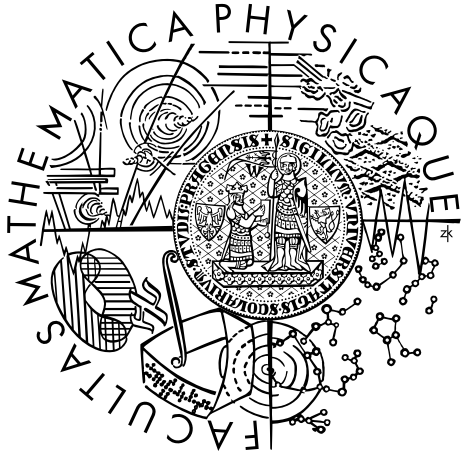
Charles University in Prague
Faculty of Mathematics and Physics

MASTER THESIS



Bc. Vojtěch Bouř

Post-selection Inference: Lasso & Group Lasso

Department of Probability and Mathematical Statistics

Supervisor of the master thesis: RNDr. Matúš Maciak, Ph.D.
Study programme: Mathematics
Study branch: Financial and Insurance Mathematics

Prague 2017

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In Prague date 11 May 2017

Title: Post-selection Inference: Lasso & Group Lasso

Author: Bc. Vojtěch Bouř

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Matúš Maciak, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The lasso is a popular tool that can be used for variable selection and estimation, however, classical statistical inference cannot be applied for its estimates. In this thesis the classical and the group lasso is described together with efficient algorithms for the solution. The key part is dedicated to the post-selection inference for the lasso estimates where we explain why the classical inference is not suitable. Three post-selection tests for the lasso are described and one test is proposed also for the group lasso. The tests are compared in simulations where finite sample properties are examined. The tests are further applied on a practical example.

Keywords: Post-selection Inference, Lasso, Group Lasso, L1 regularization, Lasso significance.

In this place I would like to express sincere thanks to RNDr. Matúš Maciak, Ph.D. for the professional supervision, valuable advices and inspiration.

Contents

Introduction **2**

1 Classical lasso method **4**
1.1 Lasso estimate 4
1.2 Computational aspects 7
1.2.1 LARS method 7
1.2.2 Coordinate descent 10
1.2.3 Additional options for solving the lasso 11

2 Group lasso modification **12**
2.1 Group lasso definition 12
2.2 Usage and computational aspects 13

3 Statistical inference for lasso **14**
3.1 Classical statistical inference 14
3.2 Consistency of the lasso estimate 15
3.3 Post-selection inference 16
3.3.1 Covariance test 17
3.3.2 Tests based on the polyhedral lemma 20
3.3.3 Other approaches 25
3.4 Post-selection test for the group lasso 25

4 Finite sample properties **27**
4.1 Simulations 27
4.1.1 All zero coefficients 27
4.1.2 One and two nonzero coefficients 28
4.1.3 Four nonzero coefficients, high-dimension 32
4.2 Simulations for the group lasso test 33
4.3 Real data example 34

Conclusion **40**

Bibliography **41**

List of Figures **43**

List of Tables **44**

Introduction

Linear regression models are still widely used because of their simplicity and often sufficient approximation. In these models we usually work with a dependent variable (response) and a set of independent variables (predictors or covariates). A normal linear regression model is a statistical model where the response is a linear function of unknown parameters, i.e., regression coefficients $\beta_1, \ldots, \beta_p$. Having some observations with the sample size n (usually $n > p$) we estimate and examine these model parameters.

In the statistical modeling there are a lot of various cases where the number of parameters to be estimated p is large relative to the sample size n. This case ($p > n$) is common, for example, in genomics where there are a lot of gene measurements but only a few samples (patients). Typically we deal with thousands of genes but only hundreds of patients. Then it is necessary (and also in other high-dimensional cases) to select from many candidate variables these we include in our model. Such model can be linear regression where there are more predictors than observations, and we need to decide, which variables are relevant. With high p we are overfitting our model and we also need to deal with large variance. For these situations we would like to somehow regularize or choose the most important variables.

Among techniques that are used for estimation in high-dimensional (that is the $p > n$ case) problems, the *lasso* belongs to the most popular ones. Lasso estimates are found similarly as the ordinary least squares estimates where we additionally put some conditions on the regularization of coefficients. An interesting property of this tool is that not only it shrinks some coefficients towards zero but also it puts others exactly equal to zero. This is a desirable feature because of simplicity and interpretability of the resulting model. Of course, we need to have some tools to verify whether the selected variables are truly the most relevant ones. Thus, from the statistical point of view we are interested in some statistical tests which could be used to test the significance of parameters — so called post-selection inference.

In the normal linear regression we are able (because of the linearity) to easily construct confidence intervals or p-values for coefficients and also perform various tests (significance of a variable, test on a submodel, etc.). The difficulty with the lasso is that the statistical inference cannot be performed in the same approach for the lasso estimates. We cannot explicitly write the lasso solution because it is an outcome of a more complex optimization problem, and the selection of the non-zero coefficients is random. In general, we cannot say what will be the selected model unless we run the whole lasso minimization. The classical inference tools can not be applied because they are too optimistic, as they only consider fixed hypothesis. However, for the lasso (and similar selection procedures) there were post-selection inference tools developed instead. They are specifically designed to handle the randomness in the selection. The theory about this statistical inference after a variable selection process, such as the lasso, is quite recent, and it is still growing. The idea of this thesis is to summarize and compare statistical tests that can be performed after the selection of relevant covariates by the lasso.

The structure of the thesis is following. Firstly, some basic methodological

background for the lasso estimation is presented together with some computational aspects. Other versions derived from the original lasso definition, such as the *group lasso*, are briefly mentioned too. The main part of the thesis is in Chapter 3 where we describe the post-selection tests and shortly speak about alternative approaches for the statistical inference. In the last chapter we present some simulation results where the post-selection tests are compared and finite sample properties are investigated. The proposed tests are also applied on a real data scenario.

1. Classical lasso method

In this chapter we give a little motivation for the lasso. We introduce the lasso method, and we describe some algorithms for solving the lasso problem. The *lasso* name stands for the *"least absolute shrinkage and selection operator"*. This method for estimation in linear models was introduced in (Tibshirani, 1996).

1.1 Lasso estimate

At the beginning, let us start with the usual linear regression set-up. Suppose we have the response $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ and the matrix of predictor variables $\mathbf{X} \in \mathbb{R}^{n \times p}$:

$$\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)\,,$$

where $\boldsymbol{x}_j = (x_{1,j}, \ldots, x_{n,j})^\top,\ j = 1, \ldots, p$, are vectors of predictors. We assume that $\boldsymbol{y}$ is a linear function of predictors with normally distributed errors:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n), \tag{1.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a vector of unknown parameters and $\boldsymbol{I}_n \in \mathbb{R}^{n \times n}$ is an identity matrix. For now we will assume $p < n$, and that the first column of $\mathbf{X}$ are all 1s. We denote by $\widehat{\boldsymbol{\beta}}^{OLS}$ the ordinary least squares estimate of the vector $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}^{OLS} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\| \boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2, \tag{1.2}$$

where $\|\boldsymbol{v}\|_2$ is the L_2 (Euclidean) norm of the vector $\boldsymbol{v}$.

The estimate $\widehat{\boldsymbol{\beta}}^{OLS}$ has the very known form $\widehat{\boldsymbol{\beta}}^{OLS} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \boldsymbol{y}$. To invert $\mathbf{X}^\top \mathbf{X}$ the full-rank $\mathbf{X}$ is needed, otherwise $\widehat{\boldsymbol{\beta}}^{OLS}$ is not unique. When the columns of $\mathbf{X}$ are almost linearly dependent then $\mathbf{X}^\top \mathbf{X}$ is close to being singular and difficult to invert. One idea for improving the inversion in the normal equations is to add some positive values on the diagonal of $\mathbf{X}^\top \mathbf{X}$. The estimate has then the form

$$\left(\mathbf{X}^\top \mathbf{X} + \lambda \boldsymbol{I}_p\right)^{-1} \mathbf{X}^\top \boldsymbol{y},$$

where $\lambda \geq 0$. The previous formula is a solution to *the ridge regression* problem (see, for example, Fu, 1998)

$$\widehat{\boldsymbol{\beta}}^{ridge} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\| \boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_2^2. \tag{1.3}$$

The ridge estimates $\widehat{\boldsymbol{\beta}}^{ridge}$ are shrunk toward zero by the penalty term $\lambda \|\boldsymbol{\beta}\|_2^2 = \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}$ in the sense that they are smaller than the corresponding least square estimates in their absolute value, but they are never exactly zero (Figure 1.2), while the amount of shrinkage depends on the value of the parameter $\lambda > 0$.

When we change the penalizing term $\lambda \|\boldsymbol{\beta}\|_2^2$ in (1.3) to $\lambda \|\boldsymbol{\beta}\|_1 = \lambda \sum_{j=1}^p |\beta_j|$ we get another shrinkage method, so called *lasso*. The lasso estimate $\widehat{\boldsymbol{\beta}}(\lambda)$ is defined by

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{2} \left\| \boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1, \tag{1.4}$$

where $\lambda \geq 0$ is sometimes called a tuning or a shrinkage parameter. The previous formula (in so called *langrangian* form) is equivalent to another definition:

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{2} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to } \|\boldsymbol{\beta}\|_1 \leq d, \tag{1.5}$$

where $d \geq 0$. If $\widehat{\boldsymbol{\beta}}(\lambda)$ is the solution of (1.4), it also minimizes (1.5) with $d = \sum_{j=1}^p |\widehat{\beta}_j(\lambda)|$ (Friedman et al., 2007). In the further notation we suppress λ in the lasso solution $\widehat{\boldsymbol{\beta}}$, and we will implicitly assume that the solution depends on the tuning parameter.

The lasso was originally defined also with the intercept parameter $\alpha \in \mathbb{R}$. For every d the lasso estimate of α is equal to the related OLS estimate. In the theoretical part we mostly assume that the columns of $\mathbf{X}$ are standardized to have a zero mean and $\frac{1}{n}\sum_{i=1}^n x_{ij}^2 = 1$, for all $j = 1, \ldots, p$. Then $\hat{\alpha}^{OLS} = \overline{y}$ (the mean of y). We assume, without loss of generality, that $\overline{y} = 0$, and we will not take α into account anymore. If the intercept would need to be included in the model, we can always center the response. Then $\overline{y} = 0$ and the effect of the intercept is taken into account.

From (1.4) and (1.5) we see that the lasso solution solves the L_1-penalized sum of squares. The name *"least absolute shrinkage and selection operator"* was given because the lasso shrinks some parameters of the vector $\boldsymbol{\beta}$ towards 0 and puts other equal to 0. Unlike the ridge regression the lasso performs, in addition to the parameter estimation, also the model selection by setting some parameters to zero. The level of shrinkage depends on the parameter d (or λ). For $d < d_0$, where $d_0 = \sum_{j=1}^p |\widehat{\beta}_j^{OLS}|$, the coefficients will be shrunk towards 0. For $d \geq d_0$ the problem equals the problem of least squares. For $\lambda = 0$ the solution of (1.4) also equals the usual linear least squares regression.

We can compare the shrinkage approach of all three methods (linear regression, the ridge regression and the lasso) in an orthonormal design case. When $\mathbf{X}^\top\mathbf{X} = \boldsymbol{I}_p$ (the identity matrix) the ridge solution of (1.3) equals $\widehat{\beta}_j^{OLS}/(1+\lambda)$ and the lasso solution can be also written in the explicit form $\text{sign}(\widehat{\beta}_j^{OLS})(|\widehat{\beta}_j^{OLS}| - \lambda)_+$, where $(x)_+ = \max(x, 0)$, for every $j = 1, \ldots, p$ (see Figure 1.1 for fixed $\lambda = 2$). The last transformation of the least square estimate is also called the soft-thresholding (see Hastie et al., 2008).

The idea why the lasso often puts coefficients equal to zero but the ridge regression does not, can be easily visualized for the situation when $p = 2$. The expression $\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ can be written as

$$\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{OLS}\right)^\top \mathbf{X}^\top\mathbf{X} \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{OLS}\right), \tag{1.6}$$

plus a constant. When $p = 2$ we have a condition $\lambda(|\beta_1| + |\beta_2|) \leq d$ for the lasso and $\lambda(\beta_1^2 + \beta_2^2) \leq d$ for the ridge regression (by rewriting the expression (1.3)). The condition for the lasso is a rotated square and a circle for the ridge regression (see Figure 1.2). The point where the elliptical contours of the function (1.6) hit the square is the lasso solution. When they touch the corner it corresponds to a zero coefficient. However, in the ridge regression case, there are no corners to hit, which means we hardly obtain zero coefficients.

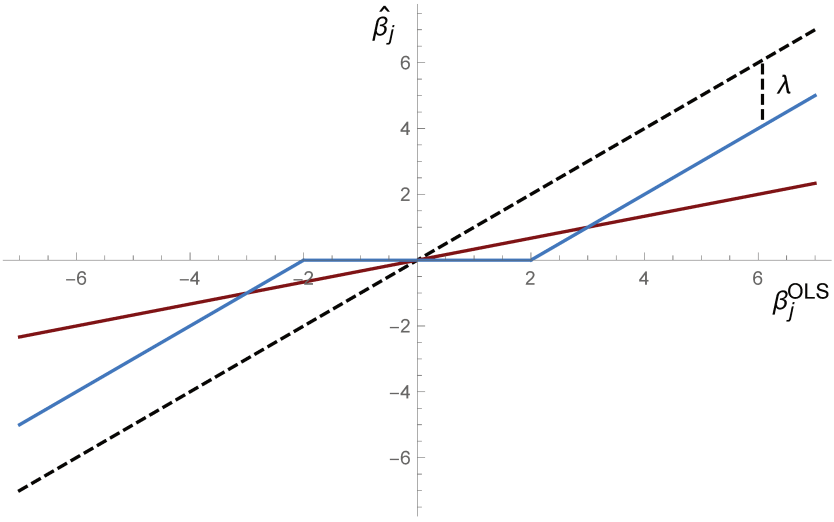


Figure 1.1: *The proportional shrinkage of the ridge regression (the brown line) and the soft-thresholding of the lasso (the blue line) in the case of orthonormal columns of* $\boldsymbol{X}$*. The black dashed line shows the unrestricted least square estimate.*

For a general predictor matrix $\mathbf{X}$ the lasso estimate $\widehat{\boldsymbol{\beta}}$ has no explicit form. However, the function we minimize in (1.4) is still convex. To show this we can write the objective function as

$$f(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + h(\boldsymbol{\beta}),$$

where $g(\boldsymbol{\beta}) = \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$, and $h(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_1$. First, we calculate the Hessian, the matrix of the second derivatives

$$\Delta^2 g(\boldsymbol{\beta}) = \left(\frac{\partial^2 g}{\partial\beta_i \partial\beta_j}\right)_{i=1,j=1}^{p,p} = \mathbf{X}^\top \mathbf{X}.$$

For any vector $l \in \mathbb{R}^p$: $l^\top \mathbf{X}^\top \mathbf{X} l \geq 0$ (the matrix $l^\top \mathbf{X}^\top \mathbf{X} l$ is positive semidefinite). This implies the convexity of $g(\boldsymbol{\beta})$.

For any $\alpha \in (0,1)$ and any $\boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2$ we have from triangle inequality:

$$\begin{aligned} h(\alpha\boldsymbol{\beta}_1 + (1-\alpha)\boldsymbol{\beta}_2) &= \lambda \|\alpha\boldsymbol{\beta}_1 + (1-\alpha)\boldsymbol{\beta}_2\|_1 \\ &\leq \lambda \|\alpha\boldsymbol{\beta}_1|_1 + \lambda \|(1-\alpha)\boldsymbol{\beta}_2\|_1 \\ &= \lambda\alpha \|\boldsymbol{\beta}_1\|_1 + \lambda(1-\alpha) \|\boldsymbol{\beta}_2\|_1 \\ &= \alpha h(\boldsymbol{\beta}_1) + (1-\alpha) h(\boldsymbol{\beta}_2), \end{aligned}$$

which means the function h is convex. A sum of two convex functions is again a convex function, which implies the convexity of $f(\boldsymbol{\beta})$.

As the lasso sets many parameters equal to zero it can be used for the model selection also in the high-dimensional case (when $p >> n$).

The equation (1.4) is a quadratic programming problem with linear inequality constraints. There are many algorithms for solving this issue, thanks to the convexity. Two of them will be described in the next section.

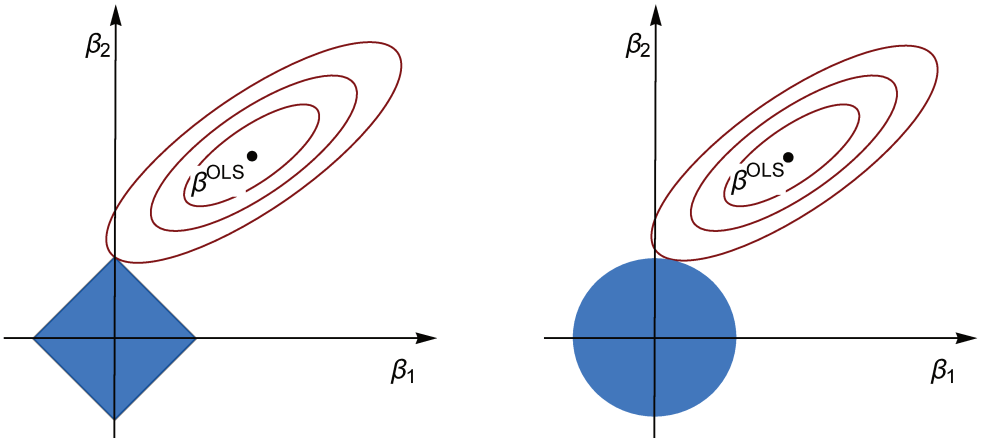


Figure 1.2: *Illustration of the geometry of Lasso (on the left) vs. Ridge regression (on the right)*

1.2 Computational aspects

As the lasso problem is convex there are several possibilities to solve it. In this part we describe two efficient algorithms for solving the lasso problem (1.4): *the LARS* method and *the coordinate descent* algorithm.

1.2.1 LARS method

The LARS (*Least Angle Regression*) algorithm was originally introduced in (Efron et al., 2004). The outcome of the algorithm is a piecewise linear function for each estimated non-zero coefficient. First, we describe the original version of the LARS algorithm, which does not correspond to the lasso method exactly, and later, we go over to the modification for solving the lasso.

The idea of the algorithm is following: We start with all coefficients equal to zero. Then we find a predictor most correlated with the response. We take the largest step possible in the direction of this predictor (meaning we are moving the coefficient of this predictor toward the least square value) until some other predictor is same correlated with the current residual $\boldsymbol{y} - \hat{\boldsymbol{\mu}}$, where $\hat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\beta}}$ is a vector of the estimated regression coefficients. We then proceed in the direction "in the middle" of these two predictors (changing coefficients of these two predictors) until a third predictor is same correlated. Next we continue in the "least angle direction" between these three predictors, and so on. Figure 1.3 shows the graphical meaning of LARS for two covariates. In the given picture, the covariances between the current residual and predictors are given by means of angles between the corresponding vectors.

At every step of the algorithm we add one new element to the nonzero set of coefficients (i.e., the active set). The result is a whole coefficient path (the lasso path) depending on λ. It begins at the point where all coefficients are zero (λ is large enough). Then as coefficients are growing, λ is decreasing, and the path ends at the point where coefficients are equal to the OLS estimates ($\lambda = 0$). The

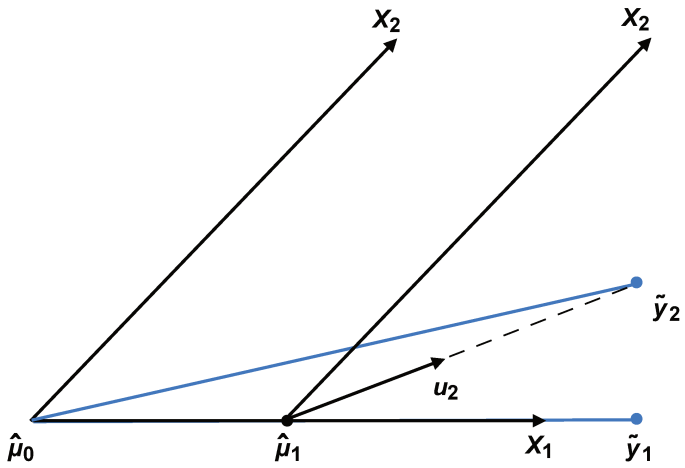


Figure 1.3: *The LARS algorithm for two covariates:* $\tilde{\boldsymbol{y}}_1$ *and* $\tilde{\boldsymbol{y}}_2$ *indicate the projection of* $\boldsymbol{y}$ *into the linear space of* $\boldsymbol{x}_1$ *and the linear space spanned by* x_1 *and* x_2*:* $\mathcal{L}(\boldsymbol{x}_1)$ *and* $\mathcal{L}(\boldsymbol{x}_1, \boldsymbol{x}_2)$*. We begin at* $\hat{\boldsymbol{\mu}}_0$*. The residual vector* $\tilde{\boldsymbol{y}}_2 - \hat{\boldsymbol{\mu}}_0$ *is more correlated with* $\boldsymbol{x}_1$ *than* $\boldsymbol{x}_2$ *(because of the smaller angle), so the next step will be in the direction of* $\boldsymbol{x}_1$*. The next LARS estimate is* $\hat{\boldsymbol{\mu}}_1$ *such that* $\tilde{\boldsymbol{y}}_2 - \hat{\boldsymbol{\mu}}_1$ *bisects the angle between* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$*. The next estimate would be in the direction of* $\boldsymbol{u}_2$*. For* $p = 2$ *we have* $\hat{\boldsymbol{\mu}}_2 = \tilde{\boldsymbol{y}}_2$*, but for* $p > 0$ *the estimate* $\hat{\boldsymbol{\mu}}_2$ *does not reach* $\tilde{\boldsymbol{y}}_2$*.*

lasso path is a piecewise linear function of λ for each predictor. This implies that only the points where the slope of the curve is changing are needed to calculate the whole path. Thus, it can be calculated in p steps.

We will now describe the algorithm more formally. We assume that the covariate columns $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are linearly independent and $\mathcal{A}$ is a subset of indexes $\{1, \ldots, p\}$. We first define quantities

$$\begin{aligned} \mathbf{X}_{\mathcal{A}} &= (\cdots s_j \boldsymbol{x}_j \cdots)_{j \in \mathcal{A}}\,, \\ N_{\mathcal{A}} &= \mathbf{X}_{\mathcal{A}}^{\top} \mathbf{X}_{\mathcal{A}}, \\ A_{\mathcal{A}} &= \left(\mathbf{1}_{\mathcal{A}}^{\top} \mathbf{N}_{\mathcal{A}}^{-\mathbf{1}} \mathbf{1}_{\mathcal{A}}\right)^{-\frac{1}{2}}, \\ w_{\mathcal{A}} &= A_{\mathcal{A}} N_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}, \end{aligned} \tag{1.7}$$

where $\mathbf{X}_{\mathcal{A}}$ is a matrix made by columns of $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ which are in $\mathcal{A}$ multiplied by the signs s_j, and $\mathbf{1}_{\mathcal{A}}$ is a vector of 1's with the length being equal to $|\mathcal{A}|$. Let

$$\boldsymbol{u}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}} w_{\mathcal{A}} \tag{1.8}$$

be so called *equiangular vector* for which

$$\mathbf{X}_{\mathcal{A}}^{\top} \boldsymbol{u}_{\mathcal{A}} = A_{\mathcal{A}} \mathbf{1}_{\mathcal{A}} \quad \text{and} \quad ||\mathbf{u}_{\mathcal{A}}|| = \mathbf{1}. \tag{1.9}$$

Further, suppose that $\hat{\boldsymbol{c}}$ is a vector of current correlations, $\hat{C}$ is the maximum

current absolute correlation, and s_j are signs of $\hat{c}_j$:

$$\begin{aligned} &\hat{\boldsymbol{c}} = \mathbf{X}^\top(\boldsymbol{y} - \hat{\boldsymbol{\mu}}), \\ &\hat{C} = \max_j\{|\hat{c}_j|\}, \\ &s_j = \operatorname{sign}\{\hat{c}_j\}, \quad \text{for} \quad j \in \mathcal{A}, \end{aligned} \tag{1.10}$$

where $\hat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ is the current LARS estimate. We define the active set $\mathcal{A}$ as a set of indexes corresponding to the covariates having maximum absolute correlation with the current residuum (which means their coefficients are nonzero at the current step):

$$\mathcal{A} = \{j \ : |\hat{c}_j| = \hat{C}\} = \{j \ : \widehat{\beta}_j \neq 0\}. \tag{1.11}$$

LARS algorithm

1. Standardize predictors to have a zero mean and $\|\boldsymbol{x}_j\|_2^2 = 1$ for all j. Start with $\hat{\boldsymbol{\mu}}_0 = 0$, $\hat{\boldsymbol{c}} = \mathbf{X}^\top\boldsymbol{y}$ and $\widehat{\boldsymbol{\beta}_0} = (0, \ldots, 0)^\top$.
2. Find the predictor $\boldsymbol{x}_j$ with the largest value of $|\hat{c}_j|$ (the most correlated with the current residuum) and define the active set $\mathcal{A} = \{j\}$.
3. Repeat the following, until all predictors from $\mathcal{A}^C$ are in the active set:
   - Compute $\hat{\boldsymbol{c}}$, $\hat{C}$, $\mathbf{X}_\mathcal{A}$, $A_\mathcal{A}$, $\boldsymbol{u}_\mathcal{A}$ by formulas (1.10), (1.7), (1.8) and also the inner product vector
   $$\boldsymbol{a} = \mathbf{X}^\top\boldsymbol{u}_\mathcal{A} = (a_1, \ldots, a_p)^\top. \tag{1.12}$$
   - Update $\hat{\boldsymbol{\mu}}$ by
   $$\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}} + \hat{\gamma}\boldsymbol{u}_\mathcal{A}, \tag{1.13}$$
   where
   $$\hat{\gamma} = \min_{j \in \mathcal{A}^C}{}^{+} \left\{ \frac{\hat{C} - \hat{c}_j}{A_\mathcal{A} - a_j}, \frac{\hat{C} + \hat{c}_j}{A_\mathcal{A} + a_j} \right\}. \tag{1.14}$$
   The symbol $\min^+$ in the previous formula denotes the minimum where we consider only positive members for each choice of j.
   - Set $\mathcal{A} = \mathcal{A} \cup \hat{j}$, where $\hat{j}$ is the minimizing index in (1.14).

That is how the algorithm works. However, for the lasso a small modification is needed. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^\top$ be a lasso solution and $\hat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$. Then the sign of any estimated coefficient β_j must be the same as the sign of the current correlation $\hat{c}_j = \boldsymbol{x}_j^\top(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ (see Hastie et al., 2008, Section 3.4.4):

$$\operatorname{sign}(\widehat{\beta}_j) = \operatorname{sign}(\hat{c}_j) = s_j. \tag{1.15}$$

The LARS as we described it does not guarantee the equation (1.15), but it can be adjusted to do so: *When a non-zero coefficient changes a sign (becomes zero), drop the variable from the active set and recalculate the current equiangular direction* (1.13).

To take this modification into account in the LARS algorithm described above, we need to add the following items to the third step:

- Define the p-vector $\hat{\boldsymbol{f}}$ to be equal to $s_j w_{\mathcal{A}_j}$ for $j \in \mathcal{A}$ and zero elsewhere. Here $w_{\mathcal{A}_j}$ denotes the element of the vector $w_{\mathcal{A}}$ corresponding to the index j.
- For $j \in \mathcal{A}$, update
$$\widehat{\beta}_j(\gamma) = \widehat{\beta}_j^{\,\text{prev}} + \gamma \hat{f}_j,$$
where $\widehat{\beta}_j^{\,\text{prev}}$ are the lasso estimates from the previous step.
- Let
$$\gamma_j = -\frac{\widehat{\beta}_j}{\hat{f}_j}, \text{ and } \quad \tilde{\gamma} = \min_{\gamma_j > 0}\{\gamma_j\}. \tag{1.16}$$
- If $\gamma = \tilde{\gamma}$, set $\mathcal{A} = \mathcal{A} - \{\tilde{j}\}$.

Details of the algorithm can be found in (Efron et al., 2004). As the variables can be dropped of from the active set and they can (possibly) return into the path later, the number of steps for this modification can be more than p.

1.2.2 Coordinate descent

The idea of this algorithm for solving the lasso problem (1.4) was proposed already in (Fu, 1998). The main principle of the algorithm is to optimize each parameter separately while we keep all the others fixed. We describe the algorithm similarly as in (Hastie et al., 2008).

Coordinate descent algorithm for the lasso

1. Standardize predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$. Define a grid of values $\lambda_1 > \lambda_2 > \ldots$. The first value λ_1 is chosen such that the lasso solution vector is a zero vector: $\widehat{\boldsymbol{\beta}}(\lambda_1) = \mathbf{0}$.
2. For each $\lambda \in \{\lambda_1, \lambda_2, \ldots\}$, repeat the following steps over $j = 1, 2, \ldots, p$ till convergence:
   - Rewrite the lasso problem (1.4) and isolate β_j:
   $$\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{l \neq j} x_{il}\tilde{\beta}_l(\lambda) - x_{ij}\beta_j\right)^2 + \lambda \sum_{l \neq j}\left|\tilde{\beta}_l(\lambda)\right| + \lambda\left|\beta_j\right|, \tag{1.17}$$
   where $\tilde{\beta}_l(\lambda)$ is the current estimate for β_l ($l \neq j$) for the given λ.
   - Compute the partial residuals $r_{ij} = y_i - \sum_{l \neq j} x_{il}\tilde{\beta}_l(\lambda)$, for all i.
   - Compute $\beta_j^* = \sum_{i=1}^{n} x_{ij} r_{ij}$ (the least squares coefficient of the partial residuals).
   - Update the current estimate $\tilde{\beta}_j$ using the soft-thresholding operator $S(.)$:
   $$\tilde{\beta}_j(\lambda) \leftarrow S\left(\beta_j^*,\ \lambda\right) = \text{sign}(\beta_j^*)(|\beta_j^*| - \lambda)_+. \tag{1.18}$$

The equation (1.17) can be viewed as the lasso problem with one variable and the response r_{ij}.

The coordinate descent algorithm for the lasso does not lead to the entire lasso path, but it can be used to compute the lasso solutions at a grid of values $\lambda_1 > \lambda_2 > \dots$. We always decrease λ a bit and cycle through the variables until convergence to the lasso estimate $\widehat{\boldsymbol{\beta}}(\lambda)$. Then decrease λ again and use the previous solution as the starting value for the new value of λ.

The described algorithm is relatively simple and can be faster than LARS when we have a lot of predictors.

1.2.3 Additional options for solving the lasso

Because of the convexity the lasso problem can be transformed into a quadratic programming problem. Two methods how to rewrite it into the quadratic programming problem, and then solve it by standard quadratic programming techniques, are described in (Tibshirani, 1996). The first is based on all possible combinations of signs of the vector $\boldsymbol{\beta}$ and the other on splitting the vector $\boldsymbol{\beta}$ into positive and negative part $\boldsymbol{\beta}^+$, $\boldsymbol{\beta}^-$. Generally, there are many tools for solving convex optimization problems.

2. Group lasso modification

From the original lasso various modifications were derived, for example, to obtain some nice features that lasso does not have (e.g., oracle properties), or to handle more complex covariates (e.g., factor variables). They usually differ in the penalizing term. For the estimation and model selection in a regression with grouped variables *the group lasso* was proposed in (Yuan and Lin, 2006). Another two lasso modifications are briefly described at the end of Section 3.2. We shortly introduce the group lasso in the following lines, and we suggest a modification of LARS which could be used for the calculation of the group lasso estimates.

2.1 Group lasso definition

In the linear regression problem with L groups (e.g., factor covariates) we define the model as

$$\boldsymbol{y} = \sum_{j=1}^{L} \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n), \tag{2.1}$$

where each matrix $\mathbf{X}_j$ of dimension $n \times p_j$ forms a group of variables, and $\boldsymbol{\beta}_j$ are the vectors of coefficients with the length p_j. The sizes p_j can be generally different for each $j = 1, \ldots, L$. The total number of parameters to be estimated is then $\sum_{j=1}^{L} p_j$. We further assume that the response $\boldsymbol{y}$ and every variable in each group are centered.

The common problem in the regression with the grouped variables is to identify the most important groups from the total number of L groups. To find out the most influential groups we can define a more general version of the lasso problem (1.4) called the group lasso. Similarly as in (Yuan and Lin, 2006), we say that vector $\widehat{\boldsymbol{\beta}}^g = \left(\widehat{\boldsymbol{\beta}}_1^\top, \ldots, \widehat{\boldsymbol{\beta}}_L^\top\right)^\top$ is the solution to the group lasso problem if

$$\widehat{\boldsymbol{\beta}}^g(\lambda) = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{p_j},\ j=1,\ldots,L}{\arg\min} \frac{1}{2} \left\| \boldsymbol{y} - \sum_{j=1}^{L} \mathbf{X}_j \boldsymbol{\beta}_j \right\|_2^2 + \lambda \sum_{j=1}^{L} \|\boldsymbol{\beta}_j\|_2, \tag{2.2}$$

where $\|\boldsymbol{\beta}_j\|_2 = \sqrt{\beta_{j1}^2 + \cdots + \beta_{jp_j}^2}$. The group lasso defined by the previous formula has the similar behavior as the classical lasso. For some $\lambda \geq 0$ it either puts a vector $\boldsymbol{\beta}_j$ to be exactly equal to zero (element-wise) or it makes all elements of some group to be nonzero. The non-relevant groups are then recognized by the related zero coefficients.

It is easy to see that the lasso problem (1.4) is a special case of (2.2) if $p_j = 1$, for all $j = 1, \ldots, L$. When $p_j = p$, that is each group contains the same number of variables, we can use a modification of the LARS algorithm for solving the problem (2.2). Further, we describe the idea of the algorithm for the simple case $p_j = p = 2$, for all j.

2.2 Usage and computational aspects

The group structure of the linear regression can be used when we include some categorical variables (e.g., factors with more levels) into the model. In the linear regression model these factors are usually coded by several dummy variables. For the model with a categorical variable the related matrix $\mathbf{X}_j$ becomes a matrix made by dummy variables to preserve the structure (2.1). When this factor is not influential we expect that the whole vector $\boldsymbol{\beta}_j$ will be zero.

Forming predictors to groups can be naturally applied when some predictors are highly correlated among themselves. Then we might want to calculate the coefficient path similar to the lasso path from LARS. In the following paragraph we suggest a version of LARS for grouped variables.

Let us, for simplicity, assume that $p_j = 2$, for every j. That is, we have just two predictors in each group. We outline a modification of the LARS algorithm for the group lasso which has the same principle as the group LARS algorithm in (Yuan and Lin, 2006). As in the classical LARS we begin with zero coefficients, an empty active set, and the current residuum equal to $\boldsymbol{y}$. In each group the pair of predictors generate a plane. For each factor we calculate an orthogonal vector to the plane $\boldsymbol{n}_j$, and by α_j we denote 90 degrees minus the angle between $\boldsymbol{n}_j$ and $\boldsymbol{y}$. We find a factor j_1 having the smallest angle α_j among all $j = 1, \ldots, L$. Such factor has also the smallest angle with $\boldsymbol{y}$. We put the factor into the active set. Then we take the largest step in the direction of projection $\boldsymbol{y}$ on the plane of the factor until some other factor from $\{1, \ldots, L\} - \{j_1\}$ has the same angle α_j with the current residuum. We add this factor into the active set and proceed in the direction of the projected current residuum on the space spanned by these two factors until a third factor comes up, and so on. As a result, we get a similar path of the coefficients as in classical LARS, but here, the whole group of variables enters the path at each knot.

Our aim is to show that the idea of the algorithm is similar as in the classical LARS. Technical details and a version of the group LARS algorithm for a different number of predictors in each group can be found in (Yuan and Lin, 2006).

3. Statistical inference for lasso

3.1 Classical statistical inference

The statistical inference in general refers to exploring a statistical model by performing various procedures, such as different statistical tests, constructing p-values, point estimates, and confidence intervals using data. Based on this inference we can discuss the properties of the model and make conclusions. The inference answers our questions about the model via statistical tests, for example, whether some assumptions are violated, or the model is inappropriate for the data.

Suppose, that our data consist of the response $\boldsymbol{y} \in \mathbb{R}^n$ and three other predictors $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, all from $\mathbb{R}^n$. We have then $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) \in \mathbb{R}^{n\times 3}$ (without an intercept term) and we want to know how $\boldsymbol{y}$ depends on these three predictors using the normal linear model (1.1). Suppose, our initial model is

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n,$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ are i.i.d. normally distributed error terms with zero mean, variance σ^2, and $\boldsymbol{\beta} = (\beta_1,\ \beta_2,\ \beta_3)^\top$ is the vector of regression parameters. We then calculate OLS estimates $\widehat{\beta}_1^{OLS}$, $\widehat{\beta}_2^{OLS}$, and $\widehat{\beta}_3^{OLS}$. Our question (among other problems related to the finding an appropriate model) is whether we need to include all three predictors in our model. To answer that we can perform various statistical tests. We usually use the residual sum of squares $\mathrm{RSS} = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2$ in F-test (when σ^2 is unknown) or the chi-square test (when σ^2 is known) to test a submodel. This test is based on the difference $\mathrm{RSS}_0 - \mathrm{RSS}$, where by RSS_0 we denote the residual sum of squares of the submodel.

The submodel with two predictors can take, for example, the form

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i.$$

Now, for testing the significance of the third predictor, we need to test the null hypothesis that $\beta_3 = 0$. Thus, we can use the t-test, equivalent to the F-test. We can also construct the p-value and the corresponding confidence interval for the estimates $\widehat{\beta}_1^{OLS}$, $\widehat{\beta}_2^{OLS}$, $\widehat{\beta}_3^{OLS}$, because we are familiar with the distribution of the test statistic.

Such statistical inference is possible for a general predictor matrix $\mathbf{X}$ only for fixed $p < n$, i.e., when the number of predictors is smaller than the sample size. If $p > n$ there is an infinite number of solutions for the parameter $\boldsymbol{\beta}$ in (1.1). The estimate $\widehat{\boldsymbol{\beta}}^{OLS}$ depends on the pseudoinverse $\left(\mathbf{X}^\top\mathbf{X}\right)^-$ and the regression parameter is not well-defined. In this case we either need some regularization of the estimated parameters or some additional assumptions on the matrix $\mathbf{X}$.

In the rest of this chapter we will focus on various statistical inference tools, which can be used for the lasso estimates. The lasso as such belongs to the L_1-regularization class of a linear model. The solution of the lasso problem (1.4) is nonlinear in $\boldsymbol{y}$, and it is not easy to characterize the distribution of the lasso estimate $\widehat{\boldsymbol{\beta}}$ (especially when $p > n$), as we cannot write an explicit form of $\widehat{\boldsymbol{\beta}}$. Actually, we are familiar only with the asymptotic distribution for fixed

p as n goes to infinity (Knight and Fu, 2000), but even in this situation we cannot use it for the construction of the confidence intervals or p-values, as it has point mass at zero (Dezeure et al., 2015). Due to this reason we cannot use the same statistical framework (confidence intervals, p-values, etc.) as in the classical linear regression. The main problem is that the model (the number of nonzero parameters) that is selected by the lasso is never known *a priori* for real data, and the model selection is, as we already mentioned, random. The classical inference techniques cannot deal with this randomness because they only consider a fixed hypothesis (Lockhart et al., 2014). In case of the lasso the tests need to be random, and they depend on the selected model. In Section 3.3 we describe the approaches used for the inference on the lasso estimates together with some examples.

Before we describe the inference itself we briefly discuss consistency property of the lasso estimates, which will turn out to be important later in this section.

3.2 Consistency of the lasso estimate

In this part we examine the asymptotic properties of the lasso estimate. Considering the lasso as a tool for the variable selection, we could also desire the lasso to identify the relevant variables, i.e., subset of coefficients that are not equal to zero in the true model. For an estimator which performs the model selection there are so called *oracle properties* (see Fan and Li, 2001) used to qualitatively judge the performance the estimation procedure. An estimator which satisfies these properties selects truly the subset of important coefficients, and the estimated coefficients follow, asymptotically, the normal distribution.

We assume the linear regression model

$$\boldsymbol{y}_n = \mathbf{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n,$$

i.e., similar as (1.1) but indexed with n. Now $\boldsymbol{\varepsilon}_n$ is the vector of i.i.d. random errors with zero mean and variance σ^2, not necessarily normally distributed.

For the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ we keep the notation without n, although we should correctly write $\boldsymbol{\beta}$ depending on n. We also assume that the model is sparse, which means some of the parameters $\boldsymbol{\beta}$ are exactly zero. These zero coefficients represent irrelevant predictors according to the response.

Next we define the set of indexes which correspond to nonzero parameters in the true model

$$\mathcal{A} = \{j; \beta_j \neq 0, j = 1, \ldots, p\}.$$

The notation here is the same as in Section 1.2.1 as the set $\mathcal{A}$ has the same meaning. Similarly, we define the set of indexes corresponding to the nonzero lasso estimates (or estimates coming from an another estimator)

$$\hat{\mathcal{A}} = \left\{j; \hat{\beta}_j \neq 0, j = 1, \ldots, p\right\}.$$

Further, we denote by Σ the covariance matrix of true nonzero predictors and by $\widehat{\boldsymbol{\beta}}^n$ the estimated regression coefficients. We say that an estimator satisfies the oracle properties when the following two conditions hold:

- *Selection consistency*: $\mathsf{P}(\hat{\mathcal{A}} = \mathcal{A}) \to 1$, as $n \to \infty$,

- *Asymptotic normality*: $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}\right) \xrightarrow{d} N(0, \Sigma),$ as $n \to \infty,$

where $\boldsymbol{\beta}_{\mathcal{A}}$ is the vector of coefficients whose indexes are in the active set, and Σ is the covariance matrix.

Even though these properties are desired for some procedure performing the model selection, the lasso, as we defined it, does not generally guarantees these oracle properties. In (Zhao and Yu, 2006) there is described so called *irrepresentable* condition under which the lasso has the oracle properties (even when we allow p to grow in a certain way as n grows). However, this condition is very strict and it cannot be easily verified in practice as it depends on signs of nonzero parameters in the true model.

Although the classical lasso itself is not consistent in terms of oracle properties or sign consistency, its other versions can be. For instance, *the adaptive lasso* (Zou, 2006) and *the elastic net* (Zou and Hastie, 2005) can be shown to satisfy the oracle properties.

The adaptive lasso estimate is defined by

$$\widehat{\boldsymbol{\beta}}^{Adap} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{2} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|,$$

where $\lambda \geq 0$, and w_j are the weights for the regression coefficients. The common estimate for the weights w_j is $\hat{w}_j = 1/|\widehat{\beta}_j^{OLS}|$, for $j =, \ldots, p$. Beside the consistency property, the adaptive lasso also overcomes the fact that all coefficients estimated by the original lasso problem are shrunk by the same amount.

Another lasso adaptation, which preserve the oracle properties, is the elastic net. The simplest version can be written as

$$\widehat{\boldsymbol{\beta}}^{Elastic} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{2} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2,$$

where $\lambda_1, \lambda_2 \geq 0$. The lasso is then the special case of the elastic net with $\lambda_2 = 0$. In a high dimensional regression, $p > n$, the lasso selects at most n variables. The elastic net can, however, choose more than n non-zero terms.

3.3 Post-selection inference

In this part we describe the *post-selection inference* for the lasso. We give an overview of various approaches in the post-selection inference (in which most of the outcomes belongs to the recent findings), and we describe some specific tests done after the model selection.

As the lasso is an adaptive procedure (meaning that the chosen set of nonzero coefficients by the procedure is never known apriori, i.e., two different models can be chosen for two different samples of a data-set) the classical statistical inference cannot be applied for lasso estimates. To be more specific, let us consider the F-test for testing the significance of some specific variable in a model and the related submodel (we assume σ to be unknown). The test statistic then is based on the difference between the residual sum of squares in the submodel and the model itself (for some fixed variable or a set of variables). However, if the variable is

picked by the lasso or another adaptive method (such as the stepwise regression) the jump in RSS will stochastically not result in the F-distribution, as it is described in (Lockhart et al., 2014). Therefore, the classical statistical inference tools are no longer valid for the adaptive procedures with the random selection in general.

In the classical inference, we usually propose some model and we gather data. Later, we validate the model by the means of some statistical test. There is some specific set of hypotheses defined and the appropriate test is used to decide. However, in the post-selection approach these hypotheses need to be random in order to account for the random process in the selection procedure. As a result, it is more difficult to describe them. In the following, we will describe three specific statistical tests which were designed to perform the inference in adaptively fitted models, for instance, by the lasso method.

The statistical theory about the post-selection inference is quite recent and it is still being developed. The theory to the post selection inference is not easy at all. The one test we further discuss more in detail is called the *covariance test* and it was the first meaningful step towards the post-selection inference. It is based on the LARS algorithm described in Section 1.2.1 The related test statistic depends on the lasso path at the points (knots) where slopes in the path are changing, and where the active set is changed. The test is asymptotic with the test statistic structure reminding the F-test (or χ^2-test).

The other two tests, namely *the TG test* (TG for "truncated gaussian") and *the spacing test*, that will be described more precisely, are based on so called *polyhedral lemma* (see further). They can be applied to any model selection method for which the selection can be written in terms of some linear inequalities on $\boldsymbol{y}$: $\{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{v}\}$, for some matrix $\boldsymbol{D}$ and a vector $\boldsymbol{v}$. In the previous notation the inequality $\leq$ is meant element-wise. The resulting distribution of the test statistic is exact and generally, the TG test does not assume any conditions on the data matrix $\mathbf{X}$ as the covariance test does. The spacing test is the exact version of the covariance test.

Different approach which can be used to find confidence intervals for the lasso estimates is, for instance, the *PoSI* method introduced in (Berk et al., 2013). This procedure fits a selected submodel, and then "corrects" the classical confidence intervals because it takes into account all possible submodels that the lasso could choose. These PoSI intervals do not assume known σ or fixed λ. On the other hand, the resulting confidence intervals can be very wide.

3.3.1 Covariance test

The test described in this section is based on the LARS algorithm and the resulting lasso path (piece-wise linear function for each non-zero estimated coefficient), and it was introduced in (Lockhart et al., 2014). Similarly, as we can test the significance of a predictor in the linear regression model, we would like to obtain the p-values for the coefficients selected by the lasso. We need to, however, choose a different approach because of the reasons described previously. The idea, which is used here, is to test variables as they are arriving to the active set step by step in the lasso path. Figure 3.1 shows an example of such path where the model was simulated with five variables but three of them being truly non-zero.

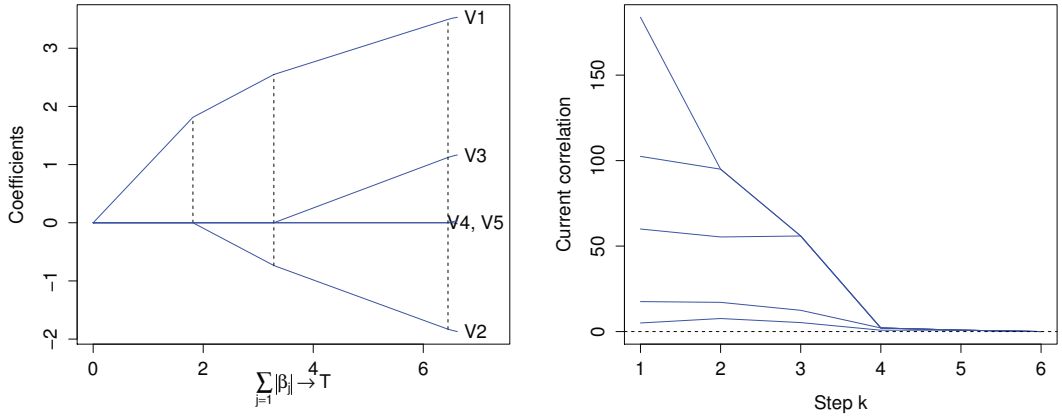


Figure 3.1: *An example of the lasso path (left) and the related graph of the decreasing covariances (right). Just three predictors ($\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$) become meaningful along the path. The covariances of the predictors with the current residuals are decreasing (in absolute value) and they are always bounded from above by the maximum current correlation $\hat{C}$.*

We keep the same notation as in Section 1.2.1. We denote by $\lambda_1 > \lambda_2 > \cdots > \lambda_r = 0$ the points (knots) where the slope of lines is changing in the lasso path, that is, where some new variable is added to the "non-zero coefficient" set $\mathcal{A}$. The first point λ_1 indicates the entry of the first variable to the model (all coefficients are still zero at λ_1), and at $\lambda_r = 0$ all coefficients are equal to their OLS estimates. Values λ_l, $l = 1, \ldots, r$ are actually equal to the maximum absolute current correlation $\hat{C}$ at each step, as can be seen in (Efron et al., 2004). According to the previous we want to test *the significance of some variable joining the active set $\mathcal{A}$ at some specific step of the LARS algorithm.*

Before we state the formal description how the test works, we provide an important condition on the model matrix $\mathbf{X}$.

Definition 1. *We say that columns $X_1, \ldots, X_p \in \mathbb{R}^n$ are in a* general position *if the affine span of any $k+1$ vectors $s_1 X_1, \ldots, s_{k+1} X_{k+1}$ does not contain any element of the set $\{\pm X_i : \ i \neq i_1, \ldots, i_{k+1}\}$ for any signs $s_1, \ldots, s_{k+1} \in \{-1, 1\}$, $k < min\{n, p\}$.*

The previous condition is quite week and it holds, for example, when the entries of $\mathbf{X}$ come from a continuous probability distribution. The general position of the columns of the model matrix ensures that the lasso path will be unique, as shown in (Tibshirani, 2013).

Now, we formally define the test statistic and explain how the p-values can be obtained. We still assume the linear regression set-up (1.1), the response with the normally distributed errors and with the variance σ^2. We further assume that $\mathbf{X}$ has its columns in the general position. Suppose, we want to test the significance of a predictor j joining the active set at λ_l (the l-th step of LARS). Let $\mathcal{A}_{l-1}$ be the active set at the previous step without the index j and let $\widehat{\boldsymbol{\beta}}(\lambda_{l+1})$ be the lasso estimate at the next knot λ_{l+1}. Further let $\bar{\boldsymbol{\beta}}_{\mathcal{A}_{l-1}}(\lambda_{l+1})$ be the lasso solution

using only variables in $\mathcal{A}_{l-1}$:

$$\bar{\boldsymbol{\beta}}_{\mathcal{A}_{l-1}}(\lambda_{l+1}) = \underset{\boldsymbol{\beta}_{\mathcal{A}_{l-1}} \in \mathbb{R}^{|\mathcal{A}_{l-1}|}}{\arg\min} \frac{1}{2} \left\| \boldsymbol{y} - \mathbf{X}_{\mathcal{A}_{l-1}} \boldsymbol{\beta}_{\mathcal{A}_{l-1}} \right\|_2^2 + \lambda_{l+1} \left\| \boldsymbol{\beta}_{\mathcal{A}_{l-1}} \right\|_1, \tag{3.1}$$

where $\mathbf{X}_{\mathcal{A}_{l-1}}$ contains the columns of $\mathbf{X}$, which corresponds to the predictors in $\mathcal{A}_{l-1}$ (the active predictors at the previous step). We define the *covariance test statistic* by

$$T_l^{cov} = \frac{\left(\boldsymbol{y}^\top \mathbf{X} \widehat{\boldsymbol{\beta}}(\lambda_{l+1}) - \boldsymbol{y}^\top \mathbf{X}_{\mathcal{A}_{l-1}} \bar{\boldsymbol{\beta}}_{\mathcal{A}_{l-1}}(\lambda_{l+1}) \right)}{\sigma^2}. \tag{3.2}$$

The name of the test statistic comes from the fact that the numerator of the expression (3.2) can be written as a difference between the empirical covariances and some small term. When the difference in the test statistic is large then the covariance of y and $\mathbf{X}\widehat{\boldsymbol{\beta}}$ is greater than the covariance of y and $\mathbf{X}_{\mathcal{A}_{l-1}} \bar{\boldsymbol{\beta}}_{\mathcal{A}_{l-1}}$ and the j-th variable is more important in the model with the active set $\mathcal{A}_l = \mathcal{A}_{l-1} \cup \{j\}$.

The covariance test statistic is evaluated at the next knot λ_{l+1} because at λ_l the j-th coefficient is still equal to zero. At $\lambda = \lambda_{l+1}$ we can see the full influence of the variable j on fitted values $\mathbf{X}\widehat{\boldsymbol{\beta}}$ right before the next variable becomes non-zero at the step $l+1$ or some other variable is removed from the active set.

The following important result allows us to calculate the corresponding p-values. For the next lines we will understand the current model to be the lasso model with the active set $\mathcal{A}_{l-1}$. Under the null hypothesis that all variables not included in the current model have zero coefficients in the true model (we can write $\mathcal{A} \subseteq \mathcal{A}_{l-1}$ where $\mathcal{A}$ denotes the true active set) the test statistic (3.2) has asymptotically exponential distribution:

$$T_l^{cov} \xrightarrow{d} \text{Exp}(1).$$

The proof can be found in (Tibshirani, 2013). Using this result we can calculate the corresponding p-value for the variable entering the non-zero set of coefficients at each step of the LARS algorithm. The null hypothesis is random in this case, conditional on the covariates selected previously, because $\mathcal{A}_{l-1}$ and j are also random. The test can be, therefore, performed at sequential steps. When all true predictors are included in the active set then an additional variable should not be significant and thus, the null hypothesis should not be rejected.

We will now introduce two other forms of the test statistic (3.2), one to see the connection to the usual chi-squared statistic and the other useful for computational reasons. Let us denote by $s_{\mathcal{A}_{l-1}}$ the signs of the lasso solution $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_{l-1}}(\lambda_l)$. We further assume that variables of the lasso solution $\bar{\boldsymbol{\beta}}_{\mathcal{A}_{l-1}}$ are all nonzero at λ_{l+1}, and the signs are matching $s_{\mathcal{A}_{l-1}}$:

$$s_{\mathcal{A}_{l-1}} = \text{sign}\left(\bar{\boldsymbol{\beta}}_{\mathcal{A}_{l-1}}(\lambda_{l+1}) \right). \tag{3.3}$$

Then the covariance test statistic can be written as

$$T_l^{cov} = \frac{\left\| \boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{OLS}_{\mathcal{A}_{l-1}} \right\|_2^2 - \left\| \boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{OLS}_{\mathcal{A}_l} \right\|_2^2}{\sigma^2} - \lambda_{l+1} \frac{(\mathbf{X}_{\mathcal{A}_l}^\top)^+ s_{\mathcal{A}_l} - (\mathbf{X}_{\mathcal{A}_{l-1}}^\top)^+ s_{\mathcal{A}_{l-1}}}{\sigma^2}, \tag{3.4}$$

where $(\mathbf{X})^+ = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is a Moore-Penrose pseudoinverse of $\mathbf{X}$, $s_{\mathcal{A}_l}$ are the signs of the lasso solution at λ_l together with the sign of the predictor j entering the active set at λ_l, and $\widehat{\boldsymbol{\beta}}^{OLS}_{\mathcal{A}_{l-1}}$ is OLS estimate using just predictors in $\mathcal{A}_{l-1}$ (analogously $\widehat{\boldsymbol{\beta}}^{OLS}_{\mathcal{A}_l}$). The first expression in the above formula is now exactly the chi-squared statistic for usual testing of the significance of predictor j where, of course, $\mathcal{A}_{l-1}$ and j are random.

The above formula leads to so called *knot form* of the covariance test statistic. We still assume that coefficients of the restricted lasso problem (3.1) are all active at λ_{l+1} and j is the variable joining the active set at λ_l. Then

$$T^{cov}_l = G(\mathcal{A}_{l-1}, s_{\mathcal{A}_{l-1}}, j)\ \frac{\lambda_l(\lambda_l - \lambda_{l+1})}{\sigma^2}, \tag{3.5}$$

where

$$G(\mathcal{A}_{l-1}, s_{\mathcal{A}_{l-1}}, j) = \left\| (\mathbf{X}^\top_{\mathcal{A}_l})^+ s_{\mathcal{A}_l} - (\mathbf{X}^\top_{\mathcal{A}_{l-1}})^+ s_{\mathcal{A}_{l-1}} \right\|^2_2.$$

The derivations of the previous two forms of the covariance test statistic can be found in (Tibshirani, 2013).

The parameter σ^2 is, however, usually unknown in practice and till now we have always defined the test statistic only for known σ^2. Fortunately, there exists also a version of the test where in (3.2) we replace σ^2 by the estimate from the full linear regression model

$$\hat{\sigma^2} = \text{MSE} = \frac{\left\| \boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{OLS} \right\|^2_2}{n-p}.$$

In (Tibshirani, 2013) there is shown that under the null hypothesis the resulting test statistic is asymptotically distributed as a random variable from the distribution $F_{2,n-p}$, i.e., F-distribution with 2 and $n-p$ degrees of freedom.

In the high-dimensional case the parameter σ^2 cannot be so easily estimated and the approaches here differ. Various methods for the estimation of the variance of errors in the lasso regression are described, for instance, in (Reid et al., 2014).

The covariance test has some limitations and we already mentioned some of them. First, certain condition are needed for the model matrix $\mathbf{X}$. For example, when a categorical variable exists among the predictors and the resulted model is described by dummy variables the assumption about $\mathbf{X}$ having columns in general position is violated. The test also does not take into account multiple comparisons. If some variable, for instance, enters the model more than once (which is permitted by the lasso modification of the LARS algorithm), we treat each situation separately and perform separate tests. At last, the test is (also for known σ^2) only asymptotic.

In the following section we introduce another test that can be applied after the model selection but this time with an exact distribution of the test statistic.

3.3.2 Tests based on the polyhedral lemma

Now, we describe two other exact tests and we also derive a generalized version of one of them that can be relatively easily used in practice. Before describing the tests and their application to the lasso estimates, we present a few important

theoretical results that will help us understand the distribution of the resulting test statistic. We will focus on the situation where the model is selected by an adaptive procedure and that can be described by linear inequalities on $\boldsymbol{y}$ as $\{\boldsymbol{y} : \boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{v}\} =: \{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{v}\}$. The latter set is in fact a polyhedron giving the name to the lemma stated below. The set $\{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{v}\}$ therefore characterizes all vectors $\boldsymbol{y}$ that yield the same selected (active) variables.

We assume that $\boldsymbol{y}$ is normally distributed, as in the last section:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n).$$

Suppose $\boldsymbol{\nu} \in \mathbb{R}^n$ is a n-vector and we are interested in the inference on $\boldsymbol{\nu}^\top \boldsymbol{y}$ conditionally on the selection. For a specific choice of $\boldsymbol{\nu}$ the term $\boldsymbol{\nu}^\top \boldsymbol{\mu}$ becomes the regression coefficient β_j of a j-th predictor, but generally by specifying $\boldsymbol{\nu}$, we can test any linear constrains on the regression coefficients. To this goal we need to study the conditional distribution of $\boldsymbol{\nu}^\top \boldsymbol{y}$. The following lemma gives us a way how to make an inference about $\boldsymbol{\nu}^\top \boldsymbol{\mu}$.

Lemma 1 (Polyhedral lemma)**.** *Let $\Sigma = \sigma^2 \boldsymbol{I}_n$ be a covariance matrix of $\boldsymbol{y}$ and $\boldsymbol{\nu} \in \mathbb{R}^n$, for which $\boldsymbol{\nu}^\top \Sigma \boldsymbol{\nu} \neq 0$. Further, let $c = \left(\boldsymbol{\nu}^\top \Sigma \boldsymbol{\nu}\right)^{-1} \boldsymbol{D} \Sigma \boldsymbol{\nu}$. Then the set $\{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{v}\}$ can be written as $\left\{\mathcal{B}^-(\boldsymbol{y}) \leq \boldsymbol{\nu}^\top \boldsymbol{y} \leq \mathcal{B}^+(\boldsymbol{y}), \mathcal{B}^0(\boldsymbol{y}) \leq 0\right\}$ where*

$$\mathcal{B}^-(\boldsymbol{y}) = \max_{j : c_j > 0} \frac{v_j - (Dy)_j + c_j \boldsymbol{\nu}^\top \boldsymbol{y}}{c_j}, \tag{3.6}$$

$$\mathcal{B}^+(\boldsymbol{y}) = \min_{j : c_j < 0} \frac{v_j - (Dy)_j + c_j \boldsymbol{\nu}^\top \boldsymbol{y}}{c_j}, \tag{3.7}$$

$$\mathcal{B}^0(\boldsymbol{y}) = \max_{j : c_j = 0} v_j - (Dy)_j. \tag{3.8}$$

In addition, if $\boldsymbol{y}$ is normally distributed, as in (1.1)*, then terms $\mathcal{B}^-$, $\mathcal{B}^+$, $\mathcal{B}^0$ are independent of $\boldsymbol{\nu}^\top \boldsymbol{y}$.*

Proof. See (Lee et al., 2016). □

The presented lemma tells us that $\{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{v}\}$ is the same as the event when $\boldsymbol{\nu}^\top \boldsymbol{y}$ lies between two values that depend on $\boldsymbol{D}$ and $\boldsymbol{v}$. We assume normality of $\boldsymbol{y}$, therefore, the conditional distribution of $\boldsymbol{\nu}^\top \boldsymbol{y} \mid \{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{v}\}$ is also normal but conditional (truncated) to lie between boundaries $\mathcal{B}^-(\boldsymbol{y})$ and $\mathcal{B}^+(\boldsymbol{y})$ (details are, for instance, in Lee et al., 2016). In the following, we will see why these results are important for the lasso.

Let us assume that $\mathbf{X}$ has columns in the general position, implying a unique lasso path. Now, the event when the lasso selects the model is characterized just by the non-zero variables (i.e., active variables) and by their signs. For instance, this event can be written as $\left\{\hat{\mathcal{A}} = \mathcal{A}_l,\ \hat{s}_{\mathcal{A}} = s_{\mathcal{A}_l}\right\}$ where $\mathcal{A}_l,\ s_{\mathcal{A}_l}$ are the active set and signs of active variables selected at the l-th step of the LARS algorithm. The key result is that this event can be also written as a set of linear inequalities for $\boldsymbol{y}$ (see Lee et al., 2016). Specifically in (Tibshirani et al., 2016) the authors show that the set has the form $\{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{0}\}$. The latter set now corresponds to the set of such vectors $\boldsymbol{y}$ for which the active set and the signs are the same (we consider fixed $\mathbf{X}$). The number of rows of $\boldsymbol{D}$ depends on p and also on l, noting the l-th step of the LARS algorithm in which we are fixing the selected model.

Obtaining the polyhedral set for the lasso arises from the general Karush-Kuhn-Tucker (KKT) conditions (see, for example, Tibshirani, 2013), when we rewrite them for the lasso solution vector $\widehat{\boldsymbol{\beta}}$.

Using the previous lemma we can express the model selection as some polyhedral restrictions on the response vector $\boldsymbol{y}$ (meaning that all y's satisfying the polyhedral restriction will yield the same model with exactly the same coefficient paths). Further, let $\Phi(x)$ be the cumulative distribution function (CDF) of the standard normal distribution $N(0,1)$, and we denote by $G^{a,b}_{\mu,\sigma^2}$ the CDF of the truncated normal distribution with the support $[a,b]$ (here μ is a parameter of the distribution $N(\mu,\sigma^2)$):

$$G^{a,b}_{\mu,\sigma^2}(x) = \frac{\Phi((x-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}. \tag{3.9}$$

Then we use the fact that the CDF of a random variable at the value of that variable has a uniform distribution together with Lemma 1 and we have

$$G^{\mathcal{B}^-,\mathcal{B}^+}_{\boldsymbol{\nu}^\top\boldsymbol{\mu},\boldsymbol{\nu}^\top\Sigma\boldsymbol{\nu}}\left(\boldsymbol{\nu}^\top\boldsymbol{y}\right) \;\mid\; \{\boldsymbol{D}\boldsymbol{y} \le \boldsymbol{0}\} \sim \mathrm{Unif}(0,1). \tag{3.10}$$

The proof of the previous statement can be found in (Lee et al., 2016). Deriving the test is now relatively straightforward. Let us assume the null hypothesis

$$H_0: \; \boldsymbol{\nu}^\top\boldsymbol{\mu} = 0 \quad \text{against} \quad H_1: \; \boldsymbol{\nu}^\top\boldsymbol{\mu} \neq 0.$$

Define the test statistic

$$T = 2\min\left\{G^{\mathcal{B}^-,\mathcal{B}^+}_{0,\boldsymbol{\nu}^\top\Sigma\boldsymbol{\nu}}\left(\boldsymbol{\nu}^\top\boldsymbol{y}\right),\; 1 - G^{\mathcal{B}^-,\mathcal{B}^+}_{0,\boldsymbol{\nu}^\top\Sigma\boldsymbol{\nu}}\left(\boldsymbol{\nu}^\top\boldsymbol{y}\right)\right\}. \tag{3.11}$$

Applying the polyhedral lemma the test statistic (3.11) can be used as a p-value for the null hypothesis $\boldsymbol{\nu}^\top\boldsymbol{\mu} = 0$ conditional on $\{\boldsymbol{D}\boldsymbol{y} \le \boldsymbol{0}\}$. When we are testing H_0 on the significance level α we are rejecting the null hypothesis if $T \le \alpha$.

TG test

We can finally describe a test for the lasso estimates that can be used for any λ, called the *TG test*. We fix λ (or a step of the LARS algorithm) and compute the appropriate matrix $\boldsymbol{D}$. The null hypothesis is, for now, H_0: $\boldsymbol{\nu}^\top\boldsymbol{\mu} = 0$, that is, we consider any linear constraint on $\boldsymbol{\mu}$ (eventually on $\boldsymbol{\beta}$). We calculate also the boundaries $\mathcal{B}^-$ and $\mathcal{B}^+$ according to the polyhedral lemma using $\boldsymbol{v} = \boldsymbol{0}$. Now, assume we are testing the null hypothesis against a one-sided alternative H_1: $\boldsymbol{\nu}^\top\boldsymbol{\mu} > 0$. We define the following test statistic

$$T^{tg} = 1 - G^{\mathcal{B}^-,\mathcal{B}^+}_{0,\sigma^2\boldsymbol{\nu}^\top\boldsymbol{\nu}}\left(\boldsymbol{\nu}^\top\boldsymbol{y}\right) = \frac{\Phi(\frac{\mathcal{B}^+}{\sigma\boldsymbol{\nu}^\top\boldsymbol{\nu}}) - \Phi(\frac{\boldsymbol{\nu}^\top\boldsymbol{y}}{\sigma\boldsymbol{\nu}^\top\boldsymbol{\nu}})}{\Phi(\frac{\mathcal{B}^+}{\sigma\boldsymbol{\nu}^\top\boldsymbol{\nu}}) - \Phi(\frac{\mathcal{B}^-}{\sigma\boldsymbol{\nu}^\top\boldsymbol{\nu}})}. \tag{3.12}$$

Then according to (3.10) the test statistic T^{tg} can be also used as a p-value for H_0, conditional on $\{\boldsymbol{D}\boldsymbol{y} \le \boldsymbol{0}\}$ (Tibshirani et al., 2016). For testing

$$H_0: \; \boldsymbol{\nu}^\top\boldsymbol{\mu} = 0 \quad \text{vs} \quad H_1: \; \boldsymbol{\nu}^\top\boldsymbol{\mu} \neq 0,$$

the test statistic has the form

$$T^{TG} = 2\,\min\left\{T^{tg}, 1 - T^{tg}\right\}.$$

When testing on the significance level α we reject the null hypothesis if $T^{TG} \leq \alpha$.

Using the latter we can also construct a conditional confidence interval with the confidence level $1-\alpha$. This confidence interval has then the form $\left[u_{\frac{\alpha}{2}},\, u_{1-\frac{\alpha}{2}}\right]$ where $u_{\frac{\alpha}{2}},\, u_{1-\frac{\alpha}{2}}$ are boundaries that satisfy

$$1 - G^{\mathcal{B}^-,\mathcal{B}^+}_{u_{\frac{\alpha}{2}},\sigma^2\boldsymbol{\nu}^\top\boldsymbol{\nu}}\left(\boldsymbol{\nu}^\top\boldsymbol{y}\right) = \frac{\alpha}{2}$$
$$1 - G^{\mathcal{B}^-,\mathcal{B}^+}_{u_{1-\frac{\alpha}{2}},\sigma^2\boldsymbol{\nu}^\top\boldsymbol{\nu}}\left(\boldsymbol{\nu}^\top\boldsymbol{y}\right) = 1 - \frac{\alpha}{2}.$$

When we are interested in the testing the significance of , say, j-th variable with the null hypothesis H_0: $\beta_j = 0$, the vector $\boldsymbol{\nu}$ becomes $\boldsymbol{\nu} = \left(\mathbf{X}^+_{\mathcal{A}}\right)^\top \boldsymbol{e}_j$. Here $\mathcal{A}$ denotes the active set at the fixed value λ and $\boldsymbol{e}_j$ is a p-dimensional vector made by zeros and 1 on the j-th place:

$$\boldsymbol{\nu}^\top\boldsymbol{\mu} = \boldsymbol{e}_j^\top\mathbf{X}^+_{\mathcal{A}}\boldsymbol{\mu} = \boldsymbol{e}_j^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \beta_j.$$

We see that the null hypothesis in the TG test is also random ($\mathcal{B}^-$ and $\mathcal{B}^+$ are random variables), as in the covariance test, and conditional on variables and their signs in the current active set. The TG test for the lasso assumes only a general position of columns of $\mathbf{X}$, which is a weak assumption. The test can be used for any fixed λ and it is an exact test.

Further, we introduce another exact test and the more general version of it, which is easy to be performed.

Spacing test

We defined the TG test for an arbitrary λ, but the previous framework can be applied also to the steps of the LARS algorithm giving so called *spacing test*. Let us denote by $\mathcal{A}_l$ the nonzero set of the coefficients at the l-th step of LARS, by j_l the index of the variable added to the active set at the l-th step, and by s_{j_l} the sign of the variable j_l. To be clear, we consider now the LARS algorithm without the lasso modification, that is, we consider only adding a variable at each step of the algorithm. Details and proofs of following statements can be found in (Tibshirani et al., 2016).

The selection of the active set of variables at each step of LARS without the lasso modification can be also written in terms of the polyhedral set $\{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{v}\}$. Actually, the matrix $\boldsymbol{D}$ now contains even less rows than in the case of the lasso selection event. Furthermore, the selection event at the l-th step of LARS can be approximated by a polyhedral set $\{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{u}\}$ where rows of the matrix $\boldsymbol{D}$ are even more reduced. The approximate representation $\{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{u}\}$ does not take into account some restrictions that often hold automatically for the original set $\{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{v}\}$, and thus by excluding them, they make $\boldsymbol{D}$ smaller. Matrix $\boldsymbol{D}$ and vector $\boldsymbol{u}$ are both specified by the approximation. They both depend on l, actually, the matrix D has just $l+1$ rows and $\boldsymbol{u}$ is random.

This approximation with the selection event $\{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{u}\}$ with D small can be applied only for some $\boldsymbol{\nu}$ when testing $\boldsymbol{\nu}^\top\boldsymbol{\mu} = 0$. The vector $\boldsymbol{\nu}$ has to lie in the column space of active variables at the l-step of LARS. Then the representation $\{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{u}\}$ can be again written as $\boldsymbol{\nu}^\top\boldsymbol{y}$ laying between two boundaries which are independent of $\boldsymbol{\nu}^\top\boldsymbol{y}$ in case of the normally distributed $\boldsymbol{y}$. In particular, for

the variable indexed by j_l, and entering the active set at the l-th step, we have $\lambda_l = \boldsymbol{\nu}^\top \boldsymbol{y}$ for

$$\boldsymbol{\nu} = \boldsymbol{\nu}_S = \frac{P^o_{\mathcal{A}_{l-1}} X_{j_l}}{s_{j_l} - X_{j_l}^\top (\mathbf{X}^+_{\mathcal{A}_{l-1}})^\top s_{\mathcal{A}_{l-1}}}. \tag{3.13}$$

Here $P^o_{\mathcal{A}} = \boldsymbol{I}_n - \mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}^\top$ is an orthogonal projection on the column space of $\mathbf{X}_{\mathcal{A}}$, for a subset $\mathcal{A} \subset \{1, \dots, p\}$. This particular choice of $\boldsymbol{\nu}$ will give us the spacing test. Testing $H_0 : \boldsymbol{\nu}^\top \boldsymbol{\mu} = 0$ is now equivalent to test $H_0 : \boldsymbol{e}_j^\top \mathbf{X}^+_{\mathcal{A}} \boldsymbol{\mu} = \beta_{j_l} = 0$. Therefore, we see that the spacing test serves for the *testing the significance of the variable added at the l-th step of the LARS algorithm.*

We define the test statistic of the spacing test as

$$T_l^{SP} = \frac{\Phi\left(\lambda_{l-1}\frac{G(\mathcal{A}_{l-1}, s_{\mathcal{A}_{l-1}}, j)}{\sigma}\right) - \Phi\left(\lambda_l \frac{G(\mathcal{A}_{l-1}, s_{\mathcal{A}_{l-1}}, j)}{\sigma}\right)}{\Phi\left(\lambda_{l-1}\frac{G(\mathcal{A}_{l-1}, s_{\mathcal{A}_{l-1}}, j)}{\sigma}\right) - \Phi\left(M_l^+ \frac{G(\mathcal{A}_{l-1}, s_{\mathcal{A}_{l-1}}, j)}{\sigma}\right)}, \tag{3.14}$$

where $G(\dots)$ is the random variable from the knot form (3.5) of the covariance test and M_l^+ is a random variable defined as a maximum of some linear functions of $\boldsymbol{y}$ (for details see Tibshirani et al., 2016, Section 5). The alternative hypothesis is in this case again $H_1 : \boldsymbol{\nu}^\top \boldsymbol{\mu} > 0$. With such choice of the test statistic we can apply the framework from the TG test where λ was fixed. When testing $H_0 : \boldsymbol{e}_j^\top \mathbf{X}^+_{\mathcal{A}} \boldsymbol{\mu} = \beta_{j_l} = 0$ we can use the test statistic defined by (3.14) as an exact p-value conditional on $\{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{u}\}$. For testing against the two-sided alternative hypothesis we can construct confidence intervals in a similar way as we described previously in TG test.

There is also a simplified version of the spacing test statistic, which does not depend on the random variable M_l^+ and is easily performed in practice. It is also defined at the l-th step of LARS as

$$T_l^{cSP} = \frac{\Phi\left(\lambda_{l-1}\frac{G(\mathcal{A}_{l-1}, s_{\mathcal{A}_{l-1}}, j)}{\sigma}\right) - \Phi\left(\lambda_l \frac{G(\mathcal{A}_{l-1}, s_{\mathcal{A}_{l-1}}, j)}{\sigma}\right)}{\Phi\left(\lambda_{l-1}\frac{G(\mathcal{A}_{l-1}, s_{\mathcal{A}_{l-1}}, j)}{\sigma}\right) - \Phi\left(\lambda_{l+1} \frac{G(\mathcal{A}_{l-1}, s_{\mathcal{A}_{l-1}}, j)}{\sigma}\right)}. \tag{3.15}$$

We see that the statistic now depends only on knots of the LARS algorithm and on the term $G(\dots)$. If we test the significance of the last variable added at the step l of LARS, i.e., $H_0 : \beta_{j_l} = 0$, the statistic (3.15) can be used as a conservative p-value (conditional on $\{\boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{u}\}$) for which

$$\mathsf{P}_{\beta_{j_l}=0}\left(T_l^{cSP} \leq \alpha | \boldsymbol{D}\boldsymbol{y} \leq \boldsymbol{u}\right) \leq \alpha.$$

We can easily see that the statistic (3.15) depends on how far the knots are apart from each other. The larger is the space between them (hence the name "spacing") the smaller the p-value.

In (Tibshirani et al., 2016), there is shown that the spacing test statistic (3.15) and the covariance statistic (3.5) are asymptotically equivalent. The covariance test is therefore an asymptotic version of the spacing test. However, the tested null hypotheses differ in the both tests. The covariance test's null actually says that all coefficients of predictors not contained in the current active set are zero at each step of LARS. The null hypothesis of the spacing test is defined also at the given step of the LARS algorithm, but it tests whether the coefficient of the

variable joining the active set is zero conditional on the other active variables. It means that at the first stage (e.g., for the first predictor to join the active set) both null hypotheses are equivalent, but they differ at subsequent steps. The TG test uses the similar approach as the spacing test, but we can fix any λ, and test any coefficient not included in the related active set.

Both the spacing and the covariance test are designed for the LARS algorithm, the spacing test in addition only for the LARS without the lasso modification, where we do not consider dropping variables from the active set. The TG test can be used also when we calculate the lasso solutions by another method, for example, via the coordinate descent.

3.3.3 Other approaches

We briefly mention other methods that can be used for the inference on the lasso estimates. One popular option for constructing the p-values or confidence intervals for the lasso estimates is so called *random sample-splitting* based on choosing variables using one part of the data and making inference with the chosen variables on the other part. The method is described more in detail, for example, in (Dezeure et al., 2015).

The idea of the procedure is following. Let us assume again that we have a response vector $\boldsymbol{y} \in \mathbb{R}^n$ and a design matrix with other variables $\mathbf{X} \in \mathbb{R}^{n \times p}$. We split the data into two halves $H1$, and $H2$, with $H1 \cap H2 = \emptyset$. We run the lasso method on one sample, say $H1$, and denote the selected variables by $S_{H1} \subset \{1, \ldots, p\}$. We then use the second half of data for finding the p-values by the ordinary least squares with the variables from S_{H1}. The latter means that in the linear regression model we use only rows of $\boldsymbol{y}$ and $\mathbf{X}$ from $H2$ and only columns of $\mathbf{X}$ that relate to S_{H1}. Finally, we can calculate the corresponding p-values (or construct some confidence intervals, for instance,) by the means of the standard inference techniques common for the ordinary linear regression framework.

Because the selected variables and the p-values are changing with different samples we can run the random sampling many times and obtain various p-values for one coefficient (the sample $H1$ is each time chosen randomly). To find just one p-value for some coefficient various approaches can be done. For instance, we can choose the median of all p-values as the final p-value.

The sample-splitting procedure is very simple, but it has some limitations. The lasso is not generally consistent for the model selection, as we noted in Section 3.2, so we have no guarantee that the chosen reduced model is the correct one.

Another method similar to the random-sample splitting is the *bootstrap* (or subsampling) described, for example, in (Efron and Tibshirani, 1993), or in (Meinhausen and Bühlmann, 2010).

Different method based on the *debiased lasso* was introduced in (Javanmard and Montanari, 2014).

3.4 Post-selection test for the group lasso

A similar idea, as in the part about the covariance test, can be used for the test of the significance of the group lasso estimates, introduced in Section 1.2.3.

We further explain how the test can be perform, and we suggest a test statistic that could be used for this type of problems. Similarly as in the covariance test, the proposed test statistic is connected to the lasso path for the group lasso coefficients $\boldsymbol{\beta}_j$, $j = 1, \ldots, L$.

Let us, for simplicity, assume that the number of variables in each group is $p_j = p$, for all $j = 1, \ldots, L$. Therefore, each matrix $\mathbf{X}_j$ of variables related to the group j has p columns. Our idea is to test the significance of a group of variables as it enters the path of the LARS for the group lasso, as described in Chapter 2. We consider a similar notation as in the previous sections about the post-selection tests. Let $\lambda_1 > \lambda_2 > \cdots > \lambda_L = 0$ be knots indicating entrances of L groups. We denote by $\mathcal{A}_l^g$ the active set of non-zero groups at the l-th step of the group LARS algorithm and by $\widehat{\boldsymbol{\beta}}_j(\lambda_l) = (\widehat{\beta}_{j1}, \ldots, \widehat{\beta}_{jp})^\top$, $j = 1, \ldots, L$ the group lasso solutions at the knot λ_l. Further, let $\bar{\boldsymbol{\beta}}(\lambda)$ be the group lasso solutions using only the groups from the previous active set $\mathcal{A}_{l-1}^g$, e.g.,

$$\bar{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^p,\ j \in \mathcal{A}_{l-1}^g}{\arg\min} \frac{1}{2} \left\| \boldsymbol{y} - \sum_{j \in \mathcal{A}_{l-1}^g} \mathbf{X}_j \boldsymbol{\beta}_j \right\|_2^2 + \lambda \sum_{j \in \mathcal{A}_{l-1}^g} \|\boldsymbol{\beta}_j\|_2. \tag{3.16}$$

Following the approach from the covariance test, we evaluate the effect of joining the group at the l-th step at λ_{l+1} because at λ_l all estimates related to that group are still zero. Then the test statistic has the form

$$T_l^g = \frac{\left(\left\langle \boldsymbol{y}, \sum_{j=1}^L \mathbf{X}_j \widehat{\boldsymbol{\beta}}_j(\lambda_{l+1}) \right\rangle - \left\langle \boldsymbol{y}, \sum_{j \in \mathcal{A}_{l-1}^g} \mathbf{X}_j \bar{\boldsymbol{\beta}}_j(\lambda_{l+1}) \right\rangle \right)}{\sigma^2}, \tag{3.17}$$

where $\langle \boldsymbol{r}, \boldsymbol{s} \rangle$ is an inner product of vectors $\boldsymbol{r}$ and $\boldsymbol{s}$, and $\bar{\boldsymbol{\beta}}_j(\lambda_{l+1}) = (\bar{\beta}_{j1}, \ldots, \bar{\beta}_{jp})^\top$. Now, similarly as in the covariance test when the difference in the numerator of the test statistic is large, the l-th group is important in the model.

We will not further investigate the properties of the proposed test statistic for the group lasso as the thesis focuses especially on the lasso. It would be desirable to present some statements about the distribution of (3.17) to justify the proposed form and also for the practical usage. We postpone the topic for some further work. Instead, we discuss some practical properties of the test statistic in the next chapter.

4. Finite sample properties

In this chapter we compare the presented post-selection tests from Sections 3.3.1 and 3.3.2 via various simulations and we apply theses inference tools on a real data scenario. In tables and figures we use following abbreviations for the tests:

- *Cov* - covariance test
- *Space* - spacing test
- *mSpace* - simplified (modified) spacing test
- *TG* - TG test

The whole simulations were performed using the `R` software (Core Development Team 2016) and the available packages (`selectiveInference, lars, gglasso, covTest`).

4.1 Simulations

We investigate the finite sample properties of the covariance test, the TG test, the spacing test, and the simplified spacing test via the simulations where we consider various models. As all tests except for the TG test (where we fix λ) are designed for the LARS algorithm we compare the results in different steps of the algorithm. To be clear, when testing the significance of some variable in some stage of LARS the variable tested in the spacing test (that is, the variable entering the non-zero set at the current step) is also tested in the TG test.

In each simulation we describe the model used for the calculation. For each model scenario we simulated 1000 independent data samples, and the random selection was performed for each data-set. Thus, we obtained the complete lasso paths, and for each path we applied the post-selection inference tools. Every time we simulated a linear regression model

$$y_i = \alpha + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n,$$

where we set $\alpha = 0.5$ and $\sigma = 1.35$. In different models scenario we considered different values of $p \in \mathbb{N}$ and various proportions of the non-zero parameters. When $n > p$ we estimate the parameter σ using the standard formula $\sqrt{MSE}$. Every test was performed on the significance level 5%.

4.1.1 All zero coefficients

Firstly, we simulate a model with $n = 40$ observations, $p = 4$ predictors where all coefficients are equal to zero. The true model has, therefore, the form

$$y_i = \alpha + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n. \tag{4.1}$$

We examine how many times the null hypothesis is rejected in each test from the total number of 1000 simulations (see Table 4.1). As the level of tests was

	Space (%)	mSpace(%)	Cov(%)	TG(%)
H_0 rejected	57 (5.7%)	57 (5.7%)	83 (8.3%)	57 (5.7%)
β_1	16 (28.1%)	16 (28.1%)	22 (26.5%)	16 (28.1%)
β_2	11 (19.3%)	11 (19.3%)	19 (22.9%)	11 (19.3%)
β_3	13 (22.8%)	13 (22.8%)	14 (16.9%)	13 (22.8%)
β_4	17 (29.8%)	17 (29.8%)	28 (33.7%)	17 (29.8%)

Table 4.1: *The estimated level of confidence and the proportion of each coefficient falsely entering the model. All coefficients were set to zero in the true model, $n = 40$, $p = 4$. The first line shows the proportions of cases when the null hypothesis was rejected at the first step of LARS in total. Further lines show the proportion of each coefficient contributing to the false rejection rate in the first line. These proportions of passing to the second step were distributed relatively equally among all four coefficients.*

set to 5 %, we expect that for the first predictor to enter the active set the null should be rejected approximately 50 times. However, as we already mentioned, the selection process is random, therefore, the first predictor (parameter) entering the model can be different across different repetitions. On the other hand, the proportions of the parameters falsely entering the model should be roughly similar.

The null hypothesis at the first step of the LARS algorithm is now same for all tests. We can interpret it that the first predictor entering the active set is zero, which actually means that all predictors are zero.

For the first predictor to enter the active set we rejected 57 times the null hypothesis in the case of the TG test and the spacing test. For the covariance test the result was somehow worse, as the null was rejected in 83 cases. The results for the spacing test and the simplified version were almost identical. We see that in all cases the results showed the rejection of the null in more than 5%. This coincides with the fact that the lasso tends to choose more non-zero coefficients into the model.

In Figure 4.1 we simulate the null distribution of the p-values for the first step by comparing the calculated p-values with theoretical quantiles of the standard uniform distribution. As stated in Section 3.3.2, the p-values of the TG test and the spacing test followed nicely the uniform distribution. In the picture the values for the TG test are hardly visible as they are covered by almost same values of the spacing test.

4.1.2 One and two nonzero coefficients

Power of tests

In the following we investigate the power of tests. For the model we use again $n = 40$ and $p = 4$, but this time we consider one parameter to be nonzero. We vary the size of the coefficient across different values from 0 to 2, and for each parameter value we perform again 1000 repetitions. We can write the true model in the form

$$y_i = \alpha + \beta_1 x_{i1} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n, \tag{4.2}$$

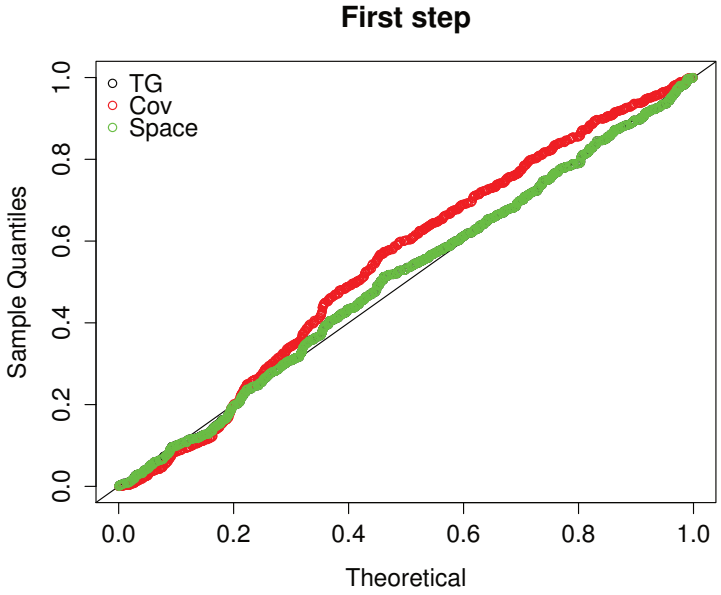


Figure 4.1: *Quantile-quantile plot (further noted as QQ-plot) of the tests for the first step of LARS. All true coefficients were set to zero.*

with β_1-values $0,\ 0.2,\ 0.4, \ldots,\ 1.8,\ 2$.

Later, besides the varying coefficient above, we also include another stronger coefficient $\beta_2 = 5$, and we measure the power of tests at the second step of LARS. Therefore, the estimated power in the latter is the portion of cases where the null is rejected at the second stage from those that passed the first step.

Figure 4.2 shows the estimated power-curves for the spacing, covariance and the TG test for both set-ups. For the modified spacing test the results were almost the same as for the spacing test, so we did not included the latter into the picture. We estimated the power of tests by the number of rejected null hypotheses out of the 1000 simulations. Firstly, we tested the entering variable at the first step of LARS, and secondly we tested the second step. We observed the p-values just for the grid of values $0,\ 0.2,\ 0.4, \ldots,\ 1.8,\ 2$, for others we did a linear approximation.

From the comparison we saw that the lines of the spacing and the covariance test were similar, but the line for the TG test lied under both of them in each case. For the value $\beta_1 = 1.4$ all tests rejected the null in more than 99% of repetitions in the first case. In the second set-up (when we measured the power at the second stage), for the same β_1, the covariance and the spacing test rejected the null in more than 99% of the cases, and the TG test in 97%. In this case we measured the highest power for the covariance test.

Similar behavior of the covariance test and the spacing test can be explain by the asymptotic equivalence between them. The reason why we measured for the TG test the lowest power is probably that in the spacing test some of the polyhedral inequalities are removed from the original LARS selection event, therefore, we condition on less than in the case of the TG test.

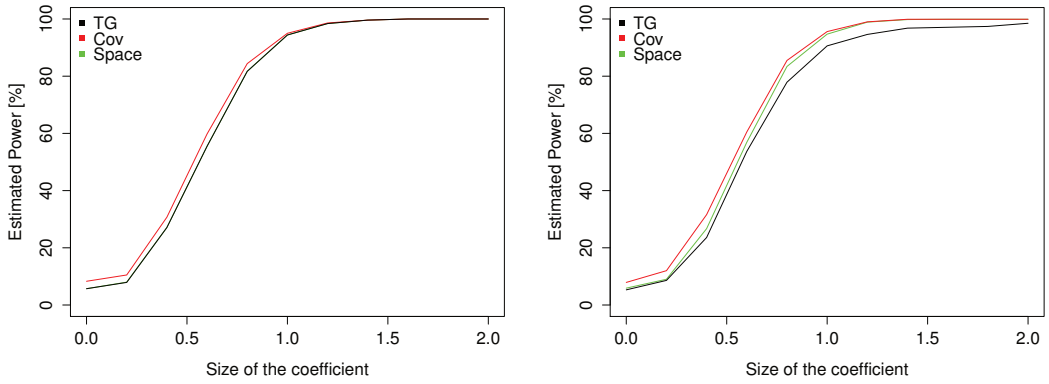


Figure 4.2: *Estimated power curves for the spacing test (green line), the covariance test (red line), and the TG test (black line). The left picture shows the set-up for the first predictor to enter the active set when others were set to zero, the right picture for the second, entering after the stronger predictor.*

Two nonzero coefficients

In the next set-up we again work with the low dimensional case $n = 40$, $p = 4$ and set two coefficients $\beta_1 = 1.5$, $\beta_2 = -1.5$ and the other two to zero. Table 4.2 summarizes the rejected null hypotheses for all tests at the first three steps of LARS.

At the first and the second step there was no case where we rejected the significance of the zero predictors. We counted the rejected hypotheses conditionally, therefore, if we did not reject the null for the nonzero coefficients at a certain step of LARS, we did not consider the simulation in the further steps.

We observed again the similar behavior of both, the spacing and the covariance test, at the first and the second step of the LARS algorithm. The TG test worked worse at the second step but better at the third step, when other tests rejected the null more times. Therefore, the TG test was the closest to the significance level of 5%. The spacing test and the modified version showed again almost the same numbers.

At the first two steps all tests were choosing only the non-zero coefficients β_1 and β_2. From the total number of 1000 simulations the spacing and the covariance test chose 786 times the true model, the TG test only 765 times.

For each predictor entering the active set at the steps 1-4 we constructed also QQ-plots of the p-values (Figure 4.3), again without the modified spacing test. This time the results are shown unconditionally, all 1000 simulations are plotted in every graph. At the first stage each test behaved again same as the other two. Because for the third and the fourth step the null hypothesis holded we observed that points of the spacing test and the TG test followed correctly the diagonal. The red points of the covariance test did not lie on the diagonal because the standard uniform distribution is not appropriate for the p-values of the covariance test under the null hypothesis. At the second step all three tests showed again a very high power.

Space	Step 1	Step 2	Step 3
	β_1: 399	β_2: 397 (99.5%)	β_j: 27 (6.8%)
	β_2: 392	β_1: 389 (99.2%)	β_j: 28 (7.2%)
H_0 rejected	791	786 (99.4%)	55 (7.0%)
H_0 not rejected	209	5 (0.6%)	731 (93.0%)

mSpace	Step 1	Step 2	Step 3
	β_1: 399	β_2: 397 (99.5%)	β_j: 27 (6.8%)
	β_2: 392	β_1: 389 (99.2%)	β_j: 27 (6.9%)
H_0 rejected	791	786 (99.4%)	54 (6.9%)
H_0 not rejected	209	5 (0.6%)	732 (93.0%)

Cov	Step 1	Step 2	Step 3
	β_1: 399	β_2: 397 (99.5%)	β_j: 30 (7.6%)
	β_2: 391	β_1: 389 (99.2%)	β_j: 43 (11.1%)
H_0 rejected	790	786 (99.5%)	73 (9.3%)
H_0 not rejected	210	4 (0.5%)	713 (90.7%)

TG	Step 1	Step 2	Step 3
	β_1: 399	β_2: 387 (97.0%)	β_j: 20 (5.2%)
	β_2: 392	β_1: 378 (96.4%)	β_j: 24 (6.3%)
	β_j: 0	-	-
H_0 rejected	791	765 (96.8%)	44 (5.8%)
H_0 not rejected	209	26 (3.2%)	721 (94.2%)

Table 4.2: *How many times the null hypothesis was rejected at the first three steps of LARS. Two coefficients β_1, β_2 were set to non-zero, $n = 40$, $p = 4$. The first line of each test shows the case when β_1 entered the active set at the first step and β_2 at the second step. The second line shows the opposite order. At the second and the third step we considered only the simulations where the null was rejected at the previous step. The percentages of these simulations are shown in brackets. No other coefficient than β_1 or β_2 was chosen at the steps 1 and 2.*

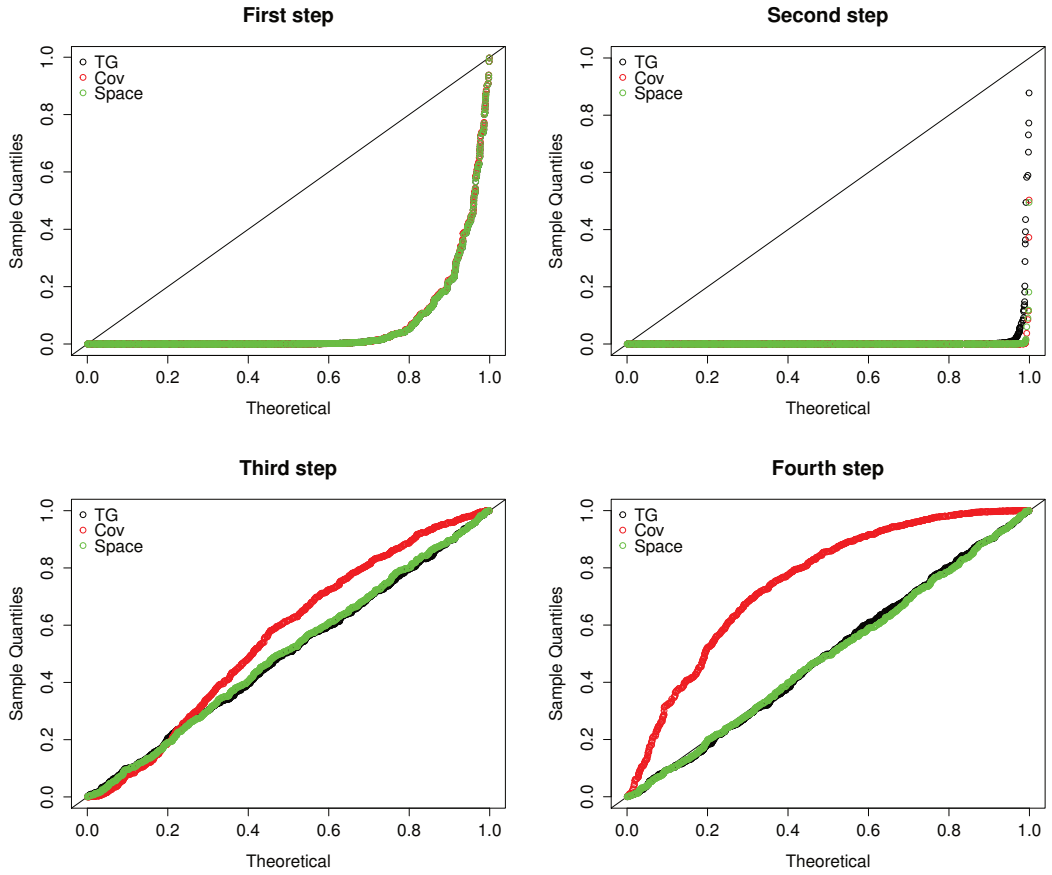


Figure 4.3: *QQ-plots for the post-selection p-values of the predictors entering the active set at certain steps of LARS. Two predictors out of four were truly nonzero,* $n = 40$.

4.1.3 Four nonzero coefficients, high-dimension

Next we simulate the linear regression model with $n = 40$ and $p = 80$. The parameter σ is again set to $\sigma = 1.35$. Among all coefficients only four of them are set to non-zero: $\beta_1 = 2,\ \beta_2 = -2,\ \beta_3 = 1.5,\ \beta_4 = -1.5$.

We cannot use the standard estimate for σ in this set-up because $p > n$. Instead, we use the function `estimateSigma` from the `selectiveInference` package of `R`. The parameter σ is, therefore, estimated in the following way. Firstly, the lasso estimates are found with λ selected by the cross-validation (the method is described, for example, in Hastie et al., 2008). Then σ is estimated by

$$\hat{\sigma} = \sqrt{\frac{1}{n-m-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where $\hat{y}_i$ are the predicted values using the cross-validated λ, and m is the number of non-zero coefficients in the lasso fit. The basic descriptive statistics about the estimate of σ are summarized in Table 4.3.

This time our goal is to measure how many times each test "chose" the right model. By choose we mean that the null hypothesis was rejected for the truly non-zero coefficients $\beta_1,\ \beta_2,\ \beta_3,\ \beta_4$ at least at the first four steps, and it was not rejected at further steps. Because we consider only LARS without the lasso

Min	1st Qu.	Median	Mean	3rd Qu.	Max	St.dev.
0.403	1.081	1.314	1.292	1.541	2.238	0.340

Table 4.3: *Descriptive statistics for the estimated parameter σ based on 1000 simulations: minimum, the empirical 25% and 75% quantiles, median, mean, maximum, and the standard deviation.*

condition (because of the comparison with the spacing test) we need to discard the cases when variables are changing signs during the steps.

The results are summarized in Table 4.4. Because for the spacing and the modified spacing test the results were same the table shows values just for the first one. We note that we did not consider any corrections of the p-values.

Among all tests the most successful test, which chose all the non-zero variables in the first four steps of LARS and stopped at the fifth step, was the spacing test. The covariance test behaved again similarly. For $n = 40$ the tests did not performed very well as they all chose in more than 50% a model with less than four variables. The lasso is known that it overfits the model, which means, it selects more variables than the true model contains (this is related to the fact that the lasso does not have the oracle properties). For such a small number of predictors this property was not visible. Therefore, we repeated the same simulation but this time with $n = 400$ (see the second part of Table 4.4). For this set-up all the post-selection tests chose in more than 65% the true model, and there was also a small portion of the cases when they selected a model with five and six variables, which supported the overfitting property.

Graphical results are drawn in QQ-plots (see Figure 4.4) for the first six steps of LARS. The subsequent steps are again plotted unconditionally, every time all 1000 points are plotted in each graph. We have to keep in mind that in the steps $2, 3, \ldots$ in the covariance test we are testing whether all other coefficients not included in the active set are zero, and in the other two tests we test whether the coefficient of the predictor entering the active set at the current step is zero. At the fifth and the sixth stage we can see the difference between the behavior of the TG test and the spacing test. Although, both tests are exact, the p-values of the TG test were closer to the diagonal, i.e., closer to the uniform distribution.

4.2 Simulations for the group lasso test

In this part we briefly examined the properties of the proposed test statistic in Section 3.4. Our aim is to simulate the null distribution of the test statistic (3.17). For this purpose we define a model with $n = 50$ observations, $p = 12$, and $L = 4$:

$$\boldsymbol{y} = \sum_{j=1}^{4} \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n),$$

where $\mathbf{X}_j \in \mathbb{R}^{(50\times 3)}$ and $\boldsymbol{\beta}_j \in \mathbb{R}^3$, for $j = 1, \ldots, 4$. Therefore we have four groups, each consisting of the three predictors. In each of the group all three parameters are set to zero. Therefore, the true model is the same as (4.1) except for n and p.

For our purpose we test only the first group of parameters, that is, the group that becomes non-zero as the first. We calculate further the test statistic (3.17),

	$n = 40$			$n = 400$		
Variables	Space	Cov	TG	Space	Cov	TG
1, 2, 3, 4 ||	28.9%	28.6%	4.7%	83.6%	82.7%	65.8%
1, 2, 3, 4, # ||	4.9%	5.2%	0.4%	3.4%	4.3%	3.1%
1, 2, 3, 4, ## ||	1.0%	1.0%	-	0.1%	0.1%	0.3%
1, 2, 3, 4, ### ||	0.4%	0.4%	-	-	-	-
1, 2, 3, 4, #### ||	0.1%	-	-	-	-	-
Less than four variables	64.6%	64.7%	94.9%	12.9%	12.9%	30.8%

Table 4.4: *The proportion of the simulations when the true model was chosen. The left part shows the high-dimensional case with 80 predictors. The right part shows the same set-up but with 400 observations. Four coefficients were truly non-zero. The table shows the cases when the null hypotheses was rejected for the non-zero variables 1, 2, 3, 4 at initial steps and was not rejected at the next step for each test. The first line expresses the proportion of cases when the non-zero variables (regardless of the order) were chosen during the first four steps, and the null was not rejected at the fifth step. The second line shows the simulations when the tests chose the non-zero variables plus one another during the first five steps, and the null was rejected at the sixth step. Similarly the third, the fourth, and the fifth line. The last line says how many times the tests chose less than four variables. It corresponds to the cases when the true model was not selected.*

which has now a simple form as the second term in the numerator of the formula is zero when testing the first group:

$$T_1^g = \frac{\left\langle \boldsymbol{y}, \sum_{j=1}^{4} \mathbf{X}_j \widehat{\boldsymbol{\beta}}_j(\lambda_2) \right\rangle}{\sigma^2}.$$

The parameter λ_2 denotes the point (knot) where the second group becomes non-zero.

We performed again 1000 simulations and examined, subsequently, the null distribution via the QQ-plots and the histograms (see Figure 4.5). The null distribution did not really correspond to the $Exp(1)$ distribution. The empirical quantiles behaved similarly as the exponential distribution but with a different parameter.

This thesis is mainly focused on the classical lasso, and we did not further examined the proposed test statistic for the group post-selection test. We leave this topic for other work.

4.3 Real data example

Finally, we compare the post-selection tests on a real data set. The data shows a breast cancer annual mortality from Castellón province in Spain between the years 1980 and 2007. Our aim is to analyze the number of changepoints of the trend in the data. Further, we explain the problem more in detail, and we show how the lasso can be applied for this kind of a problem.

As there were at least two important events during the observed period that had an impact on the trend, we expect some points in the data where the trend

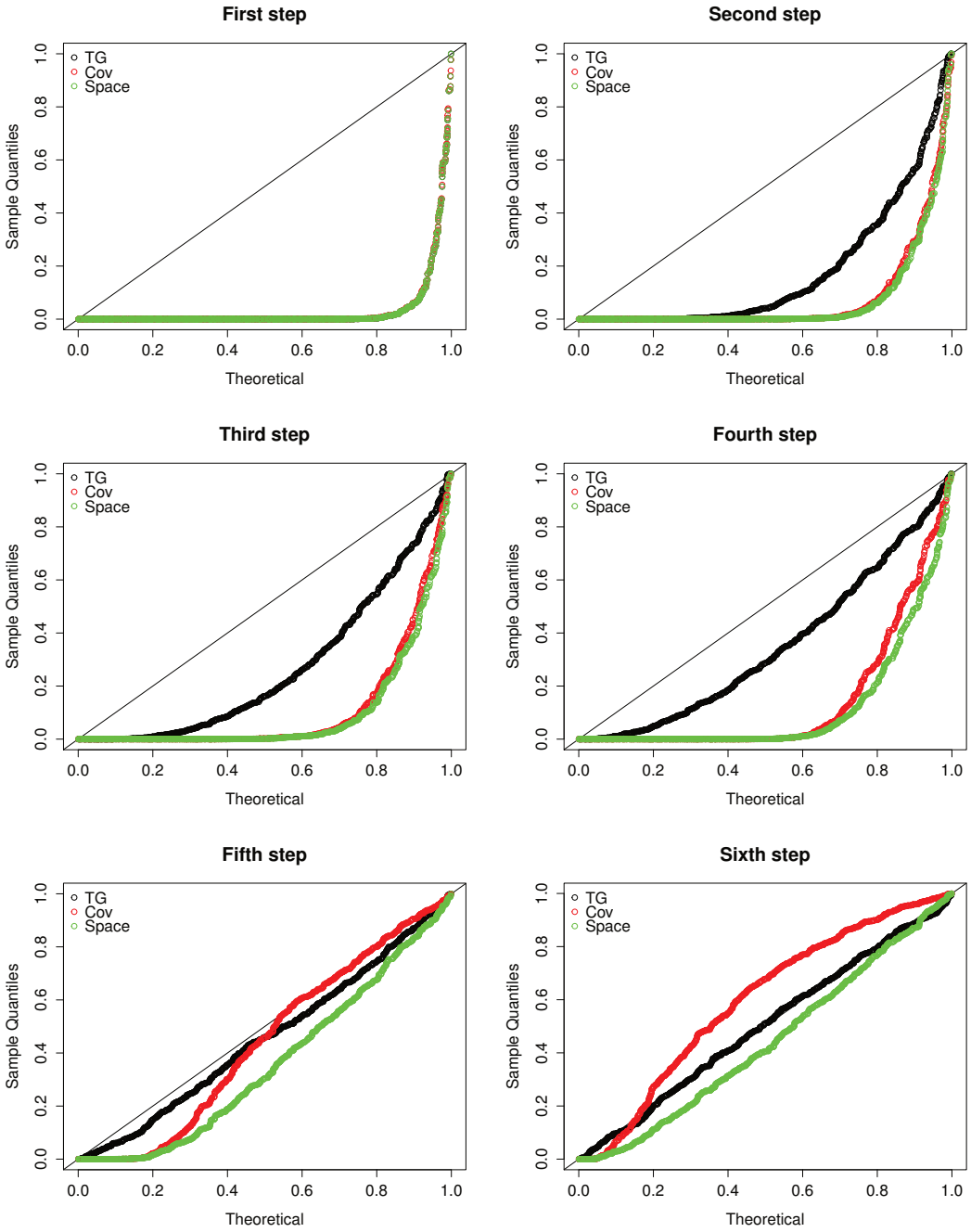


Figure 4.4: *QQ-plots for the post-selection p-values of the predictors entering the active set at steps of LARS. Four predictors were truly nonzero,* $n = 40, p = 80$.

is changing (so called *changepoints* or *joinpoints*). The problem of finding these points is the topic of the *joinpoint regression* (see Figure 4.6 for an example of one joinpoint in the mortality data).

Here we work with a special case of the joinpoint regression when the possible changepoints are represented by the observations, that is, the years. Let us denote the observed years as $x_1 < x_2 < \cdots < x_n$. Therefore, we have $x_1 = 1980$ and $x_n = 2007$. Under the condition that the possible changepoints are the years, we can formulate our problem as a minimization problem

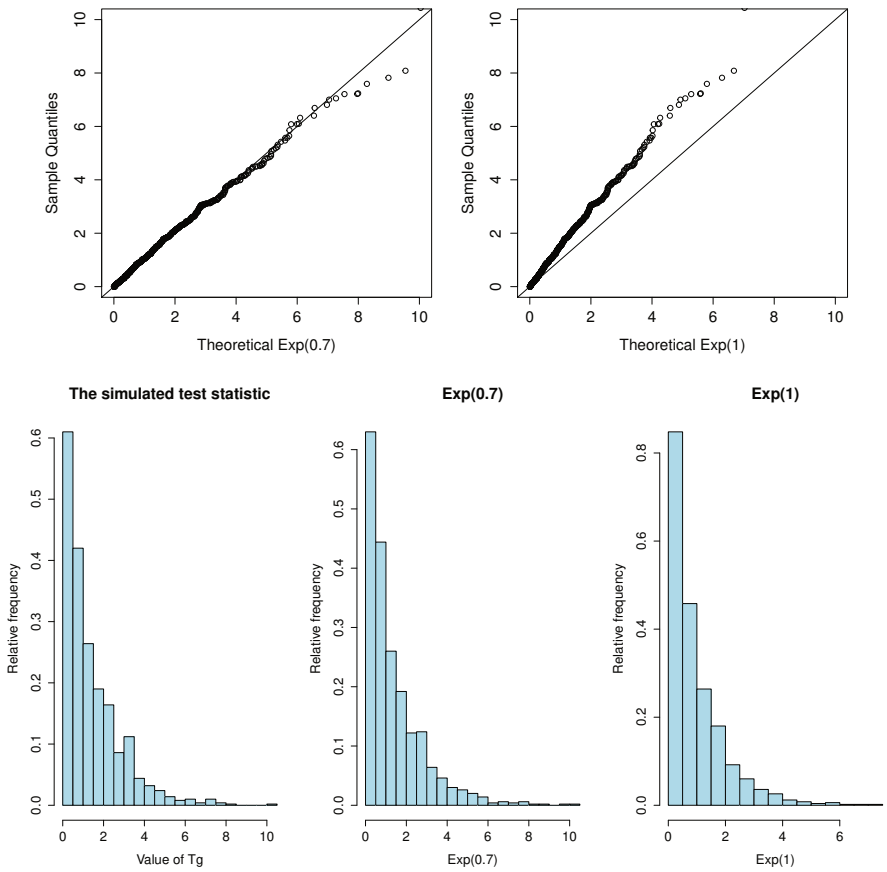


Figure 4.5: *QQ-plots and histograms for the simulated test statistics under the null distribution compared to the exponential distribution with the rate parameter 1 and 0.7. The graphs show values for the first group of predictors chosen by the group lasso in the model where all true coefficients were set to zero.*

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{n+2}}{\arg\min} \frac{1}{2} \left\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\right\|_2^2 + \lambda \left\|\boldsymbol{\beta}_{(2)}\right\|_1, \tag{4.3}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is the response vector, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{n+2})^\top$ is the vector of parameters to be estimated, and $\boldsymbol{\beta}_{(2)}$ is $\boldsymbol{\beta}$ without the first two parameters: $\boldsymbol{\beta}_{(2)} = (\beta_2, \ldots, \beta_{n+2})^\top$. The design matrix $\mathbf{X}$ is in this case defined as follows:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & 0 & 0 & \ldots & 0 \\ 1 & x_2 & (x_2 - x_1)^+ & 0 & \ldots & 0 \\ 1 & x_3 & (x_3 - x_1)^+ & (x_3 - x_2)^+ & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & x_n & (x_n - x_1)^+ & (x_n - x_2)^+ & \ldots & (x_n - x_n)^+ \end{pmatrix}, \tag{4.4}$$

where $(x)^+ = \max(x, 0)$ denotes the positive part of x. Last n columns of the design matrix $\mathbf{X}$ represent all years when the slope could be changed. For the next lines we denoted the first two column of $\mathbf{X}$ as $\mathbf{X}_{(1)}$ and the remaining columns as $\mathbf{X}_{(2)}$.

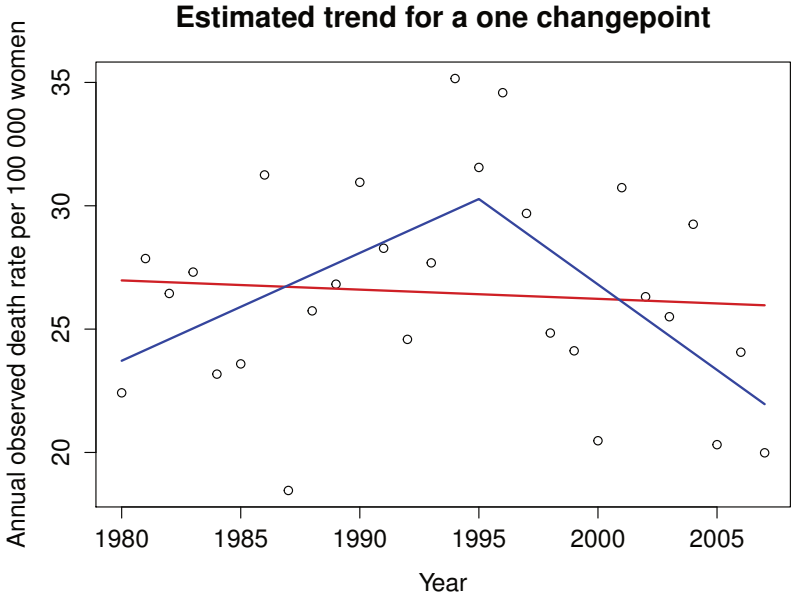


Figure 4.6: *Mortality trend in the data with one changepoint. The red line shows the estimated regression line and the blue line represents the estimated trend with one changepoint.*

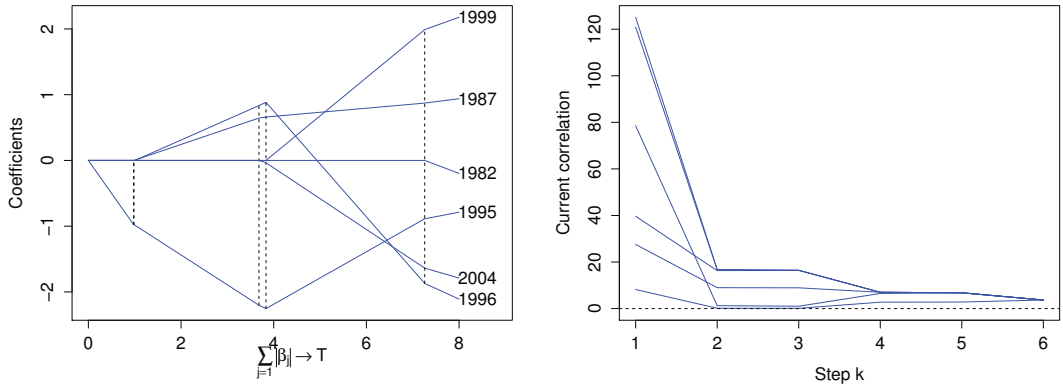


Figure 4.7: *Application of LARS to the changepoint-model of the breast cancer data. The resulting path (left) and the decreasing covariances (right) show the entrance of the first six predictors to the active set of the non-zero coefficients.*

The problem (4.3) now looks almost as some lasso problem, the only difference is that in the penalty term we have parameters from $\boldsymbol{\beta}_{(2)}$. The terms β_0 and β_1 correspond to the intercept and the slope in the usual linear regression model. The remaining part $\boldsymbol{\beta}_{(2)}$ and its non-zero elements correspond to the changepoints. In order to use the lasso for finding the non-zero terms of $\boldsymbol{\beta}_{(2)}$ we define a hat matrix

$$\mathbf{H} = \mathbf{X}_{(1)} \left(\mathbf{X}_{(1)}^\top \mathbf{X}_{(1)} \right)^{-1} \mathbf{X}_{(1)}^\top,$$

and we set $\tilde{\boldsymbol{y}} = (\boldsymbol{I}_n - \mathbf{H})\boldsymbol{y}$, and $\tilde{\mathbf{X}} = (\boldsymbol{I} - \mathbf{H})\mathbf{X}_{(2)}$. The same solution to the

Step	Var	Coef	Low TG	Up TG	TG	Space	Cov
1	1995	-1.130	-1.825	-0.417	0.002	0.002	<0.000
2	1996	-0.392	14.444	Inf	0.982	0.997	1.000
3	1987	1.124	-5.382	Inf	0.084	0.018	0.307
4	2004	-1.092	-30.225	Inf	0.860	0.965	0.868

Table 4.5: *The results of the tests in the four initial steps of LARS. The column "Var" shows which year was being tested. The third column presents the estimated coefficients, and the columns "Low TG" and "Up TG" are lower and upper boundaries of the 95% confidence interval calculated by the TG test. The last three columns represent the calculated p-values.*

vector $\boldsymbol{\beta}_{(2)}$ in (4.3) can be obtained by solving the following problem

$$\widehat{\boldsymbol{\beta}}_{(2)}(\lambda) = \underset{\boldsymbol{\beta}_{(2)} \in \mathbb{R}^n}{\arg\min} \frac{1}{2} \left\| \tilde{\boldsymbol{y}} - \tilde{\mathbf{X}} \boldsymbol{\beta}_{(2)} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta}_{(2)} \right\|_1, \tag{4.5}$$

which is now the standard lasso problem. For further details and the proof see (Maciak and Mizera, 2016).

Interpretation of the parameters in $\boldsymbol{\beta}_{(2)}$ in (4.5) is the same as in (4.3). Overall, we have $p = n = 28$ parameters to be estimated. Our approach for the usage of the post-selection tests is following. Similarly as during the simulations, we solve the introduced lasso problem (4.5) using the LARS algorithm, and in each step we test whether the entering parameter (i.e., the change in the slope) is significant by the three post-selection tests (we use just the spacing test and not the modification). As the number of testing parameters is large we estimate σ by the same approach as in Section 4.1.3.

The LARS algorithm chose 1995 to be the first changing year (see Figure 4.7). The points 1996 and 1987 entered the active set almost simultaneously. The first decrease in the current correlation was large compared to the rest of the steps, which also supports the significance of the changepoint around 1995. When we tested the significance of the entering variables in LARS we saw that all three tests chose only the first entering parameter (the year 1995) to be significant as the null hypothesis was rejected at the second step every time (see Table 4.5 for the first four stages).

The spacing test, however, rejected the null in the third step for 1987. Because 1996 and 1987 entered almost in the same step, and 1996 was not significant, it can indicate that 1996 is a correction of 1995. The relevant changepoint can lie, for example, at the end of the year 1995. We have only annual data, which is somehow limited for the changepoint location. For this purpose we tried to extend the design matrix $\mathbf{X}$ also for changes in the points 1980.5, 1981.5, 1982.5, etc., which yielded in $p = 55 > n$ parameters to be estimated. The results were, however, similar to the original problem. The first three coefficients that became nonzero were related to 1995, 1995.5, and 1987 (in that order). All tests rejected the null for the first predictor and did not reject for the second one. This approach yielded also one changepoint in the data. By using the TG test and the spacing test we chose the same model as with the covariance test, which was used in (Maciak and Mizera, 2016).

Joinpoints in the data are usually being founded by another algorithms, for example, in two stages. In the first stage the data are explored whether it contains any changepoint. At the second step the trend is estimated in each part of the data. By using the lasso and the post-selection tests we found changepoints and estimated the related slopes at once.

Conclusion

In the thesis we examined the post-selection tests that can be used for the lasso estimates and their inference. The post-selection inference is relatively new topic, and the statistical properties of such tests are still under an ongoing investigation.

Firstly, we introduced the lasso method, we presented some basic properties, and we discussed the computational aspects. As the minimization problem for finding the lasso estimates is convex, there are a lot of ways for solving it. The efficient way is, for instance, the LARS algorithm.

We dedicated one short chapter to the group lasso, which is a lasso modification for grouped variable problems. Further, we briefly mentioned another two modifications, the adaptive lasso and the elastic net.

The classical statistical inference is not suitable for the lasso estimates as it cannot deal with the randomness in the selection of the variables. The lasso is an adaptive procedure, which means, the variables are chosen randomly according to the data. In the main part of the thesis we gave more information about the appropriate (so called post-selection) inference for the lasso. We introduced and described the following post-selection tests: the covariance test, the TG test, and the spacing test. We also proposed a test statistic that could be used for the group lasso estimates.

The post-selection tests were compared in various simulations where we examined statistical properties, such as the power of tests, the level of confidence, the null distribution, and other abilities. The tests were applied on the steps of the LARS algorithm. We used various set-ups, including the high-dimensional case ($p > n$). We also simulated the null distribution of the proposed test statistic for the group lasso estimates.

The tests for the lasso were further used for the breast cancer data set from Spain where we were finding changepoints in the trend development over time. The problem of the joinpoint regression can be formulated via the lasso minimization problem where parameters to be estimated represent the points (years) when the possible change in the trend can occur. By this example we successfully showed that the post-selection tests for the lasso estimates can be applied also in the practical situation.

Bibliography

Berk R., Brown L., Buja A., Zhang K., and Zhao L. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

Dezeure R., Bühlmann P., Meier L., and Meinshausen N. High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical Science*, 30(4):533–558, 2015.

Efron B. and Tibshirani R. *An Introduction to the Bootstrap*. Chapman & Hall, London, 1993.

Efron B., Hastie T., Johnstone I., and Tibshirani R. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Fan J. and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, 96(456):1348–1360, 2001.

Friedman J., Hastie T., Höfling H., and Tibshirani R. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

Fu J. W. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

Hastie T., Tibshirani R., and Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edition. Springer, Stanford, California, 2008. ISBN 978-0-387-84857-0.

Javanmard A. and Montanari A. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Lear. Res.*, 15:2869–2909, 2014.

Knight K. and Fu W. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378, 2000.

Lee J., Sun D., Sun Y., and Taylor J. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Lockhart R., Taylor J., Tibshirani J. R., and Tibshirani R. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

Maciak M. and Mizera I. Regularization techniques in joinpoint regression. *Stat Papers*, 57:939–955, 2016.

Meinhausen N. and Bühlmann P. Stability selection. *Journal of the Royal Statistical Society Series B*, 72(4):417–473, 2010.

Reid S., Friedman J., and Tibshirani R. A study of error variance estimation in lasso regression. *arXiv: 1311.5274*, 2014.

Tibshirani J. R. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

Tibshirani J. R., Taylor J., Lockhart R., and Tibshirani R. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–602, 2016.

Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

Yuan M. and Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(Part 1):49–67, 2006.

Zhao P. and Yu B. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

Zou H. The adaptive lasso and its oracle properties. *Journal of American Statistical Association*, 101(476):1418–1429, 2006.

Zou H. and Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.

List of Figures

1.1 *The proportional shrinkage: the ridge regression and the soft-thresholding of the lasso.* 6
1.2 *Illustration of the geometry of the lasso vs. the ridge regression.* 7
1.3 *The LARS algorithm for two covariates.* 8

3.1 *An example of the lasso path.* 18

4.1 *Quantile-quantile plot of the tests for the first step of LARS.* 29
4.2 *Estimated power curves.* 30
4.3 *QQ-plots for the post-selection p-values.* 32
4.4 *QQ-plots for the post-selection p-values.* 35
4.5 *QQ-plots and histograms for the simulated test statistics under the null distribution.* 36
4.6 *Mortality trend in the data with one changepoint.* 37
4.7 *Application of LARS to the changepoint-model of the breast cancer data.* 37

List of Tables

4.1 *The estimated level of confidence and the proportion of each coefficient.* 28
4.2 *How many times the null hypothesis was rejected at the first three steps of LARS.* 31
4.3 *Descriptive statistics for the estimated parameter* σ. 33
4.4 *The proportion of the simulations when the true model was chosen for the high-dimensional set-up.* 34
4.5 *The results of the tests in the four initial steps of LARS.* 38