Dimension Reduction Models for Functional Data

Wei Yu and Jane-Ling Wang
Genentech UC Davis

4th Lehmann Symposium
May 11, 2011

Functional Data

- A sample of curves - one curve, $X(t)$, per subject.

  - These curves are usually considered realizations of a stochastic process in $L_2(I)$.

$\Rightarrow$ ∞ - dimensional

- In reality, $X(t)$ is recorded at a dense time grid, often equally spaced (regular).

$\Rightarrow$ high-dimensional.

Example: Medfly Data

- Number of eggs laid daily were recorded for each of the 1.000 female medflies until death.
- $X(t)$= # of eggs laid on day t.
- Average lifetime = 35.6 days
- Average lifetime reproduction = 759.3 eggs

Longitudinal Data

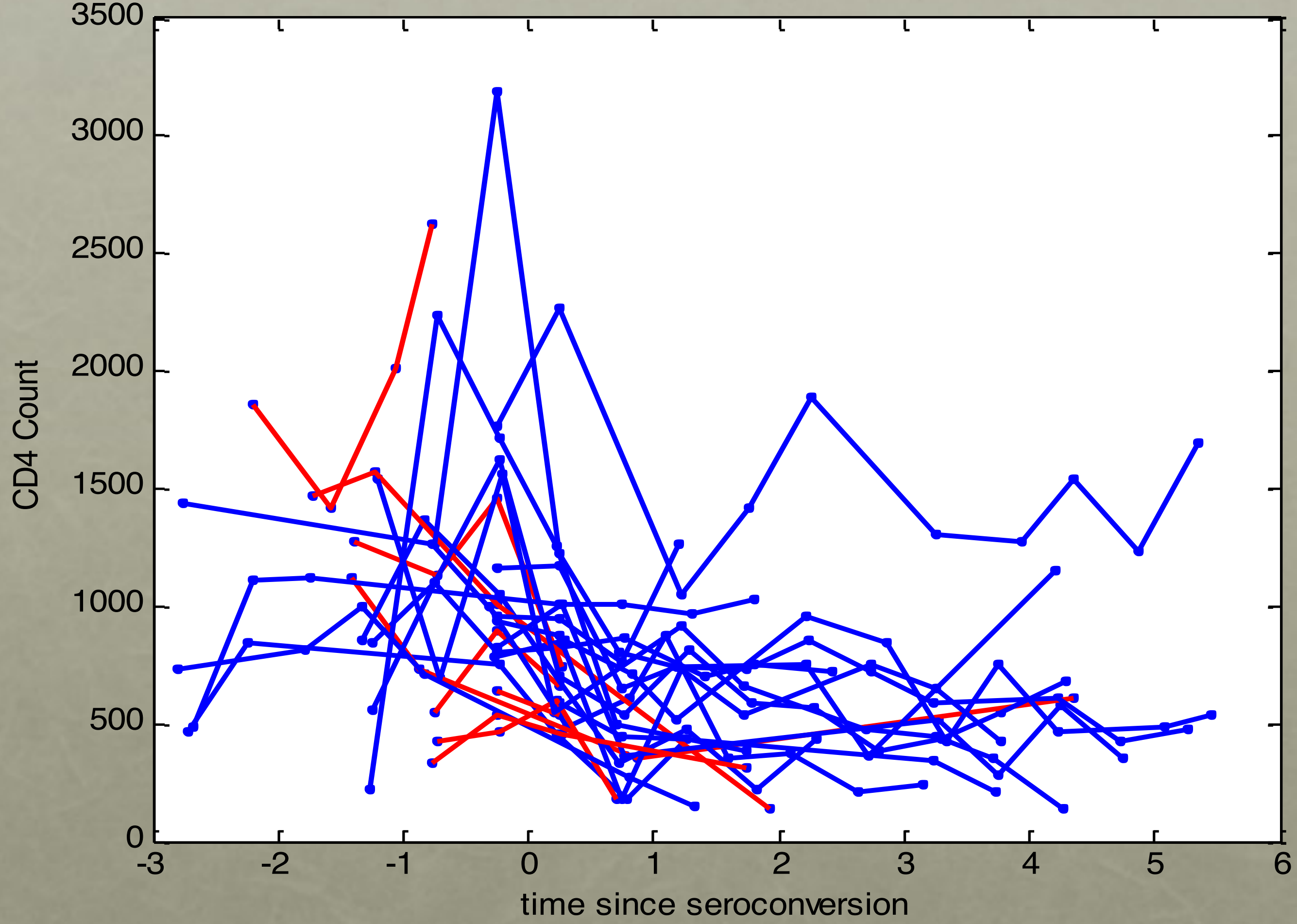
- When X(t) is recorded sparsely, often irregular in the time grid, they are referred to as longitudinal data.

Longitudinal data = sparse functional data

- "regular and sparse" functional data = panel data

They require parametric approaches and will not be considered in this talk.

CD4 Counts of First 25 Patients



Three Types of Functional Data

- **Curve data** - This is the easiest to handle in theory, as functional central limit theorem and LLN apply.

  - $\sqrt{n}$ rate of convergence can be achieved because the observed data is ∞ - dimensional.

- **Dense functional data** – could be presmoothed and inherit the same asymptotic properties as curve data.

- **Sparse functional data / longitudinal data** – hardest to handle both in methodology and theory .

Dimension Reduction

- Despite the different forms that functional data are observed, there is an infinite dimensional curve underneath all these data.

- Because of this intrinsic infinite dimensional structure, dimension reduction is required to handle functional/longitudinal data.

Dimension Reduction

- Principal Component analysis (PCA) is a standard dimension reduction tool for multivariate data. It is essentially a spectral decomposition of the covariance matrix.

- PCA has been extended to functional data and termed functional principal component analysis (FPCA).

Dimension Reduction

- FPCA leads to the Karhunan-Loeve decomposition:

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

where $\mu(t) = E(X(t))$,

ϕ_k are the eigenfunctions of the covarnaice function $\Sigma(s, t) = \mathrm{cov}\,(X(s), X(t))$.

References for FPCA

- **Dense Functional Data**

  - Rice and Silverman (1991, JRSSB)

    Hall and Housseni (2006, AOS)

- Sparse Functional data – Yao Müller and Wang (2005) Hall, Müller and Wang (2006)

- Hsing and Li (2010)

Dimension Reduction Regression

- In this talk, we focus on regression models that involves functional data.

- There are two scenarios:

  - Scalar response Y and functional/longitudinal covariate $X(t)$

  - Functional response $Y(t)$ and functional covariates, $X_1(t), \cdots, X_p(t)$, some of which may be scalars.

Univariate Response: Sliced Inverse Regression

Motivation

- Model univariate response *Y* with longitudinal covariate *X(t)*.

- Current approaches:

  * Functional linear model:

$$Y = \int \beta(t) X(t) dt + e = \ < \beta, X > + e$$

  * Completely nonparametric: $Y = g(X) + e,$

$$g : \text{functional space} \ \rightarrow \Re.$$

Motivation

* Functional single-index model:

$$Y = g(<\beta, X>) + e.$$

* Goal: Use multiple indices

$$<\beta_1, X>, \cdots, <\beta_k, X>$$

$$Y = g(<\beta_1, X>, \cdots, <\beta_k, X>) + e.$$

without any model assumption on g.

Background

$Y \in \mathbb{R}, \ \ X \in \mathbb{R}^p$

Dimension reduction model: $Y = f(\beta_1^T X, \cdots, \beta_k^T X, e),$

where f is unknown, $\ e \perp X \ , \ \ k \ll p.$

$\Leftrightarrow$ Given $(\beta_1^T X, \cdots, \beta_k^T X), \ \ Y \perp X.$

$\Leftrightarrow$ These k indices captured all the information contained in X.

Background

- Special Cases:

$$Y = f_1(\beta_1^T X) + \cdots + f_k(\beta_k^T X) + e$$

$\Rightarrow$ projection pursuit model

$$Y = f(\beta_1^T X) + e,$$

$\Rightarrow$ single-index model.

Sliced Inverse Regression (Li, 1991)

- Separate the dimension reduction stage from the nonparametric estimation of the link function.
- Stage 1 – Estimate the linear space generated by β's

  $\Downarrow$

  Effective dimension reduction (EDR) space

  * Only the EDR space can be identified , but not β.
- Stage 2 - Estimate the nonparametric link function f via a smoothing method.

How and Why does SIR work?

- Do inverse regression **$E(X|Y)$** rather than the forward regression $E(Y|X)$.
- For standardized X, **Cov[$E(X|Y)$]** is contained in the EDR space under a design condition.

⇨ Eigenvectors of Cov[$E(X|Y)$] are the EDR directions.

- Perform a principal component analysis on $E(X|Y)$.
- SIR employs a simple approach to estimate $E(X|Y)$ by slicing the range of Y into H slices and use the sample mean of X's within each slice to estimate $E(X|Y)$.

- **Algorithm**
  - Sample: $\{(Y_1, \mathbf{X}_1), \cdots, (Y_n, \mathbf{X}_n)\}$;
  - Sort the data by Y: $\{(Y_{(1)}, \mathbf{X}_{(1)}), \cdots, (Y_{(n)}, \mathbf{X}_{(n)})\}$;
  - Divide the sorted data set into H slices;
  - Within hth slice, compute the sample mean of $\mathbf{X}$,

$$\bar{\mathbf{X}}_h = \frac{1}{n_h} \sum_{(i)\in slice\ h} \mathbf{X}_{(i)};$$

  - Compute the covariance matrix of the sliced means of $\mathbf{X}$,

$$\hat{\Sigma}_e = n^{-1} \sum_{h=1}^{H} n_h (\bar{\mathbf{X}}_h - \bar{\mathbf{X}})(\bar{\mathbf{X}}_h - \bar{\mathbf{X}})';$$

  - Find the e.d.r. directions by

$$\hat{\Sigma}_e \hat{\beta}_j = \hat{\theta}_j \hat{\Sigma}_{\mathbf{x}} \hat{\beta}_j,$$

    where $\hat{\theta}_1 \geq \hat{\theta}_2 \geq \cdots \geq 0$.

When does SIR work?

- Linear design condition : For any $b \in \Re^p$

$$E(b'X \mid \beta_1 X, \cdots, \beta_k X) = \text{linear function of } \beta_1 X, \cdots, \beta_k X.$$

- The design condition is satisfied when X is elliptical symmetric, e.g. Gaussian.
- When the dimension of X is high, the conditoin is satisfied for almost all EDR spaces (Hall and Li (1993)).

End of Introduction to SIR

How to Extend SIR to Functional Data?

$$\text{Response} \quad Y \in \Re, \qquad \text{covariate} \quad X(t)$$

- Need to estimsate E$\{X(t)|Y\}$ and its covariance, Cov[E $\{X(t)|Y\}$].
- This is straightforward if the entire curve $X(t)$ can be observed.

  Therefore SIR can be employed directly at each point t.
- Ferre and Yao (2003), Ferre and Yao (2005, 2007)
- Ren and Hsing (2010)

How to Extend SIR to Functional Data?

Response $Y \in \mathbb{R}$, Covariate $X(t)$ - a function

- What if the curves are only observable at sparse and possibly irregular time points?

  Observe (Y_i, X_i) for the ith subject.

  where $X_i = (X_{i1}, \cdots, X_{ini})$, with $X_{ij} = X_i(t_{ij})$.

- We consider a unified approach that adapts to both sparse longitudinal and functional covariates.

Functional Inverse Regression (FIR)
Yu and Wang (201?)

Response $Y \in \Re$, covariate $X(t) \in L^2([a,b])$.

Observe Y_i and $X_i = (X_{i1}, \cdots, X_{in_i})$, where $X_{ij} = X_i(t_{ij})$.

- To estimate E{X(t)| Y=y} = $\mu(t, y)$, we do a 2D smoothing of $\{X_{ij}\}$ over $\{t_{ij}, Y_i\}$, for $j = 1, \cdots, n_i$; $i = 1, \cdots, n$.

- Once we have $\hat{\mu}(t,y)$, Cov [E{X(t)|Y}] can be estimated by the sample covariance

$$\hat{\Gamma}(s,t) = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}(s, Y_i) \hat{\mu}(t, Y_i).$$

- **Algorithm**
  - Sample: $\{(Y_1, \mathbf{X}_1), \cdots, (Y_n, \mathbf{X}_n)\}$;
  - Sort the data by Y: $\{(Y_{(1)}, \mathbf{X}_{(1)}), \cdots, (Y_{(n)}, \mathbf{X}_{(n)})\}$;
  - Divide the sorted data set into H slices;
  - Within hth slice, compute the sample mean of $\mathbf{X}$,

$$\bar{\mathbf{X}}_h = \frac{1}{n_h} \sum_{(i) \in slice\ h} \mathbf{X}_{(i)};$$

  - Compute the covariance matrix of the sliced means of $\mathbf{X}$,

$$\hat{\Sigma}_e = n^{-1} \sum_{h=1}^{H} n_h (\bar{\mathbf{X}}_h - \bar{\mathbf{X}})(\bar{\mathbf{X}}_h - \bar{\mathbf{X}})';$$

  - Find the e.d.r. directions by

$$\hat{\Sigma}_e \hat{\beta}_j = \hat{\theta}_j \hat{\Sigma}_{\mathbf{x}} \hat{\beta}_j,$$

    where $\hat{\theta}_1 \geq \hat{\theta}_2 \geq \cdots \geq 0$.

Theory

- Identifiability of the EDR space

  - We need to standardize the curve $X(t)$, but the covariance operator of X is not invertible!

- Under standard regularity conditions,

  cov $[E\{X(t)|Y\}]$ can be estimated at 2D rate, but

$$\| \hat{\beta}_j - \beta_j \| = o_P((nh)^{-1} + h^2)$$

- EDR directions, β's can be estimated at 1D rate.

Choice of # of Indices

- Fraction of variation explained
- AIC or BIC.
- A Chi-square test as in Li(1991).
- Ferre and Yao (2005) used an approach in Ferre (1998).
- Li and Hsing (2010) developed another procedure.

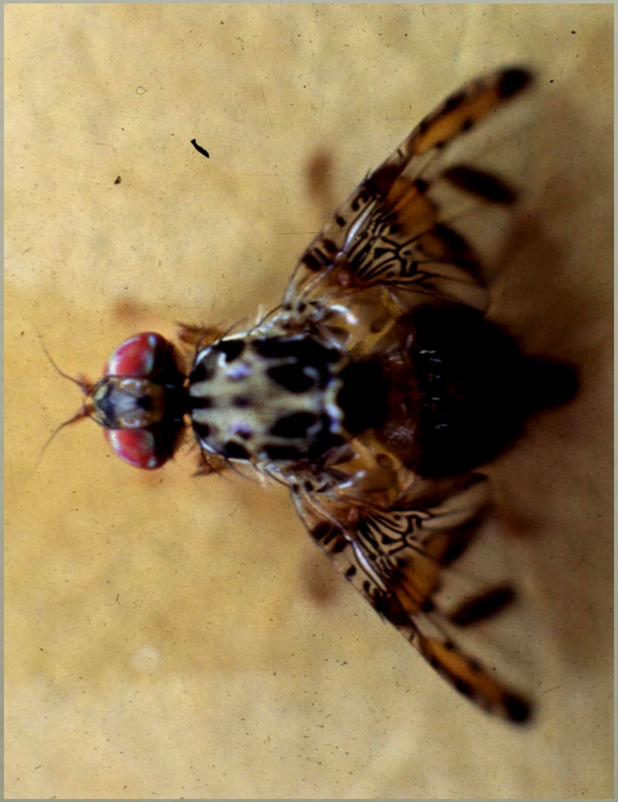
End of FIR

Fecundity Data

- Number of eggs laid daily were recorded for each of the 1.000 female medflies until death.

- Average lifetime = 35.6 days
- Average lifetime reproduction = 759.3 eggs
- 64 flies were infertile and excluded from this analysis.
- Goal : How early reproduction (daily egg laying up to day 20) relates to mortality.
- Y= lifetime (days), $X(t)$= # of eggs laid on day t, $1 \leq t \leq 20$.

Mediterranean Fruit Fly



Multivariate PCA on *X(t)*

Table 2: Importance of components

Component	Standard deviation	Proportion of Variance	Cumulative Proportion
1	76.5174	0.5597	0.5597
4	18.8990	0.0341	0.7782
5	17.2429	0.0284	0.8067
9	14.6652	0.0206	0.8960
10	14.0522	0.0189	0.9149
12	12.8043	0.0157	0.9476
13	12.1599	0.0141	0.9617
16	10.9386	0.0114	0.9993

Multivariate PCA (cont'd)

Proportion	80%	90%	95%	99%
Number of directions	5	10	13	16

- This is not surprising as reproduction is a complicated system that is subject to a lot of variations.
- Hence, a PC regression is not an effective dimension reduction tool for this data.
- However, the information it contains for lifetime may be simpler and could be summarized by much fewer EDR directions.

Comparison of PCA and FSIR

Proportion	80%	90%	95%	99%
Number of directions	5	10	13	16

Table 3: The coverage of first direction

Data type	Coverage
Complete data	0.92
Sparse data	0.992

Sparse Egg Laying Curves

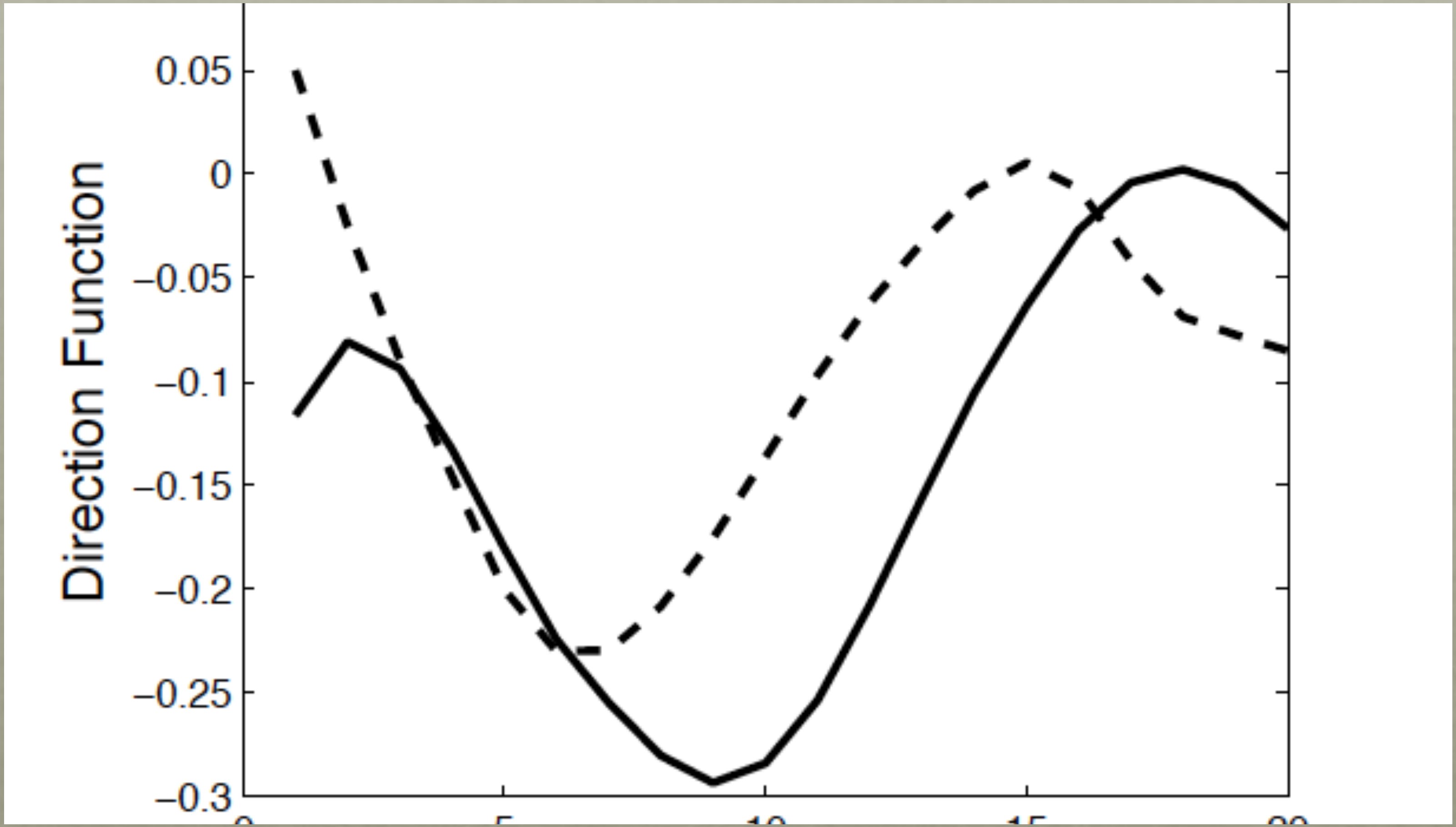
- Randomly select n_i from {1,2,…,8} and then choose n_i days from the *i*th fly.

Table 3: The coverage of first direction

Data type	Coverage
Complete data	0.92
Sparse data	0.992

- Thus, one (or two) directions suffices to summarize the information contained in the fecundity data to infer lifetime of the same fly.

Estimated Directions



Complete data (solid), Sparse data (dash)

Conclusion

- The first directions estimated from the complete and sparse data have similar pattern.
- The correlation between the effective data, using a single index $< \beta, X>$, for the complete and sparse data turns out to be 0.8852 .
- Sparse data provided similar information as the complete data, and both outperform the principal component regression for this data.

Functional Response:
Single (or Multiple) Index Model

Objectives

- Model longitudinal response *Y(t)* with longitudinal covariates, $X_1(t),\cdots,X_p(t)$, *s*ome or all of $X_i(t)$ may be scalar.

- Adopt a dimension reduction (**semiparametric**) model

AIDS Data

- CD4 counts of 369 patients were recorded.
- Five covariates, age is time-invariant but the rest four are longitudinal.

  packs of cigarettes

  Recreational drug use (1: yes, 0: no)

  number of sexual partners

  mental illness scores

Single (or Multiple) Index Model

First consider $Y \in \mathbb{R}, \quad X \in \mathbb{R}^p.$

$Y = g(\beta^T X) + \varepsilon \qquad \rightarrow$ single index

$Y = g(\beta_1^T X, \beta_2^T X, \ldots, \beta_k^T X) + \varepsilon \qquad \rightarrow$ multiple indices

$k < p$

Functional Single Index Model

Jiang and Wang (2011, AOS)

- When there is no longitudinal component.

$$Y = \mathrm{g}(\beta^{\mathrm{T}} X) + \varepsilon.$$

$$Y \to Y(t) \Rightarrow Y(t) = g(\beta^{\mathrm{T}} X) + \varepsilon$$

- However, this uses the same link function at all time t and does not properly address the role of the time factor,

Functional Single Index Model

- We consider a time dynamic link functio

$$Y(t) = g(t, \beta^{T} X) + \varepsilon.$$

Non Dynamic: $Y(t) = g(\beta^{T} X) + \varepsilon$

- Longitudinal $X(t) \Rightarrow$ $Y(t) = g(t, \beta^{T} X(t)) + \varepsilon.$

- For identifiability, we assume

$$\|\beta\| = 1 \text{ and } \beta_1 > 0.$$

Method and Theory: Estimation

$$Y(t) = g(t, \beta^{T} z(t)) + \varepsilon.$$

- We adopt an approach that estimates β and μ simultaneously by extending

  "MAVE" by Xia *et al.* (2002)

  to longitudinal data.

- The advantage is that no undersmoothing is needed to estimate β at the root-n rate.

MAVE (Xia *et al.*, 2002)

Single index model $Y = \mu(\beta^T Z) + \epsilon$.

Cond Var$= \sigma^2_\beta(\beta^T Z) = E[\{Y - E(Y|\beta^T Z)\}^2|\beta^T Z]$

$E\{\sigma^2_\beta(\beta^T Z)\} = E\{Y - E(y|\beta^T Z)\}^2$.

β could be estimated by minimizing $E\{\sigma^2_\beta(\beta^T Z)\}$.

$$\hat{\beta} = \arg \min_{|\beta|=1, a_j, b_j} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} [Y_i - \{a_j + b_j \beta^T (Z_j - Z_i)\}]^2 w_{ij} \right),$$

where $w_{ij} = K(\beta^T(Z_i - Z_j)/h) / \sum_{k=1}^{n} K(\beta^T(Z_k - Z_j)/h)$.

MAVE (Xia *et al.*, 2002)

$$\hat{\beta} = \arg \min_{|\beta|=1, a_j, b_j} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} [Y_i - \{a_j + b_j \beta^T (Z_j - Z_i)\}]^2 w_{ij} \right),$$

where $w_{ij} = K(\beta^T(Z_i - Z_j)/h) / \sum_{k=1}^{n} K(\beta^T(Z_k - Z_j)/h)$.

Here a local linear smoother is applied to

$E(Y \mid \beta^T Z) = \mu(\beta^T Z) \;\; : \;\; a + b\,(\beta^T Z)$

MAVE for Longitudinal Data

The minimizing target now becomes:

$$\sum_{j=1}^{n}\sum_{l=1}^{n_i}\sum_{i=1}^{n}\sum_{k=1}^{n_j}\left(y_{ik} - a_{jl} - b_{jl}(t_{jl} - t_{lk}) - d_{jl}\beta^T(z_{jl} - z_{lk})\right)^2 W_{ikjl},$$

where $Z_{ik} = Z_i(t_{ik}), \quad Z_{jl} = Z_j(t_{jl}),$

$$W_{ikjl} = \frac{K\left(\frac{t_{ik}-tjl}{h_t}, \frac{\beta^T(z_i - z_j)}{h_z}\right)}{\sum_{i=1}^{n}\sum_{k=1}^{n_i} K\left(\frac{t_{ik}-tjl}{h_t}, \frac{\beta^T(z_i - z_j)}{h_z}\right)}.$$

Algorithm for MAVE

1. Initialize $\boldsymbol{h_t}$ and $\boldsymbol{h_z}$, the bandwidths of $\boldsymbol{T}$ and $\boldsymbol{\beta^T Z}$ respectively.

2. Initialize the value of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}_{(\mathbf{0})}$.

3. With $\hat{\boldsymbol{\beta}}_{(i)}$ given, $(\hat{\tilde{\boldsymbol{a}}}, \hat{\tilde{\boldsymbol{b}}}, \hat{\tilde{\boldsymbol{d}}}) = \arg\min_{\tilde{a},\tilde{b},\tilde{d}} \hat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}_{(i)}}$ can be obtained by weighted LS.

4. With $(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}}, \tilde{\boldsymbol{d}})$ given from the previous procedure, $\hat{\boldsymbol{\beta}}_{(i+1)} = \arg\min_{\boldsymbol{\beta}} \hat{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}$ can be obtained by weighted LS, too.

5. Repeat step 3 and 4 till $|\hat{\boldsymbol{\beta}}_{(i+1)} - \hat{\boldsymbol{\beta}}_{(i)}| < \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is some given tolerance value.

rMAVE (Refined MAVE)

- If we iterate MAVE once to refine it, this is called rMAVE.

- Xia et al. (2002) found such an iteration improves efficiency.

- We adopted rMAVE for longitudinal data.

$\sqrt{n}$ - convergence of β

Theorem. Let $\hat{\beta}$ be the estimator of β_0 in the algorithm. Under some regularity conditions, we have

$$\sqrt{n}(\hat{\beta} - \beta_0) \longrightarrow^{\mathcal{D}} N(0, \Sigma),$$

where

$$\begin{aligned}
\Sigma =& [E(G(T, Z))]^{+} \Sigma^{*} [E(G(T, Z))]^{+}, \\
G(t, z) =& \left(\frac{d\mu(t, \beta_0^T z)}{d(\beta_0^T z)} \right)^2 \left(z z^T - m(t, z) m(t, z)^T \right), \\
G_0(t, z) =& \left(\frac{d\mu(t, \beta_0^T z)}{d(\beta_0^T z)} \right) (z - m(t, z)), \\
m(t, z) =& E(Z | T = t, \beta_0^T Z = \beta_0^T z),
\end{aligned}$$

and A^{+} is the Moore-Penrose inverse of matrix A.

$\sqrt{n}$ - convergence of β

$$\Sigma^* = \frac{EN-1}{EN} E(\{G_0(T_{11}, Z_1)\epsilon_{11}\}\{G_0(T_{12}, Z_1)\epsilon_{12})\}^T) + \frac{1}{EN} E(\{G_0(T_{11}, Z_1)\epsilon_{11}\}\{G_0(T_{11}, Z_1)\epsilon_{11}\}^T)$$

Convergence of the Mean Fucntion

$$\sqrt{n\bar{N}h_t h_z}\ [\hat{\mu}(t,u)-\mu(t,u)] \rightarrow N(\eta(t,u),\Sigma(t,u)),$$

where $\bar{N} = \Sigma\ n_i$.

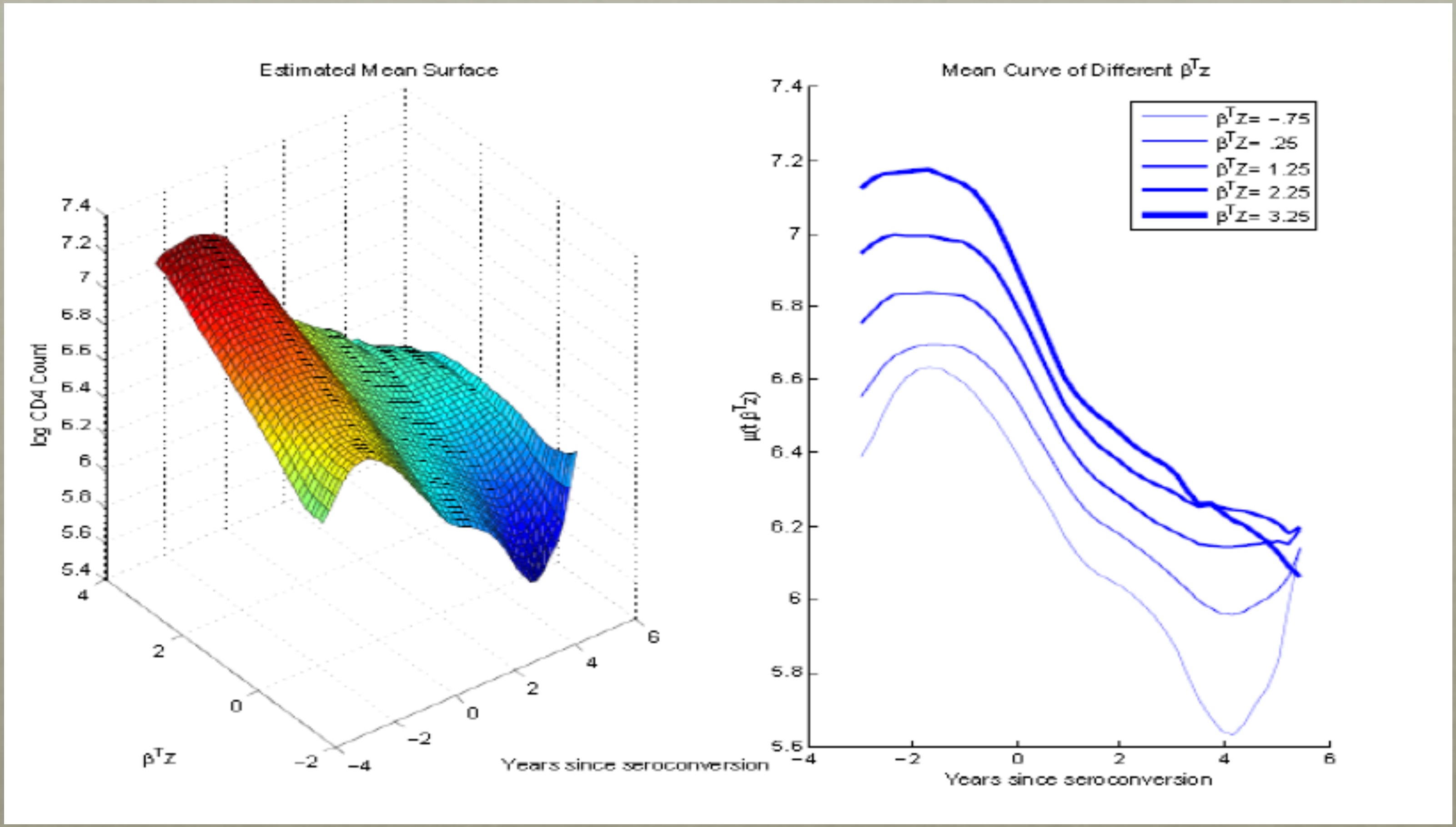
$$\sqrt{n\bar{N}h_t h_z}\ [\hat{\mu}(t,\hat{\beta}^T Z)-\mu(t,\beta^T Z)] \rightarrow$$

$$N(\eta(t,\ \beta^T Z),\ \Sigma(t,\ \beta^T Z))$$

AIDS data Analysis

3-fold CV	h_μ	(1.25, 3.00)
	$\hat{\beta}^T$	(0.0141, 0.5700, 0.8211, -0.0159, -0.0216)
	$\mathrm{Var}(\hat{\beta}) \approx \frac{\hat{\Sigma}}{\sqrt{n}}$	$\begin{pmatrix} 0.0035 & 0.0045 & 0.0210 & -0.0010 & -0.0003 \\ 0.0045 & 0.0956 & 0.2733 & -0.0049 & -0.0038 \\ 0.0210 & 0.2733 & 2.4311 & 0.0029 & -0.0214 \\ -0.0010 & -0.0049 & 0.0029 & 0.0069 & -0.0009 \\ -0.0003 & -0.0038 & -0.0214 & -0.0009 & 0.0021 \end{pmatrix}$
10-fold CV	h_μ	(1.00, 4.00)
	$\hat{\beta}^T$	(0.0128, 0.5530, 0.8326, -0.0193, -0.0225)
	$\mathrm{Var}(\hat{\beta}) \approx \frac{\hat{\Sigma}}{\sqrt{n}}$	$\begin{pmatrix} 0.0037 & 0.0070 & 0.0284 & -0.0011 & -0.0005 \\ 0.0070 & 0.1287 & 0.3744 & -0.0065 & -0.0061 \\ 0.0284 & 0.3744 & 2.7206 & -0.0018 & -0.0302 \\ -0.0011 & -0.0065 & -0.0018 & 0.0070 & -0.0008 \\ -0.0005 & -0.0061 & -0.0302 & -0.0008 & 0.0023 \end{pmatrix}$

AIDS: Mean Function



Single-index Model as an Exploratory Tool

- This suggests the possibility of a more parsimonious model.

$$Y(t) = \mu(\mathrm{t}) + \mathrm{f}(\beta^T X(t)) + \varepsilon.$$

- $\mu(\mathrm{t})$ could be parametric.
- Random effects could be added.

Conclusion

- Common marginal models for longitudinal data use the additive form, and employ parametric models for both the mean and covariance function.

  - Both parametric forms are difficult to detect for sparse and noisy longitudinal data.

- A semiparametric model, such as the single index model, may be useful as an exploratory tool to search for a parametric model.

Conclusion

- Our approach allows for multiple indices.
- Could extend the random effects model to make the eigenfunctions covariate dependent

  Jiang and Wang (2010, AOS)

- Could use an additive model instead of index model.

Thank You