

# STATS 211: Statistical Methods II

## Lecture 5: Generalized Additive Models

Babak Shahbaba

UCI, Winter 2012

## Background

- Previously, we discussed splines as a class of models to handle nonlinear relationships between the response variables and predictors.
- In this lecture, we discuss “generalized additive models” (GAM) as an alternative approach to build nonlinear regression models.
- These models have the following form:

$$y = \alpha + f_1(x_1) + \dots + f_p(x_p) + \epsilon$$

where  $f_j$  are *smooth* (nonparametric) functions (we can make some of  $f_j$  simple linear functions) and  $\epsilon$  is the error term with mean zero.

## Fitting additive models

- The functions  $f_j$  are typically estimated using a scatterplot smoother such as the cubic smoothing spline discussed in the previous lecture.
- We then minimize the following penalized residual sum of squares:

$$PRSS(\alpha, f_1, \dots, f_p) = \sum_{i=1}^N \left( y_i - \alpha - \sum_{j=1}^p f_j(x_{ij}) \right)^2 + \sum_{j=1}^p \lambda_j \int f_j''(t_j)^2 dt_j$$

- The minimizer of the above function is an additive cubic spline model, where each function  $f_j$  is a cubic spline depending on  $x_j$  only and with knots at each unique values of  $x_{ij}$ .
- To make the model identifiable, it is common to set  $\sum_{i=1}^N f_j(x_{ij}) = 0$  for all  $j$ .

# Backfitting

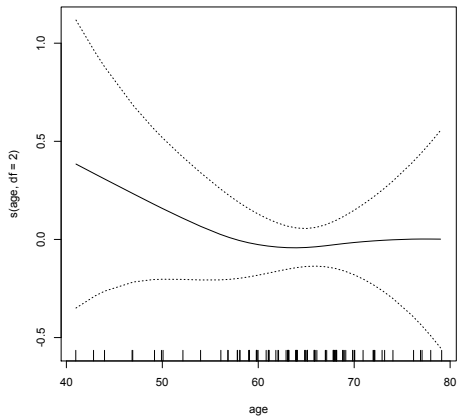
- We can use the following iterative procedure, called *backfitting*, to estimate the functions:
  1. Initialize  $\hat{\alpha} = \text{avg}(y_i)$ ,  $\hat{f}_j \equiv 0$ ,  $\forall i, j$
  2. Iteratively update the functions  $f_j$  as follows until they stabilize (i.e., they don't change substantially from one iteration to another):
    - 2.1 Apply a cubic smoothing spline  $S_j$  to model  $\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})\}_1^N$  as a function of  $x_{ij}$  to obtain a new estimate  $\hat{f}_j$
    - 2.2 Center  $\hat{f}_j$  so its mean becomes zero (i.e., subtract the mean).

## Example

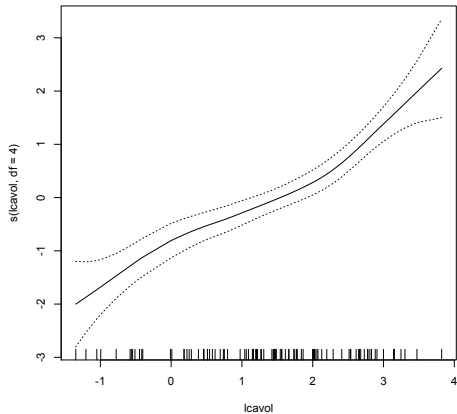
- For the prostate datasets, we use the `gam` library to build a generalized additive model for `lpsa` as a nonlinear function of `age`, `lcavol`, and `bph`.

```
gam1 <- gam(lpsa ~ s(age, df=2) + s(lcavol, df=4)  
+ s(lbph, df=3), data = Prostate)
```

# Example



# Example



# Example

