# Very high-dimensional data:

# convex and quasi-convex optimization for consistent model selection

**Peter Bühlmann**

**ETH Zürich**

# 1. High-dimensional data

$(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. or stationary

$X_i \in \mathbb{R}^p$ predictor variable

$Y_i$ univariate response variable, e.g. $Y_i \in \mathbb{R}$ or $Y_i \in \{0, 1\}$

high-dimensional: $p \gg n$

areas of application: astronomy, biology, imaging, marketing research, text classification,...

# High-dimensional linear models

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

$$p \gg n$$

includes basis expansion with highly overcomplete dictionary

goal: variable selection; but how?

approaches include:

variable selection via AIC, BIC, gMDL (in a forward manner);

Bayesian methods for regularization and variable selection; boosting; ...

Lasso; new relaxed Lasso, ...

our requirements:

- computationally feasible
- statistically accurate for selecting the correct variables and for prediction

computational feasibility for high-dimensional problems

$\rightsquigarrow$

greedy methods, heuristic search

or

convex optimization

# 3. Lasso-relaxation is "quite" good for $p \gg n$

Lasso or $\ell^1$-penalized regression (Tibshirani, 1996):

$$\hat{\beta}_{Lasso} = \text{argmin}_{\beta} n^{-1} \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p} \beta_j X_i^{(j)})^2 + \underbrace{\lambda}_{\geq 0;\ \text{penalty par.}} \sum_{j=1}^{p} |\beta_j|$$

- does variable selection: some (many) $\beta_j$'s exactly equal to 0
- does shrinkage
- involves a convex optimization only

this is convex relaxation:

replace the computationally hard/infeasible subset selection ($\ell^0$-penalty)

$$\text{argmin}_\beta n^{-1}\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p}\beta_j X_i^{(j)})^2 + \gamma \underbrace{\sum_{j=1}^{p} \mathbb{1}_{\{\beta_j \neq 0\}}}_{\|\beta\|_0}$$

e.g. AIC, BIC, ...

by the convex $\ell^1$-penalized problem

## 3.1. Prediction with convex Lasso-relaxation

consistency for prediction in high-dimensions (Greenshtein & Ritov, 2004)

- $p = p_n = O(n^{\alpha})$ for any $0 < \alpha < \infty$ (high-dimensional)
- $\sum_{j=1}^{p_n} |\beta_{j,n}| = o(n^{1/4} \log(n)^{-1/4})$ (sparse)

$\rightsquigarrow$ $\mathbf{E}_X[(\hat{f}(X) - f(X))^2] = o_P(1)$, $f, \hat{f}$ linear

Donoho, Candes, Tao, Tanner,... $\approx$ 2003-2005: many results on the $L_2$-norm (prediction) for basis pursuit and Lasso if $p = p_n = O(n)$

## 3.2. Variable selection and graphical modeling with the Lasso

goal: use the Lasso for determining presence/absence of associations between random variables ($\leadsto$ includes regression)

### Gaussian conditional independence graph

assume that $X = X^{(1)}, \ldots, X^{(p)} \sim \mathcal{N}_p(\mu, \Sigma)$

graph:

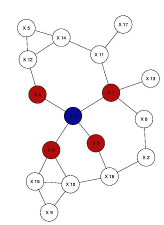
set of nodes $\Gamma = \{1, 2, \ldots, p\}$, corresponding to the $p$ random variables

set of edges $E \subseteq \Gamma \times \Gamma$ defined as:

there is an undirected edge between node $i$ and $j$

$\overset{\text{def}}{\Leftrightarrow}$ $X^{(i)}$ conditionally dependent of $X^{(j)}$ given all other $\{X^{(k)};\ k \neq i, j\}$

$\Leftrightarrow$ $\Sigma_{ij}^{-1} \neq 0$



note: $\Sigma_{ij}^{-1}$ corresponds to $\beta_j^{(i)} = \Sigma_{ij}^{-1}/\Sigma_{ii}^{-1}$, where

$$X^{(i)} = \beta_j^{(i)} X^{(j)} + \sum_{k \neq i,j} \beta_k^{(i)} X^{(k)} + \text{error}^{(i)}$$

$\rightsquigarrow$ we can infer the graph from variable selection in regression

$$\beta_j^{(i)} = 0 \Leftrightarrow \Sigma_{ij}^{-1} = 0 \; (\Leftrightarrow \beta_i^{(j)} = 0)$$

huge computational problem when using e.g. subset selection à la BIC:

worst case $p2^{p-1}$ least squares problems!

and still infeasible with up- down-dating strategies

## **Just relax!**

replace the computationally hard problem by a convex problem:

compute the Lasso estimates $\hat{\beta}_i^{(j)}$ (for all regressions)

Estimation of graph:

estimate an edge between node $i$ and $j$ if

$$\hat{\beta}_j^{(i)} \neq 0 \text{ and } \hat{\beta}_i^{(j)} \neq 0$$

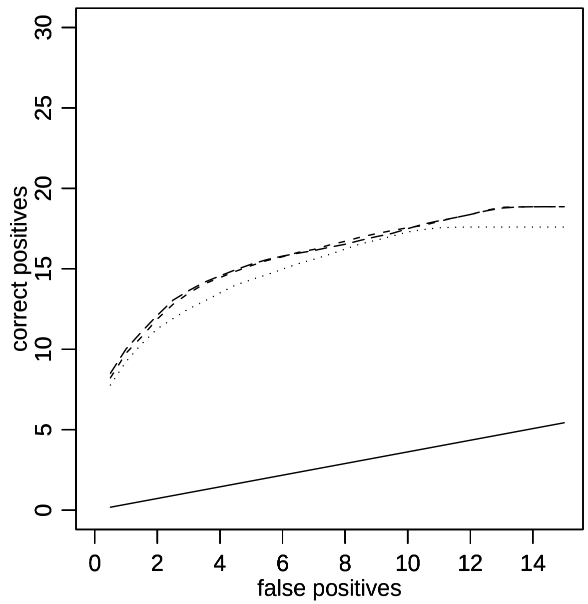
(for finite samples: it could happen that only one of the $\hat{\beta}_j^{(i)}, \hat{\beta}_i^{(j)}$ is $\neq 0$)

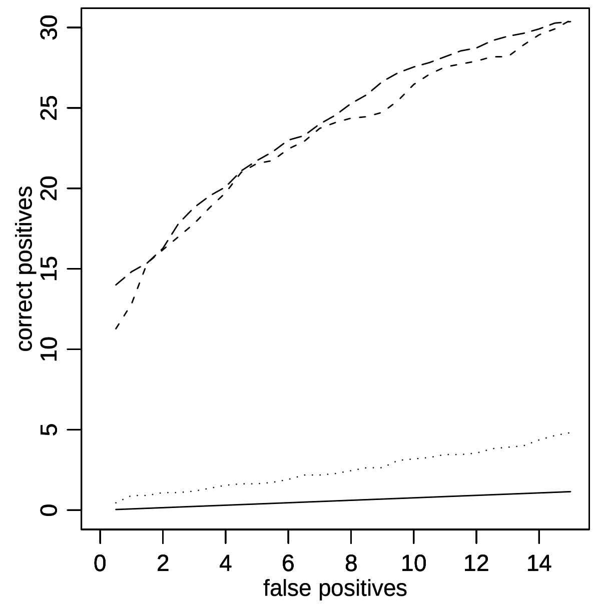
this involves only convex optimizations!

instead of checking exhaustively $p2^{p-1}$ least squares problems (e.g. using BIC)

# Comparison of Lasso and classical stepwise selection

$p = 10$ $\qquad$ $p = 30$




dotted $\cdots\cdot$ stepwise selection

dashed $\_\ \_\ \_$ Lasso

ROC-curves for estimated graphs with $p = 10, 30$ nodes and $n = 40$ obs.

true graphs are sparse, having at most 4 edges out of every node

# Some theory for high dimensions

Theorem (Meinshausen & PB, 2004)

For $\lambda_n \sim Cn^{-1/2+\delta/2}$,

$$\mathbb{P}[\text{estimated graph}(\lambda_n) \;=\; \text{true graph}] = 1 + O(\exp(-Cn^{\delta})) \quad (n \to \infty)$$

$(0 < \delta < 1)$

if

- Gaussian data
- $p = p_n = O(n^{\alpha})$ for any $\alpha > 0$ (high-dimensional)
- maximal number of edges out of a node $= O(n^{\kappa})$ $(0 < \kappa < 1)$ (sparseness)
- plus some other technical conditions (one of them being “a bit” restrictive)

justification for relaxation with computationally simple convex problems!

## Choice of $\lambda$

Theorem doesn't say much about choosing $\lambda$...

first (not so good) idea: choose $\lambda$ to optimize prediction

e.g. via some cross-validation scheme

but: for prediction oracle solution

$$\lambda^* = \arg\min_{\lambda} \mathbf{E}[(X^{(i)} - \sum_{j \neq i} \hat{\beta}_j^{(i)}(\lambda) X^{(j)})^2]$$

$\mathbb{P}[\text{estimated neighborh.}(\lambda^*)_i = \text{true neighborh.}_{;i}] \to 0 \ (p_n \to \infty, n \to \infty)$

$\mathbb{P}[\text{estimated graph}(\lambda^*) = \text{true graph}] \to 0 \ (p_n \to \infty, n \to \infty)$

asymptotically: the prediction optimal graph is too large

(Meinshausen & PB, 2004; related example by Leng et al., 2004)

# The good message

Lasso produces a set of sub-models

$$M_1 \subset \ldots \subset \ldots \qquad \underbrace{M_{pred-opt}}_{\text{optimal for prediction with Lasso}} \qquad \subset \ldots \subset M_N$$

with $N = O(\min(n, p))$

and $M_{true}$ is with probability $1 - O(\exp(-Cn^{\delta}))$ among these models

but $M_{true} \neq M_{red-opt}$

# 4. Beyond Lasso

consider linear model $Y = X\beta + \varepsilon$

for orthonormal design: $\mathbf{X}^T\mathbf{X} = I$: Lasso yields the soft-threshold estimator

**Is soft-thresholding or Lasso a good thing?**

- $\beta_1, \ldots \beta_p$ i.i.d. $\sim$ Double-Exponential,
  soft-thresholding and the Lasso yield the MAP (which often performs well)
- minimax results for soft-thresholding (Donoho & Johnstone, ...)

but: a different story in the very high-dimensional sparse case

assume:

- $p = p_n \sim C_1 \exp(C_2 n^{1-\xi})$ $(0 < \xi < 1)$
- effective number of variables is finite (finite $\ell^0$-norm)

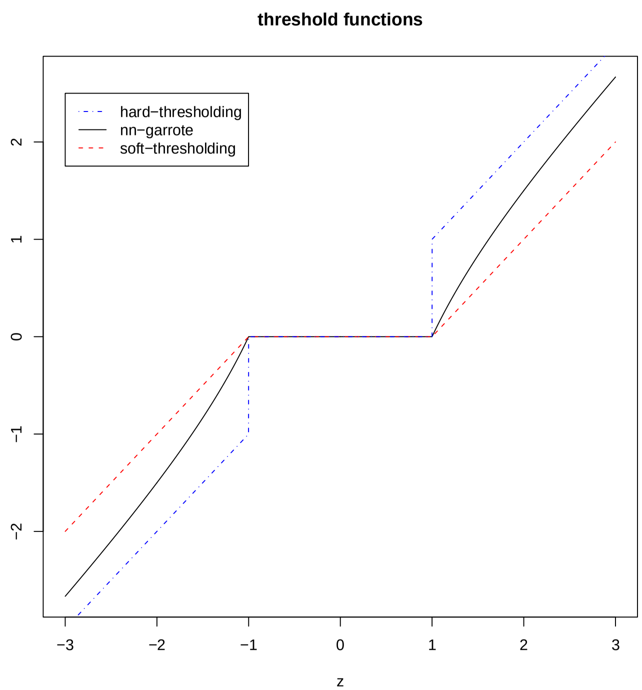
  non-effective variables are independent

Theorem (Meinshausen, 2005)

$$\mathbb{P}[\inf_{\lambda} \underbrace{L(\lambda)}_{\text{risk of Lasso}} > cn^{-r}] \to 1 \ (n \to \infty) \text{ for } r > \xi$$

while optimal rate is $n^{-1}$ (achieved e.g. by OLS with the true variables)

$\leadsto$ Lasso can have very poor convergence rate reason: need large $\lambda$ for variable selection $\rightsquigarrow$ strong bias of soft-thresholding




Better:

- SCAD (Fan and Li, 2001)

- Nonnegative Garrote (Breiman, 1995)

- Bridge estimation
  (Frank and Friedman, 1993)

they all work for general $\mathbf{X}$

for non-orthogonal $\mathbf{X}$:

- non-convex optimization for SCAD or Bridge estimation
- NN-Garrote only for $p \leq n$

## 4.1. The relaxed Lasso (Meinshausen, 2005)

for $\lambda \geq 0,\ 0 \leq \phi \leq 1$

$$\hat{\beta}_{\lambda,\phi} = \arg\min_{\beta} n^{-1} \sum_{i=1}^{n} (Y_i - \sum_{j \in \underbrace{\mathcal{M}_\lambda}_{\text{model from Lasso}(\lambda)}} \beta_j X_i^{(j)})^2 + \phi\lambda\|\beta\|_1$$

for $\phi = 0$: OLS on selected variables from Lasso($\lambda$)

for $\phi = 1$: Lasso($\lambda$)

amount of computation for finding all solutions over $\lambda$ and $\phi$:

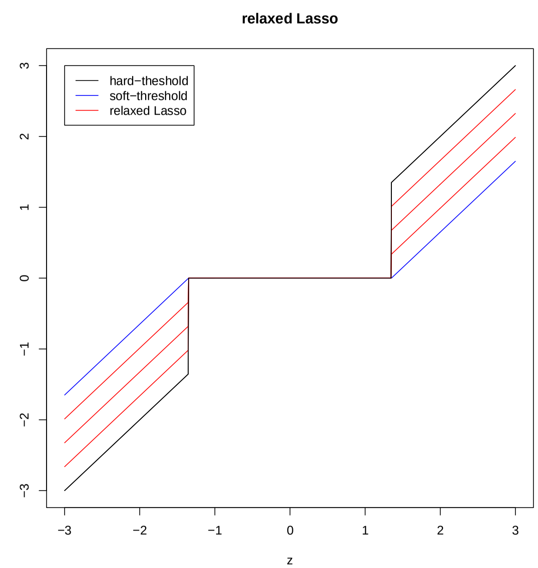
often, the same computational complexity as for Lasso/LARS (surprising):

$$O(np\min(n,p)) = O(p) \text{ if } p \gg n$$

worst case: $O(np\min(n,p)^2) = O(p)$ if $p \gg n$ still linear in $p$

this is "quasi-convex" optimization: two levels of a convex problem for orthonormal case:

$$\mathbf{X}^T\mathbf{X} = I$$




Theorem (Meinshausen, 2005)

in general, with essentially the same assumptions as for the Lasso

$$\inf_{\lambda,\phi} L(\lambda, \phi) = O_P(n^{-1}) \ (n \to \infty)$$

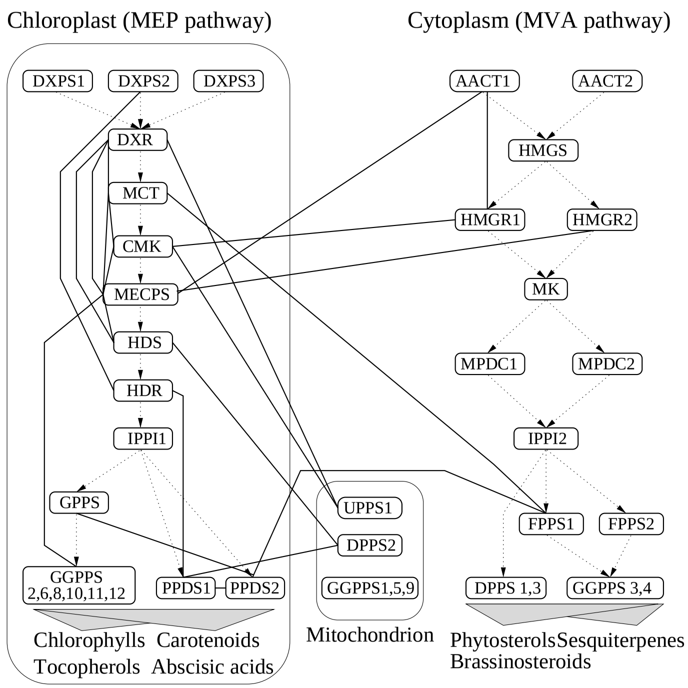
relaxed Lasso for variable selection and graphs/dependency networks

prediction optimal (or cross-validated) tuning parameters yield (for suitably regular cases) consistent variable selection and graph estimates two biosynthesis pathways in Arabidopsis

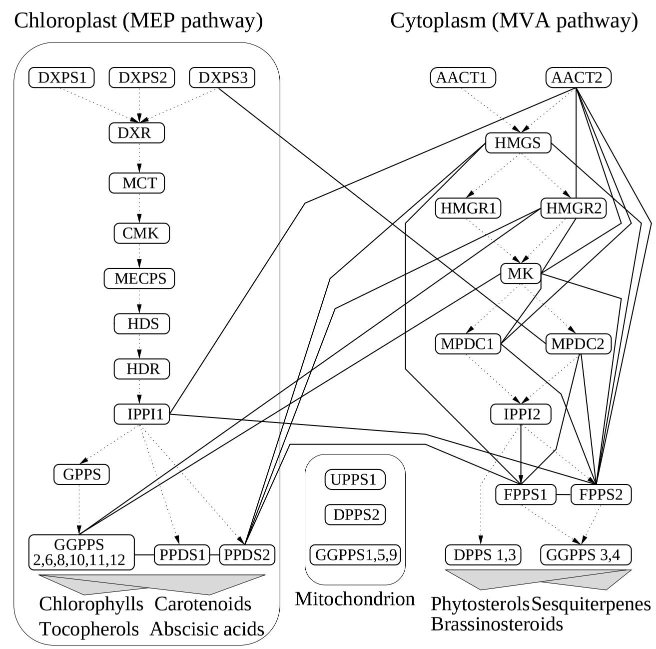
$n = 118$ Affymetrix gene expression measurements, $p = 39$ genes

plus additional biological information

$\rightsquigarrow$ the relaxed Lasso has been used as a "starting point" (Wille et al., 2004)




edges from MEP "module" to MVA

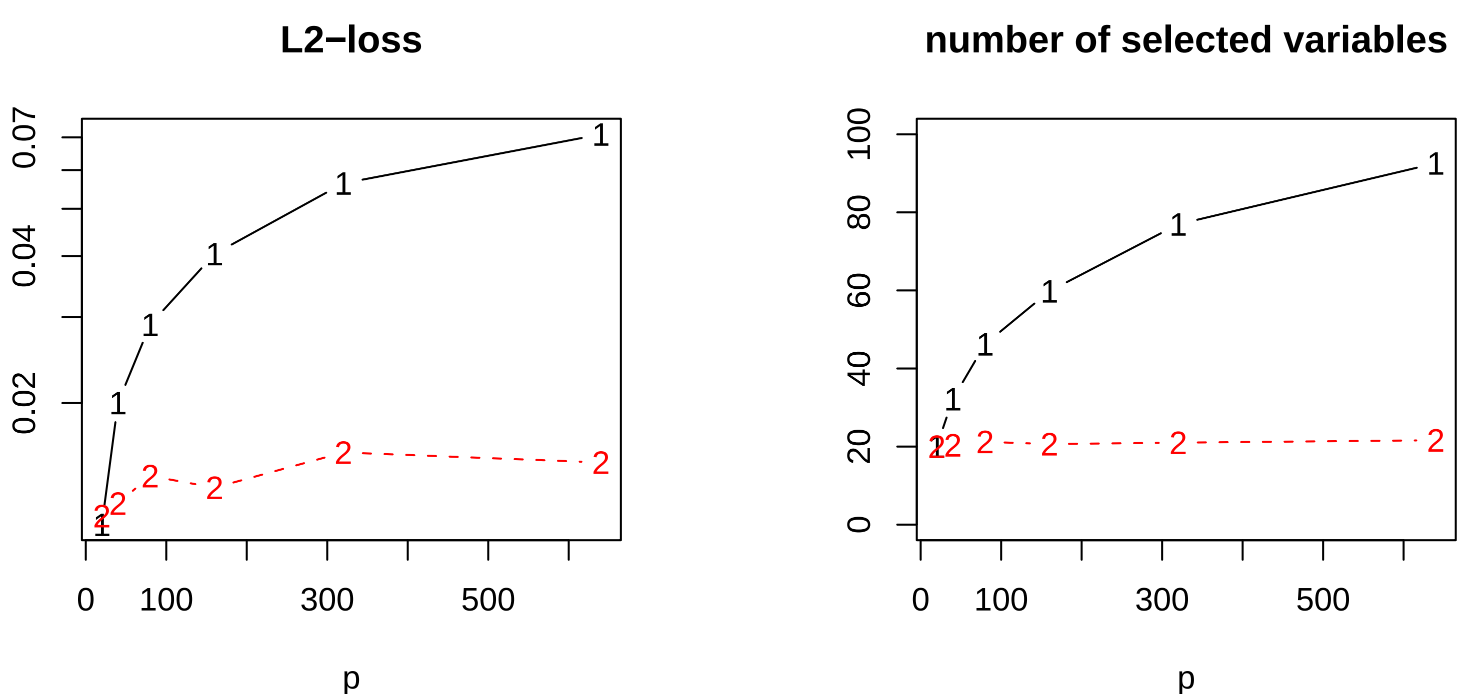



edges from MVA "module" to MEP

biologically most interesting novel connection: from IPPI1 to MVA "module"

## Regression: $n = 300, p = 20, \ldots 650, p_{eff} = 20$

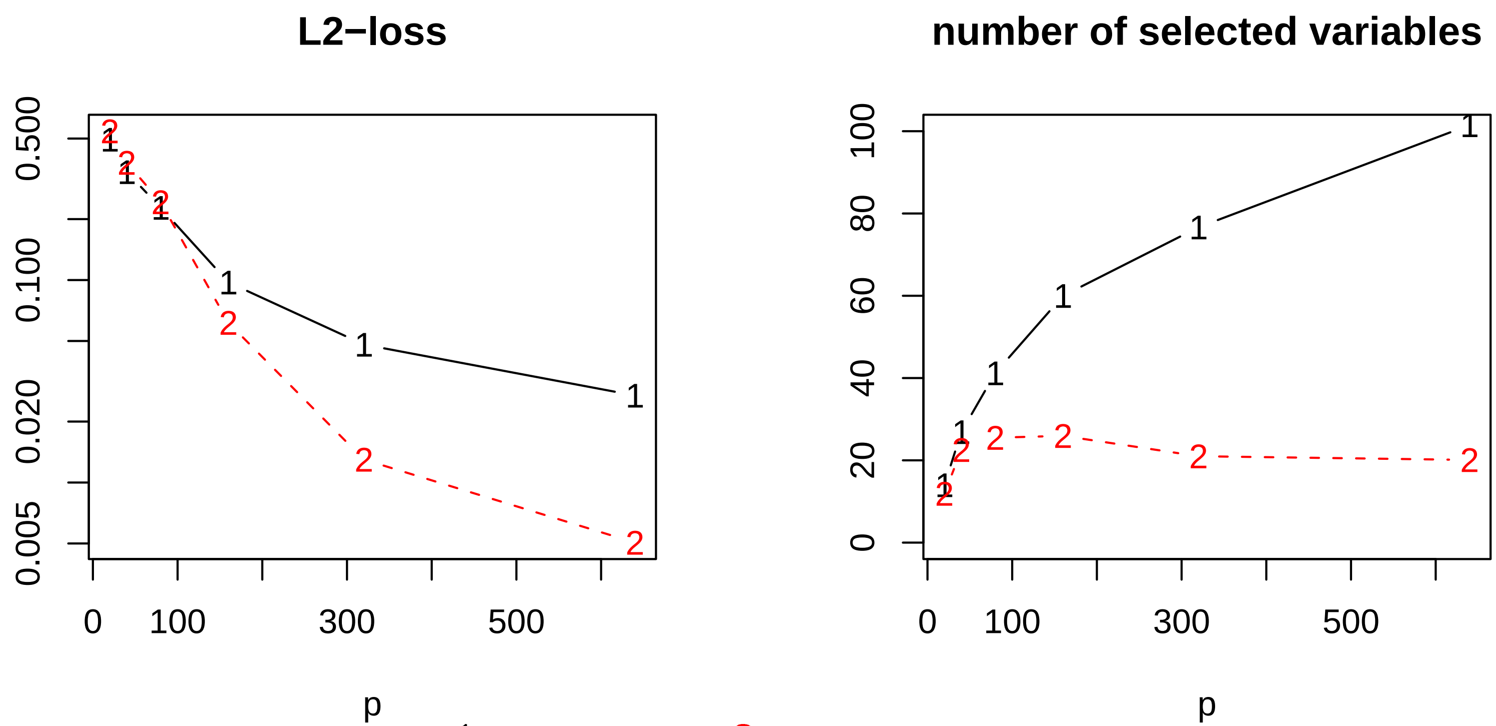
the price of collecting too many covariates




$1$: Lasso $2$: relaxed Lasso

pure noise variables are much less damaging with the relaxed Lasso than for Lasso and they are very disturbing for Ridge-type regularization (e.g. SVM)

$$n = p = 20, \ldots 650, p_{eff} = 20$$




$1$: Lasso $2$: relaxed Lasso

relaxed Lasso never substantially worse than the Lasso: the price for the flexibility of the relaxed Lasso is the larger search space $0 \leq \phi \leq 1$ (Lasso: $\phi = 1$)

relaxed Lasso is also $\underbrace{\text{better}}_{\text{for prediction and variable selection}}$ than Lasso-OLS hybrid

in particular if, e.g.

$$\beta_1, \ldots, \beta_{p_{eff}} \text{ i.i.d. } \sim \text{ Double-Exponential}$$
$$\beta_{p_{eff}+1} = \ldots = \beta_p = 0$$

and $p$ large, $p_{eff}$ not so large

## binary lymph node classification in breast cancer using gene expressions: a high noise problem

$n = 49$ samples, $p = 7129$ gene expressions

cross-validated misclassification rate:

| | |
|---|---|
| relaxed Lasso (tuned by 5-fold CV): | 16.3% |
| Lasso (tuned by 5-fold CV): | 21.0% |
| SVM: | 36.9% |
| DLDA: | 36.1% |

selected genes (on whole data set):

relaxed Lasso: 2 genes (!) Lasso: 23 genes

average from CV: 7.3 genes

the 2 genes from relaxed Lasso are also selected by Lasso

note the identifiability problem among highly correlated predictor variables short DNA motif modeling and prediction of 5' splice sites (Meier & PB, 2005)

$Y \in \{0, 1\}$: 5' is a splice site or not
$X \in \{A, C, G, T\}^9$: 9 DNA sequence positions

log-linear model with main effects and second-order interactions

but: $\ell^1$-penalized MLE depends on parameterization

Group Lasso (Yuan & Lin, 2004) helps

$\rightsquigarrow$ whole terms (e.g. an interaction term) are selected

training data $n = 10'000$ (only a fraction from Burge et al. (1999))

test data $n_{test} = 4208$

slightly better (w.r.t. ROC) than maximum entropy modeling (Yeo and Burge, 2004)

| | pred. 0 | pred. 1 |
|---|---|---|
| true 0 | 87'212 | 2505 |
| true 1 | 804 | 3404 |

could also tune for low false positives

but:

computations (of the whole path of relaxed group Lasso solutions) are subtle due to non-quadratic loss function and non-strict convexity of $\ell^1$-penalization

$\rightsquigarrow$ problem-specific implementations are required

## 5. Relations to Boosting

Boosting is “related to” Lasso

cf. Efron, Hastie, Johnstone, Tibshirani (2004)

and Boosting is much more generic than Lasso

e.g. other loss functions, nonparametric models, factors (i.e. group of variables),...

## 5.1. $L_2$Boosting

(Friedman, 2001)

specify a base procedure ("weak learner"):

$$\text{data} \xrightarrow{\text{algorithm A}} \hat{\theta}(\cdot) \quad \text{(a function estimate)}$$

e.g.: simple linear regression, tree (CART), ...

$L_2$Boosting with base procedure $\hat{\theta}(\cdot)$: repeated fitting of residuals

$$m = 1: \ (X_i, Y_i)_{i=1}^n \rightsquigarrow \hat{\theta}_1(\cdot),\ \hat{f}_1 = \underbrace{\nu}_{\text{e.g. }=0.1} \hat{\theta}_1 \rightsquigarrow \text{resid. } U_i = Y_i - \hat{f}_1(X_i)$$

$$m = 2: \ (X_i, U_i)_{i=1}^n \rightsquigarrow \hat{\theta}_2(\cdot),\ \hat{f}_2 = \hat{f}_1 + \nu\hat{\theta}_2 \rightsquigarrow \text{resid. } U_i = Y_i - \hat{f}_2(X_i)$$

$$\ldots \qquad \ldots$$

$\hat{f}_{m_{stop}}(\cdot) = \nu \sum_{m=1}^{m_{stop}} \hat{\theta}_m(\cdot)$ (greedy fitting of residuals)

Tukey (1977): twicing for $m_{stop} = 2$ and $\nu = 1$

Componentwise linear least squares base procedure for linear model fitting

linear OLS regression against the one predictor variable which reduces residual sum of squares most

$$\hat{\theta}(x) = \hat{\beta}_{\hat{\mathcal{S}}} x^{(\hat{\mathcal{S}})},$$

$$\hat{\beta}_j = \sum_{i=1}^{n} Y_i X_i^{(j)} / \sum_{i=1}^{n} (X_i^{(j)})^2, \;\; \hat{\mathcal{S}} = \arg\min_j \sum_{i=1}^{n} (Y_i - \hat{\beta}_j X_i^{(j)})^2$$

$L_2$Boosting with componentwise linear LS yields linear model fit:

first round of estimation: selected predictor variable $X^{(\hat{\mathcal{S}}_1)}$ (e.g. $= X^{(3)}$)

corresponding $\hat{\beta}_{\hat{\mathcal{S}}_1}$

use shrunken fit $\hat{f}_1(x) = \nu \hat{\beta}_{\hat{\mathcal{S}}_1} x^{(\hat{\mathcal{S}}_1)}$ (e.g. $\nu = 0.1$)

second round of estimation: selected predictor variable $X^{(\hat{\mathcal{S}}_2)}$ (e.g.$= X^{(21)}$)

corresponding $\hat{\beta}_{\hat{\mathcal{S}}_2}$

use shrunken fit $\hat{f}_2(x) = \hat{f}_1(x) + \nu \hat{\beta}_{\hat{\mathcal{S}}_2} x^{(\hat{\mathcal{S}}_2)}$

etc.

for $\nu = 1$, this is known as

Matching Pursuit (Mallat and Zhang, 1993)

Weak greedy algorithm (deVore & Temlyakov, 1997)

a version of Boosting (Schapire, 1992; Freund & Schapire, 1996)

Gauss-Southwell algorithm

C.F. Gauss in 1803

"Princeps Mathematicorum"

R.V. Southwell in 1933

Professor in engineering, Oxford

$L_2$Boosting with comp.wise linear LS is consistent for very high-dimensional, sparse linear models (PB, 2004)

properties for variable selection are not rigorously known

using the analogy to the Lasso/relaxed Lasso: instead of boosting,

$\leadsto$ boosting algorithm which is sparser than boosting

## 5.2. Sparse $L_2$Boosting

(PB and Yu, 2005)

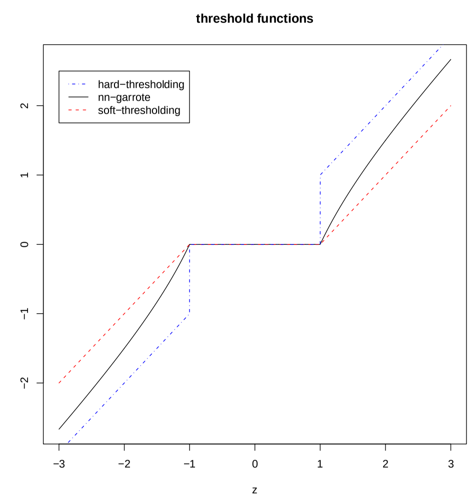
instead of minimizing RSS in every iteration,

minimize a final prediction error (FPE) criterion: we propose gMDL,

$$\hat{\theta}_m = \arg\min_{\theta(\cdot)} \sum_{i=1}^{n} (Y_i - \hat{f}_{m-1}(X_i))^2 + \underbrace{\text{gMDL-penalty}}_{\text{requires d.f. for boosting}}$$

d.f. for boosting via trace of hat-matrices for orthonormal linear model:

Sparse$L_2$Boosting with componentwise linear least squares yields Breiman's nonnegative garrote estimator (PB & Yu, 2005)

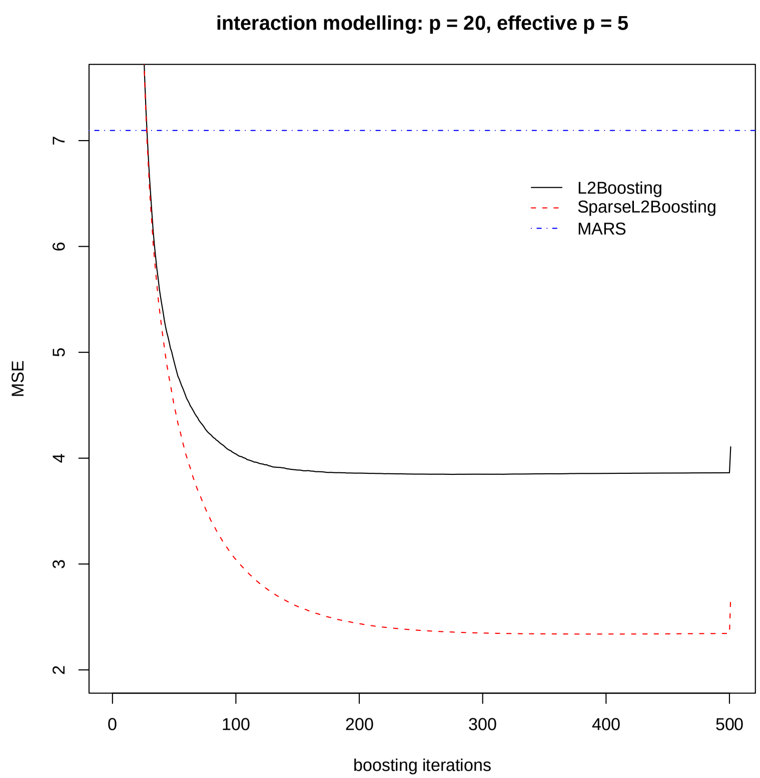



- Sparse$L_2$Boost yields sparser solutions than $L_2$Boosting
- Sparse$L_2$Boost still very generic (although less generic than $L_2$Boosting)
  e.g. nonparametric problems, non-quadratic loss functions
- no theory but lots of empirical evidence that
  Sparse$L_2$Boosting is a reasonable variable selection method

## Boosting with first-order interactions

base procedure: pairwise thin plate splines ($\mathbb{R}^2 \to \mathbb{R}$) which selects the pair of predictors such that corresponding spline smooth reduces RSS most (fixed d.f.)

$\rightsquigarrow$ nonparametric model fit with first-order interactions




Friedman #1 model:

$Y = 10\sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 +$

$10X_4 + 5X_5 + \mathcal{N}(0,1)$

$X = (X_1, \ldots, X_{20}) \sim$ Unif.$([0,1]^{20})$

Sample size $n = 50$

Dimension $p = 20, p_{eff} = 5$

# 6. Conclusions

- for variable selection and graphical modelling
  want to be sparser than prediction-optimal $\ell^1$-penalized solutions
  (or sparser than ordinary boosting)

- relaxed Lasso has the property that prediction optimal solutions yield good
  (i.e. consistent) variable selection
  $\rightsquigarrow$ can use cross-validation to determine a good model
  better to do “quasi-convex” instead of convex optimization
  (empirically similar for boosting: prediction optimal Sparse$L_2$Boosting
  often yields good variable selection scheme)