Generalized Additive Models

Yufei Wang & Zhuoqun Cheng

Contents

Generalized Additive Models

The traditional linear model has the form

$$E(Y|X_1, X_2, \cdots, X_p) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

In the regression setting, a generalized additive model has the form

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

Generalized Additive Models

We relate the mean of the binary response $\mu(X) = \Pr(Y = 1|X)$ to the predictors via a linear regression model and the logit link function:

$$\log\left(\frac{\mu(x)}{1-\mu(x)}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

The additive logistic regression model replaces each linear term by a more general functional form

$$\log\left(\frac{\mu(x)}{1-\mu(x)}\right) = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

Generalized Additive Models

In general, the conditional mean μ(X) of a response Y is related to an additive function of the predictors via a link function g:

$$g[\mu(x)]=\alpha+f_1(X_1)+f_2(X_2)+\cdots+f_p(X_p)$$

Generalized Additive Models

Examples of classical link functions are the following:

• $g(\mu) = \mu$ is the identity link, used for linear and additive models for Gaussian response data.

• $g(\mu) = \text{logit}(\mu)$ as above, or $g(\mu) = \text{probit}(\mu)$, the probit link function , for modeling binomial probabilities. The probit function is the inverse Gaussian cumulative distribution function: $\text{probit}(\mu) = \Phi^{-1}(\mu)$

• $g(\mu) = \log(\mu)$ for log-linear or log-additive models for Poisson count data.

Generalized Additive Models

How does the f_j look like ?

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X),$$

$h_m(X) : IR^p \rightarrow$ IR the mth transformation of X, m =1, . . . ,M.

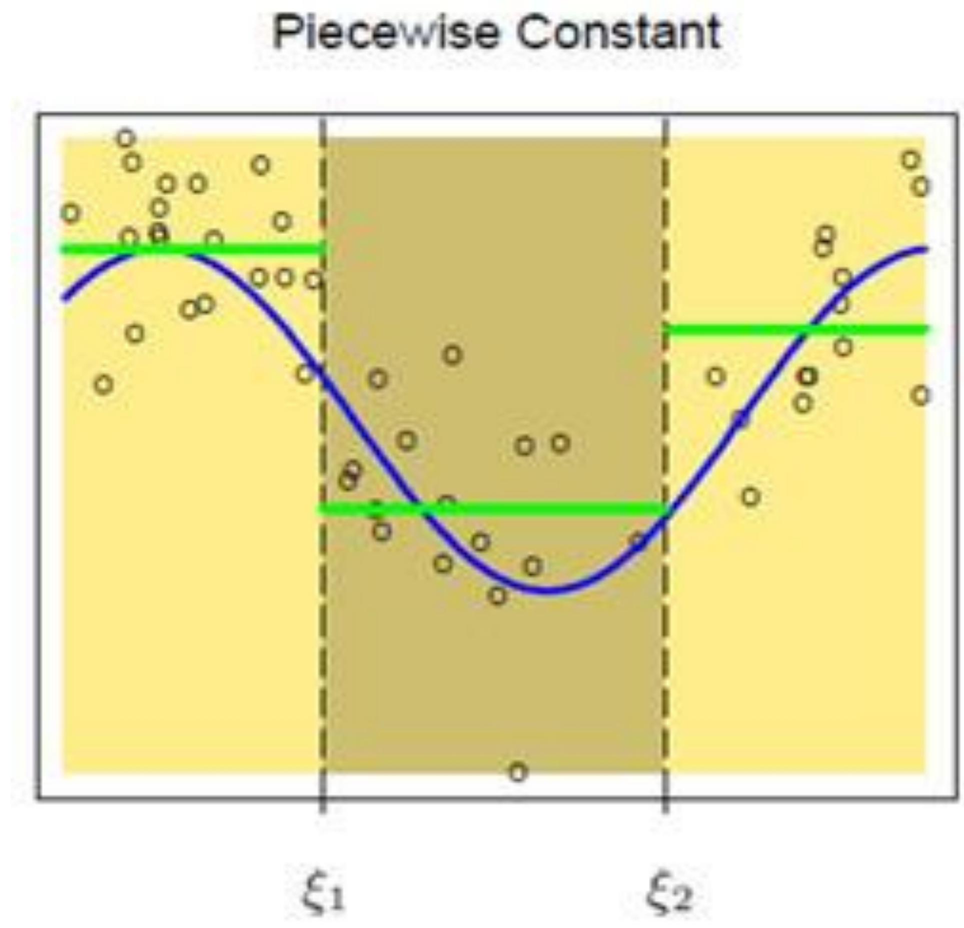
Generalized Additive Models

Some simple and widely used examples of the h_m are the following:

- $h_m(X) = X_m$, m = 1, . . . , p recovers the original linear model.
- $h_m(X) = X_j{}^2$ or $h_m(X) = X_j \cdot X_k$ allows us to augment the inputs with polynomial terms to achieve higher-order Taylor .
- $h_m(X) = log(X_j)$, $\sqrt{X_j}$, . . . , permits other nonlinear transformations of single inputs.
- $h_m(X) = I(L_m < X_k < U_m)$, an indicator for a region of X_m.

Piecewise Polynomials

$$h_1(X) = I(X < \xi_1), \quad h_2(X) = I(\xi_1 \le X < \xi_2), \quad h_3(X) = I(\xi_2 \le X).$$



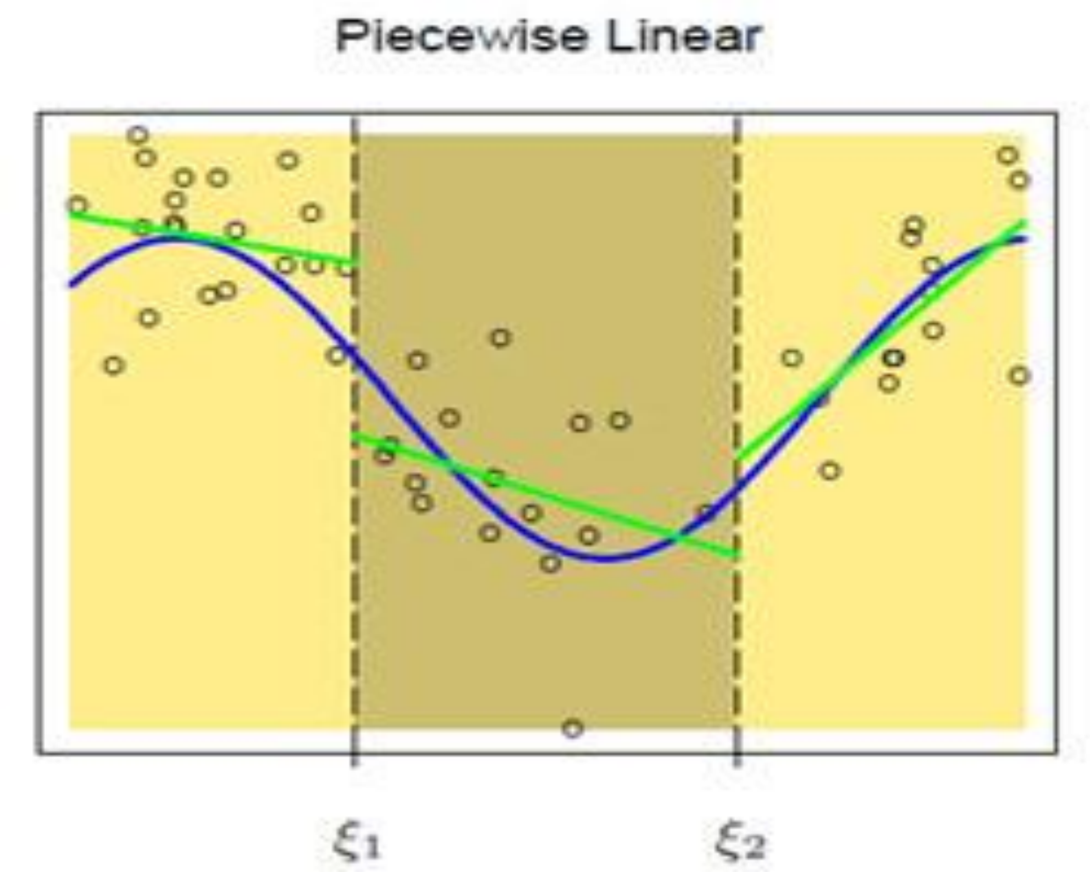
$$f(X) = \sum_{m=1}^{3} \beta_m \cdot h_m(X)$$

$$\widehat{\beta_m} = \overline{Y_m}$$

Piecewise Polynomials

We add Three additional basis functions :

$\boldsymbol{h_{m+3}(X)} = \overline{\boldsymbol{h_m}}\boldsymbol{(X)} \cdot \boldsymbol{X}$, m = 1, . . . , 3.

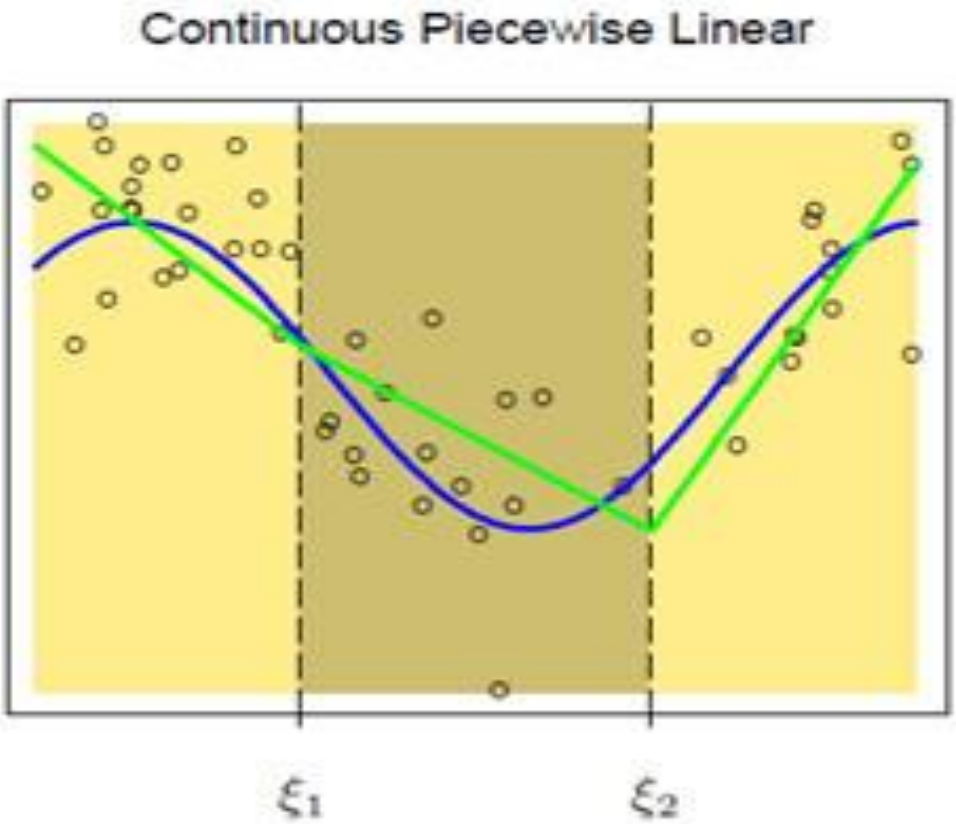


Piecewise Polynomials

Then we add two constraint conditions:

$$f(\xi_1{}^{-})=f(\xi_1{}^{+}) \iff \beta_1+\xi_1\beta_4=\beta_2+\xi_1\beta_5$$

$$f(\xi_2{}^{-})=f(\xi_2{}^{+}) \iff \beta_2+\xi_2\beta_5=\beta_3+\xi_2\beta_6$$



Piecewise Polynomials

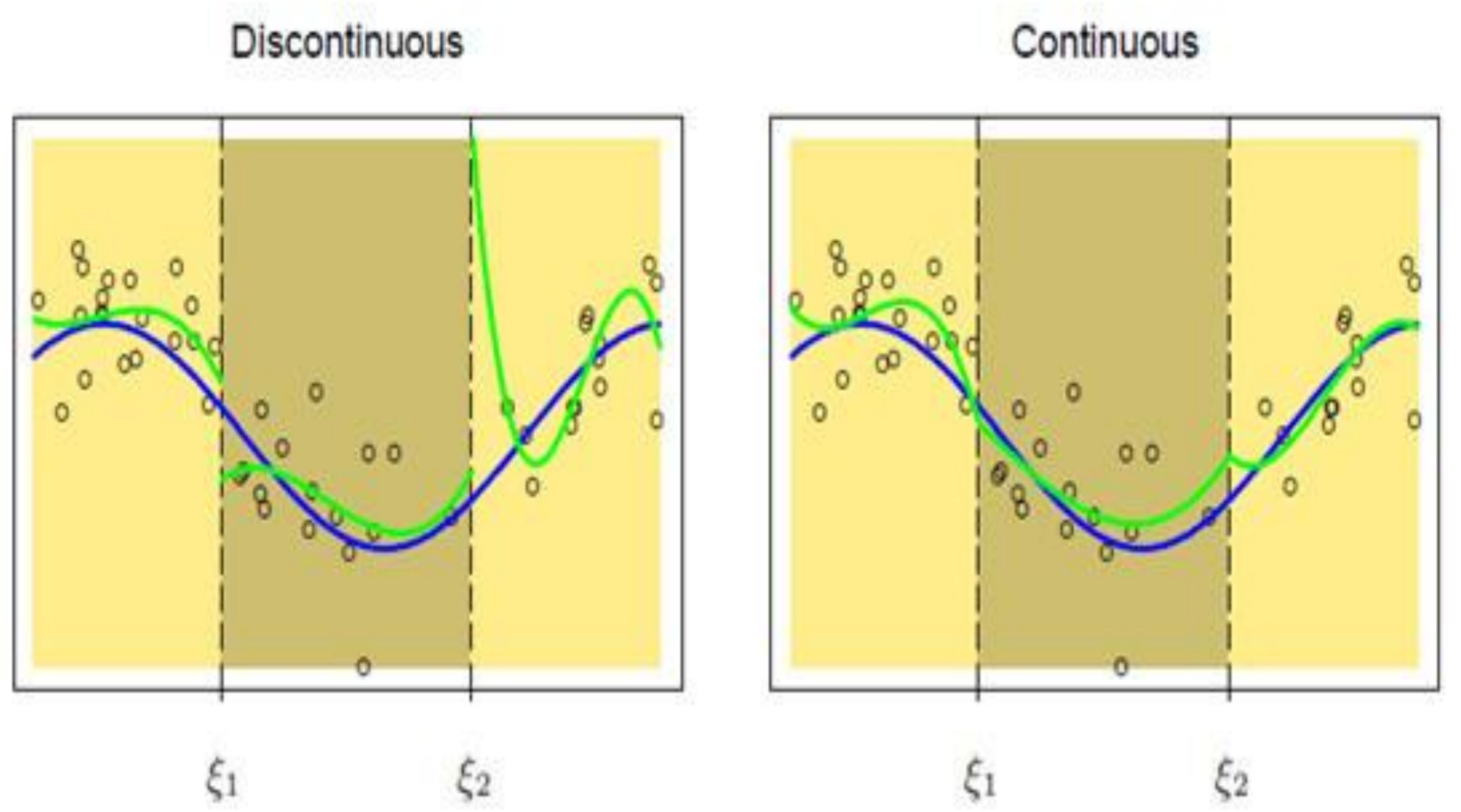
A more direct way to proceed in this case is to use a basis that incorporates the constraints:

$$h_1(X) = 1, h_2(X) = X, h_3(X) = (X - \xi_1)_+,\ h_4(X) = (X - \xi_2)_+,$$

$$(X - \xi_1)_+ = \begin{cases} X - \xi_1 & , X > \xi_1 \\ 0 & , X \leq \xi_1 \end{cases}$$

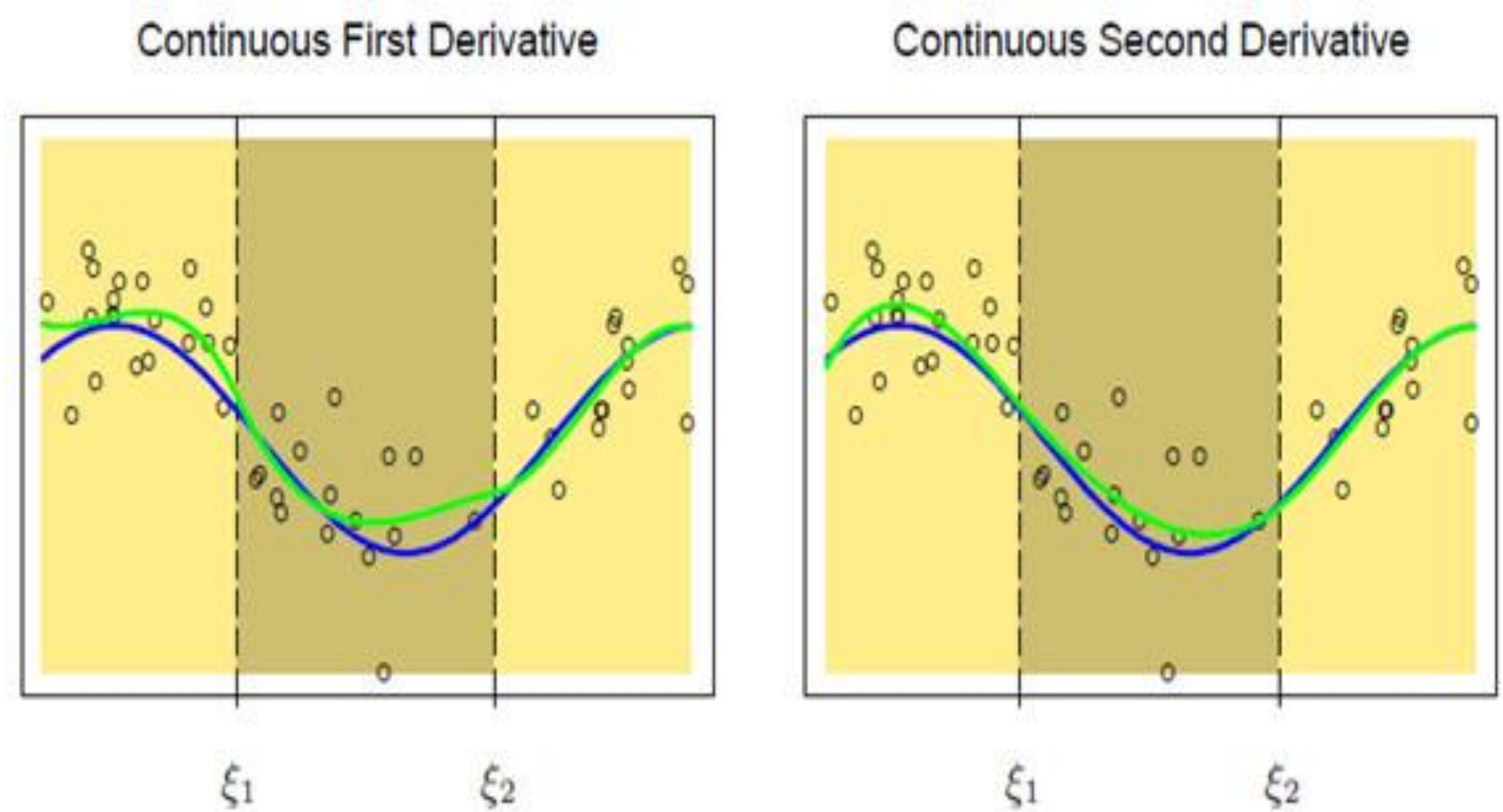
Piecewise Polynomials

We often prefer smoother functions, and these can be achieved by increasing the order of the local polynomial.



Piecewise Polynomials

The function in this figure is continuous, and has continuous first and second derivatives at the knots.



cubic spline

The function has two continuous derivatives at the knots.
It is known as a cubic spline.
It is not hard to show that the basis represents a cubic spline with knots at 1 and 2:

$$h_1(X) = 1, \quad h_3(X) = X^2, \quad h_5(X) = (X - \xi_1)_+^3,$$

$$h_2(X) = X, \quad h_4(X) = X^3, \quad h_6(X) = (X - \xi_2)_+^3.$$

Piecewise Polynomials

More generally, an order M spline with knots j , j = 1, . . . ,K is a piecewise-polynomial of order M, and has continuous derivatives up to order M − 2.

Likewise the general form for the truncated-power basis set would be:

$$
\begin{aligned}
h_j(X) &= X^{j-1},\ j=1,\ldots,M,\\
h_{M+\ell}(X) &= (X-\xi_\ell)_+^{M-1},\ \ell=1,\ldots,K
\end{aligned}
$$

Fitting Additive Models

The additive model has the form

$$Y = \alpha + \sum_{j=1}^{p} f_j(x_j) + \varepsilon$$

Consider the following problem : among all functions $f_1, f_1, f_2, \cdot\cdot\cdot, f_p$ with two continuous derivatives, find one that minimizes the penalized residual sum of squares

$$PRSS(\alpha, f_1, f_2, \cdot\cdot\cdot, f_p) = \sum_{i=1}^{N} (y_i - \alpha - \sum_{j=1}^{p} f_j(x_{ij}))^2$$
$$+ \sum_{j=1}^{p} \lambda_j \int f_i''(t_j)^2 d_{t_j}$$

Fitting Additive Models

The Backfitting Algorithm for Additive Models

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

2. Cycle: j=1,2, · · · ,p.

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

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

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

Additive Logistic Regression

In this model,

$$y = \begin{cases} 0, & no\ event \\ 1, & event\ happen \end{cases}$$

We wish to model Pr(Y =1|X), the probability of an event given values of the prognostic factors

$$X^T = (x_1, \ldots, x_p).$$

Additive Logistic Regression

The generalized additive logistic model has the form:

$$\log \frac{P_r(Y=1|X)}{P_r(Y=0|X)} = \alpha + f_1(x_1) + \ldots + f_p(x_p)$$

The functions $f_1, f_2, \cdot \cdot \cdot , f_p$ are estimated by a back fitting algorithm with in a Newton–Raphson procedure, shown in Algorithm.

Additive Logistic Regression

Algorithm Local Scoring Algorithm for the Additive Logistic Regression Model.

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

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

3. Continue step2 until the change in the functions falls below a prespecified threshold.

Additive Logistic Regression

Example : Predicting Email Spam

We apply a generalized additive model to the spam data. The data consists of information from 4601 email messages, in a study to screen email for " spam " (i.e. junk email).

Fitting Additive Models

The response variable is binary, with values email or spam, and there are 57 predictors as described below:

- 48 quantitative predictors—the percentage of words in the email that match a given word. Examples include business, address, internet, free and george. The idea was that these could be customized for Individual users.
- 6 quantitative predictors—the percentage of characters in the email That match a given character. The characters are ch;, ch(, ch[, ch!, ch$, and ch#.
- The average length of uninterrupted sequences of capital letters:CAPAVE.
- The length of the longest uninterrupted sequences of capital letters: CAPMAX.
- The sum of the length of uninterrupted sequences of capital letters: CAPTOT.

Additive Logistic Regression

In this model:

$$y = \begin{cases} 0, & email \\ 1, & spam \end{cases}$$

A test set of size 1536 was randomly chosen, leaving 3065 observations in the training set. A generalized additive model was fit, using a cubic smoothing spline with a nominal four degrees of freedom for each predictor.

The test error rates are shown in Table1; the over all error rate is 5.3%. By comparison, a linear logistic regression has a test error rate of 7.6%.

Additive Logistic Regression

Table 1

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

Table2 shows the predictors that are highly significant in the additive model.

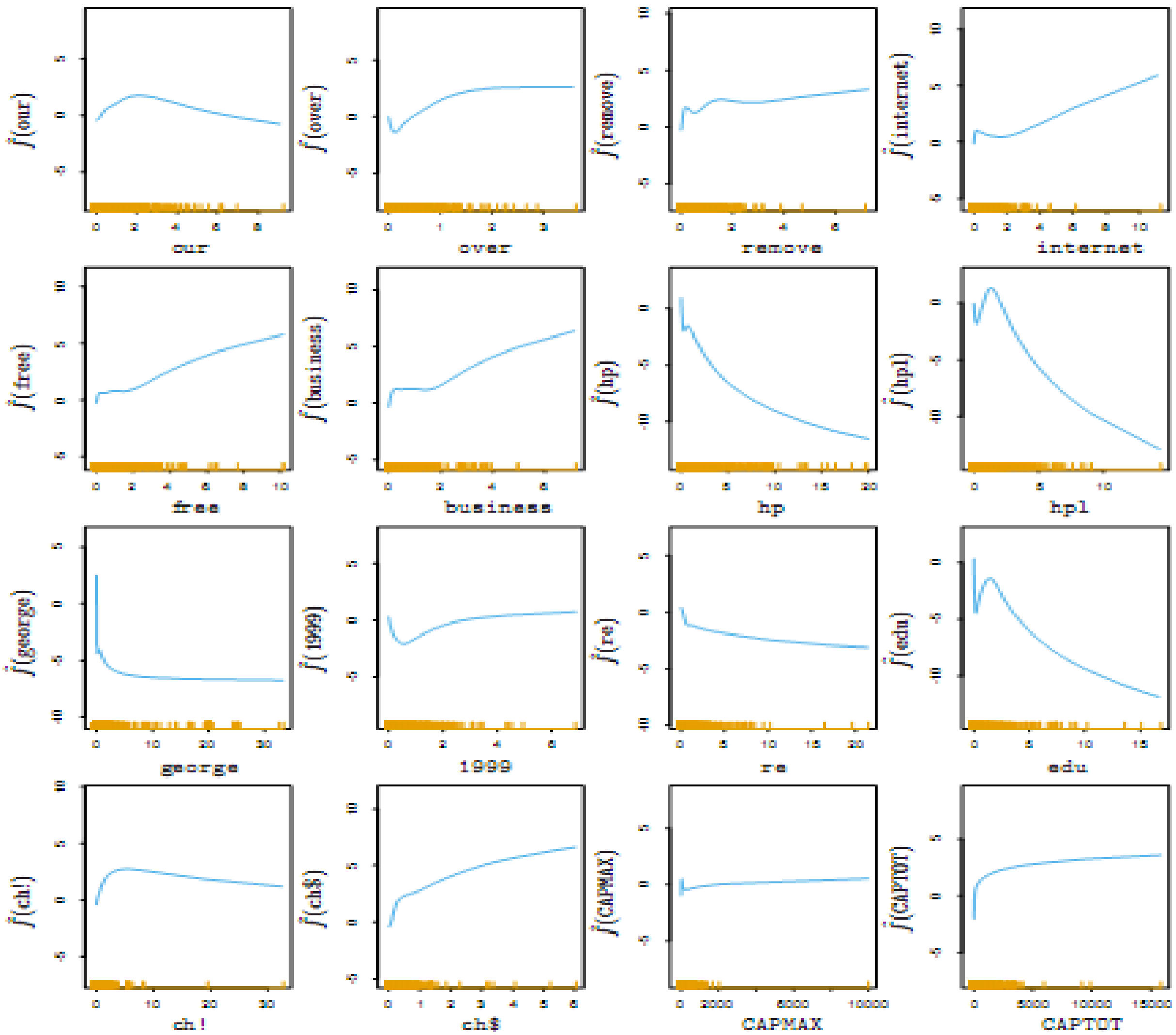
Additive Logistic Regression

Table2

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

Additive Logistic Regression

The figure shows the estimated functions for the significant predictors appearing in Table2.



Thank You !