Lecture 5: Smooth Regression Models

Will Welch (adapted from materials by Matias Salibian-Barrera)

Department of Statistics, University of British Columbia

STAT 447B: Methods for Statistical Learning

September–December 2014

Today's Learning Goals

1. Review
2. Penalized Splines
3. Natural Cubic Splines
4. Kernel Smoothers
5. Generalized Additive Models

Statistical Models: Flexibility Versus Simplicity

- Bigger, more flexible models
  - Good: can adapt to complexities in the mean (systematic) relationship between y and x
  - Good: hence reduce bias of parameter estimators and predictions
  - Bad: can overfit, increasing variance of parameter estimators and predictions
- Smaller, more restrictive models
  - Good: easier to interpret
  - Good: tend to have smaller variances of parameter estimators and predictions
  - Bad: do not adapt fully to complexities, increasing bias of parameter estimators and predictions
- "Optimal" model complexity: manage bias versus variance trade-off

Approach So Far

- Increase complexity of the basis function type
  - Linear $x \Rightarrow$ Polynomials $\Rightarrow$ linear splines $\Rightarrow$ quadratic splines $\Rightarrow$ cubic splines ...
- And/or increase the number of basis functions
  - More knots
- Manage the bias-variance trade-off by optimizing cross-validation prediction accuracy

Alternatively

- Start with a very flexible model
  - say cubic splines, large number of knots
- Rein in its complexity
  - Penalty for (excessive) complexity
- Manage the penalty size and hence complexity by cross-validation

Penalized Regression Splines

- Take a flexible model like cubic splines with many knots (K large)
- Cubic splines are fit by solving

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{K+4}} \sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}^T \mathbf{x}_i \right)^2$$

where

$$\mathbf{x}_i = \left(1, x_i, x_i^2, x_i^3, (x_i - \kappa_1)_+^3, (x_i - \kappa_2)_+^3, \cdots, (x_i - \kappa_K)_+^3 \right)^T$$

- Note that the parameters that may overfit are β_{j+4}, $j = 1, \ldots, K$

Penalized Regression Splines

- One can try to solve:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{K+4}} \sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}^T \mathbf{x}_i \right)^2$$

subject to

$$\sum_{j=1}^{K} \beta_{j+4}^2 \ \leq \ C$$

for some constant $C > 0$.

- This will typically give a less wiggly fit
- Less overfitting

Penalized Regression Splines

- Equivalent to

$$\min_{\beta \in \mathbb{R}^{K+4}} \left(\sum_{i=1}^{n} \left(y_i - \beta^T \mathbf{x}_i \right)^2 + \lambda \, \beta^T \, \mathbf{D} \, \beta \right)$$

for some constant $\lambda > 0$.

- The matrix $\mathbf{D} = \mathrm{diag}\left(\mathbf{0}_4, \mathbf{I}_K\right)$
- The solution is $\hat{\beta}_\lambda = \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{D}\right)^{-1} \mathbf{X}^T \mathbf{Y}$
- Why?

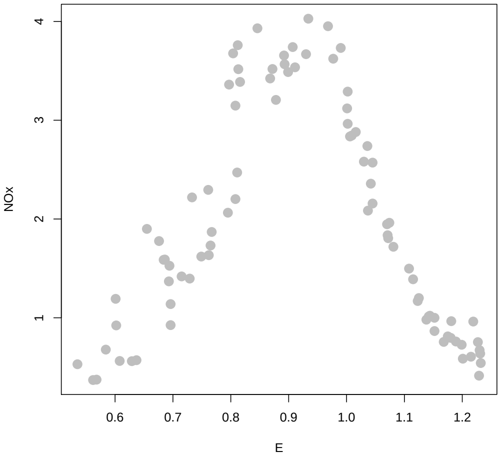
Mini Activity

- We want to minimize over $\boldsymbol{\beta}$

$$\sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}^T \mathbf{x}_i \right)^2 \; + \; \lambda \, \boldsymbol{\beta}^T \, \mathbf{D} \, \boldsymbol{\beta} = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \, \boldsymbol{\beta}^T \, \mathbf{D} \, \boldsymbol{\beta}$$

- Show that the solution is $\hat{\beta}_\lambda = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{D})^{-1} \, \mathbf{X}^T \mathbf{Y}$
- Hand in a group solution with your names and student IDs. You have 10 minutes.

Ethanol Data



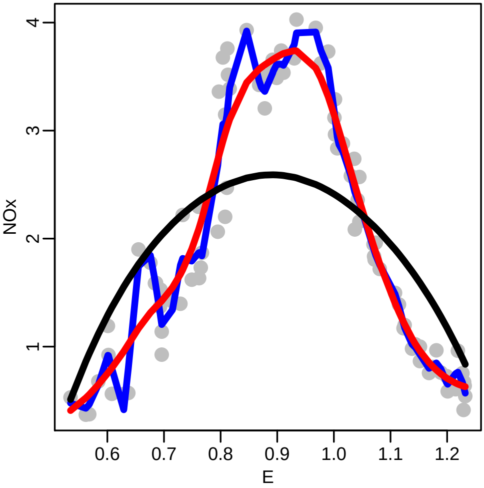
Ethanol Data

```
> data(ethanol, package='SemiPar')
> dim(ethanol)
[1] 88  3
> head(ethanol)
    NOx  C     E
1 3.741 12 0.907
2 2.295 12 0.761
3 1.498 12 1.108
4 2.881 12 1.016
5 0.760 12 1.189
6 3.120  9 1.001
```

From SemiPar documentation

- NOx - concentration of oxides of nitrogen in exhaust
- C - compression ratio
- E - richness of air/ethanol mix

Ethanol Data: Penalized Cubic Splines, 50 knots



Ethanol Data: Sizes of Penalties

- Fit using `spm` in library(SemiPar)
- Penalty set by `spar`, related to λ
- Why are the lines different? **Clicker question 1.**

Linear Smoothers

- Penalized regression splines are "linear smoothers"
- Predicted values are

$$\hat{\mathbf{Y}} = \mathbf{X}\,\hat{\beta}_\lambda = \mathbf{X}\,(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{D})^{-1}\,\mathbf{X}^T\mathbf{Y}$$

$$= \mathbf{S}_\lambda\,\mathbf{Y}$$

for some "fixed" matrix $\mathbf{S}_\lambda$ that does not depend on **Y**.

- Just like least squares!

Natural Cubic Splines

- Consider the following problem

$$\min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int \left(f^{(2)}(t) \right)^2 dt$$

- The solution is a *natural* cubic spline with n knots at $x_1, x_2, \ldots, x_n$.
- *Natural* cubic splines are cubic splines with the restriction that they are linear beyond the boundary knots.

Selecting the Size of the Penalty

- Cross-validation: consider *n*-fold CV

$$\text{CV RMSE}(\lambda) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \beta_\lambda^{(-i)} \right)^2 ,$$

where $\beta_\lambda^{(-i)}$ is the fit without using the point (y_i, x_i), and choose a value λ_0 such that

$$\text{CV RMSE}(\lambda_0) \ \leq \ \text{CV RMSE}(\lambda) \quad \forall \, \lambda \geq 0$$

Selecting the Size of the Penalty

- Computing CV RMSE(λ)...

$$\text{CV RMSE}(\lambda) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \beta_\lambda^{(-i)} \right)^2$$

We might need to re-fit the model n times

- For some smoothers and models this is not necessary. For many linear smoothers $\hat{\mathbf{Y}} = \mathbf{S}_\lambda \mathbf{Y}$ we have

$$\text{CV RMSE}(\lambda) = \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mathbf{Y}}_i}{1 - \mathbf{S}_{\lambda,i,i}} \right)^2$$

Selecting the Size of the Penalty

- Computing $\mathbf{S}_{\lambda,i,i}$, $i = 1, \ldots, n$ can be demanding
- Sometimes one uses generalized CV

$$\text{GCV RMSE}(\lambda) = \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mathbf{Y}}_i}{1 - \operatorname{tr}(\mathbf{S}_\lambda)/\, n} \right)^2 = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\mathbf{Y}}_i \right)^2}{(1 - \operatorname{tr}(\mathbf{S}_\lambda)/\, n)^2}$$

Kernel Smoothers

- We are interested in estimating

$$f(x) \;=\; E\,(\,Y|\,X = x)$$

- Given a sample (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n)

$$\hat{f}(x) \;=\; \text{average}\Big\{ y_i \;:\; x_i = x \Big\}$$

$$\hat{f}(x) \;=\; \text{average}\Big\{ y_i \;:\; x_i \text{ is close to } x \Big\}$$

Kernel Smoothers

- More formally

$$\hat{f}(x) = \text{average}\left\{ y_i \; : \; |x_i - x| \leq h \right\}$$

$$\hat{f}(x) = \frac{1}{n_x} \sum_{i \,:\, |x_i - x| \leq h} y_i$$

$$\hat{f}(x) = \frac{\sum_{i \,:\, |x_i - x| \leq h} y_i}{\sum_{i \,:\, |x_i - x| \leq h} 1}$$

Kernel Smoothers

More formally

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K(x_i, x, h)\, y_i}{\sum_{i=1}^{n} K(x_i, x, h)}$$

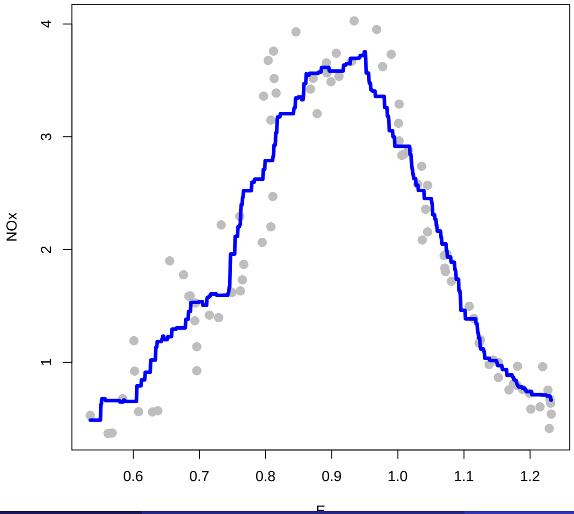
where

$$K(x_i, x, h) = W\left(\frac{x_i - x}{h}\right)$$

and

$$W(t) = \begin{cases} 1 & \text{if } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Ethanol Data: Kernel Smoother, $h = 0.07$



Ethanol Data: Why is the Kernel Smoother Not Smooth?

- The kernel smoother is discontinuous here.
- What part of this formulation leads to a non-smooth “smoother”? **Clicker question 2.**

Kernel Smoothers

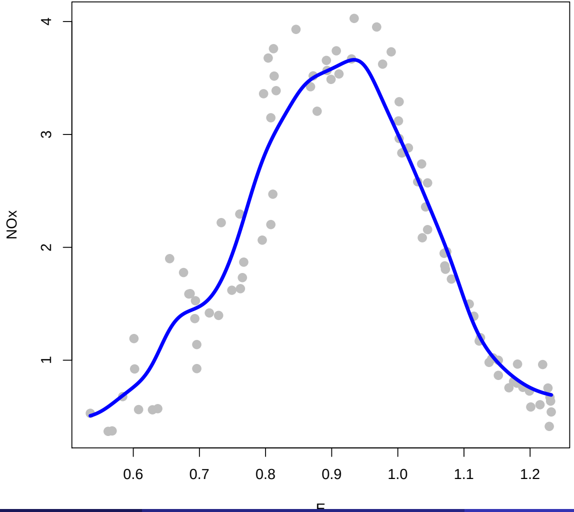
- Discontinuities come from $W(t)$
- Use a smooth kernel

$$K(x_i, x, h) \;=\; W\left(\frac{x_i - x}{h}\right)$$

with

$$W(t) \;=\; \begin{cases} 1 - t^2 & \text{if } |t| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Ethanol Data: Kernel Smoother, $h = 0.03$



Kernel Smoothers

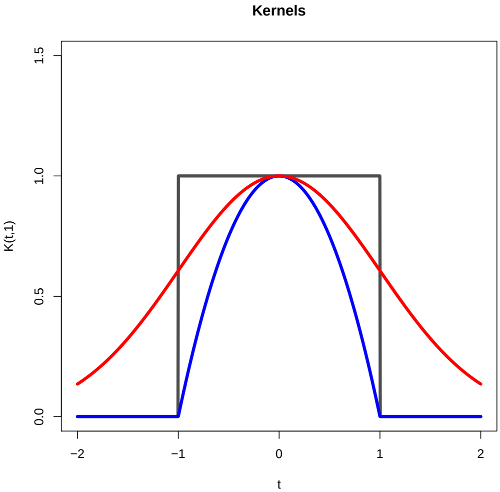
- Other kernels...

$$K(x_i, x, h) \; = \; W\left(\frac{x_i - x}{h}\right)$$

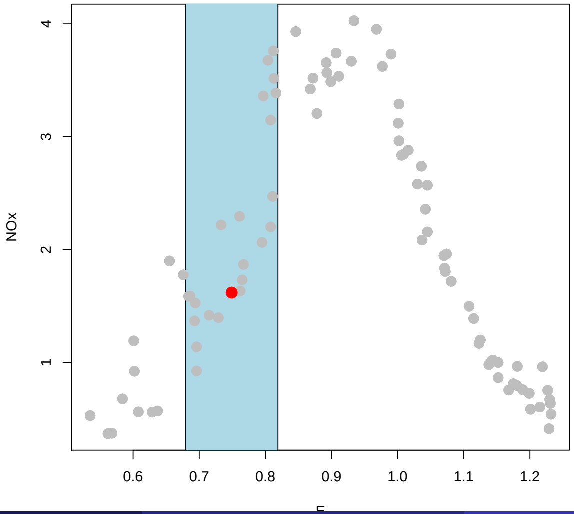
with

$$W(t) \; = \; \phi(t) \; \propto \; \exp\left(-t^2/2\right)$$

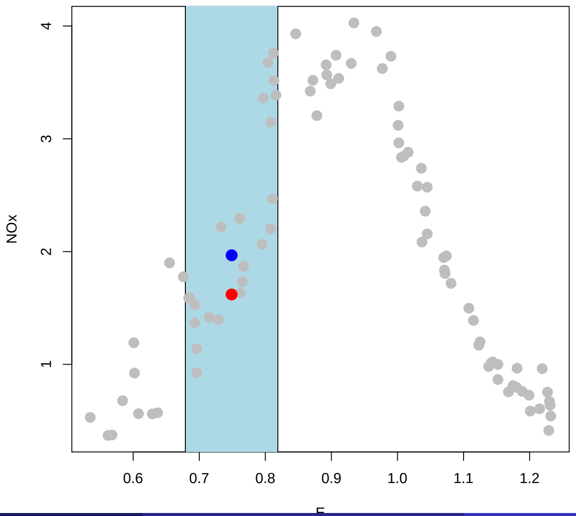
Kernel Smoothers: Kernel (Weight) Functions



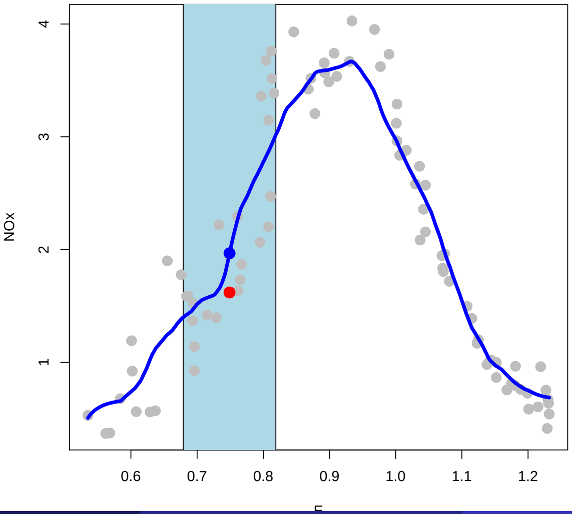
Kernel Smoother in Action



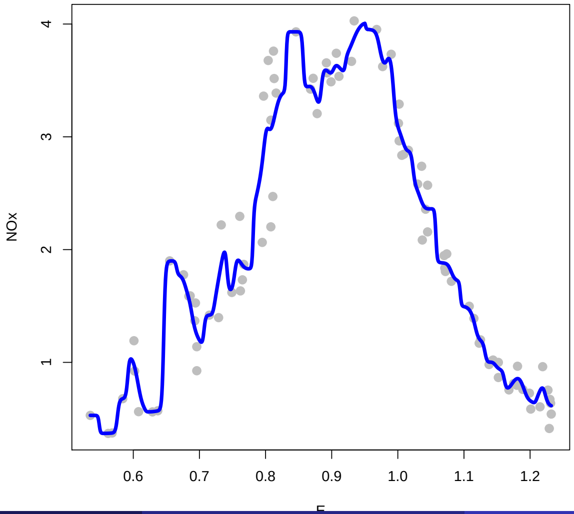
Kernel Smoother in Action



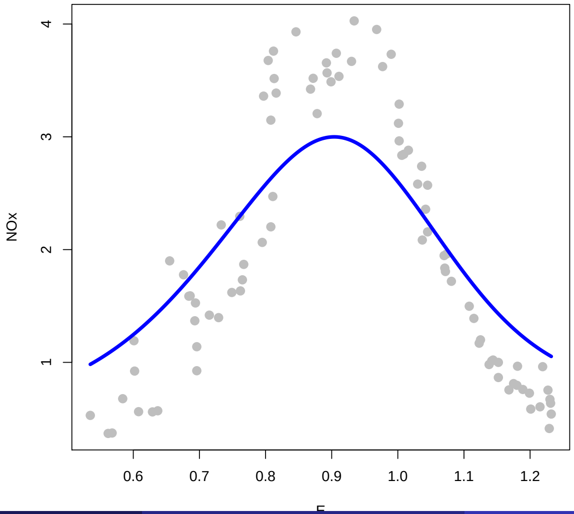
Kernel Smoother in Action



Small Bandwidth h



Larger Bandwidth h



Generalized Additive Models (GAMs)

- An automatic way of generating a flexible model with automatic smoothing

$$E(Y) = \beta_0 + f_1(x_1) + f_2(x_2) + \dots$$

  where f_1 and f_2 are smooth functions

- These functions can be estimated in `R` with `gam`

```
library(mgcv)
ethanol.gam <- gam(NOx ~ s(E), data = ethanol)
```

- `s(E)` terms: “Smooth terms are represented using penalized regression splines . . .” (see `help(gam)`)

Ethanol Data: GAM