Fast and robust procedures in high-dimensional variable selection

Wojciech Rejchel

Nicolaus Copernicus University, Toruń , Poland

Joint works with M. Bogdan (University of Wrocław) and K. Furmańczyk (Warsaw University of Life Sciences)

Single index model

- $$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n$$

Single index model

- $$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n$$
- $Y_i \in \mathbb{R}$ - response variable

Single index model

- $$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n$$
- $Y_i \in \mathbb{R}$ - response variable
- $X_i \in \mathbb{R}^p$ - predictor

Single index model

- $$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n$$
- $Y_i \in \mathbb{R}$ - response variable
- $X_i \in \mathbb{R}^p$ - predictor
- β - true parameter

Single index model

- $$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n$$
- $Y_i \in \mathbb{R}$ - response variable
- $X_i \in \mathbb{R}^p$ - predictor
- β - true parameter
- ε_i - random error

Single index model

- $$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n$$
- $Y_i \in \mathbb{R}$ - response variable
- $X_i \in \mathbb{R}^p$ - predictor
- β - true parameter
- ε_i - random error
- g - unknown function

Single index model - examples

- Linear model

$$Y_i = \beta' X_i + \varepsilon_i$$

Single index model - examples

- Linear model

$$Y_i = \beta' X_i + \varepsilon_i$$

- Modified linear model

$$Y_i = g(\beta' X_i + \varepsilon_i)$$

Single index model - examples

- Linear model

$$Y_i = \beta' X_i + \varepsilon_i$$

- Modified linear model

$$Y_i = g(\beta' X_i + \varepsilon_i)$$

-

$$Y_i = g(\beta' X_i) + \varepsilon_i$$

More examples - binary regression

- $$P(Y_i = 1|X_i) = g(\beta' X_i), \quad i = 1, \ldots, n$$

More examples - binary regression

- $$P(Y_i = 1|X_i) = g(\beta' X_i), \quad i = 1, \dots, n$$
- g - unknown function

More examples - binary regression

- $$P(Y_i = 1|X_i) = g(\beta' X_i), \quad i = 1, \ldots, n$$
- g - unknown function
- Logistic model

$$P(Y_i = 1|X_i) = \frac{\exp(\beta' X_i)}{\exp(\beta' X_i) + 1}$$

More examples - binary regression

- $$P(Y_i = 1|X_i) = g(\beta' X_i), \quad i = 1, \ldots, n$$
- g - unknown function
- Logistic model

$$P(Y_i = 1|X_i) = \frac{\exp(\beta' X_i)}{\exp(\beta' X_i) + 1}$$

- Probit model

$$P(Y_i = 1|X_i) = \Phi(\beta' X_i)$$

Variable selection

- Find

$$T = \{1 \leq j \leq p : \beta_j \neq 0\}$$

Variable selection

- Find
$$T = \{1 \leq j \leq p : \beta_j \neq 0\}$$
- Even if $p >> n$

Variable selection

- Find
$$T = \{1 \leq j \leq p : \beta_j \neq 0\}$$
- Even if $p >> n$
- Function g is unknown, so β is not identifiable

Variable selection

- Find
$$T = \{1 \leq j \leq p : \beta_j \neq 0\}$$
- Even if $p >> n$
- Function g is unknown, so β is not identifiable
- Under mild assumptions it can be recognized up to positive multiplicative constant

Single index model

- $$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n$$

Single index model

- $$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n$$
- No assumptions on distribution of errors, existence of their moments etc.

Single index model

- $$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n$$
- No assumptions on distribution of errors, existence of their moments etc.
- Function g is unknown but increasing wrt first variable

Single index model

- $$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n$$
- No assumptions on distribution of errors, existence of their moments etc.
- Function g is unknown but increasing wrt first variable
- Binary regression

$$P(Y_i = 1 | X_i) = g(\beta' X_i), \quad i = 1, \ldots, n$$

Single index model

- $$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n$$
- No assumptions on distribution of errors, existence of their moments etc.
- Function g is unknown but increasing wrt first variable
- Binary regression

$$P(Y_i = 1 | X_i) = g(\beta' X_i), \quad i = 1, \ldots, n$$

- Function g is unknown but increasing

Single index model

- $$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n$$
- No assumptions on distribution of errors, existence of their moments etc.
- Function g is unknown but increasing wrt first variable
- Binary regression

$$P(Y_i = 1 | X_i) = g(\beta' X_i), \quad i = 1, \ldots, n$$

- Function g is unknown but increasing
- Robust variable selection

Single index model

- $$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \dots, n$$
- No assumptions on distribution of errors, existence of their moments etc.
- Function g is unknown but increasing wrt first variable
- Binary regression
$$P(Y_i = 1|X_i) = g(\beta' X_i), \quad i = 1, \dots, n$$
- Function g is unknown but increasing
- Robust variable selection
- ... and computationally fast

Linear model

- Linear model

$$Y_i = \beta' X_i + \varepsilon_i$$

Linear model

- Linear model

$$Y_i = \beta' X_i + \varepsilon_i$$

- Lasso estimator

$$\arg\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \theta' X_i\right)^2 + \lambda\,|\theta|_1$$

Linear model

- Linear model

$$Y_i = \beta' X_i + \varepsilon_i$$

- Lasso estimator

$$\arg\min_{\theta} \quad \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \theta' X_i \right)^2 + \lambda \, |\theta|_1$$

- $|\theta|_1 = \sum_{j=1}^{p} |\theta_j|$

Linear model

- Linear model

$$Y_i = \beta' X_i + \varepsilon_i$$

- Lasso estimator

$$\arg\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \theta' X_i\right)^2 + \lambda \, |\theta|_1$$

- $|\theta|_1 = \sum_{j=1}^{p} |\theta_j|$
- Least absolute deviation estimator with Lasso penalty

$$\arg\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \theta' X_i\right| + \lambda \, |\theta|_1$$

Linear model

- Linear model
$$Y_i = \beta' X_i + \varepsilon_i$$
- Lasso estimator
$$\arg\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \theta' X_i\right)^2 + \lambda\,|\theta|_1$$
- $|\theta|_1 = \sum_{j=1}^{p} |\theta_j|$
- Least absolute deviation estimator with Lasso penalty
$$\arg\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \theta' X_i\right| + \lambda\,|\theta|_1$$
- Time-consuming when $n, p \sim 1000$

Ranks

- Response variables $Y_1, \ldots Y_n$

Ranks

- Response variables $Y_1, \ldots Y_n$
- Sort them $Y_{(1)} \leq Y_{(2)} \leq Y_{(3)} \leq \ldots \leq Y_{(n)}$

Ranks

- Response variables $Y_1, \ldots Y_n$
- Sort them $Y_{(1)} \leq Y_{(2)} \leq Y_{(3)} \leq \ldots \leq Y_{(n)}$
- Assign ranks $R_1, \ldots R_n$

Ranks

- Response variables $Y_1, \ldots Y_n$
- Sort them $Y_{(1)} \leq Y_{(2)} \leq Y_{(3)} \leq \ldots \leq Y_{(n)}$
- Assign ranks $R_1, \ldots R_n$
- RankLasso

$$\min_{\theta} \quad Q(\theta) + \lambda \, |\theta|_1$$

Ranks

- Response variables $Y_1, \ldots Y_n$
- Sort them $Y_{(1)} \leq Y_{(2)} \leq Y_{(3)} \leq \ldots \leq Y_{(n)}$
- Assign ranks $R_1, \ldots R_n$
- RankLasso

$$\min_{\theta} \quad Q(\theta) + \lambda \, |\theta|_1$$

$$Q(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(R_i/n - \theta' X_i \right)^2$$

Procedures

- Computationally fast

Procedures

- Computationally fast
- No new algorithmic machinery

Procedures

- Computationally fast
- No new algorithmic machinery
- Robust wrt to outliers and function g

Procedures

- Computationally fast
- No new algorithmic machinery
- Robust wrt to outliers and function g
- What is estimated?

Binary regression

- Logistic model

$$P(Y_i = 1|X_i) = \frac{\exp(\beta' X_i)}{\exp(\beta' X_i) + 1}$$

Binary regression

- Logistic model

$$P(Y_i = 1|X_i) = \frac{\exp(\beta' X_i)}{\exp(\beta' X_i) + 1}$$

- Lasso estimator

$$\arg\min_{\theta} \quad LogLike(\theta) + \lambda \, |\theta|_1$$

Binary regression

- Logistic model

$$P(Y_i = 1|X_i) = \frac{\exp(\beta' X_i)}{\exp(\beta' X_i) + 1}$$

- Lasso estimator

$$\arg\min_{\theta} \quad LogLike(\theta) + \lambda\,|\theta|_1$$

- $|\theta|_1 = \sum_{j=1}^{p} |\theta_j|$

Procedure

- Treat Y_i's as numeric and apply Lasso for linear regression

$$\min_{\theta} \quad \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \theta' X_i \right)^2 + \lambda \left| \theta \right|_1$$

Procedure

- Treat Y_i's as numeric and apply Lasso for linear regression

$$\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \theta' X_i\right)^2 + \lambda\left|\theta\right|_1$$

- Bühlmann, vd Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*

Procedures

- Computationally fast

Procedures

- Computationally fast
- No new algorithmic machinery

Procedures

- Computationally fast
- No new algorithmic machinery
- What is estimated?

Binary regression

- $$Q(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \theta' X_i \right)^2$$

Binary regression

- $$Q(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \theta' X_i\right)^2$$
- We estimate
$$\theta^0_{bin} = \arg\min_{\theta\in\mathbb{R}^p} \quad \mathbb{E}\, Q(\theta)$$

Binary regression

- $$Q(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \theta' X_i\right)^2$$
- We estimate
$$\theta^0_{bin} = \arg\min_{\theta \in \mathbb{R}^p} \quad \mathbb{E}\, Q(\theta)$$
- Relation between β and ${\theta^0}_{bin}$?

RankLasso

- $$Q(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(R_i/n - \theta' X_i\right)^2$$

RankLasso

- $$Q(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(R_i/n - \theta' X_i \right)^2$$
- We estimate
$$\theta^0_{rank} = \arg\min_{\theta \in \mathbb{R}^p} \quad \mathbb{E}\, Q(\theta)$$

RankLasso

- $$Q(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(R_i/n - \theta' X_i\right)^2$$
- We estimate
$$\theta^0_{rank} = \arg\min_{\theta\in\mathbb{R}^p} \quad \mathbb{E}\, Q(\theta)$$
- Relation between β and ${\theta^0}_{rank}$?

Assumptions

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

Assumptions

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ - i.i.d.
- $\mathbb{E} X_1 = 0$

Assumptions

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ - i.i.d.
- $\mathbb{E}X_1 = 0$
- $H = \mathbb{E}X_1 X_1'$ - positive definite

Assumptions

- **Assumption:** For each $\theta \in \mathbb{R}^p$ the conditional expectation $\mathbb{E}(\theta' X_1 | \beta' X_1)$ exists and

Assumptions

- **Assumption:** For each $\theta \in \mathbb{R}^p$ the conditional expectation $\mathbb{E}(\theta' X_1 | \beta' X_1)$ exists and

$$\mathbb{E}(\theta' X_1 | \beta' X_1) = d_\theta \beta' X_1$$

for a real number $d_\theta \in \mathbb{R}$.

Assumptions

- **Assumption:** For each $\theta \in \mathbb{R}^p$ the conditional expectation $\mathbb{E}(\theta' X_1 | \beta' X_1)$ exists and

$$\mathbb{E}(\theta' X_1 | \beta' X_1) = d_\theta \beta' X_1$$

for a real number $d_\theta \in \mathbb{R}$.

- Satisfied in simple regression model, i.e. $X_1 \in \mathbb{R}$

Assumptions

- **Assumption:** For each $\theta \in \mathbb{R}^p$ the conditional expectation $\mathbb{E}(\theta' X_1 | \beta' X_1)$ exists and

$$\mathbb{E}(\theta' X_1 | \beta' X_1) = d_\theta \beta' X_1$$

  for a real number $d_\theta \in \mathbb{R}$.
- Satisfied in simple regression model, i.e. $X_1 \in \mathbb{R}$
- X_1 comes from the elliptical distribution (multivariate normal distribution or t-distribution)

Assumptions

- **Assumption:** For each $\theta \in \mathbb{R}^p$ the conditional expectation $\mathbb{E}(\theta' X_1 | \beta' X_1)$ exists and

$$\mathbb{E}(\theta' X_1 | \beta' X_1) = d_\theta \beta' X_1$$

  for a real number $d_\theta \in \mathbb{R}$.
- Satisfied in simple regression model, i.e. $X_1 \in \mathbb{R}$
- X_1 comes from the elliptical distribution (multivariate normal distribution or t-distribution)
- Hall, Li (1993, Ann. Stat.)

Binary regression

If assumptions are satisfied and g is increasing

Binary regression

If assumptions are satisfied and g is increasing then there exists $\gamma > 0$ such that

$$(\theta^0_{bin})_j = \gamma \beta_j$$

for $j = 1, \ldots, p$

Binary regression

If assumptions are satisfied and g is increasing

then there exists $\gamma > 0$ such that

$$(\theta^0_{bin})_j = \gamma\beta_j$$

for $j = 1, \ldots, p$

Therefore, $supp(\beta) = supp(\theta^0_{bin})$.

RankLasso

If assumptions are satisfied

RankLasso

If assumptions are satisfied

g is increasing wrt first variable

RankLasso

If assumptions are satisfied

g is increasing wrt first variable

cumulative distribution function of Y_i is increasing

RankLasso

If assumptions are satisfied

g is increasing wrt first variable

cumulative distribution function of Y_i is increasing

then there exists $\gamma > 0$ such that

$$(\theta^0_{rank})_j = \gamma \beta_j$$

for $j = 1, \ldots, p$

RankLasso

If assumptions are satisfied

g is increasing wrt first variable

cumulative distribution function of Y_i is increasing

then there exists $\gamma > 0$ such that

$$(\theta^0_{rank})_j = \gamma\beta_j$$

for $j = 1, \ldots, p$

Therefore, $supp(\beta) = supp(\theta^0_{rank})$.

Properties in variable selection

- Procedures are variable selection consistent under "irrepresentable condition" - restrictive assumption, hardly satisfied in practice, common for Lasso estimators.

Properties in variable selection

- Procedures are variable selection consistent under "irrepresentable condition" - restrictive assumption, hardly satisfied in practice, common for Lasso estimators.
- Under much weaker assumption we can obtain

Properties in variable selection

- Procedures are variable selection consistent under "irrepresentable condition" - restrictive assumption, hardly satisfied in practice, common for Lasso estimators.
- Under much weaker assumption we can obtain screening

$$T \subset \hat{T}$$

Properties in variable selection

- Procedures are variable selection consistent under "irrepresentable condition" - restrictive assumption, hardly satisfied in practice, common for Lasso estimators.
- Under much weaker assumption we can obtain screening

$$T \subset \hat{T}$$

or even separability of predictors

$$\forall_{j \in T, k \notin T} \quad |\hat{\theta}_j| > |\hat{\theta}_k|$$

Properties in variable selection

- Procedures are variable selection consistent under "irrepresentable condition" - restrictive assumption, hardly satisfied in practice, common for Lasso estimators.
- Under much weaker assumption we can obtain screening

$$T \subset \hat{T}$$

or even separability of predictors

$$\forall_{j \in T, k \notin T} \quad |\hat{\theta}_j| > |\hat{\theta}_k|$$

- We do not require function g to be known (except being increasing)

Properties in variable selection

- Procedures are variable selection consistent under "irrepresentable condition" - restrictive assumption, hardly satisfied in practice, common for Lasso estimators.
- Under much weaker assumption we can obtain screening

$$T \subset \hat{T}$$

or even separability of predictors

$$\forall_{j \in T, k \notin T} \quad |\hat{\theta}_j| > |\hat{\theta}_k|$$

- We do not require function g to be known (except being increasing)
- We do not require any assumptions on error terms or function g, except being increasing wrt first variable

Properties in variable selection

- Procedures are variable selection consistent under "irrepresentable condition" - restrictive assumption, hardly satisfied in practice, common for Lasso estimators.
- Under much weaker assumption we can obtain screening
$$T \subset \hat{T}$$
or even separability of predictors
$$\forall_{j \in T, k \notin T} \quad |\hat{\theta}_j| > |\hat{\theta}_k|$$
- We do not require function g to be known (except being increasing)
- We do not require any assumptions on error terms or function g, except being increasing wrt first variable
- Using thresholding or weighted (adaptive) procedures

Consistency in variable selection

- Theoretical analysis of RankLasso is more involved than for regular Lasso, because

Consistency in variable selection

- Theoretical analysis of RankLasso is more involved than for regular Lasso, because
- $$Q(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(R_i/n - \theta' X_i\right)^2$$

  is sum of dependent random variables

Consistency in variable selection

- Theoretical analysis of RankLasso is more involved than for regular Lasso, because
- $$Q(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(R_i/n - \theta' X_i\right)^2$$

  is sum of dependent random variables
- U-statistics theory is used

Theorem

- $a \in (0, 1), q \geq 1$

Theorem

- $a \in (0,1), q \geq 1$
- predictors are subgaussian

Theorem

- $a \in (0,1), q \geq 1$
- predictors are subgaussian
- cone invertibility condition is satisfied

Theorem

- $a \in (0,1), q \geq 1$
- predictors are subgaussian
- cone invertibility condition is satisfied
- $n \geq K_1 p_0^2 \log(p/a)$

Theorem

- $a \in (0,1), q \geq 1$
- predictors are subgaussian
- cone invertibility condition is satisfied
- $n \geq K_1 p_0^2 \log(p/a)$
- $\lambda \geq K_2 \sqrt{\frac{\log(p/a)}{n}}$

Theorem

- $a \in (0,1), q \geq 1$
- predictors are subgaussian
- cone invertibility condition is satisfied
- $n \geq K_1 p_0^2 \log(p/a)$
- $\lambda \geq K_2 \sqrt{\frac{\log(p/a)}{n}}$
- With probability at least $1 - K_3 a$ we have

$$|\hat{\theta} - \theta^0|_q \leq K_4 p_0^{1/q} \lambda$$

Theorem

- $\beta_{min} = \min_{j \in T} |\beta_j|$

Theorem

- $\beta_{min} = \min_{j \in T} |\beta_j|$
- Our estimators are able to separate predictors

$$\forall_{j \in T, k \notin T} \quad |\hat{\theta}_j| > |\hat{\theta}_k|$$

Theorem

- $\beta_{min} = \min_{j \in T} |\beta_j|$
- Our estimators are able to separate predictors

$$\forall_{j \in T, k \notin T} \quad |\hat{\theta}_j| > |\hat{\theta}_k|$$

if

$$\beta_{min} \geq \frac{K_1 \lambda}{\gamma}$$

Linear model

- $$Y_i = \beta' X_i + \varepsilon_i$$

Linear model

- $$Y_i = \beta' X_i + \varepsilon_i$$
- $X_i \sim N(0, \Sigma)$

Linear model

- $$Y_i = \beta' X_i + \varepsilon_i$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma = I$ or $\Sigma_{jj} = 1, \Sigma_{jk} = 0.3$

Linear model

- $$Y_i = \beta' X_i + \varepsilon_i$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma = I$ or $\Sigma_{jj} = 1, \Sigma_{jk} = 0.3$
- $\varepsilon \sim$ Cauchy distribution

Linear model

- $$Y_i = \beta' X_i + \varepsilon_i$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma = I$ or $\Sigma_{jj} = 1, \Sigma_{jk} = 0.3$
- $\varepsilon \sim$ Cauchy distribution
- $$\beta = (\underbrace{3, \ldots, 3}_{p_0}, \underbrace{0, \ldots, 0}_{p-p_0})$$

Linear model

- $$Y_i = \beta' X_i + \varepsilon_i$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma = I$ or $\Sigma_{jj} = 1, \Sigma_{jk} = 0.3$
- $\varepsilon \sim$ Cauchy distribution
- $$\beta = (\underbrace{3, \ldots, 3}_{p_0}, \underbrace{0, \ldots, 0}_{p - p_0})$$
- $p_0 \in \{3, 10, 20\}$

Linear model

- $$Y_i = \beta' X_i + \varepsilon_i$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma = I$ or $\Sigma_{jj} = 1, \Sigma_{jk} = 0.3$
- $\varepsilon \sim$ Cauchy distribution
- $$\beta = (\underbrace{3, \ldots, 3}_{p_0}, \underbrace{0, \ldots, 0}_{p-p_0})$$
- $p_0 \in \{3, 10, 20\}$
- $n \in \{100, 200, 300, 400\}$

Linear model

- $$Y_i = \beta' X_i + \varepsilon_i$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma = I$ or $\Sigma_{jj} = 1, \Sigma_{jk} = 0.3$
- $\varepsilon \sim$ Cauchy distribution
- $$\beta = (\underbrace{3, \ldots, 3}_{p_0}, \underbrace{0, \ldots, 0}_{p-p_0})$$
- $p_0 \in \{3, 10, 20\}$
- $n \in \{100, 200, 300, 400\}$
- $p \in \{100, 400, 900, 1600\}$

Estimators

- RankLasso (rL)

Estimators

- RankLasso (rL)

$$\lambda^{rL} = 0.3\sqrt{\frac{\log p}{n}}$$

Estimators

- RankLasso (rL)

$$\lambda^{rL} = 0.3\sqrt{\frac{\log p}{n}}$$

- adaptive RankLasso (arL)

Estimators

- RankLasso (rL)

$$\lambda^{rL} = 0.3\sqrt{\frac{\log p}{n}}$$

- adaptive RankLasso (arL)
- thresholded RankLasso (rLth)

Estimators

- RankLasso (rL)

$$\lambda^{rL} = 0.3 \sqrt{\frac{\log p}{n}}$$

- adaptive RankLasso (arL)
- thresholded RankLasso (rLth)
- LAD-Lasso (LAD)

Estimators

- RankLasso (rL)

$$\lambda^{rL} = 0.3\sqrt{\frac{\log p}{n}}$$

- adaptive RankLasso (arL)
- thresholded RankLasso (rLth)
- LAD-Lasso (LAD)

$$\lambda^{LAD} = 1.5\sqrt{\frac{\log p}{n}}$$

Estimators

- RankLasso (rL)

$$\lambda^{rL} = 0.3\sqrt{\frac{\log p}{n}}$$

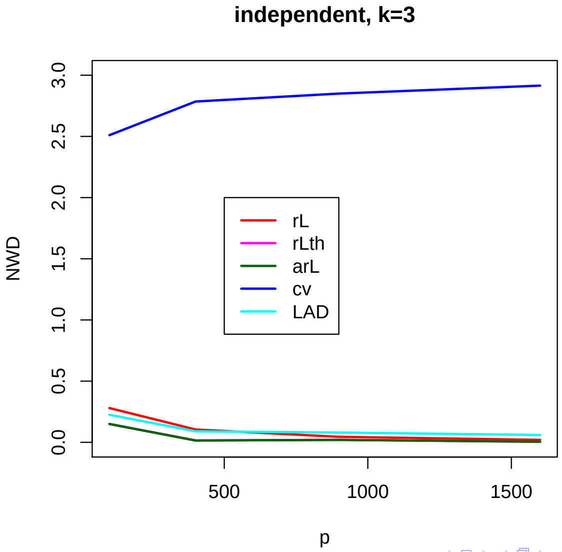
- adaptive RankLasso (arL)
- thresholded RankLasso (rLth)
- LAD-Lasso (LAD)

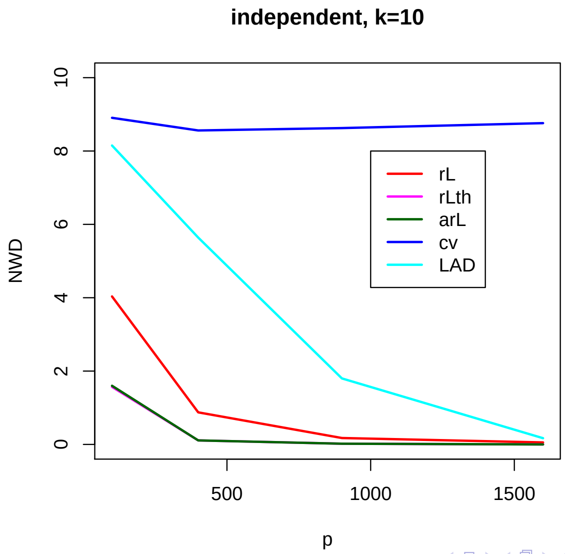
$$\lambda^{LAD} = 1.5\sqrt{\frac{\log p}{n}}$$

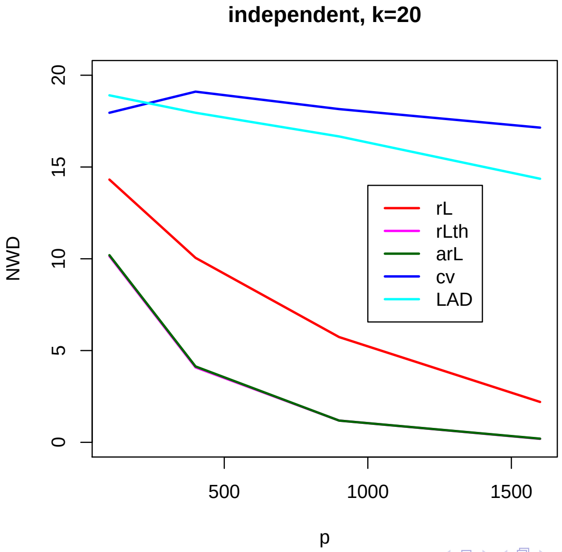
- Lasso with cross-validation (cv)

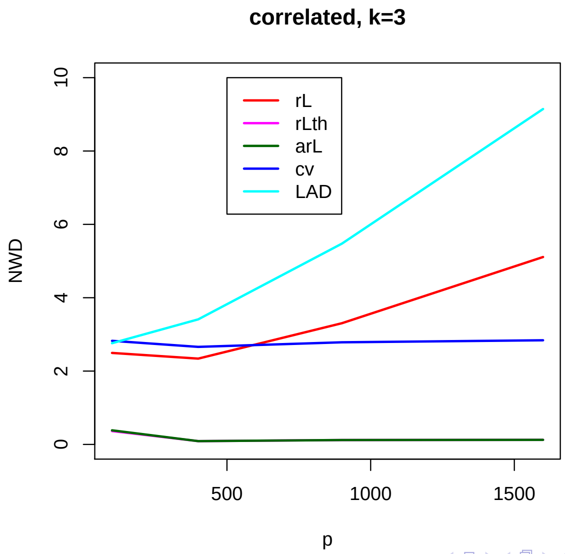
Estimators

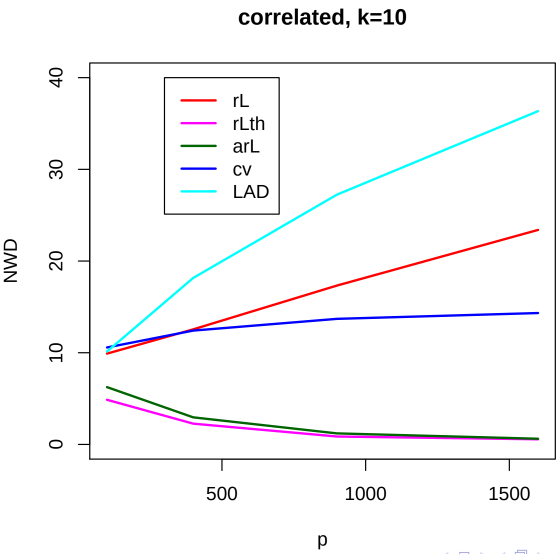
- NWD - average number of wrong decisions

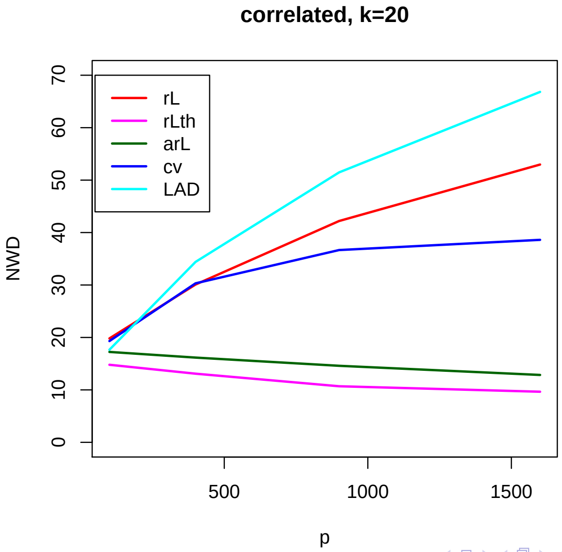

independent, k=3
NWD
0.0
0.5
1.0
1.5
2.0
2.5
3.0
500
1000
1500
p
rL
rLth
arL
cv
LAD


independent, k=10
NWD
0
2
4
6
8
10
500
1000
1500
p
rL
rLth
arL
cv
LAD


independent, k=20
NWD
p
0
5
10
15
20
500
1000
1500
rL
rLth
arL
cv
LAD


correlated, k=3
rL
rLth
arL
cv
LAD
NWD
p
0
2
4
6
8
10
500
1000
1500


correlated, k=10
rL
rLth
arL
cv
LAD
NWD
0
10
20
30
40
500
1000
1500
p


correlated, k=20
rL
rLth
arL
cv
LAD
NWD
0
10
20
30
40
50
60
70
500
1000
1500
p

Exponential model

- $$Y_i = \exp(4 + 0.05\,\beta' X_i) + \varepsilon_i$$

Exponential model

- $$Y_i = \exp(4 + 0.05\, \beta' X_i) + \varepsilon_i$$
- $X_i \sim N(0, \Sigma)$

Exponential model

- $$Y_i = \exp(4 + 0.05\,\beta' X_i) + \varepsilon_i$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.3$

Exponential model

- $$Y_i = \exp(4 + 0.05\,\beta' X_i) + \varepsilon_i$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.3$
- $\varepsilon \sim$ Cauchy distribution

Exponential model

- $$Y_i = \exp(4 + 0.05\,\beta' X_i) + \varepsilon_i$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.3$
- $\varepsilon \sim$ Cauchy distribution
- $$\beta = (\underbrace{3, \ldots, 3}_{p_0}, \underbrace{0, \ldots, 0}_{p-p_0})$$

Exponential model

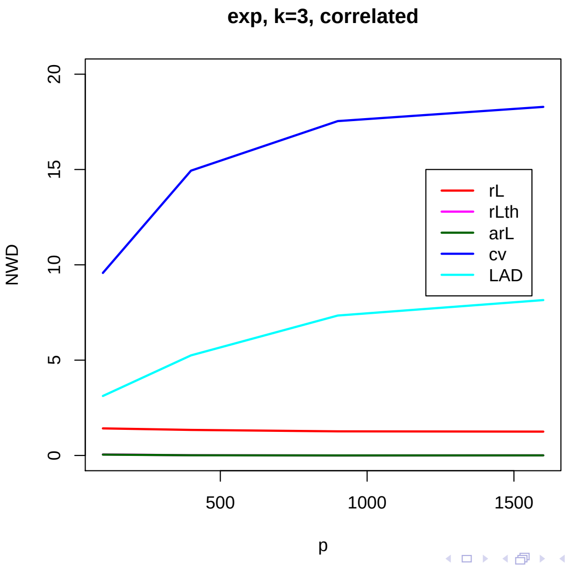
- $$Y_i = \exp(4 + 0.05\,\beta' X_i) + \varepsilon_i$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.3$
- $\varepsilon \sim$ Cauchy distribution
- $$\beta = (\underbrace{3, \ldots, 3}_{p_0}, \underbrace{0, \ldots, 0}_{p - p_0})$$
- $p_0 \in \{3, 10, 20\}$

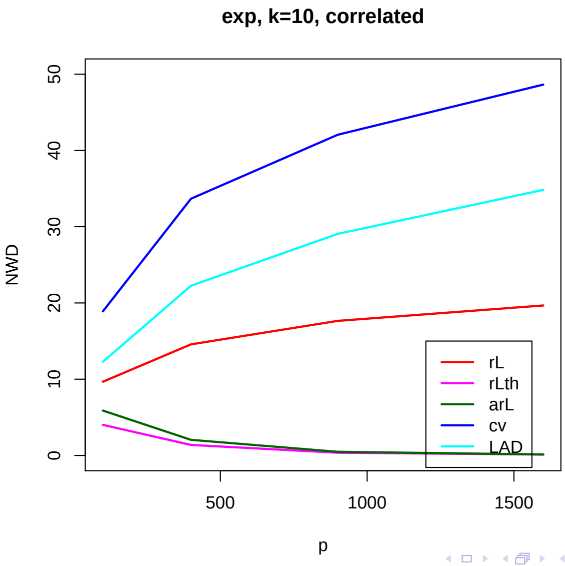
Exponential model

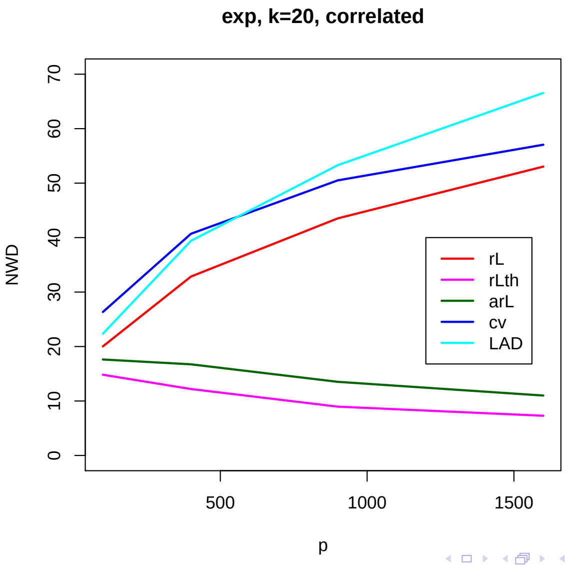
- $$Y_i = \exp(4 + 0.05\,\beta' X_i) + \varepsilon_i$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.3$
- $\varepsilon \sim$ Cauchy distribution
- $$\beta = (\underbrace{3, \ldots, 3}_{p_0}, \underbrace{0, \ldots, 0}_{p-p_0})$$
- $p_0 \in \{3, 10, 20\}$
- $n \in \{100, 200, 300, 400\}$

Exponential model

- $$Y_i = \exp(4 + 0.05\,\beta' X_i) + \varepsilon_i$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.3$
- $\varepsilon \sim$ Cauchy distribution
- $$\beta = (\underbrace{3, \ldots, 3}_{p_0}, \underbrace{0, \ldots, 0}_{p - p_0})$$
- $p_0 \in \{3, 10, 20\}$
- $n \in \{100, 200, 300, 400\}$
- $p \in \{100, 400, 900, 1600\}$


exp, k=3, correlated
NWD
0
5
10
15
20
p
500
1000
1500
rL
rLth
arL
cv
LAD


exp, k=10, correlated
NWD
0
10
20
30
40
50
500
1000
1500
p
rL
rLth
arL
cv
LAD


exp, k=20, correlated
NWD
0
10
20
30
40
50
60
70
500
1000
1500
p
rL
rLth
arL
cv
LAD

Binary regression

- $$P(Y_i = 1|X_i) = \frac{\exp(\beta' X_i)}{\exp(\beta' X_i) + 1}$$

Binary regression

- $$P(Y_i = 1|X_i) = \frac{\exp(\beta' X_i)}{\exp(\beta' X_i) + 1}$$
- $X_i \sim N(0, \Sigma)$

Binary regression

- $$P(Y_i = 1|X_i) = \frac{\exp(\beta' X_i)}{\exp(\beta' X_i) + 1}$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.5$

Binary regression

- $$P(Y_i = 1|X_i) = \frac{\exp(\beta' X_i)}{\exp(\beta' X_i) + 1}$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.5$
- $n \in \{100, 350, 600\}$

Binary regression

- $$P(Y_i = 1|X_i) = \frac{\exp(\beta' X_i)}{\exp(\beta' X_i) + 1}$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.5$
- $n \in \{100, 350, 600\}$
- $p \in \{100, 1225, 3600\}$

Binary regression

- $$P(Y_i = 1|X_i) = \frac{\exp(\beta' X_i)}{\exp(\beta' X_i) + 1}$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.5$
- $n \in \{100, 350, 600\}$
- $p \in \{100, 1225, 3600\}$
- $$\beta = (\underbrace{1, -1, -1, 1, \ldots, 1}_{10}, \underbrace{0, \ldots, 0}_{p-10}, 1)$$

Binary regression

- $$P(Y_i = 1|X_i) = \frac{\exp(\beta' X_i)}{\exp(\beta' X_i) + 1}$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.5$
- $n \in \{100, 350, 600\}$
- $p \in \{100, 1225, 3600\}$
- $$\beta = (\underbrace{1, -1, -1, 1, \dots, 1}_{10}, \underbrace{0, \dots, 0}_{p-10}, 1)$$
- signs are chosen randomly

Binary regression

- $$P(Y_i = 1|X_i) = \frac{\exp(\beta' X_i)}{\exp(\beta' X_i) + 1}$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.5$
- $n \in \{100, 350, 600\}$
- $p \in \{100, 1225, 3600\}$
- $$\beta = (\underbrace{1, -1, -1, 1, \ldots, 1}_{10}, \underbrace{0, \ldots, 0}_{p-10}, 1)$$
- signs are chosen randomly
- intercept $= 1$

Estimators

- Lasso with $Y_i \in \{0, 1\}$

$$\min_{\theta} \quad \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \theta' X_i \right)^2 + \lambda \left| \theta \right|_1$$

Estimators

- Lasso with $Y_i \in \{0, 1\}$

$$\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \theta' X_i\right)^2 + \lambda\,|\theta|_1$$

λ chosen by CV

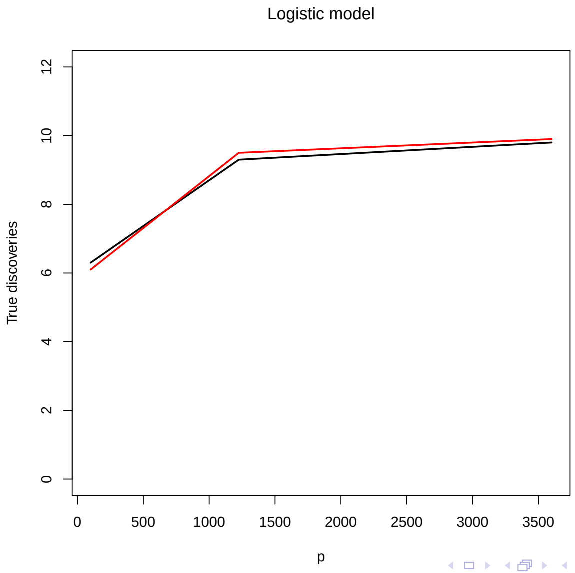
Estimators

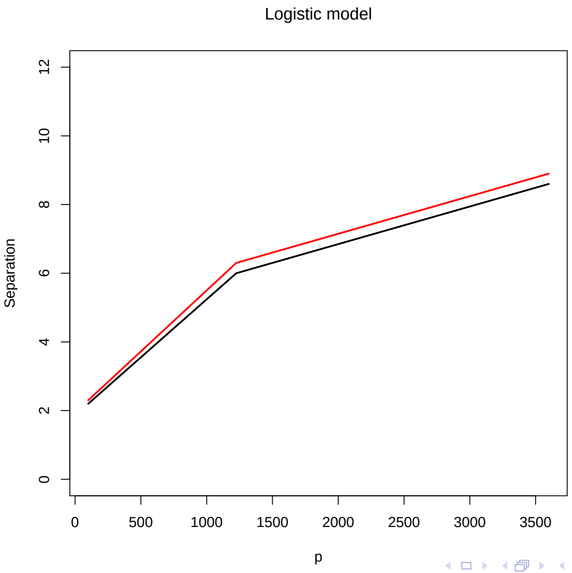
- Lasso with $Y_i \in \{0,1\}$

$$\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \theta' X_i\right)^2 + \lambda\,|\theta|_1$$

  λ chosen by CV

- Lasso for logistic regression with λ chosen by CV


Logistic model
True discoveries
0
2
4
6
8
10
12
0
500
1000
1500
2000
2500
3000
3500
p


Logistic model
Separation
0
2
4
6
8
10
12
0
500
1000
1500
2000
2500
3000
3500
p

Cauchy model

- $$P(Y_i = 1|X_i) = \arctan(\beta' X_i)/\pi + 0.5$$

Cauchy model

- $$P(Y_i = 1|X_i) = \arctan(\beta' X_i)/\pi + 0.5$$
- $X_i \sim N(0, \Sigma)$

Cauchy model

- $$P(Y_i = 1|X_i) = \arctan(\beta' X_i)/\pi + 0.5$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.5$

Cauchy model

- $$P(Y_i = 1|X_i) = \arctan(\beta' X_i)/\pi + 0.5$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.5$
- $n \in \{100, 350, 600\}$

Cauchy model

- $$P(Y_i = 1|X_i) = \arctan(\beta' X_i)/\pi + 0.5$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.5$
- $n \in \{100, 350, 600\}$
- $p \in \{100, 1225, 3600\}$

Cauchy model

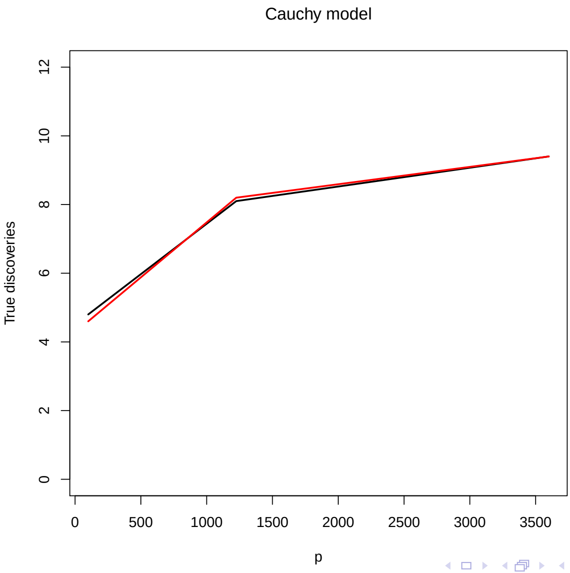
- $$P(Y_i = 1|X_i) = \arctan(\beta' X_i)/\pi + 0.5$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.5$
- $n \in \{100, 350, 600\}$
- $p \in \{100, 1225, 3600\}$
- $$\beta = (\underbrace{1, -1, -1, 1, \ldots, 1}_{10}, \underbrace{0, \ldots, 0}_{p-10}, 1)$$

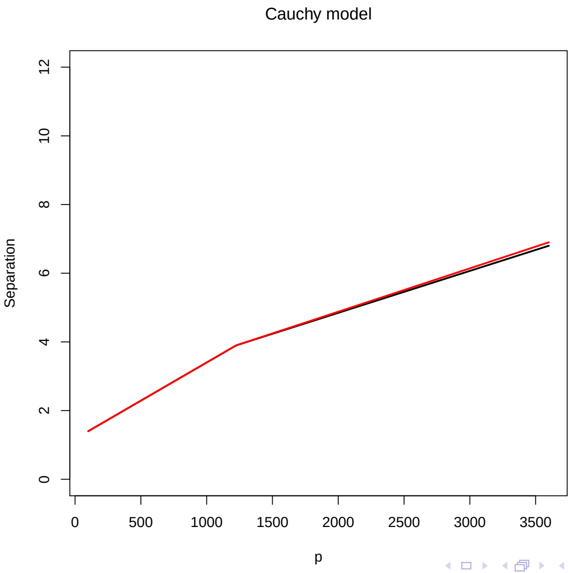
Cauchy model

- $$P(Y_i = 1|X_i) = \arctan(\beta' X_i)/\pi + 0.5$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.5$
- $n \in \{100, 350, 600\}$
- $p \in \{100, 1225, 3600\}$
- $$\beta = (\underbrace{1, -1, -1, 1, \ldots, 1}_{10}, \underbrace{0, \ldots, 0}_{p-10}, 1)$$
- signs are chosen randomly

Cauchy model

- $$P(Y_i = 1|X_i) = \operatorname{arctan}(\beta' X_i)/\pi + 0.5$$
- $X_i \sim N(0, \Sigma)$
- $\Sigma_{jj} = 1, \Sigma_{jk} = 0.5$
- $n \in \{100, 350, 600\}$
- $p \in \{100, 1225, 3600\}$
- $$\beta = (\underbrace{1, -1, -1, 1, \ldots, 1}_{10}, \underbrace{0, \ldots, 0}_{p-10}, 1)$$
- signs are chosen randomly
- intercept $= 1$


Cauchy model
True discoveries
0
2
4
6
8
10
12
0
500
1000
1500
2000
2500
3000
3500
p


Cauchy model
Separation
p
0
2
4
6
8
10
12
0
500
1000
1500
2000
2500
3000
3500

References

Fan, J., Fan, Y., and Barut, E. (2014). *Adaptive robust variable selection.* Ann. Statist., 42:324–351.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). *Regularization paths for generalized linear models via coordinate descent*. Journal of Statistical Software, 33:1–22.

Furmańczyk, K., Rejchel, W. (2020). *Prediction and variable selection in high-dimensional misspecified binary classification.* Entropy, *accepted.*

Hall, P., Li, K.-C. (1993). *On almost linearity of low dimensional projections from high dimensional data.* Ann. Statist., 21:867–889.

Li, K.-C. and Duan, N. (1989). *Regression analysis under link violation.* Annals of Statistics, 17:1009–1052.

R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2013.

Rejchel, W., Bogdan, M. (2019). *Rank-based Lasso - efficient methods for high-dimensional robust model selection*, arXiv:1905.05876v2.

Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso.* J. R. Statist. Soc. B, 58, 267–288.

Ye, F. and Zhang, C. H. (2010). *Rate minimaxity of the lasso and dantzig selector for the l_q loss in l_r balls*. J. Mach. Learn. Res., 11:3519–3540.