High-dimensional additive models

the special case with natural cubic splines

(Ch. 5.3.2 in Bühlmann and van de Geer (2011))

consider the estimation problem with the SPS penalty:

$$\hat{f}_1, \ldots, \hat{f}_p = \mathrm{argmin}_{f_1,\ldots,f_p \in \mathcal{F}} \Big(\| Y - \sum_{j=1}^{p} f_j \|_n^2 + \lambda_1 \sum_{j=1}^{p} \| f_j \|_n + \lambda_2 \sum_{j=1}^{p} I(f_j) \Big)$$

where $\mathcal{F}$ = Sobolev space of functions on $[a, b]$ that are continuously differentiable with square integrable second derivatives

Proposition 5.1 in Bühlmann and van de Geer (2011)
Let $a, b \in \mathbb{R}$ such that $a < \min_{i,j}(X_i^{(j)})$ and $b > \max_{i,j}(X_i^{(j)})$. Let $\mathcal{F}$ be as above. Then, the $\hat{f}_j$'s are natural cubic splines with knots at $X_i^{(j)},\ i = 1, \ldots, n$.

implication: the optimization over functions is exactly representable as a parametric problem with dim $\approx 3np$

the optimization over functions is exactly representable as a parametric problem with

therefore:

$$f_j = H_j\beta_j,\ \ H_j \text{ from natural cubic spline basis}$$

$$\|f_j\|_n = \|H_j\beta_j\|_2/\sqrt{n} = \sqrt{\beta_j^T H_j^T H_j\beta_j}/\sqrt{n}$$

$$I(f_j) = \sqrt{\int ((H_j\beta_j)'')^2} = \sqrt{\beta_j^T \underbrace{(H_j'')^T H_j''}_{=:W_j}\beta} = \sqrt{\beta_j^T W_j\beta_j}$$

$\leadsto$ convex problem

$$\hat{\beta} = \mathrm{argmin}_\beta \left(\|Y - H\beta\|_2^2/n + \lambda_1 \sum_{j=1}^p \sqrt{\beta_j^T H_j^T H_j\beta_j/n} + \lambda_2 \sum_{j=1}^p \sqrt{\beta_j^T W_j\beta_j} \right)$$

SPS penalty of group Lasso type

for easier computation: instead of

$$\text{SPS penalty} = \lambda_1 \sum_j \|f_j\|_n + \lambda_2 \sum_j I(fj)$$

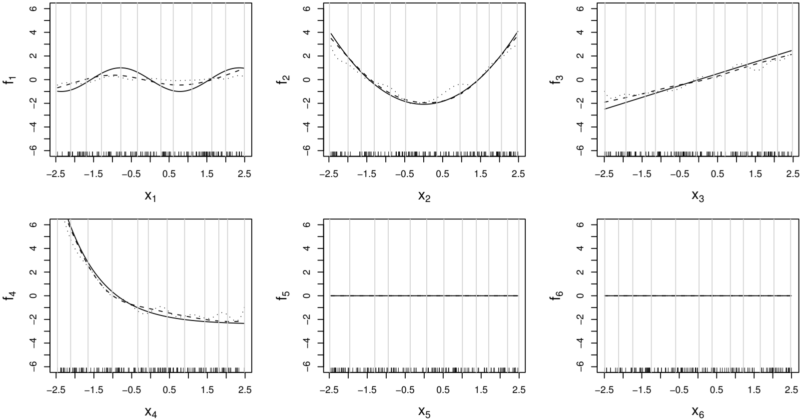
one can also use as an alternative:

$$\text{SPS Group Lasso penalty} = \lambda_1 \sum_j \sqrt{\|f_j\|_n^2 + \lambda_2^2 I^2(f_j)}$$

in parameterized form, the latter becomes:

$$\lambda_1 \sum_{j=1}^{p} \sqrt{\|H_j\beta_j\|_2^2/n + \lambda_2^2 \beta_j^T W_j \beta_j} = \lambda_1 \sum_{j=1}^{p} \sqrt{\beta_j^T (H_j^T H_j/n + \lambda_2^2 W_j)\beta_j}$$

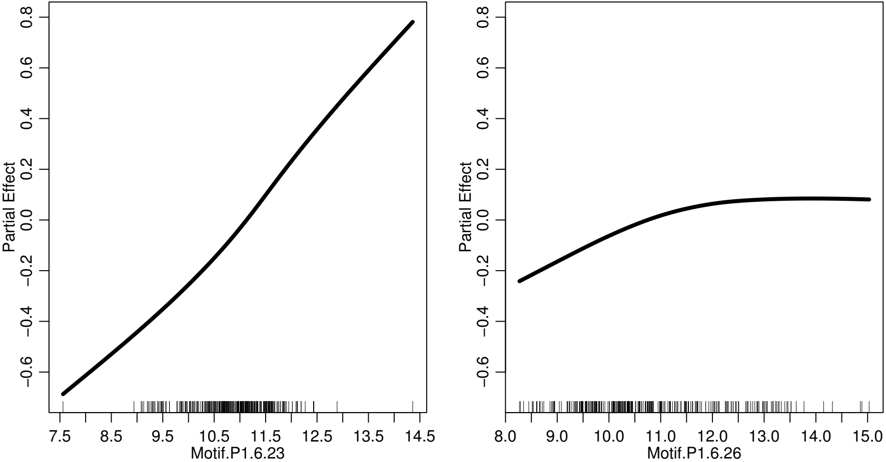
$\leadsto$ for every λ_2: a generalized Group Lasso penalty simulated example: $n = 150, p = 200$ and 4 active variables



dotted line: $\lambda_2 = 0$

$\rightsquigarrow$ λ_2 seems not so important: just consider a few candidate values (solid and dashed line)

motif regression: $n = 287, \ p = 195$



$\rightsquigarrow$ a linear model would be “fine as well”

Prediction and variable screening with additive models

(Ch. 5.6 in Bühlmann & van de Geer (2011))

most of the theory is done for SPS penalty: w.l.o.g. assume $\mu = 0$ and that each f_j^0 is twice continuously differentiable

$\hat{f}(.) = \hat{f}_{\lambda_1,\lambda_2}(.) = \sum_{j=1}^{p} \hat{f}_j(.)$

Consistency:

$$\|\hat{f} - f^0\|_n^2 = n^{-1} \sum_{i=1}^{n} |\hat{f}(X_i) - f^0(X_i)|^2 = o_P(1) \ (p \geq n \to \infty)$$

if

- Gaussian errors (for simplicity), fixed design
- $\lambda_1 \asymp n^{-2/5}, \ \lambda_2 \asymp n^{-4/5}\sqrt{\log(pn)}$ and $\log(p) = O(n^{1/5})$
- $\lambda_1 \sum_{j=1}^{p} \|f_j^0\|_n + \lambda_2 \sum_{j=1}^{p} I(f_j^0) = o(1)$

(sparsity and smoothness)

assuming in addition a compatibility-type assumption
with compatibility-type constant bounded away from zero
(and $p \gg n$):

$$\|\hat{f} - f^0\|_n^2 = O_P(s_0 \sqrt{\log(p)} n^{-4/5})$$

$$s_0 = |S_0 = \{j;\ \|f_j^0\|_n \neq 0\}| \text{ (sparsity w.r.t. additive functions)}$$

$\leadsto$ variable screening:
if for $j \in S_0$: $\|f_j^0\|_n \gg \sqrt{s_o} \log(p)^{1/4} n^{-2/5}$, then

$$\hat{S} = \{j; \|\hat{f}_j\|_n \neq 0\} \supseteq S_0 \text{ with high probability}$$

Conclusions

if the problem is sparse and smooth:
only a few $X^{(j)}$'s influence Y (only a few non-zero f_j^0) and the non-zero f_j^0 are smooth
$\rightsquigarrow$ one can often afford to model and fit additive functions in high dimensions

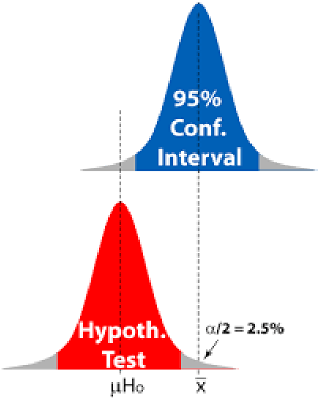
reason:

- dimensionality is of order $\mathrm{dim} = O(pn)$
  $\log(\mathrm{dim})/n = O((\log(p) + \log(n))/n)$ which is still small
- sparsity and smoothness then lead to: if each f_j^0 is twice continuously differentiable

$$\|\hat{f} - f^0\|_2^2/n = O_P(\underbrace{\text{sparsity}}_{\text{no. of non-zero } f_j^0} \sqrt{\log(p)}n^{-4/5})$$

(cf. Ch. 8.4 in Bühlmann & van de Geer (2011))

Uncertainty quantification:
p-values and confidence intervals (slides, denoted as Ch. 10)



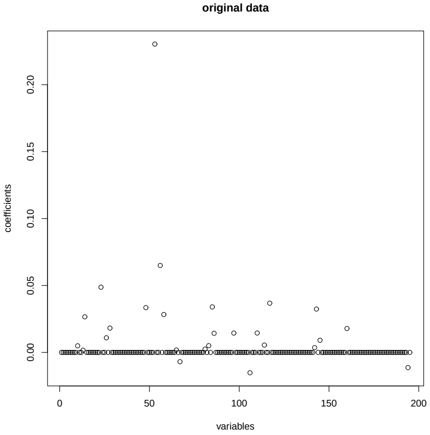
frequentist
uncertainty quantification

(in contrast to Bayesian inference)

classical concepts but in very high-dimensional settings

Toy example: Motif regression ($p = 195, n = 143$)

Lasso estimated coefficients $\hat{\beta}(\hat{\lambda}_{\mathrm{CV}})$



p-values/quantifying uncertainty would be very useful!

$$Y = X\beta^0 + \varepsilon \quad (p \gg n)$$

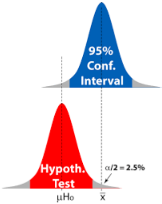
classical goal: statistical hypothesis testing

$$H_{0,j} : \beta_j^0 = 0 \text{ versus } H_{A,j} : \beta_j^0 \neq 0$$

or $$H_{0,G} : \beta_j^0 = 0 \ \forall\, j \in \underbrace{G}_{\subseteq\{1,\ldots,p\}} \text{ versus } H_{A,G} : \exists j \in G \text{ with } \beta_j^0 \neq 0$$

background: if we could handle the asymptotic distribution of the Lasso $\hat{\beta}(\lambda)$ under the null-hypothesis

$\rightsquigarrow$ could construct p-values



this is very difficult!
asymptotic distribution of $\hat{\beta}$ has some point mass at zero,...
Knight and Fu (2000) for $p < \infty$ and $n \to \infty$

because of “non-regularity” of sparse estimators
“point mass at zero” phenomenon $\rightsquigarrow$ “super-efficiency”


(Hodges, 1951)

$\rightsquigarrow$ standard bootstrapping and subsampling should not be used

$\rightsquigarrow$ de-sparsify/de-bias the Lasso instead

The de-sparsified or de-biased Lasso

Recap: if $p < n$ and $\mathrm{rank}(X) = p$, then:

$$\hat{\beta}_{\mathrm{OLS},j} = Y^T Z^{(j)} / (X^{(j)})^T Z^{(j)}$$

$$Z^{(j)} = X^{(j)} - X^{(-j)} \hat{\gamma}^{(j)}$$

$$= \text{OLS residuals from } X^{(j)} \text{ vs. } X^{(-j)} = \{X^{(k)};\ k \neq j\}$$

$$\hat{\gamma}^{(j)} = \mathrm{argmin}_{\gamma} \|X^{(j)} - X^{(-j)}\gamma\|_2^2$$

idea for high-dimensional setting:
use the Lasso for the residuals $Z^{(j)}$

The de-sparsified Lasso

consider

$$\begin{aligned} Z^{(j)} &= X^{(j)} - X^{(-j)}\hat{\gamma}^{(j)} \\ &= \text{Lasso residuals from } X^{(j)} \text{ vs. } X^{(-j)} = \{X^{(k)};\ k \neq j\} \\ \hat{\gamma}^{(j)} &= \text{argmin}_{\gamma}\|X^{(j)} - X^{(-j)}\gamma\|_2^2 + \lambda_j\|\gamma\|_1 \end{aligned}$$

build projection of Y onto $Z^{(j)}$:

$$\frac{Y^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} \underbrace{=}_{Y = X\beta^0 + \varepsilon} \beta_j^0 + \underbrace{\sum_{k \neq j} \frac{(X^{(k)})^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} \beta_k^0}_{\text{bias}} + \frac{\varepsilon^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}}$$

estimate bias and subtract it:

$$\widehat{\text{bias}} = \sum_{k \neq j} \frac{(X^{(k)})^T X^{(j)}}{(X^{(j)})^T Z^{(j)}} \quad \underbrace{\hat{\beta}_k}_{\text{standard Lasso}}$$

$\rightsquigarrow$ de-sparsified Lasso estimator

$$\hat{b}_j = \frac{Y^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} - \sum_{k \neq j} \frac{(X^{(k)})^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} \hat{\beta}_k \quad (j = 1, \ldots, p)$$

not sparse! Never equal to zero for all $j = 1, \ldots, p$

can also be represented as

$$\hat{b}_j = \underbrace{\hat{\beta}_j}_{\text{standard Lasso}} + \frac{(Y - X\hat{\beta})^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} \quad \text{"de-biased Lasso"}$$

using that

$$\frac{Y^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} = \beta_j^0 + \sum_{k \neq j} \frac{(X^{(k)})^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} \beta_k^0 + \frac{\varepsilon^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}}$$

we obtain

$$\sqrt{n}(\hat{b}_j - \beta_j^0) = \underbrace{\sqrt{n} \sum_{k \neq j} \frac{(X^{(k)})^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} (\beta_k^0 - \hat{\beta}_k)}_{\sqrt{n}\cdot \text{ (bias term of de-biased Lasso)}} + \underbrace{\sqrt{n} \frac{\varepsilon^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}}}_{\text{fluctuation term}}$$

so far, this holds for any $Z^{(j)}$

assume fixed design X, e.g. condition on X
Gaussian error $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$

fluctuation term:

$$\sqrt{n}\frac{\varepsilon^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} = \frac{n^{-1/2}\varepsilon^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}/n} \sim \mathcal{N}(0, \frac{\sigma^2 \|Z^{(j)}\|_2^2/n}{|(X^{(j)})^T Z^{(j)}/n|^2})$$

bias term of de-biased Lasso: we exploit two things

- $\|\hat{\beta} - \beta^0\|_1 = O_P(s_0\sqrt{\log(p)/n})$
- KKT condition for Lasso (on $X^{(j)}$ versus $X^{(-j)}$): $|(X^{(k)})^T Z^{(j)}/n| \leq \lambda_j/2$

therefore:

$$
\begin{aligned}
& \sqrt{n}\sum_{k\neq j} \frac{(X^{(k)})^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}}(\beta_k^0 - \hat{\beta}_k) \\
= \quad & \sqrt{n}\sum_{k\neq j} \frac{(X^{(k)})^T Z^{(j)}/n}{(X^{(j)})^T Z^{(j)}/n}(\beta_k^0 - \hat{\beta}_k) \\
\leq \quad & \sqrt{n}\max_{k\neq j} \left|\frac{(X^{(k)})^T Z^{(j)}/n}{(X^{(j)})^T Z^{(j)}/n}\right| \|\hat{\beta} - \beta^0\|_1 \\
\leq \quad & \sqrt{n}\frac{\lambda_j/2}{(X^{(j)})^T Z^{(j)}/n} O_P(s_0\sqrt{\log(p)/n}) \\
= \quad & O_P(s_0 \log(p)/\sqrt{n}) = o_P(1) \text{ if } s_0 \ll \frac{\sqrt{n}}{\log(p)}
\end{aligned}
$$

if $\lambda_j \asymp \sqrt{\log(p)/n}$ and $(X^{(j)})^T Z^{(j)}/n \asymp O(1)$

summarizing $\leadsto$
Theorem 10.1 in the notes
assume:

- $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$
- $\lambda_j = C_j \sqrt{\log(p)/n}$ and $\|Z^{(j)}\|_2^2/n \geq L > 0$
- $s_0 = o(\sqrt{n}/\log(p))$ (a bit sparse than "usual")
- $\|\hat{\beta} - \beta^0\|_1 = O_P(s_0 \sqrt{\log(p)/n})$
(i.e., compatibility constant ϕ_o^2 bounded away from zero)

Then:

$$\sigma^{-1}\sqrt{n}\frac{(X^{(j)})^T Z^{(j)}/n}{\|Z^{(j)}\|_2/\sqrt{n}}(\hat{b}_j - \beta_j^0) \Longrightarrow \mathcal{N}(0,1) \quad (j = 1, \ldots, p)$$

more precisely:

$$\sigma^{-1}\sqrt{n}\frac{(X^{(j)})^T Z^{(j)}/n}{\|Z^{(j)}\|_2/\sqrt{n}}(\hat{b}_j - \beta_j^0) = W_j + \Delta_j$$

$$(W_1, \ldots, W_p)^T \sim \mathcal{N}_p(0, \sigma^2\Omega), \ \max_{j=1,\ldots,p} |\Delta_j| = o_P(1)$$

confidence intervals for β_j^0:

$$\hat{b}_j \pm \hat{\sigma} n^{-1/2} \frac{\|Z^{(j)}\|_2/\sqrt{n}}{|(X^{(j)})^T Z^{(j)}/n} \Phi^{-1}(1 - \alpha/2)$$

$\hat{\sigma}^2 = \|Y - X\hat{\beta}\|_2^2/n$ or $\hat{\sigma}^2 = \|Y - X\hat{\beta}\|_2^2/(n - \|\hat{\beta}\|_0^0)$

can also test

$$H_{0,j}: \ \beta^0_j = 0 \text{ versus } H_{A,j}: \ \beta^0_j \neq 0$$

can also test group hypothesis: for $G \subseteq \{1, \ldots, p\}$

$$H_{0,G}: \ \beta^0_j \equiv 0 \forall j \in G$$
$$H_{A,G}: \exists j \in G \text{ such that } \beta^0_j \neq 0$$

under $H_{0,G}$:

$$\max_{j \in G} \sigma^{-1} \sqrt{n} \frac{|(X^{(j)})^T Z^{(j)}/n|}{\|Z^{(j)}\|_2/\sqrt{n}} |\hat{b}_j| = \max_{j \in G} |W_j + \Delta_j| \asymp \underbrace{\max_{j \in G} |W_j|}_{\text{distr. simulated}}$$

and plug-in $\hat{\sigma}$ for σ

Choice of tuning parameters

as usual: $\hat{\beta} = \hat{\beta}(\hat{\lambda}_{\text{CV}})$; what is the role of λ_j?

$$\text{variance} = \sigma^2 n^{-1} \frac{\|Z^{(j)}\|_2^2/n}{|(X^{(j)})^T Z^{(j)}/n|^2} \asymp \sigma^2/\|Z^{(j)}\|_2^2$$

if $\lambda_j \searrow$ then $\|Z^{(j)}\|_2^2 \searrow$, i.e. large variance

error due to bias estimation is bounded by:

$$|\ldots| \le \sqrt{n} \frac{\lambda_j/2}{|(X^{(j)})^T Z^{(j)}/n|} \|\hat{\beta} - \beta^0\|_1 \propto \lambda_j$$

assuming λ_j is not too small
if $\lambda_j \searrow$ (but not too small) then bias estimation error $\searrow$

$\leadsto$ inflate the variance a bit to have low error due to bias estimation: control type I error at the price of slightly decreasing power

How good is the de-biased Lasso?

asymptotic efficiency:
for the de-biased Lasso to “work” we require

- sparsity: $s_0 = o(\sqrt{n}/\log(p))$
  this cannot be beaten in a minimax sense
- compatibility condition for X

for optimality in terms of the lowest possible asymptotic variance achieving the “Cramer-Rao” lower bound:

- require in addition that $X^{(j)}$ versus $X^{(-j)}$ is sparse:
  $s_j \ll n/\log(p)$

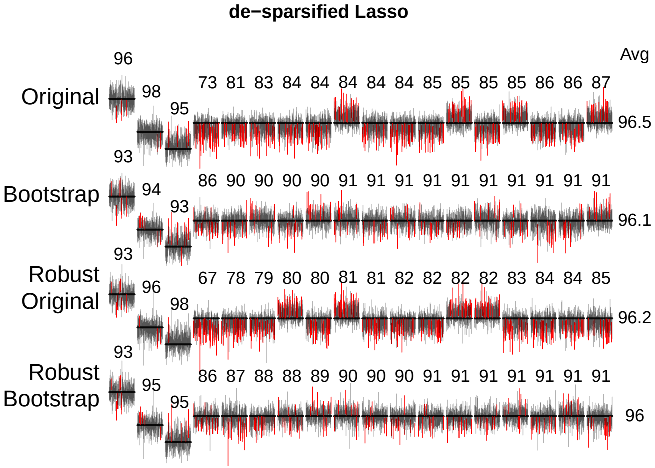
then... skipping details, the de-biased Lasso achieves (see Theorem 10.2):

$$\sqrt{n}(\hat{b}_j - \beta_j^0) \Longrightarrow \mathcal{N}(0, \underbrace{\sigma^2 \Theta_{jj}}_{\text{Cramer-Rao lower bound}})$$

$\Theta = \Sigma_X^{-1} = \mathrm{Cov}(X)^{-1} \rightsquigarrow$ as for OLS in low dimensions!

Empirical results

R-software hdi



black: confidence interval covered the true coefficient
red: confidence interval failed to cover