CKW

Statistics 8330: Data Analysis III
Beyond Linearity: Generalized Additive Models

Suggested Reading: James, Witten, Hastie, and Tibshirani (JWHT), 2013, Chapter 7

Supplemental Reading: Hastie, Tibshirani, and Friedman (HTF), 2009, Chapter 9.1

Christopher K. Wikle

University of Missouri
Department of Statistics

Generalized Additive Models

CKW

We seek a more automatic flexible method for nonlinear regression models.

These are called *"generalized additive models"* or *GAMs*. They take the form:

$$E(Y|X_1, X_2, \ldots, X_p) = \alpha + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p),$$

where $f_j(\cdot), j = 1, \ldots, p$ are unspecified smooth (nonparametric) functions.

Note, this model is *additive* because we have a separate f_j for each X_j, and they are then added together. Thus, we allow fairly complicated functions of the features, but retain the interpretability that additivity affords.

We can write the regression model in more detail as

$$y_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i.$$

Generalized Additive Models (cont.)

If each of the functions $f_j(\cdot)$ were written in terms of a basis expansion, we could fit this model by least squares. We could also take a more general approach and fit each function using a "scatterplot smoother" (e.g., cubic smoothing spline or kernel smoother) using an algorithm that simultaneously estimates all p-functions (see below).

Importantly, we can apply this to some transformation of the mean response (e.g., link function) as we did with generalized linear models. For example, consider the two-group classification problem via logistic regression. In this case, we seek to relate the mean of the binary response $\mu(X) = Pr(Y = 1|X)$ to functions of the predictors via an additive regression model using a *logit* link function:

$$\log\left(\frac{\mu(X)}{1-\mu(X)}\right) = \alpha + f_1(X_1) + \ldots + f_p(X_p),$$

where, again, each f_j is an unspecified smooth function.

Generalized Additive Models (cont.)

CKW

Thus, in general, we relate the conditional mean response to the additive function of the predictors via a link function, g:

$$g[\mu(X)] = \alpha + f_1(X_1) + \ldots + f_p(X_p).$$

For example,

- $g(\mu) = \mu$: the identity link
- $g(\mu) = \text{logit}(\mu)$ or, $g(\mu) = \text{probit}(\mu)$: as in the glm case for binary and binomial data (recall, the probit function is the inverse Gaussian cumulative distribution function: $\text{probit}(\mu) = \Phi^{-1}(\mu)$.)
- $g(\mu) = \log(\mu)$: e.g., for Poisson count data
- others (as in glm)

Generalized Additive Models (cont.)

CKW

We note that not all of the functions f_j need to be nonlinear. After fitting, we may see that there is no need for that level of complexity for nonlinear terms for some inputs. We can mix linear and other parametric forms with the nonlinear terms. One place this occurs is if some of the inputs are qualitative variables (or factors). We also note that we can have nonlinear terms in more than one variable as well. E.g.,

- $g(\mu) = X^T\beta + \alpha_k + f(Z)$: a *semi-parametric* model, where X is a vector of predictors to be modeled linearly, α_k the effect for the kth level of a qualitative input, V, and the effect of the predictor Z is modeled nonparametrically.
- $g(\mu) = f(X) + g_k(Z)$: where k indexes the levels of a qualitative input V and thus creates an interaction term with Z, $g(V, Z) = g_k(Z)$
- $g(\mu) = f(X) + g(Z, W)$: where g is a nonparametric function in two features, Z and W.

Fitting Additive Models

CKW

Recall that we choose the functions f_j to be "smoothers." For example, consider the cubic smoothing spline. Then, we can consider a penalized sum of squares criterion that we wish to minimize:

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

where the $\lambda_j \geq 0$ are tuning parameters.

As before, it can be shown that the minimizer of this penalized sum of squares is an additive cubic spline model with each of the functions f_j a cubic spline in X_j, and with knots at each of the unique values of $x_{ij}, i = 1, \ldots, n$.

Note that this solution is not unique because α is not identifiable without another constraint. We usually say $\sum_{i=1}^{n} f_j(x_{ij}) = 0$, for all j, which implies that $\hat{\alpha} = \text{ave}(y_i)$.

Fitting Additive Models (cont.)

It is also the case that if the matrix $\mathbf{X} = \{x_{ij}\}$ is of full column rank, then the minimizer is unique. If this isn't the case (i.e., $\mathbf{X}'\mathbf{X}$ is singular) then it can be shown, interestingly, that the linear portions of the components f_j cannot be uniquely determined, but the nonlinear parts can.

There is an iterative (*backfitting*) procedure that can be used to fit the model:

Let $\mathcal{S}_j$ be the cubic smoothing spline for the jth function. We apply this to the target residuals $\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})\}_1^n$, as a function of x_{ij} to obtain the new estimate $\hat{f}_j$. This is done for each predictor in turn, using the current estimates of the other functions $\hat{f}_k$ when computing the differences. This proceeds iteratively until the estimates $\hat{f}_j$ stabilize (see Algorithm 9.1 in HTF (2009) for details).

Fitting Additive Models: Backfitting Algorithm

Algorithm 9.1 *The Backfitting Algorithm for Additive Models.*

1. Initialize: $\hat{\alpha} = \frac{1}{N}\sum_1^N y_i$, $\hat{f}_j \equiv 0, \forall i, j$.

2. Cycle: $j = 1, 2, \ldots, p, \ldots, 1, 2, \ldots, p, \ldots,$

$$\hat{f}_j \leftarrow \mathcal{S}_j\left[\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})\}_1^N\right],$$

$$\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{N}\sum_{i=1}^N \hat{f}_j(x_{ij}).$$

until the functions $\hat{f}_j$ change less than a prespecified threshold.

Fitting Additive Models (cont.)

CKW

Note, we can use any smoother to take the place of $\mathcal{S}_j$ in the fitting algorithm. For example, one can use

- other univariate regression smoothers such as local polynomial regression and kernel methods;
- linear regression operators yielding polynomial fits, piecewise constant fits, parametric spline fits, Fourier series fits, etc;
- more complicated operators such as surface smoothers for second or higher-order interactions or periodic smoothers for seasonal effects.

If we are considering the smoother only at the training points, it can be represented by the $n \times n$ smoother matrix, $\mathbf{S}_j$, that we talked about previously. In that case, we can use this to estimate the degrees of freedom associated with the jth term from: $\text{trace}[\mathbf{S}_j] - 1$.

Fitting Generalized Additive Models

CKW

In the generalized additive model case, the optimization criterion that is considered is the penalized log-likelihood. This is maximized by combining the backfitting procedure with a likelihood maximizer.

For example, recall from DA II that the usual Newton-Raphson routine for maximizing log-likelihoods for generalized linear models can be recast as an iteratively reweighed least squares (IRLS) algorithm. This involves iteratively fitting a weighted linear regression of a working response variable on the covariates; each regression yielding new values of the parameter estimates, which are then used to get new working responses and weights, etc.

In the generalized additive model, the weighted linear regression component is replaced by a weighted backfitting algorithm. [See the Algorithm 9.2 in HTF (2009) and further details in Hastie and Tibshirani, 1990.]

Fitting Generalized Additive Models: Logistic Algorithm

Algorithm 9.2 *Local Scoring Algorithm for the Additive Logistic Regression Model.*

1. Compute starting values: $\hat{\alpha} = \log[\bar{y}/(1-\bar{y})]$, where $\bar{y} = \text{ave}(y_i)$, the sample proportion of ones, and set $\hat{f}_j \equiv 0 \; \forall j$.

2. Define $\hat{\eta}_i = \hat{\alpha} + \sum_j \hat{f}_j(x_{ij})$ and $\hat{p}_i = 1/[1+\exp(-\hat{\eta}_i)]$.

   Iterate:

   (a) Construct the working target variable

   $$z_i = \hat{\eta}_i + \frac{(y_i - \hat{p}_i)}{\hat{p}_i(1-\hat{p}_i)}.$$

   (b) Construct weights $w_i = \hat{p}_i(1-\hat{p}_i)$

   (c) Fit an additive model to the targets z_i with weights w_i, using a weighted backfitting algorithm. This gives new estimates $\hat{\alpha}, \hat{f}_j, \; \forall j$

3. Continue step 2. until the change in the functions falls below a prespecified threshold.

Example: HTF (2009, Section 9.1)

Consider the example in HTF (2009, Section 9.1) that applies a generalized additive model to predict email "spam" (junk email). The data consists of 4601 email messages that have been classified as email (0) or spam (1). There are 57 predictors, 48 of which are quantitative corresponding to the percentage of words in the email that match a given word; 6 are quantitative corresponding to the percentage of characters that match a given character, and 3 variables corresponding to the length of sequences of capital letters. A test set was constructed of size 1536 with 3065 used for the training set.

The GAM used cubic smoothing splines with 4 degrees of freedom (i.e., λ_j was chosen so that $\text{trace}[\mathbf{S}_j(\lambda_j)] - 1 = 4$). Each variable was log-transformed before the analysis. As shown below, the overall test error rate was 5.3%.

Example: HTF (2009, Section 9.1)

CKW

TABLE 9.1. *Test data confusion matrix for the additive logistic regression model fit to the spam training data. The overall test error rate is* 5.5%.

	Predicted Class	
True Class	`email` (0)	`spam` (1)
`email` (0)	58.3%	2.5%
`spam` (1)	3.0%	36.3%

TABLE 9.2. *Significant predictors from the additive model fit to the spam training data. The coefficients represent the linear part of $\hat{f}_j$, along with their standard errors and Z-score. The nonlinear P-value is for a test of nonlinearity of $\hat{f}_j$.*

Name	Num.	df	Coefficient	Std. Error	Z Score	Nonlinear P-value
Positive effects						
`our`	5	3.9	0.566	0.114	4.970	0.052
`over`	6	3.9	0.244	0.195	1.249	0.004
`remove`	7	4.0	0.949	0.183	5.201	0.093
`internet`	8	4.0	0.524	0.176	2.974	0.028
`free`	16	3.9	0.507	0.127	4.010	0.065
`business`	17	3.8	0.779	0.186	4.179	0.194
`hpl`	26	3.8	0.045	0.250	0.181	0.002
`ch!`	52	4.0	0.674	0.128	5.283	0.164
`ch$`	53	3.9	1.419	0.280	5.062	0.354
`CAPMAX`	56	3.8	0.247	0.228	1.080	0.000
`CAPTOT`	57	4.0	0.755	0.165	4.566	0.063
Negative effects						
`hp`	25	3.9	−1.404	0.224	−6.262	0.140
`george`	27	3.7	−5.003	0.744	−6.722	0.045
`1999`	37	3.8	−0.672	0.191	−3.512	0.011
`re`	45	3.9	−0.620	0.133	−4.649	0.597
`edu`	46	4.0	−1.183	0.209	−5.647	0.000

Example: HTF (2009, Section 9.1)

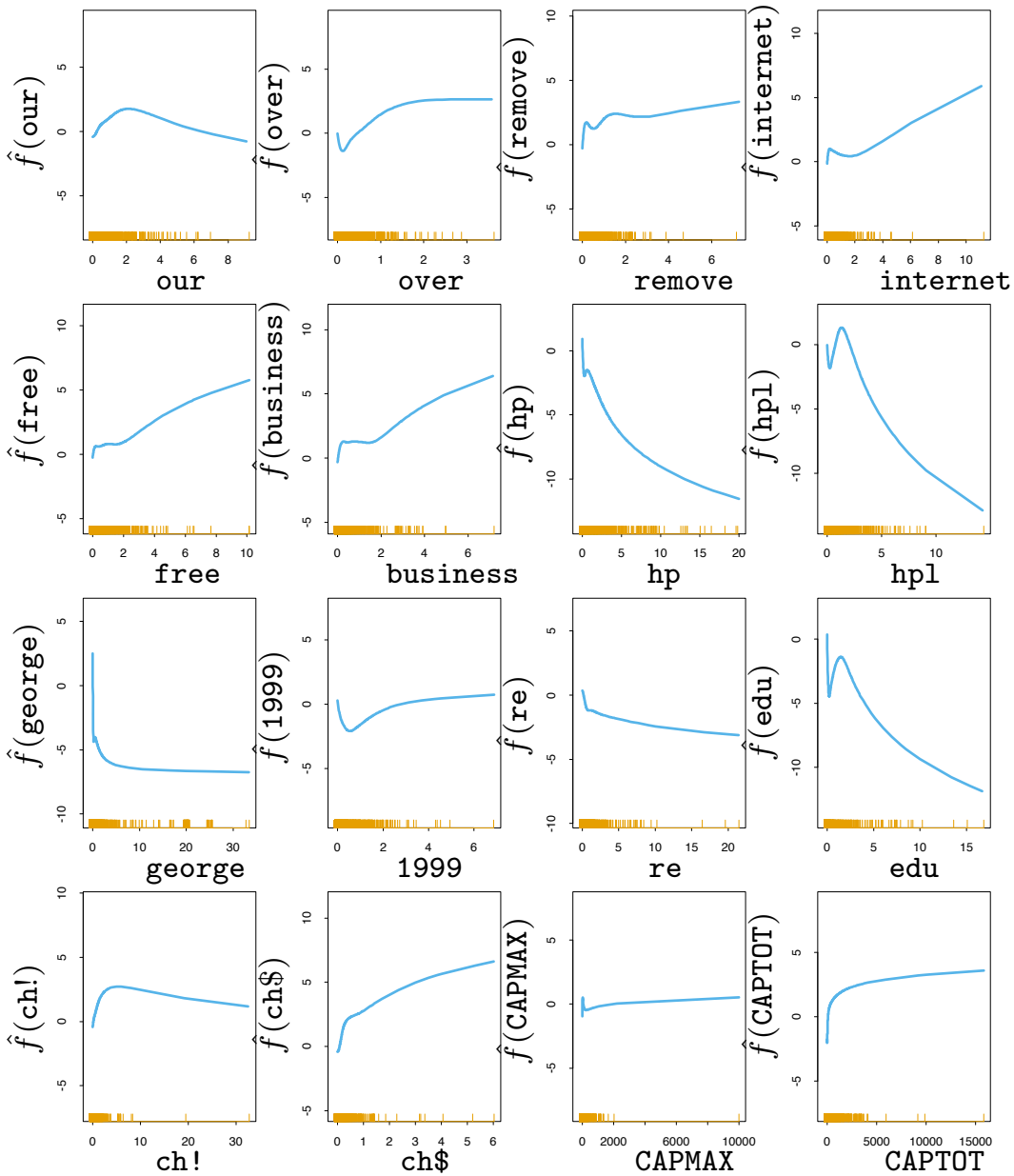


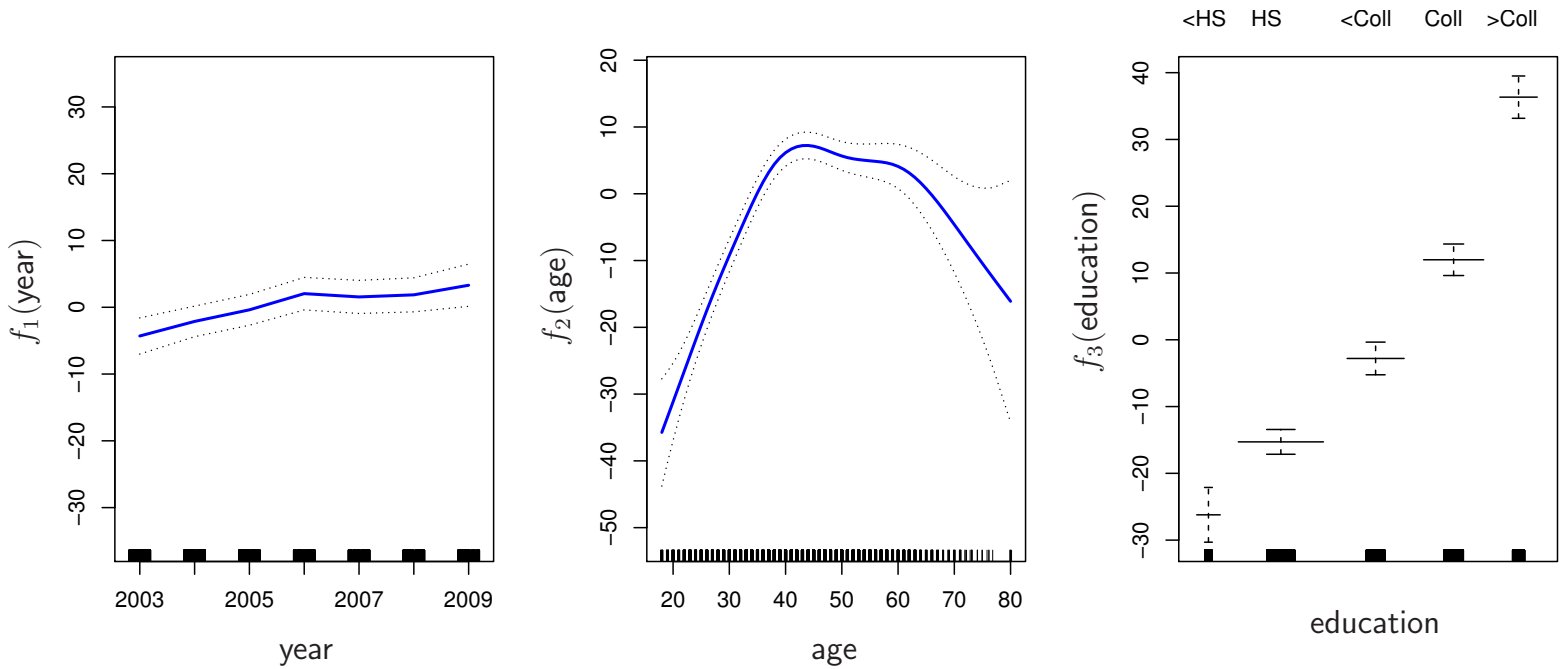
FIGURE 9.1. *Spam analysis: estimated functions for significant predictors. The* rug plot *along the bottom of each frame indicates the observed values of the corresponding predictor. For many of the predictors the nonlinearity picks up the discontinuity at zero.*

Another Example

CKW

Consider the example from JWHT (2013, Section 7.7) that fits:

$$\text{wage} = \alpha + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education}) + \epsilon$$



Pros and Cons of GAMs

CKW

JWHT summarize the pros and cons of GAMs as follows:

- GAMs allow us to fit a nonlinear f_j to each X_j, so we automatically model nonlinear relationships that could be missed by standard linear regression
- The nonlinear fits can potentially make more accurate predictions
- Because the model is additive, we can still examine the effect of each X_j on Y individually while holding all of the other variables fixed
- The smoothness of the function f_j for the variable X_j can be summarized by degrees of freedom
- The main limitation of GAMs is that the model is restricted to be additive. We can still include interactions, but have to make them manually like in LS regression.
- GAMs can be difficult to fit in high dimensions because all variables have to be fit (unless one uses a specialized algorithm)

Extensions

Note that there are many more details and extensions associated with GAMs. For example, one can include random effects in this framework (GAMMs).

I recommend the book *Generalized Additive Models: An Introduction with R (Second Edition)* by Simon N. Wood (2017; Chapman & Hall/CRC).

We will also see that many of the tree-based models that we are going to cover next can be considered extensions of GAMs.