Lecture 9: Generalised Additive Models

MATH5824 Generalised Linear and Additive Models

Richard P Mann

MATH5824 Generalised Linear and Additive Models

Reading

Course notes: Chapter 6

`www.richardpmann.com/MATH5824`

From Smoothing Splines to GAMs

So far: Single explanatory variable, normal errors.

$$y_i = f(t_i) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Now: Multiple explanatory variables, possibly non-normal responses.

Generalised Additive Models (GAMs) combine:

1. GLM framework (exponential family, link functions)
2. Non-parametric smooth functions of predictors

The GAM Framework

A GAM has three components, extending the GLM:

1. **Random component:** Y belongs to the exponential family with parameters θ and ϕ.
2. **Systematic component** (non-linear predictor):

$$\eta = \sum_{j=1}^{p} f_j(x_j)$$

where each f_j is a *smooth function* (not necessarily linear).

3. **Link function:**

$$\eta = g(\mu), \qquad \mu = g^{-1}(\eta)$$

GAM vs. GLM

	GLM	GAM
Predictor	$\eta = \sum \beta_j x_j$	$\eta = \sum f_j(x_j)$
Each term	Linear: $\beta_j x_j$	Smooth: $f_j(x_j)$
Parameters	$\boldsymbol{\beta}$ (finite)	Functions f_j
Estimation	Maximum likelihood	Penalised likelihood
Response	Exponential family	Exponential family

Note: A GLM is a special case of a GAM where each $f_j(x_j) = \beta_j x_j$.

GAMs can also include parametric (linear) terms alongside smooth terms.

Penalised Deviance

For non-Gaussian data: Replace penalised least squares with **penalised deviance**.

With m smooth terms $f_1, \ldots, f_m$ and parametric coefficients $\boldsymbol{\beta}$:

$$R_\nu = D(\mathbf{y}, f_1, \ldots, f_m, \boldsymbol{\beta}) + \sum_{h=1}^{m} \lambda_h \, J_\nu(f_h)$$

where:

- $D(\mathbf{y}, \ldots)$ is the deviance (from GLM theory)
- λ_h is the smoothing parameter for the hth smooth term
- $J_\nu(f_h)$ is the roughness penalty for f_h

Each smooth term has its *own* smoothing parameter λ_h, chosen by GCV.

GAMs in R: The mgcv Package

Key function: `gam()` from the `mgcv` package.

```r
library(mgcv)

# Fit a GAM with smooth terms
fit <- gam(y ~ s(x1, k = 10) + s(x2, k = 10),
           family = "gaussian")

# With specified smoothing parameter
fit <- gam(y ~ s(x1, k = 10, sp = 3.5))

# Without sp: lambda chosen by GCV automatically
fit <- gam(y ~ s(x1, k = 10))
```

`s()` specifies a smooth term (cubic smoothing spline). The argument `k` sets the maximum dimensionality of the spline basis.

Key gam() Output

```r
# Model summary
summary.gam(fit)

# Analysis of deviance
anova.gam(fit)

# Useful components
fit$fitted.values   # Fitted values
fit$sp              # Smoothing parameter(s)
fit$gcv.ubre        # GCV criterion value
sum(fit$hat)        # Total effective df
```

Important: The `edf` (effective degrees of freedom) for each smooth term indicates whether the relationship is approximately linear (edf ≈ 1) or genuinely non-linear (edf $\gg 1$).

Example: Coronary Heart Disease

Data: South African CHD case-control study ($n = 462$).

Variables:

- Response: CHD status (binary: 0/1)
- Explanatory: tobacco consumption, age, family history

GLM (linear effects):

$$\text{logit}(\mathbb{P}(\text{CHD})) = \beta_0 + \beta_1 \cdot \text{tobacco} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{famhist}$$

GAM (smooth effects):

$$\text{logit}(\mathbb{P}(\text{CHD})) = \beta_0 + f_1(\text{tobacco}) + f_2(\text{age}) + \beta_3 \cdot \text{famhist}$$

CHD Example: GLM Fit

```
hr <- read.table("SAheart.txt", sep = ",",
                 header = TRUE, row.names = 1)
attach(hr)

glm1 <- glm(chd ~ tobacco + age + famhist,
            family = "binomial")
```

	Estimate	Std. Error	z	p-value
(Intercept)	-3.621	0.445	-8.14	< 0.001
tobacco	0.083	0.026	3.23	0.001
age	0.049	0.009	5.16	< 0.001
famhist (Present)	0.975	0.220	4.43	< 0.001

All variables significant. **However:** assumes logit is *linear* in tobacco and age.

CHD Example: GAM Fit

```r
library(mgcv)
gam1 <- gam(chd ~ s(tobacco, k = 20) + s(age, k = 20)
            + famhist, family = "binomial")
summary.gam(gam1)
```

Smooth term results:

Term	edf	Ref.df	χ^2	p-value
`s(tobacco)`	6.08	7.57	17.89	0.018
`s(age)`	1.00	1.00	24.11	< 0.001

Key findings:

- **Tobacco:** edf ≈ 6 — genuinely non-linear relationship
- **Age:** edf ≈ 1 — approximately linear (GAM agrees with GLM)

Interpreting Effective Degrees of Freedom

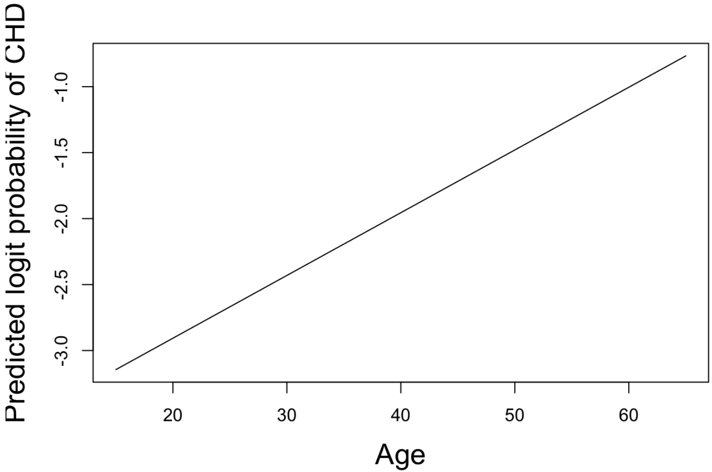
What does edf **tell us?**

edf	Interpretation
≈ 1	Approximately linear relationship
2–3	Mildly non-linear (e.g., quadratic)
> 5	Substantially non-linear

In the CHD example:

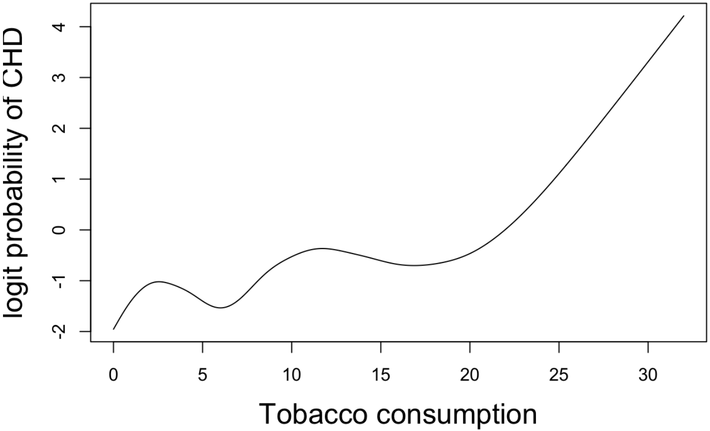
- Age effect is essentially linear — a GLM is adequate for this term
- Tobacco effect is non-linear — the smooth function captures structure that a linear term misses

CHD Example: Predicted Effect of Age



Predicted logit probability of CHD as a function of age (holding tobacco = 0, family

CHD Example: Predicted Effect of Tobacco



Predicted logit probability of CHD as a function of tobacco consumption (holding age

Plotting Smooth Effects in R

```r
# Age effect
newdat1 <- data.frame(age = seq(15, 65, by = 0.1),
                      tobacco = 0,
                      famhist = "Absent")
pred1 <- predict.gam(gam1, newdata = newdat1)
plot(newdat1$age, pred1, type = "l",
     xlab = "Age", ylab = "logit P(CHD)")

# Tobacco effect
newdat2 <- data.frame(tobacco = seq(0, 32, by = 0.1),
                      age = 40,
                      famhist = "Absent")
pred2 <- predict.gam(gam1, newdata = newdat2)
plot(newdat2$tobacco, pred2, type = "l",
     xlab = "Tobacco", ylab = "logit P(CHD)")
```

Summary

Key points:

- GAMs extend GLMs by replacing linear terms $\beta_j x_j$ with smooth functions $f_j(x_j)$
- Estimation uses penalised deviance with GCV-selected smoothing parameters
- The `mgcv` package in R provides `gam()` for fitting
- Effective degrees of freedom (edf) indicate the degree of non-linearity
- edf ≈ 1: linear (GAM reduces to GLM for that term)
- GAMs can mix smooth and parametric terms (e.g., categorical variables)
- Partial effect plots show how each predictor relates to the response

This concludes the module. GAMs provide a flexible framework for modelling non-linear relationships within the GLM paradigm.