Statistics for high-dimensional data: Introduction, and the Lasso for linear models

Peter Bühlmann and Sara van de Geer

Seminar für Statistik, ETH Zürich

May 2012

High-dimensional data

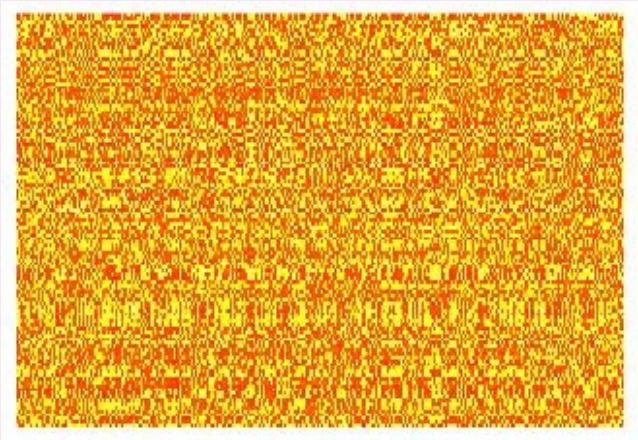
Riboflavin production with Bacillus Subtilis

(in collaboration with DSM (Switzerland))

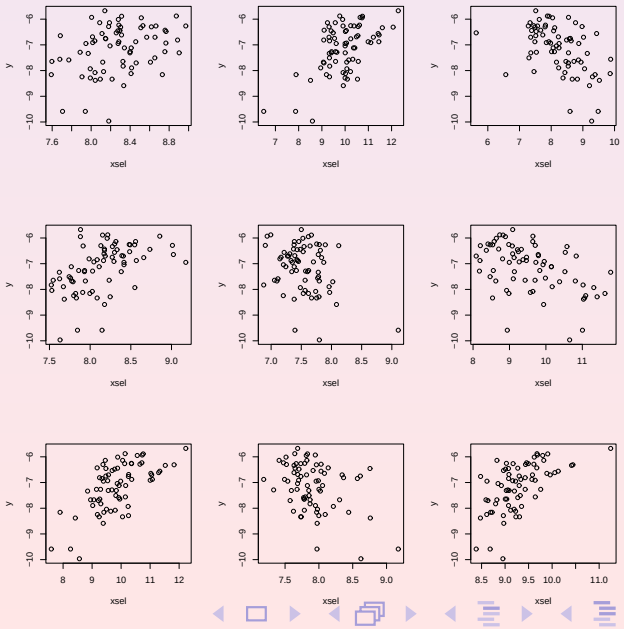
goal: improve riboflavin production rate of Bacillus Subtilis
using clever genetic engineering

response variables $Y \in \mathbb{R}$: riboflavin (log-) production rate
covariates $X \in \mathbb{R}^p$: expressions from $p = 4088$ genes
sample size $n = 115$, $p \gg n$

gene expression data



Y versus 9 “reasonable” genes



general framework:

$$Z_1, \ldots, Z_n \text{ (with some "i.i.d. components")}$$
$$\dim(Z_i) \gg n$$

for example:
$Z_i = (X_i, Y_i),\ X_i \in \mathcal{X} \subseteq \mathbb{R}^p, Y_i \in \mathcal{Y} \subseteq\in \mathbb{R}$: regression with $p \gg n$
$Z_i = (X_i, Y_i),\ X_i \in \mathcal{X} \subseteq\in \mathbb{R}^p, Y_i \in \{0, 1\}$: classification for $p \gg n$

numerous applications:
biology, imaging, economy, environmental sciences, ...

High-dimensional linear models

$$Y_i = \sum_{j=1}^{p} \beta_j^0 X_i^{(j)} + \varepsilon_i, \; i = 1, \ldots, n$$

$p \gg n$

in short: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$

goals:

- prediction, e.g. w.r.t. squared prediction error
- estimation of β^0, e.g. w.r.t. $\|\hat{\beta} - \beta^0\|_q$ $(q = 1, 2)$
- variable selection
  i.e. estimating the active set with the effective variables
  (having corresponding coefficient $\neq 0$)

we need to regularize...

and there are many proposals

- Bayesian methods for regularization
- greedy algorithms: aka forward selection or boosting
- preliminary dimension reduction
- ...

e.g. 3'190'000 entries on Google Scholar for
"high dimensional linear model" ...

we need to regularize...
and there are many proposals

- Bayesian methods for regularization
- greedy algorithms: aka forward selection or boosting
- preliminary dimension reduction
- ...

e.g. 3'190'000 entries on Google Scholar for
"high dimensional linear model" ...

we need to regularize...
and there are many proposals

- Bayesian methods for regularization
- greedy algorithms: aka forward selection or boosting
- preliminary dimension reduction
- ...

e.g. 3'190'000 entries on Google Scholar for
"high dimensional linear model" ...

Penalty-based methods

if true β^0 is sparse w.r.t.

- $\|\beta^0\|_0^0 =$ number of non-zero coefficients
  $\rightsquigarrow$ regularize with the $\|\cdot\|_0$-penalty:
  $\text{argmin}_\beta(n^{-1}\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|_0^0)$, e.g. AIC, BIC
  $\rightsquigarrow$ computationally infeasible if p is large (2^p sub-models)

- $\|\beta^0\|_1 = \sum_{j=1}^p |\beta_j^0|$
  $\rightsquigarrow$ penalize with the $\|\cdot\|_1$-norm, i.e. Lasso:
  $\text{argmin}_\beta(n^{-1}\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|_1)$
  $\rightsquigarrow$ convex optimization:
  computationally feasible and very fast for large p

The Lasso (Tibshirani, 1996)

Lasso for linear models

$$\hat{\beta}(\lambda) = \text{argmin}_\beta(n^{-1}\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \underbrace{\lambda}_{\geq 0} \underbrace{\|\beta\|_1}_{\sum_{j=1}^p |\beta_j|})$$

$\rightsquigarrow$ convex optimization problem

- Lasso does variable selection
  some of the $\hat{\beta}_j(\lambda) = 0$
  (because of "ℓ_1-geometry")
- $\hat{\beta}(\lambda)$ is a shrunken LS-estimate

more about "ℓ_1-geometry"

equivalence to primal problem

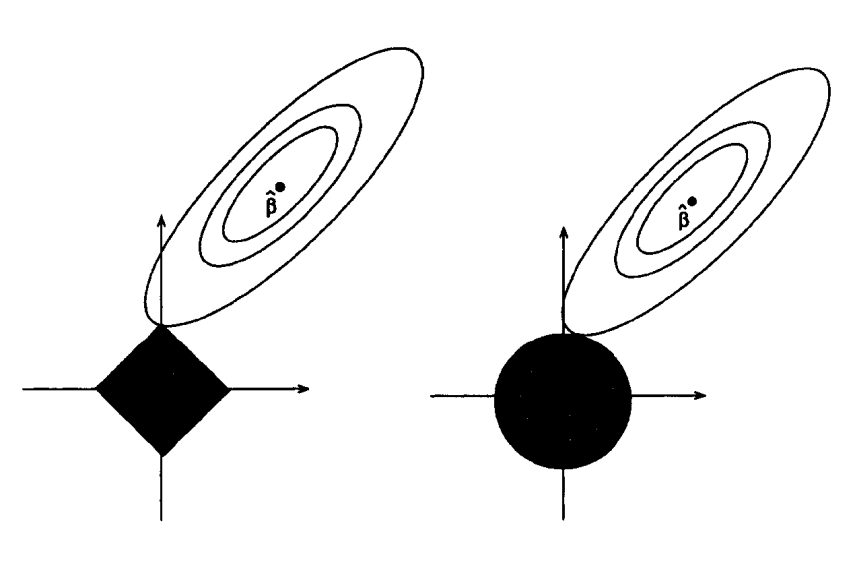
$$\hat{\beta}_{\text{primal}}(R) = \text{argmin}_{\beta;\|\beta\|_1 \le R} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n,$$

with a correspondence between λ and R which depends on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$
[such an equivalence holds since

- $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n$ is convex in β
- convex constraint $\|\beta\|_1 \le R$

see e.g. Bertsekas (1995)]

p=2



left: ℓ_1-"world"
residual sum of squares reaches a minimal value (for certain constellations of the data) if its contour lines hit the ℓ_1-ball in its corner
$\rightsquigarrow \ \hat{\beta}_1 = 0$

ℓ_2-"world" is different

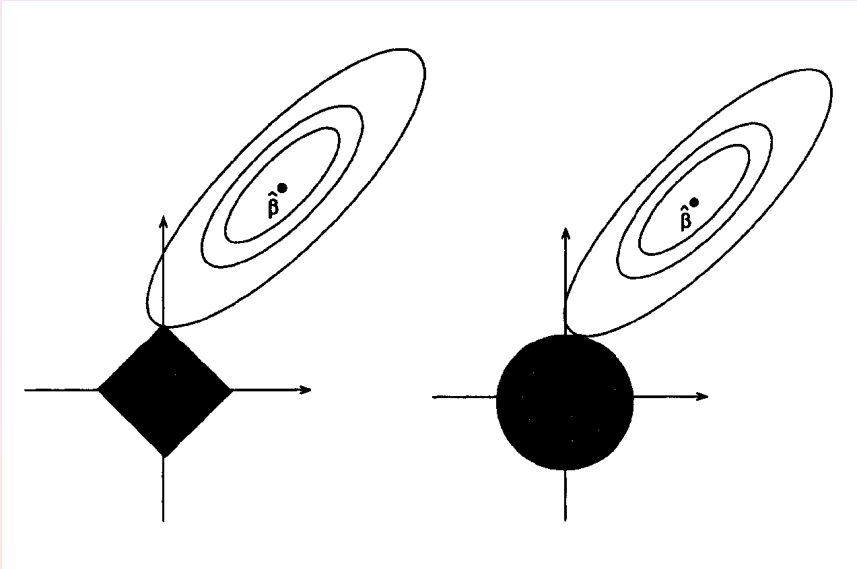
Ridge regression,

$$\hat{\beta}_{\text{Ridge}}(\lambda) = \text{argmin}_\beta \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_2^2 \right),$$

equivalent primal equivalent solution

$$\hat{\beta}_{\text{Ridge;primal}}(R) = \text{argmin}_{\beta;\|\beta\|_2 \le R} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n,$$

with a one-to-one correspondence between λ and R



A note on the Bayesian approach

model:

$$\beta_1, \ldots, \beta_p \text{ i.i.d. } \sim p(\beta)d\beta,$$
$$\text{given } \beta: \ \mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 I_n) \text{ with density } f(\mathbf{y}|\sigma^2, \beta)$$

posterior density:

$$p(\beta|\mathbf{Y}, \sigma^2) = \frac{f(\mathbf{Y}|\beta, \sigma^2)p(\beta)}{\int f(\mathbf{Y}|\beta, \sigma^2)p(\beta)d\beta} \propto f(\mathbf{Y}|\beta, \sigma^2)p(\beta)$$

and hence for the MAP (Maximum A-Posteriori) estimator:

$$\begin{aligned}\hat{\beta}_{\text{MAP}} &= \text{argmax}_\beta p(\beta|\mathbf{Y}, \sigma^2) = \text{argmin}_\beta - \log\left(f(\mathbf{Y}|\beta, \sigma^2)p(\beta)\right) \\ &= \text{argmin}_\beta \left(\frac{1}{2\sigma^2}\|\mathbf{Y} - X\beta\|_2^2 - \sum_{j=1}^{p} \log(p(\beta_j))\right)\end{aligned}$$

examples:
1. Double-Exponential prior DExp(ξ):
$p(\beta) = \frac{\tau}{2} \exp(-\tau\beta)$
$\rightsquigarrow \hat{\beta}_{\text{MAP}}$ equals the Lasso with penalty parameter $\lambda = n^{-1} 2\sigma^2\tau$

2. Gaussian prior $\mathcal{N}(0, \tau^2)$:
$p(\beta) = \frac{1}{\sqrt{2\pi}\tau} \exp(-\beta^2/(2\tau^2))$
$\rightsquigarrow \hat{\beta}_{\text{MAP}}$ equals the Ridge estimator with penalty parameter
$\lambda = n^{-1}\sigma^2/\tau^2$

but we will argue that Lasso is also good if the truth is sparse with respect to $\|\beta^0\|_0^0$, e.g. if prior is (much) more spiky around zero than Double-Exponential distribution

Orthonormal design

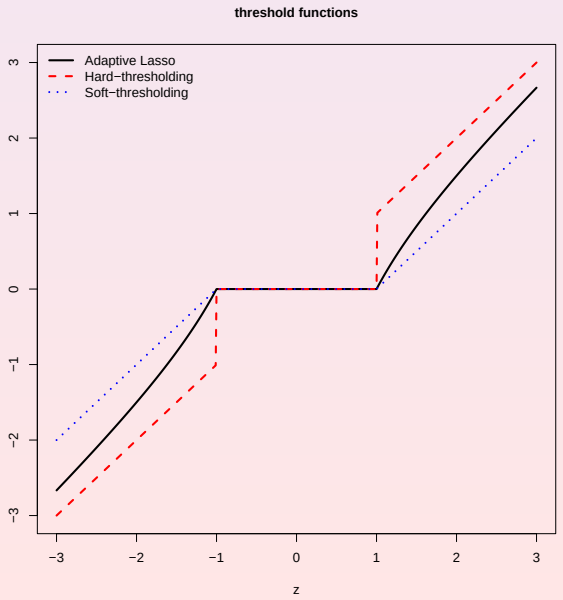
$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad n^{-1}\mathbf{X}^T\mathbf{X} = I$$

Lasso = soft-thresholding estimator

$$\hat{\beta}_j(\lambda) = \mathrm{sign}(Z_j)(|Z_j| - \lambda/2)_+, \quad \underbrace{Z_j}_{=\mathrm{OLS}} = (n^{-1}\mathbf{X}^T\mathbf{Y})_j,$$

$$\hat{\beta}_j(\lambda) = g_{\mathrm{soft}}(Z_j),$$

[this is Exercise 2.1]



Prediction

goal: predict a new observation Y_{new}
consider expected (w.r.t. new data; and random X) squared error loss:

$$\begin{aligned} & \mathbb{E}_{X_{\text{new}}, Y_{\text{new}}}[(Y_{\text{new}} - X_{\text{new}}\hat{\beta})^2] = \sigma^2 + \mathbb{E}_{X_{\text{new}}}[(X_{\text{new}}(\beta^0 - \hat{\beta}))^2] \\ = \quad & \sigma^2 + (\hat{\beta} - \beta^0)^T \underbrace{\Sigma}_{\text{Cov}(X)} (\hat{\beta} - \beta^0) \end{aligned}$$

$\rightsquigarrow$ terminology "prediction error":

for random design $\mathbf{X}$: $(\hat{\beta} - \beta^0)^T \Sigma (\hat{\beta} - \beta^0) = \mathbb{E}_{X_{\text{new}}}[(X_{\text{new}}(\hat{\beta} - \beta^0))^2]$
for fixed design $\mathbf{X}$: $(\hat{\beta} - \beta^0)^T \hat{\Sigma} (\hat{\beta} - \beta^0) = \|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2 / n$

binary lymph node classification using gene expressions:
a high noise problem
$n = 49$ samples, $p = 7130$ gene expressions

despite that it is classification: $\mathbb{P}[Y = 1|X = x] = \mathbb{E}[Y|X = x]$
$\rightsquigarrow$ $\hat{p}(x)$ via linear model; can then do classification

cross-validated misclassification error (2/3 training; 1/3 test)

Lasso	L_2Boosting	FPLR	Pelora	1-NN	DLDA	SVM
21.1%	17.7%	35.25%	27.8%	43.25%	36.12%	36.88%

with variable selection

best 200 genes (Wilcoxon test)
no additional variable selection

Lasso selected on CV-average 13.12 out of $p = 7130$ genes

from a practical perspective:
if you trust in cross-validation: can validate how good we are
i.e. prediction may be a black box, but we can evaluate it!

binary lymph node classification using gene expressions:
a high noise problem
$n = 49$ samples, $p = 7130$ gene expressions

despite that it is classification: $\mathbb{P}[Y = 1|X = x] = \mathbb{E}[Y|X = x]$
$\rightsquigarrow$ $\hat{p}(x)$ via linear model; can then do classification

cross-validated misclassification error (2/3 training; 1/3 test)

Lasso	L_2Boosting	FPLR	Pelora	1-NN	DLDA	SVM
21.1%	17.7%	35.25%	27.8%	43.25%	36.12%	36.88%

with variable selection

best 200 genes (Wilcoxon test)
no additional variable selection

Lasso selected on CV-average 13.12 out of $p = 7130$ genes

from a practical perspective:
if you trust in cross-validation: can validate how good we are
i.e. prediction may be a black box, but we can evaluate it!

and in fact: we will hear that

- Lasso is consistent for prediction assuming “essentially nothing”
- Lasso is optimal for prediction assuming the “compatibility condition” for $\mathbf{X}$

Estimation of regression coefficients

$$\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon, \ p \gg n$$

with fixed (deterministic) design $\mathbf{X}$

problem of identifiability:
for $p > n$: $\mathbf{X}\beta^0 = X\theta$
for any $\theta = \beta^0 + \xi$, ξ in the null-space of $\mathbf{X}$

$\rightsquigarrow$ cannot say anything about $\|\hat{\beta} - \beta^0\|$ without further assumptions!
$\rightsquigarrow$ we will work with the compatibility assumption (see later by Sara)
and Sara will explain: under compatibility condition

$$\|\hat{\beta} - \beta^0\|_1 \leq C\frac{s_0}{\phi_0^2}\sqrt{\log(p)/n},$$

$$s_0 = |\mathrm{supp}(\beta^0)| = |\{j; \ \beta_j^0 \neq 0\}|$$

Variable selection

Example: Motif regression
for finding HIF1α transcription factor binding sites in DNA seq.
Müller, Meier, PB & Ricci

$Y_i \in \mathbb{R}$: univariate response measuring binding intensity of HIF1α on coarse DNA segment i (from CHIP-chip experiments)
$X_i = (X_i^{(1)}, \ldots, X_i^{(p)}) \in \mathbb{R}^p$:
$X_i^{(j)}$ = abundance score of candidate motif j in DNA segment i (using sequence data and computational biology algorithms, e.g. MDSCAN)

question: relation between the binding intensity Y and the abundance of short candidate motifs?

$\rightsquigarrow$ linear model is often reasonable
"motif regression" (Conlon, X.S. Liu, Lieb & J.S. Liu, 2003)

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \; n = 287, \; p = 195$$

goal: variable selection
$\rightsquigarrow$ find the relevant motifs among the $p = 195$ candidates

Lasso for variable selection

$$\hat{S}(\lambda) = \{j;\ \hat{\beta}_j(\lambda) \neq 0\}$$

for $$S_0 = \{j; \beta_j^0 \neq 0\}$$

no significance testing involved

it's convex optimization only!

(and that can be a problem... see later)

Motif regression

for finding HIF1α transcription factor binding sites in DNA seq.

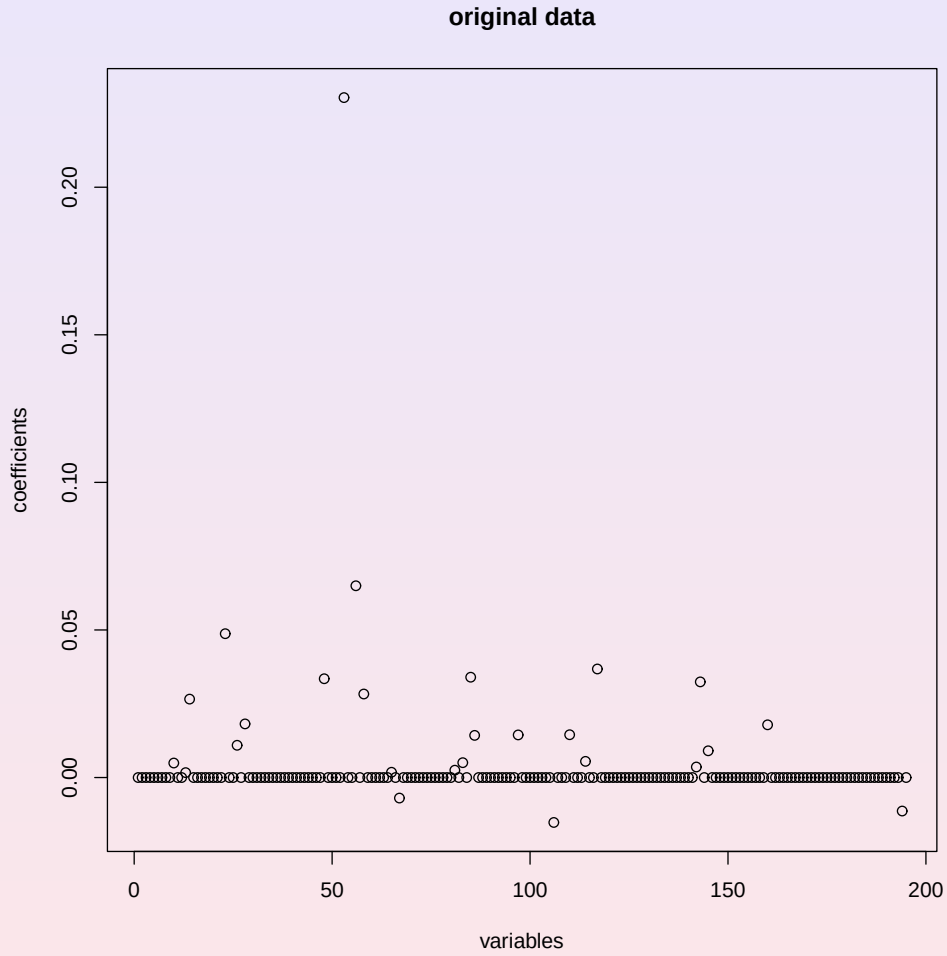
$Y_i \in \mathbb{R}$: univariate response measuring binding intensity on coarse DNA segment i (from CHIP-chip experiments)
$X_i^{(j)}$ = abundance score of candidate motif j in DNA segment i

variable selection in linear model $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i$,

$i = 1, \ldots, n = 287,\ p = 195$

$\rightsquigarrow$ Lasso selects 26 covariates and $R^2 \approx 50\%$
i.e. 26 interesting candidate motifs

motif regression: estimated coefficients $\hat{\beta}(\hat{\lambda}_{\mathrm{CV}})$



"Theory" for variable selection with Lasso

for (fixed design) linear model $\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon$ with
active set $S_0 = \{j;\ \beta_j^0 \neq 0\}$
two key assumptions

1. neighborhood stability condition for design $\mathbf{X}$
   $\Leftrightarrow$ irrepresentable condition for design $\mathbf{X}$
2. beta-min condition

$$\min_{j \in S_0} |\beta_j^0| \quad \geq \quad C\sqrt{s_0 \log(p)/n}, \quad C \text{ suitably large}$$

both conditions are sufficient and "essentially" necessary for

$$\hat{S}(\lambda) = S_0 \text{ with high probability}, \quad \lambda \gg \underbrace{\sqrt{\log(p)/n}}_{\text{larger than for pred.}}$$

already proved in Meinshausen & PB, 2004 (publ: 2006)
and both assumptions are restrictive!

"Theory" for variable selection with Lasso

for (fixed design) linear model $\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon$ with
active set $S_0 = \{j;\ \beta_j^0 \neq 0\}$
two key assumptions

1. neighborhood stability condition for design $\mathbf{X}$
   $\Leftrightarrow$ irrepresentable condition for design $\mathbf{X}$
2. beta-min condition

$$\min_{j \in S_0} |\beta_j^0| \quad \geq \quad C\sqrt{s_0 \log(p)/n}, \quad C \text{ suitably large}$$

both conditions are sufficient and "essentially" necessary for

$$\hat{S}(\lambda) = S_0 \text{ with high probability}, \quad \lambda \gg \underbrace{\sqrt{\log(p)/n}}_{\text{larger than for pred.}}$$

already proved in Meinshausen & PB, 2004 (publ: 2006)
and both assumptions are restrictive!

neighborhood stability condition $\Leftrightarrow$ irrepresentable condition

(Zhao & Yu, 2006)

$$n^{-1}\mathbf{X}^T\mathbf{X} = \hat{\Sigma}$$

active set $S_0 = \{j;\ \beta_j \neq 0\} = \{1, \ldots, s_0\}$ consists of the first s_0 variables; partition

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{S_0,S_0} & \hat{\Sigma}_{S_0,S_0^c} \\ \hat{\Sigma}_{S_0^c,S_0} & \hat{\Sigma}_{S_0^c,S_0^c} \end{pmatrix}$$

irrep. condition : $\|\hat{\Sigma}_{S_0^c,S_0}\hat{\Sigma}_{S_0,S_0}^{-1}\operatorname{sign}(\beta_1^0, \ldots, \beta_{s_0}^0)^T\|_\infty < 1$

not very realistic assumptions... what can we expect?

recall: under compatibility condition

$$\|\hat{\beta} - \beta^0\|_1 \leq C \frac{s_0}{\phi_0^2} \sqrt{\log(p)/n}$$

consider the relevant active variables

$$S_{\text{relev}} = \{j;\ |\beta_j^0| > C \frac{s_0}{\phi_0^2} \sqrt{\log(p)/n}\}$$

then, clearly,

$$\hat{S} \supseteq S_{\text{relev}} \text{ with high probability}$$

screening for detecting the relevant variables is possible!
without beta-min condition and assuming compatibility condition only

in addition: assuming beta-min condition

$$\min_{j \in S_0} |\beta_j^0| > C \frac{s_0}{\phi_0^2} \sqrt{\log(p)/n}$$

$$\hat{S} \supseteq S_0 \text{ with high probability}$$

screening for detecting the true variables

Tibshirani (1996):
LASSO = Least Absolute Shrinkage and Selection Operator

new translation:
LASSO = Least Absolute Shrinkage and Screening Operator

Practical perspective

choice of λ: $\hat{\lambda}_{CV}$ from cross-validation
empirical and theoretical indications (Meinshausen & PB, 2006) that

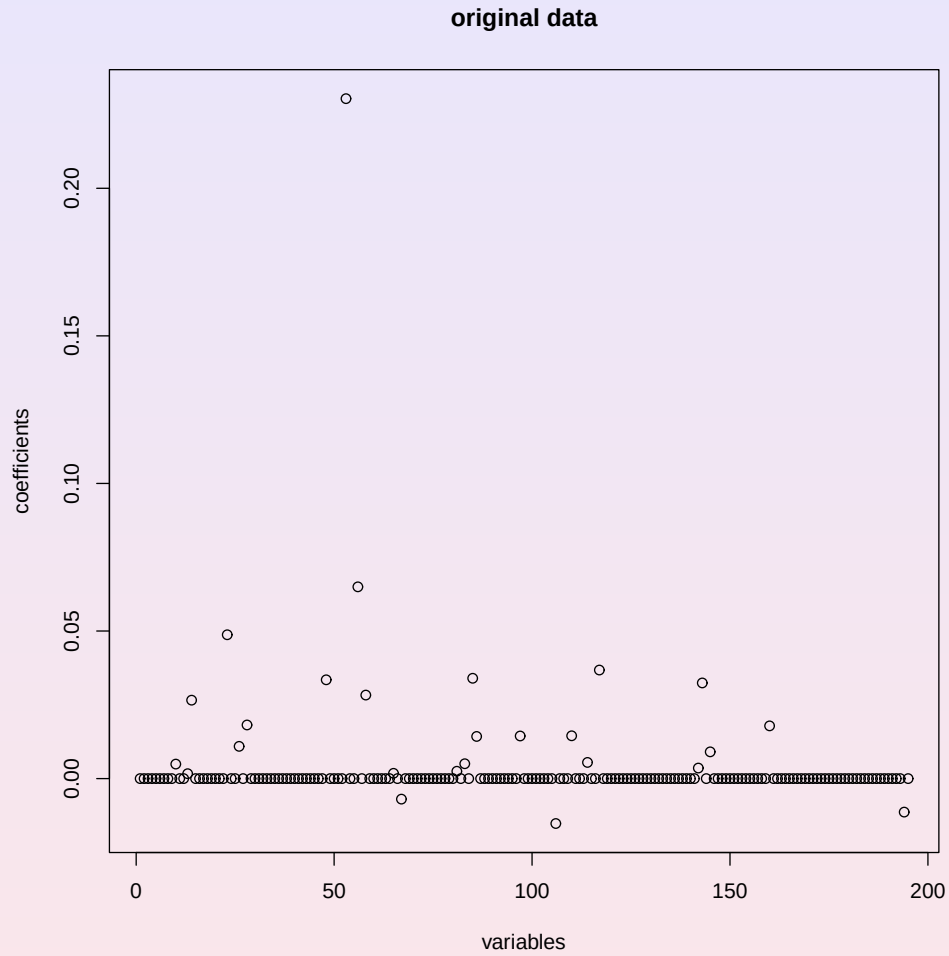
$$\hat{S}(\hat{\lambda}_{CV}) \supseteq S_0 \quad (\text{or } S_{\text{relev}})$$

moreover

$$|\hat{S}(\hat{\lambda}_{CV})| \leq \min(n, p) (= n \text{ if } p \gg n)$$

$\rightsquigarrow$ huge dimensionality reduction (in the original covariates)

motif regression: estimated coefficients $\hat{\beta}(\hat{\lambda}_{\mathrm{CV}})$



which variables in $\hat{S}$ are false positives?
(p-values would be very useful!)

recall:

$$\hat{S}(\hat{\lambda}_{CV}) \supseteq S_0 \quad (\text{or } S_{\text{relev}})$$

and we would then use a second-stage to reduce the number of false positive selections

$\rightsquigarrow$ re-estimation on much smaller model with variables from $\hat{S}$

- OLS on $\hat{S}$ with e.g. BIC variable selection
- thresholding coefficients and OLS re-estimation
- adaptive Lasso (Zou, 2006)
- ...

recall:

$$\hat{S}(\hat{\lambda}_{CV}) \supseteq S_0 \quad (\text{or } S_{\text{relev}})$$

and we would then use a second-stage to reduce the number of false positive selections

$\rightsquigarrow$ re-estimation on much smaller model with variables from $\hat{S}$

- OLS on $\hat{S}$ with e.g. BIC variable selection
- thresholding coefficients and OLS re-estimation
- adaptive Lasso (Zou, 2006)
- ...

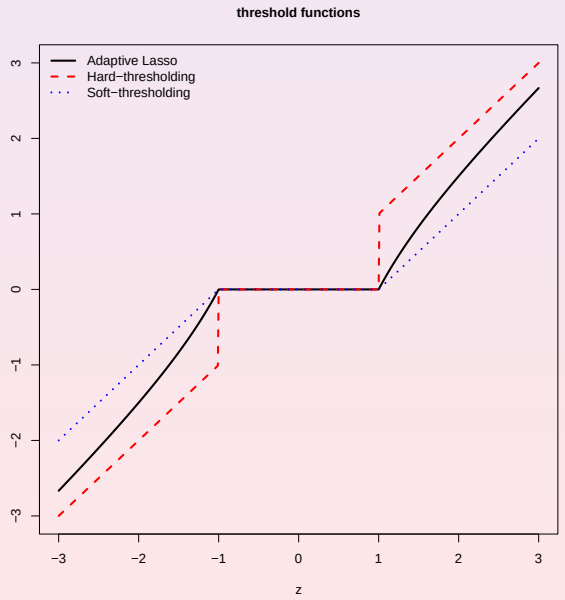
Adaptive Lasso (Zou, 2006)

re-weighting the penalty function

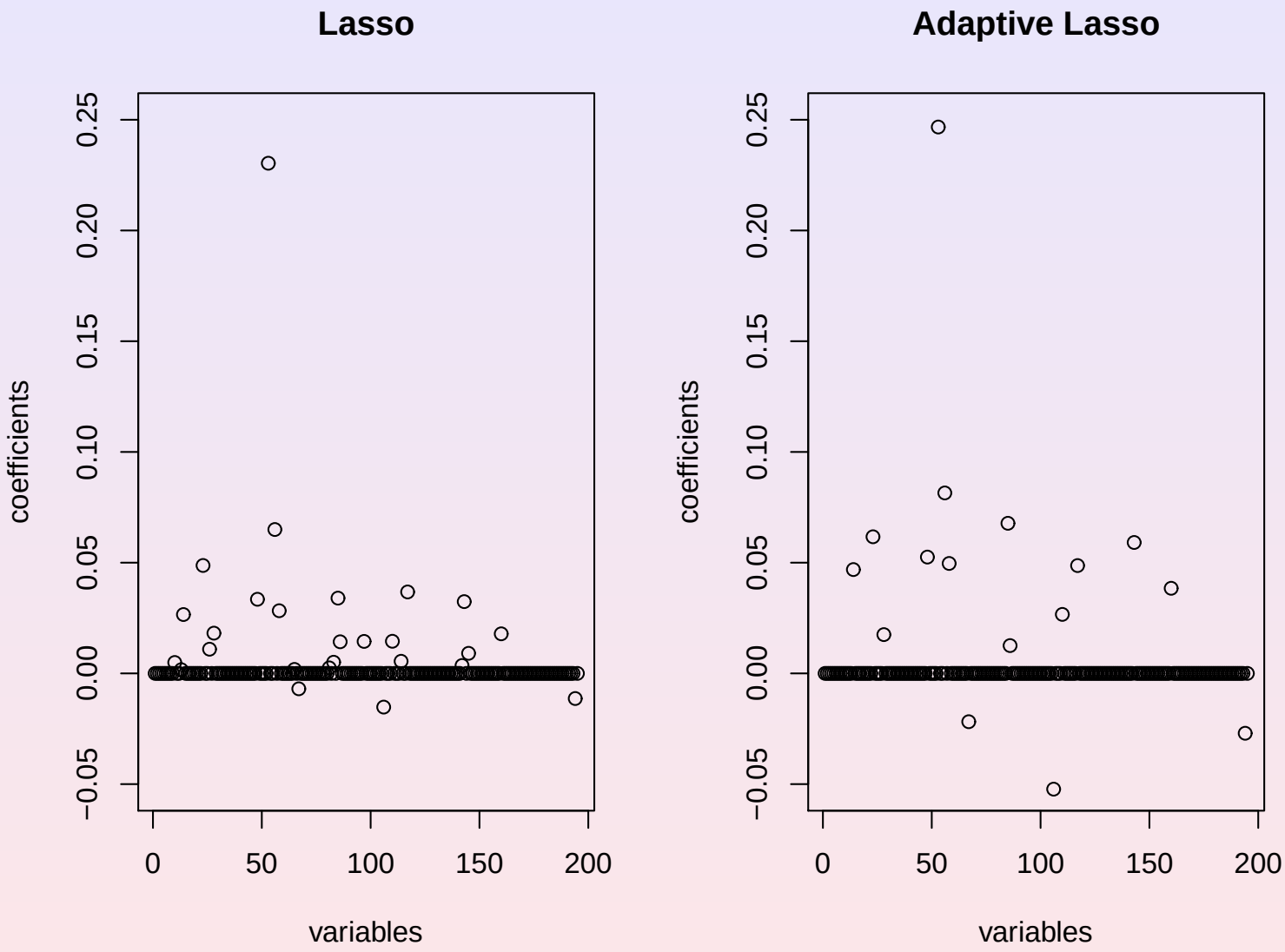
$$\hat{\beta} = \text{argmin}_{\beta}(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{init,j}|}),$$

$\hat{\beta}_{init,j}$ from Lasso in first stage $\underbrace{\text{(or OLS if } p < n\text{)}}_{\text{Zou (2006)}}$

for orthogonal design,
if $\hat{\beta}_{init}$ = OLS:
Adaptive Lasso = NN-garrote
$\rightsquigarrow$ less bias than Lasso



motif regression



Lasso selects 26 variables Adaptive Lasso selects 16 variables

KKT conditions and Computation

characterization of solution(s) $\hat{\beta}$ as minimizer of the criterion function

$$Q_\lambda(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_1$$

since $Q_\lambda(\cdot)$ is a convex function:
necessary and sufficient that subdifferential of $\partial Q_\lambda(\beta)/\partial\beta$ at $\hat{\beta}$ contains the zero element

Lemma 2.1 first part (in the book)
denote by $G(\beta) = -2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta)/n$ the gradient vector of $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n$
Then: $\hat{\beta}$ is a solution if and only if

$$G_j(\hat{\beta}) = -\text{sign}(\hat{\beta}_j)\lambda \text{ if } \hat{\beta}_j \neq 0,$$
$$|G_j(\hat{\beta})| \leq \lambda \text{ if } \hat{\beta}_j = 0$$

Lemma 2.1 second part (in the book)

If the solution of $\text{argmin}_\beta Q_\lambda(\beta)$ is not unique (e.g. if $p > n$), and if $G_j(\hat{\beta}) < \lambda$ for some solution $\hat{\beta}$, then $\hat{\beta}_j = 0$ for all (other) solutions $\hat{\beta}$ in $\text{argmin}_\beta Q_\lambda(\beta)$.

The zeroes are "essentially" unique
("essentially" refers to the situation: $\hat{\beta}_j = 0$ and $G_j(\hat{\beta}) = \lambda$)

Proof: Exercise (optional), or see in the book

Coordinate descent algorithm for computation

general idea is to compute a solution $\hat{\beta}(\lambda_{\text{grid},k})$ and use it as a starting value for the computation of $\hat{\beta}(\underbrace{\lambda_{\text{grid},k-1}}_{<\lambda_{\text{grid},k}})$

$\beta^{(0)} \in \mathbb{R}^p$ an initial parameter vector. Set $m = 0$.

REPEAT:
Increase m by one: $m \leftarrow m + 1$.
For $j = 1, \ldots, p$:

$$\text{if } |G_j(\beta_{-j}^{(m-1)})| \leq \lambda : \text{ set } \beta_j^{(m)} = 0,$$
$$\text{otherwise: } \beta_j^{(m)} = \text{argmin}_{\beta_j} Q_\lambda(\beta_{+j}^{(m-1)}),$$

β_{-j}: parameter vector setting jth component to zero
$\beta_{+j}^{(m-1)}$: parameter vector which equals $\beta^{(m-1)}$ except for jth component equalling β_j
UNTIL numerical convergence

for squared error loss: explicit up-dating formulae (Exercise 2.7)

$$G_j(\beta) = -2\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\beta)$$
$$\beta_j^{(m)} = \frac{\text{sign}(Z_j)(|Z_j| - \lambda/2)_+}{\hat{\Sigma}_{jj}},$$
$$Z_j = \mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\beta_{-j})/n, \quad \hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}.$$

$\rightsquigarrow$ componentwise soft-thresholding

this is very fast if true problem is sparse
active set strategy: can do non-systematic cycling, visiting mainly the active (non-zero) components

riboflavin example, n=71, p=4088
0.33 secs. CPU using `glmnet`-package in `R`
(Friedman, Hastie & Tibshirani, 2008)

coordinate descent algorithm converges to a stationary point (Paul Tseng $\approx$ 2000)
$\rightsquigarrow$ convergence to a global optimum, due to convexity of the problem

main assumption:

$$\text{objective function} = \text{smooth function} + \underbrace{\text{penalty}}_{\textit{separable}}$$

here: "separable" means "additive", i.e., $\text{pen}(\beta) = \sum_{j=1}^{p} p_j(\beta_j)$

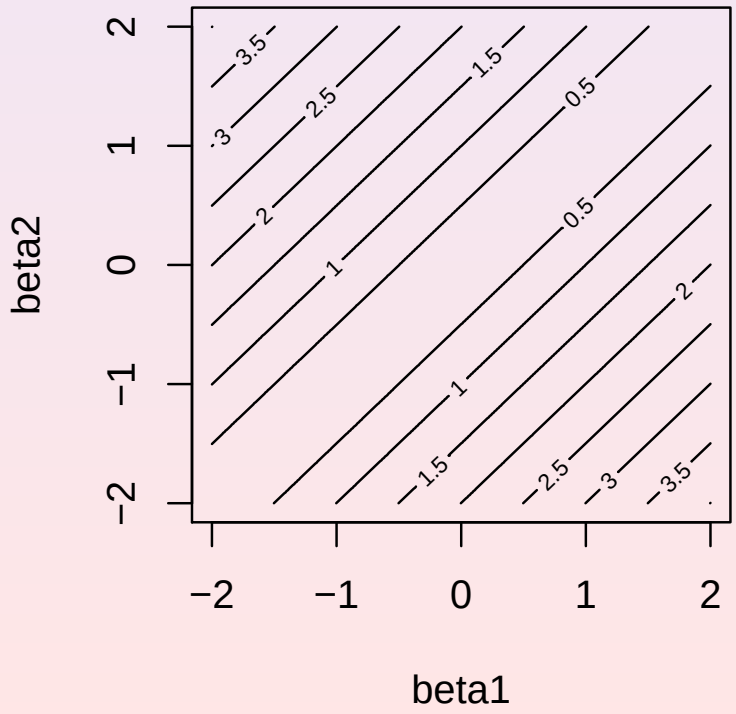
failure of coordinate descent algorithm:
Fused Lasso

$$\hat{\beta} = \text{argmin}_\beta \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| + \lambda_2 \|\beta\|_1$$

but $\sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$ is non-separable

contour lines of penalties for $p = 2$

|beta1 − beta2|



|beta1| + |beta2|