Lecture 3. FDA Basic - Further Topics

Functional Data Analysis

Jun Song

Department of Statistics
Korea University

Outline

We begin Chapter 2 - Further Topics.

- Derivatives
- Penalized Smoothing
- Curve Alignment

Derivatives

The ability to examine derivatives is another distinction of FDA (as compared to MVA). Once a function, $X_n(t)$ is represented using a basis, calculating its derivative is relatively straightforward.

$$X_n(t) = \sum_{k=1}^{K} c_{nk} B_k(t) \Longrightarrow X_n'(t) = \sum_{k} c_{nk} B_k'(t).$$

Example - Matérn

The Matérn process is a Gaussian process with a covariance function defined as

$$C(t,s) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|t-s|}{\rho} \right) K_\nu \left(\sqrt{2\nu} \frac{|t-s|}{\rho} \right)$$

- σ^2 : Point-wise variance.
- ρ: Range parameter, controls how quickly dependence falls off.
- ν: Smoothness parameter, controls the number of derivatives of the resulting process ($< \nu - 1$).
- K_ν: Bessel function of the second kind.

A good process for simulating "realistic" data. Popular in spatial statistics.
A Gaussian process with Matérn covariance is $\lceil \nu \rceil - 1$ times differentiable in the mean-square sense

Example - Matérn

A few examples.

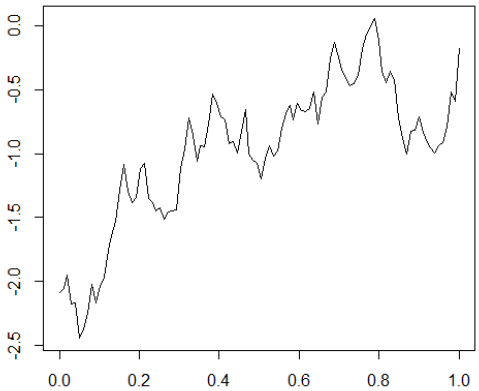
- If $\nu = 1/2$ then
$$C(t,s) = \sigma^2 \exp\{-|t-s|/\rho\}.$$
- If $\nu = 3/2$ then
$$C(t,s) = \sigma^2 \left(1 + \sqrt{3}|t-s|/\rho\right) \exp\{-\sqrt{3}|t-s|/\rho\}.$$
- If $\nu = \infty$ then
$$C(t,s) = \sigma^2 \exp\{-|t-s|^2/(2\rho^2)\}$$

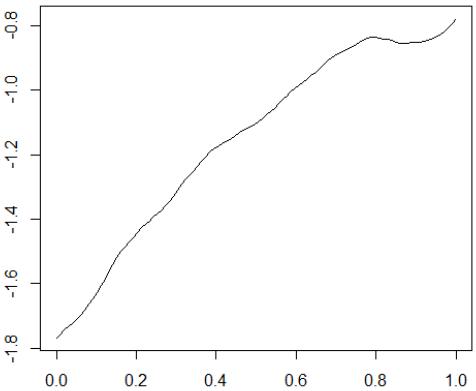
The top one has a special name, what is it?

Example - Matérn - Code

```
set.seed(201609); library(fields); library(expm)
m<-100; times<-seq(0,1,length=m)
range<-1; nu1=1/2; sig2<-1; nu2=3/2
Matern(.5,range=range,nu=nu1)

## [1] 0.6065307

d_mat<-abs(outer(times,times,"-"))
C_1<-apply(d_mat,c(1,2),FUN=Matern,range=range,nu=nu1)
C_1<-C_1*sig2
C_1_sq<-sqrtm(C_1)
C_2<-apply(d_mat,c(1,2),FUN=Matern,range=range,nu=nu2)
C_2<-C_2*sig2
C_2_sq<-sqrtm(C_2)
Z<-rnorm(m)
X1<-C_1_sq%*%Z; X2<-C_2_sq%*%Z
```

Example - Matérn - Plot

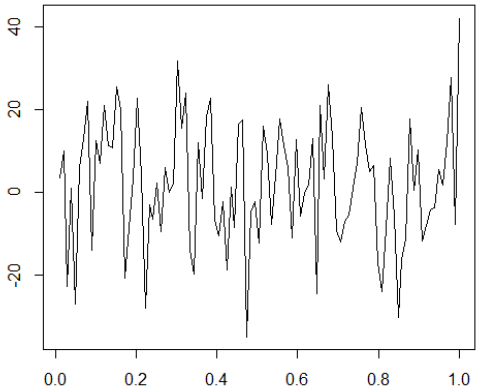


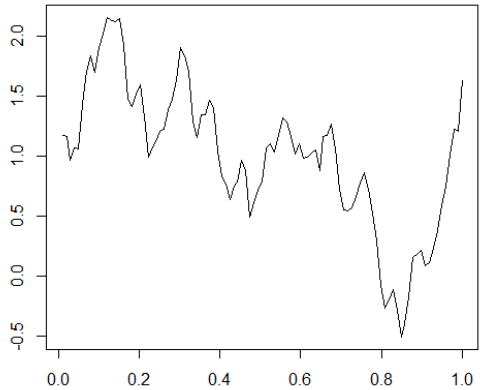


Example - Numeric Derivative

If we observe $\{X(t_i)\}$ then we can approximate its derivative:

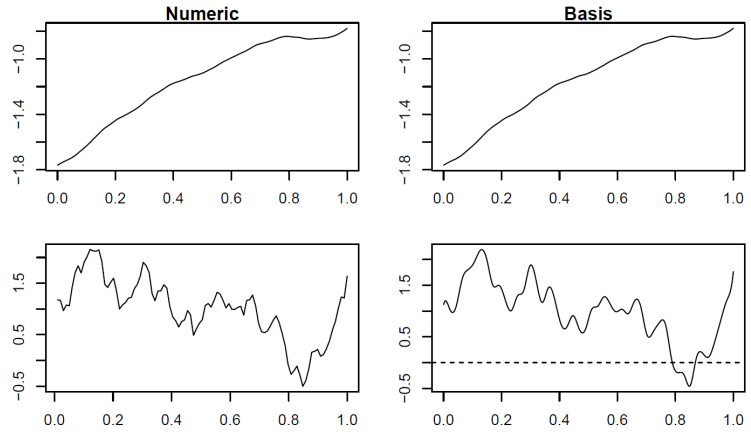
$$X'(t_i) \approx \frac{X(t_i) - X(t_{i-1})}{t_i - t_{i-1}}.$$





Example - Compare

```
mybasis<-create.bspline.basis(c(0,1),nbasis=50)
X2.f<-Data2fd(times,X2,mybasis)
X2d.f<-deriv.fd(X2.f)
```



Penalized smoothing

- Taking derivatives of functions must be done with care.
- If the data hasn't been properly smoothed then the functions can look funny or be misleading.
- We will now develop a strategy for more carefully smoothing the curves.
- As a by-product, the choice of the number of basis functions will no longer be too crucial (as long as it is relatively large).

Suppose that $\{Y_{jn}\}$ are the function values observed at time points t_{nj}. Our aim is to reconstruct $X_n(t)$, the underlying curve, using a penalized smoothing method.

Penalized smoothing

Fitting a basis expansion is basically done with least squares:

$$S(\mathbf{c}_n) = \sum_j \left(Y_{jn} - \sum_k c_{nk} B_k(t_{nj}) \right)^2 .$$

A penalized approach penalizes the resulting function to promote different properties (such as smoothness):

$$S_\lambda(\mathbf{c}_n) = \sum_j \left(Y_{jn} - \sum_k c_{nk} B_k(t_{nj}) \right)^2 + \lambda \int_0^1 [L(\tilde{X}_n)(t)]^2 \, dt,$$

where we use $\tilde{X}_n(t) = \sum_k c_{nk} B_k(t_{nj})$ for notational convenience. L is some specified linear differential operator (usually the second derivative).

Penalized smoothing - L

Consider the L is the second derivative.

$$L(x)(t) = x^{(2)}(t)$$

If $\lambda \to \infty$ it will be a curve with $x^{(2)}(t) = 0$ i.e. straight line

There is quite a bit of flexibility in choosing L. In general, one can specify it as

$$L(\cdot) := \sum_{m=0}^{M} \alpha_m D^{(m)}(\cdot).$$

Common examples are first and second derivatives. A popular choice for periodic data is the harmonic acceleration operator

$$L(x)(t) = \frac{4\pi^2}{T^2} x^{(1)}(t) + x^{(3)}(t).$$

Penalized smoothing - Basis

The solution to the penalized regression equation can be written down in a closed form. Notice that

$$L(\tilde{X}_n)(t) = \sum_k c_{nk} L(B_k)(t) \Longrightarrow \int_0^1 [L(\tilde{X}_n)(t)]^2 = \mathbf{c}_n^\top \mathbf{W} \mathbf{c}_n,$$

where

$$W_{ij} = \int_0^1 L(B_i)(t) L(B_j)(t)\ dt.$$

Penalized smoothing - Basis

So, define the vector $\mathbf{Y}_n = \{Y_{nj}\}$ and matrix $\mathbf{B}_n = \{B_k(t_{nj})\}$, then the penalized regression equation can be expressed as

$$S(\mathbf{c}_n) = (\mathbf{Y}_n - \mathbf{B}_n\mathbf{c}_n)^\top(\mathbf{Y}_n - \mathbf{B}_n\mathbf{c}_n) + \lambda\mathbf{c}_n^\top\mathbf{W}\mathbf{c}_n.$$

This is just a ridge regression, so the solution is given by

$$\hat{\mathbf{c}}_n = (\mathbf{B}_n^\top\mathbf{B}_n + \lambda\mathbf{W})^{-1}\mathbf{B}_n^\top\mathbf{Y}_n.$$

Notice that if all curves are observed at the same time points, then the $\mathbf{B}_n \equiv \mathbf{B}$. Once we have $\hat{\mathbf{c}}_n$ we can predict $\hat{X}_n(t)$ for any value of t. The fitted values are $\hat{\mathbf{Y}}_n = \mathbf{B}_n\hat{\mathbf{c}}_n$.

Penalized smoothing - Degrees of Freedom

To choose λ it helps to have the "degrees of freedom" from the model fit which is usually defined as

$$df = \text{trace}(\mathbf{B}_n(\mathbf{B}_n^\top \mathbf{B}_n + \lambda \mathbf{W})^{-1}\mathbf{B}_n^\top).$$

Notice that if $\lambda = 0$ then

$$df = \text{trace}((\mathbf{B}_n^\top \mathbf{B}_n)^{-1}\mathbf{B}_n^\top \mathbf{B}_n) = K.$$

Penalized smoothing - Choosing λ

To choose λ there are a number of ways. First, define

$$RSS = (\mathbf{Y}_n - \hat{\mathbf{Y}}_n)^\top (\mathbf{Y}_n - \hat{\mathbf{Y}}_n).$$

- $GCV(\lambda) = \frac{J}{(J-df)^2} RSS.$
- $AIC(\lambda) = J \log(J^{-1} RSS) + 2df$
- $BIC(\lambda) = J \log(J^{-1} RSS) + \log(J) df$
- Cross-validation

One would choose the λ which minimizes one of the above. GCV is maybe the most popular in FDA. Regardless, one must ALWAYS visually inspect the results to see if they make sense.

Penalized smoothing - Choosing λ

Consider a vector of outcomes, $\mathbf{Y} \in \mathbb{R}^N$, from some arbitrary model but whose predicted values can be expressed as

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y},$$

where $\mathbf{H}$ is some matrix which does not depend on $\mathbf{Y}$. Then the GCV is defined as

$$GCV = \frac{N|\mathbf{Y} - \hat{\mathbf{Y}}|^2}{(N - \text{trace}(\mathbf{H}))^2},$$

this part of why it is so popular in FDA, very easy to compute for a variety of models.

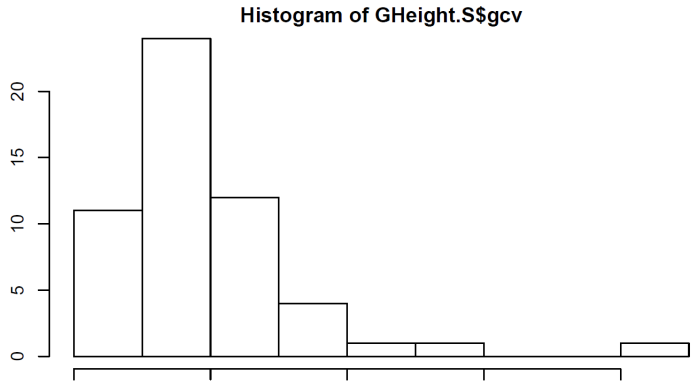
Penalized smoothing - R

We have been using `Data2fd` for function conversion, and it allows for smoothing, but it doesn't report the GCV. For more controlled smoothing we will use `smooth.basis`.

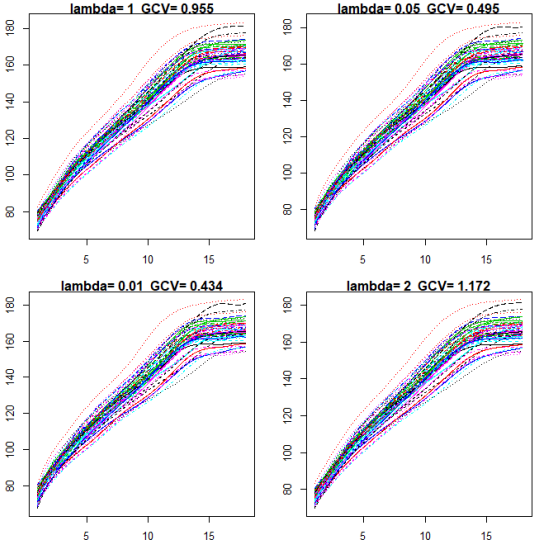
```
times = growth$age; GHeight = growth$hgtf
my_basis<-create.bspline.basis(c(1,18),
            nbasis=10,norder=5)
my_par<-fdPar(my_basis,Lfdobj=2,lambda=1)
GHeight.S<-smooth.basis(times,GHeight,my_par)
names(GHeight.S)[1:5]; names(GHeight.S)[6:9]

## [1] "fd"    "df"    "gcv"   "beta" "SSE"
## [1] "penmat"  "y2cMap"  "argvals" "y"
```

Penalized smoothing - R

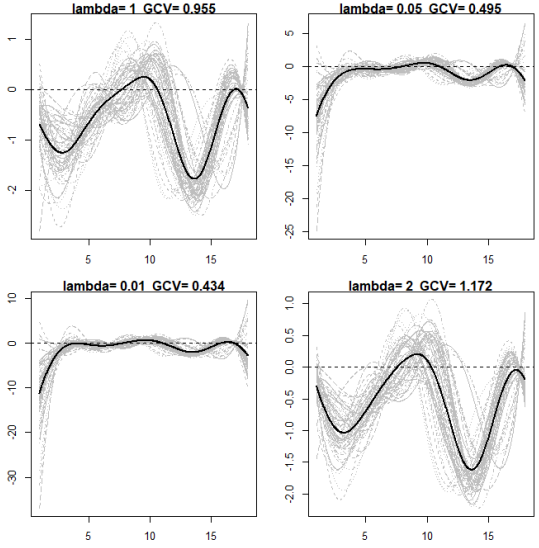
```
mean(GHeight.S$gcv)

## [1] 0.9545315

hist(GHeight.S$gcv)
```



Penalized smoothing - R



Berkeley - 2nd Derivatives



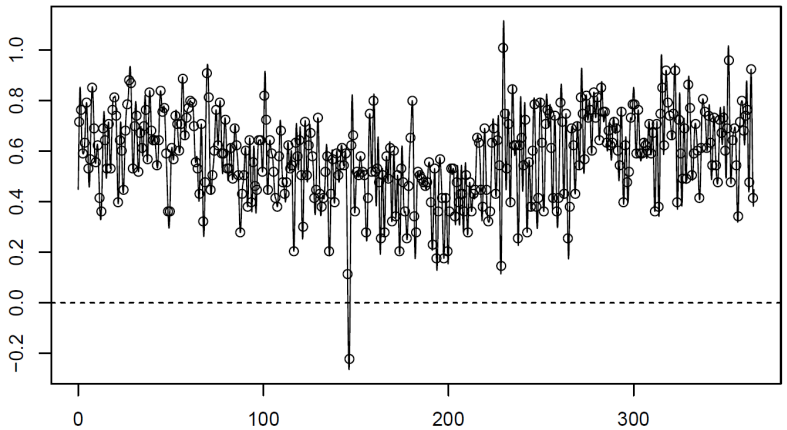
Example - Canadian Weather

The `CanadianWeather` data is also part of the `fda` package. It consists of daily temperature and precipitation data at 35 locations in Canada. Each daily measurement is an average across the years 1960 to 1994, so each station produces only one curve.

```
nbasis = 365; yearRng = c(0,365)
daybasis = create.fourier.basis(yearRng, nbasis)
logprecav = CanadianWeather$dailyAv[,,'log10precip']
dayprecfd <- with(CanadianWeather, smooth.basis(day.5,
        logprecav, daybasis,
        fdnames=list("Day","Station","log10(mm)"))$fd)
names(CanadianWeather)[1:3];names(CanadianWeather)[4:6

## [1] "dailyAv"  "place"    "province"
## [1] "coordinates" "region"      "monthlyTemp"
## [1] "monthlyPrecip" "geogindex"
```

Example - Canadian Weather

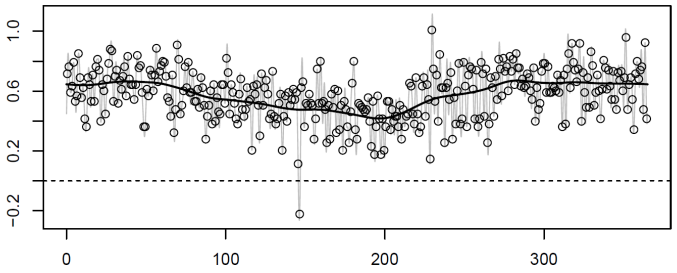
```
par(mar=c(3,3,1,1))
plot(dayprecfd[1]);points(day.5,logprecav[,1])
```



Example - Canadian Weather

Lets smooth it up some

```
par(mar=c(3,3,1,1))
daybasis = create.fourier.basis(yearRng, nbasis)
mypar = fdPar(daybasis,Lfdobj=2,lambda=10000)
Y.f<-smooth.basis(day.5,logprecav,mypar)
plot(dayprecfd[1],col='grey');points(day.5,logprecav[,
plot(Y.f$fd[1],add=TRUE,lwd=2)
```



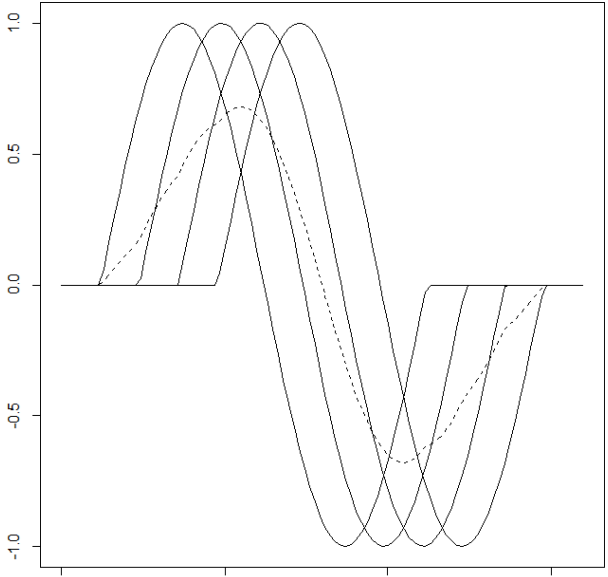
Example - Canadian Weather

```
mypar = fdPar(daybasis,Lfdobj=2,lambda=.001)
Y.f2<-smooth.basis(day.5,logprecav,mypar)
mean(Y.f2$gcv)

## [1] 0.08145941

mean(Y.f$gcv)

## [1] 0.04038778
```

Curve alignment

Often samples of curves may be "out of alignment", and inference can be more accurate/interpretable if the curves are properly aligned. The most common reason for being out of alignment is that each unit has its own "time scale". For example, subjects will hit puberty/growth spurts at slightly different ages, but we might want these events to be aligned across subjects.

Example



Modes of variation

A sample of curves can be decomposed into two sources of variation:

- amplitude variation - variation between curves in terms of height/amplitude,
- phase variation - variation between curves in terms of the domain.

Often one wishes to eliminate the phase variation and focus exclusively on the amplitude variation (but this is not always the case).

Modeling phase variation

We model phase variation by introduciont *warping functions*:

$$X_n(t) = X_n^*(h_n(t)).$$

Here X_n is the observed (unaligned) curve, X_n^* is the unobserved aligned curve, and h_n is a subject-specific warping function that is also unobserved. Variation between $X_n(t)$ and $X_n^*(t)$ describes the phase variation, while variation between different $X_n^*(t)$ describes the amplitude variation.

Warping functions

There are a few properties that make sense for the warping functions $h_n(t)$:

- $h(0) = 0$,
- $h(T) = T$,
- $h(t)$ is monotonically increasing.

Any concerns? If we can estimate the warping functions $h_n(t)$ then the aligned curves will be

$$X_n^*(t) = X_n(h_n^{-1}(t)).$$

Landmark Registration

The first approach to alignment is called landmark registration. The idea is to choose landmarks from each function to align. For example, the maximum, minimum, and so on. Often this involves manually choosing the location of the landmarks through some point and click function, so it can be very time-consuming.

Let $t_{n1}, \ldots, t_{nJ}$ be the selected time points, and suppose $t_1, \ldots, t_n$ are the selection made off of the mean function (or some function from the sample). Then the goal is to find h_n such

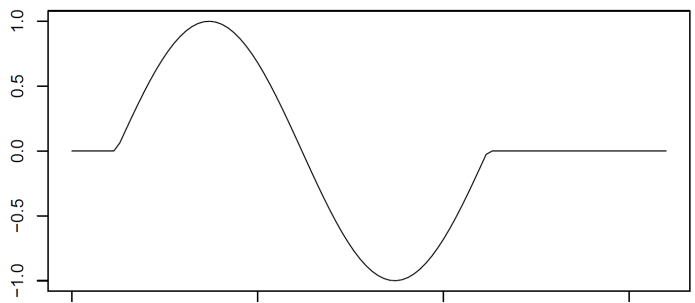
$$h_n(t_j) = t_{nj},$$

which can be done using fitting a polynomial or some basis expansion.

Landmark - R

Using the `locator` function, you can click on plots to find the coordinates of landmarks. You can then use the `landmarkreg` function to align the functions.

```
plot(times,x1,type="l",ylab="",xlab="")
num_loc<-2
t_loc<-locator(num_loc)
```



Continuous Registration

Continuous registration is much easier to automate. The idea is that you choose a specific curve (usually the mean function) and align "all" time points to that curve. So basically, one would choose h_n such that

$$\int (X(h_n^{-1}(t)) - \bar{X}(t))^2 dt,$$

is as small as possible. Typically the h_n^{-1} is expanded using some basis, while imposing our other discussed constraints. This can be carried out using the `register.fd` function in `R`.

Example - Berkeley

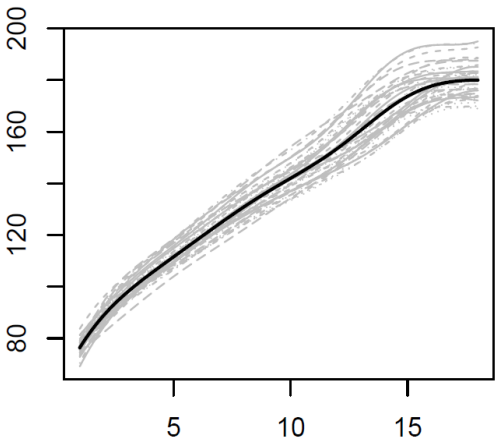
Let's return to the Berkeley growth data, and try out the alignment. Why does registering such data make sense?

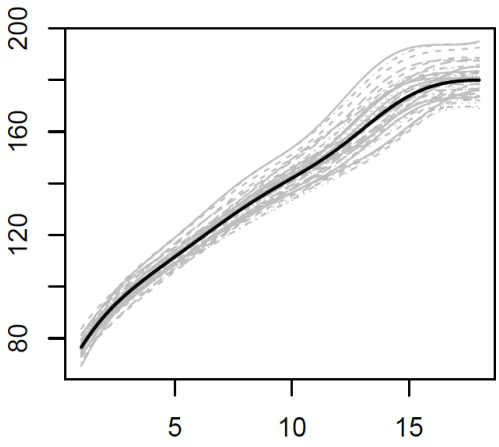
```
hgtbasis <- create.bspline.basis(c(1,18), norder = 6,
growfdPar <- fdPar(hgtbasis, 4, 10^(-.5))
X.f<-smooth.basis(growth$age,growth$hgtm,growfdPar)$fd
X.reg<-register.fd(X.f)
```

```
names(X.reg)

## [1] "regfd"  "warpfd" "Wfd"    "shift"  "y0fd"
## [6] "yfd"
```

Example - Berkeley

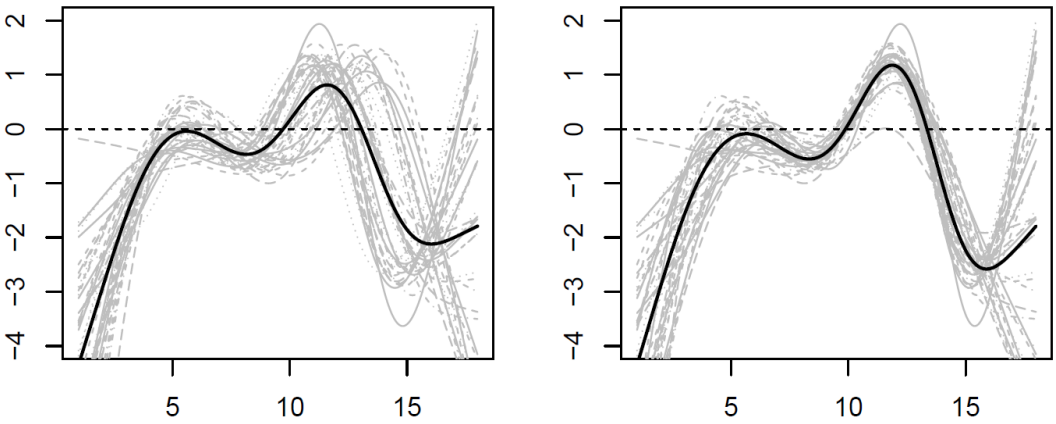
Let's return to the Berkeley growth data, and try out the alignment. Why does registering such data make sense?





Example - Berkeley

Let's take a look at the 2nd derivatives. First, we take the derivatives, then we align them.



Further Topics with References-1

We do not cover all directions of FDA research. (curve alignment, classification, clustering, etc). You can further study

- **Registration/Alignment**
  - Marron, J. S., Ramsay, J. O., Sangalli, L. M. and Srivastava, A. (2015). Functional data analysis of amplitude and phase variation. *Statistical Science*, 30, 468–484.
  - Earls, Cecilia and Hooker, Giles (2016). Adapted variational bayes for functional data registration, smoothing, and prediction. *Bayesian Analysis*.
- **Learning/Clustering/Classification**: some parts of
  - Wang, J-L., Chiou, J-M. and Muller, H-G. (2016). Review of functional data analysis. *The Annual Review of Statistics and Its Application*, 3, 257–295.
  - Cuevas, Antonio (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147, 1–23.
  - Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8, 231–255.

Further Topics with References-2

- **Classification**: some parts of
  - Delaigle, Aurore and Hall, Peter (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2), 267–286.
  - Grorecki, T., Krzysko, M. and Wolynski, W. (2015). Classification problems based on regression models for multidimensional functional data. *Statistics in Transition*, 16, 97–110.
- **Dimension Reduction**: my papers & literature therin