SLOPE for GLM

Michal Kos, Malgorzata Bogdan, Sangkyun Lee

This is a preliminary report.
Please do not cite or publish without permission.

November 24, 2017

Contents

1 Introduction

1.1 Basics

In this paper we consider generalized linear models with natural link functions. Let's assume we have a response vector $y = (y_1, ..., y_n)^T$ whose elements are independent random variables from a distribution with density:

$$f(y_i|X_i b) = c(y_i)exp\{y_i X_i b - m(X_i b)\} \quad i = 1, 2, ...$$

where:
c - non-negative measurable function,
$b \in \mathbb{R}$ - unknown vector of parameters,
X_i - i-th row in design matrix,
m - smooth function,
Based on the above relation, we can express the log-likelihood function as follows:

$$s(b) = \sum_{i=1}^{n} (y_i X_i b - m(X_i b)) - C$$

One can show that:

$$\mu(\theta) = E_\theta y = \frac{dm(\theta)}{d\theta}$$
$$\Sigma(\theta) = cov_\theta y = \frac{d^2 m(\theta)}{d\theta^2}$$

1.2 Formulation of a problem

We are testing n hypotheses:

$$H_i \; : \; b_i = 0, \; i = 1, ..., p$$

and we reject $H_i \;\Leftrightarrow\; \hat{b}_i \neq 0$.
Let's denote V and R as the number of false rejections and all rejections, respectively. Moreover without loss of generality we reorganize vector b in a way that its first p_0 elements are truly equal to 0. For such setting the false discovery rate (FDR) is defined in a following way:

$$FDR := \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \sum_{r=1}^{p} \frac{1}{r} \sum_{i=1}^{p_0} P\left(\{H_i \; is \; rejected \; and \; R = r\}\right)$$

We would like to find a procedure of selection of a positive and non-increasing vector λ for which the following optimization problem would control the FDR at a predetermined level q:

$$\hat{b} = \min_b \left(\underbrace{-l(b) + \sum \lambda_i |b_{(i)}|}_{f(b)} \right)$$

where $l(b) = s(b)/n$ and $b_{(i)}$ is the element of vector b with i-th largest absolute value ($|b_{(1)}| \geqslant |b_{(2)}| \geqslant ... \geqslant |b_{(p)}|$).

2 Asymptotic control of FDR

2.1 Theorem

Assumptions:

1. for all i, j x_{ij}, are independent and $Ex_{ij} = 0$
2. elements in each column of design matrix are i.i.d.
3. $\frac{1}{n}\mathbb{E}[X^T X] = I$
4. p - const.
5. y_i are independent and have conditional density given X_i:
   $f(y_i|X_i b) = c(y_i) exp\{y_i X_i b - m(X_i b)\} \quad i = 1, 2, ...$
6. $\lambda_1 \geqslant ... \geqslant \lambda_p$ and $\lambda_i = \frac{\sqrt{\Sigma(0)}}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{qi}{2p}\right)$,
7. the parameter vector $b^0(n) = \tilde{b}^0/n^s$ where $\tilde{b}^0 = const.$ and $s \in (0, 0.5]$, or we consider linear regression and $b^0(n) = const.$
8. $\lim_{n\to\infty} F_n(b^0)/n = G(b^0)$ and $\infty > \lambda_{max}(G(b^0)) \geqslant \lambda_{min}(G(b^0)) > 0$ where $F_n(b) = \sum_{i=1}^n X_i' \Sigma(X_i b) X_i$
9. There exist $\delta > 0$ and $d > 0$ such that for any element of $\{b : \|b - b^0\| < \delta\}$ and sufficiently large n:
   $F_n(b) - dF_n(b^0)$ is positive semidefinite,

With the above assumptions the SLOPE procedure asymptotically controls FDR in a sense that the following inequality holds:

$$FDR_n \leqslant D \to \frac{qp_0}{p}$$

where p_0 is number of zero elements in b^0.

3 Simulations for SNIPS

In this section we present results of simulations for logistic regression - one of the generalized linear models in genetic context.

3.1 Design matrix X

The elements of design matrix X had following properties:

- X_{ij} were independent, and i.i.d. in each column,
- $E(X_{ij}) = 0$
- $\frac{1}{n}E(X^T X) = I$

Simulations were performed for SNIPS - for each column we generated probability $q_j \in [0.1, 0.5]$ for which:

$$P(X_{ij} = 0) = q_j^2$$
$$P(X_{ij} = 1) = 2q_j(1 - q_j)$$
$$P(X_{ij} = 2) = (1 - q_j)^2$$

The distribution of x_{ij} can be obtain as a sum of two i.i.d. variables drawn from Bernoulli distribution with probability of success equal q_j. This is related to the fact that chromosome has two chromatids.
In all simulations number of columns was fixed p = 500. On the other hand the number of rows varied - $n = 250, 500, 1000, 2000$.

3.2 Response vector y

The elements of vector y were generated from following distribution:

$$P(y_i = 1) = 1 - P(y_i = -1) = p_i = \frac{1}{1 + exp(-x_i^T b)}$$

where x_i^T was the i-th row in design matrix X and b is parameter vector.

3.3 Estimation of vector b

In order to estimate the unknown vector b we used several methods.

3.3.1 SLOPE

The estimator is obtained via optimization of following problem:

$$\hat{b} = arg\min_{b}\left(\frac{1}{n}\sum_{i=1}^{n} log(1+exp(-y_i x_i^T b)) + 0.5\frac{1}{\sqrt{n}}\sum_{j=1}^{p}\lambda_j|b|_{(j)}\right)$$

where:

$$\lambda_j = \Phi^{-1}\left(1-\frac{qj}{2p}\right)$$

3.3.2 MLE

The estimator is the maximum likelihood estimator.

3.3.3 SLOPE MLE

The estimator is the maximum likelihood estimator constructed on a basis of columns of design matrix associated with nonzero elements of slope estimator (for other columns the estimator was 0).

3.3.4 χ^2 MLE

In order to obtain this estimator in first step we calculated p-values of χ^2 independence test for each column of design matrix relatively to the response vector y. Next the Benjamini-Hochber procedure was applied to the set of p-values. The estimator was the maximum likelihood estimator constructed on columns chosen by Benjamini-Hochber procedure (for other columns the estimator was 0).

3.3.5 Lasso 1

The estimator is the Lasso estimator with:

$$\lambda = 0.5\frac{1}{\sqrt{n}}\Phi^{-1}\left(1-\frac{q}{2p}\right) = 0.5\frac{1}{\sqrt{n}}\lambda_1$$

3.3.6 Lasso cv

The estimator is the Lasso estimator with λ chosen via 10-fold cross validation. We used Matlab function lassoglm with changed parameters: NumLambda = 25 (default 100) and RelTol = 0.05 (Convergence threshold for the coordinate descent algorithm, default 0.0001). We have changed parameters to obtain results in reasonable time (approx. 16 h simulations for one value of k - number of nonzero elements in true parameter vector).

3.4 Examined features

FDR - False discovery rate,
Power - ratio of expected number of correctly rejected hypothesis to k (number of nonzero elements in true parameter vector),
MSE - mean squared error of estimator,
Mean squared error of prediction $\mathbb{E}(Y-\hat{Y})^2$,
Mean squared error of probability $\mathbb{E}(p-\hat{p})^2$,

3.5 Design of simulations

For each configuration of k and n we simulated the design matrix X and response vector y. Next we calculated estimators and examined features. For predictive analysis ("mean squared error of probability" and "mean squared error of prediction") we simulated additional 200 testing observations. In order to obtain averaged results we repeated simulations 500 times for each configuration of k and n.

3.6 Parameter vector b

The vector of parameters b had following form:

$$b = (\underbrace{b_1, b_1, ..., b_1}_{p-p_0\ elem.}, \underbrace{0, 0, ..., 0}_{p_0\ elem.})^T$$

where $b_1 = 9/\sqrt{n}$ (local alternatives) and $b_1 = 5/\sqrt{250}$ (classical situation).
Moreover, we changed the number of nonzero elements in b:

$$k = p - p_0 = 0, 1, 5, 10, 20, 50$$

3.7 Results for local alternatives $b_1 = 9/\sqrt{n}$

For this scenario of simulations SLOPE, χ^2 and Lasso 1 controls FDR on predefined level ($q = 0.1$). The FDR level for SLOPE and χ^2 for different configurations is close to 0.1. The Lasso 1 is the most conservative method and FDR seems to decrease towards 0 as k or n increases. The Lasso CV on the other hand do not control FDR and its values are quite high (in most cases $FDR > 0.5$).
The highest power has Lasso CV. This is quite natural since it is the only method that do not control FDR. Out of methods that control FDR the highest power has SLOPE. Its power is about 10% higher then power of χ^2 (second best method).
Lasso 1 and slope has globally the highest MSE. It is natural consequence of biasness of this estimators. It is noticeable that Slope mle has always smaller MSE than

χ^2 and that both methods are very strong for small k. The lasso cv is the best method when k is big and n small (simultaneously). However when n grows Slope mle and χ^2 outperform the Lasso cv.
Behavior of mean squared error of probability MSEP is analogous to that for MSE. The mean squared error of prediction for methods that control FDR is smallest for Slope mle. Next best is χ^2.The worst are Slope and Lasso 1. For this methods the error is close to maximal value 2. The Lasso cv predicts best for small n and big k. However for large enough n Slope mle and χ^2 would probably outperform the Lasso cv.

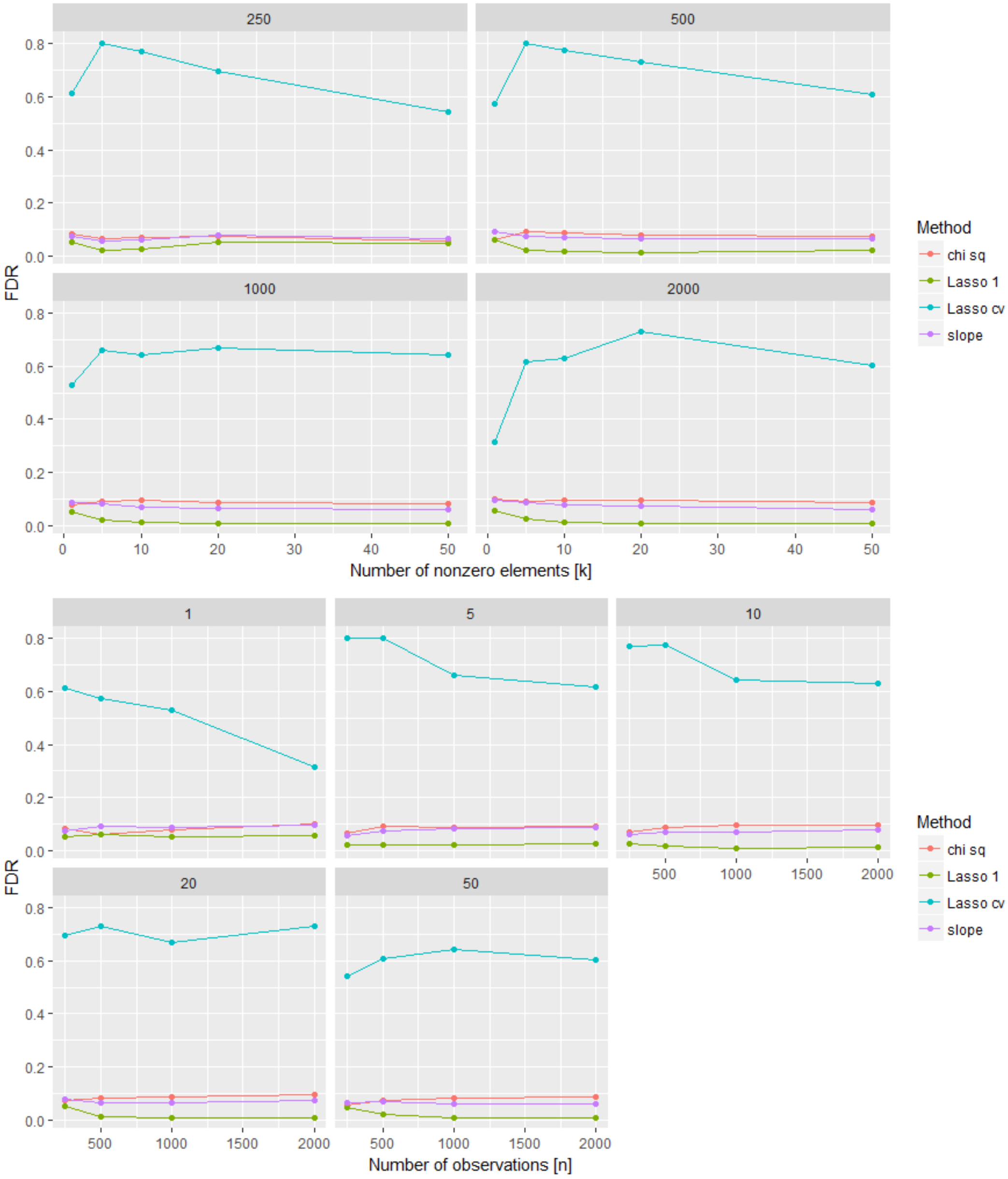


Figure 1: FDR as a function of k and n.

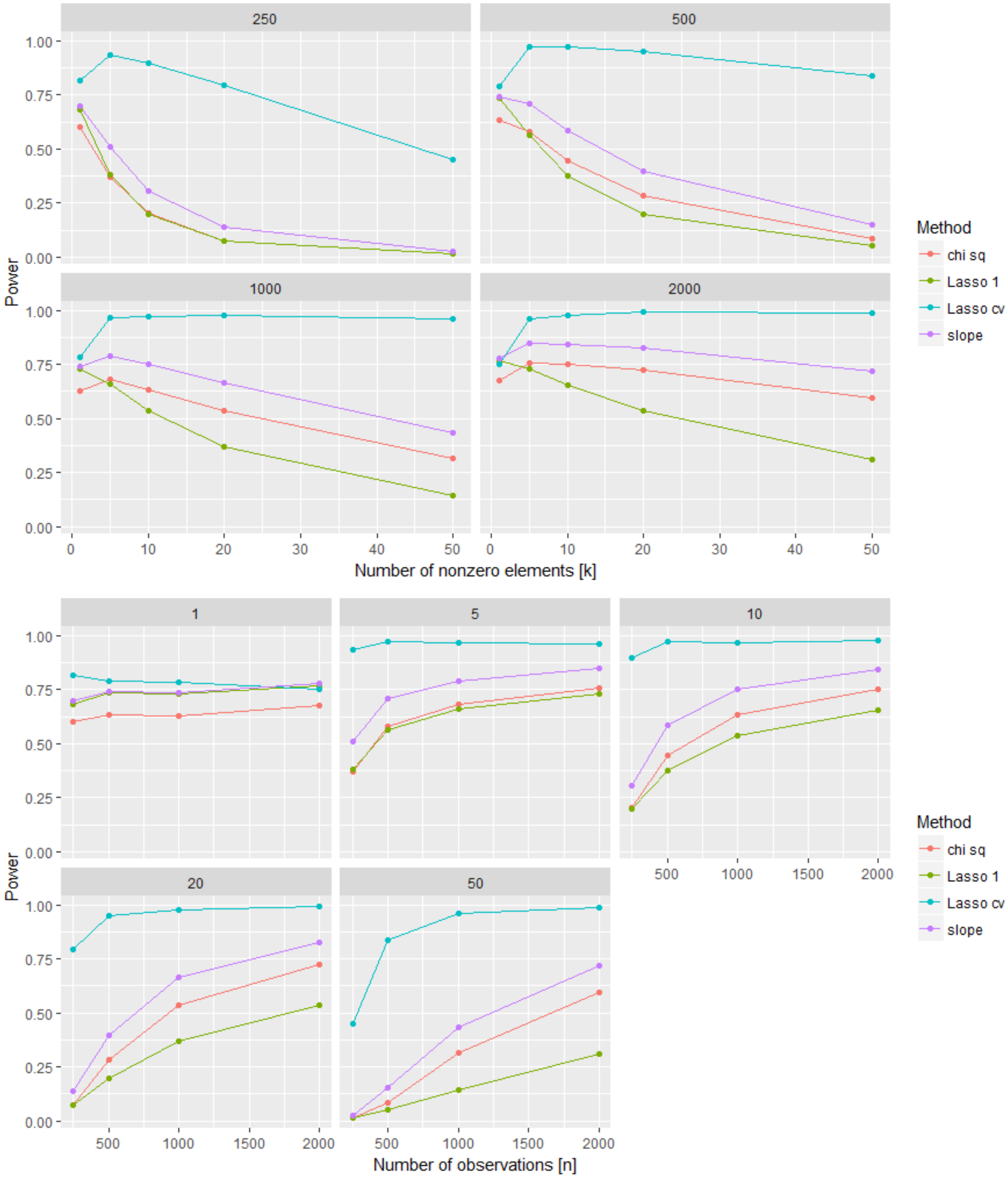


Figure 2: Power

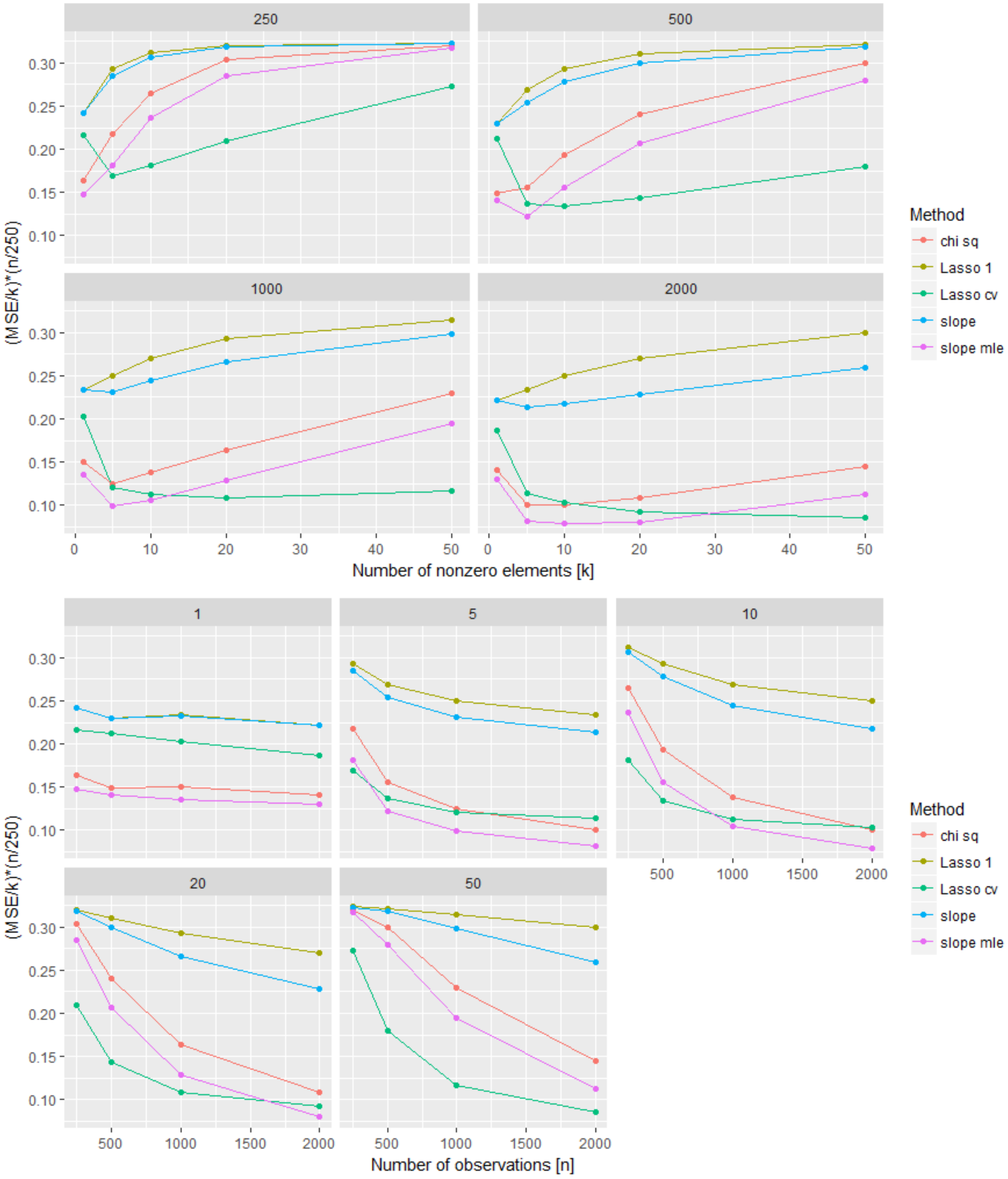


Figure 3: Modified MSE

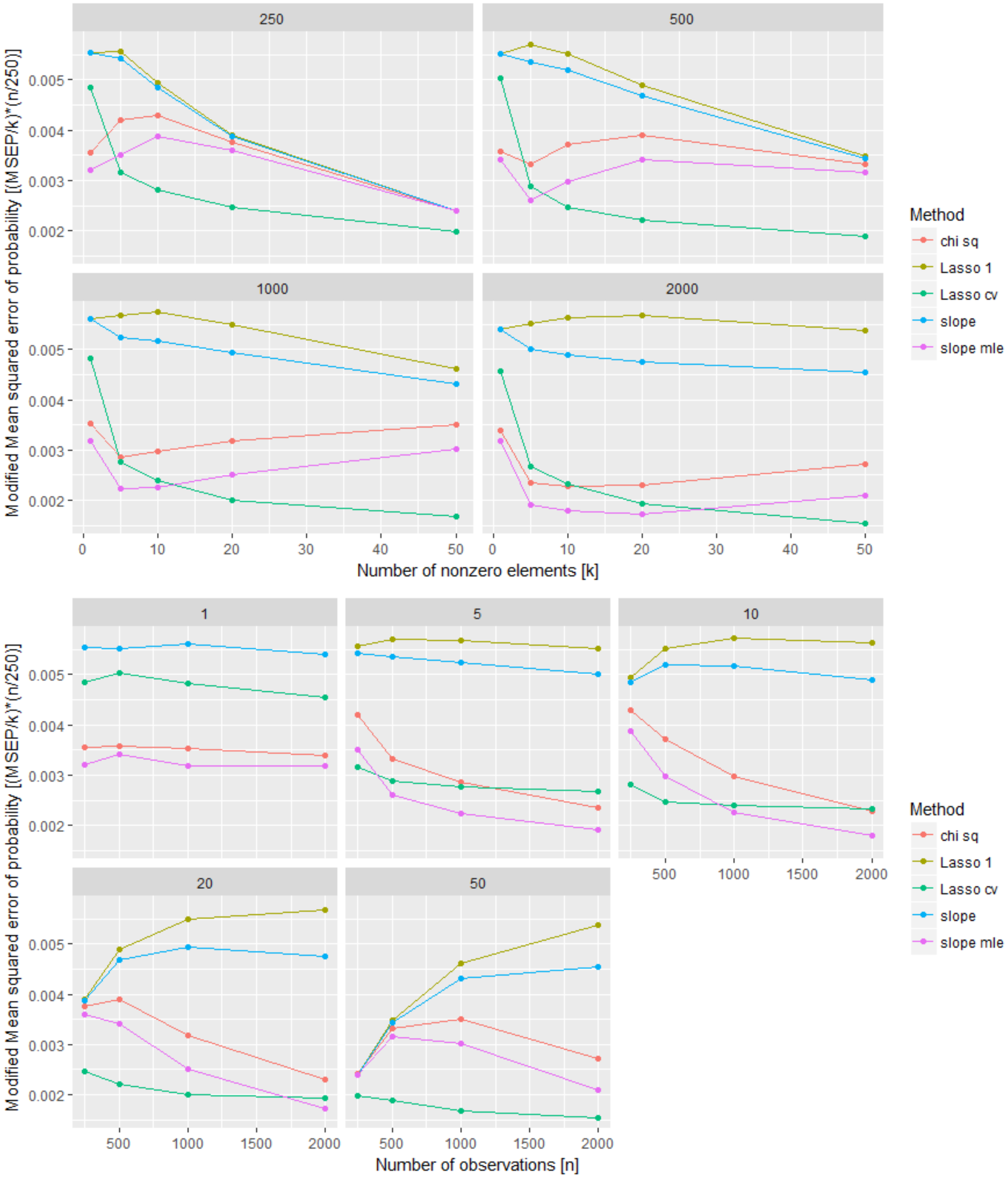


Figure 4: Modified MSEP.

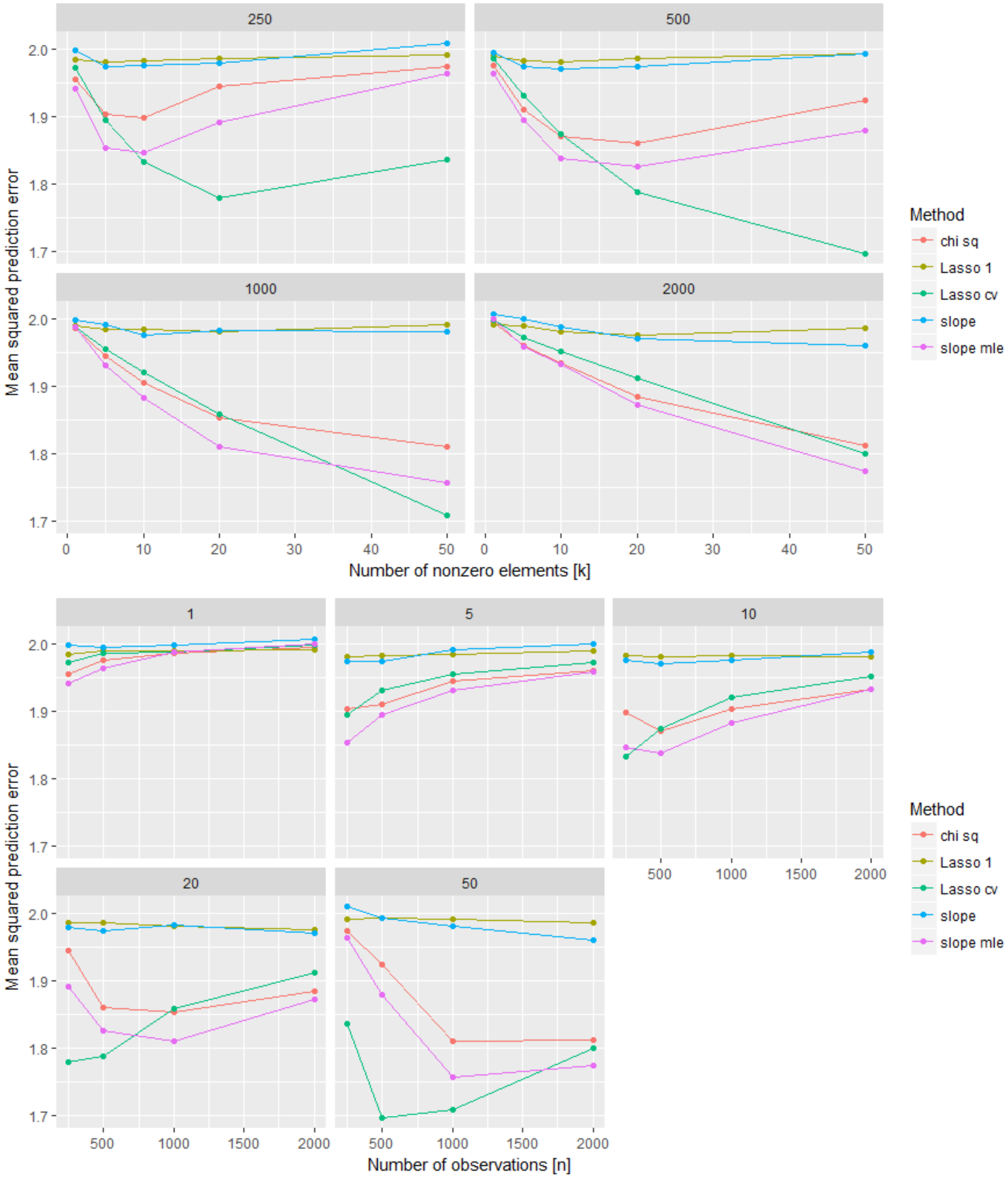


Figure 5: Mean squared error of prediction

3.8 Results for classic assumptions $b_1 = 5/\sqrt{250}$

The mutual behavior of methods for different features is analogous to that from previous section.
However, there are several differences between results for classical assumptions and local alternatives. First of all the Power of methods for classical case rise much faster - for $n = 2000$ many of them detects almost all nonzero elements of true parameter vector. There is also noticeable that mean squared error of prediction is lower for classical case. The results for MSE and MSEP are hard to compare due to differences in "normalization".

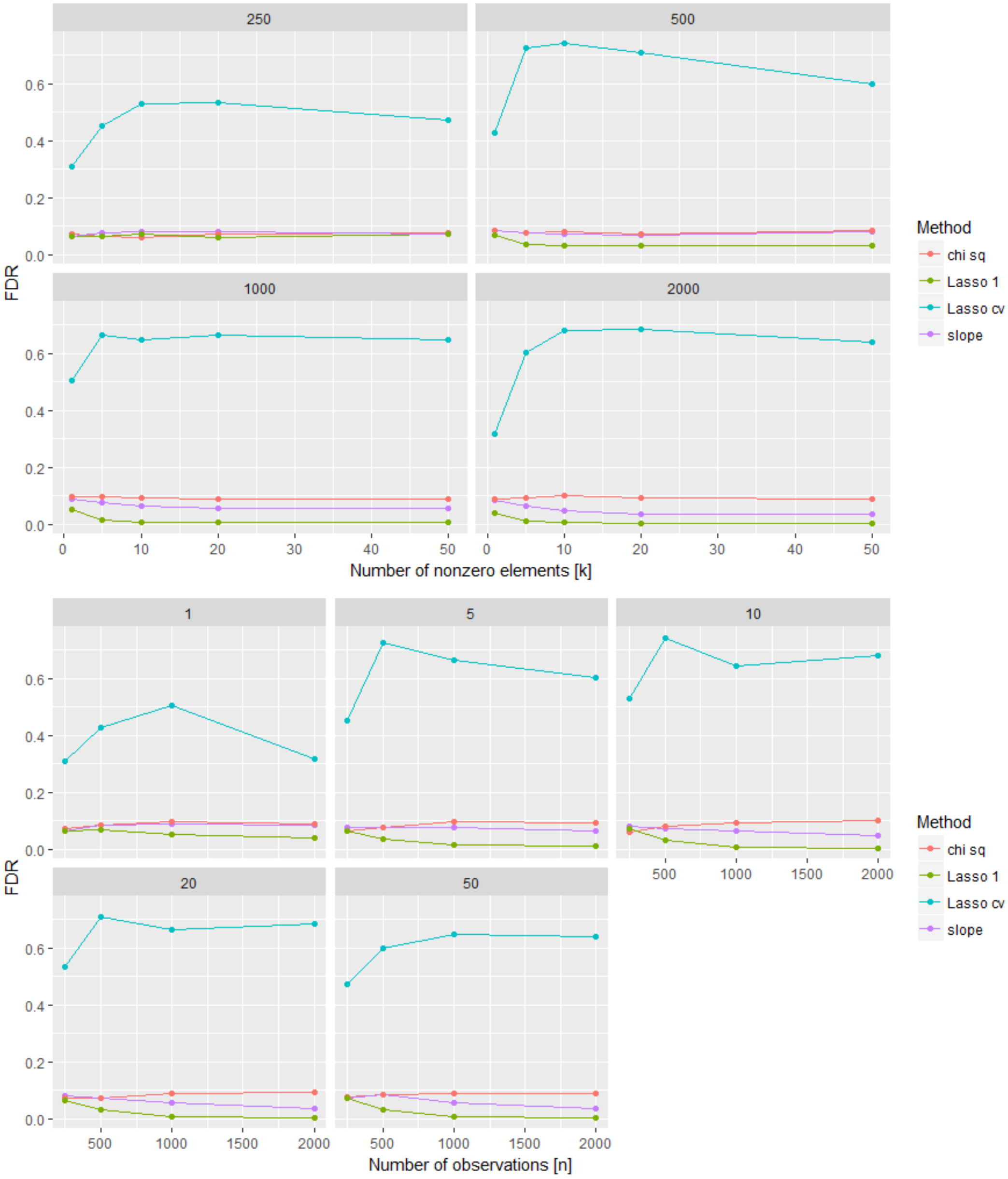

250
500
1000
2000
FDR
0.0
0.2
0.4
0.6
0
10
20
30
40
50
Number of nonzero elements [k]
Method
chi sq
Lasso 1
Lasso cv
slope
1
5
10
20
50
500
1000
1500
2000
Number of observations [n]

Figure 6: FDR

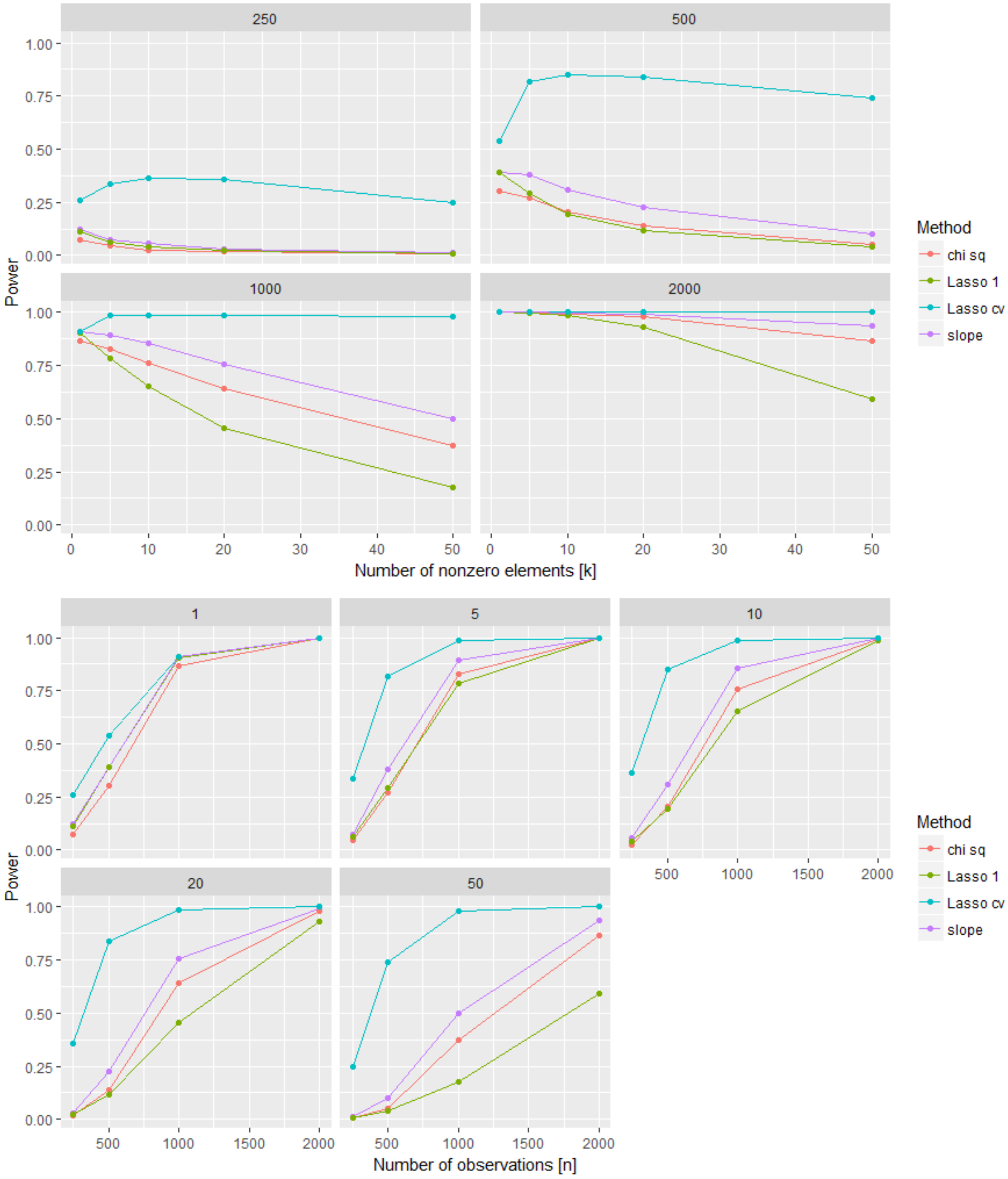


Figure 7: Power

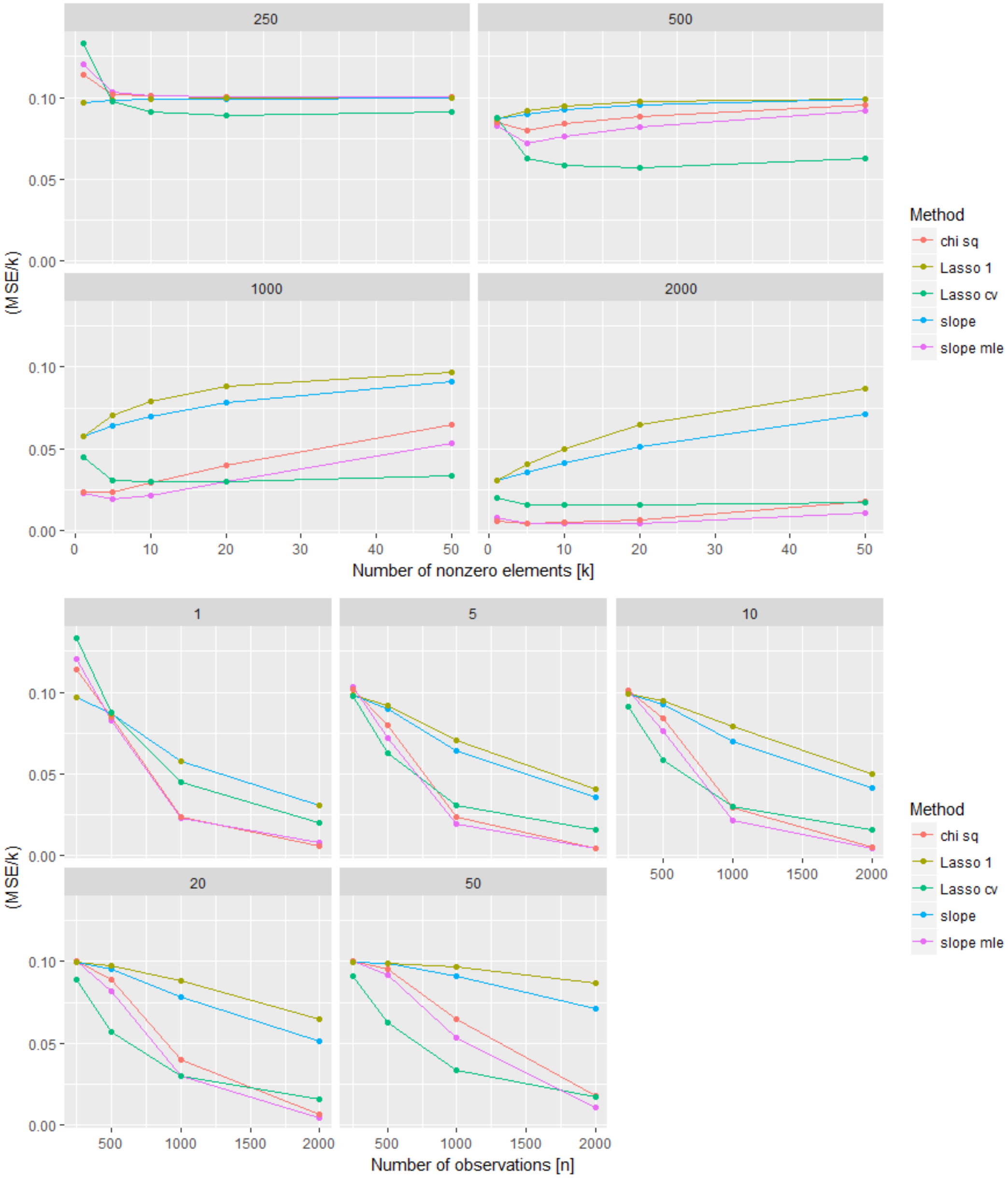


Figure 8: Modified MSE.

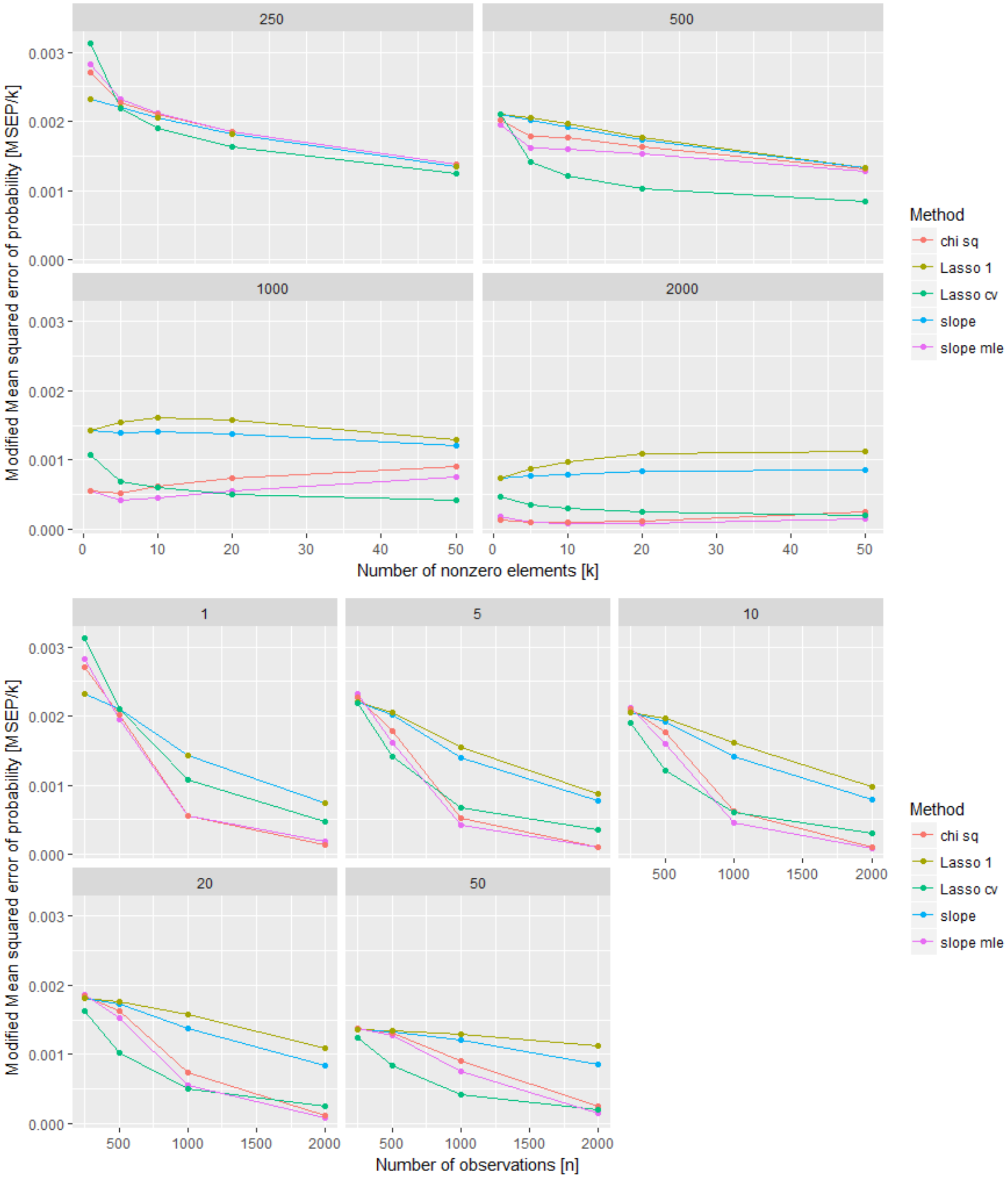
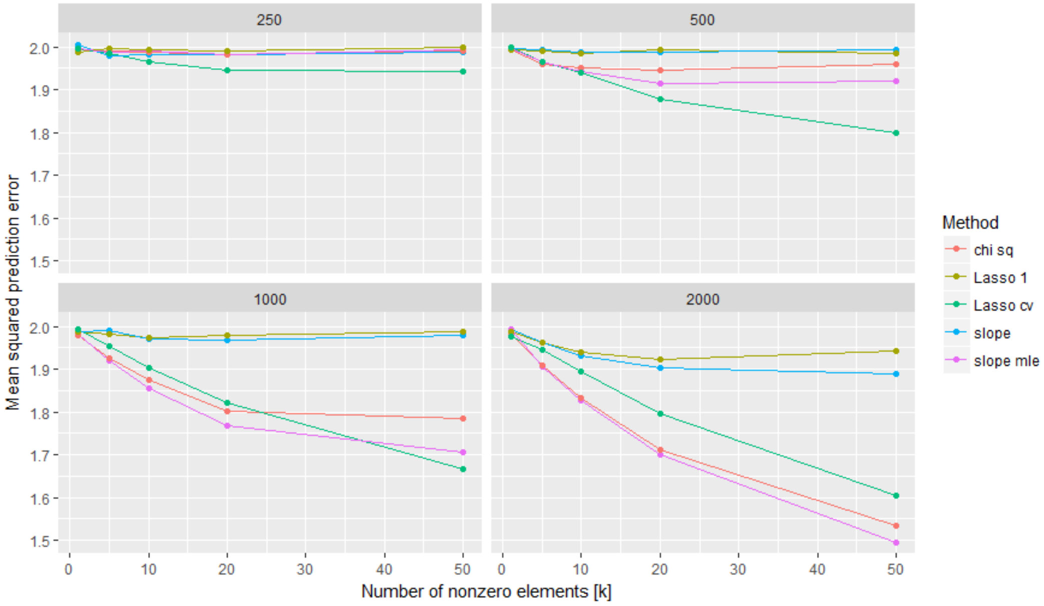


Figure 9: Modified MSEP.



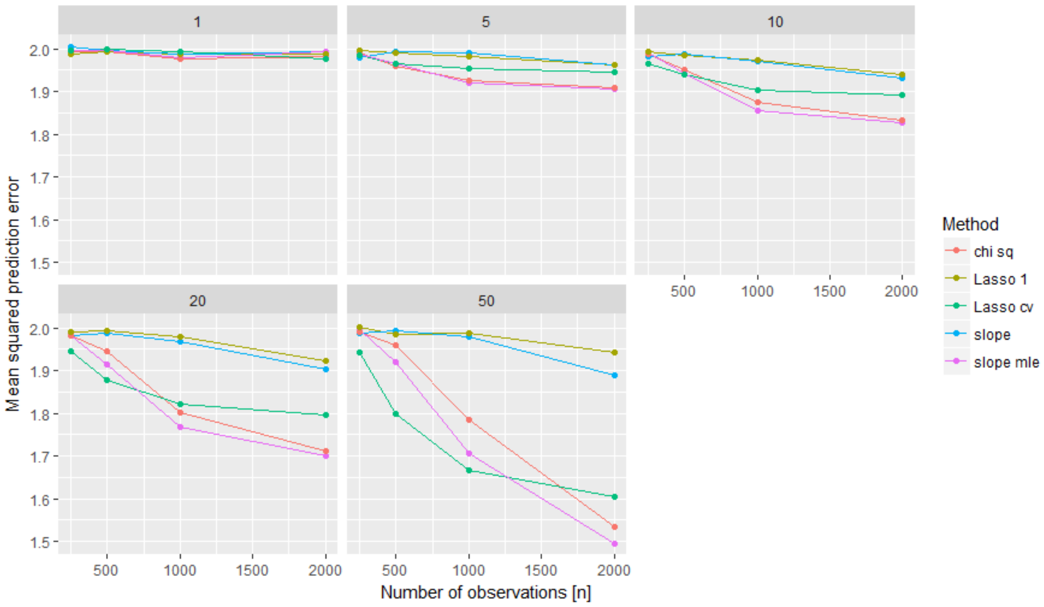


Figure 10: Mean squared error of prediction

4 Appendix

In order to prove the theorem we will use following lemmas (the proofs one can find further in an appendix). Let's introduce a following notation on the vector $U(b) = (U_1(b), ..., U_p(b))^T$ where $U_j(b) = \dfrac{\partial l(b)}{\partial b_j}$. Furthermore let's define SLOPE estimator as:

$$\hat{b} = \min_b \left(-l(b) + \sum \lambda_i |b_{(i)}|\right)$$

Statement
If $|\hat{b}_j| > |\hat{b}_k|$ then $|U_j(\hat{b})| > |U_k(\hat{b})|$.

Let's introduce a notation $U_{(j)}(\hat{b})$ on the statistic associated with the element $\hat{b}_{(j)}$. From the above statement we have that if $|\hat{b}_{(j)}| > |\hat{b}_{(k)}|$ then $|U_{(j)}(\hat{b})| > |U_{(k)}(\hat{b})|$. On the other hand if $|\hat{b}_{(j)}| = |\hat{b}_{(j+1)}|$ the indexing is ambiguous and therefore we organize it in a way that associated statistics have following property:

$$|U_{(j)}(\hat{b})| \geqslant |U_{(j+1)}(\hat{b})|$$

This way we obtain that $|U_{(1)}(\hat{b})| \geqslant ... \geqslant |U_{(p)}(\hat{b})|$.

Lemma 1
When $R = r$, than for any sequence $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_p \geqslant 0$, arbitrary n and arbitrary $a > 0$, it holds:

$$\left[\hat{b}_i \neq 0\right] \Leftrightarrow \left[\lambda_r < |U_i(\hat{b}) + a\hat{b}_i|\right]$$

Lemma 1 enables characterization of event $\{H_i\ is\ rejected\ and\ R = r\}$

Corollary
For any sequence $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_p \geqslant 0$, arbitrary n and arbitrary $a > 0$, it holds:
$\{H_i\ is\ rejected\ and\ R = r\} = \{\lambda_r < |U_i(\hat{b}) + a\hat{b}_i|\ and\ R = r\}$

Lemma 2
For any sequence $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_p \geqslant 0$ and arbitrary, strictly positive (all elements bigger than 0) sequence $\{a_i\}_{i=1}^p$, the event $\{R = r\}$ occurs if and only if both following conditions are satisfied simultaneously:
(i) For any $j \leqslant r$

$$\sum_{i=j}^{r} \lambda_i < \sum_{i=j}^{r} |U_{(i)}(\hat{b}) + a_i\hat{b}_{(i)}|$$

(ii) and for any $j \geqslant r+1$

$$\sum_{i=r+1}^{j} \lambda_i \geqslant \sum_{i=r+1}^{j} |U_{(i)}(\hat{b}) + a_i \hat{b}_{(i)}|$$

This lemma enables characterization of event $\{R = r\}$ in terms of properties of scoring vector.

Lemma 3
Assumptions:

1. for all i, j x_{ij}, are independent and $Ex_{ij} = 0$
2. elements in each column of design matrix are i.i.d.
3. $\frac{1}{n}\mathbb{E}[X^T X] = I$
4. p - const.
5. y_i are independent and have conditional density given X_i:
   $f(y_i|X_i b) = c(y_i) exp\{y_i X_i b - m(X_i b)\} \quad i = 1, 2, ...$
6. $\lambda_1 \geqslant ... \geqslant \lambda_p$ and $\lambda_i = \frac{c_i}{\sqrt{n}}$ where c_i is constant,
7. $\lim_{n\to\infty} F_n(b^0)/n = G(b^0)$ and $\infty > \lambda_{max}(G(b^0)) \geqslant \lambda_{min}(G(b^0)) > 0$ where $F_n(b) = \sum_{i=1}^{n} X_i' \Sigma(X_i b) X_i$
8. There exist $\delta > 0$ and $d > 0$ such that for any element of $\{b : \|b - b^0\| < \delta\}$ and sufficiently large n:
   $F_n(b) - dF_n(b^0)$ is positive semidefinite,

With the above assumptions, $\|\sqrt{n}(\hat{b} - b^0)\|$ is bounded in probability.

Lemma 4.1
Let's assume that real parameter vector $b^0(n) = \tilde{b}^0/n^s$ where $\tilde{b}^0 = const.$ and $s \in (0, 0.5]$. Furthermore assume all conditions from Lemma 3. Under this assumptions:

$$Z = \frac{\sqrt{n}}{\sqrt{\Sigma(0)}}(U(\hat{b}) + \Sigma(0)\hat{b}) \; is \; AN(\sqrt{n}\sqrt{\Sigma(0)}b^0, \mathbb{I})$$

Lemma 4.2
For linear regression under assumptions of lemma 3, thesis of lemma 4.1 holds for $s = 0$ (real parameter vector $b^0(n) = const.$)

Lemma 4.1 and 4.2 together with lemma 1 and 2 show that for carefully chosen sequence $a = (a_1, ..., a_p)^T$ asymptotically event $\{H_i\ is\ rejected\ and\ R = r\} = \{\lambda_r < |U_i(\hat{b}) + a\hat{b}_i|\ and\ R = r\}$ do not depend on $\hat{b}$.

4.1 proof of Theorem

Let us consider method SLOPE applied to the data modified in reference to original by exclusion of $j-th$ column from the design matrix for which the $b_j = 0$ (rest of the data remains unmodified). In consequence of absence of the $j-th$ element in vector b we consider new vector $\tilde{\lambda}$, whose relationship with original vector has following form $\tilde{\lambda}_i = \lambda_{i+1}$. Let's denote the number of rejections for the procedure defined on modified data by $\tilde{R}$. Suppose $r \geqslant 1$ and $\lambda_i = \frac{\sqrt{\Sigma(0)}}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{qi}{2p}\right)$ (sequence satisfying assumptions of Lemma 1 and 2). Let's assume that assumptions of Lemma 3 are satisfied and lemma 4.1 or 4.2 for original data. Then the following inequality holds:

$$FDR_n \leqslant D \to \frac{qp_0}{p}$$

From definition of FDR and Lemma 1 we have:

$$FDR_n = \sum_{r=1}^{p} \frac{1}{r} \sum_{j=1}^{p_0} P\left(|U_j(\hat{b}) + \Sigma(0)\hat{b}_j| \geq \lambda_r\ and\ R = r\}\right) =$$

$$= \sum_{r=1}^{p} \sum_{j=1}^{p_0} \sum_{l=1}^{p} \frac{1}{r} P\left(|U_j(\hat{b}) + \Sigma(0)\hat{b}_j| \geq \lambda_r\ and\ R = r\ and\ \tilde{R} = l\}\right) \leqslant$$

$$\leqslant \sum_{r=1}^{p} \sum_{j=1}^{p_0} \sum_{\substack{l=1 \\ l \neq r-1}}^{p} \frac{1}{r} P\left(|U_j(\hat{b}) + \Sigma(0)\hat{b}_j| \geq \lambda_r\ and\ R = r\ and\ \tilde{R} = l\}\right) +$$

$$+ \sum_{r=1}^{p} \sum_{j=1}^{p_0} \frac{1}{r} P\left(|U_j(\hat{b}) + \Sigma(0)\hat{b}_j| \geq \lambda_r\ and\ \tilde{R} = r-1\}\right)$$

From the fact that $\tilde{R} = r-1$ characterize properties of distribution of vector $\tilde{U}$ and a fact that asymptotically vector $\tilde{U}$ is independent from statistic Z_j (lemma 4.1 or 4.2) we obtain:

$$P\left(|U_j(\hat{b}) + \Sigma(0)\hat{b}_j| \geq \lambda_r\ and\ \tilde{R} = r-1\}\right) \to P\left(|Z_j| \geq \sqrt{\frac{n}{\Sigma(0)}}\lambda_r\ and\ \tilde{R} = r-1\}\right)$$

$$= P\left(|Z_j| \geq \sqrt{\frac{n}{\Sigma(0)}}\lambda_r\right) P\left(\tilde{R} = r-1\}\right) = \frac{qr}{p} P\left(\tilde{R} = r-1\}\right)$$

Therefore we have:

$$\sum_{r=1}^{p}\sum_{j=1}^{p_0}\frac{1}{r}P\left(|U_j(\hat{b})+\Sigma(0)\hat{b}_j| \geq \lambda_r \text{ and } \tilde{R} = r-1\}\right) \to \sum_{r=1}^{p}\sum_{j=1}^{p_0}\frac{1}{r}\frac{qr}{p}P\left(\tilde{R} = r-1\}\right) = \frac{qp_0}{p}$$

To prove the thesis it remains to show that:

$$P\left(|U_j(\hat{b})+\Sigma(0)\hat{b}_j| \geq \lambda_r \text{ and } R = r \text{ and } \tilde{R} = l\}\right) \to 0 \ \ for \ \ l \neq r-1$$

Without lost of generality we can assume that $j = 1$.
Let's assume that $\tilde{R} = k-1$ and $k < r$, then based on Lemma 2 and 3 we have:

$$\sqrt{\frac{n}{\Sigma(0)}}\sum_{i=k}^{r-1}\tilde{\lambda}_i \geqslant \sqrt{\frac{n}{\Sigma(0)}}\sum_{i=k}^{r-1}|\tilde{U}_{(i)}(\hat{b})+\Sigma(0)\hat{\tilde{b}}_{(i)}| \to \sum_{i=k}^{r-1}|\tilde{Z}_{(i)}|$$

Using the relation $\tilde{\lambda}_i = \lambda_{i+1}$ and a fact that $|Z_{(i)}| \leqslant |\tilde{Z}_{(i-1)}|$ we obtain:

$$\sqrt{\frac{n}{\Sigma(0)}}\sum_{i=k+1}^{r}\lambda_i = \sqrt{\frac{n}{\Sigma(0)}}\sum_{i=k}^{r-1}\lambda_{i+1} = \sqrt{\frac{n}{\Sigma(0)}}\sum_{i=k}^{r-1}\tilde{\lambda}_i \geqslant \sum_{i=k}^{r-1}|\tilde{Z}_{(i)}| = \sum_{i=k+1}^{r}|\tilde{Z}_{(i-1)}| \geqslant \sum_{i=k+1}^{r}|Z_{(i)}|$$

Simultaneously, because $R = r$ we have:

$$\sqrt{\frac{n}{\Sigma(0)}}\sum_{i=k+1}^{r}\lambda_i < \sqrt{\frac{n}{\Sigma(0)}}\sum_{i=k+1}^{r}|U_{(i)}(\hat{b})+\Sigma(0)\hat{b}_{(i)}| \to \sum_{i=k+1}^{r}|Z_{(i)}|$$

in consequence the above probability converge to 0.
Let's assume now that $\tilde{R} = k-1$ and $k > r$. On the basis of lemma 2 and 3 we have:

$$\sqrt{\frac{n}{\Sigma(0)}}\sum_{i=r}^{k-1}\tilde{\lambda}_i < \sqrt{\frac{n}{\Sigma(0)}}\sum_{i=r}^{k-1}|\tilde{U}_{(i)}(\hat{b})+\Sigma(0)\hat{\tilde{b}}_{(i)}| \to \sum_{i=r}^{k-1}|\tilde{Z}_{(i)}|$$

Again using the relation between $\tilde{\lambda}$ and λ and a fact that $|Z_{(i)}| = |\tilde{Z}_{(i-1)}|$ for $i \geqslant r$ which is consequence of the assumption that asymptotically $|Z_j| \geqslant \sqrt{\frac{n}{\Sigma(0)}}\lambda_r$ we obtain:

$$\sqrt{\frac{n}{\Sigma(0)}}\sum_{i=r+1}^{k}\lambda_i = \sqrt{\frac{n}{\Sigma(0)}}\sum_{i=r+1}^{k}\tilde{\lambda}_{i-1} = \sqrt{\frac{n}{\Sigma(0)}}\sum_{i=r}^{k-1}\tilde{\lambda}_i < \sum_{i=r}^{k-1}|\tilde{Z}_{(i)}| = \sum_{i=r+1}^{k}|\tilde{Z}_{(i-1)}| = \sum_{i=r+1}^{k}|Z_{(i)}|$$

On the other hand because $R = r$ we have:

$$\sqrt{\frac{n}{\Sigma(0)}}\sum_{i=r+1}^{k}\lambda_i \geqslant \sqrt{\frac{n}{\Sigma(0)}}\sum_{i=r+1}^{k}|U_{(i)}(\hat{b})+\hat{b}_{(i)}| \to \sum_{i=r+1}^{k}|Z_{(i)}|$$

which again show that probability of above event converge to 0 and eventually we obtain a the thesis of theorem.

4.2 proof of Lemma 1

$(\Rightarrow)$
From the optimality of $\hat{b}$ we have following inequality for any vector b:

$$0 \geqslant f(\hat{b}) - f(b) = l(b) - l(\hat{b}) + \sum_{j=1}^{p} \lambda_j |\hat{b}_{(j)}| - \sum_{j=1}^{p} \lambda_j |b_{(j)}| \tag{1}$$

In first step we will prove that for vector b of form $b = \hat{b} + h(0,...,0,\underbrace{(-1)sgn(\hat{b}_i)}_{i-th\ pos.},0,...,0)^T = \hat{b} + hL_i$ and small enough, positive h we have:

$$\sum_{j=1}^{p} \lambda_j |\hat{b}_{(j)}| - \sum_{j=1}^{p} \lambda_j |b_{(j)}| \geqslant h\lambda_r$$

Let's consider first simplest case where $|\hat{b}_i| \neq |\hat{b}_j|$ for any $i \neq j$. In this situation if we take

$$h < \min_{j,\ i \neq j} \left(\big| |\hat{b}_i| - |\hat{b}_j| \big| \right)$$

then the ordering in vector b and $\hat{b}$ will be the same. Furthermore two vectors differ only in j-th position and from the form of vector b we know that the absolute value of j-th element in vector b is smaller then in vector $\hat{b}$. In consequence we obtain for certain index k that:

$$\sum_{j=1}^{p} \lambda_j |\hat{b}_{(j)}| - \sum_{j=1}^{p} \lambda_j |b_{(j)}| = h\lambda_k \geqslant h\lambda_r$$

The last inequality is a consequence of assumptions $R = r$ and $\hat{b}_j \neq 0$ and a fact that λ_r is the smallest element of vector λ associated with non-zero elements of vector $\hat{b}$. Let's consider now case where two or more elements of vector $\hat{b}$ have the same absolute value as $|\hat{b}_i|$. The reasoning is analogous however in this situation value $|\hat{b}_i|$ will be associated with a set of elements of vector λ. Of course the power of associated set will be the same as the number of elements of vector $\hat{b}$ with absolute value the same as $|\hat{b}_i|$. Moreover the associated set of lambdas will contain consecutive elements of whole sequence. It is also straightforward that the indexing in a group of elements of vector $\hat{b}$ with the same module as $|\hat{b}_i|$ is ambiguous. For such situation if we take

$$h < \min_{j,\ |\hat{b}_i| \neq |\hat{b}_j|} \left(\big| |\hat{b}_i| - |\hat{b}_j| \big| \right)$$

we obtain for certain index k that:

$$\sum_{j=1}^{p} \lambda_j |\hat{b}_{(j)}| - \sum_{j=1}^{p} \lambda_j |b_{(j)}| = h\lambda_k \geqslant h\lambda_r$$

however this time λ_k is the smallest element of vector λ associated with $|\hat{b}_{(i)}|$. The justification for the last inequality is analogous to the previously discussed situation. It is worth mentioning that proven property do not depend in any way from function $l(b)$.
Using proven relation in first inequality we obtain:

$$\lambda_r \leqslant \frac{l(b) - l(\hat{b})}{-h}$$

Via transformation of expression of right side of inequality we obtain:

$$\lambda_r \leqslant \frac{l(b) - l(\hat{b})}{-h} = sgn(\hat{b}_i)\frac{l(\hat{b} + hL_i) - l(\hat{b})}{hL_i} \rightarrow sgn(\hat{b}_i)U_i(\hat{b})$$

for $h \rightarrow 0$. By adding to both sides factor $a|\hat{b}_i|$ we obtain:

$$\lambda_r < \lambda_r + a|\hat{b}_i| \leqslant sgn(\hat{b}_i)(U_i(\hat{b}) + a\hat{b}_i) = |U_i(\hat{b}) + a\hat{b}_i|$$

This ends proof in one direction.
$(\Leftarrow)$
The proof of second implication will be shown via contradiction. Let's assume that $\hat{b}_i = 0$ and $\lambda_r < |U_i + a\hat{b}_i|$ and consider vector $b = \hat{b} + h(0, ..., 0, \underbrace{1}_{i-th\ pos.}, 0, ..., 0)^T = \hat{b} + hL_i$. Let's recall that λ_{r+1} is the largest element of vector λ associated with elements of vector $\hat{b}$ with value 0. It is easy to notice that for small enough, positive h ($h < \min_i\{|\hat{b}_i|,\ \hat{b}_i \neq 0\}$) we obtain:

$$\sum_{j=1}^{p} \lambda_j |\hat{b}_{(j)}| - \sum_{j=1}^{p} \lambda_j |b_{(j)}| = -h\lambda_{r+1}$$

and in consequence following inequality holds:

$$0 \geqslant f(\hat{b}) - f(b) = l(b) - l(\hat{b}) - h\lambda_{r+1}$$

Transforming above inequality we obtain:

$$\lambda_{r+1} \geqslant \frac{l(\hat{b} + hL_i) - l(\hat{b})}{h} \rightarrow U_i(\hat{b})$$

Analogical reasoning for vector $b = \hat{b} + h(0, ..., 0, \underbrace{-1}_{i-th\ pos.}, 0, ..., 0)^T = \hat{b} + hL_i$ provides us:

$$\lambda_{r+1} \geqslant -U_i(\hat{b})$$

In consequence we obtain that $|U_i(\hat{b})| \leqslant \lambda_{r+1}$. On the other hand we assumed that $\lambda_r < |U_i(\hat{b}) + a\hat{b}_i| = |U_i(\hat{b})|$ ($\hat{b}_i = 0$) which leads to contradiction and ends the proof.
Corollary 1
From the proof of Lemma 1 it is easy to notice that when $a = 0$:
(i) $\left[\hat{b}_i \neq 0\right] \Rightarrow \left[\lambda_r \leqslant |U_i(\hat{b})|\right]$
(ii) if additionally $\lambda_r > \lambda_{r+1}$ then $\left[\hat{b}_i \neq 0\right] \Leftarrow \left[\lambda_r \leqslant |U_i(\hat{b})|\right]$
Corollary 2
From the proof of Lemma 1 it is also straightforward that when $\hat{b}_i \neq 0$ then $U_i(\hat{b})$ and $\hat{b}_i$ have the same sign.

4.3 proof of Statement

The proof is analogical to proof of Lemma 1.
Let's consider a vector $b = \hat{b} + h(0,...,0,\underbrace{(-1)sgn(\hat{b}_j)}_{j-th\ \ pos.},0,...,0,\underbrace{sgn(\hat{b}_k)}_{k-th\ \ pos.},0,...,0)^T = \hat{b} + hL$. For small enough positive h:

$$0 \geqslant f(\hat{b}) - f(b) = l(b) - l(\hat{b}) - h\lambda_l + h\lambda_m$$

where λ_l and λ_m are the smallest and the biggest element of vector λ associated with values $\hat{b}_j$ and $\hat{b}_k$, respectively. From assumption $|\hat{b}_j| > |\hat{b}_k|$ we know that $0 < \lambda_l - \lambda_k$. Therefore, after transformation we obtain the thesis:

$$0 < \lambda_l - \lambda_m \leqslant (-1)\frac{l(b) - l(\hat{b})}{h} \rightarrow (-1)\nabla_L l(b) = sgn(\hat{b}_j)U_j(\hat{b}) - sgn(\hat{b}_k)U_k(\hat{b})$$

The last equitation is a consequence of differentiability of function $l(b)$. Using relation $sgn(\hat{b}_i)U_i(\hat{b}) = |U_i(\hat{b})|$ (see corollary 2) we obtain:

$$0 < |U_j(\hat{b})| - |U_k(\hat{b})|$$

This ends the proof.

4.4 proof of Lemma 2

Let's assume $R = r$ and determine $j \leqslant r$.
Furthermore, let I be a set of indexes for which the rank of absolute value of the element of vector $\hat{b}$ is between j and r. Similarly to the proof of Lemma 1 we consider vector b defined in following way:

$$b = \begin{cases} \hat{b}_i - h\, sgn(\hat{b}_i) & i \in I \\ \hat{b}_i \quad otherwise & \end{cases}$$

Again, for h positive and small enough we obtain:

$$0 \geqslant f(\hat{b}) - f(b) = l(b) - l(\hat{b}) + h\sum_{i=j}^{r}\lambda_i$$

This is a consequence of construction of vector b. It is obtained via pulling $r-(j-1)$ smallest nonzero elements in vector $\hat{b}$ towards zero by the same factor h.
Transforming above inequality we obtain:

$$\sum_{i=j}^{r}\lambda_i \leqslant \frac{l(b)-l(\hat{b})}{-h} \longrightarrow \sum_{i=j}^{r} sgn(\hat{b}_{(i)})U_{(i)}(\hat{b}) = \sum_{i=j}^{r} sgn(\hat{b}_{(i)})U_{(i)}(\hat{b})$$

for $h \to 0$. By adding to both sides of the inequality $\sum_{i=j}^{r} a_i|\hat{b}_{(i)}|$ and using corollary 2 we obtain:

$$\sum_{i=j}^{r}\lambda_i < \sum_{i=j}^{r}(\lambda_i + a_i|\hat{b}_{(i)}|) \leqslant \sum_{i=j}^{r} sgn(\hat{b}_{(i)})\left(U_{(i)}(\hat{b}) + a_i\hat{b}_{(i)}\right) = \sum_{i=j}^{r}|U_{(i)}(\hat{b}) + a_i\hat{b}_{(i)}|$$

which ends the proof that $R = r$ ensure (i).
Now let's determine $j \geqslant r+1$ and introduce set J of indexes for which the rank of absolute value of the element of vector $\hat{b}$ is between $r+1$ and j. Let's consider following set of vectors b:

$$b = \begin{cases} hE_i & i \in J \\ \hat{b}_i & otherwise \end{cases}$$

where $E_i = \pm 1$ and $h > 0$.
For small enough h and for any vector b we obtain:

$$0 \geqslant f(\hat{b}) - f(b) = l(b) - l(\hat{b}) - h\sum_{i=r+1}^{j}\lambda_i$$

Again it is associated with a form of vector b. However this time $j-r$ elements with value 0 of vector $\hat{b}$ are pushed away from 0 by factor h.
By transformation of above inequality we obtain:

$$\sum_{i=r+1}^{j}\lambda_i \geqslant \frac{l(b)-l(\hat{b})}{h} \longrightarrow_{h\to 0} \sum_{i=r+1}^{j} E_{(i)}U_{(i)}(\hat{b})$$

This inequality is valid for any sequence of $\{E_i\}$. Therefore:

$$\sum_{i=r+1}^{j}\lambda_i \geqslant \sum_{i=r+1}^{j}|U_{(i)}(\hat{b})| = \sum_{i=r+1}^{j}|U_{(i)}(\hat{b}) + a_i\hat{b}_{(i)}|$$

The last equality is a consequence of a fact that $\hat{b}_{(i)} = 0$ for $i > r$.
This ends the proof of second state.
To prove the equivalence in other direction we will show that there are at least and at most r nonzero elements in $\hat{b}$.
Suppose that conditions (i) and (ii) are fulfilled and that there are $R = j < r$ nonzero element in $\hat{b}$. From the first part of the proof we know that when $R = j$ for $k > j$ $\hat{b}_{(k)} = 0$ and:

$$\sum_{i=j+1}^{r} \lambda_i \geqslant \sum_{i=j+1}^{r} |U_{(i)}(\hat{b})|$$

On the other hand from (i) we have:

$$\sum_{i=j+1}^{r} \lambda_i < \sum_{i=j+1}^{r} |U_{(i)}(\hat{b})|$$

which leads to a contradiction.
Now let's assume that $R = j > r$. Again this implies that:

$$\sum_{i=r}^{j} \lambda_i < \sum_{i=r}^{j} |U_{(i)}(\hat{b}) + a_i \hat{b}_{(i)}|$$

On the other hand from (ii) we obtain:

$$\sum_{i=r}^{j} \lambda_i \geqslant \sum_{i=r}^{j} |U_{(i)}(\hat{b}) + a_i \hat{b}_{(i)}|$$

which ends the proof of Lemma 2.

4.5 proof of Lemma 3

Let's recall that optimization problem has following form:

$$\hat{b} = arg \min_{b} \underbrace{\left(-l(b) + \sum_{j=1}^{p} \lambda_j |b_{(j)}| \right)}_{g(b)}$$

Let's consider following expression:

$$h(b) = g(b) - g(b^0) = l(b^0) - l(b) + \sum_{j=1}^{p} \lambda_j |b_{(j)}| - \sum_{j=1}^{p} \lambda_j |b^0_{(j)}|$$

Naturally $h(0) = 0$. Consequently, if $h(b) > 0$ for b from a sphere with center point b^0 then $\hat{b}$ must lie inside the sphere.
If we expand $l(b)$ near $b = b^0$ we obtain:

$$g(b) - g(b^0) = \frac{1}{n}\sum_{j=1}^{p}\big((y-\mu(b^0))X\big)_j(b_j - b_j^0) + \frac{1}{n}(b-b^0)'X'\Sigma(\tilde{b})X(b-b^0) + \sum_{j=1}^{p}\lambda_j|b_{(j)}| - \sum_{j=1}^{p}\lambda_j|b_{(j)}^0| =$$

$$\frac{1}{n}\sum_{j=1}^{p}\big((y-\mu(b^0))X\big)_j(b_j - b_j^0) + \frac{1}{n}(b-b^0)'(F_n(\tilde{b}) - dF_n(b^0))(b-b^0) +$$

$$(b-b^0)'\frac{1}{n}dF_n(b^0)(b-b^0) + \sum_{j=1}^{p}\lambda_j|b_{(j)}| - \sum_{j=1}^{p}\lambda_j|b_{(j)}^0|$$

From last two assumptions we know that:

$$\frac{1}{n}(b-b^0)'(F_n(\tilde{b}) - dF_n(b^0))(b-b^0) > 0$$

$$(b-b^0)'\frac{d}{n}F_n(b^0)(b-b^0) \to d(b-b^0)'G(b^0)(b-b^0) \geqslant C_1\|b-b^0\|^2$$

where $C_1 = const$.
It is easy to show that:

$$E[\frac{1}{n}\sum_{j=1}^{p}\big((y-\mu(b^0))X\big)_j(b_j - b_j^0)] = 0$$

This is obtained by conditioning the expected value on design matrix X. On the other hand:

$$Var[\frac{1}{n}\sum_{j=1}^{p}\big((y-\mu(b^0))X\big)_j(b_j - b_j^0)] = \frac{1}{n^2}Var[\sum_{i=1}^{n}\sum_{j=1}^{p}(y_i - \mu(X_i b^0))X_{ij}(b_j - b_j^0)] =$$

$$\frac{1}{n^2}\sum_{i=1}^{n}E[(y_i - \mu(X_i b^0))^2(X(b-b^0))_i^2] = \frac{1}{n^2}\sum_{i=1}^{n}E[(X(b-b^0))_i^2 E[(y_i - \mu(X_i b^0))^2|X_i]] =$$

$$\frac{1}{n^2}\sum_{i=1}^{n}E[(X(b-b^0))_i^2\Sigma(X_i b^0)] = \frac{1}{n}E[\frac{1}{n}\sum_{i=1}^{n}((b-b^0)'X')_i\Sigma(X_i b^0)(X(b-b^0))_i] =$$

$$\frac{1}{n}E\underbrace{[(b-b^0)'\frac{1}{n}F_n(b^0)(b-b^0)]}_{\to(b-b^0)'G(b^0)(b-b^0)} \to 0$$

We know that $J_\lambda(b) = \sum_{j=1}^{p} \lambda_j |b_{(j)}|$ is a norm of vector b, therefore by using triangular inequality and norms equivalence in $\mathbb{R}^p$, we obtain:

$$|J_\lambda(b) - J_\lambda(b^0)| \leqslant J_\lambda(b - b^0) \leqslant \frac{C_2}{\sqrt{n}} \|b - b^0\|$$

where $C_2 = const$.
Summarizing we obtain that for large enough n, with probability tending to 1:

$$g(b) - g(b^0) \geqslant C_1 \|b - b^0\|^2 - \frac{C_2}{\sqrt{n}} \|b - b^0\|$$

From above we can always find large enough $M > 0$ such that for $\|b - b^0\| = \frac{M}{\sqrt{n}}$ we obtain:

$$g(b) - g(b^0) \geqslant C_1 \|b - b^0\|^2 - \frac{C_2}{\sqrt{n}} \|b - b^0\| = C_1 \frac{M^2}{n} - C_2 \frac{M}{n} > 0$$

This leads to conclusion that $\|\sqrt{n}(\hat{b} - b^0)\|$ is bounded in probability.

4.6 proof of Lemma 4.1

proof:
Let's expand $U(\hat{b})$ near real parameter vector b^0:

$$\sqrt{n}\left(U(\hat{b}) + \Sigma(0)\hat{b}\right) = \sqrt{n}\left(U(b^0) - \frac{1}{n}F_n(b^*)(\hat{b} - b^0) + \Sigma(0)\hat{b}\right) =$$

$$\underbrace{\sqrt{n}U(b^0)}_{\to N(0,\Sigma(0)\mathbb{I})\ (MCLT)} + \sqrt{n}b^0\Sigma(0) + \underbrace{(\Sigma(0) - \frac{1}{n}F_n(b^*))}_{\to 0\ in\ prob.} \underbrace{\sqrt{n}(\hat{b} - b^0)}_{bound.\ in\ prob.}$$

4.7 proof of Lemma 4.2

The proof is analogous to that for lemma 4.1. The most important observation is that for linear regression $\frac{1}{n}F_n(b) \to \Sigma(0)$ for any b.