Lecture 14. Nonparametric GLMs (cont.)

Nan Ye

School of Mathematics and Physics
University of Queensland

Recall: Nonparametric Models

Parametric models

- Fixed structure and number of parameters.
- Represent a fixed class of functions.

Nonparametric models

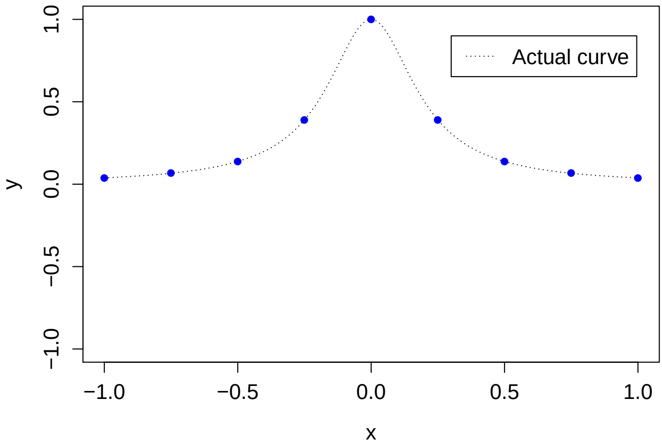
- Flexible structure where the number of parameters usually grow as more data becomes available.
- The class of functions represented depends on the data.
- Not models without parameters, but nonparametric in the sense that they do not have fixed structures and numbers of parameters as in parametric models.

This Lecture

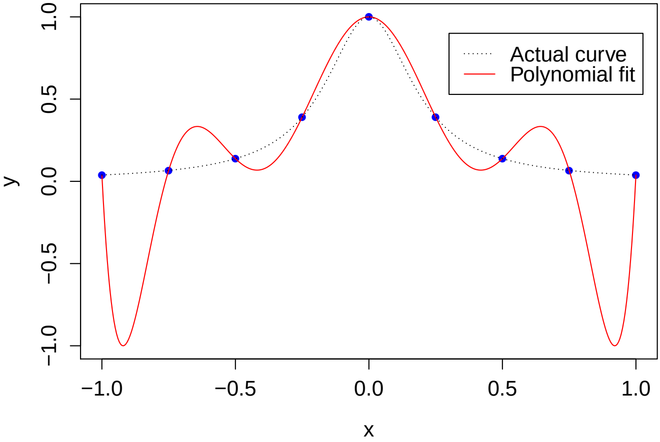
- Smoothing splines
- Generalized additive models

Smoothing Splines

If we fit a degree 8 polynomial on these 9 points, will the polynomial be a good fit?



No...



Runge phenomenon: polynomial fits can be very unstable.

Trade-off between smoothness and quality of fit

- We want to find a curve $f(x)$ that fits data well, and is sufficiently smooth at the same time.
- This can be formulated as finding f to minimize

$$R(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda J(f),$$

  where $J(f)$ is a measure of the roughness of f, and $\lambda > 0$ is a parameter controlling the tradeoff between the smoothness and the quality of fit.
- $J(f)$ is also called a regularizer.

Measuring roughness

- For a quadratic function $f(x) = cx^2$, large $f''(x)$ indicates that the curve is very wiggly.
- In general, for any function f, if $f''(x)$ is usually large, then f looks very wiggly.
- We can use

$$J(f) = \int_a^b f''(x)^2 dx$$

as a measure for overall roughness of f over $[a, b]$.

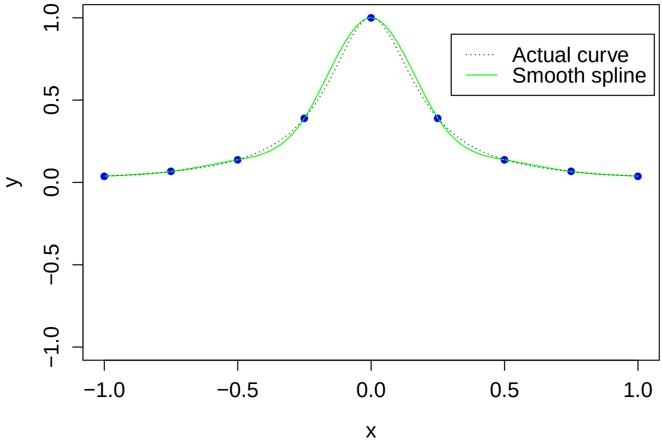
Smoothing splines

- Assume that $a < \min_i x_i$, and $b > \max_i x_i$.
- Consider the problem of finding a function f minimizing

$$R(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b f''(x)^2 dx.$$

- When $\lambda = 0$, f can be any function passing through the data.
- When $\lambda = \infty$, f is the OLS fit.
- When $0 < \lambda < \infty$, f is a natural cubic spline with knots at the unique x_i values.

Revisiting the example



A smoothing spline can fit the data well and is smooth!

A basis for natural cubic spline

- Recall: natural splines are linear at two ends.
- Assume that the knots are $t_1, \ldots, t_m$.
- A natural cubic spline is a linear combination of the following m basis functions

$$n_1(x) = 1, \quad n_2(x) = x,$$
$$n_{2+i}(x) = d_i(x) - d_{m-1}(x), \quad i = 1, \ldots, m-2,$$

where $d_i(x) = \frac{(x-t_i)_+^3 - (x-t_m)_+^3}{t_m - t_i}$.

Fitting a smoothing spline

- Training data: $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbf{R} \times \mathbf{R}$.
- A smoothing spline is fitted by minimizing

$$\hat{\beta} = \sum_{i=1}^{n} (\beta^\top \mathbf{z}_i - y_i)^2 + \lambda \beta^\top \Omega \beta,$$

  where $\mathbf{z}_i = (n_1(x_i), \ldots, n_n(x_i))$, n_i's use x_i's as the knots, and $\Omega_{ij} = \int n_i''(x) n_j''(x) dx$.
- The fitted spline is

$$f(x) = \sum_i \hat{\beta}_i n_i(x).$$

Matrix form

- Let $\mathbf{Z}$ be the $n \times n$ matrix with $\mathbf{z}_i$ as the i-th row.
- Then $\hat{\beta}$ can be written as

$$\hat{\beta} = (\mathbf{Z}^\top \mathbf{Z} + \lambda\Omega)^{-1}\mathbf{Z}^\top \mathbf{y}.$$

- We thus have

$$\hat{\mathbf{y}} = \mathbf{Z}\hat{\beta} = S_\lambda \mathbf{y},$$

  where S_λ is the smoother matrix

$$S_\lambda = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z} + \lambda\Omega)^{-1}\mathbf{Z}^\top.$$

Effective degree of freedom

- The effective degree of freedom of a smoothing spline is

$$\mathrm{df}_\lambda = \mathrm{trace}(S_\lambda),$$

  where the trace of a matrix is the sum of its diagonal elements.
- The effective degree of freedom can be considered as a generalization of the concept of the number of free parameters.

Selection of smoothing parameters

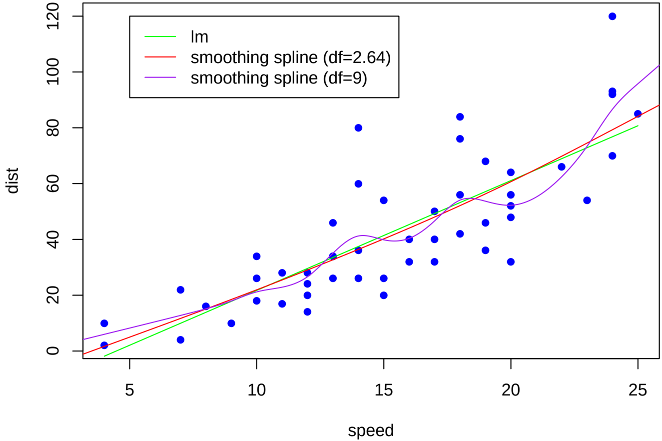
- The effective degree of freedom df_λ provides an intuitive way to manually specify the smoothing parameter λ.
- There are various procedures used for automatically determining the λ values, such as cross-validation, generalized cross validation.

Smoothing splines in R

```
> fit.spline.df <- smooth.spline(cars$speed, cars$dist, df=9)
Smoothing Parameter  spar= 0.3858413  lambda= 0.0001576001 (11
    iterations)
Equivalent Degrees of Freedom (Df): 8.998755
Penalized Criterion (RSS): 2054.319
GCV: 262.3012

> fit.spline.gcv <- smooth.spline(cars$speed, cars$dist)
Smoothing Parameter  spar= 0.7801305  lambda= 0.1112206 (11
    iterations)
Equivalent Degrees of Freedom (Df): 2.635278
Penalized Criterion (RSS): 4187.776
GCV: 244.1044
```

- By default, the smoothing parameter λ is determined using generalized cross validation.



lm
smoothing spline (df=2.64)
smoothing spline (df=9)
dist
0
20
40
60
80
100
120
5
10
15
20
25
speed

Generalized Additive Models

- Smoothing spline is a nonparametric analogue of OLS.
- We can extend the approach to GLM.

Idea

- Replace the linear predictor by $\beta_0 + h_1(x_1) + \ldots + h_d(x_d)$.
- Maximize roughness penalized log-likelihood instead of log-likelihood.

Generalized additive model (GAM)

- Recall: A GLM has the following structure

$$\begin{aligned} &\text{(systematic)} \quad \mathbb{E}(Y \mid \mathbf{x}) = h(\beta^\top \mathbf{x}), \\ &\text{(random)} \quad Y \mid \mathbf{x} \text{ follows an exponential family distribution.} \end{aligned}$$

- A *generalized additive model* has the following structure

$$\begin{aligned} &\text{(systematic)} \quad \mathbb{E}(Y \mid \mathbf{x}) = h(\beta_0 + \sum_i h_i(x_i)) \\ &\text{(random)} \quad Y \mid \mathbf{x} \text{ follows an exponential family distribution.} \end{aligned}$$

This defines a conditional probability model

$$p(y \mid \mathbf{x}, \beta_0, h_1, \ldots, h_d)$$

Roughness penalty approach for GAM

- We want to choose β_0, $h_1, \ldots, h_d$ to maximize

$$\sum_i \ln p(y_i \mid \mathbf{x}_i, \beta_0, h_1, \ldots, h_d) - \sum_j \lambda_j \int h_j''(x_j)^2 dx_j.$$

- Again, if each $\lambda_j > 0$, then each h_j must be a natural cubic spline with knots at the unique values of x_j.
- This reduces the problem to a finite-dimensional parametric regression problem.

Remarks

- Higher order derivatives may be used in the regularizer (smoothness penalty).
- We can also use regression splines instead of smoothing splines to represent h_i's.
- h_i's may use a mix of different representations.

  e.g. $h_1(x_1) = x_1$, $h_2(x_2)$ *a regression spline,* $h_3(x_3)$ *a smoothing spline...*

What You Need to Know

- Smoothing splines
  - The roughness penalty approach
  - Natural cubic splines as smoothing splines
  - Smoothing parameter and effective degree of freedom
- Generalized additive model
  - GAM as a generalization of GLM
  - Roughness penalty approach for GAM