Functional Data Perspectives for Traffic Monitoring and Forecasting

Kehui Chen

Department of Statistics, University of Pittsburgh

Joint work with Hans-Georg Müller at UC Davis

November 11, 2014, INFORMS

Traffic Monitoring and Forecasting

- Dedicated equipment: loop detectors, cameras and radars.
- GPS-enabled phone based traffic monitoring system
- New types of data and new data analysis approaches: Functional data perspectives for traffic forecasting.
- "Mobile Century" Experiment: Joint UC Berkeley - Nokia project.
  J. Herrera, D. Work, R. Herring, X. Ban, Q. Jacobson and A. Bayen (2010)
- The follow-up project 'Mobile Millennium' is generating more data. `http://traffic.berkeley.edu`.


Union Square
Oakland
San Francisco
Canyon
Danville
Alameda
San Leandro
San Francisco Bay
Hayward
Daly City
South San Francisco
Millbrae
Burlingame
San Mateo
Belmont
Redwood City
Union City
Half Moon Bay
Palo Alto
680
13
80
880
280

Individual Trip Data

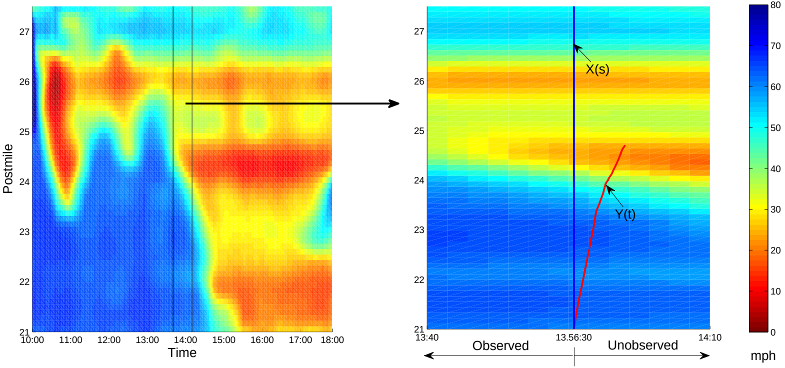
- Decoto Road to the south (Postmile 21) and Winton Avenue to the north (Postmile 27.5)
- Combine data

$$(\{t_l, s_l\}, V_l)_{l=1,\ldots,N},$$

where $N = \sum_i N_i$.

- One can apply a two-dimensional smoothing procedure for these combined data to recover a smooth random velocity field $V(t,s)$ along the highway as an exploratory step.

Observed and Future Velocity Field



Functional Data Perspective

- Assume underlying latent smooth random process that generates the data. Data points could be densely and regularly spaced, or sparsely and irregularly sampled. Measurements may be contaminated with noises.
- Recover the underlying $X(s)$ based on functional principal component analysis not individual smoothing: borrow strength from entire sample.
- Modeling conditional distributions of $Y(t)$ given $X(s)$: predicted curve and global prediction bands.

Functional Principal Component Analysis

- $X(s)$ is a second order random process,
  mean function $\mu(s) \in L^2(\mathscr{T})$,
  continuous covariance function $G(s_1, s_2) = \text{cov}(X(s_1), X(s_2))$.
- $G(s_1, s_2) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s_1)\phi_k(s_2)$, eigenvalues $\lambda_1 \geq \lambda_2, \cdots, \lambda_k, \cdots \geq 0$, and eigenfunctions $\phi_k(t)$.
- Karhunen-Loève expansion: double orthogonal.

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

- Best linear expansion with K components:

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

Estimation

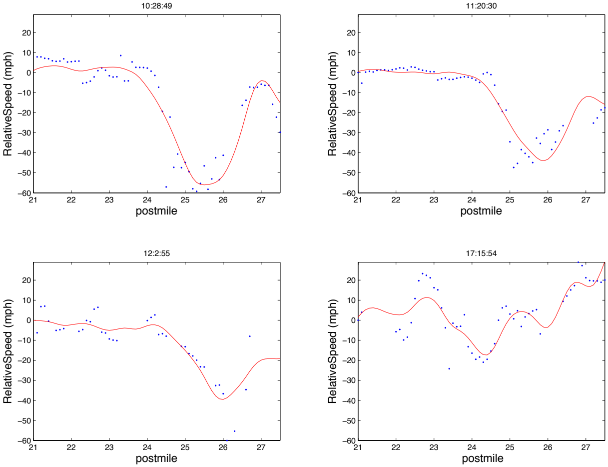
- $$\hat{X}(s) = \hat{\mu}(s) + \sum_{k=1}^{K} \hat{\xi}_k \hat{\phi}_k(s)$$

- Pool all the sample. Smoothing of mean and covariance functions leads to eigenfunctions/eigenvalues.

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

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

Predictor Functions



Prediction for Response Functions

- $X(s) \approx \mu_X(s) + \sum_{k=1}^{K} \xi_k \phi_k(s)$
  $Y(t) \approx \mu_Y(t) + \sum_{j=1}^{P} \zeta_j \psi_j(t)$

Prediction for Response Functions

- $X(s) \approx \mu_X(s) + \sum_{k=1}^{K} \xi_k \phi_k(s)$
  $Y(t) \approx \mu_Y(t) + \sum_{j=1}^{P} \zeta_j \psi_j(t)$
- Prediction of Mean Function, FAM (Müller and Yao 2008)
  $E(Y(t)|X) \approx \mu_Y(t) + \sum_{j=1}^{P} \sum_{k=1}^{K} f_{jk}(\xi_k) \psi_j(t)$

Prediction for Response Functions

- $X(s) \approx \mu_X(s) + \sum_{k=1}^{K} \xi_k \phi_k(s)$
  $Y(t) \approx \mu_Y(t) + \sum_{j=1}^{P} \zeta_j \psi_j(t)$
- Prediction of Mean Function, FAM (Müller and Yao 2008)
  $E(Y(t)|X) \approx \mu_Y(t) + \sum_{j=1}^{P} \sum_{k=1}^{K} f_{jk}(\xi_k) \psi_j(t)$
- $\quad \mathrm{cov}(Y(t_1), Y(t_2) \mid X)$
  $\approx \sum_{j=1}^{P} \mathrm{var}(\zeta_j \mid X) \psi_j(t_1) \psi_j(t_2)$
  $\approx G_{YY}(t_1, t_2) + \sum_{j=1}^{P} \sum_{k=1}^{K} \{g_{jk}(\xi_k) - f_{jk}^2(\xi_k)\} \psi_j(t_1) \psi_j(t_2)$

Prediction for Response Functions

- $X(s) \approx \mu_X(s) + \sum_{k=1}^{K} \xi_k \phi_k(s)$
  $Y(t) \approx \mu_Y(t) + \sum_{j=1}^{P} \zeta_j \psi_j(t)$
- Prediction of Mean Function, FAM (Müller and Yao 2008)
  $E(Y(t)|X) \approx \mu_Y(t) + \sum_{j=1}^{P} \sum_{k=1}^{K} f_{jk}(\xi_k) \psi_j(t)$
- $\quad \mathrm{cov}(Y(t_1), Y(t_2) \mid X)$
  $\approx \sum_{j=1}^{P} \mathrm{var}(\zeta_j \mid X) \psi_j(t_1) \psi_j(t_2)$
  $\approx G_{YY}(t_1, t_2) + \sum_{j=1}^{P} \sum_{k=1}^{K} \{g_{jk}(\xi_k) - f_{jk}^2(\xi_k)\} \psi_j(t_1) \psi_j(t_2)$
- If $X(s)$ is a Gaussian process,
  $f_{jk}(\xi_k) = E(\zeta_j \mid \xi_k)$, $g_{jk}(\xi_k) = E(\zeta_j^2 - \gamma_j \mid \xi_k)$

Global Prediction Bands

- $Y_X(t) \approx \mu_{Y|X}(t) + \sum_{j=1}^{P} \zeta_j(X)\psi_j(t)$.
- $$\Omega_{X,\alpha} = \{(\zeta_1(X), \dots, \zeta_P(X)) : \quad \sum_{j=1}^{P} \frac{\zeta_j(X)^2}{\gamma_j(X)} \leq \mathscr{C}_{X,\alpha}^2\},$$

  such that

  $$P(\zeta_X \in \Omega_{X,\alpha}) = 1 - \alpha.$$

- The upper bound function $U(t)$ is found by solving the maximization problems

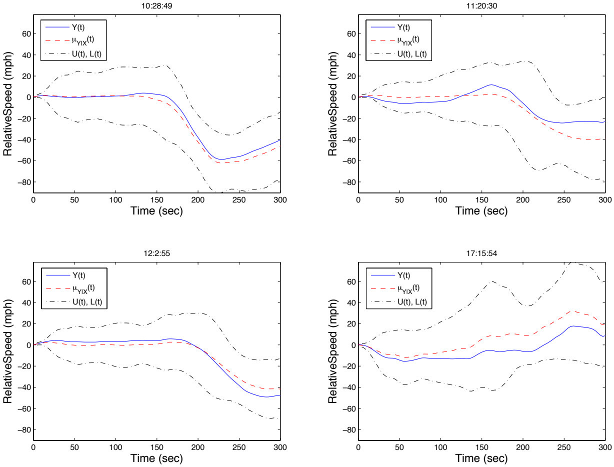
$$\max_{\zeta_X \in \Omega_{X,\alpha}} \left\{ \mu_{Y|X}(t) + \sum_{j=1}^{P} \zeta_j(X)\psi_j(t) \right\}, \qquad \text{for all } 0 < t < 1.$$

Global Prediction Bands

- $\hat{U}(t) = \hat{\mu}_{Y|X}(t) + \left\{ \mathscr{C}_{X,\alpha}^2 \sum_{j=1}^{P} \hat{\gamma}_j(X) \hat{\psi}_j^2(t) \right\}^{1/2}$
$= \hat{\mu}_{Y|X}(t) + \mathscr{C}_{X,\alpha} \{ \widehat{\text{var}}(Y_X(t)) \}^{1/2}$.
- In the case that $(\zeta_1(X), \ldots, \zeta_P(X))$ are jointly Gaussian, $\mathscr{C}_{X,\alpha} = \mathscr{C}_\alpha = \sqrt{\chi^2_{P,1-\alpha}}$.
- In general case: Find a constant $\mathscr{C}_\alpha$ and regions $\Omega_{X,\alpha}$,

$$E[P(\zeta_X \in \Omega_{X,\alpha})] = 1 - \alpha.$$

Estimated 90% Prediction Regions



Extensions

- Other types of data from GPS-enabled phones: VTLs.
- Dynamic updating: prediction based on the current time and location.
- larger networks of roads: divide and conquer.
- Other functional data methods.

- K. Chen and H.G. Müller (2014), " Modeling conditional distributions for functional responses, with application to traffic monitoring via GPS-enabled mobile phones", *Technometrics*, **56(3)**, 347-358.
- Code available (written in Matlab), PACE package version 2.17, `http://www.stat.ucdavis.edu/PACE/`
- J. Herrera, D. Work, R. Herring, X. Ban, Q. Jacobson and A. Bayen (2010), "Evaluation of Traffic Data Obtained via GPS-Enabled Mobile Phones: The Mobile Century Field Experiment," *Transportation Research C*, **18**, 568-583.

THANK YOU!