LECTURES
ON FUNCTIONAL DATA ANALYSIS

Hans-Georg Müller
UC Davis

Oviedo

August 2016

Supported by NSF

Lectures on FDA – Part IV

FUNCTIONAL REGRESSION

FUNCTIONAL REGRESSION MODELS

X	$\mapsto$	Y	
$\mathbb{R}^d$		$\mathbb{R}$	Multiple Regression, GLM
$\mathbb{R}^{d_1}$		$\mathbb{R}^{d_2}$	Multivariate Regression
L^2		$\mathbb{R}$	"Functional Predictor Models"
$\mathbb{R}^d$		L^2	"Functional Response Models"
L^2		L^2	"Function to Function Regression"

MODELING FUNCTIONAL PREDICTORS

1. Functional Linear Regression

Idea: Extending the multivariate linear regression model $E(Y|X) = BX$ to functional data $(X(t), Y)$ or $(X(t), Y(t))$:

$$E(Y|X) = \mu_Y + \int (X(s) - \mu_X(s))\beta(s)\, ds,$$

the functional linear regression model with regression parameter function β and scalar responses (Grenander 1950) (also generalized version by including link function (GFLM));

$$E(Y(t)|X) = \mu_Y(t) + \int (X(s) - \mu_X(s))\beta(s,t)\, ds,$$

model with functional responses (Ramsay & Dalzell 1991)

2. Functional Nonparametric Regression

$$E(Y|X) \;=\; \mu_Y + g(X)$$

for "smooth" function g, in analogy to nonparametric regression (Ferraty & Vieu 2006)

Problem: Curse of dimensionality, as predictor is infinite-dimensional. The infinite-dimensional curse can be quantified by using results on small ball probabilities for stochastic processes (Hall, M, Yao 2009).

$\Rightarrow$ Require functional regression models that fall between these extremes

PRINCIPAL COMPONENT REPRESENTATION OF FUNCTIONAL LINEAR REGRESSION

With predictor representations

$$X(s) = \mu_X(s) + \sum_{k=1}^{\infty} A_k \phi_k(s)$$

obtain from normal equations for the functional linear model (FLM) $E(Y|X) = \mu_Y + \int \beta(s)(X(s) - \mu_X(s))ds$:

$$\beta(s) = \sum_{k=1}^{\infty} \frac{E(A_k Y)}{E(A_k^2)} \phi_k(s) = \sum_{k=1}^{\infty} \beta_k \phi_k(s),$$

implying

$$E(Y|X) = \sum_k \beta_k A_k$$

FLM FOR FUNCTIONAL PREDICTORS AND RESPONSES

Extending the multivariate linear regression model $E(Y|X) = BX$ to functional data $(X(t), Y(t))$:

$$E(Y(t)|X) = \mu(t) + \int (X(s) - \mu_X(s))\beta(s,t)\, ds.$$

Estimation of the parameter function $\beta(\cdot,\cdot)$ is an inverse problem.

- Idea: Extending the least squares normal equation $\mathrm{cov}(X, Y) = \mathrm{cov}(X)B$.
- "Functional Normal Equation" (He et al. 2000,2003) For auto-covariance operator A_G of predictors X and

$$r_{XY}(s,t) = \mathrm{cov}\left[X(s), Y(t)\right] : \qquad r_{XY} = A_G\beta.$$

- Since A_G is a compact operator in L^2, equation is not invertible. Require functional generalized inverse: Well-defined under regularity conditions and obtained by regularization – truncation of included components or penalty (Cai & Hall 2006, Hall & Horwitz 2007).

Solution of the functional normal equation:

$$\beta^*(s,t) = \sum_{j,k=1}^{K} \frac{\mathrm{cov}(A_j, B_k)}{\mathrm{var}(A_j)} \phi_j(s)\psi_k(t).$$

Requires truncation with $K = K(n) \to \infty$ as $n \to \infty$

Existence of solution in image space of A_G.

REPRESENTATIONS OF FLR

With predictor and response representations

$$X(s) = \mu_X(s) + \sum_{k=1}^{\infty} A_k \phi_k(s),\ Y(t) = \mu_Y(t) + \sum_{m=1}^{\infty} B_m \psi_m(t)$$

obtain from normal equations for the model
$E(Y(t)|X) = \mu_Y(t) + \int \beta(s,t)(X(s) - \mu_X(s))ds$
the representation

$$\beta(s,t) = \sum_{m=1}^{\infty}\sum_{k=1}^{\infty} \frac{E(A_k B_m)}{E(A_k^2)} \phi_k(s)\psi_m(t) = \sum_{m=1}^{\infty}\sum_{k=1}^{\infty} \beta_{mk}\phi_k(s)\psi_m(t)$$

which implies $E(B_m|X) = \sum \beta_{mk} A_k$ and (as A_k are uncorrelated)

$$E(B_m|A_k) = E[E(B_m|A_1, A_2, \ldots)|A_k] = E[E(B_m|X)|A_k] = \beta_{mk}A_k.$$

- Other basis representations (wavelets, B-splines) have been considered, eigen-representation has advantages due to uncorrelatedness of scores (independence in Gaussian case) and (relative) sparseness of representation but the eigen-representation is not connected to the response.

- For densely sampled functional data, use estimated FPC scores $\hat{A}_k, \hat{B}_m$ for $\hat{\beta}_{mk} = \widehat{\mathrm{cov}}(\hat{A}_k, \hat{B}_m)/\hat{\lambda}_k$, ie, decompose functional regression into a series of simple linear regressions through the origin.

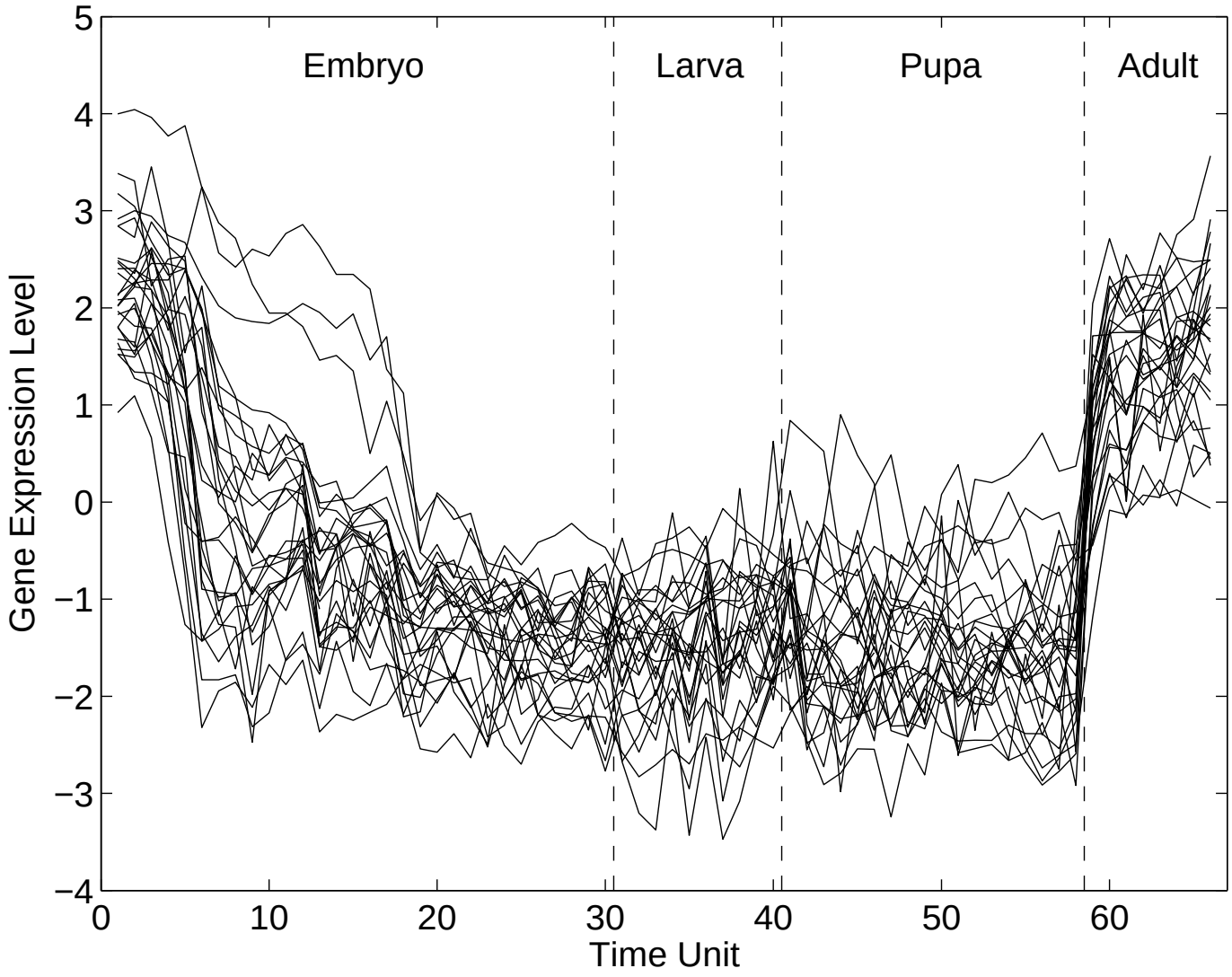
- Inference: Separately sample predictor and response data for randomly resampled subjects under H_0 : no regression relation, then recalculate functional regression and obtain bootstrap distribution of suitable test statistics such as functional R^2.

DROSOPHILA LIFE CYCLE GENE EXPRESSION

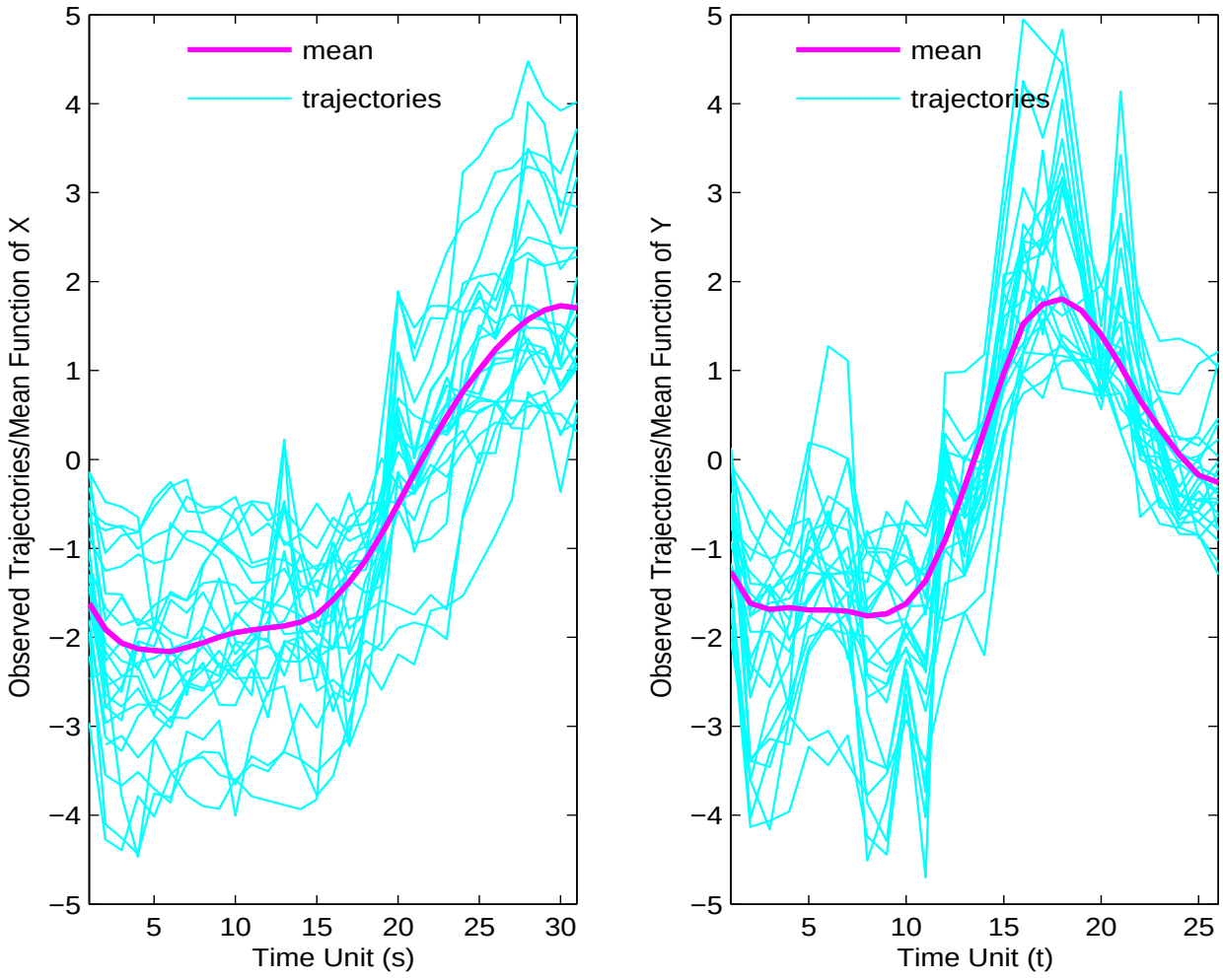
Consider gene time course data, where gene expression is repeatedly measured for:

- 23 "muscle specific" genes: tissue-specific, muscle development
- 22 "skeleto-neural" genes

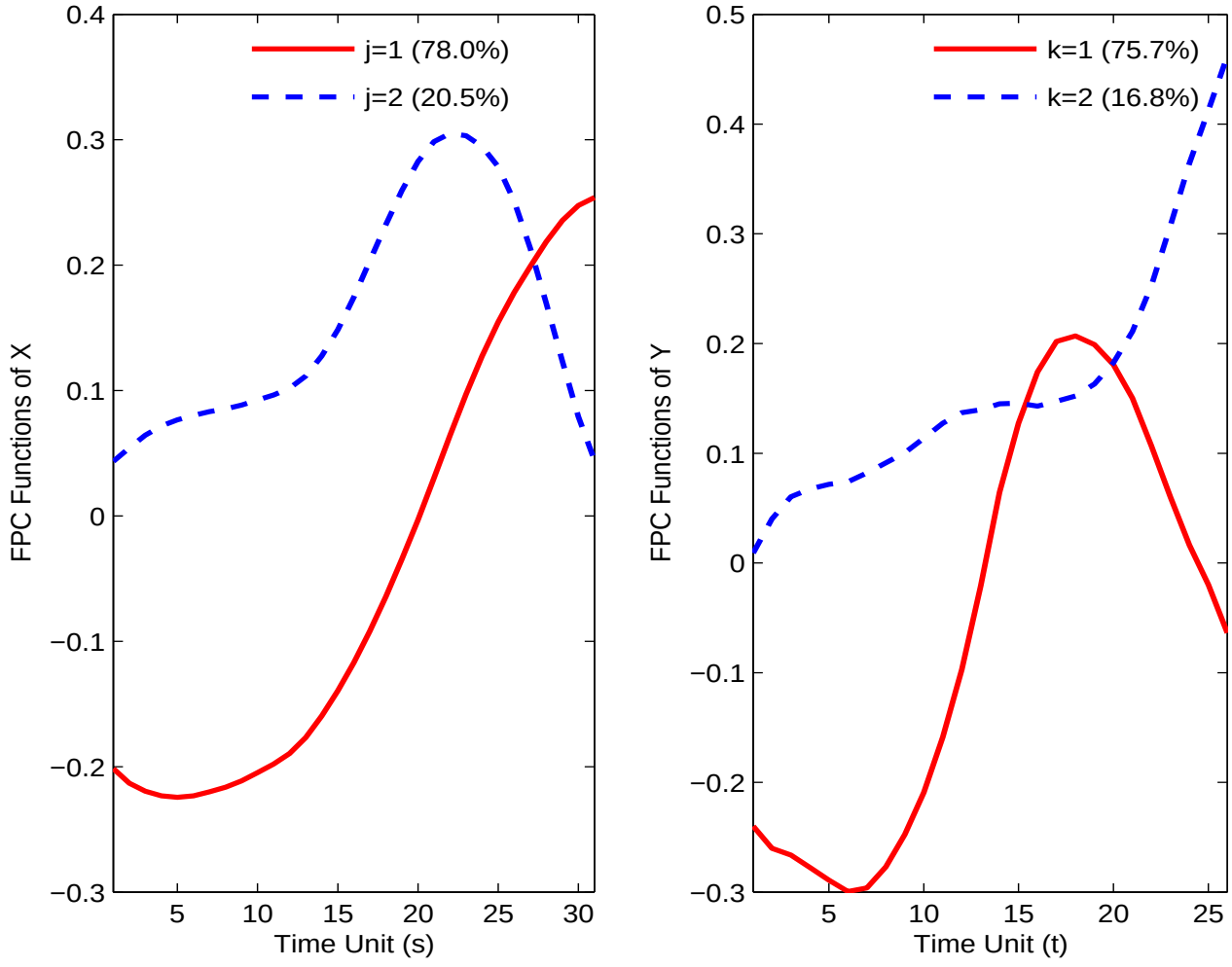
Müller & Chiou 2007 CSDA Müller, Chiou, Leng 2008 BMC Bioinformatics



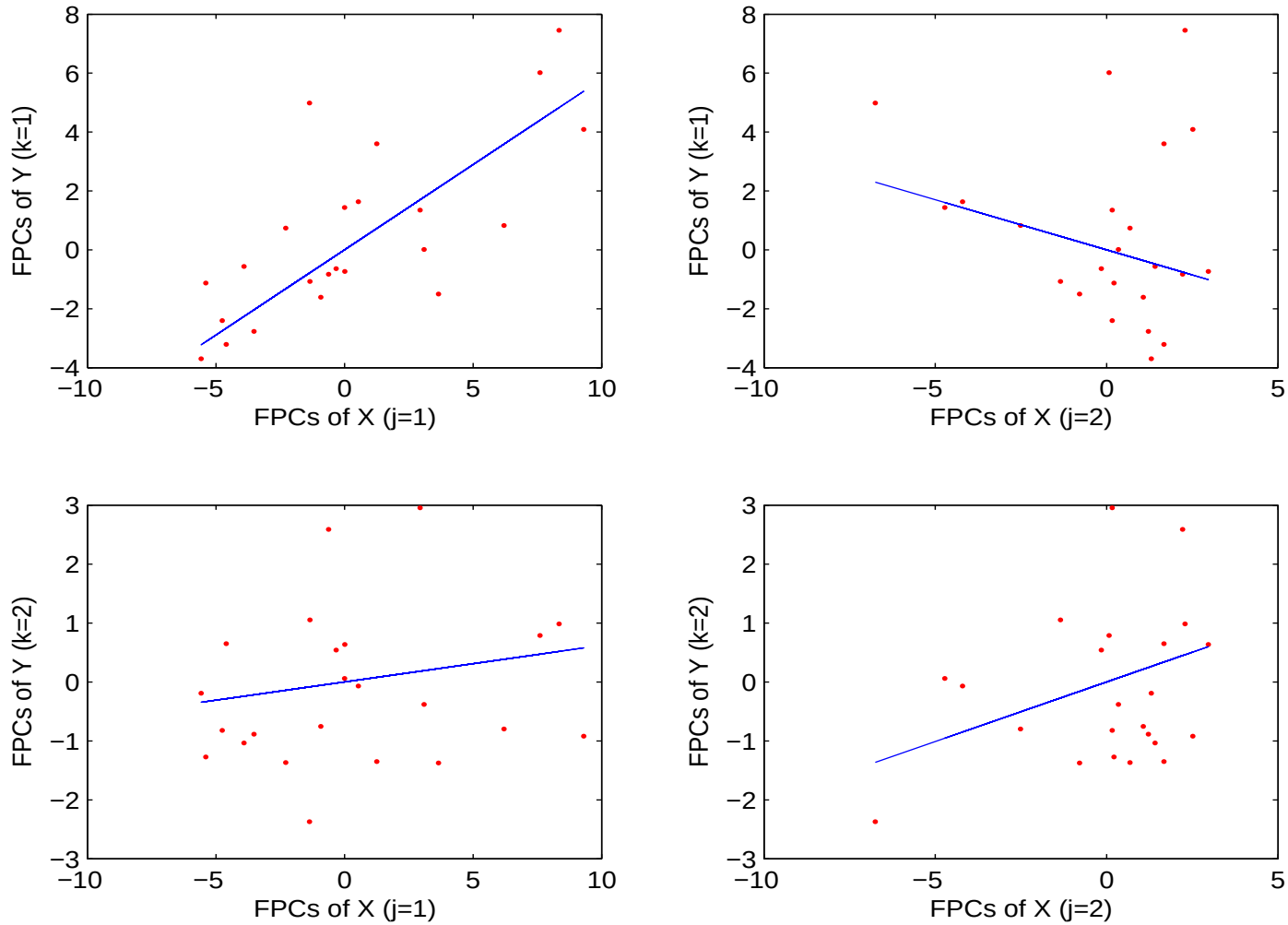
A subset of observed gene expression profiles (strict maternal genes). Each profile (or curve) is composed of expression levels of one gene at different time points.



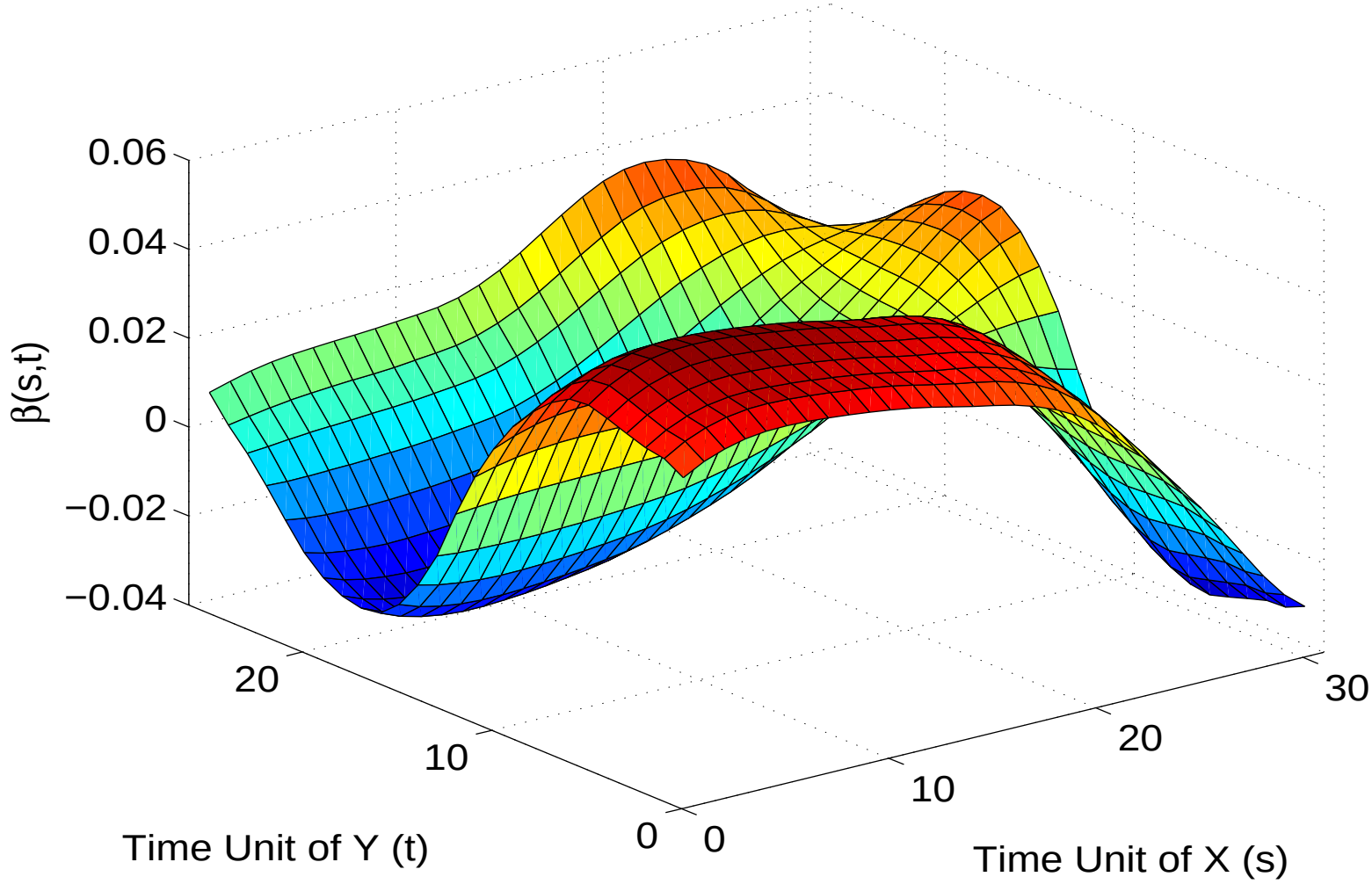
Observed trajectories and estimated mean function for muscle-specific genes for predictor profiles X (corresponding to gene expression profiles in embryo phase, left panel) and for response profiles Y (profiles for pupa-adult phase, right panel)



First two estimated eigenfunctions for temporal gene expression trajectories for the muscle-specific genes in embryo phase (predictors X, left panel) and pupa-adult phase (responses Y, right panel).



Scatterplots of functional principal component scores B_k of response trajectories versus A_j of predictor trajectories, for $j, k = 1, 2$, for muscle-specific genes



Estimated regression parameter function $\hat{\beta}(s,t)$ for muscle-specific genes with embryo phase as predictor $X(s)$ (plotted towards the right) and pupa-adult phase as response $Y(t)$ (plotted towards the left)

FUNCTIONAL COEFFICIENT OF DETERMINATION AND DIAGNOSTICS

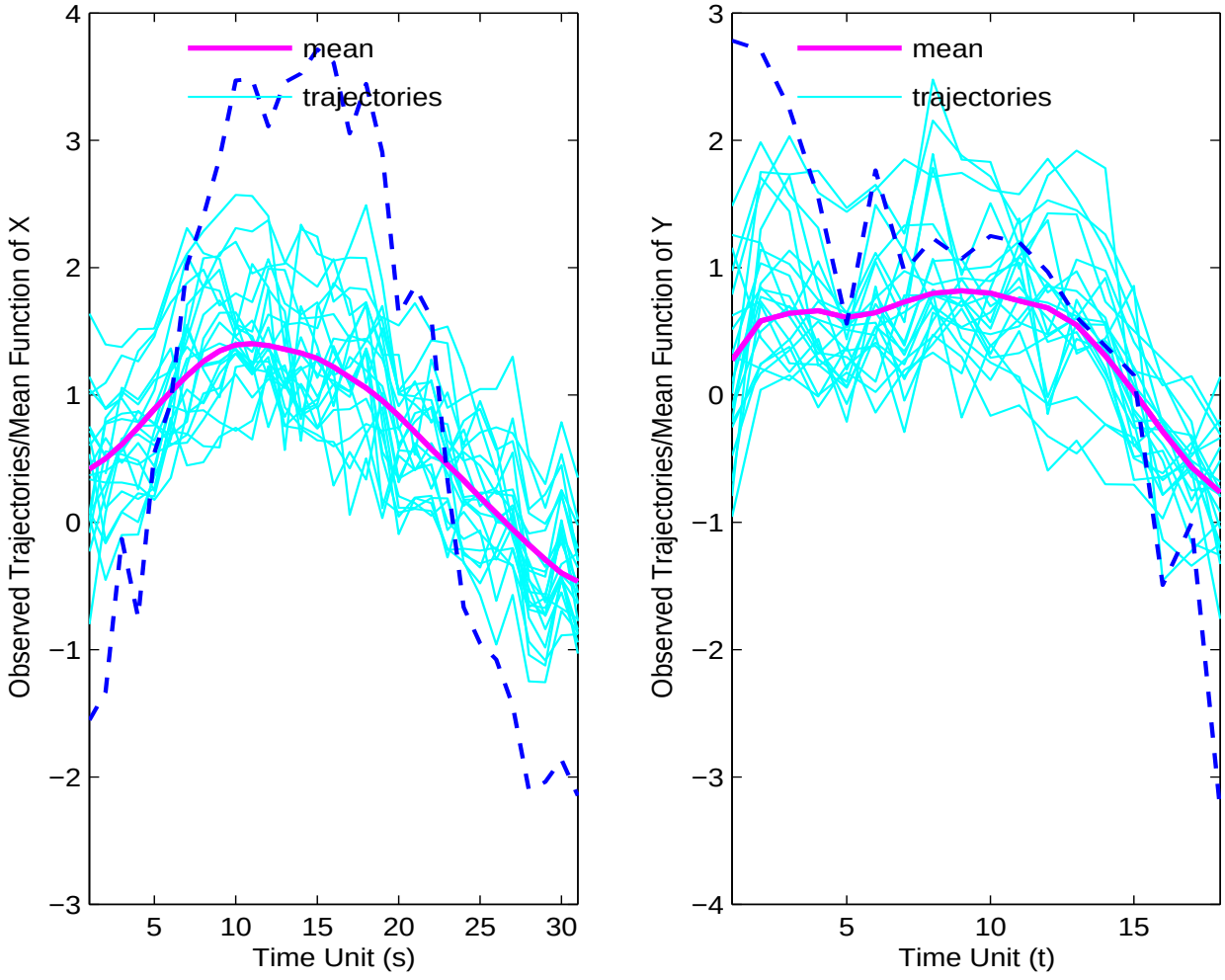
Extension from the multiple linear regression case:

$$R^2 = \frac{\int_{\mathcal{T}} \text{var}(E[Y(t)|X])dt}{\int_{\mathcal{T}} \text{var}(Y(t))dt} = \sum_{j=1}^{\infty} \frac{\sum_{k=1}^{\infty} R_{kj}^2 \tau_k}{\sum_{k=1}^{\infty} \tau_k},$$

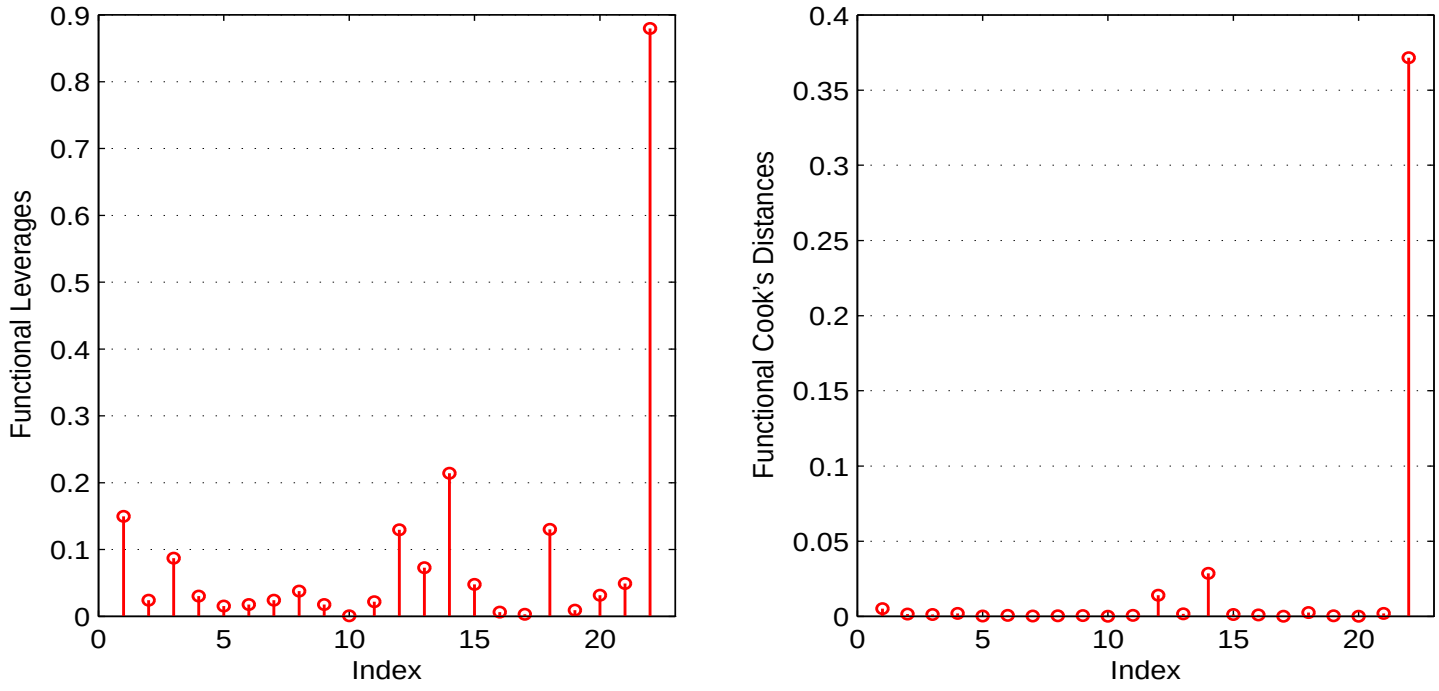
where

$$R_{kj}^2 = \frac{[\text{cov}(A_j, B_k)]^2}{\lambda_j \tau_k}$$

are the coefficients of determination for the simple linear regressions of B_k on A_j. Obtain estimate $R^2 = 0.85$ for muscle-specific genes ($p = 0.0010$ from bootstrap test)
Functional diagnostics can be obtained by a similar weighting scheme: Functional hat matrix, functional Cook's distance, etc.



Observed trajectories and estimated mean function for cytoskeleton/neural genes in embryo phase (for predictor X, left panel) and pupa phase (for response Y, right panel), respectively. Trajectories of gene CG2198 are dashed.



Functional leverages obtained as diagonal elements of functional hat matrix H (left panel) and functional Cook's distances (right panel) for the functional regression of cytoskeleton/neural genes.

FUNCTIONAL LINEAR MODEL FOR LONGITUDINAL DATA

Regress processes $Y(\cdot)$ on processes $X(\cdot)$ under sparse data situation. Notation:
$X_i(s)$ on $[0, \mathcal{S}]$: smooth predictor curve

U_{il}: measurements of $X_i(\cdot)$ at S_{il}, $1 \leq i \leq n, 1 \leq l \leq L_i$

$Y_i(t)$ on $[0, \mathcal{T}]$: smooth response curve

V_{ij}: measurements of $Y_i(\cdot)$ at T_{ij}, $1 \leq j \leq N_i$

Functional Regression Model

$$E[Y(t)|X(\cdot)] = \mu_Y(t) + \int_0^{\mathcal{S}} \beta(s,t)(X(s) - \mu_X(s))ds.$$

$\beta(s,t)$: smooth regression function, $\int_0^{\mathcal{T}} \int_0^{\mathcal{S}} \beta^2(s,t)dsdt < \infty$.

Modelling Predictor and Response Curves:

$$U_{il} = X_i(S_{il}) + e_{il} = \mu_X(S_{il}) + \sum_{m=1}^{\infty} A_{im}\phi_m(S_{il}) + e_{il},$$

$$V_{ij} = Y_i(T_{ij}) + \epsilon_{ij} = \mu_Y(T_{ij}) + \sum_{k=1}^{\infty} B_{ik}\psi_k(T_{ij}) + \epsilon_{ij}.$$

BASIS REPRESENTATION

$$\beta(s,t) = \sum_{k,m=1}^{\infty} \frac{E[A_m B_k]}{E[A_m^2]} \phi_m(s)\psi_k(t)$$

Estimating $E[A_m B_k]$:

$$\widehat{E}[A_m B_k] = \int_0^{\mathcal{T}} \int_0^{\mathcal{S}} \hat{\phi}_m(s)\widehat{\Gamma}_{XY}(s,t)\hat{\psi}_k(t)dsdt,$$

where $\widehat{\Gamma}_{XY}(s,t)$ is local linear smoothing estimate of the covariance surface $\Gamma_{XY}(s,t) = \mathrm{cov}(X(s), Y(t))$.

CONDITIONAL METHOD

Objective: Predict trajectory Y^* of a new subject, given observations $U^* = (U_1^*, \cdots, U_{L^*}^*)^T$ of $X^*(\cdot)$.

$$\begin{aligned} E[Y^*(t)|X^*(\cdot)] &= \mu_Y(t) + \int_0^{\mathcal{S}} \beta(s,t)X^*(s)ds \\ &= \mu_Y(t) + \sum_{k,m=1}^{\infty} \frac{E[A_m B_k]}{E[A_m^2]} A_m^* \psi_k(t) \end{aligned}$$

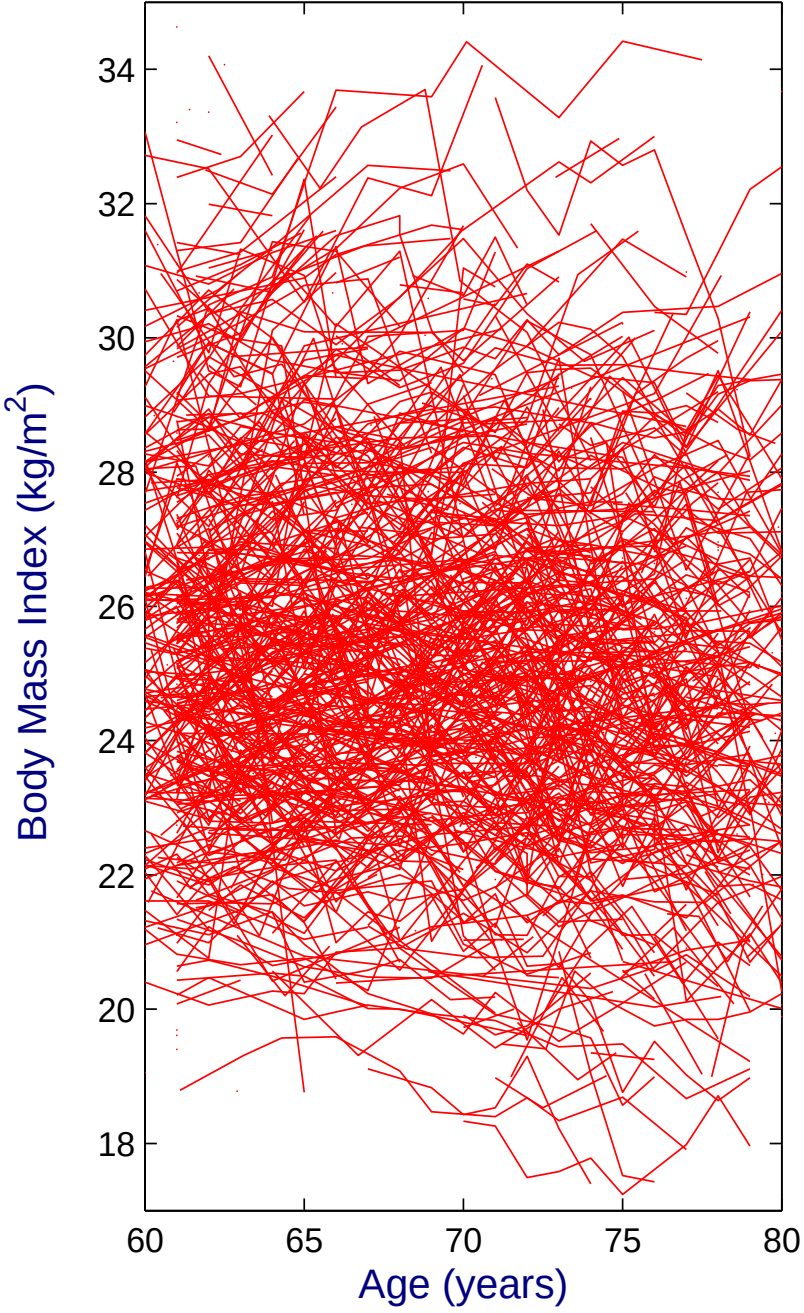
Constraint: $\mu_Y(t) = \int_0^{\mathcal{S}} \beta(s,t)\mu_X(s)ds$.

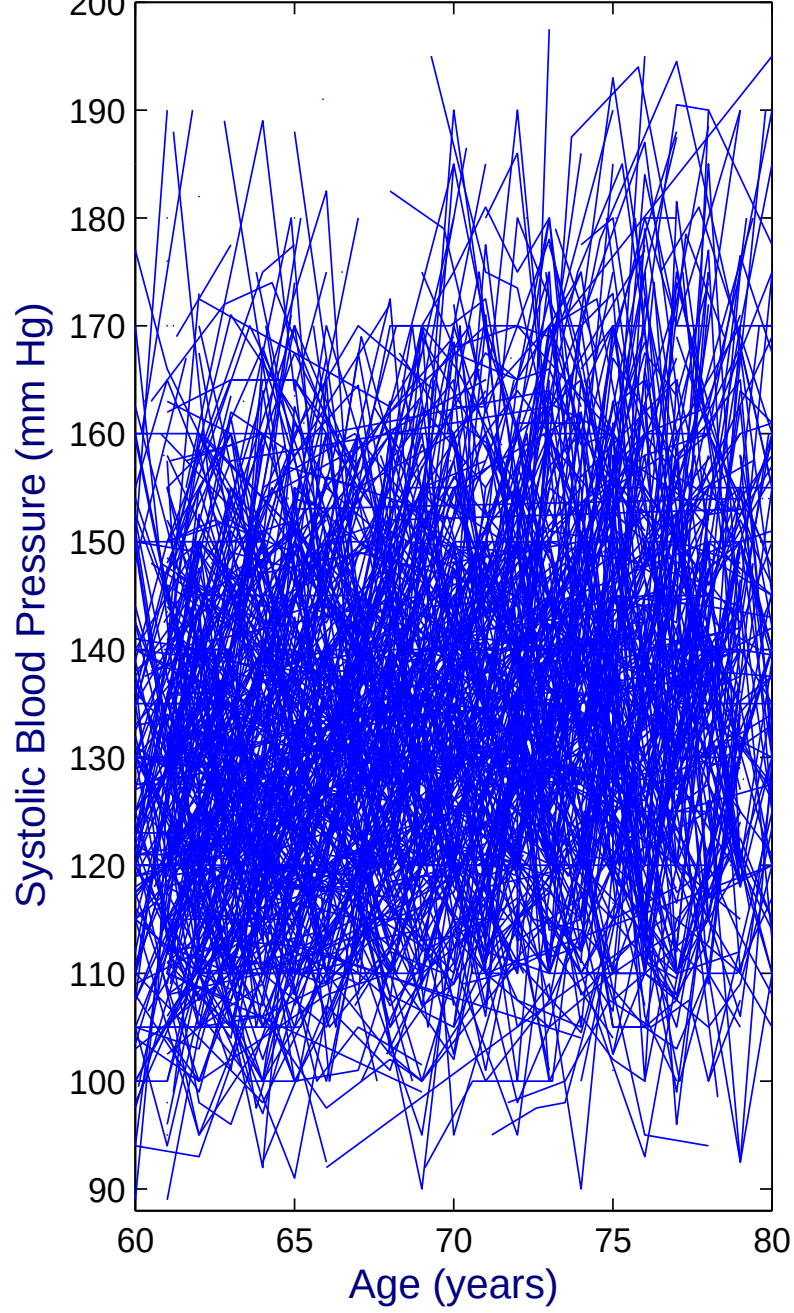
APPLICATION

Functional Regression of Systolic Blood Pressure on Body Mass Index

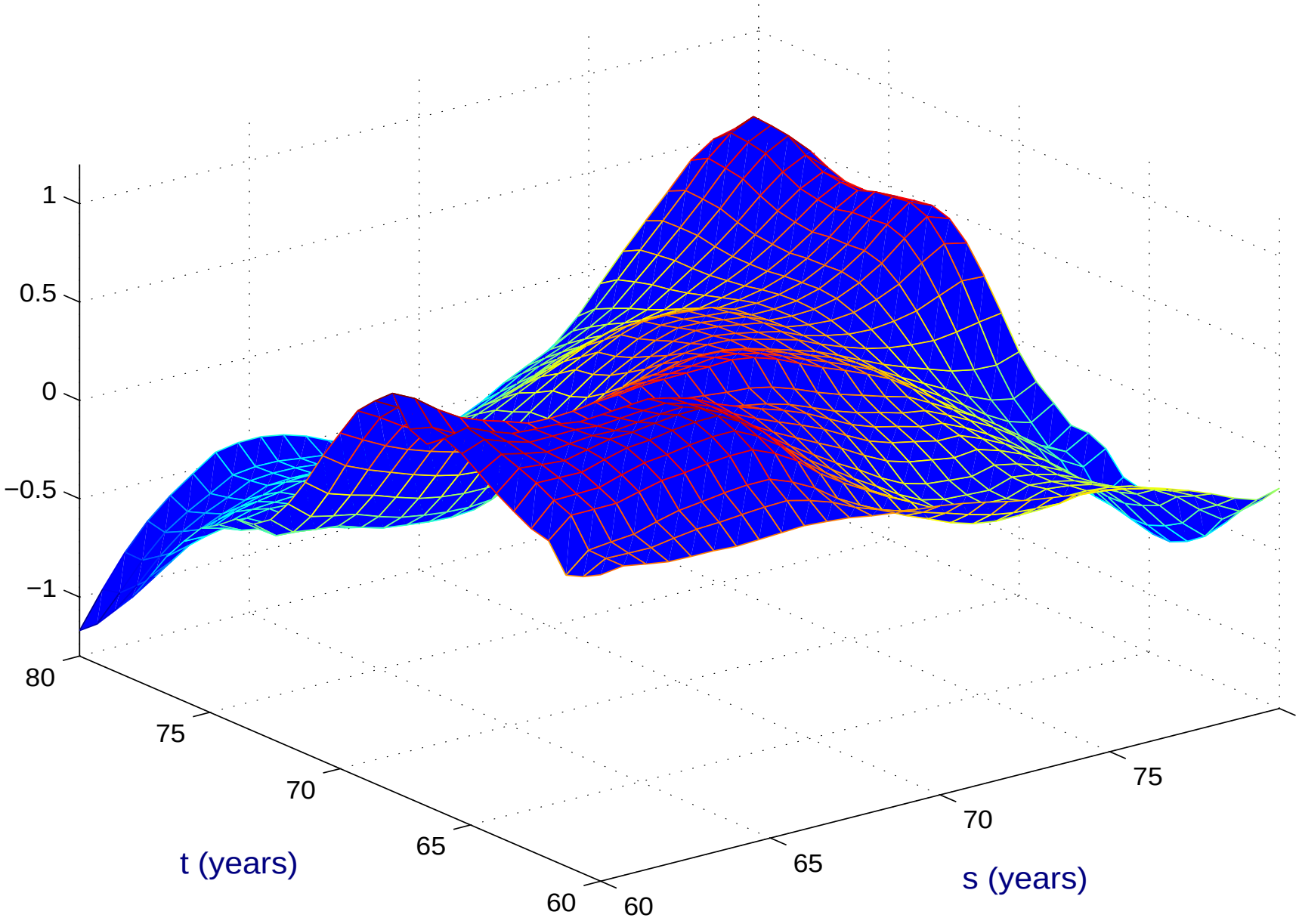
Data: Body mass index (BMI) and systolic blood pressure (SBP) for 812 participants in the Baltimore Longitudinal Study on Aging

Irregular and Sparse Measurements
$R^2 = 0.13$

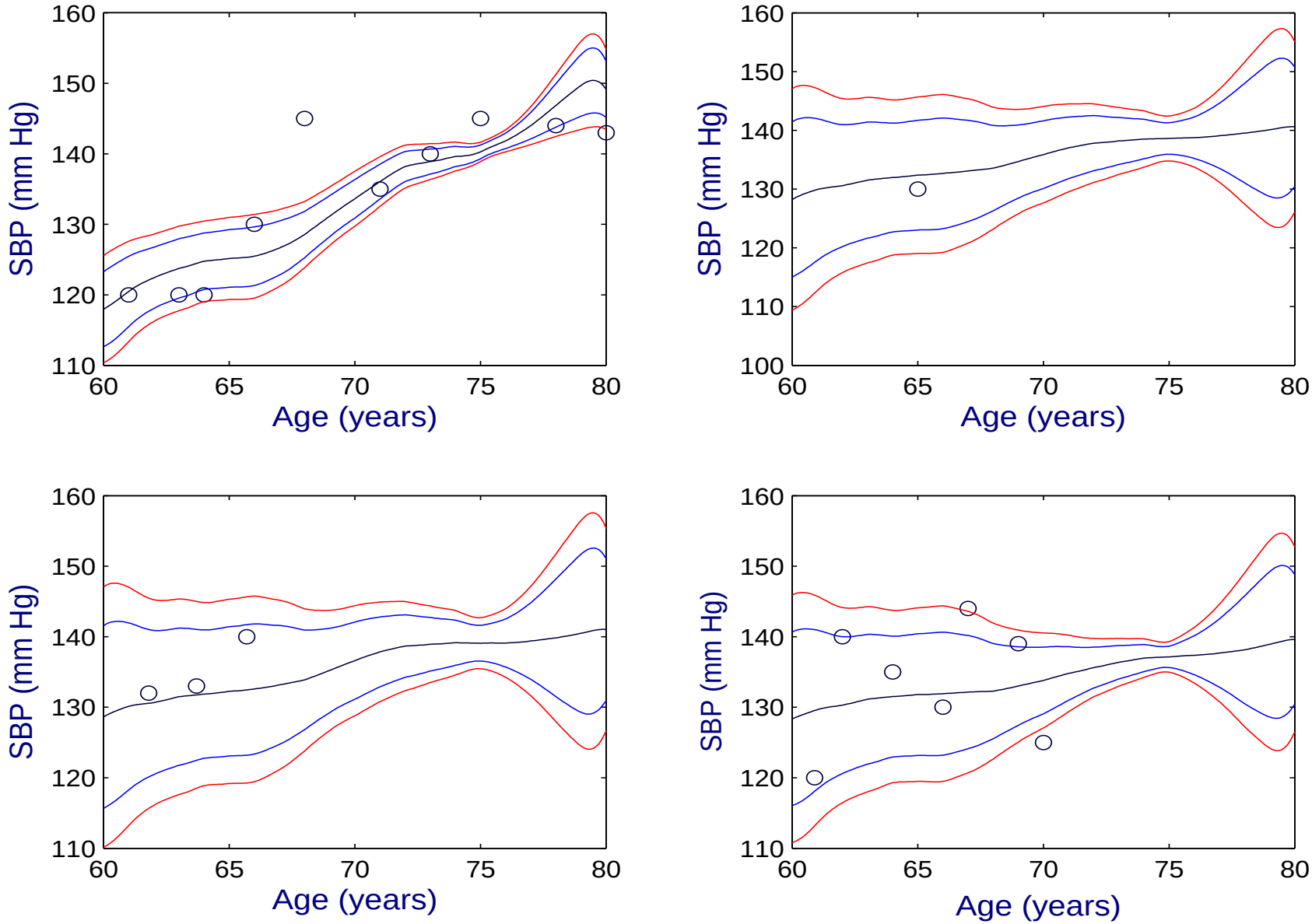




Observed paths of Body Mass Index (left) and Systolic Blood Pressure (right) for 812 participants.



Estimated regression function $\hat{\beta}(s,t)$, where the predictor (BMI) time is s (in years), and the response (SBP) time is t (in years).



Observed data (circles), predicted trajectories (black), 95% pointwise (blue) and simultaneous (red) bands obtained by one-leave-out analysis.

GENERALIZED FUNCTIONAL LINEAR MODEL

- Predictors $X(t) \in L^2$, Response $Y \in \mathbb{R}$
- Components: Parameter Function $\beta(\cdot)$, Link Function $g(\cdot)$, Variance Function $\sigma^2(\cdot)$

$$\begin{aligned} \eta_i &= \alpha + \int \beta(t) X_i(t)\, d(t) \quad \text{linear predictors} \\ Y_i &= g(\eta_i) + e_i = \mu_i + e_i, \quad i = 1, \ldots, n, \end{aligned}$$

with i.i.d. errors e_i, means $E(Y_i) = \mu_i = g(\eta_i)$ and $E(e|X(\cdot)) = 0, \mathrm{var}(e|X(\cdot)) = \sigma^2(\mu)$.

- If link function $g(\cdot)$ and variance function $\sigma^2(\cdot)$ are unknown and smooth, they can be estimated from the data.
- Applications of generalized functional linear model (GFLM): Functional logistic regression and classification, when Y denotes class membership and the logistic link function is used.
- With orthonormal basis ϕ_j, $j \geq 1$,

$$X(t) = \sum_{j=1}^{\infty} A_j \phi_j(t), \; \beta(t) = \sum_{j=1}^{\infty} \beta_j \phi_j(t), \; \int \beta(t) X(t)\, dt = \sum_{j=1}^{\infty} \beta_j A_j.$$

- Under regularity conditions, can obtain asymptotic consistency of β and of $E(Y|X)$ – this is an active area of research

FURTHER EXTENSIONS OF THE FLM

"Classic" extensions: linear $\Rightarrow$ quadratic $\Rightarrow$ polynomial

The polynomial functional regression model (Yao & M 2010)

$$\begin{aligned} E(Y|X) = \alpha &+ \int_{\mathcal{T}} \beta(t) X^c(t) dt + \int_{\mathcal{T}^2} \gamma(s,t) X^c(s) X^c(t) ds dt \\ &+ \int_{\mathcal{T}^3} \gamma_3(t_1,t_2,t_3) X^c(t_1) X^c(t_2) X^c(t_3) dt_1 dt_2 dt_3 + \ldots \\ &+ \int_{\mathcal{T}^p} \gamma_p(t_1,\ldots,t_p) X^c(t_1) \ldots X^c(t_p) dt_1 \ldots dt_p, \end{aligned}$$

with α as intercept and β, γ, γ_j, $3 \leq j \leq p$, as linear, quadratic and jth order regression parameter functions. In terms of FPCs,

$$\begin{aligned} E(Y|X) = \alpha &+ \sum_{j_1 \geq 1} \beta_{j_1} A_{j_1} + \sum_{j_1 \leq j_2} \gamma_{j_1 j_2} A_{j_1} A_{j_2} + \sum_{j_1 \leq j_2 \leq j_3} \gamma_{j_1 j_2 j_3} A_{j_1} A_{j_2} A_{j_3} \\ &+ \ldots + \sum_{j_1 \leq \ldots \leq j_p} \gamma_{j_1 \ldots j_p} A_{j_1} \ldots A_{j_p}, \end{aligned}$$

model includes all interaction effects up to p time points.

FUNCTIONAL QUADRATIC REGRESSION

$$E(Y|X) = \alpha + \sum_{k=1}^{\infty} \beta_k A_k + \sum_{k=1}^{\infty} \sum_{\ell=1}^{k} \gamma_{k\ell} A_k A_\ell,$$

Quadratic diagonal case

$$E(Y|X) = \alpha + \sum_{k} \beta_k A_k + \sum_{k} \gamma_{kk} A_k^2.$$

With eigenvalues λ_k for X and covariance functions

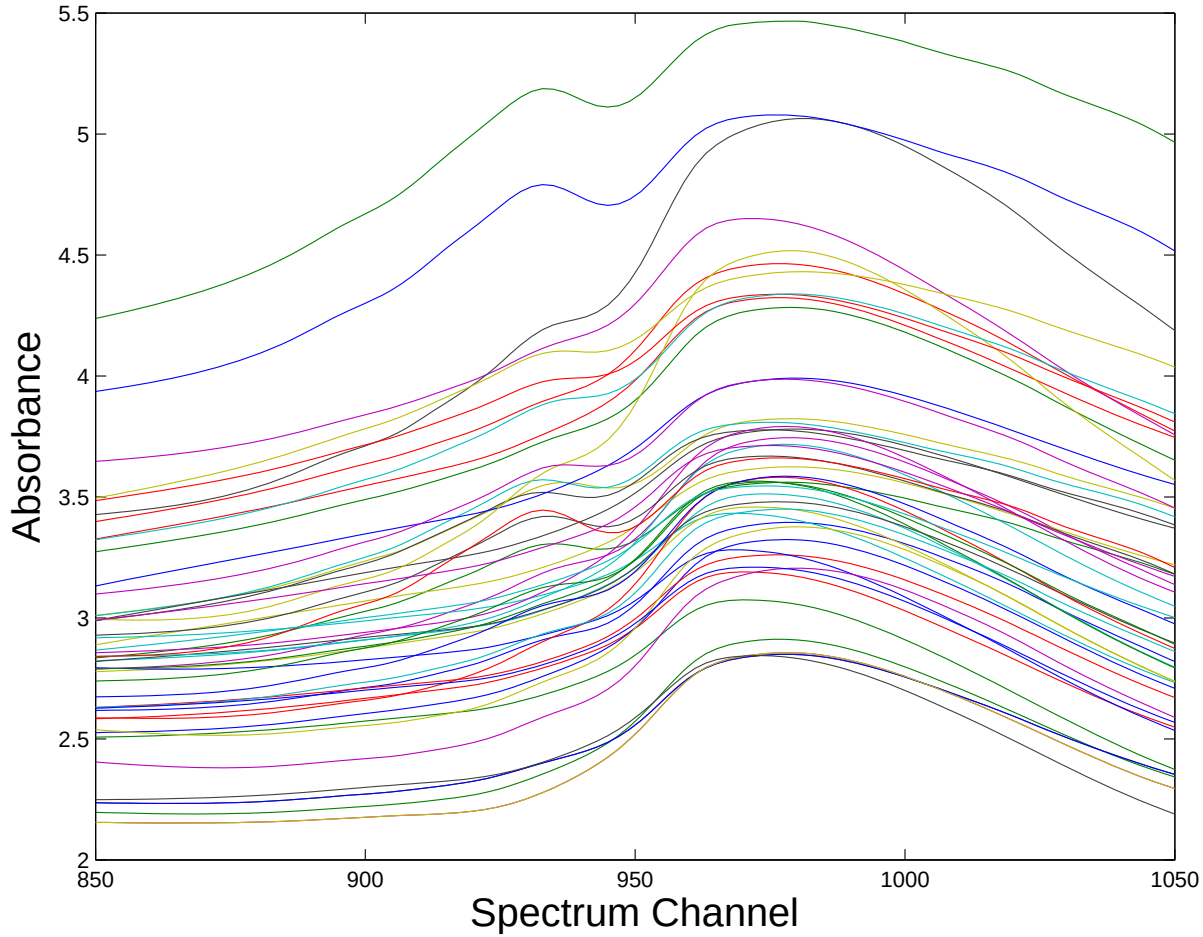
$$\begin{aligned} C_1(t) &= \text{cov}\{X(t), Y\} = \sum_{k=1}^{\infty} \eta_k \phi_k(t), \\ C_2(s,t) &= E\{X(s)X(t)Y\} = \sum_{k,\ell=1}^{\infty} \rho_{k\ell} \phi_k(s)\phi_k(t), \end{aligned}$$

least squares estimators are obtained via the representations
$\alpha = \mu_Y - \sum_k \gamma_{kk}\lambda_k, \quad \beta_k = \eta_k/\lambda_k, \quad \gamma_{k\ell} = \rho_{k\ell}/(\lambda_k\lambda_\ell)$,
for $k < \ell, \quad \gamma_{kk} = (\rho_{kk} - \mu_Y\lambda_k)/(E(A_k^4) - \lambda_k^2)$.

- Can easily be implemented with PACE (`quadreg`)

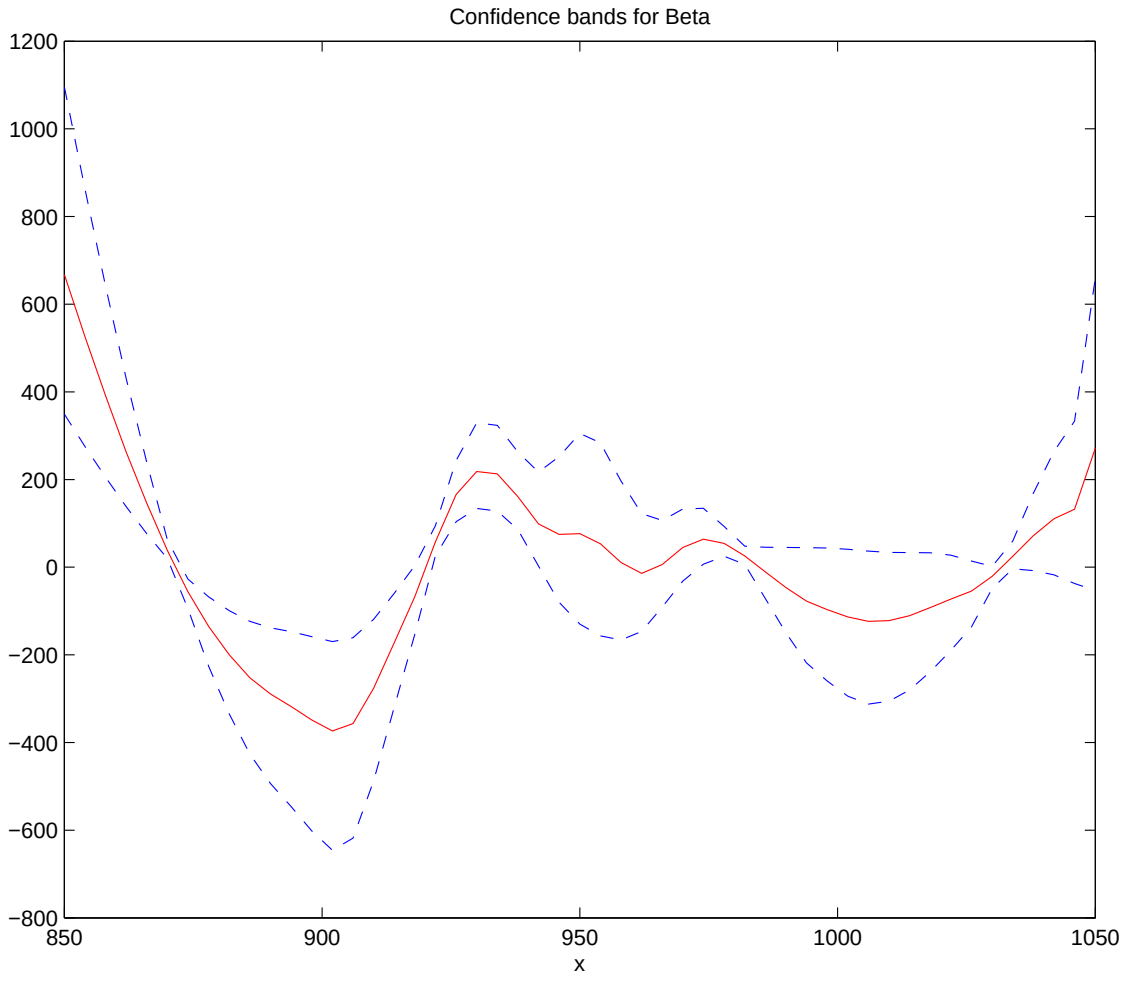
- Obtain consistent estimates and rates of convergence for parameter functions $\hat{\alpha} - \alpha = O_p(\alpha_n)$, $\|\hat{\beta} - \beta\| = O_p(\beta_n)$, $\|\hat{\gamma} - \gamma\| = O_p(\gamma_n)$ and for predicting new responses under either one of two assumptions:
  - Gaussian assumption on predictor processes X: Convergence rates for sparse irregular designs
  - Densely observed functional predictors with noise; Gaussian assumption not needed for convergence rates
  - Note: The proofs for the two designs are quite different.

Predictor Functions: Tecator Spectral Data



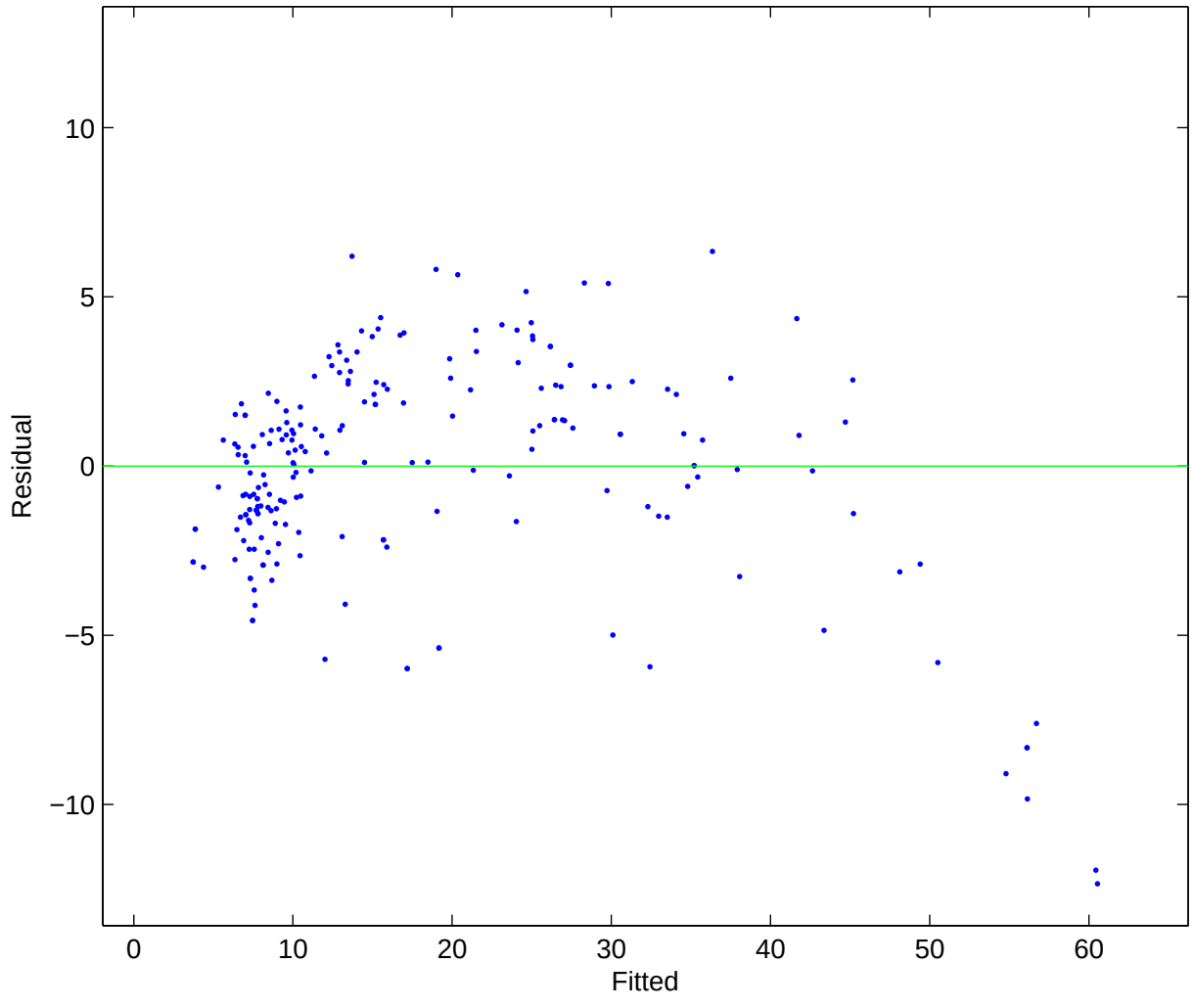
Log-transformed absorbance spectra for Tecator fat contents data, for subset of 50 meat specimen

Functional Linear Regression

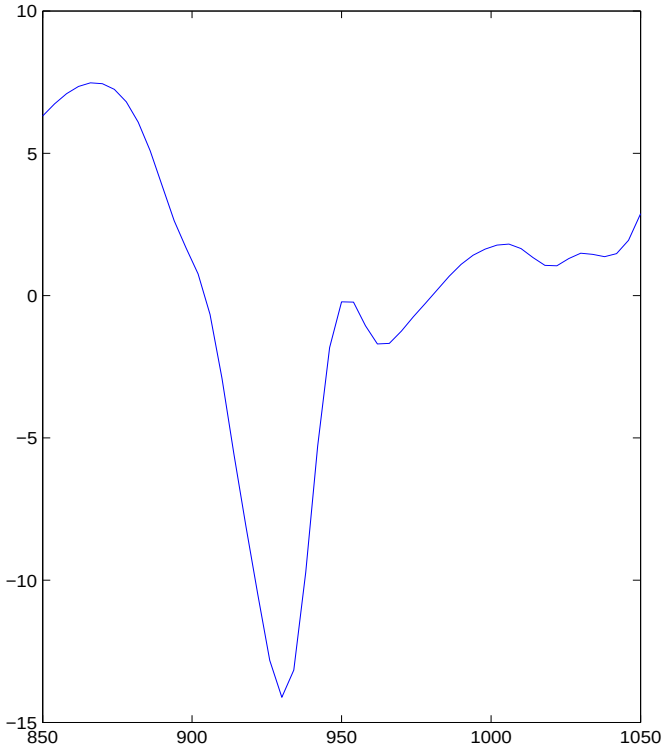


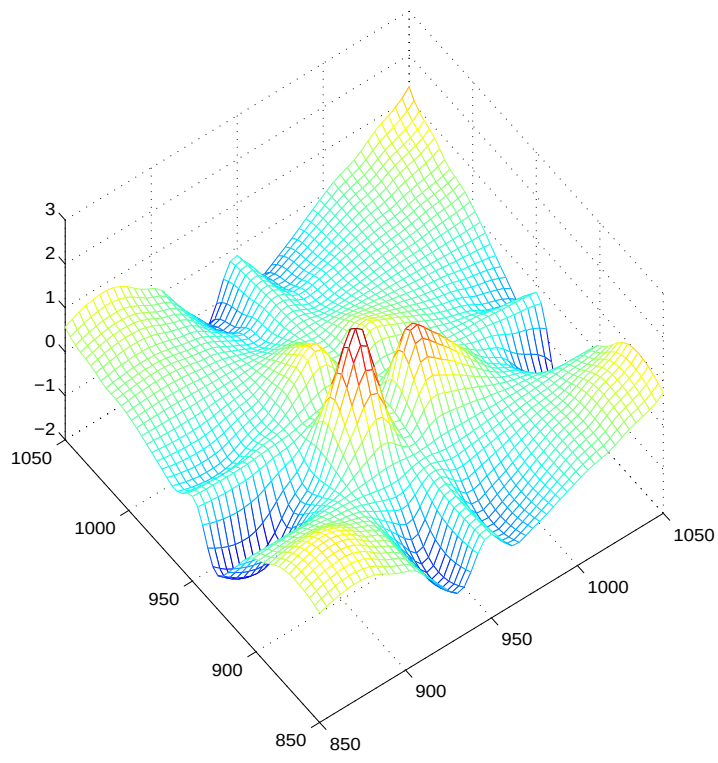
$$E(Y|X) = \alpha + \int X^c(t)\beta(t)dt$$

Residuals for Functional Linear Regression



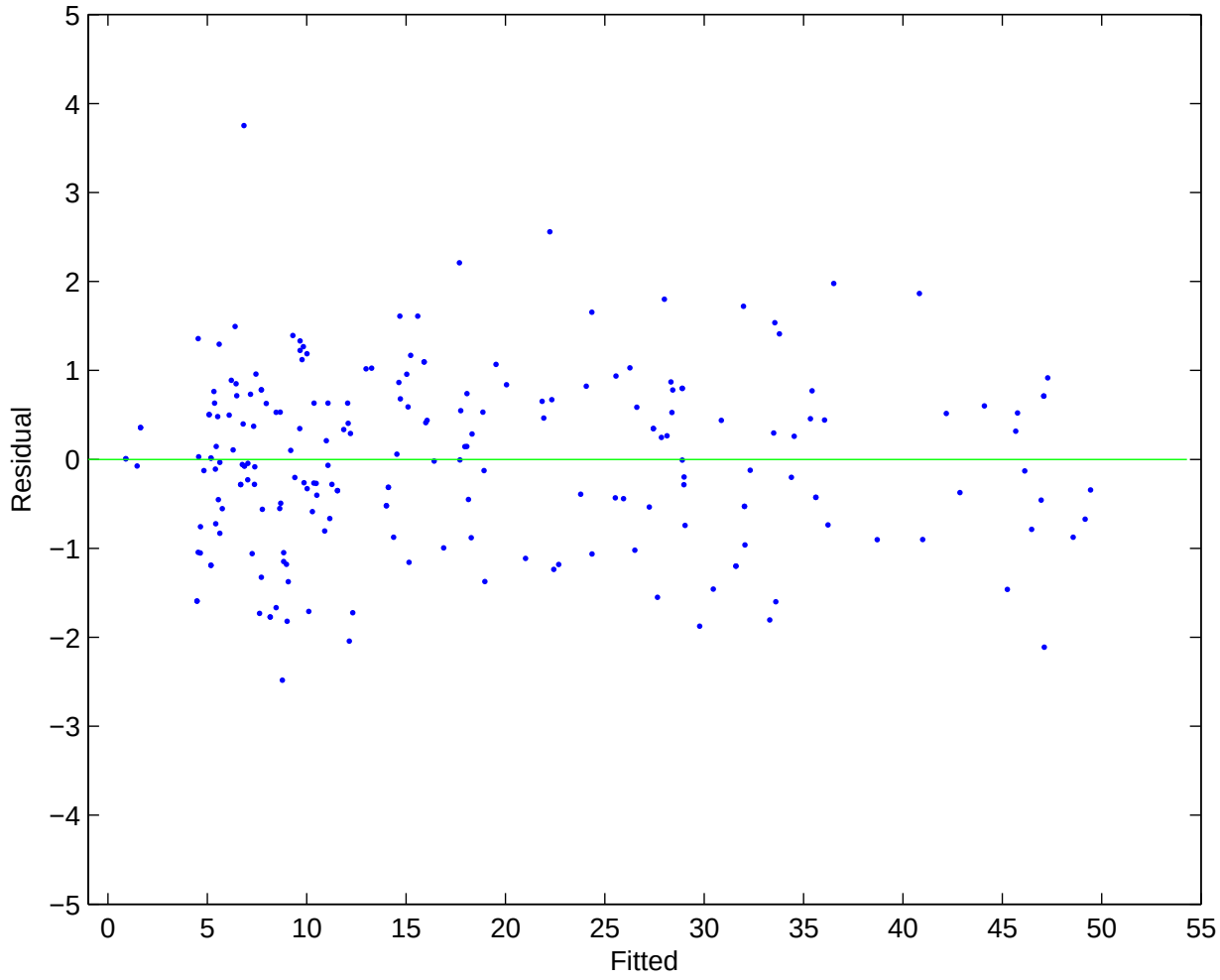
Functional Quadratic Regression

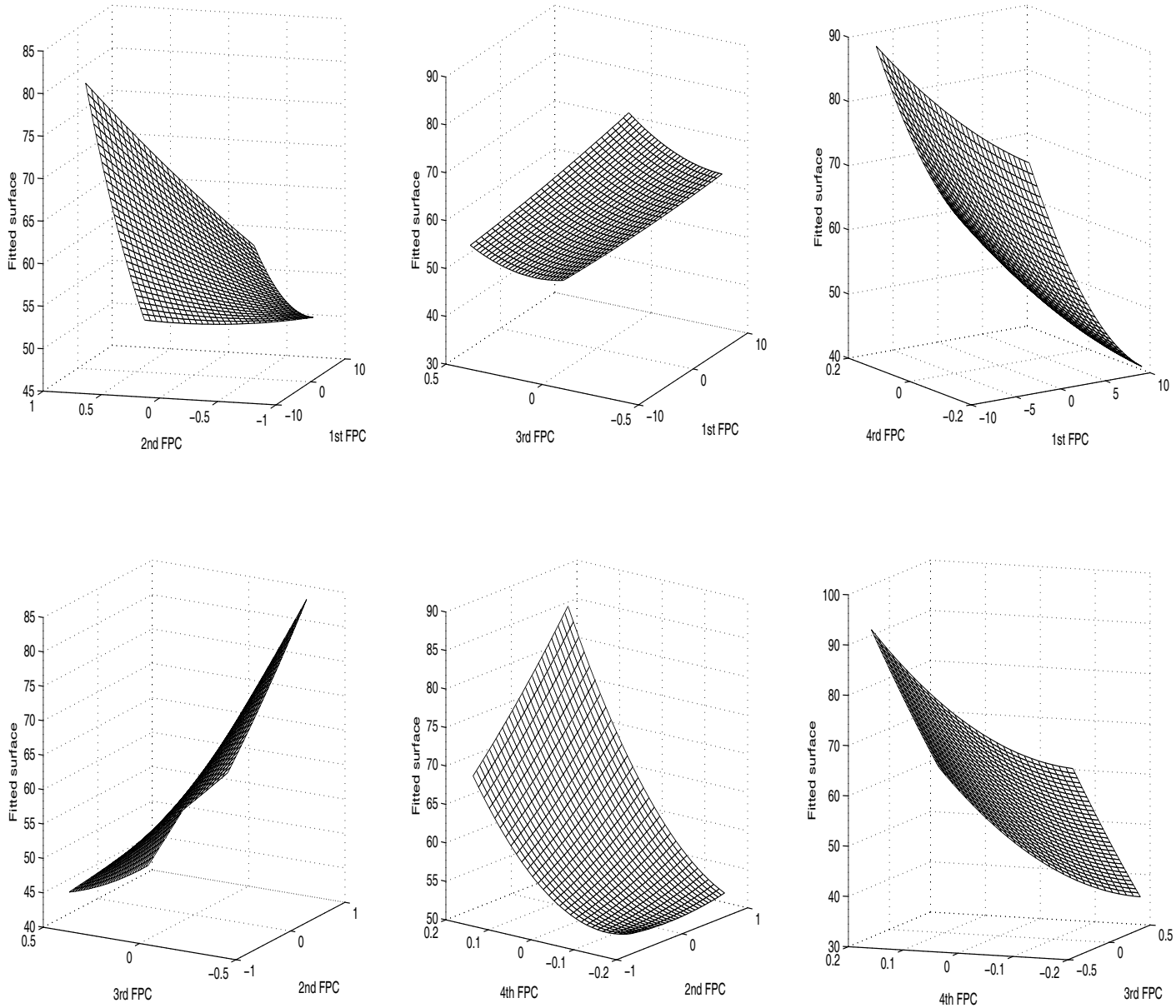




$$E(Y|X) = \alpha + \int X^c(t)\beta(t)dt + \iint \gamma(s,t)X^c(s)X^c(t)dsdt$$

Residuals for Functional Quadratic Regression





Sections through the fitted model $E(Y|A_1, A_2, A_3, A_4)$.

AN ADDITIVE EXTENSION OF THE FUNCTIONAL LINEAR MODEL (FLM)

The least squares parameter function in the FLM

$$E(Y|X) = \mu_Y + \int \beta(s)(X(s) - \mu_X(s))ds$$

has the representation

$$\beta(s) = \sum_m \sum_k \beta_k \phi_k(s) \text{ with } \beta_k = E(A_k Y)/E(A_k^2),$$

yielding

$$E(Y|X) = \sum_k \beta_k A_k.$$

This motivates the following extension:
Functional Additive Model

$$E(Y|X) = \sum_k f_k(A_k),$$

where f_k are smooth nonparametric functions; analogously for functional responses.

FUNCTIONAL ADDITIVE MODEL (FAM)

Assuming independent predictor scores A_j (automatically implied in the Gaussian case) we find

$$E(Y|A_k) = E\{E(Y|X)|A_k\} = E\{\sum_{j=1}^{\infty} f_j(A_j)|A_k\} = f_k(A_k).$$

Consequence: Functional Additive Model can be implemented simply by 1-d scatterplot smoothing of Y vs $\hat{A}_{ik}$ to obtain the defining functions f_k.

No backfitting iteration is needed: Fast and straightforward implementation with PACE. Analogously for functional regression model with scalar responses. For situations with several predictor functions within subjects: Can apply common additive model to ensemble of selected FPCs for all predictor functions.

ASYMPTOTICS FOR FAM

Employing PACE, one may show under regularity conditions that $\hat{f}_k$ is consistent for f_k and the prediction $\hat{E}(Y|X^*)$ is consistent for $E(Y|X^*)$ (M & Yao 2008)

Key steps for proof:

- Differences between A_{ik} and $\hat{A}_{ik}$ are asymptotically small enough to be negligible for the FAM smoothing steps.
- Perturbation analysis for linear operators, bounding the difference between operators A_G and $A_{\hat{G}}$.
- In the dense design case, obtain essentially 1-d rates of convergence for the component functions $\hat{f}_k$.

ADDITIVE EXTENSION OF THE FUNCTIONAL RESPONSE MODEL

Consider FLM with functional responses, with FPC representation

$$Y(t) = \mu_Y(t) + \sum_m B_m \psi_m(t).$$

Then the least squares parameter function in the FLM

$$E(Y(t)|X) = \mu_Y(t) + \int \beta(s,t)(X(s) - \mu_X(s))ds$$

has the representation

$\beta(s,t) = \sum_m \sum_k \beta_{km}\phi_k(s)\psi_m(t)$ with $\beta_{km} = E(A_k B_m)/E(A_k^2)$

yielding

$$E(Y(t)|X) = \sum_m \sum_k \beta_{mk} A_k \psi_m(t).$$

This motivates the following extension:

Functional Additive Model

$$E(Y(t)|X) = \sum_m \sum_k f_{km}(A_k)\psi_m(t),$$

where f_{km} are smooth nonparametric functions.

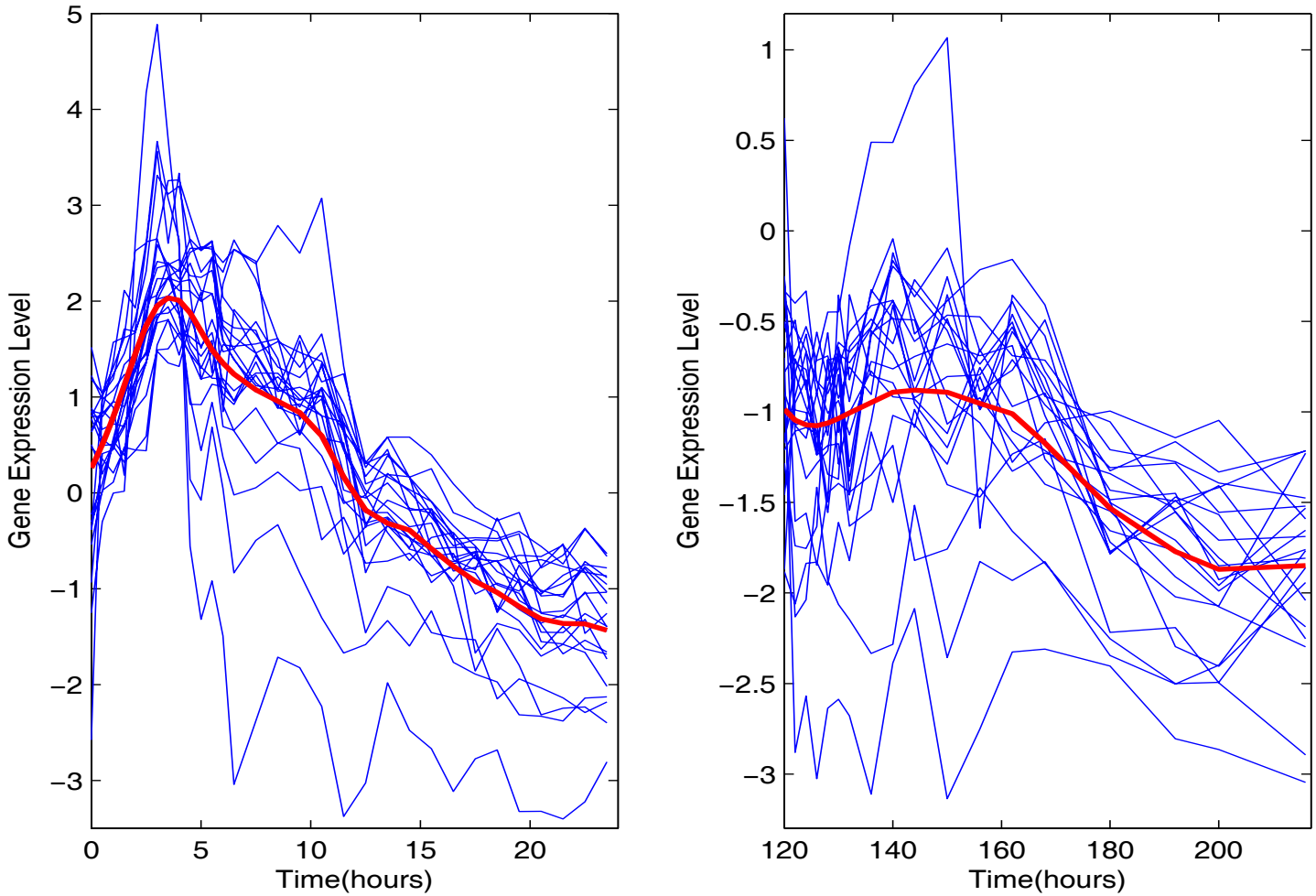
FAM FOR FUNCTIONAL RESPONSES

Assuming independent predictor scores A_j (automatically implied in the Gaussian case) we find

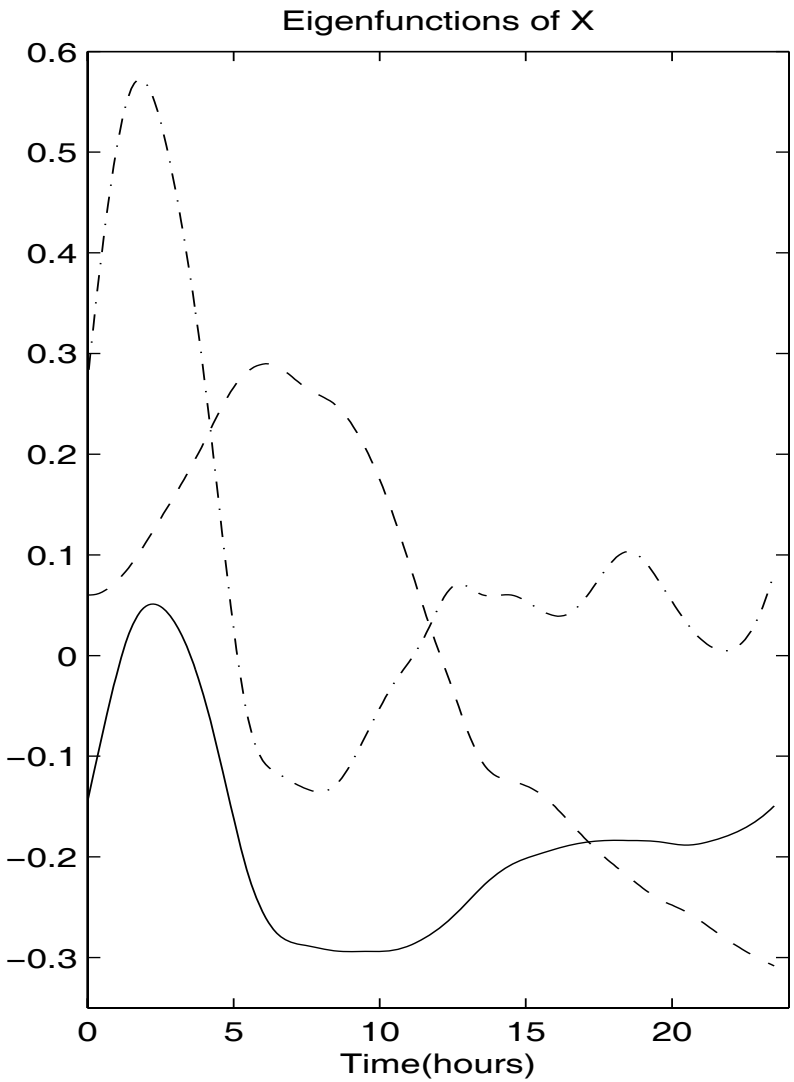
$$E(B_m|A_k) = E\{E(B_m|X)|A_k\} = E\{\sum_{j=1}^{\infty} f_{jm}(A_j)|A_k\} = f_{km}(A_k).$$

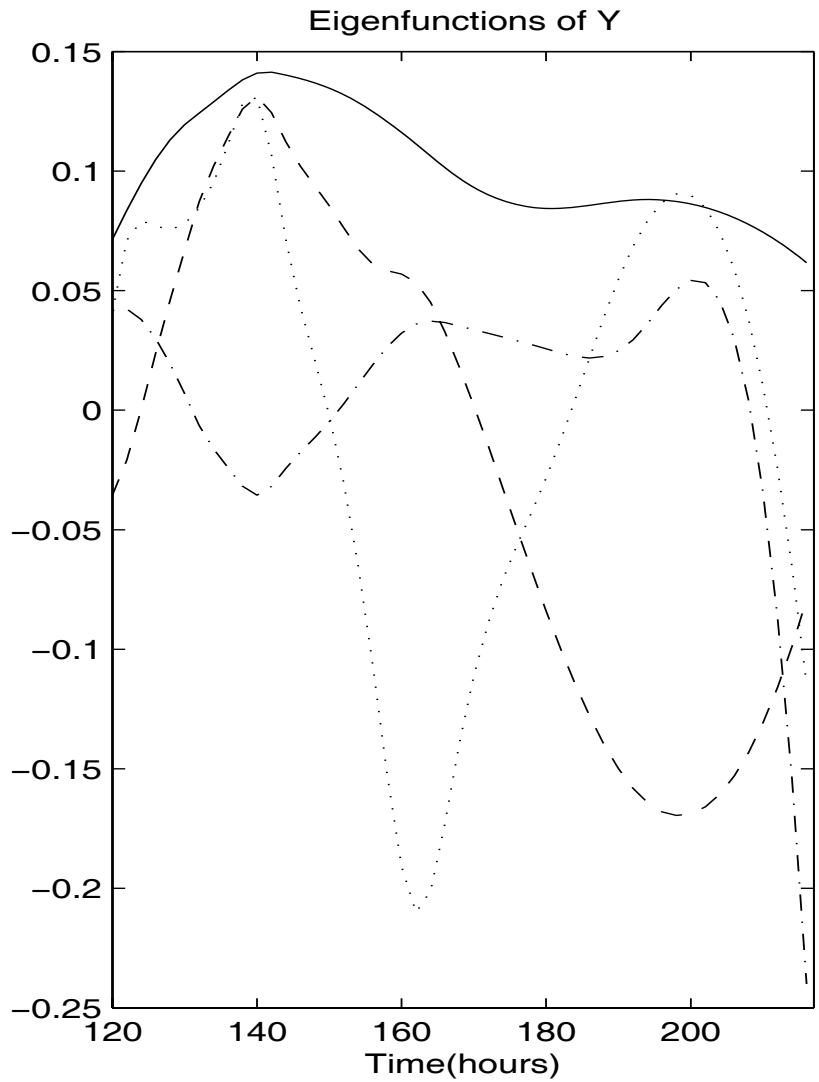
Consequence: Functional Additive Model can be implemented simply by 1-d scatterplot smoothing of $\hat{B}_{im}$ vs $\hat{A}_{ik}$ to obtain the defining functions f_{km}.

No backfitting iteration is needed: Fast and straightforward implementation with PACE. Analogously for functional regression model with scalar responses. For situations with several predictor functions within subjects: Can apply common additive model to ensemble of selected FPCs for all predictor functions.



Gene time course data, zygotic genes for Drosophila for embryo phase (left) and pupa phase (right).

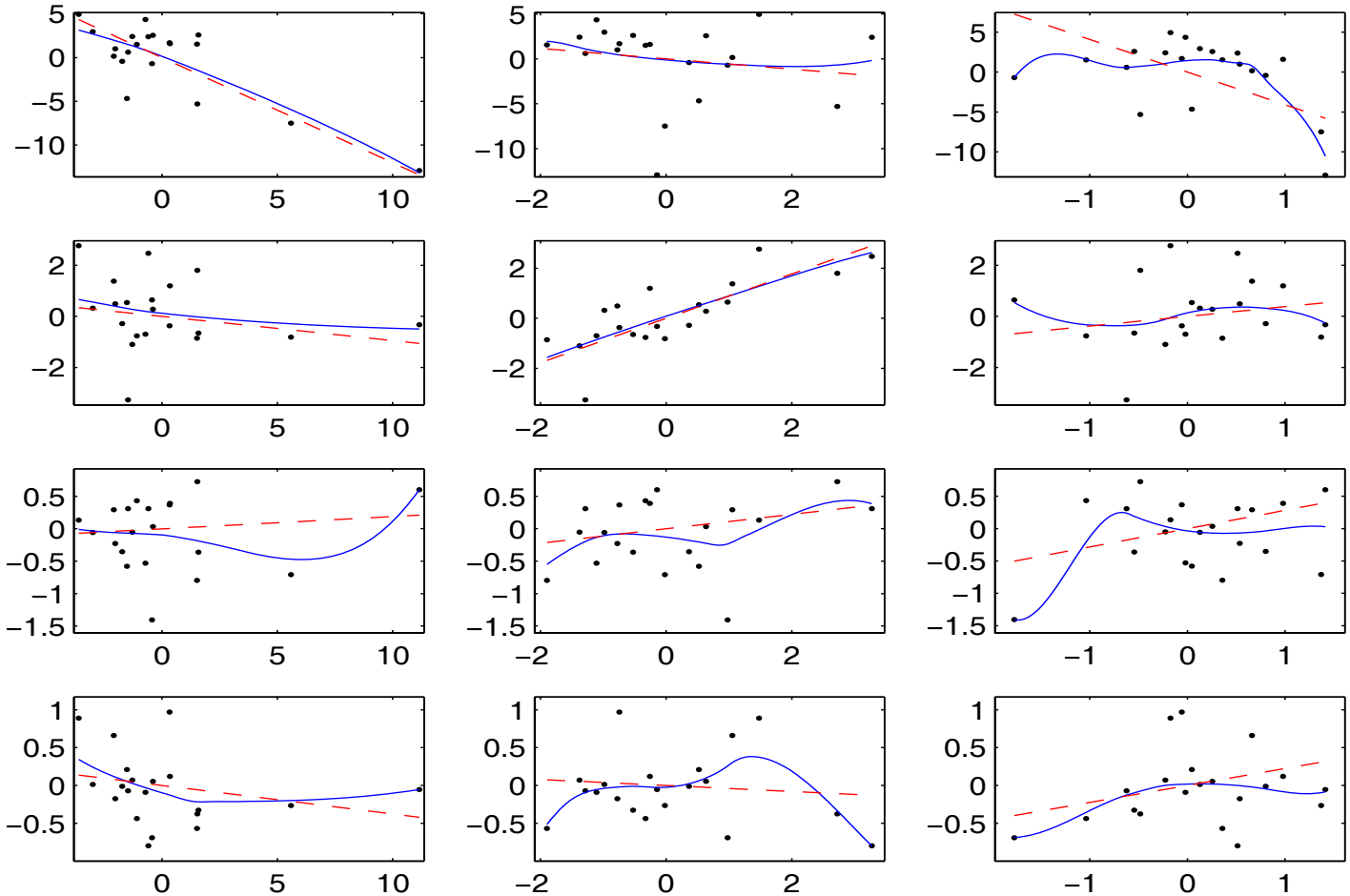




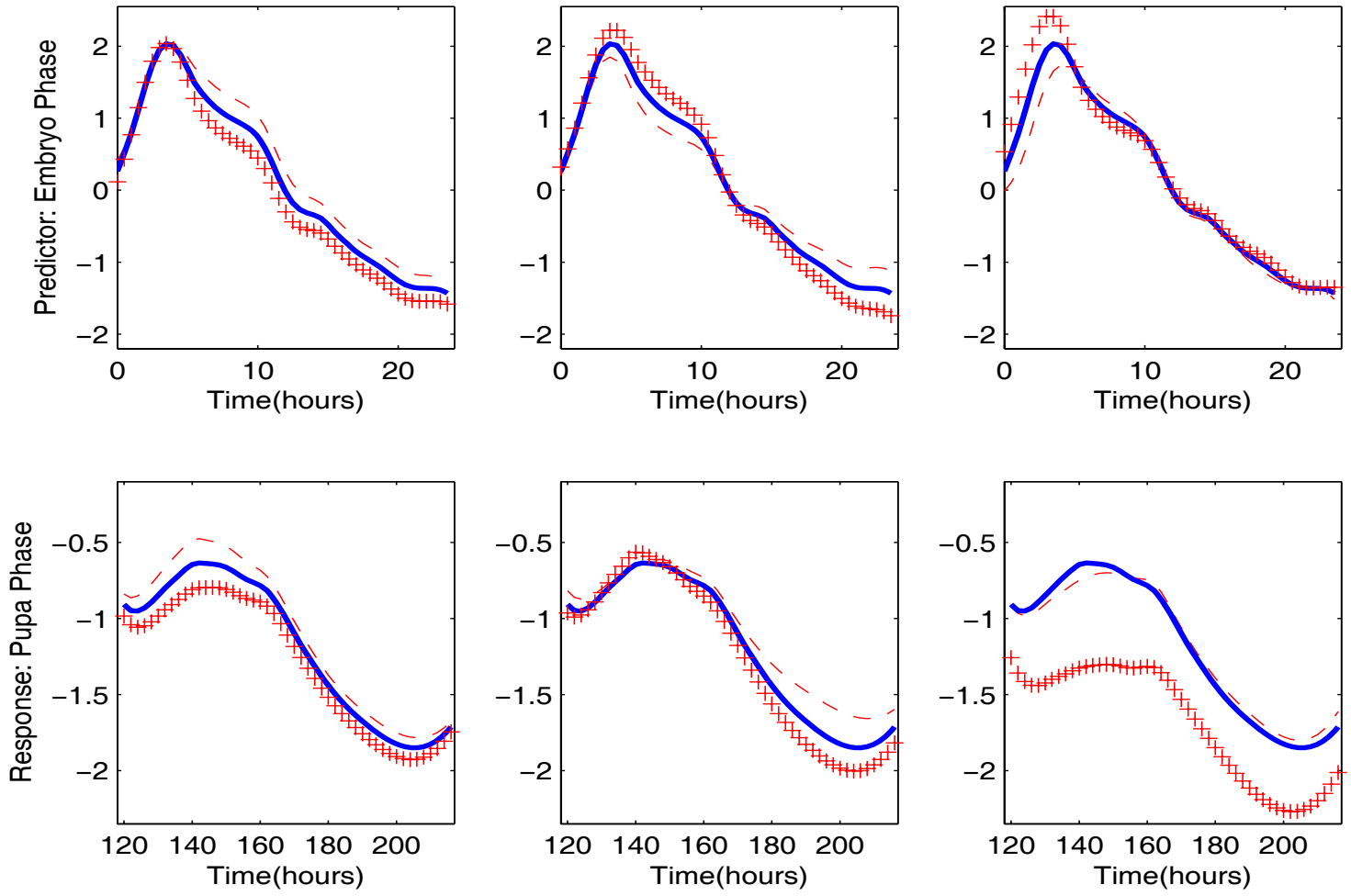
First three eigenfunctions for embryo phase (predictor) and first four eigenfunctions for pupa phase (response).

Table: Functional R^2, 25th, 50th and 75th percentiles and mean of the cross-validated observed relative prediction errors, $\text{RPE}_{(-i),f}$, comparing FAM and functional linear regression models for zygotic data.

	25th	50th	75th	Mean	R^2
FAM	.0506	.0776	.1662	.1301	0.19
LIN	.0479	.0891	.1727	.1374	0.16



Scatterplots (dots), local polynomial (solid) and linear (dashed) estimates for the regressions of estimated FPC scores of the pupa phase (y-axis) versus those for the embryo phase (x-axis).



Changes of response functions as predictor functions change in the directions of the first three eigenfunctions when fitting the Functional Additive Model.

ASYMPTOTICS FOR FAM

Employing PACE, one may show under regularity conditions that $\hat{f}_{km}$ is consistent for f_{km} and the prediction $\hat{E}(Y(t)|X^*)$ is consistent for $E(Y(t)|X^*)$ (M & Yao 2008)

Key steps for proof:

- Differences between B_{im} and $\hat{B}_{im}$ and A_{ik} and $\hat{A}_{ik}$ are asymptotically small enough to be negligible for the FAM smoothing steps.

- Perturbation analysis for linear operators, bounding the difference between operators A_G and $A_{\hat{G}}$.
- In the dense design case, one may obtain essentially 1-d rates of convergence for the component functions $\hat{f}_{km}$.

Continuous Additive Model

- FAM is additive in the functional principal components A_k, can be characterized as frequency-additive
- Is it possible to construct a time-additive functional regression model?
- Difficulty: Time domain is uncountable

- Solution is continuous additive model, which adds smoothness: and considers the limit of a sequence of additive regression models for increasingly dense time grids with additive regression functions $f_j(\cdot) = g(t_j, \cdot)$ with $E\{g(t_j, X(t_j)\} = 0$.

- This leads to
  $E\{Y \mid X(t_1), \ldots, X(t_m)\} = EY + \frac{1}{m} \sum_{j=1}^{m} g\{t_j, X(t_j)\}$
  with limit
  $E(Y \mid X) = EY + \int g\{t, X(t)\}\, dt$
  (M, Yao, Wu 2013, Mc Lean et al 2014)

- For smooth transformations ζ of $X(t)$, this model includes
  $E(Y \mid X) = EY + \int \beta(t)[\zeta\{X(t)\} - E\zeta\{X(t)\}]\, dt.$

Scalar Quantile Regression with Functional Predictors

- Functional Regression Quantiles, extending Koenker's linear regression quantile approach to functional predictors (Cardot 2005)

- Nonparametric regression quantile approach: Extend the mean regression approach to a conditional distribution target. With binary link function g,

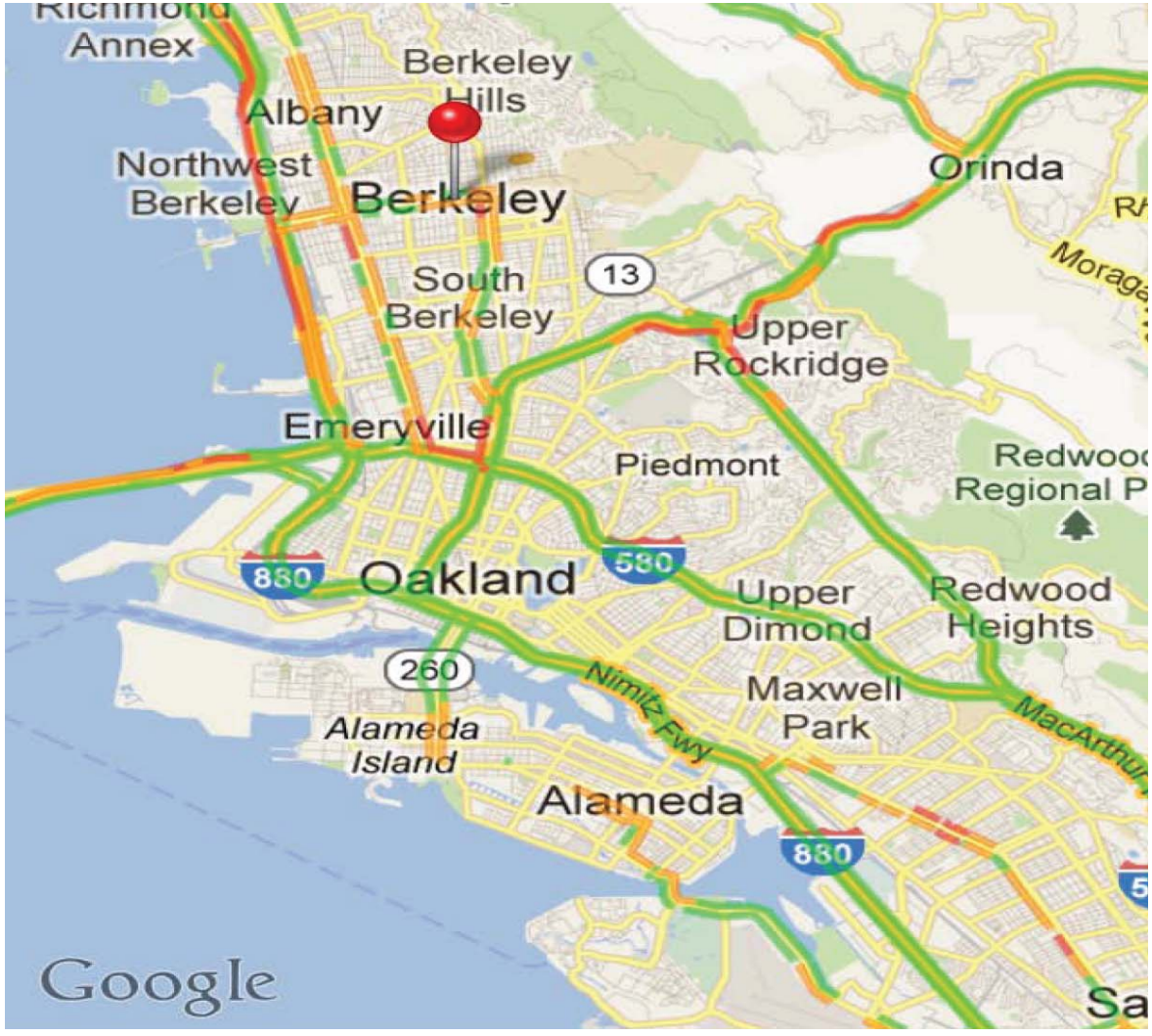
$$P(Y \leq y|X) = E(I(Y \leq y)|X) = g^{-1}(\alpha(t) + \int X^c(t)\beta(y,t)dt)$$

- Then take inverse of the conditional distribution to obtain conditional quantile function (Chen & M, 2012)

Functional Conditional Regression

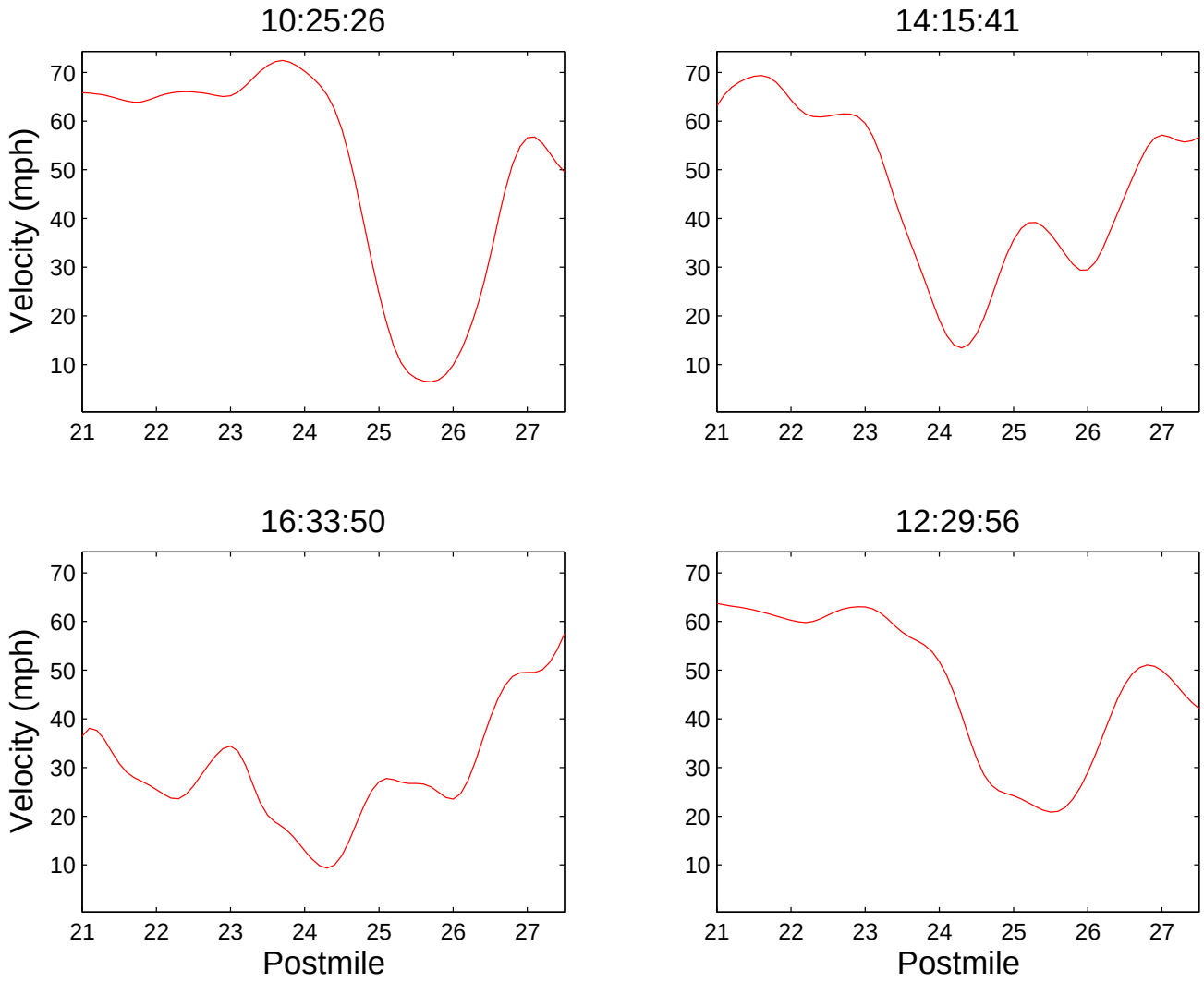
- Can the extension of mean regression to target conditional distributions be extended to the case of functional responses?
- Difficulty: Distributions for stochastic processes hard to specify
- But there is one case where this works: Gaussian processes

Illustrative Example: Traffic Data



Velocity on I-880

X = currently available velocity profile along a stretch of highway,
Y = future velocity profile at drive-through, which will determine actual travel time



Prediction for Response Functions

- Y and X are both functions
- *FPCfam*: $E(Y(t)|X) = \mu_Y(t) + \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} f_{jk}(A_k)\psi_j(t)$
- *FPCpredBands* (Chen and M 2012): Global prediction bands for Y conditional on X
- For Gaussian process: $E(Y|X)$ and $\mathrm{cov}(Y|X)$
- Common principal component assumption
  Additive assumption

$$\mathrm{cov}(Y(t_1), Y(t_2) \mid X)$$
$$= G_{YY}(t_1, t_2) + \sum_{j=1}^{\infty}[\sum_{k=1}^{\infty} g_{jk}(A_k) - (\sum_{k=1}^{\infty} f_{jk}(A_k))^2]\psi_j(t_1)\psi_j(t_2)$$

Modeling the Prediction Bands

- Global prediction bands for Gaussian case:

$$P(\mu(t) - D_X(t) \leq Y_X(t) \leq \mu(t) + D_X(t) \mid X) \geq 1 - \alpha$$

  where $D_X(t) = \mathcal{C}_\alpha \left\{ \mathrm{var}(Y(t)|X) \right\}^{1/2}$

- For more general random processes:

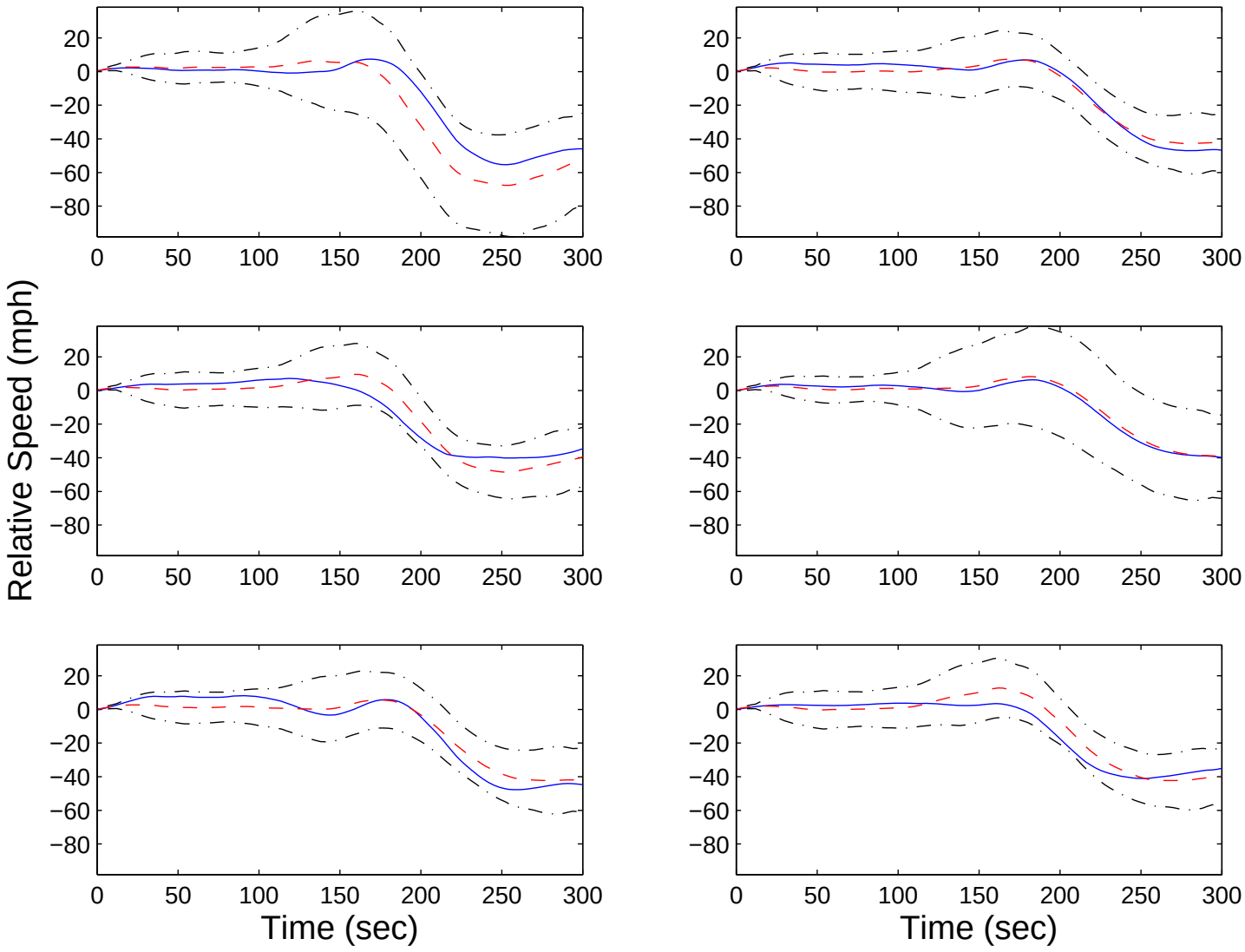
$$E\left\{P(L_X(t) \leq Y_X(t) \leq U_X(t) \mid X)\right\} \geq 1 - \alpha$$

- Find $\mathcal{C}_\alpha$ by empirical coverage on test sets

'Mobile Century' Data

- Joint UC Berkeley - Nokia project (Herrera et al., 2010)
- Students were hired to drive on a segment of highway I-880 and send data (time, location, and speed) back through GPS enabled mobile phones.
- The follow-up project 'Mobile Millennium' is generating more data.

Estimated 90% Prediction Regions



Lectures on FDA – Part V

Derivatives

FUNCTIONAL GRADIENTS

For functional linear regression with scalar responses:
Derivative of an operator $\Psi : L^2 \to \mathcal{R}$ at $x = \sum_k A_{xk}\phi_k$ is a linear operator $\Psi_x^{(1)}$: For functions u and scalars δ,

$$\Psi(x + \delta u) = \Psi(x) + \delta\, \Psi_x^{(1)}(u) + o(\delta) \quad \text{as } \delta \to 0.$$

The functional derivative operator at x is characterized by the functional directional derivatives

$$\Psi_x^{(1)}(\phi_k) = \gamma_{xk} \in \mathcal{R},\ k = 1, 2, \ldots$$

in the directions of the basis functions ϕ_k. This is a data analytic implementation of Gâteaux derivatives.

Representation

$$\Psi_x^{(1)} = \sum_{k=1}^{\infty} \gamma_{xk}\, \Phi_k,$$

where $\gamma_{xk} = \Psi_x^{(1)}(\phi_k)$ is a scalar, and Φ_k denotes the linear projection operator with

$$\Phi_k(u) = A_{uk} = \int u(t)\phi_k(t)dt, \quad \text{for all} \quad u \in L^2(\mathcal{T}).$$

Example: Functional linear model. Representing the regression parameter function β in the eigenbasis ϕ_k, $\beta(t) = \sum_k \beta_k \phi_k(t)$, $t \in \mathcal{T}$, leads to

$$\Psi_L(X) = \mu_Y + \sum_{k=1}^{\infty} \beta_k A_{Xk} = \mu_Y + \sum_{k=1}^{\infty} \beta_k \Phi_k(X).$$

For any δ and arbitrary square integrable functions with representations $u = \sum_k A_{uk}\phi_k$ and $x = \sum_k A_{xk}\phi_k$,

$$\Psi_L(x + \delta u) = \mu_Y + \sum_k \beta_k (A_{xk} + \delta A_{uk}) = \Psi_L(x) + \delta \sum_k \beta_k A_{uk}.$$

Then $\Psi_x^{(1)} = \sum_{k=1}^{\infty} \beta_k \Phi_k \quad \Rightarrow \quad \gamma_{xk} = \beta_k$. The functional derivative does not depend on x, as $\Psi_x^{(1)}(\phi_k) = \beta_k$.

ADDITIVE MODELING OF FUNCTIONAL GRADIENTS

Consider additive functional operator

$$\Psi_A(X) = E(Y^c|X) = \sum_{k=1}^{\infty} f_k(A_{Xk}),$$

subject to $Ef_k(A_{Xk}) = 0$, $k = 1, \ldots,$ for FPC scores A_{Xk}.

For functions $x = \sum_k A_{xk}\phi_k$ and $u = \sum_k A_{uk}\phi_k$,

$$\Psi_A(x + \delta u) = \sum_k f_k(A_{xk} + \delta A_{uk}) = \Psi_A(x) + \delta \sum_k f_k^{(1)}(A_{xk})A_{uk} + o(\delta),$$

so that for the functional additive model,

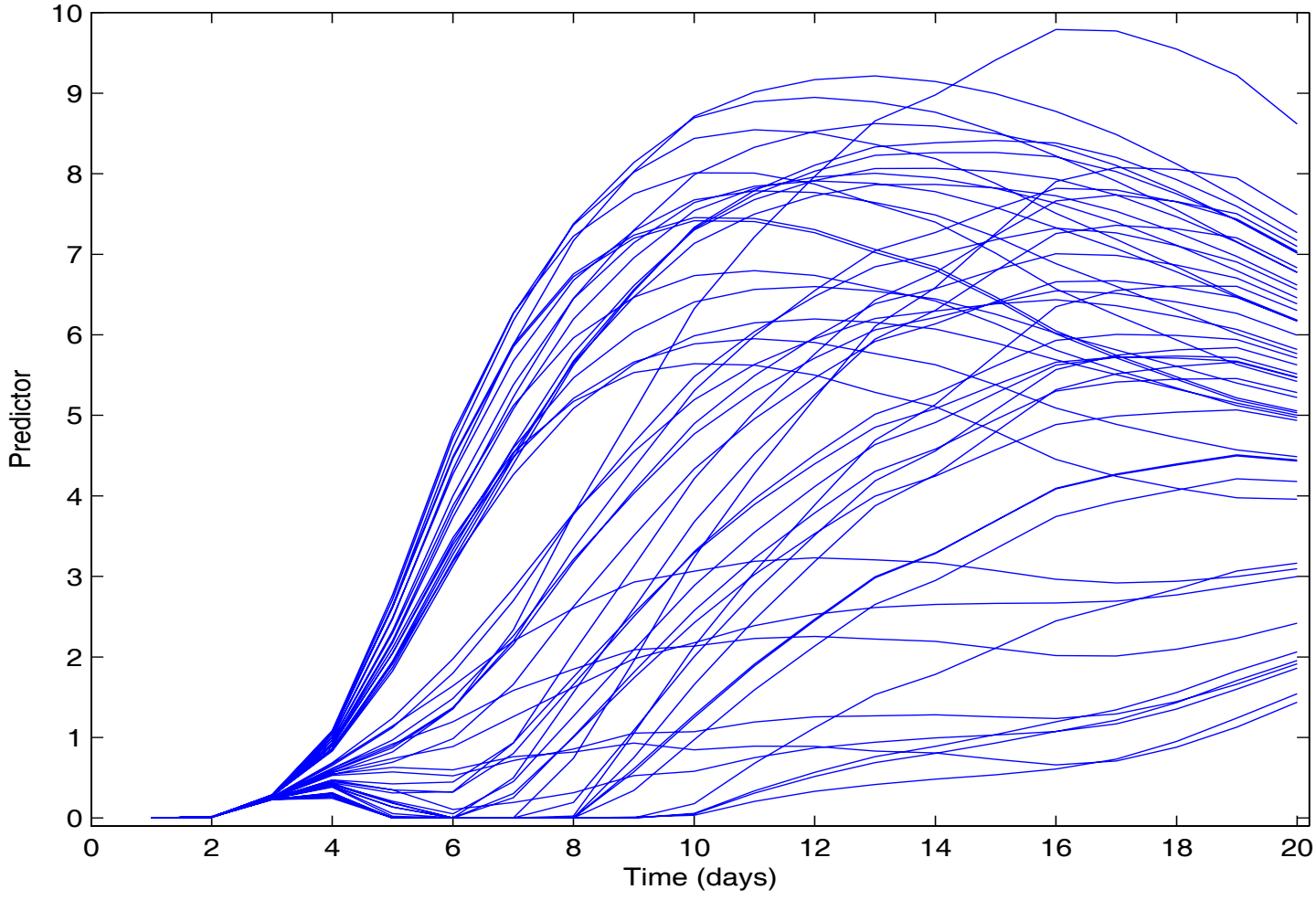
$$\Psi_{A,x}^{(1)}(u) = \sum_{k=1}^{\infty} f_k^{(1)}(A_{xk})A_{uk} = \sum_{k=1}^{\infty} \gamma_{A,xk}\Phi_k(u), \quad \gamma_{A,xk} = f_k^{(1)}(A_{xk}).$$

- Can easily extend to higher order derivatives due to additive structure

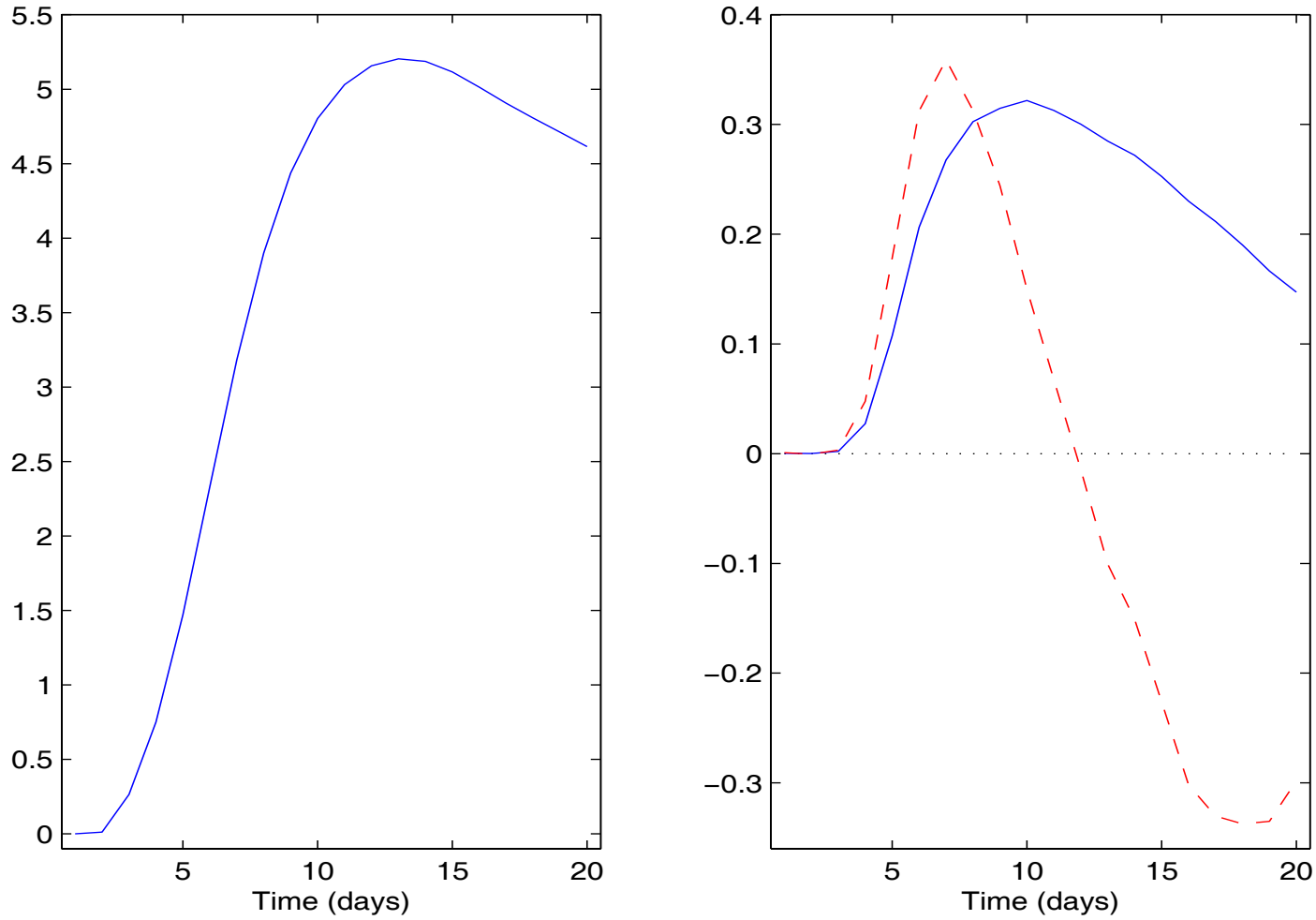
- Asymptotics: For densely sampled functions, may obtain derivatives through derivative estimates of the additive functions, with the 1-d rates of convergence for derivative estimation.

GRADIENTS FOR EGG-LAYING

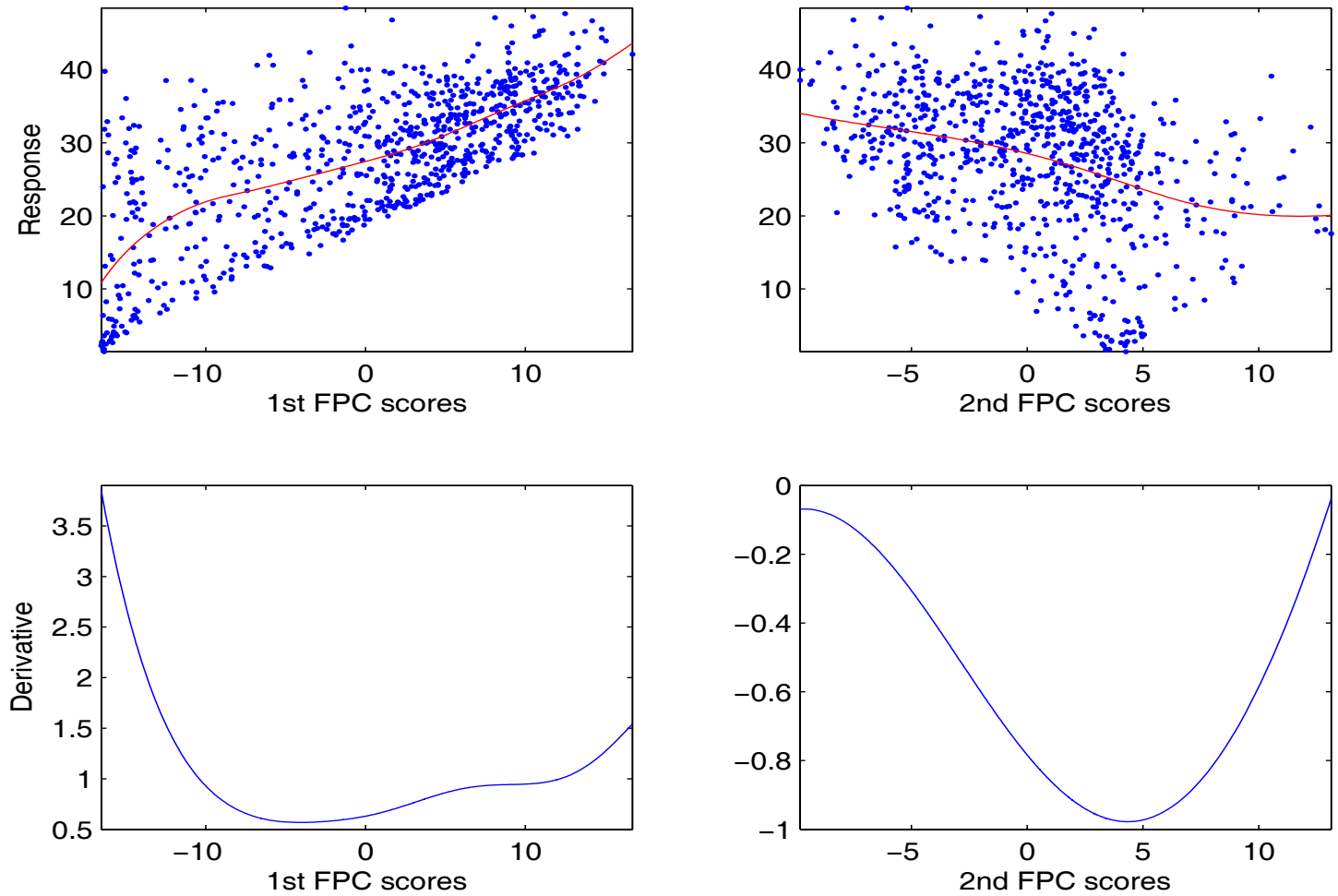
- Predictor functions: Egg-laying trajectories (daily egg counts) for cohort of 818 female medflies (Carey et al. 98) that live ≤ 20 days.
- Response: Lifetime fertility = total number of eggs laid over lifetime
- Preprocessing: Square root transformation of egg counts
- Question: How do early reproductive trajectories influence overall reproductive success.
- Tools: Gradient field and its visualization



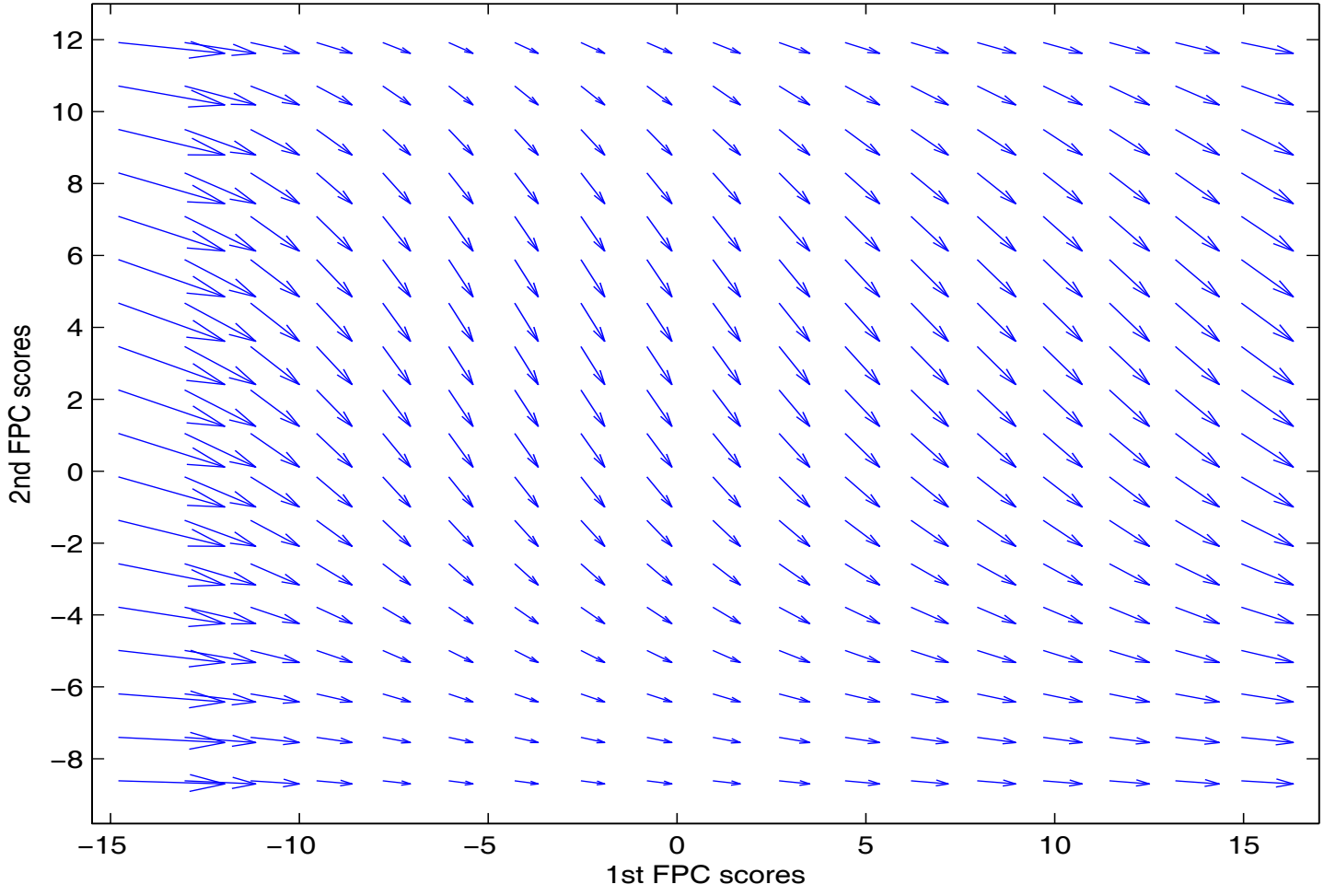
Egg-laying trajectories (predictors, smoothed) for 50 randomly selected flies, from a total of 818 medflies, for the first 20 days of their lifespan. Response=total eggs (reproductive success)



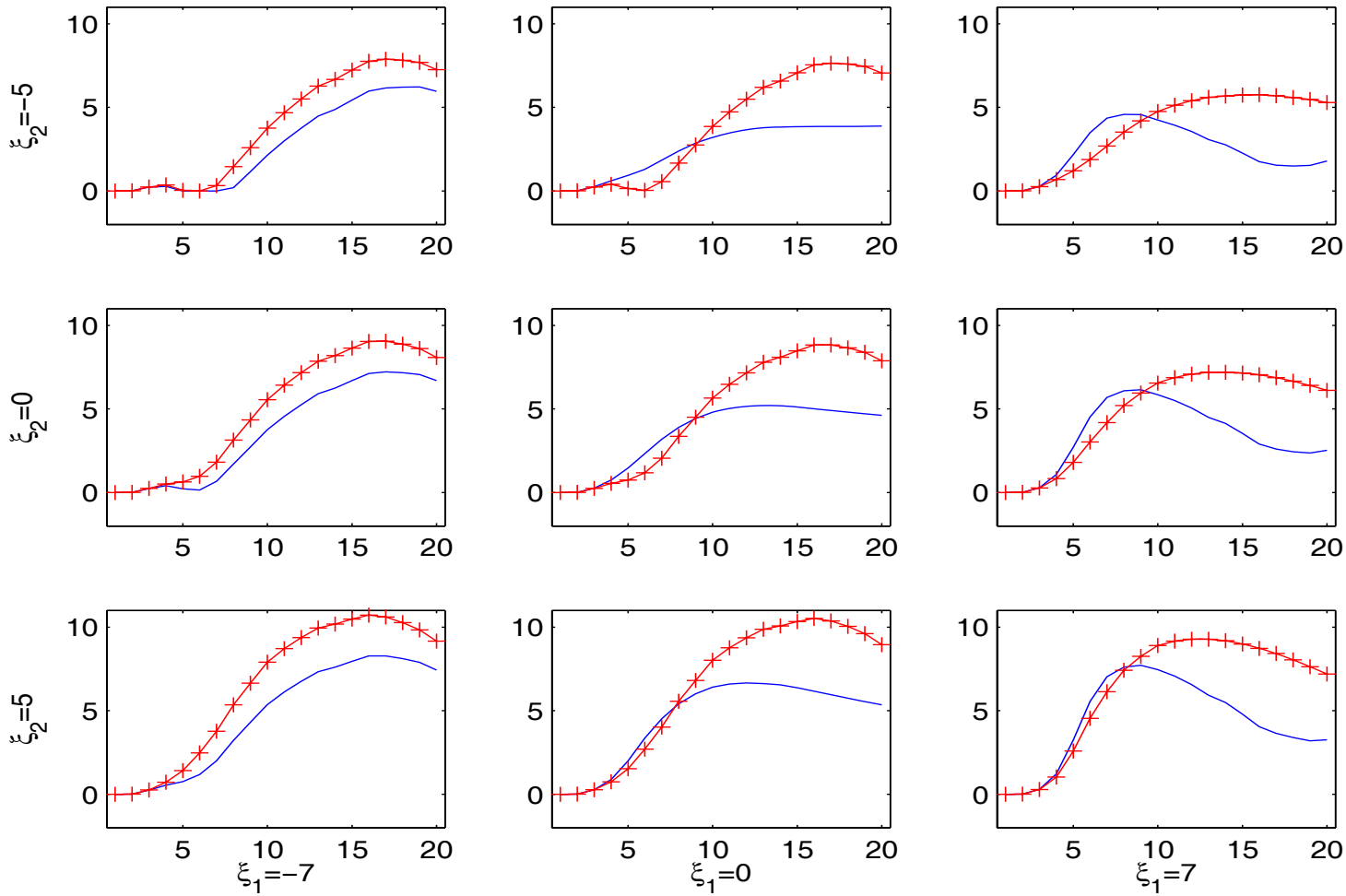
Smooth estimates of mean function (left panel) and first two eigenfunctions (right panel) of the predictor trajectories, explaining 72.1% (solid) and 18.6% (dashed) of the total variation, respectively.



Top panels: Nonparametric regression of the response (total fertility) on the first (left) and second (right) FPC scores of predictor processes. Bottom panels: Estimated derivatives of the smooth regression functions.



Estimated functional gradient field for total fertility, differentiated against the predictor process, expressed in terms of gradients of the response with respect to first (x-axis) and second (y-axis) FPC scores.



Visualization of the shape changes in fertility trajectories along the gradients: Bases are blue trajectories (9 combinations of the FPC scores $A_1 = \{-7, 0, 7\}$ and $A_2 = \{-5, 0, -5\}$, tip of the arrows red trajectories.

ESTIMATING DERIVATIVES FROM SPARSE DATA

Differentiating Karhunen-Loève representation:

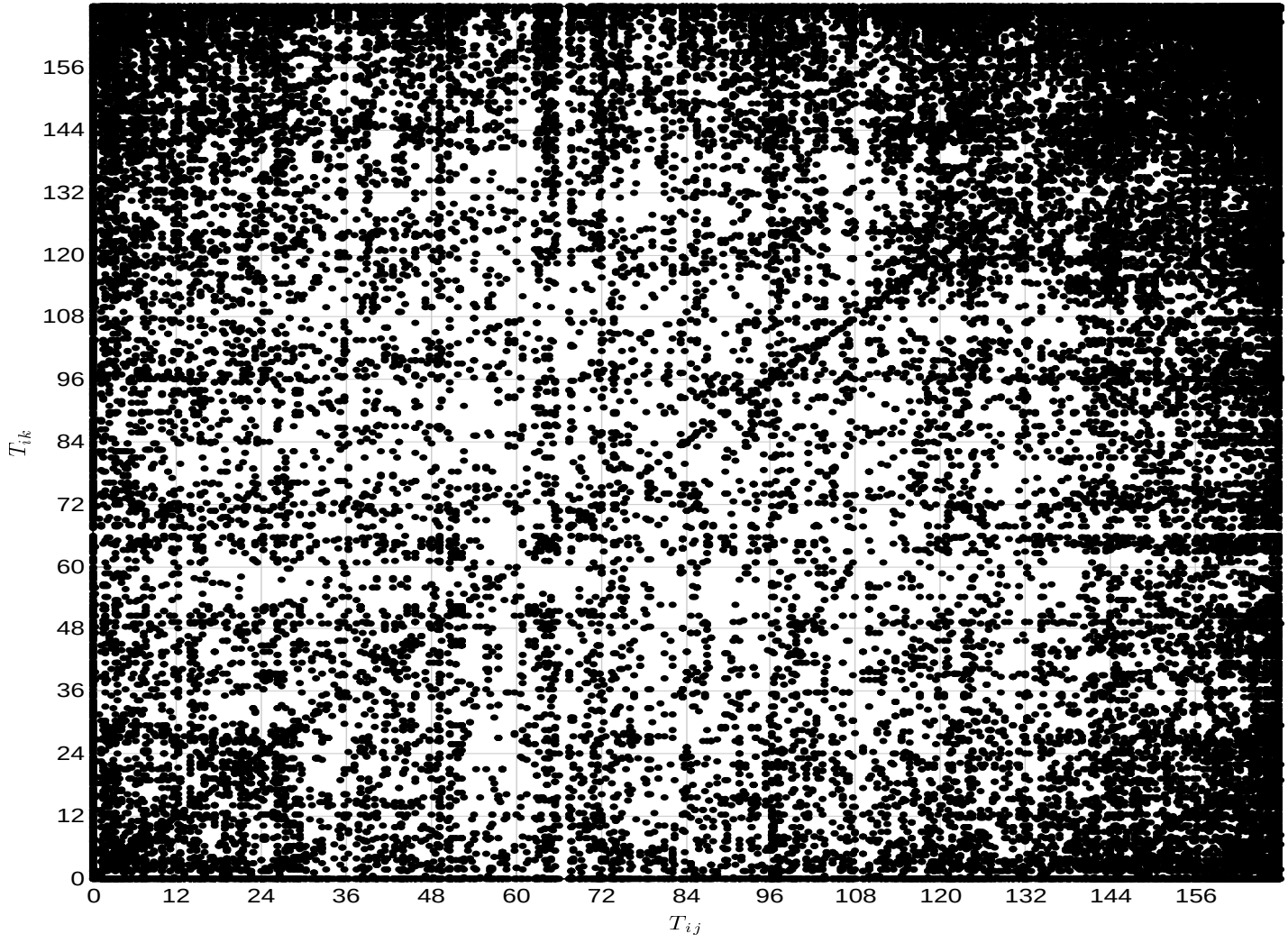
$$X_i^{(\nu)}(t) = \mu^{(\nu)}(t) + \sum_{k=1}^{\infty} A_{ik}\phi_k^{(\nu)}(t), \quad \nu = 0, 1, \ldots.$$

- Obtain estimated random effects A_{ik} by conditioning as before
- Estimate $\mu^{(\nu)}(t)$ by known nonparametric 1-d differentiation, applied to pooled scatterplots.
- How to obtain $\phi_k^{(\nu)}$? Observe

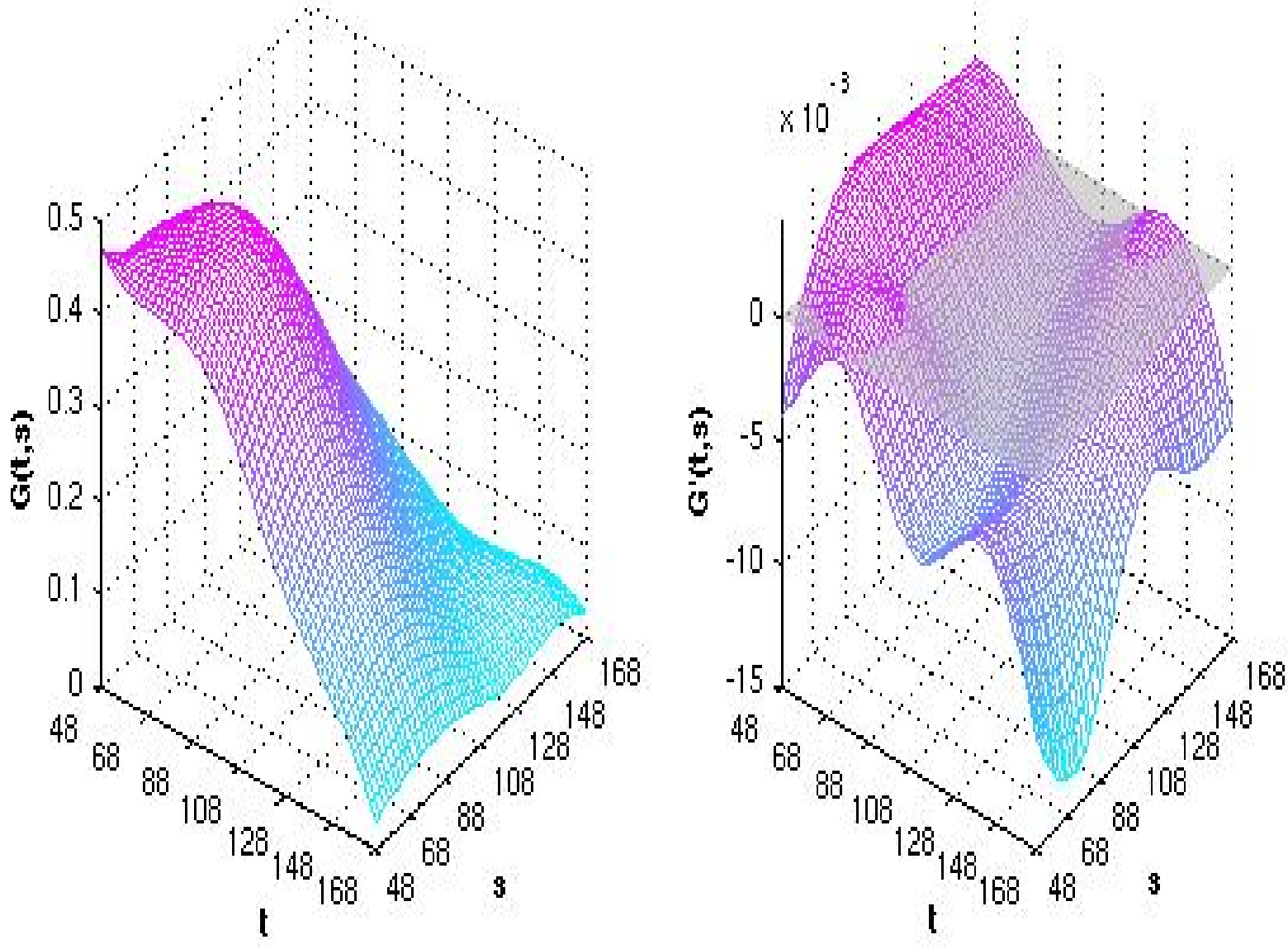
$$\frac{d^\nu}{dt^\nu}\int_{\mathcal{T}} G(t,s)\phi_k(s)ds = \lambda_k \frac{d^\nu}{dt^\nu}\phi_k(t),$$

implying

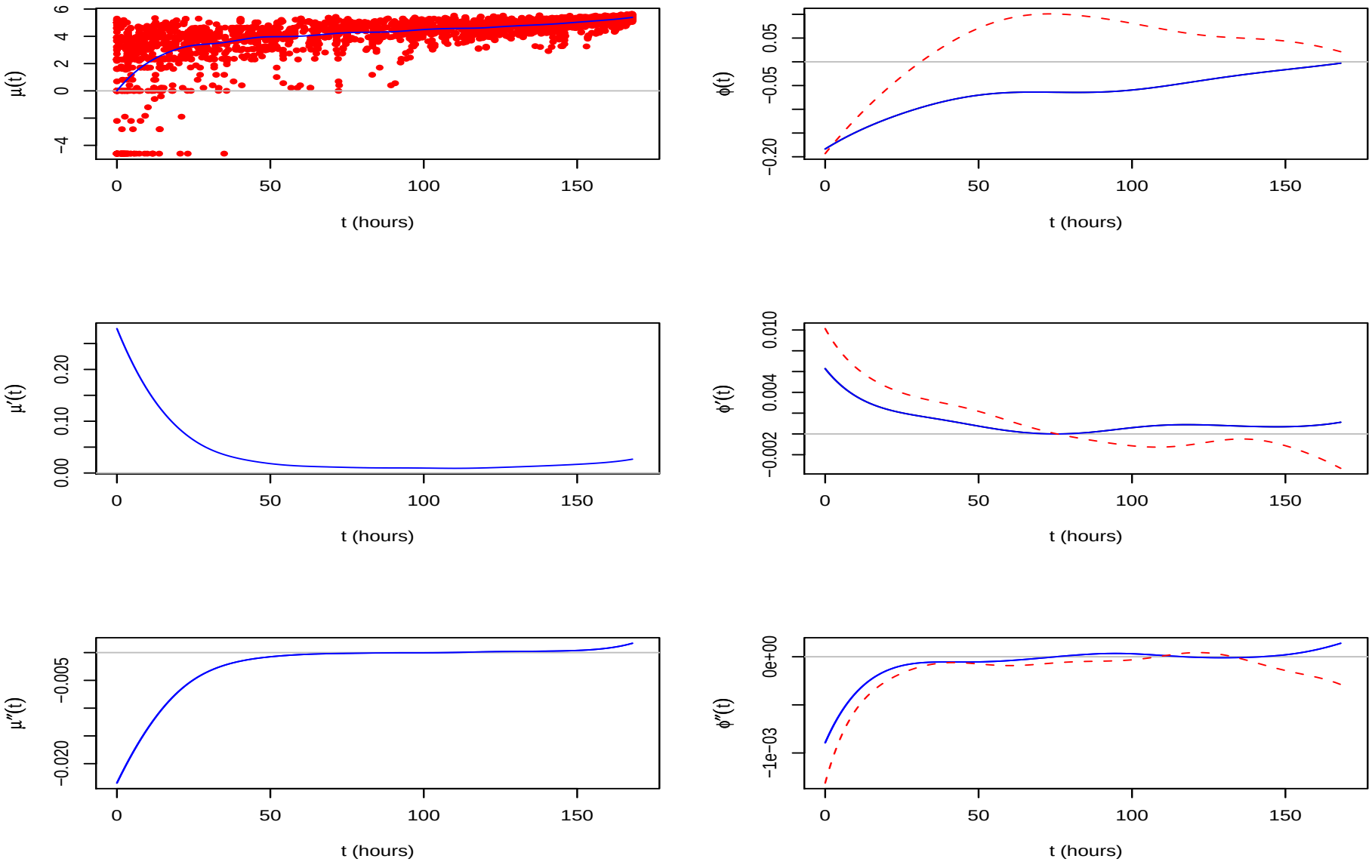
$$\phi_k^{(\nu)}(t) = \frac{1}{\lambda_k}\int_{\mathcal{T}} \frac{\partial^\nu}{\partial t^\nu} G(t,s)\phi_k(s)ds.$$



Locations of all pairs of points where bids are recorded for auction data.



Estimated covariance surface from all pairs and estimated partial derivative surface for auction data.



Estimates of mean and first two eigenfunctions and their first two derivatives for auction data.

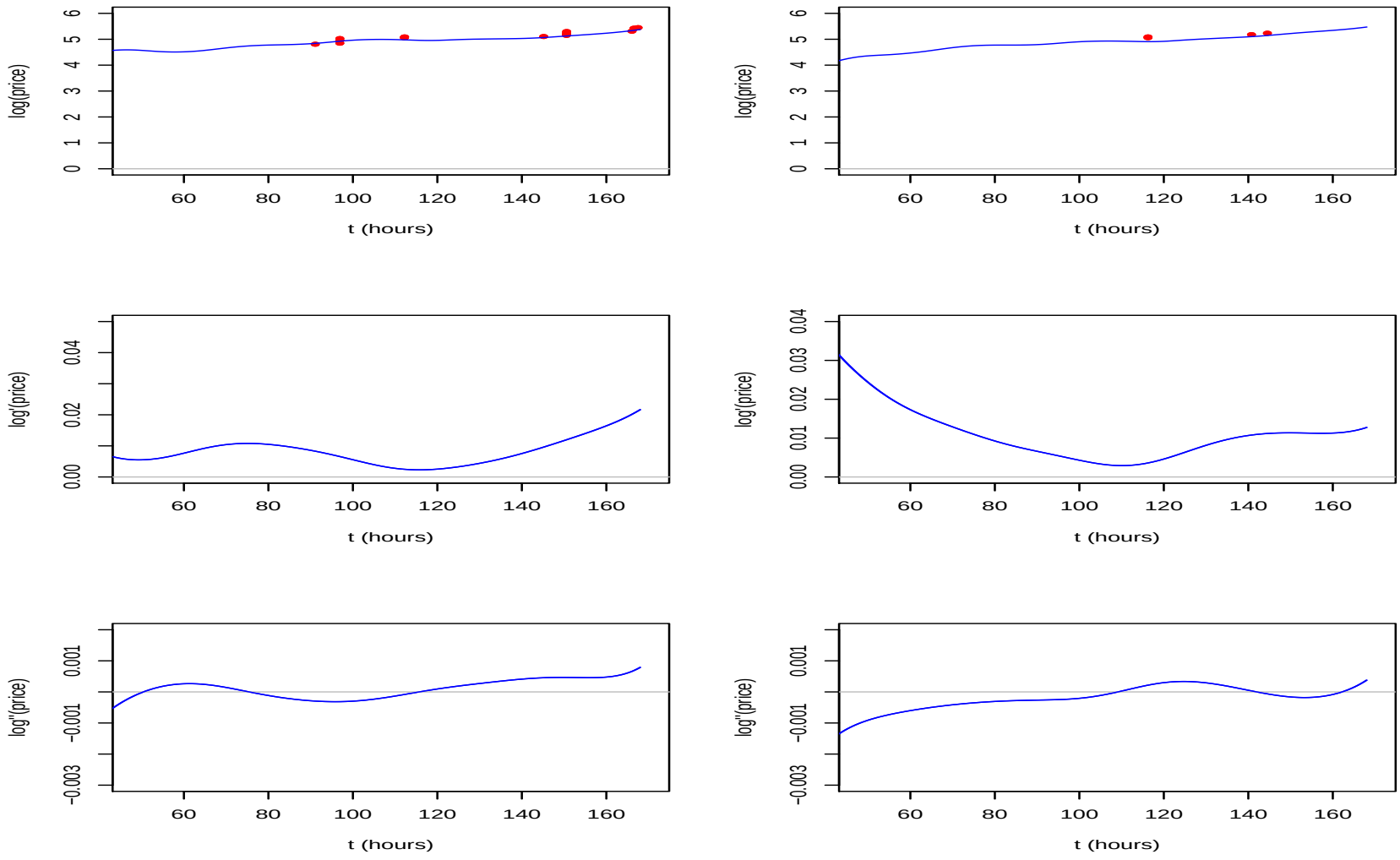
DERIVATIVES OF TRAJECTORIES

- Obtain

$$\hat{X}_{i,K}^{(\nu)}(t) = \hat{\mu}^{(\nu)}(t) + \sum_{k=1}^{K} \hat{A}_{ik} \hat{\phi}_k^{(\nu)}(t).$$

  for the derivatives of the random trajectories X_i.

- Choosing the number of included components K: e.g. by Fraction of variance explained
- Asymptotic convergence results and confidence intervals for the case of a Gaussian process
- In simulations, this differentiation method works much better than single curve derivative estimation (splines, kernels, ...)



Fitted price trajectories and their first two derivatives for two auctions.

DYNAMICS OF GAUSSIAN PROCESSES

From the Karhunen-Loève representation of processes X, obtain for the covariance function for derivatives

$$\operatorname{cov}\{X^{(\nu_1)}(t), X^{(\nu_2)}(s)\} = \sum_{k=1}^{\infty} \lambda_k \phi_k^{(\nu_1)}(t) \phi_k^{(\nu_2)}(s),\ \nu_1, \nu_2 \in \{0, 1\},\ s, t \in \mathcal{T}$$

Assuming Gaussianity of X,

$$\begin{pmatrix} X^{(1)}(t) - \mu^{(1)}(t) \\ X(t) - \mu(t) \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{\infty} A_k \phi_k^{(1)}(t) \\ \sum_{k=1}^{\infty} A_k \phi_k(t) \end{pmatrix}$$

$$\sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sum_{k=1}^{\infty} \lambda_k \phi_k^{(1)}(t)^2 & \sum_{k=1}^{\infty} \lambda_k \phi_k^{(1)}(t) \phi_k(t) \\ \sum_{k=1}^{\infty} \lambda_k \phi_k^{(1)}(t) \phi_k(t) & \sum_{k=1}^{\infty} \lambda_k \phi_k(t)^2 \end{pmatrix} \right)$$

EMPIRICAL DIFFERENTIAL EQUATION

Population level: $E\{X^{(1)}(t) - \mu^{(1)}(t) \mid X(t)\} = \beta(t)\{X(t) - \mu(t)\}$

Subject level:

$$X^{(1)}(t) - \mu^{(1)}(t) = \beta(t)\{X(t) - \mu(t)\} + Z(t), \ t \in \mathcal{T},$$

with varying coefficient function

$$\begin{aligned} \beta(t) &= \frac{\mathrm{cov}\{X^{(1)}(t), X(t)\}}{\mathrm{var}\{X(t)\}} = \frac{\sum_{k=1}^{\infty} \lambda_k \phi_k^{(1)}(t)\phi_k(t)}{\sum_{k=1}^{\infty} \lambda_k \phi_k(t)^2} \\ &= \frac{1}{2}\frac{d}{dt}\log[\mathrm{var}\{X(t)\}], \ t \in \mathcal{T}, \end{aligned}$$

and Gaussian drift process Z.

DRIFT PROCESS

Gaussian drift process is such that
(i) $Z(t)$, $X(t)$ are independent at each $t \in \mathcal{T}$; (ii) $E\{Z(t)\} = 0$;
(iii) Z has the representation

$$Z(t) = \sum_{k=1}^{\infty} \sqrt{\frac{\lambda_k}{2T^3}}(2k-1)\pi \int_0^T \sin\{\frac{(2k-1)\pi}{2T}\, u\} \times\{\phi_k^{(1)}(t) - \beta(t)\phi(t)\}\, dW(u)$$

Integral equation version

$$\begin{aligned} X(t) &= X(s) + \{\mu(t) - \mu(s)\} \\ &+ \int_s^t \beta(u)\{X(u) - \mu(u)\}\, du + \int_s^t Z(u)\, du, \end{aligned}$$

for any $s, t \in \mathcal{T}$, $s < t$.

LEARNING GAUSSIAN DYNAMICS

- For varying coefficient function β use plug-in estimates

$$\hat{\beta}(t) = \frac{\sum_{k=1}^{K} \hat{\lambda}_k \hat{\phi}_k^{(1)}(t) \hat{\phi}_k(t)}{\sum_{k=1}^{K} \hat{\lambda}_k \hat{\phi}_k^2(t)}.$$

- dynamic regression to the mean (negative β)
- dynamic exponential growth (positive β)
- Interpretation within population model
  $E\{X^{(1)}(t) - \mu^{(1)}(t) \mid X(t)\} = \beta(t)\{X(t) - \mu(t)\}$

For drift process Z

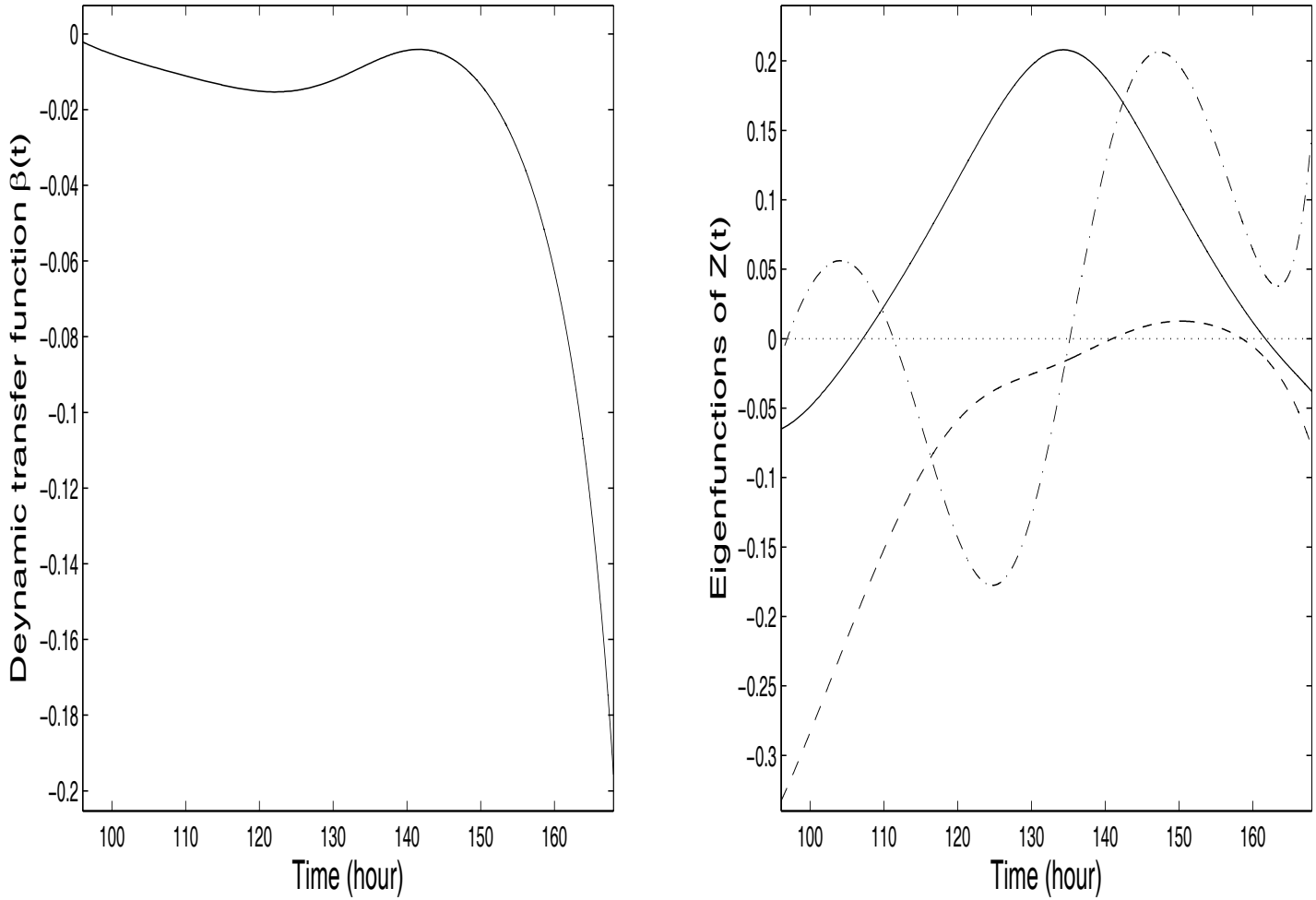
$$\mathrm{var}(Z(t)) =$$
$$\left(\sum_k \lambda_k(\phi_k^{(1)}(t))^2 \sum_k \lambda_k \phi_k^2(t) - \{\sum_{k=1}^{\infty} \lambda_k \phi_k^{(1)}(t)\phi_k(t)\}^2\right) / \sum_k \lambda_k \phi_k^2(t)$$

and

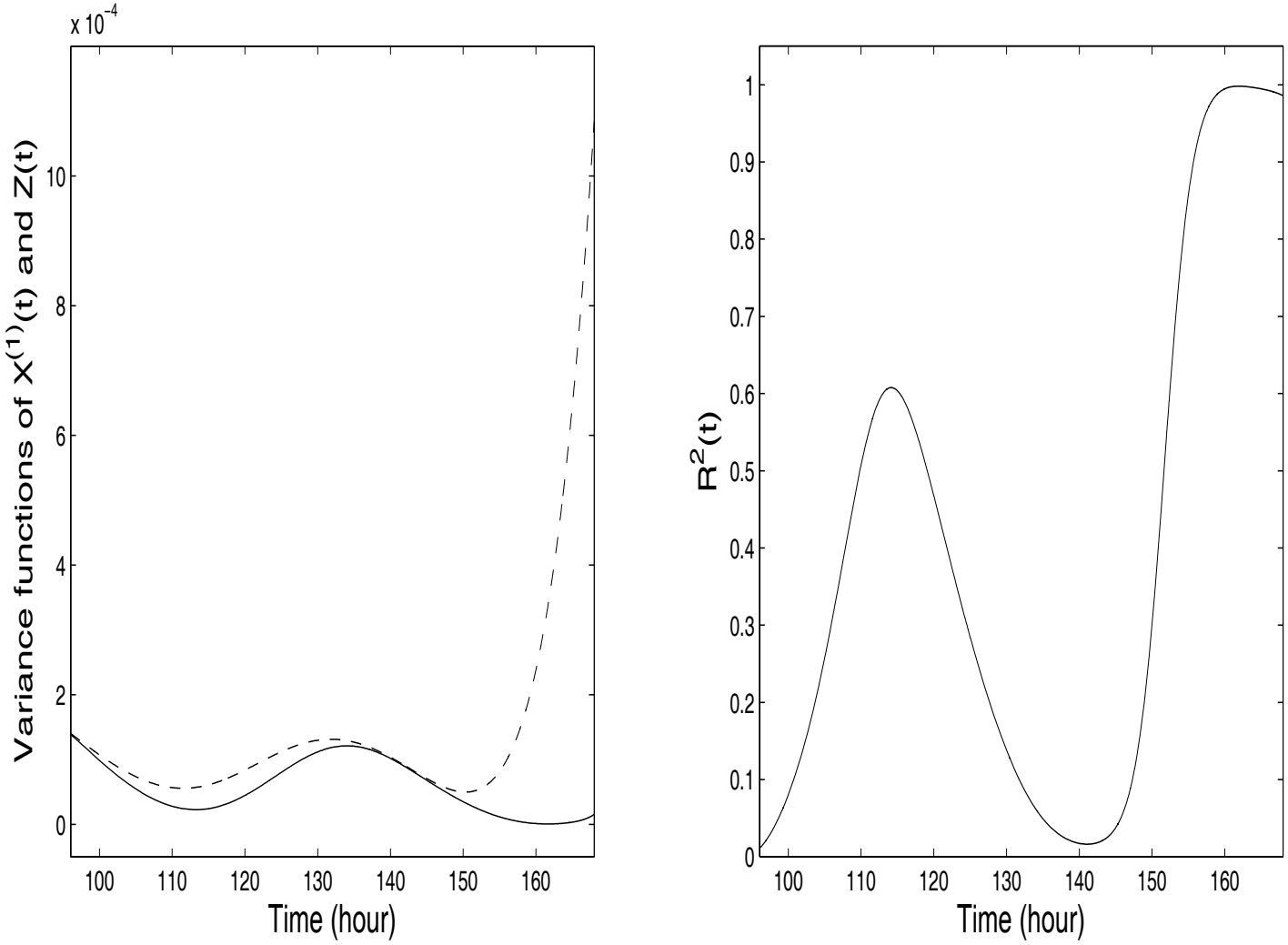
$$\mathrm{var}\{X^{(1)}(t)\} = \beta(t)^2 \mathrm{var}\{X(t)\} + \mathrm{var}\{Z(t)\}.$$

Then the fraction of the variance of $X^{(1)}(t)$ explained by the deterministic part of the differential equation is given by:

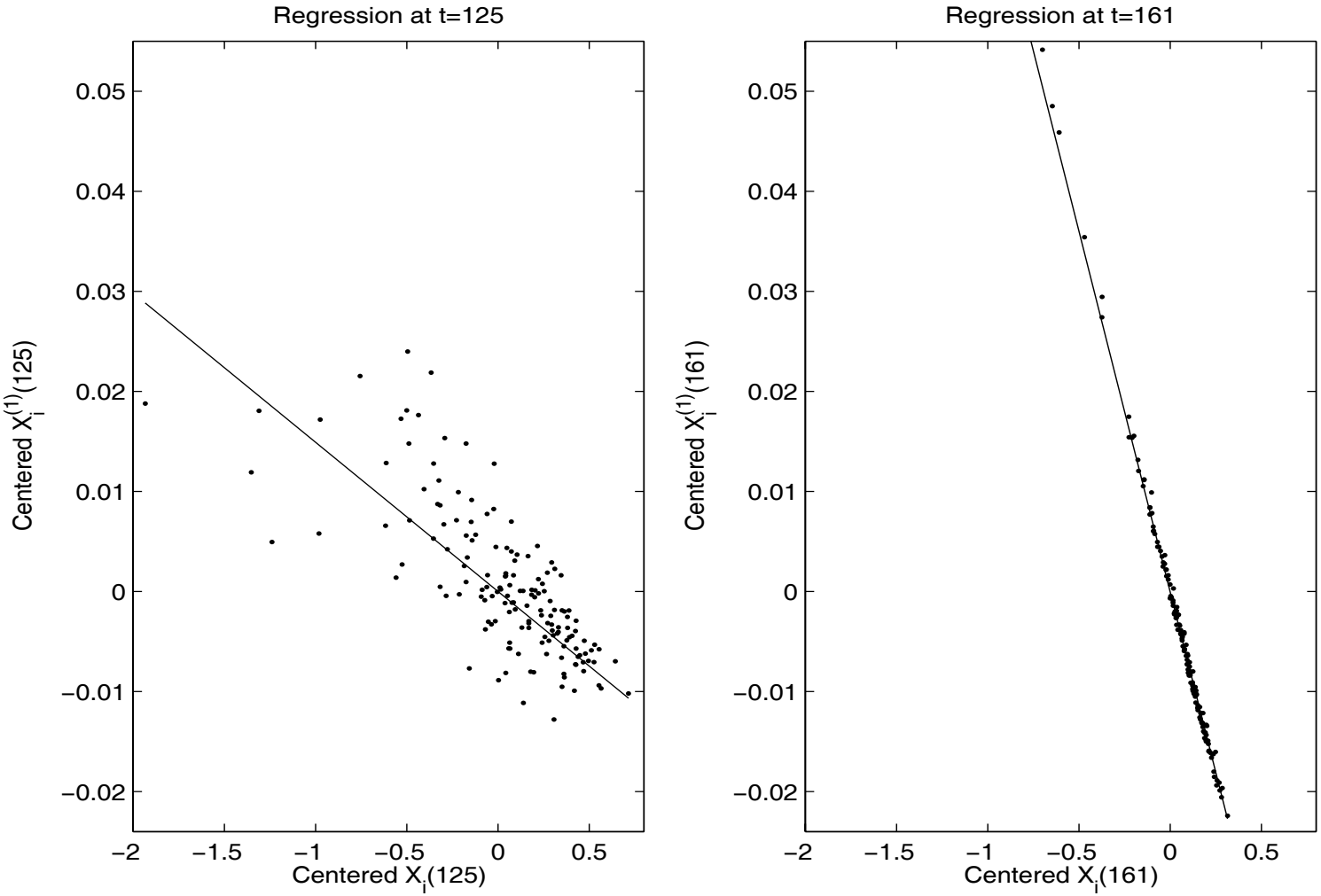
$$R^2(t) = \frac{\mathrm{var}\{\beta(t)X(t)\}}{\mathrm{var}\{X^{(1)}(t)\}} = \frac{\{\sum_{k=1}^{\infty} \lambda_k \phi_k^{(1)}(t)\phi_k(t)\}^2}{\sum_{k=1}^{\infty} \lambda_k \phi_k(t)^2 \sum_{k=1}^{\infty} \lambda_k \phi_k^{(1)}(t)^2}.$$



Left: Smooth estimate of the dynamic varying coefficient function β for auction data. Right: Smooth estimates of the first (solid), second (dashed) and third (dash-dotted) eigenfunction of drift process Z.



Left: Smooth estimates of the variance functions of $X^{(1)}(t)$ (dashed) and $Z(t)$ (solid). Right: Smooth estimate of $R^2(t)$, the variance explained by the deterministic part of the dynamic equation at time t.



Regression of $X_i^{(1)}(t)$ on $X_i(t)$ (both centered) at $t = 125$ hours (left panel) and $t = 161$ hours (right panel), respectively, with regression slopes $\beta(125) = -.015$ and coefficient of determination $R^2(125) = 0.28$, respectively, $\beta(161) = -.072$ and $R^2(161) = 0.99$.

LEARNING DYNAMICS – NON-GAUSSIAN CASE

- Data Model. For n realizations X_i of an underlying process X, have N_i measurements Y_{ij} $(i = 1, \ldots, n,\ j = 1, \ldots, N_i)$,

$$Y_{ij} = Y_i(t_{ij}) \quad = \quad X_i(t_{ij}) + \epsilon_{ij},$$

  with iid zero mean finite variance measurement errors ϵ_{ij}.

- Linear Gaussian Dynamics. As before, with varying coefficient function β,

$$X'(t) \quad = \quad \mu_{X'}(t) + \beta(t)\{X(t) - \mu_X(t)\} + Z_2(t),$$

  where Z_2 is a zero mean drift process with $\mathrm{cov}\{Z_2(t), X(t)\} = 0$.

- General Dynamics. There always exists a function f with

$$E\{X'(t) \mid X(t)\} = f\{t, X(t)\}, \quad X'(t) = f\{t, X(t)\} + Z(t)\ ,$$

  with $E\{Z(t) \mid X(t)\} = 0$ almost surely and where f is unknown. Learning dynamics corresponds to inferring f.

- Special Case: Autonomous Dynamics.

$$E\{X'(t) \mid X(t)\} = f_1(X(t)), \quad f_1 \text{ unknown}$$

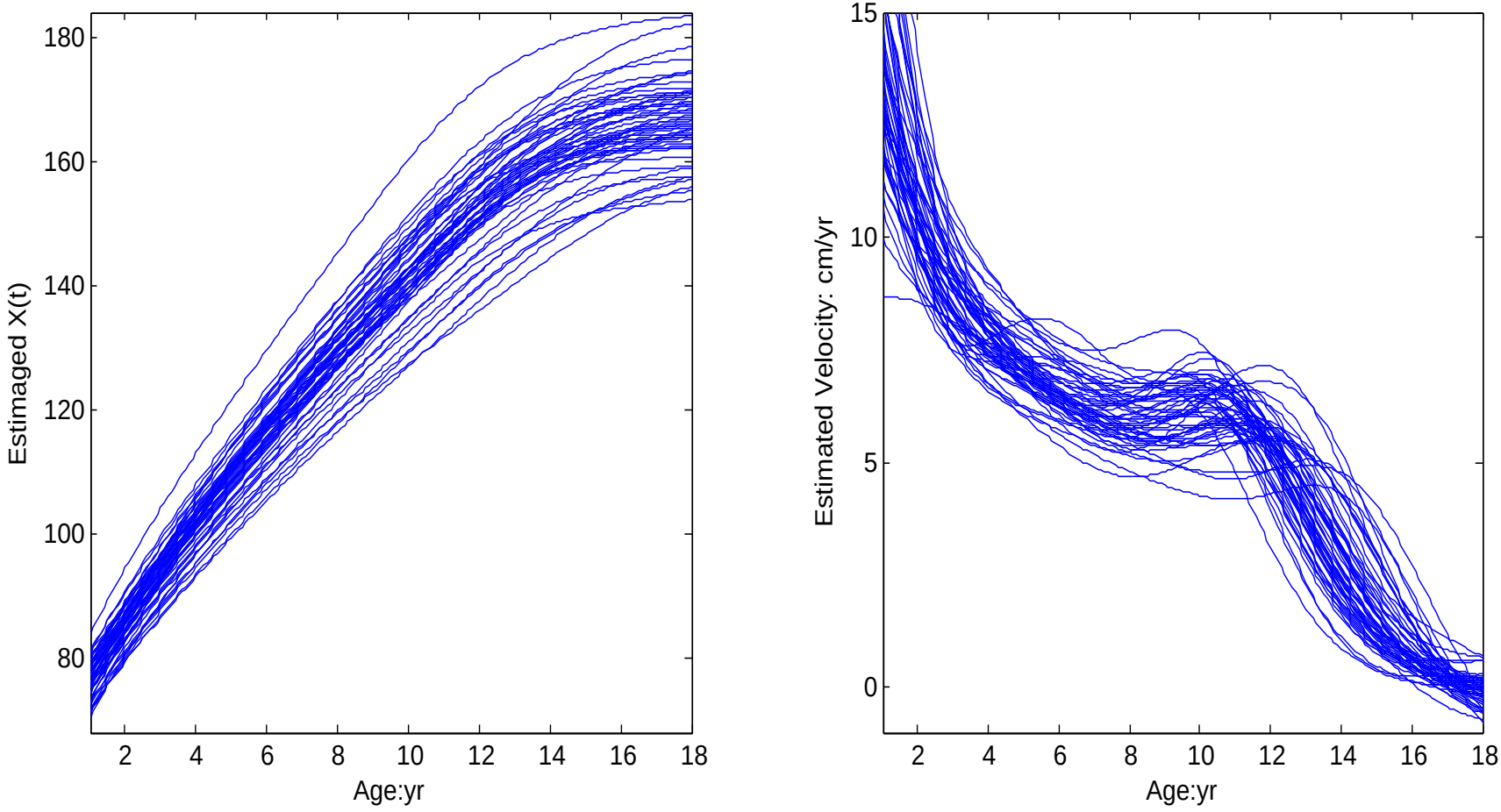
- Parametric Dynamics. Parametric differential equations

$$X_i'(t) = g\{t, X_i(t), \theta_i\}$$

require extensive knowledge of underlying system – often incorrect and hard to fit. Not much known for incorporating random effects θ_i.

BERKELEY LONGITUDINAL GROWTH STUDY

- Dynamics of Human Growth of Interest
- Nonlinear Parametric Models: Preece-Baines, Triple-Logistic Subject-by-subject fitting, limited efficiency
- Berkeley Growth Study – 54 girls with 31 height measurements for ages 1 to 18, recorded at different time intervals, ranging from three months (from 1 to 2 years old), six months (from 8 to 18 years old), to one year (from 3 to 8 years old).
- Learning dynamics:
  – Gain a better understanding of the growth process.
  – Distinguish between normal and pathological patterns of development.



Left panel: Estimated growth curves for 54 girls. Right panel: Estimated growth velocity trajectories for 54 girls.

ESTIMATING THE DRIVING FUNCTION f

Adopt a two-step kernel smoothing approach to obtain an estimator for f in $E\{X'(t) \mid X(t)\} = f\{t, X(t)\}$:

- Step 1: Obtaining estimates for $X(t)$ and $X'(t)$:

$$\widehat{X}_i(t) = \frac{1}{h_X} \sum_{j=1}^{N_i} \int_{s_{j-1}}^{s_j} Y_{ij} K\left(\frac{u-t}{h_X}\right) du,$$

$$\widehat{X'}_i(t) = \frac{1}{h_{X'}^2} \sum_{j=1}^{N_i} \int_{s_{j-1}}^{s_j} Y_{ij} K_2\left(\frac{u-t}{h_{X'}}\right) du,$$

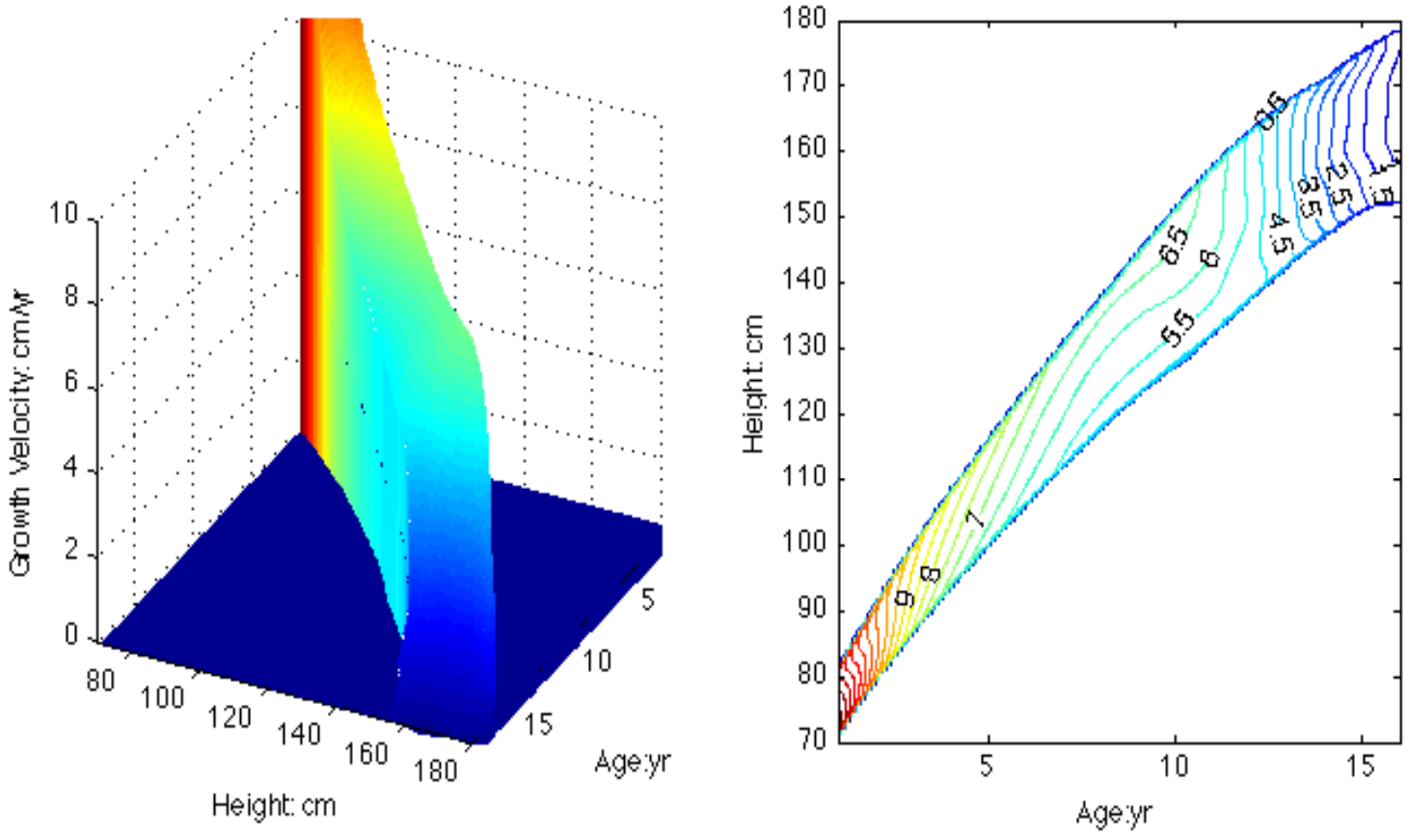
  where $s_j = (t_{ij} + t_{i,j+1})/2$ and $h_X > 0$ and $h_{X'} > 0$ are smoothing bandwidths.

- Step 2: Trajectory estimates $\widehat{X}(t)$ and $\widehat{X'}(t)$ from Step 1 are combined to obtain a Nadaraya–Watson kernel estimator for f,

$$\widehat{f}(t,x) = \frac{\sum_{i=1}^{n} K\{\frac{\widehat{X}_i(t)-x}{b_X}\}\widehat{X'}_i(t)}{\sum_{i=1}^{n} K\{\frac{\widehat{X}_i(t)-x}{b_X}\}}.$$

  utilizing bandwidths $b_X > 0$.

- Under regularity conditions, this gives consistent estimators.



Left panel: Estimated surface $\hat{f}(t, x)$ on a curved domain, characterizing the deterministic part of the nonlinear dynamic model. Right panel: Contour plot of the surface $\hat{f}(t, x)$.

- Linear concurrent model. Relating two stochastic processes $X(t)$ and $U(t)$ at each time $t \in \mathcal{T}$, the linear concurrent model captures a linear relationship between X and U through a deterministic function $\beta(t)$,

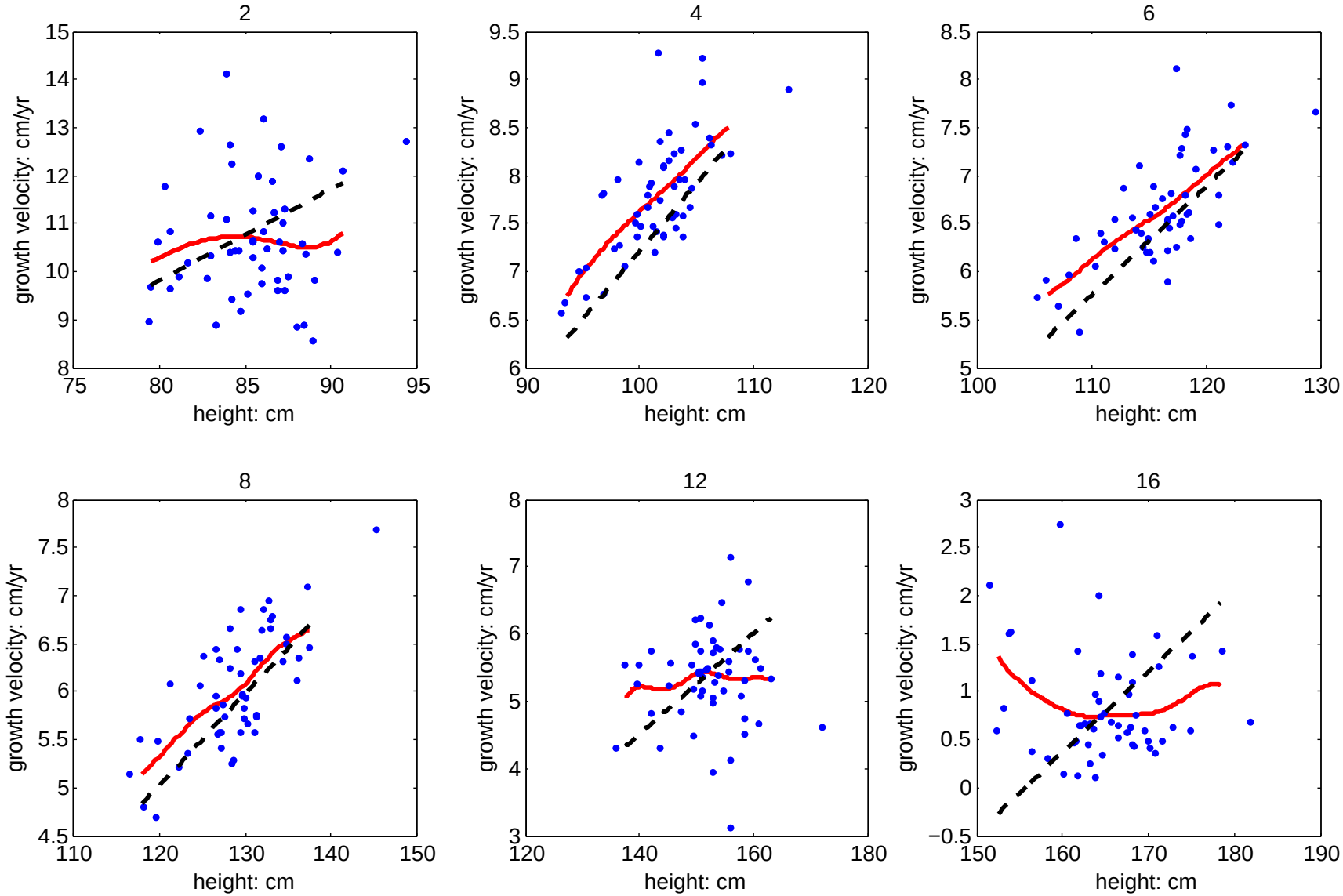
$$U(t) = \mu_U(t) + \beta(t)\{X(t) - \mu_X(t)\} + Z_2(t),$$

where $Z_2(t)$ is a zero mean drift process with $\mathrm{cov}\{Z_2(t), X(t)\} = 0$.

- Nonlinear concurrent model. Proposed methodology covers the case where the link between $U(t)$ and $X(t)$ is nonlinear,

$$U(t) = f\{t, X(t)\} + Z(t) \ ,$$

with $E\{Z(t) \mid X(t)\} = 0$ almost surely and $f\{t, X(t)\} = E\{U(t) \mid X(t)\}$. Can establish consistency and rates of convergence for two-step estimators.



Each of the panels, arranged for ages $t = 2, 4, 6, 8, 12$, from left to right and top to bottom, respectively, illustrates estimates $\hat{f}(t, \cdot)$ of the deterministic part of the nonlinear dynamic model (solid), the linear estimates (dashed) and the scatterplot of observed data pairs $(x(t), x^{(1)}(t))$.

Lectures on FDA – Part VI

Correlation, Connectivity and Densities

BOLD (Blood Oxygen Level Dependent) Brain Signals

- Resting State fMRI: Subjects are told to relax and let their mind flow freely while in the scanner

- BOLD Signals:
  - Time courses that are measured at 240 time points (spaced 2 seconds apart) at each voxel
  - One recording every two seconds
  - Recording is more or less simultaneous over all voxels
  - Signal reflects oxygen metabolism
  - Preprocessing needed to improve alignment in time and space and to eliminate interference of physiological signals such as breathing and heartbeat

- Brain can be organized into "hubs" of voxels that are situated around "seed voxels" and are relatively highly connected

- Data recorded at the UC Davis Neuroimaging Center (Owen Carmichael) – we focus on 20 hubs that were identified by Buckner et al (2009)

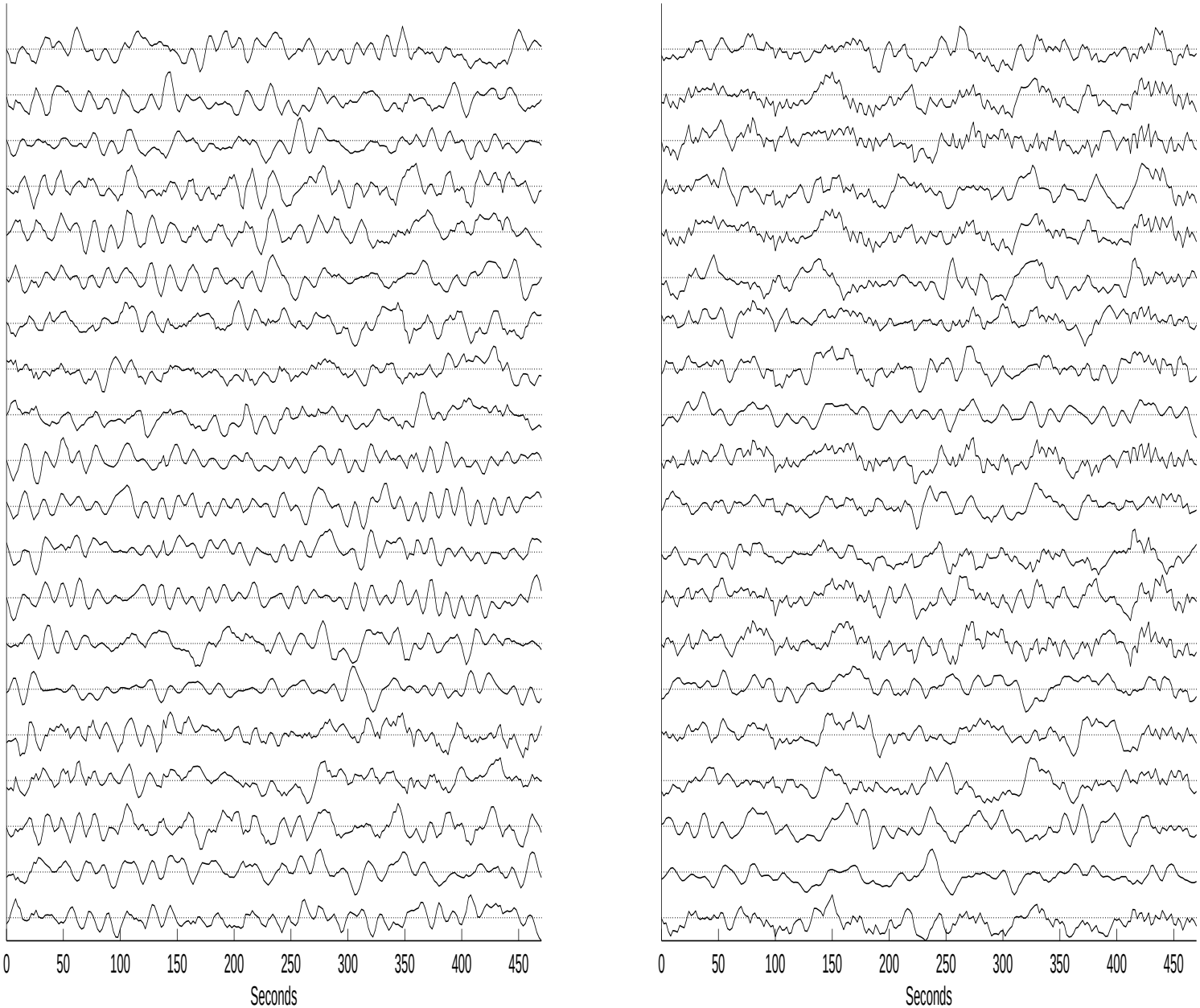


Figure: BOLD signals for the seed voxels of 20 hubs identified in Buckner et al (2009), normal (left) and demented (right) subject

Quantifying Brain Connectivity

- How to quantify the similarity of BOLD signals across voxels
- Subject-specific "Pearson correlation" between voxels j and l:

$$\rho_{jl} = \frac{\sum_{k=1}^{K} \left(S_{jk} - K^{-1} \sum_{k=1}^{K} S_{jk} \right) \left(S_{lk} - K^{-1} \sum_{k=1}^{K} S_{lk} \right)}{\sqrt{\sum_{k=1}^{K} (S_{jk} - K^{-1} \sum_{k=1}^{K} S_{jk})^2 \sum_{k=1}^{K} (S_{lk} - K^{-1} \sum_{k=1}^{K} S_{lk})^2}},$$

  where S_{jk} is the signal for the j-th voxel at the k-th timepoint.
- Corresponding correlation matrix is $C = \{\rho_{jl}\}_{j,l=1,\ldots,L}$
- The "Pearson correlation" is a special case of a functional correlation measure

Functional Correlation Measures

- Many such measures have been proposed over the years
- First proposed measure: Functional Canonical Correlation (Leurgans et al 1993)
- For pairs of functions (X, Y), with L^2 inner product $\langle X, Y \rangle = \int X(s)Y(s)ds$, define $\rho_{\mathsf{FCC}} = \sup_{\|u\|=\|v\|=1} \operatorname{corr}(\langle u, X \rangle, \langle v, Y \rangle)$
- Requires to solve an inverse problem, which is bad news for functional data (as inverses of compact operators are unbounded): This makes functional canonical correlation a very difficult exercise in regularization.

- Functional Singular Correlation. By changing the target criterion to $(u_0, v_0) = \text{argsup}_{\|u\|=\|v\|=1} \text{cov}(\langle u, X\rangle, \langle v, Y\rangle)$, one can avoid the inverse problem.

- The functional singular correlation (Yang et al 2010) $\rho_{\text{FSC}} = \text{cov}(\langle u_0, X\rangle, \langle v_0, Y\rangle)/\sqrt{\text{var}(\langle u_0, X\rangle)\text{var}(\langle u_0, Y\rangle)}$ is obtained from the singular components of (X, Y)

- Dynamical Correlation. Define correlation as a cosine between standardized curves (Dubin and Müller, 2005)

$$X^*(t) = \frac{X(t) - \langle X, 1\rangle}{(\int (X(t) - \langle X, 1\rangle)^2 dt)^{1/2}}, \; Y^*(t) = \frac{Y(t) - \langle Y, 1\rangle}{(\int (Y(t) - \langle Y, 1\rangle)^2 dt)^{1/2}}$$

leads to the dynamical correlation $\rho = E\langle X_k^*, X_l^*\rangle$.

- The above version of dynamical correlation is the target of the fMRI Pearson correlation measure in the limit for smooth signals as the measurement times get denser.

Connectivity Density Analysis

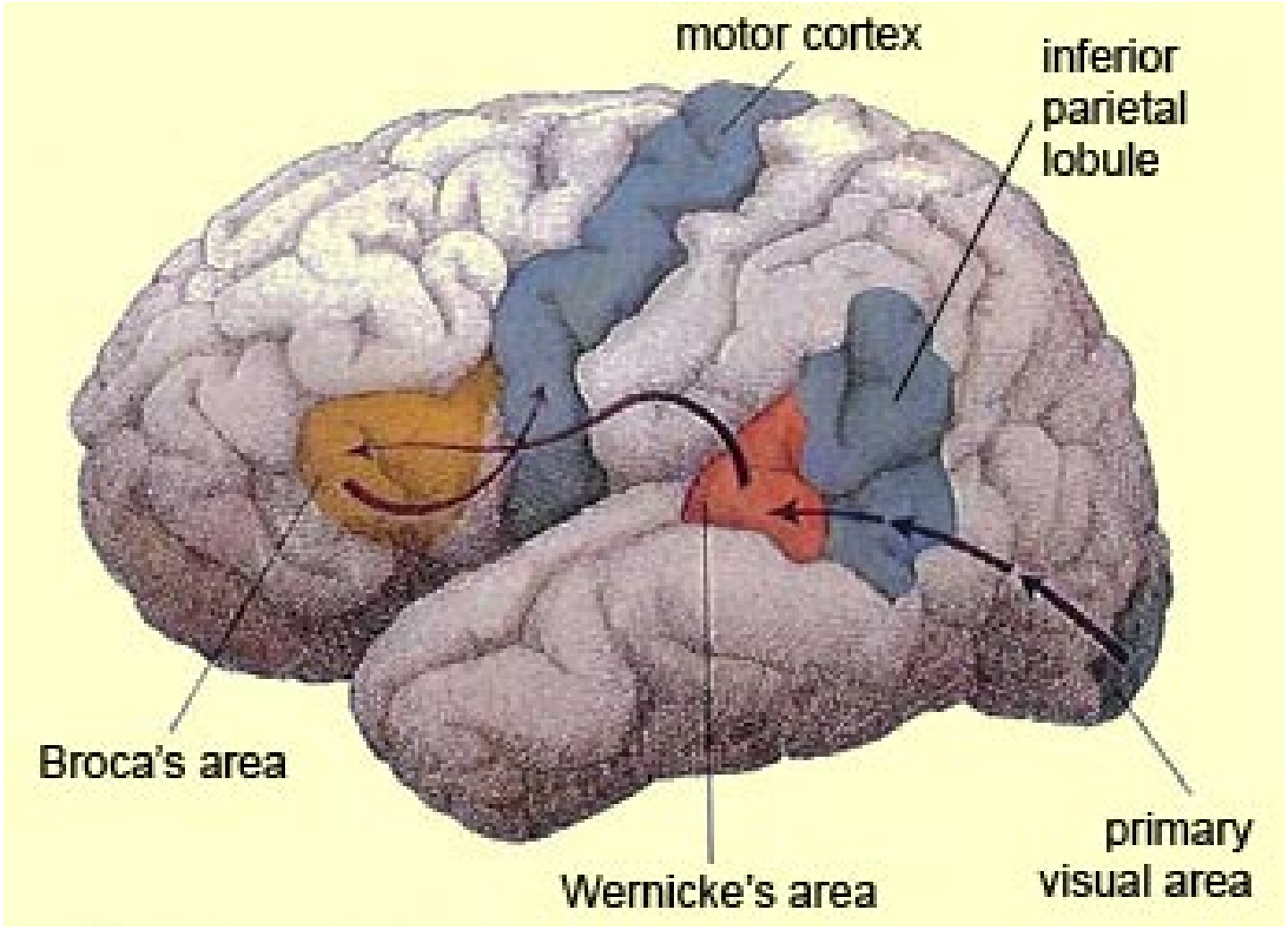
- Adopting dynamic/“Pearson" correlation, we aim at intra-hub analysis where the correlations between the voxels in a $11 \times 11 \times 11$ cube around a centrally located seed voxel and the seed voxel are considered

- Aim to describe intra-hub connectivity through the density of the correlations observed within the hub. The intra-hub connectivity for each subject is thus summarized by a density function

- Goal: Modeling density functions as functional data

Data Analysis

Data from $n = 68$ Alzheimer's patients recorded at the UC Davis Alzheimer Center for the hub that was identified within the parietal lobule by Buckner et al (2009).



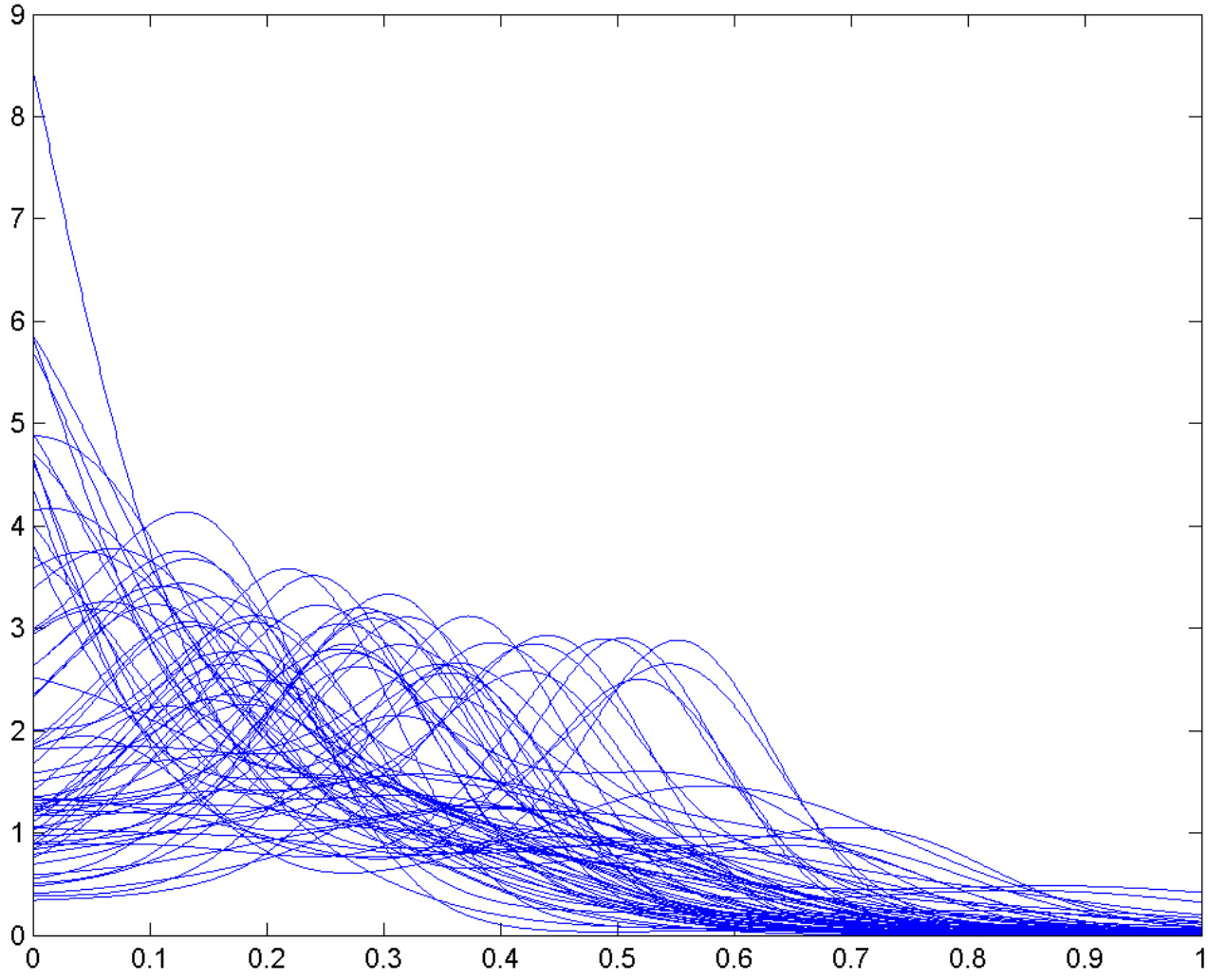


Figure: Kernel density estimates for the time course correlations between voxels in the parietal lobule hub for $n = 68$ patients

Densities as Functional Data

- Random densities f (one density observed for each subject) are constrained functional data since $\int f = 1,\ f \geq 0$.
- Therefore, density functions do not live in a linear space and linear methods such as Functional Principal Component Analysis (FPCA, Kneip & Utikal 2001) are suboptimal.
- How to address this problem?
  Transform densities f so that $\psi(f)$ is in a linear functional space, then perform FPCA there and transform back.

The Transformation Approach

- Data are $f \sim_{i.i.d.} \mathfrak{F}$ for a density valued process $\mathfrak{F}$
- Observe n (random) density functions $f_1, \ldots, f_n$ defined on a common interval $[0, 1]$. Denote the space of continuous and strictly positive densities on $[0, 1]$ by $\mathcal{G}$.
- Find a suitable continuous and invertible transformation $\psi : \mathcal{G} \to L^2(\mathcal{T})$

Sample density estimates

- For a random density f in the sample: Usually do not observe f, only i.i.d. sample $W_1, \ldots, W_N$ drawn from f. Then f needs to be estimated from this sample of random size N.

- Consistency of the transformation approach via construction of modified kernel estimator that produces uniformly consistent bona fide density estimates, satisfying $sup_{f \in \mathcal{G}} E[d_2(\hat{f}, f)] = O(h)$, where h is the smoothing bandwidth and d_2 the L^2 distance.

Examples for Transformations $\mathcal{G} \leftrightarrow L^2$

- Log hazard transformation

$$\psi_H(f)(t) = \log(h(t)) = \log\left\{\frac{f(t)}{1-F(t)}\right\}, \quad t \in [0, 1-\delta].$$

Special care needs to be taken for the inverse on $x \in (1-\delta, 1]$.

- Log quantile transformation

$$\psi_Q(f)(t) = \log(q(t)) = -\log\{f(Q(t))\}, \quad \text{where} \quad q = Q' = F^{-1'}.$$

Inverse from $X(t) = \psi_Q(f)(t)$,

$$F^{-1}(t) = \left\{\int_0^1 e^{X(s)}\, ds\right\}^{-1} \int_0^t e^{X(s)}\, ds$$

Transformation Modes of Variation

- Construct modes of variation in the transformed space for processes $X = \psi(f)$ and map back to density space, i.e.,

$$\psi^{-1}\left(\nu + \alpha\sqrt{\lambda_k}\phi_k\right), \quad \alpha \in \mathcal{R},$$

  where (λ_k, ϕ_k) are k-th eigenvalue/eigenfunction of X.

- This leads to the transformation modes of variation

$$g_k(x, \alpha, \psi) = \psi^{-1}\left(\nu + \alpha\sqrt{\lambda_k}\phi_k\right)(x).$$

- The resulting representations of the original densities in the sample are

$$f_i(x, K, \psi) = \psi^{-1}\left(\nu + \sum_{k=1}^{K} A_{ik}\phi_k\right)(x),$$

  substituting suitable estimates.

- Choosing the truncation point K by fraction of variance explained.

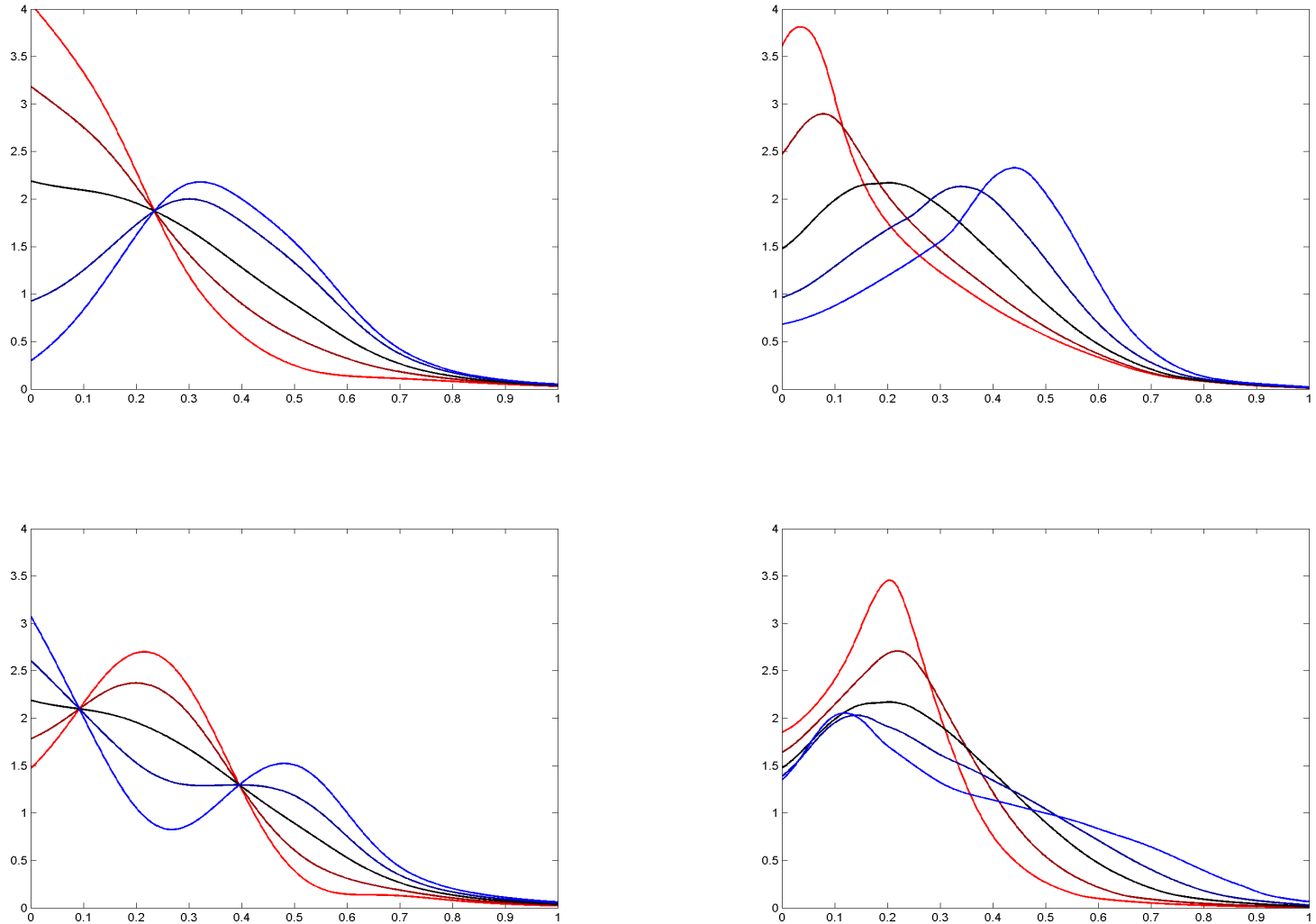


Figure: First (above) and second (below) modes of variation ($\alpha_1 = .1, .25, 0.25, .75, 0.9$) for the distributions of seed voxel correlations for $n = 68$ patients from the Davis study. FPCA (left) and Log Quantile Density Transformations (right). Black line is the mean.

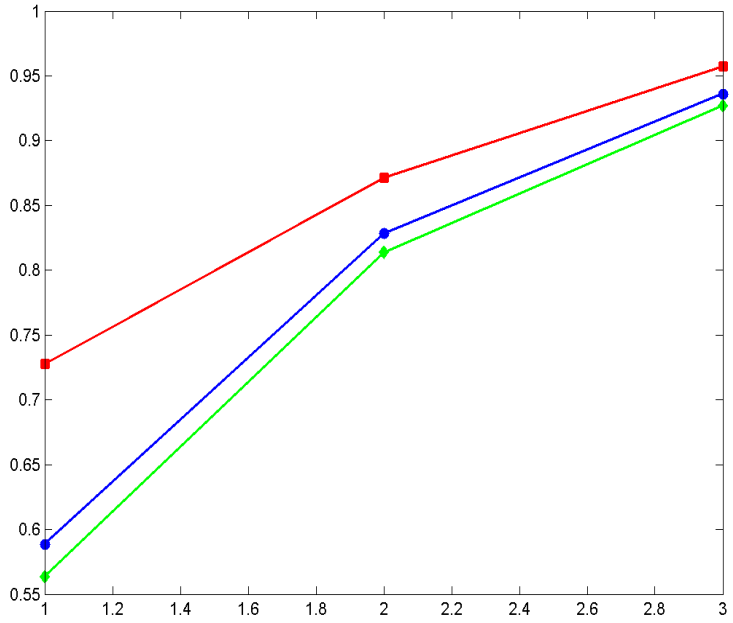


Figure: Fraction of variance explained for $K = 1, 2, 3$ components, for Log Quantile Density transformation (red) and FPCA (blue)

Number of Components K	1	2	3	4
FPCA	0.180	0.185	0.193	0.201
LQD	0.180	0.176	0.169	0.173

Table: Estimated mean squared cross-validation prediction errors for functional linear model, predicting Executive Cognitive Test Score

Lectures on FDA – Part VII

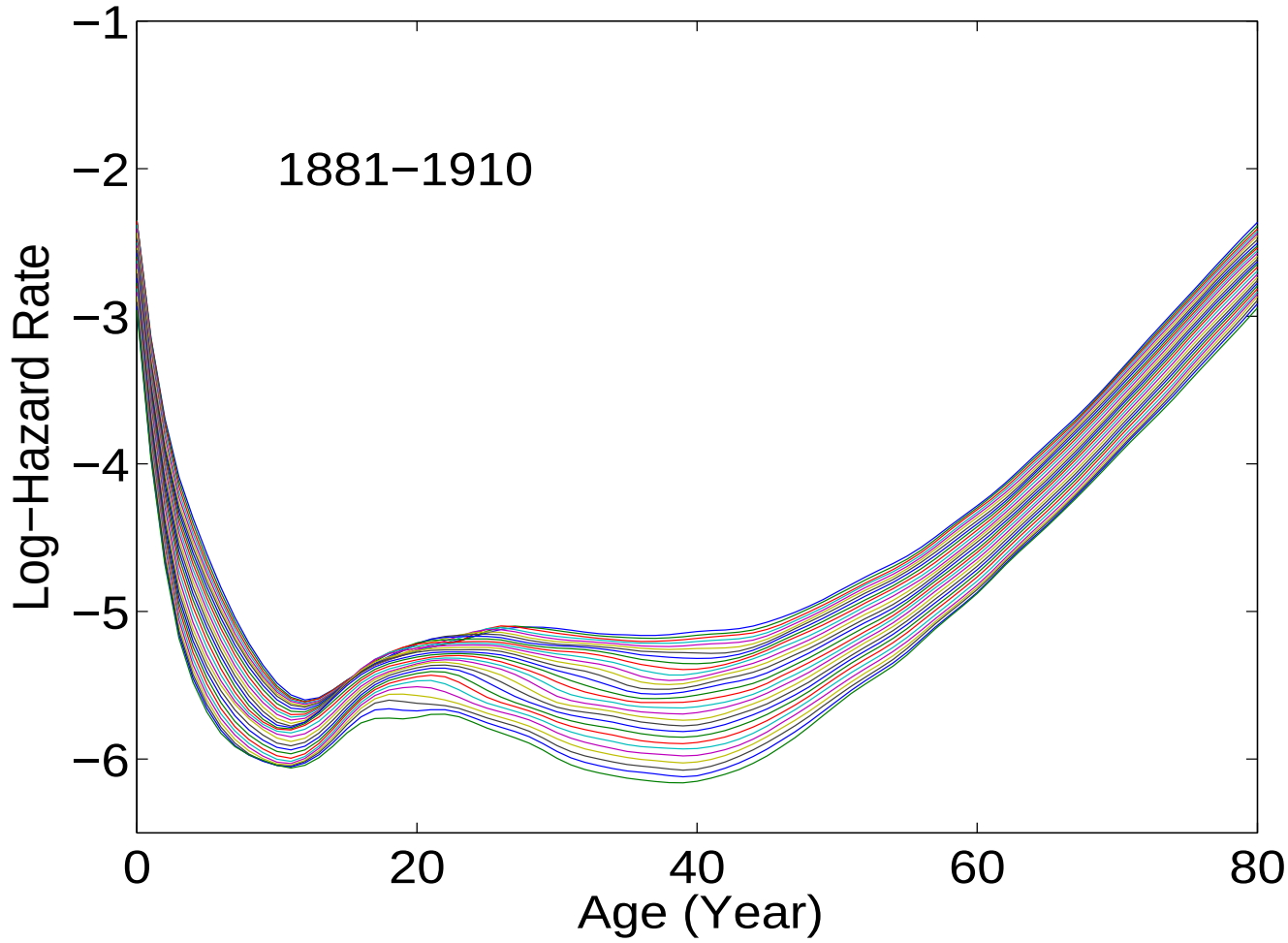
Function-valued stochastic processes

Chen & Müller JASA 2014

Chen, Delicado & Müller JASA in press

Dependent Functional Data

- Spatial Dependency (spatially indexed functions considered eg. in Nerini et al 2010, Delicado et al 2010, Gromenko et al 2013)
- Temporal Dependency: Linear models for random functions observed repeatedly over time (eg Greven et al 2010; Zipunnikov et al 2012; Horvath and Kokoszka (book) 2012)
- Hierarchical linear functional models, functional ANOVA (eg Morris et al 2006; Crainiceanu et al 2009)
- Examples for samples of repeatedly observed functions: Mortality profiles, repeatedly observed over calendar years for many countries; Daily movement profiles for subject tracking.



Changes in mortality curves (quantified as log-hazard functions derived from lifetables) for Swedish women born 1881-1910 (Chiou & Müller 2010): Repeatedly observed functional trajectories per country, for a sample of countries

Function-Valued Stochastic Processes

- Traditional:

$$\mathcal{T} \times \Omega \to X(t, \omega) \in \mathcal{R}, \quad X \in L^2, \quad X \text{ is smooth}$$

- Now the value of the process at each $t \in \mathcal{T}$ is a random function $X(t, \cdot)$ on a domain $\mathcal{S}$:

$$\mathcal{T} \times \Omega \to X(t, \cdot, \omega) \in L^2(\mathcal{T} \times \mathcal{S}),$$

with

$$E(X(t, s)) = \mu(t, s), \quad C(t_1, s_1, t_2, s_2) = \mathrm{cov}(X(t_1, s_1), X(t_2, s_2)).$$

- Data can be viewed as functional data in two arguments, the time index t of the stochastic process and the argument s of the observed functions. These repeatedly observed functions are dependent within subject.

Karhunen-Loève (KL) Representation for Function-Valued Stochastic Processes

$$X(s,t) = \mu(s,t) + \sum_{r=1}^{\infty} Z_r \gamma_r(s,t), \quad s \in \mathcal{S},\ t \in \mathcal{T}.$$

Here $\{\gamma_r : r \geq 1\}$ are the orthonormal eigenfunctions of the linear operator with kernel C, forming a basis on $L^2(\mathcal{S} \times \mathcal{T})$, $\{Z_r = \int \gamma_r(s,t) X^c(s,t) dsdt : r \geq 1\}$ are the (uncorrelated) functional principal components (FPCs) with $\mathrm{E}(Z_r) = 0$.

- First K components explain as least as much variance as any other $K-$dimensional representation, for all K
- Arguments are symmetric and the effects of s and t hard to separate: For many applications separation is crucial, and the KL representation hard to interpret. In demographic applications, t is calendar time and s the age of a cohort.
- Covariance estimation for sparse designs requires 4-dimensional smoothing.

A Tensor Product Representation

- For an arbitrary orthonormal basis $\{\psi_j : j \geq 1\}$ of $L^2(\mathcal{S})$, one always has

$$X(s,t) = \mu(s,t) + \sum_{j=1}^{\infty} \xi_j(t)\psi_j(s)$$

with mean function μ and random coefficient functions $\{\xi_j : j \geq 1\}$

- Applying the KL representations of the random functions ξ_j,

$$\xi_j(t) = \sum_{k=1}^{\infty} \chi_{jk}\phi_{jk}(t),$$

with eigenfunctions ϕ_{jk} and FPCs χ_{jk}, leads to

$$X(s,t) = \mu(s,t) + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \chi_{jk}\phi_{jk}(t)\psi_j(s).$$

Is there a best basis ψ_j?

Define the the marginal covariance function,

$$G(s,u) = \int_{\mathcal{T}} C((s,t),(u,t))\, dt, \quad \text{with } s,u \in \mathcal{S},$$

i.e., averaging the autocovariance function over t.

Then the eigenfunctions/eigenvalues (ψ_j, τ_j) of the linear operator with symmetric non-negative kernel G are the best basis in the following sense: For any $K \geq 1$,

$$(\psi_1, \ldots, \psi_K) = \underset{(z_1,\ldots,z_K)}{\text{argmin}} \; \mathrm{E} \left(\int_{\mathcal{T}} \| X^c(\cdot,t) - \sum_{j=1}^{K} \langle X^c(\cdot,t), z_j \rangle_{\mathcal{S}} z_j \|_{\mathcal{S}}^2 dt \right),$$

i.e., the basis ψ_j explains most of the variance on average when averaging over t (Chen et al 2014, preprint).

Marginal KL Representation and Marginal FPCA

- Using the optimal basis ψ_j leads to the representations

$$\begin{aligned} X(s,t) &= \mu(s,t) + \sum_{j=1}^{\infty} \xi_j(t)\psi_j(s) \\ &= \mu(s,t) + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \chi_{jk}\phi_{jk}(t)\psi_j(s). \end{aligned}$$

- This marginal KL representation differs from the standard KL representation
- The roles of s and t are distinguished
- ϕ_{jm}, ϕ_{kp} not necessarily orthogonal for $j \neq k$, and χ_{jm}, χ_{kp} not necessarily uncorrelated
- Contains tracking functions $\xi_j(t)$, $j \geq 1$, that describe the change in shapes of the repeatedly observed functions as time t is increasing.

Estimation

- Pool the data of all subjects to obtain an estimate $\hat{\mu}(s,t)$ of the mean function $\mu(s,t)$ and use this to obtain centered data (cross-sectional averaging and interpolation or smoothing)

- Center the pooled data and obtain estimates $\hat{G}(s_1,s_2)$ of the marginal covariance function $G(s_1,s_2)$, by (a) cross-sectional averaging and interpolation, omitting the data for the diagonal $s_1 = s_2$ that are contaminated; (b) then averaging over the observation grid in direction t.

- Obtain the eigenfunctions $\hat{\psi}_j$ and eigenvalues $\hat{\tau}_j$ associated with $\hat{G}$ by standard methods and the FPC function estimates
$$\hat{\xi}_{i,j}(t) = \int \hat{X}_i^c(s,t)\hat{\psi}_j(s)ds.$$

- For each fixed j, obtain estimates for the FPCs χ_{ijk} and eigenfunctions $\{\phi_{jk}(t) : k \geq 1\}$ for the FPC function estimates $\{\xi_{i,j}(t), j \geq 1\}$.

- Overall representation

$$\hat{X}_i(s,t) = \hat{\mu}(s,t) + \sum_{j=1}^{P} \hat{\xi}_{i,j}(t)\hat{\psi}_j(s)$$

$$= \hat{\mu}(s,t) + \sum_{j=1}^{P}\sum_{k=1}^{K_j} \hat{\chi}_{i,jk}\hat{\phi}_{jk}(t)\hat{\psi}_j(s).$$

- The included number of components P is selected by the fraction of variance explained (FVE) criterion, finding the smallest P such that $\sum_{j=1}^{P} \hat{\tau}_j / \sum_{j=1}^{M} \hat{\tau}_j > 1 - p$, where M is large and we choose $p = 0.15$.

- The number of included components K_j is determined by a second application of FVE, where the variance explained by each term (j,k) is

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\chi}_{i,jk}^2 / \frac{1}{n}\sum_{i=1}^{n} ||X(s,t) - \hat{\mu}(s,t)||_{\mathcal{S}\times\mathcal{T}}^2.$$

Consistency

Under regularity conditions, perturbation results such as those of Bosq (2000) imply for $1 \leq j \leq P$:

$$\|\hat{G}(s,u) - G(s,u)\| = O_p((1/n)^{1/2})$$

$$|\hat{\tau}_j - \tau_j| = O_p((1/n)^{1/2})$$

$$\|\hat{\psi}_j(s) - \psi_j(s)\| = O_p((1/n)^{1/2})$$

$$\sup_{1\leq m\leq M} |\hat{\xi}_{i,j}(t_{im}) - \xi_{i,j}(t_{im})| = O_p((\log n/n)^{1/2})$$

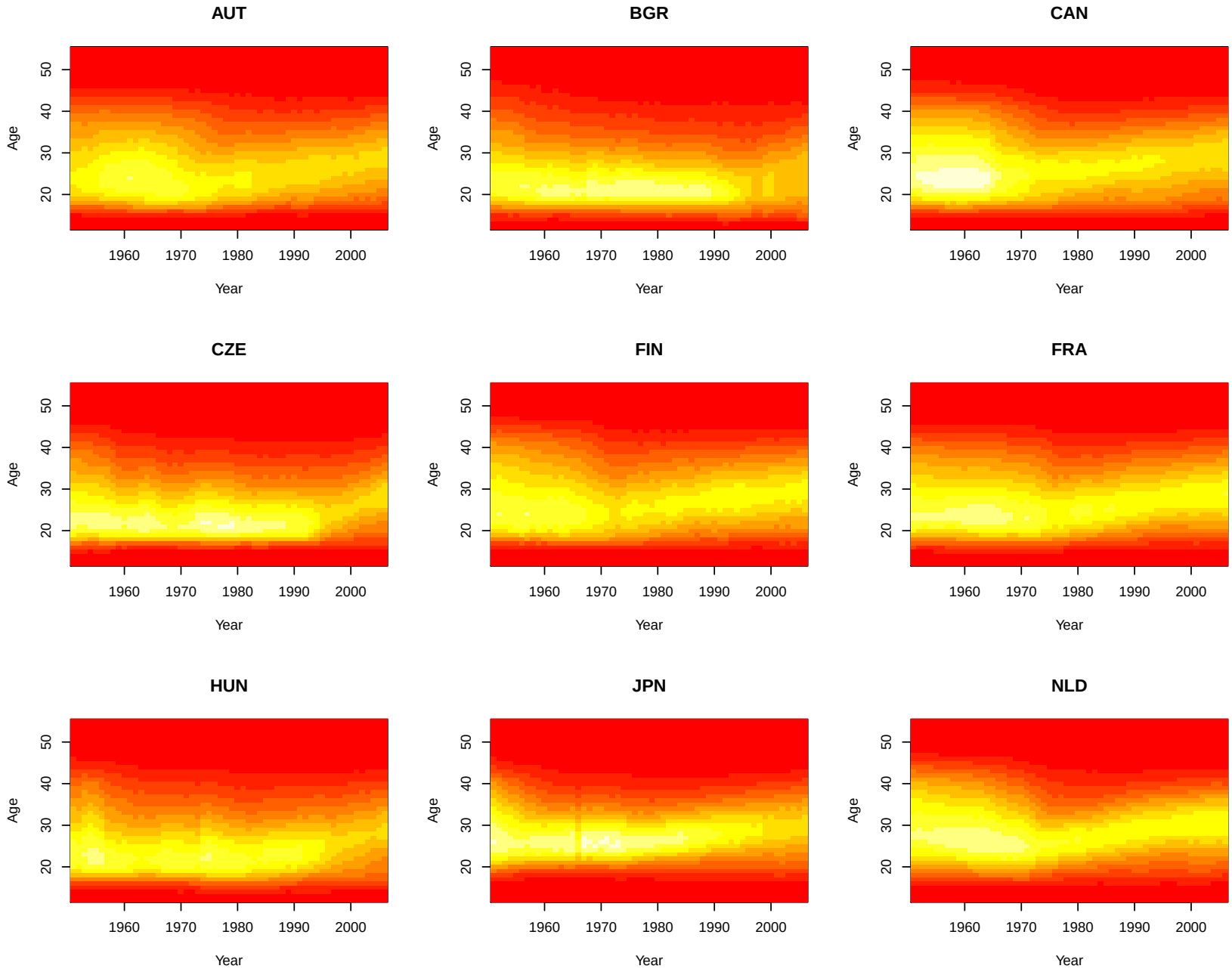
Marginal FPCA for Fertility Data

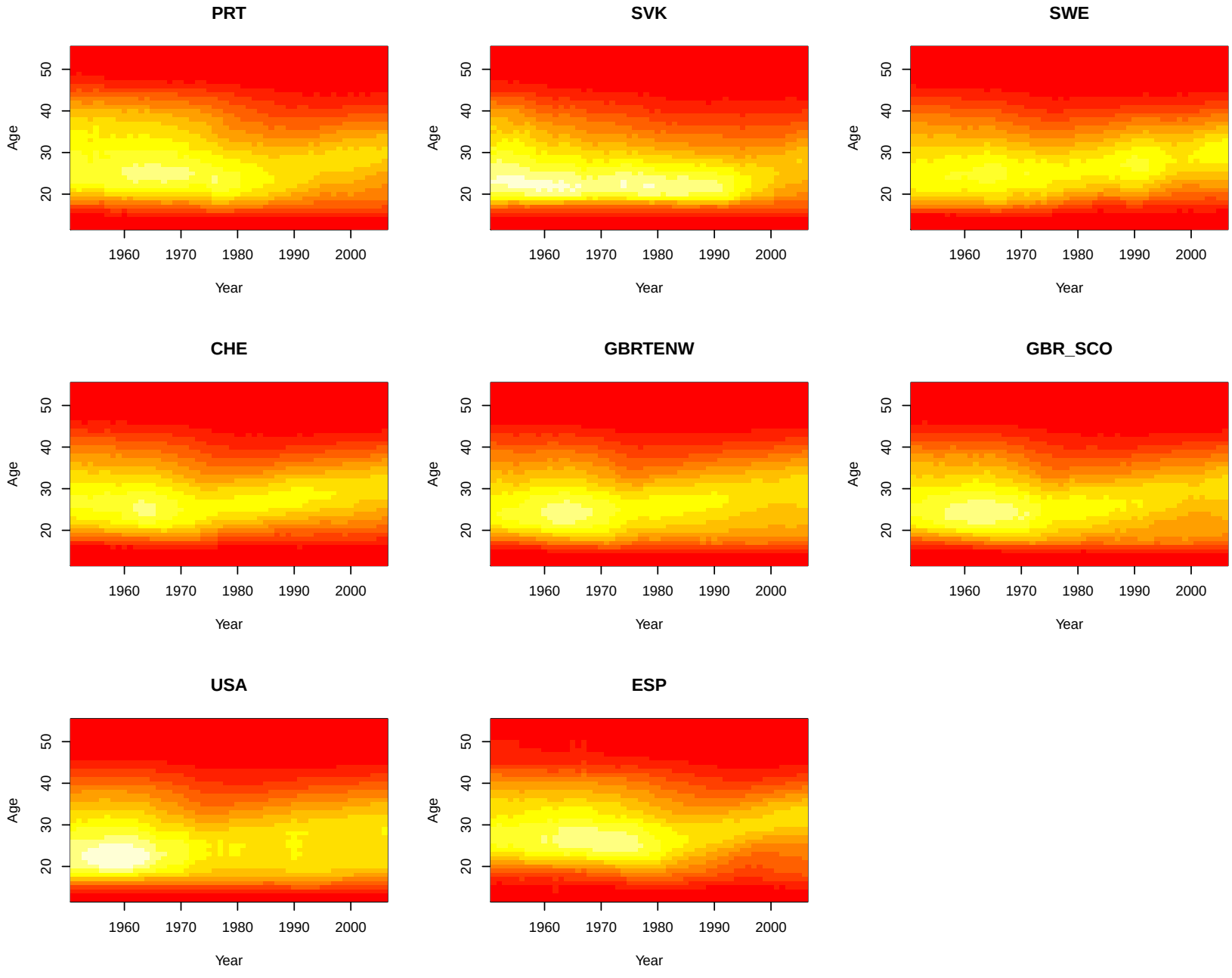
- Age-Specific Fertility Rate (AFSR) for 17 countries, 1951 to 2006 (Human Fertility Database 2013 (HFD-2013))
- AFSR for age s (expressed in years) and calendar year t:

$$\text{ASFR}(s,t) = \frac{\text{Births during the year } t \text{ to women of age } s}{\text{Person-years lived for the year } t \text{ by women of age } s}.$$

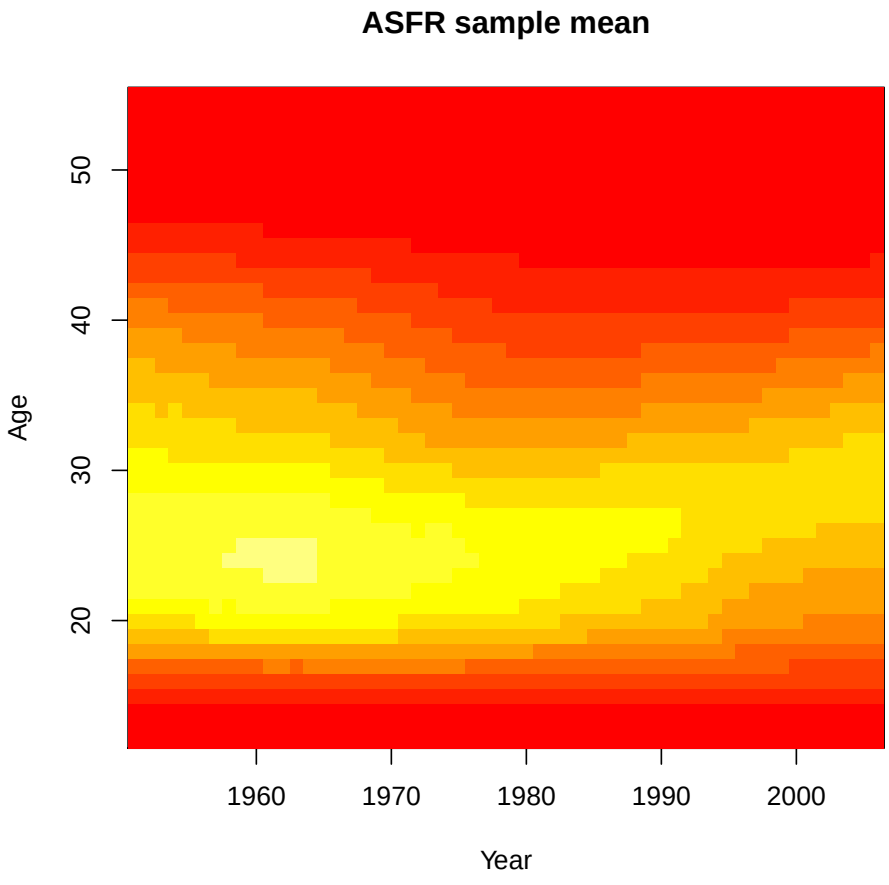
- Ages of mothers s range from 12 to 55 years old.
- Data: 17 independent units (countries), corresponding to a realization of the function valued stochastic process $\text{ASFR}(\cdot, t)$ at each year t. Observation grid (age, calendar-year) has 44×56 equidistant points.

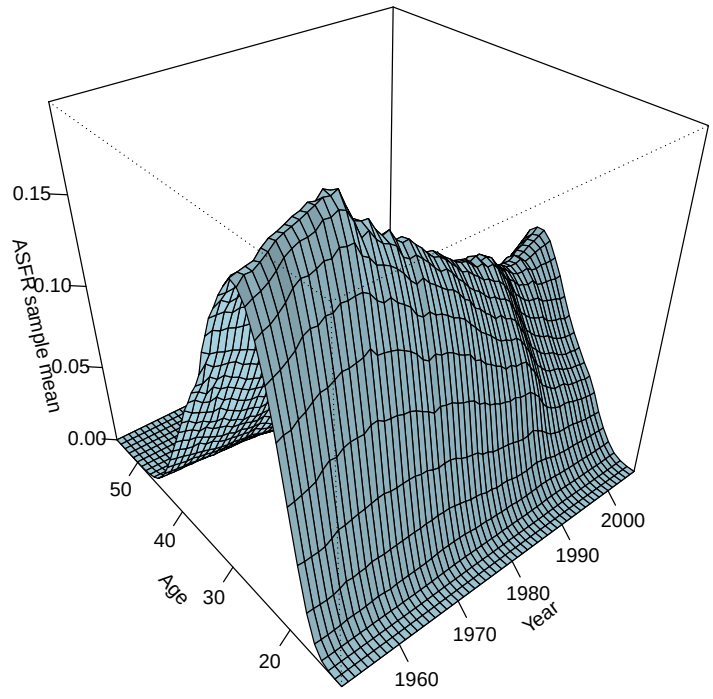
Abbreviation	Country name
AUT	Austria
BGR	Bulgaria
CAN	Canada
CHE	Switzerland
CZE	Czech Republic
FIN	Finland
FRA	France
GBR_SCO	U.K., Scotland
GBRTENW	U.K., England and Wales
HUN	Hungary
JPN	Japan
NLD	Netherlands
PRT	Portugal
SVK	Slovakia
SWE	Sweden
USA	U.S.A.
ESP	Spain


AUT
BGR
CAN
CZE
FIN
FRA
HUN
JPN
NLD
Age
Year
20
30
40
50
1960
1970
1980
1990
2000


PRT
SVK
SWE
CHE
GBRTENW
GBR_SCO
USA
ESP
Age
Year
20
30
40
50
1960
1970
1980
1990
2000

Mean function $\mu(s, t)$

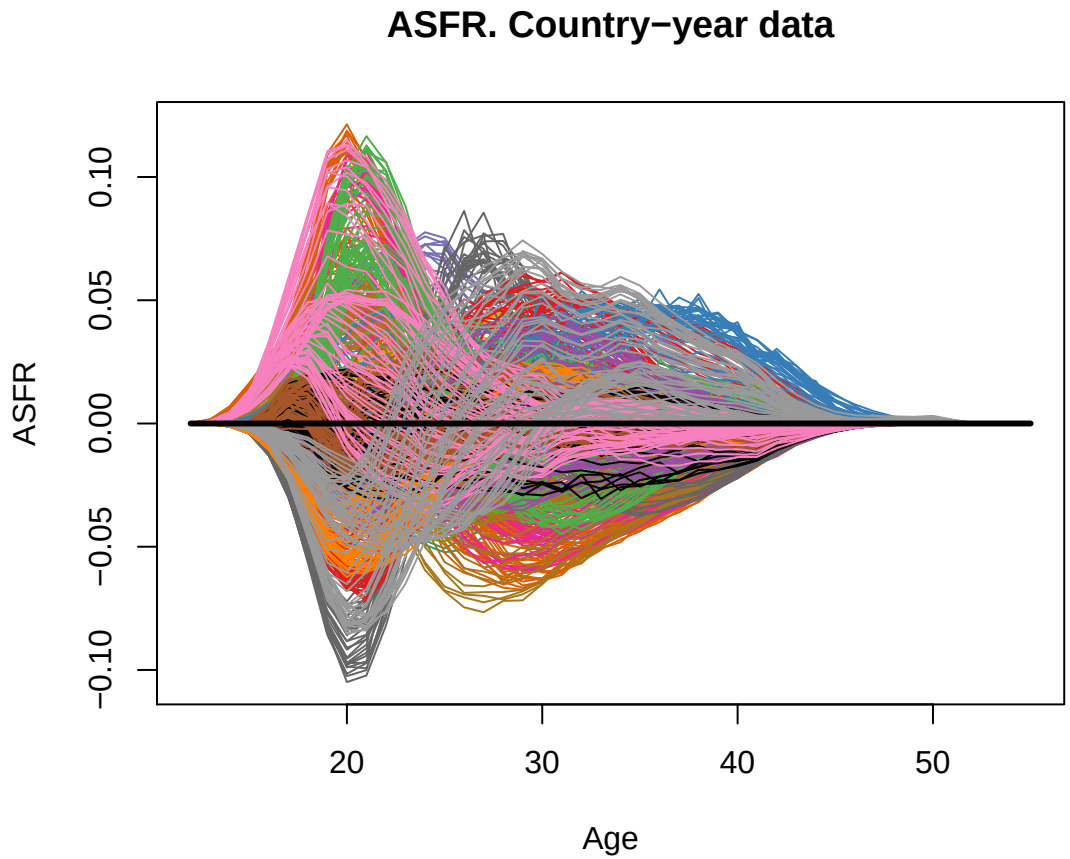




Empirical observations

- Fertility rates differ across countries (e.g., Sweden/Austria vs USA).
- Mother's ages at the fertility peak differ (Slovakia/Czech Republic vs Japan/Spain).
- The baby boom in the 1950's-60's was more expressed in USA and Canada than in other countries.
- The timing of fertility changes in calendar time differs across countries (Netherlands/Finland vs Spain/Portugal).

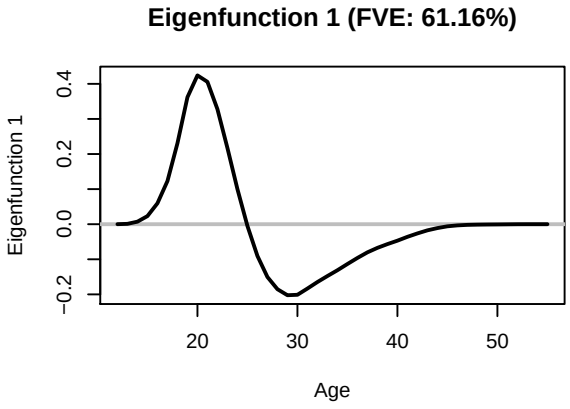
Marginal FPCA

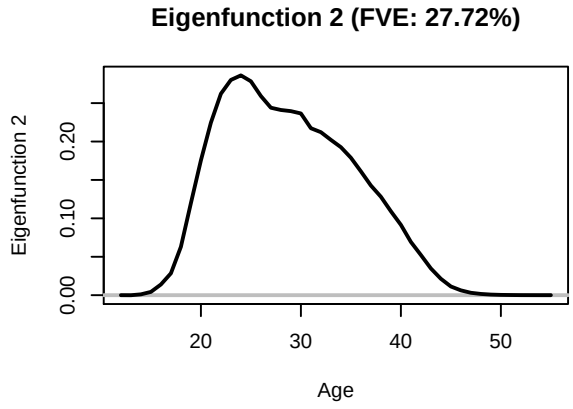


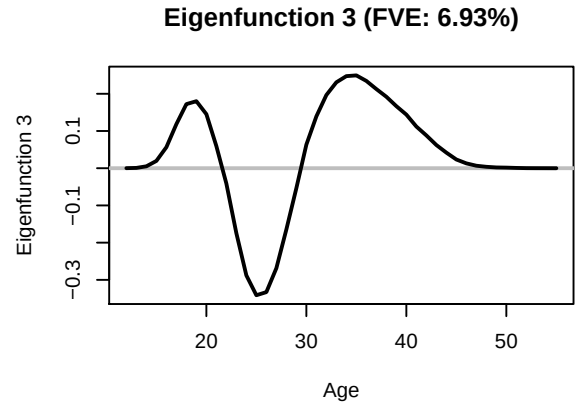
The $nM = 17 \times 56 = 952$ centered functional data

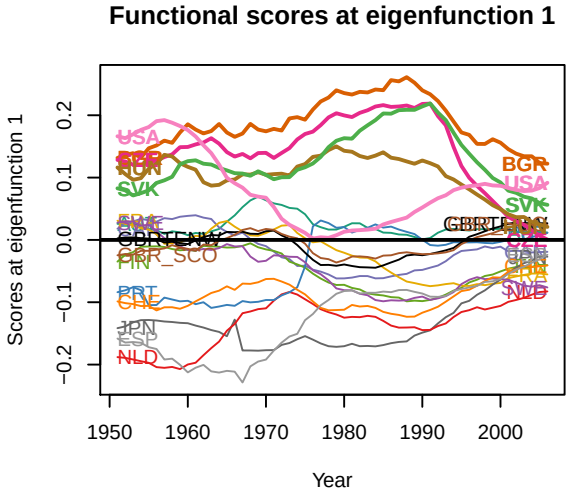
$$\text{ASFR}_i^c(s, t_m) = \text{ASFR}_i(s, t_m) - \overline{\text{ASFR}}(s, t_m)$$

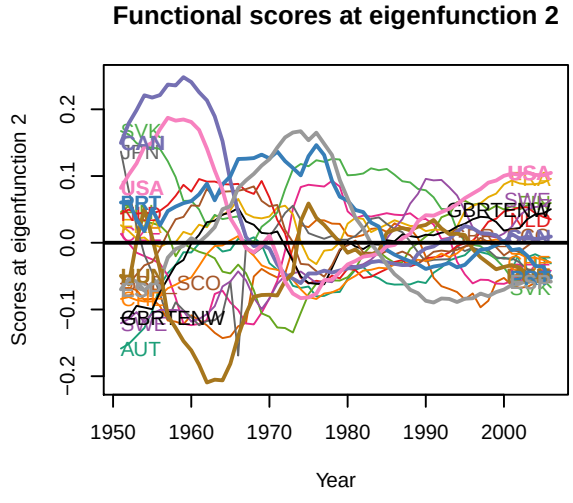
First three eigenfunctions $\hat{\psi}_j(s)$, $j = 1, 2, 3$, for fertility$= f($age$)$, with FVE of 95.8%.

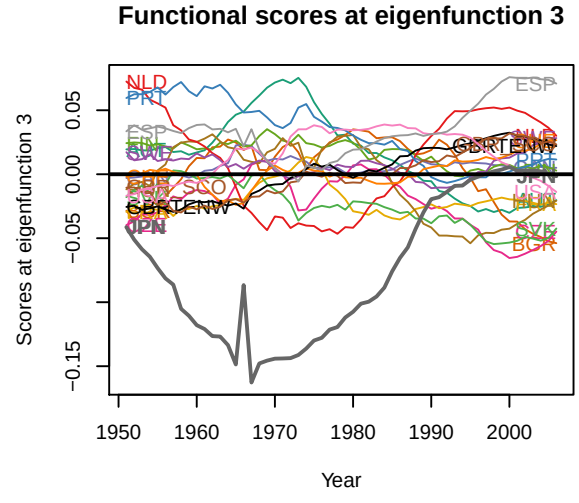








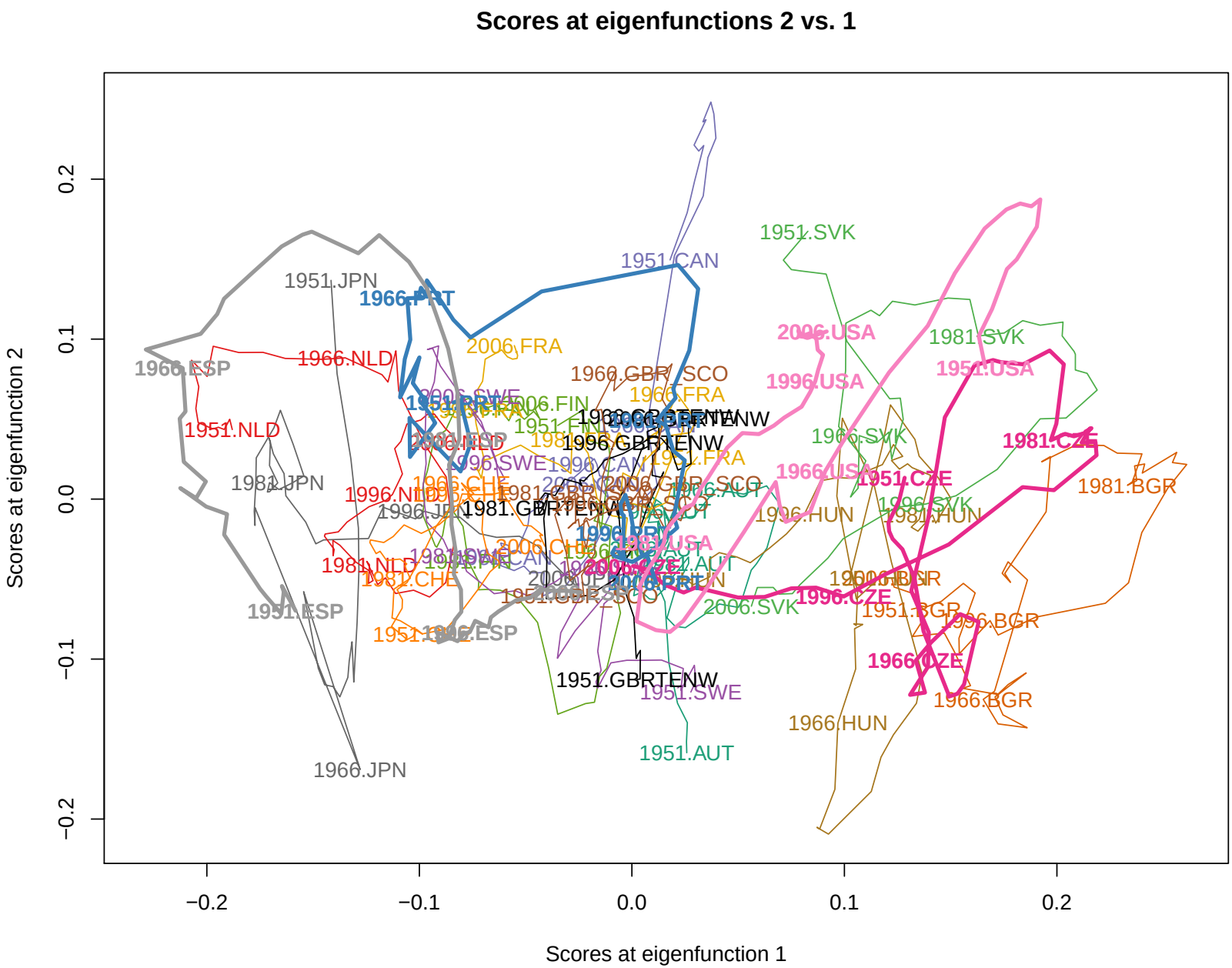


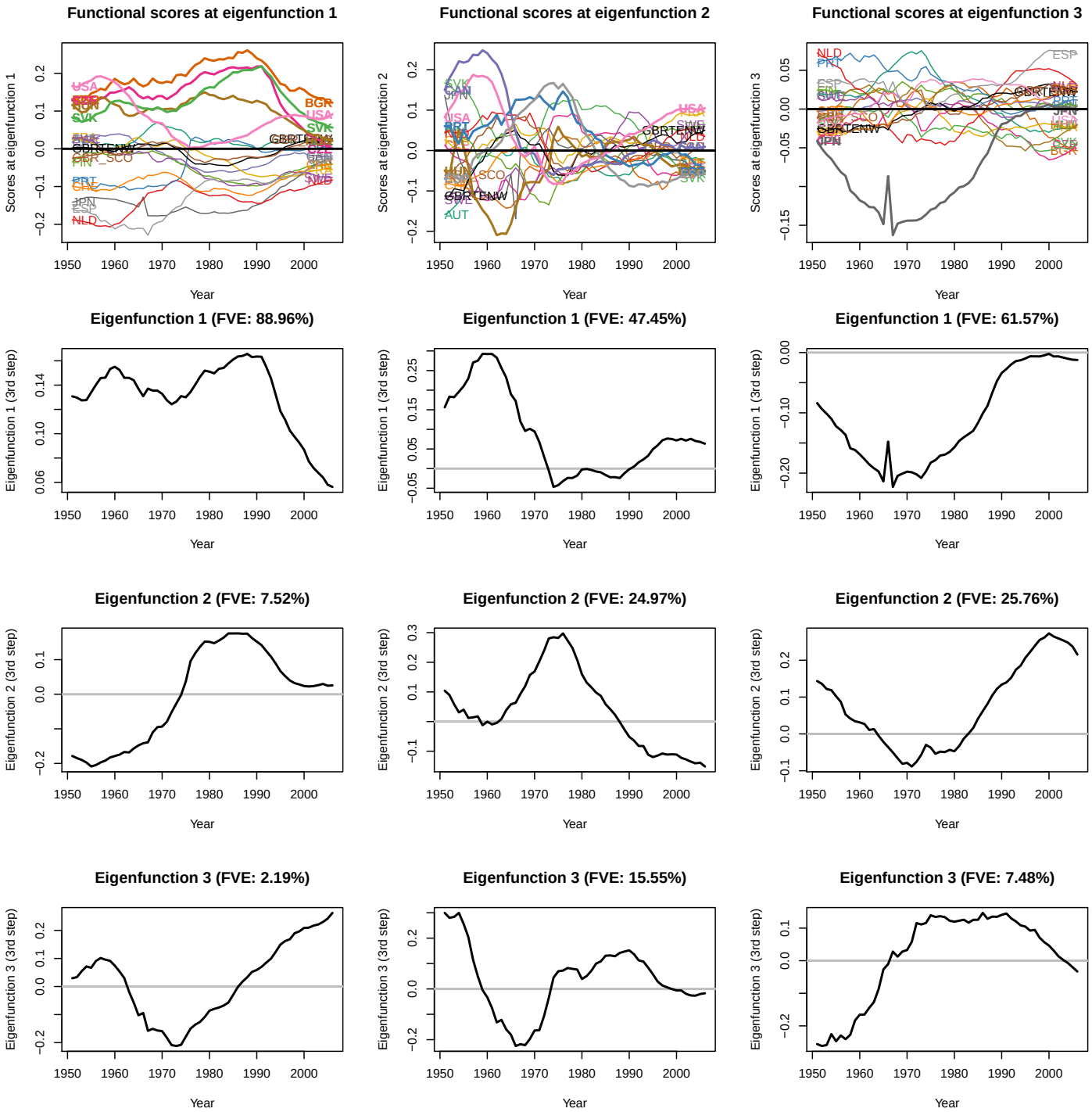


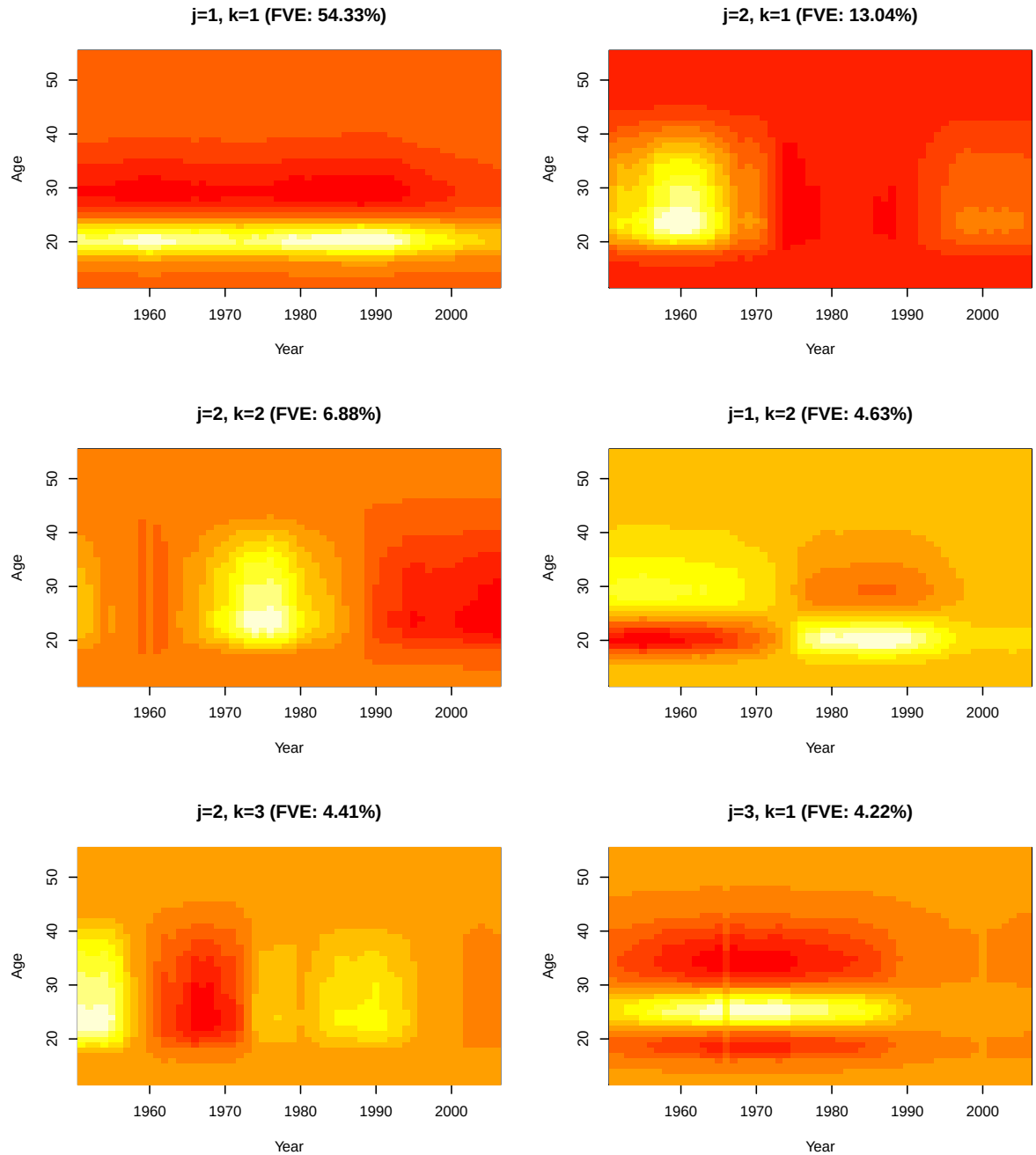
Japan, 1966

- In the third FPC function, Japan (grey) shows an anomalous behavior at year 1966.
- This is reflected in eigenfunction 1 of the third FPC function.
- We checked for the reason of this anomaly:
  In 1966 the total fertility in Japan suddenly declined to the lowest value ever recorded – 1966 was the year of the *Hinoe-Uma* (Fire Horse, a calendar event that occurs every 60 years) – there is a belief that girls born during Hinoe-Uma are unlucky.
- About a quarter of normal annual births were either lost in 1966 or were shifted to 1965 or 1967.

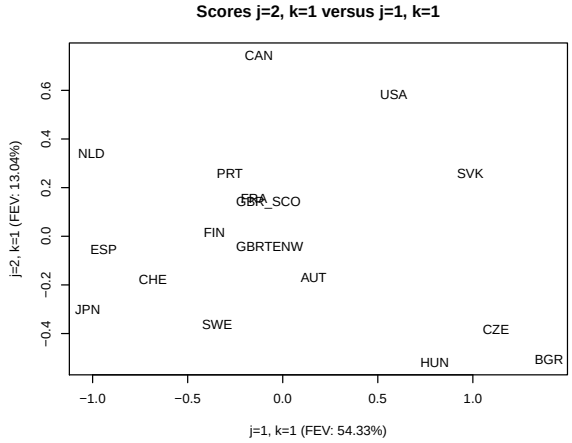
Trackplot $\{(\xi_1(t), \xi_2(t)),\ t \in \mathcal{T}\}$

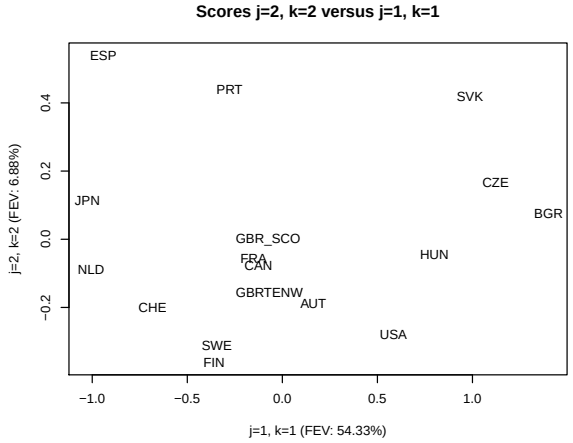


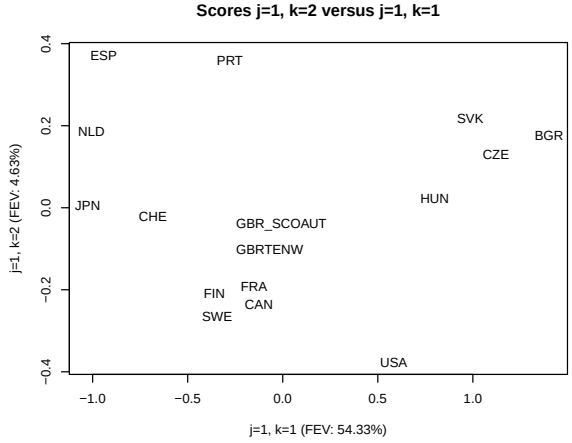

Functional scores at eigenfunction 1
Scores at eigenfunction 1
Year
Functional scores at eigenfunction 2
Scores at eigenfunction 2
Year
Functional scores at eigenfunction 3
Scores at eigenfunction 3
Year
Eigenfunction 1 (FVE: 88.96%)
Eigenfunction 1 (3rd step)
Year
Eigenfunction 1 (FVE: 47.45%)
Eigenfunction 1 (3rd step)
Year
Eigenfunction 1 (FVE: 61.57%)
Eigenfunction 1 (3rd step)
Year
Eigenfunction 2 (FVE: 7.52%)
Eigenfunction 2 (3rd step)
Year
Eigenfunction 2 (FVE: 24.97%)
Eigenfunction 2 (3rd step)
Year
Eigenfunction 2 (FVE: 25.76%)
Eigenfunction 2 (3rd step)
Year
Eigenfunction 3 (FVE: 2.19%)
Eigenfunction 3 (3rd step)
Year
Eigenfunction 3 (FVE: 15.55%)
Eigenfunction 3 (3rd step)
Year
Eigenfunction 3 (FVE: 7.48%)
Eigenfunction 3 (3rd step)
Year

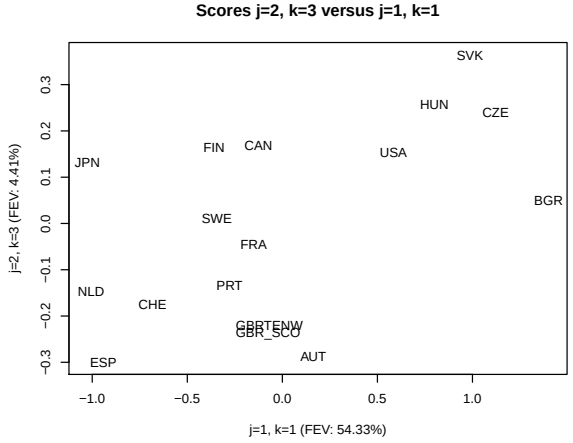


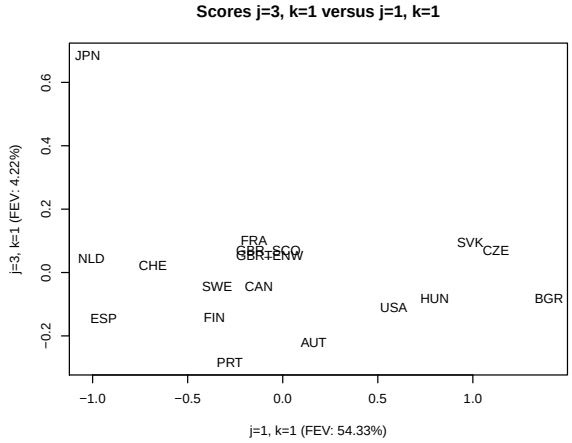
Product functions $\hat{\phi}_{jk}(t)\hat{\psi}_j(s)$ corresponding to the six terms contributing highest FVE in the marginal FPCA representation

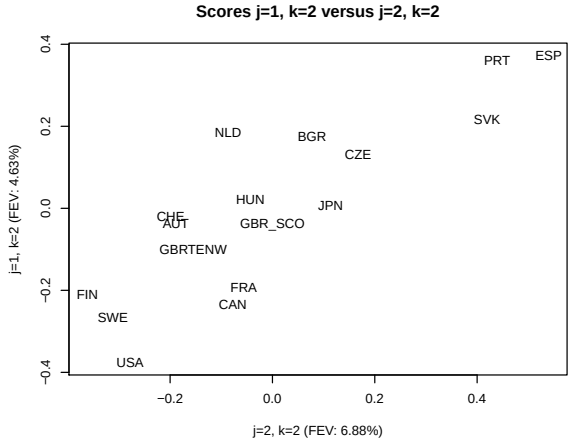

Scores j=2, k=1 versus j=1, k=1
CAN
USA
NLD
PRT
SVK
FRA
GBR_SCO
FIN
ESP
GBRTENW
CHE
AUT
JPN
SWE
CZE
HUN
BGR
j=2, k=1 (FEV: 13.04%)
j=1, k=1 (FEV: 54.33%)


Scores j=2, k=2 versus j=1, k=1
ESP
PRT
SVK
CZE
JPN
BGR
GBR_SCO
HUN
FRA
CAN
NLD
GBRTENW
AUT
CHE
USA
SWE
FIN
j=2, k=2 (FEV: 6.88%)
j=1, k=1 (FEV: 54.33%)


Scores j=1, k=2 versus j=1, k=1
ESP
PRT
SVK
NLD
BGR
CZE
HUN
JPN
CHE
GBR_SCOAUT
GBRTENW
FIN
FRA
CAN
SWE
USA
j=1, k=2 (FEV: 4.63%)
j=1, k=1 (FEV: 54.33%)


Scores j=2, k=3 versus j=1, k=1
SVK
HUN
CZE
FIN
CAN
USA
JPN
BGR
SWE
FRA
NLD
PRT
CHE
GBRTENW
GBR_SCO
AUT
ESP
j=2, k=3 (FEV: 4.41%)
j=1, k=1 (FEV: 54.33%)


Scores j=3, k=1 versus j=1, k=1
JPN
FRA
GBR_SCO
GBRTENW
SVK
CZE
NLD
CHE
SWE
CAN
HUN
BGR
USA
ESP
FIN
AUT
PRT
j=3, k=1 (FEV: 4.22%)
j=1, k=1 (FEV: 54.33%)


Scores j=1, k=2 versus j=2, k=2
PRT
ESP
SVK
NLD
BGR
CZE
HUN
JPN
CHE
AUT
GBR_SCO
GBRTENW
FIN
FRA
CAN
SWE
USA
j=1, k=2 (FEV: 4.63%)
j=2, k=2 (FEV: 6.88%)

Comparing multidimensional FPCA and marginal FPCA

- As expected, standard FPCA is more parsimonious than marginal FPCA for the fertility data: Four FPCs derived from the 2-dimensional Karhunen-Loève expansion explain about the same amount of variance as 6 terms from the marginal FPCA.

- Marginal FPCA represents the functional data as a sum of terms that are products of two functions, each depending on only one argument. This provides for much better interpretability and feature discovery.

- For instance, the second eigenfunction ψ_2 of the marginal FPCA corresponds to a level of fertility component, with a country-specific time-varying multiplier $\xi_2(t)$. Standard FPCA does not pinpoint these key features.

- Marginal FPCA makes it much easier than standard FPCA to analyze the time dynamics of the fertility process.

Common Principal Component Case

- This is a special case, where in the marginal FPCA model

$$X(s,t) = \mu(s,t) + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \chi_{jk}\phi_{jk}(t)\psi_j(s)$$

  the eigenfunctions $\phi_{jk}(t)$ in the Karhunen-Loève expansion of the random functions $\xi_j(t)$ do not depend on j.

- This leads to the simplified tensor product representation

$$X(s,t) = \mu(s,t) + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \chi_{jk}\phi_{k}(t)\psi_j(s).$$

- Application of this Common Principal Component model yielded similar results for the fertility data

- A special case of the marginal FPCA model, less parsimonious, but with improved interpretability. Can be consistently fitted by adopting the same trick as used in the marginal model twice, namely to base eigenanalysis on average covariances.

- Conditional FPCA. A previous approach for function-valued stochastic processes (Chen & Müller 2012),

$$X_i(s|t) = \mu(s|t) + \sum_{k=1}^{\infty} \eta_k(t)\varrho_k(s|t), \quad \eta_k(t) = \sum_{p=1}^{\infty}\sum_{p=1}^{\infty} \zeta_{ikp}\varsigma_{kp}(t)$$

  where $\varrho_k(\cdot|t)$ is the k-th eigenfunction of the conditional process at longitudinal time t and $\eta_k(t)$ are the random expansion coefficient functions, that are further expanded in their eigenfunction basis $\{\varsigma_{kp},\, p \geq 1\}$.

- Adopts two repeated FPCAs, the first (conditional) one in direction s for each fixed t and the second for the random functions $\eta_j(t)$, leading to challenges for theory and practice.

Outlook

- There are 4+ models to choose from:
  - Higher-dimensional Karhunen-Loève expansion and unrestricted FPCA – most parsimonious and least interpretable
  - Marginal FPCA
  - Marginal FPCA with Common Principal Components
  - Conditional FPCA
- Which model is most useful/preferred depends on the application, comparative model evaluation is good practice.

- Combining Stringing with Models for Function-Valued Stochastic Processes. High-dimensional functional data (gene expression time courses for p genes and n subjects, $p >> n$) can be stringed, thereby creating observations of a function-valued process. Then apply one of the models that have been discussed (research in progress).

Lectures on FDA – Part VIII

Nonlinear Methods for the analysis of functional data

Chen & Müller AS 2012

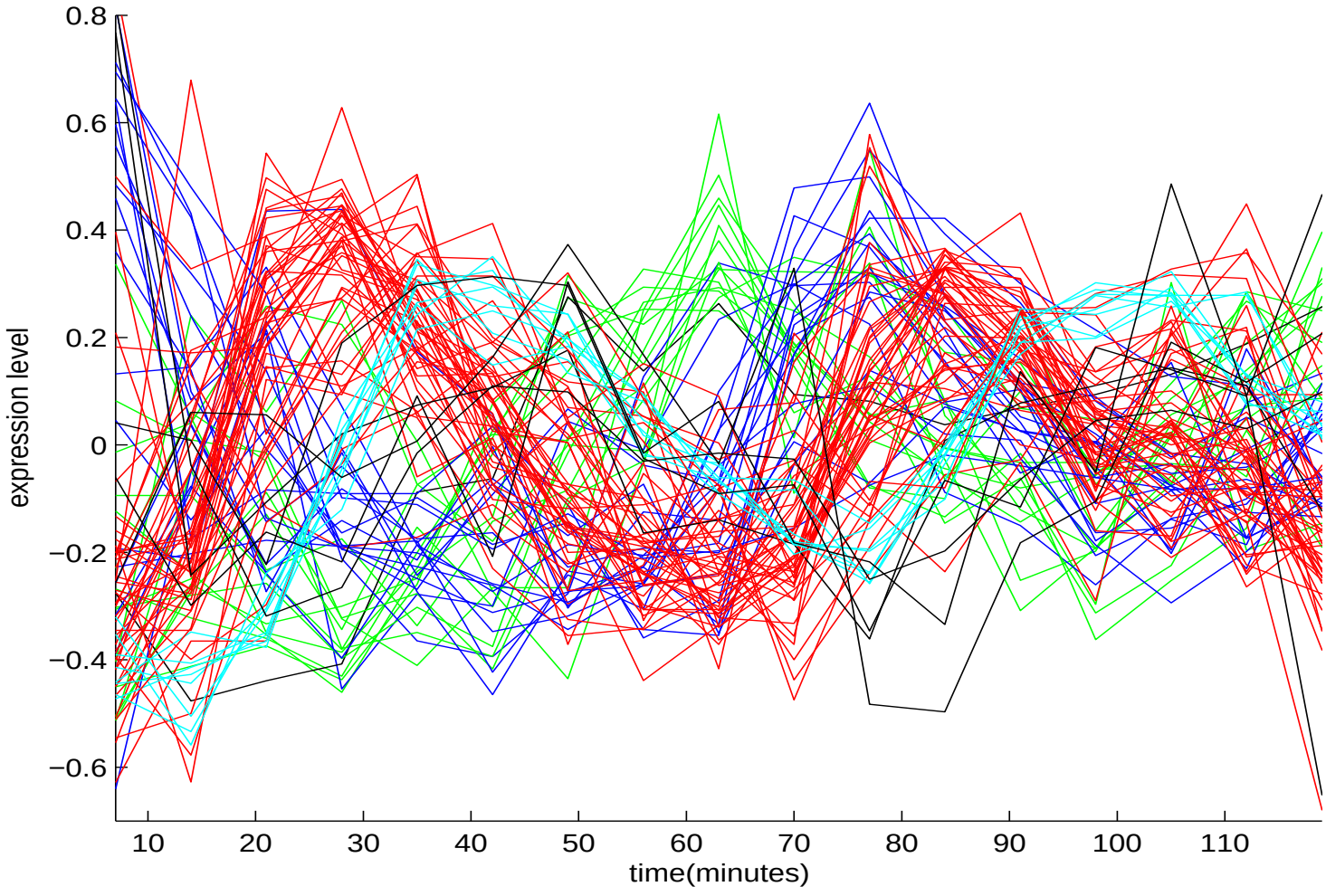
TIME WARPING FOR FUNCTIONAL DATA

- Functional data may contain amplitude and time variation
- Time variation is a nonlinear component
- Usually time and amplitude variation cannot be separated: Identifiability Problem
- Square integrable functional data can always be fully expanded in a complete basis of L^2, such as eigenbasis, Legendre basis or Fourier basis, regardless of whether there is time variation or not
- Motivation for modeling warping/alignment/registration:
  - Obtain more parsimonious models with fewer components
  - Better data interpretation
  - Better prediction

APPROACHES FOR TIME WARPING AND CURVE SYNCHRONIZATION

- Dynamic Time Warping (Sakoe & Chiba 1978)
- Shape-invariant modeling (Stützle, Gasser et al. 1980; Kneip & Gasser 1988)
- Landmark method (Structural Analysis, Kneip & Gasser 1992; Gasser & Kneip 1995)
- Nonparametric MLE (Rønn 2001, Gervini & Gasser 2005)
- Registration and Fitting (Kneip & Ramsay 2008)
- Pairwise Curve Synchronization (Tang & Müller 2008)
- Clustering and Warping (Tang & Müller 2009)
- Density synchronization (Bolstad et al 2003, Zhang & Müller 2011)

Yeast Cell Cycle Data



EXAMPLE: SIMPLE TIME-SHIFT WARPING

- Assume

$$X_i(t) = X_i(t, \tau_i) = \mu(t - \tau_i) + \delta Z_i(t - \tau_i),$$

for random time shifts τ_i.

- For a pair of random curves $X_i(t)$ and $X_j(t)$, the relative time shift is $s_{ij} = \tau_i - \tau_j$, $i, j = 1, \ldots, K$.
- For $\tilde{s}_{ij} = \arg\min_s d(X_j(t - s), X_i(t))$, with d the L^2 distance, and

$$\Delta_{ij}(s) = E\left(\int_{\mathcal{T}} \left(X_i(t, \tau_i) - X_j(t - s, \tau_j)\right)^2 \mathrm{dt} \,\middle|\, \tau_i, \tau_j\right),$$

under regularity conditions,

$$\tilde{s}_{ij} = \arg\min_s \Delta_{ij}(s) = s_{ij} + O(\delta) = \tau_i - \tau_j + O(\delta).$$

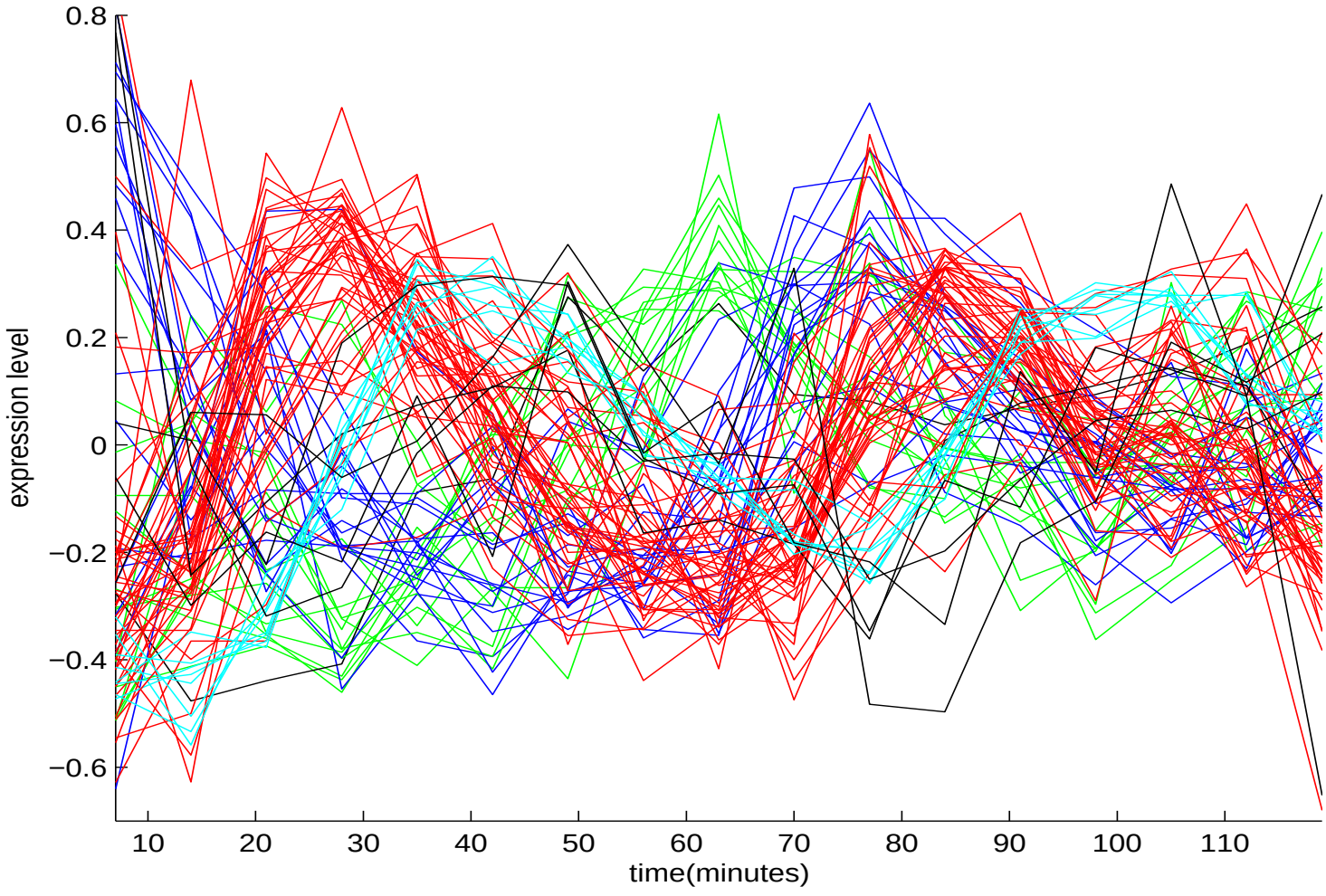
(Leng & M 2004), motivating $\tilde{s} = A\tau + \varepsilon$ for a design matrix A, and

$$\hat{\tau} = (A^T A)^{-1} A^T \tilde{s}.$$

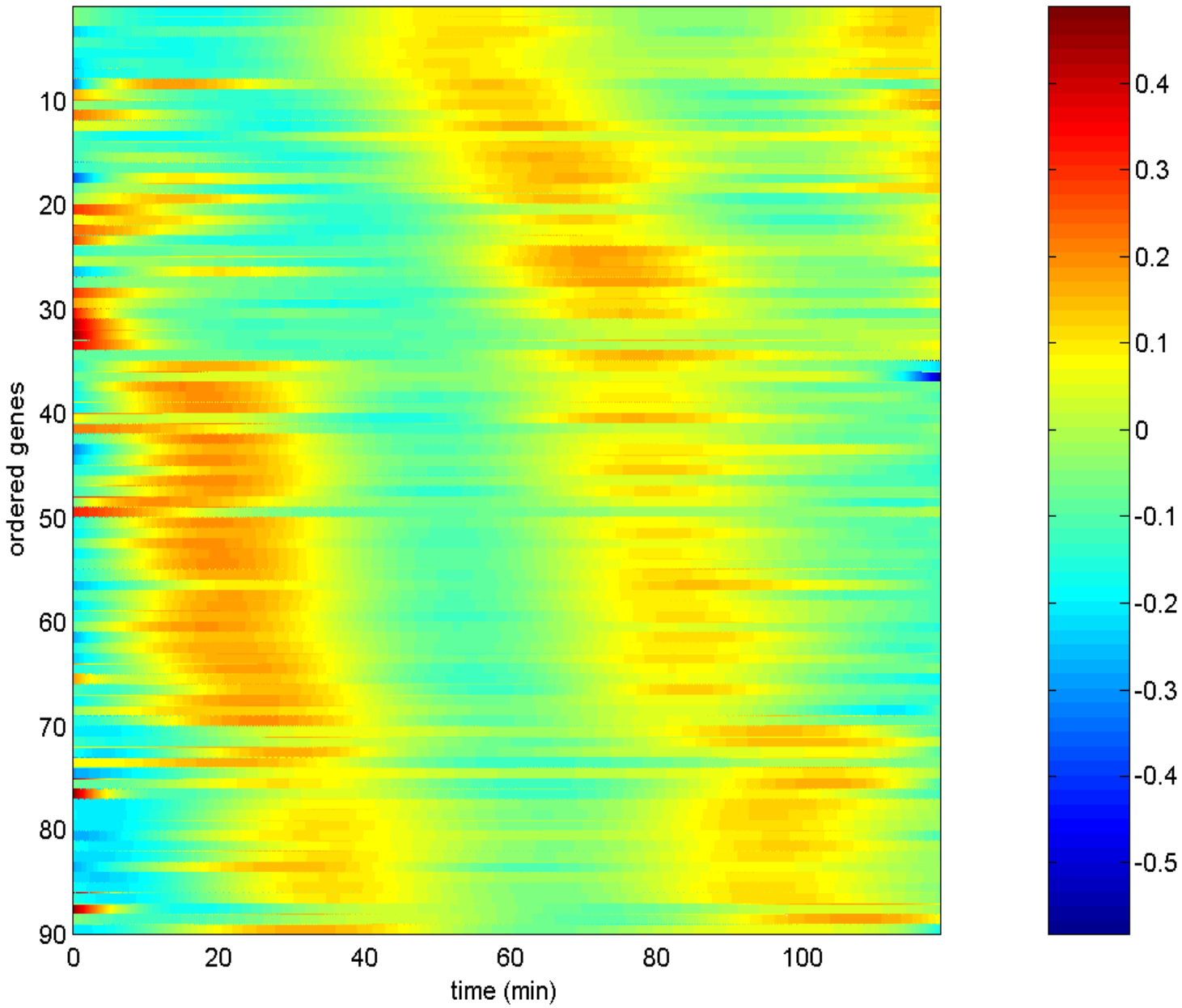
TIME ORDERING FOR YEAST GENE EXPRESSION PROFILES

- Yeast RNA Expression Level Data (Alter, Brown & Botstein 2000)
- mRNA levels of 6108 cell-cycle genes from *Saccharomyces cerevisiae* obtained from α-factor synchronized yeast cells
- Consider subset of 82 gene profiles with expression measured at 7-minute intervals for a total of 119 minutes, recorded as normalized log_2 ratios
- Cell phases: G1, S, G2, M

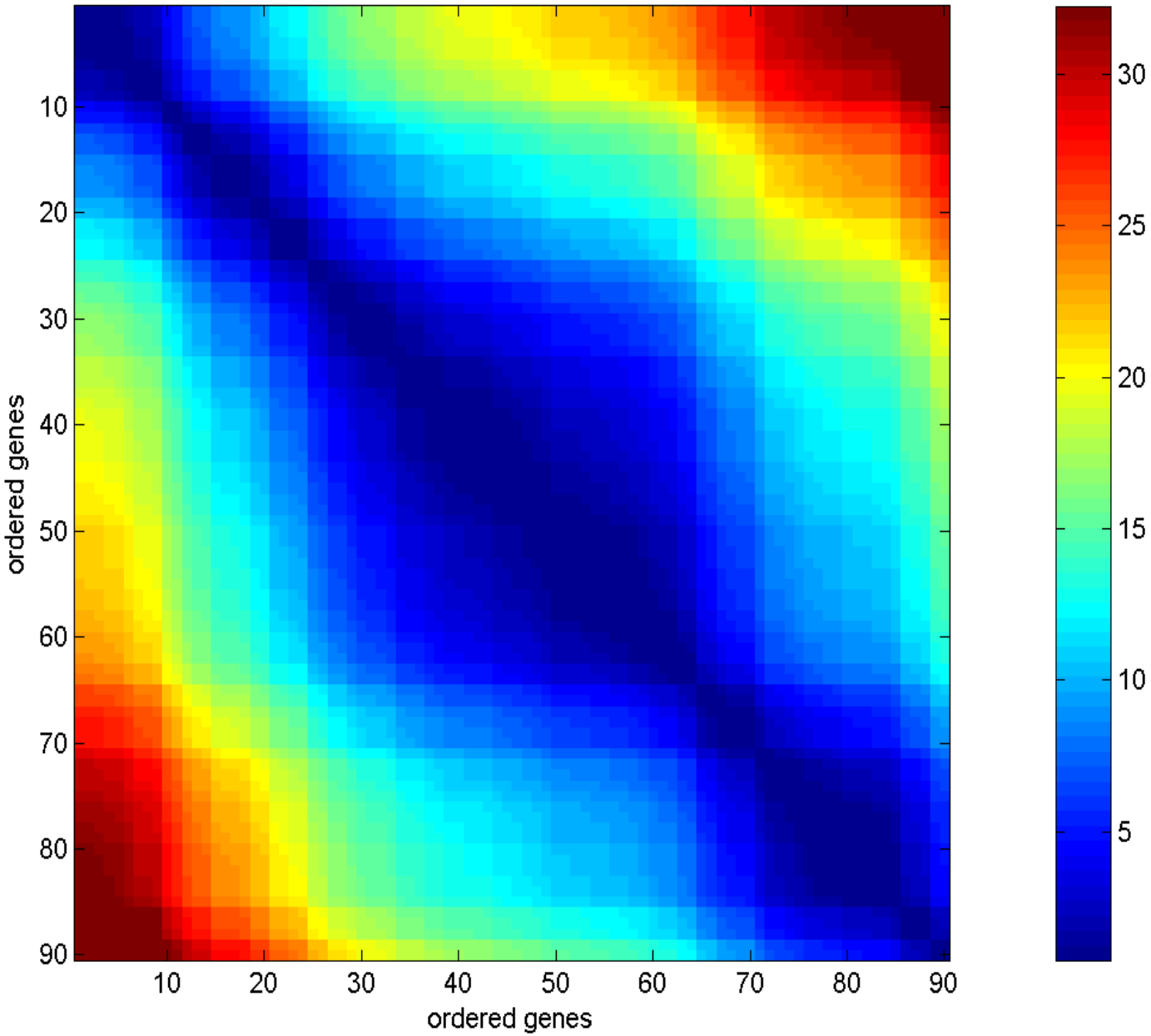
Yeast Cell Cycle Data



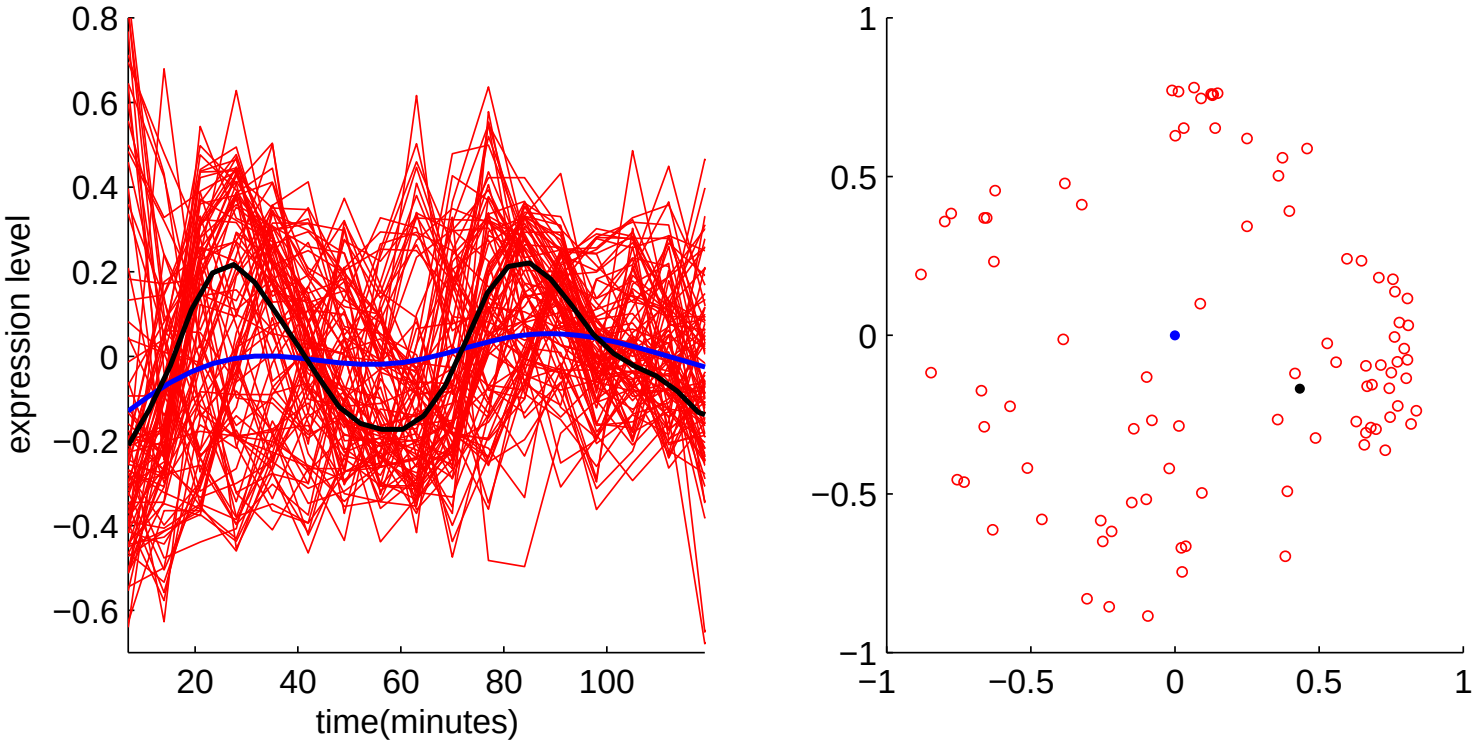
Heat Map


ordered genes
10
20
30
40
50
60
70
80
90
0
20
40
60
80
100
time (min)
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5

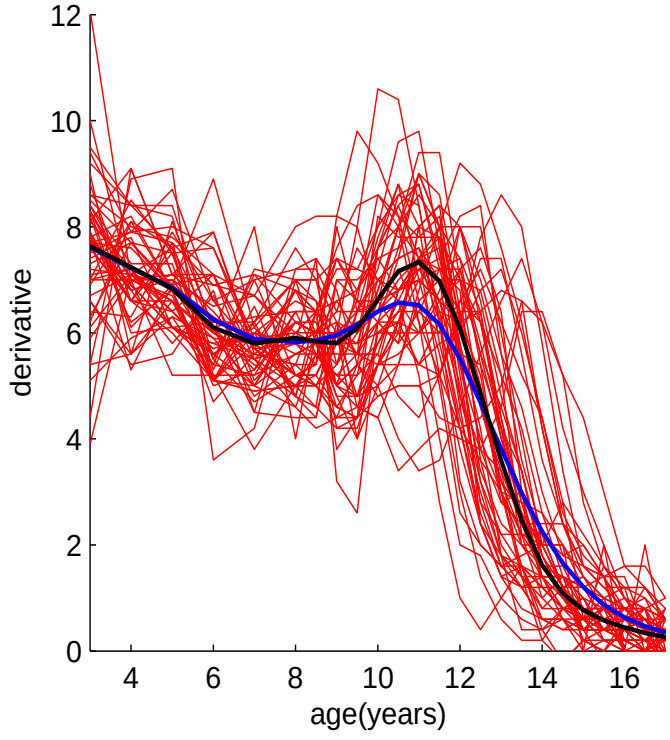
Clustering Time Shifts

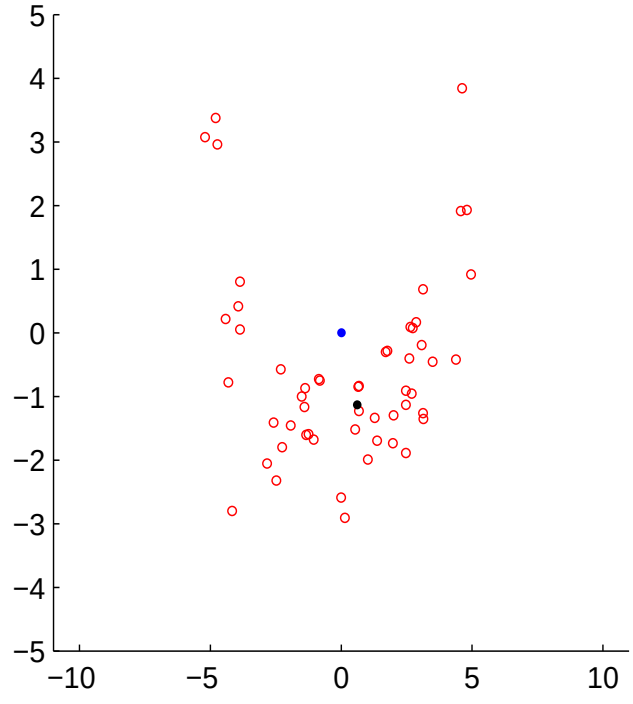


FPC2 vs FPC1 for Yeast Cell Cycle Data



Berkeley Growth Study: Girls





MANIFOLDS IN FUNCTION SPACE

"Simple" functional manifolds $\mathcal{M}$ in L^2 space that are isomorphic to a subspace of the Euclidean space

Represented by coordinate map $\psi^{-1} : \mathbb{R}^d \to \mathcal{M} \subset L^2$, which is bijective, such that ψ, ψ^{-1} are continuous and isometric.

d is the intrinsic dimension of the manifold $\mathcal{M}$.

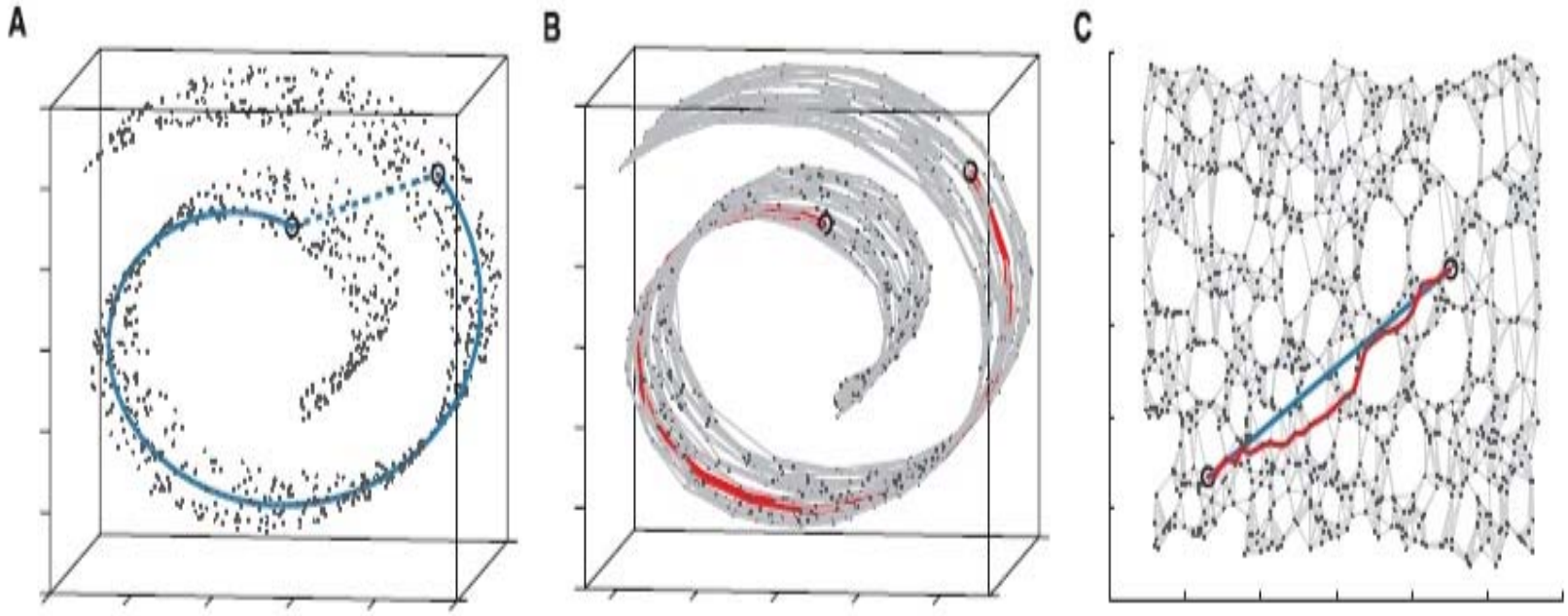
Probability measure Q for random vectors $\boldsymbol{\vartheta} \in \mathbb{R}^d$ leads to induced probability measure Q_ψ in L^2 by $Q_\psi(A) = Q(\psi(A))$.

Distances on manifold $\mathcal{M}$: L^2 distance – not adapted to nonlinearity;
Geodesic distance $d_g(x_1, x_2)$ is the length of the shortest path on $\mathcal{M}$ connecting the two points, and therefore is adapted to $\mathcal{M}$.

GLOBAL MANIFOLD LEARNING

Geodesic distances and ISOMAP (Tenenbaum 2000) under isometry assumption

The Swiss Roll



FUNCTIONAL MANIFOLD MEAN

Define

$$\boldsymbol{\mu} = \mathsf{E}\{\psi(X)\}, \qquad \mu^{\mathcal{M}} = \psi^{-1}(\boldsymbol{\mu}),$$

where $\boldsymbol{\mu}$ is the mean in the d-dimensional representation space, and $\mu^{\mathcal{M}}$ is the manifold mean in L^2 space.

Can show: $\boldsymbol{\mu}$ is the Fréchet mean with regard to geodesic distance and does not depend on specific choice of map φ.

Traditional cross-sectional means for functional data in L^2 can be far away from the data cloud and then do not represent the data in a meaningful way.

Going beyond the mean: Analogous problems when linearly representing random functions in an orthonormal L^2 basis.

MANIFOLD MODES OF VARIATION

For functions in $\mathcal{M} \subset L^2$: eigenfunction expansion

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} A_k \phi_k(t),\ t \in \mathcal{T}, \quad A_k = \int_{\mathcal{T}} (X(t) - \mu(t))\phi_k(t)dt$$

will not generally provide a parsimonious representation.

Similarly, eigenfunction-based modes of functional variation

$$X_{j,\alpha} = \mu + \alpha \lambda_j^{1/2} \phi_j,\ j = 1, 2, \ldots,\ \alpha \in \mathbb{R},$$

generally will not satisfy $X_{j,\alpha} \in \mathcal{M}$.

Functional manifold component (FMC) vectors $e_j^{\mathcal{M}} \in \mathbb{R}^d$, $j = 1, \ldots, d$, are defined by the eigenvectors of the covariance matrix of $\psi(X) \in \mathbb{R}^d$, i.e.,

$$\mathrm{Cov}(\psi(X)) = \sum_{j=1}^{d} \lambda_j^{\mathcal{M}} (e_j^{\mathcal{M}})(e_j^{\mathcal{M}})^T$$

where $\lambda_1^{\mathcal{M}} \geq \ldots \geq \lambda_d^{\mathcal{M}}$ are the eigenvalues of $\mathrm{Cov}(\psi(X))$.

Manifold modes of functional variation

$$X_{j,\alpha}^{\mathcal{M}} = \psi^{-1}\big(\boldsymbol{\mu} + \alpha(\lambda_j^{\mathcal{M}})^{\frac{1}{2}} e_j^{\mathcal{M}}\big),\ j = 1, \ldots, d,\ \alpha \in \mathbb{R},$$

where $\boldsymbol{\mu}$ is the mean in the d-dimensional representation space.

- Only finitely many modes of variation, at most d.
- The manifold modes of variation $X_{j,\alpha}^{\mathcal{M}}$ are uniquely defined
- Functions $X \in \mathcal{M}$ can be uniquely represented by a d-vector of FMCs $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_d) \in \mathbb{R}^d$, in terms of

$$X = \psi^{-1}(\boldsymbol{\mu} + \sum_{j=1}^{d} \vartheta_j e_j^{\mathcal{M}}), \quad \vartheta_j = \langle \psi(X) - \boldsymbol{\mu}, e_j^{\mathcal{M}} \rangle,\ j = 1, \ldots, d$$

IMPLEMENTATION

Tenenbaum et al (ISOMAP), Science (2000): Estimate ψ by

$$\hat{\psi} = \text{argmin}\psi \sum_{i,j=1}^{n} \{||\psi(X_i) - \psi(X_j)|| - d_g(X_i, X_j)\}^2,$$

where the infimum is taken over all functions $\psi : L^2 \to \mathbb{R}^d$ and $d_g(\cdot, \cdot)$ is the geodesic distance.

- Can be interpreted as a modified version of MDS.
- In practice: Construct ψ only over finite sample points X_i.
- Starting point: Observed data are $Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}$.
  $\Rightarrow$ Need to recover functions and their L^2 distances:
  Pre-smoothing (for dense designs only) or PACE method (for both dense and sparse designs)

- Require small distances for ISOMAP, since only near neighbor relations matter (Dijkstra's local graph algorithm)
  - FPCA representation with large number of included components
  - Directly track L^2 distances with conditioning for noisy or sparse data (Peng & M 2008, AOAS)
- Recovered functions are not exactly located on the manifold $\mathcal{M}$, due to measurement errors, sparse measurements, etc., even if true functions are on manifold: Modify Dijkstra's algorithm to allow for local connection paths to occur only in relatively dense areas of the data, by adding a penalty.
- Interpolation of the embedding map ψ^{-1}: For $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\hat{\psi}^{-1}(\boldsymbol{\theta}) = \frac{\sum_i \kappa\big(H^{-1}(\hat{\psi}(X_i) - \boldsymbol{\theta})\big)\hat{X}_i}{\sum_i \kappa\big(H^{-1}(\hat{\psi}(X_i) - \boldsymbol{\theta})\big)},$$

where κ is a d-dimensional kernel, H a smoothing parameter.

- Choice of intrinsic dimension d:
  Fraction of distance explained (Tenenbaum et al. 2000)

- Auxiliary parameters (including smoothing bandwidths) selected by cross-validation

- **Asymptotics:** Under regularity conditions, the estimated manifold mean and manifold modes $\hat{X}_{j,\alpha}^{\mathcal{M}}$ are consistent as $n \to \infty$. Rates of convergence have been recently derived in Chen & M (2012, AS). They depend on convergence behavior of Isomap.

- Proof: Extend local convergence to global convergence of the manifold embedding map ψ^{-1}. Use properties of d-dimensional smoothers.

SIMULATION I

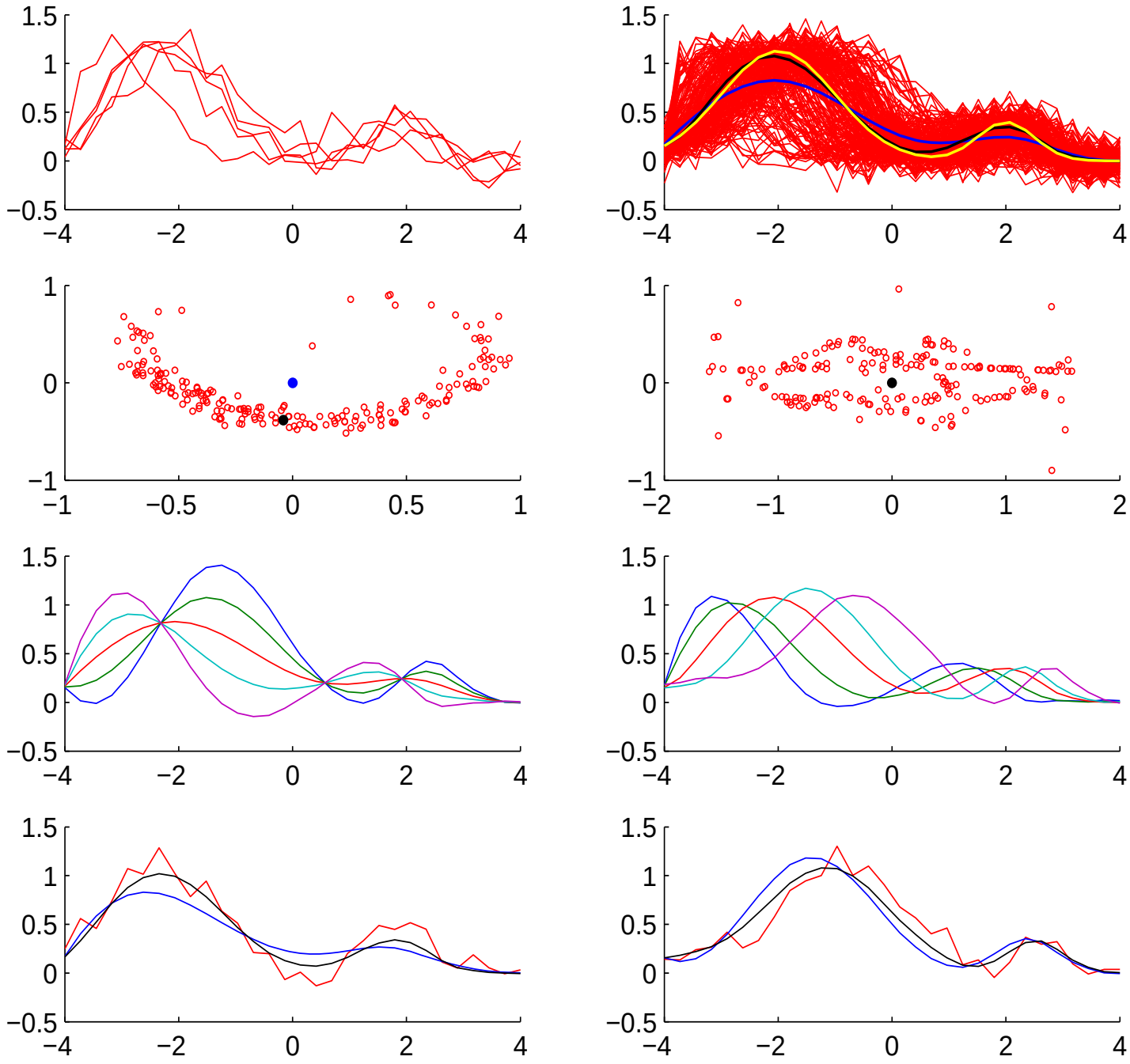
Simulating functional manifolds:
$n = 200$, 30 equi-spaced observations per function, signal-to-noise ratio is 2.

$$\mathcal{M}_1 = \{X \in L^2([-4,4]) : X(t) = \mu(h_\alpha(t))\},$$

where $\mu(t) = \frac{2}{\sqrt{\pi}} \exp\{-\frac{1}{2}(t+2)^2\} + \frac{1}{\sqrt{2\pi}} \exp\{-2(t-2)^2\}$.

Random warping of a common shape function μ, which has two peaks, where the time warping function h_α is generated from the cumulative Beta distribution family and α is a random parameter, $\alpha = \max(-1, Z)$, where $Z \sim \mathrm{N}(0, 0.09)$.

Dimension $d = 1$

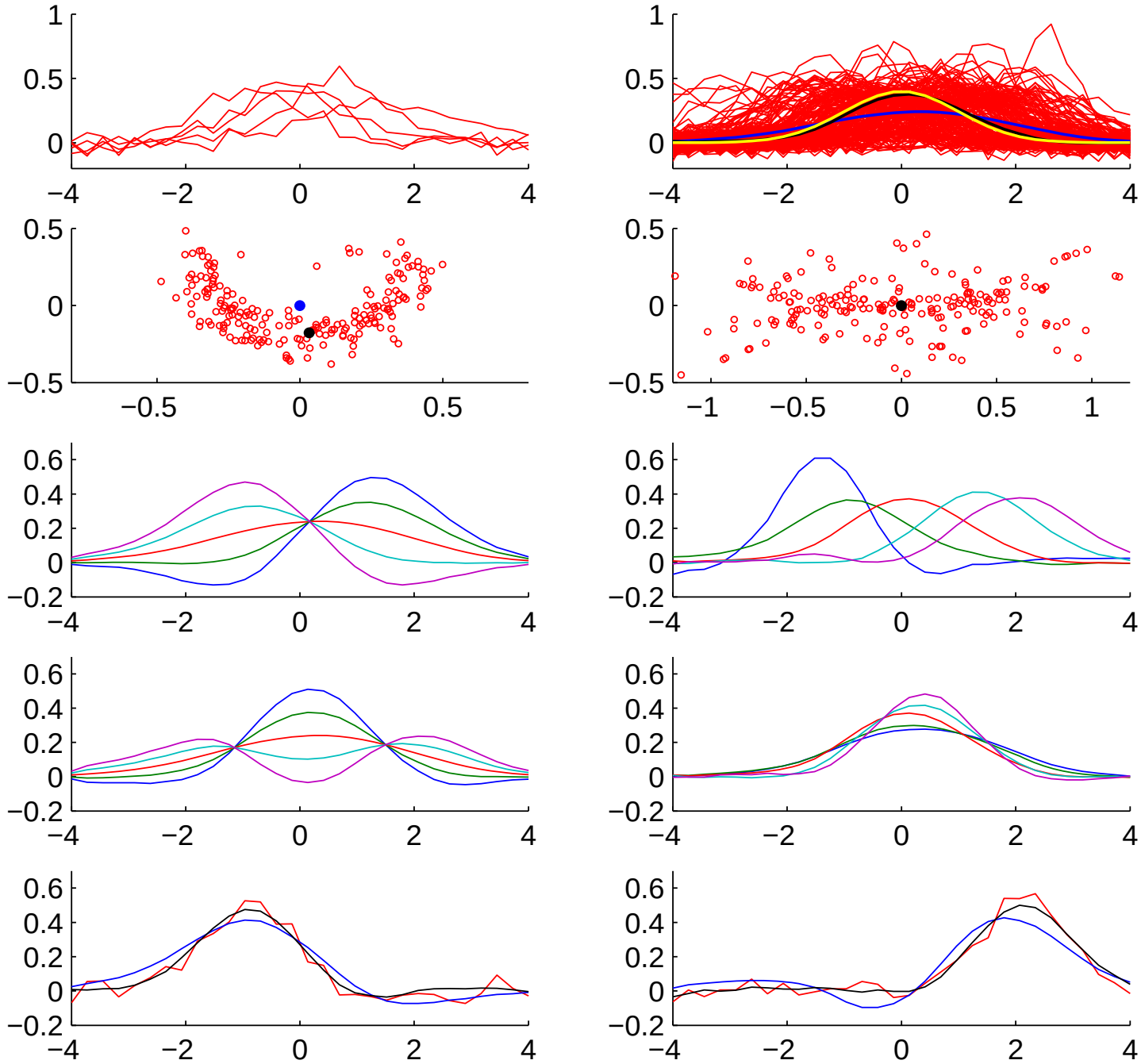

1.5
1
0.5
0
−0.5
−4
−2
0
2
4
1.5
1
0.5
0
−0.5
−4
−2
0
2
4
1
0
−1
−1
−0.5
0
0.5
1
1
0
−1
−2
−1
0
1
2
1.5
1
0.5
0
−0.5
−4
−2
0
2
4
1.5
1
0.5
0
−0.5
−4
−2
0
2
4
1.5
1
0.5
0
−0.5
−4
−2
0
2
4
1.5
1
0.5
0
−0.5
−4
−2
0
2
4

SIMULATION II

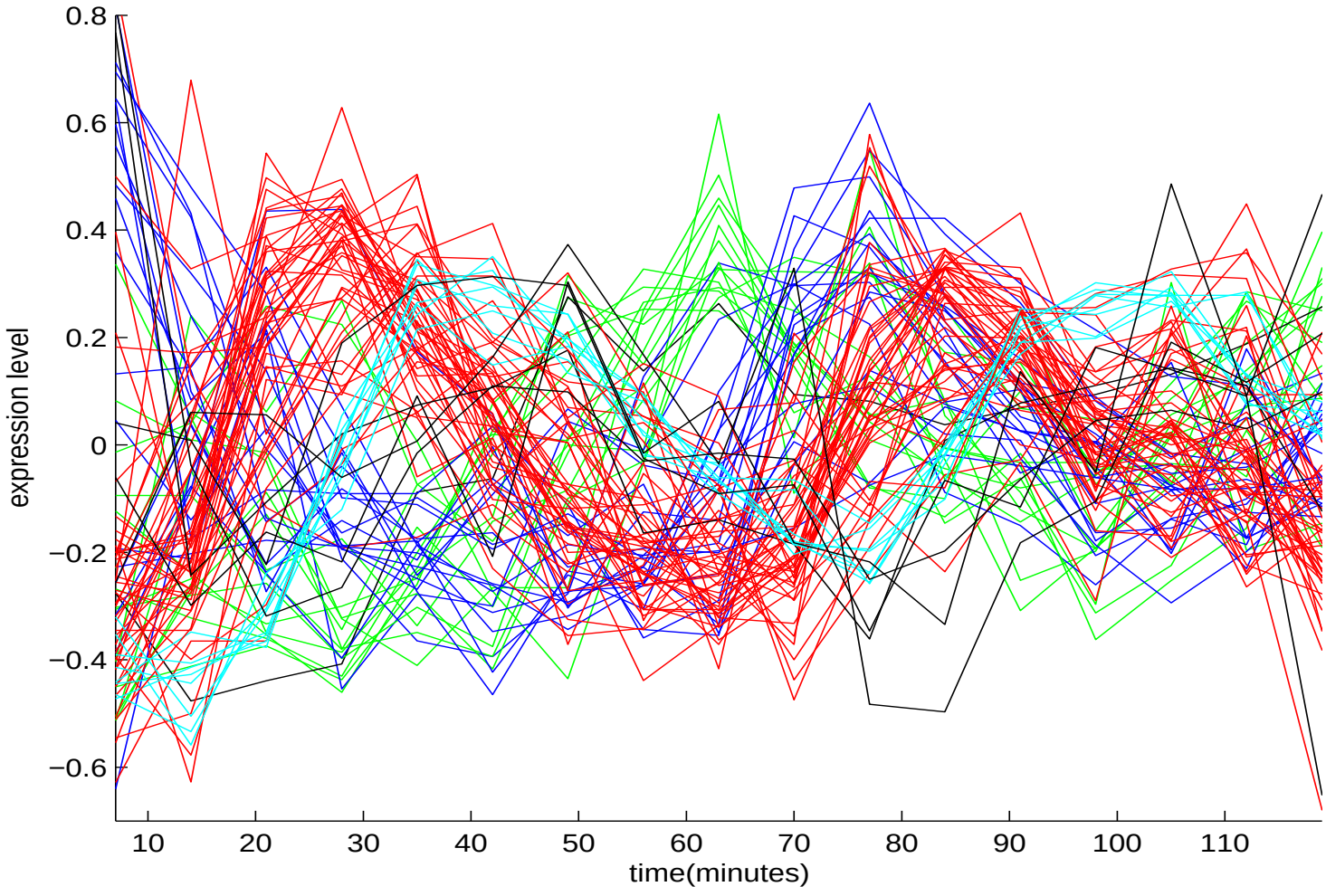
$$\mathcal{M}_2 = \{X \in L^2([-4,4]) : \\ X(t) = \frac{1}{\sqrt{2\pi\alpha^2}} \exp[-\frac{1}{2\alpha^2}(t-\beta)^2],\ \alpha > 0,\ \beta \in \mathbb{R}\}.$$

Collection of Gaussian densities, corresponding to a shift-scale family, where $\alpha = \max(0, Z)$, $Z \sim \mathrm{N}(1, 0.04)$ and $\beta \sim \mathrm{N}(0, 1)$.

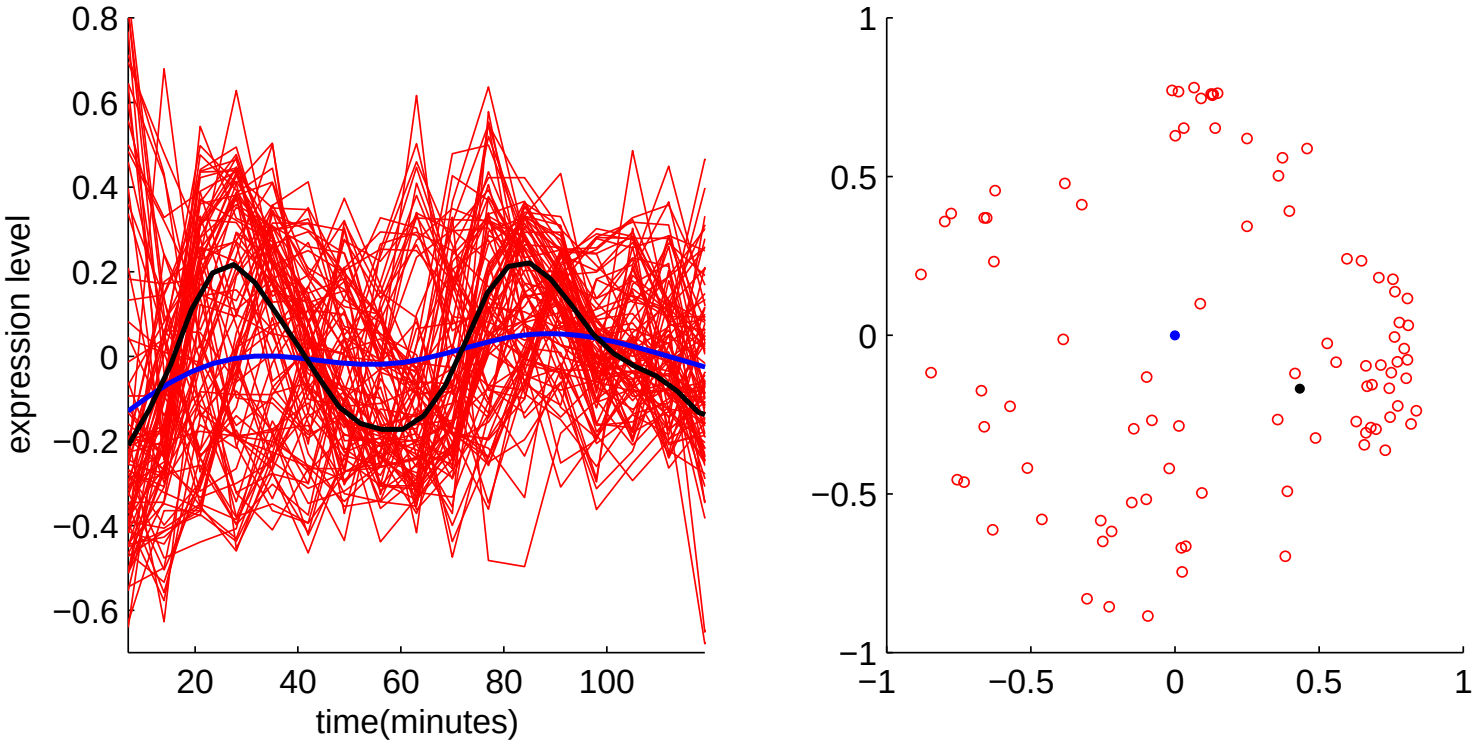
Dimension $d = 2$


1
0.5
0
−4
−2
0
2
4
0.5
0
−0.5
−0.5
0
0.5
−1
1
0.6
0.4
0.2
−0.2

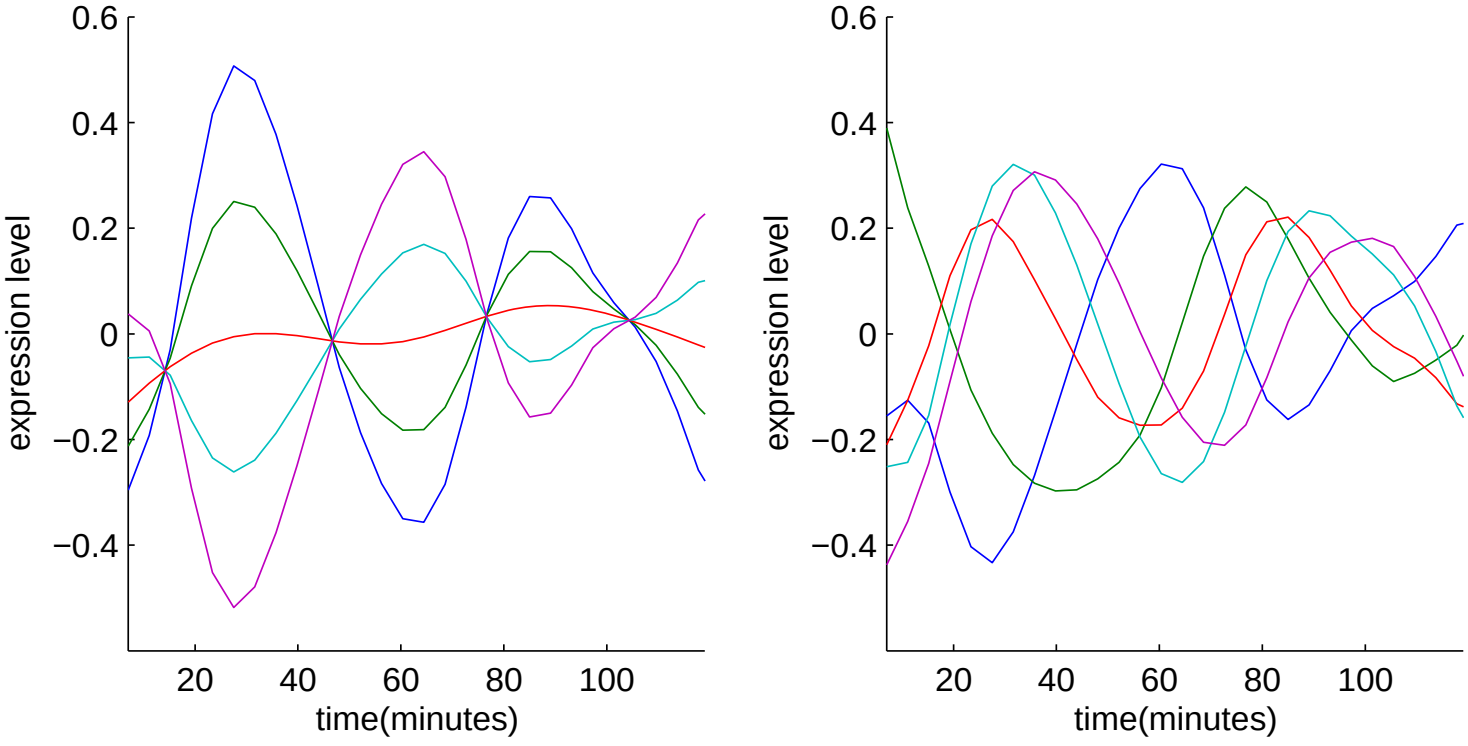
Yeast Cell Cycle Data



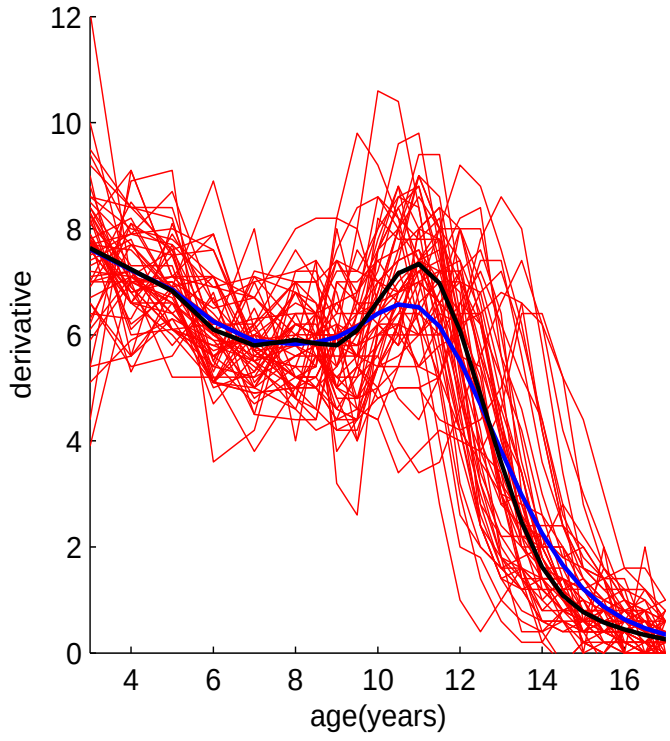
FPC2 vs FPC1 for Yeast Cell Cycle Data

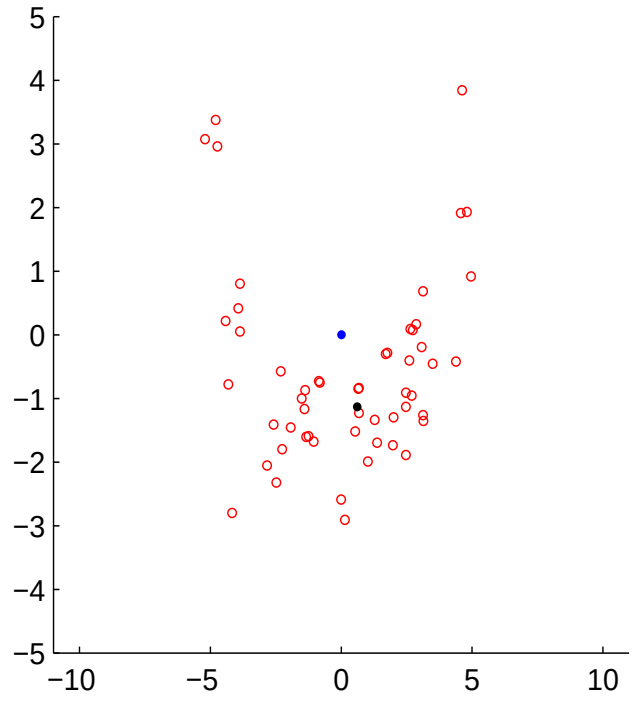


First Principal Mode of Variation and Manifold Mode of Variation for Yeast Cell Cycle Data

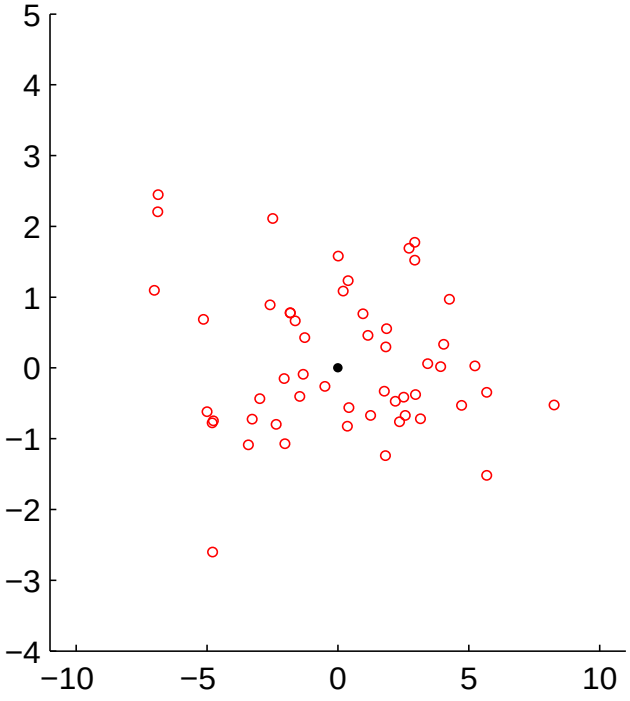


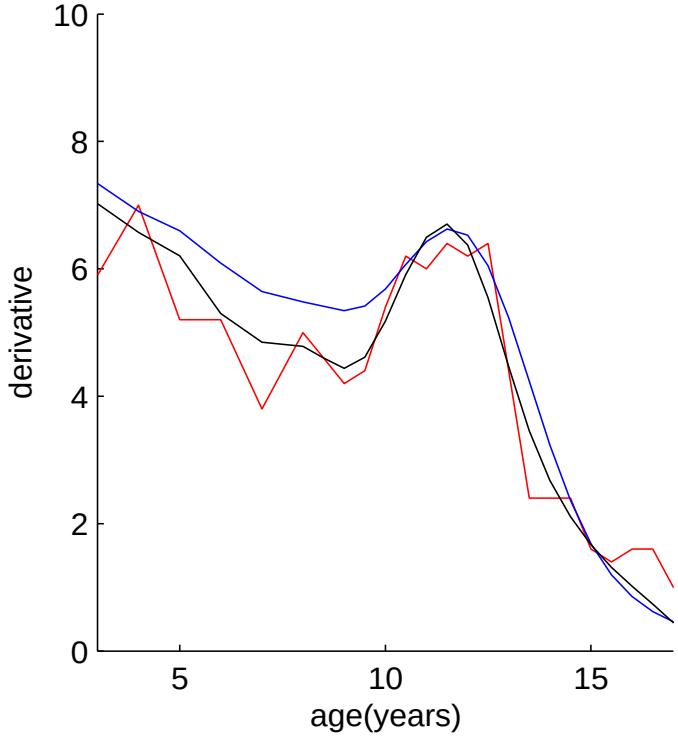
FPC2 vs FPC1 for Berkeley Growth Curves



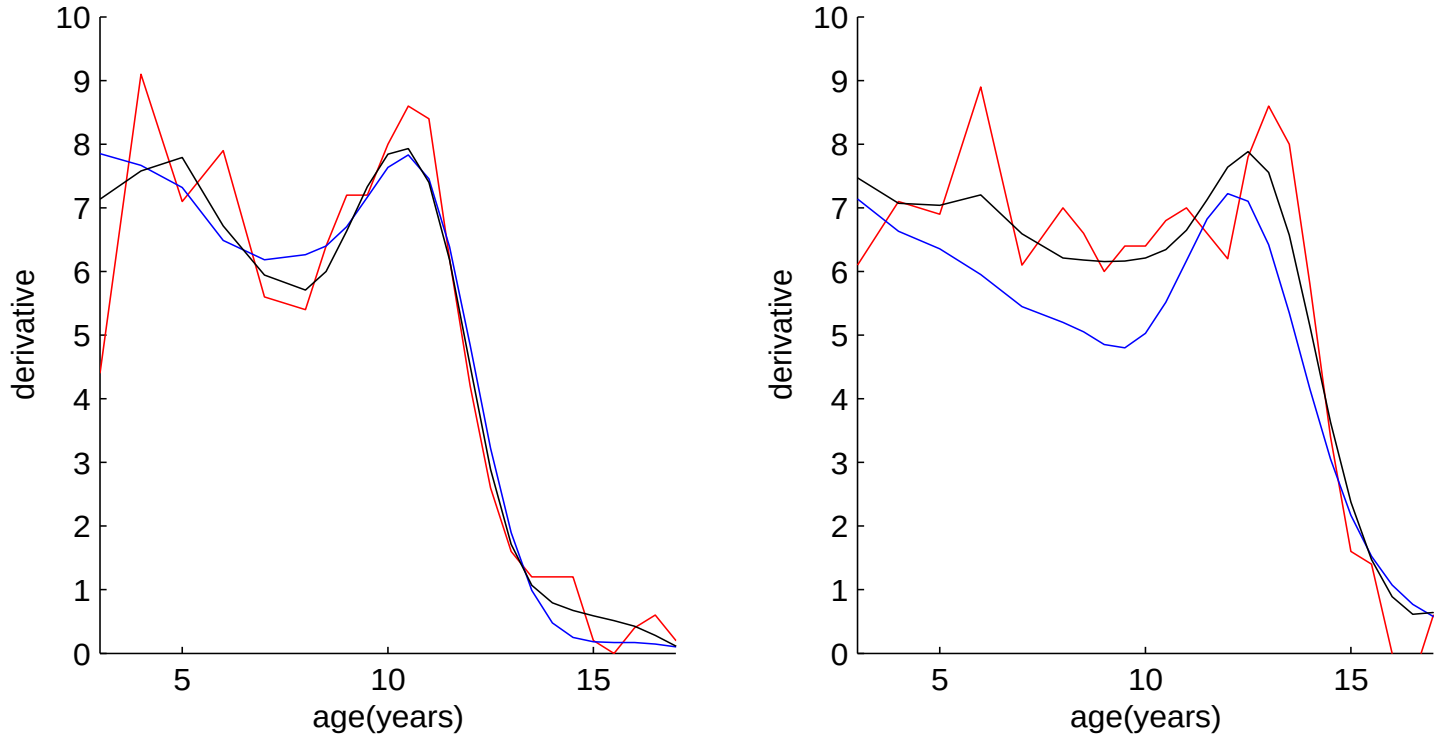


Growth Curves: Representation Space and Trajectory Fits with Principal and Manifold Components

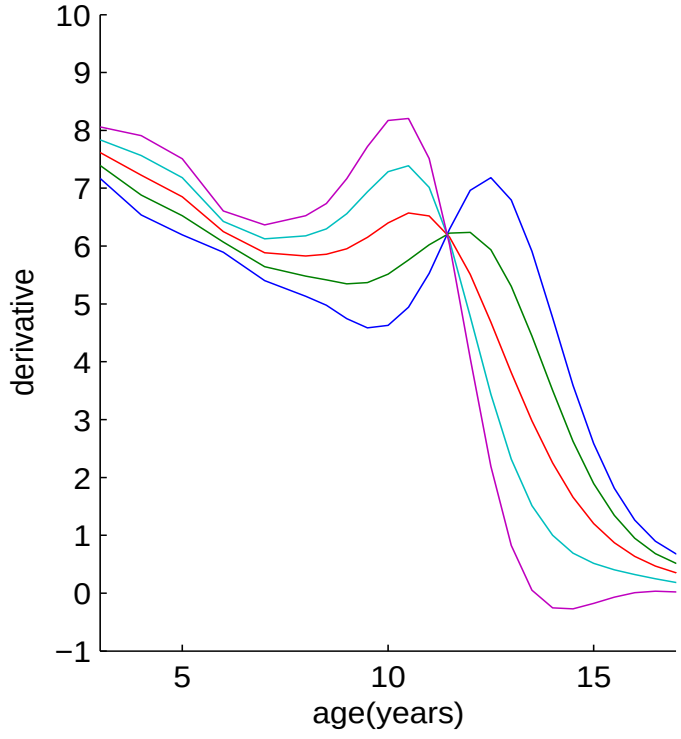


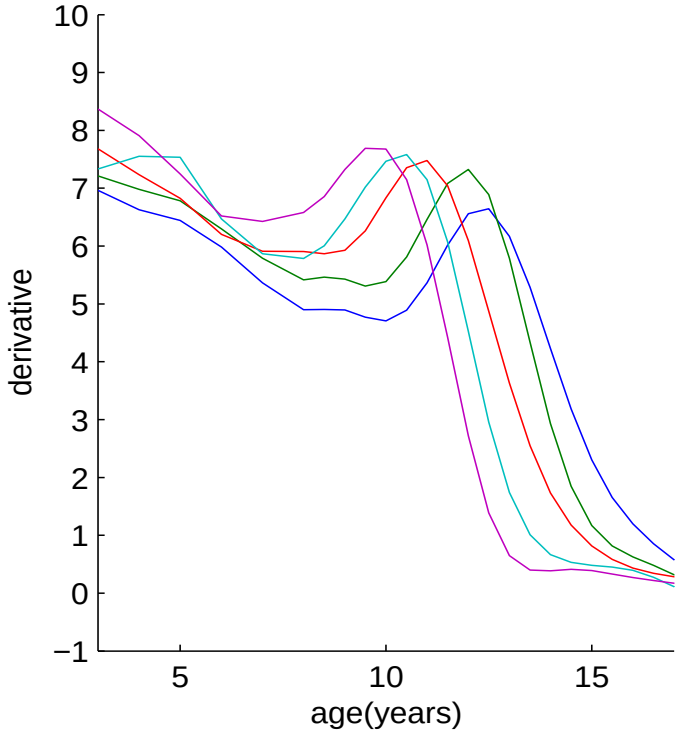


Growth Curves: Trajectory Fits with Principal and Manifold Components

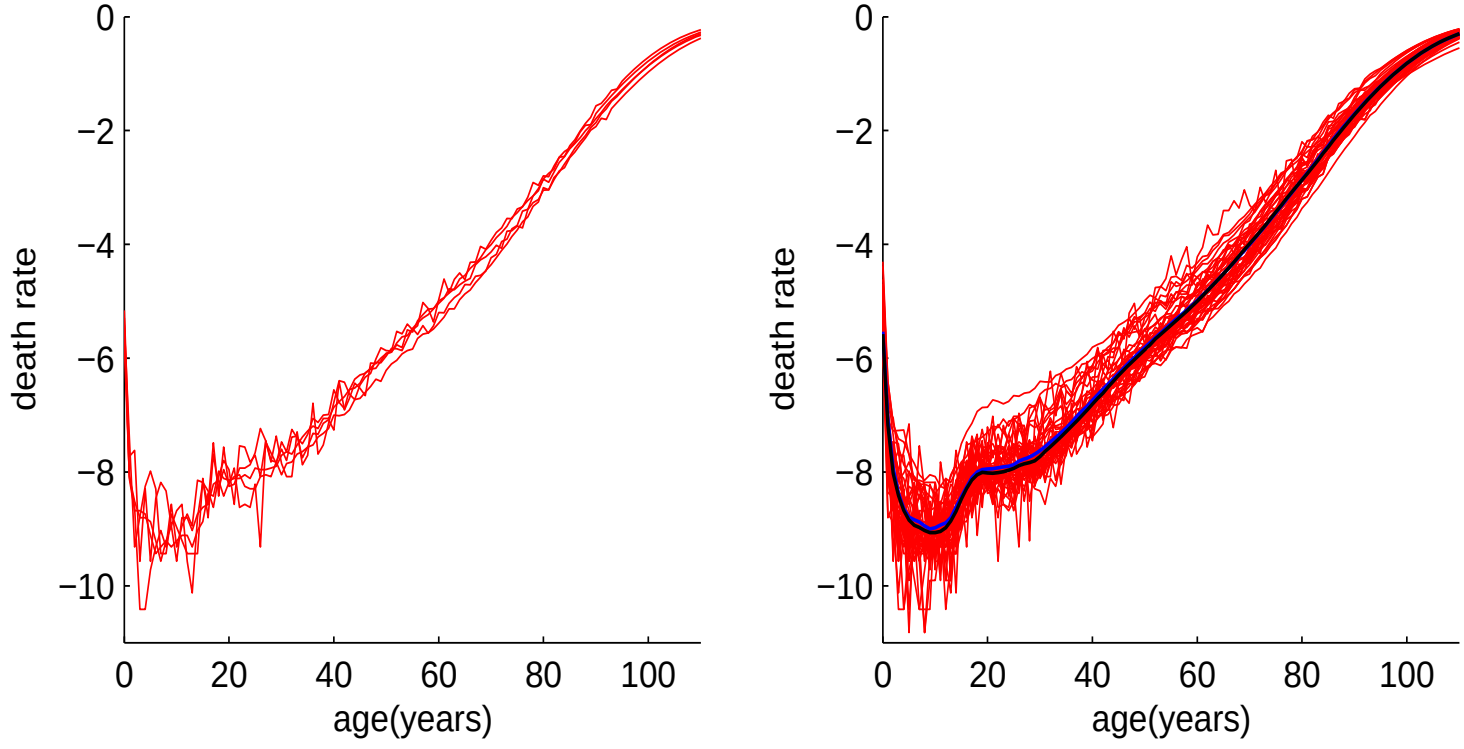


Growth Curves: Principal and Manifold Modes of Variation

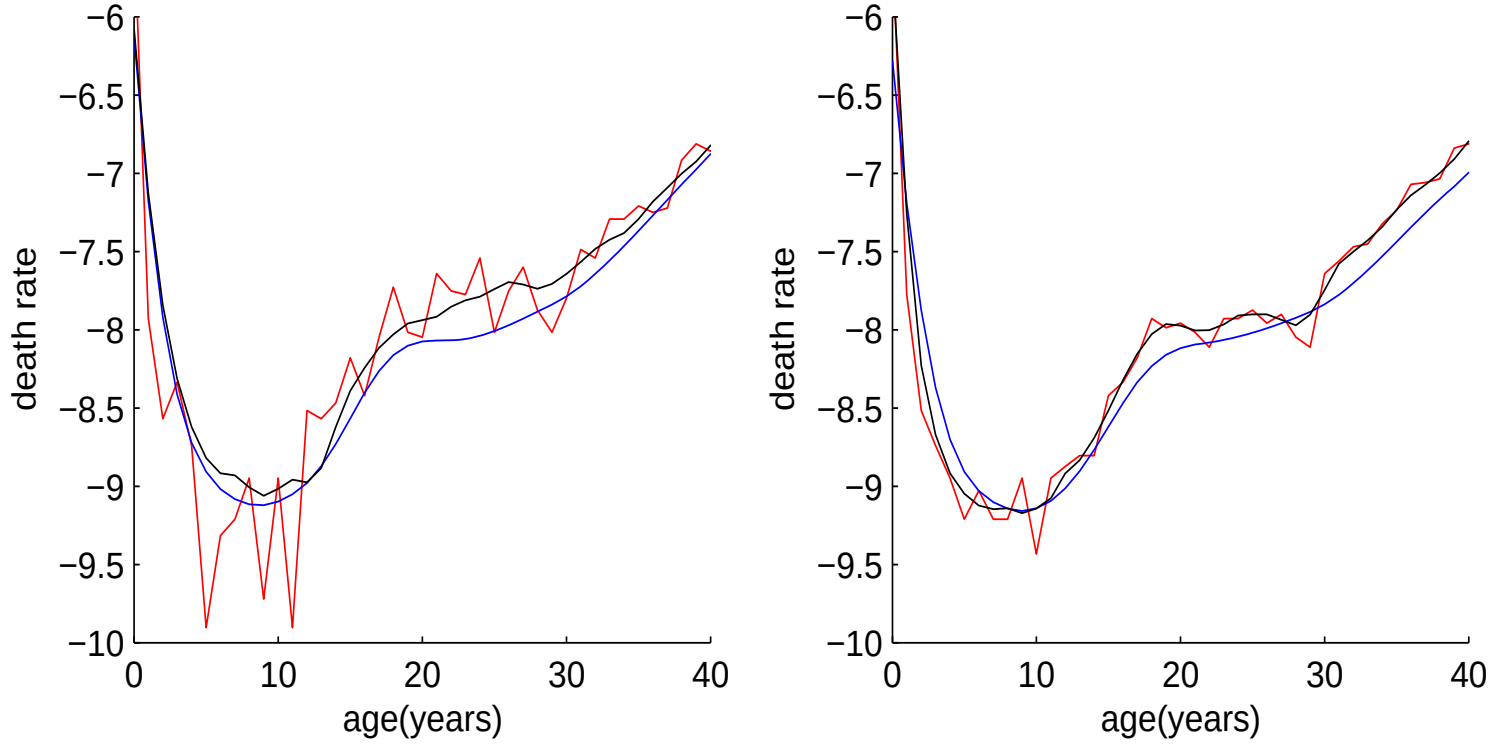




Log Hazard Rates Reflecting Human Mortality for 44 Countries



Trajectory Fits with Principal and Manifold Components for Human Mortality



FUNCTIONAL MANIFOLD REGRESSION

- Apply manifold embedding to functional regression problems, especially when time warping or other nonlinear features may play a role.
- For $Y \in \mathbb{R}$, $X \in \mathcal{M} \subset L^2$:

$$E(Y \mid X) = E(Y \mid \psi(X)) = E(Y \mid \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \mathbb{R}^d.$$

  Can use linear, additive or single index regression.
- Advantage: Predictor is low-dimensional due to the nonlinear dimension reduction. No need to worry about functional asymptotics where number of included basis functions increases.
- Application: Predict adult height from available height measurements on $[0, 12]$ for growth data. Functional linear manifold regression reduces cross-validation prediction error by about 15% compared to functional linear regression.