Functional Generalized Additive Models

A new model for regression with functional predictors

Mathew McLean

PhD Candidate
Cornell University
School of Operations Research and Information Engineering
mwm79@cornell.edu

September 22, 2012

Collaborators

Joint work with

- David Ruppert - Cornell University
- Giles Hooker - Cornell University
- Ana-Maria Staicu - NC State University
- Fabian Scheipl - LMU Munich

Outline

Functional Regression - Setup and Notation

- Goal: From N samples predict Y using smooth function $X : \mathcal{T} \to \mathcal{X}$;
- $\mathcal{T}$ closed interval. $\mathcal{T} = [0, 1]$ w.l.o.g. Often, t is time
- For now, r.v Y is continuous and normally distributed
- $X(t)$ observed at finite number of points and presmoothed

Functional Linear Model (FLM)

The most commonly used functional regression model:

FLM

$$E(Y_i|X_i) = \beta_0 + \int_{\mathcal{T}} \beta(t) X_i(t)\, dt \qquad i = 1, \ldots, N$$

- $\beta(\cdot)$ is unknown smooth coefficient function
- $\mathrm{Var}(Y_i|X_i) = \sigma^2$

- Effect of X on Y is linear for each t (Easy to interpret)

FLM as limit of LM

Back to multivariate data

- Observe each function at J equally-spaced points

$$x_{ij} \equiv X_i(t_j); \qquad \beta_j \equiv \beta(t_j); \qquad j = 1\ldots, J$$

FLM as limit of LM

Back to multivariate data

- Observe each function at J equally-spaced points

$$x_{ij} \equiv X_i(t_j); \qquad \beta_j \equiv \beta(t_j); \qquad j = 1\ldots, J$$

- Linear Model:

$$E(Y_i|X_{i1},\ldots,X_{iK}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = \beta_0 + \sum_{j=1} \frac{\beta_j(t_j)}{J} X_i(t_j) J^{-1}$$

(think Riemann sum)

FLM as limit of LM

Back to multivariate data

- Observe each function at J equally-spaced points

$$x_{ij} \equiv X_i(t_j); \qquad \beta_j \equiv \beta(t_j); \qquad j = 1 \ldots, J$$

- Linear Model:

$$E(Y_i | X_{i1}, \ldots, X_{iK}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = \beta_0 + \sum_{j=1} \frac{\beta_j(t_j)}{J} X_i(t_j) J^{-1}$$

(think Riemann sum)

- Letting $J \to \infty$ we arrive at

$$E(Y_i | X_i) = \beta_0 + \int_{\mathcal{T}} \beta(t) X_i(t)\, dt$$

Functional Linear Model (FLM)

The most commonly used functional regression model:

FLM

$$E(Y_i|X_i) = \beta_0 + \int_{\mathcal{T}} \beta(t) X_i(t)\, dt \qquad i = 1, \ldots, N$$

- $\beta(\cdot)$ is unknown smooth coefficient function
- $\mathrm{Var}(Y_i|X_i) = \sigma^2$

- Effect of X on Y is linear for each t (Easy to interpret)
- Linear Model with an infinite number of predictors (limit of Riemann sum approximation)

Functional Linear Model (FLM)

The most commonly used functional regression model:

FLM

$$E(Y_i|X_i) = \beta_0 + \int_{\mathcal{T}} \beta(t) X_i(t)\, dt \qquad i = 1, \ldots, N$$

- $\beta(\cdot)$ is unknown smooth coefficient function
- $\mathrm{Var}(Y_i|X_i) = \sigma^2$

- Effect of X on Y is linear for each t (Easy to interpret)
- Linear Model with an infinite number of predictors (limit of Riemann sum approximation)
- Coefficient function commonly estimated in one of two ways
  1) Using B-splines and roughness penalty
  2) Using functional principal components analysis (fPCA)

Extending the FLM

- As with LM for multivariate data, FLM not general enough
  - Need model for more complex response-predictor relationships

Extending the FLM

- As with LM for multivariate data, FLM not general enough
  - Need model for more complex response-predictor relationships
- We desire a model that is:
  1) Highly flexible (low bias)
  2) Avoids curse of dimensionality (low variance)
  3) Easy to interpret (not a "black box")
  4) Has FLM as a special case

An Additive Model With Functional Predictor - FGAM

The model we propose is

FGAM

$$E(Y_i|X_i) = \theta_0 + \int_{\mathcal{T}} F\{X_i(t), t\}\, dt$$

unknown bivariate function $F : \mathcal{X} \times \mathcal{T} \to \mathbb{R}$

An Additive Model With Functional Predictor - FGAM

The model we propose is

FGAM

$$E(Y_i|X_i) = \theta_0 + \int_{\mathcal{T}} F\{X_i(t), t\}\, dt$$

unknown bivariate function $F : \mathcal{X} \times \mathcal{T} \to \mathbb{R}$

- Need to impose smoothness of $F(\cdot, \cdot)$ in x and t
  - Two parameters, λ_x and λ_t control function complexity

An Additive Model With Functional Predictor - FGAM

The model we propose is

FGAM

$$E(Y_i|X_i) = \theta_0 + \int_{\mathcal{T}} F\{X_i(t), t\}\, dt$$

unknown bivariate function $F : \mathcal{X} \times \mathcal{T} \to \mathbb{R}$

- Need to impose smoothness of $F(\cdot, \cdot)$ in x and t
  - Two parameters, λ_x and λ_t control function complexity
- If $F(x, t) = \beta(t)x$, we get the FLM

An Additive Model With Functional Predictor - FGAM

$$E(Y_i|X_i) = \theta_0 + \int_{\mathcal{T}} F\{X_i(t), t\}\, dt$$

- Define

$$x_{ij} \equiv X_i(t_j) \quad f_j(\cdot) \equiv F(\cdot, t_j) J^{-1}$$

and consider the additive model

$$E(Y_i|X_{i1}, \ldots, X_{iJ}) = \theta_0 + \sum_{j=1}^{J} f_j\{x_{ij}\} = \theta_0 + \sum_{j=1}^{J} F\{x_{ij}, t_j\} J^{-1}$$

An Additive Model With Functional Predictor - FGAM

$$E(Y_i|X_i) = \theta_0 + \int_{\mathcal{T}} F\{X_i(t), t\}\, dt$$

- Define

$$x_{ij} \equiv X_i(t_j) \quad f_j(\cdot) \equiv F(\cdot, t_j) J^{-1}$$

and consider the additive model

$$E(Y_i|X_{i1}, \ldots, X_{iJ}) = \theta_0 + \sum_{j=1}^{J} f_j\{x_{ij}\} = \theta_0 + \sum_{j=1}^{J} F\{x_{ij}, t_j\} J^{-1}$$

An Additive Model With Functional Predictor - FGAM

$$E(Y_i|X_i) = \theta_0 + \int_{\mathcal{T}} F\{X_i(t), t\}\, dt$$

- Define

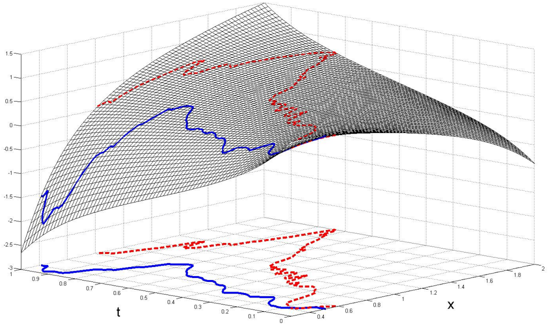
$$x_{ij} \equiv X_i(t_j) \quad f_j(\cdot) \equiv F(\cdot, t_j) J^{-1}$$

  and consider the additive model

$$E(Y_i|X_{i1}, \ldots, X_{iJ}) = \theta_0 + \sum_{j=1}^{J} f_j\{x_{ij}\} = \theta_0 + \sum_{j=1}^{J} F\{x_{ij}, t_j\} J^{-1}$$

- Let $J \to \infty$ arrive at FGAM

Example Estimated Surface



Estimated surface $\hat{F}(x,t)$ and two predictor curves.

- The solid curve belongs to a control and the dashed curve belongs to an MS patient.

Identifiability

- Define $F^*(x,t) = F(x,t) + g(t)$, where $\int_{\mathcal{T}} g(t)\,dt = 0$
  Notice that
  $$\int_{\mathcal{T}} F^*(x,t)\,dt = \int_{\mathcal{T}} F(x,t)\,dt$$

Need constraints to ensure identifiability and interpretability

Identifiability

- Define $F^*(x,t) = F(x,t) + g(t)$, where $\int_{\mathcal{T}} g(t)\,dt = 0$
  Notice that
  $$\int_{\mathcal{T}} F^*(x,t)\,dt = \int_{\mathcal{T}} F(x,t)\,dt$$

Need constraints to ensure identifiability and interpretability

- Also check for numerical rank deficiency during fitting
- Different constraints are possible, will affect c. bands

Model for $F(x, t)$

We use bivariate tensor product B-splines for $F(x, t)$

$$F(x, t) = \sum_{j=1}^{K_x} \sum_{k=1}^{K_t} \theta_{jk} B_j^X(x) B_k^T(t)$$

- $\{B_j^X(x) : j = 1, \ldots, K_x\}$ and $\{B_k^T(x) : k = 1, \ldots, K_t\}$ are low-rank, univariate B-spline bases
- Equally spaced knots, must specify degree of the spline and number of basis functions

Putting It Together

$$E(Y_i|X_i) = \theta_0 + \int_{\mathcal{T}} F\{X_i(t), t\}\, dt$$

$$F(x, t) = \sum_{j=1}^{K_x} \sum_{k=1}^{K_t} \theta_{jk} B_j^X(x) B_k^T(t)$$

- The model becomes

$$E(Y_i|X_i) = \theta_0 + \sum_{j=1}^{K_x} \sum_{k=1}^{K_t} \theta_{jk} Z_{jk}(i) = \mathbb{Z}\boldsymbol{\theta}$$

  - $Z_{jk}(i) = \int_{\mathcal{T}} B_j^X\{X_i(t)\} B_k^T(t)\, dt$
  - $\mathbb{Z}$ is $N \times (1 + K_x K_t)$ matrix of $Z_{jk}(i)$ with first column $\mathbf{1}$

Some Notes On Penalized Regression Splines

- Multiple smoothing parameters estimated simultaneously
- Fast, numerically stable fitting methods
- Easily extends to additional scalar or functional predictors
- Response can be from any exponential family of distributions

Some Notes On Penalized Regression Splines

- Multiple smoothing parameters estimated simultaneously
- Fast, numerically stable fitting methods
- Easily extends to additional scalar or functional predictors
- Response can be from any exponential family of distributions
- We use specific type known as P-splines (Marx & Eilers, 1996)
    - Other types of splines and penalties possible

Penalized Likelihood Estimation

Penalized least squares objective function is:
(ignoring constraints)

$$(\mathbf{Y} - \mathbb{Z}\boldsymbol{\theta})^T(\mathbf{Y} - \mathbb{Z}\boldsymbol{\theta}) + \lambda_x\boldsymbol{\theta}^T\mathbb{P}_x\boldsymbol{\theta} + \lambda_t\boldsymbol{\theta}^T\mathbb{P}_t\boldsymbol{\theta}$$

- $\mathbb{P} = \lambda_x\mathbb{P}_x + \lambda_t\mathbb{P}_t$ incorporates difference penalties on X and t
- Solution is

$$\widehat{\boldsymbol{\theta}} = (\mathbb{Z}^T\mathbb{Z} + \mathbb{P})^{-1}\mathbb{Z}^T\mathbf{Y}$$

Penalized Likelihood Estimation

Penalized least squares objective function is:
(ignoring constraints)

$$(\mathbf{Y}-\mathbb{Z}\boldsymbol{\theta})^T(\mathbf{Y}-\mathbb{Z}\boldsymbol{\theta})+\lambda_x\boldsymbol{\theta}^T\mathbb{P}_x\boldsymbol{\theta}+\lambda_t\boldsymbol{\theta}^T\mathbb{P}_t\boldsymbol{\theta}$$

- $\mathbb{P}=\lambda_x\mathbb{P}_x+\lambda_t\mathbb{P}_t$ incorporates difference penalties on X and t
- Solution is
$$\widehat{\boldsymbol{\theta}}=(\mathbb{Z}^T\mathbb{Z}+\mathbb{P})^{-1}\mathbb{Z}^T\mathbf{Y}$$
- Hat Matrix
$$\mathbb{H}=\mathbb{Z}(\mathbb{Z}^T\mathbb{Z}+\mathbb{P})^{-1}\mathbb{Z}^T$$
$\mathrm{tr}\,(\mathbb{H})=$ effective degrees of freedom.
  - Measures effective number of parameters in fit
  - Value of λ not informative for quantifying amount of smoothing
  - $1+d_xd_t\le\mathrm{tr}(\mathbb{H})\le 1+K_xK_t$ where d_x, d_t are order of penalties

Penalized Likelihood Estimation

Penalized least squares objective function is:
(ignoring constraints)

$$(\mathbf{Y} - \mathbb{Z}\boldsymbol{\theta})^T(\mathbf{Y} - \mathbb{Z}\boldsymbol{\theta}) + \boldsymbol{\theta}^T\mathbb{P}\boldsymbol{\theta}$$

- $\mathbb{P} = \lambda_x\mathbb{P}_x + \lambda_t\mathbb{P}_t$ incorporates difference penalties on X and t
- Note: Can handle "$p > N$"
  - No automatic model selection, though this is possible
- Must check for rank deficiency during fitting

Penalized Likelihood Estimation

Penalized least squares objective function is:
(ignoring constraints)

$$(\mathbf{Y} - \mathbb{Z}\boldsymbol{\theta})^T(\mathbf{Y} - \mathbb{Z}\boldsymbol{\theta}) + \boldsymbol{\theta}^T\mathbb{P}\boldsymbol{\theta}$$

- $\mathbb{P} = \lambda_x\mathbb{P}_x + \lambda_t\mathbb{P}_t$ incorporates difference penalties on X and t
- Note: Can handle "$p > N$"
  - No automatic model selection, though this is possible
- Must check for rank deficiency during fitting
- Possible to formulate as a mixed effects model

Penalized Likelihood Estimation

Penalized least squares objective function is:
(ignoring constraints)

$$(\mathbf{Y} - \mathbb{Z}\boldsymbol{\theta})^T(\mathbf{Y} - \mathbb{Z}\boldsymbol{\theta}) + \boldsymbol{\theta}^T\mathbb{P}\boldsymbol{\theta}$$

- $\mathbb{P} = \lambda_x\mathbb{P}_x + \lambda_t\mathbb{P}_t$ incorporates difference penalties on X and t
- Note: Can handle "$p > N$"
    - No automatic model selection, though this is possible
- Must check for rank deficiency during fitting
- Possible to formulate as a mixed effects model
- Smoothing parameters selected using GCV

Variance of Estimated Surface

- We use generalization of Bayesian estimator of Wabha (1983)

$$\mathrm{var}(\widehat{\boldsymbol{\theta}}) = \hat{\sigma}^2(\mathbb{Z}^T\mathbb{Z} + \mathbb{P})^{-1}$$

  $\hat{\sigma}^2$ is a consistent estimator of σ^2

- Accounts for bias, but assumes fixed smoothing parameter
- Variance of estimated surface $\widehat{\mathbf{F}} = \mathbb{B}\widehat{\boldsymbol{\theta}}$ is

$$\mathrm{var}(\widehat{\mathbf{F}}) = \mathbb{B}\mathrm{var}(\widehat{\boldsymbol{\theta}})\mathbb{B}^T$$

  - $\mathbb{B}$ is matrix of B-spline evaluations over grid of x, t values

Variance of Estimated Surface

- We use generalization of Bayesian estimator of Wabha (1983)

$$\mathrm{var}(\widehat{\boldsymbol{\theta}}) = \hat{\sigma}^2(\mathbb{Z}^T\mathbb{Z} + \mathbb{P})^{-1}$$

$\hat{\sigma}^2$ is a consistent estimator of σ^2

- Accounts for bias, but assumes fixed smoothing parameter
- Variance of estimated surface $\widehat{\mathbf{F}} = \mathbb{B}\widehat{\boldsymbol{\theta}}$ is

$$\mathrm{var}(\widehat{\mathbf{F}}) = \mathbb{B}\mathrm{var}(\widehat{\boldsymbol{\theta}})\mathbb{B}^T$$

  - $\mathbb{B}$ is matrix of B-spline evaluations over grid of x, t values
- C.I. using this estimator has good "average" performance
  - Coverage close to nominal when averaged across all $\mathbf{x}$ and $\mathbf{t}$,
  - Exception is when bias becomes too large (over-smooth)
  - Coverage can still be poor at individual x_i and t_j values
- Accounting for identifiability constraints improves coverage

Testing for Constant Surface or FLM

- Notice $\frac{\partial^2}{\partial x^2}F(x,t) = 0$ for all x and t implies

$$F(x,t) = \beta(t)x$$

- Can construct confidence bands for $\frac{\partial^2}{\partial x^2}F(x,t)$ to check FLM
- Easy to do since derivatives of B-splines are easy to compute

Testing for Constant Surface or FLM

- Notice $\frac{\partial^2}{\partial x^2} F(x,t) = 0$ for all x and t implies

$$F(x,t) = \beta(t)x$$

- Can construct confidence bands for $\frac{\partial^2}{\partial x^2} F(x,t)$ to check FLM
- Easy to do since derivatives of B-splines are easy to compute
- Can test $H_0 : \boldsymbol{\theta} = \mathbf{0}$ and $H_0 : \mathbf{F} = \mathbf{0}$ using sandwich estimator

Transforming the Functional Predictor

Idea: Transform functional predictor, $X(t)$, to say, $G_t(x) = G \circ X(t)$

- The new surface to be estimated is $F(g, t)$
- Estimation procedure is the same

Transforming the Functional Predictor

Idea: Transform functional predictor, $X(t)$, to say, $G_t(x) = G \circ X(t)$

- The new surface to be estimated is $F(g, t)$
- Estimation procedure is the same
- Why transform $X(t)$?
  - Improve predictive performance
  - Improve numerical stability (no $\mathbf{0}$ columns in design matrix)
  - Ensure new data falls inside range of X marginal basis
  - Ex. Quantile transformation: $\widehat{G}_t(x) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{X_i(t) < x\}}$

Implementation

- λ's chosen using algorithms in `mgcv` package of S. Wood
- Our code will soon be available in R package `refund`

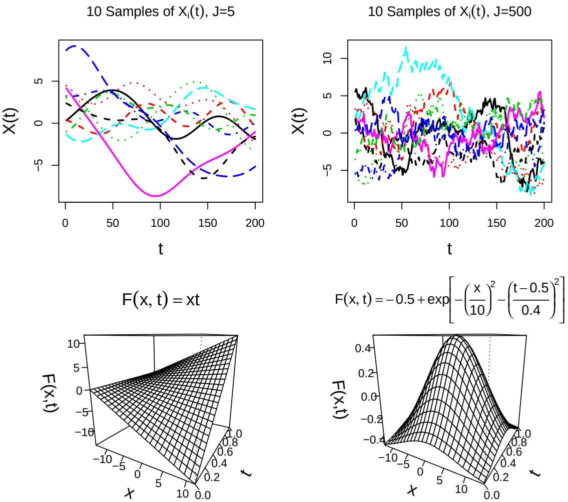
Functional Regression Models Used

- FLM with roughness penalty (FLM1)
- FLM with fPCA (FLM2)
- Functional Additive Model of Yao+Müller:
  - GAM in f.p.c. scores (FAM)
- Fully nonparametric kernel estimator of Ferraty+Vieu (FV):

$$\widehat{r}(X) = \frac{\sum_{i=1}^{N} Y_i \, K\left\{h^{-1} d(X, X_i)\right\}}{\sum_{i=1}^{N} K\left\{h^{-1} d(X, X_i)\right\}},$$

  - $K(\cdot)$ is an asymmetrical kernel with bandwidth h,
  - d is a semimetric
- FGAM
  - Using original predictor functions
  - Using quantile transformed predictor functions

Simulation Setup



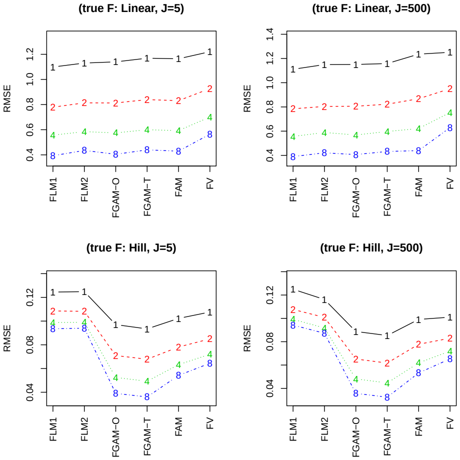
- σ^2 chosen so that SNR$=\dfrac{\text{var}\left[\int_T F\{X(t),t\}\,dt\right]}{\sigma^2}$ is constant

Predictive Performance - Median RMSE

- 1000 simulations with 67 training samples, 33 test samples
- Four different empirical signal to noise ratios: 1, 2, 4, 8

Predictive Performance - Median RMSE

- 1000 simulations with 67 training samples, 33 test samples
- Four different empirical signal to noise ratios: 1, 2, 4, 8



Diffusion Tensor Imaging

- Interested in how Multiple Sclerosis affects cognitive function
- Compare brains of MS patients with healthy controls
- $X(t)$ is measure of diffusion of water in particular brain tract
- 3 functional measurements summarizing the diffusion:
  - Parallel diffusivity - largest eigenvalue
  - Perpendicular diffusivity - average of two other eigenvalues
  - Fractional anisotropy (=0 if isotropic diffusion)

Diffusion Tensor Imaging

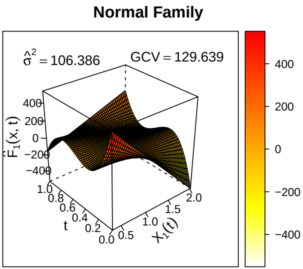
- Interested in how Multiple Sclerosis affects cognitive function
- Compare brains of MS patients with healthy controls
- $X(t)$ is measure of diffusion of water in particular brain tract
- 3 functional measurements summarizing the diffusion:
  - Parallel diffusivity - largest eigenvalue
  - Perpendicular diffusivity - average of two other eigenvalues
  - Fractional anisotropy (=0 if isotropic diffusion)
- Response is PASAT score: a cognitive test scored from 0-60,
  - Administered to cases only
  - MS patients known to perform poorly on this test

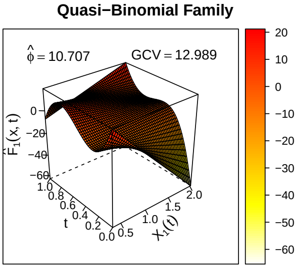
Estimated Surface

Two fits: assuming 1) normal and 2) quasi-binomial family

- $K_x = 6, \; K_t = 7,$
  $d_x = d_t = 2,$
  $\gamma = 1.4,$
  Cubic B-splines
- $X(t)$: Parallel Diff.

Estimated Surface

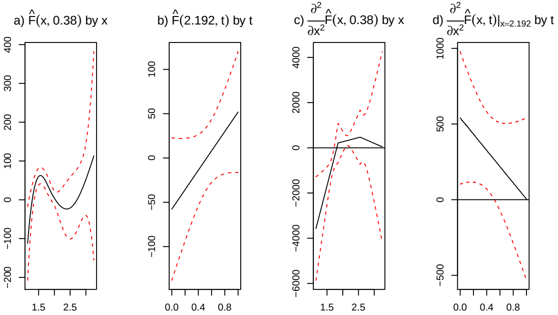




Two fits: assuming 1) normal and 2) quasi-binomial family

- $K_x = 6,\ K_t = 7,$
  $d_x = d_t = 2,$
  $\gamma = 1.4,$
  Cubic B-splines
- $X(t)$: Parallel Diff.
- Both models have similar out-of-sample prediction performance

Fixed slices of $\widehat{F}$ and $\partial^2/\partial x^2 \widehat{F}(x, t)$



- Untransformed parallel diffusivity as functional predictor

Leave-One-Curve-Out Prediction Error

- RMSE$=\left[N^{-1}\sum_{i=1}^{N}(y_i-\widehat{y}_{(i)})^2\right]^{1/2}$,

$\widehat{y}_{(i)}$ is the prediction for y_i when it is left out of training set

Leave-One-Curve-Out Prediction Error

- RMSE$=\left[N^{-1}\sum_{i=1}^{N}(y_i-\widehat{y}_{(i)})^2\right]^{1/2}$,

$\widehat{y}_{(i)}$ is the prediction for y_i when it is left out of training set

Measurement	FGAM-O	FGAM-T	FLM1	FLM2	FV	FAM
Perp. Diffusivity	12.22	10.46	10.98	11.27	11.16	11.71
Frac. Anisotropy	12.55	11.60	11.87	11.91	12.11	12.70
Para. Diffusivity	11.94	12.09	12.32	12.24	11.97	11.86

- FGAM with transformation seems to perform best here

Future Work on FGAM

- Work with sparse, noisy predictor measurements
- Formal test for checking for FLM
- Ability to fit models with many more parameters than data
- Model selection with several functional predictors
- Alternatives to GCV for choosing smoothing parameters

Summary

Proposed the Functional Generalized Additive Model:
A new model for predicting scalar using functional predictors.

FGAM is

- Intuitive extension of additive models to functional data
- Highly flexible AND highly interpretable
- Easily estimated using penalized regression splines
- Serves as useful diagnostic for checking FLM

Advertising

- For further details and a list of references see

M.W. McLean, G. Hooker, A.-M. Staicu, F. Scheipl, D. Ruppert. Functional Generalized Additive Models. *Journal of Computational and Graphical Statistics, to appear*

- A copy of the paper and R code can be obtained from

`http://people.orie.cornell.edu/mwm79/`

- See package **`refund`** available on CRAN to fit FGAM in R
- Please send questions and feedback to

`mwm79@cornell.edu`