Recap: Group Lasso

partition the parameter into groups $\mathcal{G}_j$:

$$\bigcup_{j=1}^{q} \mathcal{G}_j = \{1, \ldots, p\},\ \mathcal{G}_j \cap \mathcal{G}_k = \emptyset\ (j \neq k)$$

Group Lasso for linear models:

$$\hat{\beta}(\lambda) = \text{argmin}_\beta \left(\|Y - X\beta\|_2^2/n + \lambda \sum_{j=1}^{q} m_j \|\beta_{\mathcal{G}_j}\|_2 \right)$$

typically $m_j = \sqrt{T_j},\ T_j = |\mathcal{G}_j|$

fact: the Group Lasso estimator is sparse w.r.t. groups

$$\begin{array}{rl} \text{either} & \hat{\beta}_{\mathcal{G}_j} \equiv 0, \text{ i.e. } (\hat{\beta}_{\mathcal{G}_j})_r = 0 \ \forall r \in \mathcal{G}_j \\ \text{or} & (\hat{\beta}_{\mathcal{G}_j})_r \neq 0 \ \forall r \in \mathcal{G}_j \end{array}$$

reason: sparsity occurs at points of non-stationarity

Group Lasso for categorical variables

Parameterization of model matrix

example: 4 levels, $p = 2$ variables

main effects only

```
> xx1
[1] 0 1 2 3 3 2 1 0
Levels: 0 1 2 3
> xx2
[1] 3 3 2 2 1 1 0 0
Levels: 0 1 2 3

> model.matrix(~xx1+xx2,
contrasts=list(xx1="contr.sum",xx2="contr.sum"))
  (Intercept) xx11 xx12 xx13 xx21 xx22 xx23
1           1    1    0    0   -1   -1   -1
2           1    0    1    0   -1   -1   -1
3           1    0    0    1    0    0    1
4           1   -1   -1   -1    0    0    1
5           1   -1   -1   -1    0    1    0
6           1    0    0    1    0    1    0
7           1    0    1    0    1    0    0
8           1    1    0    0    1    0    0
attr(,"assign")
[1] 0 1 1 1 2 2 2
attr(,"contrasts")
attr(,"contrasts")$xx1
[1] "contr.sum"

attr(,"contrasts")$xx2
[1] "contr.sum"
```

with interaction terms

```
> model.matrix(~xx1*xx2,
contrasts=list(xx1="contr.sum",xx2="contr.sum"))
  (Intercept) xx11 xx12 xx13 xx21 xx22 xx23 xx11:xx21 xx12:xx21 xx13:xx21
1           1    1    0    0   -1   -1   -1        -1         0         0
2           1    0    1    0   -1   -1   -1         0        -1         0
3           1    0    0    1    0    0    1         0         0         0
4           1   -1   -1   -1    0    0    1         0         0         0
5           1   -1   -1   -1    0    1    0         0         0         0
6           1    0    0    1    0    1    0         0         0         0
7           1    0    1    0    1    0    0         0         1         0
8           1    1    0    0    1    0    0         1         0         0
  xx11:xx22 xx12:xx22 xx13:xx22 xx11:xx23 xx12:xx23 xx13:xx23
1        -1         0         0        -1         0         0
2         0        -1         0         0        -1         0
3         0         0         0         0         0         1
4         0         0         0        -1        -1        -1
5        -1        -1        -1         0         0         0
6         0         0         1         0         0         0
7         0         0         0         0         0         0
8         0         0         0         0         0         0
attr(,"assign")
 [1] 0 1 1 1 2 2 2 3 3 3 3 3 3 3 3 3
attr(,"contrasts")
attr(,"contrasts")$xx1
[1] "contr.sum"

attr(,"contrasts")$xx2
[1] "contr.sum"
```

Prediction of DNA splice sites (Ch. 4.3.1 in Bühlmann and van de Geer (2011))

want to predict donor splice site where coding and non-coding regions in DNA start/end

$$\underbrace{\cdot\ \cdot\ \cdot\ \cdot}_{\text{exon: coding}} \quad GT \quad \underbrace{\cdot\ \cdot\ \cdot}_{\text{intron: non-coding}}$$

seven positions around "GT"

training data:

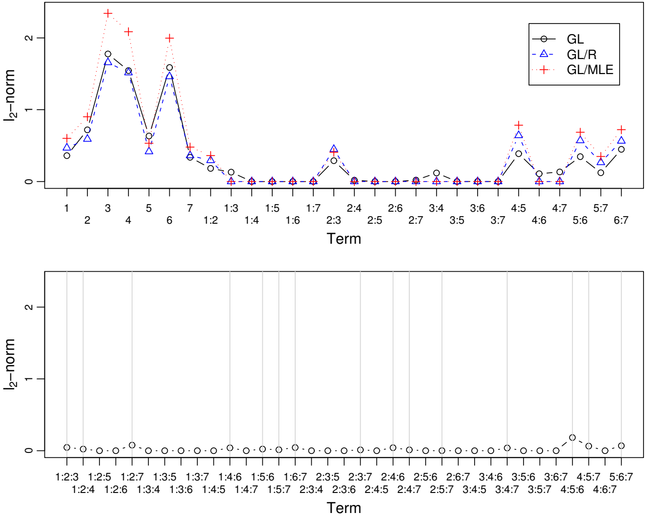
$$Y_i \in \{0, 1\} \text{ true donor site or not}$$
$$X_i \in \{A, C, G, T\}^7 \text{ positions}$$
$$i = 1, \ldots, \approx 188'000$$

unbalanced: $Y_i = 1$: 8415; $Y_i = 0$: 179'438

model: logistic linear regression model with intercept, main effects and interactions up to order 2 (3 variables interact)
$\leadsto$ dimension = 1155

methods:

- Group Lasso
- MLE on $\hat{S} = \{j;\ \hat{\beta}_{\mathcal{G}_j} \neq 0\}$
- as above but with Ridge regularized MLE on $\hat{S}$



mainly main effects (quite debated in computational biology...)

Theoretical guarantees for Group Lasso

follows “similarly” but with more complicated arguments as 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

$$\begin{aligned} \nabla\rho(\hat{\beta})_{\mathcal{G}_j} + \lambda m_j \frac{\hat{\beta}_{\mathcal{G}_j}}{\|\hat{\beta}_{\mathcal{G}_j}\|_2} \qquad & \text{if } \hat{\beta}_{\mathcal{G}_j} \not\equiv 0, \\ \|\nabla\rho(\hat{\beta})_{\mathcal{G}_j}\|_2 \leq \lambda m_j \qquad & \text{if } \hat{\beta}_{\mathcal{G}_j} \equiv 0 \end{aligned}$$

block coordinate descent

Algorithm 1 Block Coordinate Descent Algorithm

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

2: **repeat**

3: 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 $\mathscr{S}^{[m]}$.

4: if $\|(-\nabla\rho(\beta^{[m-1]}_{-\mathscr{G}_j})_{\mathscr{G}_j}\|_2 \le \lambda m_j$: 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).

5: **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:

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

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

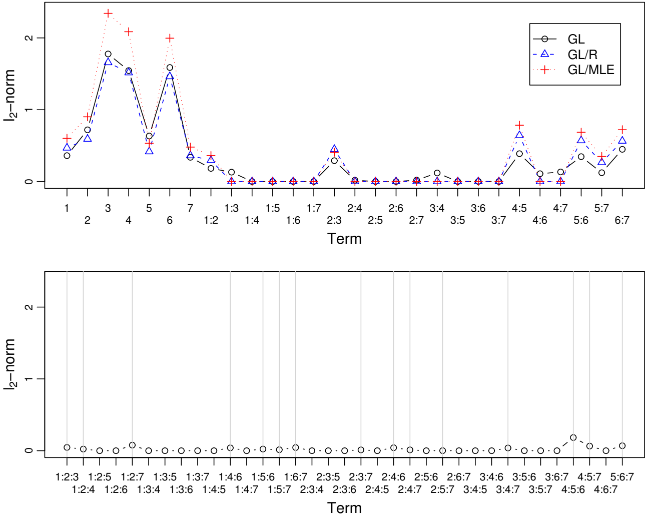
can simply solve the "tilde" problem: $\leadsto \hat{\tilde{\beta}} \leadsto \hat{\beta}_{\mathcal{G}_j} = A_j^{-1/2} \hat{\tilde{\beta}}_{\mathcal{G}_j}$

special but important case: groupwise prediction penalty

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

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

- penalty is invariant under arbitrary reparameterizations withing 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:

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

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 $\approx 3np$

SSP penalty of group Lasso type

for easier computation: instead of

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

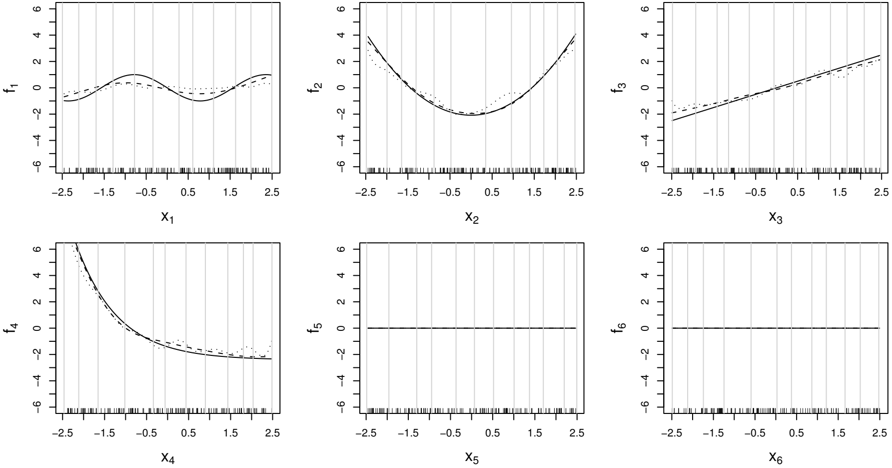
one can also use as an alternative:

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

in parameterized form, the latter becomes:

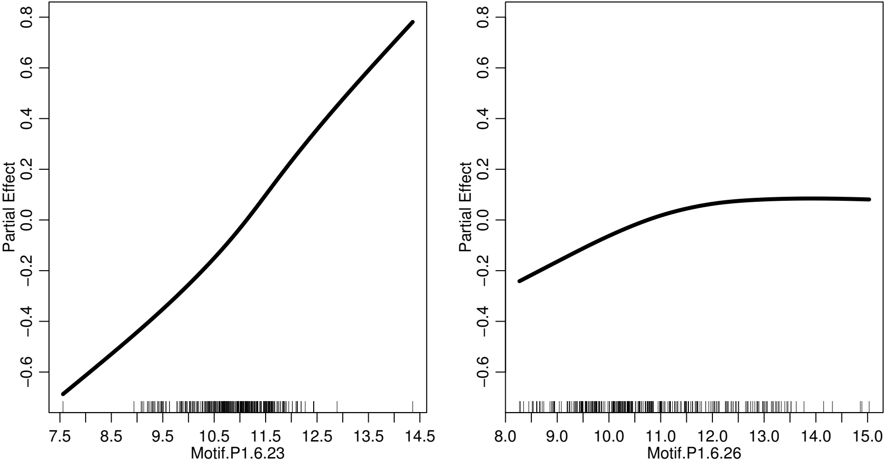
$$\lambda_1 \sum_{j=1}^{p} \sqrt{\|H_j \beta_j\|_2^2 / n + \lambda_2^2 \beta_j^T W_j \beta_j} = \lambda_1 \sum_{j=1}^{p} \sqrt{\beta_j^T (H_j^T H_j / n + \lambda_2^2 W_j) \beta_j}$$

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



$\rightsquigarrow$ a linear model would be "fine as well"