Turbo for multiple regression

For simplicity consider case of 2 predictors.

Given: $(\underline{x}_1, y_1) \ldots (\underline{x}_n, y_n)$ obs of $(\underline{X}, Y)$

$\underline{X} = (X_1, X_2)$ $\quad \underline{x}_i = (x_{i1}, x_{i2})$

Goal: Estimate $E(Y \mid \underline{x}) = f(\underline{x})$

Simplifying assumption $f(\underline{x})$ is additive

$$f(\underline{x}) = f_1(x_1) + f_2(x_2)$$

Note:

Linear functions are additive

$$\ell(\underline{x}) = a_0 + \underbrace{a_1 x_1}_{\ell_1(x_1)} + \underbrace{a_2 x_2}_{\ell_2(x_2)}$$

Additivity is strong assumption: "Most" functions are not additive

For example $f(\underline{x}) = x_1 * x_2$ is not additive:

There are no univariate function $g_1, g_2 \; \mathbb{R} \to \mathbb{R}$ such that $f(\underline{x}) = x_1 * x_2 = g_1(x_1) + g_2(x_2)$

Generalizing Turbo

Define basis functions

$B_i^1(\underline{x}) = (x_1 - x_{i1})_+ \quad i = 1 \ldots n$

$B_i^2(\underline{x}) = (x_2 - x_{i2})_+ \quad i = 1 \ldots n.$

$B_0(\underline{x}) = 1$

Note: $B_1^1 \ldots B_n^1$ depend on $\underline{x}$ only throug x_1

$B_1^2 \ldots B_n^2$ $\underline{x}$ x_2

Any linear combination of these basis functions is additive: It is the sum of functions that depend only on x_1 and of functions that depend only on x_2

To fit additive model we run Turbo on dictionary $\{ B_1^1(\underline{x}) \ldots B_n^1(\underline{x}), B_1^2(\underline{x}) \ldots B_n^2(\underline{x}), B_0 \}$

There are more basis functions than observations but we are doing variable selection, so operationally that does not matter.

Note

- Easy to incorporate categorical predictors via dummy variables
- Lots of room for "engineering": smaller dictionaries, higher order splines ....

Shrinkage: An alternative to selection

Example: order 2 smoothing splines

TS: $(x_1, y_1) \ldots (x_n, y_n)$ $x_1 < x_2 \ldots < x_n$

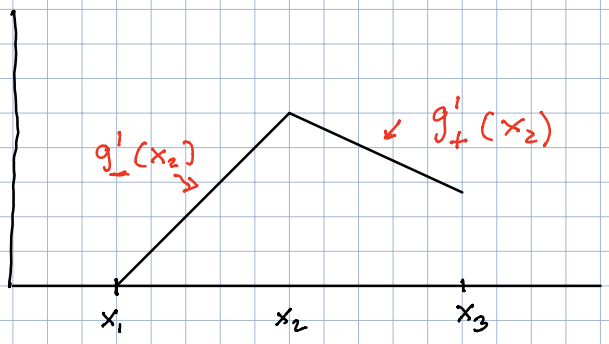
$B_j(x) = (x - x_j)_+ \quad j = 1 \dots n-1$

$B_n(x) = 1$

$$g(x) = \sum_{j=1}^{n} a_j B_j(x) \quad \text{order 2 spline}$$

Measure roughness of $g(x)$ by

$$\varphi(\underline{a}) = \sum_{i=2}^{n-1} \left(g'_-(x_i) - g'_+(x_i) \right)^2 \quad (*)$$



$$* = \sum_{i=2}^{n-1} \left(\sum_{r=1}^{i-1} a_r - \sum_{r=1}^{i} a_r \right)^2$$

$$= \sum_{i=2}^{n-1} a_i^2$$

Smoothing spline

$$\hat{\underline{a}}_\lambda = \underset{\underline{a}}{\operatorname{argmin}} \left[\| y - X\underline{a} \|^2 + \lambda \sum_{i=2}^{n+1} a_i^2 \right]$$

↑ residual sum of squares

↑ penalty term

What will the spline look like

- For $\lambda = \infty \Rightarrow a_2 \dots a_{n-1} = 0$ no penalty for

basis functions B_1 and B_n

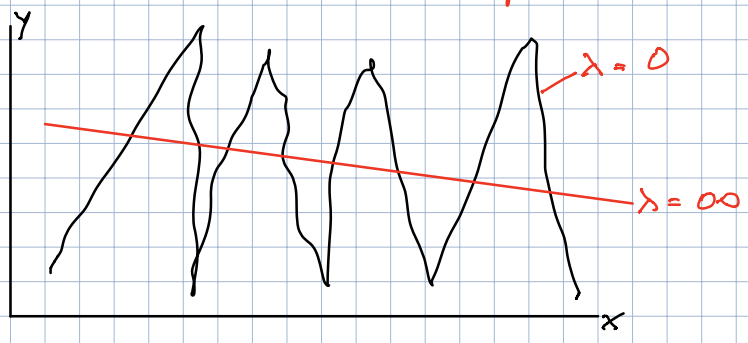
$B_1 = (x - x_1)_+$ linear over the range of data

$B_n = 1$

The smoothing spline for $\lambda = \infty$ will be the least squares line

- For $\lambda = 0$ RSS will be 0

The smoothing spline we be the piecewise linear function that interpolates the data.



Contrast Turbo and order 2 smoothing splines

Turbo forces some of the coefficients $a_1 \ldots a_n$ to be 0 and chooses the values of the others by least squares

Spline smoothing shrinks the coefficients towards 0

Kernel smoother with small bandwidth

Variable selection " with large λ

Shrinkage (spline smoothing) with small λ

More about shrinkage

Given TS $(\underline{x}_1, y_1) \dots (\underline{x}_n, y_n)$

Assume $y_i = \langle \underline{a}, \underline{x}_i \rangle + \varepsilon_i$

ε_i uncorrelated

$E(\varepsilon_i) = 0$

$V(\varepsilon_i) = \sigma^2$

Then the LS estimate

$\hat{\underline{a}} = \arg\min_{\underline{a}} \| \underline{y} - X\underline{a} \|^2$ has minimum variance among all linear unbiased estimates (Gauss-Markov theorem)

Contrast to ridge regression estimator

$$\hat{\underline{a}}_R(\lambda) = \arg\min_{\underline{a}} \| \underline{y} - X\underline{a} \|^2 + \lambda \| \underline{a} \|^2$$

and to the Lasso estimator

$$\hat{\underline{a}}_L(\lambda) = \arg\min_{\underline{a}} \| \underline{y} - X\underline{a} \|^2 + \lambda \sum |a_i|$$

RR and Lasso estimators will be biased if $\lambda > 0$ but for appropriate choice of λ

will have smaller expected squared error.