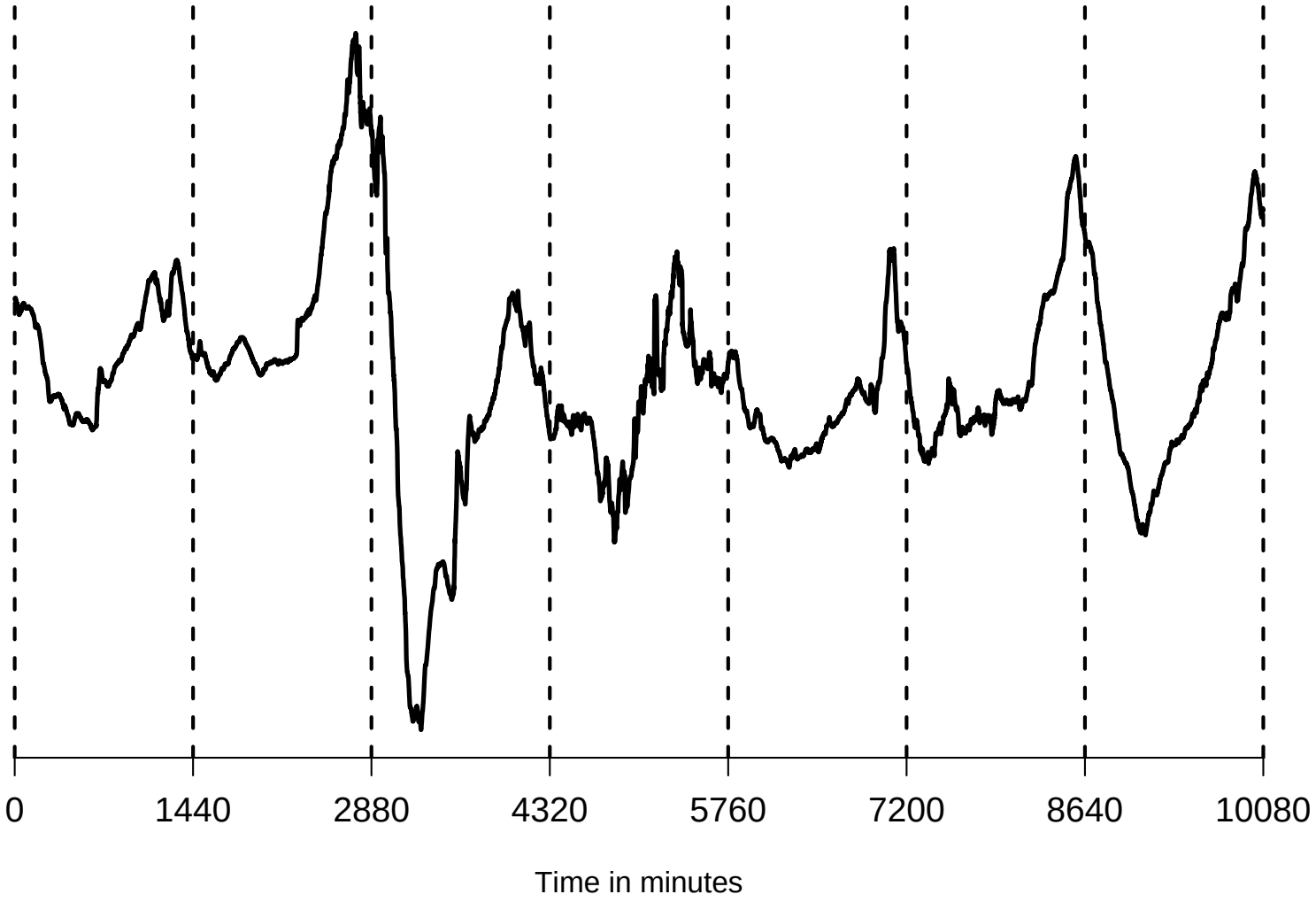
Fundamental concepts of functional data analysis

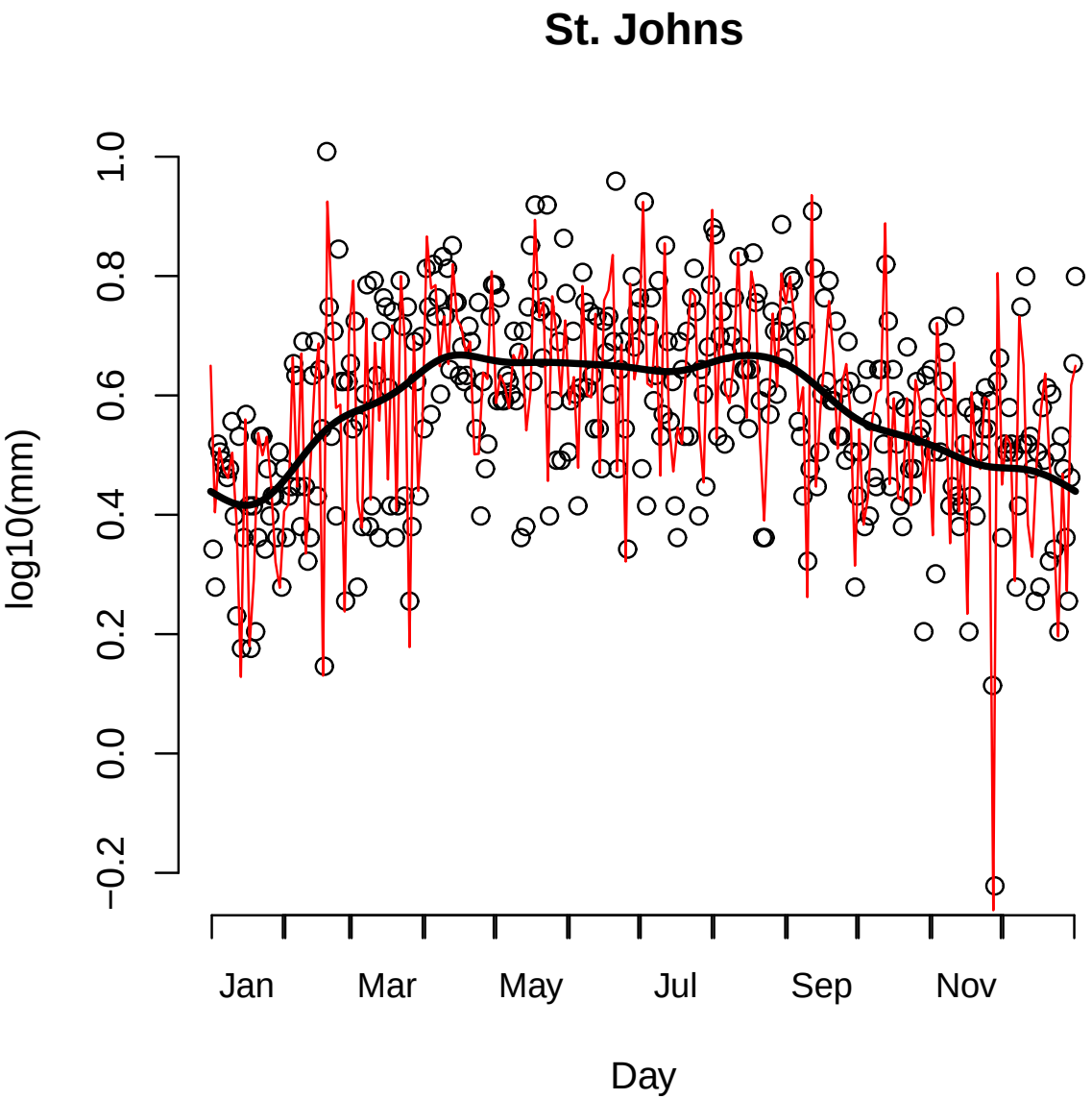
Piotr Kokoszka

Department of Statistics, Colorado State University

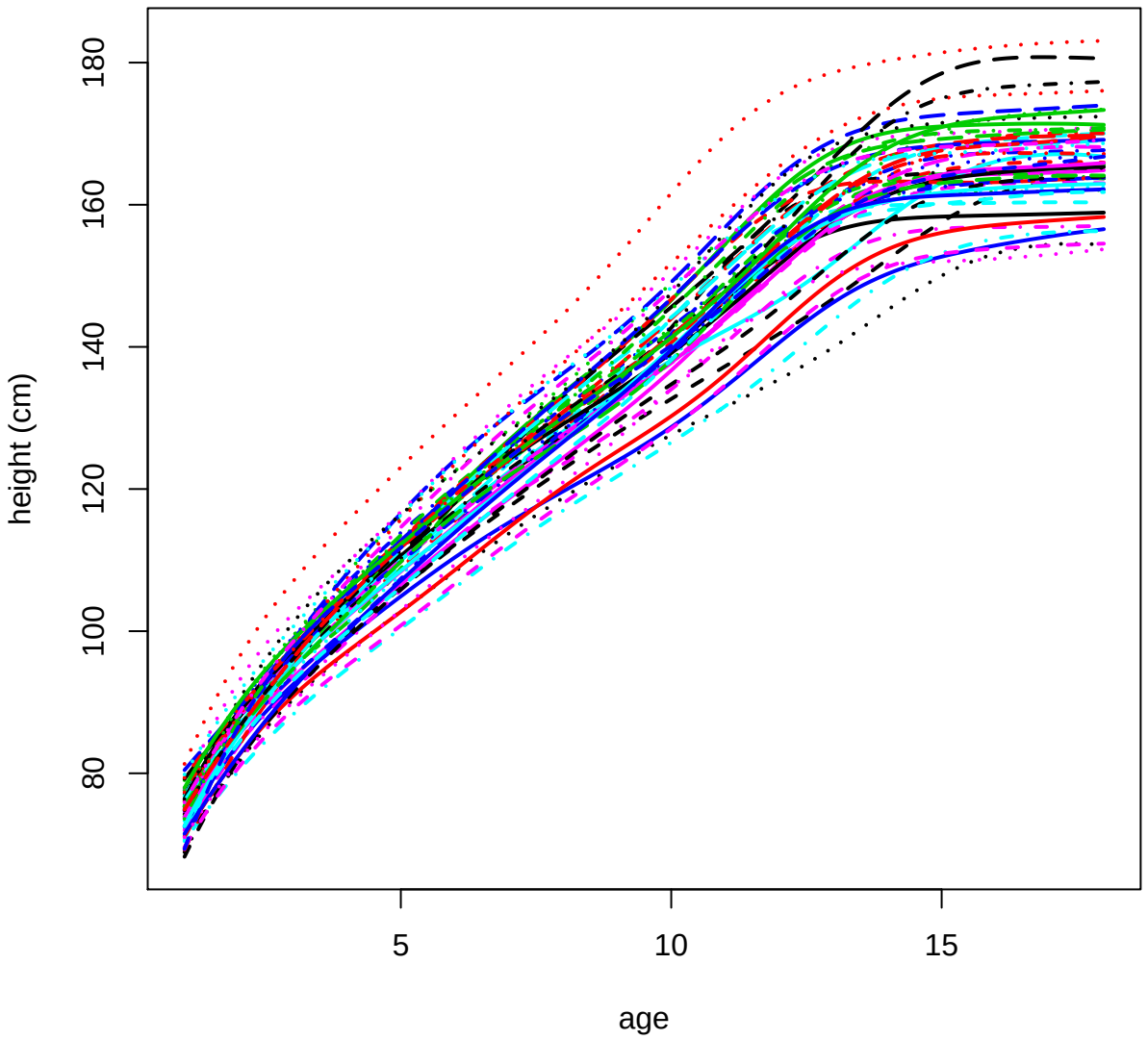
Examples of functional data



The horizontal component of the magnetic field measured at Honolulu magnetic observatory.



Average log–precipitation at St. Johns, Canada. The thick line shows a smooth.



Smoothed growth curves of 54 girls age 1–18.

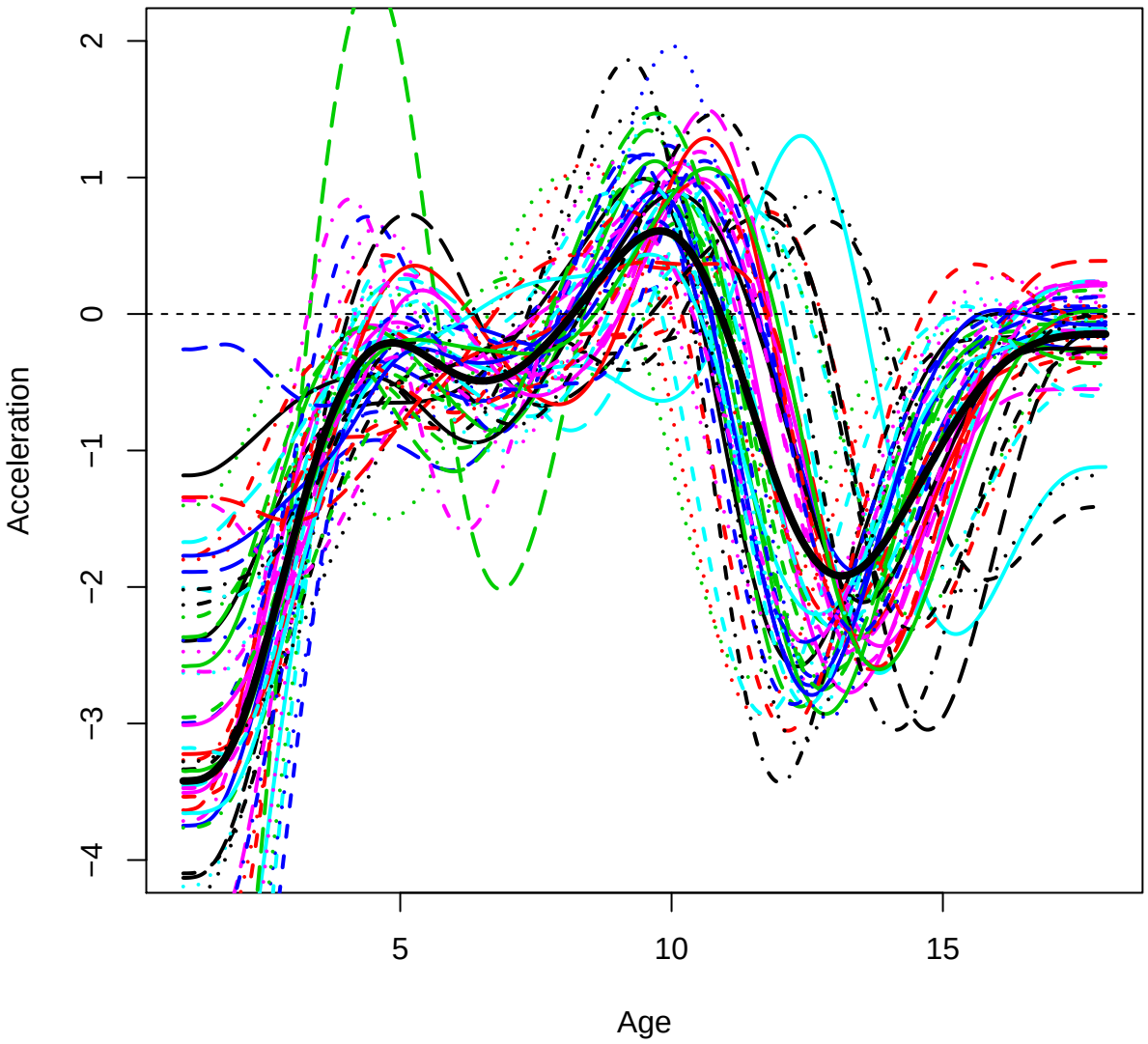
Expansions

Smoothing and dimension reduction have been major themes of FDA research.

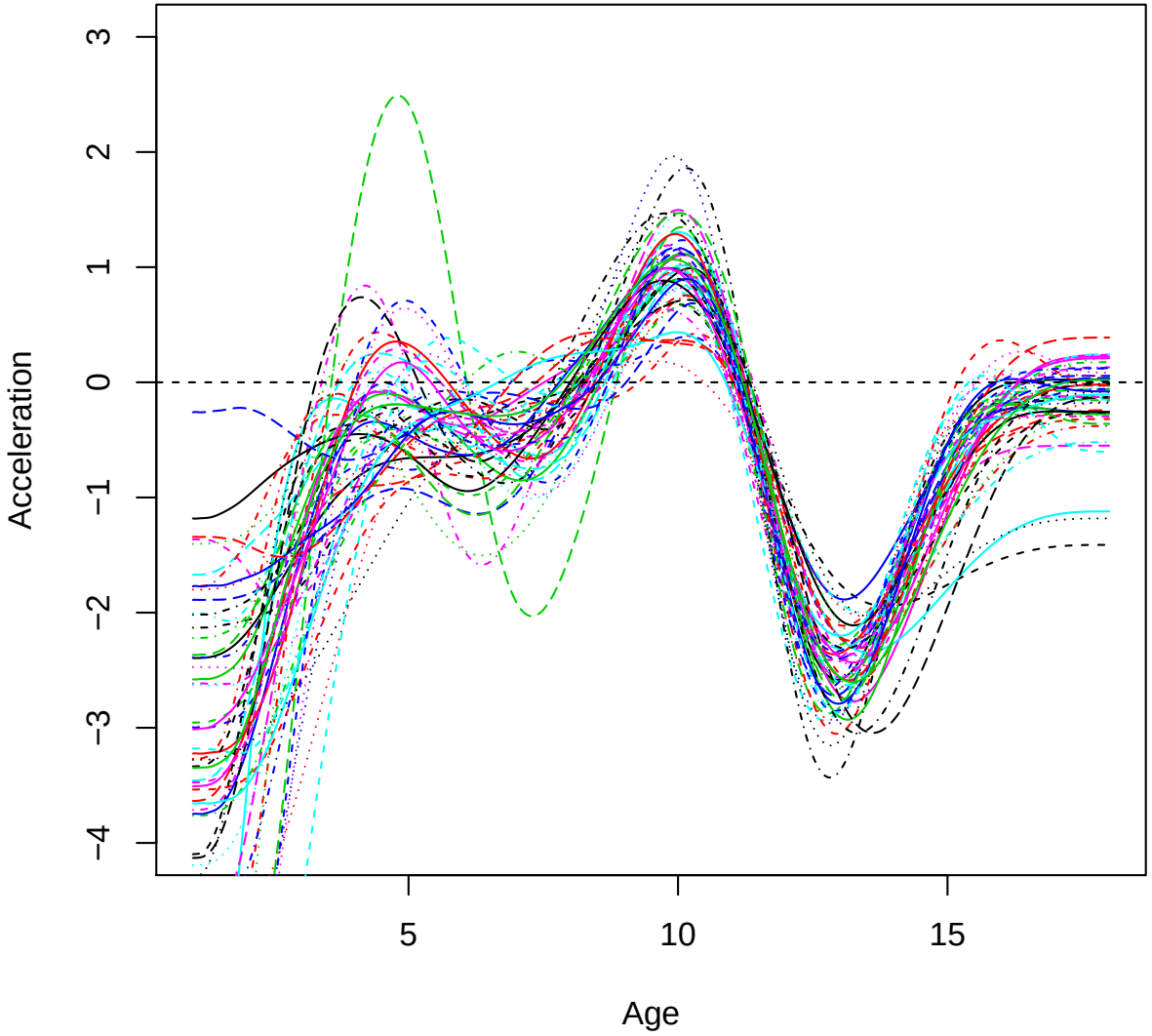
Basis expansion:

$$X_n(t) \approx \sum_{m=1}^{M} c_{nm} B_m(t), \quad 1 \leq n \leq N.$$

Derivatives can be computed.



Acceleration curves of 54 girls with the mean function in bold.



Continuously registered acceleration curves of the 54 girls.

Statistical models

Functional response model:

$$Y_i = \int \psi(s) X_i(s) ds + \varepsilon_i, \quad 1 \leq i \leq N.$$

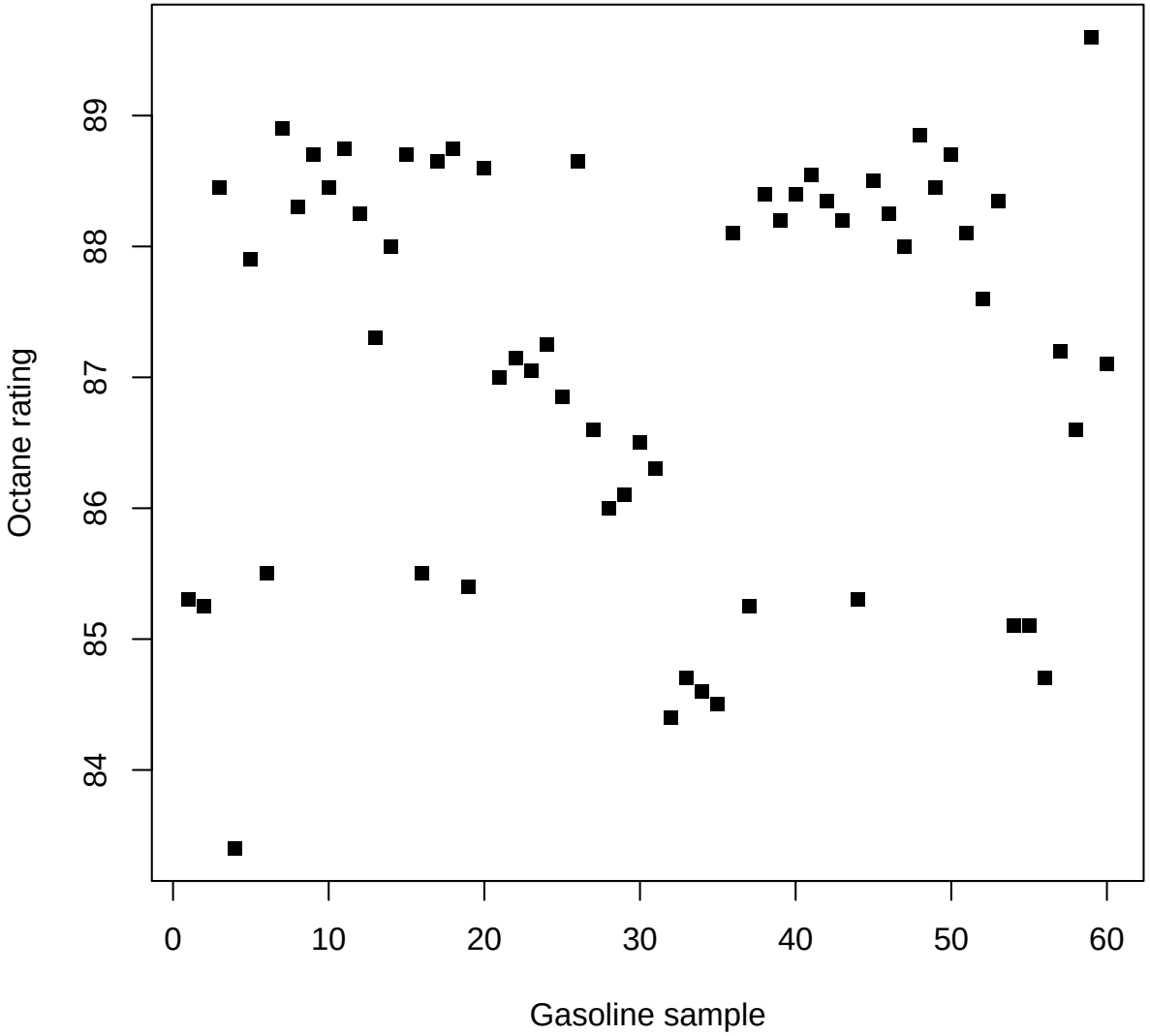
Function–on–function regression:

$$Y_i(t) = \int \psi(t,s) X_i(s) ds + \varepsilon_i(t).$$

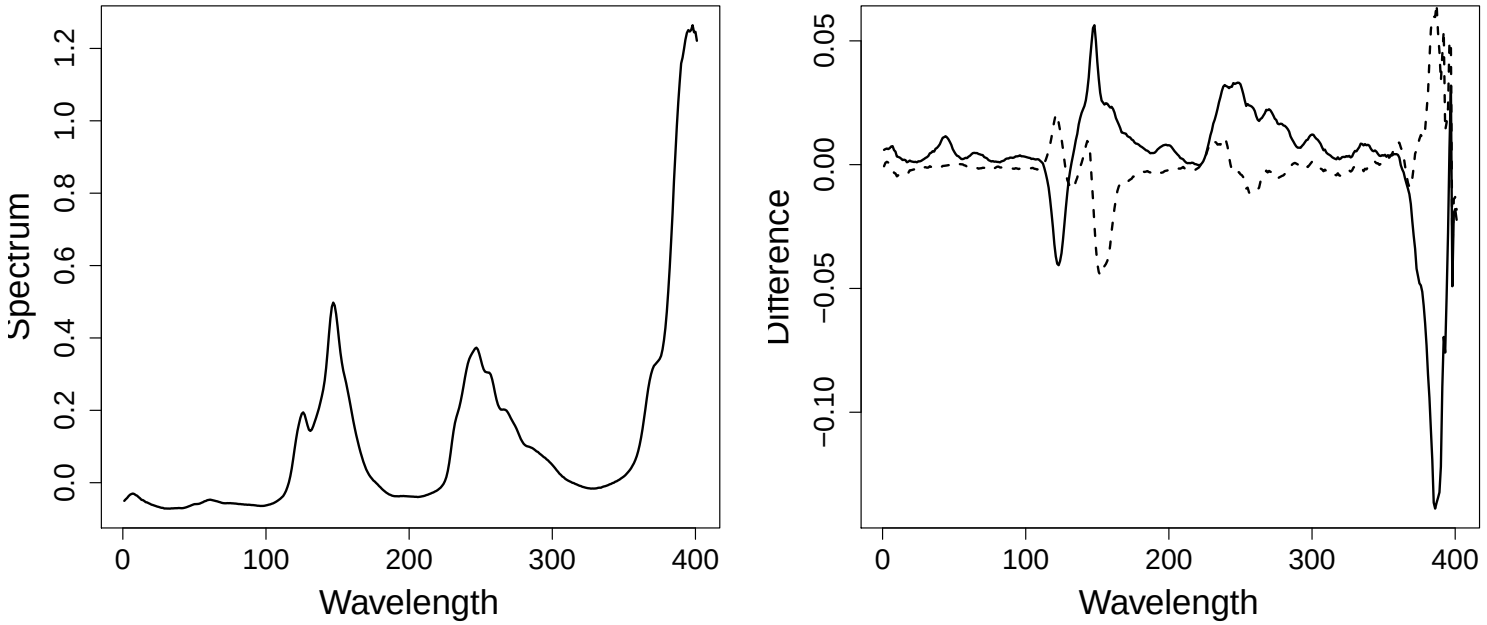
Autoregression:

$$X_i(t) = \int \psi(t,s) X_{i-1}(s) ds + \varepsilon_i(t).$$

Dimension of functional parameters, e.g. ψ, is larger than the sample size N. Regularization has been a major theme of FDA research.

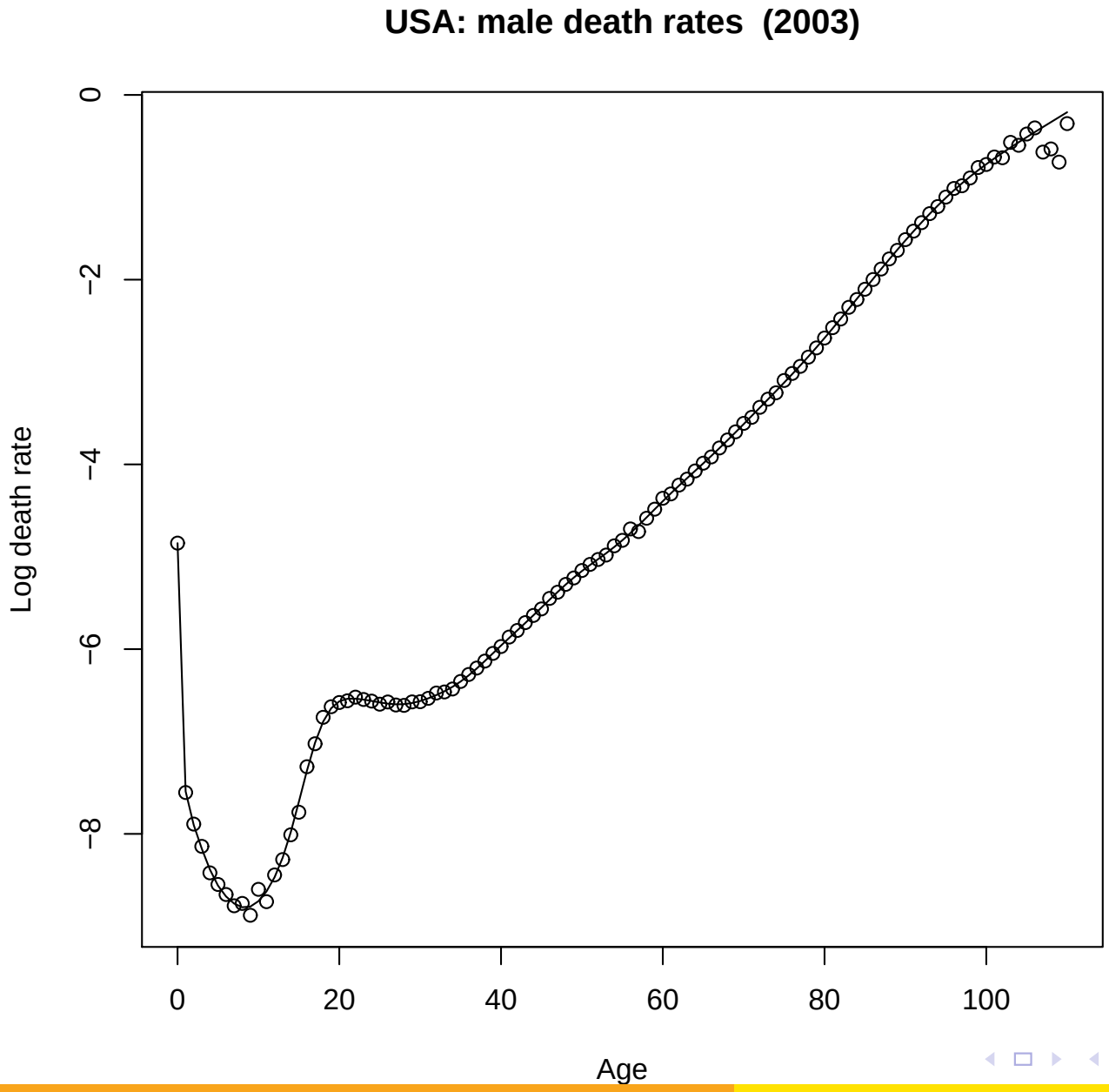


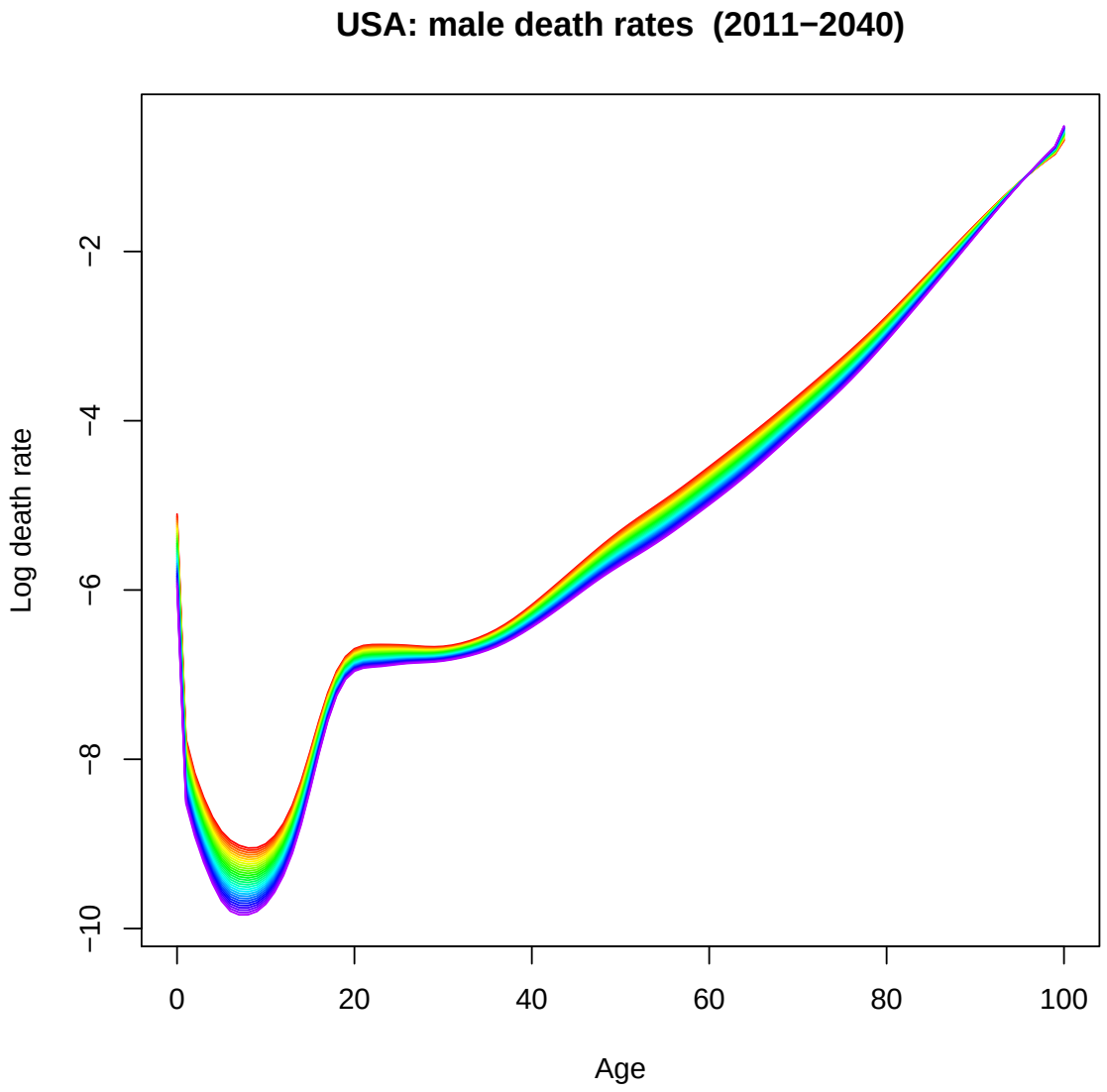
Octane ratings of 60 gasoline samples.



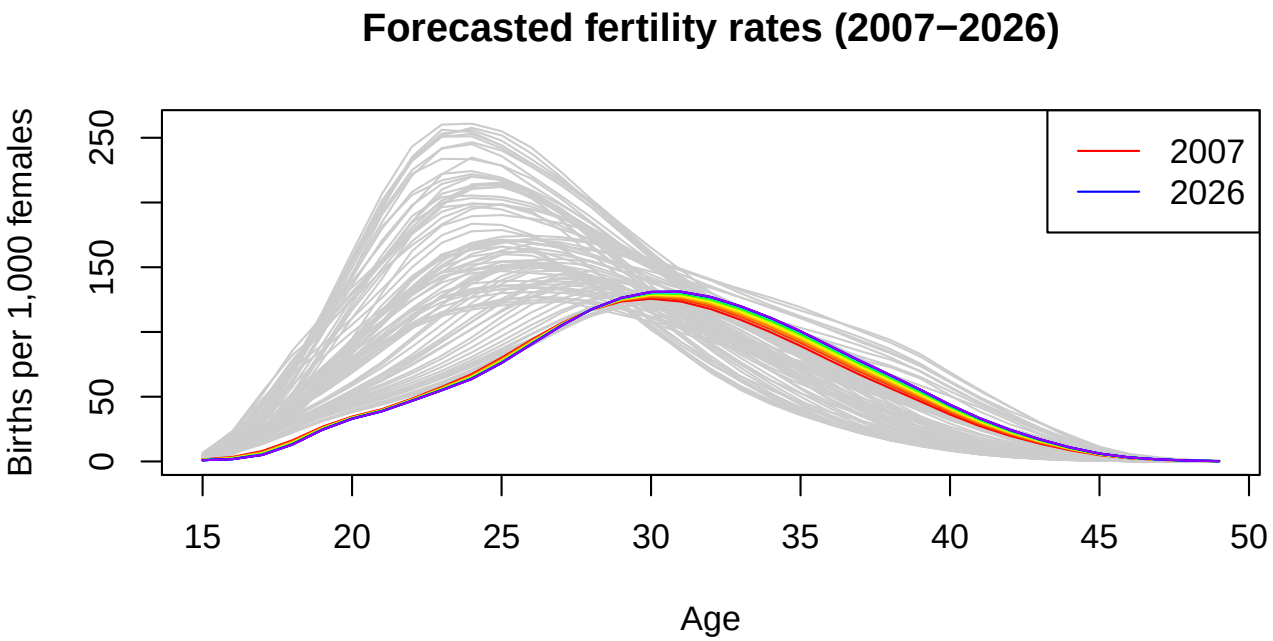
Left: Near infrared spectrum of a gasoline sample with index 1; Right: differences between the spectrum of the samples with indexes 2 and 1 (continuous) and 5 and 1 (dashed).

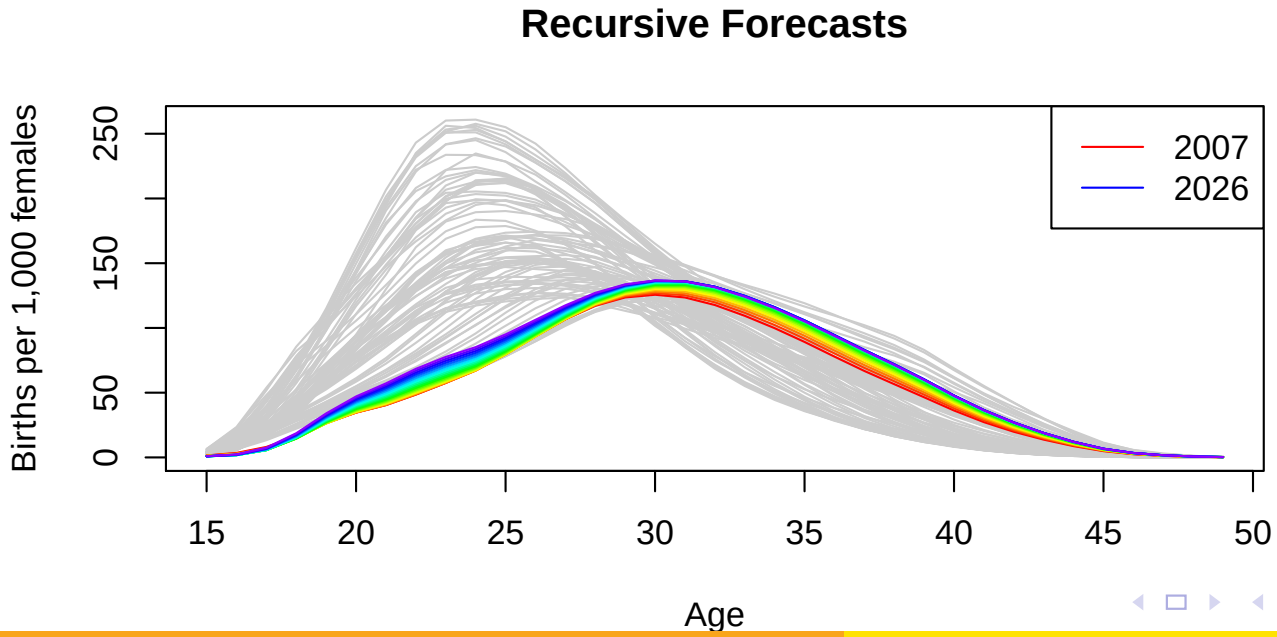
Time series of functions



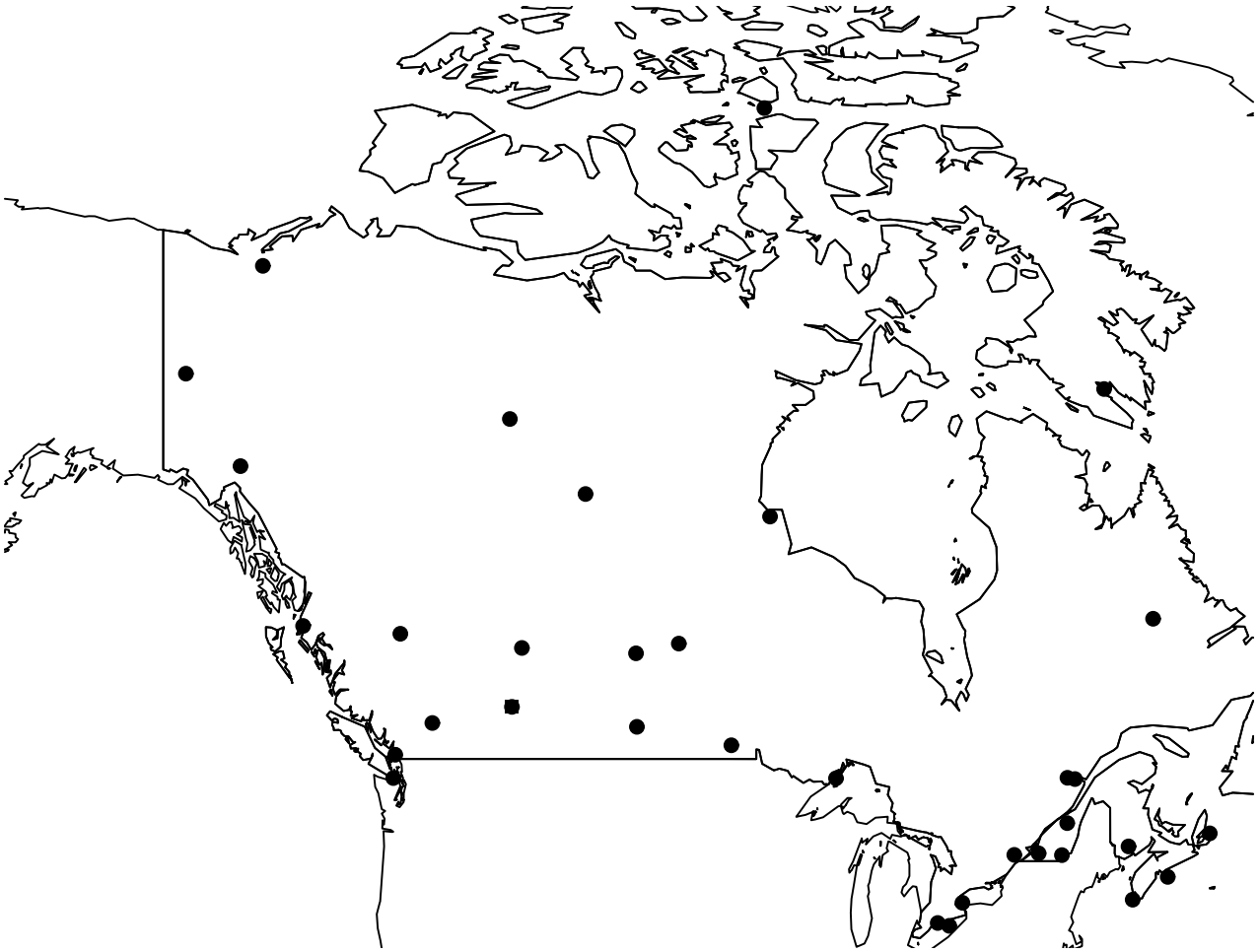


Predicted US male log mortality rate curves in rainbow code; years close to 2011 are in red, those close to 2040 in violet.

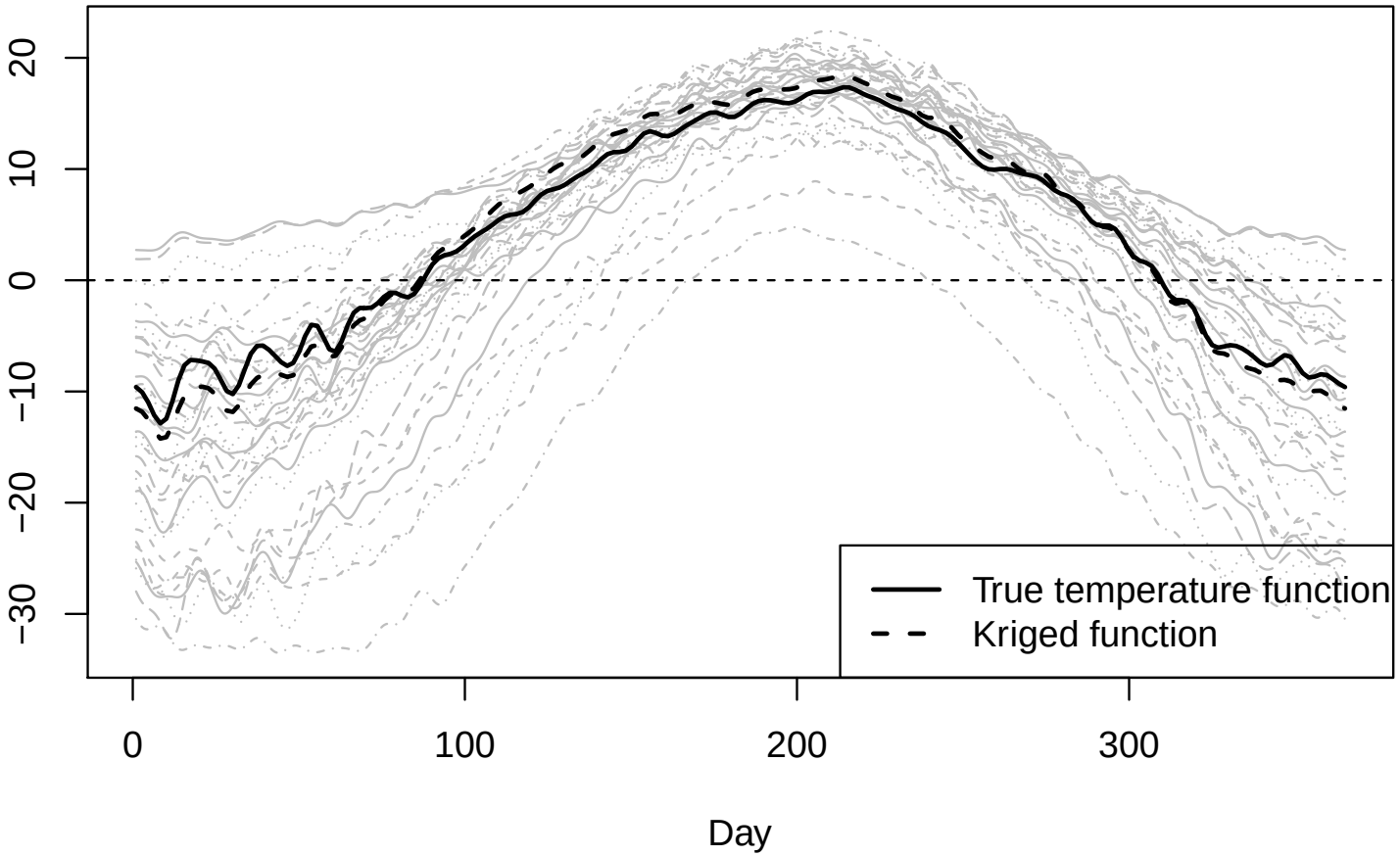

Forecasted fertility rates (2007−2026)
Births per 1,000 females
0
50
150
250
15
20
25
30
35
40
45
50
Age
2007
2026


Recursive Forecasts
Births per 1,000 females
0
50
150
250
15
20
25
30
35
40
45
50
Age
2007
2026

Spatial functional data



35 Canadian weather stations, Calgary marked with a square.



True and predicted (kriged) temperature functions at Calgary.

Some new directions

- Clustering and classification of huge sets of functions
(spectral profiles of stars, expression profiles of proteins)
- High dimensional functional regressions
($Y_i(t)$ protein expressions, $X_g(s)$ gene expressions, $1 \leq g \leq G$, G can be several thousand for one i.)
- Manifold domains (second generation FD)
(physical domains (brain), domains generated by restrictions)
- Multilevel dependence
(patient, visit, records from brain regions, scalar covariates)
- Connectivity and Network identification (possibly multilevel)
(Several types of measurements of brain activity, which regions interact at which levels?)
- Extreme events in set of functions indexed by time and space
($X_n(\mathbf{s}, t)$, temperature in year n at location $\mathbf{s}$ on day t, probabilities of heat waves or droughts).

Examples of asymptotic techniques

Functional observations are assumed to be functions in an L^2 space (so that variance-like objects can be defined).

$$\nu_p(X) = \left(E \left\{ \int X^2(t)dt \right\}^{p/2} \right)^{1/p} < \infty.$$

If $\nu_2^2(X) = E \int X^2(t)dt < \infty$, we say that X is *square integrable*.

We must often assume that $\nu_4(X) < \infty$.

Dependence in time series

Bernoulli shifts:

$$X_n = f(\varepsilon_n, \varepsilon_{n-1}, \ldots),$$

the ε_i are iid elements taking values in a measurable space S, and f is a measurable function $f : S^\infty \to L^2$.
Suppose $\{\varepsilon_i'\}$ are independent copies of $\{\varepsilon_i\}$ and set

$$X_n^{(m)} = f(\varepsilon_n, \varepsilon_{n-1}, \ldots, \varepsilon_{n-m+1}, \varepsilon'_{n-m}.\varepsilon'_{n-m-1}, \ldots)$$

Approximability condition:

$$\sum_{m=1}^{\infty} \nu_p\big(X_n - X_n^{(m)}\big) < \infty.$$

Estimation of second order structure

Covariance function ($EX_i = 0$):

$$c(t,s) = \mathrm{cov}(X_1(t), X_1(s))$$

Sample covariance function:

$$\hat{c}(t,s) = \frac{1}{N}\sum_{n=1}^{N}(X_n(t) - \bar{X}_N(t))(X_n(s) - \bar{X}_N(s)).$$

Under approximability with $p = 4$,

$$E\|\widehat{C} - C\|_{\mathcal{S}}^2 = O(N^{-1}).$$

(Hilbert–Schmidt norm)
Consistency, with the same rate, of the eigenfunctions follows.
These eigenfunctions, called *functional principal components* often form an optimal basis for expansions.

Invariance principle

Functional partial sum process:

$$V_N(x,t) = \frac{1}{\sqrt{N}} \sum_{n=1}^{\lfloor Nx \rfloor} X_n(t), \qquad 0 \le t, x \le 1.$$

Under approximability with $p = 2 + \delta$, there are Gaussian processes $\Gamma_N(x,t)$ such that for every N

$$\sup_{0 \le x \le 1} \| V_N(x,\cdot) - \Gamma_N(x,\cdot) \|_2 = o_p(1).$$

$$\{\Gamma_N(x,t), 0 \le x, t \le 1\} \stackrel{\mathcal{D}}{=} \{\Gamma(x,t), 0 \le x, t \le 1\},$$

$$\Gamma(x,t) = \sum_{i=1}^{\infty} \lambda_i^{1/2} W_i(x) \phi_i(t).$$

The W_i are independent standard Wiener processes, λ_i and ϕ_i are eigenvalues and the eigenfunction of the *long–run covariance function* (different from c above).