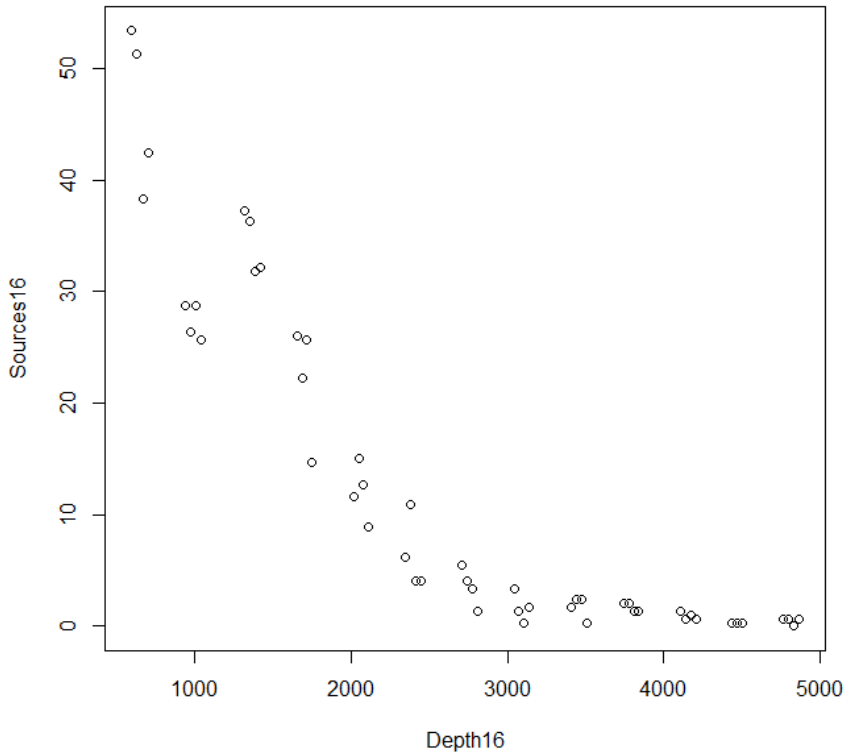
Biostatistics-Lecture 14
Generalized Additive Models

Ruibin Xi
Peking University
School of Mathematical Sciences

Generalized Additive models

- Gillibrand et al. (2007) studied the bioluminescence-depth relationship



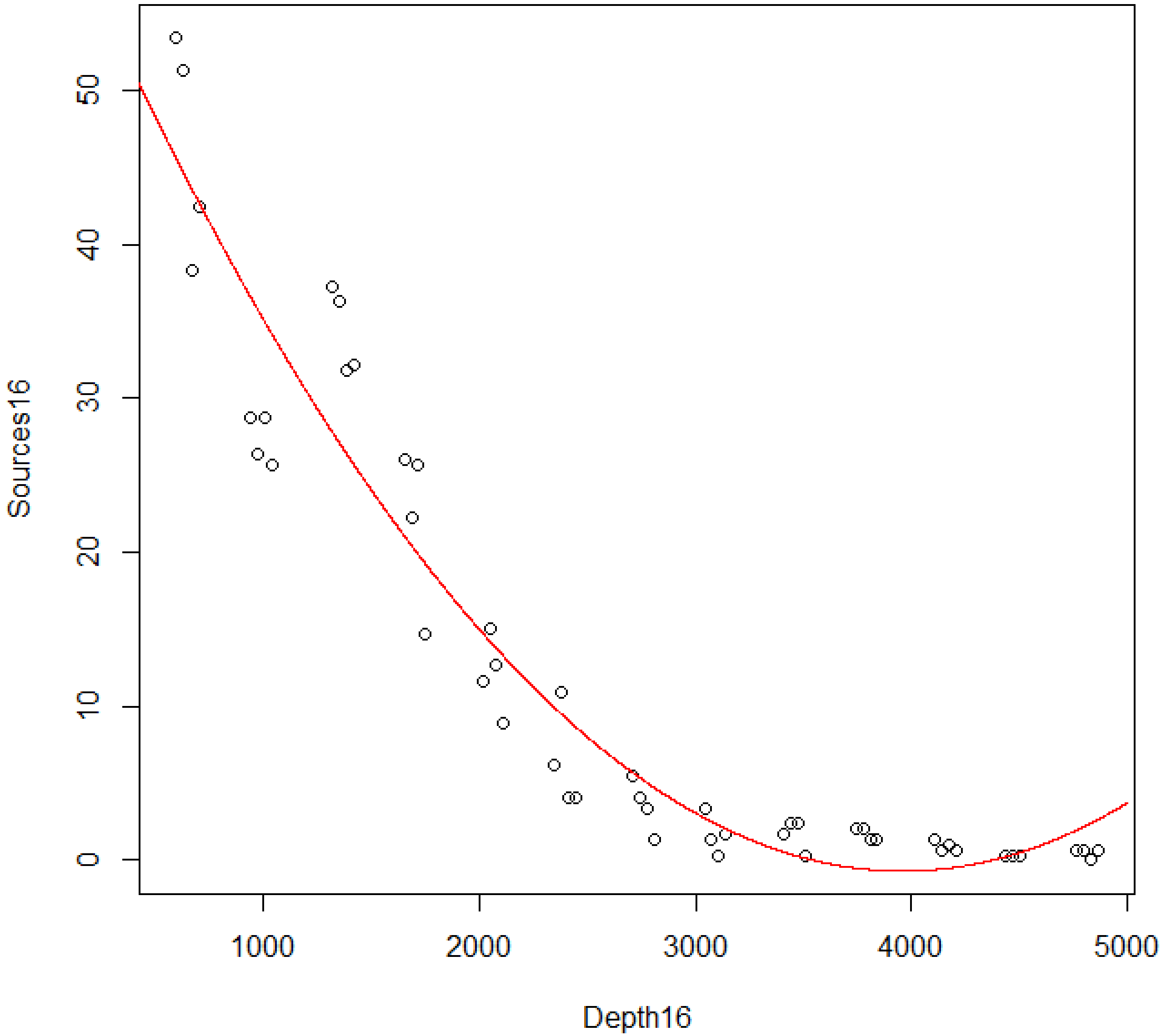
Generalized Additive models

- Gillibrand et al. (2007) studied the bioluminescence-depth relationship
- We may first consider the model

$$Y = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3 + \varepsilon$$

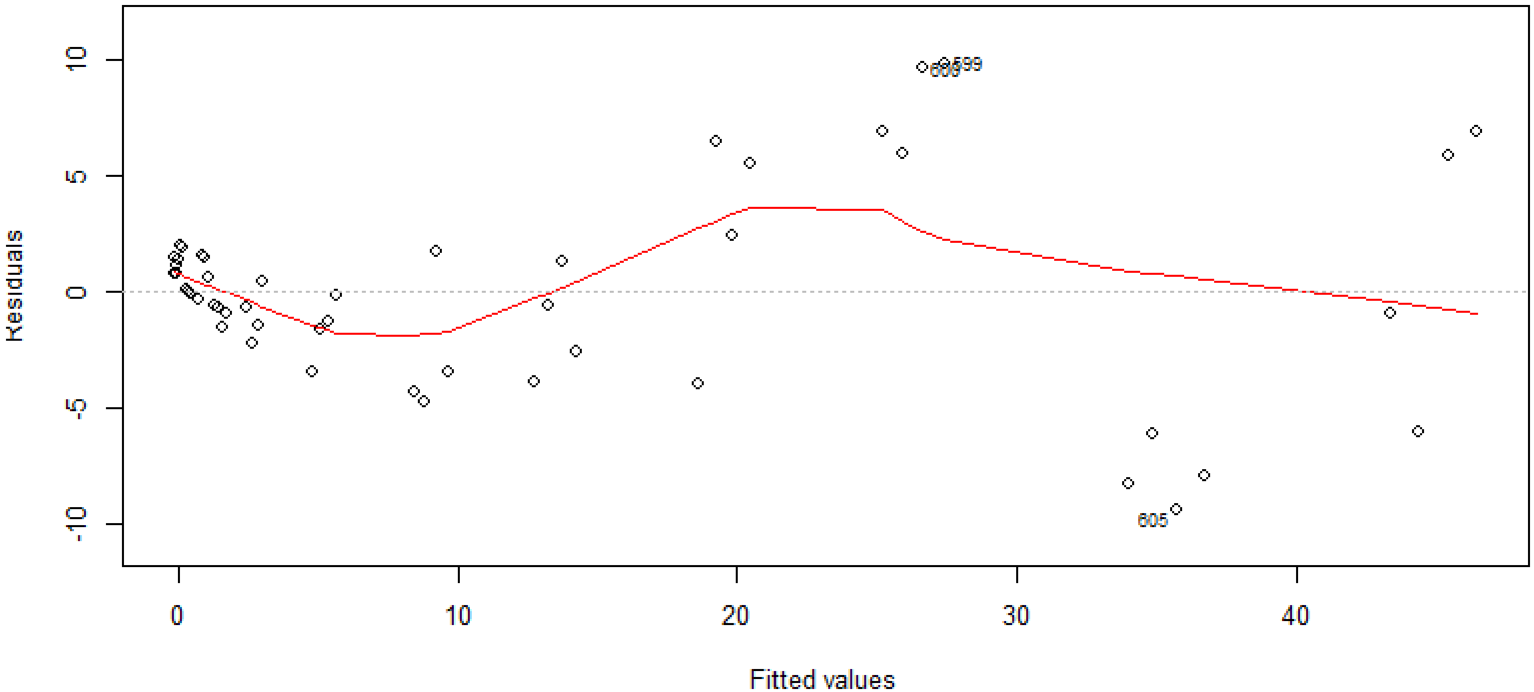
(

- Gilli
  biol
- We



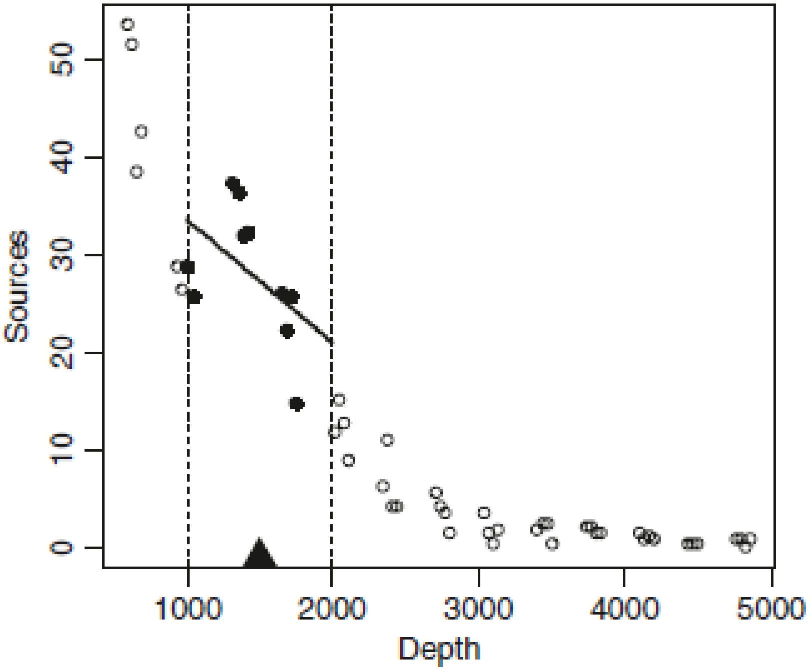
Generalized Additive models

- Gillibrand et al. (2007) studied the

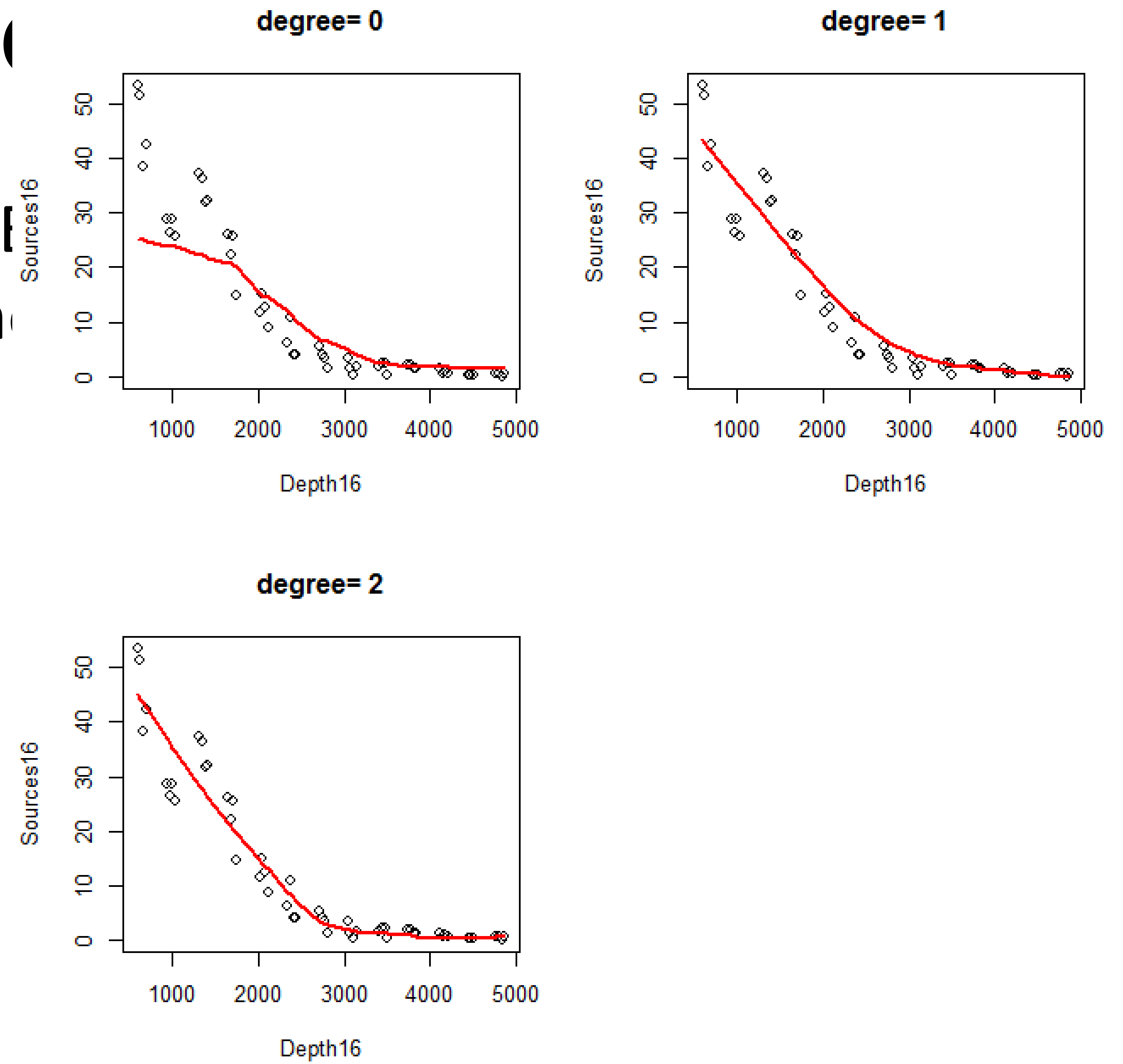


Generalized Additive models

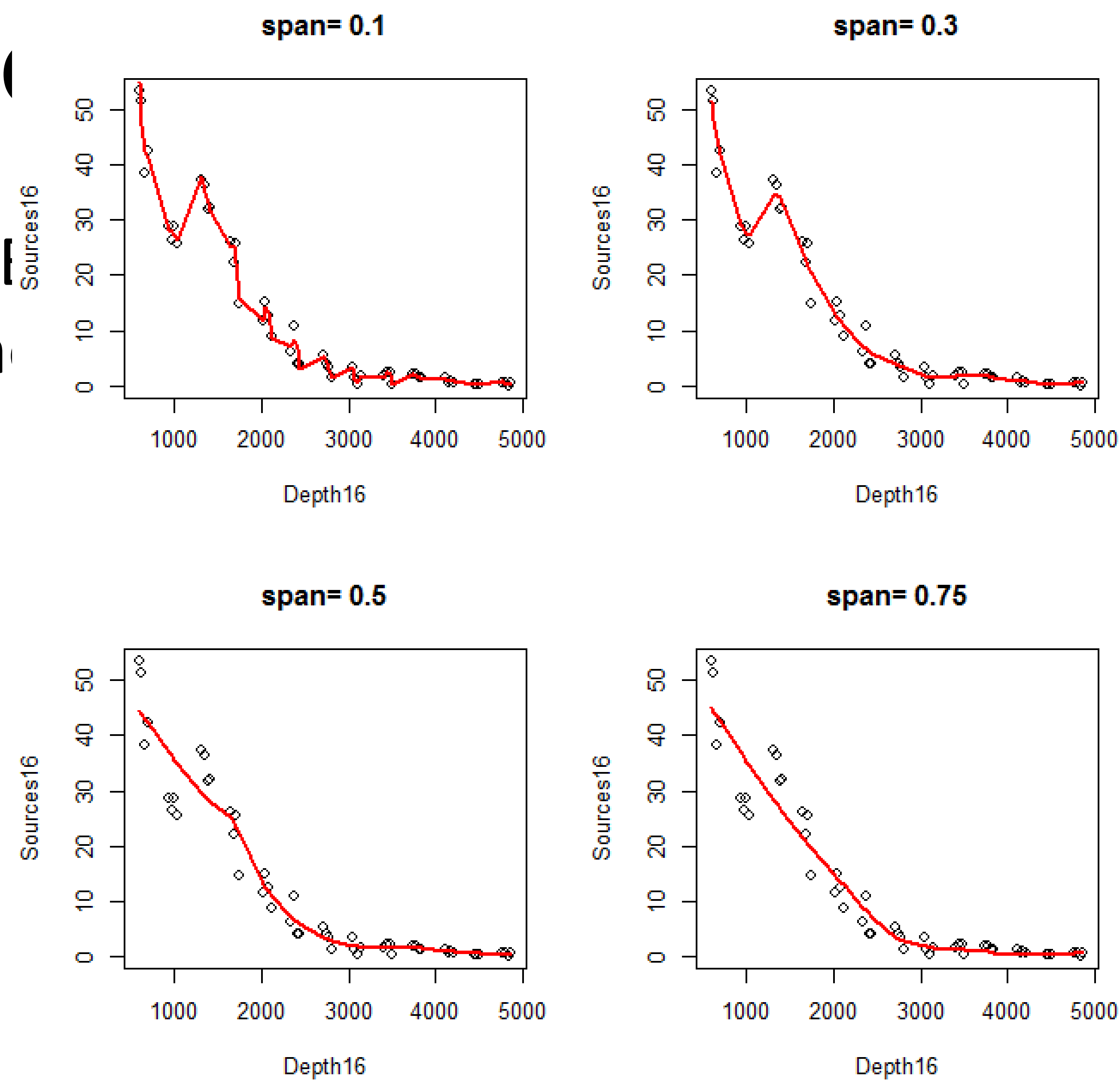
- LOESS (locally weighted scatterplot smoothing)-based



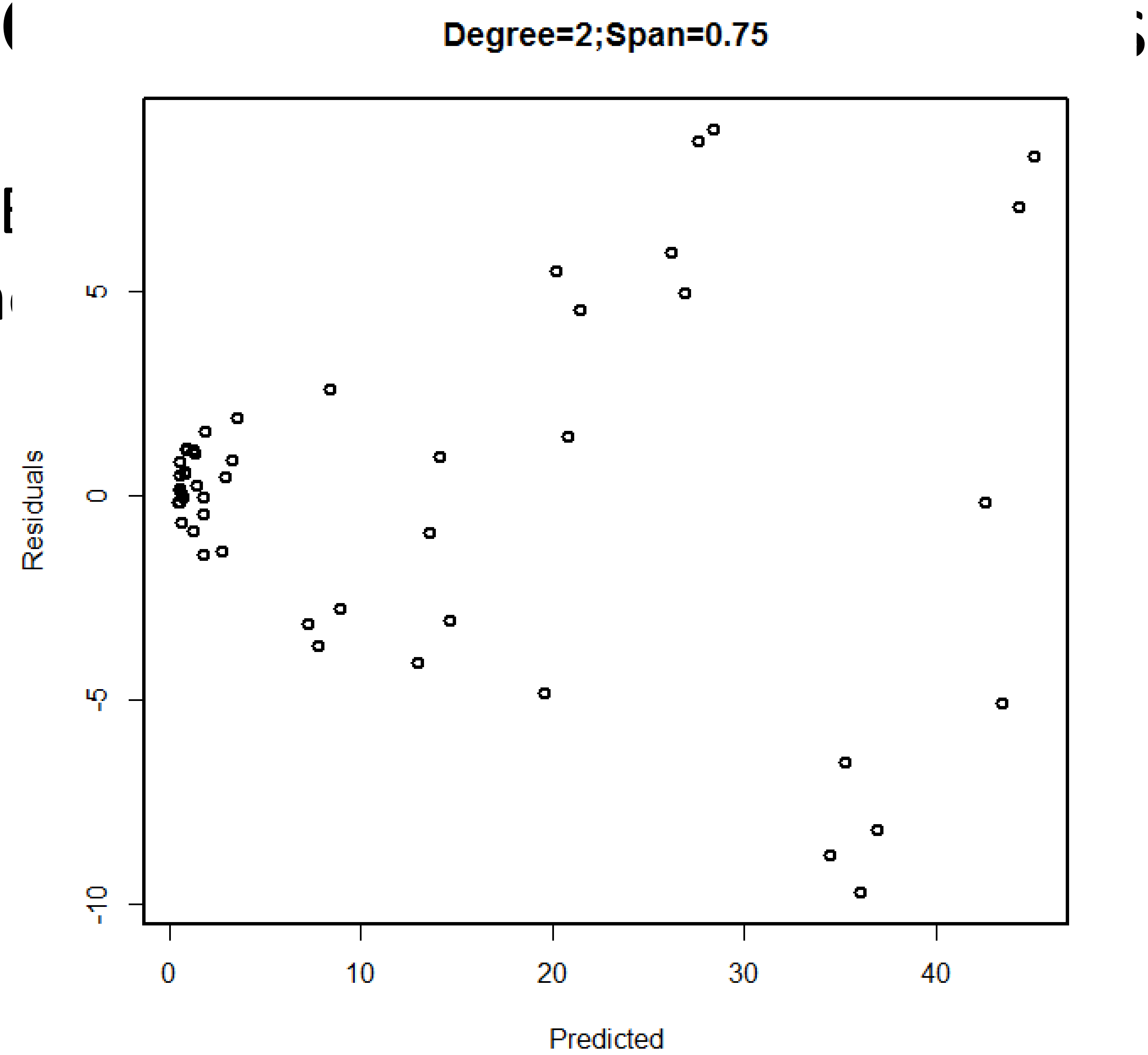
- LO
  sm


degree= 0
degree= 1
degree= 2
Sources16
Depth16
0
10
20
30
40
50
1000
2000
3000
4000
5000

• LO
sm


span= 0.1
span= 0.3
span= 0.5
span= 0.75
Sources16
Depth16
0
10
20
30
40
50
1000
2000
3000
4000
5000

- LO
sm

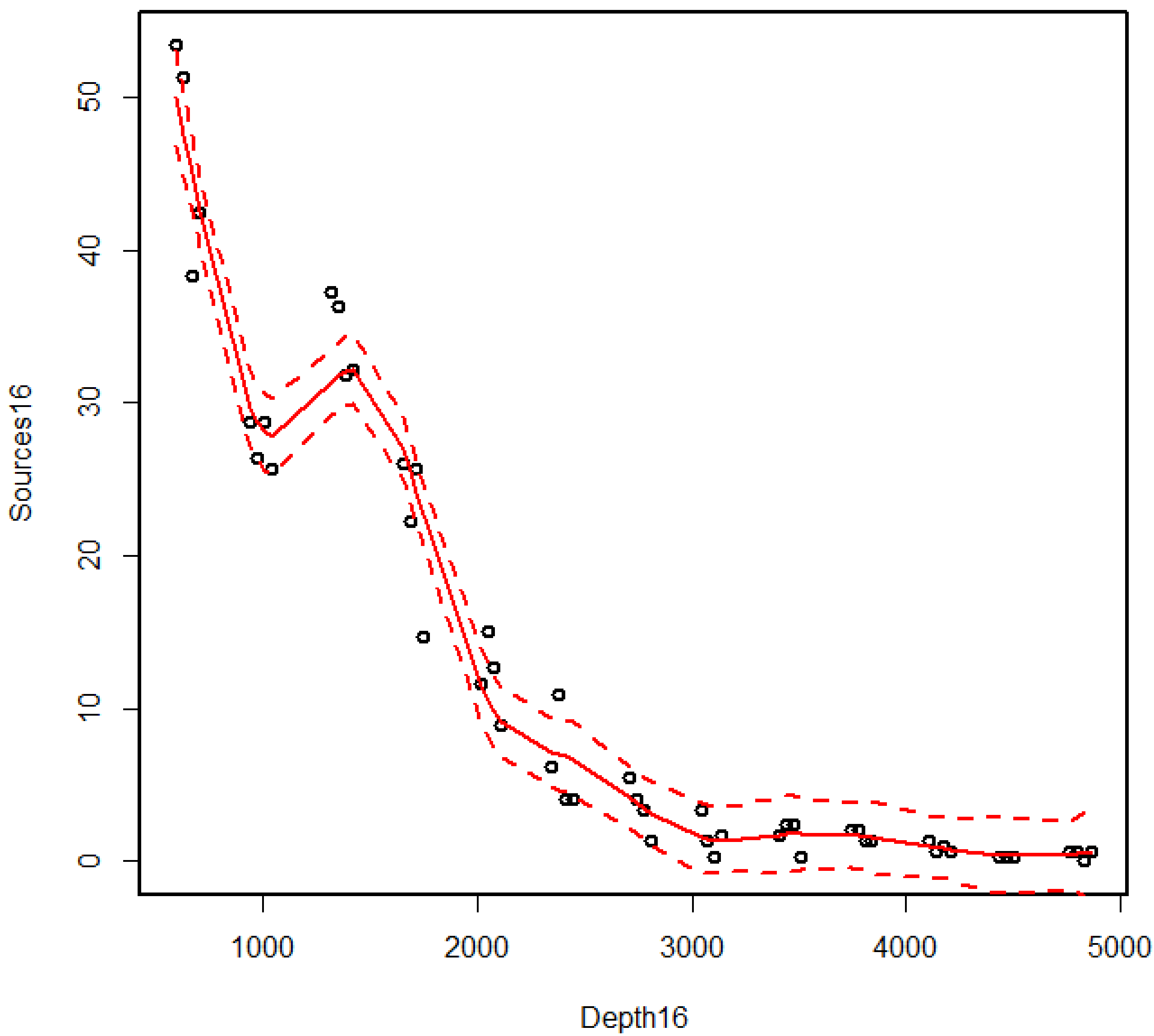


Generalized Additive models

- LOESS (locally weighted scatterplot smoothing)-based

- Spline-based

- LO
  sm
- Spl



Generalized Additive models

- Suppose that the model is

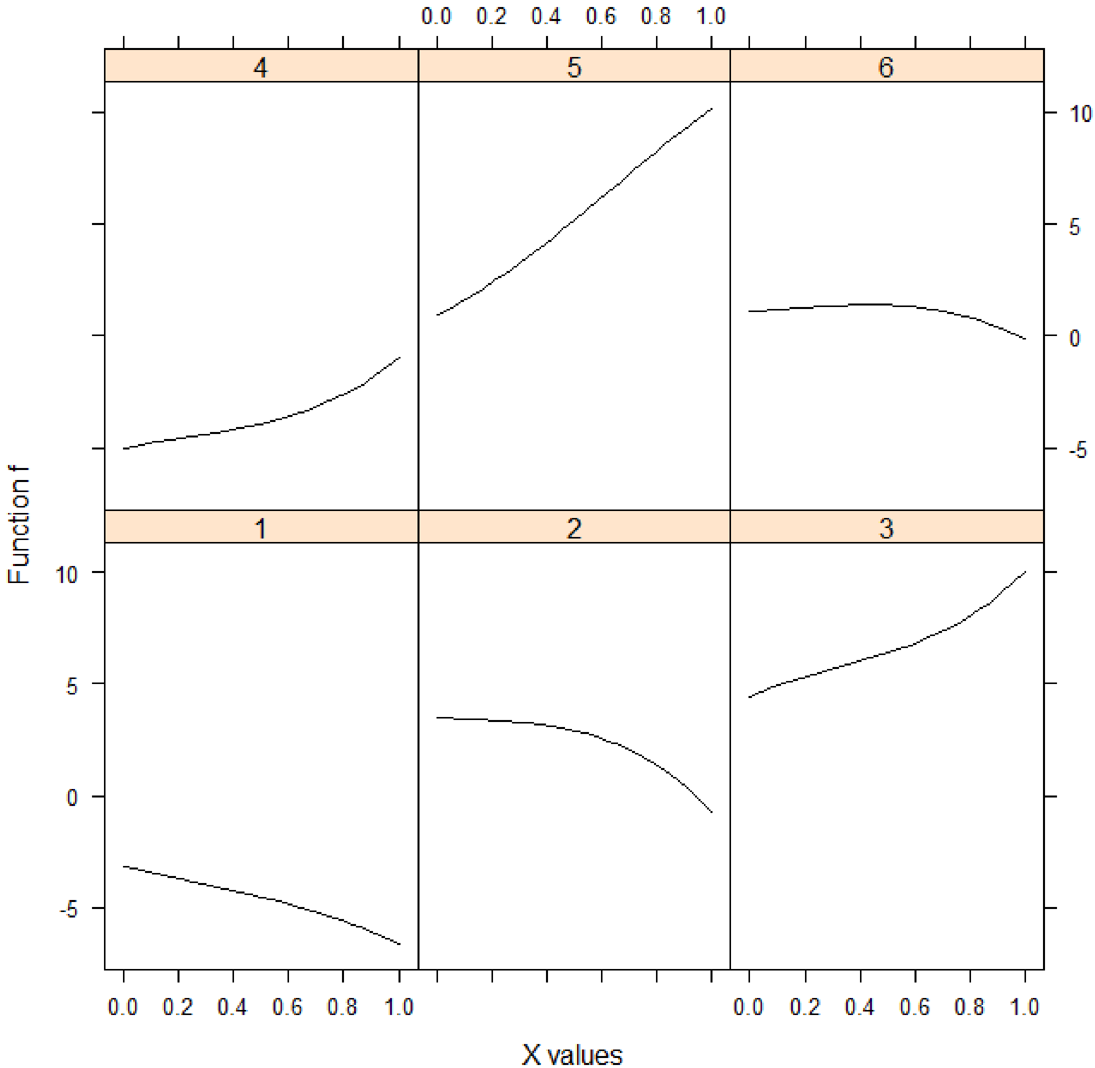
$$Y_i = \alpha + f(X_i) + \varepsilon_i \qquad \text{where} \quad \varepsilon_i \sim N(0, \sigma^2)$$

- As discussed before, we may approximate the unknown function f with polynomials

$$f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3$$

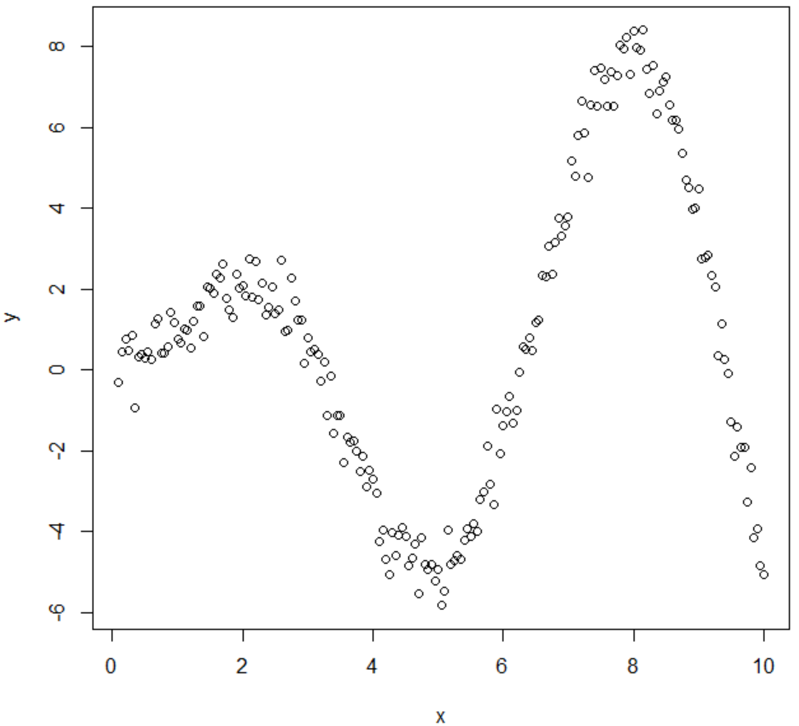
- But the problem is this representation is not very flexible

Generalized Additive models



Generalized Additive models

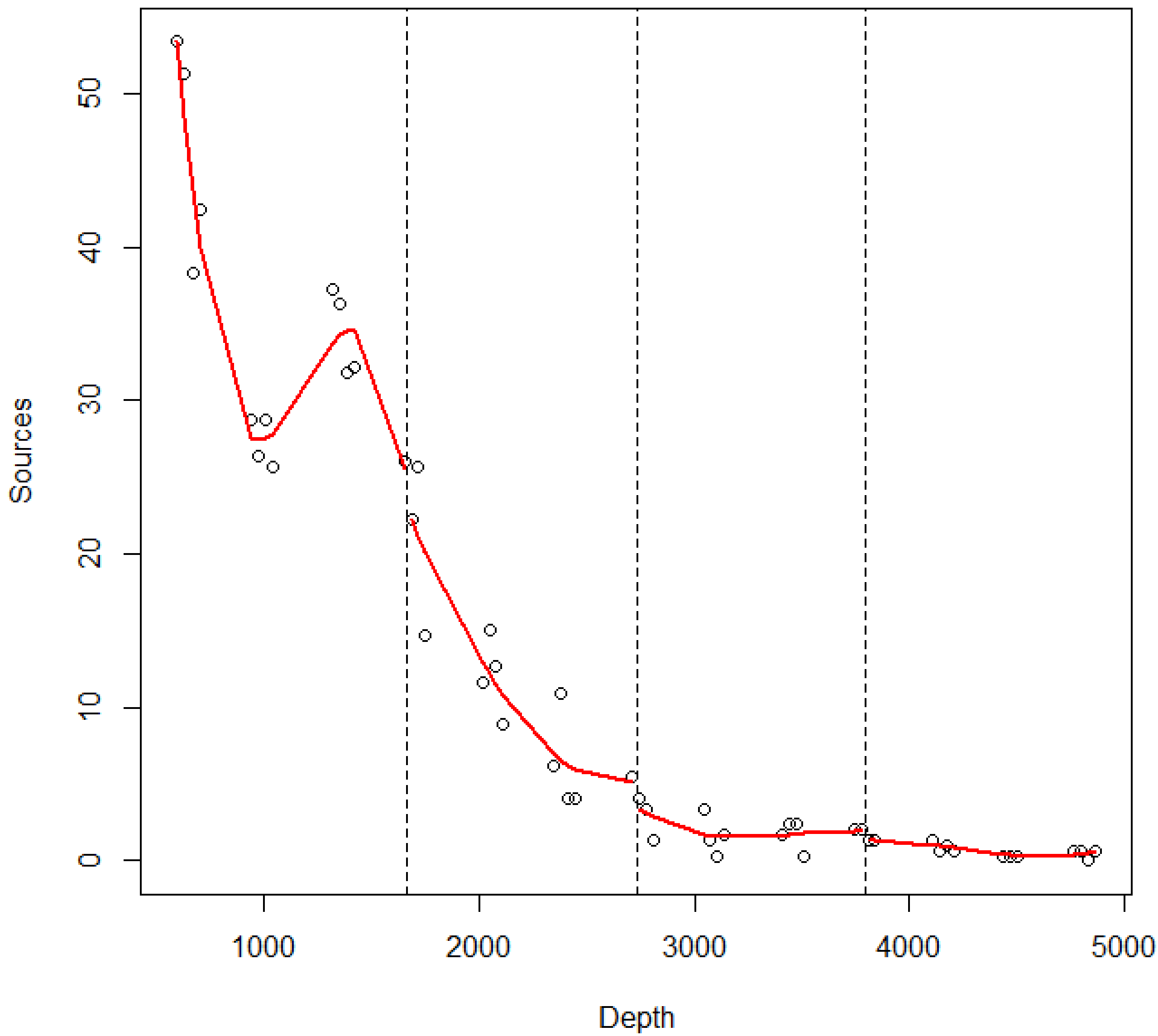
- What if the data looks like



Generalized Additive models

- One way to overcome this difficulty is to divide the range of the x variable to a few segment and perform regression on each of the segments

- On div seg the of



Generalized Additive models

- One way to overcome this difficulty is to divide the range of the x variable to a few segment and perform regression on each of the segments

- The cubic spline ensures the line looks smooth

Generalized Additive models

- Cubic spline bases (assuming 0<x<1)

$$b_1(x) = 1, b_2(x) = x$$

$$b_{i+2} = R(x, x_i^*) \text{ for } i = 1 \ldots q-2$$

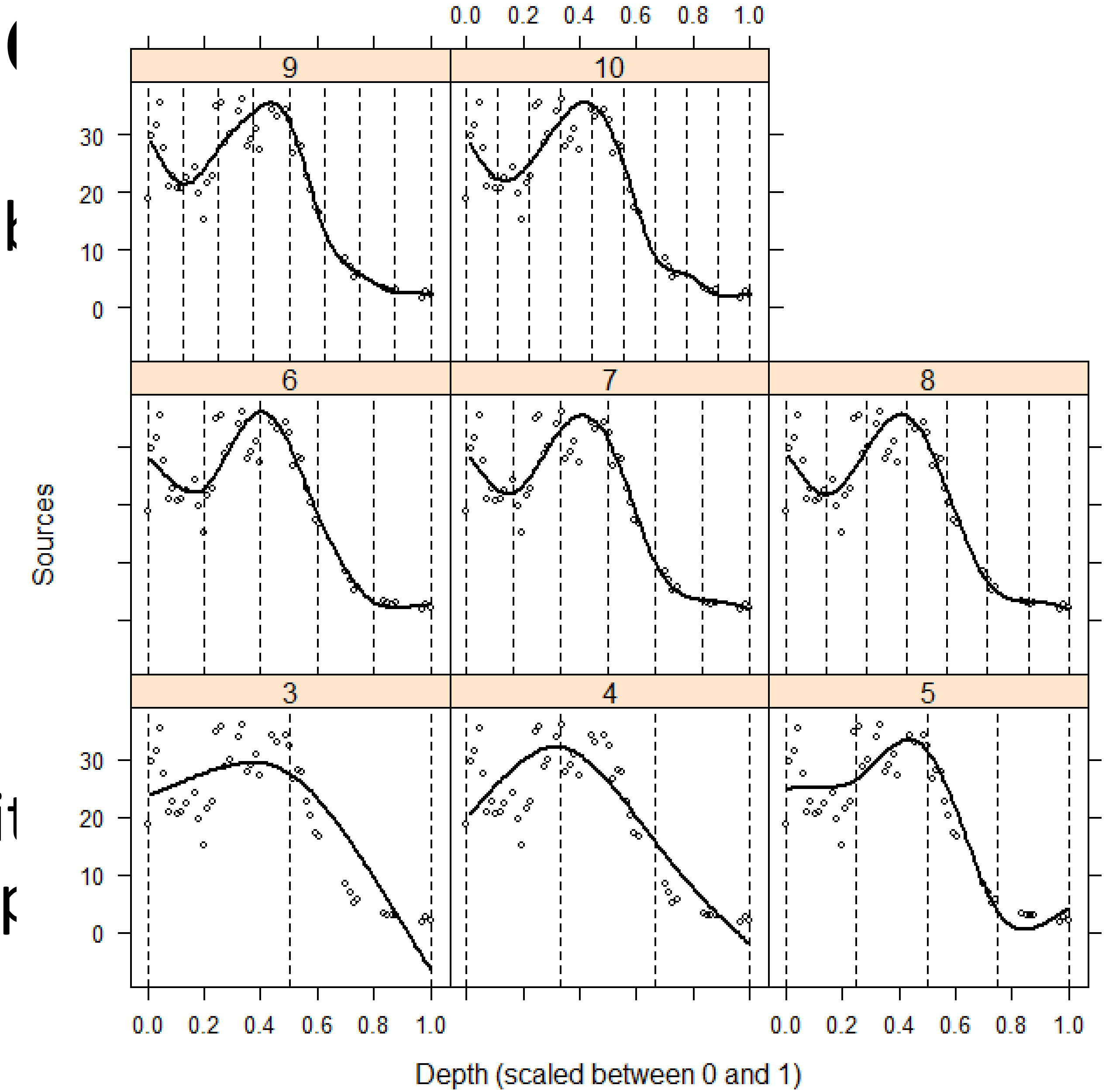
knots

$$R(x,z) = \left[(z-1/2)^2 - 1/12\right]\left[(x-1/2)^2 - 1/12\right]/4$$
$$- \left[(|x-z| - 1/2)^4 - 1/2\,(|x-z| - 1/2)^2 + 7/240\right]/24.$$

- With this bases, the model may be approximated by

$$Y_i = \alpha + \sum_{j=1}^{p} \beta_j \times b_j(X_i) + \varepsilon_i \qquad \text{where} \quad \varepsilon_i \sim N(0, \sigma^2)$$

- Cu



9
10
6
7
8
3
4
5
Sources
Depth (scaled between 0 and 1)
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30

- Wit
  app

Generalized Additive models

- How to determine the number of knots?
  - Use model selection methods?
  - But this is problematic
- Instead we may use the penalized regression spline
  - Rather than minimizing $\|\mathbf{y} - \mathbf{X}\beta\|^2$
  - We could minimize

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \int_0^1 [f''(x)]^2 dx$$

Generalized Additive models

- Because f is linear in the parameters β_i , the penalty can be written as

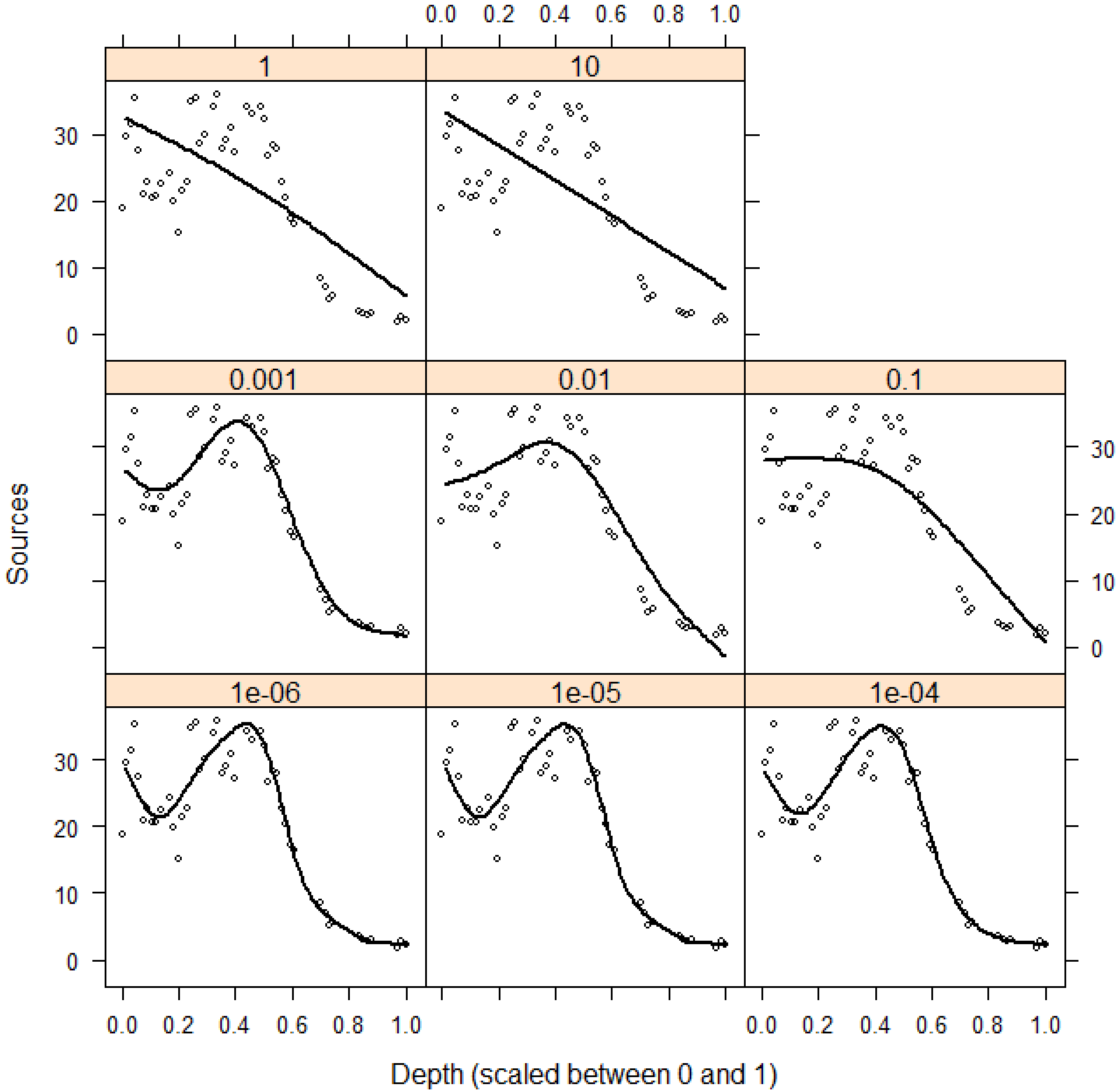
$$\int_0^1 [f''(x)]^2 dx = \boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta}$$

  - $S_{i+2,j+2} = R(x_i^*, x_j^*) \quad i,j = 1,\ldots,q-2$
  - The first two rows and columns are 0
- The penalized regression spline can be written as

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta}$$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

- Bec ne
  per
  - –
  - – .
- The itten
  as



Generalized Additive models

- How to choose λ?

- Ordinary cross-validation (OCV)
  - We my try to minimize

$$M = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \hat{f}(X_i))^2$$

  - Instead, we may minimize

$$\mathcal{V}_o = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - y_i)^2$$

$$\mathbb{E}(\mathcal{V}_o) \approx \mathbb{E}(M) + \sigma^2$$

Generalized Additive models

- It is inefficient to use the OCV by leaving out one datum at a time
- But, it can be shown that

$$\mathcal{V}_o = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_i)^2/(1 - A_{ii})^2$$

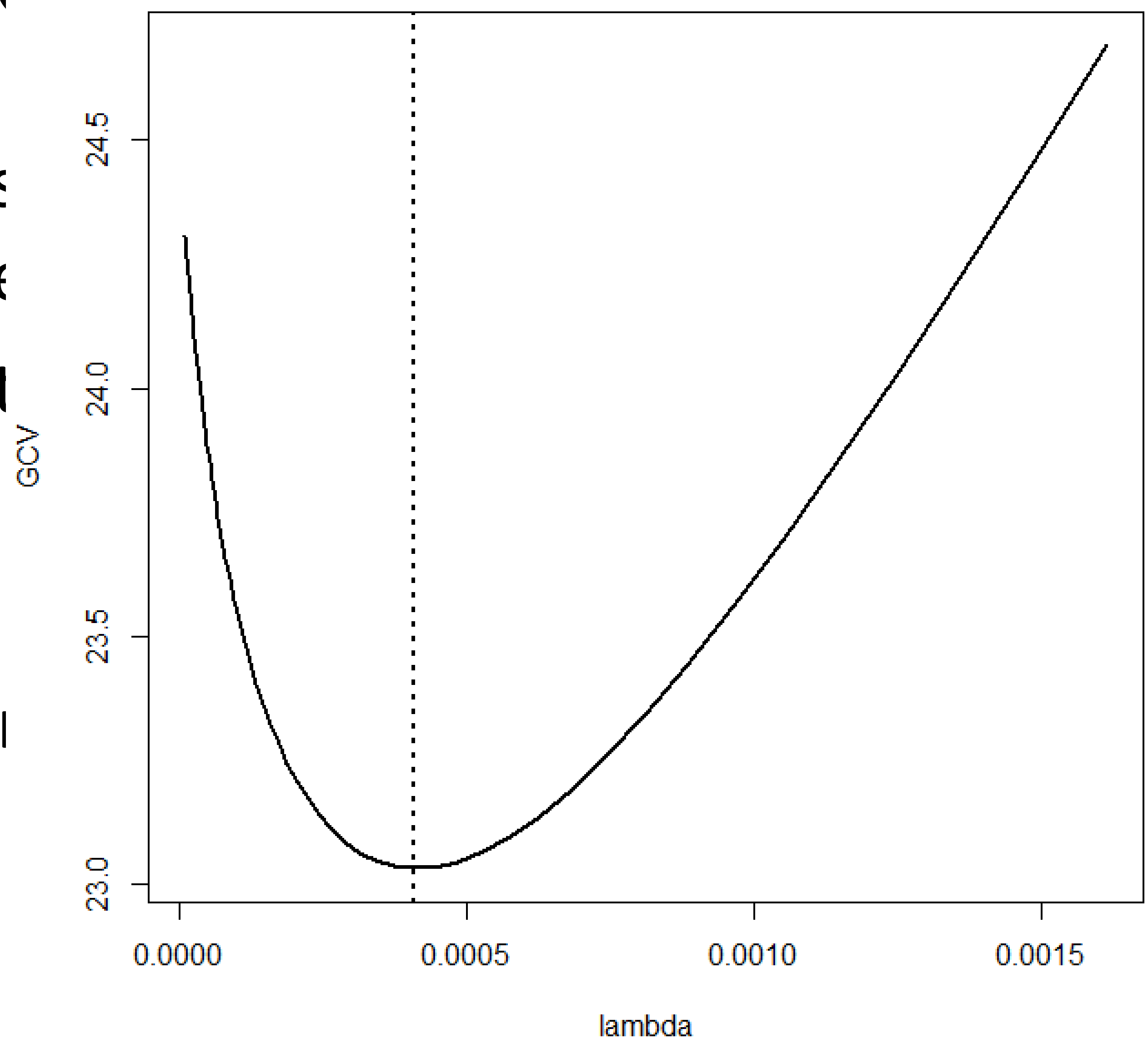
$$\mathbf{A} = \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\mathbf{X}^{\mathsf{T}}$$

- Generalized cross-validation (GCV)

$$\mathcal{V}_g = \frac{n\sum_{i=1}^{n}(y_i - \hat{f}_i)^2}{[tr(\mathbf{I} - \mathbf{A})]^2}.$$

(s

- It is out
  one
- But
- Ge



Generalized Additive models

- Additive models with multiple explanatory variables

$$Y_i = \alpha + f_1(X_i) + f_2(Z_i) + \varepsilon_i \qquad \text{where} \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$f_1(X_i) = \sum_{j=1}^{p} \beta_j \times b_j(X_i) \qquad f_2(Z_i) = \sum_{j=1}^{p} \gamma_j \times b_j(Z_i)$$

$$\|\mathbf{Y} - \mathbf{X} \times \boldsymbol{\beta}\|^2 + \lambda_1 \int f_1''(x)^2 dx + \lambda_2 \int f_2''(x)^2 dx$$

$$\|\mathbf{Y} - \mathbf{X} \times \boldsymbol{\beta}\|^2 + \boldsymbol{\beta}' \times \mathbf{S} \times \boldsymbol{\beta}$$

- We may also consider the model like

$$Y_i = \alpha + f_1(X_i) + \beta \times Z_i + \mathit{factor}(W_i) + \varepsilon_i \qquad \text{where} \quad \varepsilon_i \sim N(0, \sigma^2)$$