LEARNING DYNAMICS FROM FUNCTIONAL DATA

Hans-Georg Müller
UC Davis

RED RAIDER SYMPOSIUM LUBBOCK

26 October 2012

Based on joint work with
Wenwen Tao, UC Davis
Nicolas Verzelen, University of Montpellier
Fang Yao, University of Toronto

INTRODUCTION

Data: Longitudinal studies, e.g. Baltimore Longitudinal on Aging; e-Bay online auction data

Model: Sample of irregularly measured realizations of an underlying stochastic process, assumed to be smooth

Goals: Estimating derivatives for irregularly sampled random trajectories

Learning the underlying dynamics – empirical differential equation

Methods: Functional principal component analysis; Smoothing and differentiation (local least squares); Representations of stochastic processes

STOCHASTIC PROCESS PERSPECTIVE

Assume observed data are generated by underlying stochastic process $X \in L^2(T)$ with finite second moments:

$\mu(t) = E(X(t))$ mean function
$G(s,t) = \mathrm{cov}\{X(s), X(t)\}$ covariance function.

Define auto-covariance operator $(A_G f)(t) = \int f(s)G(s,t)\,ds$ with orthonormal eigenfunctions ϕ_k and ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$,

$$(A_G \phi_k)(t) = \lambda_k \phi_k(t)$$

FUNCTIONAL PRINCIPAL COMPONENTS (FPC)

KARHUNEN-LOÈVE REPRESENTATION USING FPCs

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} A_k \phi_k(t),$$

where $A_k = \int_0^T \{X(t) - \mu(t)\}\phi_k(t)dt$, are uncorrelated r.v. with $EA_k = 0$, $EA_k^2 = \lambda_k$, the functional principal components.

Some key papers:

- Grenander 1950: Basic ideas (following up on Karhunen 1949)
- C.R. Rao 1958: Preliminary version for growth curves
- Castro, Lawton & Sylvestre 1987: Modes of Variation in industrial applications
- Rice & Silverman 1991, Rice & C. Wu 2001: B-splines and systematic study

- Book: Ramsay & Silverman 2005: Presmoothing (usually inefficient)
- Bali, Boente, Tyler & J.L. Wang 2012: Systematic study of robust FPCA

Why Functional Principal Components?

- Parsimonious description of longitudinal/functional data as it is the unique linear representation which explains the highest fraction of variance in the data with a given number of components.
- Main attraction is equivalence $X \equiv \{A_1, A_2, \ldots\}$ so that X can be expressed in terms of mean function μ and the countable sequence of eigenfunctions and uncorrelated FPC scores A_k.
- For modeling functional regression: Functions $f(X)$ have an equivalent function $g(A_1, A_2, \ldots)$ so that

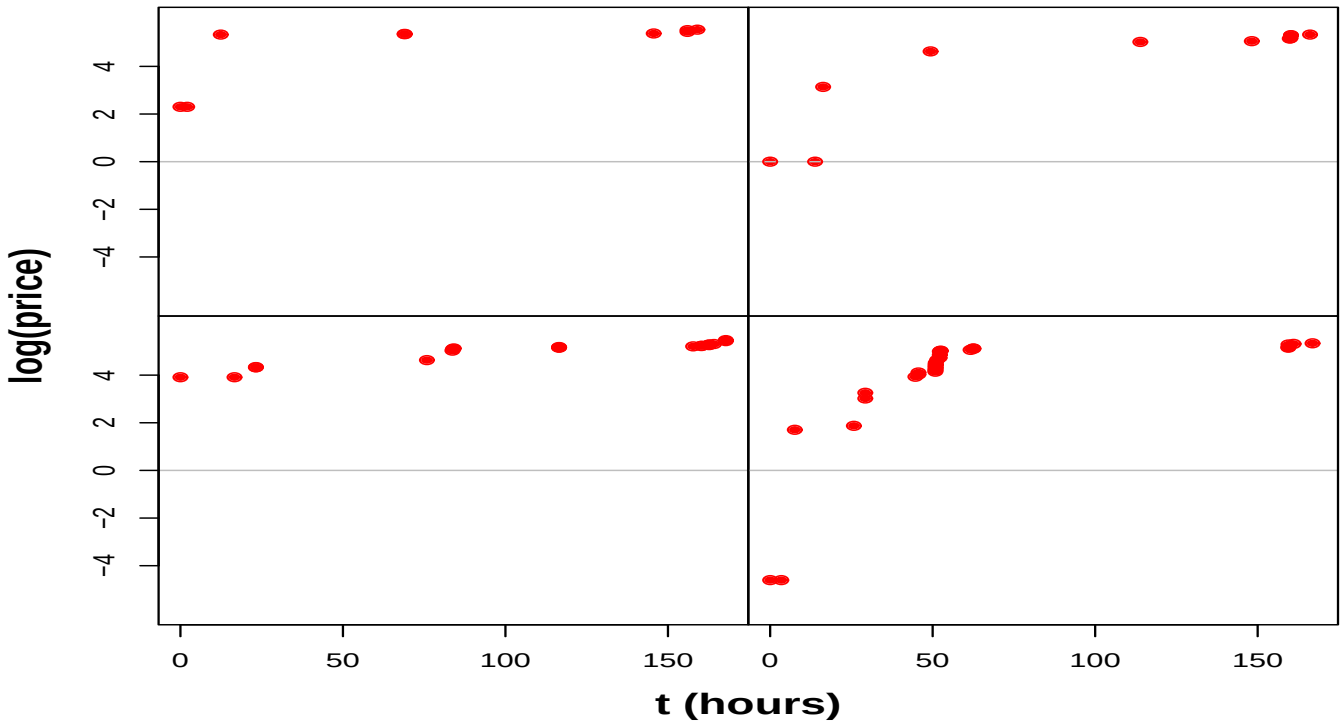
$$f(X) \equiv g(A_1, A_2, \ldots)$$

FUNCTIONAL DATA DESIGNS

- Fully observed functions without noise at arbitrarily dense grid
  Measurements $Y_{it} = X_i(t)$ available for all $t \in T$,
  $i = 1, \ldots, n$:
  Often unrealistic but mathematically convenient

- Dense design with noisy measurements
  Measurements $Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij}$, where T_{ij} are recorded on a regular grid, $T_{i1}, \ldots, T_{iN_i}$, and $N_i \to \infty$:
  Applies to typical functional data

- Sparse design with noisy measurements = Longitudinal data
  Measurements $Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij}$, where T_{ij} are random times and their number N_i per subject is random and finite.

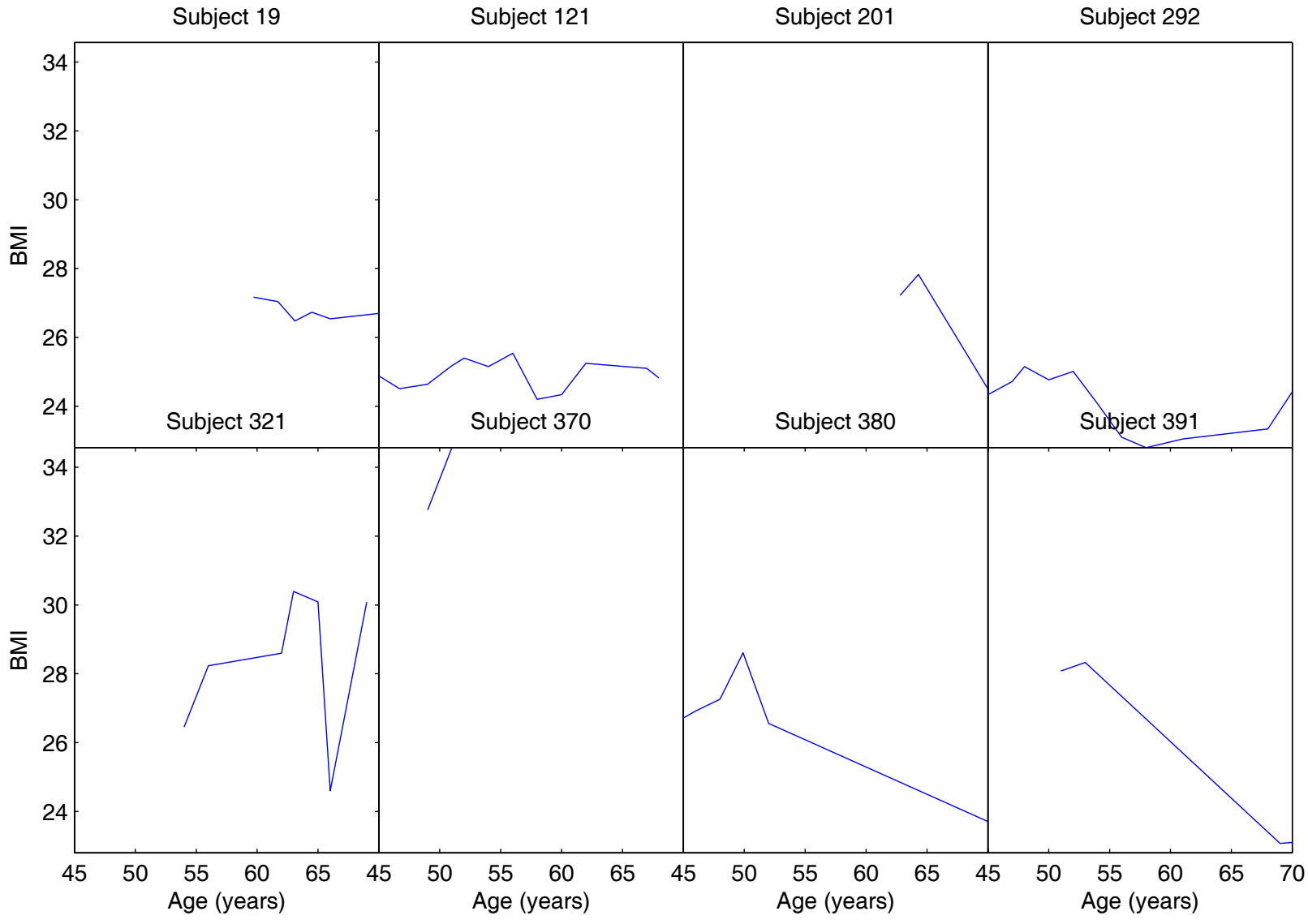


Four eBay auctions: willing-to-pay prices (log-transformed) recorded against time (in hours). Selected from 156 same-item auctions – data from W. Jank

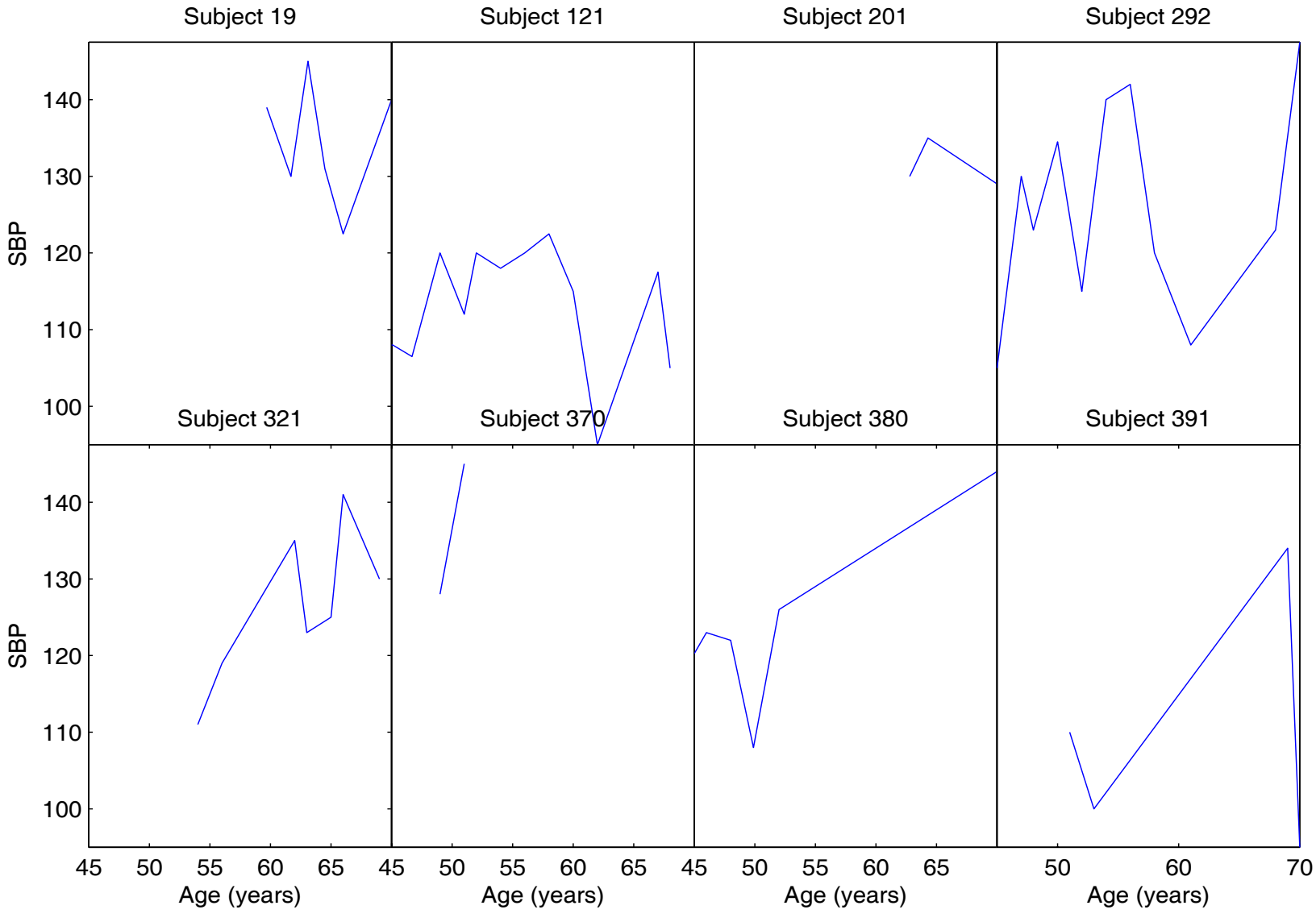
BALTIMORE LONGITUDINAL STUDY ON AGING

- Subset of $n = 507$ males whose Body Mass Index (BMI) and Systolic Blood Pressure (SBP) were measured at least twice between ages 45 and 70 and who survived beyond age 70.

- Measurements are both noisy and spaced irregularly, with both the measurement times and the number of available measurements varying from subject to subject.



Observations of BMI for eight randomly selected subjects



Observations of SBP for eight randomly selected subjects

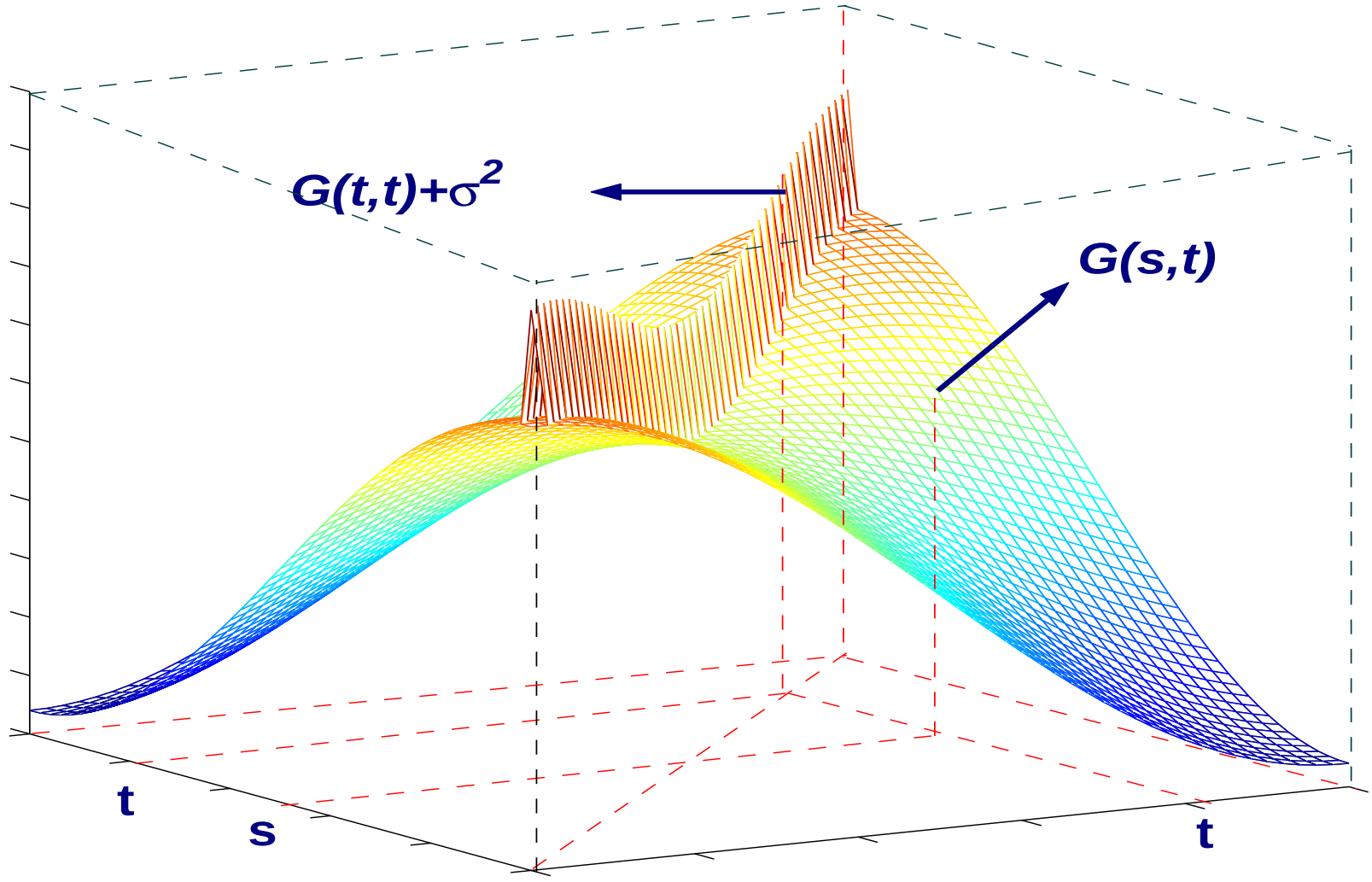
PACE

Principal Analysis by Conditional Expectation (Yao, M, Wang 2005ab, Liu & M 2009) to obtain components of the functional principal component representation for all of these designs.

Idea: Borrowing strength from entire sample for estimation of individual trajectories

Implementation steps:

- Mean function: Smoothing across all pooled observations
- Covariance surface: Pooling products for pairs of observations from the same subject, then smoothing – denoising is achieved by separating out the diagonal (Staniswalis & Lee 1998)



Relationship between the covariance surface and variances on the diagonal: Decomposing diagonal into error and covariance components.

IMPLEMENTATION ISSUES

- Obtain eigenvalues/eigenfunctions:

  For k-th eigenvalue/eigenfunction pair (λ_k, ϕ_k) use discretized versions of eigenequations,

$$\int_0^T \text{cov}(X(s), X(t))\phi_k(s)ds = \lambda_k \phi_k(t),$$

  s.t. $\int_0^T \phi_k(t)^2 dt = 1, \quad \int_0^T \phi_k(t)\phi_m(t)dt = 0, \quad m \neq k,$
  substituting smoothed estimates for the covariance surface.

- Project initial smoothed covariance estimates on space of non-negative definite covariance matrices: (Hall, M, Yao 2008)

$$\hat{\text{cov}}(X(s), X(t)) = \sum_{k=1, \hat{\lambda}_k > 0}^{K} \hat{\lambda}_k \hat{\phi}_k(s)\hat{\phi}_k(t).$$

- Obtain Functional principal components (the random effects):
  - Conditioning $E(A_k|U_i)$, where U_i is the vector of available data for the i-th subject (random dimension)
  - Best linear predictor for conditional expectation (best predictor under Gaussian assumptions)
  - Substitute estimates for eigenvalues, eigenfunctions, covariances
  - Regularization for inverses of cova matrices at random locations
  - Choice of regularization parameters (number of included components, smoothing parameters: GCV, FVE, BIC,...)
- Implementation of FPCA and functional regression models: PACE 2.16 at: http://anson.ucdavis.edu/∼mueller/data/programs.html

ESTIMATING DERIVATIVES FROM SPARSE DATA

Differentiating Karhunen-Loève representation:

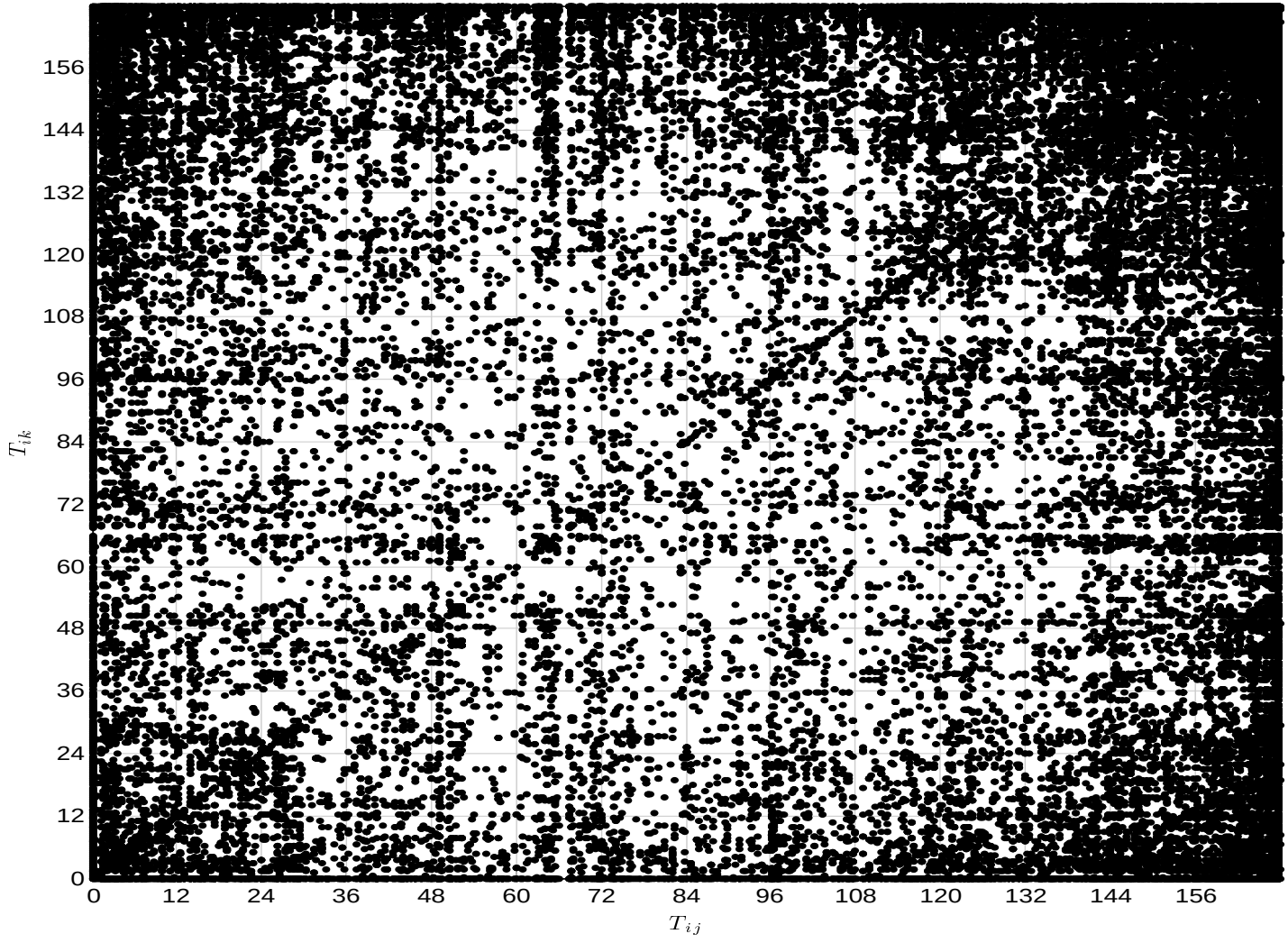
$$X_i^{(\nu)}(t) = \mu^{(\nu)}(t) + \sum_{k=1}^{\infty} A_{ik}\phi_k^{(\nu)}(t), \quad \nu = 0, 1, \ldots.$$

- 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

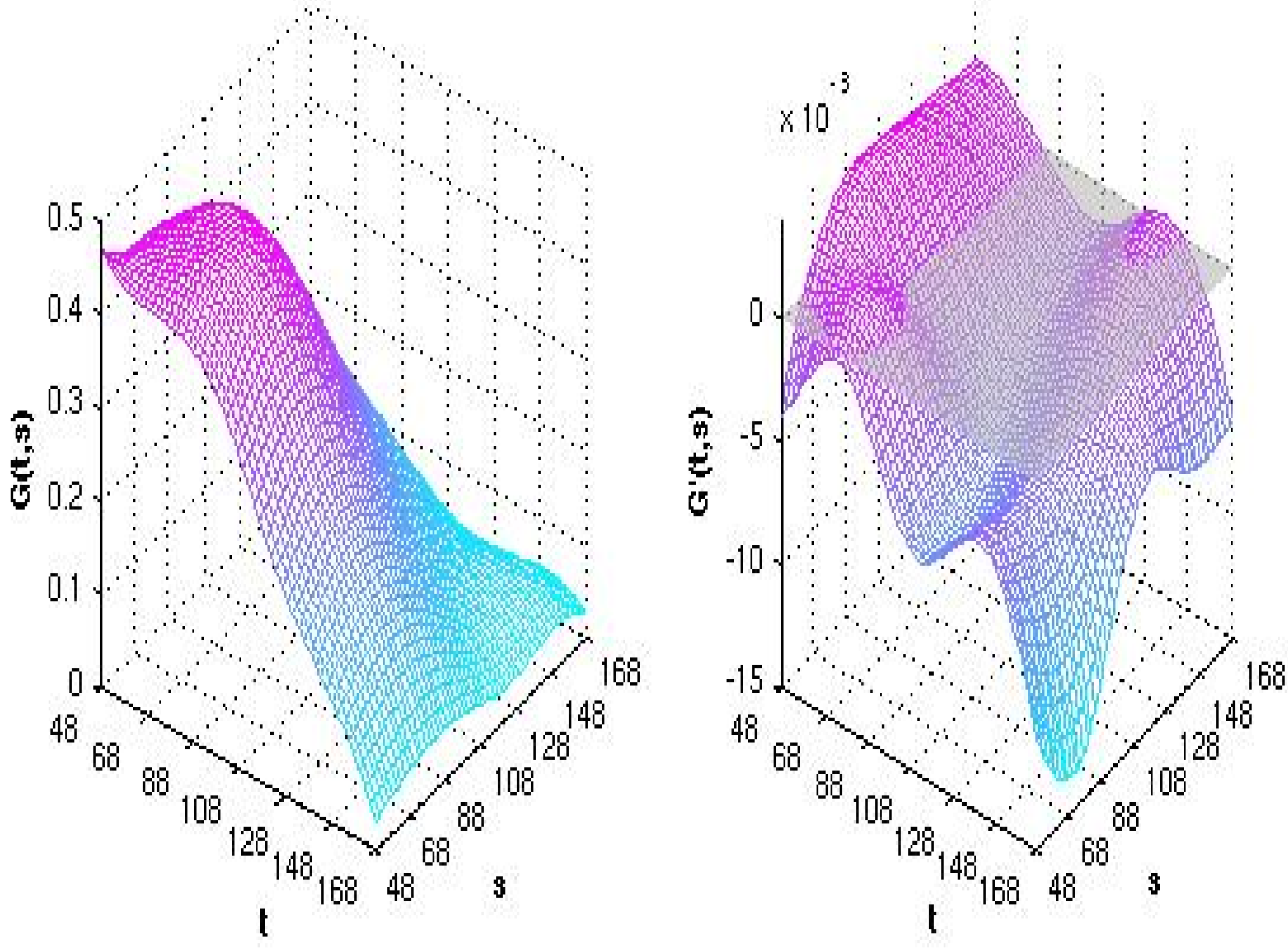
$$\frac{d^\nu}{dt^\nu}\int_{\mathcal{T}} G(t,s)\phi_k(s)ds = \lambda_k \frac{d^\nu}{dt^\nu}\phi_k(t),$$

implying

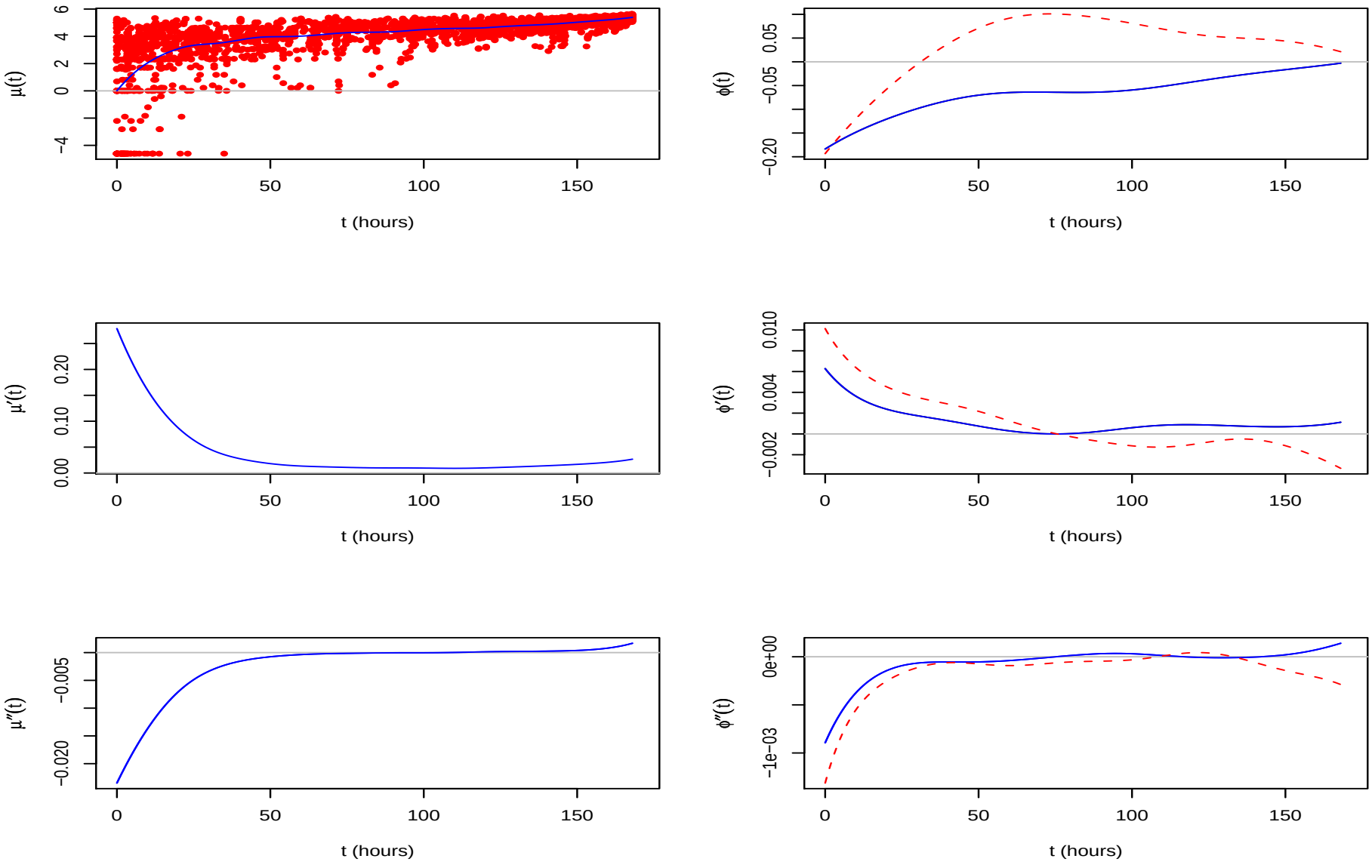
$$\phi_k^{(\nu)}(t) = \frac{1}{\lambda_k}\int_{\mathcal{T}} \frac{\partial^\nu}{\partial t^\nu} G(t,s)\phi_k(s)ds.$$



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.

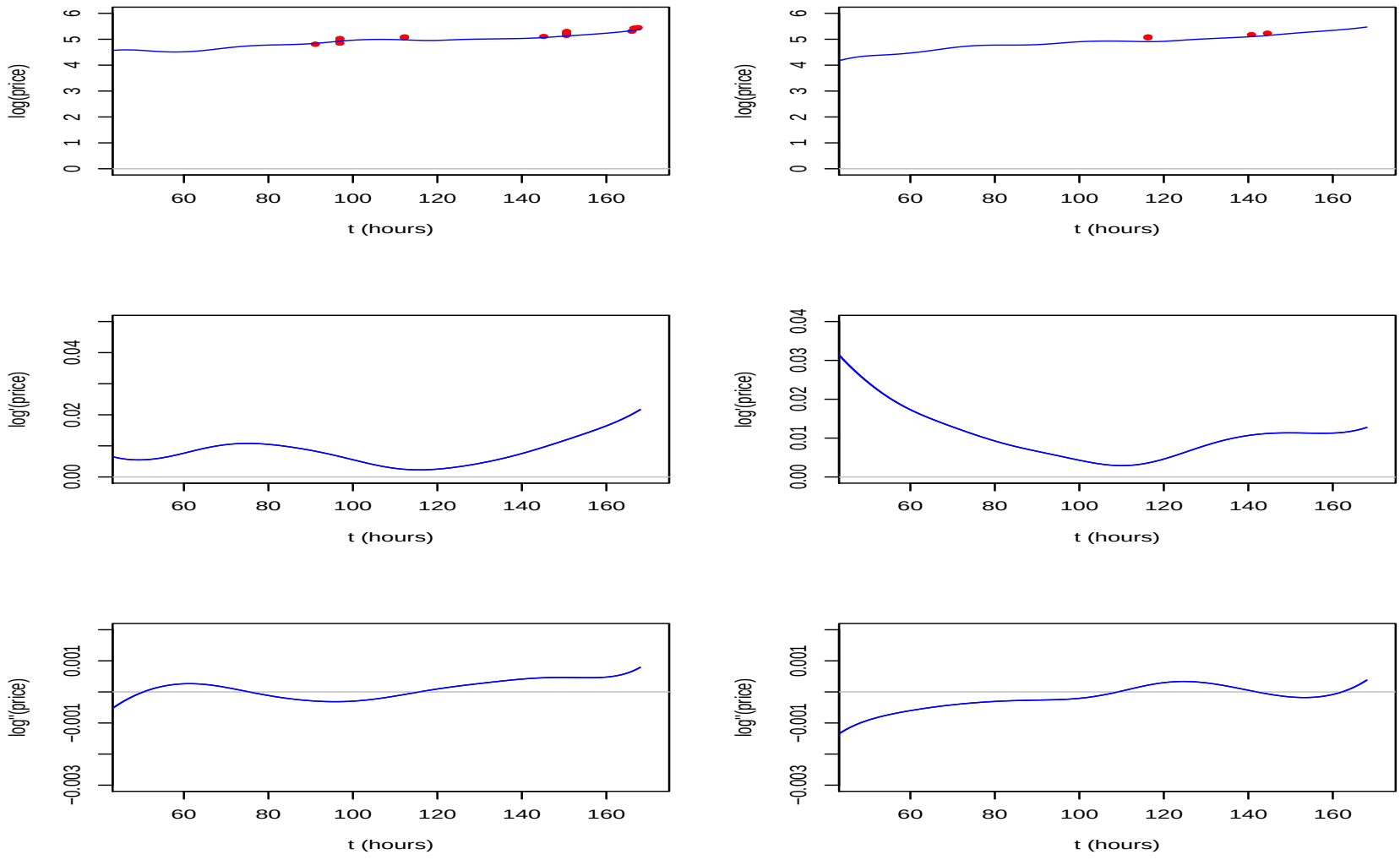
DERIVATIVES OF TRAJECTORIES

- 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

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

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

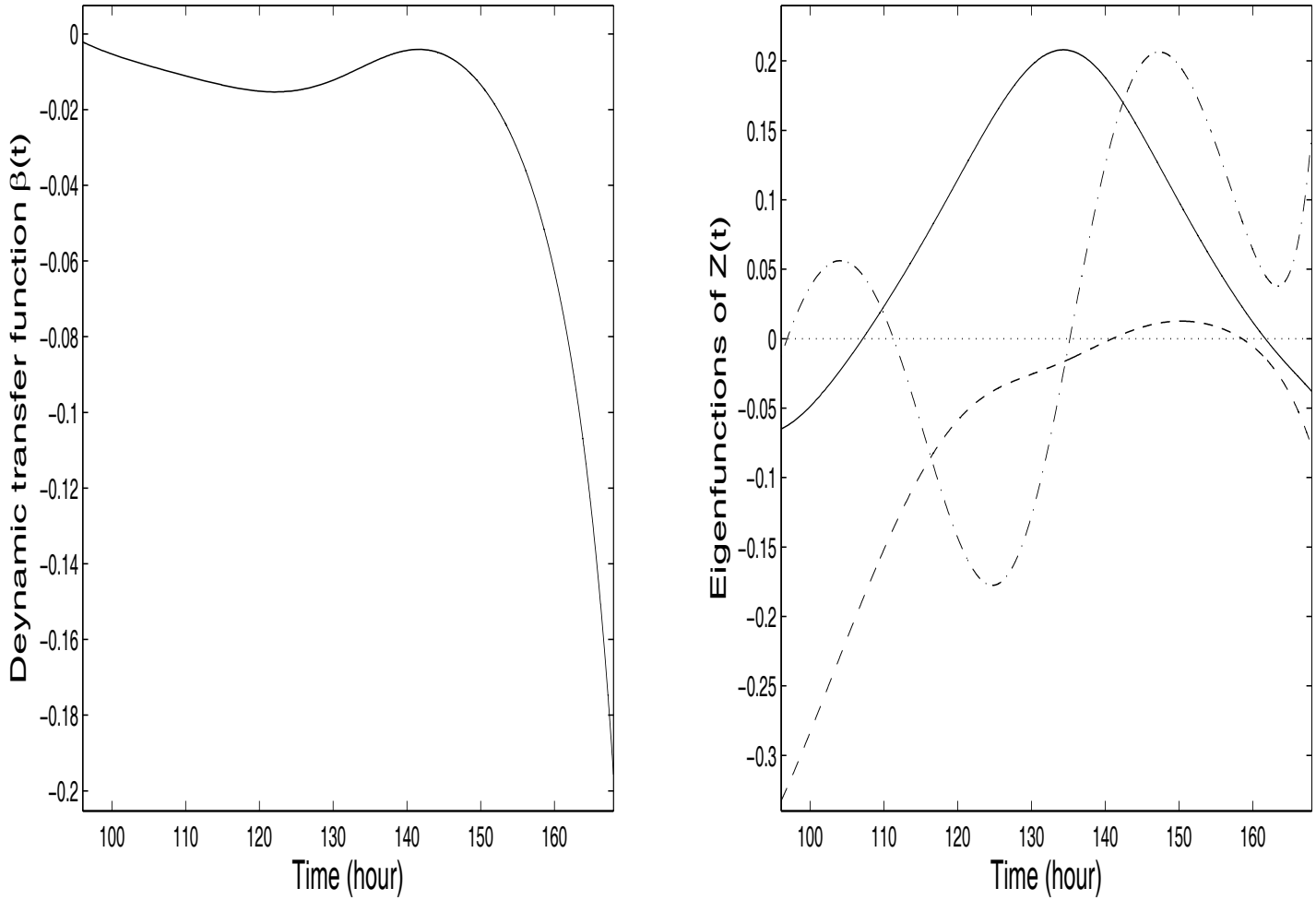
$$\mathrm{var}(Z(t)) =$$
$$\left(\sum_k \lambda_k(\phi_k^{(1)}(t))^2 \sum_k \lambda_k \phi_k^2(t) - \{\sum_{k=1}^{\infty} \lambda_k \phi_k^{(1)}(t)\phi_k(t)\}^2\right) / \sum_k \lambda_k \phi_k^2(t)$$

and

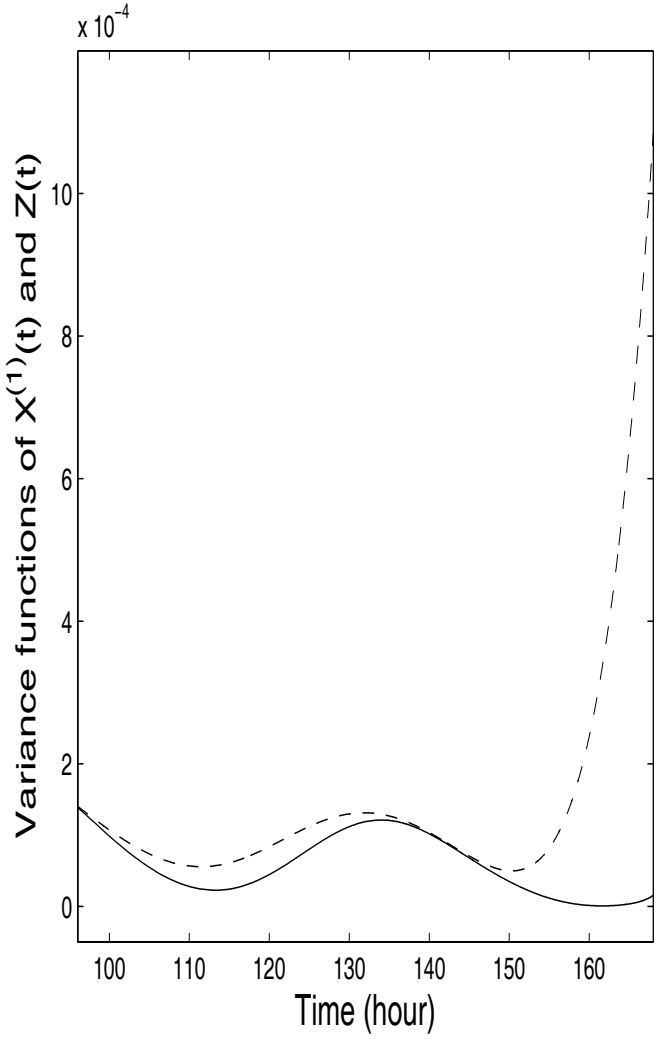
$$\mathrm{var}\{X^{(1)}(t)\} = \beta(t)^2 \mathrm{var}\{X(t)\} + \mathrm{var}\{Z(t)\}.$$

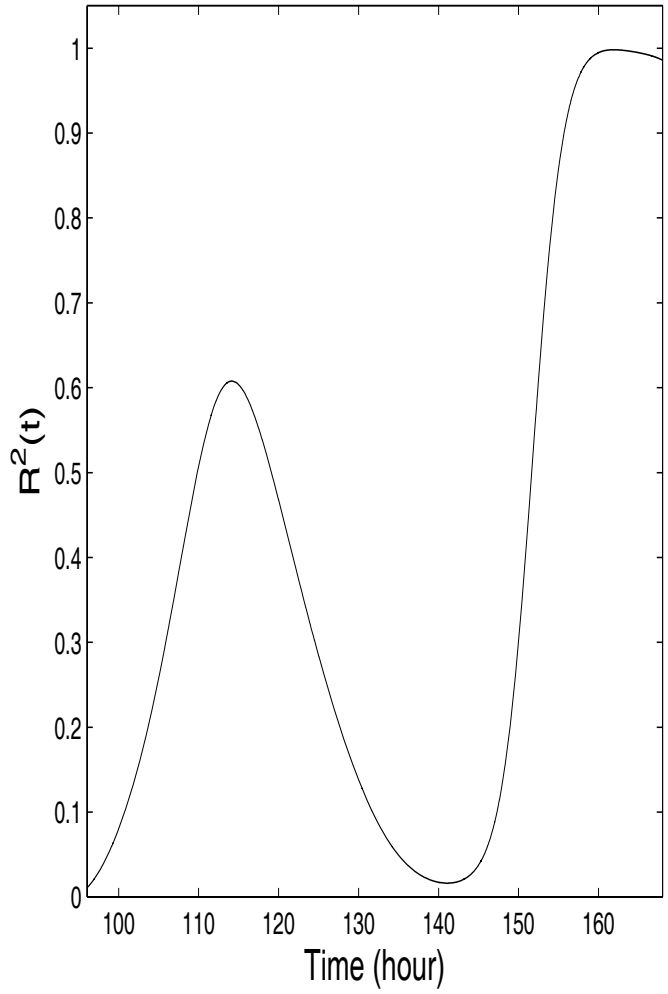
Then the fraction of the variance of $X^{(1)}(t)$ explained by the deterministic part of the differential equation is given by:

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

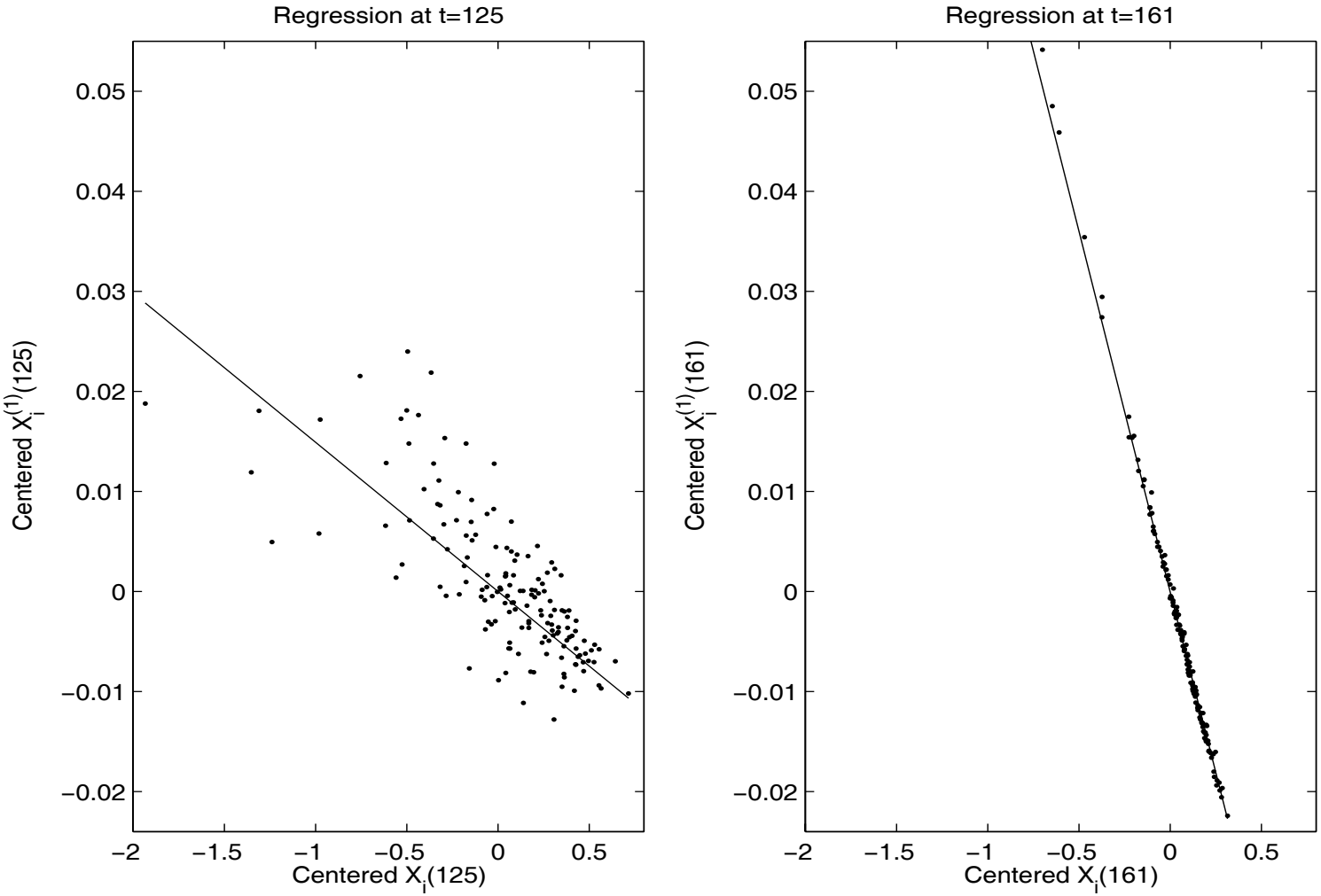


Left: Smooth estimate of the dynamic varying coefficient function β for auction data. Right: Smooth estimates of the first (solid), second (dashed) and third (dash-dotted) eigenfunction of drift process Z.





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.

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

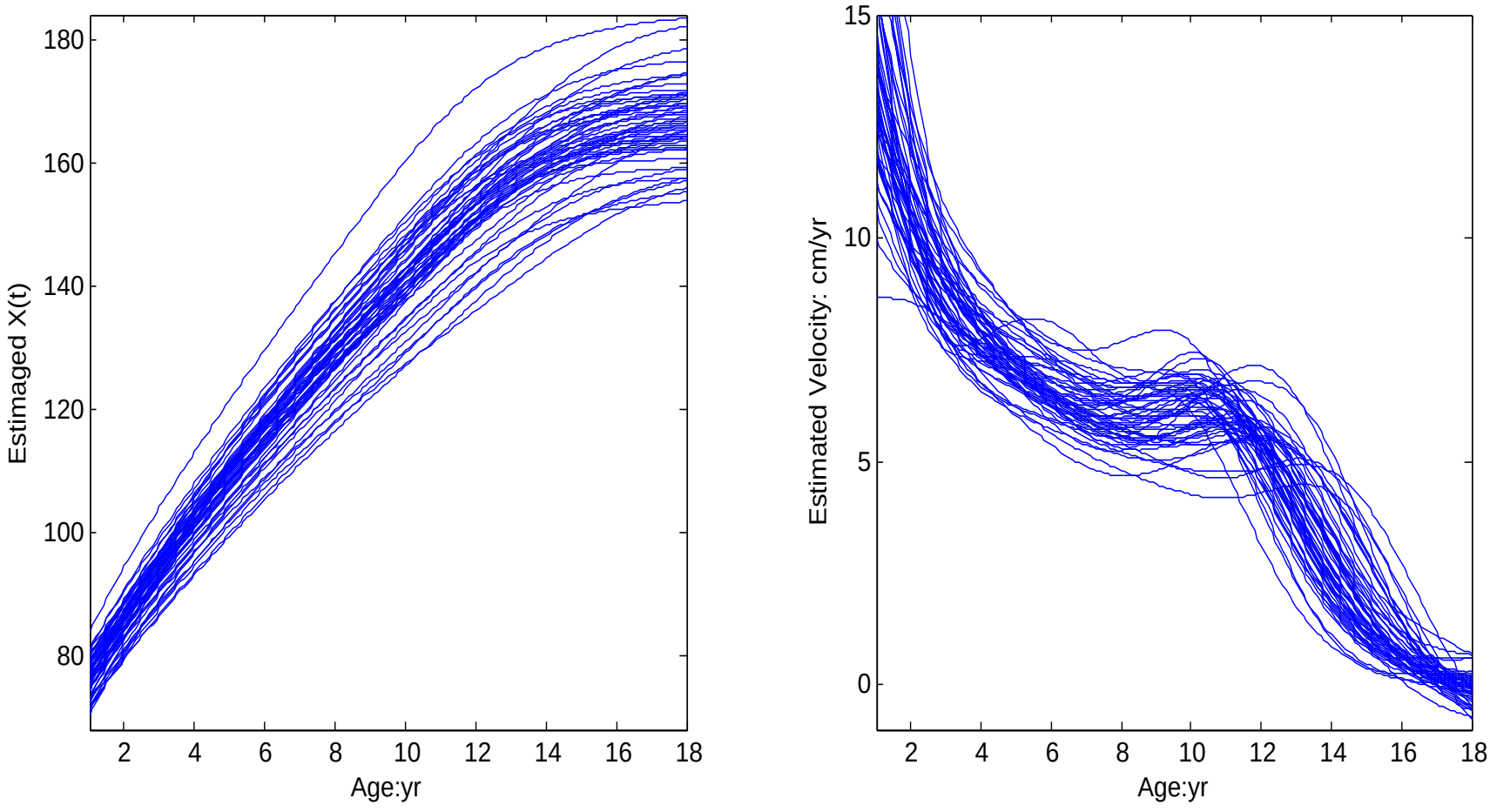
- Parametric Dynamics. Parametric differential equations

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

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 $X'(t)$:

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

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

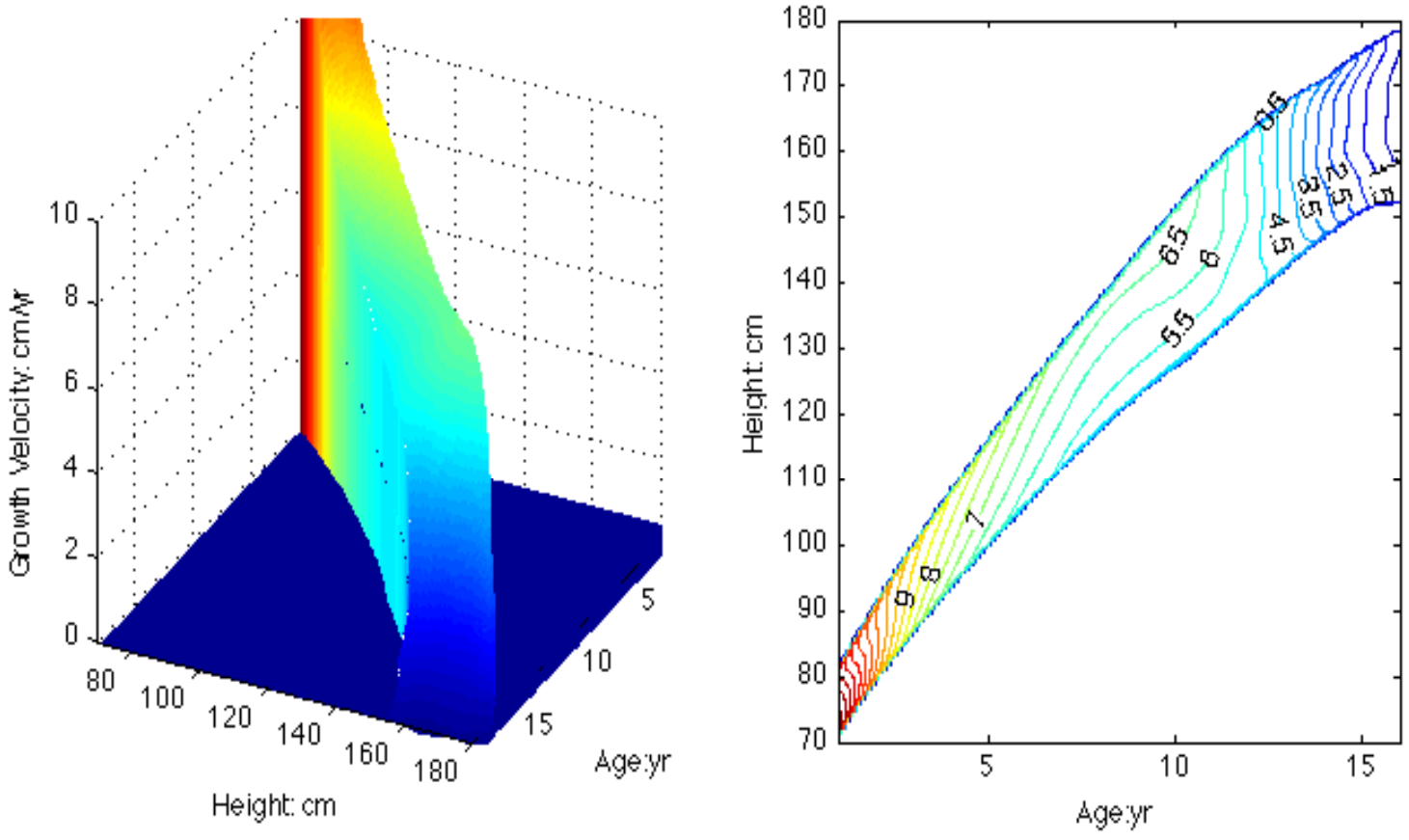
  where $s_j = (t_{ij} + t_{i,j+1})/2$ and $h_X > 0$ and $h_{X'} > 0$ are smoothing bandwidths.

- 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 $\hat{f}(t, x)$.

DECOMPOSING VARIANCE

- Since $\mathrm{var}\{X'(t)\} = \mathrm{var}[f\{t, X(t)\}] + \mathrm{var}\{Z(t)\}$, on subdomains where the variance of the drift process $\mathrm{var}\{Z(t)\}$ is small, the deterministic approximation

$$X'(t) = f\{t, X(t)\} \quad (t \in \mathcal{T}),$$

  is reasonable. Then future changes of individual trajectories are easily predictable.

- Fraction of the variance of $X'(t)$ that is explained by the deterministic part

$$R^2(t) = \frac{\mathrm{var}[f\{t, X(t)\}]}{\mathrm{var}\{X'(t)\}} = 1 - \frac{\mathrm{var}\{Z(t)\}}{\mathrm{var}\{X'(t)\}} \ .$$

- Quantify predictability by

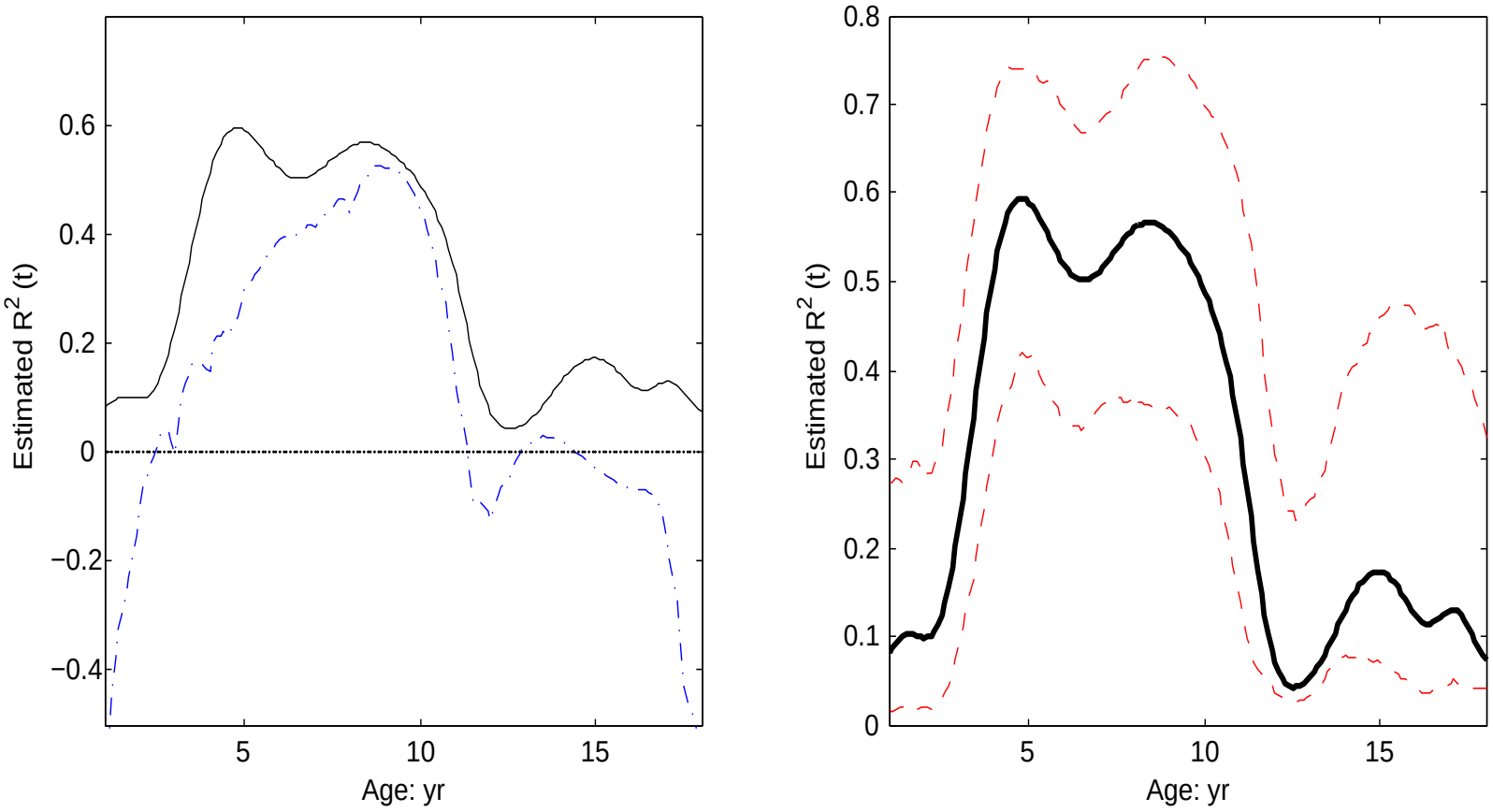
$$S(t,x) = \frac{f^2(t,x)}{E\{X'^2(t) \mid X(t)=x\}} = \frac{f^2(t,x)}{f^2(t,x) + \mathrm{var}\{Z(t) \mid X(t)=x\}}$$

When $S(t,x)$ is close to one, then $f^2(t,x)$ is large compared to $\mathrm{var}\{Z(t) \mid X(t)=x\}$ and the process is well predictable when $X(t)=x$.

- Diagnostics for linearity. For the coefficient of determination for the linear dynamic model

$$R_L^2(t) = \frac{\mathrm{var}\,\{\beta(t)X(t)\}}{\mathrm{var}\{X'(t)\}}$$

one expects that $R^2(t) \geq R_L^2(t)$ On subdomains of $\mathcal{T}$ where $R(t)$ is close to $R_L(t)$, one may infer that the data-driven differential equation is reasonably linear.



Left panel: Estimated coefficients of determination $\widehat{R}^2(t)$, corresponding to the fraction of variance explained by the deterministic part of the nonlinear dynamic model (solid), in comparison with the corresponding fractions of variance $\widehat{R}^2_L(t)$ explained by linear dynamics (dot-dashed). Right panel: 95% bootstrap confidence interval for $R^2(t)$.

- 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 $\beta(t)$,

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

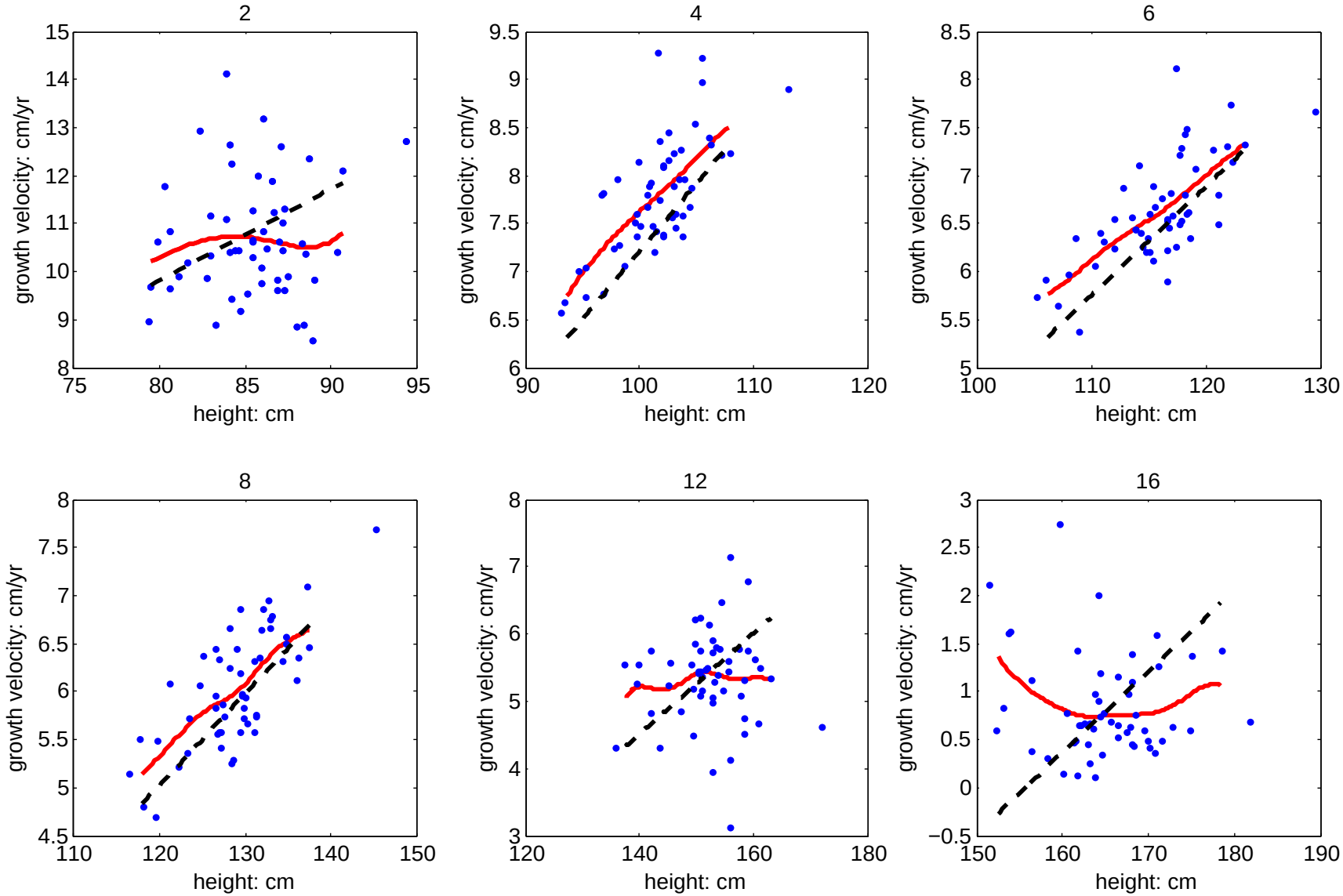
where $Z_2(t)$ is a zero mean drift process with $\mathrm{cov}\{Z_2(t), X(t)\} = 0$.

- 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.

- Learning Gaussian dynamics works for sparse data, learning non-Gaussian dynamics is viable only for dense data



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 $(x(t), x^{(1)}(t))$.