STAT 824 sp 2023 Lec 07 slides

Additive model for nonparametric multiple regression

Karl B. Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Table of Contents

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

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

where ε is independent of X with $\mathbb{E}\varepsilon = 0$ and $\mathbb{E}\varepsilon^2 = \sigma^2$.

$\mathbb{E}[Y|X] = m(X)$.

Multivariate Nadaraya-Watson estimator

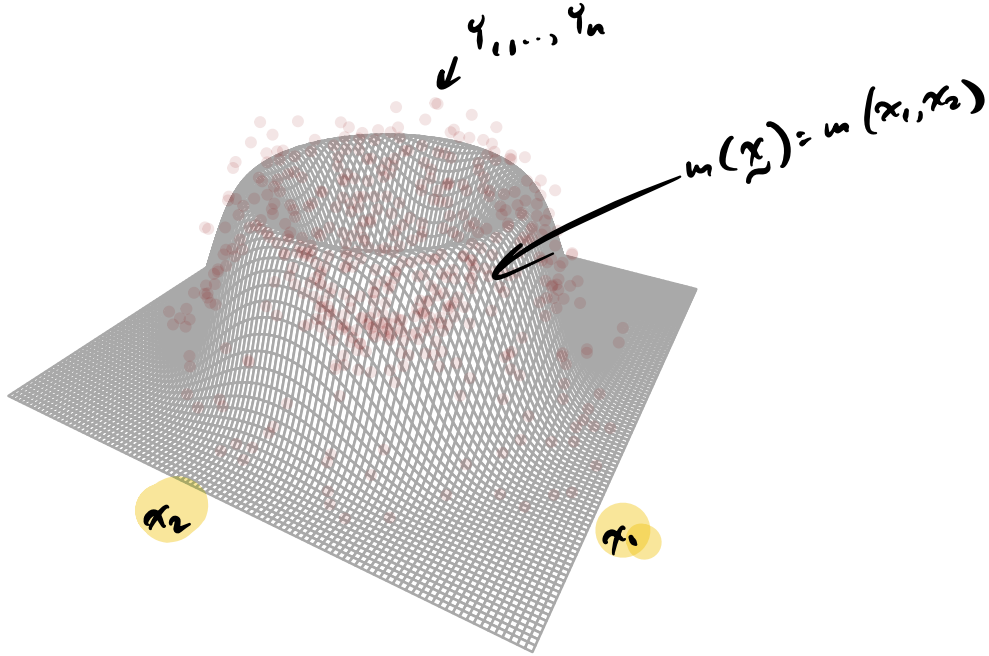
A multivariate version of the Nadaraya-Watson estimator is given by

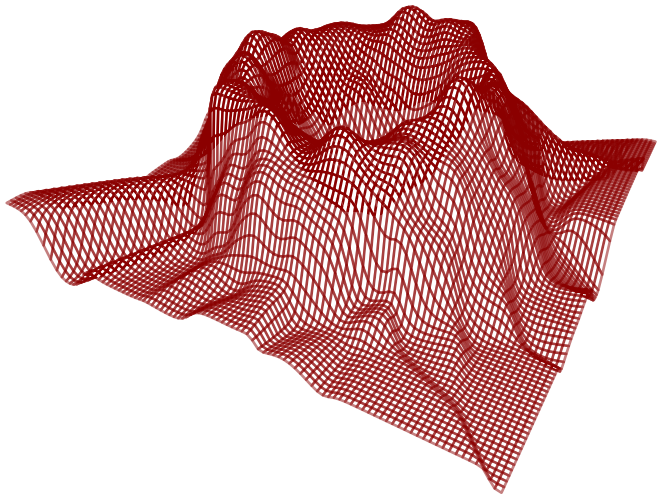
$$\hat{m}_n^{\text{NW}}(x) = \frac{\sum_{i=1}^n Y_i K(h^{-1}(X_i - x))}{\sum_{j=1}^n K(h^{-1}(X_j - x))} \qquad \text{for all } x \in [0,1]^p,$$

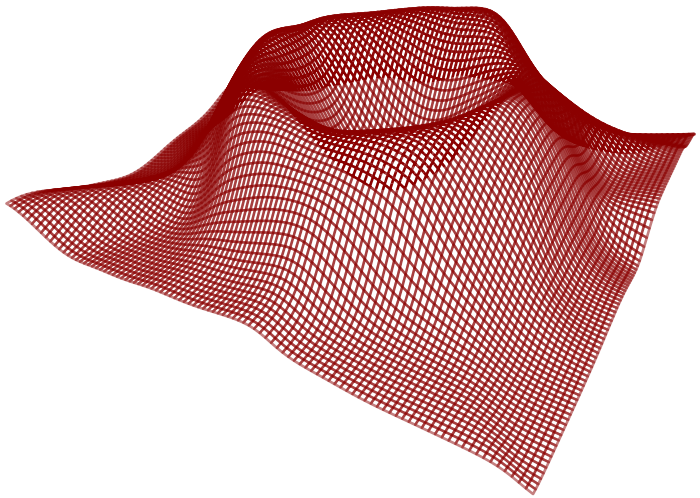
for some kernel function $K : \mathbb{R}^p \to \mathbb{R}$ and bandwidth $h > 0$.

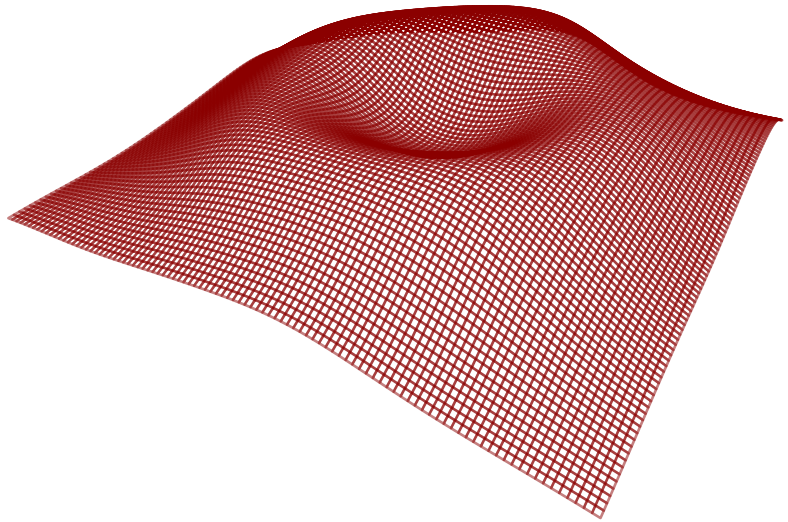
Kernel often like $K(u) = \prod_{j=1}^p G(u_j)$ where G a univariate kernel like

$$G(z) = \phi(z) \quad \text{or} \quad G(z) = \frac{3}{4}(1 - z^2)\mathbf{1}(|z| \leq 1).$$


$y_1, \ldots, y_n$
$m(\underset{\sim}{x}) = m(x_1, x_2)$
x_2
x_1







Consider the variance $\text{Var}\, \hat{m}_n^{\text{NW}}(x_0)$. We make the following assumptions.

(K1) Let $K(u) \leq K_{\max} < \infty \ \forall \ u \in \mathbb{R}^p$.

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

$$0 < c_1 \leq \frac{1}{nh^p} \sum_{i=1}^n K(h^{-1}(X_i - x)) \leq c_1^{-1}$$
$$0 < c_2 \leq \frac{1}{nh^p} \sum_{i=1}^n K^2(h^{-1}(X_i - x)) \leq c_2^{-1}$$

for some c_1, c_2, for all $x \in [0,1]^p$, for all $n \geq n_0$.

Bounds on $\text{Var}\, \hat{m}_n^{\text{NW}}(x_0)$

Under (K1) and (D1), for all $n \geq n_0$, we have

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

Exercise: Prove the above and interpret.

We could also consider local polynomial estimators in the multivariate setting.

Local linear multivariate regression estimator

A multivariate local linear estimator $\hat{m}_n^{\text{LP-1}}(x)$ of $m(x)$ is given by $\hat{\theta}_0(x)$, where

$$(\hat{\theta}_0, \hat{\boldsymbol{\theta}}_1)(x) = \underset{\theta_0 \in \mathbb{R}, \boldsymbol{\theta} \in \mathbb{R}^p}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \theta_0 - \boldsymbol{\theta}_1^T (X_i - x))^2 K(h^{-1}(X_i - x)).$$

This is also subject to the curse of dimensionality.

A way to mitigate the curse of dimensionality is by assuming *additivity*.

The *additivity* assumption is that $m : [0,1]^p \to \mathbb{R}$ may be written

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

for all $x \in [0,1]^p$ for some functions $m_1, \ldots, m_p : [0,1] \to \mathbb{R}$.

Stone (1985) argued that lots of real-world regression functions could be well-approximated by additive functions [5].

Additive model

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

$$Y = \mu + m_1(X_1) + \cdots + m_p(X_p) + \varepsilon, \quad \text{for some } m_1, \ldots, m_p : [0,1] \to \mathbb{R},$$

where ε is independent of X with $\mathbb{E}\varepsilon = 0$ and $\mathbb{E}\varepsilon^2 = \sigma^2$.

Discuss: Is the additive model identifiable? Examples.

$$Y = \mu + m_1(x_1) + m_2(x_2) + \varepsilon$$

① $Y = 1 + x_1^2 + \sin(x_2) + \varepsilon$

② $Y = 0 + (1 + x_1^2) + \sin(x_2) + \varepsilon$

③ $Y = 0 + x_1^2 + (1 + \sin(x_2)) + \varepsilon$

① $Y = 1 + (x_1^2 - \mathbb{E}x_1^2) + (\sin(x_2) - \mathbb{E}\sin(x_2)) + \varepsilon$

$$Y = \mu + m_1(x_1) + \underline{m_2(x_2)} + \varepsilon$$

$\mathbb{E}Y = \mu$

Identifiability condition for additive model

For the sake of identifiability we will assume, without loss of generality, that

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

We will make our estimators satisfy the identifiability condition empirically, i.e.

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

for any estimators $\hat{m}_1, \ldots, \hat{m}_p$.

We will always estimate μ with $\bar{Y}_n$.

From now on, assume $\mu = 0$ and that $Y_1, \ldots, Y_n$ are centered, so our model is just

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

There are **multitudes** of ways to estimate $m_1, \ldots, m_p$. One is this:

A least-squares splines estimator for the additive model

Least-squares spline estimators $\hat{m}_1^{\text{spl}}, \ldots, \hat{m}_p^{\text{spl}}$ of $m_1, \ldots, m_p$ may be defined as

$$\left(\hat{m}_1^{\text{spl}}, \ldots, \hat{m}_p^{\text{spl}}\right) = \underset{g_j \in \bar{\mathcal{M}}_{nj}}{\text{argmin}} \sum_{i=1}^n \left[Y_i - \sum_{j=1}^p g_j(X_{ij})\right]^2,$$

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

$$\bar{\mathcal{M}}_{nj} = \text{span}\{\bar{b}_{j1}, \ldots, \bar{b}_{jd}\}, \text{ with } \bar{b}_{jl}(x) = b_{jl}(x) - n^{-1} \sum_{i=1}^n b_{jl}(X_{ij}),$$

for $l = 1, \ldots, d$, where $b_{j1}, \ldots, b_{jd}$ are cubic B-spline basis functions.

Note that I have defaulted to cubic splines (we can use splines of other orders).

Exercise:

1. Verify that each $\hat{m}_j^{\text{spl}}$ will satisfy $n^{-1} \sum_{i=1}^n \hat{m}_j^{\text{spl}}(X_{ij}) = 0$.
2. Write the objective function in matrices. Give normal equations.
3. Check whether $\bar{\mathbf{B}}_{jn} = \left(\bar{b}_{j\ell}(X_{ij})\right)_{1 \le i \le n, 1 \le \ell \le d}$ has full rank.

$$m_j(x) \approx \sum_{k=1}^{d_j} \alpha_{jk} b_{jk}(x)$$

$$g_j(x) = \sum_{k=1}^{d_j} \alpha_{jk} \bar{b}_{jk}(x)$$

need $\frac{1}{n}\sum_{i=1}^{n} g_j(X_{ij}) = 0$ for each j.

To enforce, modify basis functions:

$$\bar{b}_{jk}(x) = b_{jk}(x) - \frac{1}{n}\sum_{i=1}^{n} b_{jk}(X_{ij})$$

(empirically center our basis functions).

minimize over $g_1, \dots, g_p$

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} g_j(X_{ij})\right)^2$$

$$= \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\sum_{k=1}^{d_j} \alpha_{jk}\bar{b}_{jk}(X_{ij})\right)^2$$

$$= \left\| \underset{\sim}{Y} - \sum_{j=1}^{p} \bar{\mathbb{B}}_j \underset{\sim}{\alpha}_j \right\|_2^2$$

$$\bar{\mathbb{B}}_j = \begin{pmatrix} \bar{b}_{j1}(X_{j1}) & \cdots & \bar{b}_{jd_j}(X_{j1}) \\ \vdots & & \\ \bar{b}_{j1}(X_{jn}) & \cdots & \bar{b}_{jd_j}(X_{jn}) \end{pmatrix}$$

$$= \left\| \underset{\sim}{Y} - \underbrace{\left[\bar{\mathbb{B}}_1 \cdots \bar{\mathbb{B}}_p\right]}_{\bar{\mathbb{B}}} \underbrace{\begin{bmatrix} \underset{\sim}{\alpha}_1 \\ \vdots \\ \underset{\sim}{\alpha}_p \end{bmatrix}}_{\underset{\sim}{\alpha}} \right\|_2^2$$

$$\underset{\sim}{\alpha}_j = \begin{pmatrix} \alpha_{j1} \\ \vdots \\ \alpha_{jd_j} \end{pmatrix} \qquad \underset{\sim}{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

$$= \| \underset{\sim}{Y} - \bar{\mathbb{B}}\,\underset{\sim}{\alpha} \|_2^2$$

$$\hat{\alpha} = (\bar{\mathbb{B}}^T\bar{\mathbb{B}})^{-1}\bar{\mathbb{B}}^T\underset{\sim}{Y}.$$

$$\hat{m}_j(x) = \sum_{k=1}^{d_j} \hat{\alpha}_{jk}\bar{b}_{jk}(x)$$

$[\mathbb{B}_1\ \mathbb{B}_2]$ is not full rank, i.e. $(\mathbb{B}_1\ \mathbb{B}_2)^T(\mathbb{B}_1\ \mathbb{B}_2)$ is not invertible.

$\hat{m}_j$ evaluated at design points

$$\underset{\sim}{\hat{m}}_j = \bar{\mathbb{B}}_j \underset{\sim}{\hat{\alpha}}_j$$

To center the columns of a nxd matrix $\mathbb{B}$, do

$$\overline{\mathbb{B}} = \left(\mathbb{I}_n - \frac{1}{n}\underset{n\times 1}{\mathbb{1}_n}\underset{1\times n}{\mathbb{1}_n^T}\right)\mathbb{B} \quad , \underset{n\times 1}{\mathbb{1}_n} \text{ of ones.}$$

$$= \mathbb{B} - \mathbb{1}_n \frac{1}{n}\mathbb{1}_n^T\mathbb{B}$$

1xd vector of column means

$$Y_i = m_1(X_{1i}) + m_2(X_{2i}) + \varepsilon_i$$

$$Y_i - m_2(X_{2i}) = m_1(X_{1i}) + \varepsilon_i$$

$$Y_i - m_1(X_{1i}) = m_2(X_{2i}) + \varepsilon_i$$

Make plots

Plot

$Y_i - \hat{m}_2(X_{2i}) \sim X_{1i}$ ← should look like shape of m_1

$Y_i - \hat{m}_1(X_{1i}) \sim X_{2i}$ ← should show me m_2

---

any p

$$Y_i = \sum_{j=1}^{p} m_j(X_{ji}) + \varepsilon_i$$

$$Y_i - \sum_{k\neq j} m_k(X_{ki}) = m_j(X_{ji}) + \varepsilon_i$$

Plot

$$Y_i - \sum_{k\neq j} \hat{m}_k(X_{ki}) \quad \text{against} \quad X_{ji}$$

For $\bar{\mathbf{B}}_n = [\bar{\mathbf{B}}_{n1}, \ldots, \bar{\mathbf{B}}_{np}]$ with $\bar{\mathbf{B}}_{nj} = (\bar{b}_{jk}(X_{ij}))_{1 \leq i \leq n, 1 \leq k \leq d}$, the solution to

$$(\bar{\mathbf{B}}_n^T \bar{\mathbf{B}}_n)\alpha = \bar{\mathbf{B}}_n^T \mathbf{Y}$$

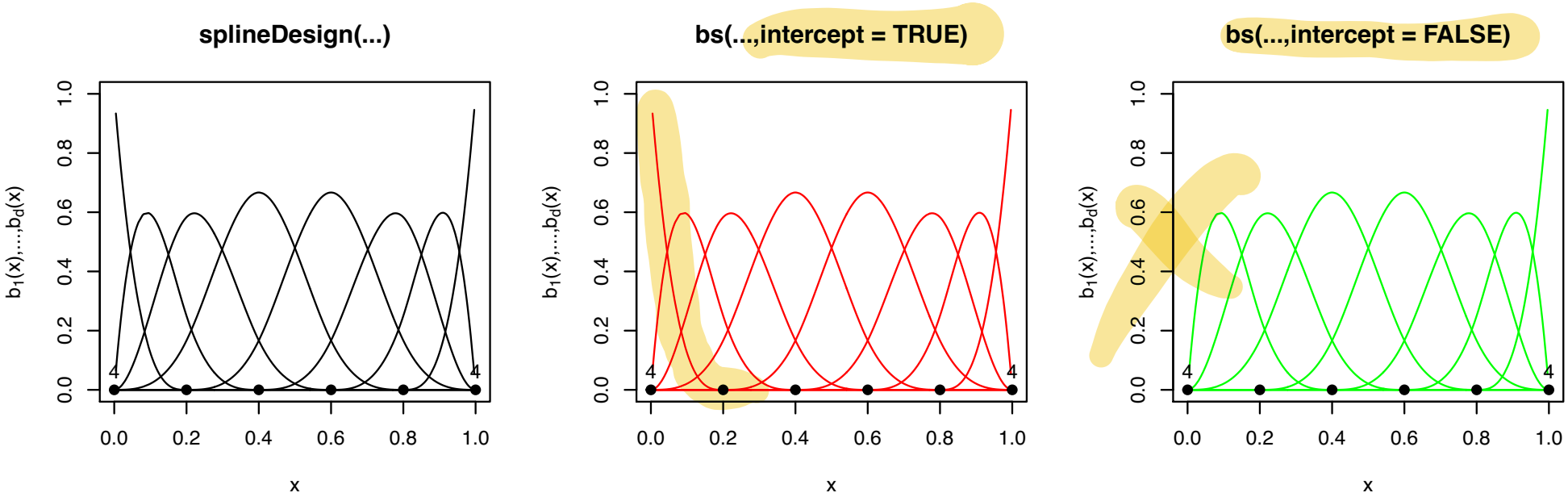
is not unique, because the $\bar{\mathbf{B}}_{n1}, \ldots, \bar{\mathbf{B}}_{np}$ do not have full-column rank—due to:

The B-splines have the property that $\sum_{\ell=1}^{d} b_{j\ell}(x) = 1$ for all $x \in [0, 1]$.

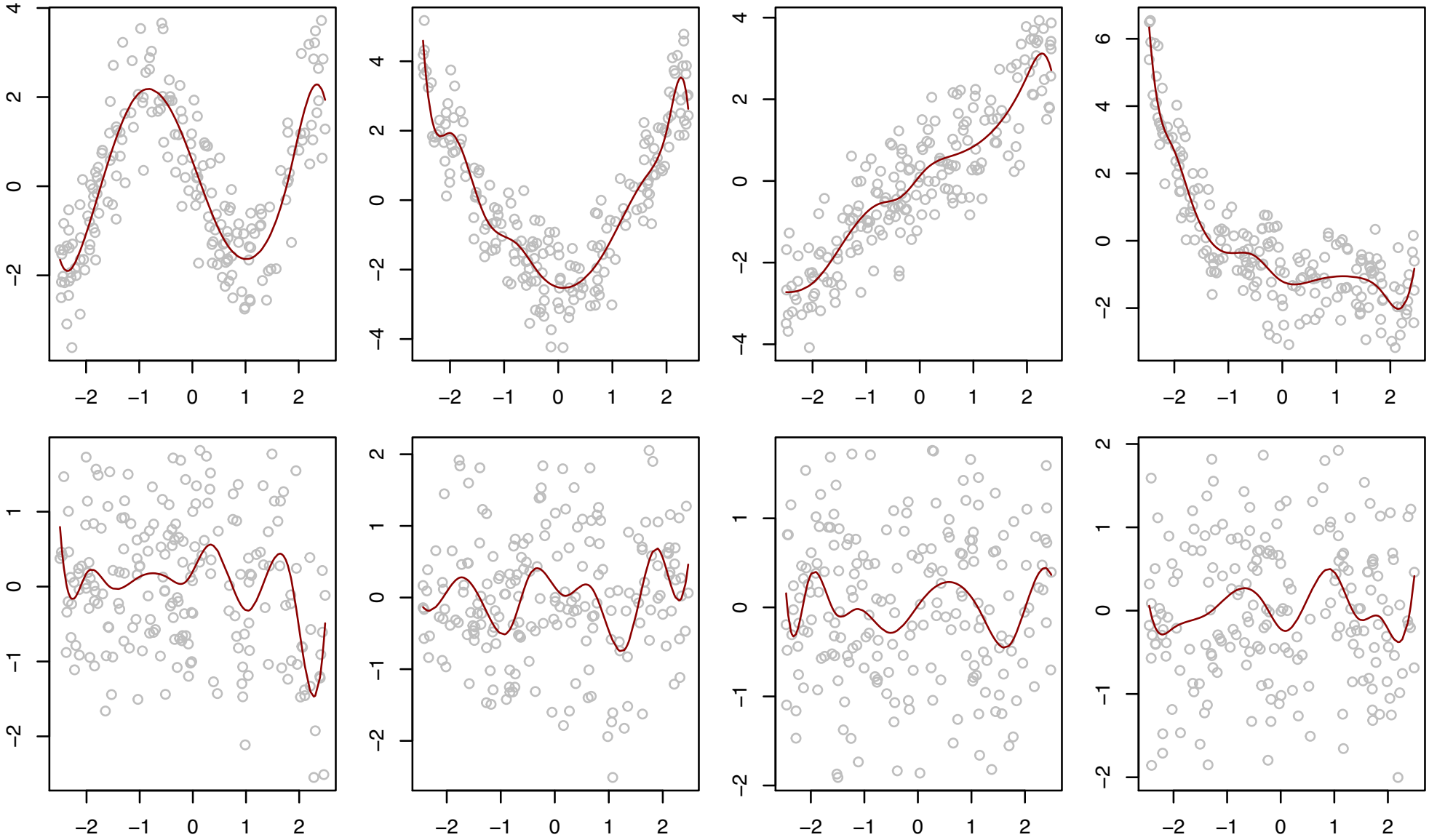
One fix is to discard the first basis function $b_{11}, \ldots, b_{p1}$ for each component...

Illustrate: Write up some code for fitting LS splines in the additive model.

The `bs()` function with `intercept = FALSE` removes the first basis function:



Least–squares splines estimator (directly computed with basis functions centered, 1 removed)



$= \hat{\alpha}$: $\underset{\alpha \in \mathbb{R}^p}{\operatorname{argmin}} \| Y - B\alpha \|_2^2 = (B^T B)^{-1} B^T Y$

Exercise: Let $\mathbf{B} = [\mathbf{B}_1, \mathbf{B}_{-1}]$ have full column rank and let

$$\mathbf{P}_{-1} = \mathbf{B}_{-1}(\mathbf{B}_{-1}^T\mathbf{B}_{-1})^{-1}\mathbf{B}_{-1} \quad \text{and} \quad \mathbf{B}_{1\setminus -1} = (\mathbf{I} - \mathbf{P}_{-1})\mathbf{B}_1.$$

← the part of B_1 not explained by B_{-1}.

$= B_1 - P_{-1} B_1$

1. Show that
$$\begin{bmatrix} \mathbf{B}_1^T\mathbf{B}_1 & \mathbf{B}_1^T\mathbf{B}_{-1} \\ \mathbf{B}_{-1}^T\mathbf{B}_1 & \mathbf{B}_{-1}^T\mathbf{B}_{-1} \end{bmatrix} \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1^T\mathbf{Y} \\ \mathbf{B}_{-1}^T\mathbf{Y} \end{bmatrix}$$
if and only if $\hat{\alpha}_1 = (\mathbf{B}_{1\setminus -1}^T\mathbf{B}_{1\setminus -1})^{-1}\mathbf{B}_{1\setminus -1}\mathbf{Y}$.
2. If $\text{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$, give $\text{Var}(\mathbf{b}_1^T\hat{\alpha}_1)$.

Discuss: The purpose of this exercise.

$$\| \underset{\sim}{y} - \mathbf{B}\underline{\alpha} \|_2^2 = \underset{\sim}{y}^T\underset{\sim}{y} - 2\underset{\sim}{y}^T\mathbf{B}\underline{\alpha} + \underline{\alpha}^T\mathbf{B}^T\mathbf{B}\underline{\alpha}$$

$$\frac{\partial}{\partial \underline{\alpha}}[\cdot] = -2\mathbf{B}^T\underset{\sim}{y} + 2\mathbf{B}^T\mathbf{B}\underline{\alpha} \overset{\text{set}}{=} 0$$

$$\mathbf{B}^T\mathbf{B}\,\hat{\underline{\alpha}} = \mathbf{B}^T\underset{\sim}{y}$$

$\mathbf{B} = [\mathbf{B}_1, \mathbf{B}_{-1}]$

$$\begin{pmatrix} \mathbf{B}_1^T\mathbf{B}_1 & \mathbf{B}_1^T\mathbf{B}_{-1} \\ \mathbf{B}_{-1}^T\mathbf{B}_1 & \mathbf{B}_{-1}^T\mathbf{B}_{-1} \end{pmatrix} \begin{pmatrix} \hat{\underline{\alpha}}_1 \\ \hat{\underline{\alpha}}_{-1} \end{pmatrix} = \begin{bmatrix} \mathbf{B}_1^T\underset{\sim}{y} \\ \mathbf{B}_{-1}^T\underset{\sim}{y} \end{bmatrix}$$

Solve for $\hat{\underline{\alpha}}_1$

We can also penalize the wiggliness of the fitted functions:

A penalized splines estimator for the additive model

Penalized spline estimators $\hat{m}_1^{\text{pspl}}, \ldots, \hat{m}_p^{\text{pspl}}$ of $m_1, \ldots, m_p$ may be defined as

$$\left(\hat{m}_1^{\text{pspl}}, \ldots, \hat{m}_p^{\text{pspl}}\right) = \underset{g_j \in \bar{\mathcal{M}}_{nj}}{\text{argmin}} \sum_{i=1}^n \left[Y_i - \sum_{j=1}^p g_j(X_{ij})\right]^2 + \lambda \sum_{j=1}^p \int_0^1 [g_j''(x)]^2 dx,$$

for some $\lambda > 0$, where for $j = 1, \ldots, p$,

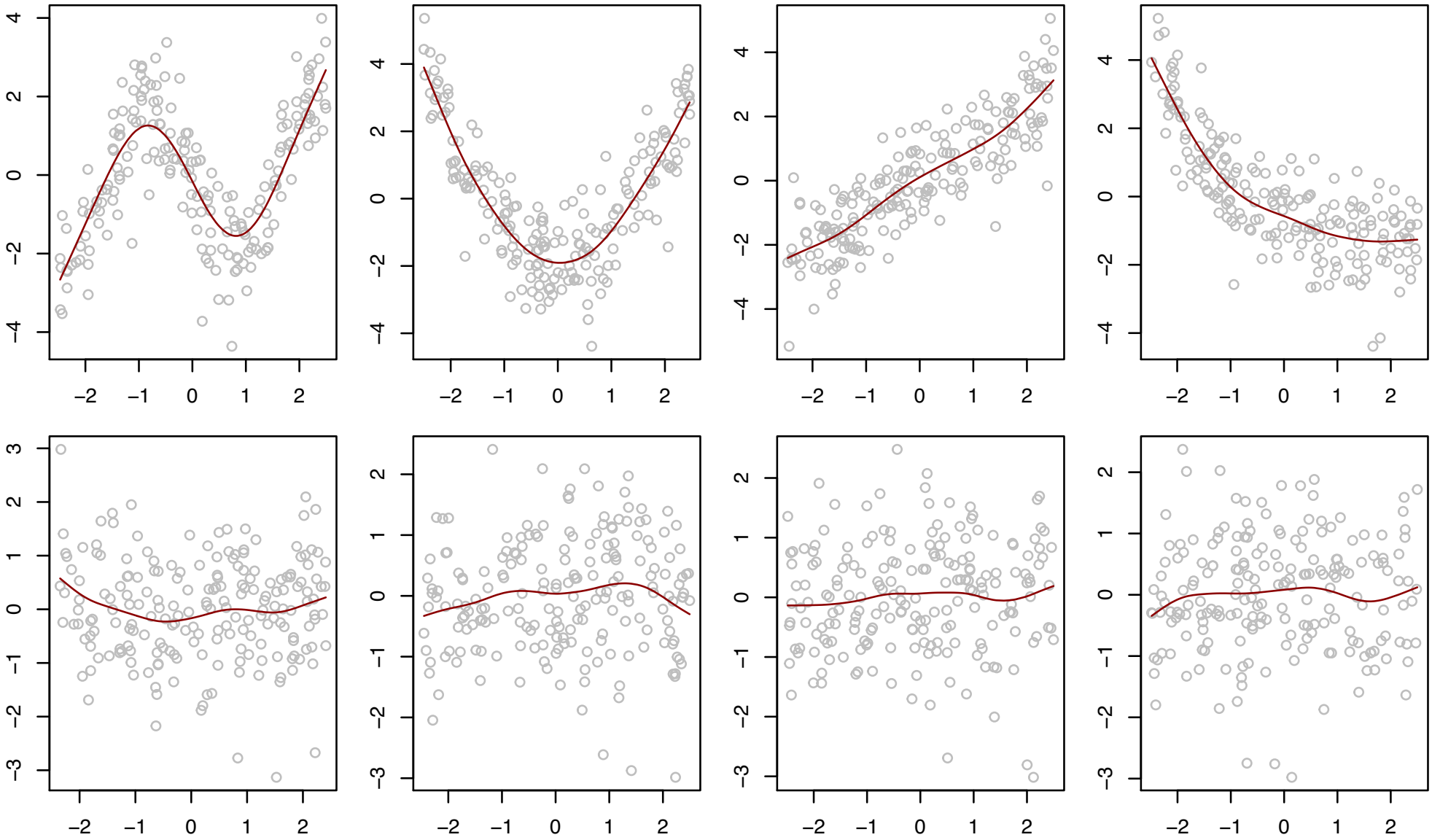
$$\bar{\mathcal{M}}_{nj} = \text{span}\{\bar{b}_{j1}, \ldots, \bar{b}_{jd}\}, \text{ with } \bar{b}_{jl}(x) = b_{jl}(x) - n^{-1} \sum_{i=1}^n b_{jl}(X_{ij}),$$

for $l = 1, \ldots, d$, where $b_{j1}, \ldots, b_{jd}$ are cubic B-spline basis functions.

Exercise:

1. Write the objective function in matrices. Give normal equations. Issues?
2. Show and run sample *R code* for fitting the penalized splines estimator.

Penalized splines estimator (directly computed with basis functions centered, 1 removed)



Backfitting is a fast and simple way to compute estimators in the additive model.

It also spares us from the numerical issues encountered in the last few slides.

The components of the model $Y = \sum_{j=1}^{p} m_j(X_j) + \varepsilon$ have the interpretation

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

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

Now, letting Π_j represent conditional expectation given X_j, we have

$$m_j = \Pi_j Y - \sum_{k \neq j} \Pi_j m_k \text{ for } j = 1, \ldots, p \quad (\text{with } m_k := m_k(X_k)).$$

We can write this system of equations as

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

$$Y = \sum_{j=1}^{p} m_j(X_j) + \varepsilon \quad , \quad \mathbb{E}[\varepsilon \mid \underset{\sim}{X}] = 0$$

$$\mathbb{E}[Y \mid X_j] = \mathbb{E}\left[\sum_{j=1}^{p} m_j(X_j) + \varepsilon \mid X_j \right]$$

$$= m(X_j) + \sum_{k \neq j} \mathbb{E}\left[m_k(X_k) \mid X_j \right]$$

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

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

$$\mathbb{E}[Y \mid X_1] = \mathbb{E}[m(X_1) \mid X_1] + \cdots + \mathbb{E}[m(X_p) \mid X_1] + \mathbb{E}[\varepsilon \mid X_1]$$

$$= m_1(X_1) + \cdots + \mathbb{E}[m(X_p) \mid X_1]$$

$$= m_1(X_1) + \sum_{j \neq 1} \mathbb{E}[m(X_j) \mid X_1]$$

$$\mathbb{E}[Y \mid X_2] = m_2(X_1) + \sum_{j \neq 2} \mathbb{E}[m(X_j) \mid X_2]$$

$$\vdots$$

$$\mathbb{E}[Y \mid X_k] = m_k(X_k) + \sum_{j \neq k} \mathbb{E}[m(X_j) \mid X_k] \qquad k = 1, \ldots, p$$

$$\begin{bmatrix} \mathbb{E}[Y \mid X_1] \\ \mathbb{E}[Y \mid X_2] \\ \vdots \\ \mathbb{E}[Y \mid X_p] \end{bmatrix} = \begin{bmatrix} \overbrace{m_1(X_1)}^{m_1} + \sum_{j \neq 1} \mathbb{E}[\overbrace{m(X_j)}^{m_j} \mid X_1] \\ \underbrace{m_2(X_2)}_{m_2} + \sum_{j \neq 2} \mathbb{E}[m(X_j) \mid X_2] \\ \\ \\ m_p(X_p) + \sum_{j \neq p} \mathbb{E}[m(X_j) \mid X_p] \end{bmatrix}$$

Let Π_j be like $\mathbb{E}[\cdot \mid x_j]$. Then the above is

$$\begin{bmatrix} \Pi_1 y \\ \Pi_2 y \\ \vdots \\ \Pi_p y \end{bmatrix} = \begin{bmatrix} m_1 + \sum_{j \neq 1} \Pi_1 m_j \\ m_2 + \sum_{j \neq 2} \Pi_2 m_j \\ \vdots \\ m_p + \sum_{j \neq p} \Pi_p m_j \end{bmatrix}$$

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

Remember the smoother matrix:

$$\begin{bmatrix} \hat{m}(x_1) \\ \vdots \\ \hat{m}(x_n) \end{bmatrix} = \underset{n \times n}{S} \; \underset{\sim}{y}$$

$\hat{m}$ at design points

The *backfitting algorithm* solves an empirical version

$$\begin{bmatrix} \mathbf{I}_n & \mathbf{S}_1 & \dots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I}_n & \dots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p & \mathbf{S}_p & \dots & \mathbf{I}_n \end{bmatrix} \begin{bmatrix} \hat{\mathbf{m}}_1 \\ \hat{\mathbf{m}}_2 \\ \vdots \\ \hat{\mathbf{m}}_p \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1\mathbf{Y} \\ \mathbf{S}_2\mathbf{Y} \\ \vdots \\ \mathbf{S}_p\mathbf{Y} \end{bmatrix}.$$

of the system of equations on the previous slide, where

- $\mathbf{S}_1, \dots, \mathbf{S}_p$ are smoother matrices associated with univariate smoothers.
- $\hat{\mathbf{m}}_1, \dots, \hat{\mathbf{m}}_p$ are $n \times 1$ with evaluations of estimators at design points.
- $\mathbf{Y} = (Y_1, \dots, Y_n)^T$.

Backfitting algorithm (Gauss–Seidel). See Buja et al. (1989), [1].

Initialize: $\hat{\mathbf{m}}_1, \dots, \hat{\mathbf{m}}_p = \mathbf{0}$. Then iterate: For $j = 1, \dots, p$

1. $\hat{\mathbf{m}}_j \leftarrow \mathbf{S}_j(\mathbf{Y} - \sum_{k \neq j} \hat{\mathbf{m}}_k)$
2. $\hat{\mathbf{m}}_j \leftarrow \hat{\mathbf{m}}_j - n^{-1}\mathbf{1}_n^T\hat{\mathbf{m}}_j$ (centering step for identifiability)

until $\hat{\mathbf{m}}_1, \dots, \hat{\mathbf{m}}_p$ no longer change.

To implement the backfitting algorithm, we just need the smoother matrices.

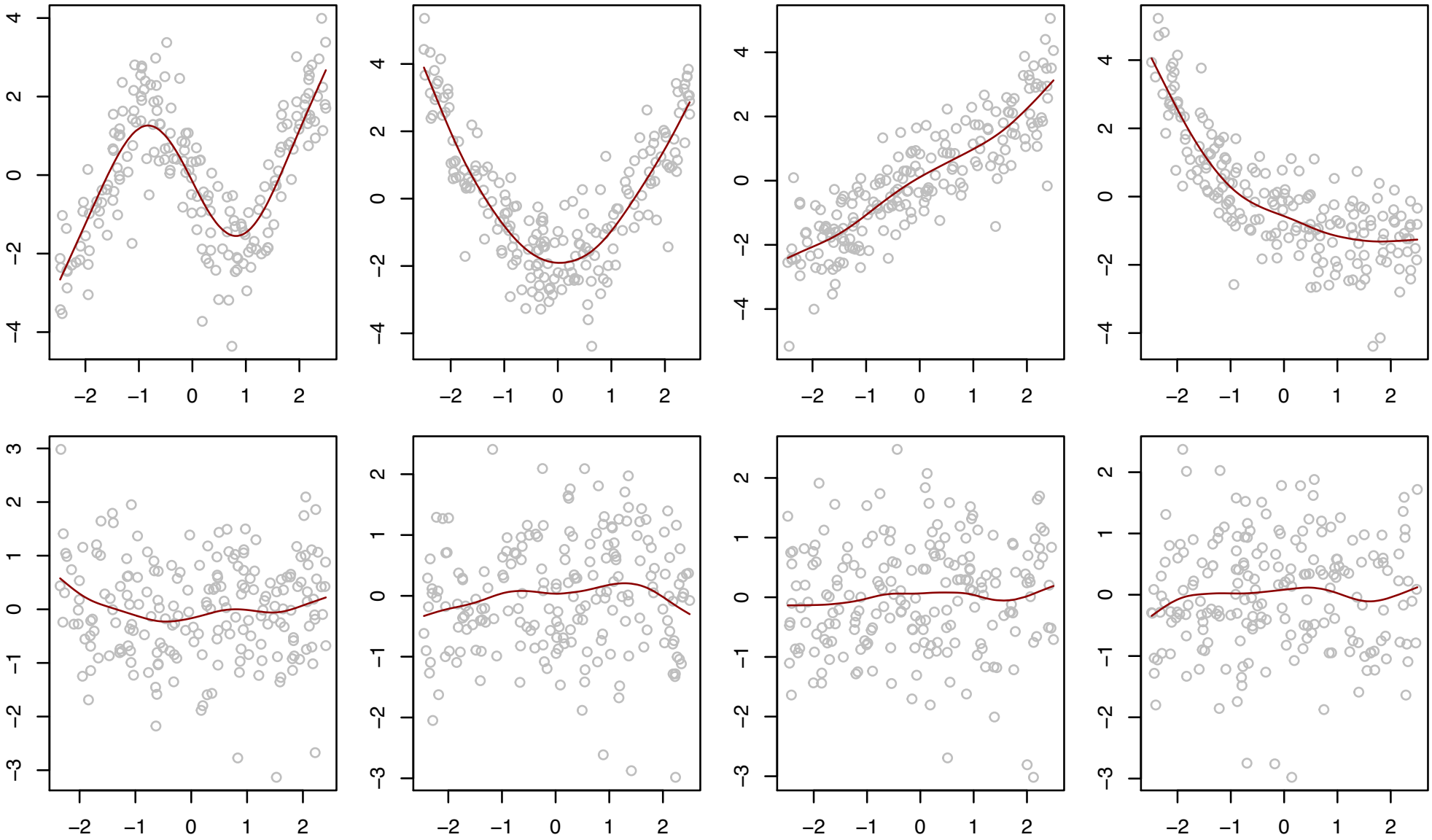
- Least-squares splines: $\mathbf{S}_j = \mathbf{B}_{nj}(\mathbf{B}_{nj}^T\mathbf{B}_{nj})^{-1}\mathbf{B}_{nj}^T$
- Penalized splines: $\mathbf{S}_j = \mathbf{B}_{nj}(\mathbf{B}_{nj}^T\mathbf{B}_{nj} + \lambda\mathbf{\Omega}_j)^{-1}\mathbf{B}_{nj}^T$

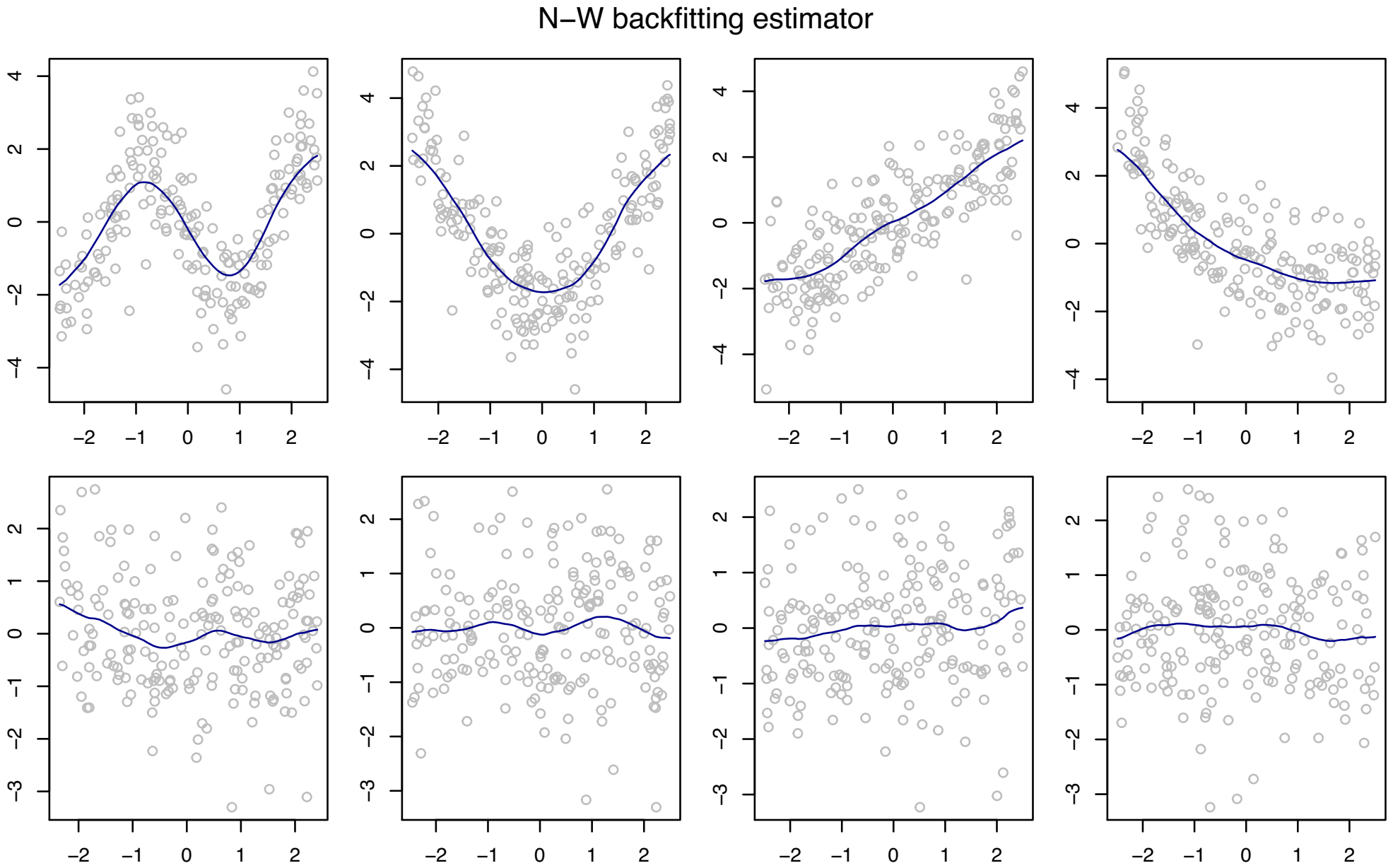
  (Note that there is no need to center the basis functions.)

- Nadaraya-Watson: $\mathbf{S}_j = \left(\frac{K(h^{-1}(X_{kj} - X_{ij}))}{\sum_{\ell=1}^{n} K(h^{-1}(X_{\ell j} - X_{ij}))} \right)_{1 \leq i \leq n, 1 \leq k \leq n}$

Exercise: Demonstrate backfitting with penalized splines. Use *R code*.

Penalized splines backfitting estimator




N–W backfitting estimator

$$\underset{n\times 1}{\tilde{Y}} = \underset{n\times 1}{\tilde{X}_1}\beta_1 + \cdots + \underset{n\times 1}{\tilde{X}_p}\beta_p + \underset{n\times 1}{\tilde{\varepsilon}}$$

Let $\mathbb{X} = [\tilde{X}_1, \ldots, \tilde{X}_p]$, then $\hat{\tilde{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\tilde{Y}$

Exercise: Try backfitting for multiple linear regression, $m_1(x) = \beta_1 x$ $\quad m_p(x) = \beta_p x$

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

with $\mathbb{E}X_j = 0$ for $j = 1, \ldots, p$.

1. What are the smoother matrices $\mathbf{S}_1, \ldots, \mathbf{S}_p$ in this context?
2. Derive and write down the details of the backfitting algorithm.
3. How do you obtain the least-squares estimators $\hat{\beta}_1, \ldots, \hat{\beta}_p$ from the output?
4. Implement on some made-up data.

Make $S_1, \ldots, S_p$ for simple linear reg.

$$\underset{n\times 1}{\hat{\tilde{Y}}} = \underset{n\times 1}{\tilde{X}_1}\hat{\beta}_1 + \cdots + \underset{n\times 1}{\tilde{X}_p}\hat{\beta}_p = \hat{\tilde{m}}_1 + \cdots + \hat{\tilde{m}}_p.$$

$$\underset{n\times n}{S_1} = \tilde{X}_1(\tilde{X}_1^T\tilde{X}_1)^{-1}\tilde{X}_1^T \qquad (\tilde{Y} = \tilde{X}_1\beta + \tilde{\varepsilon})$$

$$\underset{n\times n}{S_p} = \tilde{X}_p(\tilde{X}_p^T\tilde{X}_p)^{-1}\tilde{X}_p^T$$

Now consider the performance of nonparametric estimators in the additive model.

Least-squares splines performance in additive model, Stone (1985), [5]

Suppose $m = m_1 + \cdots + m_p$, where $m_j \in \mathcal{H}(\beta, L)$ for $j = 1, \ldots, p$, and let $\hat{\mathbf{m}}_{1,r}^{\text{spl}}, \ldots, \hat{\mathbf{m}}_{p,r}^{\text{spl}}$ be $n \times 1$ vectors with the fitted values of the least-squares splines estimators of order $r \geq \beta - 1$ and the true functions. Then, provided $X_1, \ldots, X_n$ have a "nice" distribution and $K_n = \alpha n^{\frac{1}{2\beta+1}}$ for some $\alpha > 0$, we have

MISE →

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

← the optimal univariate rate.

C ↑ p

each $j = 1, \ldots, p$ for some constant $C > 0$ for large enough n.

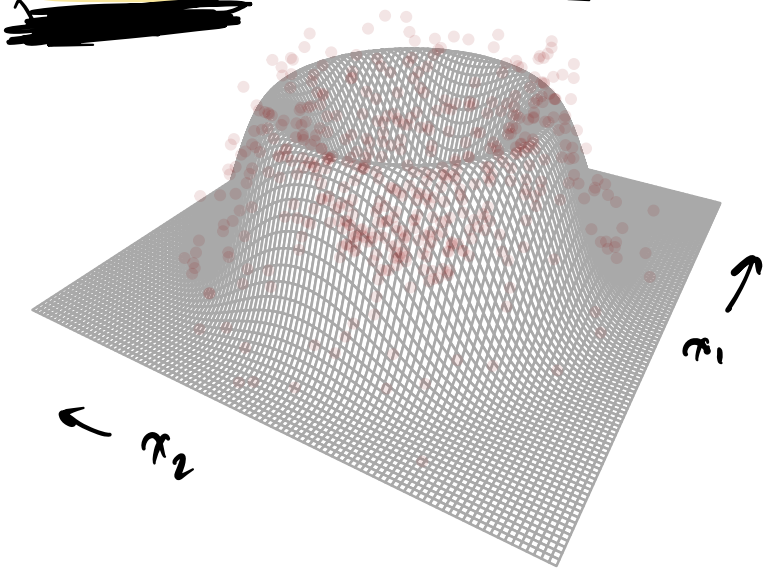
The $\mathbf{m}_1, \ldots, \mathbf{m}_p$ are $n \times 1$ with evaluations of the true functions at design points.

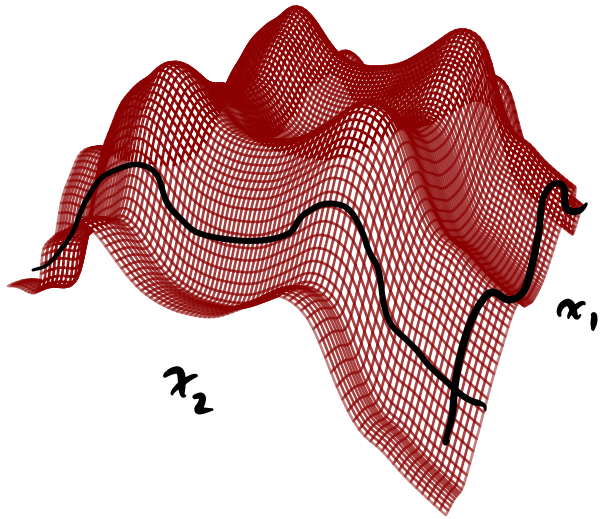
We estimate the additive model components at the univariate nonparametric rate!

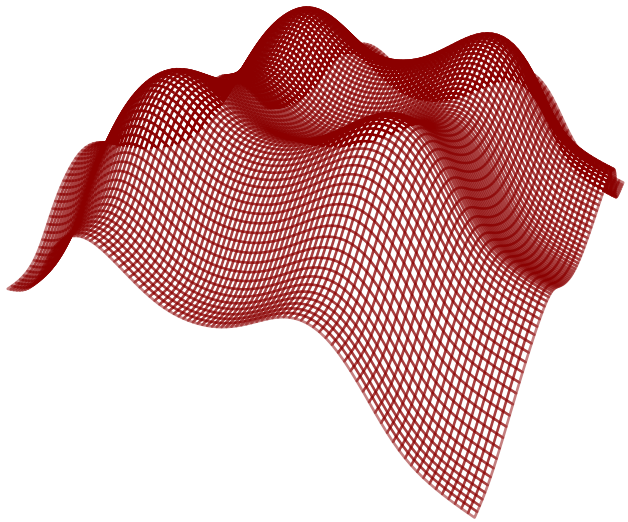
Discuss: What if the additivity assumption is false?

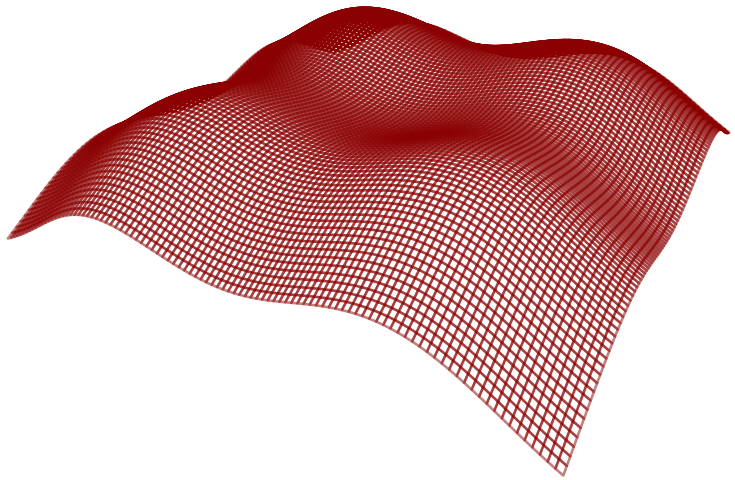
Additivity assumes

$$m(x_1, x_2) = m_1(x_1) + m_2(x_2)$$









Stone (1985) treated p as a constant, absorbing it into C. If we track p, we get

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

So we see that the dimension, i.e. # of covariates, affects estimation.

The sparse additive model

$Y = m(x_1) + \cdots + m_p(x_p) + \varepsilon$

In large-p settings, we often make a *sparsity assumption*; we assume

$$s := |\mathcal{A}| < p \quad \text{where} \quad \mathcal{A} = \{j : m_j \neq 0\}.$$

$m_j = 0$ for $j \notin \mathcal{A}$

This means that some of the functions are equal to zero, giving

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

$$= \sum_{j=1}^{p} m_j(x_j) + \varepsilon$$

The covariates with indices in $\mathcal{A}$ are sometimes called "active".

Many estimators have been proposed in this setting.

Sparsity via the group lasso. See Huang et al. (2010), [2].

Group lasso estimators $\hat{m}_1^{\mathsf{L}}, \ldots, \hat{m}_1^{\mathsf{L}}$ of $m_1, \ldots, m_p$ can be defined as

$$\hat{m}_j^{\mathsf{L}}(x) = \sum_{k=1}^{d} \hat{\alpha}_{jk}^{\mathsf{L}} \bar{b}_{jk}(x), \quad j = 1, \ldots, p,$$

where $\hat{\boldsymbol{\alpha}}_j^{\mathsf{L}} = (\hat{\alpha}_{j1}^{\mathsf{L}}, \ldots, \hat{\alpha}_{jd}^{\mathsf{L}})^T$, $j = 1, \ldots, p$ are given by

group lasso penalty.

$$(\hat{\boldsymbol{\alpha}}_1^{\mathsf{L}}, \ldots, \hat{\boldsymbol{\alpha}}_p^{\mathsf{L}}) = \underset{\boldsymbol{\alpha}_j \in \mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{Y} - \sum_{j=1}^{p} \bar{\mathbf{B}}_{nj} \boldsymbol{\alpha}_j\|_2^2 + \lambda \sum_{j=1}^{p} \|\boldsymbol{\alpha}_j\|_2,$$

with $\bar{\mathbf{B}}_{nj} = (\bar{b}_{jk}(X_{ij}))_{1 \le i \le n, 1 \le k \le d}$, $j = 1, \ldots, p$.

Can get adaptive lasso estimators $\hat{m}_j^{\mathsf{AL}}(x) = \sum_{k=1}^{d} \hat{\alpha}_{jk}^{\mathsf{AL}} \bar{b}_{jk}(x), \quad j = 1, \ldots, p$, with

$$(\hat{\boldsymbol{\alpha}}_1^{\mathsf{AL}}, \ldots, \hat{\boldsymbol{\alpha}}_p^{\mathsf{AL}}) = \underset{\boldsymbol{\alpha}_j \in \mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{Y} - \sum_{j=1}^{p} \bar{\mathbf{B}}_{nj} \boldsymbol{\alpha}_j\|_2^2 + \lambda_{\mathsf{A}} \sum_{j=1}^{p} \frac{1}{\|\hat{\boldsymbol{\alpha}}_j^{\mathsf{L}}\|_2} \cdot \|\boldsymbol{\alpha}_j\|_2,$$

A sparse penalized splines estimator. See Meier et al. (2009), [3].

Sparse pen. spline estimators $\hat{m}_1^{\text{spspl}}, \ldots, \hat{m}_p^{\text{spspl}}$ of $m_1, \ldots, m_p$ may be defined as

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

for some $\lambda > 0, \xi \geq 0$, where for $j = 1, \ldots, p$,

$$\bar{\mathcal{M}}_{nj} = \text{span}\{\bar{b}_{j1}, \ldots, \bar{b}_{jd}\}, \text{ with } \bar{b}_{jl}(x) = b_{jl}(x) - n^{-1} \textstyle\sum_{i=1}^{n} b_{jl}(X_{ij}),$$

for $l = 1, \ldots, d$, where $b_{j1}, \ldots, b_{jd}$ are cubic B-spline basis functions.

Exercise: Show how this can be formulated as a group lasso problem.

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

Soft-thresholded backfitting algorithm. Ravikumar et al. (2009), [4]

Initialize: $\hat{\mathbf{m}}_1, \ldots, \hat{\mathbf{m}}_p = \mathbf{0}$. Then iterate: For $j = 1, \ldots, p$

1. $\hat{\mathbf{m}}_j \leftarrow \mathbf{S}_j(\mathbf{Y} - \sum_{k \neq j} \hat{\mathbf{m}}_k)$
2. $\hat{\mathbf{m}}_j \leftarrow \begin{cases} \hat{\mathbf{m}}_j \cdot (\|\hat{\mathbf{m}}_j\|_n - \lambda)/\|\hat{\mathbf{m}}_j\|_n, & \text{if } \|\hat{\mathbf{m}}_j\|_n \geq \lambda \\ \mathbf{0}, & \text{if } \|\hat{\mathbf{m}}_j\|_n < \lambda \end{cases}$
3. $\hat{\mathbf{m}}_j \leftarrow \hat{\mathbf{m}}_j - n^{-1}\mathbf{1}_n^T \hat{\mathbf{m}}_j$ (centering step for identifiability)

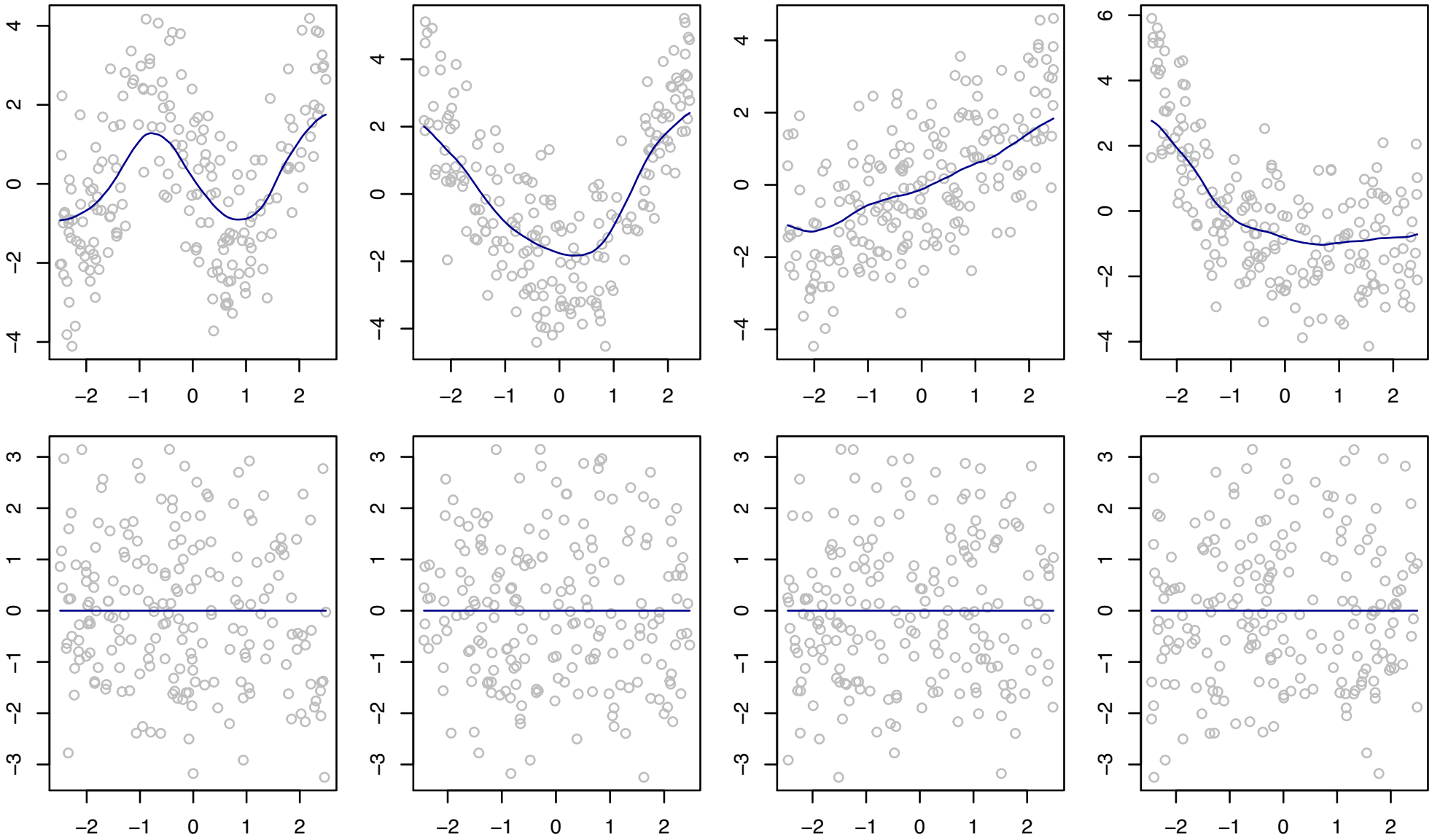
until $\hat{\mathbf{m}}_1, \ldots, \hat{\mathbf{m}}_p$ no longer change.

We can apply this to any linear smoother.

$\|\hat{m}_j\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n m_j^2(x_{ji})}$

$\|\hat{m}_j\|_n^2$ = mean of squared values of $\hat{m}_j$ at design points

Nadaraya–Watson soft–thresholded backfitting estimator




Andreas Buja, Trevor Hastie, and Robert Tibshirani.
Linear smoothers and additive models.
The Annals of Statistics, pages 453–510, 1989.


Jian Huang, Joel L Horowitz, and Fengrong Wei.
Variable selection in nonparametric additive models.
The Annals of Statistics, 38(4):2282, 2010.


Lukas Meier, Sara Van de Geer, Peter Bühlmann, et al.
High-dimensional additive modeling.
The Annals of Statistics, 37(6B):3779–3821, 2009.


Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman.
Sparse additive models.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(5):1009–1030, 2009.


Charles J Stone.
Additive regression and other nonparametric models.
The Annals of Statistics, pages 689–705, 1985.