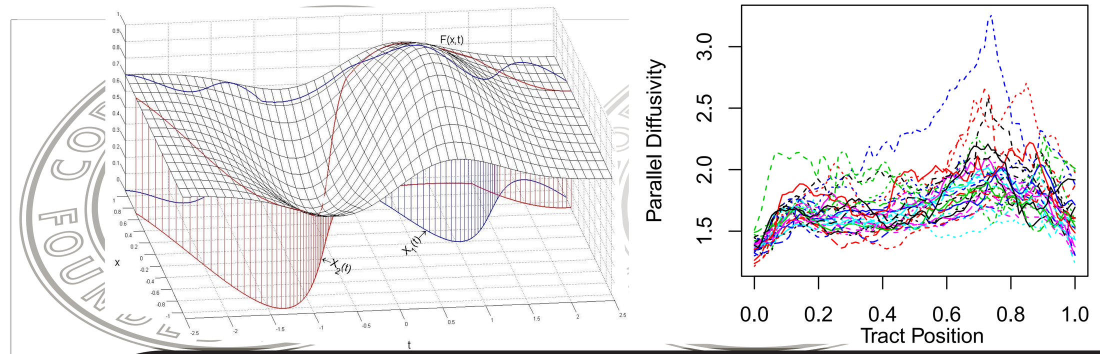


Functional Generalized Additive Models

Mathew McLean†¶, Giles Hooker†, Ana-Maria Staicu‡, Fabian Scheipl♯, David Ruppert†

†Cornell University, ‡NC State University, ♯LMU Munich, ¶mwm79@cornell.edu

Motivating Dataset

- Diffusion Tensor Imaging (DTI) study comparing the white matter tracts of Multiple Sclerosis (MS) patients with healthy controls
- Each scan produces a signal/function $X(t)$ of tract position t
- Goal: use $X(t)$ to predict (scalar) health outcome Y and identify which portions of the tract are most important for the prediction

Standard Model: FLM

- Assume that the predictor, $X(\cdot)$, is observed at a dense grid of points on closed interval $\mathcal{T}$.
- Most commonly used model is

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

where $\beta(\cdot)$ is an unknown smooth regression function and $X(t)$ is a functional covariate.
- For each t effect of $X_i(t)$ on Y_i is linear.

New Model: FGAM

- Linearity assumption of FLM is often too strong
- We desire a more flexible model that remains easy to understand and estimate
- We propose the model

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

where $F(x,t)$ is an unknown, smooth function
- Two tuning parameters control complexity of F
- Need identifiability constraints

Intuition From Multivariate Data

- Let $x_{ij} = X_i(t_j)$; $\quad \beta_j = \beta(t_j)$; $\quad j = 1, \ldots, J$
- Thinking of a Riemann sum, consider

$$E(Y_i|\mathbf{X}_i) = \sum_{j=1}^{p} \beta_j x_{ij} = \sum_{j=1} \frac{\beta_j(t_j)}{J} X_i(t_j) J^{-1}$$

- FLM obtained as limit of this LM as $J \to \infty$.
- Now define $f_j(\cdot) = F(\cdot, t_j)J^{-1}$ and consider AM

$$E(Y_i|\mathbf{X}_i) \sum_{j=1}^{J} f_j\{x_{ij}\} = \sum_{j=1}^{J} F\{x_{ij}, t_j\}J^{-1}$$

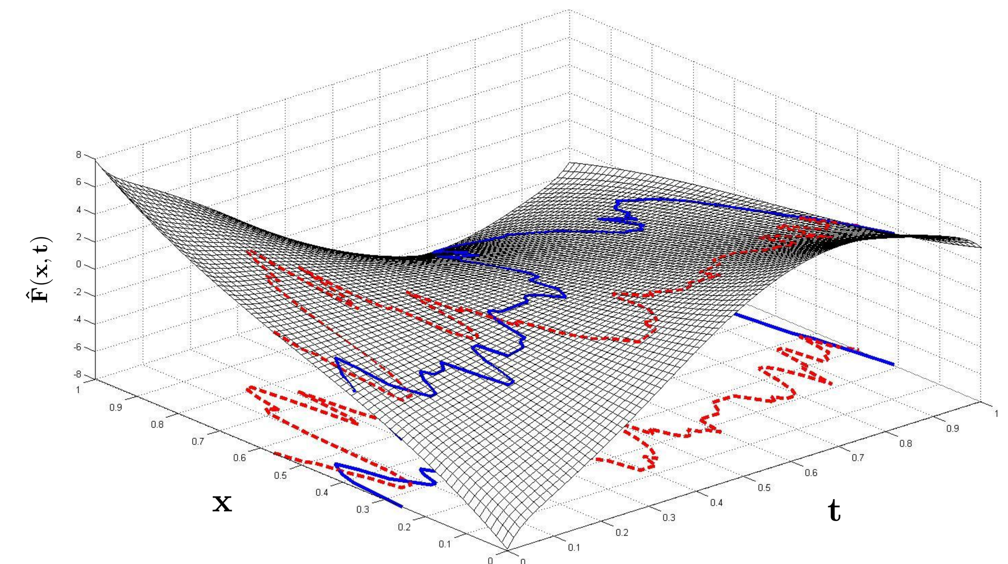
- FGAM is limit of this AM as $J \to \infty$.

Contributions

FGAM: A novel regression model for association studies between a scalar response and a functional predictor that can model complex response-predictor relationships while remaining interpretable.

Our model
- is highly flexible (low bias)
- avoids curse of dimensionality (low variance)
- is easy to interpret (not a "black box")
- Easily fit using fast, polished R package mgcv
- has FLM as a special case ($F(x,t) = \beta(t)x$)

Example Estimated Surface



- Interpretation: Small t values appear most influential; subject with red predictor curve gets higher predicted response
- Nonlinearity in x suggests that an FLM may be inadequate

R Implementation

- FGAM implemented in R package `refund`
- Function `fgam` acts as a wrapper for `gam` in `mgcv`
- E.g. $g(Y_i) = \theta_0 + f(z_1) + \int_{\mathcal{T}_1} F(X_{i1}(t),t)dt + \int_{\mathcal{T}_2} \beta(t)X_{i2}(t)dt$ can be fit by specifying

```
fgam(y~s(z1)+af(X1)+lf(X2),...)
```

- Extends `mgcv`; therefore, it can handle alternative penalties and bases, generalized responses, random effects, automatic model selection, multivariate smooth terms, etc.

Funding

My work was supported in part by NIH grant R01NS060910, NSF grant DMS-0805975, and an NSERC PGS-D.

Estimation

- The surface $F(\cdot,\cdot)$ is parameterized using tensor products of B-splines:

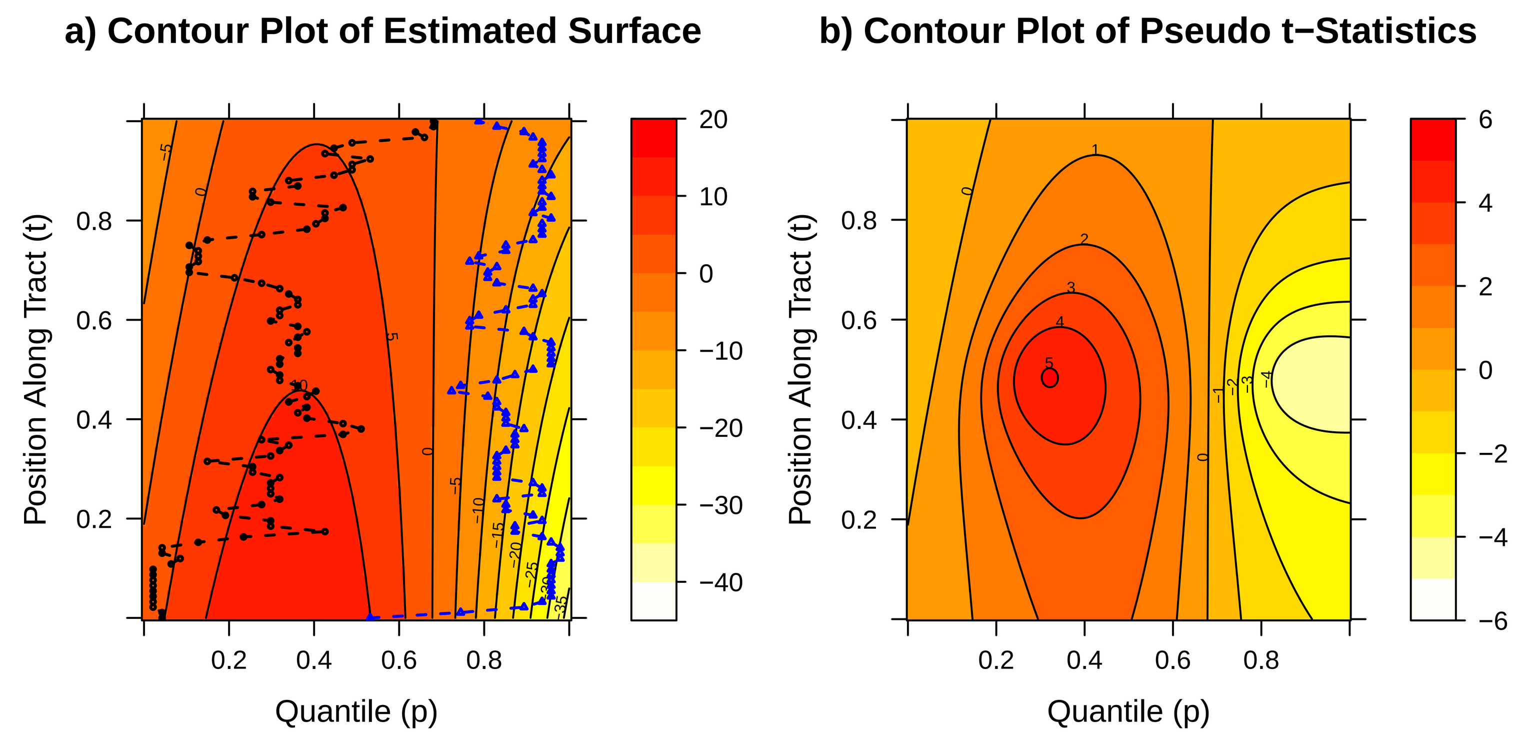
$$F(x,t) = \sum_{j=1}^{K_x}\sum_{k=1}^{K_t} \theta_{j,k} B_j^X(x) B_k^T(t)$$

- We use the P-splines of Eilers & Marx (1996)
- $\{B_j^X(x) : j = 1, \ldots, K_x\}$ and $\{B_k^T(x) : k = 1, \ldots, K_t\}$ are spline bases.
- Y can be from any exponential family distribution
- Smoothing parameters are chosen using generalized cross validation
- Modularity of P-splines: including multiple functional or scalar predictors in the model is simple
- Use generalization of Bayesian confidence bands of Wahba (1983) to construct approximate confidence bands for the true surface to account for bias due to smoothing
- Also obtain confidence bands for the estimated second derivative surface w.r.t. x:

  Significant differences from $\partial^2/\partial x^2 F(x,t) = 0$ at a particular (x,t) suggests FLM does not hold there
- Can transform functional predictor to possibly improve predictions or numerical stability

Results

- Outcome variable is score on a cognitive test called PASAT, which takes integer values between $0-60$
- Predictor is transformed parallel diffusivity: largest eigenvalue summarizing the diffusion at position t
  - Transformation used is the empirical cdf: $\widehat{F}(p,t)$ is effect of $X(t)$ being at its pth quantile
- a) Plot of $\widehat{F}(p,t)$ including two subjects' transformed predictor curves
- Black curve in a) will have higher predicted response than the blue
- b) is contour plot of $\widehat{F}(p,t)$ divided by its estimated standard error
- Middle values of t appear to be very influential on PASAT score
- Good out-of-sample RMSE performance compared with other popular scalar on function regression models



- Simulation studies show
  - FGAM performs nearly as well at out-of sample prediction as FLM when the true model is an FLM
  - FGAM offers substantial gains in predictive performance when FLM is not the true model
  - FGAM is very competitive with other non-FLM scalar on function regression models
  - The proposed Bayesian confidence intervals have good average coverage probabilities

Current Work

- Developing formal test of H_0:FLM vs. H_1:FGAM
  - Use mixed model representation and test particular variance components
- Extending to sparsely observed predictor curves
  - Bayesian MM - estimation via MCMC or VB

For Further Details

- McLean et al. (2012), Functional Generalized Additive Models, *Journal of Computational and Graphical Statistics*, to appear.
- Visit `courses2.cit.cornell.edu/mwmclean`
- R code available in `refund` package on CRAN