Very high-dimensional data: prediction and variable selection

Peter Bühlmann

Seminar für Statistik, ETH Zürich

substantial part: work with Nicolai Meinshausen

High-dimensional data

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

X_i p-dimensional 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 data

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

X_i p-dimensional 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

$$Y_i = (\beta_0 +) \sum_{j=1}^{p} \beta_j X_i^{(j)} + \epsilon_i, \; i = 1, \ldots, n$$

$p \gg n$

in short: $Y = X\beta + \epsilon$

goals:

- prediction, e.g. squared prediction error
- variable selection
  i.e. estimating the effective variables
  (having corresponding coefficient $\neq 0$)

High-dimensional linear models

$$Y_i = (\beta_0 +) \sum_{j=1}^{p} \beta_j X_i^{(j)} + \epsilon_i,\ i = 1, \ldots, n$$

$p \gg n$

in short: $Y = X\beta + \epsilon$

goals:

- prediction, e.g. squared prediction error
- variable selection
  i.e. estimating the effective variables
  (having corresponding coefficient $\neq 0$)

The best approach: ask a clever mind...

Other approaches include:

Ridge regression (Tikhonov regularization) for prediction

variable selection via AIC, BIC, (g)MDL (in a forward manner)

Bayesian methods for regularization, ...

computational feasibility for high-dimensional problems $\rightsquigarrow$

- (quasi-) convex optimization: (relaxed) Lasso

- greedy methods: Boosting
- "hierarchical" methods: PC-algorithm
  (Spirtes, Glymour, Scheines; for Graphical Modeling)

Other approaches include:

Ridge regression (Tikhonov regularization) for prediction

variable selection via AIC, BIC, (g)MDL (in a forward manner)

Bayesian methods for regularization, ...

computational feasibility for high-dimensional problems $\rightsquigarrow$

- (quasi-) convex optimization: (relaxed) Lasso

- greedy methods: Boosting
- "hierarchical" methods: PC-algorithm
  (Spirtes, Glymour, Scheines; for Graphical Modeling)

Lasso for linear models

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

$\rightsquigarrow$ convex optimization problem

- Lasso does variable selection
  some of the $\hat{\beta}_j(\lambda) = 0$
  (because of "ℓ^1-geometry")
- $\hat{\beta}(\lambda)$ is (typically) a shrunken LS-estimate

The prediction problem

Theorem (Greenshtein & Ritov, 2004)

- linear model with $p = p_n = O(n^\alpha)$ for some $\alpha < \infty$ (high-dimensional)
- $\|\beta\|_1 = \|\beta_n\|_1 = \sum_{j=1}^{p_n} |\beta_{j,n}| = o((n/\log(n))^{1/4})$ (sparse)
- other minor conditions

Then, for suitable $\lambda = \lambda_n$,

$$\mathbb{E}_X[(\underbrace{\hat{f}(X)}_{\hat{\beta}(\lambda)^T X} - \underbrace{f(X)}_{\beta^T X})^2] \longrightarrow 0 \text{ in probability } (n \to \infty)$$

and Lasso performs “quite well” for prediction

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

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%

multivariate gene selection

best 200 genes (Wilcoxon test)
no additional gene selection

Lasso selected on CV-average 13.12 out of $p = 7129$ genes

The variable selection problem

$$Y_i = (\beta_0 +) \sum_{j=1}^{p} \beta_j X_i^{(j)} + \epsilon_i,\ i = 1, \ldots, n$$

goal: find the effective predictor variables
i.e. the set $\mathcal{E}_{true} = \{j;\ \beta_j \neq 0\}$

ℓ^0-penalty methods, e.g. BIC, AIC,...

$$\hat{\beta}(\lambda) = \text{argmin}_\beta (n^{-1} \|Y - X\beta\|^2 + \lambda \underbrace{\|\beta\|_0}_{\sum_{j=1}^{p} I(\beta_j \neq 0)})$$

- computationally infeasible
  ad-hoc heuristic optimization such as forward-backward etc...
- often "instable" $\rightsquigarrow$ poor prediction (Breiman (1996, 1998))

convexization of computationally hard problem $\rightsquigarrow$ Lasso

$$\hat{\beta}(\lambda) = \text{argmin}_{\beta}(n^{-1}\|Y - X\beta\|^2 + \lambda\|\beta\|_1)$$

which does variable selection, i.e. $\hat{\beta}_j(\lambda) = 0$ for some j's

selected variables

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

$\rightsquigarrow$ can be computed efficiently for all λ's using the LARS algorithm (Efron, Hastie, Johnstone, Tibshirani, 2004)

$O(np\min(n,p))$ operation counts

linear in p if $p \gg n$

CPU time
lymph node classification example: $p = 7129$, $n = 49$

computing Lasso solutions for all λ's

2.603 seconds using `lars` in `R` (with `use.gram=F`)

with L_2Boosting (i.e. Gauss-Southwell)
for large range of solutions

it's less than a second!

0.906 seconds using `mboost` in `R` (PB & Hothorn, 2006)

C.F. Gauss in 1803

"Princeps Mathematicorum"

R.V. Southwell in 1933

Professor in Oxford

CPU time
lymph node classification example: $p = 7129$, $n = 49$

computing Lasso solutions for all λ's

2.603 seconds using `lars` in `R` (with `use.gram=F`)

with L_2Boosting (i.e. Gauss-Southwell)
for large range of solutions

it's less than a second!

0.906 seconds using `mboost` in `R` (PB & Hothorn, 2006)

C.F. Gauss in 1803

"Princeps Mathematicorum"

R.V. Southwell in 1933

Professor in Oxford

CPU time
lymph node classification example: $p = 7129$, $n = 49$

computing Lasso solutions for all λ's

2.603 seconds using `lars` in `R` (with `use.gram=F`)

with L_2Boosting (i.e. Gauss-Southwell)
for large range of solutions

it's less than a second!

0.906 seconds using `mboost` in `R` (PB & Hothorn, 2006)

C.F. Gauss in 1803

"Princeps Mathematicorum"

R.V. Southwell in 1933

Professor in Oxford

CPU time
lymph node classification example: $p = 7129$, $n = 49$

computing Lasso solutions for all λ's

2.603 seconds using `lars` in `R` (with `use.gram=F`)

with L_2Boosting (i.e. Gauss-Southwell)
for large range of solutions

it's less than a second!

0.906 seconds using `mboost` in `R` (PB & Hothorn, 2006)

C.F. Gauss in 1803

"Princeps Mathematicorum"

R.V. Southwell in 1933

Professor in Oxford

Properties of $\hat{\mathcal{E}}(\lambda)$

Theorem (Meinshausen & PB, 2004)

- Y, $X^{(j)}$'s Gaussian (not crucial)
- sufficient and almost necessary LfV condition
  (LfV = **L**asso **f**or **V**ariable selection); see also Zhao & Yu (2006)
- if $p = p(n)$ is growing with n
  - $p(n) = O(n^{\alpha})$ for some $0 < \alpha < \infty$ (high-dimensionality)
  - $|\mathcal{E}_{true,n}| = O(n^{\kappa})$ for some $0 < \kappa < 1$ (sparsity)
  - the non-zero β_j's are outside the $n^{-1/2}$-range

Then: if $\lambda = \lambda_n \sim const.n^{-1/2-\delta/2}$ $(0 < \delta < 1/2)$,

$$\mathbb{P}[\hat{\mathcal{E}}(\lambda) = \mathcal{E}_{true}] = 1 - O(\exp(-Cn^{1-\delta}))$$

statistical (asymptotic) justification of convexization of computationally hard problem for variable selection

Properties of $\hat{\mathcal{E}}(\lambda)$

Theorem (Meinshausen & PB, 2004)

- Y, $X^{(j)}$'s Gaussian (not crucial)
- sufficient and almost necessary LfV condition (LfV = **L**asso **f**or **V**ariable selection); see also Zhao & Yu (2006)
- if $p = p(n)$ is growing with n
  - $p(n) = O(n^{\alpha})$ for some $0 < \alpha < \infty$ (high-dimensionality)
  - $|\mathcal{E}_{true,n}| = O(n^{\kappa})$ for some $0 < \kappa < 1$ (sparsity)
  - the non-zero β_j's are outside the $n^{-1/2}$-range

Then: if $\lambda = \lambda_n \sim const.n^{-1/2-\delta/2}$ $(0 < \delta < 1/2)$,

$$\mathbb{P}[\hat{\mathcal{E}}(\lambda) = \mathcal{E}_{true}] = 1 - O(\exp(-Cn^{1-\delta}))$$

statistical (asymptotic) justification of convexization of computationally hard problem for variable selection

LfV condition is restrictive

sufficient and necessary for consistent model selection with Lasso

it fails to hold if design matrix is “too correlated”

$\Rightarrow$ Lasso is not consistent anymore for selecting the true model

The LfV condition: a condition on the covariance of X

$$\underbrace{\text{LfV condition}}_{\text{Meinshausen \& PB (2004)}} \quad \Leftrightarrow \quad \underbrace{\text{Irrepresentable condition}}_{\text{Zhao \& Yu (2006)}}$$

$''\Leftrightarrow''$ Lasso is consistent for variable selection

Irrepresentable condition $\Leftrightarrow |\hat{\Sigma}_{noise;eff}\hat{\Sigma}^{-1}_{eff;eff}\text{sign}(\beta_{eff})| \leq 1 - \eta$

it holds for

- $\hat{\Sigma}_{ij} \leq \rho^{|i-j|}$ $(0 \leq \rho < 1)$ power decay correlations
- dictionaries with $\underbrace{\text{coherence}}_{\text{max. correlation}} < (2p_{eff} - 1)^{-1}$
  (notion of coherence: Donoho, Elad & Temlyakov (2004))
- easy to construct examples where condition fails to hold

Choice of λ

first (not so good) idea: choose λ to optimize prediction
e.g. via some cross-validation scheme

but: for prediction oracle solution

$$\lambda^* = \text{argmin}_\lambda \mathbb{E}[(Y - \sum_{j=1}^{p} \hat{\beta}_j^{(}\lambda) X^{(j)})^2]$$

$\mathbb{P}[\hat{\mathcal{E}}(\lambda^*) = \mathcal{E}_{true}] < 1 \ (n \to \infty)$ (or $= 0$ if $p_n \to \infty \ (n \to \infty)$)

asymptotically: prediction optimality yields too large models
(Meinshausen & PB, 2004; related example by Leng et al., 2004)

in summary:

- prediction optimal solution yields asymptotically too large models
- if LfV condition fails to hold (and assuming weaker conditions)
  Lasso yields models which contain the true model

$\rightsquigarrow$ Lasso as a
"filter for variable selection"

i.e. true model is contained in selected models from Lasso

Binary lymph node classification in breast cancer: $n = 49$ $p = 7130$

5-fold CV tuned Lasso selects 23 genes (on whole data set)

note (in practice): identifiability problem among highly correlated predictor variables

$\rightsquigarrow$ an ad-hoc approach:
keep the 23 plus all its highly correlated genes for further modeling, interpretation etc...

From filtering to selection of variables

with Lasso, we obtain sequence of sub-models

$$\widehat{\mathcal{SUB}} = \{\hat{\mathcal{E}}(\lambda_r);\ 1 \leq r \leq \underbrace{r_{max}}_{=O(\min(n,p))}\},\ \lambda_1 = 0 < \lambda_2 < \ldots < \lambda_{max}$$

typically

$$\hat{\mathcal{E}}(\lambda_{max}) \subset \ldots \subset \hat{\mathcal{E}}(\lambda_2) \subset \hat{\mathcal{E}}(\lambda_1)$$

assuming the LfV and other conditions:
with high probability,

$$\mathcal{E}_{true} \in \widehat{\mathcal{SUB}},$$
$$(\text{and } \mathcal{E}_{true} \subseteq\ \hat{\mathcal{E}}(\lambda^*))$$

$\rightsquigarrow$ we only need a good selector within $\widehat{\mathcal{SUB}}$

first (empirically not so good idea): choose best model in $\widehat{\mathcal{SUB}}$ using BIC or related method

better:
use the Lasso again for the models in $\widehat{\mathcal{SUB}}$:

$$\underbrace{\hat{\mathcal{E}}(\lambda_{max})}_{\rightsquigarrow \text{Lasso again}} \quad \underbrace{\hat{\mathcal{E}}(\lambda_{r_{max}-1})}_{\rightsquigarrow \text{Lasso again}} \quad \cdots \quad \underbrace{\hat{\mathcal{E}}(\lambda_2)}_{\rightsquigarrow \text{Lasso again}} \quad \underbrace{\hat{\mathcal{E}}(\lambda_1)}_{\rightsquigarrow \text{Lasso again}}$$

this is the Relaxed Lasso (Meinshausen, 2005)

first (empirically not so good idea): choose best model in $\widehat{\mathcal{SUB}}$ using BIC or related method

better:
use the Lasso again for the models in $\widehat{\mathcal{SUB}}$:

$$\underbrace{\hat{\mathcal{E}}(\lambda_{max})}_{\rightsquigarrow \text{Lasso again}} \quad \underbrace{\hat{\mathcal{E}}(\lambda_{r_{max}-1})}_{\rightsquigarrow \text{Lasso again}} \quad \cdots \quad \underbrace{\hat{\mathcal{E}}(\lambda_2)}_{\rightsquigarrow \text{Lasso again}} \quad \underbrace{\hat{\mathcal{E}}(\lambda_1)}_{\rightsquigarrow \text{Lasso again}}$$

this is the Relaxed Lasso (Meinshausen, 2005)

Relaxed Lasso

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

$$\hat{\beta}_{\lambda,\phi} = \text{argmin}_{\beta}(n^{-1}\sum_{i=1}^{n}(Y_i - \sum_{j \in \hat{\mathcal{E}}(\lambda)} \beta_j X_i^{(j)})^2 + \phi\lambda\|\beta\|_1)$$

for $\phi = 0$: OLS on selected variables from Lasso(λ)
for $\phi = 1$: Lasso(λ)

amount of computation for finding all solutions over λ and ϕ:
often, the same computational complexity as for Lasso/LARS:

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

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

Relaxed Lasso

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

$$\hat{\beta}_{\lambda,\phi} = \text{argmin}_{\beta}(n^{-1}\sum_{i=1}^{n}(Y_i - \sum_{j \in \hat{\mathcal{E}}(\lambda)} \beta_j X_i^{(j)})^2 + \phi\lambda\|\beta\|_1)$$

for $\phi = 0$: OLS on selected variables from Lasso(λ)
for $\phi = 1$: Lasso(λ)

amount of computation for finding all solutions over λ and ϕ:
often, the same computational complexity as for Lasso/LARS:

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

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

Relaxed Lasso

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

$$\hat{\beta}_{\lambda,\phi} = \text{argmin}_{\beta}(n^{-1}\sum_{i=1}^{n}(Y_i - \sum_{j \in \hat{\mathcal{E}}(\lambda)} \beta_j X_i^{(j)})^2 + \phi\lambda\|\beta\|_1)$$

for $\phi = 0$: OLS on selected variables from Lasso(λ)
for $\phi = 1$: Lasso(λ)

amount of computation for finding all solutions over λ and ϕ:
often, the same computational complexity as for Lasso/LARS:

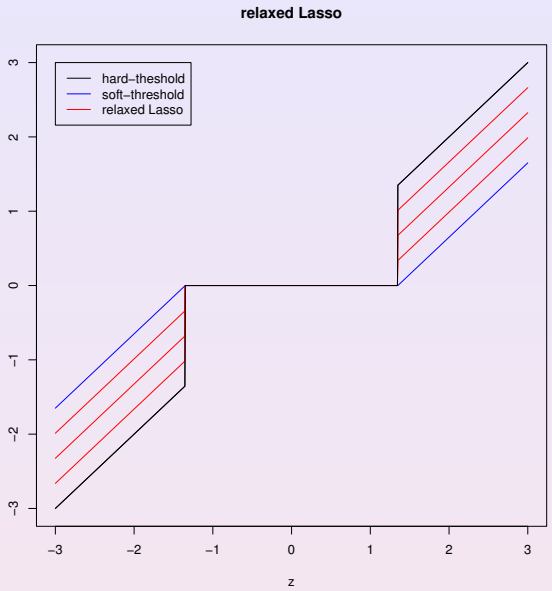
$$O(np\min(n,p)) = O(p) \text{ if } p \gg n$$

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

Properties of the relaxed Lasso

for orthonormal case:
$\mathbf{X}^T\mathbf{X} = I$

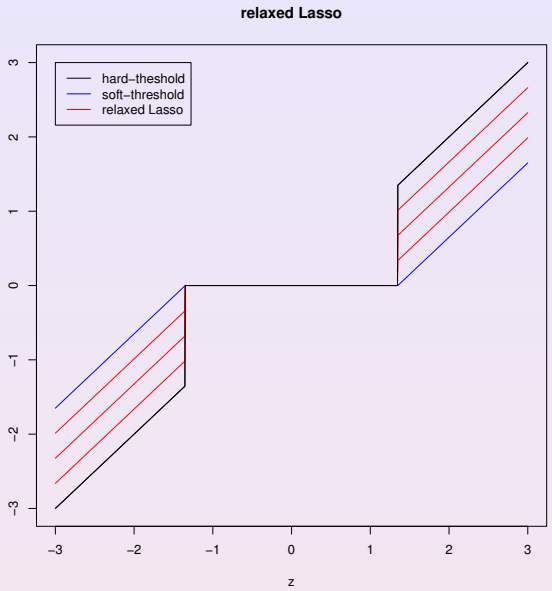


for general case:
assume the LfV and other “Lasso conditions”

prediction optimal tuned relaxed Lasso
is consistent for variable selection
(Meinshausen, 2005)

Properties of the relaxed Lasso

for orthonormal case:
$\mathbf{X}^T\mathbf{X} = I$



for general case:
assume the LfV and other “Lasso conditions”

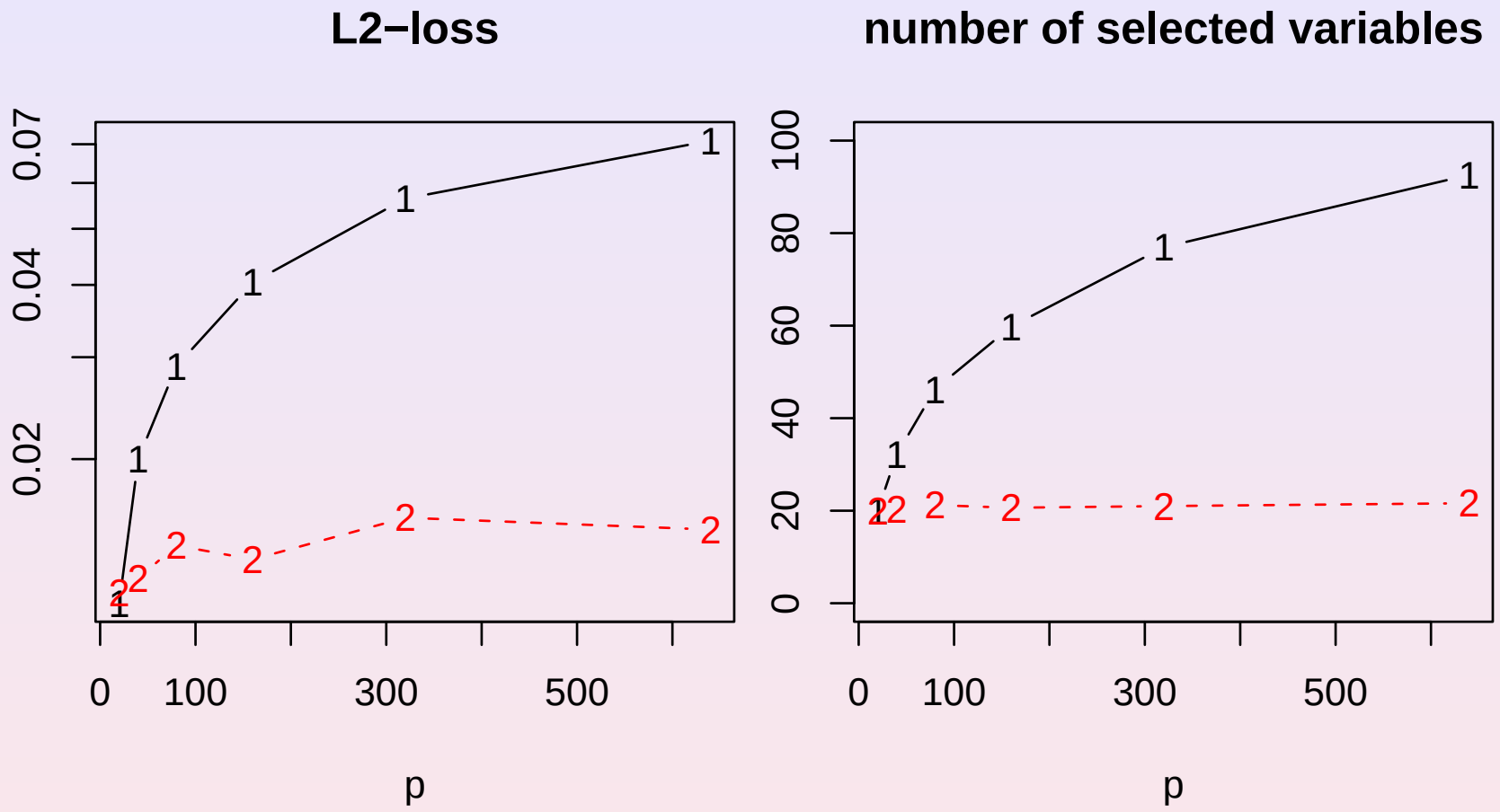
prediction optimal tuned relaxed Lasso
is consistent for variable selection
(Meinshausen, 2005)

for very high-dimensional case:

- $p = p_n \sim C_1 \exp(C_2 n^{1-\xi})$ $(0 < \xi < 1)$
- effective number of variables is finite (finite ℓ^0-norm)
  non-effective variables are independent
- "Lasso assumptions" from before

Lasso has very slow MSE convergence rate (depending on ξ)
Relaxed Lasso has MSE rate $O(n^{-1})$
(Meinshausen, 2005)

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



1: Lasso 2: relaxed Lasso

additional pure noise variables are much less damaging with the relaxed Lasso than for Lasso

for prediction:

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: $\phi = 1$)

for variable selection:

Relaxed Lasso almost always sparser than Lasso

Binary lymph node classification in breast cancer: $n = 49$ $p = 7130$

5-fold CV tuning for each method

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

	misclassif. error	number of selected genes
Lasso	21.1%	13.12
Relaxed Lasso	20.1%	7.3

Binary lymph node classification in breast cancer: $n = 49$ $p = 7130$

5-fold CV tuning for each method

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

	misclassif. error	number of selected genes
Lasso	21.1%	13.12
Relaxed Lasso	20.1%	7.3

DNA splice site detection

DNA sequence

$$\ldots ACGGC \ldots \quad \underbrace{NNN \quad \underbrace{GC}_{\text{potential donor site}} \quad NNNN}_{\text{3 positions exon } GC \text{ 4 positions intron}} \quad \ldots AAC \ldots$$

response $Y \in \{0, 1\}$: splice or non-splice site
predictor variables: 7 factors each having 4 levels
(full dimension: $4^7 = 16'384$)
data: $p = 16'384$, $n = 11'220$

logistic regression:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \text{ main effects} + \text{first order interactions} + \ldots$$

with sum-to-zero constraints

use "Lasso" which selects whole terms
instead of selection of dummy indicator variables
e.g. the interaction term between factor 2 and 5 (which is encoded with 9 free parameters/dummy indicators)

$\rightsquigarrow$ Group Lasso (Yuan and Lin (2006), for Gaussian regression)

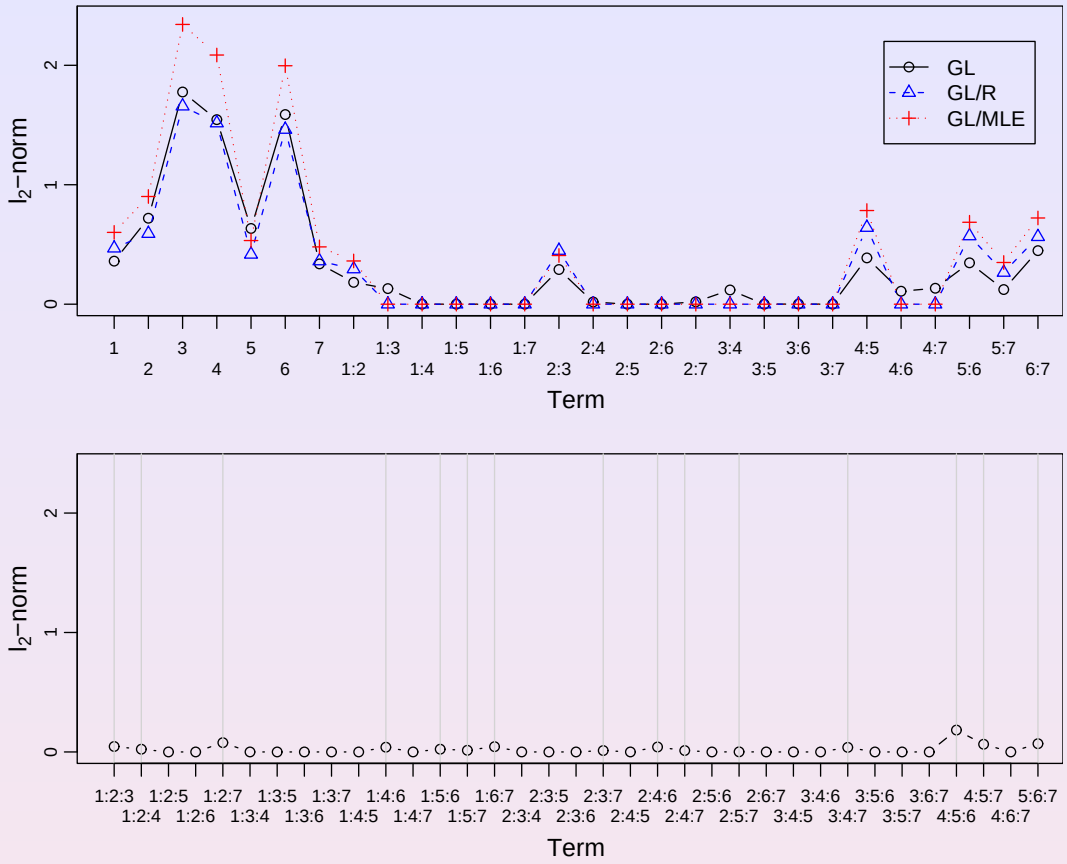
$$\text{penalty: } \lambda \sum_{\text{term } j} \|\beta_j\|_2$$

- invariant under orthonormal parameter transformations
- if term j is of dimension 1: $\|\beta_j\|_2 = \|\beta_j\|_1$

- new efficient algorithms are needed for Group Lasso with binomial likelihood
  $\rightsquigarrow$ Block gradient descent with tight approximations for the Hessian
- theory and methodology for high-dimensions: "similar" as for the Lasso

(Meier, v.d. Geer & PB, 2006)

Group Lasso/Ridge: in spirit of the Relaxed Lasso
1st stage: Group Lasso for logistic regression
2nd stage: Ridge logistic regression on models from 1st stage
$\rightsquigarrow$ allows for hierarchical model fitting
$\rightsquigarrow$ better term selection and better prediction than Group Lasso



- mainly neighboring DNA positions show interactions (has been “known” and “debated”)
- no interaction among exon and intron positions (with Group Lasso/Ridge method)
- no second-order interactions (with Group Lasso/Ridge method)

predictive power:
competitive with “state to the art” maximum entropy modeling from Yeo and Burge (2004)

correlation between true and predicted class

Logistic Group Lasso/Ridge	0.6593
max. entropy (Yeo and Burge)	0.6589

- our model is simple (not necessarily the method/algorithm) and has clear interpretation
- it is as good or better than many of the complicated non-Markovian stochastic process models (e.g. Zhao, Huang and Speed (2004))

Conclusions

especially for high-dimensional data:

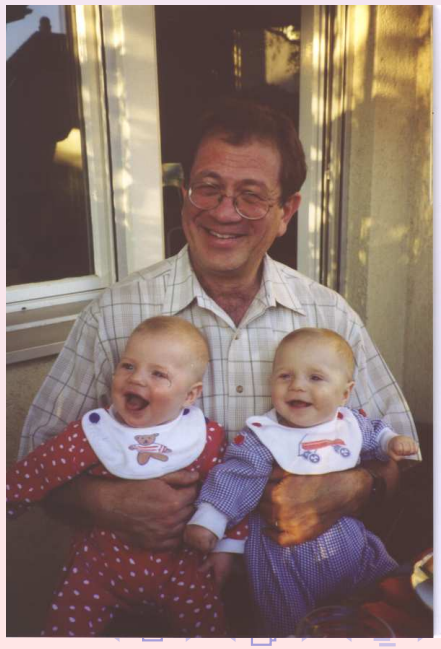
- Lasso useful for variable filtering
  it is computationally attractive: linear in dimensionality p
  the "true model" is contained in the solution set of Lasso

- Relaxed Lasso (or similar two stage procedures):
  - often better prediction than Lasso
  - optimal prediction penalty yields consistent model selection
  - sparser solutions than Lasso

Thank you Peter

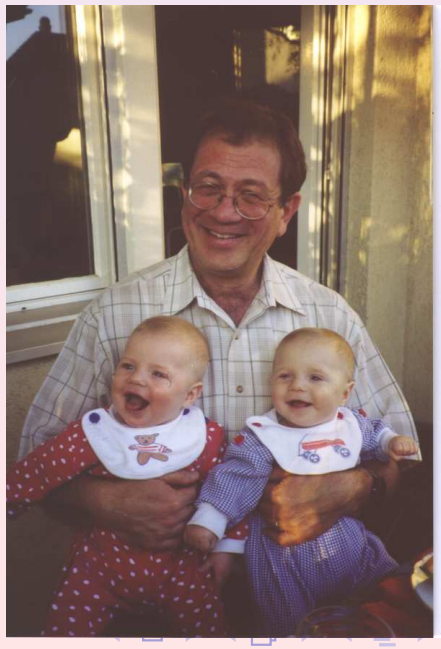
- for having been a great

  (my greatest!) post-doc mentor
- for your genuine interest in understanding what others (including myself) are thinking and doing
- that Nancy and you have always been like parents and friends to my whole family



Thank you Peter

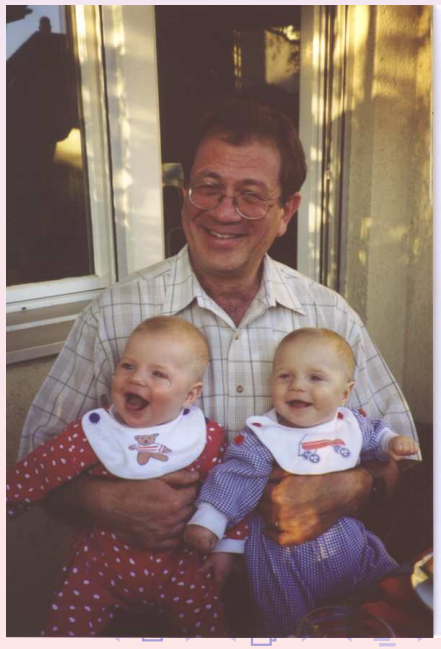
- for having been a great (my greatest!) post-doc mentor
- for your genuine interest in understanding what others (including myself) are thinking and doing
- that Nancy and you have always been like parents and friends to my whole family



Thank you Peter

- for having been a great

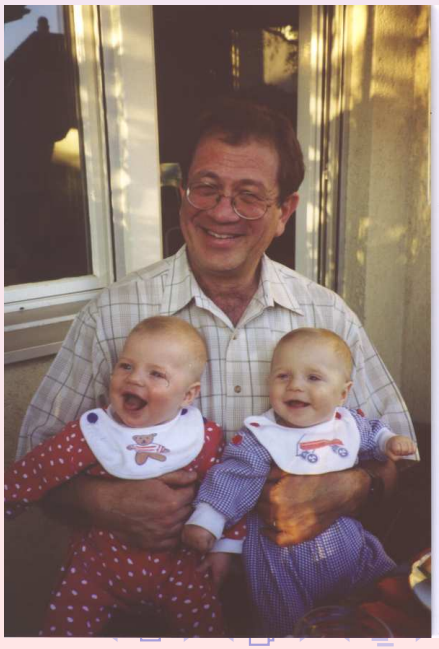
  (my greatest!) post-doc mentor
- for your genuine interest in understanding what others (including myself) are thinking and doing
- that Nancy and you have always been like parents and friends to my whole family



Thank you Peter

- for having been a great

  (my greatest!) post-doc mentor
- for your genuine interest in understanding what others (including myself) are thinking and doing
- that Nancy and you have always been like parents and friends to my whole family



Thank you Peter

- for having been a great

  (my greatest!) post-doc mentor
- for your genuine interest in understanding what others (including myself) are thinking and doing
- that Nancy and you have always been like parents and friends to my whole family