta_k + \epsilon_i
\end{aligned}
$$

How to choose K?

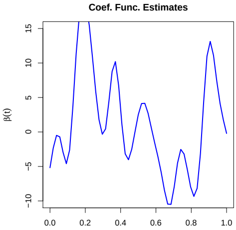
$K = 2$



$K = 7$



$K = 10$



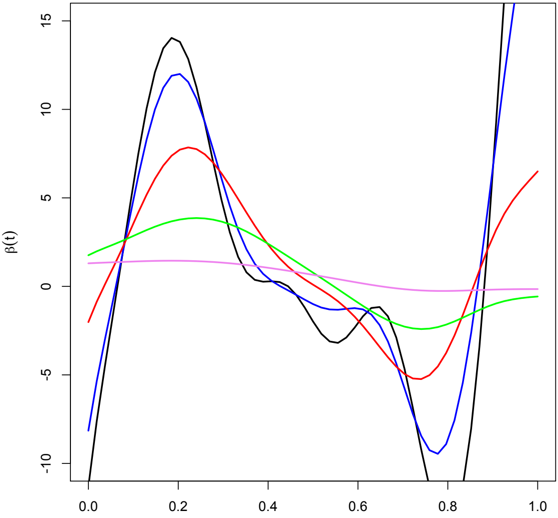
Regularization with roughness penalties

Could choose α and β to minimize

$$\sum_{i=1}^{n}\left(Y_i - \alpha - \int x_i(s)\beta(s)\,dt\right)^2 + \lambda \int \left(\beta''(s)\right)^2\,dt$$

- First term: (proportional to) mean squared error (MSE): measures fidelity to the data (how well the model “fits” the data)
- Second term: measures the smoothness of the coefficient function

Example fits with a range of tuning parameters



How to choose λ?

The tuning parameter λ controls the tradeoff between these.

- If λ is too large, it will result in smooth estimates at the expense of large MSE (underfitting).
- If λ is too small, the MSE will be small but the estimated β function will be very wiggly (overfitting).
- Neither one of these will provide good "out of sample" predictions.

Could choose λ by cross-validation:

$$CV(\lambda) = \sum_{i=1}^{n} \left(Y_i - \alpha_\lambda^{(i)} - \int x_i(t) \beta_\lambda^{(i)}(t)\, dt \right)^2$$

Choose λ to minimize $CV(\lambda)$

Also: generalized cross-validation (GCV), restricted maximum likelihood (REML) …

Function on scalar regression: Example scenarios

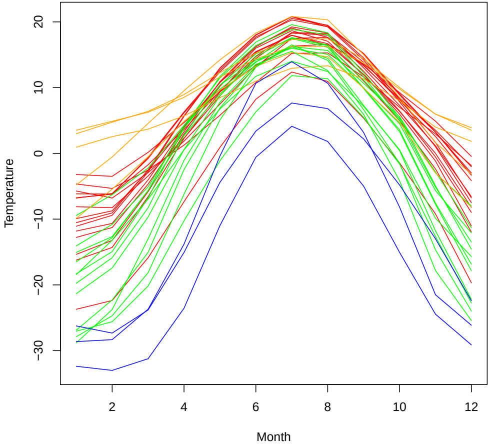
$X =$ climate zone
$Y =$ temperature (over time)

$X =$ age
$Y =$ activity level (over time)

$X =$ diagnosis
$Y =$ brain image

Canadian weather data

X = region (Arctic, Atlantic, Continental, Pacific)
Y = temperature (degrees Celsius) over time



Function on scalar regression

A "functional ANOVA" model:

$$Y_{ij}(s) = \mu(s) + \alpha_i(s) + \epsilon_{ij}(s), \; i = 1, \ldots, n$$

For identifiability, could constrain that $\sum_i \alpha_i(s) = 0$ for all t.

More generally, given data $(\boldsymbol{x}_1, \{Y_1(s), s \in \mathcal{S}\}), \ldots, (\boldsymbol{x}_n, \{Y_n(s), s \in \mathcal{S}\})$, where $\boldsymbol{x}_i$ is a p-vector, the function-on-scalar regression model is

$$Y_i(s) = \boldsymbol{x}_i^T \boldsymbol{\beta}(s) + \epsilon_i(s),$$

where $\boldsymbol{\beta}(s) = (\beta_1(s), \ldots, \beta_p(s))$.

Function on scalar regression: data representation

If the functional observations are observed at a grid of points, say, $s_1, \ldots, s_N$, then let

$$Y : n \times N = [Y_i(s_j)],\ i = 1. \ldots, n, j = 1, \ldots, N.$$

We could also think about expressing the β functions on the same grid, i.e., let

$$B : p \times N = [\beta_i(s_j)],\ i = 1, \ldots, p; j = 1, \ldots, N.$$

Expressing the ϵ's the same way and writing the X matrix as usual, the discrete version of the model becomes

$$Y = XB + E.$$

This has the same form as multivariate analysis of variance (MANOVA).

Function on scalar regression: basis function representation

Given basis functions $\psi_1(s), \ldots, \psi_K(s)$, we could express

$$\begin{aligned} Y_i(s) &= \sum_{k=1}^{K} c_{ik}\psi_k(s) \\ \beta_j(s) &= \sum_{k=1}^{K} \theta_{jk}\psi_k(s) \end{aligned}$$

The model then becomes

$$C = X\Theta + E$$

Fitting by penalizing roughness

Could choose $\boldsymbol{\beta}$ to minimize

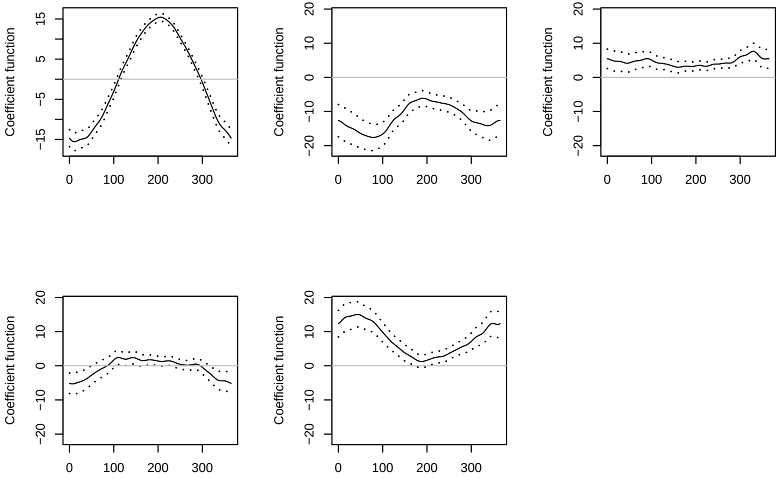
$$\sum_{i=1}^{n} \int \left(Y_i(s) - \boldsymbol{x}_i^T \boldsymbol{\beta}(s) \right)^2 dt + \lambda \sum_{j=1}^{p} \int \left(\beta_j''(s) \right)^2 dt$$

More generally, in the discretized space, we could minimize

$$||Y - XB|| + \lambda \sum_{j=1}^{p} B_j^T P B_j,$$

where B_j is the jth row of B.

Application to Canadian weather data



Function on function regression: Example scenarios

X = temperature (over time)
Y = precipitation (over time)

X = fractional anisotropy along corpus callosum tract
Y = fractional anisotropy along corticospinal tract

X = hip angle through a gait cycle
Y = knee angle through a gait cycle

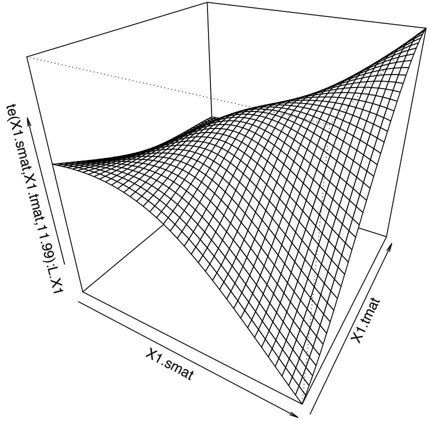
Function on function regression: the model

Given functional data $(\{x_1(s), s \in \mathcal{S}\}, \{Y_1(t), t \in \mathcal{T}\}), \ldots, (\{x_n(s), s \in \mathcal{S}\}, \{Y_n(t), t \in \mathcal{T}\})$, the model could be expressed

$$Y_i(t) = \int \beta(s,t) x_i(s)\, ds + \epsilon_i(t)$$

The coefficient function in this case is a (two-dimensional) surface.

Function on function regression: Example



Software

- `refund` package
- `fda` package
- `fda.usc` package

Stuff we haven't even mentioned

- Inference on functional model parameters
- Model selection, model building
- Alternative penalties
- Model diagnostics and goodness of fit
- "Generalized" versions of functional linear models
- Hierarchical models for functional data
- Supervised/unsupervised classification of functional data
- Functional "depth" and functional boxplots
- Many other topics …

Useful references

- Ferraty and Vieu (2006). *Nonparametric Functional Data Analysis*. Springer.
- Ramsay and Silverman (2005). *Functional Data Analysis, Second Edition*. Springer.
- Ramsay and Silverman (2002). *Appled Functional Data Analysis*. Springer.
- Sørensen, Goldsmith, and Sangalli (2013). An introduction with medical applications to functional data analysis. *Statistics in Medicine* **32**:5222-5240.