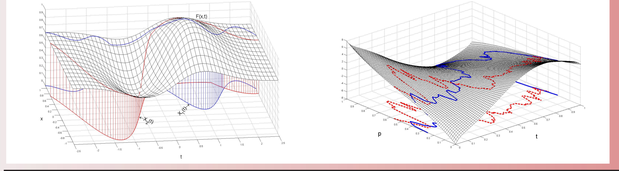


Functional Generalized Additive Models

Mathew McLean†, Giles Hooker†, Ana-Maria Staicu‡, David Ruppert†

†Cornell University ‡NC State University

Cornell University
Operations Research and
Information Engineering

Introduction

We introduce the functional generalized additive model (FGAM), a novel regression model for association studies between a scalar response and a functional predictor. Rather than having an additive model in a finite number of principal components as in Muller and Yao (2008), our model incorporates the functional predictor directly and we regard our model as the natural functional extension of generalized additive models. Our model is more flexible than the functional linear model (FLM), while retaining its ease of interpretation.

Setup and Model

Suppose one observes data $\{(X_i(t), Y_i) : t \in \mathcal{T}\}$ for $i = 1, \ldots, N$, where X_i is a real-valued, random curve on the compact interval $\mathcal{T}$ and Y_i is a scalar. We assume that the predictor, $X(\cdot)$, is observed at a dense grid of points. The FGAM is given by

$$g\{E(Y_i|X_i)\} = \theta_0 + \int_{\mathcal{T}} F\{X_i(t), t\}\, dt, \qquad (1)$$

where $F(\cdot,\cdot)$ is an unknown regression function and $X(t)$ is a functional covariate.

- For identifiability, we use the constraints $\sum_{i=1}^{N} F\{X_i(t), t\} = 0$ for all observation times, t.
- To avoid potentially having a tensor product of B-splines with no observed data on its support, we also consider transforming $X(t)$ by its empirical cdf for each value of t. $F(p,t)$ is now the effect of $X(t)$ being at its pth quantile.
- The model is invariant to transformations of the functional predictor. If an FLM holds for any transformation, the FGAM still holds.

Estimation

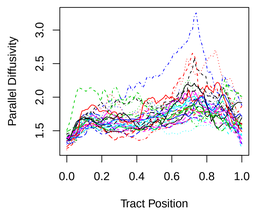
We model $F(\cdot,\cdot)$ using tensor products of B-splines:

$$F(x,t) = \sum_{j=1}^{K_x}\sum_{k=1}^{K_t} \theta_{j,k} B_j^X(x) B_k^T(t) \qquad (2)$$

where $\{B_j^X(x) : j = 1, \ldots, K_x\}$ and $\{B_k^T(x) : k = 1, \ldots, K_t\}$ are spline bases.

- We use the P-splines of Eilers & Marx (1996)
- The model can be fit in R using the mgcv package. P-IRLS is used to maximize the penalized log-likelihood for each choice of the smoothing parameters, which are chosen to minimize GCV.
- Including multiple functional predictors as well as scalar predictors in the model is simple.
- In the identity link case, we use sandwich estimators of the variance of the estimated surface to construct approximate 95% confidence bands and compute pseudo-t statistics.
- Also obtain confidence bands for the estimated second derivative surface w.r.t. x and check for significant differences from 0 to roughly assess if an FLM would be sufficient.

DTI Data



- We apply our model to a study comparing white matter tracts of multiple sclerosis (MS) patients with control subjects using diffusion tensor imaging.
- MS is a central nervous system disorder leading to lesions in white matter which disrupts the ability of cells in the brain to communicate with each other.
- We study the corpus callosum tract here due to its importance in cognition.
- We use three different functions of the eigenvalues from the diffusion tensors as functional predictors: parallel diffusivity, perpendicular diffusivity, and fractional anisotropy.

Funding

This work was supported in part by NIH grant R01NS060910 and NSF grant DMS-0805975.

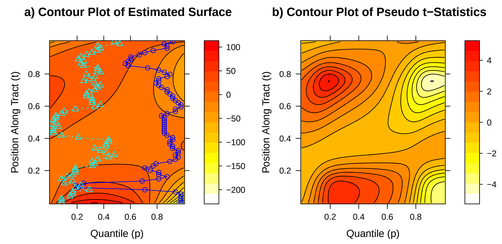
References

[1] M.W. McLean, G. Hooker, A.-M. Staicu, D. Ruppert. Functional Generalized Additive Models *Submitted.*

Results

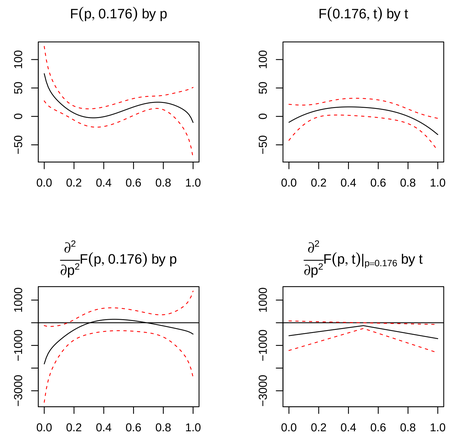
We evaluate our model by predicting the score on a cognitive test called the PASAT. The outcome takes integer values between 0 and 60. We compare the out of sample predictive performance of the FGAM (assuming the response is Gaussian) with the FLM fit using roughness penalties (FLM1) and functional principal component analysis (FLM2) and a functional kernel regression model (FKRM).

Sample FGAM fit for a transformed predictor (see second 3D plot in top left as well):

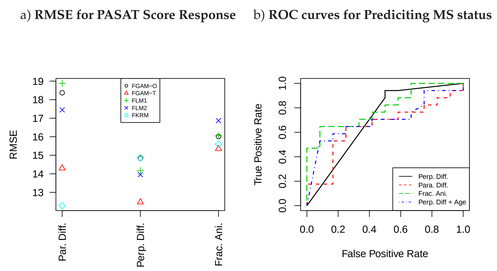


The results from considering a logistic link function assuming a binomial response were similar.

A sample of slices of the estimated surface (first row) and estimated second derivative surface:

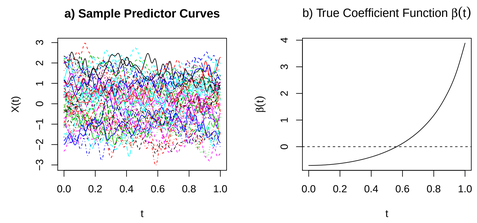


We also fit a logistic link FGAM to predict the MS status of the subjects in the study.

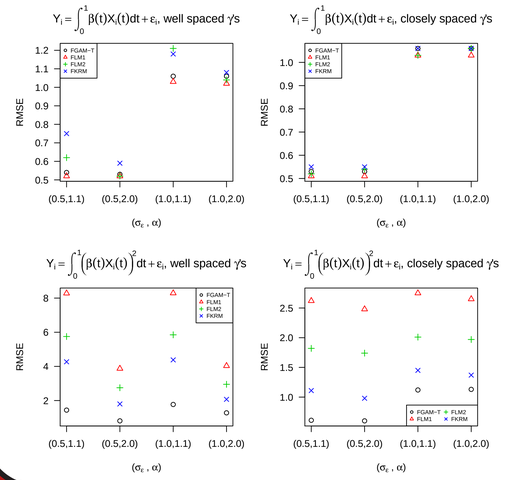


Simulation Study

For each of eight scenarios, we generated 1000 replicate datasets each consisting of 100 curves sampled at 200 equally-spaced points in $[0, 1]$. Two cases each are considered for the spacing of the eigenvalues, γ, of the covariance function of $X(\cdot)$; the rate of decay of the eigenvalues, α; and the error variance, σ_ϵ^2. We perform one group of simulations where the FLM is the true model and another group where it is not.



We see that the FGAM performs nearly identical to the FLM when the FLM is the true model and provides substantial improvements in the case when the FLM is not the true model:



Open Questions

- The dataset contains several more tracts and several more measurements for each tract. Faster fitting methods allowing for all these possible functional predictors to be modelled at once and ways for determining which predictors are truly significant are needed.
- Many of the subjects had multiple scans performed. This longitudinal aspect of the study was not considered, but potentially could be using mixed models.