Variable selection based on multiple, high-dimensional genomic data: from the Lasso to the smoothed adaptive Lasso

Peter Bühlmann

Seminar für Statistik, ETH Zürich

January 2007

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

high-dimensional: $p \gg n$

areas of application: biology, astronomy, 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 response variable, e.g. $Y_i \in \mathbb{R}$ or $Y_i \in \{0, 1\}$

high-dimensional: $p \gg n$

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

Some examples from biology

1. Classification of cancer sub-types based on microarray gene expression data
$X =$ gene expression profile
$Y \in \{0, 1, \ldots, J-1\}$ the class-label of cancer sub-type
$n \approx 10 - 100, \; p \approx 3'000 - 25'000$

2. Motif regression:
search for transcription factor binding site on DNA sequence using gene expressions and DNA sequence data

motif = (overrepresented) pattern on DNA sequence
(transcription factor binding site)

data:
$X =$ motif scores for motifs up-stream of single genes based on sequence data only (e.g. MDscan from Liu et al.)
$Y =$ gene expression for single genes, over multiple time points
$p \approx 4'000, \; n \approx 20 \times 4'000 = 80'000$

Some examples from biology

1. Classification of cancer sub-types based on microarray gene expression data
$X =$ gene expression profile
$Y \in \{0, 1, \ldots, J-1\}$ the class-label of cancer sub-type
$n \approx 10 - 100, \; p \approx 3'000 - 25'000$

2. Motif regression:
search for transcription factor binding site on DNA sequence using gene expressions and DNA sequence data

motif = (overrepresented) pattern on DNA sequence
(transcription factor binding site)

data:
$X =$ motif scores for motifs up-stream of single genes based on sequence data only (e.g. MDscan from Liu et al.)
$Y =$ gene expression for single genes, over multiple time points
$p \approx 4'000, \; n \approx 20 \times 4'000 = 80'000$

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 $\neq 0$)

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

(2^p sub-models) $\rightsquigarrow$

$$\Leftrightarrow \quad \begin{array}{l}\text{(quasi-) convex optimization}\\ \text{(adaptive)} \qquad \underbrace{\text{Lasso}}_{\text{Tibshirani (1996)}}\end{array}$$

Lasso for linear models

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

$\rightsquigarrow$ 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 $p = 7130$ 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: 2^p sub-models
  ad-hoc heuristic optimization such as forward-backward
- often “instable” (Breiman (1996, 1998))

convexization of computationally hard problem $\rightsquigarrow$

use the Lasso for variable selection : $\hat{\mathcal{E}}(\lambda) = \{j;\ \hat{\beta}_j(\lambda) \neq 0\}$

$\rightsquigarrow$ 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: $p = 7130$, $n = 49$

computing Lasso solutions for all λ's

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

Properties of Lasso for variable selection

Theorem (Meinshausen & PB, 2004 (publ: 2006))

- $Y, \ X^{(j)}$'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 Lasso for variable selection

Theorem (Meinshausen & PB, 2004 (publ: 2006))

- $Y, \ X^{(j)}$'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

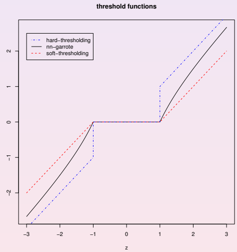
it fails to hold if design matrix is "too correlated"

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

The “reason”

too much bias – shrinkage even for large values

for orthogonal design:



Bias in soft-thresholding
is disturbing (at least sometimes)

better:
Nonnegative Garrote (Breiman, 1995)
and similar proposals

The LfV condition: a condition on the covariance of X

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

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

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

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

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

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

asymptotically: prediction optimality yields too large models
(Meinshausen & PB, 2004; related example by Leng et al., 2006)

If LfV condition fails to hold:

Meinshausen & Yu (2006): for suitable $\lambda = \lambda_n$

$$\|\hat{\beta} - \beta\|_2^2 = \sum_{j=1}^{p} (\hat{\beta}_j - \beta_j)^2 = o_P(1)$$

under much weaker conditions than LfV

- maximal and minimal sparse eigenvalues of empirical covariance matrix
- number of effective variables in relation to sparse eigenvalues of empirical covariance matrix

implication: Lasso yields too large models
(for fixed coefficients $\beta_j,\ j = 1, 2, \ldots$)

in summary: asymptotically,

- prediction optimal solution yields too large models
- if LfV condition fails to hold
  Lasso yields too large models

⇝ Lasso as a
“filter for variable selection”

i.e. true model is contained in selected models from Lasso in summary: asymptotically,

- prediction optimal solution yields too large models
- if LfV condition fails to hold
  Lasso yields too large models

⇝ 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...

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

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

Adaptive Lasso

recap: under “weak” assumptions,

$$\mathcal{E}_{true} \subseteq \hat{\mathcal{E}}(\underbrace{\hat{\lambda}}_{\text{pred. optim.}})$$

quite many non-zero, “small” $\hat{\beta}_j$’s from the Lasso

$\rightsquigarrow$ various possibilities to improve:

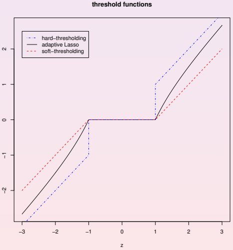
- hard-thresholding of coefficients
  (using prediction optimality)
- thresholding of coefficients and re-estimation of non-zero coefficients with least squares (using prediction optimality)

Adaptive Lasso (Zou, 2006): re-weighting the penalty function

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

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

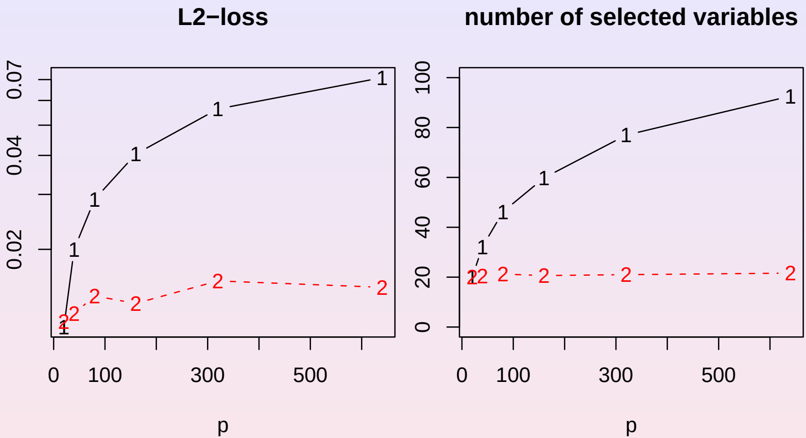
for orthogonal design,
if $\hat{\beta}_{init}$ = OLS:
Adaptive Lasso = NN-garrote



furthermore:

- Zou (2006): adaptive Lasso is consistent for variable selection “in general”
  (proof for low-dimensional problems only)
- Huang, Ma & Zhang (2006): as above but for sparse, high-dimensional problems

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



1: Lasso 2: adaptive Lasso

additional pure noise variables are much less damaging with the adaptive Lasso than for 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
Adaptive 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
Adaptive Lasso	20.1%	7.3

Bacillus Subtilis for vitamin production (project with DSM)

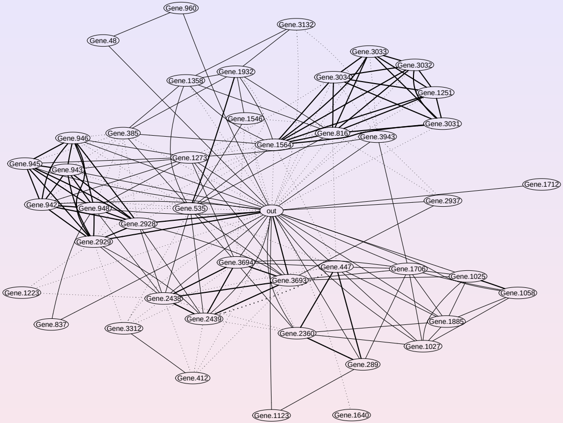
data: response Y, $p = 4088$ gene expressions, $n = 115$

goal: find important genes for Y

statistically:
regression problem Y versus $p = 4088$ gene expressions
find the variables (genes) which are important for regression

identifiability problem due to high correlation (collinearity) among genes
$\rightsquigarrow$ elastic net (Zou & Hastie, 2005) which encourages to select non or all among highly correlated predictor variables

Adaptive elastic net (builds on the idea of the adaptive Lasso)



regression coefficients ⇔ partial correlations

⇝ Gaussian graphical model

useful visualization: neighbours of Y (selected genes) and their conditional dependencies (w.r.t. all variables (genes)) we did not insist on sparsity

but aimed for all relevant and their highly correlated variables/genes

⇝ more “false positives”, but sometimes desirable in exploratory stage

Can we improve?

adaptive Lasso yields pretty good solutions for variable selection and prediction

but note the limitation: $p \gg n$ ...

“strategy”:

make “sample size” larger by integrating other suitable data-sets

a simple model:

data-sets $D(t),\ t = 1, 2, \ldots, N$

each measuring $Y(t), X(t)$ with sample size $n(t)$

$Y(t) = X(t)\beta(t) + \epsilon(t)$,

$$\beta(t) \quad \underbrace{\text{smoothly changing over } t}_{\text{topology for indexing data-sets}}$$

Can we improve?

adaptive Lasso yields pretty good solutions for variable selection and prediction

but note the limitation: $p \gg n$ ...

“strategy”:

make “sample size” larger by
integrating other suitable data-sets

a simple model:

data-sets $D(t),\ t = 1, 2, \ldots, N$

each measuring $Y(t), X(t)$ with sample size $n(t)$

$$Y(t) = X(t)\beta(t) + \epsilon(t),$$

$$\beta(t) \quad \underbrace{\text{smoothly changing over } t}_{\text{topology for indexing data-sets}}$$

Time course experiments

$t = 1, 2, \ldots, N$ represents time
$Y(t)$ and $X(t)$ measurements for the same variables
$\rightsquigarrow$ use usual metric on $\mathbb{R}^+$

use smoothed Lasso (Meier & PB (in progress))

$$\hat{\beta}(\tau) = \text{argmin}_{\beta} \sum_{t=1}^{T} \underbrace{K(\frac{t-\tau}{h})}_{\text{weight } w(t,\tau)} \; (n^{-1}\|Y(t) - X(t)\beta\|_2^2 + \lambda\|\beta\|_1)$$

(if $n(t) \equiv n$)

Time course experiments

$t = 1, 2, \ldots, N$ represents time
$Y(t)$ and $X(t)$ measurements for the same variables
$\leadsto$ use usual metric on $\mathbb{R}^+$

use smoothed Lasso (Meier & PB (in progress))

$$\hat{\beta}(\tau) = \text{ argmin}_{\beta} \sum_{t=1}^{T} \underbrace{K(\frac{t-\tau}{h})}_{\text{weight } w(t,\tau)} \; (n^{-1}\|Y(t) - X(t)\beta\|_2^2 + \lambda\|\beta\|_1)$$

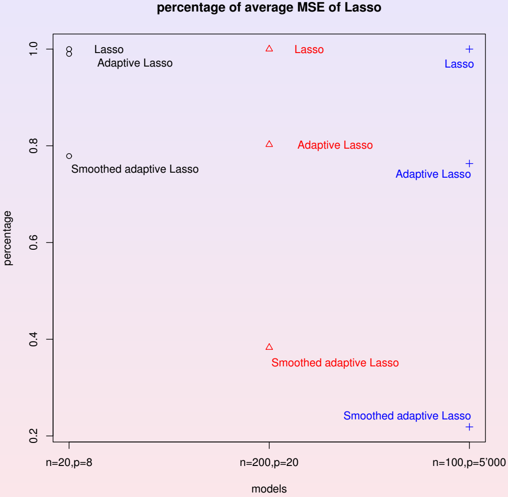
(if $n(t) \equiv n$)

results for the smoothed (adaptive) Lasso (Meier & PB):
if $h = h_N \to 0$ suitably slowly and $\beta(\cdot)$ is smooth: for suitable $\lambda = \lambda(n, N, h)$:

- improved convergence rate for $\|\hat{\beta} - \beta\|_2^2$ by a factor $(Nh)^{-a}$ (for smoothed Lasso and smoothed adaptive Lasso)

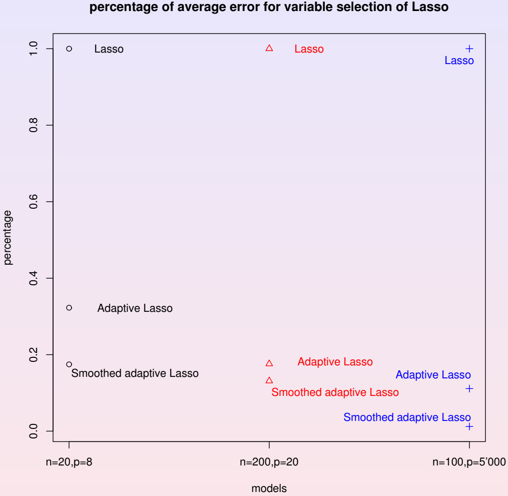
  $a = 1$ and $(Nh_{opt})^{-1} = N^{-4/5}n^{1/5}$ for low-dimensional case
  i.e. improvement if N not too small w.r.t. n ($N/n^{1/4} \to \infty$)
  e.g: $n^{1/4} \approx 2.7$ if $n = 50$ and $n^{1/4} \approx 8$ if $n = 4'000$
- for the smoothed adaptive Lasso:
  asymptotic consistency for variable selection (as for non-smoothed case), but better empirical performance

$$\mathbb{E}[N^{-1}\sum_{t=1}^{N}\|\hat{\beta}(t)-\beta(t)\|_2^2]$$



$N = 9$ $(n = 20)$ or $N = 18$ $(n = 100, 200)$

$$\mathbb{E}[|\hat{p}_{eff} - p_{eff}|]$$



and smoothed adaptive Lasso is sparsest

$N = 9$ ($n = 20$) or $N = 18$ ($n = 100, 200$)

Motif regression for time-course experiments

goal: find transcription factor binding sites
  (for a set of co-regulated genes)
fact: a transcription factor tends to recognize
  a conserved pattern (a “motif”) in DNA sequence

$\leadsto$ search for “overrepresented patterns” such as TCTATTGTTT occurring in up-stream region of gene(s)

MotifRegressor which integrates sequence and gene expression data (Conlon, Liu, Lieb & Liu, 2003):

- for highly expressed genes:
  up-stream of each gene, search for p candidate motifs with MDscan, based on DNA sequence data only
- compute motif-score for all n genes and all p candidate motifs
  (score $\approx$ occurrences of candidate motif in gene's up-stream region)
  based on DNA sequence data only
- $n \approx 4'000 - 25'000$ genes and their expression Y,
  $p \approx 4'000$ candidate motif-scores for each gene
  gene expression and DNA sequence data
- do regression and determine the significant variables (i.e. candidate motifs which are significant):

$$Y_i = \text{ gene-expression of gene}_i = \sum_{j=1}^{p} \beta_j \underbrace{X_{i,j}}_{\text{motif score}} + \text{error}_i$$

this approach is very competitive in comparison to other

“always”: very noisy data
after variable selection: $R^2 \approx 0.05 - 0.15$

nevertheless:
MotifRegressor seems often better than competitive algorithms

further improvement with adaptive Lasso over forward variable selection in MotifRegressor
and it yields meaningful or even true findings (work in progress with Liu and collaborators)

“always”: very noisy data
after variable selection: $R^2 \approx 0.05 - 0.15$

nevertheless:
MotifRegressor seems often better than competitive algorithms

further improvement with adaptive Lasso over forward variable selection in MotifRegressor
and it yields meaningful or even true findings (work in progress with Liu and collaborators)

“always”: very noisy data
after variable selection: $R^2 \approx 0.05 - 0.15$

nevertheless:
MotifRegressor seems often better than competitive algorithms

further improvement with adaptive Lasso over forward variable selection in MotifRegressor
and it yields meaningful or even true findings (work in progress with Liu and collaborators)

Time-course experiments (e.g. from cell-cycle)

for every time point t:

- gene-expression vector/profile $Y(t)$
- motif-scores for every gene and every motif-candidate: always the same, i.e. $X(t) \equiv X$

multivariate regression:

$$Y(t) = X\beta(t) + \text{ error}(t)$$

and reasonable assumption that $\beta(\cdot)$ changes smoothly w.r.t. time

$\rightsquigarrow$ use the smoothed adaptive Lasso

Spellman et al.'s cell-cycle experiment for yeast

$N = 9$ time points
$n = 4443, \; p = 2155$

cross-validated mean squared prediction error:

time point	adaptive Lasso	smoothed adaptive Lasso
1	40.5	41.7
2	95.2	95.5
3	60.0	61.4
4	59.3	59.0
5	32.3	32.4
6	36.5	36.5
7	37.3	37.3
8	27.1	27.1
9	29.9	29.9

$\leadsto$ essentially the same predictive performance
but note: high noise $\Rightarrow$ similar prediction performance

number of selected variables (motifs):

time point	adaptive Lasso	smoothed adaptive Lasso
1	500	438
2	73	73
3	43	20
4	77	46
5	53	41
6	0	0
7	0	0
8	45	16
9	0	0

⇝ smoothed adaptive Lasso often substantially sparser
fewer false positives expected

interpretation of significant motifs (via TRANSFAC):
e.g.

$$\text{GACGCG} \xrightarrow{\textit{TRANSFAC}} \underbrace{\text{MCB}}_{\text{trans. factor}}$$

some well known cell-cycle regulators:
STE12, SCB, MCB, PH04, SW15, MCM1

and in addition: ROX1, M3B, XBP1

- substantial overlap of findings with Conlon, Liu, Lieb & Liu (2003)
- our method is much more stable than (non-smoothed) forward variable selection used in Conlon, Liu, Lieb & Liu (2003)
  ⇝ fewer false positives expected with smoothed adaptive Lasso

currently working on motif finding in Arabidopsis Thaliana
(much less explored organism then yeast)

with Gruissem lab at ETH Zürich

Conclusions

1. Lasso is computationally attractive for variable selection in high-dimensional generalized linear models
   (including e.g. Cox's partial likelihood for survival data)
   but: it yields too large models
2. Adaptive Lasso is an elegant, effective way to correct Lasso's overestimation behavior
3. Smoothed adaptive Lasso is potentially powerful for time-course experiments
   (or multivariate structures, i.e. "multiple data-sets")

software packages are available in `R`:
`lars`, `glmppath`, `grplasso`