Theoretical guarantees for Group Lasso

follows "similarly" but with more complicated arguments than for the Lasso

Algorithm for Group Lasso

consider the KKT conditions for the objective function

$$Q_\lambda(\beta) = \underbrace{n^{-1}\sum_{i=1}^{n}\rho_\beta(X_i, Y_i)}_{\text{e.g. } \|Y - X\beta\|_2^2/n} + \lambda\sum_{j=1}^{q} m_j\|\beta_{\mathcal{G}_j}\|_2$$

Lemma (Lemma 4.3 in Bühlmann and van de Geer (2011))
Assume $\rho_\beta = n^{-1}\sum_{i=1}^{n}\rho_\beta(X_i, Y_i)$ is differentiable and convex (in β). Then, a necessary and sufficient condition for $\hat{\beta}$ to be a solution is

$$\nabla\rho(\hat{\beta})_{\mathcal{G}_j} = -\lambda m_j \frac{\hat{\beta}_{\mathcal{G}_j}}{\|\hat{\beta}_{\mathcal{G}_j}\|_2} \qquad \text{if } \hat{\beta}_{\mathcal{G}_j} \not\equiv 0,$$

$$\|\nabla\rho(\hat{\beta})_{\mathcal{G}_j}\|_2 \le \lambda m_j \qquad \text{if } \hat{\beta}_{\mathcal{G}_j} \equiv 0$$

block coordinate descent

Algorithm 1 Block Coordinate Descent Algorithm

Let $\beta^{[0]} \in \mathbb{R}^p$ be an initial parameter vector. Set $m = 0$.

repeat

Increase m by one: $m \leftarrow m + 1$.
Denote by $\mathscr{S}^{[m]}$ the index cycling through the block coordinates $\{1, \ldots, q\}$:
$\mathscr{S}^{[m]} = \mathscr{S}^{[m-1]} + 1 \bmod q$. Abbreviate by $j = \mathscr{S}^{[m]}$ the value of $\mathscr{S}^{[m]}$.

if $\|(-\nabla\rho(\beta^{[m-1]}_{-\mathscr{G}_j})_{\mathscr{G}_j}\|_2 \leq \lambda m_j :$ set $\beta^{[m]}_{\mathscr{G}_j} = 0$,
otherwise: $\beta^{[m]}_{\mathscr{G}_j} = \arg\min_{\beta_{\mathscr{G}_j}} Q_\lambda(\beta^{[m-1]}_{+\mathscr{G}_j})$,

where $\beta^{[m-1]}_{-\mathscr{G}_j}$ is defined in (4.14) and $\beta^{[m-1]}_{+\mathscr{G}_j}$ is the parameter vector which equals $\beta^{[m-1]}$ except for the components corresponding to group $\mathscr{G}_j$ whose entries are equal to $\beta_{\mathscr{G}_j}$ (i.e. the argument we minimize over).

until numerical convergence

block-updates where the blocks correspond to the groups

The generalized Group Lasso penalty

Chapter 4.5 in Bühlmann and van de Geer (2011)

$$\text{pen}(\beta) = \lambda \sum_{j=1}^{q} m_j \sqrt{\beta_{\mathcal{G}_j}^T A_j \beta_{\mathcal{G}_j}},$$

A_j positive definite

can do the computation with standard group Lasso by transformation:

$$\tilde{\beta}_{\mathcal{G}_j} = A_j^{1/2} \beta_{\mathcal{G}_j} \rightsquigarrow \text{pen}(\tilde{\beta}) = \lambda \sum_{j=1}^{q} m_j \|\tilde{\beta}_{\mathcal{G}_j}\|_2$$

$$X\beta = \sum_{j=1}^{q} \tilde{X}_{\mathcal{G}_j} \tilde{\beta}_{\mathcal{G}_j} =: \tilde{X}\tilde{\beta},\ \tilde{X}_{\mathcal{G}_j} = X_{\mathcal{G}_j} A_j^{-1/2}$$

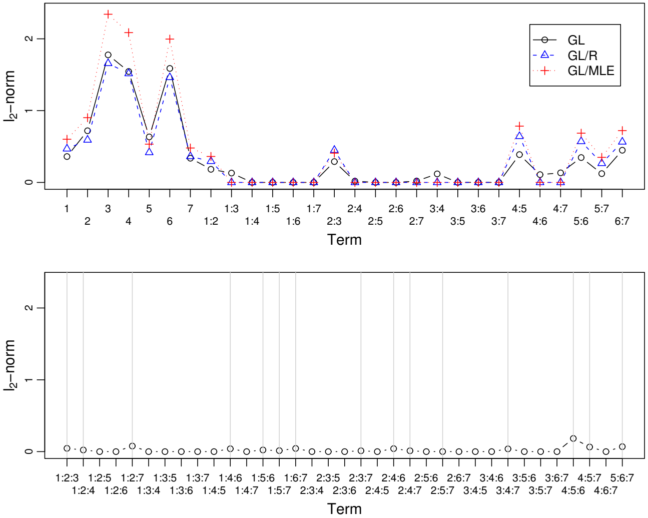
can simply solve the “tilde” problem: $\rightsquigarrow \hat{\tilde{\beta}} \rightsquigarrow \hat{\beta}_{\mathcal{G}_j} = A_j^{-1/2} \hat{\tilde{\beta}}_{\mathcal{G}_j}$

special but important case: groupwise prediction penalty

$$\mathrm{pen}(\beta) = \lambda \sum_{j=1}^{q} m_j \|X_{\mathcal{G}_j}\beta_{\mathcal{G}_j}\|_2 = \lambda \sum_{j=1}^{q} m_j \sqrt{\beta_{\mathcal{G}_j}^T X_{\mathcal{G}_j}^T X_{\mathcal{G}_j}\beta_{\mathcal{G}_j}}$$

$X_{\mathcal{G}_j}^T X_{\mathcal{G}_j}$ typically positive definite for $|\mathcal{G}_j| < n$

- penalty is invariant under arbitrary reparameterizations within every group $\mathcal{G}_j$: important!
- when using an orthogonal parameterization such that $X_{\mathcal{G}_j}^T X_{\mathcal{G}_j} = I$: it is the standard Group Lasso
  with categorical variables: this is in fact what one has in mind (can use groupwise orthogonalized design) or one should use the groupwise prediction penalty



is with groupwise orthogonalized design matrices

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 wit the SSP penalty:

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

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 $\approx 3np$ (namely cubic splines) the optimization over functions is exactly representable as a parametric problem (with ciubic splines)

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

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

SSP penalty of group Lasso type

for easier computation: instead of

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

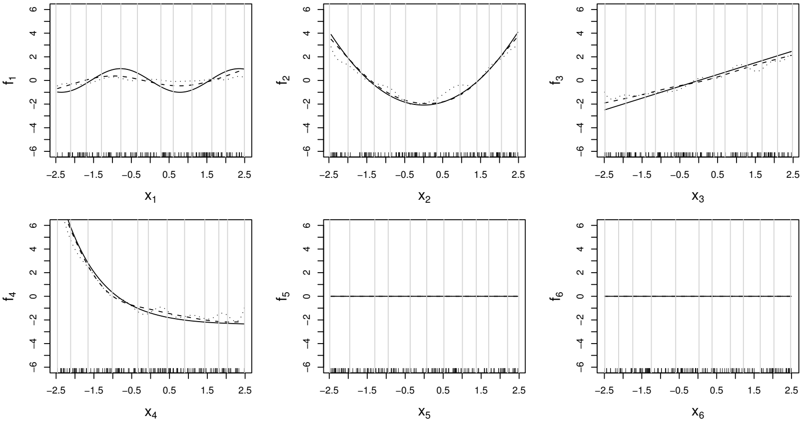
one can also use as an alternative:

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

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

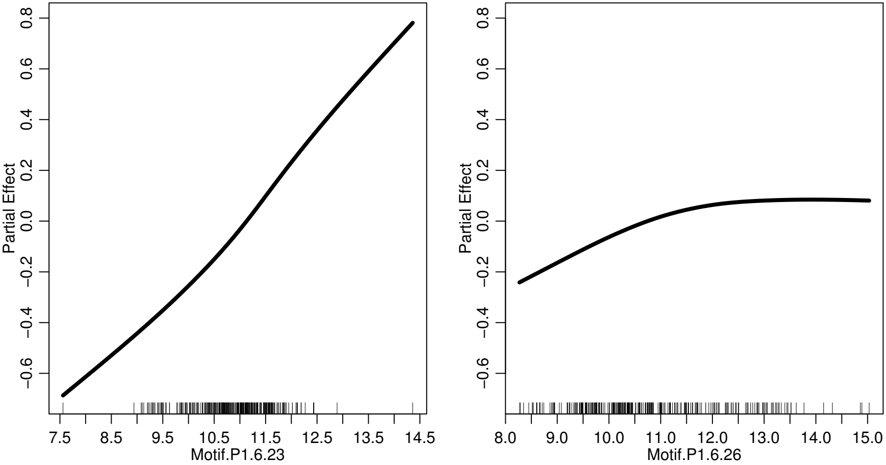


dotted line: $\lambda_2 = 0$

$\rightsquigarrow \lambda_2$ 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"

Theoretical properties of high-dimensional additive models

- prediction and function estimation:
  compatibility-type assumption for the functions f_j^0
- screening property:
  beta-min analogue assumption for non-zero functions f_j^0

see Chapters 5.6 and 8.4 in Bühlmann and van de Geer (2011)

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
$\leadsto$ 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))