Functional Data Analysis with PACE

Kehui Chen

Department of Statistics,

University of California, Davis

JSM, 2012

Outline

- General introduction of PACE
- Illustrative examples for various functional regression programs

Overview of PACE

- Implements various methods of Functional Data Analysis (FDA).
- Provides analysis for sparsely or densely sampled random trajectories and time courses.
- The core program is based on the Principal Analysis by Conditional Expectation (PACE) algorithm.
- The most updated version is PACE 2.15, written in Matlab, along with an R version in development.

Development of PACE

- Supported by various *NSF* grants.
- Coordinated by Hans-Georg Müller and Jane-Ling Wang.
- *PACE* 1.0 was written by Fang Yao in 2005, and subsequent major improvements were made by Bitao Liu.
- Contributors and developers include (alphabetical order):

  Dong Chen, Kehui Chen, Jeng-Min Chiou, Joel Dubin, Andrew Farris, Andrea Gottlieb, Jinjiang He, Ci-Ren Jiang, Yu-Ru Su, Rona Tang, Wenwen Tao, Shuang Wu, Cong Xu, Matt Yang, Wenjing Yang, Xiaoke Zhang.

Functional Principal Component Analysis

- $X(t)$ is a second order random process,
  mean function $\mu(t) \in L^2(\mathscr{T})$,
  continuous covariance function $G(s,t) = \mathrm{cov}(X(s), X(t))$.
- $G(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t)$, eigenvalues $\lambda_1 \geq \lambda_2, \cdots, \lambda_k, \cdots \geq 0$,
  eigenfunctions $\phi_k(t)$ form an orthogonal basis.
- Karhunen-Loève expansion

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

- Best linear expansion with p components:

$$X(t) \approx \mu(t) + \sum_{k=1}^{p} \xi_k \phi_k(t).$$

Dense and Sparse Designs

- Very densely and regularly observed data: empirical mean and covariance, and $\xi_k = \int_{\mathcal{T}} (X(t) - \mu(t))\phi_k(t)dt$.
- Densely recorded but irregular design, or contaminated with error: pre-smoothing for individual curves.
- Sparse random design (longitudinal data): pre-smoothing is problematic.
- PACE works for both dense and sparse data.

The Core Program FPCA

- Pool all the sample $Y_{ij} = X_i(t_{ij}) + \varepsilon_{ij}$, $1 \le i \le n, 1 \le j \le m_i$, and estimate mean and covariance by local linear smoothing. One (two) dimensional nonparametric rate for sparse data, and $\sqrt{n}$ rate for dense data.

- Conditional expectation method to estimate the components ξ_{ik}. For sparse case, best linear unbiased prediction; for dense data, it is asymptotically equivalent to the numerical approximation of $\xi_{ik} = \int_{\mathcal{T}} (X_i(t) - \mu(t))\phi_k(t)dt$.

- Yao et al. (2005), Hall et al. (2006), Li and Hsing (2010), Cai and Yuan (2010).

Local Linear Smoothing Estimators

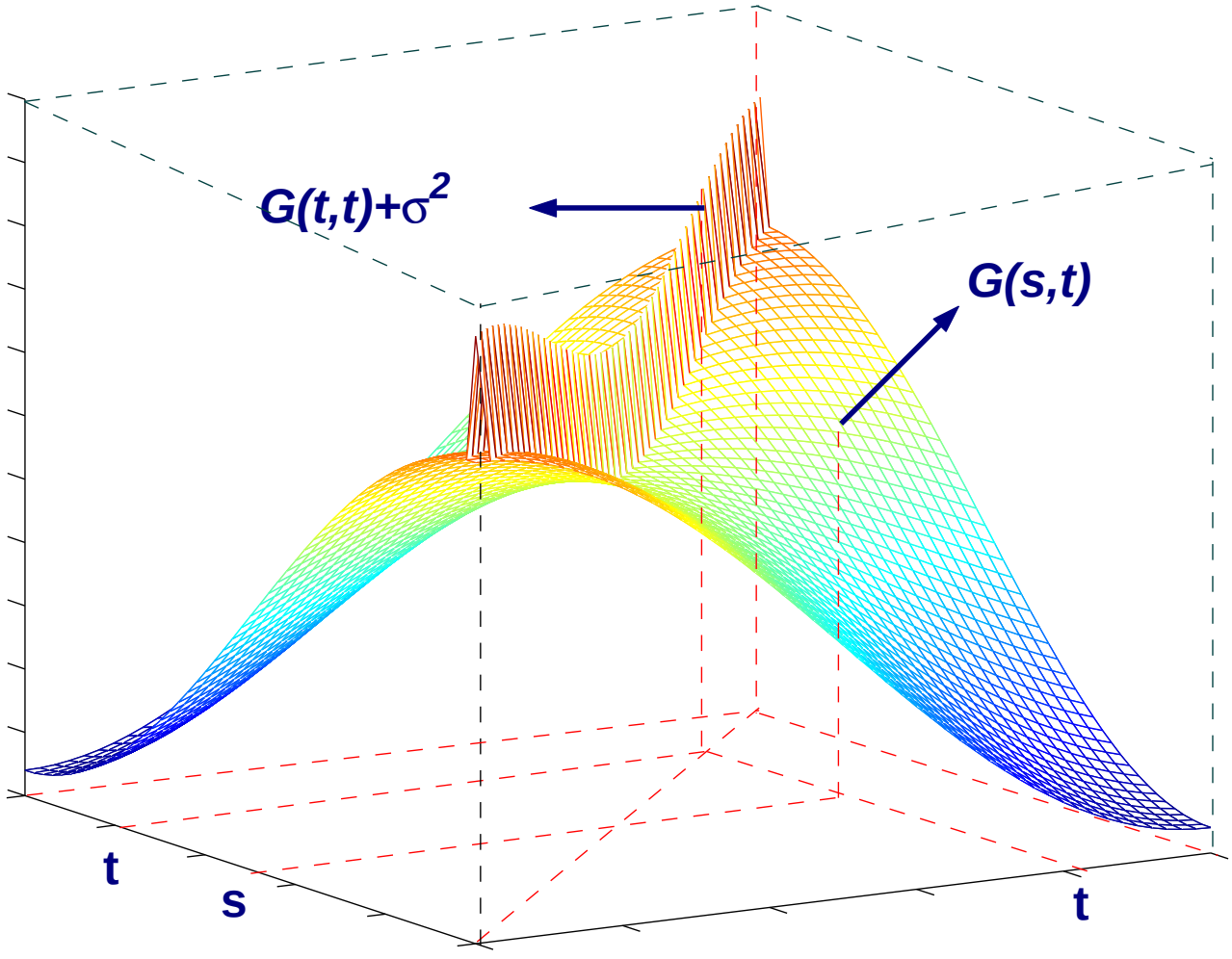
- Mean function is given by $\hat{\mu}(t) = \hat{a}_0$, where

$$(\hat{a}_0, \hat{a}_1) = \arg\min \sum_{i=1}^{n} \sum_{j=1}^{m_i} \{[Y_{ij} - a_0 - a_1(t_{ij} - t)]^2 \times K_h(t_{ij} - t)\}.$$

- Covariance function is given by $\hat{G}(t_1, t_2) = \hat{a}_0$, where

$$(\hat{a}_0, \hat{a}_1, \hat{a}_2) = \arg\min \sum_{i=1}^{n} \sum_{j \neq l} \{[Y_{ij}^c Y_{il}^c - a_0 - a_1(t_{ij} - t_1) - a_2(t_{il} - t_2)]^2 \times K_b(t_{ij} - t_1) K_b(t_{il} - t_2)\}.$$

Covariance Estimation



Principal Analysis by Conditional Expectation

- $\mathbf{X}_i = (X_i(t_{i1}), \ldots, X_i(t_{im_i}))^T$, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$, $\mu_i = (\mu(t_{i1}), \ldots, \mu(t_{im_i}))^T$, $\phi_{ik} = (\phi_k(t_{i1}), \ldots, \phi_k(t_{im_i}))^T$, by Gaussianity

$$\mathrm{E}[\xi_{ik}|\mathbf{Y}_i] = \lambda_k \phi_{ik}^T \Sigma_{\mathbf{Y}_i}^{-1} (\mathbf{Y}_i - \mu_i),$$

  where $\Sigma_{\mathbf{Y}_i} = \mathrm{cov}(\mathbf{Y}_\mathrm{i}, \mathbf{Y}_\mathrm{i}) = \mathrm{cov}(\mathbf{X}_\mathrm{i}, \mathbf{X}_\mathrm{i}) + \sigma^2 \mathbf{I}_{\mathrm{m_i}}$.

- The method is robust and works well for non-Gaussian data.

Functional Regression in PACE

- Linear regression and diagnostics
- Quadratic (Polynomial) regression
- Additive modeling
- Generalized responses
- Quantile and conditional distribution modeling
- Function to scalar; function to function

Illustrative Example: Meat Spectral Data

- *FPCreg*, *FPCdiag*: Let $X^c(t) = X^c(t) - \mu(t)$

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

Illustrative Example: Meat Spectral Data

- *FPCreg*, *FPCdiag*: Let $X^c(t) = X^c(t) - \mu(t)$

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

- *FPCQuadReg*: (Yao and Müller 2010, Horvath and Reeder, 2012)

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

Illustrative Example: Meat Spectral Data

- *FPCreg*, *FPCdiag*: Let $X^c(t) = X^c(t) - \mu(t)$

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

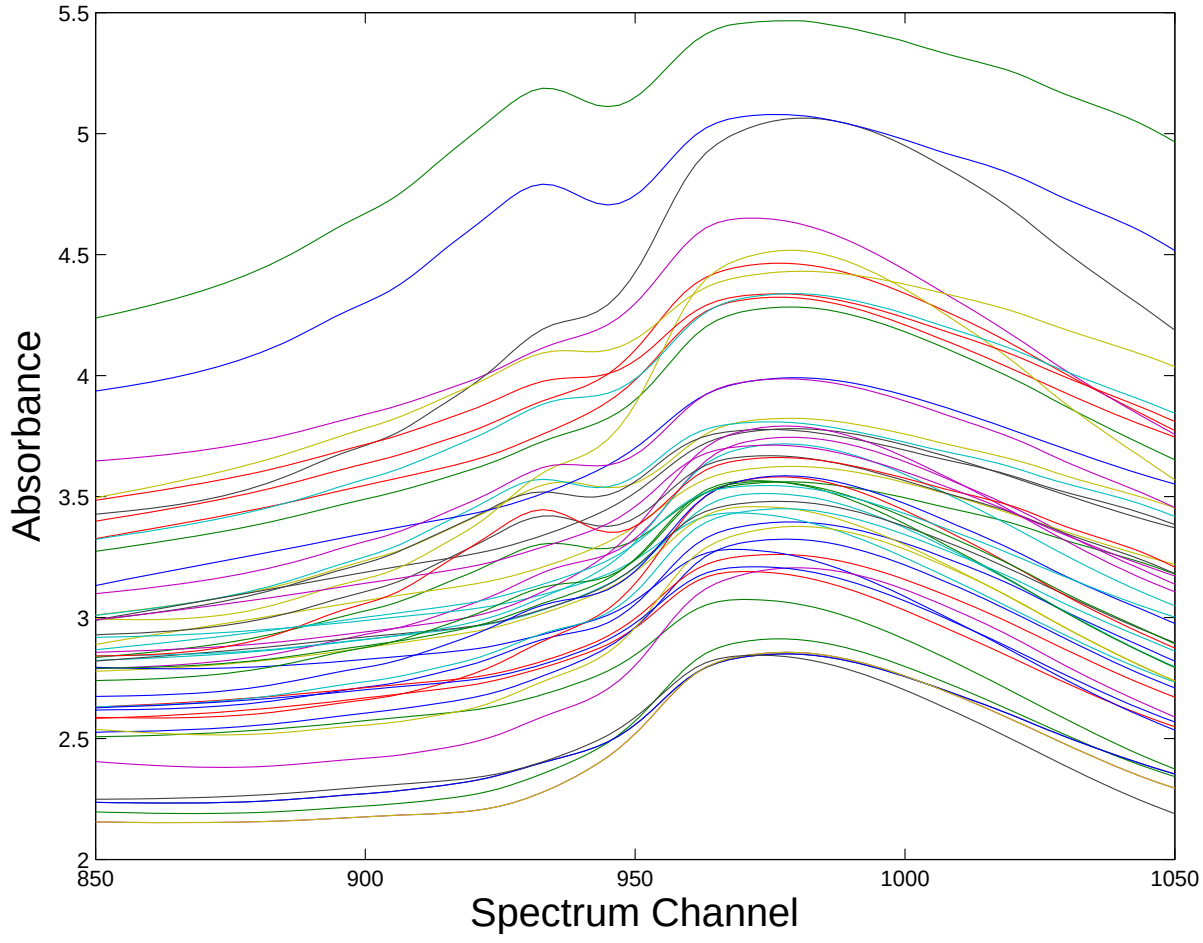
- *FPCQuadReg*: (Yao and Müller 2010, Horvath and Reeder, 2012)

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

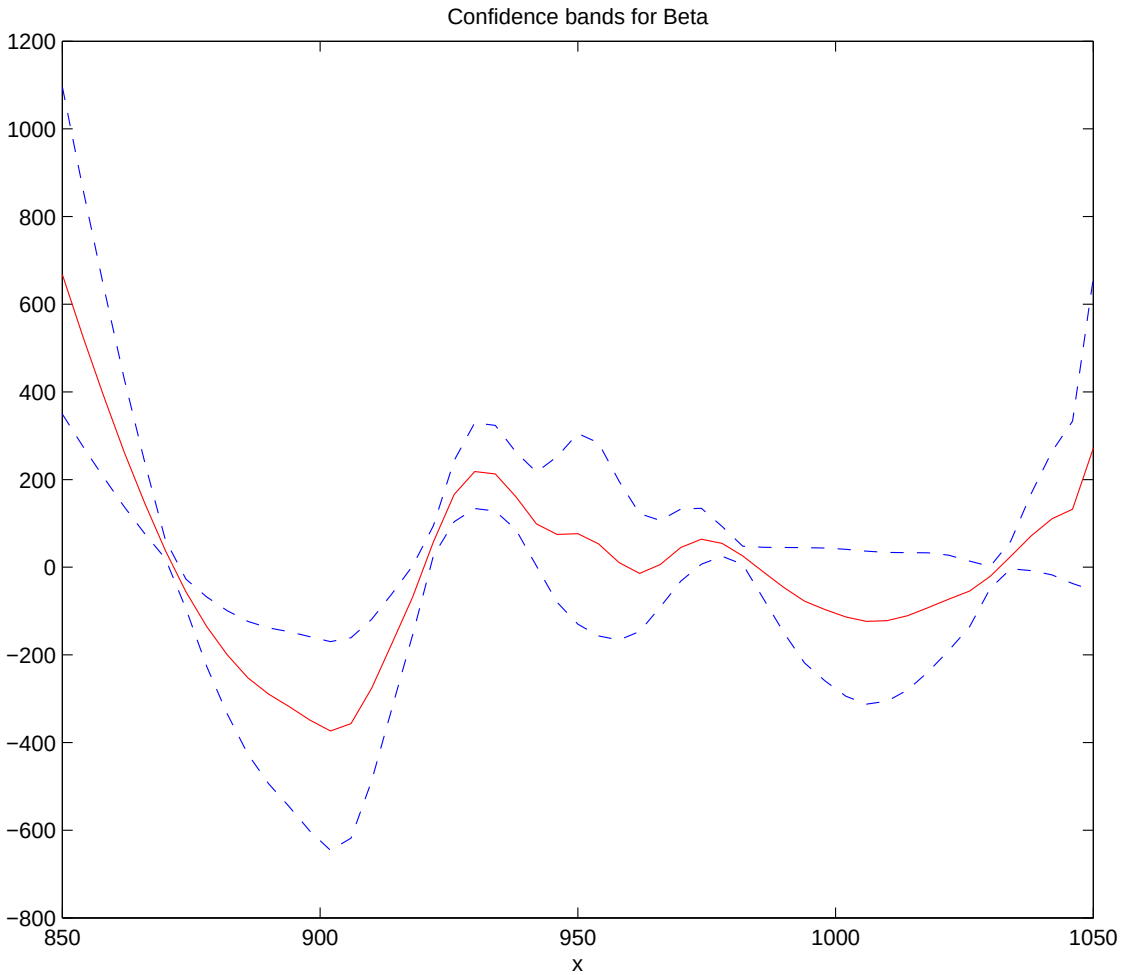
- *FPCquantile* (Chen and Müller 2012. *JRSSB*.)

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

Predictor Functions: Spectral Data

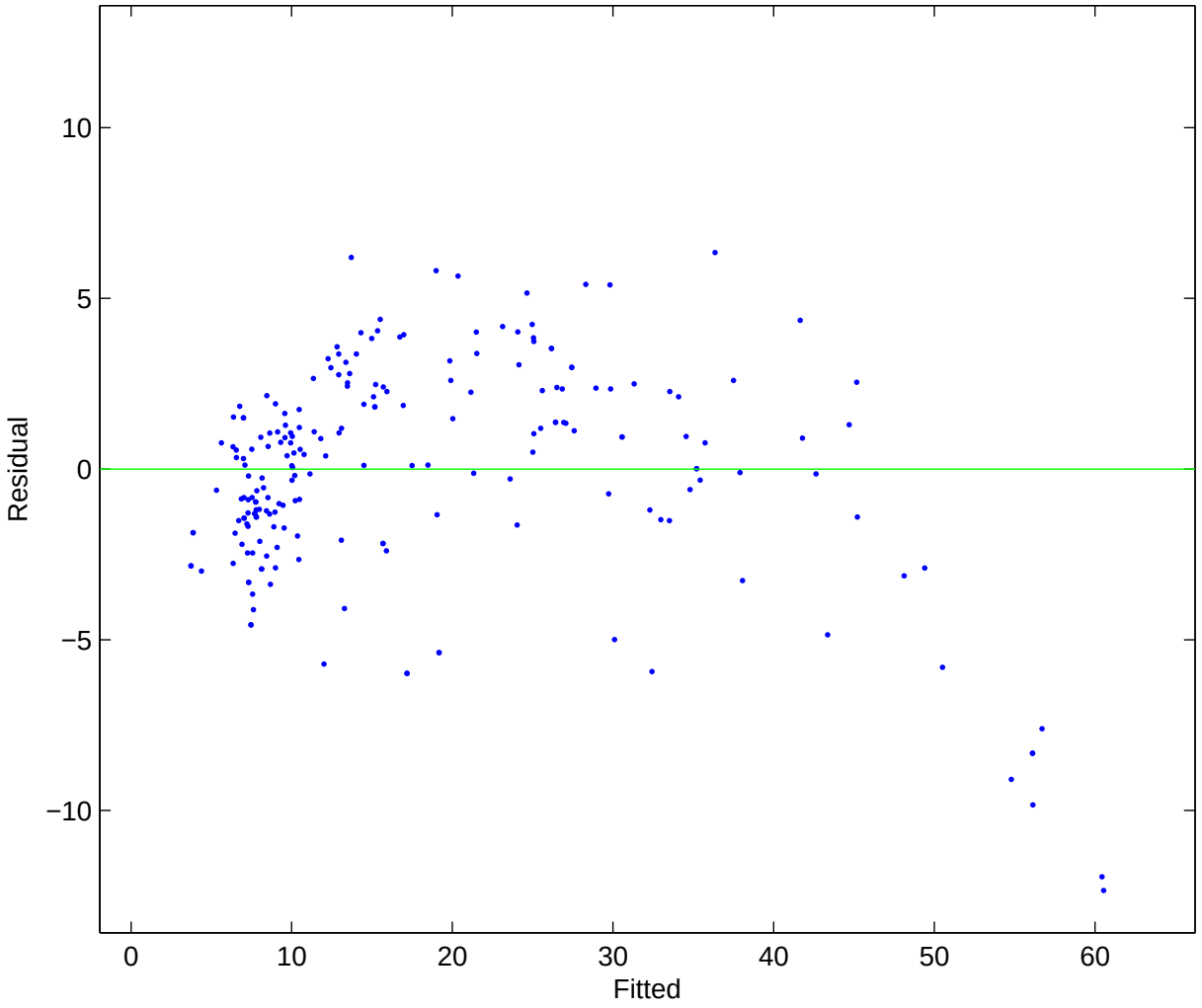


Coefficient of Linear Regression

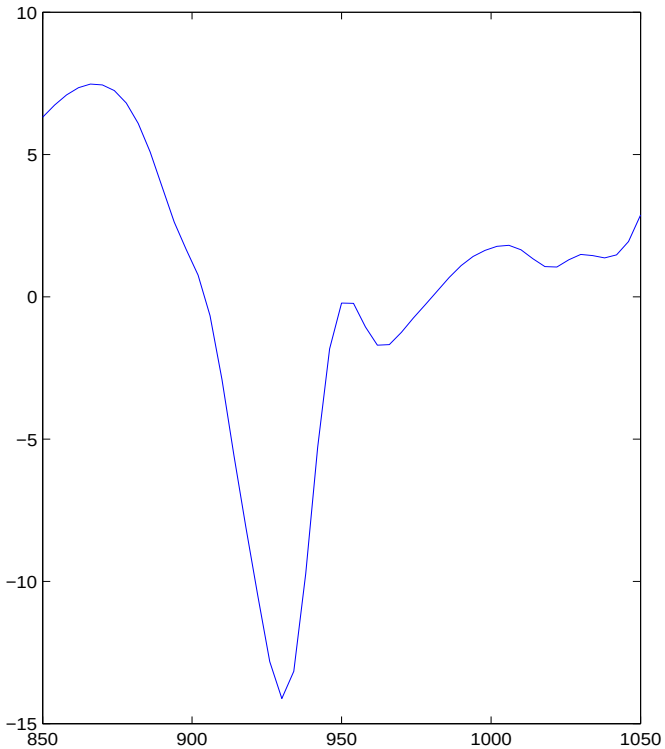


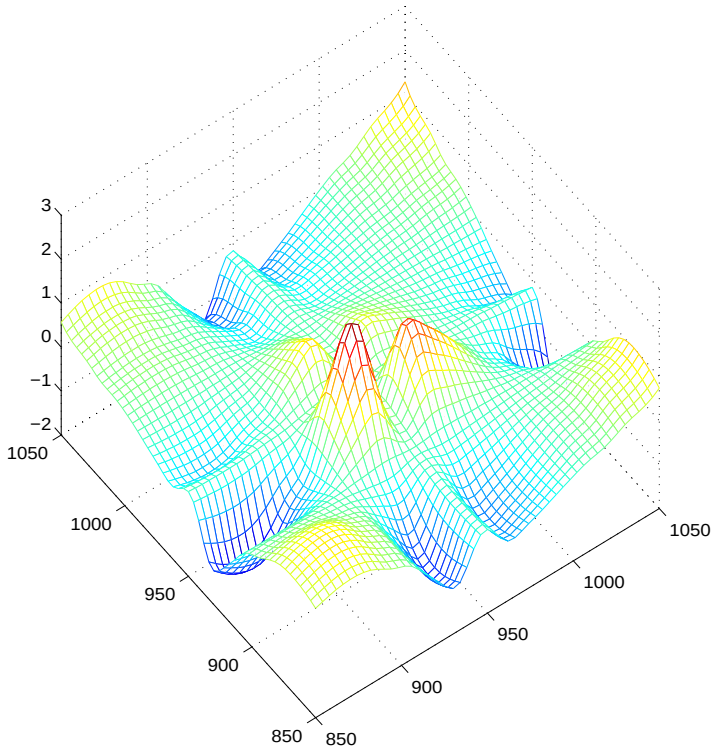
$$E(Y|X) = \alpha + \int X^c(t)\beta(t)dt$$

Residual Plot for Linear Regression



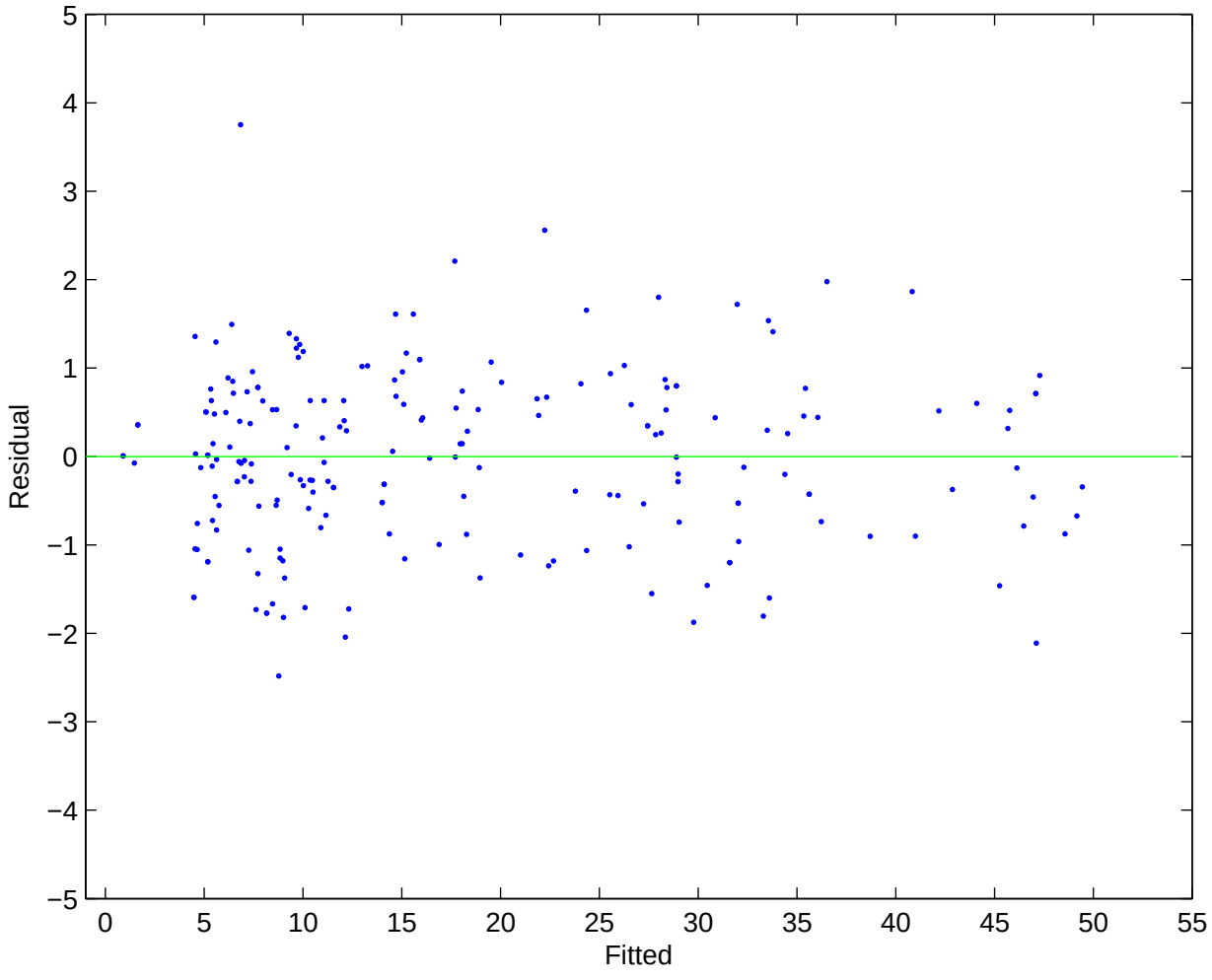
Coefficients of Quadratic Regression



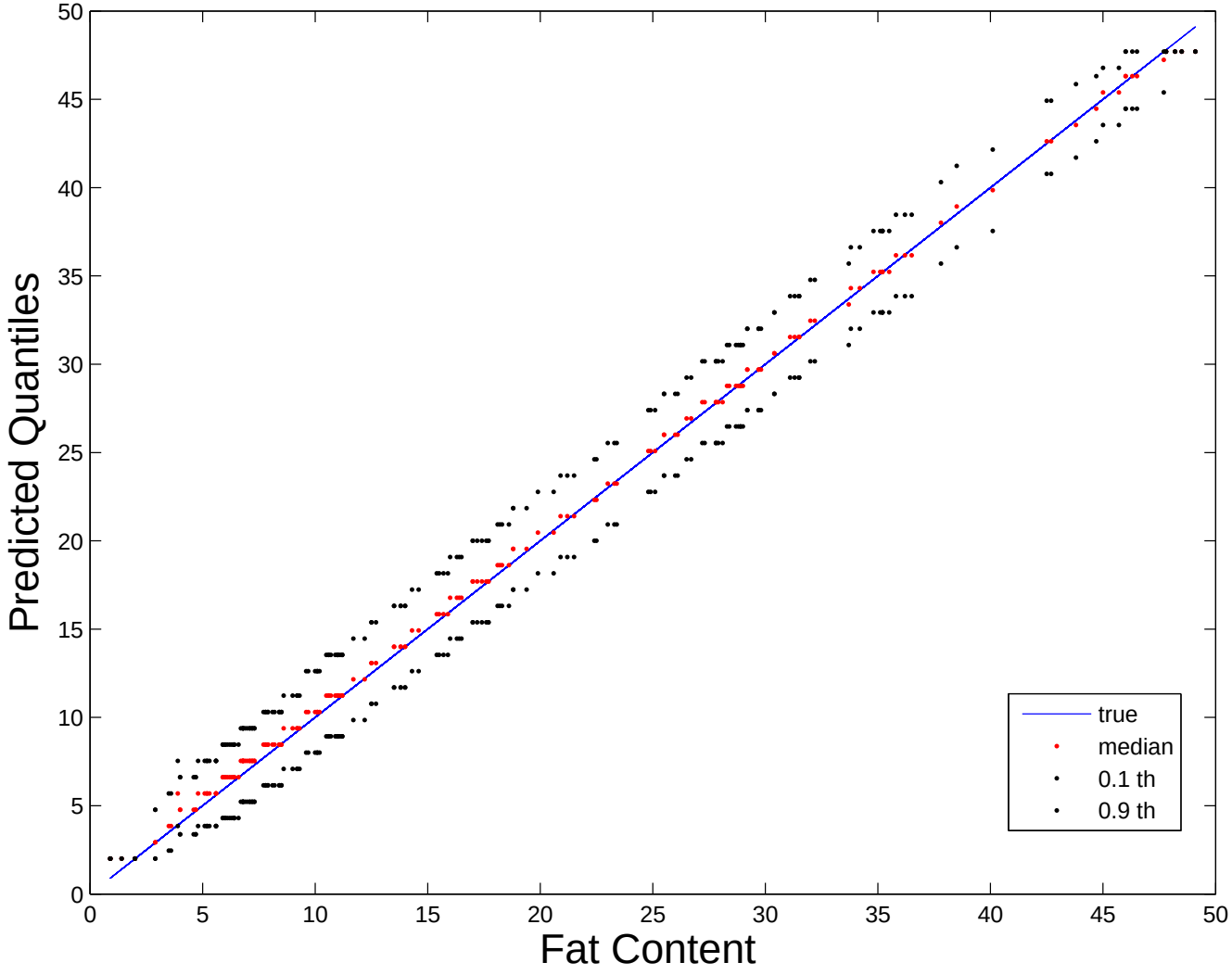


$$E(Y|X) = \alpha + \int X^c(t)\beta(t)dt + \iint \gamma(s,t)X^c(s)X^c(t)dsdt$$

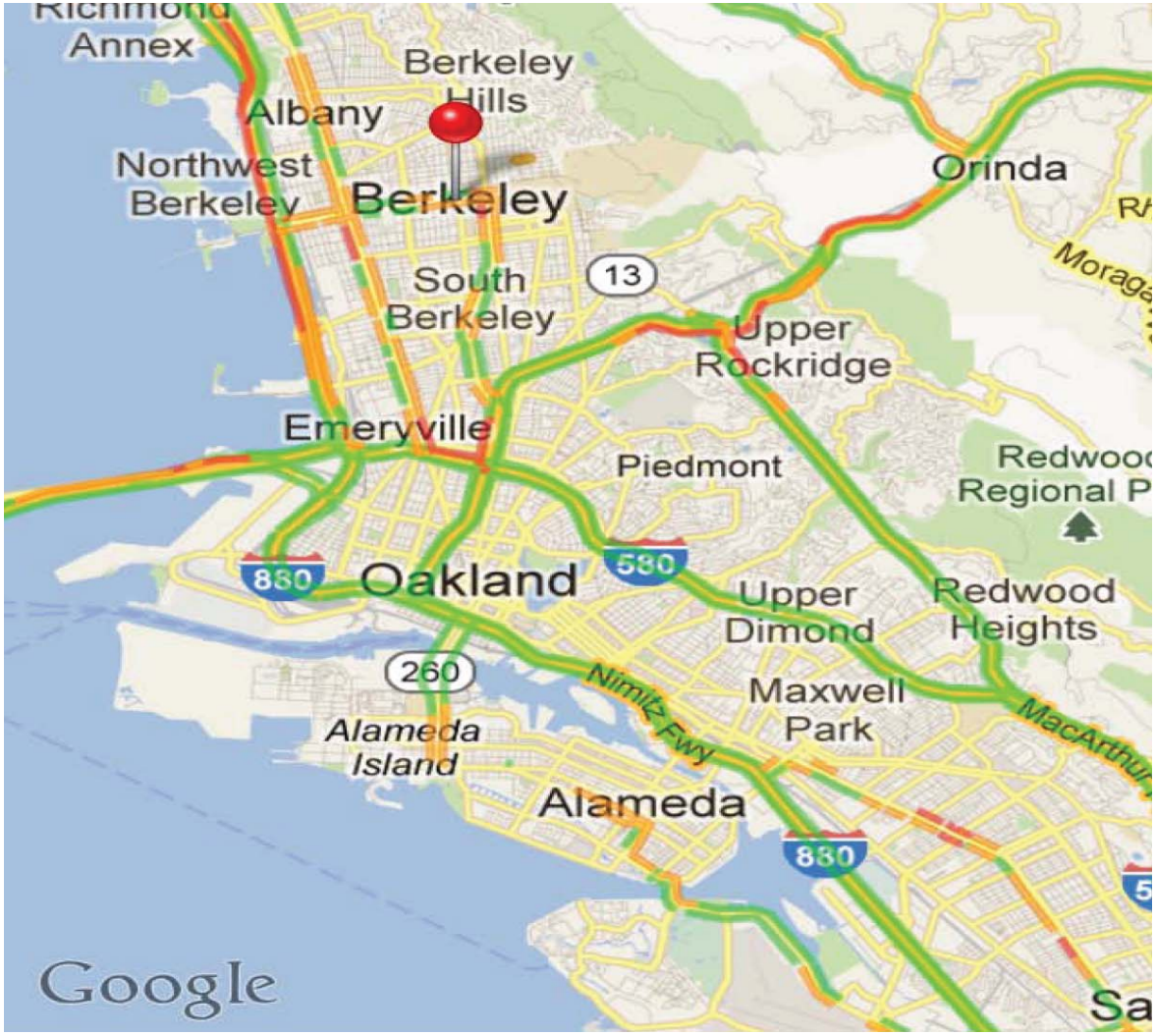
Residual Plot for Quadratic Regression



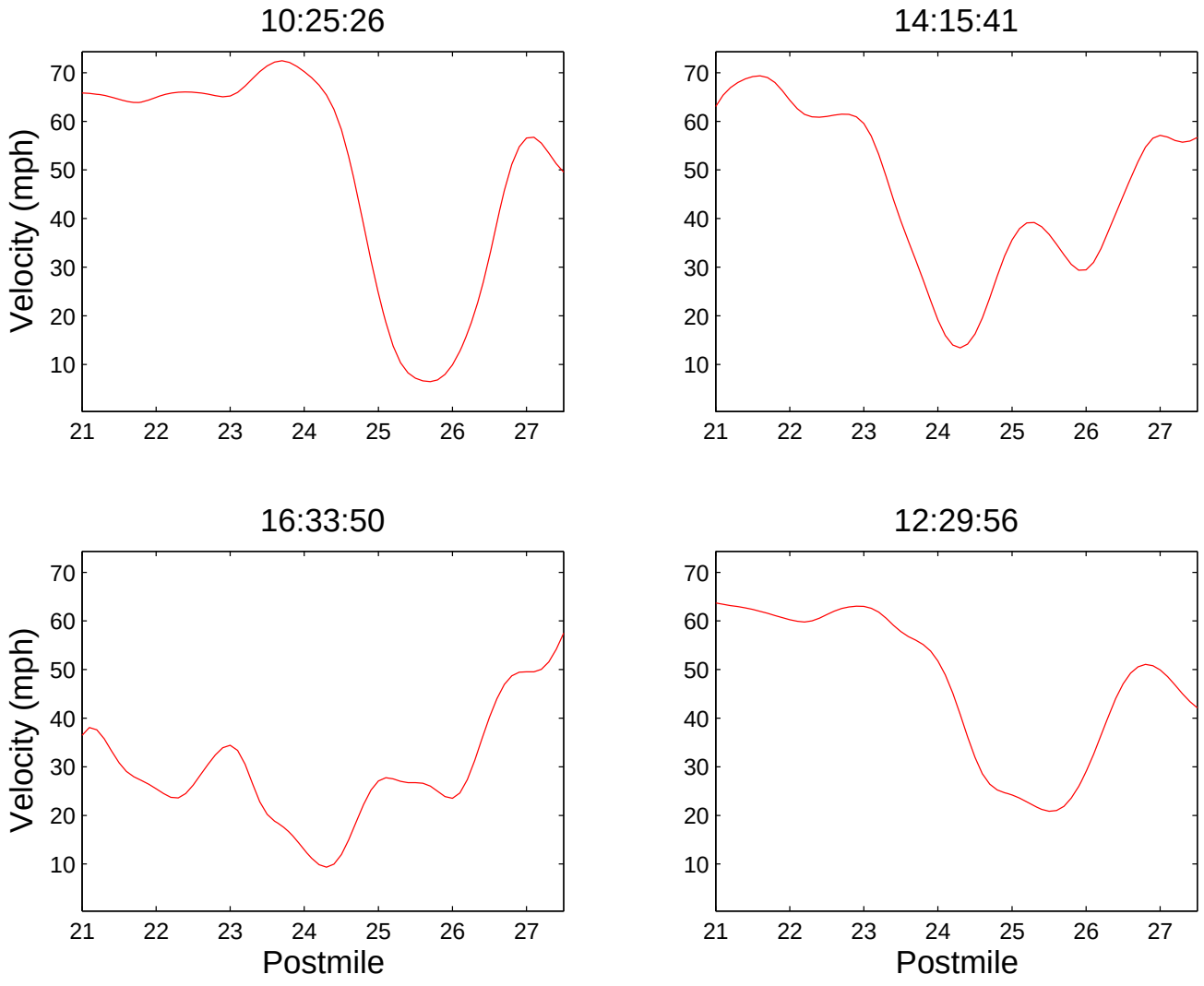
Quantiles


Predicted Quantiles
Fat Content
0
5
10
15
20
25
30
35
40
45
50
true
median
0.1 th
0.9 th

Illustrative Example: Traffic Data



Velocity on I-880



Prediction for Response Functions

- Y and X are both functions

Prediction for Response Functions

- Y and X are both functions
- *FPCfam*: $E(Y(t)|X) = \mu_Y(t) + \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} f_{jk}(\xi_k)\psi_j(t)$

Prediction for Response Functions

- Y and X are both functions
- *FPCfam*: $E(Y(t)|X) = \mu_Y(t) + \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} f_{jk}(\xi_k)\psi_j(t)$
- *FPCpredBands* (Chen and Müller 2012): Global prediction bands for Y conditional on X

Prediction for Response Functions

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

Prediction for Response Functions

- Y and X are both functions
- *FPCfam*: $E(Y(t)|X) = \mu_Y(t) + \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} f_{jk}(\xi_k)\psi_j(t)$
- *FPCpredBands* (Chen and Müller 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

$$\begin{aligned}&\mathrm{cov}(Y(t_1), Y(t_2) \mid X)\\ &= G_{YY}(t_1,t_2) + \sum_{j=1}^{\infty}\Big\{\sum_{k=1}^{\infty} g_{jk}(\xi_k) - \Big(\sum_{k=1}^{\infty} f_{jk}(\xi_k)\Big)^2\Big\}\psi_j(t_1)\psi_j(t_2)\end{aligned}$$

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 $D_X(t) = \mathscr{C}_\alpha \{\text{var}(Y(t)|X)\}^{1/2}$

- For more general random processes:

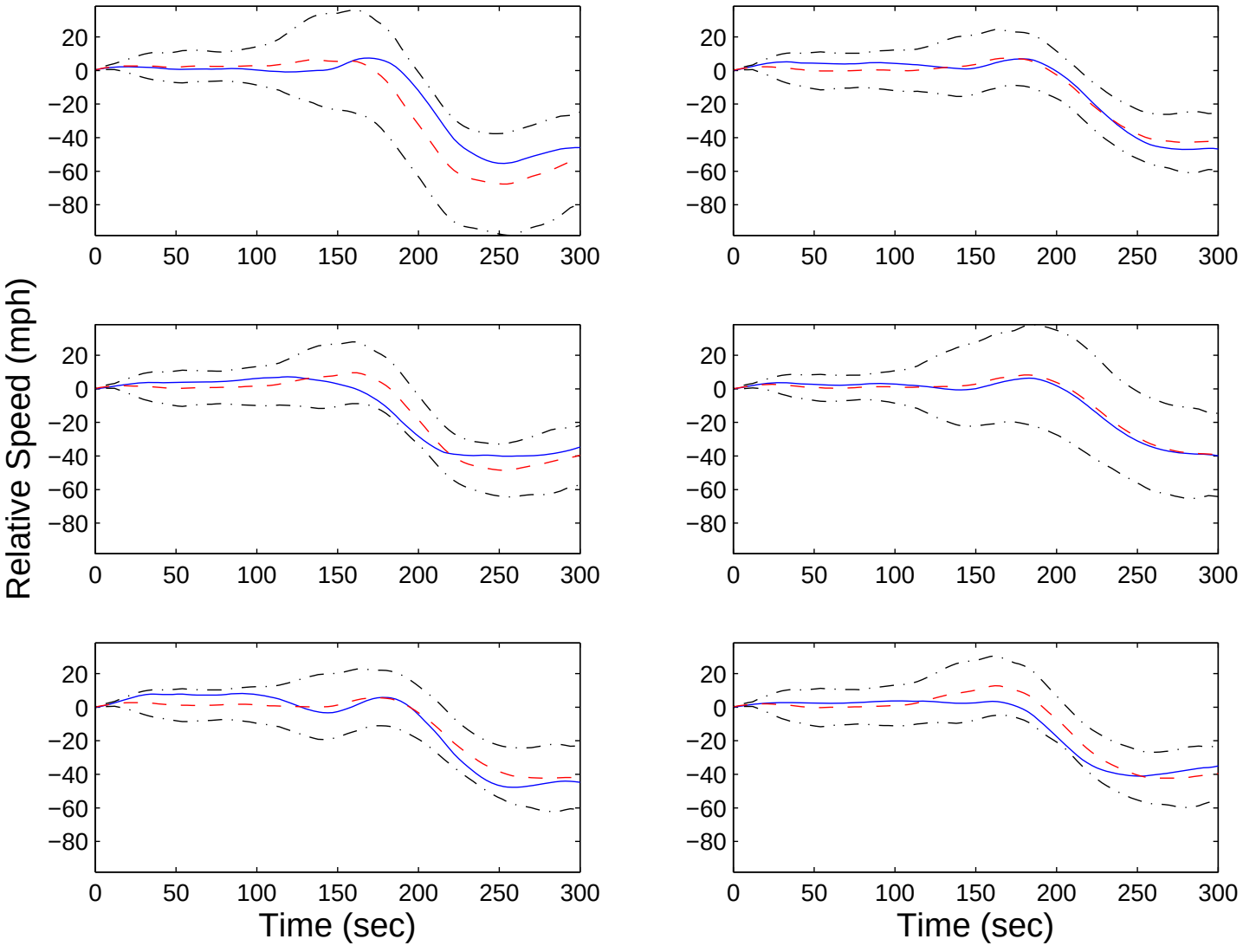
$$E\{P(L_X(t) \leq Y_X(t) \leq U_X(t) \mid X)\} \geq 1 - \alpha$$

- Find $\mathscr{C}_\alpha$ by the empirical coverage

'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



Other Important Tools in PACE

- Modeling of derivatives (linear and nonlinear empirical dynamics)
- Modeling of functional errors (variance processes, volatility processes)
- Time-synchronization based on pairwise warping
- Functional manifold analysis
- Modeling of functional correlations
- Distance based methods (curve clustering)
- Stringing method

Get Started with PACE

```
%[res] = FPCA(y,t,p)
%This function calls PCA.m.
%======
%Input:
%======
%       y:          1*n cell array, y{i} is the vector of measurements for the ith subject,
%                   i=1,...,n.
%       t:          1*n cell array, t{i} is the vector of time points for the ith subject on which
%                   corresponding measurements y{i} are taken, i=1,...,n.
%       p:          a struct obtained from setOptions.m sets the rest of arguments for PCA.m
%                   ex:
%                   >> p = setOptions();
%=======
%Output:
%=======
%       res:         a cell array that contains all returned values from PCA.m
%                  where the last element contains the names of res
%
%    To see the names for res, type names(res)
%    To get individual value back, type getVal(res,varname)
%    To see an example, check with example.m
%    See also PCA, example, names
%
function [res] = FPCA(y,t,p)

  if nargin == 2
    p = setOptions();
```

Get Started with PACE

- User Friendly: help files, examples, documentation, references.
- $\gg \quad p = setOptions()$
  $\gg \quad p2 = setOptions('bwmu', 3)$
- Various options for bandwidth selection, number of components, different designs, errors, pre-binning options.
- The code and descriptions can be downloaded from `http://anson.ucdavis.edu/~mueller/data/programs.html`.

THANK YOU!

- Yao, F., Müller, H.G., Wang, J.L. (2005), Functional data analysis for sparse longitudinal data. *J. American Statistical Association*, 100, 577-590.
- Yao, F., Müller, H.G., Wang, J.L. (2005), Functional Linear Regression Analysis for Longitudinal Data. *Annals of Statistics*, 33, 2873-2903.
- Chiou, J., Müller, H.G. (2007), Diagnostics for functional regression via residual processes. *Computational Statistics and Data Analysis*, 51, 4849-4863.
- Müller, H.G., Yao, F. (2010), Functional quadratic regression. *Biometrika* 97, 49-64.
- Müller, H.-G. and Yao, F. (2008), Functional additive models, *J. of the American Statistical Association*, 103, 1534-1544.
- Müller, H.-G. and Stadtmüller, U. (2005), Generalized functional linear models, *Annals of Statistics*, 33, 774–805.

- Chen, K. and Müller, H.-G. (2012), Conditional quantile analysis when covariates are functions, with application to growth data, *J. of the Royal Statistical Society: Series B*, 74, 67-89.
- Liu, B., Müller, H.G. (2009), Estimating derivatives for samples of sparsely observed functions, with application to on-line auction dynamics. *J. American Statistical Association*, 104, 704-717.
- Müller, H.G., Yao, F. (2010), Empirical dynamics for longitudinal data. *Annals of Statistics*, 38, 3458C3486.
- Müller, H.G., Stadtmüller, U., Yao, F. (2006), Functional variance processes. *J. of the American Statistical Association*, 101, 1007-1018.
- Müller, H.G., Sen, R., Stadtmüller, U. (2011), Functional Data Analysis for Volatility. *J. Econometrics* 165, 233-245.

- Tang, R., Müller, H.G. (2008), Pairwise curve synchronization for high-dimensional data.*Biometrika*, 95, 875-889
- Chen, D., Müller, H.G. (2012), Nonlinear manifold representations for functional data. *Annals of Statistics*, 40, 1-29.
- Yang, W., Mller, H.G. Müller, H.G., Stadtmller, U. (2011), Functional singular component analysis. *J. Royal Statistical Society B*, 73, 303C-324.
- Dubin, J., Müller, H.G. (2005), Dynamical correlation for multivariate longitudinal data. *J. American Statistical Association*, 100, 872-881.
- Peng, J., Müller, H.G. (2008), Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Annals of Applied Statistics*, 2, 1056-1077.
- Chen, K., Chen, K., Müller, H.G., Wang, J.L. (2011), Stringing high-dimensional data for functional analysis. *J. American Statistical Association*, 106, 275-284.