Functional Latent Feature Models With A Single-Index Interaction

Naisyin Wang
Department of Statistics
University of Michigan

Yehua Li and Raymond J. Carroll
University of Georgia and Texas A&M University

PL Hsu 100 Conference, Peking University

July 4, 2010

Our Goals

- Generalize the Standard FDA Model to allow for low-dimensional interactions
- Quantify the cost due to estimating functional principal components
- Analyze a nutrition data set

Outline

- The standard FDA model (Ramsay & Silverman, 2003)
- Allowing for low-dimensional interactions
- The Procedure
- Results with Known Basis Functions
- Results with Estimated Basis Functions
- Simulation
- Data Analysis

Some Basic Models

- Ordinary nonparametric regression of Y on X:
  $E(Y|X) = \mu\{\theta(X)$
- The generalized partially linear model:
  $E(Y|X, Z) = \mu\{\theta(X) + Z^{\mathrm{T}}\beta\}$.
- The single-index model (Stoker and Härdle, 89):
  $E(Y|X) = \mu\{\theta(X^{\mathrm{T}}\kappa)\}$
- The generalized partially linear single index model (GPLSIM, Carroll, Fan, Gijbels, Wand, 97):
  $E(Y|X, Z) = \mu\{\theta(X^{\mathrm{T}}\kappa) + Z^{\mathrm{T}}\beta\}$.
- We want to do the GPLSIM for functional data and also allow for interactions

Data Structure

- Scalar Response Y_i
- Longitudinal covariate $X_i(t)$, with mean $\mu_X(t)$
- Possibly observed covariate $W_i(t) = X_i(t) + U_i(t)$, with $U_i(\cdot)$ white noise.
- Fixed covariate Z_i, includes a 1.0 for an intercept
- Relate Y to $\{X(\cdot), Z\}$

Standard FDA Model

- The standard FDA model says that there is an unknown function $\mathfrak{A}(\cdot)$ such that

$$Y_i = \int \mathfrak{A}(t) X_i(t) dt + Z_i^{\mathrm{T}} \beta + \epsilon_i.$$

- We start with the simplest case that $X_i(\cdot)$ is entirely observed.

Standard FDA Model

- Let $\Psi(t) = \{\psi_1(t), \psi_2(t), \cdots, \psi_p(t)\}^{\mathrm{T}}$ be p orthonormal functions.
- A standard model is that

$$\mathfrak{A}(t) = \Psi^{\mathrm{T}}(t)\alpha.$$

- Define for $j = 1, ..., p$,

$$\xi_{ij} = \int \psi_j(t)\{X_i(t) - \mu_X(t)\}dt.$$

- Then we have that with $\xi_i = (\xi_{i1}, ..., \xi_{ip})^{\mathrm{T}}$,

$$\int \mathfrak{A}(t)X_i(t)dt = \alpha_0 + \xi_i^{\mathrm{T}}\alpha_1.$$

Functional Latent Features

- ξ is a vector of latent variables, which we refer as the latent feature of the functional data.

- Two typical structural considerations in functional data analysis.
  - Fixed bases: (ψ_j) are known basis functions, such as Fourier or wavelet basis functions.
  - Data-driven bases, e.g., the (ψ_j) are the leading principal components of $X(t)$.

Allowing Interactions

- Let Z_1 be a subset of Z. The standard FDA model has no interactions between $X(\cdot)$ and Z.
- We propose a simple alternative, namely that instead of $\mathfrak{A}(t) = \Psi^{\mathrm{T}}(t)\alpha$, we have

$$\mathfrak{A}(t, Z_1, \theta) = \Psi^{\mathrm{T}}(t)\alpha(Z_1, \theta).$$

- To obtain dimension reduction, we further write

$$\mathfrak{A}(t, Z_1^{\mathrm{T}}\theta) = \Psi^{\mathrm{T}}(t)\{\alpha_1 + \mathcal{S}(Z_1^{\mathrm{T}}\theta)\alpha_2\}$$

where $\mathcal{S}(\cdot)$ is of one dimension.

Thoughts Behind the Interaction Structure

- Recall that we reduce $\Psi_1(t)\alpha_1(Z_1,\theta) + \cdots + \Psi_p(t)\alpha_p(Z_1,\theta)$ to
$$\Psi^{\mathrm{T}}(t)\{\alpha_1 + \mathcal{S}(Z_1^{\mathrm{T}}\theta)\alpha_2\}.$$
- We consider only ONE $\mathcal{S}(\cdot)$ instead of p of them.
- It would have been over-ambitious to estimate p $\mathcal{S}(\cdot)$ due to the unstability of estimation of a higher PCA direction.
- $Z_1^{\mathrm{T}}\theta$ explains the main "Z-direction" on which $X(t)$ interact with Z.
- We use α_2 (parametric) to accommodate differences in interactions among different PCA directions.

Allowing Interactions

- With the single index model and the known ξ's, we are left with the model

$$Y = \xi^{\mathrm{T}}\{\alpha_1 + \mathcal{S}(Z_1^{\mathrm{T}}\theta)\alpha_2\} + Z^{\mathrm{T}}\beta + \epsilon$$

- This is a semiparametric single index model with an unknown function $\mathcal{S}(\cdot)$.

- There are various things needed for identifiability, we use $\|\theta\| = \|\alpha_2\| = 1$, $\mathrm{E}\{\mathcal{S}(\theta^T Z_1)\} = 0$.

Allowing Interactions With Non-Gaussian Data

- The material is easily generalized to quasilikelihood.
- Let the mean and variance of Y given (X, Z) be given as

$$E(Y_i|X_i, Z_i) = \mu\left[\xi_i^{\mathrm{T}}\{\alpha_1 + \mathcal{S}(Z_{1i}^{\mathrm{T}}\theta)\alpha_2\} + Z_i^{\mathrm{T}}\beta\right];$$

$$\mathrm{var}(Y_i|X_i, Z_i) = V\left(\mu\left[\xi_i^{\mathrm{T}}\{\alpha_1 + \mathcal{S}(Z_{1i}^{\mathrm{T}}\theta)\alpha_2\} + Z_i^{\mathrm{T}}\beta\right]\right).$$

Algorithm

- The basic mean model is that

$$E(Y_i|X_i, Z_i) = \mu\left[\xi_i^{\mathrm{T}}\{\alpha_1 + \mathcal{S}(Z_{1i}^{\mathrm{T}}\theta)\alpha_2\} + Z_i^{\mathrm{T}}\beta\right].$$

- The terms ξ_i are still known to us, for now.
- There are options in fitting quasilikelihood, we follow the use of the **MAVE** method of Xia and Härdle (06).
- The MAVE idea is to do two steps.
- **Step 1**: local quasilikelihood backfitting with full multivariate kernel weights to get a consistent estimate of θ, say $\widetilde{\theta}$.
- **Step 2**: local quasilikelihood backfitting with univariate kernel weights with arguments $Z_1^{\mathrm{T}}\widetilde{\theta}$ to get efficient estimates.

Results when the Basis Functions are Known

- We have shown that the **final estimated function** has the **usual type of bias-variance decomposition**. Under given conditions, $\|\widetilde{\Theta} - \Theta_0\| \to 0$ with probability 1.
- We have shown that the final estimates of all the parameters has $\sqrt{n}$ **rate of convergence**, is **asymptotically normal**, and has an asymptotic covariance matrix of the form $A^{-1}BA^{-1}$, where A^{-1} is a generalized inverse. Under some conditions,

$$\sqrt{n}(\widehat{\Theta} - \Theta_0) \to \text{Normal}(\mathbf{0}, \mathcal{A}^{-}\mathcal{B}\mathcal{A}^{-}),$$
$$\sqrt{nh}\{\widehat{\mathcal{S}}(u) - \mathcal{S}_0(u) - h^2\mathcal{S}^{(2)}(u)\sigma_k^2/2\} \to \text{Normal}\{0, \sigma_{\mathcal{S}}^2(u)\},$$

What if the Basis Functions are Unknown?

- The most widely used data driven method uses **functional principal components** (FPCA).
- Define $\mu_X(t) = E\{X(t)\}$,

$$R(s,t) = \text{cov}\{X(s), X(t)\} = \sum_{k=1}^{\infty} \omega_k \psi_k(s)\psi_k(t).$$

- The **Karhunen-Loève expansion** says that

$$X_i(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_{ij}\psi_j(t),$$

where $\mathrm{E}(\xi_j) = 0$, $\text{cov}(\xi_{ij}, \xi_{ik}) = I(j = k)\omega_j$.

Discrete functional data

- We **do not observe the process** $X_i(\cdot)$.
- Instead, **we observe discrete observations with noise**, i.e., at observation times T_{ij}, we observe

$$W_{ij} = X_i(T_{ij}) + U_{ij}, \quad j = 1, \cdots, m_i,$$

where U_{ij} are independent zero-mean errors independent of $X_i(\cdot)$ and Z_i, with $\text{var}\{U_i(t)\} = \sigma_u^2$.

Discrete functional data

- It is commonly assumed that there are infinite number non-zero eigenvalues.
- A more realistic assumption is that Y only depends on a **finite number** p of the leading principal components.
- To push this through, we have to estimate $\text{cov}\{X(s), X(t)\}$; we use kernel methods and W to do this.
- We have $\widehat{R}(s,t) = \widehat{\sigma}_{XX}(s,t) - \widehat{\mu}_X(s)\widehat{\mu}_X(t)$, $\sigma_{XX}(s,t) = \mathrm{E}\{X(s)X(t)\}$.
- We then get $\widehat{\sigma}_u^2$ based on $\sigma_w^2(t) = \text{var}\{W(t)\} = R(t,t) + \sigma_u^2$.
- Estimate of PC-scores: numerical integration (NI) method (Müller, 2007).

Effect of Estimating Basis Function

- Remember our basic model

$$E(Y_i|X_i, Z_i) = \mu\left[\xi_i^{\mathrm{T}}\{\alpha_1 + \mathcal{S}(Z_{1i}^{\mathrm{T}}\theta)\alpha_2\} + Z_i^{\mathrm{T}}\beta\right].$$

- We now substitute estimates $\widehat{\xi}_i$ for ξ_i.
- We have shown **under certain technical conditions** that estimating the PC scores **will not** impact the variability of the estimate of $\mathcal{S}(\cdot)$ but **will increase** the variability of the parameter estimates.

Simulation

- We generate Gaussian longitudinal process $X(t)$ for $t \in [0, 1]$, with mean function $\mu_X(t) = (t - 0.6)^2 - 0.1$.
- The covariance function of the process had 2 principal components, $\psi_1(t) = 1$, $\psi_2(t) = \sqrt{2}\sin(2\pi t)$, and the eigenvalues were $\omega_1 = 1.0$ and $\omega_2 = 0.6$.
- Assume $m = 30$ discrete observations on each curve, with random observation time points being uniformly distributed on the interval $[0, 1]$.
- Discrete observations on X are contaminated with zero-mean Gaussian error with variance $\sigma_u^2 = 0.1$.

Simulation

- We generated a binary response Y from the logistic model

$$\text{pr}(Y = 1|Z, \xi) = H\{\alpha_1^{\mathrm{T}}\xi + \mathcal{S}(\theta^{\mathrm{T}}Z_1)\alpha_2^{\mathrm{T}}\xi + \beta^{\mathrm{T}}Z\}.$$

- $Z = (1, Z_1^{\mathrm{T}}, Z_2)^{\mathrm{T}}$, Z_1 is a 2-dimensional random vector with a uniform distribution on $[0, 1]^2$, and Z_2 is a binary variable with $\text{pr}(Z_2 = 1) = 0.5$.

- We let $\mathcal{S}(\cdot)$ be a sine bump function similar to that used in Carroll et al. (1997), $\mathcal{S}(t) = 2\sin\{(t - c_1)/(c_2 - c_1)\}$, where $c_1 = 1/\sqrt{2} - 1.645/\sqrt{12}$ and $c_2 = 1/\sqrt{2} + 1.645/\sqrt{12}$.
- We let $n = 100$ and repeated the simulation 200 times.

Results of the Simulation

- We found some variance inflation due to estimating the PC scores.
- There was a 50% variance inflation for estimating the parameters associated with the second principal component
- Main effect parameters were badly biased if the interaction was ignored.

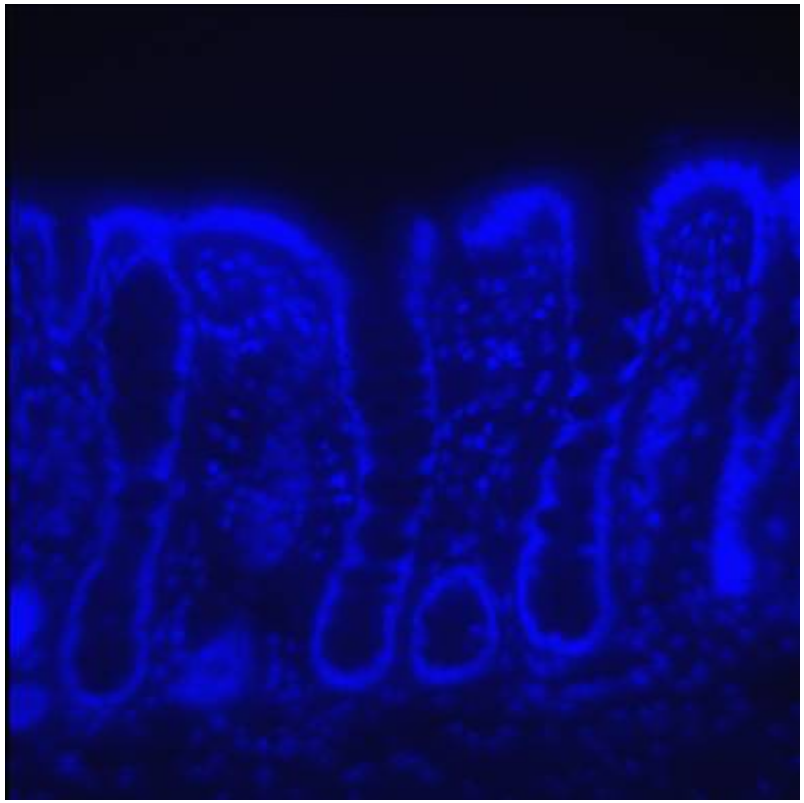
Simulation results

	Full Model				Reduced Model			
	Truth	Mean	SD	Bias		Mean	SD	Bias
β_0	-1	-1.038	0.418	-0.038	β_0	-0.939	0.368	0.061
β_1	2	2.062	0.494	0.062	β_1	1.757	0.429	-0.243
β_2	-2	-2.021	0.466	-0.021	β_2	-1.741	0.406	0.259
β_3	2	2.076	0.299	0.076	β_3	1.888	0.270	-0.112
α^*_{11}	1.7986	1.755	0.194	-0.044	α_1	1.497	0.171	-0.301
α^*_{12}	-0.0014	-0.096	0.207	-0.095	α_2	-0.168	0.185	-0.167
α_{21}	0.7071	0.678	0.188	-0.029				
α_{22}	0.7071	0.666	0.249	-0.041				
θ_1	0.7071	0.699	0.116	-0.008				
θ_1	0.7071	0.696	0.116	-0.011				

Colon Carcinogenesis Data

- Beneath the colon tissue, there are pore structures called 'colonic crypts', see next slide.
- A crypt typically contains 25-30 cells. Functional covariate, $X(t)$,: *p27* measured at cell level, and t: relative cell location within the crypt.
- *p27* is a protein that inhibits the cell cycle.
- We sampled about 20 crypts from each of the 12 rats, with a total of $n = 249$ crypts.
- There are 2 diet groups (corn oil diet or fish oil diet) and 2 treatment groups (with/without butyrate supplement).

Colonic crypts



Colon Carcinogenesis Data

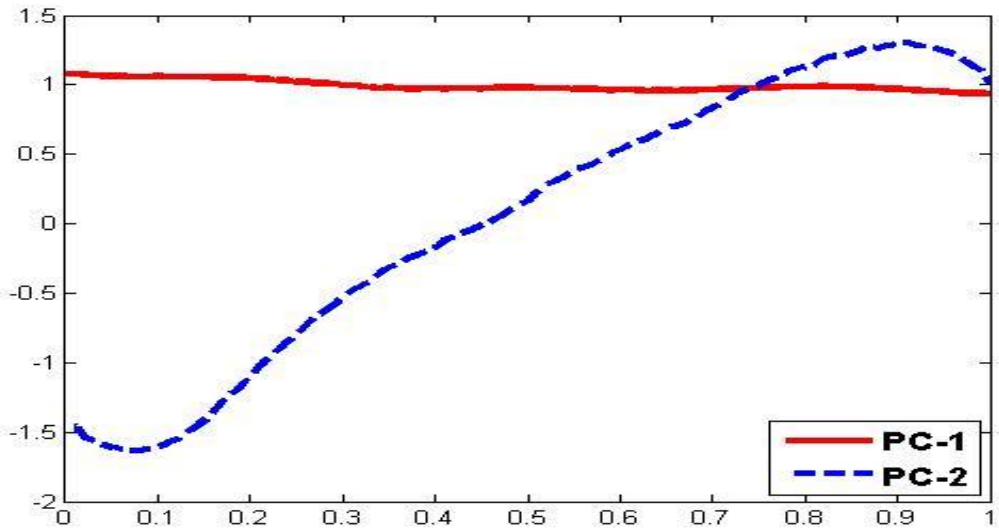
Goal of the study: to build a regression model between $Y =$ mean apoptotic (programmed cell death) rate within a crypt and

- X=*p27* profile curve within a crypt
- Z=environment variables (diet, treatment), and mean proliferation rate in each crypt
- interaction between X and Z. so that Z_1 is the same as Z but without the intercept.

PCA for *p27* data

- The first 3 eigenvalues are 0.871, 0.019 and 0.005 respectively.
- In our regression, we will only use the first 2 PC's.

PCA for *p27* data



The Shape of the Interaction Surface

- Our model is

$$\begin{aligned} Y_i &= \int \mathfrak{A}(t, Z_{i1}^{\mathrm{T}}\theta) X_i(t) dt + Z_i^{\mathrm{T}}\beta + \epsilon_i \\ \mathfrak{A}(t, Z_1^{\mathrm{T}}\theta) &= \Psi^{\mathrm{T}}(t)\{\alpha_1 + \mathcal{S}(Z_1^{\mathrm{T}}\theta)\alpha_2\} \end{aligned}$$

- The shape of $\mathfrak{A}(t, Z_1^{\mathrm{T}}\theta)$ varies dramatically based on where the cell is located, see next.

Estimated functional coefficient function

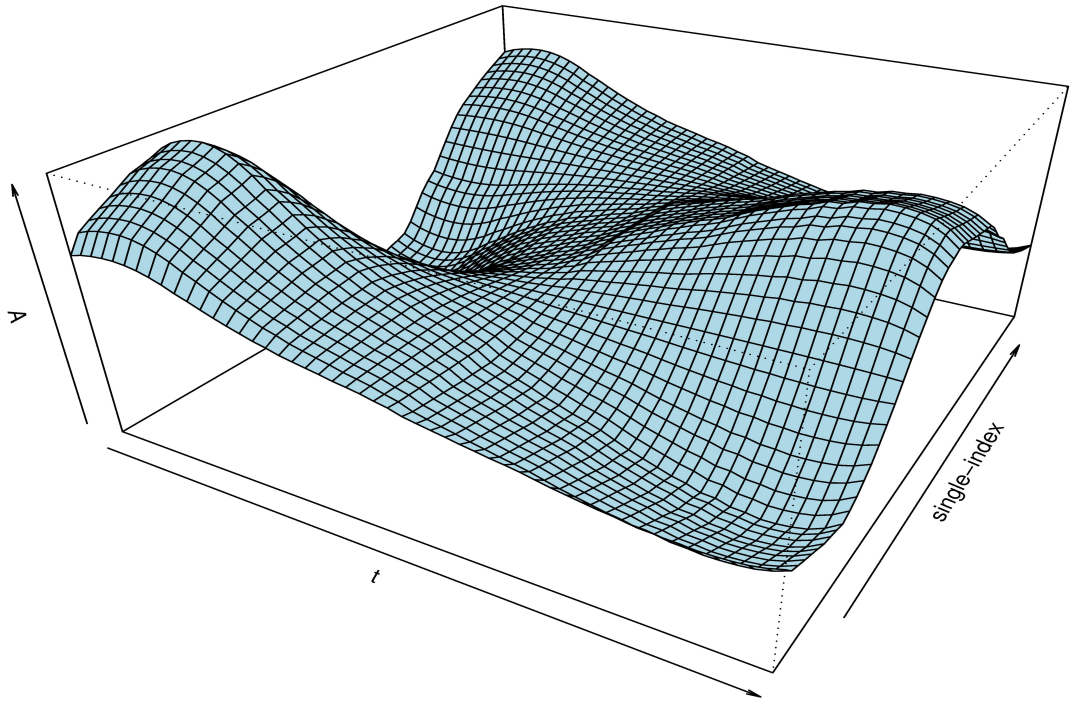


Figure: Semiparametric estimator of $\mathfrak{A}(t, \theta^{\mathrm{T}} Z_1)$ in the colon carcinogenesis p27 data.

The Single Index

- The main effects for Z in $Z^{\mathrm{T}}\beta$ were modest but dominated by an increase in Apoptosis for the fish oil diet.
- The interaction θ was dominated by a highly significant effect of butyrate exposure and the proliferative index.
- If we ignore the possible interaction, NOTHING is statistically significant in the resulting model, including fish oil intake.

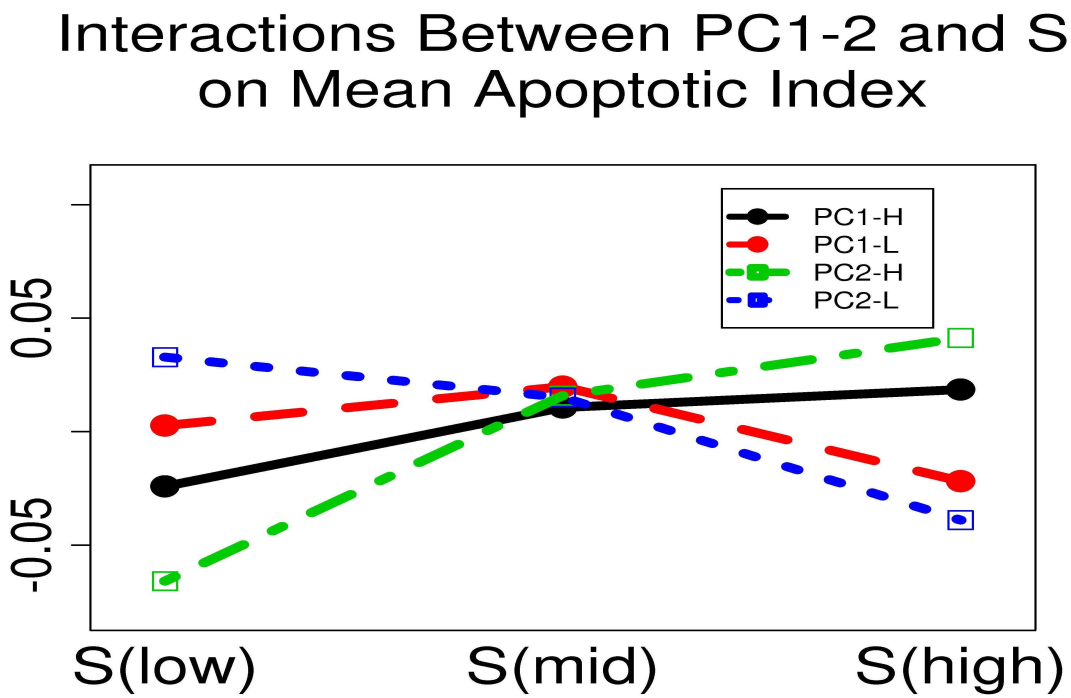
Interaction

- If $\mathcal{S}(\cdot)$ is constant, then there should be no relationship between it and X, the p27 biomarker or on Y
- By implication, there should thus be no relationship between $\mathcal{S}(\cdot)$ and the PC scores.
- We thus first divide the function estimates $\widehat{\mathcal{S}}(\cdot)$ into three subgroups:
  - high values of $\mathcal{S}(\cdot)$, (S_{high}: $\mathcal{S} > 1.5$);
  - low values of $\mathcal{S}(\cdot)$ (S_{low}: $\mathcal{S} < -1.5$);
  - and the ones that are in between (S_{mid}).

Interaction

- We then dichotomized each of the two p27 PC scores according whether they belong to the top or bottom 50% of the scores.
- This produces 4 groups in the data: PC1-Low, PC1-High, PC2-Low and PC2-High.
- If $\mathcal{S}(\cdot)$ is constant, there there should be no systematic relationship between $\widehat{\mathcal{S}}(\cdot)$ and the average apoptotic index of these four PC groups.
- The next graph shows a clear relationship.

Interpretation of interaction (Cont.)



Summary

- We proposed a new class of functional generalized linear models, which allows multivariate covariates and their interaction with functional latent features.
- The interaction between the functional and multivariate predictor is modeled with a single-index structure, which is flexible yet parsimonious.
- We propose a MAVE-type of estimating procedure based on local quasi-likelihood.
- We show an asymptotic variance inflation on parameter estimates due to estimating PC scores and basis functions. This may be an important issue in practice.

Fitted model

	α_{11}	α_{12}	α_{21}	α_{22}
Estimate	-0.0004	0.0235	-0.0480	0.9988
SE	0.0108	0.1003	0.1653	0.0767
p-value	0.9719	0.8145	0.7714	0.0000
	β_0	β_{fish}	β_{buty}	β_{prolif}
Estimate	0.2627	0.0514	0.0223	-0.0062
SE	0.0247	0.0193	0.0201	0.0135
p-value	0.0000	0.0078	0.2667	0.6484
	θ_{fish}	θ_{buty}	θ_{prolif}	
Estimate	0.4208	-0.7143	-0.5592	
SE	0.2847	0.2419	0.2005	
p-value	0.1394	0.0031	0.0053	