apts.ac.uk
Academy for PhD Training in Statistics

Flexible Regression

Session 5 - Flexible regression extensions

Please see the full notes for full explanation and details. The slides will help to signpost and guide you through the main points of the notes.

Claire Miller & Tereza Neocleous

Overview

- Gaussian processes for regression;
- Functional data analysis;
- Other related topics.

6. Extensions

Approaches for modelling random functions (collections of random functions):

Gaussian processes (GPs)

A Bayesian nonparametric model for function estimation in regression.

Functional Data Analysis (FDA)

The analysis of information on functions or curves - interested in the combined information over a set of functions.

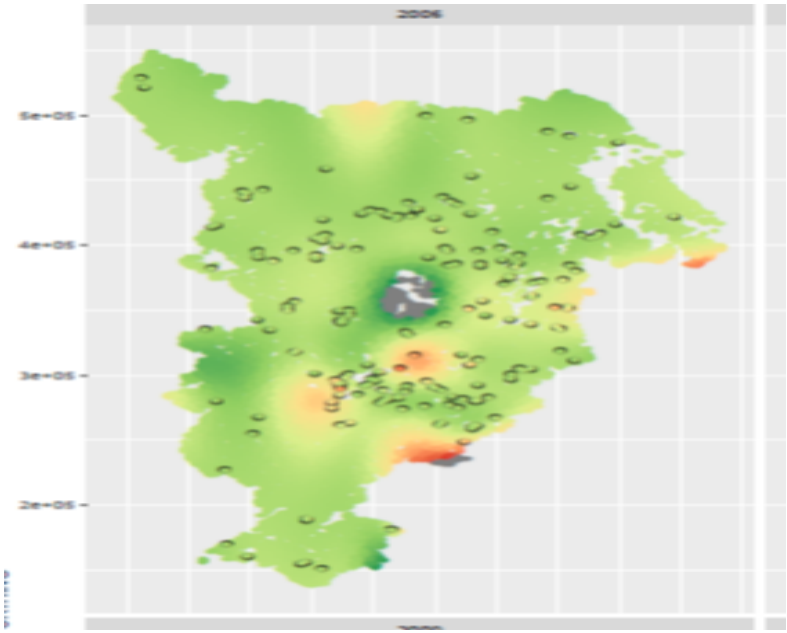
6.1 Gaussian processes (GP)

Definition of a GP

A collection of random variables $Y_i = Y(\mathrm{x}_i)$ $(i = 1, 2, 3, \ldots)$ depending on covariates x_i such that any *finite* subset of random variables $\mathrm{y} = (Y_1, \ldots, Y_n) = (Y(\mathrm{x}_1), \ldots, Y(\mathrm{x}_n))$ has a multivariate Normal distribution.

Spatial context:

In geostatistics this model is known as a **kriging model**.



6.1 Gaussian processes (GP)

$$\mathsf{y} \sim \mathsf{N}\left(0, \mathsf{K} + \sigma^2 \mathsf{I}\right) \qquad \text{with} \qquad \mathsf{K} = \begin{bmatrix} k(\mathsf{x}_1, \mathsf{x}_1) & \dots & k(\mathsf{x}_1, \mathsf{x}_n) \\ \vdots & \ddots & \vdots \\ k(\mathsf{x}_n, \mathsf{x}_1) & \dots & k(\mathsf{x}_n, \mathsf{x}_n) \end{bmatrix}$$

Compare to earlier in the week:

Using $Y_i = f_i + \varepsilon_i$ with $f_i = f(\mathsf{x}_i)$ and $\varepsilon_i \sim \mathsf{N}(0, \sigma^2)$ we can rewrite this as

$$\mathsf{y}|\mathsf{f} \sim \mathsf{N}(\mathsf{f}, \sigma^2 \mathsf{I}) \qquad\qquad \mathsf{f} \sim \mathsf{N}\left(0, \mathsf{K}\right),$$

i.e. $\mathrm{Cov}(f_i, f_j) = k(\mathsf{x}_i, \mathsf{x}_j)$

6.1 Gaussian processes (GP)

Comparisons with GAMs

- GAMs are easy to fit and interpret;
- Computation can become challenging at higher dimensions;
- GPs allow the response to depend on all inputs simultaneously;
- The covariance function elements play a similar role to the smoothing parameters in the classic GAM;
- The type of structure captured by a GP model is mainly determined by its kernel;
- One of the difficulties can be in choosing a kernel which can represent structure in the data.

6.1 GP: covariance functions

Covariance function

$k(\mathrm{x}_i, \mathrm{x}_j) = \mathrm{Cov}(f_i, f_j)$ is called the covariance function / kernel function.

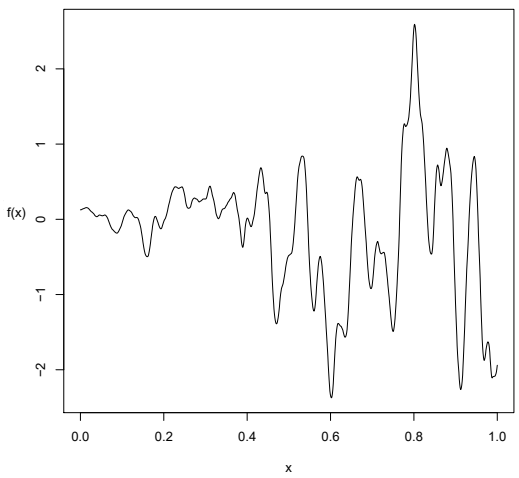
- k can be choosen freely as long as
  - k needs to be **symmetric** , i.e. $k(\mathrm{x}_i, \mathrm{x}_j) = k(\mathrm{x}_j, \mathrm{x}_i)$, and
  - the matrix

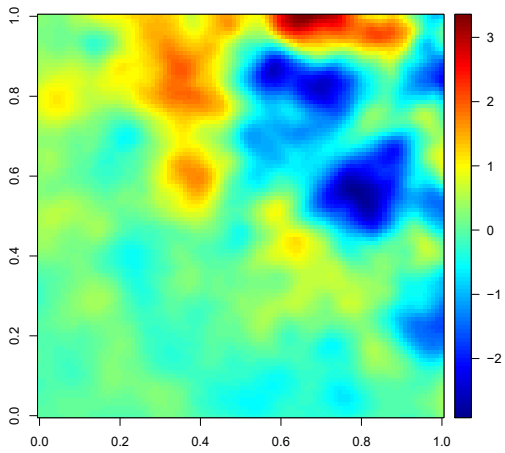
$$\mathrm{K} = \begin{bmatrix} k(\mathrm{x}_1, \mathrm{x}_1) & \dots & k(\mathrm{x}_1, \mathrm{x}_n) \\ \vdots & \ddots & \vdots \\ k(\mathrm{x}_n, \mathrm{x}_1) & \dots & k(\mathrm{x}_n, \mathrm{x}_n) \end{bmatrix}$$

    is positive (semi-)definite.

6.1 GP: stationary processes

- k is called **stationary** if $k(\mathrm{x}_i, \mathrm{x}_j) = k(\mathrm{x}_i - \mathrm{x}_j)$
- "Only relative positions matter"
- Stationarity implies homogeneous variance.

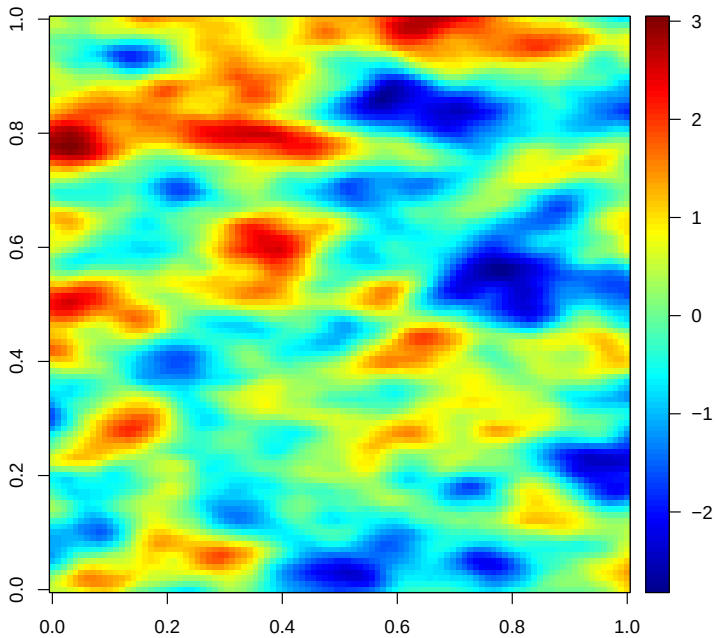




These proceses are not stationary.

6.1 GP: isotropic processes

- k is called **isotropic** if $k(\mathrm{x}_i, \mathrm{x}_j) = k(\|\mathrm{x}_i - \mathrm{x}_j\|)$
- "Only distance, but not directions matter."
- We will assume isotropy for the remainder of this session.



This proceses is not isotropic.

6.1 GP: separable processes

- A process is called **separable** if

$$k(\mathrm{x}_i, \mathrm{x}_j) = k_1(x_{i1} - x_{j1}) \cdot k_2(x_{i2} - x_{j2}) \cdots k_p(x_{ip} - x_{jp}).$$

- If the covariance function is separable and the data are on a regular grid, then

$$\mathrm{K} = \mathrm{K}_1 \otimes \mathrm{K}_2 \otimes \ldots \otimes \mathrm{K}_m,$$

  (K_j is the covariance matrix constructed using the unique values of the j-th block of covariate only)

- **Separability can also be used as a model assumption.**

6.1 GP: separable processes

Application: Spatio-temporal processes

- Data observed over space (s_i) and time t_i.
  $\rightsquigarrow \mathrm{x}_i = (s_{i1}, s_{i2}, t_i) = (\mathrm{s}_i, t_i)$

- Spatio-temporal models are often assumed to be separable

$$k((\mathrm{s}_i, t_i), (\mathrm{s}_j, t_j)) = k_1(\mathrm{s}_i, \mathrm{s}_j) k_2(t_i, t_j)$$

6.1.2 GP: predictions

Reminder: Conditional distributions for a Gaussian

Assume that

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

Then the conditional distribution of y_2 given y_1 is

$$y_2|y_1 \sim N\left(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

6.1.2 GP: predictions

- **New observation** y_0 with covariates x_0.

- Look at joint distribution

$$\begin{pmatrix} \mathsf{y} \\ y_0 \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathsf{K} + \sigma^2 \mathsf{I} & \mathsf{k}_0 \\ \mathsf{k}_0' & k_{00} + \sigma^2 \end{pmatrix} \right),$$

  with $\mathsf{k}_0 = (k(\mathsf{x}_0, \mathsf{x}_1), \ldots, k(\mathsf{x}_0, \mathsf{x}_n))$ and $k_{00} = k(\mathsf{x}_0, \mathsf{x}_0)$.

- Formula for the conditional distribution of a Gaussian yields

$$y_0 | \mathsf{y} \sim \mathsf{N}\left(\mathsf{k}_0' \left(\mathsf{K} + \sigma^2 \mathsf{I} \right)^{-1} \mathsf{y}, \left(k_{00} - \mathsf{k}_0' \left(\mathsf{K} + \sigma^2 \mathsf{I} \right)^{-1} \mathsf{k}_0 \right) + \sigma^2 \right)$$

6.1.3 GP: examples of covariance functions

Squared exponential (SE) / Gaussian kernel

$$k(\mathrm{x}_i, \mathrm{x}_j) = \tau^2 \exp(-\rho \|\mathrm{x}_i - \mathrm{x}_j\|^2)$$

Very smooth process: infinitely differentiable

Exponential / OU process

$$k(\mathrm{x}_i, \mathrm{x}_j) = \tau^2 \exp(-\rho \|\mathrm{x}_i - \mathrm{x}_j\|)$$

Very rough process: continuous, but not diffferentiable
OU process is continuous equivalent of $AR(1)$ process.

τ^2 controls prior variance, ρ controls (inverse) correlation length

6.1.3 GP: examples of covariance functions

γ-exponential

$$k(\mathrm{x}_i, \mathrm{x}_j) = \tau^2 \cdot \exp(-\rho \|\mathrm{x}_i - \mathrm{x}_j\|^\gamma)$$

with $0 < \gamma \leq 2$.

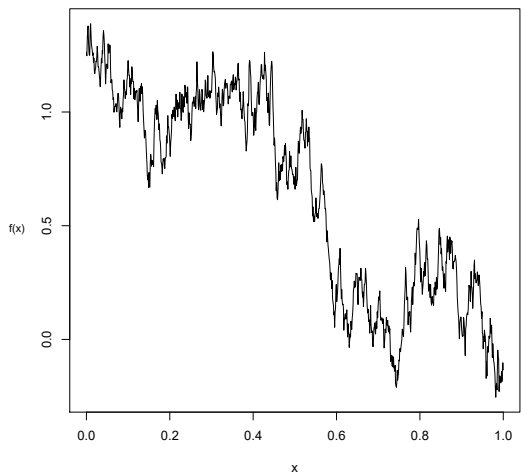
Matérn class

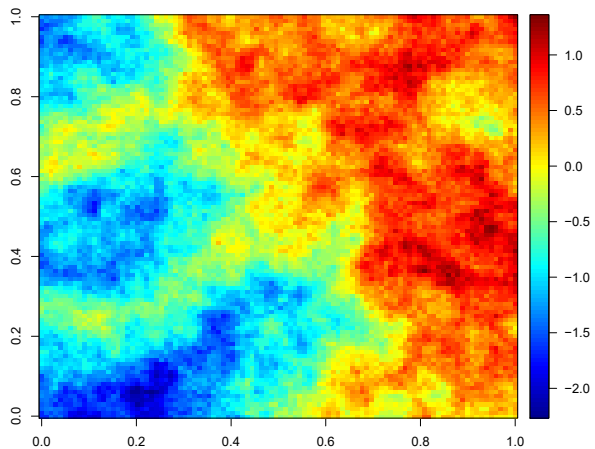
$$k(\mathrm{x}_i, \mathrm{x}_j) = \tau^2 \cdot \frac{1}{\Gamma(\kappa) 2^{\kappa-1}} (\sqrt{2\kappa}\rho \|\mathrm{x}_i - \mathrm{x}_j\|)^\kappa K_\kappa \left(\sqrt{2\kappa}\rho \|\mathrm{x}_i - \mathrm{x}_j\| \right),$$

Special cases: OU process ($\kappa = \frac{1}{2}$) and the squared exponential ($\kappa \to +\infty$).

6.1.3 GP: examples of covariance functions

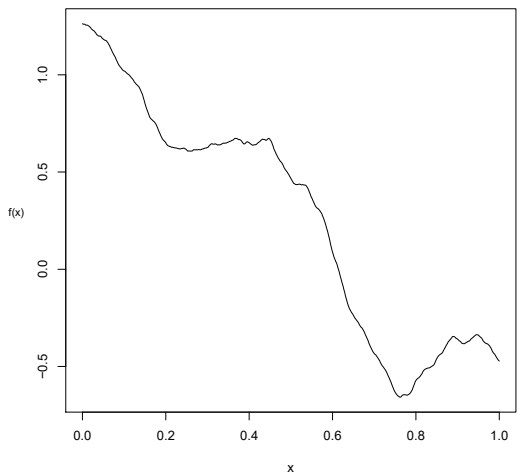
Matérn: $\kappa = 0.5$ (OU process)

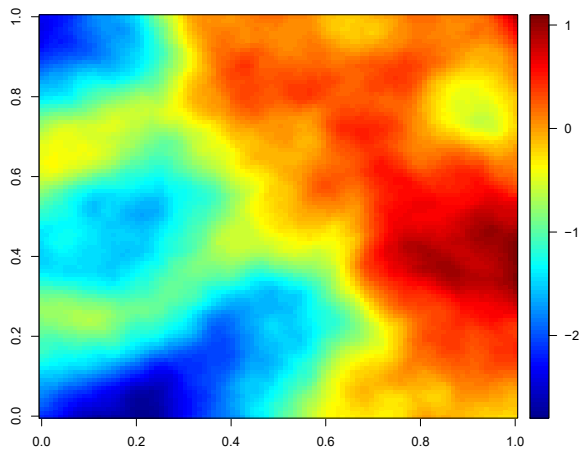




6.1.3 GP: examples of covariance functions

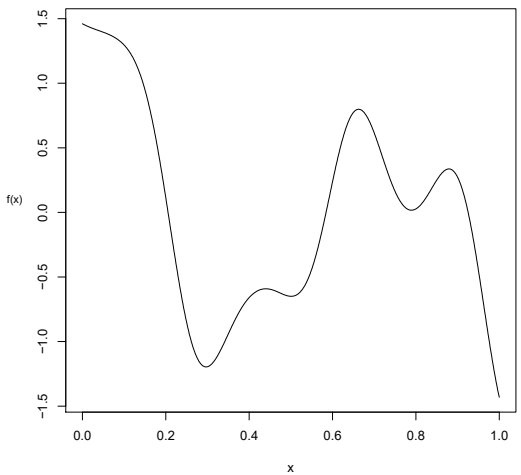
Matérn: $\kappa = 1.5$

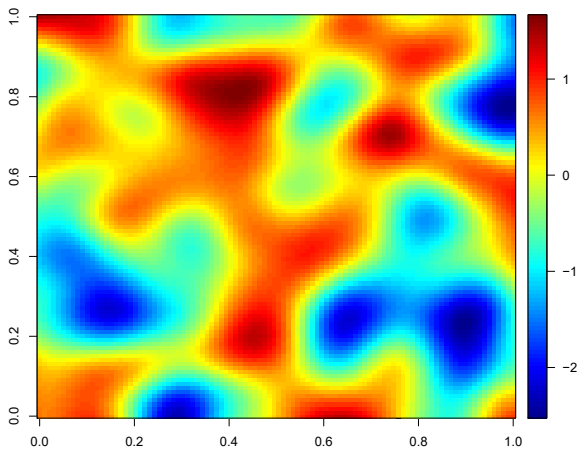




6.1.3 GP: examples of covariance functions

Matérn: $\kappa = +\infty$ (SE/Gaussian)





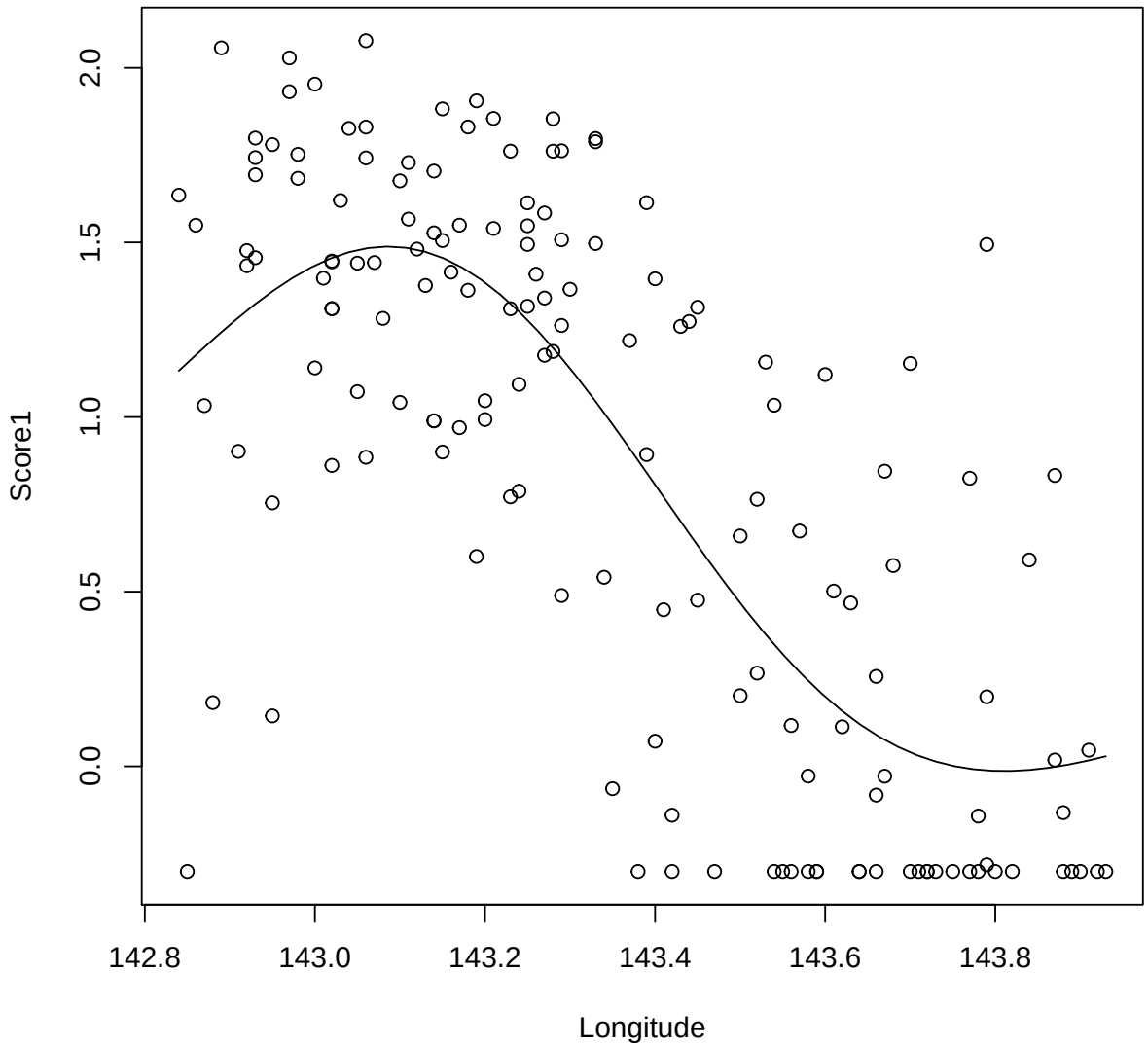
6.1.5 Gaussian processes in R

Great Barrier Reef data

```
library(mlegp)
fit <- mlegp(trawl$Longitude , trawl$Score1)
newdata <- data.frame(Longitude = seq(min(trawl$Longitude),
                                      max(trawl$Longitude), len=50))
predictions <- predict(fit, newdata)
plot(Score1~Longitude, data=trawl)
lines(newdata$Longitude, predictions)
```

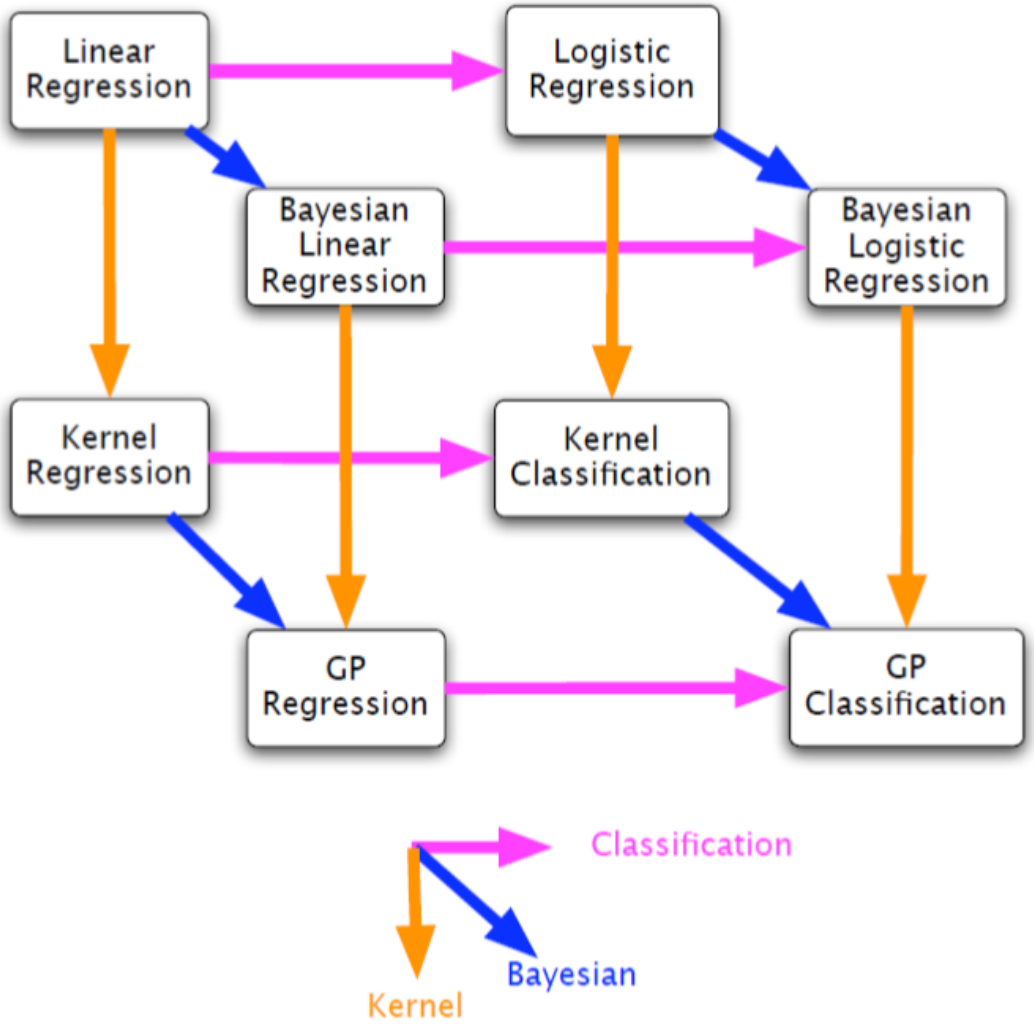
6.1.5 Gaussian processes in R

Great Barrier Reef data



6.1 Gaussian processes

A view from:
`mlss2011.comp.nus.edu.sg/uploads/Site/lect1gp.pdf` as to how different flexible regression methods fit together:



6.1 Gaussian processes - additional references

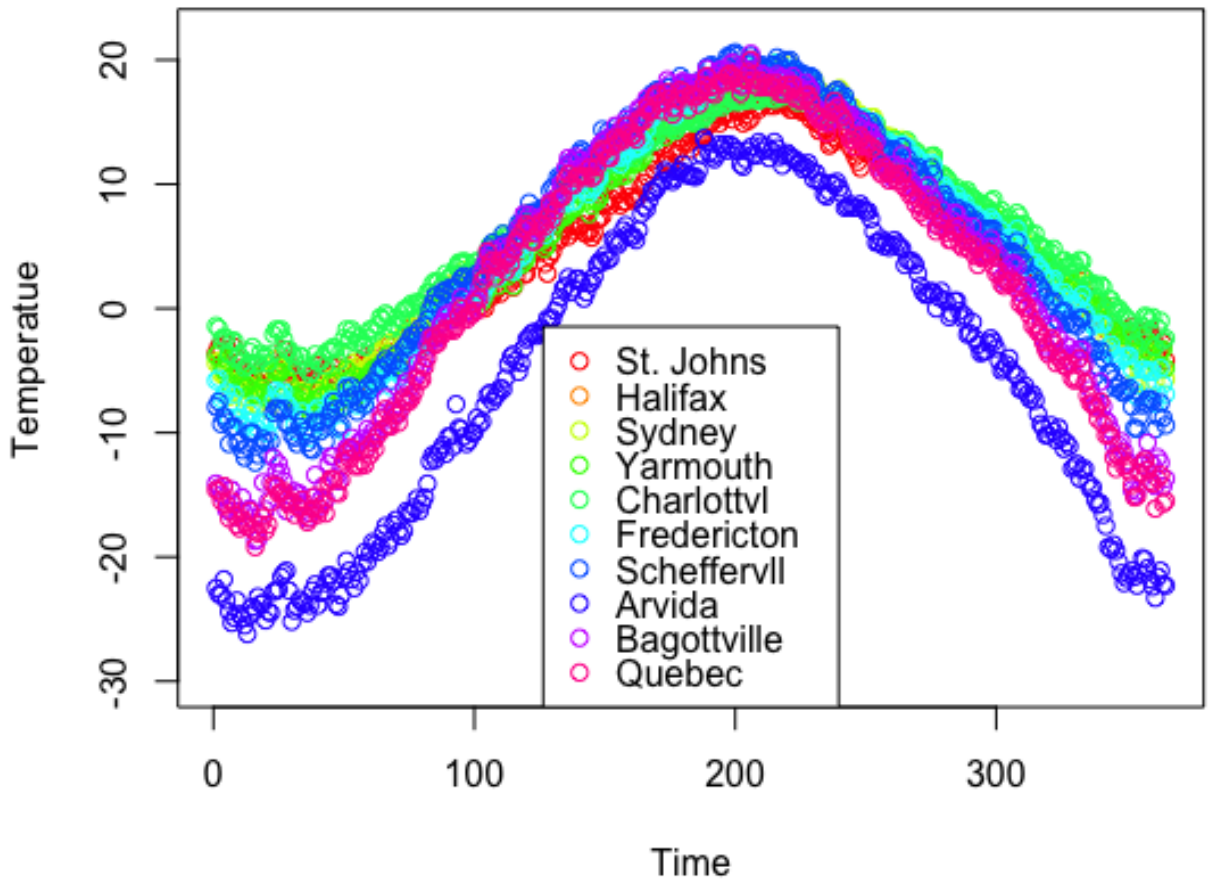
Neal, R. M. (1996) Bayesian learning for neural networks. Springer Verlag.

Rasmussen, C.E. and Williams, C.K.I. (2006) Gaussian Processes for Machine Learning. MIT Press

Rasmussen CE (2011) "The Gaussian Process Website" `www.gaussianprocess.org`

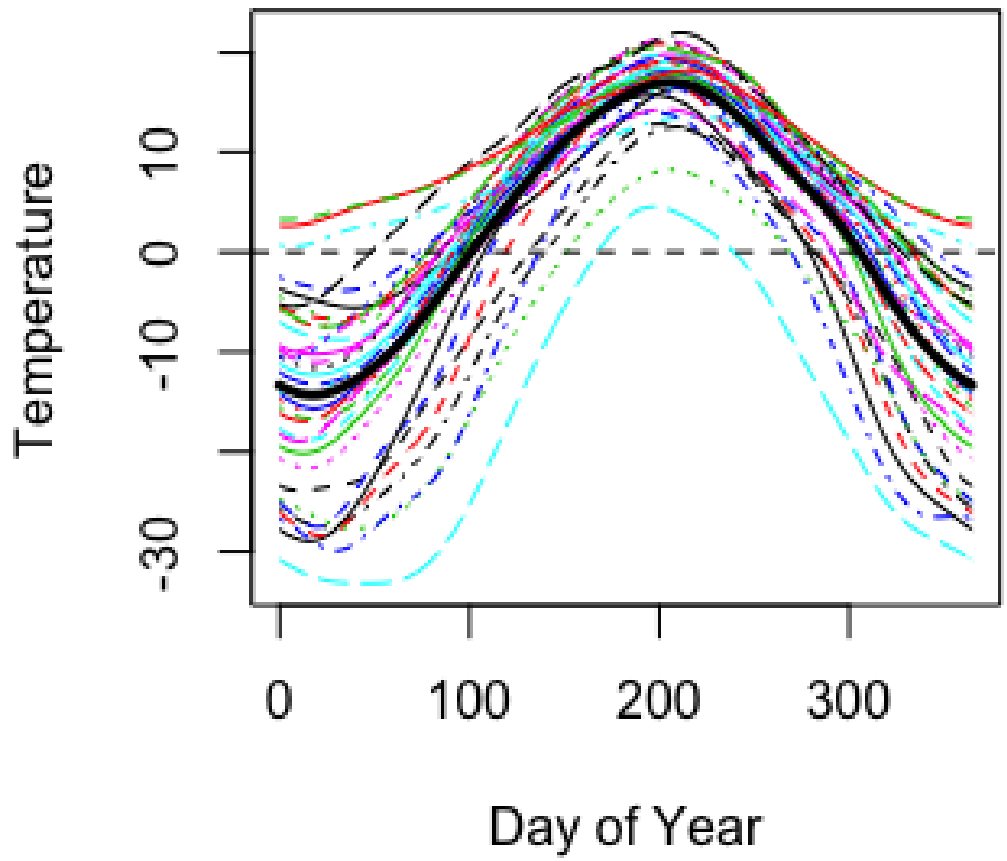
6.2 Functional Data Analysis (FDA) - example

Extending the previous approaches to summarise temperature across multiple seasonal patterns.



6.2 Functional Data Analysis (FDA) - example

Extending the previous approaches to **summarise temperature across multiple seasonal functions**.



6.2 Functional data analysis (FDA)

For example, a response might be in the form of a function collected by a monitoring device which effectively collects data continuously over time at several different locations.

- An application monitoring air pollution at 40 points over a city;
- The sensor records every 15 mins;
- For each sensor/location we have a time series, and hence we have 40 time series';

Although in practice the data may be discretised on a grid of time points for each location, it can be helpful to think of this as representing a function.

For this example, we would have 40 curves to analyse (one at each location).

6.2 FDA

- The functions are smooth, usually meaning that one or more derivatives can be estimated and are useful.

- No assumptions, such as stationarity, low dimensionality, equally spaced sampling points, etc, are made about the functions or the data.

6.2.1 Functional data methods

There are functional counterparts to standard statistial approaches:

- summary statistics;
- analysis of variance;
- multiple regression analysis;
- principal components analysis;
- canonical correlation analysis;
- cluster and classification analysis;

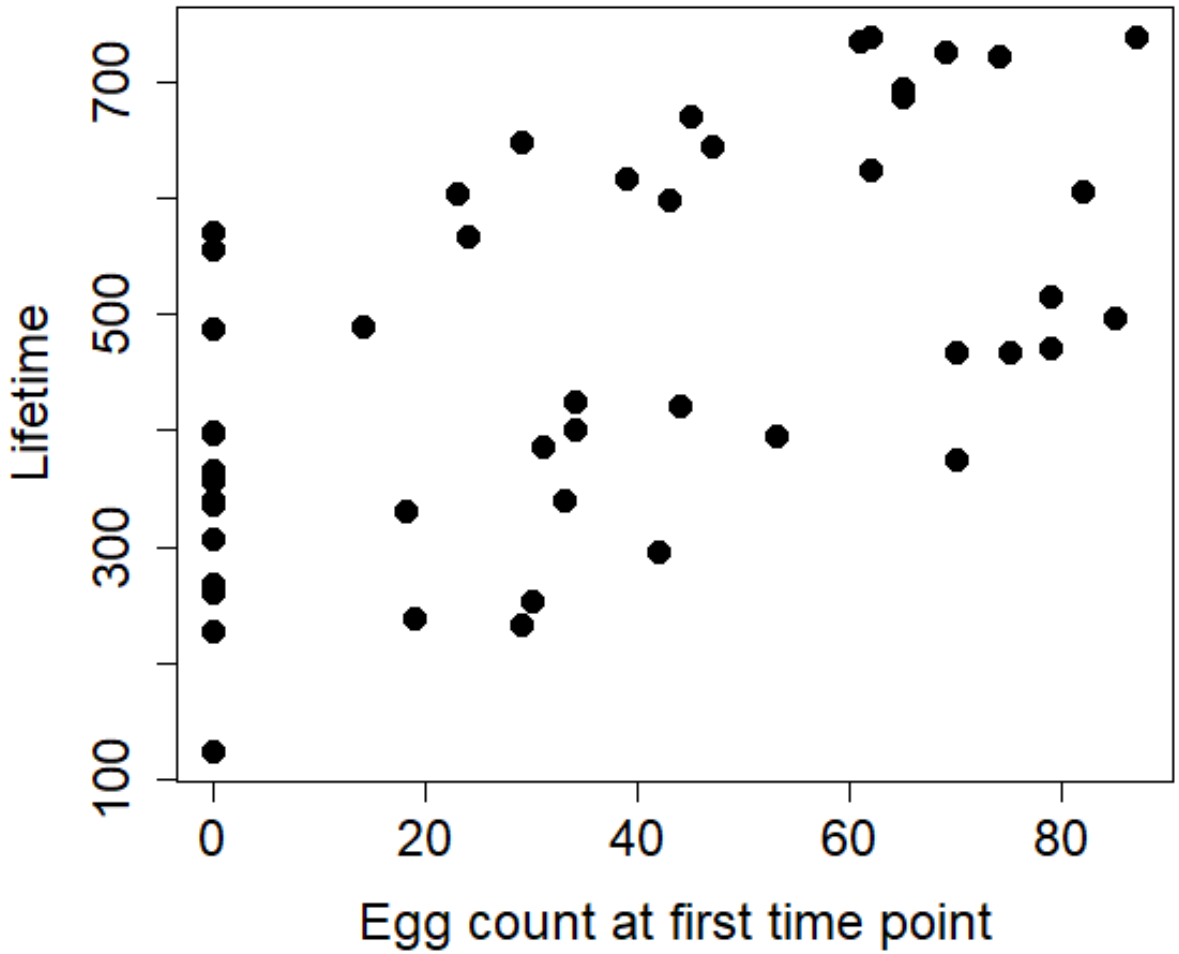
one way to think of functional data analysis is that it combines ideas of smoothing and multivariate statistics.

6.2.1 FDA methods: example 6.2.1

Mediterranean fruit flies

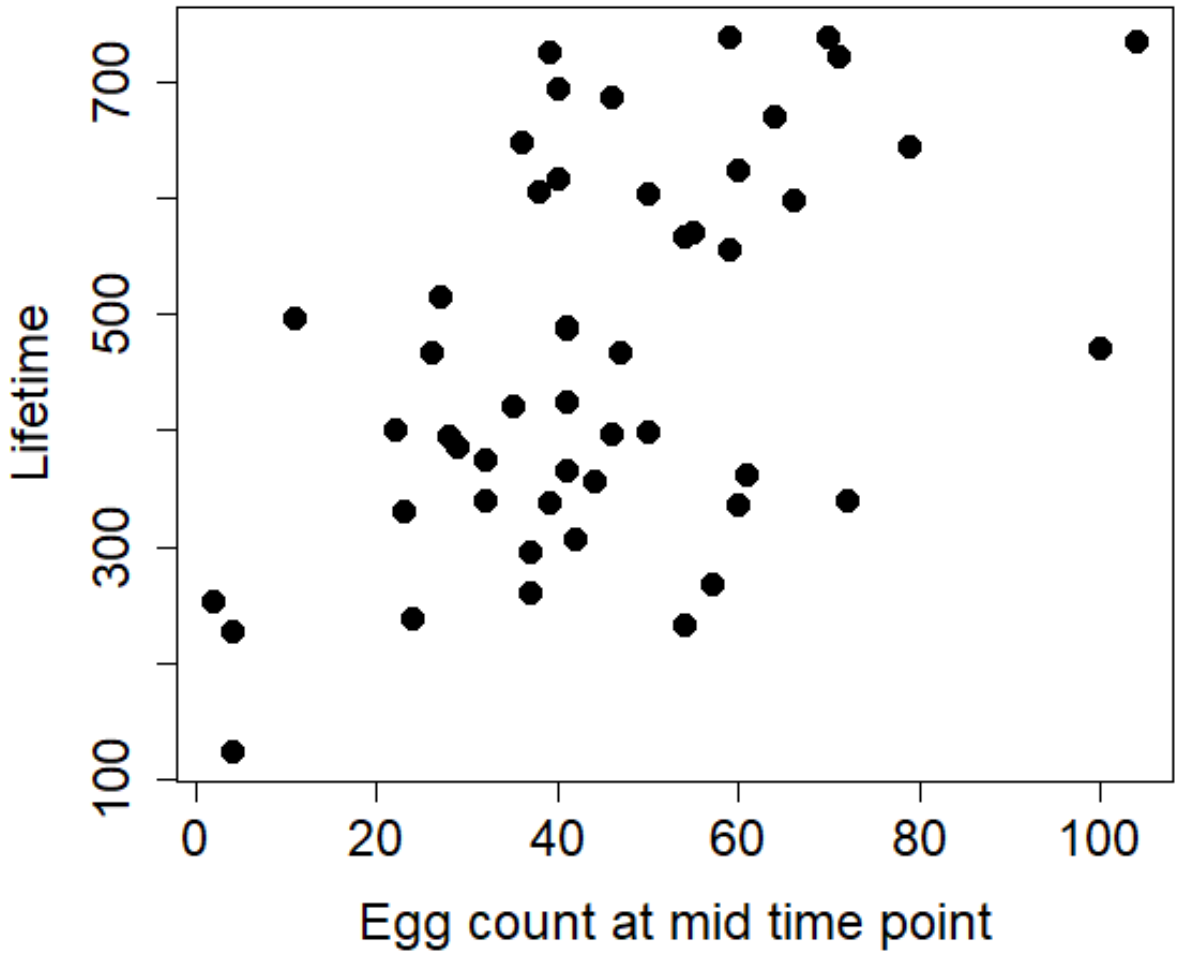
- We want to predict the future life time of flies.
- Data available: the record of number of eggs laid in the last 25 days for 50 flies.
- Can we use these to predict life time of flies?

6.2.1 FDA methods: example 6.2.1



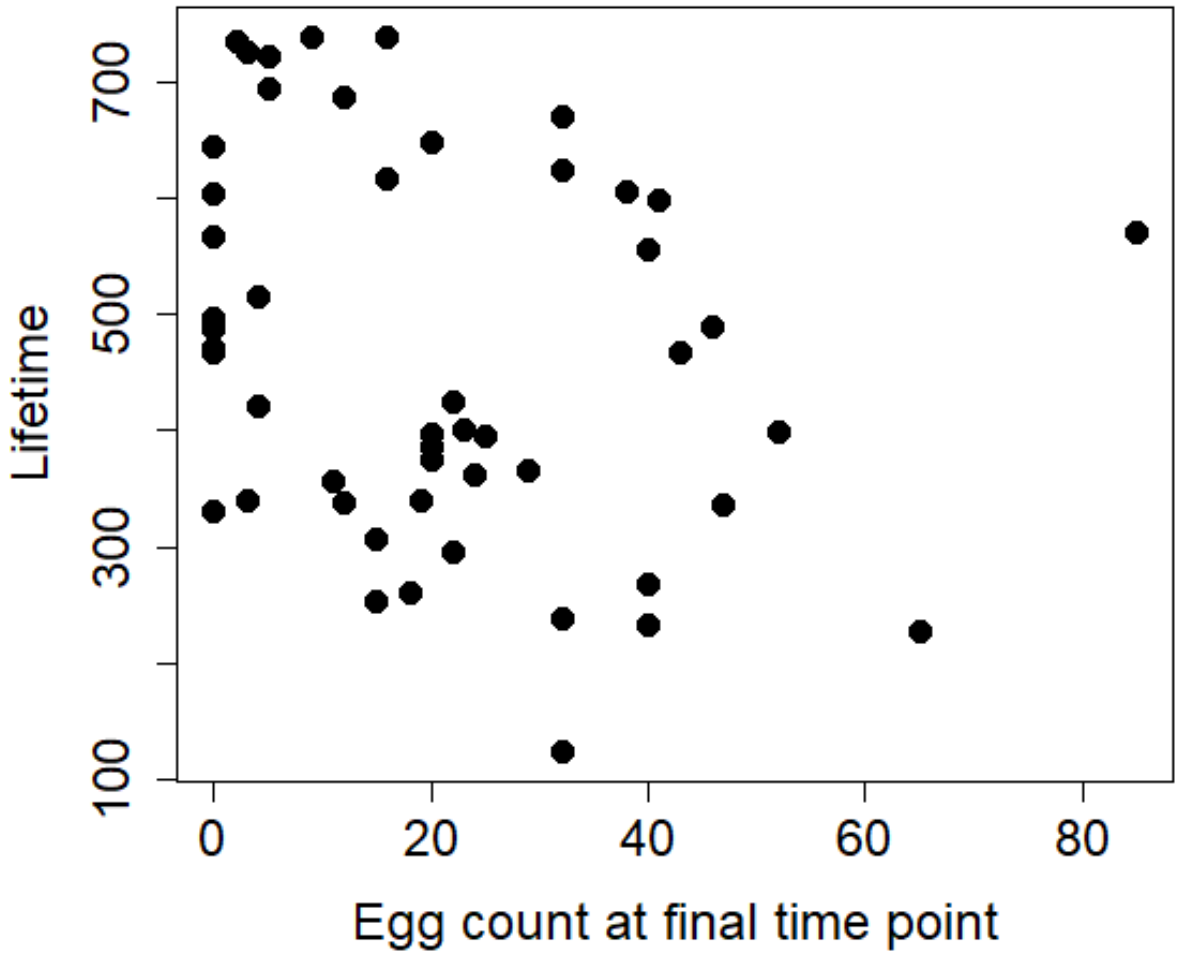
Lifetime of fly (y) for each of 50 flies against number of eggs laid by each fly on day 1 of the 25 days (x).

6.2.1 FDA methods: example 6.2.1



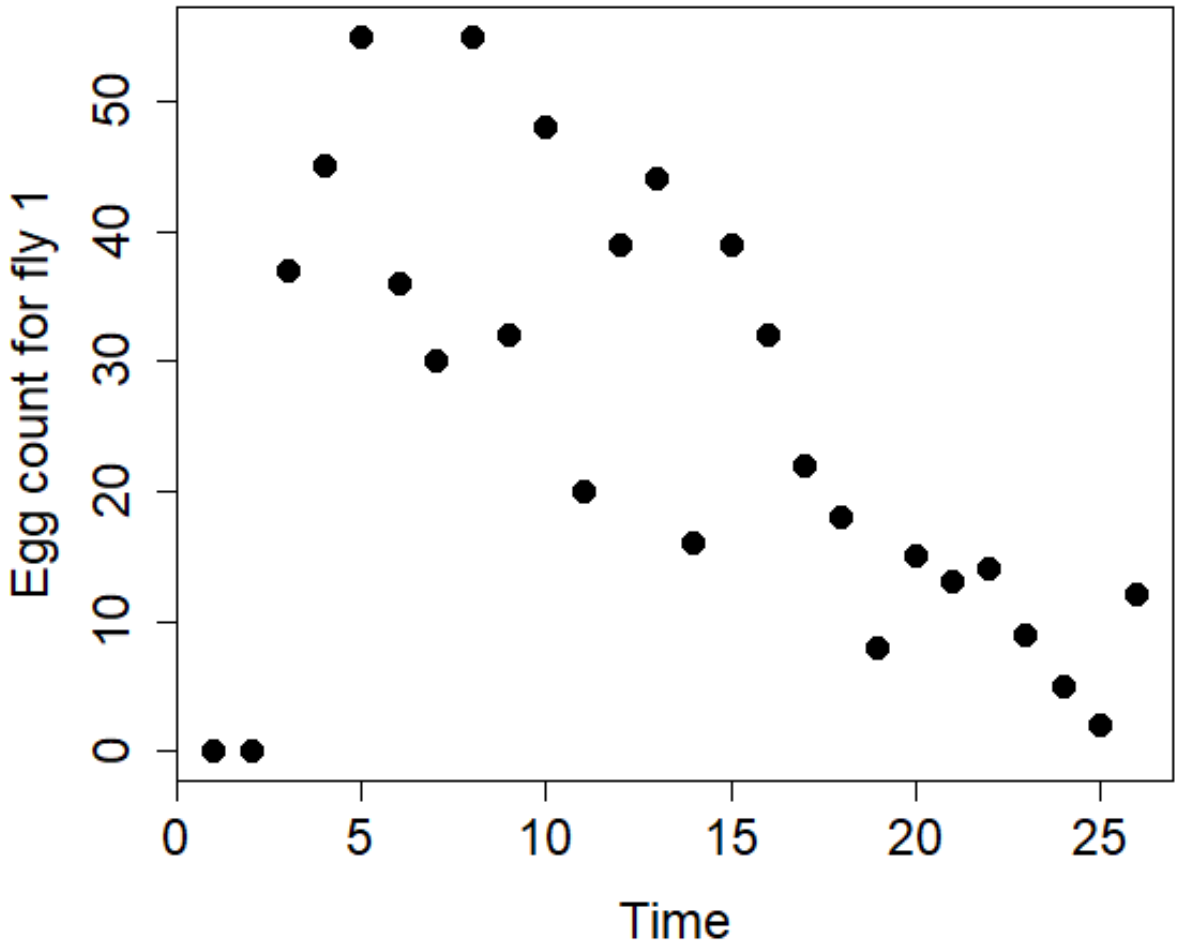
Lifetime of fly (y) for each of 50 flies against number of eggs laid by each fly at mid time point of the 25 days (x).

6.2.1 FDA methods: example 6.2.1



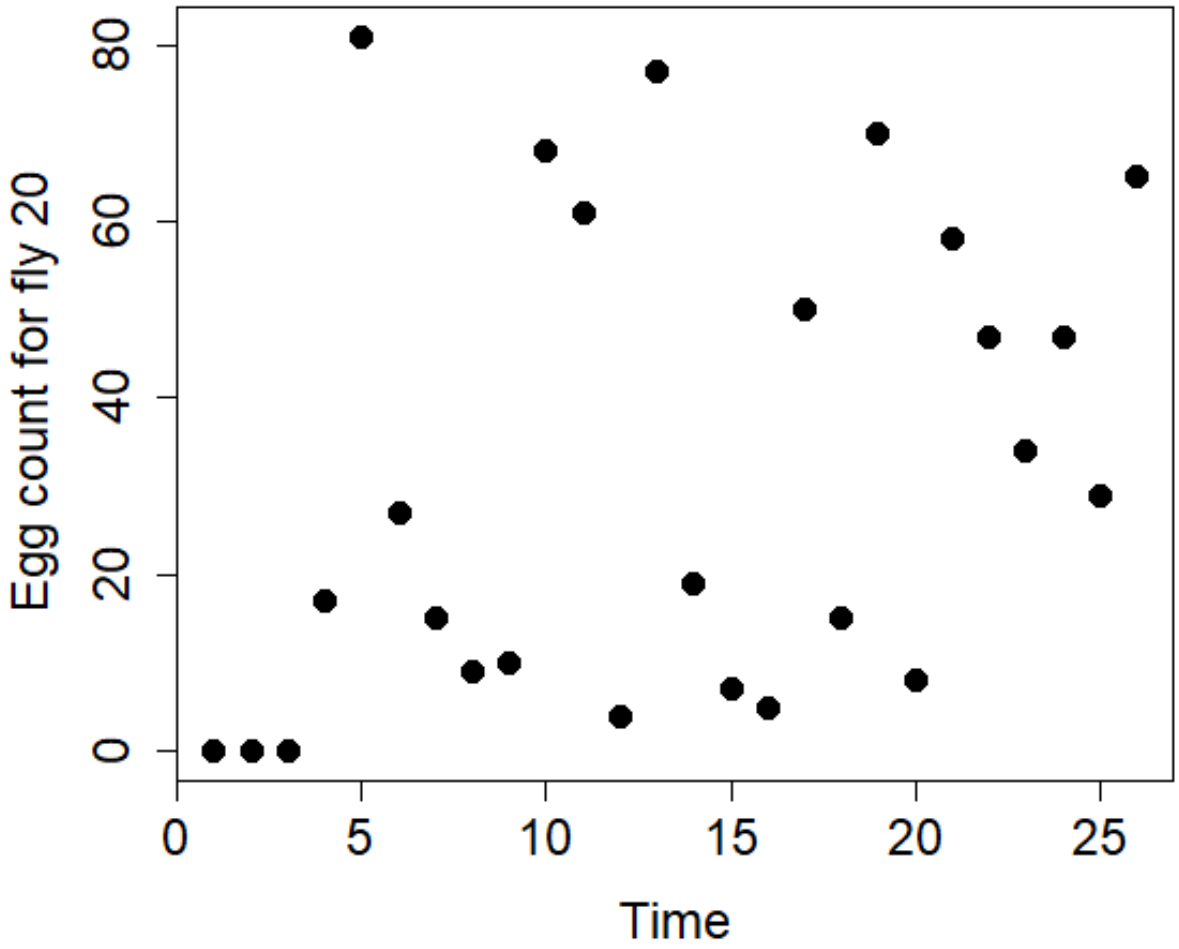
Lifetime of fly (y) for each of 50 flies against number of eggs laid by each fly on day 25 of the 25 days (x).

6.2.1 FDA methods: example 6.2.1



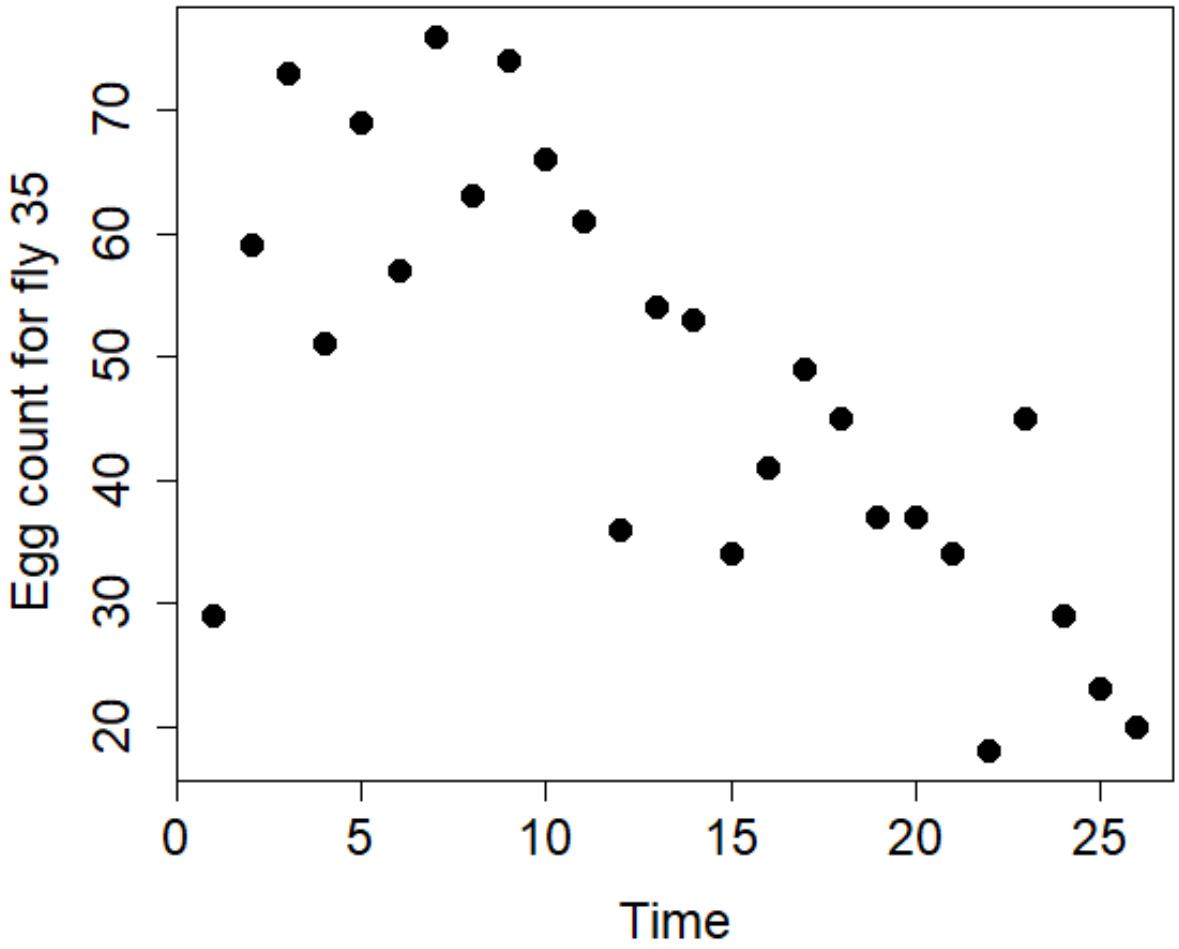
Number of eggs laid by fly 1 (y) on each of 25 days (x).

6.2.1 FDA methods: example 6.2.1



Number of eggs laid by fly 20 (y) on each of 25 days (x).

6.2.1 FDA methods: example 6.2.1



Number of eggs laid by fly 35 (y) on each of 25 days (x).

6.2.1 FDA methods: example 6.2.1

In this example the covariate is functional. Rather than having a single egg count we have a time series of 25 counts, i.e. our covariate is a function of time $x_i(t)$ for each of fifty fruit flies.

This suggests using a regression model of the form

$$\mathbb{E}(Y_i) = \int_0^{25} x_i(t)\beta(t)\, dt$$

to predict the future lifetime of the fly.

6.2.1 FDA methods

- estimate smooth functions from discrete noisy data;
- use basis function expansions to model functions, and impose smoothness using roughness penalties;
- $\rightsquigarrow$ a set of curves, one for each 'individual' in your study.
- derivatives can be used to investigate velocity and acceleration;
- often interested in separating phase and amplitude.

6.2.2 FDA: curve fitting

The usual starting poing is to estimate a smooth curve for each 'individual' using basis functions. Earlier we had that:

$$f(x) = \sum_j \beta_j B_j(x).$$

with basis functions B and basis coeffcients $\boldsymbol{\beta}$.

The main choices of basis functions are B-splines (introduced in chapter 2 and the basis of choice for most non-periodic data) and Fourier series (best for periodic data).

6.2.2 FDA: curve fitting

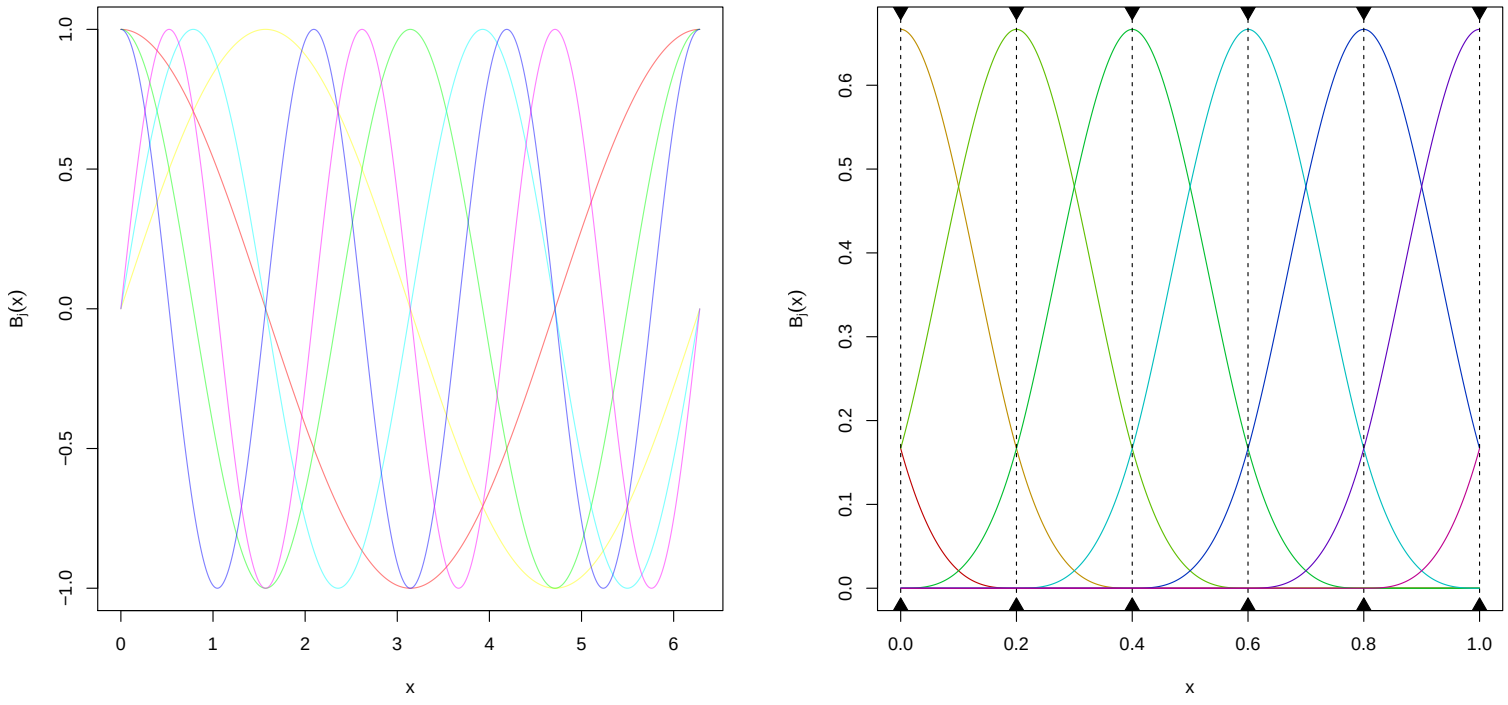


Figure: Examples of a Fourier basis and a B-spline basis.

6.2.2 FDA: curve fitting

Fourier basis

$$f(x_i) \approx \frac{a_0}{2} + \sum_{j=1}^{r} a_j \cos\left(\frac{2\pi j x_i}{P}\right) + b_j \sin\left(\frac{2\pi j x_i}{P}\right),$$

where $x_i \in (0, P)$. This approximation corresponds to using the design matrix

$$\mathrm{B} = \begin{pmatrix} \frac{1}{2} & \cos\left(\frac{2\pi x_1}{P}\right) & \sin\left(\frac{2\pi x_1}{P}\right) & \dots & \cos\left(\frac{2\pi r x_1}{P}\right) & \sin\left(\frac{2\pi r x_1}{P}\right) \\ & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{2} & \cos\left(\frac{2\pi x_n}{P}\right) & \sin\left(\frac{2\pi x_n}{P}\right) & \dots & \cos\left(\frac{2\pi r x_1}{P}\right) & \sin\left(\frac{2\pi r x_1}{P}\right) \end{pmatrix}$$

with $\boldsymbol{\beta} = (a_0, a_1, b_1, \dots, a_r, b_r)$.

6.2.3 FDA: summary statistics

Functional mean and covariance

mean $\bar{x}(t) = \frac{1}{n} \sum x_i(t)$

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

6.2.3 Functional PCA

Instead of a covariance matrix Σ we have a surface $\sigma(x, t)$ for functions and the eigendecomposition is re-interpreted through the Karhunen-Loève decomposition:

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

with the ξ_i orthonormal, and providing the principal components, and the d_i providing the variance.

6.2.3 Functional PCA

The principal component scores are:

$$f_{ij} = \int \xi_i(t)[x_j(t) - \bar{x}(t)]dt.$$

The best way to obtain an idea of the variation for each component is to plot:

$$\bar{x}(t) \pm 2\sqrt{d_i}\xi_i(t).$$

6.2.3 Functional PCA

Functional regression models

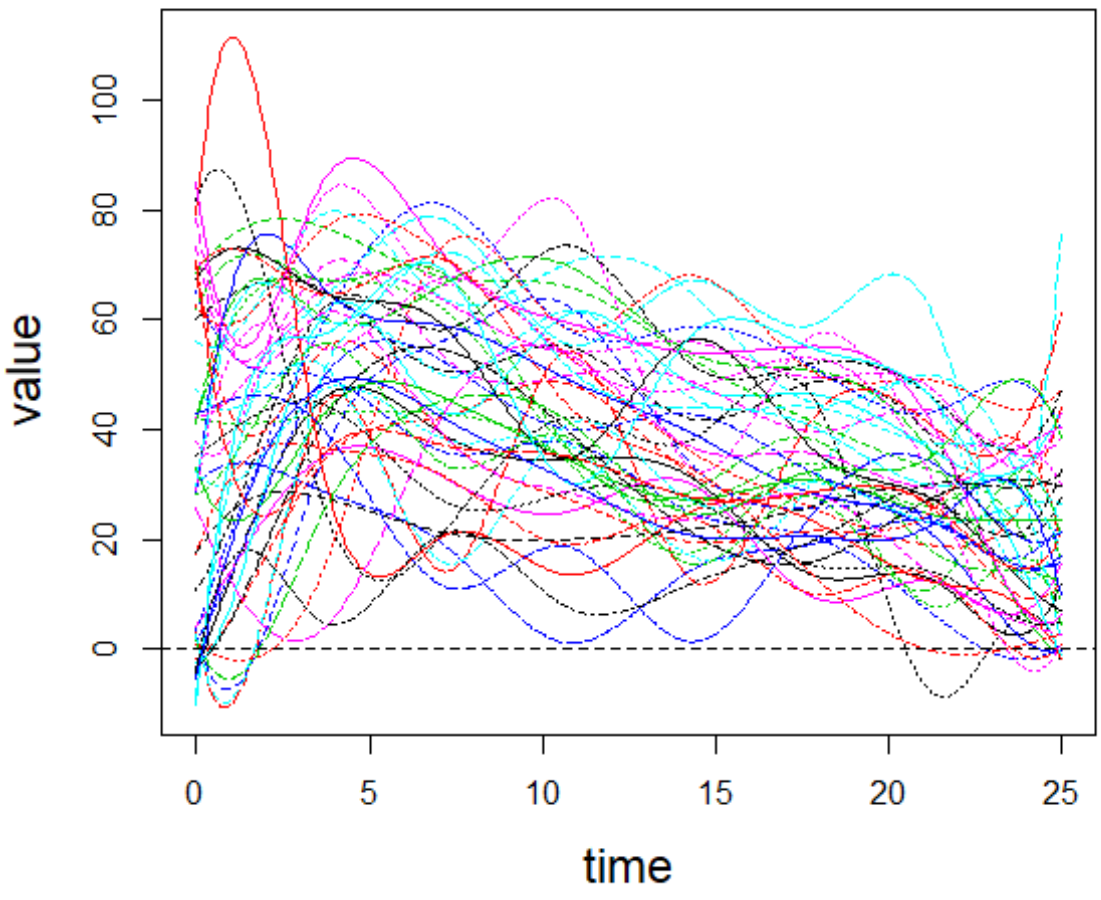
There are three types of model to consider:

1. Response is a function; covariates are multivariate.
2. Response is scalar or multivariate; covariates are functional.
3. Both response and covariates are functional.

6.2.4 FDA in R

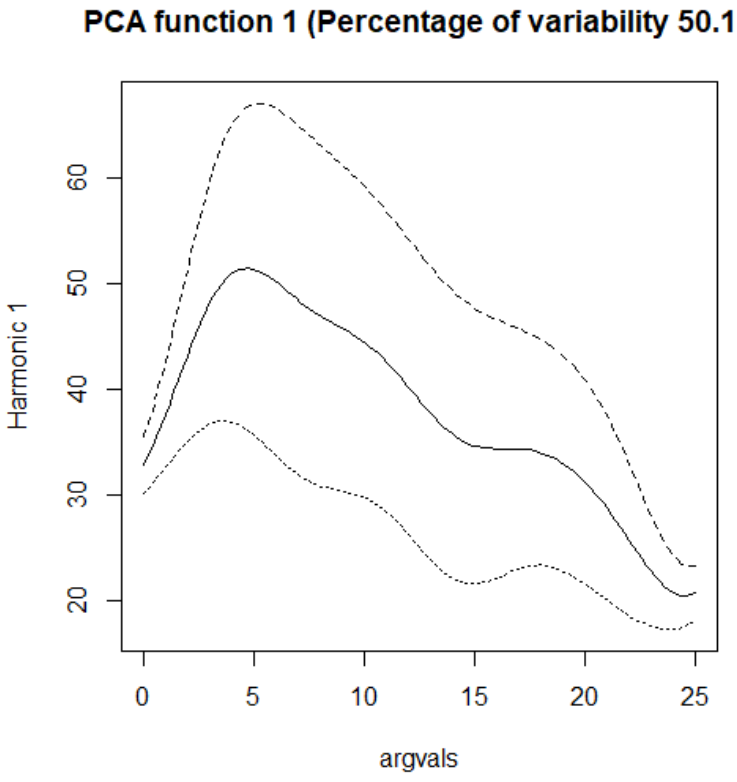
The `fda` packages in R can be used to implement functional summary statistics, functional pca and functional regression.

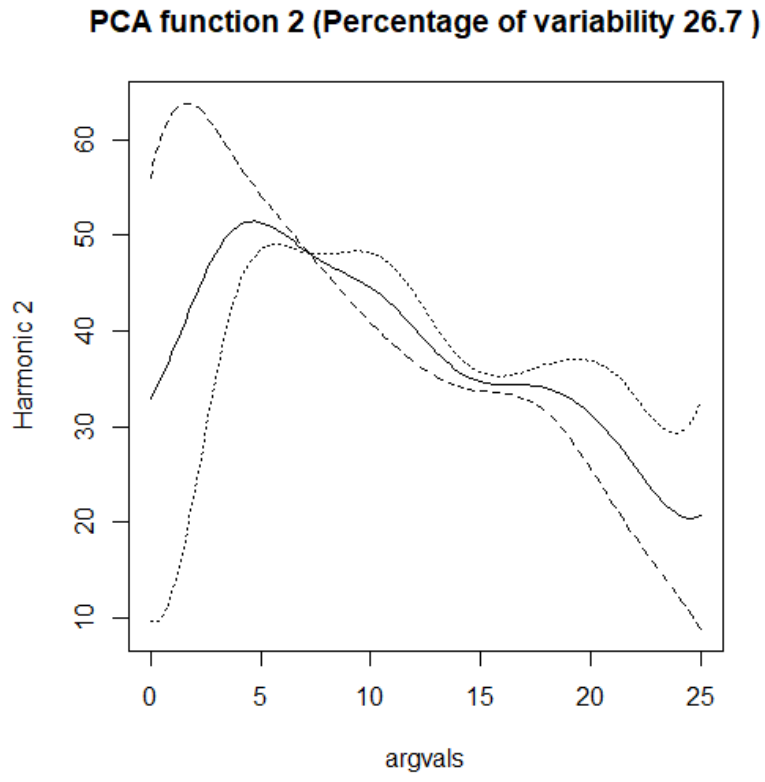
Smooth curves for number of eggs laid by each fly



6.2.4 FDA in R

Functional principal components





6.2.4 FDA in R

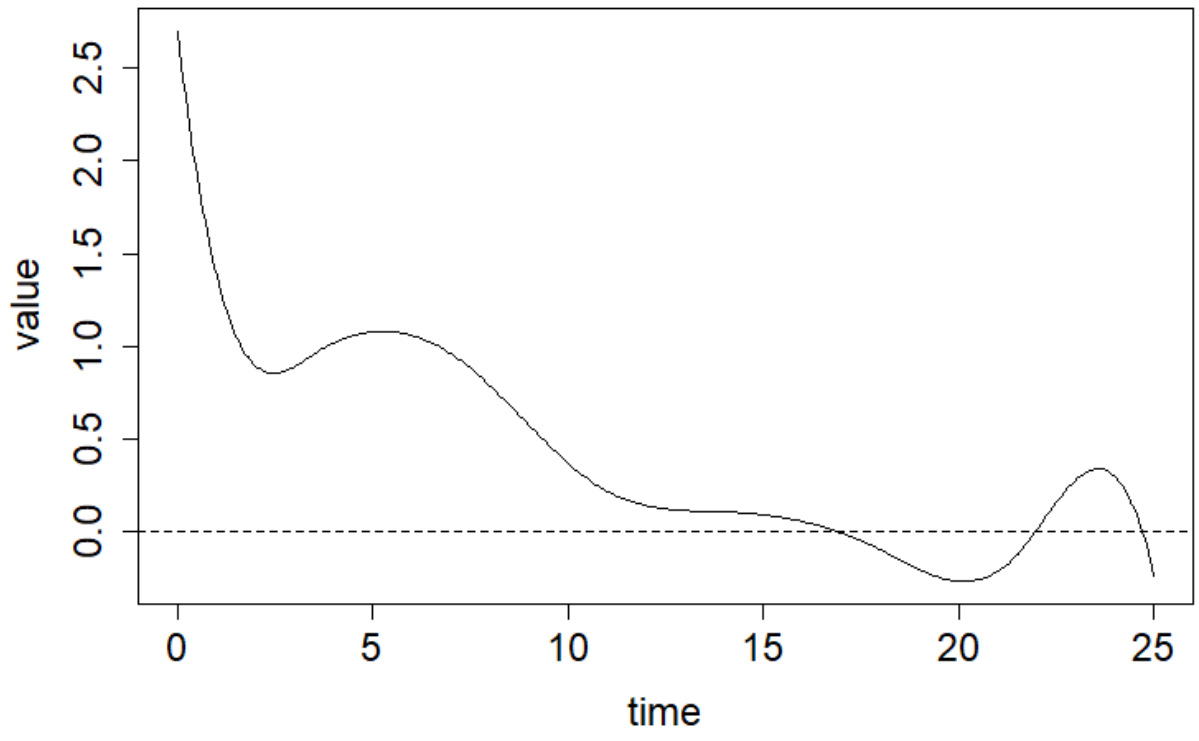
Functional regression

Finally we fit the functional regression model. Note that the regression coefficient is now itself a function (represented as a B-spline). In this example, the response *lifetime* is a scalar and the covariate is functional and so we have a model of the form:

$$Y_i = \beta_0 + \int_0^t \beta(t) x_i(t) dt + \varepsilon_i$$

6.2.4 FDA in R

Coefficients from a functional regression model

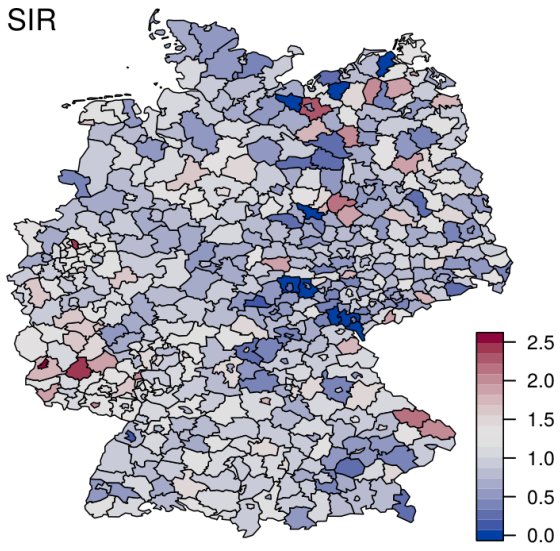


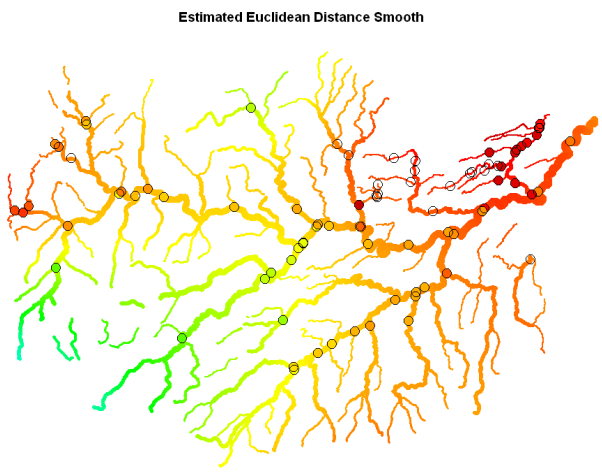
Regression coefficients are positive before around 15 days.

6.3 Other flexible regression models

For example, spatial data

- point processes or discrete/areal units - the latter of which are very often investigated using CAR models.
- connected network data e.g. from a river.





6.3 Other flexible regression models

Mixture models

In the preliminary material we introduced estimation using density functions, this can be extended along with work on Dirichlet processes to provide a nonparametric representation for mixture models.

6.3 Other flexible regression models

Neural networks

From a computational viewpoint the artifical intelligence approach of neural networks can be seen as an alternative (similar and sometimes more flexible) approach to fitting additive models.

Neural networks are basically nonlinear models.

6.3 Other flexible regression models

Therefore, there is very much more to explore and develop in this field for those of you that are interested.....