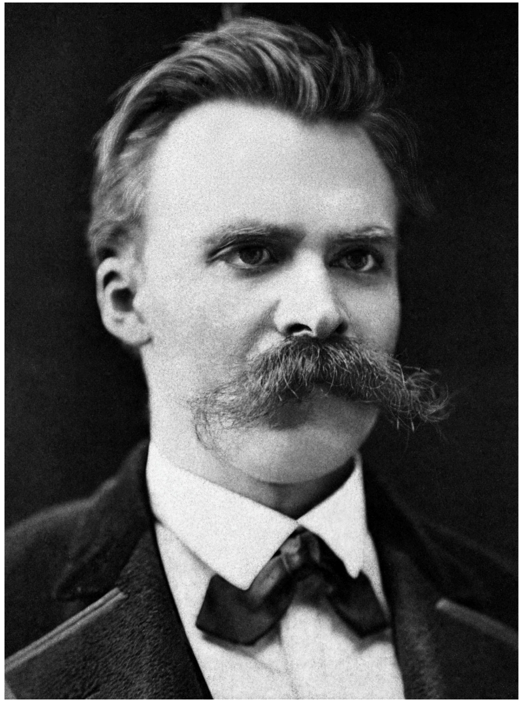
Functional Data Analysis

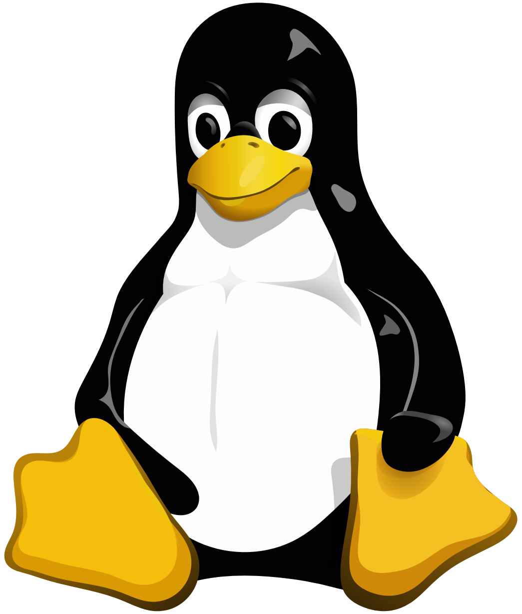
EAI Math Reading Group

25/08/2024

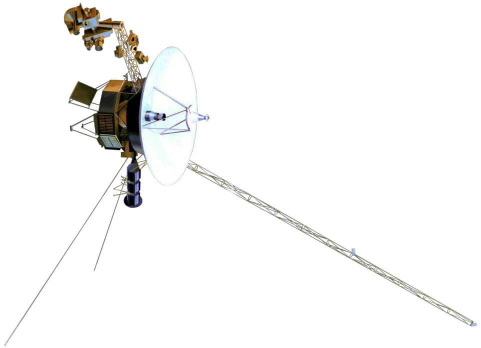
Notable anniversaries



(1900) Death of Friedrich Nietzsche



(1991) First announcement of Linux



(2012) Voyager 1 exits solar system

https://en.wikipedia.org/wiki/August_25

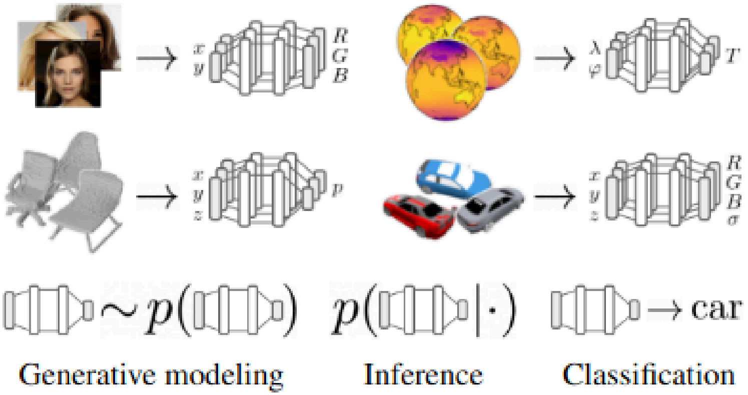
Motivation

From data to functa: Your data point is a function and you can treat it like one

Emilien Dupont [*1] **Hyunjik Kim** [*2] **S. M. Ali Eslami** [2] **Danilo Rezende** [2] **Dan Rosenbaum** [3 2]

Abstract

It is common practice in deep learning to represent a measurement of the world on a discrete grid, e.g. a 2D grid of pixels. However, the underlying signal represented by these measurements is often continuous, e.g. the scene depicted in an image. A powerful continuous alternative is then to represent these measurements using an *implicit neural representation*, a neural function trained to output the appropriate measurement value for any input



Outline

1. Functional Data
2. Basic Functional Statistics
3. Functional PCA and Regression

FDA software

Sadly, most of the existing software for FDA is in R

- https://cran.r-project.org/web/views/FunctionalData.html

There is one Python package for it, though

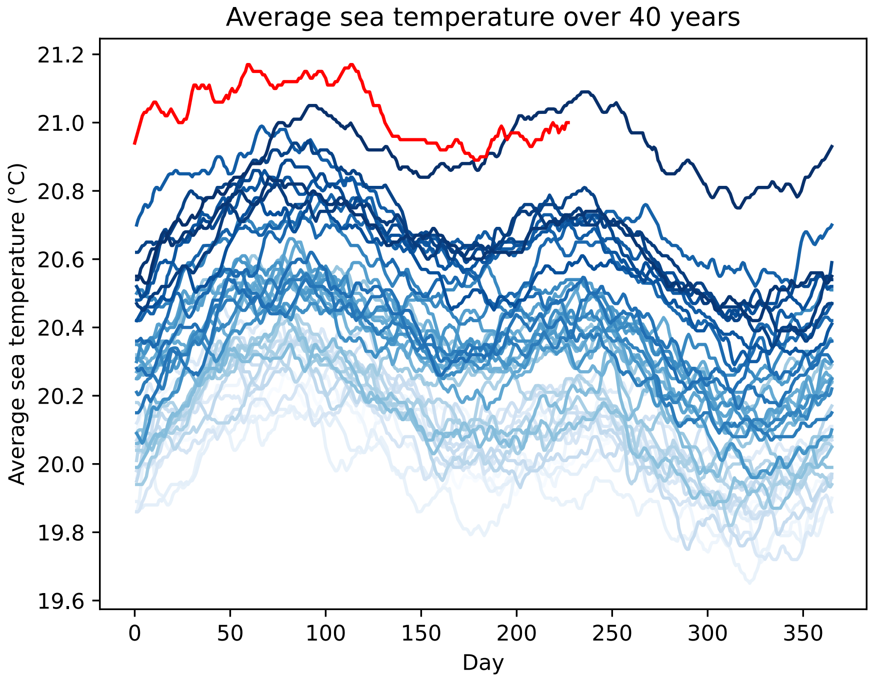
- https://fda.readthedocs.io/en/latest/index.html

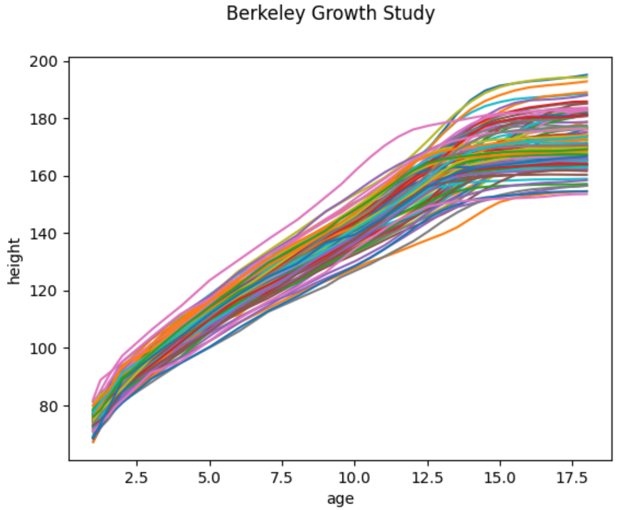
Functional Data and where to find it

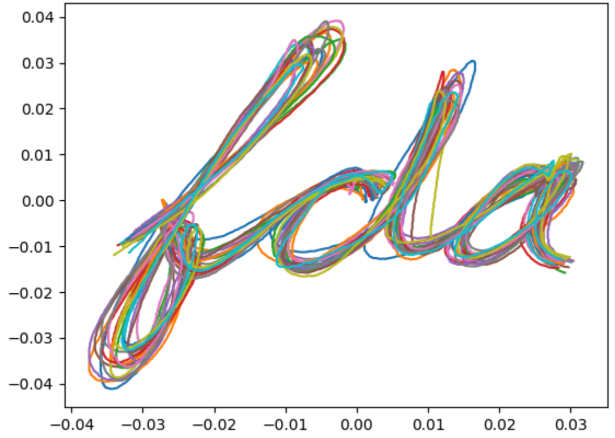
In many applications, data can be thought as a function $T \to \mathbb{R}^d$ on a *continuous* domain T.

Examples

- Time series $f : [0, 1] \to \mathbb{R}$ (e.g. Nvidia stocks)
- Spatial data $f : \mathbb{S}^2 \to \mathbb{R}$ (e.g. temperature at every point on earth)
- Images $f : [0, 1] \times [0, 1] \to \mathbb{R}^3$
- ODEs/PDEs


Average sea temperature over 40 years
Average sea temperature (°C)
21.2
21.0
20.8
20.6
20.4
20.2
20.0
19.8
19.6
0
50
100
150
200
250
300
350
Day


Berkeley Growth Study
height
200
180
160
140
120
100
80
2.5
5.0
7.5
10.0
12.5
15.0
17.5
age


0.04
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
−0.04
−0.03
−0.02
−0.01
0.00
0.01
0.02
0.03

Functional variables

A random element X is a *functional variable* if it takes values in a function space $\mathcal{F}$. It is denoted as $\{X(t) \mid t \in T\}$.

A *realization* of X will be denoted $x = \{x(t) \mid t \in T\}$

Examples of function spaces

- $C[0,1]$ (continuous functions on $[0,1]$)
- $L^2[0,1] = \left\{f : [0,1] \to \mathbb{R} \mid \int_0^1 f^2(t)dt < \infty\right\}$

Observing Functional Data

Problem We usually can't observe functional data directly

Instead we observe discrete samples $\{x(t_1\}, ..., x(t_n)\}$ at times $t_1 < ... < t_n$.

More problems

Data can be

- densely sampled ($t_k = k\Delta t$)
- sparsely sampled (arbitrary t_k)
- irregularly sampled (different sample times for each observation)
- noisy ($x(t_k) = X(t_k) + \varepsilon(t_k)$)

Why not just multivariate data?

It is tempting to just treat e.g. time series data as just high-dimensional multivariate observations.

This is a bad idea:

- Only works with densely, regularly sampled data
- Nearby datapoints are usually highly correlated
- Want to take “derivatives” of observations

Let functions be functions!

Function approximation

Given samples $\{y(t_1), ..., y(t_N)\}$ we want to obtain a curve x such that

$$y(t) = x(t) + \varepsilon(t)$$

Generic framework: solve the optimization problem

$$\min_{\theta} \|x_\theta - y\|,$$

where $\{x_\theta\}_\theta$ is a parametric family of functions, and $\|.\|$ is a norm (or seminorm)

Ideally, we want the parametrization to be easy to manipulate (e.g. in a vector space)

Basis function approach

Let $\mathcal{F}$ be a function space with norm $\|.\|_{\mathcal{F}}$. A set $\{\varphi_k\} \subset \mathcal{F}$ is called a *basis* of $\mathcal{F}$ if it is a set of linearly independent functions such that for any $f \in \mathcal{F}$, there exists constants $\{c_{k,K}\} \subset \mathbb{R}$ such that

$$\left\| f - \sum_{k=1}^{K} c_{k,K} \varphi_k \right\|_{\mathcal{F}} \longrightarrow 0$$

as $K \to \infty$.

Intuition: We can approach f arbitrarily well using a finite number of basis functions.

Examples of function bases: Polynomials

For $\mathscr{F} = C[0,1]$, with norm $\|f\|_\infty = \sup_{t\in[0,1]}|f(t)|$

- *Monomials* $\{t^k\}$ are a basis. (but they're terrible numerically)
- *Bernstein polynomials*

$$b_{k,K} = C_k^K t^k (1-t)^{K-k}$$

  are the preferred alternative (but numerically unstable for large K)

Examples of function bases: Fourier

For $\mathcal{F} = L^2[0,1]$, with L^2-norm, the Fourier basis

$$\varphi_0(t) = 1, \qquad \varphi_{2k}(t) = \sqrt{2}\cos(2\pi kt), \qquad \varphi_{2k+1}(t) = \sqrt{2}\sin(2\pi kt)$$

is an *orthonormal* basis.

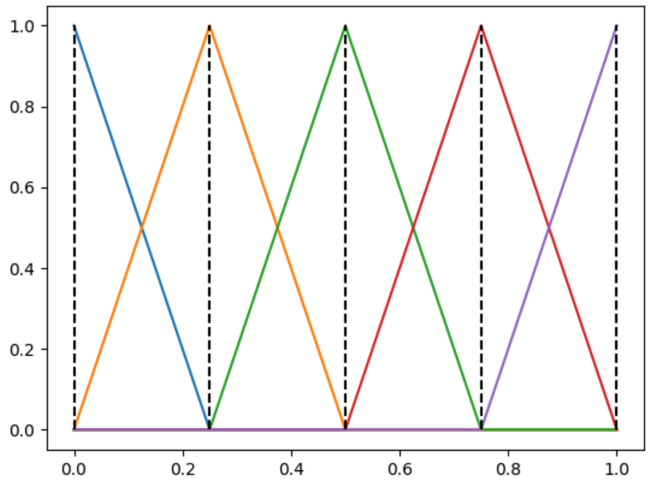
For regularly sampled data, it can be fitted *very* fast using FFT.

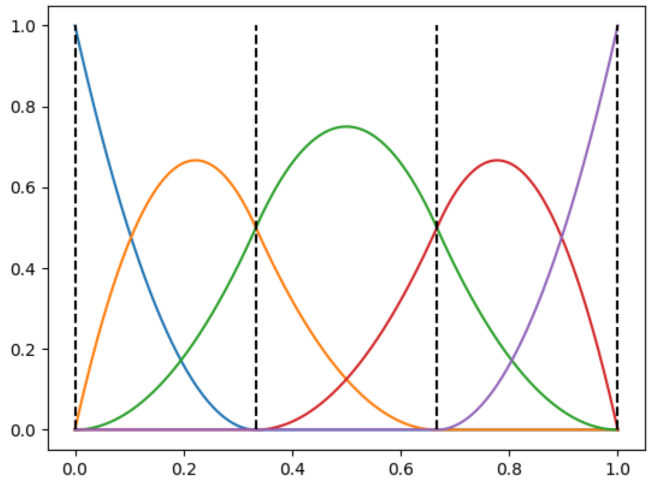
See also: orthonormal polynomials (Chebyshev, Lagrange, Hermite, ...)

B-splines

Idea: use local polynomial approximation over each interval $[t_k, t_{k+1}]$

Spline basis: $B_{k,d}(t) = \frac{t - t_k}{t_{k+d} - t_k} B_{k,d-1}(t) + \frac{t_{k+d+1} - t}{t_{k+d+1} - t_{k+1}} B_{k+1,d-1}(t)$, with $B_{k,0}(t) = 1$





Fitting splines

Suppose $y = x + \varepsilon$, the simplest way to fit a spline to the data y is to minimize

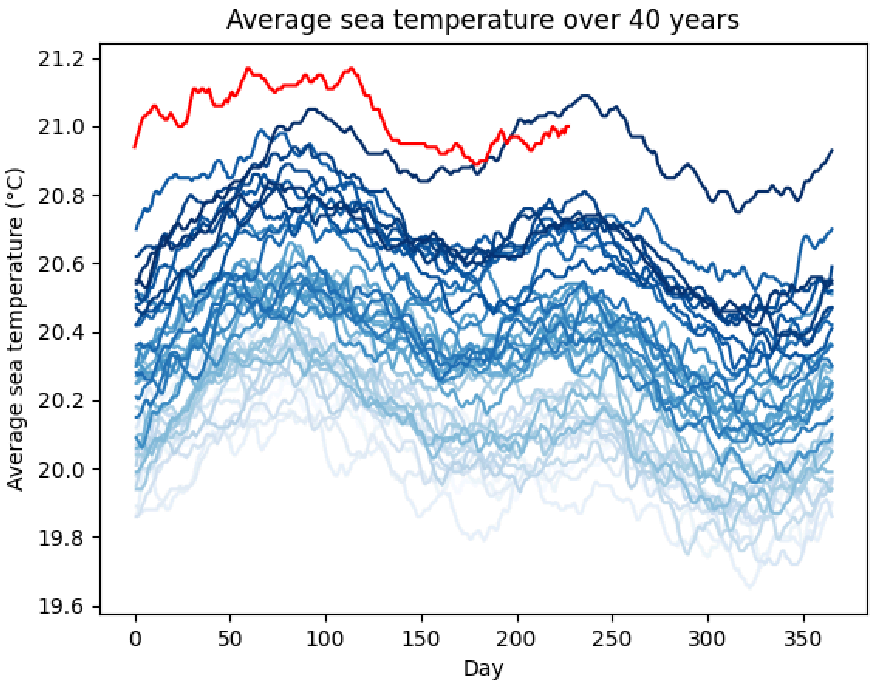
$$\mathrm{SSR} = \sum_{i=1}^{n} (y(t_i) - \hat{x}(t_i))^2$$

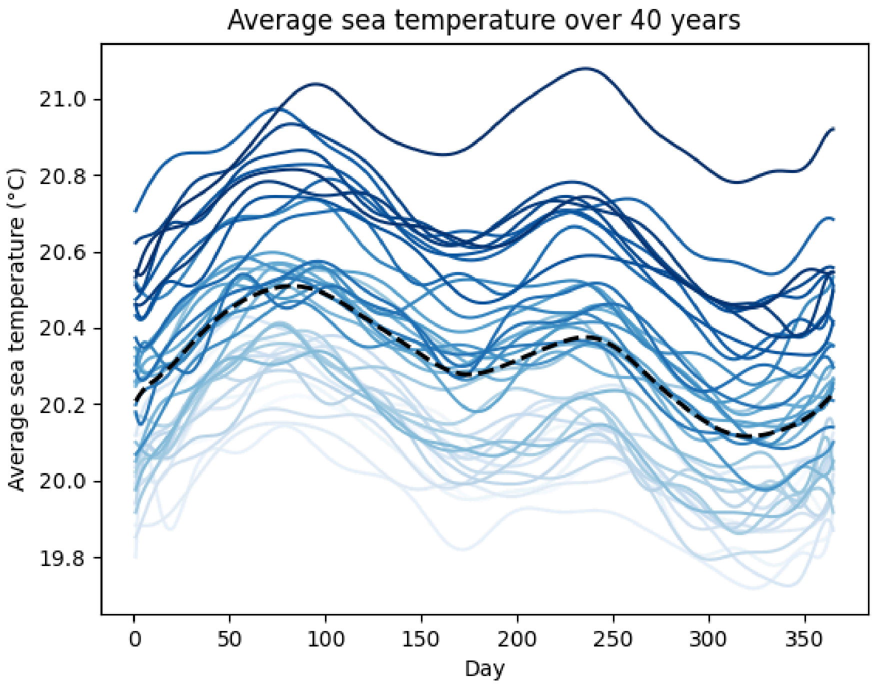
where $\hat{x} = \sum_{k=1}^{K} \alpha_k \varphi_k$, ($\{\varphi_k\}$ is the basis)

This is just a leasts square. Construct $\Phi = [\varphi_k(t_i)]_{ki}$ and $\boldsymbol{y} = [y(t_1) ... y(t_n)]'$, then

$$\Phi\alpha = \boldsymbol{y}$$

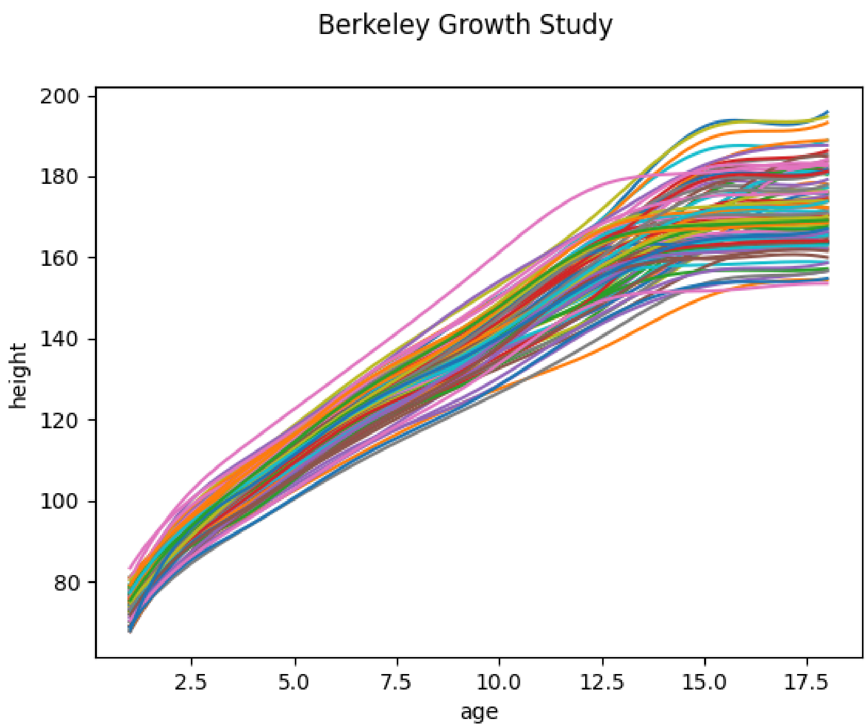
Example: Fitting climate time series data

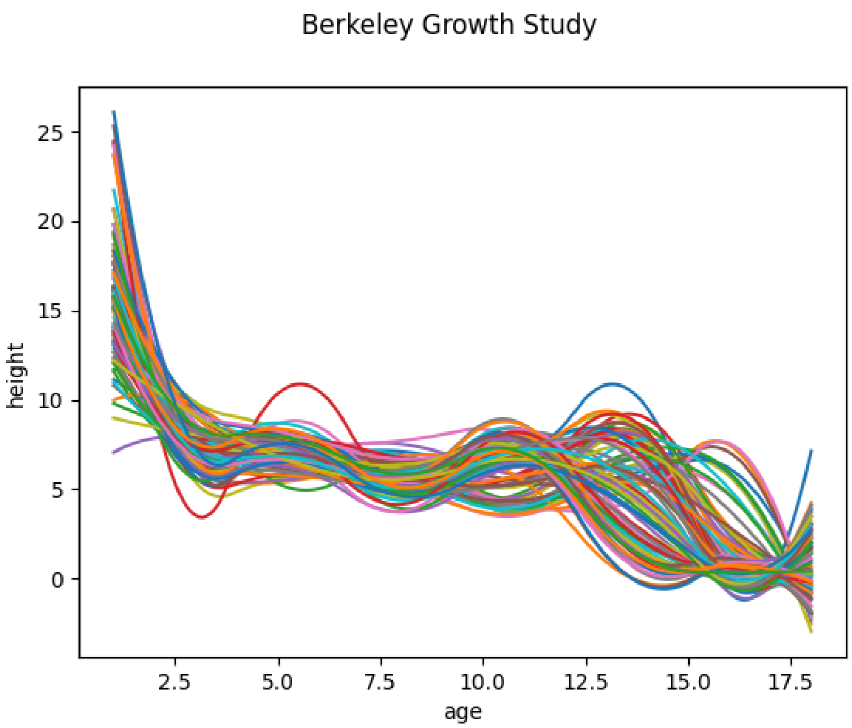




https://climatereanalyzer.org/

Example: Children growth dataset





Basic Functional Statistics

Functional Analysis setup

We generally prefer to work with a *Hilbert space* i.e. a vector space of functions H with an *inner product* $\langle , \rangle$ that satisfies

- $\langle x, y \rangle = \langle y, x \rangle$
- $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$
- $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \Leftrightarrow x = 0$

The inner product defines a norm $\|x\| = \sqrt{(\langle x, x \rangle)}$

Topological condition: H is *complete* (every Cauchy sequence is convergent)

Sample and population mean (in finite dimensions)

Let $X_1, X_2, \ldots$ be i.i.d variables in $\mathbb{R}^d$ In classical statistics, we have

- The sample mean $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ as an estimator for $\mu = E[X_1]$
- The sample covariance matrix

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{k=1}^{n}\left(X_k - \overline{X}_n\right)\left(X_k - \overline{X}_n\right)'$$

as an estimator for $\Sigma = E\left[(X_1 - \mu)(X_1 - \mu)'\right]$

Can we generalize this to functional variables? Intuitively, yes, but there are a couple of grisly mathematical details.

Population mean in Hilbert spaces

The random function X is *weakly integrable* if $\exists \mu \in H$ such that

$$E[\langle X, y \rangle] = \langle \mu, y \rangle \qquad \forall y \in H$$

Then μ is called the *expectation* of X.

X is said to be integrable ($X \in L^1_H$) if $E[\|X\|] < \infty$

Lemma:

- If $X \in L^1_H$, then X us weakly integrable
- If X is weakly integrable, then μ is unique

Example: In $L^2[0,1]$, $(E[X])(t) = E[X(t)]$ (almost everywhere)

Covariance operator

Let $X \in L^2_H$ (i.e. $E\left[\|X\|^2\right] < \infty$), then the linear operator $C : H \to H$

$$Cy = E[\langle X - E[X], y \rangle (X - E[X])]$$

is called the *covariance operator* of X.

NB: In $\mathbb{R}^d$,

$$Cy = E\left[(X - E[X])(X - E[X])' y\right] = E[(X - E[X])(X - E[X])]y = \Sigma y$$

Covariance kernel

Let $H = L^2[0,1]$, wlog let $E[X] = 0$, then

C is a kernel operator with kernel $c(t,s) = E[X(t)X(s)]$, i.e.

$$(Cy)(t) = \int_0^1 c(t,s)y(s)ds$$

Properties:

- $c(t,s)$ is symmetric $\Rightarrow C$ is a symmetric operator
- C is positive semidefinite: $\langle Cy, y\rangle \geq 0, \forall y \in H$
- C is a compact operator

$\Rightarrow$ *Many* very nice properties, including an orthonormal basis of eigenfunctions.

Covariance kernel

Let $H = L^2[0,1]$, wlog let $E[X] = 0$, then

C is a kernel operator with kernel $c(t,s) = E[X(t)X(s)]$, i.e.

$$(Cy)(t) = \int_0^1 c(t,s)y(s)ds$$

Properties:

- $c(t,s)$ is symmetric $\Rightarrow C$ is a symmetric operator
- C is positive semidefinite: $\langle Cy, y\rangle \geq 0, \forall y \in H$
- C is a compact operator

$\Rightarrow$ *Many* very nice properties, including an orthonormal basis of eigenfunctions.

Limit theorems (multivariate case)

Let X_k be an i.i.d sequence of random variables in $\mathbb{R}^d$, such that $E[\![|X_k\|]\!] < \infty$, then

- $$\frac{1}{n}\sum_{k=1}^{n} X_k \underset{n\to\infty}{\longrightarrow} E[X_1]$$
- If also $E[\![|X_k\|^2]\!] < \infty$, then

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(X_k - E[X_k]) \xrightarrow{\mathcal{D}} \mathcal{N}(0, C),$$

where $C = \mathrm{Cov}(X_1)$.

Limit Theorems (Functional case)

Let X_k be an i.i.d sequence with values in a (separable) Hilbert space H, such that $X_1 \in L^1_H$, then

- $$\left\| \frac{1}{n} \sum_{k=1}^{n} X_k - E[X_1] \right\| \to 0 \qquad \text{almost surely}$$
- If also $E\left[\|X_1\|^2\right] < \infty$, then

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} (X_k - E[X_k]) \xrightarrow{\mathcal{D}} \mathcal{N}(0, C)$$

where $\mathcal{N}(0, C)$ is a *gaussian element* in H with covariance operator C, the covariance operator of X_1.

Functional PCA and Functional Regression

PCA in multivariate stats

The covariance matrix C of X taking values in $\mathbb{R}^d$ (wlog $E[X]=0$) is a symmetric positive semidefinite matrix and thus has positive eigenvalues $\lambda_1 \geq ... \geq \lambda_d \geq 0$ and orthonormal eigenvectors $v_1, ..., v_d$.

The inner products $Y_i = \langle X, v_i \rangle$ define the *Principal Component scores* and the truncated projection

$$Y^{[k]} = \sum_{i=1}^{k} Y_i v_i$$

maximizes variance of $Y = AX$ over all linear maps $A : \mathbb{R}^d \to \mathbb{R}^k$

Other interpretation: $\{v_i\}$ is the “optimal basis” for X.

Functional Principal Components

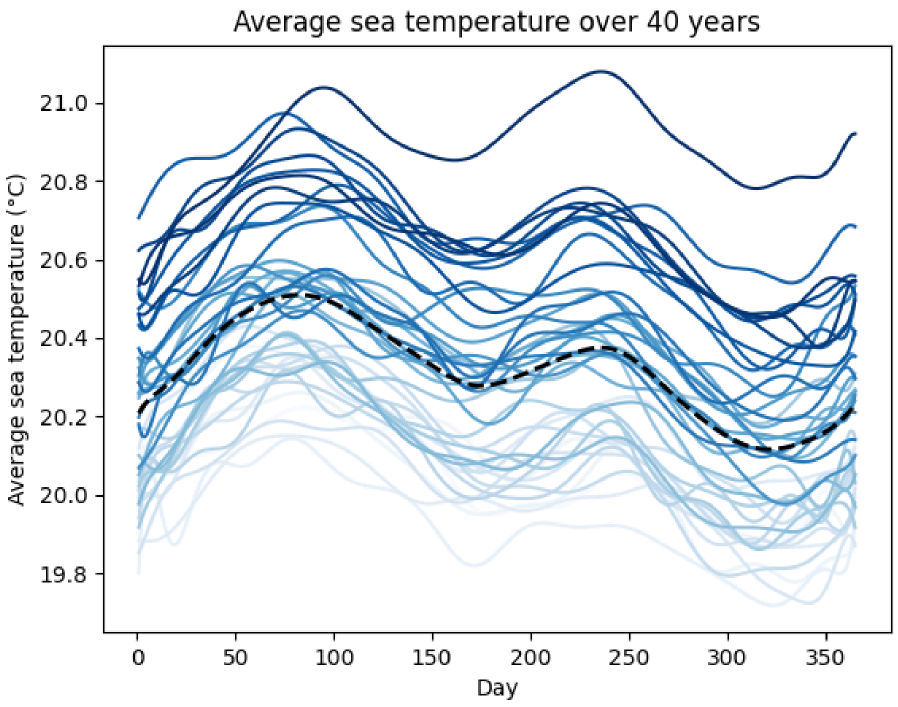
Let H be a (separable) Hilbert space and let $X \in L^2_H$, with $E[X] = 0$. Let C be the covariance operator of X and suppose it has eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$, with eigenfunctions $v_1, v_2, \ldots$, then

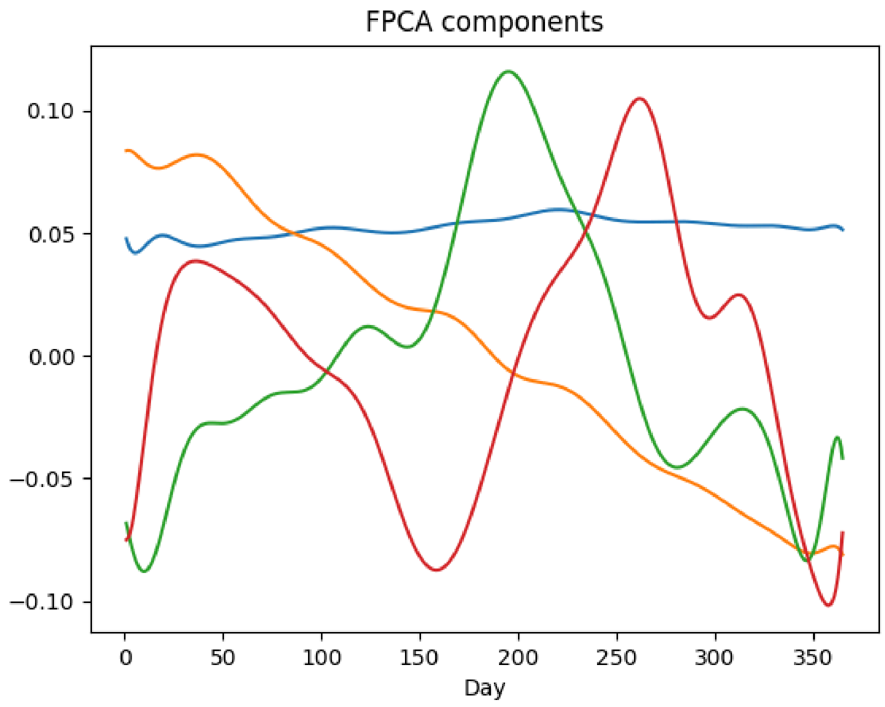
$$Y_i = \langle X, v_i \rangle$$

is the i-th *functional principal score*, and v_i is its corresponding *functional principal component.*

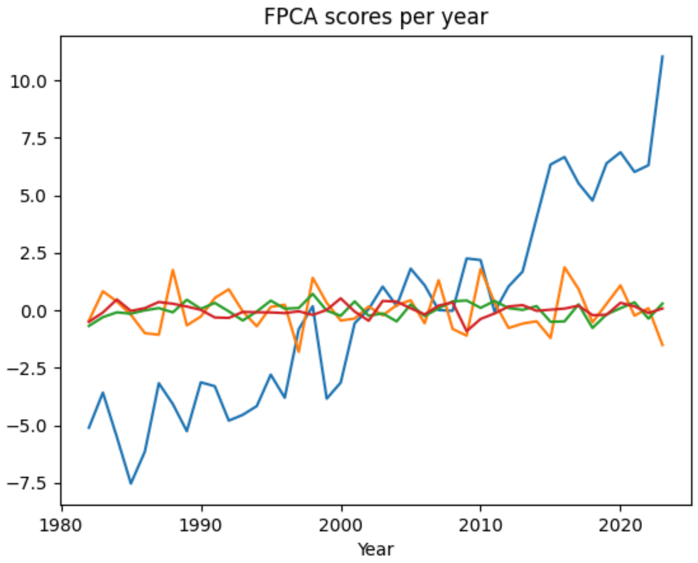
In practice: we only have the sample covariance $\hat{C}y = \frac{1}{n} \sum_{\{k=1\}}^{n} \langle X_k, y \rangle X_k$, and can thus only estimate up to n eigenpairs.

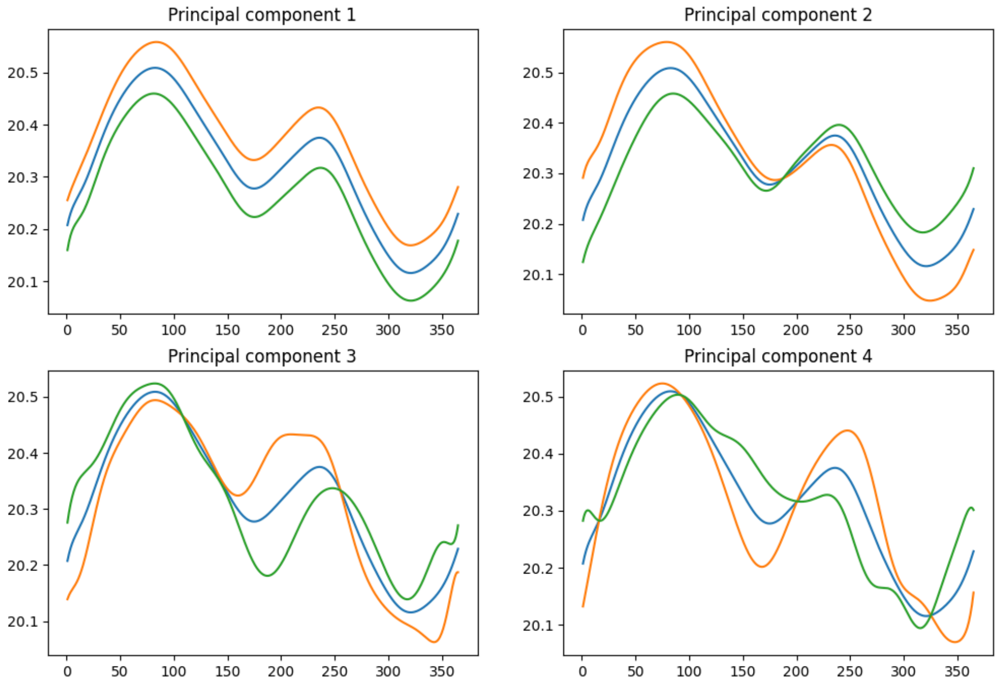
Example: FPCA on climate data





Example: FPCA on climate data





Functional Regression (at a glance)

We want to do simple data analysis using functional data. In particular, linear regression.

Different flavours

- Function on scalar

$$Y_{kg}(t) = \mu(t) + \alpha_g(t) + \varepsilon_{kg}(t)$$

- Scalar on function

$$Y_k = \alpha + \int_T X_k \beta(t) dt + \varepsilon_k$$

Function on function

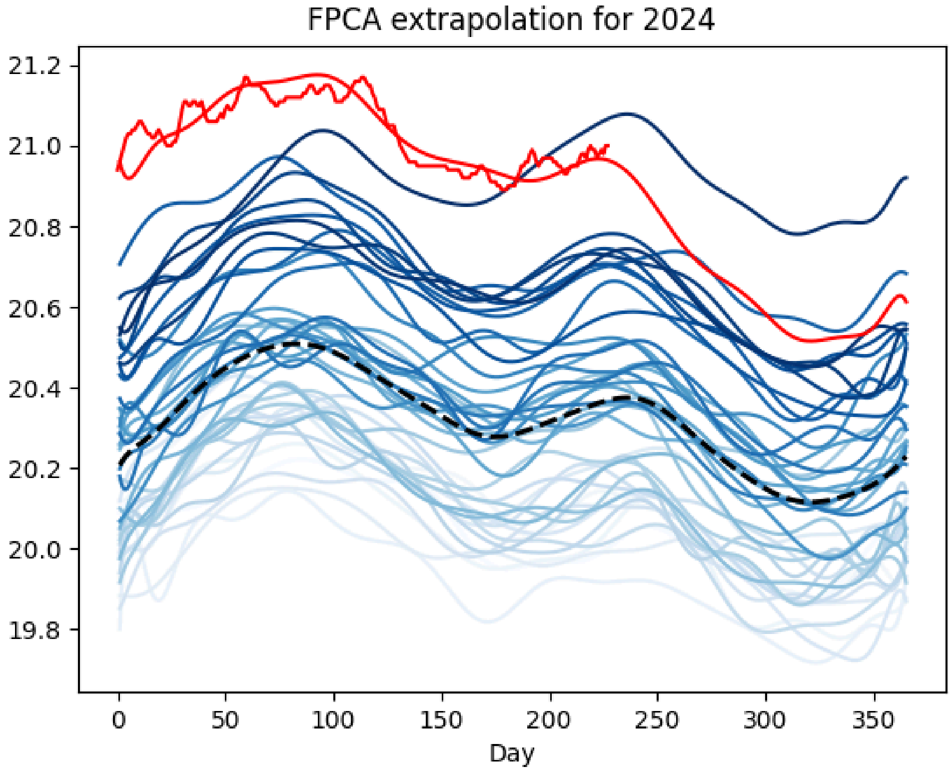
$$Y_k(t) = \alpha(t) + \int_T \beta(t, s) X_k(s) ds + \varepsilon_k(t)$$

Example

Given partially observed data $\{x(t_i) \mid 1 \leq t_i \leq t_M\}$, we want to predict future measurements using PCA components

$$\min \left\| x - \mu - \sum_{k=1}^{K} \alpha_k \varphi_k \right\|_M^2$$

where $\|f\|_M^2 = \int_0^M f(t)^2 dt$



Other ML approaches

Beyond linear models, there are other ML techniques we can apply to Functional data

- Vector space representation $\rightarrow$ Use basis coefficients as vector features (e.g. Fourier Neural Operators)
- Metric space $\rightarrow$ can use methods like k-NN, k-means, ...

In general, however, FDA techniques are not especially GPU friendly :(

Most of the value is in the basis function representation/FPCA as features.

Conclusion

- Statistics in infinite dimensions!
- Somewhat versatile data transform (but limited as domain dimension increases)
- Can still do many things we know from finite dimensions (PCA wins again)