ISyE 6416: Computational Statistics
Spring 2017

Lecture 10: Spline

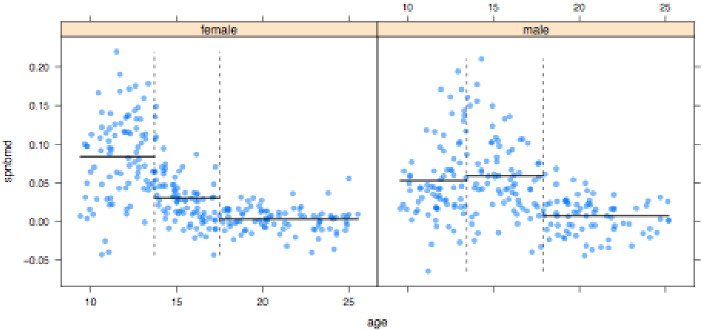
Prof. Yao Xie

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

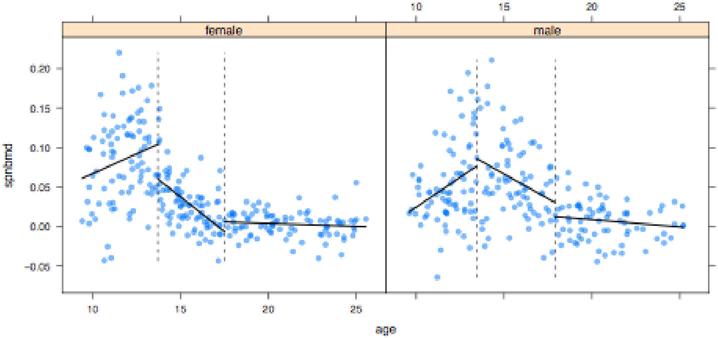
Motivation: non-linear regression

- Bone mineral density versus age for male versus female.
- To deal with non-linearity: split the data into a number a parts; perform a regression on each part.
- Splitting either via evenly spaced "knots", or via known locations based on external information.

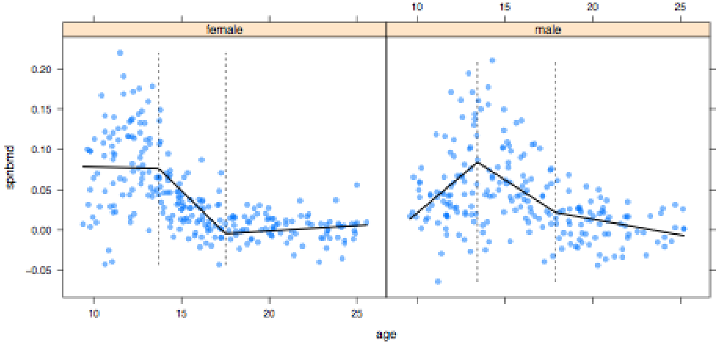
Piecewise constant model



Piecewise linear model



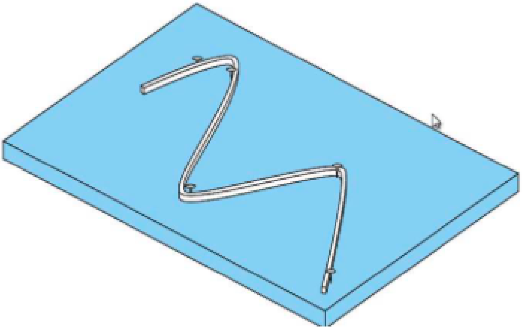
Continuous piecewise linear model



Spline

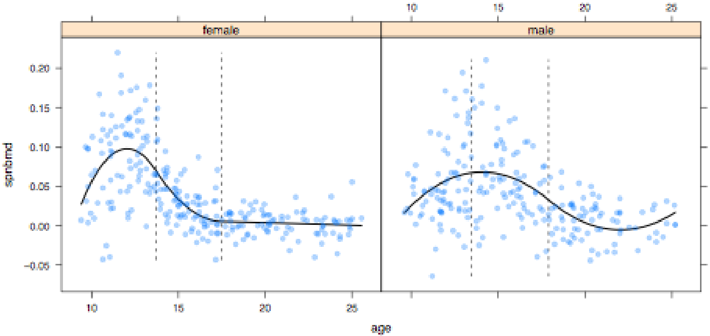
- A spline is a piecewise polynomial function.
- A cubic spline is 3rd order polynomial.
- Fit piecewise continuous splines to noisy data.

The concept of spline is using a thin , flexible strip (called a spline) to draw smooth curves through a set of points.

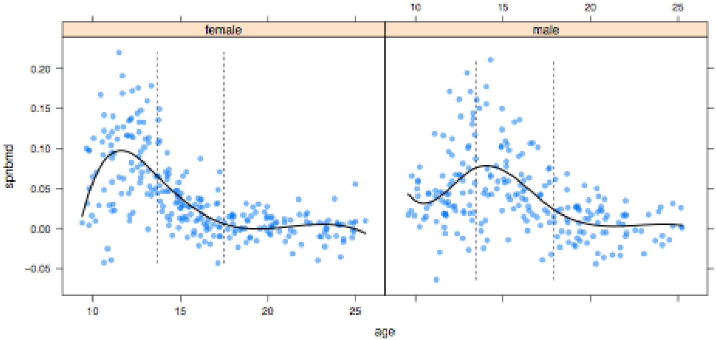


https://www.youtube.com/watch?v=LgodR7pwwW8

Quadratic splines

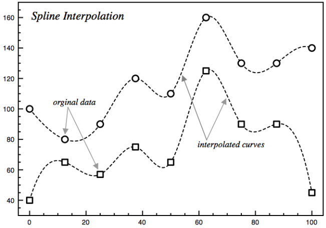


Cubic splines

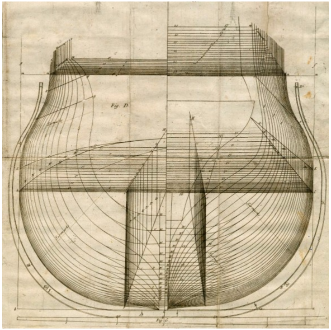


Formal definition

- Assume $f(x_i) = f_i$ of the function $f(x)$ at the points $x_0 < x_1 < \ldots < x_n$.
- A cubic interpolating spline $s(x)$ is a function on the interval $[x_0, x_n]$ satisfying
  - $s(x)$ is a cubic polynomial on each node-to-node interval $[x_i, x_{i+1}]$
  - $s(x_i) = f_i$ at each node x_i
  - the second order derivative $s''(x)$ exists and is continuous throughout the entire interval $[x_0, x_n]$
  - at the terminal nodes, $s''(x_0) = s''(x_n) = 0$



- Cubic splines are derived from the physical laws that govern bending of thin beams.
- An approximate solution of the minimum energy bending equation, valid when the amount of bending is small.



Properties of spline

- There is exactly one function $s(x)$ on $[x_0, x_n]$ satisfying these properties.
- Intuitively, these requirements leads to well-defined math problems.
- For n knots, the number of parameters can be $4n$
- At the same time,
  - $2n$ zeroth-order condition $s(x_i) = f_i$
  - $n-1$ first order condition $s'(x)$ continuous at knots
  - $n+1$ second order conditions

  Number of unknowns $=$ number of parameters (necessary condition)

Computation for a spline

- inter-knot distances $h_i = x_{i+1} - x_i$
- second order derivative $\sigma_i = s''(x_i)$ ($n+1$ parameters to parameterize the cubic spline function)
- we can derive the following

$$M\sigma = Qf$$

$$M = \begin{bmatrix} \frac{1}{3}(h_0+h_1) & \frac{h_1}{6} & 0 & \cdots & 0 & 0 \\ \frac{h_1}{6} & \frac{1}{3}(h_1+h_2) & \frac{h_2}{6} & \cdots & 0 & 0 \\ 0 & \frac{h_2}{6} & \frac{1}{3}(h_2+h_3) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{3}(h_{n-3}+h_{n-2}) & \frac{h_{n-2}}{6} \\ 0 & 0 & 0 & \cdots & \frac{h_{n-2}}{6} & \frac{1}{3}(h_{n-2}+h_{n-1}) \end{bmatrix}$$

$$\sigma = [\sigma_1, \cdots, \sigma_{n-1}], \quad f = [f_0, f_1, \ldots, f_n]$$

$$Q = \begin{bmatrix} 1/h_0 & -1/h_0 - 1/h_1 & 1/h_1 & & & & \\ & 1/h_1 & -1/h_1 - 1/h_2 & 1/h_2 & & & \\ & & \ddots & \ddots & & \ddots & \\ & & & 1/h_{n-2} & -1/h_{n-2} - 1/h_{n-1} & 1/h_{n-1} \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n+1)}$$

Solving the linear system of equations

- Matrix M is symmetric and positive definite, and tridiagonal
- Cholesky factorization

$$M = LDL^T$$

where

$$L = \begin{bmatrix} 1 & & & & \dots & 0 \\ a_1 & 1 & & & & \vdots \\ & \ddots & \ddots & & & \\ \vdots & & \ddots & \ddots & & \\ 0 & \dots & & & a_{n-2} & 1 \end{bmatrix}$$

and D is a diagonal matrix.

This enables efficient inverse of the matrix

$$\sigma = M^{-1}Qf = (L^T)^{-1}D^{-1}L^{-1}Qf$$

inversion of L and D has $\mathcal{O}(n)$ complexity.

Final expressions for splines

$$
\begin{aligned}
s_i(x) = {} & \frac{\sigma_i}{6h_i}(x_{i+1} - x)^3 + \frac{\sigma_{i+1}}{6h_i}(x - x_i)^3 \\
& + \left(\frac{f_{i+1}}{h_i} - \frac{\sigma_{i+1} h_i}{6} \right)(x - x_i) + \left(\frac{f_i}{h_i} - \frac{\sigma_i h_i}{6} \right)(x_{i+1} - x) \\
& i = 0, 1, \ldots, n-1.
\end{aligned}
$$

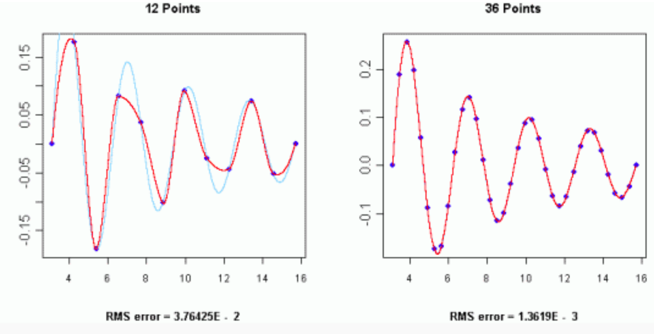
Minimum energy property

- Why spline? For any other twice continuously differentiable function
$$\int_{x_0}^{x_n} [g''(x)]^2 dx \geq \int_{x_0}^{x_n} [s''(x)]^2 dx$$

Error bound

Suppose that $f(x)$ is twice continuously differentiable and $s(x)$ is the spline interpolating $f(x)$ at the knots $x_0 < x_1 < \cdots < x_n$. If $h = \max_{0 \le i \le n-1}(x_{i+1} - x_i)$ then

$$\max_{x_0 \le x \le x_n} |f(x) - s(x)| \le h^{3/2}[\int_{x_0}^{x_n} f''(y)^2 dy]^{1/2}.$$



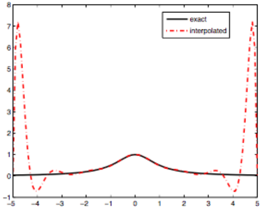
$f(x) = \sin(2x)/x.$

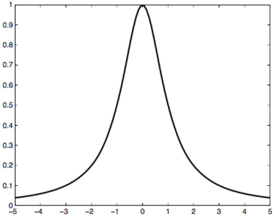
Problem with fitting a global polynomial

Runge's example

$$f(x) = \frac{1}{1+x^2}$$

High order interpolation using a global polynomial often exhibit these oscillations





- $f(x)$ interpolated using 15th order polynomial based on equidistant sample points.
- $f(x)$ interpolated using cubic spline based on 15 equidistant samples.

Example

i	0	1	2	3
x_i	0.9	1.3	1.9	2.1
y_i	1.3	1.5	1.85	2.1
$h_i = x_{i+1} - x_i$	0.4	0.6	0.2	

The equation for solving σ becomes

$$\begin{bmatrix} 2.0 & 0.4 \\ 0.4 & 1.6 \end{bmatrix} \begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.4 \end{bmatrix}$$

$\Rightarrow \sigma_1 = 0.2105, \sigma_2 = 0.1974$
$\Rightarrow$

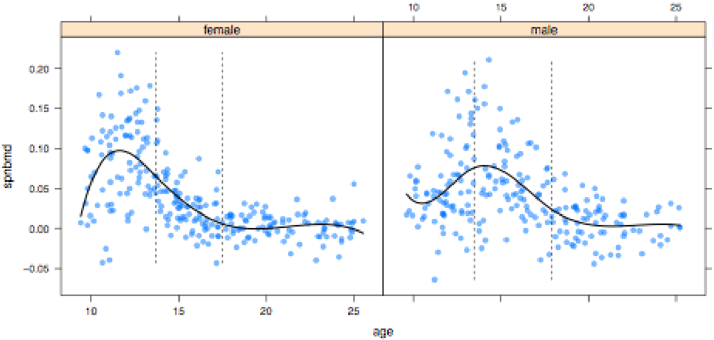
$$S_0(x) = 0.0877(x - 0.9)^3 + 3.736(x - 0.9) + 3.25(1.3 - x)$$
$$S_1(x) = 0.0585(x - 1.3)^3 + 0.0548(1.9 - x)^3 + 3.0636(x - 1.3) + 2.4790(1.9 - x)$$
$$S_2(x) = 0.1645(x - 1.9)^3 + 10.5(x - 1.9) + 9.2434(2.1 - x)$$

Nonlinear regression

- Given responses y_i, and variables x_i

$$y_i = f(x_i) + \epsilon_i, \quad i = 0, \ldots, n$$

f: unknown regression function

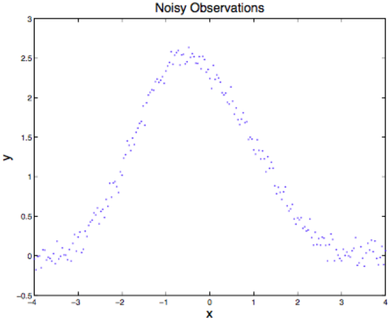


Nonlinear regression

- Given weights $w_0, w_1, \ldots, w_n$, $w_i > 0$, minimize

$$J_\alpha(s) = \alpha \sum_{i=0}^{n} w_i [y_i - s(x_i)]^2 + (1-\alpha) \int_{x_0}^{x_n} [s''(x)]^2 dx$$

- tradeoff between smoothness of s and goodness of fit
  $\alpha \in (0,1)$



Matrix-vector parameterization

- One can show
$$\int_{x_0}^{x_n} s''(x)^2 dx = \sigma^T M \sigma$$
$$J_\alpha(f) = \alpha(y - f)^T W(y - f) + (1 - \alpha) f^T Q^T M^{-1} Q f$$
where $W = \mathrm{diag}\{w_0, \ldots, w_n\}$
- spline function s parameterized by f
- solution
$$\hat{f} = [\alpha W + (1 - \alpha) Q^T M^{-1} Q]^{-1} \alpha W y$$
- one can show
$$\hat{\sigma} = [\alpha M + (1 - \alpha) Q^T W^{-1} Q]^{-1} \alpha Q y$$

Cross validation

- For notational convenience, we reformulate the optimization problem

$$J_\lambda(s) = \sum_{i=0}^{n} w_i[y_i - s(x_i)]^2 + \lambda \int_{x_0}^{x_n} [s''(x)]^2 dx$$

$\lambda = (1-\alpha)/\alpha$

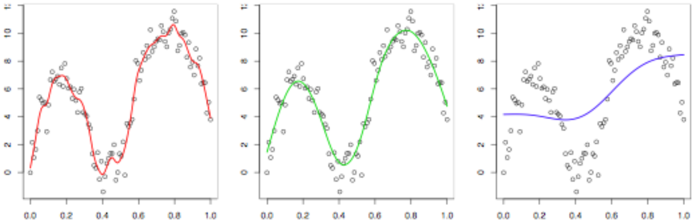
- Define leave-one-out cost function, for $1 \leq k \leq n$

$$h_\lambda^{(-k)}(x) = \arg\min_s \sum_{i=0, i\neq k}^{n} w_i[y_i - s(x_i)]^2 + \lambda \int_{x_0}^{x_n} [s''(x)]^2 dx$$

- Define cross-validation criterion function

$$\mathsf{CV}(\lambda) = \sum_{k=0}^{n} [y_k - h_\lambda^{(-k)}(x_k)]^2$$

Example with $n = 100$ points:



λ too small $\qquad$ λ just right $\qquad$ λ too big

One can show

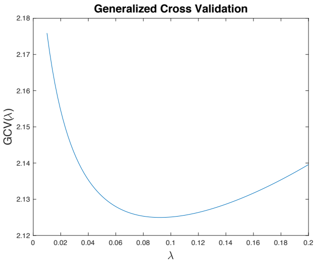
$$\text{CV}(\lambda) = \sum_{k=0}^{n} \frac{[y_k - \hat{f}(\lambda)_k]^2}{[1 - [S(\lambda)]_{kk}]^2}$$

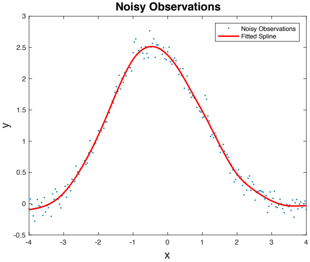
Generalized CV (GCV): replace $[S(\lambda)]_{kk}$ by its average, since it can get close to 1.

$$\text{GCV}(\lambda) = \sum_{k=0}^{n} \frac{[y_k - \hat{f}(\lambda)_k]^2}{[1 - \frac{\text{Tr}(S(\lambda))}{(n+1)}]^2}$$

where

$$S(\lambda) = [W + \lambda Q^T M^{-1} Q]^{-1} W$$





Bi-cubic interpolation