An Introduction to Functional Data Analysis

Chongzhi Di
Fred Hutchinson Cancer Research Center
`cdi@fredhutch.org`

Biotat 578A: Special Topics in (Genetic) Epidemiology
November 10, 2015

- Textbook
  – Ramsay and Silverman (2005), *Functional Data Analysis*, 2nd edition, Springer.
  – Ruppert, Wand and Carroll (2003), *Semiparametric Regression*, Cambridge University Press.
- Software: R packages
  fda: Functional Data Analysis
  refund: Regression with Functional Data
  SemiPar: Semiparametric Regression
- Acknowledgements
  Prof. Giles Hooker, Cornell University
  `http://faculty.bscb.cornell.edu/~hooker`

What Is Functional Data?

Functional data is multivariate data with an ordering on the dimensions. (Müller, (2006))

Key assumption is *smoothness*:

$$y_{ij} = x_i(t_{ij}) + \epsilon_{ij}$$

with t in a continuum (usually time), and $x_i(t)$ smooth

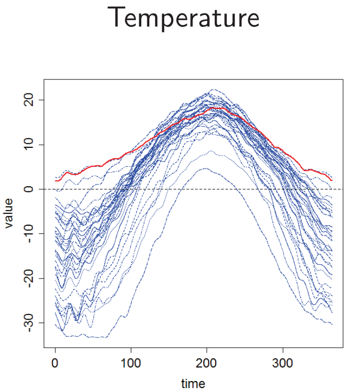
Functional data = the functions $x_i(t)$.

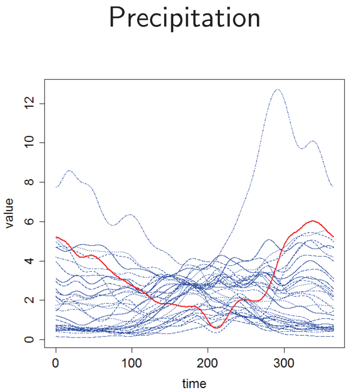
- Optical tracking equipment (eg handwriting data, but also for physiology, motor control,...)
- Electrical measurements (EKG, EEG and others)
- Spectral measurements (astronomy, materials sciences)

But, noisier and less frequent data can also be used.

Canadian Weather Data

Average daily temperature and precipitation records in 35 weather stations across Canada

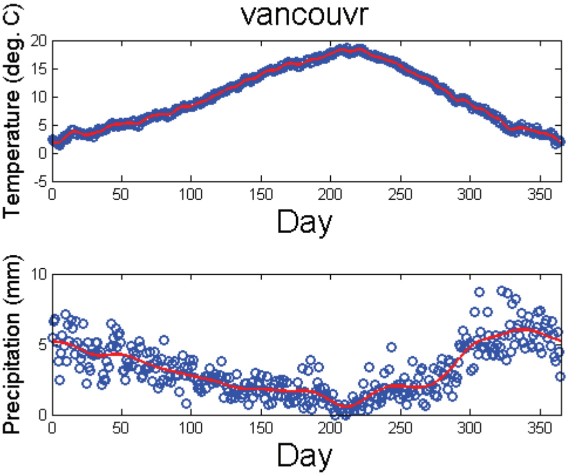




Interest is in variation in and relationships between smooth, underlying processes.

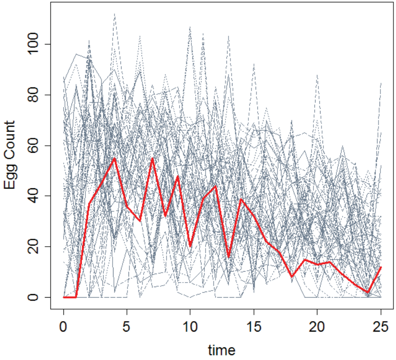
Weather In Vancouver

Measure of climate: daily precipitation and temperature in Vancouver, BC averaged over 40 years.



Medfly Data

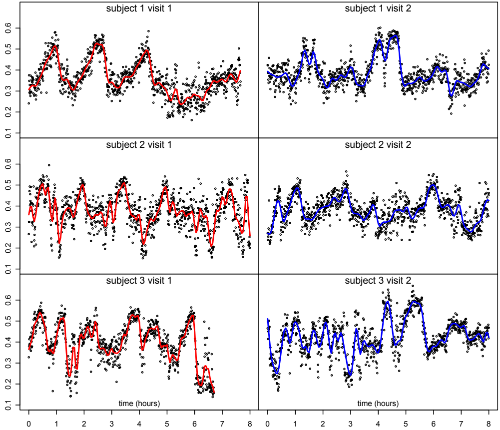
Records of number of eggs laid by Mediterranean Fruit Fly (Ceratitis capitata) in each of 25 days (courtesy of H.-G. Müller).



- Total of 50 flies
- Assume eggcount measurements relate to smooth process governing fertility

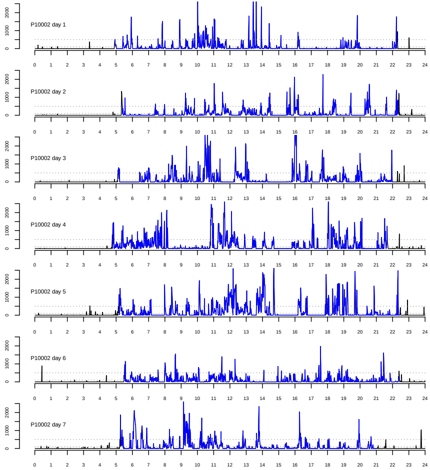
SHHS: Sleep Heart Health Study

- More than $3,000$ subjects, two visits per subject
- $Y_{ij}(t)$: normalized EEG $\delta-$power series



Accelerometry data

- Activity Count Data: three dimensional time series per subject
- 1-minute resolution: 10080 time points in 7 days



What Are We Interested In?

- Representations of distribution of functions
    - mean
    - variation
    - covariation
- Relationships of functional data to
    - covariates
    - responses
    - other functions
- Relationships between derivatives of functions.
- Timing of events in functions.

What Are The Challenges?

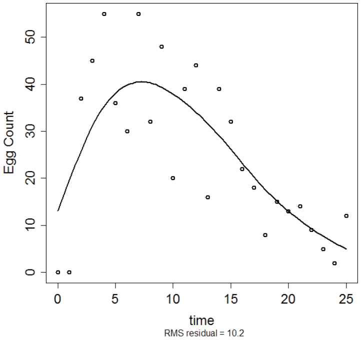
- Estimation of functional data from noisy, discrete observations.
- Numerical representation of infinite-dimensional objects
- Representation of variation in infinite dimensions.
- Description of statistical relationships between infinite dimensional objects.
- $n < p = \infty$, and use of smoothness.

Representing Functional Data

From Discrete to Functional Data

Represent data recorded at discrete times as a continuous function in order to

Medfly record 1



- Allow evaluation of record at any time point (especially if observation times are not the same across records).
- Evaluate rates of change.
- Reduce noise.

From Discrete to Functional Data

1. Representing non-parametric continuous-time functions.
   - Basis-expansion methods:

$$x(t) = \sum_{i=1}^{K} \phi_i(t) c_i$$

   for pre-defined $\phi_i(t)$ and coefficients c_i.

2. Reducing noise in measurements
   - Smoothing penalties:

$$c = \text{argmin} \sum_{i=1}^{n} (y_i - x(t_i))^2 + \lambda \int [Lx(t)]^2 \, dt$$

   - $Lx(t)$ measures "roughness" of x
   - λ a "smoothing parameter" that trades-off fit to the y_i and roughness; must be chosen.

Basis Expansions

$$y_i = x(t_i) + \epsilon_i$$

represent $x(t)$ as

$$x(t) = \sum_{j=1}^{K} c_j \phi_j(t) = \Phi(t)\mathbf{c}$$

We say $\Phi(t)$ is a *basis system* for x.

Terms for curvature in linear regression

$$y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \cdots + \epsilon_i$$

implies

$$x(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \cdots$$

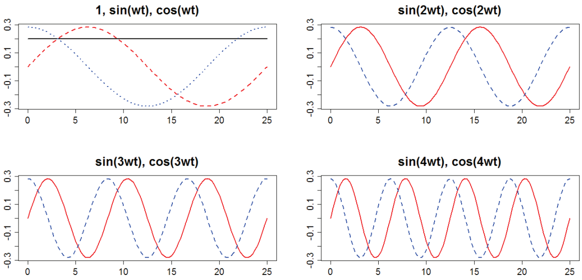
Polynomials are unstable; Fourier bases and B-splines will be more useful.

The Fourier Basis

- basis functions are sine and cosine functions of increasing frequency:

$$1, sin(\omega t), cos(\omega t), sin(2\omega t), cos(2\omega t), \ldots$$
$$sin(m\omega t), cos(m\omega t), \ldots$$

- constant $\omega = 2\pi/P$ defines the period P of oscillation of the first sine/cosine pair.



B-spline Bases

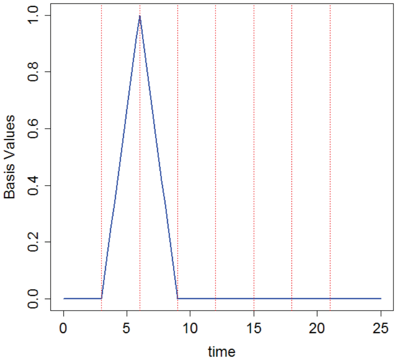
- Splines are polynomial segments joined end-to-end.
- Segments are constrained to be smooth at the joins.
- The points at which the segments join are called *knots*.
- System defined by
  - The order m (order $=$ degree$+1$) of the polynomial
  - the location of the knots.

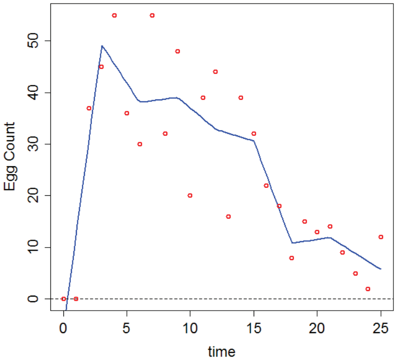
See de Boor, 2001, “A Practical Guide to Splines”, Springer.

Splines

Medfly data with knots every 3 days.

Splines of order 2: piecewise linear, continuous

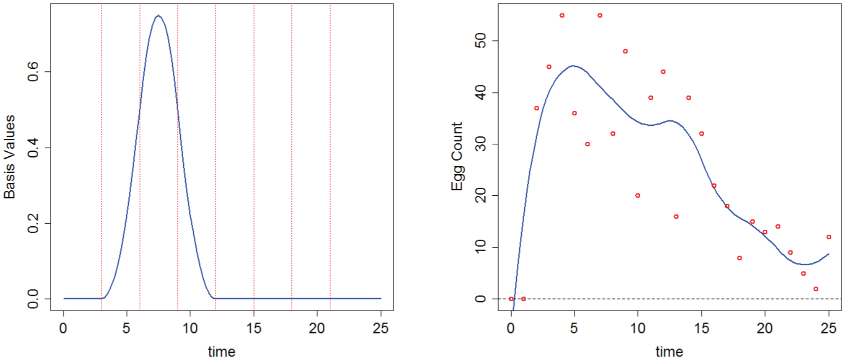




Splines

Medfly data with knots every 3 days.

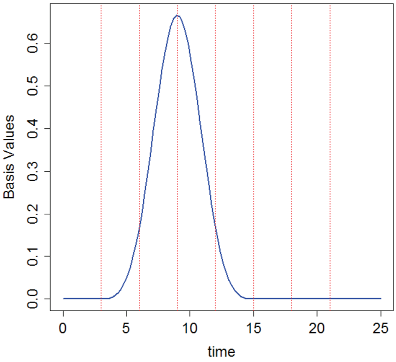
Splines of order 3: piecewise quadratic, continuous derivatives

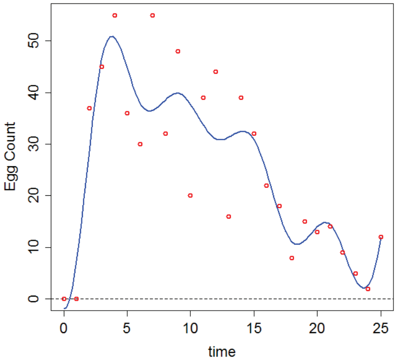


Splines

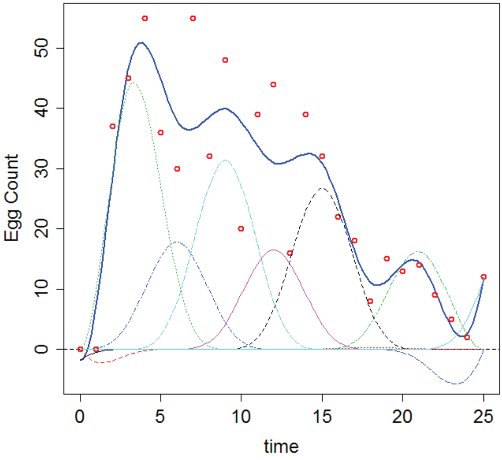
Medfly data with knots every 3 days.

Splines of order 4: piecewise cubic, continuous 2nd derivatives





An illustration of basis expansions for B-splines



Sum of scaled basis functions results in fit.

Ordinary Least-Squares Estimates

Assume we have observations for a single curve

$$y_i = x(t_i) + \epsilon$$

and we want to estimate

$$x(t) \approx \sum_{j=1}^{K} c_j \phi_j(t)$$

Minimize the sum of squared errors:

$$SSE = \sum_{i=1}^{n} (y_i - x(t_i))^2 = \sum_{i=1}^{n} (y_i - \Phi(t_i)\mathbf{c})^2$$

This is just linear regression!

Linear Regression on Basis Functions

- If the N by K matrix $\mathbf{\Phi}$ contains the values $\phi_j(t_k)$, and $\mathbf{y}$ is the vector $(y_1, \ldots, y_N)$, we can write

$$SSE(\mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})^T (\mathbf{y} - \mathbf{\Phi c})$$

- The error sum of squares is minimized by the *ordinary least squares estimate*

$$\hat{\mathbf{c}} = \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{y}$$

- Then we have the estimate

$$\hat{x}(t) = \Phi(t)\hat{\mathbf{c}} = \Phi(t) \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{y}$$

Smoothing Penalties

- Problem: how to choose a basis? Large affect on results.
- Finesse this by specifying a very rich basis, but then imposing smoothness.
- In particular, add a penalty to the least-squares criterion:

$$\text{PENSSE} = \sum_{i=1}^{n} (y_i - x(t_i))^2 + \lambda J[x]$$

- $J[x]$ measures "roughness" of x.
- λ represents a continuous tuning parameter (to be chosen):
    - $\lambda \uparrow \infty$: roughness increasingly penalized ,$x(t)$ becomes smooth.
    - $\lambda \downarrow 0$: penalty reduces, $x(t)$ fits data better.

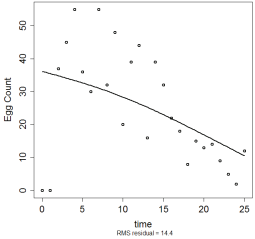
What do we mean by smoothness?

Some things are fairly clearly smooth:

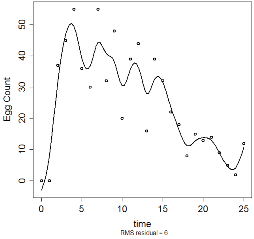
- constants
- straight lines

What we really want to do is eliminate small "wiggles" in the data while retaining the right shape

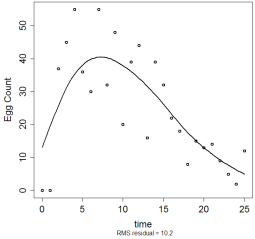
Too smooth



Too rough



Just right



The D Operator

We use the notation that for a function $x(t)$,

$$Dx(t) = \frac{d}{dt}x(t)$$

We can also define further derivatives in terms of powers of D:

$$D^2x(t) = \frac{d^2}{dt^2}x(t), \ldots, D^kx(t) = \frac{d^k}{dt^k}x(t), \ldots$$

- $Dx(t)$ is the instantaneous *slope* of $x(t)$; $D^2x(t)$ is its *curvature*.
- We measure the size of the curvature for all of x by

$$J_2[x] = \int \left[D^2x(t)\right]^2 dt$$

Calculating the Penalized Fit

When $x(t) = \Phi(t)\mathbf{c}$, we have that

$$\int \left[D^2 x(t)\right]^2 dt = \int \mathbf{c}^T \left[D^2\Phi(t)\right] \left[D^2\Phi(t)\right]^T \mathbf{c} dt = \mathbf{c}^T R_2 \mathbf{c}$$

$[R_2]_{jk} = \int [D^2\phi_j(t)][D^2\phi_k(t)]dt$ is the *penalty matrix*.

The penalized least squares estimate for $\mathbf{c}$ is n

$$\hat{\mathbf{c}} = \left[\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda R_2\right]^{-1} \boldsymbol{\Phi}^T \mathbf{y}$$

This is still a linear smoother:

$$\hat{\mathbf{y}} = \boldsymbol{\Phi} \left[\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda R_2\right]^{-1} \boldsymbol{\Phi}^T \mathbf{y} = S(\lambda)\mathbf{y}$$

Linear Smooths and Degrees of Freedom

- In least squares fitting, the degrees of freedom used to smooth the data is exactly K, the number of basis functions

- The smoothing penalty reduces the flexibility of the smooth
- The degrees of freedom are controlled by λ. A natural measure turns out to be

$$df(\lambda) = \text{trace}\left[S(\lambda)\right], \; S(\lambda) = \mathbf{\Phi}\left[\mathbf{\Phi}^T\mathbf{\Phi} + \lambda R_L\right]^{-1}\mathbf{\Phi}^T$$

- Medfly data fit with 25 basis functions, $\lambda = e^4$ resulting in $df = 4.37$.

Choosing Smoothing Parameters: Cross Validation

There are a number of data-driven methods for choosing smoothing parameters.

- Ordinary Cross Validation: leave one point out and see how well you can predict it:

$$\text{OCV}(\lambda) = \frac{1}{n}\sum\left(y_i - x_\lambda^{-i}(t_i)\right)^2 = \frac{1}{n}\sum\frac{(y_i - x_\lambda(t_i))^2}{(1 - S(\lambda)_{ii})^2}$$

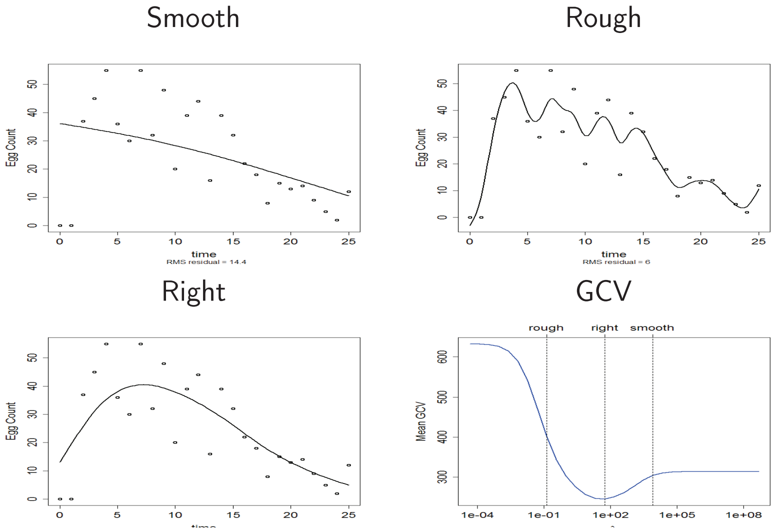
- Generalized Cross Validation tends to smooth more:

$$\text{GCV}(\lambda) = \frac{\sum(y_i - x_\lambda(t_i))^2}{[\text{trace}(\mathbb{I} - S(\lambda))]^2}$$

  will be used here.
- Other possibilities: AIC, BIC,...

Generalized Cross Validation

Use a grid search, best to do this for $\log(\lambda)$



Alternatives: Smoothing and Mixed Models

Connection between the smoothing criterion for $\mathbf{c}$:

$$\text{PENSSE}(\lambda) = \sum_{i=1}^{n} (y_i - \mathbf{c}^T \Phi(t_i))^2 + \lambda \mathbf{c}^T R \mathbf{c}$$

and negative log likelihood if $\mathbf{c} \sim N(0, \tau^2 R^{-1})$:

$$\log L(\mathbf{c}|\mathbf{y}) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{c}^T \Phi(t_i))^2 + \frac{1}{2\tau^2} \mathbf{c}^T R \mathbf{c}$$

(note that R is singular – must use generalized inverse).

Suggests using ReML estimates for σ^2 and τ^2 in place of λ.

Summary

1. Basis Expansions

$$x_i(t) = \Phi(t)\mathbf{c}_i$$

   - Good basis systems approximate any (sufficiently smooth) function arbitrarily well.
   - Fourier bases useful for periodic data.
   - B-splines make efficient, flexible generic choice.
2. Smoothing Penalties used to penalize roughness of result
   - $Lx(t) = 0$ defines what is "smooth".
   - Commonly $Lx = D^2x \Rightarrow$ straight lines are smooth.
   - Departures from smoothness traded off against fit to data.
   - GCV used to decide on trade off; other possibilities available.

These tools will be used throughout the rest of FDA.

Once estimated, we will treat smooths as fixed, observed data (but see comments at end).

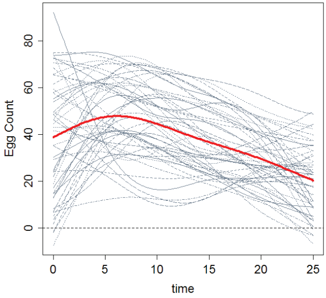
Exploratory Data Analysis

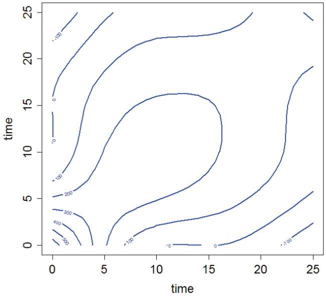
Mean and Variance

Summary statistics:

- mean $\bar{x}(t) = \frac{1}{n}\sum x_i(t)$
- covariance
  $\sigma(s,t) = \mathrm{cov}(x(s), x(t)) = \frac{1}{n}\sum(x_i(s) - \bar{x}(s))(x_i(t) - \bar{x}(t))$

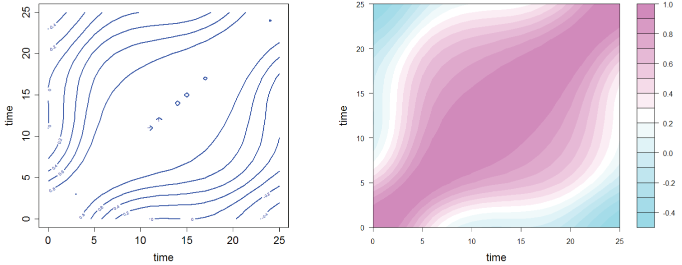
Medfly Data:





Correlation

$$\rho(s,t) = \frac{\sigma(s,t)}{\sqrt{\sigma(s,s)}\sqrt{\sigma(t,t)}}$$



From multivariate to functional data: turn subscripts j, k into arguments s, t.

Functional PCA

- Instead of covariance matrix Σ, we have a surface $\sigma(s,t)$.
- Would like a low-dimensional summary/interpretation.
- Multivariate PCA, use Eigen-decomposition:

$$\Sigma = U^T D U = \sum_{j=1}^{p} d_j u_j u_j^T$$

and $u_i^T u_j = I(i=j)$.

- For functions: use Karhunen-Loève decomposition:

$$\sigma(s,t) = \sum_{j=1}^{\infty} d_j \xi_j(s) \xi_j(t)$$

for $\int \xi_i(t)\xi_j(t)dt = I(i=j)$

PCA and Karhunen-Loève

$$\sigma(s,t) = \sum_{i=1}^{\infty} d_i \xi_i(s)\xi_i(t)$$

- The $\xi_i(t)$ maximize $\text{Var}\left[\int \xi_i(t)x_j(t)dt\right]$.
- $d_i = \text{Var}\left[\int \xi_i(t)x_j(t)dt\right]$
- $d_i/\sum d_i$ is proportion of variance explained
- Principal component scores are

$$f_{ij} = \int \xi_j(t)[x_i(t) - \bar{x}(t)]dt$$

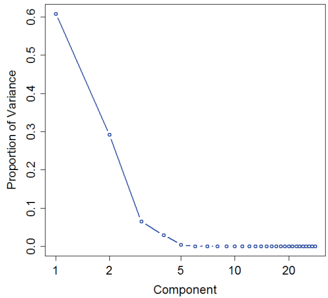
- Reconstruction of $x_i(t)$:

$$x_i(t) = \bar{x}(t) + \sum_{j=1}^{\infty} f_{ij}\xi_j(t)$$

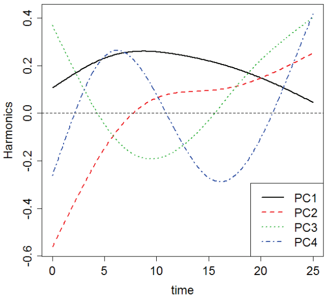
functional Principal Components Analysis

fPCA of Medfly data

Scree Plot



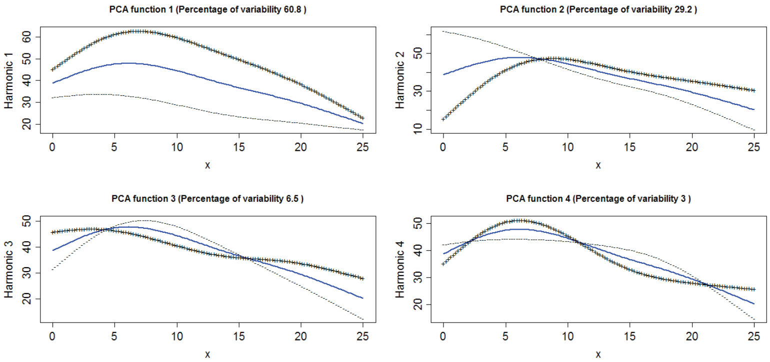
Components



Usual multivariate methods: choose # components based on percent variance explained, screeplot, or information criterion.

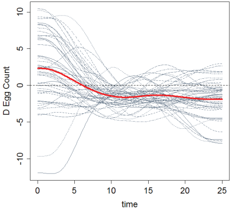
functional Principal Components Analysis

Interpretation often aided by plotting $\bar{x}(t) \pm 2\sqrt{d_i}\xi_i(t)$

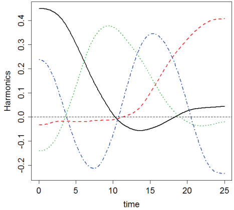


Derivatives

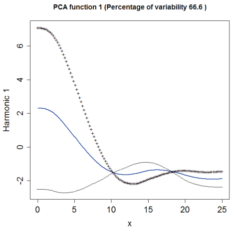
Derivatives



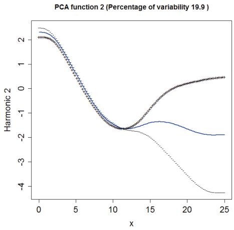
PCs



Component 1



Component 2



- Often useful to examine a rate of change.
- Examine first derivative of medfly data.
- Variation divides into fast or slow either early or late.