

## Cubic Smoothing Spline

Consider a regression problem  $Y_i = \eta(x_i) + \epsilon_i$ ,  $i = 1, \dots, n$ , where  $x_i \in [0, 1]$  and  $\epsilon_i \sim N(0, \sigma^2)$ . A *cubic smoothing spline* is the minimizer  $\eta_\lambda$  of a penalized least squares score,

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \eta(x_i))^2 + \lambda \int_0^1 \ddot{\eta}^2 dx.$$

Slide 1

It is a piecewise cubic polynomial, with the third derivative jumping at the *knots*  $\xi_1 < \xi_2 < \dots < \xi_q$ , the ordered distinctive  $x_i$ ; it is linear beyond the first and the last knots.

Alternatively, one may solve a constrained LS problem,

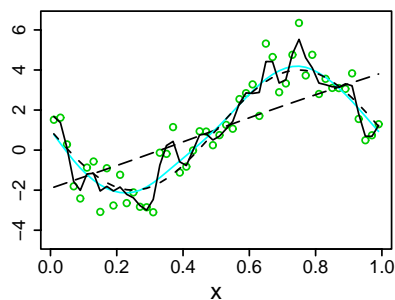
$$\min \frac{1}{n} \sum_{i=1}^n (Y_i - \eta(x_i))^2, \quad \text{subject to} \quad \int_0^1 \ddot{\eta}^2 dx \leq \rho.$$

## Role of Smoothing Parameter

With  $\lambda = \infty$ , one has  $\eta(x) = \beta_0 + \beta_1 x$ . With  $\lambda = 0$ , one has the minimum curvature interpolant.

The *smoothing parameter*  $\lambda$  controls the trade-off between the goodness-of-fit and smoothness of  $\eta_\lambda$ . Its proper selection is crucial to the practical performance.

Slide 2



Consider a simple simulation with  $x_i = (i - 0.5)/50$ ,  $i = 1, \dots, 50$ ,  $\eta(x) = 1 + 3 \sin(2\pi x - \pi)$ , and  $\sigma^2 = 1$ . The estimate  $\eta_\lambda$  was calculated at  $\log_{10} n\lambda = 0, -3, -6$ .

### Penalized Likelihood Method

To estimate  $\eta(x)$  on a generic domain  $\mathcal{X}$  using stochastic data, one may use the minimizer of

$$L(\eta|\text{data}) + (\lambda/2)J(\eta),$$

Slide 3

where  $L(\eta|\text{data})$  is the minus log likelihood and  $J(f)$  is a quadratic roughness penalty, with a null space  $\mathcal{N}_J = \{f : J(f) = 0\}$ .

Alternatively, one may use constrained maximum likelihood,

$$\min L(\eta|\text{data}), \quad \text{subject to } J(\eta) \leq \rho.$$

The solution usually falls on the boundary  $\{\eta : J(\eta) = \rho\}$ , and by the Lagrange method, the solution can be obtained through penalized likelihood.

### Penalized Likelihood Method

♣ REGRESSION: Observing  $(x_i, Y_i)$  from exponential family

$$Y|x \sim \exp\{(y\eta(x) - b(\eta(x)))/a(\phi) + c(y, \phi)\},$$

one may estimate  $\eta(x)$  using the minimizer

$$-\frac{1}{n} \sum_{i=1}^n \{Y_i \eta(x_i) - b(\eta(x_i))\} + \frac{\lambda}{2} J(\eta).$$

Slide 4

♣ DENSITY ESTIMATION: Observing  $X_i$  from  $f(x) = e^\eta / \int_{\mathcal{X}} e^\eta dx$ , one may estimate  $\eta(x)$  by the minimizer of

$$-\frac{1}{n} \sum_{i=1}^n \left\{ \eta(X_i) - \log \int_{\mathcal{X}} e^\eta dx \right\} + \frac{\lambda}{2} J(\eta).$$

A side condition, say  $\int_{\mathcal{X}} \eta dx = 0$ , shall be imposed on  $\eta$  for a one-to-one transform  $f \leftrightarrow e^\eta / \int_{\mathcal{X}} e^\eta dx$ .

### Penalized Likelihood Method

♣ HAZARD ESTIMATION: Let  $T$  be the lifetime of an item with survival function  $S(t|u) = P(T > t|u)$ , possibly dependent on a covariate  $U$ . The hazard rate is given by

$$e^{\eta(t,u)} = \lim_{\Delta t \rightarrow 0} \frac{P(t + \Delta t \geq T > t|u)}{(\Delta t)P(T > t|u)} = -\frac{\partial \log S(t|u)}{\partial t}.$$

Slide 5

Let  $Z$  be the left-truncation time and  $C$  be the right-censoring time, independent of  $T$  and of each other.

Observing  $(U_i, Z_i, X_i, \delta_i)$ , where  $X = \min(T, C)$ ,  $\delta = I_{[T \leq C]}$ , and  $Z < X$ , one may estimate the log hazard  $\eta$  by the minimizer of

$$-\frac{1}{n} \sum_{i=1}^n \left\{ \delta_i \eta(X_i, U_i) - \int_{Z_i}^{X_i} e^{\eta(t, U_i)} dt \right\} + \frac{\lambda}{2} J(\eta).$$

### ANOVA Decomposition

Consider one-way ANOVA  $Y_{ij} = \mu_i + \epsilon_{ij}$ ,  $i = 1, \dots, K$ , where  $\mu_i = \mu + \alpha_i$ . The identifiability is through a side condition such as  $\alpha_1 = 0$  or  $\sum_{i=1}^K \alpha_i = 0$ .

Recasting the model as  $Y_j = f(x_j) + \epsilon_j$ , where  $f(x)$  is defined on  $\mathcal{X} = \{1, \dots, K\}$ . The decomposition  $\mu_i = \mu + \alpha_i$  becomes

$$f(x) = Af + (I - A)f = f_\emptyset + f_x,$$

where  $A$  “averages out”  $x$  to return a constant; examples include  $Af = f(1)$  ( $\alpha_1 = 0$ ) and  $Af = \sum_{x=1}^K f(x)/K = \bar{f}$  ( $\sum_{i=1}^K \alpha_i = 0$ ).

On  $\mathcal{X} = [a, b]$ , one may use  $Af = f(a)$  or  $Af = \int_a^b f dx / (b - a)$ .

• The *averaging operator*  $A$  satisfies  $A(Af) = Af$ . The “*treatment effect*”  $f_x$  satisfies  $Af_x = 0$ .

Slide 6

### ANOVA Decomposition

For  $f(x) = f(x_{\langle 1 \rangle}, \dots, x_{\langle \Gamma \rangle})$  on  $\mathcal{X} = \prod_{\gamma=1}^{\Gamma} \mathcal{X}_{\gamma}$ , one has

$$f = \left\{ \prod_{\gamma=1}^{\Gamma} (I - A_{\gamma} + A_{\gamma}) \right\} f = \sum_{\mathcal{S}} \left\{ \prod_{\gamma \in \mathcal{S}} (I - A_{\gamma}) \prod_{\gamma \notin \mathcal{S}} A_{\gamma} \right\} f = \sum_{\mathcal{S}} f_{\mathcal{S}},$$

Slide 7

where  $\mathcal{S} \subseteq \{1, \dots, \Gamma\}$ ;  $f_{\emptyset} = \prod A_{\gamma} f$  is the constant,

$f_{\gamma} = f_{\{\gamma\}} = (I - A_{\gamma}) \prod_{\alpha \neq \gamma} A_{\alpha} f$  is the  $x_{\langle \gamma \rangle}$  main effect,

$f_{\gamma, \delta} = f_{\{\gamma, \delta\}} = (I - A_{\gamma})(I - A_{\delta}) \prod_{\alpha \neq \gamma, \delta} A_{\alpha} f$  is the  $x_{\langle \gamma \rangle} - x_{\langle \delta \rangle}$

interaction. The terms satisfy side conditions  $A_{\gamma} f_{\mathcal{S}} = 0, \forall \mathcal{S} \ni \gamma$ .

◇ **Proposition:** For any two sets of averaging operators  $A_{\gamma}$  and  $\tilde{A}_{\gamma}$ ,  $\prod_{\gamma \in \mathcal{I}} (I - A_{\gamma}) f = 0$  if and only if  $\prod_{\gamma \in \mathcal{I}} (I - \tilde{A}_{\gamma}) f = 0$ , where  $\mathcal{I}$  is any index set.

♣  $f = f_{\emptyset} + f_1 + f_2$  is intrinsic,  $f = f_{\emptyset} + f_1 + f_{1,2}$  is not.

### ANOVA Decomposition

♣ On  $\mathcal{X} = [0, 1]^2$ , with  $A_1 f = f(0, x_{\langle 2 \rangle})$  and  $A_2 f = f(x_{\langle 1 \rangle}, 0)$ , one has

$$f_{\emptyset} = A_1 A_2 f = f(0, 0),$$

$$f_1 = (I - A_1) A_2 f = f(x_{\langle 1 \rangle}, 0) - f(0, 0),$$

$$f_2 = A_1 (I - A_2) f = f(0, x_{\langle 2 \rangle}) - f(0, 0),$$

$$f_{1,2} = (I - A_1)(I - A_2) f$$

$$= f(x_{\langle 1 \rangle}, x_{\langle 2 \rangle}) - f(x_{\langle 1 \rangle}, 0) - f(0, x_{\langle 2 \rangle}) + f(0, 0).$$

Slide 8

With  $A_{\gamma} f = \int_0^1 f dx_{\langle \gamma \rangle}$ ,  $\gamma = 1, 2$ , one has

$$f_{\emptyset} = A_1 A_2 f = \int_0^1 \int_0^1 f dx_{\langle 1 \rangle} dx_{\langle 2 \rangle},$$

$$f_1 = (I - A_1) A_2 f = \int_0^1 (f - \int_0^1 f dx_{\langle 1 \rangle}) dx_{\langle 2 \rangle},$$

$$f_2 = A_1 (I - A_2) f = \int_0^1 (f - \int_0^1 f dx_{\langle 2 \rangle}) dx_{\langle 1 \rangle},$$

$$f_{1,2} = (I - A_1)(I - A_2) f$$

$$= f - \int_0^1 f dx_{\langle 2 \rangle} - \int_0^1 f dx_{\langle 1 \rangle} + \int_0^1 \int_0^1 f dx_{\langle 1 \rangle} dx_{\langle 2 \rangle}.$$

### Curse of Dimensionality

Consider uniformly distributed data on  $\mathcal{X} = [0, 1]^k$ . A cube with sides of lengths 0.5 has volume  $0.5^k$ , containing 50%, 25%, 12.5% of the data for  $k = 1, 2, 3$ . The *sparsity* of high dimensional space makes “local estimation,” the essence of nonparametric methods, infeasible.

Alternatively, consider the flexibility afforded by a 5-piece piecewise polynomial in 1-D. To achieve the same flexibility in 3-D, one has 125 pieces by taking products.

To control model complexity, additive models with the possible addition of second-order interactions are popular, as with the classical discrete ANOVA models.

Slide 9

### Conditional Independence and Graphical Models

For joint density  $f(x, y)$  of r.v.'s  $(X, Y)$  on  $\mathcal{X} \times \mathcal{Y}$ , write

$$f(x, y) = \frac{e^{\eta(x,y)}}{\int_{\mathcal{X}} dx \int_{\mathcal{Y}} e^{\eta(x,y)} dy} = \frac{e^{\eta_x + \eta_y + \eta_{x,y}}}{\int_{\mathcal{X}} dx \int_{\mathcal{Y}} e^{\eta_x + \eta_y + \eta_{x,y}} dy},$$

where  $\eta_x$ ,  $\eta_y$ , and  $\eta_{x,y}$  are ANOVA terms;  $\eta_0$  is eliminated for a one-to-one

transform.  $X \perp Y \iff \eta_{x,y} = 0 \iff (I - A_x)(I - A_y)\eta = 0.$

The conditional density only depends on  $\eta_y + \eta_{x,y}$ ,

$$f(y|x) = \frac{e^{\eta(x,y)}}{\int_{\mathcal{Y}} e^{\eta(x,y)} dy} = \frac{e^{\eta_y + \eta_{x,y}}}{\int_{\mathcal{Y}} e^{\eta_y + \eta_{x,y}} dy}.$$

Substituting  $(y, z)$  for  $y$  above, one has

$$f(y, z|x) = \frac{e^{\eta_y + \eta_z + \eta_{y,z} + \eta_{x,y} + \eta_{x,z} + \eta_{x,y,z}}}{\int_{\mathcal{Y}} dy \int_{\mathcal{Z}} e^{\eta_y + \eta_z + \eta_{y,z} + \eta_{x,y} + \eta_{x,z} + \eta_{x,y,z}} dz}.$$

$(Y \perp Z)|X \iff \eta_{y,z} + \eta_{x,y,z} = 0 \iff (I - A_y)(I - A_z)\eta = 0.$

Slide 10

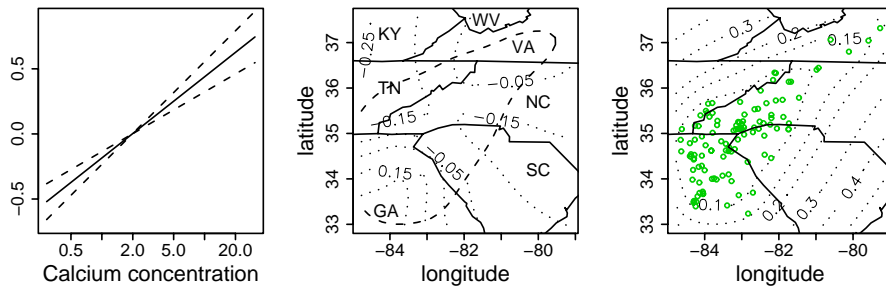
### Water Acidity in Lakes

Water acidity measurements were recorded for 112 lakes in the Blue Ridge along with calcium concentrations in a 1984 EPA survey. A model of the following form was fitted.

$$\text{pH} = \eta_{\emptyset} + \eta_c(\text{cal}) + \eta_g(\text{geo}) + \eta_{c,g}(\text{cal}, \text{geo}) + \epsilon.$$

Slide 11

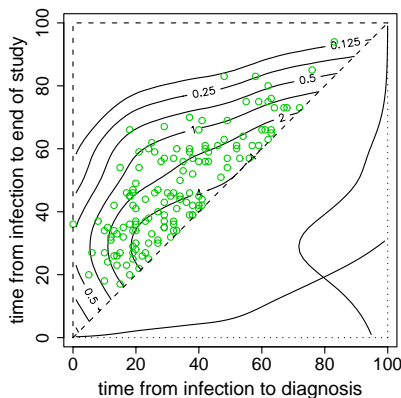
The geography was treated in an isotropically invariant manner.



### AIDS Incubation

Data concerning HIV infection by blood transfusion were collected by CDC, which included the time  $X$  from transfusion to AIDS diagnosis and the time  $Y$  from transfusion to the end of collection; the data are naturally truncated with  $X \leq Y$ .

Slide 12



Assuming pre-truncation independence of  $X$  and  $Y$ , the density is given by

$$f(x, y) = \frac{e^{\eta_x(x) + \eta_y(y)}}{\int_0^{100} dy \int_0^y e^{\eta_x(x) + \eta_y(y)} dx}.$$

The estimate on the left was based on 141 “elderly patients” of age 60 or above.