Statistics for High-Dimensional Data: Selected Topics

Peter Bühlmann

Seminar für Statistik, ETH Zürich

June 2014
High-dimensional linear model: basic concepts in methodology, theory and computation
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 $n = 115$, $p \gg n$

Y versus 9 “reasonable” genes

gene expression data
general framework:

\[ Z_1, \ldots, Z_n \text{ (with some "i.i.d. components")} \]

\[ \text{dim}(Z_i) \gg n \]

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

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: \( Y = X\beta + \varepsilon \)

goals:

- prediction, e.g. w.r.t. squared prediction error
- estimation of \( \beta^0 \), 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. 4’160’000 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. 4’160’000 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. 4’160’000 entries on Google Scholar for “high dimensional linear model” ...
Penalty-based methods

if true $\beta^0$ is sparse w.r.t.

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

- $\|\beta^0\|_1 = \sum_{j=1}^{p} |\beta_j^0|$
  - $\sim$ penalize with the $\| \cdot \|_1$-norm, i.e. Lasso:
    $\arg\min_{\beta} (n^{-1}\|Y - X\beta\|^2 + \lambda\|\beta\|_1)$
  - $\sim$ convex optimization:
    computationally feasible and very fast for large $p$
as we will see: regularization with the $\ell_1$-norm is good ("nearly optimal") even if truth is sparse w.r.t. $\|\beta^0\|_0$.

\(\leadsto\) can exploit computational advantage of $||.||_1$-norm regularization even if the problem has a different sparsity structure.
The Lasso (Tibshirani, 1996)

Lasso for linear models

\[ \hat{\beta}(\lambda) = \arg\min_\beta (n^{-1} \| Y - X\beta \|^2 + \lambda \left( \sum_{j=1}^p |\beta_j| \right) \geq 0) \]

\( \sim \) convex optimization problem

- Lasso does variable selection
  some of the \( \hat{\beta}_j(\lambda) = 0 \)
  (because of “\( \ell_1 \)-geometry”)

- \( \hat{\beta}(\lambda) \) is a shrunken LS-estimate
more about “ℓ₁-geometry”

equivalence to primal problem

\[ \hat{\beta}_\text{primal}(R) = \arg\min_{\beta; \|\beta\|_1 \leq R} \| Y - X\beta \|_2^2 / n, \]

with a correspondence between \( \lambda \) and \( R \) which depends on the data \((X_1, Y_1), \ldots, (X_n, Y_n)\)

[such an equivalence holds since

▶ \( \| Y - X\beta \|_2^2 / n \) is convex in \( \beta \)

▶ convex constraint \( \|\beta\|_1 \leq R \)

see e.g. Bertsekas (1995)]
left: $\ell_1$-“world”
residual sum of squares reaches a minimal value (for certain constellations of the data) if its contour lines hit the $\ell_1$-ball in its corner
$\leadsto \hat{\beta}_1 = 0$
$\ell_2$-"world" is different

Ridge regression,

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

equivalent primal equivalent solution

$$\hat{\beta}_{\text{Ridge; primal}}(R) = \arg\min_\beta_{\| \beta \|_2 \leq R} \| Y - X\beta \|^2_2 / n,$$

with a one-to-one correspondence between $\lambda$ and $R$
A note on the Bayesian approach

model:

$$\beta_1, \ldots, \beta_p \text{ i.i.d. } \sim p(\beta) d\beta,$$

given $\beta$ : $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$ with density $f(y|\sigma^2, \beta)$

posterior density:

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

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

$$\hat{\beta}_{MAP} = \arg\max_{\beta} p(\beta|Y, \sigma^2) = \arg\min_{\beta} -\log \left(f(Y|\beta, \sigma^2)p(\beta)\right)$$

$$= \arg\min_{\beta} \left(\frac{1}{2\sigma^2} \|Y - X\beta\|^2_2 - \sum_{j=1}^{p} \log(p(\beta_j))\right)$$
examples:
1. Double-Exponential prior DExp(ξ):
\[ p(\beta) = \frac{\tau}{2} \exp(-\tau \beta) \]
\[ \sim \hat{\beta}_{MAP} \text{ equals the Lasso with penalty parameter } \lambda = n^{-1} 2\sigma^2 \tau \]

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

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

\[ Y = X\beta + \varepsilon, \quad n^{-1}X^TX = I \]

Lasso = soft-thresholding estimator

\[ \hat{\beta}_j(\lambda) = \text{sign}(Z_j)(|Z_j| - \lambda/2)_+, \quad Z_j = (n^{-1}X^TY)_j, \]

\[ = \text{OLS} \]

\[ \hat{\beta}_j(\lambda) = g_{\text{soft}}(Z_j), \]

[Exercise]
Prediction

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

\[
\mathbb{E}_{X_{\text{new}}, Y_{\text{new}}}[ (Y_{\text{new}} - X_{\text{new}}\hat{\beta})^2 ] = \sigma^2 + \mathbb{E}_{X_{\text{new}}}[ (X_{\text{new}}(\beta^0 - \hat{\beta}))^2 ] \\
= \sigma^2 + (\hat{\beta} - \beta^0)^T \Sigma \text{Cov}(X) \text{(hat beta - beta0)} \\
= \sigma^2 + (\hat{\beta} - \beta^0)^T \Sigma (\hat{\beta} - \beta^0)
\]

\( \leadsto \) terminology “prediction error”:

for random design \( X \): \( (\hat{\beta} - \beta^0)^T \Sigma (\hat{\beta} - \beta^0) = \mathbb{E}_{X_{\text{new}}}[ (X_{\text{new}}(\hat{\beta} - \beta^0))^2 ] \)
for fixed design \( X \): \( (\hat{\beta} - \beta^0)^T \hat{\Sigma} (\hat{\beta} - \beta^0) = \|X(\hat{\beta} - \beta^0)\|_2^2 / n \)
binary lymph node classification using gene expressions:
a high noise problem

\[ n = 49 \text{ samples, } p = 7130 \text{ gene expressions} \]

despite that it is classification: 
\[ \mathbb{P}[Y = 1|X = x] = \mathbb{E}[Y|X = x] \]

\[ \leadsto \hat{p}(x) \text{ via linear model; can then do classification} \]

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Lasso</th>
<th>L_2Boosting</th>
<th>FPLR</th>
<th>Pelora</th>
<th>1-NN</th>
<th>DLDA</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error</td>
<td>21.1%</td>
<td>17.7%</td>
<td>35.25%</td>
<td>27.8%</td>
<td>43.25%</td>
<td>36.12%</td>
<td>36.88%</td>
</tr>
</tbody>
</table>

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

\[ n = 49 \text{ samples, } p = 7130 \text{ gene expressions} \]

despite that it is classification:

\[ \mathbb{P}[Y = 1|X = x] = \mathbb{E}[Y|X = x] \sim \hat{\rho}(x) \text{ via linear model; can then do classification} \]

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Lasso</th>
<th>L₂Boosting</th>
<th>FPLR</th>
<th>Pelora</th>
<th>1-NN</th>
<th>DLDA</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misclassification</td>
<td>21.1%</td>
<td>17.7%</td>
<td>35.25%</td>
<td>27.8%</td>
<td>43.25%</td>
<td>36.12%</td>
<td>36.88%</td>
</tr>
</tbody>
</table>

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

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

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

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

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

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

\[ s_0 = |\text{supp}(\beta^0)| = |\{j; \beta_j^0 \neq 0\}| \]
⇒ let’s turn to the blackboard!
various conditions and their relations (van de Geer & PB, 2009)

oracle inequalities for prediction and estimation

RIP $\rightarrow$ weak $(S,2s)$- RIP $\rightarrow$ adaptive $(S, 2s)$- restricted regression $\rightarrow$ $(S,2s)$-restricted eigenvalue $\rightarrow$ $S$-compatibility

$|S*| \leq s$ $\rightarrow$ coherence $\rightarrow$ adaptive $(S, s)$- restricted regression $\rightarrow$ $(S,s)$-restricted eigenvalue

weak $(S, 2s)$- irrepresentable $\leftarrow$ $(S,2s)$-irrepresentable $\leftarrow$ $(S,s)$-uniform irrepresentable

$|S*| = 0 \rightarrow S* = S$
Variable selection

Example: Motif regression for finding HIF1\(\alpha\) transcription factor binding sites in DNA seq. Müller, Meier, PB & Ricci (never published...)

\(Y_i \in \mathbb{R}\): univariate response measuring binding intensity of HIF1\(\alpha\) 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?

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

$$Y = X\beta + \epsilon, \ n = 287, \ p = 195$$

goal: variable selection
$\sim$ find the relevant motifs among the $p = 195$ candidates
Lasso for variable selection

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

for \( S_0 = \{ j; \beta^0_j \neq 0 \} \)

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 
\[
Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i,
\]

\( i = 1, \ldots, n = 287, \ p = 195 \)

\( \sim \) Lasso selects 26 covariates and \( R^2 \approx 50\% \)
i.e. 26 interesting candidate motifs
motif regression: estimated coefficients $\hat{\beta}(\hat{\lambda}_{CV})$
“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

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

both conditions are sufficient and “essentially” necessary for

\( \hat{S}(\lambda) = S_0 \) with high probability, \( \lambda \gg \sqrt{\log(p)/n} \)
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 $Y = X\beta^0 + \varepsilon$ with active set $S_0 = \{j; \beta_j^0 \neq 0\}$
two key assumptions

1. neighborhood stability condition for design $X$
   $\iff$ irrepresentable condition for design $X$

2. beta-min condition

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

both conditions are sufficient and “essentially” necessary for

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

already proved in Meinshausen & PB, 2004 (publ: 2006)
and both assumptions are restrictive!
neighborhood stability condition $\iff$ irrepresentable condition (Zhao & Yu, 2006)

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

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

$$S_{relev} = \{ j; | \beta_j^0 | > C \frac{S_0}{\phi_0^2} \sqrt{\log(p)/n} \}$$

then, clearly,

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

screening for detecting the relevant variables is possible!

without beta-min condition and assuming compatibility condition only
in addition: assuming beta-min condition

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

\(\hat{S} \supseteq S_0\) 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
choice of $\lambda$: $\hat{\lambda}_{CV}$ from cross-validation
empirical and theoretical indications (Meinshausen & PB, 2006) that

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

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 $\hat{\beta}(\hat{\lambda}_{CV})$

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

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

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

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

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

recall:

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

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

\[ \leadsto \text{re-estimation on much smaller model with variables from } \hat{S} \]

- 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

\[ \hat{\beta} = \arg\min_{\beta} \left( \|Y - X\beta\|_2^2 / n + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{init,j}|} \right), \]

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

Zou (2006)

for orthogonal design,

if \( \hat{\beta}_{init} = \text{OLS} \):
Adaptive Lasso = NN-garrote

\( \sim \) 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

\[
Q_\lambda(\beta) = \|Y - X\beta\|_2^2/n + \lambda\|\beta\|_1
\]

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 (first part)
denote by \( G(\beta) = -2X^T(Y - X\beta)/n \) the gradient vector of \( \|Y - X\beta\|_2^2/n \)

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 (second part)
If the solution of 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 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 by Bühlmann and van de Geer (2011; Lemma 2.1)
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 \( \hat{\beta}(\lambda_{\text{grid}}, k-1) \)

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

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

\[
\text{if } |G_j(\beta_{-j}^{(m-1)})| \leq \lambda : \text{ set } \beta_j^{(m)} = 0, \\
\text{otherwise: } \beta_j^{(m)} = \arg\min_{\beta_j} Q_{\lambda}(\beta_{+j}^{(m-1)}),
\]

\( \beta_{-j} \): parameter vector setting \( j \)th component to zero
\( \beta_{+j}^{(m-1)} \): parameter vector which equals \( \beta^{(m-1)} \) except for \( j \)th component equalling \( \beta_j \)
UNTIL numerical convergence
for squared error loss: explicit up-dating formulae (Exercise)

\[
G_j(\beta) = -2X_j^T(Y - X\beta)/n
\]

\[
\beta_j^{(m)} = \frac{\text{sign}(Z_j)(|Z_j| - \lambda/2)_+}{\hat{\Sigma}_{jj}},
\]

\[
Z_j = X_j^T(Y - X\beta_{-j})/n, \quad \hat{\Sigma} = n^{-1}X^TX.
\]

\[\leadsto\text{componentwise soft-thresholding}\]

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 \(\mathbb{R}\)

(Friedman, Hastie & Tibshirani, 2008)
coordinate descent algorithm converges to a stationary point (Paul Tseng $\approx$ 2000)\n$\Rightarrow$ convergence to a global optimum, due to convexity of the problem

main assumption:

\[
\text{objective function} = \text{smooth function} + \underbrace{\text{penalty}}_{\text{separable}}
\]

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 \( p = 2 \)

\( |\beta_1 - \beta_2| \)

\( |\beta_1| + |\beta_2| \)
Thank you!

References: most of the material is covered in