MULTIVARIATE NONPARAMETRIC REGRESSION

Let $(X_1,Y_1),\dots,(X_n,Y_n)$ be indep. realizations of $(X,Y)\in[0,1]^p\times\mathbb{R}$, s.t.

$$Y = m(X) + \varepsilon \quad , \qquad m:[0,1]^p\to\mathbb{R},$$

where ε is independent of X and $\mathbb{E}\varepsilon=0$, $\mathbb{E}\varepsilon^2=\sigma^2<\infty$. We wish to estimate the unknown function m.

A Nadaraya-Watson type estimator of m is given by

$$\hat{m}_n^{NW}(x) = \frac{\sum_{i=1}^n Y_i\, K\left(h^{-1}(X_i-x)\right)}{\sum_{j=1}^n K\left(h^{-1}(X_j-x)\right)} \qquad \text{for all } x\in[0,1]^p$$

for some kernel function $K:\mathbb{R}^p\to\mathbb{R}$ and bandwidth $h>0$.

Often K is chosen such that $K(u)=\prod_{j=1}^p G(u_j)$ where G is some univariate kernel like

$$G(z)=\phi(z) \qquad \text{or} \qquad G(z)=\tfrac{3}{4}(1-u^2)\,\mathbb{1}(|u|\le 1).$$

Let's investigate the variance of this estimator under the assumptions

(K1) $K(u)\le K_{max}<\infty$ for all $u\in\mathbb{R}^p$

(D1) Let $X_1,\dots,X_n\in[0,1]^p$ be deterministic such that for some $n_0>0$

$$0<c_1\le\frac{1}{nh^p}\sum_{i=1}^n K\left(h^{-1}(X_i-x)\right)\le\frac{1}{c_1}$$

$$0<c_2\le\frac{1}{nh^p}\sum_{i=1}^p K^2\left(h^{-1}(X_i-x)\right)\le\frac{1}{c_2}$$

for some $c_1, c_2>0$ for all $n\ge n_0$.

Assumption (D1) is believable— think about the laws of large numbers.

Result: Under (K1) and (D1), we have

$$\operatorname{Var} \hat{m}_n^{NW}(x_0) \in \left(\frac{\sigma^2}{nh^p} \cdot \frac{c_1}{c_2} , \frac{\sigma^2}{nh^p} \cdot \frac{c_2}{c_1} \right) \quad \text{for all } x_0 \in [0,1]$$

for all $n \geq n_0$.

Remark: We have encountered again the curse of dimensionality! The variance explodes with the dimension.

Exercise: Prove the above result.

Solution: Write

$$\hat{m}_n^{NW}(x_0) = \sum_{i=1}^n W_{ni}(x_0) Y_i , \quad \text{with} \quad W_{ni}(x_0) = \frac{K\left(h^{-1}(X_i - x_0)\right)}{\sum_{j=1}^n K\left(h^{-1}(X_j - x_0)\right)} .$$

Then we have, for $n \geq n_0$,

$$\operatorname{Var} \hat{m}_n^{NW}(x_0) = \sum_{i=1}^n W_{ni}^2(x_0) \sigma^2$$

$$= \sigma^2 \sum_{i=1}^n \frac{K^2\left(h^{-1}(X_i - x_0)\right)}{\left[\sum_{j=1}^n K\left(h^{-1}(X_j - x_0)\right)\right]^2}$$

$$= \frac{\sigma^2}{nh^p} \frac{\frac{1}{nh^p} \sum_{i=1}^n K^2\left(h^{-1}(X_i - x_0)\right)}{\left[\frac{1}{nh^p} \sum_{j=1}^n K\left(h^{-1}(X_j - x_0)\right)\right]^2}$$

$$\in \left(\frac{\sigma^2}{nh^p} \cdot \frac{c_2}{c_1} , \frac{\sigma^2}{nh^p} \cdot \frac{c_1}{c_2} \right).$$

A multivariate local linear estimator is given by $\hat{m}_n^{LL}(x) = \hat{\alpha}_x$, where

$$\left(\hat{\alpha}, \underset{\sim}{\hat{\beta}}\right)_x = \underset{\alpha\in\mathbb{R},\beta\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^n \left(Y_i - \alpha - \beta^T(X_i - x)\right)^2 K\left(h^{-1}(X_i - x)\right).$$

This estimator is also subject to the curse of dimensionality.

One way to mitigate the curse of dimensionality is to make simplifying assumptions about m.

An assumption known as "Additivity" is very often used...

THE ADDITIVE MODEL

Assume that $m : [0,1]^p \to \mathbb{R}$ is of the form

$$m(x) = m_1(x_1) + \dots + m_p(x_p)$$

for some functions $m_j : [0,1] \to \mathbb{R}$, $j = 1, \dots, p$, which we call "additive components."

Stone (1985) argued that a good many functions $m : [0,1]^p \to \mathbb{R}$ that are likely to arise in multivariate regression could be well-approximated by an additive function.

The additive model for independent realizations $(X_1, Y_1), \dots, (X_n, Y_n)$ of $(X, Y) \in [0,1]^p \times \mathbb{R}$ is

$$Y = \mu + m_1(X_1) + \dots + m_p(X_p) + \varepsilon_i.$$

We must immediately consider the question of identifiability: Are the model components uniquely defined by the model, or can I change them and still get the same model?

Example: Consider two models:

model 1: $$Y = 10 + \underbrace{(1 + \sin(X_1))}_{m_1(X_1)} + \underbrace{(X_2^2 - 1)}_{m_2(X_2)} + \varepsilon$$

model 2: $$Y = 10 + \underbrace{\sin(X_1)}_{m_1(X_1)} + \underbrace{X_2^2}_{m_2(X_2)} + \varepsilon$$

Note that under both models,

$$\mathbb{E}[Y \mid X] = 10 + \sin(X_1) + X_2^2 .$$

Unless we make an identifiability assumption, we will not be able to say what the "true" functions are.

We will assume, for the sake of identifiability, that

$$\mathbb{E}\, m_j(X_j) = 0 \qquad \text{for} \quad j = 1, \dots, p.$$

Under this identifiability condition, we estimate μ with $\bar{Y}_n$, and then proceed to estimate $m_1, \dots, m_p$ using centered response values.

From now, we thus assume, WLOG, that $\mu = 0$ and that $Y_1, \dots, Y_n$ are centered.

A penalized spline estimator for the additive model:

We impose the identifiability assumption so that it holds empirically. That is, we require

$$\frac{1}{n} \sum_{i=1}^{n} \hat{m}_j(X_{ij}) = 0 \qquad \text{for} \quad j = 1, \dots, p.$$

To achive this, we use empirically centered basis functions.

Let $b_{j1},\dots,b_{jd}$ be the cubic B-spline basis functions on the set of knots

$$0,0,0,0,\frac{1}{K},\dots,(K-1)/K,1,1,1,1 \qquad \left(\begin{array}{l}\text{Could also choose the corresponding} \\ \text{empirical quantiles of} \\ X_{1j},\dots,X_{nj}.\end{array}\right)$$

and then define the empirically centered basis functions as

$$\bar{b}_{j\ell}(x) = b_{j\ell}(x) - \frac{1}{n}\sum_{i=1}^{n} b_{j\ell}(X_{ij}), \qquad \ell = 1,\dots,d.$$

Now, think for a moment about making a design matrix from these.

Exercise: Give the rank of the matrix $\left(\bar{b}_{j\ell}(X_{ij})\right)_{1\le i\le n,\,1\le \ell\le d}$ supposing

$$B_{nj} = \left(b_{j\ell}(X_{ij})\right)_{1\le i\le n,\,1\le \ell\le d}$$

has full-column rank and has rows that sum to 1.

Solution: We have

$$\left(\bar{b}_{j\ell}(X_{ij})\right)_{1\le i\le n,\,1\le \ell\le d} = \left(I_n - \frac{1}{n}1_n 1_n^T\right) B_{nj}.$$

Since the rows of B_{nj} sum to 1, we have

$$B_{nj}\, 1_d = 1_n.$$

Now,

$$\left(I_n - \frac{1}{n}1_n 1_n^T\right) B_{nj}\, 1_d = \left(I_n - \frac{1}{n}1_n 1_n^T\right) 1_n = 0,$$

so that $1_d \in \mathcal{N}\left(\left(I_n - \frac{1}{n}1_n 1_n^T\right)B_{nj}\right)$. This means

$$\operatorname{rank}\left(\left(\bar{b}_{j\ell}(X_{ij})\right)_{1\le i\le n,\,1\le \ell\le d}\right) < d.$$

[Recall that if a matrix A has full-column rank, $Ax=0 \iff x=0$.]

In order to have full-column-rank design matrices, a convention is to toss out the first basis function (bs function of splines package). Then the centered design matrix will have full rank.

Suppose that this has been done such that $K+1$ knots have been used, resulting in $K+1+3$ cubic B-spline basis functions, of which the first has been removed, resulting in d centered basis functions $\bar b_{j1},\dots,\bar b_{jd}$ for each $j=1,\dots,p$.

Then, for each $j=1,\dots,p$, set

$$\overline{\mathcal{M}}_{nj,3} = \text{span}\{\bar b_{j1},\dots,\bar b_{jd}\} \quad \text{and} \quad \bar B_{nj} = \left(\bar b_{j\ell}(X_{ij})\right)_{1\le i\le n,\, 1\le \ell\le d}.$$

Now, penalized spline estimators $\hat m_1,\dots,\hat m_p$ of $m_1,\dots,m_p$ are given by

$$(\hat m_1,\dots,\hat m_p) = \underset{g_j\in\overline{\mathcal{M}}_{nj,3}}{\text{argmin}} \sum_{i=1}^{n}\Big(Y_i + \sum_{j=1}^{p} g_j(X_{ij})\Big)^2 + \lambda\sum_{j=1}^{p}\int_0^1 [g_j''(x)]^2\,dx.$$

<u>Exercise:</u> Give in matrix form $\underset{\sim}{\hat\alpha}_1,\dots\underset{\sim}{\hat\alpha}_p$ such that $\hat m_j(x) = \bar b_{j,x}^T\underset{\sim}{\hat\alpha}_j$, where

$$\bar b_{j,x} = \left(\bar b_{j1}(x),\dots,\bar b_{jd}(x)\right)^T.$$

<u>Solution:</u> We have

$$\left(\underset{\sim}{\hat\alpha}_1,\dots,\underset{\sim}{\hat\alpha}_p\right) = \underset{\underset{\sim}{\alpha}_j\in\mathbb{R}^d}{\text{argmin}} \Big\|\underset{\sim}{Y} - \sum_{j=1}^{p}\bar B_{nj}\underset{\sim}{\alpha}_j\Big\|_2^2 + \lambda\sum_{j=1}^{p}\underset{\sim}{\alpha}_j\,\Omega_j\,\underset{\sim}{\alpha},$$

where $\bar B_{nj} = \left(\bar b_{j\ell}(X_{ij})\right)_{1\le i\le n,\, 1\le\ell\le d}$ and

$$\Omega_j = \left(\int_0^1 b''_{j\ell}(x)\, b''_{j\ell'}(x)\, dx \right)_{1\le \ell, \ell' \le d}.$$

Setting $\underset{\sim}{\hat{\alpha}} = (\underset{\sim}{\hat{\alpha}}_1^T, \dots, \underset{\sim}{\hat{\alpha}}^T)^T$, $\bar{B}_n = (B_{n1}, \dots, B_{np})$, and

$$\Omega = \text{blockdiag}\,(\Omega_1, \dots, \Omega_p),$$

we may write

$$\underset{\sim}{\hat{\alpha}} = \underset{\underset{\sim}{\alpha} \in \mathbb{R}^{pd}}{\text{argmin}} \; \| Y - \bar{B}_n \underset{\sim}{\alpha} \|_2^2 + \lambda\, \underset{\sim}{\alpha}^T \Omega\, \underset{\sim}{\alpha}$$

$$= (\bar{B}_n^T \bar{B}_n + \lambda \Omega)^{-1} \bar{B}_n^T \underset{\sim}{Y}$$

Note that we can only take the inverse $(\bar{B}_n^T \bar{B}_n + \lambda \Omega)^{-1}$ if each $\bar{B}_{n1}, \dots, \bar{B}_{np}$ has full-column rank.

Moreover, if $\lambda = 0$, we will have numerical issues when $p^* d$ is close to n.

For $\lambda > 0$, we can have $p^* d > n$.

Computing the penalized spline estimator in this way causes some head aches because of potential numerical issues (ranks of matrices) and is, moreover, slow, since one has to invert a $pd \times pd$ matrix.

In the next section we introduce an estimation strategy called backfitting, which will be faster and less liable to numerical issues.

BACKFITTING

For the additive model

$$Y = m_1(X_1) + \ldots + m_p(X_p) + \varepsilon$$

the additive components have the interpretation

$$m_j(x_j) = \mathbb{E}\Big[Y - \sum_{k \neq j} m_k(X_k) \;\Big|\; X_j = x_j \Big]$$

$$= \mathbb{E}[Y \mid X_j = x_j] - \sum_{k \neq j} \mathbb{E}\big[m_k(X_k) \mid X_j = x_j\big]$$

for $j = 1, \ldots, p$.

Letting Π_j represent condition expectation given X_j, for $j=1,\ldots,p$, we have

$$m_j = \Pi_j Y - \sum_{k \neq j} \Pi_j m_k$$

where m_ℓ represents $m_\ell(X_\ell)$, $\ell = 1, \ldots, p$. Then we may write the set of equations

$$\begin{bmatrix} I & \Pi_1 & \cdots & \Pi_1 \\ \Pi_2 & I & \cdots & \Pi_2 \\ \vdots & \vdots & \ddots & \vdots \\ \Pi_p & \Pi_p & \cdots & I \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_p \end{bmatrix} = \begin{bmatrix} \Pi_1 Y \\ \Pi_2 Y \\ \vdots \\ \Pi_p Y \end{bmatrix} .$$

The idea of the backfitting algorithm is to REPLACE the

- conditional expectation operators $\Pi_1, \ldots, \Pi_p$ with $n \times n$ smoother matrices $S_1, \ldots, S_p$ from some univariate smoothers.
- random variable $m_1, \ldots, m_p$ with the $n \times 1$ vectors $\underset{\sim}{\hat{m}}_1, \ldots, \underset{\sim}{\hat{m}}_p$ of the estimators $\hat{m}_1, \ldots, \hat{m}_p$ evaluated at the design points.
- response Y with the $n \times 1$ vector $\underset{\sim}{Y} = (Y_1, \ldots, Y_n)^T$.

Then we have the set of np equations

$$\underset{np \times np}{\begin{bmatrix} I & S_1 & \cdots & S_1 \\ S_2 & I & \cdots & S_2 \\ \vdots & \vdots & \ddots & \vdots \\ S_p & S_p & \cdots & I \end{bmatrix}} \underset{np \times 1}{\begin{bmatrix} \underset{\sim}{\hat{m}}_1 \\ \underset{\sim}{\hat{m}}_2 \\ \vdots \\ \underset{\sim}{\hat{m}}_p \end{bmatrix}} = \underset{np \times 1}{\begin{bmatrix} S_1 \underset{\sim}{Y} \\ S_2 \underset{\sim}{Y} \\ \vdots \\ S_p \underset{\sim}{Y} \end{bmatrix}} .$$

The smoother matrices could come from any linear nonparametric estimator — kernel smoothing (N-W) or local-polynomial estimators, least-square splines or penalized splines or smoothing splines.

The backfitting algorithm, which is really just an algorithm called the Gauss-Seidel algorithm, can be used to solve for $\underset{\sim}{\hat{m}}_1, \ldots, \underset{\sim}{\hat{m}}_p$ without an inversion of the big $np \times np$ matrix.

Backfitting / Gauss-Seidel Algorithm:

```
Initialize  $\underset{\sim}{\hat{m}}_1 = \underset{\sim}{\hat{m}}_2 = \cdots = \underset{\sim}{\hat{m}}_p = \underset{\sim}{0}$.
Do:  For  $j = 1, \ldots, p$
        $\underset{\sim}{\hat{m}}_j \leftarrow S_j \left( Y - \sum_{k \neq j} \underset{\sim}{\hat{m}}_k \right)$
        $\underset{\sim}{\hat{m}}_j \leftarrow \underset{\sim}{\hat{m}}_j - \left( I - \frac{1}{n} 1_n 1_n^T \right) \underset{\sim}{\hat{m}}_j$     (Centering step)
Until  $\underset{\sim}{\hat{m}}_1, \ldots, \underset{\sim}{\hat{m}}_p$ no longer change.
```

Recall that Y should be centered.

Note: If the columns of S_j sum to 1, the centering step is unecessary.

RATES OF CONVERGENCE IN THE ADDITIVE MODEL

From Stone (1985), we have the following result:

Result: Suppose $m = m_1 + \dots + m_p$, where $m_j \in \mathcal{H}(\beta, L)$ for $j=1,\dots,p$, and let $\hat{\underset{\sim}{m}}{}^{spl}_{1,r}, \dots, \hat{\underset{\sim}{m}}{}^{spl}_{p,r}$ be the fitted value vectors from least-squares splines backfitting with splines of order $r \geq \beta - 1$. Then, provided $X_1, \dots, X_n$ have a "nice" distribution, and $K_n = \alpha n^{\frac{1}{2\beta+1}}$ for some $\alpha > 0$, we have

$$\mathbb{E}\left(\frac{1}{n} \left\| \hat{\underset{\sim}{m}}{}^{spl}_{j,r} - \underset{\sim}{m}_j \right\|_2^2 \right) \leq C \cdot n^{-\frac{2\beta}{2\beta+1}}$$

for some constant $C > 0$ for large enough n.

This result means that we can estimate each additive component at the same rate as in the univariate nonparametric regression model!

Stone goes further than the stated result, proving that even if the true regression function is not additive, this rate applies to the estimator of the closest additive approximation to the true function.

The additive model thus helps mitigate the curse of dimensionality.

We here give some conditions under which we may bound MSE $\hat{m}^{spl}_{j,r}(x_0)$ and derive a bound. This is like a sketch of how to prove the result of Stone (1985).

To introduce the conditions, let

$$B_j = \left(b_{j\ell}(X_{ij}) \right)_{1 \leq i \leq n, 1 \leq \ell \leq d_n}, \quad j = 1, \dots, p,$$

$$B = [B_1, \dots, B_p]$$

and B_{-j} be the matrix B after removing B_j.

Then define $B_{j\backslash -j} = \left[I - P_{-j}\right] B_j$, where $P_{-j} = B_{-j}\left(B_{-j}^T B_{-j}\right)^{-1} B_{-j}$ for $j=1,\dots,p$.

<u>Exercise</u>: In the system of equations

$$(B_1\ B_{-1})^T(B_1\ B_{-1})\begin{pmatrix}\underset{\sim}{\hat{\alpha}}_1 \\ \underset{\sim}{\hat{\alpha}}_{-1}\end{pmatrix} = \begin{pmatrix} B_1^T B_1 & B_1^T B_{-1} \\ B_{-1}^T B_1 & B_{-1}^T B_{-1}\end{pmatrix}\begin{pmatrix}\underset{\sim}{\hat{\alpha}}_1 \\ \underset{\sim}{\hat{\alpha}}_{-1}\end{pmatrix} = \begin{bmatrix} B_1^T Y \\ B_{-1}^T Y\end{bmatrix},$$

show that $\hat{\alpha}_1 = \left(B_{1\backslash -1}{}^T B_{1\backslash -1}\right)^{-1} B_{1\backslash -1}{}^T Y$, provided $(B_1\ B_{-1})^T(B_1\ B_{-1})$ is invertible.

<u>Hint</u>: Make use of the block matrix inversion formula

$$\begin{bmatrix} A & B \\ C & D\end{bmatrix} = \begin{bmatrix} F^{-1} & -F^{-1}BD^{-1} \\ -D^{-1}CF^{-1} & D^{-1} + D^{-1}CF^{-1}BD^{-1}\end{bmatrix},$$

where $F = A - BD^{-1}C$.

<u>Solution</u>: We have

$$\begin{pmatrix}\underset{\sim}{\hat{\alpha}}_1 \\ \underset{\sim}{\hat{\alpha}}_{-1}\end{pmatrix} = \begin{pmatrix} B_1^T B_1 & B_1^T B_{-1} \\ B_{-1}^T B_1 & B_{-1}^T B_{-1}\end{pmatrix}^{-1}\begin{bmatrix} B_1^T Y \\ B_{-1}^T Y\end{bmatrix},$$

so that

$$\underset{\sim}{\hat{\alpha}}_1 = \left(B_1^T B_1 - B_1^T B_{-1}\left(B_{-1}^T B_{-1}\right)^{-1} B_{-1}^T B_1\right)^{-1} B_1^T Y$$

$$- \left(B_1^T B_1 - B_1^T B_{-1}\left(B_{-1}^T B_{-1}\right)^{-1} B_{-1}^T B_1\right)^{-1} B_1^T B_{-1}\left(B_{-1}^T B_{-1}\right)^{-1} B_{-1}^T Y$$

$$= \left(B_1^T (I - P_{-1}) B_1\right)^{-1} B_1^T (I - P_{-1}) Y$$

$\left(I - P_{-1} \text{ is idempotent} \right)$

$$= \left(\left((I-P_{-1})B_1\right)^T (I-P_{-1}) B_1\right)^{-1} \left((I-P_1)B_1\right)^T Y$$

$$= \left(B_{1\backslash -1}^T B_{1\backslash -1}\right)^{-1} B_{1\backslash -1}^T Y.$$

We see from this exercise, that for least-square splines we have

$$\hat{m}_{j,r}^{spl}(x_0) = \underset{\sim}{b}_{j,x_0}^T \left(B_{j\backslash -j}^T B_{j\backslash -j}\right)^{-1} B_{j\backslash -j}^T Y$$

for $\underset{\sim}{b}_{x_0} = \left(b_{j1}(x_0), \ldots, b_{jd_n}(x_0)\right)^T.$

<u>**Conditions:**</u> Let $m = m_1 + \ldots + m_p$, where $m_j \in \mathcal{H}(\beta, L)$ for $j = 1, \ldots, p$

Let $m_{nj,r}^{spl} \in \mathcal{M}_{n,r}$ satisfy $\| m_j - m_{nj,r}^{spl} \|_\infty \le C \cdot K_n^{-\beta}$ (We have this by de Boor (1968))

Let $X_1, \ldots, X_n \in [0,1]$ be deterministic such that for some $n_0 > 0$,

$$(C.1) \quad K_n^{-1} c_1 \le \lambda_{min}\left(\frac{1}{n} B_{j\backslash -j}^T B_{j\backslash -j}\right) \le \lambda_{max}\left(\frac{1}{n} B_{j\backslash -j}^T B_{j\backslash -j}\right) \le C_1 \cdot K_n^{-1}$$

$$(C.2) \quad \left\| \left(\frac{1}{n} B_{j\backslash -j}^T B_{j\backslash -j}\right)^{-1} \right\|_\infty \le C_2 \cdot K_n$$

$$(C.3) \quad \left\| \frac{1}{n} B_{j\backslash -j}^T \sum_{k=1}^{p} \left(\underset{\sim}{m}_j - \underset{\sim}{m}_{nj,r}^{spl}\right) \right\|_\infty \le C_3 \cdot K_n^{-1-\beta},$$

where, for $j = 1, \ldots, p$,

$$\underset{\sim}{m}_j = \left(m_j(X_{1j}), \ldots, m_j(X_{nj})\right)^T \quad \text{and} \quad \underset{\sim}{m}_{nj,r}^{spl} = \left(m_{nj,r}^{spl}(X_{1j}), \ldots, m_{nj,r}^{spl}(X_{nj})\right)^T$$

for all $n \ge n_0$, where K_n is the number of subintervals $[0,1]$ is divided into, and C, c_1, C_1, C_2, and C_3 are positive constants.

Decomposition of $\hat{m}^{spl}_{j,r}(x_0) - m_j(x_0)$ / bounding MSE $\hat{m}^{spl}_{j,r}(x_0)$:

Under (C1), (C2), and (C3), we have

$$\hat{m}^{spl}_{j,r}(x_0) - m_j(x_0) = \hat{m}^{spl}_{j,r}(x_0) - \mathbb{E}\,\hat{m}^{spl}_{j,r}(x_0) + \mathbb{E}\,\hat{m}^{spl}_{j,r}(x_0) - m_j(x_0)$$

$$= \underset{\sim}{b}_{j,x_0}^T \left(B_{j|-j}^T B_{j|-j}\right)^{-1} B_{j|-j}^T \varepsilon$$

$$+ \underset{\sim}{b}_{j,x_0}^T \left(B_{j|-j}^T B_{j|-j}\right)^{-1} B_{j|-j}^T \sum_{k=1}^{p} \underset{\sim}{m}_k - m_j(x_0)$$

$B_{j|-j}^T \underset{\sim}{m}^{spl}_{nk,r} = 0$ for $k \neq j$

$\underset{\sim}{b}_{j,x_0}^T \left(B_{j|-j}^T B_{j|-j}\right)^{-1} B_{j|-j}^T \underset{\sim}{m}^{spl}_{nj,r} = m^{spl}_{nj,r}(x_0)$

$$= \underset{\sim}{b}_{j,x_0}^T \left(B_{j|-j}^T B_{j|-j}\right)^{-1} B_{j|-j}^T \varepsilon$$

$$+ \underset{\sim}{b}_{j,x_0}^T \left(B_{j|-j}^T B_{j|-j}\right)^{-1} B_{j|-j}^T \sum_{k=1}^{p} \left(\underset{\sim}{m}_k - \underset{\sim}{m}^{spl}_{nk,r}\right)$$

$$\underset{\sim}{m}^{spl}_{nj,r}(x_0) - m_j(x_0) ,$$

where

$$\operatorname{Var}\left[\underset{\sim}{b}_{j,x_0}^T \left(B_{j|-j}^T B_{j|-j}\right)^{-1} B_{j|-j}^T \varepsilon\right] = \sigma^2\, \underset{\sim}{b}_{j,x_0}^T \left(B_{j|-j}^T B_{j|-j}\right)^{-1} \underset{\sim}{b}_{j,x_0}$$

$$= \frac{\sigma^2}{n} \underset{\sim}{b}_{j,x_0}^T \left(\frac{1}{n} B_{j|-j}^T B_{j|-j}\right)^{-1} \underset{\sim}{b}_{j,x_0}$$

$$\leq \frac{\sigma^2}{n} \|\underset{\sim}{b}_{j,x_0}\|_2^2 \frac{K_n}{c_1}$$

$\|\underset{\sim}{b}_{j,x_0}\|_2^2 \leq r+1$

$$\leq \frac{K_n}{n} \frac{\sigma^2 (r+1)}{c_1} ,$$

and

$$\underset{\sim}{b}_{j,x_0}^{T}\left(B_{j|-j}^{T}B_{j|-j}\right)^{-1}B_{j|-j}^{T}\sum_{k=1}^{p}\left(\underset{\sim}{m}_{k}-\underset{\sim}{m}_{nk,r}^{spl}\right)$$

$$=\underset{\sim}{b}_{j,x_0}^{T}\left(\frac{1}{n}B_{j|-j}^{T}B_{j|-j}\right)^{-1}\frac{1}{n}B_{j|-j}^{T}\sum_{k=1}^{p}\left(\underset{\sim}{m}_{k}-\underset{\sim}{m}_{nk,r}^{spl}\right)$$

$$\leq\left\|\underset{\sim}{b}_{j,x_0}^{T}\left(\frac{1}{n}B_{j|-j}^{T}B_{j|-j}\right)^{-1}\right\|_1\left\|\frac{1}{n}B_{j|-j}^{T}\sum_{k=1}^{p}\left(\underset{\sim}{m}_{k}-\underset{\sim}{m}_{nk,r}^{spl}\right)\right\|_\infty$$

$$\leq\underbrace{\|\underset{\sim}{b}_{j,x_0}\|_1}_{=1}\left\|\left(\frac{1}{n}B_{j|-j}^{T}B_{j|-j}\right)^{-1}\right\|_\infty\left\|\frac{1}{n}B_{j|-j}^{T}\sum_{k=1}^{p}\left(\underset{\sim}{m}_{k}-\underset{\sim}{m}_{nk,r}^{spl}\right)\right\|_\infty$$

$$\leq C_2\cdot K_n\cdot C_3\cdot K_n^{-1-\beta}$$

and

$$\underset{\sim}{m}_{nj,r}^{spl}(x_0)-m_j(x_0)\leq\left\|m_{nj,r}^{spl}-m_j\right\|_\infty\leq C\cdot K_n^{-\beta}.$$

Putting everything together, we have

$$\text{MSE}\;\hat{m}_{j,r}^{spl}(x_0)\leq C\cdot\left(K_n^{-2\beta}+\frac{K_n}{n}\right)$$

for some constant C.

Choosing $K_n=\alpha\, n^{\frac{1}{2\beta+1}}$ leads to a result like that of Stone's.

SPARSE HIGH-DIMENSIONAL ADDITIVE MODEL

As long as the # covariates p is fixed, the convergence rates discussed above for the additive model hold.

However, if we kept track of p, not absorbing it into our constants, we would see that the bias of our additive model estimators **is** scaled by the number of covariates in the model.

If we tracked the effects of p, the result of Stone (1985) would be

$$\mathbb{E}\left(\frac{1}{n}\left\|\hat{\underset{\sim}{m}}^{spl}_{j,r} - \underset{\sim}{m}_j\right\|_2^2\right) \le C \cdot p \cdot n^{-\frac{2\beta}{2\beta+1}}$$

We see that if p is very large, our estimators will perform poorly.

In order to construct good estimators when p is large, we sometimes make sparsity assumptions. In the additive model, we could assume that the set

$$\mathcal{A} = \{ j = 1, \dots, p : m_j \neq 0 \}$$

of "active" covariates has cardinality smaller than p, so that

$$Y = \sum_{j \in \mathcal{A}} m_j(X_j) + \varepsilon ,$$

with only a small number of covariates contributing to the response.

Adaptations of sparse estimators in the linear regression setting have been proposed for the sparse high-dimensional additive model.

Group lasso / adaptive group lasso:

Estimators $\hat{m}_1^L, \dots, \hat{m}_p^L$ given by $\hat{m}_j^L(x) = \sum_{\ell=1}^{d} \hat{\gamma}_{j\ell} \bar{b}_{j\ell}(x)$, $j = 1, \dots, p$, where $\underset{\sim}{\hat{\gamma}}_j = (\gamma_{j1}, \dots, \gamma_{jd})^T$ $j = 1, \dots, p$ are given by

$$\left(\underset{\sim}{\hat{\gamma}}_1, \dots, \underset{\sim}{\hat{\gamma}}_p \right) = \underset{\underset{\sim}{\gamma}_j \in \mathbb{R}^d}{\operatorname{argmin}} \left\| \underset{\sim}{Y} - \sum_{j=1}^{p} \bar{B}_{nj} \underset{\sim}{\gamma}_j \right\|_2^2 + \lambda \sum_{j=1}^{p} \| \underset{\sim}{\gamma}_j \|_2 ,$$

where $\bar{B}_{n1}, \dots, \bar{B}_{np}$ are design matrices of basis function evaluations.

This is in the form of the group lasso; the penalty sets some $\underset{\sim}{\hat{\gamma}}_j = 0$, so that the corresponding functions are equal to zero.

This can be solved very quickly with the grpreg package of Breheny.

We can also define an adaptive version of this; in a second step, obtain

$$\left(\underset{\sim}{\hat{\gamma}}_1^A, \dots, \hat{\gamma}_p^A \right) = \underset{\underset{\sim}{\gamma}_j \in \mathbb{R}^d}{\operatorname{argmin}} \left\| \underset{\sim}{Y} - \sum_{j=1}^{p} \bar{B}_{nj} \underset{\sim}{\gamma}_j \right\|_2^2 + \lambda_A \sum_{j=1}^{p} \frac{1}{\| \underset{\sim}{\hat{\gamma}}_j \|_2} \| \underset{\sim}{\gamma}_j \|_2 .$$

Then the adaptive group lasso estimator of m_j is given by

$$\hat{m}_j^{AL}(x) = \sum_{\ell=1}^{d} \underset{\sim}{\hat{\gamma}}_{j\ell}^A \bar{b}_{j\ell}(x) \quad \text{for} \quad j = 1, \dots, p.$$

The second step is called the adaptive step, and the penalty in the adaptive step promotes more sparsity while at the same time reducing the bias with which the nonzero components are estimated.

The bias coming from shrinking the estimates toward zero – not the bias from approximating the unknown functions with splines.

Tuning parameters involved in the estimators $\hat{m}_j^{AL}$ and $\hat{m}_j^{AL}$ are λ, λ_A, and the number of knots K_n on which the approximating spline functions are based.

Sparsity/smoothness penalty via group lasso: Meier (2009)

In order to penalize the wiggliness and the number of nonzero functions in the model, one may consider the estimators

$$(\hat{m}_1,\dots,\hat{m}_p) = \underset{g_j \in \bar{\mathcal{M}}_{nj,3}}{\operatorname{argmin}} \sum_{i=1}^{n}\Big(Y_i - \sum_{j=1}^{p} g_j(X_{ij})\Big)^2 + \lambda \sum_{j=1}^{p} \sqrt{\|g_j\|_n^2 + \xi \int_0^1 [g_j''(x)]^2\,dx}\,, \tag{$\bigstar$}$$

where $\bar{\mathcal{M}}_{n1,3},\dots,\bar{\mathcal{M}}_{np,3}$ are spaces of empirically centered cubic splines and

$$\|g_j\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} g_j^2(X_{ij})\,.$$

The penalty encourages sparsity and smoothness.

Exercise: Put the objective function of ($\bigstar$) into matrix form and describe how we can solve it as a group lasso problem.

We can also impose sparsity by soft-thresholding the backfitting algorithm:

Backfitting for Sparse High-dimensional Additive Model:

```
Initialize: $\underset{\sim}{\hat{m}}_1 = \underset{\sim}{\hat{m}}_2 = \dots = \underset{\sim}{\hat{m}}_p = \underset{\sim}{0}$.
Do:  For $j = 1,\dots,p$

        $\underset{\sim}{\hat{m}}_j \leftarrow S_j\big(Y - \sum_{k\neq j} \underset{\sim}{\hat{m}}_k\big)$

        $\underset{\sim}{\hat{m}}_j \leftarrow \begin{cases} (\|\underset{\sim}{\hat{m}}_j\|_n - \lambda)\dfrac{\underset{\sim}{\hat{m}}_j}{\|\underset{\sim}{\hat{m}}_j\|_n} & \text{if } \|\underset{\sim}{\hat{m}}_j\|_n > \lambda \\ 0 & \text{if } \|\underset{\sim}{\hat{m}}_j\|_n \leq \lambda \end{cases}$   (Soft-thresholding step)

        $\underset{\sim}{\hat{m}}_j \leftarrow \underset{\sim}{\hat{m}}_j - (I - \frac{1}{n}1_n 1_n^T)\underset{\sim}{\hat{m}}_j$   (Centering step)

Until: $\underset{\sim}{\hat{m}}_1,\dots,\underset{\sim}{\hat{m}}_p$ no longer change.
```