Variable selection in high-dimensional regression problems

Jan Mielniczuk

Polish Academy of Sciences and Politechnika Warszawska

Based on joint research with P. Pokarowski, A. Prochenka, P. Teisseyre and M. Kubkowski

LASSO

Least **A**bsolute **S**hrinkage and **S**election Operator:

$$\widehat{\beta}_L \equiv \widehat{\beta}_L(\lambda) = \text{argmin}_\beta \left\{ \|y - X\beta\|^2 + 2\lambda|\beta|_1 \right\}$$

Lasso based procedures of selecting predictors under high dimensionality

Outline

- Introduction: Penalized empirical risk minimisation
- Variable selection for linear and logistic regression
- Variable selection for misspecified logistic model - some comments

Penalized Empirical Risk Minimization (PERM)

Data form: (y, x^T): y- response (quantitative or nominal), $x = (x_1, \ldots, x_p)^T \in R^p$: vector of predictors.
Penalized risk minimization framework:

$$\text{Data} = \{(y_1, x_{1\cdot}^T), \ldots, (y_n, x_{n\cdot}^T)\} = \text{Train} \oplus \text{Valid} \oplus \text{Test}$$

β- model parameter, λ - penalty

Fitting: $\widehat{\beta}(\lambda) = \arg\min_{\beta}\{\text{err}(\beta, \text{Train}) + \text{penalty}(\beta, \lambda)\}$

Penalized Empirical Risk Minimization (PERM)

Data form: (y, x^T): y- response (quantitative or nominal), $x = (x_1, \ldots, x_p)^T \in R^p$: vector of predictors.
Penalized risk minimization framework:

$$\text{Data} = \{(y_1, x_{1\cdot}^T), \ldots, (y_n, x_{n\cdot}^T)\} = \text{Train} \oplus \text{Valid} \oplus \text{Test}$$

β- model parameter, λ - penalty

Fitting: $\widehat{\beta}(\lambda) = \arg\min_{\beta}\{\text{err}(\beta, \text{Train}) + \text{penalty}(\beta, \lambda)\}$

Selection: $\widehat{\lambda} = \arg\min_{\lambda} \overline{\text{err}}\left(\widehat{\beta}(\lambda), \text{Valid}\right)$

Penalized Empirical Risk Minimization (PERM)

Data form: (y, x^T): y- response (quantitative or nominal), $x = (x_1, \ldots, x_p)^T \in R^p$: vector of predictors.
Penalized risk minimization framework:

$$\text{Data} = \{(y_1, x_{1\cdot}^T), \ldots, (y_n, x_{n\cdot}^T)\} = \text{Train} \oplus \text{Valid} \oplus \text{Test}$$

β- model parameter, λ - penalty

Fitting: $\widehat{\beta}(\lambda) = \arg\min_{\beta}\{\text{err}(\beta, \text{Train}) + \text{penalty}(\beta, \lambda)\}$

Selection: $\widehat{\lambda} = \arg\min_{\lambda} \overline{\text{err}}\left(\widehat{\beta}(\lambda), \text{Valid}\right)$

Assessment: $\widehat{\text{err}} = \overline{\text{err}}\left(\widehat{\beta}(\widehat{\lambda}), \text{Test}\right)$

Penalized Empirical Risk Minimization

Empirical risk *err* is generalization of prediction error and negative log-likelihood

$$\mathrm{err}(\beta, Train) = \sum_{i=1}^{n} L(y_i, f(x_{i\cdot}, \beta))$$

which is (usually) a **convex** function of β. $L(y, f)$: loss function.

Penalized Empirical Risk Minimization

Empirical risk *err* is generalization of prediction error and negative log-likelihood

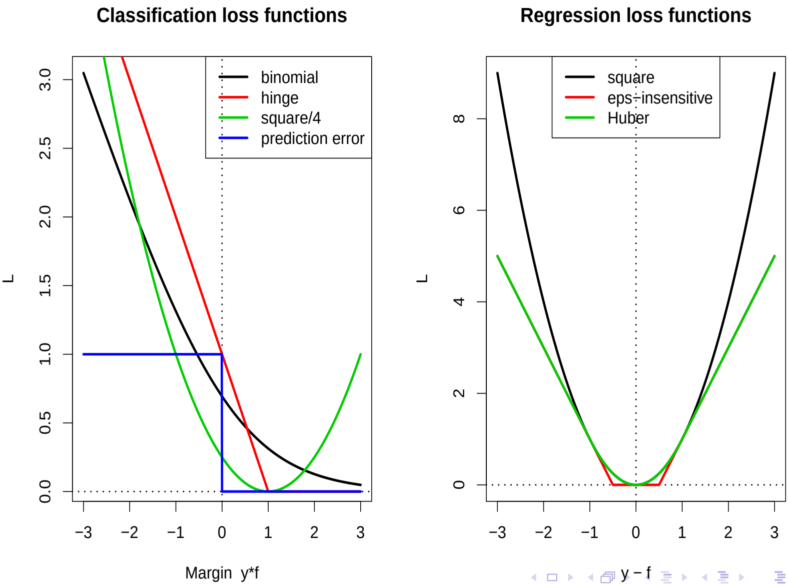
$$\mathrm{err}(\beta, Train) = \sum_{i=1}^{n} L(y_i, f(x_{i\cdot}, \beta))$$

which is (usually) a **convex** function of β. $L(y, f)$: loss function.

$$\mathrm{penalty}(\beta, \lambda) = \sum_{j=1}^{p} P_\lambda(|\beta_j|)$$

$\beta = (\beta_1, \ldots, \beta_p)^T$

$$\lambda \mathbb{1}(t > 0) \preceq P_\lambda(t) \preceq \lambda t^2$$


Classification loss functions
binomial
hinge
square/4
prediction error
L
Margin y*f
Regression loss functions
square
eps−insensitive
Huber
L
y − f

Classical Penalty Functions

Ridge Regression $\equiv \ell_2$-penalty (Hoerl and Kennard (1970))

$$P_\lambda(t) = \lambda t^2$$

Ridge Regression $\equiv \ell_2$-penalty (Hoerl and Kennard (1970))

$$P_\lambda(t) = \lambda t^2$$

Generalized Information Criterion (GIC $\ni$ AIC,BIC)
$\equiv \ell_0$-penalty (Nishi (1970))

$$P_\lambda(t) = 2\lambda \mathbb{1}(t > 0)$$

Chen, Donoho, 1995, Tibshirani, 1996:

Classical Penalty Functions

Ridge Regression $\equiv \ell_2$-penalty (Hoerl and Kennard (1970))

$$P_\lambda(t) = \lambda t^2$$

Generalized Information Criterion (GIC $\ni$ AIC,BIC) $\equiv \ell_0$-penalty (Nishi (1970))

$$P_\lambda(t) = 2\lambda \mathbb{1}(t > 0)$$

Chen, Donoho, 1995, Tibshirani, 1996: *Lasso* $\equiv \ell_1$-penalty

$$P_\lambda(t) = \lambda t$$

Important for high-dimensional problems: **sparseness of the solution** for Lasso induced by $P'_\lambda(0^+) > 0$.

Validation criteria

Choice of penalty:

$$\widehat{\lambda} = \arg\min_{\lambda} \overline{\text{err}}\left(\widehat{\beta}(\lambda), \text{Valid}\right)$$

- $\overline{\text{err}}(\widehat{\beta}) = \widehat{E}||\widehat{\beta} - \beta||^2$ (estimation error)
- $\overline{\text{err}}(\widehat{\beta}) = \widehat{E}||X(\widehat{\beta} - \beta)||^2$ (prediction error)
- $\overline{\text{err}}(\widehat{\beta}) = \widehat{P}(yx^T\hat{\beta} < 0)$ (classification error)
- $\overline{\text{err}}(\widehat{\beta}) = \widehat{P}(\mathrm{supp}\widehat{\beta} \neq \mathrm{supp}\beta)$ (**selection error**)
- others: FDR control etc.

Selection consistency

$$P(\hat{T} \neq T) \text{ is negligible for large } n$$

or equivalently
Type I and II errors negligible for large n.

- Explanatory value;
- Fundamental property for correctness of **post-model-selection inference**.

Why linear regression is so important ?

Linear predictive model is the **cornerstone od prediction**

$$\hat{Y} = g(X^T\hat{\beta})$$

examples: neural nets, compressed sensing, generalized linear models (GLM), ARMA models etc.
Linear model solution for two class classification problem works well..

It is not a fluke !

Linear model

$$y = (y_1, \ldots, y_n)^T, \quad X = [x_{1\cdot}, \ldots, x_{n\cdot}]^T = [x_{\cdot 1}, \ldots, x_{\cdot p}].$$

$y^T 1_n = 0$ and the columns are standardized:
$x_{\cdot j}^T 1_n = 0, \quad x_{\cdot j}^T x_{\cdot j} = 1$ for $j = 1, \ldots, p$.

Linear Regression Model

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \qquad i = 1, \ldots, n$$

$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \in R^n$ iid zero-mean errors.

High dimensionality and sparsity

Aim. Operational algorithms of risk minimisation which work in high-dimensional setting.
Two features of the problem:

- **High-dimensionality:** $p > n$ or $p >> n$
  NP-dimensionality $p \sim \exp(n^{\alpha})$ for some $\alpha > 0$;

Aim. Operational algorithms of risk minimisation which work in high-dimensional setting.
Two features of the problem:

- **High-dimensionality:** $p > n$ or $p >> n$
  NP-dimensionality $p \sim \exp(n^\alpha)$ for some $\alpha > 0$;
- **Sparsity:** active set $T = \{i : \beta_i \neq 0\}$ satisfies

$$|T| << \min(n, p)$$

(**bet on sparsity**)

Bet on sparsity

Bet on sparsity (statistical insight)

Consider $\hat{\beta}_T^{OLS}$ as an oracle benchmark. Then

$$E||X\hat{\beta}_T^{OLS} - X\beta||^2 = \sigma^2|T|.$$

Useless when $|T| \approx n$.
Simple approaches as OLS for all predictors $p > n$: not working (**perfect fit on training data**).
Penalized approaches valuable as they can yield sparsity of the solution.

LASSO estimator in linear model

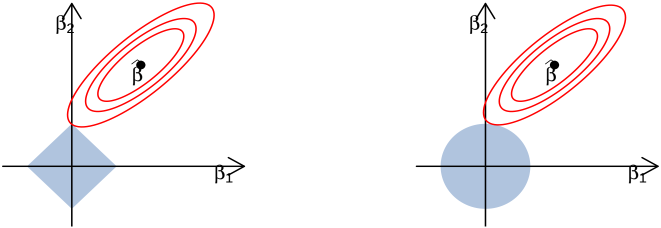
Least **A**bsolute **S**hrinkage and **S**election Operator:

$$\widehat{\beta}_L \equiv \widehat{\beta}_L(\lambda) = \text{argmin}_\beta \left\{ \|y - X\beta\|^2 + 2\lambda|\beta|_1 \right\}$$

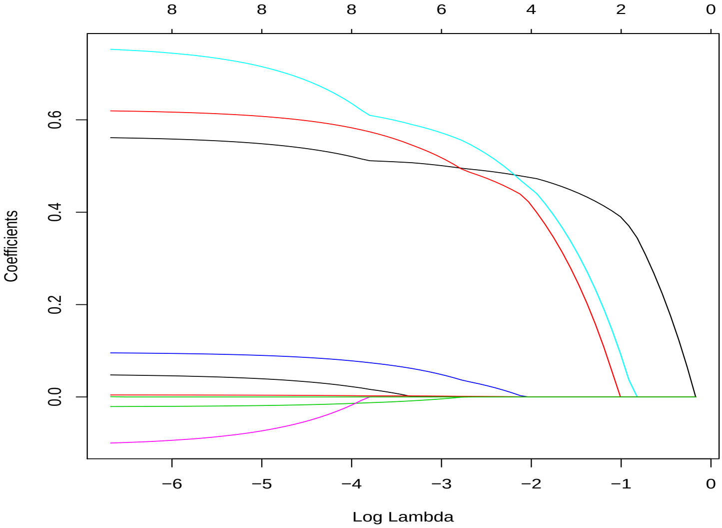
Dual (constrained) version:

$$\widehat{\beta}_L = \text{argmin}_{\beta:|\gamma|_1 \le t(\lambda)} \left\{ \|y - X\beta\|^2 \right\}$$

Penalty Functions: Lasso versus Ridge



Inclusion of predictors by Lasso for prostate data



6.1 Lasso as Soft Thresholding

One-dimensional linear regression $y = x\beta + \varepsilon$.

Focus on $y^T 1_n = x^T 1_n = 0$ and $x^T x = 1$. We have

$$\widehat{\beta} := \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = x^T y,$$

$$\widehat{\beta}_L := \arg\min_{\beta} \{\sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + 2\lambda|\beta|\} = S_\lambda(\widehat{\beta}),$$

where $S_\lambda(\widehat{\beta}) = sign(\beta)(|\beta| - \lambda)_+$ is *soft-thresholding* function.

7.1 Coordinate Descent (CD)

Algorithm 1 Minimization $f(\beta)$ via CD

$\beta = \beta^{\text{start}}$

repeat

  for $j = 1, \ldots, p$

  $\beta_j = \arg\min_b f(\beta_1, \ldots, \beta_{j-1}, b, \beta_{j+1}, \ldots, \beta_p)$

until OK

7.3 Coordinate Descent for LASSO

Algorithm 2 CD for linear LASSO

$\beta = \beta^{\text{start}}, r = y - X\beta^{\text{start}}$

for $\lambda = \lambda_k, \ldots, \lambda_1$ **do**

  repeat

    for $j = 1, \ldots, p$

      $\beta_j^{new} = S_\lambda(\beta_j^{old} + x_{\cdot j}^T r)$

      $r = r + x_{\cdot j}\beta_j^{old} - x_{\cdot j}\beta_j^{new}$

  until OK

  $\beta(\lambda) = \beta$

end for;

Three properties of Lasso

which can be used (at a price of conditions !)

- **Selection Consistency** ($T = \{i : \beta_i \neq 0\}$)

$$\hat{T}_L = T \equiv \min_{i \in T} |\hat{\beta}_{L,i}| > \max_{i \in \bar{T}} |\hat{\beta}_{L,i}| = 0$$

Never satisfied under realistic assumptions.

Three properties of Lasso

which can be used (at a price of conditions !)

- **Selection Consistency** ($T = \{i : \beta_i \neq 0\}$)

$$\hat{T}_L = T \equiv \min_{i \in T} |\hat{\beta}_{L,i}| > \max_{i \in \bar{T}} |\hat{\beta}_{L,i}| = 0$$

  Never satisfied under realistic assumptions.

- **Separation:** Lasso separates T from $\bar{T}$:

$$\min_{i \in T} |\hat{\beta}_{L,i}| > \max_{i \in \bar{T}} |\hat{\beta}_{L,i}|$$

  May fail for strongly correlated predictors (Su et al (2015)).

Three properties of Lasso

which can be used (at a price of conditions !)

- **Selection Consistency** ($T = \{i : \beta_i \neq 0\}$)

$$\hat{T}_L = T \equiv \min_{i \in T} |\hat{\beta}_{L,i}| > \max_{i \in \bar{T}} |\hat{\beta}_{L,i}| = 0$$

Never satisfied under realistic assumptions.

- **Separation:** Lasso separates T from $\bar{T}$:

$$\min_{i \in T} |\hat{\beta}_{L,i}| > \max_{i \in \bar{T}} |\hat{\beta}_{L,i}|$$

May fail for strongly correlated predictors (Su et al (2015)).

- **Screening:** Lasso yields screening:

$$\hat{T}_L \supset T \equiv \min_{i \in T} |\hat{\beta}_{L,i}| > 0$$

Holds under much milder conditions, Zou, 2006.

Folded Concave Penalties (FCP):

- $P_\lambda(t)$ is increasing, concave and $P_\lambda(0) = 0$;
- $P'_\lambda(0^+) > 0$;
- $P_\lambda(t)$= constant for $t > \gamma\lambda$ for some $\gamma > 1$;
- ...

Much more difficult algorithmically, but some approximate solutions such as LLA exist.

$$SCAD, MCP, capped - \ell_1 \in FCP$$

Folded Concave Penalties (FCP):

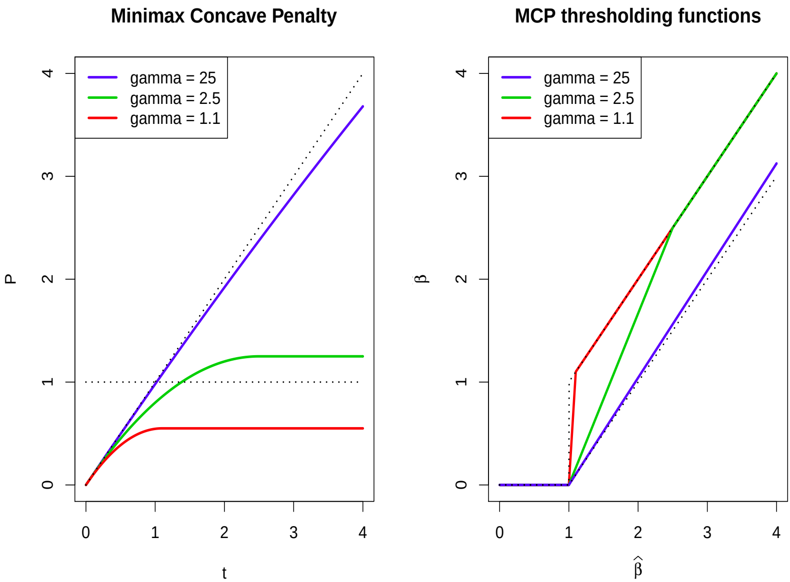
- $P_\lambda(t)$ is increasing, concave and $P_\lambda(0) = 0$;
- $P'_\lambda(0^+) > 0$;
- $P_\lambda(t)$= constant for $t > \gamma\lambda$ for some $\gamma > 1$;
- ...

Much more difficult algorithmically, but some approximate solutions such as LLA exist.

$$SCAD, MCP, capped - \ell_1 \in FCP$$

$$GIC \preceq MCP \preceq Lasso \preceq RR$$

MCP approximates more closely ℓ_0 penalty then Lasso.


Minimax Concave Penalty
MCP thresholding functions
gamma = 25
gamma = 2.5
gamma = 1.1
gamma = 25
gamma = 2.5
gamma = 1.1
P
t
β
$\hat{\beta}$
0
1
2
3
4

Screening-Selection (SS) procedure

Version of SOS (JMLR (2015)) with 'O' removed ..

Algorithm 3 SS

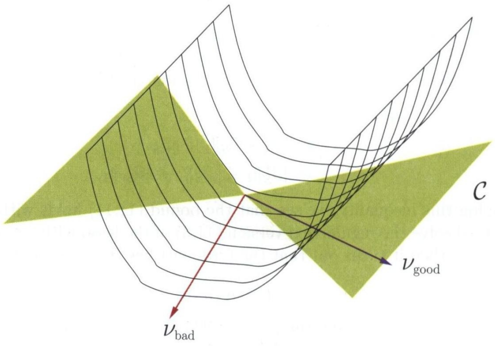
Input: y, X and λ
Screening (Lasso)
$\widehat{\beta} \equiv \widehat{\beta}(\lambda) = \text{argmin}_\gamma \left\{ \|y - X\gamma\|^2 + 2\lambda|\gamma|_1 \right\}$;
order nonzero coefficients:
$|\widehat{\beta}_{j_1}| \geq |\widehat{\beta}_{j_2}| \geq \ldots \geq |\widehat{\beta}_{j_s}|$, where $s = |\text{supp}\widehat{\beta}|$;
set $\mathcal{J} = \{\{j_1\}, \{j_1, j_2\}, \ldots, \{j_1, \ldots, j_s\}\}$;
Selection (GIC)
$\widehat{T} = \text{argmin}_{J \in \mathcal{J}} \left\{ SSE_J + \lambda^2 |J| \right\}$
Output: $\widehat{\beta}^{SS} = (X_{\widehat{T}}^T X_{\widehat{T}})^{-1} X_{\widehat{T}}^T y$

Limitations on selection consistency (statistical insight)

To detect active set: dependence between active set and its complement has to be **not too strong**, or

$$X^T X = \frac{\partial^2}{\partial\beta\partial\beta^T} E||y - X\beta||^2/2$$

is not too degenerate.
What does this mean for $p >> n$?.



Rysunek: Strict convexity of risk over a certain cone $\mathcal{C}$ (Tibshirani et al 2015))

Certain cones $\mathcal{C}$ appear naturally...

$$\delta = \widehat{\beta}_L - \beta.$$

Dual definition of Lasso implies

$$\delta \in \mathcal{C} = \{w : |w_{T^c}|_1 \leq |w_T|_1\}.$$

Namely, with $t(\lambda) = |\beta|_1$ we have

$$|\beta|_1 = |\beta_T|_1 \geq |\widehat{\beta}_L|_1 = |\beta + \delta|_1 =$$

$$= |(\beta + \delta)_T|_1 + |\delta_{T^c}|_1 \geq |\beta_T|_1 - |\delta_T|_1 + |\delta_{T^c}|_1$$

Feasibility parameters

Sign-restricted identifiability factor (SCIF)

$$\zeta_{T,a} = \inf_{\nu \in \mathcal{C}_{T,a}} \frac{|X^T X \nu|_\infty}{|\nu|_\infty}$$

where $\mathcal{C}_{T,a}$ for $a \in (0,1)$ is a certain cone. Restriction to $\mathcal{C}_{T,a}$ ensures $\zeta_{T,a} > 0$ for many high-dimensional designs.

Feasibility parameters

Sign-restricted identifiability factor (SCIF)

$$\zeta_{T,a} = \inf_{\nu \in \mathcal{C}_{T,a}} \frac{|X^T X \nu|_\infty}{|\nu|_\infty}$$

where $\mathcal{C}_{T,a}$ for $a \in (0,1)$ is a certain cone. Restriction to $\mathcal{C}_{T,a}$ ensures $\zeta_{T,a} > 0$ for many high-dimensional designs.

Scaled K-L distance

Scaled K-L distance between T and its submodels is

$$\tilde{\delta}_T = \min_{J \subset T} \frac{||(I - H_J) X_T \beta_T||^2}{|T \setminus J|}.$$

Bound for $P(\hat{T}_{SS} \neq T)$ (PP & JM, 2015)

Theorem

Under mild assumptions on feasibility parameters we have

$$P(\hat{T}_{SS} \neq T) \leq p \exp\left(-\frac{\lambda^2}{2\sigma^2}\right)$$

(some constants are omitted)
For true regressors to be distinguishable from the noise

$$\beta_{min} = \min_{i \in T} |\beta_i|$$

has to be sufficiently large. Thus the condition

$$\zeta^2_{T,a}\beta^2_{min} \geq C > 0$$

Algorithm 4 SSnet (Screening Selection algorithm on a net of λs)

Input: y, X and $(\lambda_0, \lambda_1, \ldots, \lambda_m)^T$
Screening (Lasso)
for $k = 1$ **to** m **do**
$\widehat{\beta}^{(k)} \equiv \widehat{\beta}(\lambda_k) = \text{argmin}_{\gamma} \left\{ \|y - X\gamma\|^2 + 2\lambda_k|\gamma| \right\}$;
order nonzero coefficients:
$|\widehat{\beta}^{(k)}_{j_1}| \geq |\widehat{\beta}^{(k)}_{j_2}| \geq \ldots \geq |\widehat{\beta}^{(k)}_{j_{s_k}}|$,
where $s_k = |\text{supp}\widehat{\beta}^{(k)}|$;
set $J_k(y) = \left\{ \{j_1\}, \{j_1, j_2\}, \ldots, \text{supp}\widehat{\beta}^{(k)} \right\}$
end for
Selection (GIC)
$J(y) = \bigcup_{k=1}^{m} J_k(y)$
$\widehat{T} = \text{argmin}_{J \in J(y)} \left\{ R_J + \lambda_0^2 |J| \right\}$
Output: $\widehat{T}$, $\widehat{\beta}^{SSnet} = (X^T_{\widehat{T}} X_{\widehat{T}})^{-1} X^T_{\widehat{T}} y$

SOSnet algorithm

- Use Lasso with $\lambda_i = 0, 1, \ldots, m$ to choose set of predictors I_i;
- **Fit linear model** $y \sim x_{I_i}, i = 0, 1, \ldots, m$**;**
- **Order predictors according to (t-statistics)**^{2}**;**
- Construct $\mathcal{M} = \cup$ nested models ;
- Use GIC on $\mathcal{M}$ to choose a final model.

Delete and Merge Regressors (DMR) algorithm

p predictors being factors:

- (i) Initial screening using **group Lasso**
  ℓ_1/ℓ_2 penalty : $\sum_{i=1}^{p} \lambda_i ||\beta_i||$
- For each factor separately perform tests

$$H_{kl} : \beta_{i,k} = \beta_{i,l}$$

  t^2_{kl}: dissimilarity measure between levels within factor;
- Perform clustering on each factor using $D = (t^2_{k,l})$:
  $\mathbf{h}$: vector of cutting heights;
- Order vector $[\mathbf{h}_1, \ldots, \mathbf{h}_p]$ yielding nested family $\mathcal{M}$ of models;
- Perform GIC on $\mathcal{M}$.

Numerical experiments

Four groups of algorithms

- SS, SSnet, SOSnet
- MCP calibrated by GIC (sparsenet)
- MCP calibrated by CV (sparsenet, two settings)
- MCP ($a = 1,5$ and $a = 3$) (PLUS)

$\lambda = \sigma\sqrt{2\log(p)}$,
Penalization term for GIC: $c\lambda^2$ with three values of $c \in \{1, 1.5, 2, 2.5, 3, 3.5, 4\}$.

Experiments cont'd

M I: $\beta_1 = (3, 1.5, 0, 0, 2, 0^T_{p-5})^T$ from Wang et al (2013) $(p = 3000)$
M II: $\beta_2 = (0^T_{p-10}, \pm 2, \cdots, \pm 2)^T$ Wang et al (2014) $(p = 2000)$
signs $\pm$ chosen separately for every run.
$x_1, \ldots, x_p$: normal with autoregressive (exp. a: $\rho = 0.5$,b: $\rho = 0.7$) or equicorrelated (exp. c: $\rho = 0.5$,d: $\rho = 0.7$) structure.
$n = 100$ (**M I**) and $n = 200$ (**M II**).

Tablica: True Model selection (TM) (%).

	Exp 1a	Exp 1b	Exp 1c	Exp 1d	Exp 2a	Exp 2b	Exp 2c	Exp 2
SS c_1	92.6	69.4	81.8	45.5	8.8	0.6	11.5	0.2
SS c_2	95.7	81.9	80.1	45.4	6.5	0.5	4.8	0.1
SS c_3	91.6	74.3	76.4	38.7	4.0	0.3	1.0	0.1
SSnet c_1	89.1	57.8	83.1	42.9	54.4	4.5	84.8	28.9
SSnet c_2	95.2	76.9	83.2	48.2	54.6	5.8	90.2	35.2
SSnet c_3	91.3	72.2	79.3	42.0	54.4	5.9	89.3	31.5
SOSnet c_1	85.7	45.6	83.9	39.0	74.1	7.0	85.5	34.6
SOSnet c_2	94.8	73.3	86.5	52.8	74.7	10.1	96.1	53.8
SOSnet c_3	91.2	71.0	82.8	46.6	73.0	8.9	94.7	44.2
spnet c_1	81.9	28.8	83.2	36.0	68.5	0.4	86.4	36.3
spnet c_2	91.2	39.1	86.3	51.7	68.4	0.5	96.6	49.8
spnet c_3	89.3	39.7	82.7	47.2	67.6	0.3	95.1	43.9
spnet p.1se	76.4	29.1	71.3	30.7	32.6	0.0	88.8	30.6
spnet p.min	48.7	16.0	55.4	24.2	19.4	0.0	70.4	14.5
mcp 1.5	81.0	23.5	77.5	6.3				
mcp 3	73.1	21.9	75.6	7.5	9.2	0.0	32.5	

Tablica: Relative Mean Squared Error (MSE)

	Exp 1a	Exp 1b	Exp 1c	Exp 1d	Exp 2a	Exp 2b	Exp 2c	Exp 2
SS c_1	1.5	2.7	4.2	9.8	20.0	19.8	13.2	21.1
SS c_2	1.6	3.3	4.6	10.0	22.3	20.8	19.1	24.1
SS c_3	2.5	4.8	5.1	10.6	25.0	21.9	24.9	25.9
SSnet c_1	1.7	3.3	3.9	10.4	7.0	15.2	1.5	4.8
SSnet c_2	1.7	3.5	4.1	9.8	7.6	15.5	1.4	5.2
SSnet c_3	2.5	5.1	4.7	10.3	8.5	16.6	1.6	6.6
SOSnet c_1	2.0	4.6	3.7	11.7	4.9	15.5	1.4	4.2
SOSnet c_2	1.7	4.0	3.6	9.2	4.7	15.5	1.2	3.9
SOSnet c_3	2.6	5.3	4.0	9.5	5.6	16.6	1.3	5.3
spnet c_1	2.7	12.5	3.7	11.4	4.2	26.1	1.3	4.3
spnet c_2	2.4	10.5	3.6	9.1	4.8	24.8	1.2	4.4
spnet c_3	2.9	10.3	4.1	9.5	6.0	24.7	1.3	5.9
spnet p.1se	5.7	10.9	6.8	11.5	3.7	23.9	2.0	5.7
spnet p.min	3.7	6.4	5.1	10.0	2.8	20.5	1.6	4.7
mcp 1.5	2.9	13.3	5.5	20.1				
mcp 3	7.6	14.6	8.6	19.7	25.9	28.2	16.8	

8.2 SOSnet in Regression Experiment

Tablica: Methylation data set: $n = 656, p = 193870$. Cross-validated mean root mean square error of prediction (RMSE) and mean model dimension (MD).

algorithm	RMSE	MD
SOSnet cv	5.1	336
sparsenet cv	4.8	485
SOSnet gic $c = 2.5$	5.6	40
sparsenet gic $c = 2.5$	7.2	44

Comments on results

- SOSnet: higher correct selection probability and lower MSE *simultaneously* in almost all experimental setups.
- The difference is most pronounced for higher correlations.
- Times for SOSnet $>$ 2 times shorter than for sparsenet + GIC, $>$4 times shorter than for sparsenet + CV , $>$ 20 times shorter than for PLUS implementation.
- Sparsenet tuned by GIC works much better than tuned by CV.

Binary case: remarks

- Conceptually the same. Change of a loss function, usually to logistic. More difficult algorithmically.

Binary case: remarks

- Conceptually the same. Change of a loss function, usually to logistic. More difficult algorithmically.
- Theoretical analysis more difficult due to **heteroscedasticity** of response.

Binary case: remarks

- Conceptually the same. Change of a loss function, usually to logistic. More difficult algorithmically.
- Theoretical analysis more difficult due to **heteroscedasticity** of response.
- NP-dimensional case:
  Filtering based on ranking of univariate fits (e.g.SIS, Fan et al (2009)) and then PERM analysis to chosen subset.

Binary case: remarks

- Conceptually the same. Change of a loss function, usually to logistic. More difficult algorithmically.
- Theoretical analysis more difficult due to **heteroscedasticity** of response.
- NP-dimensional case:
  Filtering based on ranking of univariate fits (e.g.SIS, Fan et al (2009)) and then PERM analysis to chosen subset.
- Fitting univariate (e.g. logistic) models to multivariate logistic data is an ultimate type of **model misspecification**.

Misspecified logistic model

Different angle:

Logistic loss in empirical risk minimisation $\equiv$ fitting a logistic model.

Data comes from binary model

$$P(Y = 1|X) = q(\beta_0 + \beta^T X)$$

X is random variable in R^p and q response function $q \neq p$,

$$p(\beta_0 + \beta^T x) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)}$$

is most frequently used tool to model dependence of binary outcome on attributes.

Important special cases: Omission of (some) valid predictors from logistic model itself, **filters** in particular.

- What happens when we misspecify response function and use logistic response p instead of q ?
- Some bias in estimation of β surely occurs, but how important is an error ?
- It is obvious that we cannot learn $||\beta||$ when q is arbitrary, but what about **direction of** β ?
- Can we learn suppβ ?

Yes, we can (frequently, at least)

Simpler framework: minimization of empirical risk ($p < n$)

$$(\hat{\beta}_0^{ML}, \hat{\beta}^{ML}) = \arg\min_{\gamma_0,\gamma} err(\gamma_0, \gamma).$$

Using $(\hat{\beta}_0^{ML}, \hat{\beta}_0^{ML})$ we estimate not β_0 and β but β_0^* and β^* such that

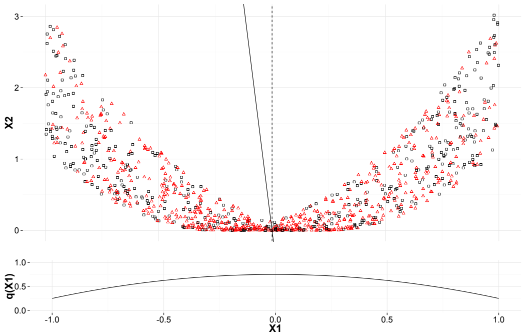
$$(\beta_0^* \beta^*) = \mathrm{argmin}_{b_0, b \in \mathbb{R}^p} E_X KL(q(\beta_0 + X^T\beta), p(b_0 + X^T b)),$$

where

$$KL(q, p) = q \log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right)$$

is Kullback-Leibler distance between two Bernoulli distributions with probabilities of success q and p.

$$X_2 \sim (X_1 + \varepsilon)^2, \quad P(y = 1|x) = q(x_1)$$



Rysunek: Squares and triangles correspond to $Y = 1$ and $Y = 0$. Solid line shows the direction of $\hat{\beta}$

Positive result: Ruud's theorem (1983)

Assume that distribution of X is nondegenerate and such that regressions with respect to $\beta^T X$ are linear

$$E(X|\beta^T X) = u\beta^T X + u_0. \qquad (R)$$

Then there exists η such that

$$\beta^* = \eta\beta$$

Important:

$$\eta \neq 0?$$

Relevance for selection of predictors ($p < n$)

$$Dev_\omega = \frac{LRT_f}{LRT_\omega}$$

- Order variables according to their residual deviances

$$Dev_{f\setminus\{i_1\}} \geq Dev_{f\setminus\{i_2\}} \geq .. \geq Dev_{f\setminus\{i_p\}}$$

and minimize GIC in the corresponding nested family.
Then if (R) is satisfied, q is **strictly monotone** and ..
$\hat{T}_{GIC}$ is consistent (P. Teisseyre, JM (2015))

Relevance for selection of predictors ($p < n$)

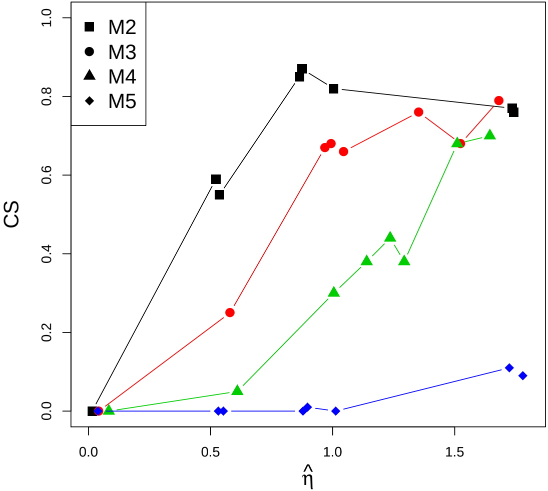
$$Dev_\omega = \frac{LRT_f}{LRT_\omega}$$

- Order variables according to their residual deviances

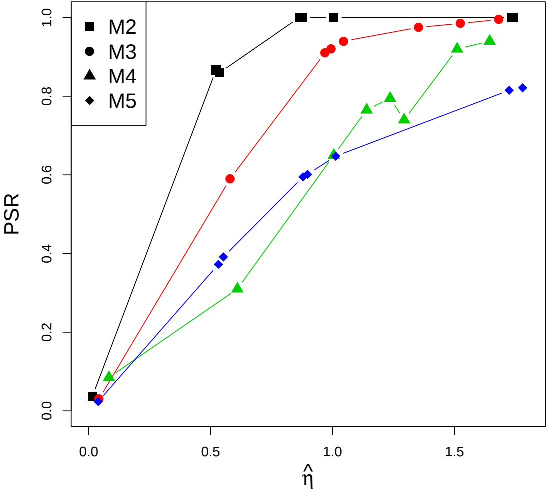
  $$Dev_{f\setminus\{i_1\}} \geq Dev_{f\setminus\{i_2\}} \geq .. \geq Dev_{f\setminus\{i_p\}}$$

  and minimize GIC in the corresponding nested family. Then if (R) is satisfied, q is **strictly monotone** and .. $\hat{T}_{GIC}$ is consistent (P. Teisseyre, JM (2015))
- For the case when $|\eta| > 1$ we are frequently better off when misspecifing the model then fitting the correct one...

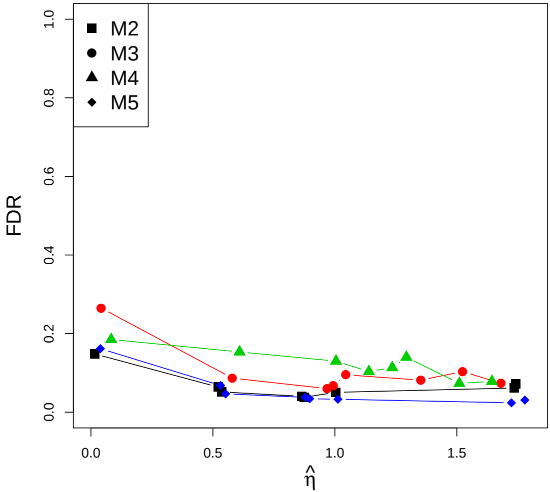
Correct selection versus η



PSR versus η



FDR versus η



For **normal predictors** we have

(M. Kubkowski, JM 2016)

$$\eta = \frac{Eq'(\beta_0 + \beta^T X)}{Ep'(\beta_0^* + \beta^{*T} X)} = \frac{Eq'(\beta_0 + \beta^T X)}{Ep'(\beta_0^* + \eta\beta^T X)}$$

For **normal predictors** we have

(M. Kubkowski, JM 2016)

$$\eta = \frac{Ep'(\beta_0 + \beta^T X)}{Ep'(\beta_0^* + \beta^{*T} X)} = \frac{Eq'(\beta_0 + \beta^T X)}{Ep'(\beta_0^* + \eta\beta^T X)}$$

(Y, X) follow **logistic** model and β^*_{lin} is a projection on a **linear** model. Then

$$\beta^*_{lin} = Ep'(\beta_0 + \beta^T X)\beta$$

i.e. direction $\beta/||\beta||$ of β can be recovered by fitting a **linear** model.

Some relevant papers

- A. Maj-Kańska, P. Pokarowski, A. Prochenka, et al. Delete or merge regressors for linear model selection. Electronic Journal of Statistics, 2015.
- P. Pokarowski, J. Mielniczuk,Combined ℓ_1 and Greedy ℓ_0 Penalized Least Squares for Linear Model Selection, Journal of Machine Learning Research, 2015
- Bach, F. et al. Optimization with sparsity-inducing penalties, 2011
- P. Ruud, Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models, Econometrica, 1983
- T. Hastie, R. Tibshirani, M. Wainwright, Statistical Learning with Sparsity, CRC 2015

Machine Learning or Statistics ?

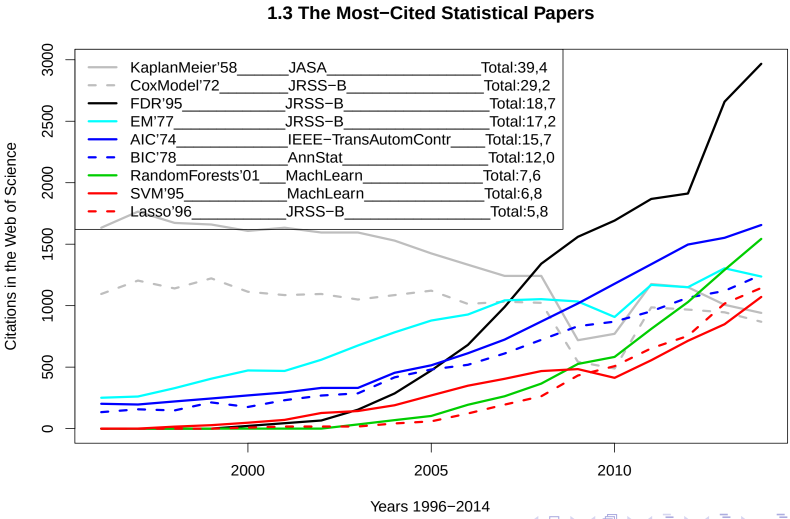
Kilka prac z JMLR ..

- P. Bellec and A. Tsybakov, Sharp Oracle Bounds for **Monotone and Convex Regression** Through Aggregation, JMLR 2015
- J. Jin and C-H. Zhang and Q. Zhang, Optimality of Graphlet Screening in **High Dimensional Variable Selection**, JMLR 2014
- X. Li and T. Zhao and Yuan and H. Liu The flare Package for **High Dimensional Linear Regression** and Precision Matrix Estimation in R, JMLR 2015
- P. Pokarowski and J. Mielniczuk Combined l1 and Greedy l0 Penalized Least Squares for **Linear Model Selection**, JMLR 2015
- M. Tan and I. W. Tsang and L. Wang Towards **Ultrahigh Dimensional Feature Selection** for Big Data, JMLR 2014

Kilka prac z Annals of Statistics ..

- P. Sherwood and L. Wang, Partially **linear additive quantile regression** in ultra-high dimension, AS 2016
- R. Barber and E. Candes Controlling the **false discovery rate** via knockoffs AS 2015
- Y. Yang and S. Tokdar Minimax-optimal **nonparametric regression** in high dimensions, AS 2015
- B. Jiang and J. S. Liu **Variable selection** for general index models via sliced inverse regression, AS 2014
- J. Fan, L. Xue, and H. Zou Strong oracle optimality of **folded concave penalized estimation**, AS 2014

Most cited statistical papers (Pokarowski, 2015)



Computational considerations

- Lasso regularized path solution requires

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

flops using LARS;

- Selection step requires

$$ns^2, \qquad s = |\mathrm{supp}\hat{\beta}_L|$$

flops . Use QR decomposition. This follows since $\mathcal{J}$ is **nested** !

Screening step is the most expensive in this and other algorithms ($s < n$)