Given: TS $(x_1, y_1) \ldots (x_n, y_n)$ iid obs of (X, Y)

Goal: Generate prediction rule $f(x)$ that predicts Y for query predictor x.

Optimal rule $f(x) = E(Y \mid x)$

Assume single predictor

Approach 1: Kernel smoothing

Approach 2: Expansion based rules

- Choose dictionary of "basis functions" $B_1(x) \ldots B_K(x)$ for which (hopefully)

$$f(x) \approx \sum_{j=1}^{K} a_j B_j(x) \Leftarrow$$

How to find $a_1 \ldots a_K$?

Obvious approach: Find $\hat{\underline{a}}$ to minimize resubstitution error:

$$\hat{\underline{a}} = \underset{\underline{a}}{\operatorname{argmin}} \| y - X \underline{a} \|^2$$

where $X_{ij} = B_j(x_i)$.

Can we do better?

Look at extreme case: $K = n$, $B_1 \ldots B_K$ linearly independent over the data (rank $(X) = n$)

$$\hat{y} = Hy = X(X^T X)^{-1} X^T y$$

$$\hat{a} = (X^T X)^{-1} X^T y$$
$$= X^{-1} X^{-T} X^T y$$
$$= X^{-1} y$$

$$= \underbrace{X X^{-1} X^{-T} X^T}_{I} y$$

$$= y$$

What's the EEE $= E\left(\frac{1}{n} \| f - Hy \|^2\right)$

$y_i = f(x_i) + \varepsilon_i \quad E(\varepsilon_i) = 0 \quad V(\varepsilon_i) = \sigma^2 \quad \varepsilon_i$ inde

$$EEE_{est} = \frac{1}{n}\Big(\underbrace{\|(I-H)f\|^2}_{\text{squared bias}} + \sigma^2 \, \text{trace } H\Big)$$

in HW01 trace$(W^T W)$

$\text{squared bias} = 0$ $\quad$ $\text{trace } H = n$

$$= EEE_{est} = 0 + \sigma^2 = \sigma^2$$

If we had K basis functions, then

$$EEE_{est} = \frac{1}{n}\left(\|(I-H)f\|^2 + \sigma^2 K\right)$$

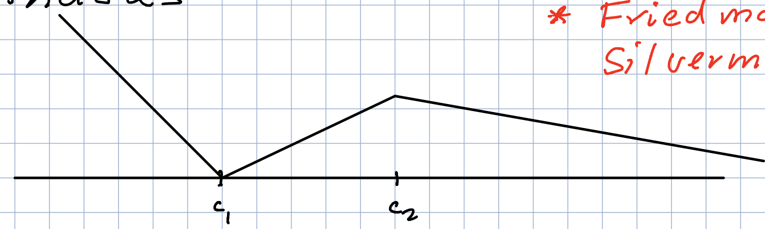
Suppose $f(x) = \sum_{i=1}^{n} a_i B_i(x)$ but some of the a_i are 0. Then it would be better to remove those basis functions from the dictionary, because this would reduce the variance component of EEE and would not increase the bias component.

Turbo$^{(x)}$ Procedure to fit linear splines

Linear spline with knots $c_1 < c_2 \cdots < c_k$ is

- Piecewise linear in $(-\infty, c_1], (c_1, c_2] \ldots (c_k, \infty)$
- Continuous

* Friedman & Silverman 1989



Any linear spline with knots $c_1 \ldots c_k$ is a linear combination of basis functions

$$B_i(x) = (x - c_i)_+ \quad i = 1 \ldots k$$

$$B_{k+1}(x) = 1$$

$$B_{k+2}(x) = x$$

Given training sample $(x_1\ y_1) \ldots (x_n\ y_n)$
Turbo finds

- # of knots
- knot positions
- slopes and intercepts of linear pieces

to mimize an estimate of the expected squared prediction error.

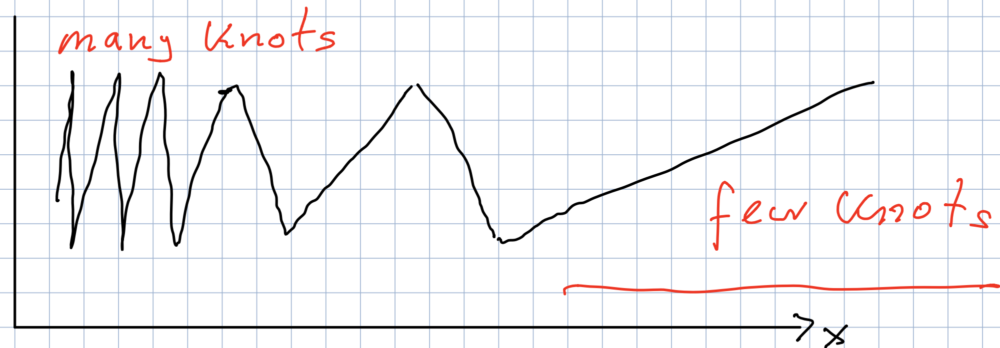
Hard problem.

Suppose $k = 1$

$$\underset{\underline{a},\ c_1}{\text{minize}} \sum_{i=1}^{n} \left(y_i - a_1 B_1(x_i) - a_2 - a_3 x_i\right)^2$$

$$B_1(x) = (x - c_1)_+$$

Finding c_1 is a nonlinear optimization problem.
No explicit solution.



Design choices made by FRS:

For given knot positions choose $\underline{a}$ by least squares

Assume $x_1 < x_2 \quad < x_n$

Dictionary $B_i = (x - x_i)_+ \quad i = 1..n$

$B_{n+1} = 1$

Note: B_1 is linear over the range of data
$\Rightarrow$ Don't need an extra linear basis function.

Fit model by stepwise forward selection

- Find best model with one basis function using exhaustive search.

$$j_1 = \text{argmin} \ \text{argmin} \sum (y_i - a B_i(x_i))^2$$

- Find the best model with two basis functions, given we already have chosen the first one,

$$\hat{j}_2 = \underset{j}{\operatorname{argmin}}\ \underset{\underline{a}}{\operatorname{argmin}} \sum \left(y_i - a_1 B_{j_1}(x_i) - a_2 B_j(x_i)\right)^2$$

⋮ etc