WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

MASTER THESIS

On variable selection of the Elastic Net in presence of multicollinearity

under supervision of

Univ.Prof. Dipl.-Ing. Dr.techn. Kurt Hornik

submitted by

Dipl.-Ing. Hanna Sophia Wutte

Jacquingasse 16/51
1030 Wien

Vienna, July 31, 2018

WIRTSCHAFTSUNIVERSITÄT WIEN
Vienna University of Economics and Business

Master Thesis

Title of Master Thesis:	
Author (last name, first name):	
Student ID number:	
Degree program:	
Examiner (degree, first name, last name):	

I hereby declare that:

1. I have written this Master thesis myself, independently and without the aid of unfair or unauthorized resources. Whenever content has been taken directly or indirectly from other sources, this has been indicated and the source referenced.

2. This Master Thesis has not been previously presented as an examination paper in this or any other form in Austria or abroad.

3. This Master Thesis is identical with the thesis assessed by the examiner.

4. (only applicable if the thesis was written by more than one author): this Master thesis was written together with

   The individual contributions of each writer as well as the co-written passages have been indicated.

Date

Signature

Abstract

We analyse the variable selection properties of the Elastic Net (EN) penalisation method in the framework of generalized linear models (GLMs) in presence of multicollinearity. Existing literature on variable selection properties of $\ell_1 + \ell_2$-penalized GLMs such as model selection consistency or the grouping effect is summarized. Inspired by the latter, for linear models in particular, the joint selection of correlated variables is analysed in more detail. Specific focus is laid on settings involving highly correlated regressors, where correct variable selection of the EN often breaks down. As a possible remedy, we discuss the concept of Stability Selection (cp. Meinshausen and Bühlmann (2010)). In form of a simulation study, we investigate, whether Stability Selection leads to improved variable selection for an EN-penalized logistic regression.

Contents

Introduction **1**

1 Variable selection in linear models **3**
1.1 Model selection consistency 4
1.2 Variable selection in finite samples 9
1.3 The Lasso and Elastic Net on correlated variables 14
1.3.1 Grouping Effect 14
1.3.2 Least Angle Regression 23

2 Stability Selection **28**
2.1 The method 28
2.2 Choice of regularisation and error control 29
2.3 Computational requirements and pointwise control 31
2.4 Consistent variable selection for the Lasso 31

3 Variable selection in GLMs **34**
3.1 Model selection consistency for logistic regression 35
3.2 A simulation study on logistic regression 40
3.2.1 Data generating process 40
3.2.2 Probability of success for the Elastic Net 40
3.2.3 Probability of success for the Elastic Net with Stability Selection 42

Appendix **45**

A Subgradients for convex optimisation **45**

Bibliography **47**

List of Figures **48**

Introduction

In predictive modelling it often is of interest to obtain interpretable approximations of some underlying data generating process. To this end, a first step one commonly resorts to is to conduct penalized regression analyses. These methods are usually easily fit and, for certain choices of penalisation, provide sparse solutions, i.e. models that depend on a reduced number of explanatory variables. In a "large-p"-setting, where an ever growing amount of (possibly noisy) data is available, such a dimension reduction to a great extend enhances interpretability of an approximating model as well as the actual process of interest.
Naturally, when relying on such penalisation methods, one would like to know how closely the approximated model resembles the true one. In particular, we like to assess if the chosen model identifies the variables that are truly relevant to the underlying data generating process. This leads to the question of correct variable selection, which, for generalized linear models amounts to asking whether the fitted model assigns non-zero coefficients to truly relevant variables, while estimated coefficients of superfluous regressors are set to zero.
In this thesis, we like to study specifically the Elastic Net penalisation methods variable selection properties in the framework of generalized linear models. Particular focus is laid on settings in which explanatory variables display some sort of correlation structure, i.e. if there is multicollinearity in the design.

A note on multicollinearity

We speak of multicollinearity when two or more explanatory variables in a regression model are highly linearly related, or in other words, when there is an approximate linear relationship among two or more of those variables. This phenomenon does not reduce the predictive power of the regression model (at least for data within the training set). However, in presence of multicollinearity, a fitted model might not give valid results about which predictors are redundant with respect to others. Hence, in particular correct variable selection may be impaired.
Situations in which some regressor can to a high degree be linearly predicted from others therefore are of specific interest when variable selection properties are discussed. Throughout this thesis, we will mostly think of settings in which the pairwise correlations among at least two of the regressors exceed a certain threshold. This includes scenarios with groups of highly correlated variables as well as other correlation structures of the design such as the well-known Toeplitz structure.

The setting and further notation

The *generalized linear model* will serve as basis for subsequent discussions and shall be introduced in the following. We consider $n \in \mathbb{N}$ i.i.d. observations of both a response variable Y_i and p predictors $X_{i,j}$, $j = 1, \ldots, p$, gathered in the design matrix $X \in \mathbb{R}^{n \times p}$. We denote by X_i and $X^{(j)}$

the i$^{\text{th}}$ row and j$^{\text{th}}$ column of X respectively. For the sake of simplicity, throughout this thesis we use capital letters for random variables as well as their realisations. The meaning of those notions will be clear from the context. We further assume that the variables $X^{(j)}$ are standardized such that $\sum_{i=1}^{n} X_{i,j}/n = 0$ and $\sum_{i=1}^{n} X_{i,j}^2/n = 1$, $\forall j \in \{1, \ldots, p\}$. The conditional distribution of Y_i given the values of the explanatory variables X_i is assumed to be a member of the exponential dispersion family. The corresponding density is of the form

$$f_Y(y|\theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right), \qquad \theta \in \mathbb{R}, \phi > 0.$$

We refer to θ and ϕ as the *canonical* and *dispersion parameters*; $c(\cdot), b(\cdot)$ are specified functions such that b satisfies

$$\mu_i \equiv \mathbb{E}[Y_i] = b'(\theta) \quad \text{and} \quad Var(Y_i) = b''(\theta)\phi,$$

for $b(\cdot)$ sufficiently differentiable. Provided that $b'(\cdot)$ is invertible, θ may be represented as a function of μ and hence the same holds true for $b''(\cdot)$. We write $b''(\theta) = V(\mu)$, where $V(\cdot)$ defines the *Variance function* of the family. The response is then linked to the explanatory variables using a smooth and invertible *link function* $g(\cdot)$ that transforms the expectation of the response variable μ_i to the so-called *linear predictor* $\eta_i := \beta_0 + \beta^T X_i$, $(\beta_0, \beta) \in \mathbb{R}^{p+1}$. In short we note:

$$\boxed{Y_i \overset{id}{\sim} f_Y, \quad g(\mu_i) = \beta_0 + \beta^T X_i, \quad \mu_i \equiv \mathbb{E}[Y_i] \qquad i = 1, \ldots, n.} \tag{1}$$

The model is assumed to be sparse, meaning that some explanatory variables are in fact irrelevant to the response variable. Let $I^* \subset \{1, \ldots, p\}$ be the indices corresponding to the true set of covariates. Then naturally $\beta_i = 0$, $i \notin I^*$. Furthermore, we denote by $k^* = |I^*|$ the number of non-zero entries in β.
As mentioned before, we like to put specific focus on the case where predictors are potentially highly correlated. In such a setting, we also allow for correlations among the true covariates indexed by I^*.

The rest of this thesis will be structured as follows. We start by discussing variable selection for linear models in Chapter 1. This entails an overview of model selection consistency as well as a detailed analysis of the Elastic Net on correlated variables including e.g. the grouping effect introduced in Zou and Hastie (2005). This analysis in particular shows, that the Elastic Net in linear models may fail to correctly select true variables and discard irrelevant ones if these regressors are highly correlated. As possible remedy to improve any selection methods stability Meinshausen and Bühlmann (2010) introduce Stability Selection. The method, that has been shown to drastically improve the selection properties of the Lasso in linear models for various settings is introduced in Chapter 2. Further results on Stability Selection as obtained from Meinshausen and Bühlmann (2010) are collected as well. An application of Stability Selection to a generalized linear model framework is given in chapter 3, where the selection properties of the EN-penalized logistic regression for settings of different correlation structures in the design are discussed by means of a simulation study. Moreover, existing findings on model selection consistency for Elastic Net-penalized GLMs, that reduce to findings for logistic regression only, are collected.

Chapter 1

Variable selection in linear models

There is a vast variety of work on theoretical aspects of penalisation methods for linear models, much of which focuses on their selection properties. In the subsequent sections, we intend to summarize the findings on variable selection in linear models for regularization methods involving ℓ_1 and $\ell_1 + \ell_2$ penalties i.e. for the popular Lasso and its generalisation the Elastic Net. In doing so, we will put particular emphasis on analysing in what way multicollinearity affects the estimated model.
We start by collecting results on model selection consistency for both penalization methods in Section 1.1. Thereafter, Section 1.2 discusses their finite sample selection properties. Finally, in Section 1.3, we further summarise and extensively analyse findings on the methods' selection properties considering (groups of) correlated variables. Below, the Lasso and Elastic Net estimators shall be briefly introduced.

The Setting
Throughout this chapter we assume our data is generated by a linear regression model. In consistency with the setting introduced in the previous chapter, such a model is obtained for $g(\cdot) = Id(\cdot)$ and assuming the conditional distribution of the response given the covariates to be Gaussian. Formally, for $\sigma^2 > 0$, this reads as

$$\boxed{Y_i \overset{id}{\sim} \mathcal{N}(\mu_i, \sigma^2), \quad \mu_i = \beta_0 + \beta^T X_i, \qquad i = 1, \dots, n.}$$

The Lasso Estimator
Amongst all penalisation techniques, the Lasso, introduced in Tibshirani (1996), has received most attention. It involves an ℓ_1-penalty, by the nature of which sparse approximations of the truly underlying model can be recovered. The Lasso estimate $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T \in \mathbb{R}^p$ is obtained as the solution to

$$\min_{(\beta_0, \beta)} \|Y - \beta_0 - X\beta\|_2^2 \quad \text{s.t. } \|(\beta_0, \beta)\|_1 \leq t, \tag{1.1}$$

with $t \geq 0$ determining the amount of regularisation. Note that the solution for β_0 is $\hat{\beta}_0 = \bar{Y}$, thus w.l.o.g. we set $\bar{Y} = 0$ and omit β_0 in the minimisation. In what follows we will however work with the Lagrangian form

$$\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1, \tag{1.2}$$

with tuning parameter $\lambda \geq 0$, which controls the shrinkage applied to the estimates. Typical

implementations of the Lasso determine the solution to (1.2) for a specified grid of values for λ. This results in a path of solutions among which the model with the smallest prediction error in terms of some measure of deviation such as the squared loss

$$SL(\hat{\beta}(\lambda)) = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}(\lambda)^T X_i\right)^2$$

is commonly selected as the final model. One often relies on k-fold cross validation for minimizing the squared error for each choice of λ.

The Elastic Net Estimator

The Elastic Net, introduced in Zou and Hastie (2005) and hereafter frequently referred to as EN, is a penalized least squares method using a more general penalty consisting of both ℓ_1 and ℓ_2 term. For non-negative λ_1, λ_2 the naive Elastic Net estimator $\hat{\beta}$ is the minimizer of

$$\min_{\beta} \|Y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2. \tag{1.3}$$

or equivalently

$$\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\left((1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2\right), \tag{1.4}$$

with $\alpha \in [0,1]$, $\lambda \geq 0$.

1.1 Model selection consistency

The Lasso and Elastic Net's ability to perform automatic variable selection undoubtedly adds to their popularity. However, when using one of these methods for such purposes, one would like to assess under which conditions and to what extent the estimated model represents the true one. As we will subsequently see, consistent model selection (in a sense to be clarified) can be guaranteed under a condition that depends on the covariates' correlations and which is not satisfied should those attain even moderately high values.
In view of studying the penalisation methods' selection properties, it is important to differ between the notions of consistency in terms of parameter estimation and model selection. Formally, an estimate $\hat{\beta}$ is said to be consistent if

$$\hat{\beta} \xrightarrow[n\to\infty]{p} \beta.$$

Consistent model selection however is said to be in place if

$$\mathbb{P}\left(\{i : \hat{\beta}_i \neq 0\} = \{i : \beta_i \neq 0\}\right) = \mathbb{P}\left(\hat{I} = I^*\right) \underset{n\to\infty}{\longrightarrow} 1.$$

Consistency w.r.t. parameter estimation does commonly not imply consistent model selection and vice versa. In general we like to see both consistencies fulfilled.

ℓ_1 Penalty

The question of how well the model resulting from Lasso relates to the true model has been the topic of discussion in inter alia Zhao and Yu (2006). The following section is aimed at collecting the most important results given in therein, whilst noting related findings on the side. In particular,

the setting in Zhao and Yu (2006) allows for the parameter β to change for growing sample size n. Throughout this thesis, we will not account for this dependency in the notation and write $\hat{\beta}$ instead of $\hat{\beta}^n$.

It was shown in Knight and Fu (2000) that for p and β fixed, i.e. independent of the sample size n, the Lasso estimate $\hat{\beta}$ is consistent provided that λ_n (which may depend on the sample size) grows less than linear and assuming regularity conditions

$$C^n := \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T \underset{n\to\infty}{\longrightarrow} C \quad \text{and} \quad \frac{1}{n}\max_{1\le i\le n} \langle X_i | X_i \rangle \underset{n\to\infty}{\longrightarrow} 0, \tag{1.5}$$

on the design matrix, with C non-negative definite. In other words, provided $\lambda_n = o(n)$ and (1.5) we have $\hat{\beta}(\lambda_n) \underset{n\to\infty}{\xrightarrow{p}} \beta$.

Interestingly, as Leng et al. (2006) point out, even in orthogonal designs and p fixed the Lasso estimate (nonetheless being consistent in terms of parameter estimation) does not consistently select the true model when the tuning parameter is chosen to maximize prediction accuracy.

For the purpose of further analysing the selection properties of the Lasso, the concept of *Sign Consistency* is introduced in (Zhao and Yu, 2006, Section 2). Sign consistency does not assume estimates to be consistent with respect to parameter estimation. Moreover, it is a stronger property than consistency w.r.t. model selection which merely requires the estimate to correctly identify the zero-entries; the regression parameters' signs need not be matched. It is argued in Zhao and Yu (2006) that an estimated model, that does not correctly identify the parameters' signs, can be misleading and hence barely qualify as correctly selected model. The reason for using sign consistency however, is said to be of technical nature. Below, we collect the terminology used in Zhao and Yu (2006).

Definition 1.1 (Equality in Sign) *An estimate $\hat{\beta}$ is* equal in sign *with the true model β, formally $\hat{\beta} =_s \beta$, if and only if*

$$sign(\hat{\beta}) = sign(\beta).$$

Here, the equation holds component-wise and

$$sign(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}.$$

Definition 1.2 (Strong Sign Consistency) *If there exists a function f of n, that does not depend on either Y_n or X_n, with $\lambda_n = f(n)$ such that*

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{\beta}(\lambda_n) =_s \beta\right) = 1,$$

we say the Lasso is strongly sign consistent.

Definition 1.3 (General Sign Consistency) *The Lasso is* general sign consistent *if*

$$\lim_{n\to\infty} \mathbb{P}\left(\exists \lambda > 0 : \hat{\beta}(\lambda) =_s \beta\right) = 1.$$

Strong sign consistency implies that consistent model selection can be achieved via the Lasso with a preselected, deterministic penalty λ_n. General sign consistency by contrast ensures that for a random realization there exists a correct amount of regularisation that selects the true model with

probability tending to one. While the former implies the latter, it is shown in Zhao and Yu (2006), that both consistencies are almost equivalent to a so-called *irrepresentable condition*, a notion that is almost necessary and sufficient for model selection consistency and will be defined in the following.

Hereafter, let $I^* = \{1, \dots, q\}$ for some $q < p$ i.e. $\beta_i \neq 0$ for $i = 1, \dots, q$ and $\beta_i = 0$ for $i = q+1, \dots, p$. We further write $\beta_{(1)} = (\beta_1, \dots, \beta_q)$, $\beta_{(2)} = (\beta_{q+1}, \dots, \beta_p)$ and denote by $X(1)$ and $X(2)$ the first q and last $p-q$ columns of X respectively. The correlation matrix C^n as defined in (1.5), can be expressed as

$$C^n = \begin{pmatrix} C^n_{11} & C^n_{12} \\ C^n_{21} & C^n_{22} \end{pmatrix}$$

with $C^n_{ij} = \frac{1}{n} X(i)^T X(j)$, $j = 1, 2$. In what follows, we assume that C^n_{11} is invertible.

Condition 1.4 (Strong Irrepresentable Condition) *There exists a constant $\eta > 0$ such that*

$$|C^n_{21} \left(C^n_{11}\right)^{-1} sign(\beta_{(1)})| \leq (1 - \eta) \cdot \mathbf{1}$$

where $\mathbf{1}$ is a $p-q$-dimensional vector of ones. The inequality is to be understood element-wise.

Condition 1.5 (Weak Irrepresentable Condition) *The inequality*

$$|C^n_{21} \left(C^n_{11}\right)^{-1} sign(\beta_{(1)})| \leq \mathbf{1},$$

with $\mathbf{1}$ a $p-q$-dimensional vector of ones, holds element-wise.

Remark 1.6 *Conditions 1.4 and 1.5 closely relate to a regularization constraint on the OLS-estimate obtained when regressing any of the superfluous covariates $X(2)$ on the relevant covariates $X(1)$. When the signs of β are unknown, for Condition 1.4 to hold true, any component of the regression estimates needs to be smaller than one. To see this note that in order for 1.4 to hold true for all possible signs, we need*

$$\left| \left(C^n_{21} \left(C^n_{11}\right)^{-1} sign(\beta(1)) \right)_i \right| \leq (1 - \eta), \quad \forall i \in \{1, \dots, p-q\}$$

for any value of $sign(\beta(1))$. Hence we demand for each $i = 1, \dots, p-q$

$$\sum_{j=1}^{q} \left| \left(C^n_{21} \left(C^n_{11}\right)^{-1} \right)_{i,j} \right| |sign(\beta(1)_j)| \leq \sum_{j=1}^{q} \left| \left(C^n_{21} \left(C^n_{11}\right)^{-1} \right)_{i,j} \right| \leq 1 - \eta,$$

meaning that the sums of the rows of $C^n_{21} \left(C^n_{11}\right)^{-1}$, or equivalently the columns of $\left(C^n_{21} \left(C^n_{11}\right)^{-1} \right)^T$ are bounded by $1 - \eta$. This however implies that

$$\left| \left(X(1)^T X(1)\right)^{-1} X(1)^T X(2)_{\cdot,i} \right| \leq (1 - \eta)\mathbf{1}, \quad \forall i \in \{1, \dots, p-q\}.$$

We now have everything at hand to state the main results obtained in Zhao and Yu (2006) (see proofs given therein), which relate Conditions 1.4 and 1.5 to strong and general sign consistency. We initially focus on a setting when p, q and β remain unchanged with growing sample size. In such a framework, the regularity Conditions 1.5 are assumed to be in place, whereby the covariance matrix C shall be positive definite. In the case of a random design, we require those conditions to hold

almost surely. For this setting, the following results on sufficiency and necessity of irrepresentable conditions 1.4 respectively 1.5 have been shown.

Theorem 1.7 *For q, p and β fixed, let regularity Conditions* (1.5) *hold true with C positive definite and assume 1.4 to be in place. Then the Lasso is strongly sign consistent. More precisely, $\forall \lambda_n$ such that $\frac{\lambda_n}{n} \underset{n\to\infty}{\longrightarrow} 0$ and $\frac{\lambda_n}{n^{\frac{1+c}{2}}} \underset{n\to\infty}{\longrightarrow} \infty$ for any $0 \leq c < 1$, we have*

$$\mathbb{P}\left(\hat{\beta}(\lambda_n) =_s \beta\right) = 1 - o\left(\exp\left(n^c\right)\right).$$

Proof. Cp. (Zhao and Yu, 2006, Theorem 1). □

Hence, the probability of the Lasso selecting the true model tends to 1 exponentially fast, provided Condition 1.4 holds true. Recall that by Knight and Fu (2000), for $\lambda_n = o(n)$ the Lasso is consistent w.r.t. parameter estimation as well. Thus, both consistencies are allowed for simultaneously under the strong irrepresentable condition when choosing the regularisation parameter accordingly.

Theorem 1.8 *For q, p and β fixed, let regularity Conditions* (1.5) *hold true with C positive definite and assume the Lasso is general sign consistent. Then there exists an $N \in \mathbb{N}$ such that the weak irrepresentable condition 1.5 is satisfied for all $n > N$.*

Proof. Cp. (Zhao and Yu, 2006, Theorem 2). □

Combining the above we obtain:

$$\text{strong IC} \implies \text{strong sign consistency} \implies \text{general sign consistency} \implies \text{weak IC}.$$

In case both the number of overall covariates p as well as the one of true variables q is allowed to grow with increasing sample size n, regularity conditions (1.5) are not sensible any more. Instead, we assume the following:

Assumption 1.9 *There exist constants $0 \leq c_1 < c_2 \leq 1$ and $M_i > 0, i = 1, \ldots, 3$ such that*

$$\frac{1}{n}\langle X^{(i)} | X^{(i)} \rangle \leq M_1, \quad \forall i = 1, \ldots p, \tag{1.6}$$

$$\alpha^T C_{11}^n \alpha \geq M_2, \quad \forall \|\alpha\|_2 = 1, \tag{1.7}$$

$$q = \mathcal{O}(n^{c_1}), \tag{1.8}$$

$$n^{\frac{1-c_2}{2}} \min_{i=1,\ldots,q} |\beta_i| \geq M_3. \tag{1.9}$$

By normalizing the covariates, condition (1.6) immediately is satisfied. Condition (1.7) imposes a lower bound on the eigenvalues of the correlation matrix corresponding to the relevant covariates. The third assumption (1.8) characterizes the required sparsity of the true model. Finally, condition (1.9) ensures that the smallest entry of β decreases at most at a rate of $n^{-\frac{1-c_2}{2}}$. Since noise terms aggregate at a rate of $n^{-\frac{1}{2}}$, this additional gap of at least size n^{c_2} prevents estimation from being dominated by these noise terms.
In the present setting, Zhao and Yu (2006) showed the following.

Theorem 1.10 *Under Assumption 1.9, let there exist a non negative constant c_3 such that $c_3 < c_2 - c_1$ and $p = \mathcal{O}(\exp(n^{c_3}))$. Then strong irrepresentable condition 1.4 implies that the Lasso is strongly sign consistent. In particular, for $\lambda_n \propto n^{\frac{1+c_4}{2}}$ with $c_3 < c_4 < c_2 - c_1$,*

$$\mathbb{P}(\hat{\beta}(\lambda_n) =_s \beta) \geq 1 - o(\exp(n^{c_3})).$$

In other words, given some regularity conditions, strong IC still suffices for the probability of the Lasso to select the true model to converge to 1 at an exponential rate, even if p is large. Necessity of IC however, could not be obtained in this setting.

In either case, for consistent model selection, we hence like to see strong IC in place for every possible value of $\text{sign}(\beta)$. At this point, the covariates' correlation structure comes into play. Condition 1.4 depends on both the (sample) correlations between relevant and superfluous variables, C_{21}^n, and amongst the true variables themselves, C_{11}^n. Sufficient conditions for strong IC given in Zhao and Yu (2006) for different correlation structures generally put restrictions on the size of both correlation matrices' indices. To obtain an upper bound smaller than one to every component of $|C_{21}^n \, (C_{11}^n)^{-1} \, \text{sign}(\beta_{(1)})|$, such constraints on the entries of C_{11}^n are needed in order to obtain a sufficiently large lower bound on its eigenvalues, which in turn restricts the eigenvalues of the inverse $(C_{11}^n)^{-1}$ to smaller sizes. Entries of C_{21}^n, which 1.4 linearly depends on, should naturally be kept low to guarantee a low upper bound. Exemplarily, we state two of the sufficient conditions for strong IC for specific designs given in (Zhao and Yu, 2006, Corollary 3 & Corollary 4).

Corollary 1.11 (Sufficient Conditions for Strong IC)

- *Let $C_{i,j}^n = \rho_n^{|i-j|}$, $i, j = 1, \ldots, p$ with $|\rho_n| \leq c < 1$. Then strong irrepresentable condition 1.4 holds.*
- *Consider the block-wise design*

$$C^n = \begin{pmatrix} B_1^n & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & B_k^n \end{pmatrix}$$

with regression coefficient $\beta = (b_1, \ldots, b_k)$ corresponding to the different blocks. Then we have:

$$\textit{strong IC holds} \iff \exists 0 < \eta \leq 1 : \textit{ strong IC holds for all } B_j^n, b_j, j = 1, \ldots, k.$$

The need of such restrictions however highlights one potential problem: if a superfluous predictor is highly correlated with a true covariate, the Lasso may not be able to uncover its irrelevance. This already becomes evident in the simple simulation study (Zhao and Yu, 2006, 3.1) with correlations of block-wise design.

$\ell_1 + \ell_2$ Penalty

The selection consistency of the Elastic Net has been studied much less intensively. Bunea (2008) give an asymptotic result for both the Lasso and Elastic Net in linear and logistic regression models. Their results are based on a study of variable selection properties for said settings in finite samples, which shall be summarized in the subsequent Section 1.2. To conclude this section, we merely state the asymptotic result for the EN-penalized linear regression (cp. (Bunea, 2008, Corollary 3.6.)).

Corollary 1.12 *Let $\lambda_1 = O(\sqrt{\frac{\log(n)}{n}})$ and assume $\min_{j \in I^*} |\beta| = O(\sqrt{\frac{\log(n)}{n}})$. Assume further that the conditions of part 2. of Theorem 1.20 are met. Then, we have*

$$\lim_{n \to \infty} \mathbb{P}\left(\hat{I}^{EN} = I\right) = 1.$$

1.2 Variable selection in finite samples

The finite sample accuracy of variable selection via ℓ_1 and $\ell_1+\ell_2$ penalisation in linear and logistic regression models has been the topic of interest in Bunea (2008). The upcoming sections seek to separately summarize the findings given therein for both penalisation methods in linear models. Note that the setting considered in Bunea (2008) allows for both p and I^* (the index-set of true covariates) to depend on the sample size n. As before, this dependence will not be indicated in the notation. Moreover, in alignment with Bunea (2008), we henceforth assume that there is a constant $L > 0$ such that $|X_{ij}| < L$ for all i, j almost-surely.
The question addressed in Bunea (2008) is then:

> Given a level of confidence $1-\gamma$, the number of variables p and the sample size n, under which assumptions on the design matrix, for which strength of the signal, and for what values of the tuning parameters do we identify the true model at the given level of confidence?

Formally, if $\hat{I}$ is an estimate of I^*, conditions that yield $\mathbb{P}\left(\hat{I} = I^*\right) \geq 1-\gamma$ are of interest. Since $\mathbb{P}(I^* = \hat{I}) \geq 1 - \mathbb{P}(I^* \nsubseteq \hat{I}) - \mathbb{P}(\hat{I} \nsubseteq I^*)$, we want to find $\hat{I}$ such that

$$\mathbb{P}\left(I^* \subseteq \hat{I}\right) \geq 1-\gamma_1 \text{ and } \mathbb{P}\left(\hat{I} \subseteq I^*\right) \geq 1-\gamma_2, \tag{1.10}$$

with $\gamma = \gamma_1 + \gamma_2$. In other words, we seek to bound from below by a large margin the probabilities of correctly including all of the true variables in the selected set and selecting a subset of the truly relevant regressors. (Bunea, 2008, Lemma 3.1.) gives the following result on what governs the former for both the Lasso and Elastic Net and which follows directly from the definitions of $\hat{I}$ and I^*.

Lemma 1.13 *Let $\hat{\beta}$ be either the Lasso or Elastic Net estimate as introduced in Chapter 1 and $\hat{I}$ be the index set corresponding to the non-zero components of $\hat{\beta}$. Then*

$$\mathbb{P}\left(I^* \nsubseteq \hat{I}\right) \leq \mathbb{P}\left(\|\hat{\beta}-\beta\|_1 \geq \min_{l \in I^*} |\beta_l|\right).$$

The probability bound given above in Lemma 1.13 is needed in a later part, where conditions for correct inclusion of all true variables is linked to specific conditions on the regression parameter.

In preparation of discussing the accuracy of variable selection in such terms, Bunea (2008) start off by giving results on the predictive performance of the estimator. More precisely, an upper bound for the ℓ_1-distance between the respective estimate $\hat{\beta}$ and the true parameter β is established. Intuitively, an estimate should be in near proximity to the true β, in order to recover the true coefficient set I^* with high probability. Bunea (2008) especially note however, that given conditions on the design matrix and as long as this distance can be controlled for in some sense, the true subset I^* can be estimated correctly. This in particular sets apart the problem of correct variable selection from correct parameter estimation.
In order to establish an upper bound for $\|\hat{\beta}-\beta\|_1$, (Bunea, 2008, Section 2) first introduces conditions on the design matrix somewhat similar to the irrepresentable conditions given in Zhao

and Yu (2006). Recall the definition of

$$C_{kj}^n := \frac{1}{n}\sum_{i=1}^{n} X_{ki}X_{ij}^T, \quad 1 \leq j,k \leq p.$$

The first condition guarantees separation of the true variables from one another as well as the irrelevant ones, in terms of the corresponding correlation coefficients' sizes.

Condition 1.14 (Condition Identif) *There exists a constant* $0 < d \leq 1$ *such that*

$$\mathbb{P}\left(\max_{j\in I^*, k\neq j} |C_{kj}^n| \leq \frac{d}{k^*}\right) = 1.$$

Define further the set

$$V_{\alpha,\epsilon} = \left\{ v \in \mathbb{R}^p : \sum_{j\notin I^*} |v_j| \leq \alpha \sum_{j\in I^*} |v_j| + \epsilon \right\}.$$

A relaxation of Condition Identif is given by:

Condition 1.15 (Condition Stabil) *Let* $\alpha, \epsilon > 0$ *be given. There exists* $0 < b \leq 1$ *such that*

$$\mathbb{P}\left(v^T C^n v \geq b\sum_{j\in I^*} v_j^2 - \epsilon\right) = 1, \quad \forall v \in V_{\alpha,\epsilon}.$$

One possible interpretation of Condition Stabil becomes evident when setting $\epsilon = 0$. If the matrix $D^n - bD$ is positive definite almost surely, formally $\mathbb{P}\left(v^T(D^n - bD)v \geq 0, \forall v \in \mathbb{R}^p\right) = 1$ where D is a $p \times p$ diagonal matrix with ones at positions I^* and zeros else, Condition 1.15 holds true. In other words, this states that the correlation matrix remains non-negative definite, if the diagonal elements corresponding to the true variables are slightly decreased and may be regarded as a stability requirement on the correlation structure. Note that Condition 1.15 is even less strict, as it merely demands $\mathbb{P}\left(v^T(D^n - bD)v \geq 0, \forall v \in V_{\alpha,\epsilon}\right) = 1$.
As is pointed out in Bunea (2008), for the special case of the Lasso in the linear regression setting the two Conditions 1.14 and 1.15 are linked as follows: if Condition 1.14 holds for some d, then Condition 1.15 holds for $0 < b \leq 1 - 7d$. Note however, that this imposes the restriction $0 < d < \frac{1}{7}$ and hence allows for only very little correlation among the covariates (cp. Condition 1.14).
Hereafter, let $\hat{\beta}^L$ and $\hat{\beta}^{EN}$ denote the Lasso and Elastic Net estimator respectively. Moreover we write $\hat{I}^L$ and $\hat{I}^{EN}$ whenever we specifically speak of the index sets corresponding to the selected variables of said methods.

Sparse ℓ_1-balls

A central result is then given by (Bunea, 2008, Theorem 2.2.) that specifies the upper bound to the ℓ_1 ball $\|\hat{\beta}^L - \beta\|_1$ and shall be given below. Recall, that $\sigma^2 = Var(Y)$.

Theorem 1.16 *Assume Condition 1.15 corresponding to* $V_{\alpha,\epsilon}$ *with* $\epsilon = 0$ *and* $\alpha = 3$ *is satisfied for some* $0 < b \leq 1$. *If we choose*

$$\frac{\lambda}{2} \geq \max\left\{4L\sigma\sqrt{\frac{\log\left(\frac{4p}{\delta}\right)}{n}}, 8L\frac{\log\left(\frac{4p}{\delta}\right)}{n}\right\},$$

then the Lasso estimate $\hat{\beta}^L$ *satisfies*

$$\mathbb{P}\left(\|\hat{\beta}^L - \beta\|_1 \leq \frac{2}{b}\lambda k^*\right) \geq 1 - \delta.$$

Qualitatively, if Condition Stabil holds true for only very small values of b, the bound on $\|\hat{\beta}^L - \beta\|_1$ becomes large. As was shown in Bunea (2008) the Elastic Net estimator resulting from least squares minimisation including ℓ_1 and ℓ_2 penalisation terms already is less affected by such small values for b. The upcoming theorem gives the corresponding ℓ_1-bound for said EN-estimate.

Theorem 1.17 *Assume Condition 1.15 corresponding to* $V_{\alpha,\epsilon}$ *with* $\epsilon = 0$ *and* $\alpha = 4$ *is satisfied for some* $0 < b \leq 1$. *If there is some* $B > 0$ *independent of the sample size* n *such that* $\max_{j\in I^*} |\beta_j| \leq B$ *and if we choose*

$$\frac{\lambda_1}{2} \geq \max\left\{4L\sigma\sqrt{\frac{\log\left(\frac{4p}{\delta}\right)}{n}}, 8L\frac{\log\left(\frac{4p}{\delta}\right)}{n}\right\}, \quad \lambda_2 = \frac{\lambda_1}{4B}$$

then the EN estimate $\hat{\beta}^{EN}$ *satisfies*

$$\mathbb{P}\left(\|\hat{\beta}^{EN} - \beta\|_1 \leq \frac{2.125}{b + \lambda_2}\lambda_1 k^*\right) \geq 1 - \delta.$$

We observe, that even if $b \approx 0$, the bound on $\|\hat{\beta}^{EN} - \beta\|_1$ stays finite for any given p, n. In dependence of λ_1 and B however, λ_2 may still become quite small and the resulting bound rather large. In particular, λ_2 can not be chosen too big, for then the ℓ_2 penalty would become prevalent resulting in the selection of all given variables.
We may now analyse when it is possible to find estimates of β close to this true value as in Theorems 1.16 and 1.17 such that additionally we have $\mathbb{P}(I^* = \hat{I}) \geq 1 - \gamma$ for some $\gamma > 0$. In view of (1.10), we start by summarizing conditions such that with high probability all true variables are included in the estimated index-set $\hat{I}$ (cp. (Bunea, 2008, Section 3.1.)).

Correct inclusion of true variables

With Lemma 1.13 and Theorems 1.16-1.17 we obtain:

Corollary 1.18 *Let* $0 < \gamma_1 < 1$ *be fixed. Assume further Condition 1.15 holds for the parameters specified in Theorems 1.16-1.17.*

- ℓ_1*-penalty:*
  If $\min_{j\in I^*} |\beta_j| \geq \frac{4}{b}\lambda k^*$ *with* λ *as in Theorem 1.16, then* $\mathbb{P}(I^* \subseteq \hat{I}^L) \geq 1 - \gamma_1$.
- $\ell_1 + \ell_2$*-penalty:*
  If there exists some $B > 0$ *such that*

$$\frac{2.125}{b + \lambda_2}\lambda_1 k^* \leq \min_{j\in I^*} |\beta_j| \leq \max_{j\in I^*} |\beta_j| \leq B$$

  with λ_1, λ_2 *given by Theorem 1.17, then* $\mathbb{P}(I^* \subseteq \hat{I}^{EN}) \geq 1 - \gamma_1$.

Note that the lower bounds on the minimum size of the true coefficients are of the form $C\lambda k^*$ ($C\lambda_1 k^*$) for some constant C. If $C \approx 1$ and k^* is rather low, satisfying $\lambda k^* < 1$ ($\lambda_1 k^* < 1$), then moderately sized signals can be correctly identified. In general however, C and k^* may take large

values, resulting in lower bounds on the coefficient size that are too conservative.
Up to now, we have made use of Condition Stabil only to derive the above results. As is shown in Bunea (2008), these lower bounds on the signal strength can be weakened if one imposes more conditions on the design of the matrix. More precisely, weaker signals can be detected when assuming the more restrictive Condition Identif instead. Hence, the price to pay for improved accuracy of variable selection in finite samples is to allow for less correlation among the true and irrelevant, and the true variables themselves. Intuitively, if a signal is very weak and correlations involving true variables are high, one can not hope to unveil the truly underlying model with high probability.
In more detail, Bunea (2008) derive the following: If Condition 1.14 is met, coefficients of sizes above the noise level $\sqrt{n}$ can be recovered. Specifically, consider $0 < \delta < 1$ fixed and K be an upper bound on k^* (one may choose $p = K$ in case k^* is unknown as a conservative bound).

Proposition 1.19

1. *ℓ_1-penalty:*
   Let
   $$\frac{\lambda}{2} \geq \max\left\{4L\sigma\sqrt{\frac{\log\left(\frac{4pK}{\delta}\right)}{n}}, 8L\frac{\log\left(\frac{4pK}{\delta}\right)}{n}\right\},$$
   and assume that
   $$\min_{j\in I^*}|\beta_j| \geq \lambda.$$
   If Condition Identif is satisfied for $d \leq \frac{1}{15}$ then
   $$\mathbb{P}\left(I^* \subseteq \hat{I}^L\right) \geq 1 - \delta - \frac{\delta}{p}.$$

2. *$\ell_1 + \ell_2$-penalty:*
   Let
   $$\frac{\lambda_1}{2} \geq \max\left\{4L\sigma\sqrt{\frac{\log\left(\frac{4pK}{\delta}\right)}{n}}, 8L\frac{\log\left(\frac{4pK}{\delta}\right)}{n}\right\},$$
   and assume that
   $$\min_{j\in I^*}|\beta_j| \geq \lambda_1.$$
   Assume further, that $\max_{j\in I^*}|\beta_j| \leq B$ for some $B > 0$ and choose $\lambda_2 = \frac{\lambda_1}{4B}$. If Condition Identif is satisfied for $d \leq \frac{1+\lambda_2}{17.5}$, then
   $$\mathbb{P}\left(I^* \subseteq \hat{I}^{EN}\right) \geq 1 - \delta - \frac{\delta}{p}.$$

Note that assuming Condition Identif instead of Condition Stabil yields a substantial relaxation of the lower bound on the size of the true coefficients. In particular, the lower bound λ_1 now no longer depends on either the possibly large k^* or the possibly small b.
As Bunea (2008) point out, Proposition 1.19 allows immediate comparison of the Lasso and Elastic Net in terms of variable selection. The difference lies in the restriction on the constant d that becomes of importance in Condition 1.14. We observe that slightly larger values are allowed for

with the EN-estimate. Hence correct variable inclusion can be guaranteed for the EN under less restrictive assumptions on the correlations of the design than for the Lasso. In other words, if the correlations involving true variables attain slightly larger values than is permitted for the Lasso, the EN may provide an alternative. Note however, that although one would like to increase the value of λ_2 in order to allow for a higher degree of correlation, this would ultimately lead to not setting any of the estimated coefficients to zero.

Correct Subset Selection

We now state conditions (almost identical to the ones given in Proposition 1.19) that guarantee $\mathbb{P}(\hat{I} \subseteq I^*) \geq 1 - \gamma_2$, thereby guaranteeing $\mathbb{P}(\hat{I} = I^*) \geq 1 - \gamma$ (see (Bunea, 2008, Theorem 3.5.)).

Theorem 1.20 *Let K be an upper bound on k^*.*

1. *ℓ_1-penalty:*
   Let
   $$\frac{\lambda}{2} \geq \max\left\{4L\sigma\sqrt{\frac{\log\left(\frac{4pK}{\delta}\right)}{n}}, 8L\frac{\log\left(\frac{4pK}{\delta}\right)}{n}\right\},$$
   and assume that
   $$\min_{j\in I^*}|\beta_j| \geq \lambda.$$
   If Condition Identif is satisfied for $d \leq \frac{1}{15}$ then
   $$\mathbb{P}\left(\hat{I}^L = I^*\right) \geq 1 - 3\delta - \frac{\delta}{p}.$$

2. *$\ell_1 + \ell_2$-penalty:*
   Let
   $$\frac{\lambda_1}{2} \geq \max\left\{4L\sigma\sqrt{\frac{\log\left(\frac{4pK}{\delta}\right)}{n}}, 8L\frac{\log\left(\frac{4pK}{\delta}\right)}{n}\right\},$$
   and assume that
   $$\min_{j\in I^*}|\beta_j| \geq \lambda_1.$$
   Assume further, that $\max_{j\in I^*}|\beta_j| \leq B$ for some $B > 0$ and choose $\lambda_2 = \frac{\lambda_1}{4B}$. If Condition Identif is satisfied for $d = \frac{1+\lambda_2}{17.5}$, then
   $$\mathbb{P}\left(\hat{I}^{EN} = I^*\right) \geq 1 - 3\delta - \frac{\delta}{p}.$$

As derived in the previous Sections 1.1 and 1.2, both the Lasso's and the Elastic Net's ability to select the true model (in finite samples and asymptotically) can be characterized by a condition that depends on the sample correlation among relevant and superfluous covariates. The results all show, that correct variable selection is not guaranteed, should these correlations be too high. We now like to gain further insight on which regressors are in fact selected among a group of correlated variables. The following section intends to collect the findings on the penalisation methods' variable selection properties in presence of multicollinearity.

1.3 The Lasso and Elastic Net on correlated variables

One of the Lasso penalty's main benefits is that it causes coefficients to be set exactly zero thus providing simple estimated models. The Elastic Net similarly induces sparsity, even though its penalty is somewhat less aggressive. Intuitively, this becomes evident considering the geometry of the penalisation methods. Figure 1.1 is a simple visualisation of the setting in (1.1) in two dimensions with $t = 1$.

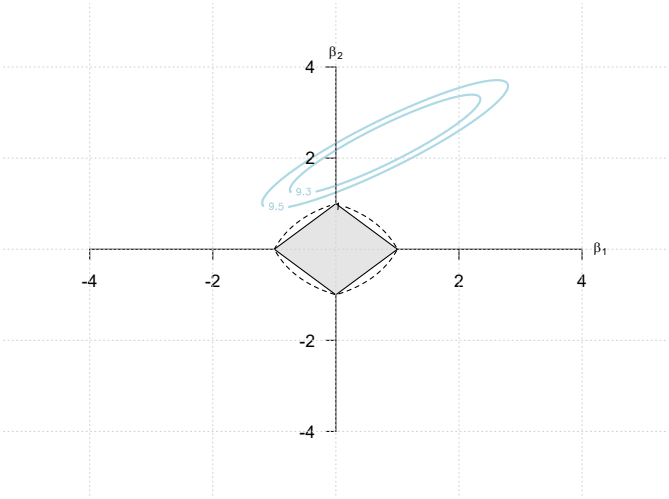

Figure 1.1: Two-dimensional estimation picture for the Lasso and Elastic Net. For $t = 1$, the constraint region of the Lasso is depicted in grey, while that of the EN is indicated by the dashed black line. The contours of the objective are given in blue.

We now like to further investigate settings in which groups of variables are correlated and compare the penalisation methods w.r.t. their selection properties. Frequently, in the existing literature it is claimed that: *Out of a group of variables whose pairwise correlations are very high the Lasso tends to select only one variable, essentially at random.* (see i.a. Zou and Hastie (2005),Grave et al. (2011)). The EN on the other hand is able to jointly select groups of correlated variables. This phenomenon, referred to as 'grouping effect' by Zou and Hastie (2005), shall be made precise below.

1.3.1 Grouping Effect

A regression model exhibits the grouping effect, if among a group of highly correlated regressors all variables tend to be assigned regression coefficients, that are equal up to a change of sign if negatively correlated. In particular, the coefficients of exactly identical variables should be identical as well. In concordance with Zou and Hastie (2005) we consider the following linear regression model with a universal penalty $J(\cdot)$ and tuning parameter $\lambda \geq 0$

$$\min \|Y - X\beta\|_2^2 + \lambda J(\beta). \tag{1.11}$$

The penalty $J(\cdot)$ is assumed to be positive for $\beta \neq 0$. Denote by $\hat{\beta}(\lambda)$ the corresponding minimizer. In such a setting, for the extreme case of identical variables, we have the following result (see (Zou and Hastie, 2005, Lemma 2.)).

Lemma 1.21 *Assume that two variables are identical i.e. there exist* $i, j \in \{1, \dots, p\}$ *such that* $X^{(i)} = X^{(j)}$.

- *If* $J(\cdot)$ *is strictly convex, then* $\hat{\beta}_i(\lambda) = \hat{\beta}_j(\lambda)$, $\forall \lambda > 0$.
- *If* $J(\beta) = \|\beta\|_1$, *then* $\hat{\beta}_i \hat{\beta}_j \geq 0$ *and for any* $s \in [0,1]$, $\tilde{\beta}(s)$ *is another minimizer of* (1.11), *whereby*

$$\tilde{\beta}_k(s) = \begin{cases} \left(\hat{\beta}_i + \hat{\beta}_j\right) s & k = i. \\ \left(\hat{\beta}_i + \hat{\beta}_j\right)(1-s) & k = j. \\ \hat{\beta}_k & \textit{else.} \end{cases}$$

Hence, by the above lemma, penalisation methods involving strictly convex penalties such as the Elastic Net are guaranteed to display the grouping effect as desired for two identical variables. The Lasso however, does not even have a unique solution. In particular, the set of valid solutions in the Lasso case includes estimates where either one or none of the coefficients corresponding to the identical variables is set to zero.

In theory, by the above, for every $\alpha \in (0, 1]$, the EN should display said grouping effect for any choice of penalty $\lambda \geq 0$. What can be observed in practice however, is the following: while for pure ridge (i.e. $\alpha = 1$) the estimated coefficients of identical variables indeed almost coincide, this effect wears off for smaller values of α. As α tends to zero and the Lasso penalty's influence on the parameter estimates rises, the estimated coefficients increasingly differ. This phenomenon can be observed for any value of λ, however, the difference is of increasing magnitude for lower values of the penalty parameter λ.

To see this, consider the following simple simulation. We generate $N = 1000$ data points from the linear model $Y = X_1 + X_2 + \epsilon$ with regressors X_1, X_2 being realistaions of independent standard normal random variables and noise term $\epsilon \sim \mathcal{N}(0, 1)$. Then, we like to fit Y regressing on $X = (X_1, X_2, X_2)$ using as penalisation methods the pure Lasso, pure ridge and some EN-versions. We do so using *cv.glmnet()* of the R-package *glmnet()* that performs cross-validated elastic net penalisation. For given values of α and λ, a solution to the penalized minimisation is obtained using cyclical coordinate descent (see Friedman et al. (2010)). K-fold cross validation is then applied to obtain a mean cross-validation error including a confidence bound for a number of λ-values. Among those, one can identify the amount of penalisation that minimizes the mean cross-validation error, referred to as $\lambda_{\min}$. Within this simulation study, we inspected the estimated coefficients for $\alpha \in \{0, 0.01, 0.1, 0.2, 0.5, 1\}$ and three different values for λ chosen from the grid suggested by *cv.glmnet()*. More precisely, the smallest and largest given values (denoted by λ_{L} and λ_{S} respectively) and $\lambda_{\min}$ were chosen. Table 1.1 shows the absolute difference between the estimated coefficients $\hat{\beta}_2, \hat{\beta}_3$ of the identical regressors X_2.

Moreover, according to Lemma 1.21, the Lasso solution for settings where two variables are identical is not unique and could vary from either coefficient being zero to both being equal. Implemented algorithms that yield a Lasso solution however seem to favour sparse solutions, i.e. they mostly set one of the coefficients to (almost) zero. The actual values for the coefficients $\hat{\beta}_2$ and $\hat{\beta}_3$ obtained by fitting a pure Lasso on the simulated data from above using *cv.glmnet()* are given in Table 1.2.

α	$\lambda_{\min}$	λ_{L}	λ_{S}
(lasso) 0	0.851	0.085	1.002
(ridge) 1	0.001	0.000	0.006
0.5	0.007	0.000	0.031
0.2	0.030	0.003	0.080
0.1	0.063	0.008	0.136
0.01	0.444	0.083	0.556

Table 1.1: Absolute difference of estimated coefficients of identical regressors for different values of EN and penalty parameters α respectively λ.

	$\lambda_{\min}$	λ_{L}	λ_{S}
$\hat{\beta}_2$	0.851	0.085	1.002
$\hat{\beta}_3$	0.000	0.000	0.000

Table 1.2: Estimated Lasso coefficients of identical regressors for different values of penalty parameter λ.

In general, we are more interested in scenarios where two regressors are highly correlated rather than exactly identical. Zou and Hastie (2005) give a quantitative description of the difference (in a sense to be subsequently defined) between the EN-coefficient paths of two covariates. More precisely, an upper bound to said distance is established, that depends on the regressors' sample correlation. The result (cp. (Zou and Hastie, 2005, Theorem 1.)) is stated below. As usual, we assume the response Y to be centred and the predictors X to be standardized.

Theorem 1.22 *Let $\hat{\beta}^{EN}(\lambda_1, \lambda_2)$ be the EN-estimate for regularisation parameters $\lambda_1, \lambda_2 \geq 0$. Assume for some $i, j \in \{1, \ldots, p\}$ and some λ_1, λ_2 that $\hat{\beta}_i^{EN}(\lambda_1, \lambda_2)\hat{\beta}_j^{EN}(\lambda_1, \lambda_2) > 0$ and define*

$$D_{i,j}(\lambda_1, \lambda_2) := \frac{1}{\|Y\|_1}|\hat{\beta}_i^{EN}(\lambda_1, \lambda_2) - \hat{\beta}_j^{EN}(\lambda_1, \lambda_2)|.$$

Then

$$D_{i,j}(\lambda_1, \lambda_2) \leq \frac{1}{\lambda_2}\sqrt{2(1 - \rho_{i,j})},$$

where $\rho_{i,j} = {X^{(i)}}^T X^{(j)}$ is the sample correlation.

Hence, if predictors i and j are highly correlated with $\rho_{i,j} \approx 1$, the difference of their coefficient paths is almost zero. In case $\rho_{i,j} < 0$, we analogously obtain a bound on the difference in the absolute values of the estimated coefficients considering -$X^{(j)}$ in the above theorem. Thus, for $\rho_{i,j} \approx -1$ the distance of the absolute value of the coefficients in the above sense will become arbitrarily small. Note however, that the strength of the bound also depends on how powerful the ℓ_2-norm is in the minimisation. The more weight the ℓ_2-penalty receives, i.e. the larger λ_2, the stronger is the grouping effect.
Moreover, Theorem 1.22 requires the estimated coefficients to be non-zero. Therefore, we do not gain information on the distance between the estimates for regularisation parameters that set either one coefficient of correlated variables to zero. Nonetheless, by the geometry of the specific problem (1.3), this may happen. However, depending on the amount of correlation, this effect can only occur for some choices of penalty parameters λ_1 and λ_2. As $\rho_{i,j} \to 1$, such a selection may happen only if the ℓ_1-penalty is given enough weight in expense of the ℓ_2 penalty.

Numerically, this can be validated as well. To that end, we simulate $N = 10^3$ data points according to $Y = 2X_1 + X_2 + 4Z + \epsilon$, with X_1, X_2 realisations of centred Gaussian variables with correlation $\rho = 0.9$ and Z, ϵ both independent standard normal realisations. For different values of elastic net parameter α, we then perform the optimisation where cross-validation is used to optimize the penalty parameter λ. Table 1.3.1 shows the resulting coefficient estimates $\hat{\beta}_1, \hat{\beta}_2$ for X_1 and X_2 for both $\lambda_{\min}$ and a more restrictive λ_{L}. We observe that for any value of $\alpha < 1$ there is a value of λ only one of the coefficients of the correlated variables is set to zero. When the penalty consists of the ℓ_2-term only, the estimated coefficients again tend to be of similar size.

$\hat{\beta}_i(\alpha)$	$\lambda_{\min}$	λ_{L}
$\hat{\beta}_1(1)$	1.538	0.004
$\hat{\beta}_2(1)$	1.197	0.004
$\hat{\beta}_1(0.5)$	1.966	0.176
$\hat{\beta}_2(0.5)$	0.911	0.000
$\hat{\beta}_1(0.2)$	1.997	0.237
$\hat{\beta}_2(0.2)$	0.841	0.000
$\hat{\beta}_1(0.1)$	2.013	0.254
$\hat{\beta}_2(0.1)$	0.818	0.000

Table 1.3: Coefficient estimates of highly correlated variables for $\lambda_{\min}$ and a comparably larger penalty parameter λ_{L} and each of $\alpha \in \{0.1, 0.2, 0.5, 1\}$.

Joint selection of correlated variables

Motivated by the above, we like to further elaborate on whether or not the EN jointly selects correlated variables. More precisely, we investigate in what way the correlation between two regressors influences their selection when an EN is fitted with fixed penalty parameters. For our analysis we consider the following setting: let $X = (X_1, X_2) \in \mathcal{L}_2(\mathbb{R}^2)$ be a two-dimensional random vector of explanatory variables with $\rho := Cor(X_1, X_2)$, $Var(X_i) = 1$ and $\mathbb{E}[X_i] = 0$, $i = 1, 2$. $Y \in \mathcal{L}_2(\mathbb{R})$ be the centred response variable. Our goal is to obtain EN-estimates $\hat{\beta}_1(\rho), \hat{\beta}_2(\rho)$ for $\beta = (\beta_1, \beta_2)^T$ as minimizers to

$$\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \left\{ (1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2 \right\},$$

or equivalently

$$\min_{\beta} Var(Y) - 2\beta^T \begin{pmatrix} Cov(X_1, Y) \\ Cov(X_2, Y) \end{pmatrix} + \beta^T \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \beta + \lambda \left\{ (1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2 \right\}, \tag{1.12}$$

with $\lambda > 0, \alpha > 0$ fixed in dependence of ρ. In what follows, we often will refer to the objective function of problem (1.12) as *cost function* and write

$$C(\beta, \rho) := Var(Y) - 2\beta^T \begin{pmatrix} Cov(X_1, Y) \\ Cov(X_2, Y) \end{pmatrix} + \beta^T \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \beta + \lambda \left\{ (1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2 \right\}. \tag{1.13}$$

For some fixed value of ρ and $\alpha > 0$ the objective function $C(\cdot, \rho)$ is strictly convex. Only for $\alpha = 1$ however, it is differentiable on $\mathbb{R}^2$. Whenever the ℓ_1-penalty receives positive weight,

differentiability breaks down at $\beta = 0$ and standard results from constraint optimisation can not be applied. For convex objectives however, one may resort to subgradient methods. See chapter A of the appendix for a collection of the results on subgradients and their use in convex optimization that we refer to in the following.

By Theorem A.6, the estimate $\hat{\beta}$ is optimal for problem (1.12) with ρ fixed iff zero is included in the cost function's subdifferential at this point, that is $0 \in \partial C(\hat{\beta}, \rho)$. With Remark A.2 and Lemma A.3 we obtain the set

$$\partial C(\beta, \rho) = -2 \begin{pmatrix} Cov(X_1, Y) \\ Cov(X_2, Y) \end{pmatrix} + 2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \lambda \left\{ (1-\alpha) \begin{pmatrix} g_{\ell_1}(\beta)_1 \\ g_{\ell_1}(\beta)_2 \end{pmatrix} + 2\alpha \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \right\}, \quad \beta \in \mathbb{R}^2,$$

whereby $g_{\ell_1}(\cdot)$ defines the subdifferential of the ℓ_1-norm according to Example A.5. Setting $\partial C(\beta, \rho) = 0$ we obtain the following representations of the minimizers $\hat{\beta}(\rho)$:

$$\hat{\beta}_2 = -\frac{1+\lambda\alpha}{\rho}\hat{\beta}_1 + \frac{1}{\rho}\left[Cov(X_1, Y) - \frac{\lambda(1-\alpha)}{2} g_{\ell_1}(\hat{\beta})_1 \right], \tag{1.14}$$

$$\hat{\beta}_1 = -\frac{1+\lambda\alpha}{\rho}\hat{\beta}_2 + \frac{1}{\rho}\left[Cov(X_2, Y) - \frac{\lambda(1-\alpha)}{2} g_{\ell_1}(\hat{\beta})_2 \right]. \tag{1.15}$$

Consider now the following: for $\rho = 1$, by Lemma 1.21 we have that $\hat{\beta}_1(\rho) = \hat{\beta}_2(\rho)$. Without loss of generality, assume that $\hat{\beta}_i(\rho) > 0$, for $i = 1, 2$ (and $\rho = 1$). We then ask the following: is there some value $\rho^* < 1$ such that either one of the coefficients $\hat{\beta}_i(\rho^*), i = 1, 2$ is set to zero while the other is not?

In order to examine the parameter estimates $\hat{\beta}_i(\rho), i = 1, 2$ for $\rho < 1$, we derive a continuity result on Lemma 1.23. For simplicity, we hereafter suppress the optimal parameter's dependency on ρ in the notation and only indicate the corresponding value for ρ when necessary.

Lemma 1.23 *Let $C(\cdot, \cdot)$ be the cost function as defined in* (1.13)*, that is*

$$\begin{aligned} C : [-1, 1] \times \mathbb{R}^2 &\to \mathbb{R} \\ (\rho, \beta) &\mapsto Var(Y) - 2\beta^T \begin{pmatrix} Cov(X_1, Y) \\ Cov(X_2, Y) \end{pmatrix} + \beta^T \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \beta + \lambda \left\{ (1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2 \right\}. \end{aligned}$$

The function

$$\begin{aligned} F : [-1, 1] &\to \mathbb{R}^2 \\ \rho &\mapsto \underset{\beta \in D(t)}{\arg\min}\, C(\beta, \rho) \end{aligned}$$

with $D(t) = \left\{ \beta \in \mathbb{R}^2 : (1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2 \leq t \right\}$ is continuous.

Proof. For any $\rho \in [-1, 1]$ the set $\underset{\beta \in D(t)}{\arg\min}\, C(\beta, \rho)$ is non-empty and consists of one element only, as $C(\rho, \cdot)$ is strictly convex. Therefore, F is well-defined.

To show continuity, let $(\rho_n)_{n \in \mathbb{N}}$ be a convergent sequence in $[-1, 1]$ such that $\rho_n \underset{n \to \infty}{\longrightarrow} \rho$. Denote by $\hat{\beta}(\rho_n), \hat{\beta}(\rho)$ the minimizers of $C(\rho_n, \cdot)$ and $C(\rho, \cdot)$ respectively. We like to show that $F(\rho_n) \underset{n \to \infty}{\longrightarrow} F(\rho)$, or equivalently $\hat{\beta}(\rho_n) \underset{n \to \infty}{\longrightarrow} \hat{\beta}(\rho)$. Since for any n, $\hat{\beta}(\rho_n) \in D(t)$, $\hat{\beta}(\rho_n)_{n \in \mathbb{N}}$ is a bounded sequence. By Bolzano-Weierstrass, there exists a convergent subsequence and hence the set of accumulation points of $\hat{\beta}(\rho_n)_{n \in \mathbb{N}}$ is non-empty. Let $\hat{\beta}^*$ be an accumulation point. Then there

exists a subsequence $\hat{\beta}(\rho_{n_k})_{k\in\mathbb{N}}$ such that $\hat{\beta}(\rho_{n_k}) \underset{k\to\infty}{\longrightarrow} \hat{\beta}^*$. By continuity of the objective function we obtain

$$C(\rho, \hat{\beta}^*) = \lim_{k\to\infty} C(\rho_{n_k}, \hat{\beta}(\rho_{n_k})) = \lim_{k\to\infty} \min_{\beta} C(\rho_{n_k}, \beta) = \min_{\beta} C(\rho, \beta),$$

and thus $\hat{\beta}^*$ is a minimizer of $C(\rho, \cdot)$. Since the argument minimizing problem (1.12) is unique we obatin $\hat{\beta}^* = \hat{\beta}(\rho)$. Hence we have shown, that the non-empty set of accumulation points is a singleton and thus

$$\liminf_{n\to\infty} \hat{\beta}(\rho_n) = \limsup_{n\to\infty} \hat{\beta}(\rho_n) = \lim_{n\to\infty} \hat{\beta}(\rho_n) = \hat{\beta}(\rho).$$

□

Proceeding with our analysis, by Lemma 1.23, there must be some $\rho^* \in [-1, 1)$ such that $\hat{\beta}(\rho) > 0$ component-wise $\forall \rho \in (\rho^*, 1]$. Moreover, by continuity, ρ^* is the largest possible correlation such that $\hat{\beta}_i(\rho^*) = 0$ for either one or both of $i = 1, 2$.

We first elaborate on the case where only one of the coefficients is set to zero. In other words, we are interested to know whether there exists some $\rho^{*2} \in [-1, 1)$ such that $\hat{\beta}_2(\rho^{*2}) = 0$ and $\hat{\beta}_1(\rho^{*2}) > 0$, and if so, how it can be characterized. Inserting (1.15) into (1.14) yields

$$\hat{\beta}_2 = \frac{\rho\left[Cov(X_1, Y) - \frac{\lambda(1-\alpha)}{2} g_{\ell_1}(\hat{\beta})_1\right] - (1+\lambda\alpha)\left[Cov(X_2, Y) - \frac{\lambda(1-\alpha)}{2} g_{\ell_1}(\hat{\beta})_2\right]}{\rho^2 - (1+\lambda\alpha)^2}. \tag{1.16}$$

Letting $\hat{\beta}_2 \to 0$ with $\hat{\beta}_1 > 0$ in the above, we have

$$\gamma = \lim_{\hat{\beta}_2\to 0} g_{\ell_1}(\hat{\beta})_2 = \frac{Cov(X_2, Y) - \frac{\rho}{(1+\lambda\alpha)}\left[Cov(X_1, Y) - \frac{\lambda(1-\alpha)}{2}\right]}{\frac{\lambda(1-\alpha)}{2}}.$$

The above limit is only valid if $|\gamma| < 1$ or equivalently

$$U_2 > \rho > L_2 \quad , \text{ if } Cov(X_1, Y) > \tfrac{\lambda(1-\alpha)}{2}, \tag{1.17}$$

$$L_2 > \rho > U_2 \quad , \text{ if } Cov(X_1, Y) < \tfrac{\lambda(1-\alpha)}{2}, \tag{1.18}$$

$$|Cov(X_2, Y)| < \frac{\lambda(1-\alpha)}{2} \quad , \text{ if } Cov(X_1, Y) = \tfrac{\lambda(1-\alpha)}{2}, \tag{1.19}$$

with

$$U_2 := \frac{(1+\lambda\alpha)\left[Cov(X_2,Y) + \frac{\lambda(1-\alpha)}{2}\right]}{Cov(X_1,Y) - \frac{\lambda(1-\alpha)}{2}},$$

$$L_2 := \frac{(1+\lambda\alpha)\left[Cov(X_2,Y) - \frac{\lambda(1-\alpha)}{2}\right]}{Cov(X_1,Y) - \frac{\lambda(1-\alpha)}{2}}.$$

Hence, for ρ^{*2} satisfying either of the three conditions (1.17)-(1.19), and provided that $|\rho^{*2}| \leq 1$, we have

$$\lim_{\rho\downarrow\rho^{*2}} \hat{\beta}_2(\rho) = 0,$$

$$\lim_{\rho\downarrow\rho^{*2}} \hat{\beta}_1(\rho) = \frac{\left[Cov(X_1, Y) - \frac{\lambda(1-\alpha)}{2}\right]}{1+\lambda\alpha}.$$

Analogous results are obtained, for $\hat{\beta}_1 \to 0$. The corresponding boundaries for ρ^{*1} are given by:

$$\begin{aligned} U_1 &= \frac{(1+\lambda\alpha)\left[Cov(X_1,Y)+\frac{\lambda(1-\alpha)}{2}\right]}{Cov(X_2,Y)-\frac{\lambda(1-\alpha)}{2}}, \\ L_1 &= \frac{(1+\lambda\alpha)\left[Cov(X_1,Y)-\frac{\lambda(1-\alpha)}{2}\right]}{Cov(X_2,Y)-\frac{\lambda(1-\alpha)}{2}}. \end{aligned}$$

With ρ^{*1} such that $|\rho^{*1}| < 1$ and either one of

$$U_1 > \rho > L_1 \quad , \text{ if } Cov(X_2,Y) > \tfrac{\lambda(1-\alpha)}{2}, \tag{1.20}$$

$$L_1 > \rho > U_1 \quad , \text{ if } Cov(X_2,Y) < \tfrac{\lambda(1-\alpha)}{2}, \tag{1.21}$$

$$|Cov(X_1,Y)| < \frac{\lambda(1-\alpha)}{2} \quad , \text{ if } Cov(X_2,Y) = \tfrac{\lambda(1-\alpha)}{2}, \tag{1.22}$$

hold true, we similarly obtain

$$\begin{aligned} \lim_{\rho\downarrow\rho^{*1}} \hat{\beta}_1(\rho) &= 0, \\ \lim_{\rho\downarrow\rho^{*1}} \hat{\beta}_2(\rho) &= \frac{\left[Cov(X_2,Y)-\frac{\lambda(1-\alpha)}{2}\right]}{1+\lambda\alpha}. \end{aligned}$$

Note, that for either ρ^{*1} or ρ^{*2} to be a valid correlation $|\rho^{*1}| < 1$ or $|\rho^{*2}| < 1$ are necessary. The condition, say $|\rho^{*1}| < 1$, is met for some ρ^{*1} satisfying one of (1.20)-(1.21) if and only if the boundaries satisfy $(L_1, U_1) \subseteq [-1, 1]$ or $(U_1, L_1) \subseteq [-1, 1]$ respectively. Hence conditions $U_1 < 1$ and $L_1 > -1$ or $L_1 < 1$ and $U_1 > -1$ are required. Those equivalently yield

$$Cov(X_1,Y) + \frac{\lambda(1-\alpha)}{2} < \underbrace{\frac{Cov(X_2,Y)-\frac{\lambda(1-\alpha)}{2}}{1+\lambda\alpha}}_{>0, \text{ if } L_1<U_1} \tag{1.23}$$

$$Cov(X_1,Y) - \frac{\lambda(1-\alpha)}{2} > \underbrace{-\frac{Cov(X_2,Y)-\frac{\lambda(1-\alpha)}{2}}{1+\lambda\alpha}}_{<0, \text{ if } L_1<U_1}, \tag{1.24}$$

for $L_1 < U_1$ and

$$Cov(X_1,Y) + \frac{\lambda(1-\alpha)}{2} < \underbrace{-\frac{Cov(X_2,Y)-\frac{\lambda(1-\alpha)}{2}}{1+\lambda\alpha}}_{>0, \text{ for } U_1<L_1} \tag{1.25}$$

$$Cov(X_1,Y) - \frac{\lambda(1-\alpha)}{2} > \underbrace{\frac{Cov(X_2,Y)-\frac{\lambda(1-\alpha)}{2}}{1+\lambda\alpha}}_{<0, \text{ for } U_1<L_1}, \tag{1.26}$$

whenever $U_1 < L_1$. In any case, two equations that heavily depend on the choice of penalty parameters need to be fulfilled. For strong penalisation (λ large) $U_1 < 1$ and $L_1 > -1$, and hence also scenarios in which one coefficient is set to zero while the other is not are less likely to happen. Likewise, larger values of α contribute to less occurrences of such phenomena. In other words, the Lasso ($\alpha = 0$) is more likely to set only one coefficient to zero than any EN. Moreover, $\hat{\beta}_1$ is more easily set to zero while $\hat{\beta}_2$ is not, if the correlation of X_2 and Y is significantly larger than that

between X_1 and Y.
Whenever condition (1.22) is met, this implies an indirect condition on the correlation ρ^{*1}. Analogous results hold for $|\rho^{*2}| < 1$.

Setting both coefficients to zero is equivalent to equations

$$\begin{aligned} 0 &= \frac{1}{\rho}\left[Cov(X_1,Y) - \frac{\lambda(1-\alpha)}{2}\underbrace{g_{\ell_1}(0)_1}_{=:\gamma_1}\right] \iff \gamma_1 = Cov(X_1,Y)\frac{2}{\lambda(1-\alpha)}, \\ 0 &= \frac{1}{\rho}\left[Cov(X_2,Y) - \frac{\lambda(1-\alpha)}{2}\underbrace{g_{\ell_1}(0)_2}_{=:\gamma_2}\right] \iff \gamma_2 = Cov(X_2,Y)\frac{2}{\lambda(1-\alpha)}, \end{aligned}$$

being satisfied. As all elements of $g_{\ell_1}(0)$ must be smaller than one in absolute value, we require $|\gamma_i| < 1$ for $i = 1, 2$ or equivalently

$$|Cov(X_i, Y)| < \frac{\lambda(1-\alpha)}{2}, \quad i = 1, 2. \tag{1.27}$$

Although this condition does not seem to explicitly depend on ρ the correlation between X_1 and X_2 still has an impact on the covariances in (1.27). Intuitively speaking, when both covariates are positively (or negatively) correlated with the target, letting their correlation tend to -1 will reduce these correlations. Similarly, given either one of the regressors is positively or negatively correlated with the dependent variable, changing the correlation $\rho \to 1$ will reduce the initial correlations $Cor(X_i, Y)$, $i = 1, 2$. Simulation 2 below describes such a scenario.

Simulation 1

In the following, we consider a simple simulation to verify the above results numerically. For an equidistant sequence of values for $\rho \in [-1, 1]$ with step-size $\delta = 0.05$, we simulate $N = 10^3$ data-points according to

$$(X_1, X_2)^T \sim \mathcal{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$

and set the explained variable to $Y = X_1 + \epsilon$ with independent noise $\epsilon \sim \mathcal{N}(0, 1)$. For each value of ρ, we then fit an EN with $\lambda = \alpha = 0.5$. Figure 1.2 shows the estimated coefficients in dependence of the correlation ρ. We observe the following: first, in line with Lemma 1.21, for $\rho = \pm 1$ the coefficients (almost) coincide in absolute value. Second, as can be quickly verified, the boundary conditions on the distance of the absolute values of the estimated coefficients according to Theorem 1.22 are satisfied. Third, we like to verify whether for values of the correlation ρ in between the boundaries L_2, U_2 (in this simulation $Cov(X_1, Y) > 0.5^3$ and thus equation (1.17) should be satisfied) the coefficient β_2 is indeed set to zero. In Figure 1.2, the orange vertical lines indicate whether correlations that lie within those respective boundaries have been used for estimation. We note that indeed, for any such $\rho \in (L_2, U_2)$ the estimate $\hat{\beta}_2$ is set to zero while $\hat{\beta}_1$ is not. The blue and green lines identify correlations that are close to either margin L_2 or U_2 with a tolerance of 0.01 and 0.1 respectively. Interestingly, although the above result does not cover such cases, also for correlations that are close to lying in their respective interval we have $\hat{\beta}_2 = 0$

and $\hat{\beta}_1 > 0$.

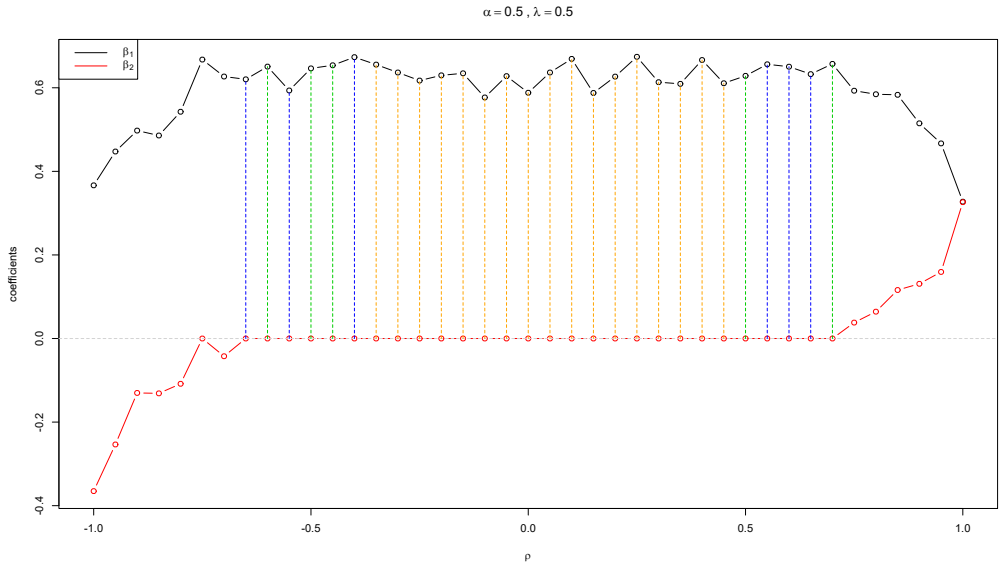

Figure 1.2: Estimated coefficients in dependence of the correlation in the explanatory variables. The vertical lines in orange indicate correlations that lie within the optimal interval $(L_2, U_2) \subseteq [-1, 1]$. Correlations, which are close to either one boundary with a tolerance of 0.05 and 0.1 are identified by the blue and green vertical lines respectively.

Simulation 2

As a second example, consider for ρ, X and ϵ as before the target $Y = X_1 + X_2 + \epsilon$ which now includes both regressors with equal weight. Figure 1.3 again shows the estimated coefficients corresponding to the sequence of values for the correlation ρ with step-size $\delta = 0.05$, as obtained by fitting an EN with $\lambda = \alpha = 0.5$. Within this setting, none of the correlations considered lies within the respective optimal interval (L_2, U_2) or (L_1, U_1). Consequently, none of the coefficients is set to zero while the other one is (significantly) different from zero. In general, we observe that for any value of ρ the estimated parameters do not differ much in size. Moreover, for $\rho \in [-1, -0.8]$, both coefficients are set to (almost) zero. The vertical lines indicate for which of those correlations the conditions for both coefficients to be zero (1.27) are satisfied (assuming a tolerance of 0.05 and 0.1 for the green and blue lines respectively).

Within the previous section, situations in which exactly one among the estimated coefficients of two correlated covariates is set to zero were characterized. Therefore it may happen that the coefficients corresponding to correlated regressors have differing signs and the bound on their difference in Theorem 1.21 can not be obtained. Nonetheless, scenarios in which the correlations satisfy the conditions for said phenomenon can not take arbitrarily high values. In particular, as a direct consequence of the continuity result 1.23, there will always be a certain threshold such that for $\rho \approx 1$ both estimates share the same sign. This can also observed in Simulation 1, where for any $\rho > 0.5$ neither one of the conditions (1.17)-(1.19) is met and the estimated coefficients' signs are matching.

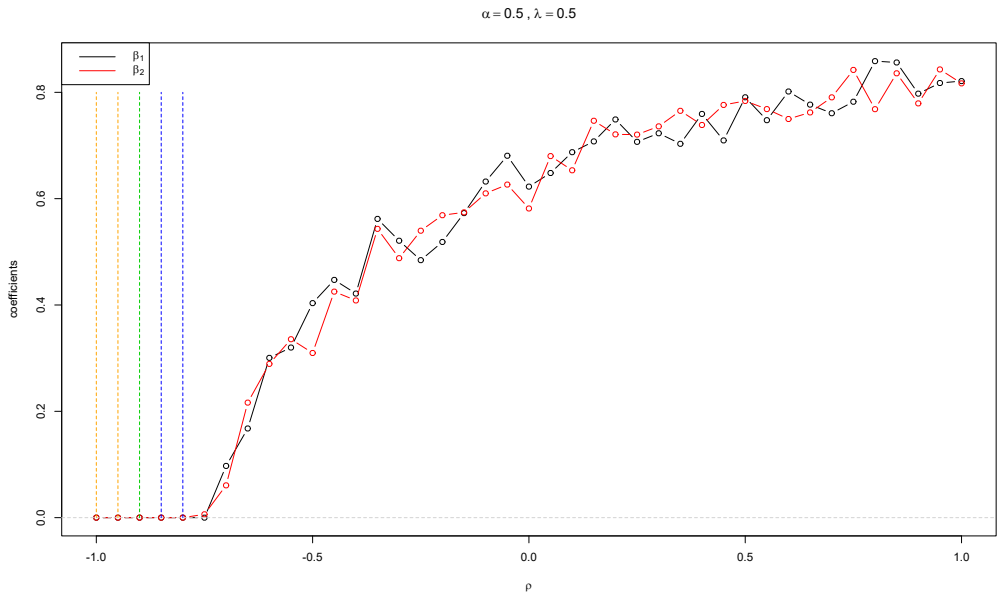

Figure 1.3: Estimated coefficients in dependence of the correlation in the explanatory variables. The orange vertical lines indicate correlations for which conditions (1.27) hold true. The green and blue lines show correlations for which those conditions are satisfied with a tolerance of 0.05 and 0.1 respectively.

For the pure Lasso penalty, an analogous bound as in Theorem 1.22 can not be easily obtained, in fact, as is claimed in Zou and Hastie (2005), the Lasso in general does not assign similar coefficients to correlated variables. A theoretical explanation is said to be obtained from Efron et al. (2004), that introduces a computationally feasible variable selection algorithm, a simple modification of which yields the entire Lasso solution path. In the following section, we collect the related findings given therein.

1.3.2 Least Angle Regression

Least angle regression is a model selection algorithm similar to forward stepwise regression, in which variable selection is less greedy. As in forward stepwise, least angle regression starts with the intercept, identifies the variable most correlated with the response, and moves the coefficient of this variable continuously toward its least squares estimate. The difference however being, that the largest step possible is taken until some other predictor has as much correlation with the current residual and enters the so-called active set. Then, the algorithm proceeds in a direction equiangular between all elements of the updated active set. In further steps, the algorithm sequentially

1. identifies the variable having as much correlation with the current residual as the variables in the active set and

2. moves the coefficient of this variable continuously along the updated equiangular direction until some other predictor joins the active set.

This way, only "as much information about a predictor as deserved" is added into the model. Moreover, the equiangular directions guarantee that coefficients of variables in the active set are moved in a way that keeps their correlations tied and decreasing.

Formally, a LARS step can be described as follows. Let $\mathcal{A}_{k-1} \subset \{1,\dots,p\}$ and $\hat{\mu}_{k-1}$ denote the active set and the LARS estimate at the end of the $k-1^{\text{th}}$ step respectively.

1. Considering the $p-$dimensional vector of so-called *current correlations*[1] $c = X^T(Y - \hat{\mu}_{k-1})$, we determine the new active set $\mathcal{A}_k$ as the index set of variables with maximal absolute correlation w.r.t. the current residual, i.e.

$$\mathcal{A}_k = \left\{ j \in \{1,\dots,p\} : |c_j| = \max_i |c_i| =: C \right\}.$$

2. The next update $\hat{\mu}_k$ of $\hat{\mu}_{k-1}$ is then obtained as

$$\hat{\mu}_k := \hat{\mu}_{k-1} + \hat{\gamma} u_{\mathcal{A}_k}, \quad \text{with} \quad u_{\mathcal{A}_k} := A_{\mathcal{A}_k} \cdot X_{\mathcal{A}_k}(X_{\mathcal{A}_k}^T X_{\mathcal{A}_k})^{-1}\mathbf{1}_{\mathcal{A}_k}, \tag{1.28}$$

with

- $n \times p$ matrix $X_{\mathcal{A}_k} = \left(\dots, s_j X^{(j)}, \dots\right)_{j\in\mathcal{A}_k}$ where $s_j = \text{sign}(c_j)$,
- normalizing constant $A_{\mathcal{A}_k} = \left(\mathbf{1}_{\mathcal{A}_k}^T (X_{\mathcal{A}_k}^T X_{\mathcal{A}_k})^{-1}\mathbf{1}_{\mathcal{A}_k}\right)^{-\frac{1}{2}}$ and
- $|\mathcal{A}_k|$-dimensional vector of ones $\mathbf{1}_{\mathcal{A}_k}$.

Before we clarify the value of $\hat{\gamma}$, we show that $u_{\mathcal{A}_k}$ as defined above is normed and equiangular w.r.t. the columns of $X_{\mathcal{A}_k}$.

Lemma 1.24 *The vector $u_{\mathcal{A}_k}$ of* (1.28) *confines equal angles less than* $90°$ *with $X_{\mathcal{A}_k}^{(j)}$, $j \in \mathcal{A}_k$. Moreover, it is normed.*

Proof. By

$$\begin{aligned} \|u_{\mathcal{A}_k}\|_2^2 &= A_{\mathcal{A}_k}^2 \left((X_{\mathcal{A}_k}(X_{\mathcal{A}_k}^T X_{\mathcal{A}_k})^{-1}\mathbf{1}_{\mathcal{A}_k})^T X_{\mathcal{A}_k}(X_{\mathcal{A}_k}^T X_{\mathcal{A}_k})^{-1}\mathbf{1}_{\mathcal{A}_k} \right) \\ &= A_{\mathcal{A}_k}^2 \left(\mathbf{1}_{\mathcal{A}_k}^T (X_{\mathcal{A}_k}^T X_{\mathcal{A}_k})^{-1}\mathbf{1}_{\mathcal{A}_k} \right) = 1, \end{aligned}$$

$u_{\mathcal{A}_k}$ is normed. Since $X_{\mathcal{A}_k}^T u_{\mathcal{A}_k} = A_{\mathcal{A}_k}\mathbf{1}_{\mathcal{A}_k}$, we have for each column $X_{\mathcal{A}_k}^{(j)}$, $j \in \mathcal{A}_k$ that ${X_{\mathcal{A}_k}^{(j)}}^T u_{\mathcal{A}_k} = A_{\mathcal{A}_k} \in (0,1)$. Hence the enclosed angle $\alpha := \arccos({X_{\mathcal{A}_k}^{(j)}}^T u_{\mathcal{A}_k}) \in (0, \frac{\pi}{2})$. □

The variable $\hat{\gamma}$ in (1.28) is defined as

$$\hat{\gamma} := \min_{j\in\mathcal{A}_k^c} \left\{ \frac{C - c_j}{A_{\mathcal{A}_k} - a_j}, \frac{C + c_j}{A_{\mathcal{A}_k} + a_j} \right\}^+, \quad \text{where} \quad a = X^T u_{\mathcal{A}_k}. \tag{1.29}$$

The $+$ indicates that the minimum is taken over positive components only. Defining $\hat{\gamma}$ as above, in fact yields the smallest possible value of $\gamma > 0$ such that some new index joins the active set. This result is shown next.

Lemma 1.25 *In the course of a LARS update as in* (1.28)*, the smallest possible step into the equiangular direction $u_{\mathcal{A}_k}$ such that a new variable joins the current active set is given by $\hat{\gamma}$, as defined in* (1.29)*.*

[1] Note that by the definition of c, its components c_j are *proportional to* the correlation between the covariate $X^{(j)}$ and the current residual vector $Y - \hat{\mu}_{k-1}$, meaning that $Cor(X^{(j)}, Y - \hat{\mu}_{k-1}) = \frac{c_j}{\|Y - \hat{\mu}_{k-1}\|_2}$.

Proof. Consider the updated LARS estimate in dependence of $\gamma > 0$

$$\mu_k(\gamma) = \hat{\mu}_{k-1} + \gamma u_{\mathcal{A}_k}.$$

The current correlations are then given by

$$c_j(\gamma) = X^{(j)^T}(Y - \mu_k(\gamma)) = c_j - \gamma a_j, \quad j \in \{1, \dots, p\}.$$

For $j \in \mathcal{A}_k$, we have that

$$|c_j(\gamma)| = |s_j C - \gamma \underbrace{X^{(j)^T} u_{\mathcal{A}_k}}_{=s_j A_{\mathcal{A}_k}}| = C - \gamma A_{\mathcal{A}_k}.$$

For $j \in \mathcal{A}_k^c$ we naturally have $|c_j(\gamma)| = |c_j - \gamma a_j| \leq C - \gamma A_{\mathcal{A}_k}$. Hence, $c_j(\gamma)$ equals its maximal value at $\gamma = (C - c_j)/(A_{\mathcal{A}_k} - a_j)$. Similarly, $-c_j(\gamma)$, the current correlation of $-X^{(j)}$, is maximal at $\gamma = (C + c_j)/(A_{\mathcal{A}_k} + a_j)$. □

As emphasized in Efron et al. (2004), LARS even permits highly correlated predictors to be included in the estimated model. In order to analyse the selection properties of the Lasso in correlated settings, we first give a modification of the LARS algorithm that yields the Lasso solution path.

The Lasso Modification

It was shown in (Efron et al., 2004, Section 3.1.), that via a simple adjustment to the LARS algorithm the full set of Lasso solutions can be obtained. Below, we collect the results given therein.
For some fixed $\lambda > 0$, consider the lasso estimate $\hat{\beta}^L$ as a solution to (1.2). Moreover, let $c_j = X^{(j)^T}(Y - X\hat{\beta}^L)$ be the correlation between the j$^{\text{th}}$ predictor and the current residual. As was derived in (Efron et al., 2004, Lemma 8), the sign of any non-zero component of $\hat{\beta}^L$ and the one of c_j must match, i.e.

$$\text{sign}(\hat{\beta}_j^L) = \text{sign}(c_j) =: s_j, \quad \forall j : \hat{\beta}_j^L \neq 0. \tag{1.30}$$

This restriction however, is not enforced by the LARS algorithm per se. Consider once more a LARS step as described earlier in this section. Suppose we have completed the $k-1^{\text{th}}$ step with active set $\mathcal{A}_{k-1}$ and LARS estimate $\hat{\mu}_{k-1}$ that now corresponds to a Lasso solution, i.e. $\hat{\mu}_{k-1} = X\hat{\beta}^L$. Assume further, that we have identified a covariate that joins the active set next, to give $\mathcal{A}_k$. The updated LARS estimate in dependence of the step size $\gamma > 0$ is then given by

$$\mu(\gamma) = \hat{\mu}_{k-1} + \gamma u_{\mathcal{A}_k} \overset{(1.28)}{=} X\hat{\beta}^L + \gamma X_{\mathcal{A}_k}\omega_{\mathcal{A}_k}, \tag{1.31}$$

with $\omega_{\mathcal{A}_k} = A_{\mathcal{A}_k}(X_{\mathcal{A}_k}^T X_{\mathcal{A}_k})^{-1}\mathbf{1}_{\mathcal{A}_k}$ a $|\mathcal{A}_k|$-dimensional vector. Defining p-dimensional vectors d and $\beta(\gamma)$ as

$$d_j := \begin{cases} s_j(\omega_{\mathcal{A}_k})_j & j \in \mathcal{A}_{k-1}, \\ 0 & \text{else}, \end{cases} \quad \text{and} \quad \beta_j(\gamma) := \hat{\beta}_j^L + \gamma d_j,$$

equation (1.31) can be further refined to give

$$\mu(\gamma) = X\beta(\gamma).$$

By definition, some component β_j will change sign at $\gamma_j = -\hat{\beta}_j^L/d_j$. The first such change will occur at

$$\tilde{\gamma} := \min_{\gamma_j > 0} \gamma_j,$$

whereby $\tilde{\gamma}$ is set to infinity whenever $\{j \in \{1, \dots, p\} : \gamma_j > 0\} = \emptyset$.
Recall that the sign s_j remains unchanged within a single LARS step. Thus, whenever the minimal step size for which the coefficient of some covariate (say $X^{(\tilde{j})}$) changes sign ($\tilde{\gamma}$) is less than the one taken by a LARS step ($\hat{\gamma}$), the sign of $\beta_{\tilde{j}}(\gamma)$ changes while the one of $c_{\tilde{j}}$ does not. Hence for $\gamma > \tilde{\gamma}$ the sign restriction (1.30) is violated and $\beta(\gamma)$ can not be a Lasso solution. This motivates the following Lasso modification:

2. (Lasso modification): If $\tilde{\gamma} < \hat{\gamma}$ with minimizing index $\tilde{j}$, instead of (1.28) set

$$\hat{\mu}_k := \hat{\mu}_{k-1} + \tilde{\gamma} u_{\mathcal{A}_k} \quad \text{and} \quad \mathcal{A}_k := \mathcal{A}_{k-1} \backslash \{\tilde{j}\}.$$

In other words, within the Lasso adaptation, if a non-zero coefficient becomes zero, the corresponding variable is dropped from the active set and the equiangular direction is recomputed.
Within the original LARS framework, the active set $\mathcal{A}$ grows monotonically larger as the algorithm progresses. The Lasso modification introduced above allows $\mathcal{A}$ to decrease as well. In order for the modified algorithm to yield the entire solution path of the Lasso, one needs the additional assumption that those increases and decreases involve one single index only. Below we restate the corresponding finding from Efron et al. (2004).

Theorem 1.26 *Under the Lasso modification, assuming that changes to the active set emerge by the inclusion or exclusion of a single index only, the LARS algorithm yields all Lasso solutions.*

By the above theorem, the modified LARS algorithm can be considered to view Lasso's selection properties from a different angle. Among correlated variables, the LARS-Lasso algorithm in fact first selects the one that shows higher correlation with the current residual. This may indeed happen at random as, depending on the noise, either regressor could appear to be said candidate. The algorithm then proceeds to obtain as much information as possible from the chosen predictor, leaving the other correlated predictor's coefficient at zero. Only for smaller penalty parameter will it be included in the model. This is different to the behaviour of the EN and even more so to the pure ℓ_2-penalized optimisation, in which the coefficients of correlated predictors are shrunk towards each other. Thereby, covariates that are not included in the true model may borrow strength from relevant correlated variables.

This instability of the Lasso has often been criticised. In fact, the property of the EN to increase stability of estimated coefficients was ranked among the method's main advantages in its introductory paper Zou and Hastie (2005). Besides the EN, many alternatives to the Lasso have been introduced each of which are proven to enhance stability of variable selection. As such an alternative, Meinshausen and Bühlmann (2010) introduces *Stability Selection* being a generic method to stabilize and improve variable selection in regression. This method yields a promising approach to improve stability of the Lasso in linear models, however it can be applied to more general settings and introduces a novel possibility to perform variable selection in penalized GLMs for any kind of regularisation that introduces sparsity. In particular, we will analyse Stability Selection for the Elastic Net in generalized linear models in chapter 3.

The details on Stability Selection shall be covered in chapter 2. Thereafter, we turn to the analysis of variable selection in generalized linear models in chapter 3. For said framework existing findings are collected and an application of Stability Selection is discussed.

Chapter 2

Stability Selection

In this chapter, we elaborate on *Stability Selection*, a generic subsampling approach to perform regularisation for structure estimation introduced in Meinshausen and Bühlmann (2010). Although the technique is applicable to estimation of discrete structure in a number of different contexts, we focus on its application to variable selection in regression.

2.1 The method

For the ensuing analysis we consider the following setting. We assume to be given data $Z^{(1)}, \ldots, Z^{(n)}$ as i.i.d. realisations of $Z = (X, Y)$, with p-dimensional covariate X and univariate response Y. Let β be a p-dimensional unknown vector of parameters that the variable of interest Y depends on[1]. We further assume that β is sparse, with $q < p$ non-zero components. We denote the set of indices corresponding to these non-zero ('true') and vanishing ('redundant') components by $I^* := \{k \in \{1, \ldots, p\} : \beta_k \neq 0\}$ and $N := \{k \in \{1, \ldots, p\} : \beta_k = 0\}$ respectively.
Furthermore, in a generic variable selection technique, we have a tuning parameter $\lambda \in \Lambda \subseteq \mathbb{R}^+$ that determines the amount of regularisation. For any $\lambda \in \Lambda$ a universal selection method yields a structure estimate $\hat{I}^\lambda \subseteq \{1, \ldots, p\}$. Note, that for any regularisation parameter $\lambda \in \Lambda$ the selected set $\hat{I}^\lambda$ depends on the sample $S = \{1, \ldots, n\}$ considered for estimation. To avoid complex notation, we only refer to this dependence and write $\hat{I}^\lambda = \hat{I}^\lambda(S)$ when necessary. The declared aim is then to determine whether there exists a $\lambda \in \Lambda$ such that $\hat{I}^\lambda = I^*$ with high probability and to identify this corresponding regularisation parameter.
To that end, Meinshausen and Bühlmann (2010) start by defining the notion of *stability paths*, that are *the probability for each variable $X^{(1)}, \ldots, X^{(p)}$ to be selected when randomly subsampling from the data.* Just as a regularisation path consists of the values of the estimated coefficients over all regularisation parameters $\lambda \in \Lambda$, i.e. $\left\{\hat{\beta}_k^\lambda, \lambda \in \Lambda, k = 1, \ldots, p\right\}$, the stability path shows the probability of being selected (see definition below) for each variable over all possible $\lambda \in \Lambda$, i.e. $\left\{\hat{\Pi}_k^\lambda, \lambda \in \Lambda, k = 1, \ldots, p\right\}$.

Definition 2.1 (Selection probabilites) *Let S be a random subsample of $\{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$, drawn without replacement, and $\lambda \in \Lambda$ be some penalisation parameter. For every set*

[1] In general, any kind of regression can be thought of here. We like to keep in mind representation (1), as it will be the setting in chapter 3.

$K \subseteq \{1, \ldots, p\}$, *the probability of being selected is*

$$\hat{\Pi}_K^\lambda := \mathbb{P}\left(K \subseteq \hat{I}^\lambda\right).$$

Note that the probability $\mathbb{P}$ in Definition 2.1 accounts for all possible sources of randomness that co-determine $\hat{I}^\lambda$ including e.g. the random subsampling as well as possibly random algorithms.
With the above, one could also define paths for groups of variables. Considering, for instance, the probability of at least one variable being selected among a group of related covariates per penalisation parameter $\lambda \in \Lambda$.
The idea of Stability Selection then is motivated as follows: in traditional variable selection methods, one model (one value for the penalty parameter λ) is usually chosen from the set

$$\{\hat{I}^\lambda, \lambda \in \Lambda\} \tag{2.1}$$

according to some optimality criterion. This can be problematic in two regards. First, the correct model I^* might not be included in (2.1). Second, even if the correct model is included, it is not clear how to determine the corresponding correct amount of penalisation λ^* or the amount necessary to select a close approximation of I^*.
By contrast, with Stability Selection one does not select a model from (2.1). Instead the data are perturbed (e.g. via subsampling) a number of times, and for each of these perturbed sets the selection method in question is performed. Those variables that occur in a large fraction of the resulting selection sets are then finally selected. Definition 2.2 gives a more precise characterisation of the variables that are chosen with Stability Selection.

Definition 2.2 *Consider a set of regularisation parameters Λ and some cutoff $\pi_{thr} \in (0, 1)$. The set of stable variables is defined as*

$$\hat{I}^{stable} := \left\{k \in \{1, \ldots, p\} : \max_{\lambda \in \Lambda} \hat{\Pi}_k^\lambda \geq \pi_{thr}\right\}.$$

In other words, a variable is kept in the model, if there exists at least one penalty parameter among the ones considered, such that its probability of being selected exceeds a certain threshold. Variables that exhibit low selection probabilities for all of the penalty parameters in question are not considered. The threshold π_{thr} is a tuning parameter that needs to be optimized. However, as Meinshausen and Bühlmann (2010) show, the results vary little for sensible choices of thresholds within a certain range.

2.2 Choice of regularisation and error control

In (Meinshausen and Bühlmann, 2010, Section 2.4), finite sample error control is addressed. In particular it is shown how to arrive at the correct amount of regularisation such that the expected number of falsely selected variables can be conservatively controlled. The aim of this section is to summarize these findings. We start by providing the notation necessary for the analysis.

Definition 2.3 *The set of selected variables when varying the regularisation parameter λ in the region Λ be defined as*

$$\hat{I}^\Lambda := \bigcup_{\lambda \in \Lambda} \hat{I}^\lambda.$$

Denote further the average number of selected variables by $q_\Lambda := \mathbb{E}(|\hat{I}^\Lambda(S)|)$*. The number of falsely selected variables with Stability Selection is given by*

$$V = \left| N \cap \hat{I}^{stable} \right|.$$

We are now interested in controlling $\mathbb{E}[V]$. In general however, the distribution of the estimator $\hat{\beta}$ is not readily available, as it depends on many unknown quantities. Under simplifying assumptions, Meinshausen and Bühlmann (2010) arrive at the following:

Theorem 2.4 (Error control) *Assume that the distribution of* $\left\{\mathbb{1}_{k\in\hat{I}^\lambda}, k \in N\right\}$ *is exchangeable for any* $\lambda \in \Lambda$*, i.e. that for any* $\lambda \in \Lambda$ *it holds that*

$$\left\{\mathbb{1}_{k\in\hat{I}^\lambda}, k \in N\right\} \overset{(d)}{=} \left\{\mathbb{1}_{k\in\hat{I}^\lambda}, k \in \sigma(N)\right\},$$

for all finite permutations σ*. Moreover, assume that the original selection method is not worse than random guessing, that is*

$$\frac{\mathbb{E}(|I^* \cap \hat{I}^\lambda|)}{\mathbb{E}(|N \cap \hat{I}^\lambda|)} \geq \frac{\mathbb{E}(|I^*|)}{\mathbb{E}(|N|)}.$$

The expected number of falsely selected variables is then bounded by

$$\mathbb{E}[V] \leq \frac{1}{2\pi_{thr} - 1} \frac{q_\Lambda^2}{p}. \tag{2.2}$$

Remark 2.5 *The expected number of falsely selected variables* $\mathbb{E}[V]$ *sometimes is referred to as* per-family error rate (PFER) *in multiple testing. Similarly, the* per-comparison error rate (PCER) *in multiple testing is given by* $\mathbb{E}[V]/p$*.*

Meinshausen and Bühlmann (2010) further elaborate on the assumptions required for the error bound (2.2). First, it is required for the original procedure to be no worse than random guessing, a quite weak assumption, that, as is argued, should be fulfilled anyhow. Second and more importantly, exchangeability of $\mathbb{1}_{k\in\hat{I}^\lambda}$ for $k \in N$ is needed. As Meinshausen and Bühlmann (2010) claim, this assumption is fulfilled for any standard variable selection technique in linear regression if the design X is random and the distribution of $(X_k)_{k\in N}$ is exchangeable. In particular, the latter is satisfied for $(X_k)_{k\in N}$ independent. Another setting in which exchangeability holds is $(X_{k_1}, \ldots, X_{k_{p-q}}) \sim \mathcal{N}_{p-q}$ with $Cov(X_k, X_l) = \rho$ for all $k \neq l, k, l \in N$ with $\rho \in (0, 1)$. In general however, for real data or more general models, there is no guarantee that the assumption of exchangeability is satisfied. Nonetheless, Meinshausen and Bühlmann (2010) show that the bound (2.2) holds up quite well for real data in linear models.

As already mentioned, the threshold parameter π_{thr} has to be tuned. Meinshausen and Bühlmann (2010) show however, that its influence is very small, that is for reasonable values such as $\pi_{thr} \in (0.6, 0.9)$ results tend to be similar. Once the value of the cutoff π_{thr} is fixed, one may choose the region of regularisation parameters Λ such that the error bound (2.2) is of some desired form. More precisely, choosing Λ such that $q_\Lambda = \sqrt{(2\pi_{thr} - 1)p}$ will control $\mathbb{E}[V] \leq 1$. Conversely, one could first fix the regularisation region Λ and then proceed to choose π_{thr} accordingly to achieve the desired error bound.

In order to perform the tuning of the parameters π_{thr} and Λ as suggested, knowledge of the generally unknown quantity q_Λ is necessary. Depending on the selection procedure however, the number s of selected samples may depend on the domain of regularisation in some way that is $s = s(\Lambda)$.

With ℓ_1 or $\ell_1 + \ell_2$ penalties, the number s is given by the number of variables that enter the regularisation path for $\lambda \in \Lambda$ $s = \bigcup_{\lambda \in \Lambda} \hat{I}^\lambda$.

2.3 Computational requirements and pointwise control

With Stability Selection, it is necessary to perform a given selection method for every $\lambda \in \Lambda$ multiple times, depending on the number of subsamples considered. In practice, 100 subsamples seem sufficient to evaluate selection probabilities according to Meinshausen and Bühlmann (2010). For the example of the Lasso with $p > n$ it is shown that Stability Selection is roughly three times more expensive than 10-fold cross validation. For $p < n$ this factor even increases to approximately 5.5.

In some cases, repeated evaluation of $\hat{I}^\lambda$ for different subsets can already be computationally demanding for a single value of $\lambda \in \Lambda$. The results from section 2.2 immediately apply to $\Lambda = \lambda$. For selection methods that incrementally select structures, i.e. for which $\hat{I}^\lambda \subseteq \hat{I}^{\lambda'}$ for all $\lambda \geq \lambda'$ pointwise control is equivalent to setting $\Lambda := [\lambda, \infty)$, as $\hat{\Pi}^\lambda$ is monotonically increasing as λ decreases. Meinshausen and Bühlmann (2010) argue that pointwise control is especially successful if λ is chosen such that the set $\hat{I}^\lambda$ is rather to large, in that with high probability $I^* \subseteq \hat{I}^\lambda$.

2.4 Consistent variable selection for the Lasso

As is successfully argued in (Meinshausen and Bühlmann, 2010, Chapter 3) by the example of the linear model, Stability Selection can markedly improve consistent variable selection. Below, we collect their results. Throughout this section, we resume the setting of Chapter 1, i.e.

$$\boxed{Y_i \overset{id}{\sim} \mathcal{N}(\mu_i, \sigma^2), \quad \mu_i = \beta_0 + \beta^T X_i, \qquad i = 1, \ldots, n.}$$

with $\sigma^2 > 0$ and normalized predictors collected in the $n \times p$-matrix X. In particular, $p \gg n$ is allowed for.

In Section 1.1 of Chapter 1 we elaborated on necessary and sufficient conditions for the Lasso to consistently select the correct model. This boiled down to the need for irrepresentable conditions (1.5) and (1.4), that put strong restrictions on the correlations among the covariates. As Meinshausen and Bühlmann (2010) show, these conditions can be circumvented by using Stability Selection. Moreover, it is argued that "adding some extra randomness" to the optimisation can lead to consistent variable selection even though the ICs are violated. To achieve this, Meinshausen and Bühlmann (2010) propose the *randomised Lasso*, a new generalisation of standard Lasso that shall be introduced next.

Definition 2.6 (Randomised Lasso) *The randomised Lasso with weakness $\alpha \in (0, 1]$ is a regularisation method for which one solves*

$$\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} \frac{|\beta_k|}{W_k}, \tag{2.3}$$

with W_k i.i.d. random variables in $[\alpha, 1]$ for $k = 1, \ldots, p$ and penalisation parameter $\lambda \in \mathbb{R}$.

While standard Lasso penalizes the absolute value of every component β_k using the same penalty-parameter λ, randomised Lasso applies a custom penalty randomly chosen from $[\lambda, \lambda/\alpha]$. In Meinshausen and Bühlmann (2010) the following distribution for the weights W_k is proposed and shall be assumed throughout this section: W_k is sampled as $W_k = \alpha$ with some probability $p_W \in (0,1)$ and $W_k = 1$ otherwise. Since with $D = \text{diag}(W_1, \dots, W_p)$

$$\begin{aligned} & \min_{\beta \in \mathbb{R}^p} \quad \|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} \frac{|\beta_k|}{W_k} \\ \iff \quad & \min_{\beta \in \mathbb{R}^p} \quad \|Y - XDD^{-1}\beta\|_2^2 + \lambda \sum_{k=1}^{p} \left| \frac{\beta_k}{W_k} \right| \\ \iff \quad & \min_{\tilde{\beta} \in \mathbb{R}^p} \quad \|Y - XD\tilde{\beta}\|_2^2 + \lambda \sum_{k=1}^{p} |\tilde{\beta}| \end{aligned}$$

(2.3) is solved using algorithms for the standard Lasso on the re-scaled predictors XD. As Meinshausen and Bühlmann (2010) surprisingly show, this *random* re-scaling can be very powerful when combined with resampling. The condition for selection consistency derived therein makes use sparse eigenvalues, a notion that is introduced below.

Definition 2.7 *For some $K \subseteq \{1, \dots, p\}$, let $X_K = (X^{(i)})_{i \in K}$ be the restriction of X to columns in the index set K. The minimal sparse eigenvalue ϕ_{min} is then defined for $k \leq p$ as*

$$\phi_{min}(k) = \inf_{a \in \mathbb{R}^{\lceil k \rceil}, K \subseteq \{1,\dots,p\} : |K| \leq \lceil k \rceil} \frac{\|X_k a\|_2}{\|a\|_2}.$$

Similarly, the maximal sparse eigenvalue ϕ_{max} is for $k \leq p$ given by

$$\phi_{min}(k) = \sup_{a \in \mathbb{R}^{\lceil k \rceil}, K \subseteq \{1,\dots,p\} : |K| \leq \lceil k \rceil} \frac{\|X_k a\|_2}{\|a\|_2}.$$

In order to give the main result, we need the following assumption:

Assumption 2.8 (Sparse eigenvalues) *There exists some $C > 1$ and some $\kappa \geq 9$ such that*

$$\frac{\phi_{max}(Cq^2)}{\phi_{min}^{3/2}(Cq^2)} < \frac{\sqrt{C}}{\kappa}, \quad q = |I^*|.$$

Within this chapter's setting, that is for the linear model, Meinshausen and Bühlmann (2010) then obtain the following result on variable selection for the randomised Lasso (cp. (Meinshausen and Bühlmann, 2010, Theorem 2)). It is shown, that for specific choice of weakness α Stability Selection with the randomised Lasso satisfies that no noise variables are selected for λ exceeding some specified threshold $\lambda_{\min}$ on a set with probability tending to one as the sample size increases. That is there exists some $\delta = \delta_q \in (0,1)$ such that for all $\pi_{thr} \geq 1 - \delta$ one has

$$N \cap \hat{I}_\lambda^{stable} = \emptyset,$$

with $\hat{I}_\lambda^{stable} = \{k : \hat{\Pi}_k^\lambda \geq \pi_{thr}\}$, $\lambda \geq \lambda_{\min}$ (on a set Ω_A with $\mathbb{P}(\Omega_A) \underset{n \to \infty}{\to} 1$). Moreover, on the same set Ω_A, all variables with sufficiently large coefficient are selected, i.e.

$$(I^* I^*_{small,\lambda}) \subseteq \hat{I}_\lambda^{stable},$$

where $I^*_{small,\lambda} = \{k : |\beta_k| \leq 0.3(Cq)^{3/2}\lambda\}$.

Remark 2.9

- *Consistent variable selection can be achieved for the randomised Lasso, even if irrepresentable condition 1.4 is violated. Provided that the minimal non-zero coefficient is bounded from below by* $\min_{k\in I^*} |\beta_k| \geq (Cq)^{3/2}0.3\lambda$ *one has*

$$\mathbb{P}\left(I^* = \hat{I}^{stable}_{\lambda}\right) \longrightarrow 1, \qquad as\ p \to \infty\ or\ n \to \infty.$$

  If such a lower bound does not hold, the method might miss variables of the set $I^*_{small,\lambda}$.

- *Randomised Lasso with Stability Selection is rather conservative towards selecting false positives.(Meinshausen and Bühlmann, 2010, Theorem 2) holds for all* $\lambda \geq \lambda_{min}$*, and hence even if* λ *is chosen considerably larger than* λ_{min}*, no noise variables will be selected.*

Within a simulation example the enhanced selection accuracy of randomised Lasso is demonstrated. Moreover, (Meinshausen and Bühlmann, 2010, Chapter 4) numerically assesses the effects of Stability Selection. In various settings, the probability of selecting a certain proportion of the true variables while leaving the coefficients of all redundant variables at zero is determined for regression and classification combined with each of pure Lasso, Lasso with Stability Selection and randomised Lasso with Stability Selection. The analysed scenarios include different structures of correlations. In all cases, Stability Selection was able to identify at least as many of the correct variables as the underlying methods themselves. In fact, most of the probabilities resulting from Stability Selection surpassed the ones obtained by the ordinary procedures, only for independent predictor variables were the approaches almost equivalent.
This promising result for the Lasso in linear models serves as motivation to deploy Stability Selection jointly with EN-penalisation in generalized linear models. In the following chapter, we like to numerically assess by means of a simple simulation, whether Stability Selection leads to improved variable selection within the framework of EN-penalized logistic regression. Prior to this simulation, the first part of the ensuing chapter aims at collecting existing findings on variable selection in GLMs. The particular choice of logistic regression for the simulation study is on one hand due to its frequent use in predictive modelling, and on the other, most of the results on variable selection in GLMs cover logistic regression only.

Chapter 3

Variable selection in GLMs

While the literature is rich on variable selection properties of $\ell_1+\ell_2$-penalized linear models, hardly any theoretical results on EN-penalized generalized linear models are to be found. Studying the grouping effect in a more general set-up, with non-quadratic loss functions quickly becomes more intricate. In particular, obtaining a similar bound on the difference of estimated coefficients in dependence of their correlation as in Section 1.3 is not as straightforward as it is for quadratic loss. While Lemma 1.21 still is applicable to GLMs, it is unclear how to obtain conditions on the correlations that characterize scenarios in which one or both of two correlated regressors is set to zero as in Section 1.3.1. Nonetheless, theoretical results on selection consistency have been obtained in Bunea (2008) for specifically ℓ_1- and $\ell_1 + \ell_2$-penalized logistic regression. These findings will be subsequently summarized. Moreover, we will follow a simulation-based approach to assess the variable selection properties for the specific choice of logistic regression in settings with different covariance structures in the design. Below, we briefly give the definitions of the Lasso and EN estimates in the GLM framework.
Recall the GLM setting (1) given in the introduction:

$$\boxed{Y_i \overset{id}{\sim} f_Y, \quad g(\mu_i) = \beta_0 + \beta^T X_i, \quad \mu_i \equiv \mathbb{E}[Y_i] \qquad i = 1, \ldots, n,}$$

with g a suitable link function and density f_Y a member of the exponential dispersion family. The Lasso estimator in this general framework is obtained as solution to

$$\min_\beta \ell(\beta; X, Y) + \lambda \|\beta\|_1,$$

for some $\lambda > 0$, where ℓ denotes the negative log-likelihood function corresponding to density f_Y. Likewise, the EN estimator minimizes

$$\min_\beta \ell(\beta; X, Y) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2,$$

with non-negative penalty parameters λ_1, λ_2. Equivalently, this minimisation can be re-parametrized as

$$\min_\beta \ell(\beta; X, Y) + \lambda \left((1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2\right),$$

with $\alpha \in [0, 1]$, $\lambda \geq 0$.

3.1 Model selection consistency for logistic regression

Similarly to Chapter 1, we start noting the findings on selection consistency. As existing literature focuses on logistic regression within this framework, we restrict ourselves to the following setting:

$$\boxed{Y_i \overset{id}{\sim} Ber(\mu_i), \quad g(\mu_i) = \beta_0 + \beta^T X_i, \quad \mu_i \equiv \mathbb{E}[Y_i] \qquad i = 1, \dots, n,}$$

with link function

$$g : (0,1) \to \mathbb{R}, \quad g(x) = \log\left(\frac{x}{1-x}\right).$$

The negative log-likelihood is in this case given by

$$\ell(\beta; X, Y) = \frac{1}{n}\sum_{i=1}^{n} \big(- Y_i X_i^T \beta + \log\left(1 + \exp\left(X_i^T \beta\right)\right) \big).$$

As in Section 1.2, we will in what follows collect the results on accuracy of variable selection in finite samples including asymptotic results for ℓ_1- and $\ell_1 + \ell_2$-penalized logistic regression as presented in Bunea (2008). Recall, that the setting therein allows for both the number of regressors p and the index set of true covariates I^* to be dependent on the sample size n. We again note, that the notation will not account for this dependence. Moreover, we once more require the covariates to be a.s.-bounded by a common constant, i.e. we assume:

Assumption 3.1 *There is a constant $L > 0$ such that $|X_{ij}| < L$ for all i, j with probability 1.*

Sparse ℓ_1-balls

In what follows, let $\hat{\beta}^L$ and $\hat{\beta}^{EN}$ be the Lasso- and EN-penalized GLM-estimates and β be the true parameter vector. Analogously to Section 1.2, we start by giving conditions for which the ℓ_1-ball $\|\hat{\beta}^L - \beta\|$ can be bounded by a quantity that only depends on the (unknown) number of truly non-zero coefficients k^*. To that end, we further assume that the true coefficient is ℓ_1-bounded:

Assumption 3.2 *There exists some $D > 0$ such that $\|\beta\|_1 < D$.*

Assumptions 3.1 and 3.2 in particular imply that $\mu_i(X_i)$ is bounded away from zero and one for all realisations X_i and $i = 1, \dots, n$.

Theorem 3.3 *Assume Condition 1.15 corresponding to $V_{\alpha,\epsilon}$ with*

$$\epsilon = \frac{\log(2)}{2^{(p \vee n)+1}} \frac{2}{\lambda}$$

and $\alpha = 3$ is satisfied for some $0 < b \leq 1$. If we choose

$$\frac{\lambda}{2} \geq (6 + 4\sqrt{2}) L \sqrt{\frac{2\log(2(p \vee n))}{n}} + 2L\sqrt{\frac{2\log(\frac{1}{\delta})}{n}} + \frac{1}{4(p \vee n)},$$

then the Lasso estimate $\hat{\beta}^L$ satisfies

$$\mathbb{P}\left(\|\hat{\beta}^L - \beta\|_1 \leq \frac{2}{sb}\lambda k^* + \left(1 + \frac{2}{\lambda}\right)\epsilon \right) \geq 1 - \delta.$$

Here, $s = s(L, D)$ is a constant which decreases with D.

Once more, the bound on $\|\hat{\beta}^L - \beta\|$ becomes large, if Condition Stabil is satisfied for small values of b. This serves as motivation to study the $\ell_1 + \ell_2$-penalty in the framework of logistic regression as well. Moreover, Bunea (2008) remark that for large values of D, the constant $\frac{1}{s}$ can be very large too, leading to a less stringent upper bound. It is shown that an improved bound can be obtained at least within an asymptotic statement, where $\frac{1}{s}$ is replaced by a constant that is arbitrarily close to one. To obtain this asymptotic result, a condition needs to be fulfilled that in essence requires Condition 1.15 to hold for a weighted version of C^n with

$$C^n_{kj} := \frac{1}{n}\sum_{i=1}^{n} X_{ki}X^T_{ij}, \quad 1 \leq j,k \leq p.$$

Consider the weighting function

$$g : \mathbb{R} \to [0,1], \quad g(z) = \frac{\exp(z)}{1+\exp(z)}.$$

Define further the $p \times p$-matrix $\tilde{C}^n$ via

$$\tilde{C}^n_{kj} := \frac{1}{n}\sum_{i=1}^{n} g'(X_i^T\beta)X_{ki}X_{ij}, \quad 1 \leq j,k \leq p.$$

Condition 3.4 (Condition LStabil) *Let $\alpha > 0, \beta > 0$ be given. There exists $0 < b \leq 1$ such that*

$$\mathbb{P}\left(v^T\tilde{C}^n v \geq b\sum_{j\in I^*} v_j^2 - \epsilon\right) = 1, \quad \forall v \in V_{\alpha,\epsilon}.$$

Finally, the following was shown (cp. (Bunea, 2008, Theorem 2.6.)):

Theorem 3.5 *Let Condition 3.4 be satisfied for some $0 < b \leq 1$ and $\lambda > 0$ such that*

$$\frac{\lambda}{2} \geq (6+4\sqrt{2})L\sqrt{\frac{2\log(2(p\vee n))}{n}} + 2L\sqrt{\frac{2\log(\frac{1}{\delta_n})}{n}} + \frac{1}{4(p\vee n)},$$

for any sequence $\delta_n \underset{n\to\infty}{\longrightarrow} 0$. If there is some $B > 0$ independent of n such that $\max_{j\in I^*}|\beta_j| \leq B$ and $\lambda k^* \longrightarrow 0$, then

$$\mathbb{P}\left(\|\hat{\beta}^L - \beta\|_1 \leq \frac{2}{wb}\lambda k^* + \left(1+\frac{2}{\lambda}\right)\epsilon\right) \underset{n\to\infty}{\longrightarrow} 1,$$

for a constant w arbitrarily close to one.

Similar to Section 1.2, for the EN-estimate Bunea (2008) obtain a bound on $\|\hat{\beta}^{EN} - \beta\|$ that is less affected by small values of b and s. Note, that we still require Assumptions 3.1 and 3.2 to hold true.

Theorem 3.6 *Assume Condition 1.15 corresponding to $V_{\alpha,\epsilon}$ with*

$$\epsilon = \frac{\log(2)}{2^{(p\vee n)+1}}\frac{2}{\lambda}$$

and $\alpha = 4$ is satisfied for some $0 < b < 1$. Let $s > 0$ be the constant given in Theorem 3.3.

Furthermore, let $B > 0$ *such that* $\max_{j \in I^*} |\beta_j| \leq B$ *and take*

$$\frac{\lambda_1}{2} \geq (6 + 4\sqrt{2})L\sqrt{\frac{2\log(2(p \vee n))}{n}} + 2L\sqrt{\frac{2\log(\frac{1}{\delta})}{n}} + \frac{1}{4(p \vee n)}, \quad \lambda_2 = \frac{\lambda_1}{4B}.$$

Then the EN estimate $\hat{\beta}^{EN}$ *satisfies*

$$\mathbb{P}\left(\|\hat{\beta}^{EN} - \beta\|_1 \leq \frac{2.125}{sb + \lambda_2}\lambda_1 k^* + \left(1 + \frac{2}{\lambda_1}\right)\epsilon\right) \geq 1 - \delta.$$

Note that here too, the penalisation parameter λ_2 corresponding to the ℓ_2-norm should not be chosen too big, to prevent losing sparsity in the resulting estimate. Moreover, Bunea (2008) remark, that asymptotic results almost identical to 3.5 can be obtained for the EN-estimate as well.

Correct inclusion of true variables

The introductory remarks in Section 1.2 including Lemma 1.13 apply in the same manner for GLMs. Having specified upper bounds on both $\|\hat{\beta}^L - \beta\|_1$ and $\|\hat{\beta}^{EN} - \beta\|_1$, Bunea (2008) obtain the following corollary that gives conditions on when with high probability all true variables are included in the estimated index-sets $\hat{I}^L$ and $\hat{I}^{EN}$ respectively (cp. Corollary 1.18).

Corollary 3.7 *Let* $0 < \gamma_1 < 1$ *be fixed. Assume further Condition 1.15 holds for the parameters specified in Theorems 3.3-3.6.*

- ℓ_1*-penalty:*
  If $\min_{j \in I^*} |\beta_j| \geq \frac{4}{sb}\lambda k^* + \left(1 + \frac{2}{\lambda}\right)\epsilon$ *with* s, λ *and* ϵ *as in Theorem 3.3, and if Assumptions 3.1 and 3.2 hold true, then* $\mathbb{P}(I^* \subseteq \hat{I}^L) \geq 1 - \gamma_1$.

- $\ell_1 + \ell_2$*-penalty:*
  If there exists some $B > 0$ *such that*

  $$\frac{2.125}{sb + \lambda_2}\lambda_1 k^* + \left(1 + \frac{2}{\lambda_1}\right)\epsilon \leq \min_{j \in I^*} |\beta_j| \leq \max_{j \in I^*} |\beta_j| \leq B$$

  with λ_1, λ_2 *and* ϵ *chosen as in Theorem 3.6, and if Assumptions 3.1 and 3.2 hold true, then* $\mathbb{P}(I^* \subseteq \hat{I}^{EN}) \geq 1 - \gamma_1$.

Up to a small additive ϵ-term, the lower bounds on the minimum size of the true coefficients are of the form $C\lambda k^*$ ($C\lambda_1 k^*$) for some constant C, just as in Corollary 1.18. Thus, the previous remarks apply here as well.
Moreover, as Bunea (2008) show, the lower bound on the minimum signal size can also be improved within the framework of logistic regression. To do so, one requires an adjusted identifiability condition. Similarly to when Condition Stabil was replaced by Condition LStabil, we now require that a weighted correlation matrix exhibits the separation properties introduced for the ordinary correlation matrix in Condition Identif.

In what follows, let $\lambda > 0$[1], $\epsilon > 0$ be such that

$$\frac{\lambda}{2} \geq (6+4\sqrt{2})L\sqrt{\frac{2\log(2(p\vee n))}{n}} + 2L\sqrt{\frac{2\log(\frac{2p}{\delta})}{n}} + \frac{1}{4(p\vee n)}, \tag{3.1}$$

$$\epsilon = \frac{\log(2)}{2^{(p\vee n)+1}}\frac{2}{\lambda}, \tag{3.2}$$

for $0 < \delta \leq 1$, p and n given. Let further d be as required by Condition 1.14. By (Bunea, 2008, Lemma 2.1.), for any such $d \in (0,1)$, there exists some $b \in (0,1)$ for which Condition 1.15 holds. For this $b \in (0,1)$ and $s > 0$ as in Theorem 3.3 define the set

$$U := \left\{ a \in \mathbb{R}^n : \max_{1\leq i\leq n} \left\| a_i - \beta^T X_i \right\| \leq \frac{2\lambda L k^*}{sb} + L\left(1+\frac{2}{\lambda}\right)\epsilon \right\}$$

The condition then reads as folows:

Condition 3.8 (Condition Lidentif) *Let d be the constant required by Condition 1.14 and consider the weighting function*

$$g : \mathbb{R} \to [0,1], \quad g(z) = \frac{\exp(z)}{1+\exp(z)}.$$

We assume that

$$\sup_{a\in U} \mathbb{P}\left(\max_{j\in I^*, k\neq j} |\frac{1}{n}\sum_{i=1}^{n} g'(a_i) X_{ij} X_{ik}| \leq \frac{d}{k^*} \right) = 1.$$

Note that if the weighting function g is the linear link, i.e. if $g(x) = x$ for $x \in \mathbb{R}$, then Condition Lidentif reduces to Condition Identif. Finally, we have everything at hand to state the result on detection of weaker signals for logistic regression obtained in Bunea (2008).

Proposition 3.9 *Let $\lambda > 0$ ($\lambda_1 > 0$ respectively) be chosen to satisfy* (3.1) *and $\epsilon > 0$ be as in* (3.2)*. Furthermore, consider $s > 0$ as specified in Theorem 3.3 and let Assumptions 3.1 and 3.2 hold true.*

1. *ℓ_1-penalty:*
   Assume Conditions 1.14 and 3.8 are met with $d \leq \frac{s}{16+2s(7+\epsilon)}$ and set U with $b \leq 1-d(7+\epsilon)$. If

   $$\min_{j\in I^*} |\beta_j| \geq 1.75\lambda + 3\left(1+\frac{2}{\lambda}\right)\epsilon,$$

   then

   $$\mathbb{P}\left(I^* \subseteq \hat{I}^L\right) \geq 1-3\delta.$$

2. *$\ell_1 + \ell_2$-penalty:*
   Let $B > 0$ and choose $\lambda_2 = \frac{\lambda_1}{B}$. Furthermore, assume Conditions 1.14 and 3.8 are met with $d \leq \frac{s+\lambda_2}{17+2s(8+\epsilon)}$ and set U with $b \leq 1-d(8+\epsilon)$.If

   $$\min_{j\in I^*} |\beta_j| \geq 1.75\lambda_1 + \left(1+\frac{2}{\lambda_1}\right)\epsilon,$$

   then

   $$\mathbb{P}\left(I^* \subseteq \hat{I}^{EN}\right) \geq 1-3\delta.$$

[1] In the EN case, we require the same condition on the penalisation parameter $\lambda_1 > 0$.

Bunea (2008) remark, that (neglecting the exponentially small ϵ) the minimum size of the true coefficients required by the above theorem for bounding the probability of correctly including true variables in the estimated set essentially is $\min_{j\in I^*}|\beta_j| \geq 1.75\lambda$ ($\min_{j\in I^*}|\beta_j| \geq 1.75\lambda_1$ respectively). As with linear models, we hence observe that one can detect weaker signals under the more restrictive conditions 1.14 and 3.8 on the correlation structure. However, for appropriate choice of penalisation λ_2, correct variable inclusion can be guaranteed for the EN under less limiting assumptions on the correlations of the design than for the Lasso. Moreover, Bunea (2008) remark that replacing the possibly small s by a term close to 1 and thereby refining the result is possible if one passes to asymptotic statements.

Correct subset selection

The results on how to guarantee $\mathbb{P}\left(\hat{I} \subseteq I^*\right) \geq 1-\gamma_2$ for some estimated index set $\hat{I}$ and some $\gamma_2 \in (0,1)$ collected in section 1.2 continue to hold for penalized logistic regression. As Bunea (2008) show, this requires the previously introduced conditions on the correlation matrix that are tailored to the logistic loss function.

Theorem 3.10 *Under the assumptions of Proposition 3.9 the following holds:*

1. *ℓ_1-penalty:*
$$\mathbb{P}\left(I^* = \hat{I}^L\right) \geq 1-5\delta.$$
2. *$\ell_1+\ell_2$-penalty:*
$$\mathbb{P}\left(I^* = \hat{I}^{EN}\right) \geq 1-5\delta.$$

Finally, we re-state the asymptotic result on consistent variable selection (cp. (Bunea, 2008, Corollary 3.8.))

Corollary 3.11 *Assume that* $\min_{j\in I^*}|\beta| = O(\sqrt{\frac{\log(n)}{n}})$ *and let conditions 3.1 and 3.2 hold true.*

1. *ℓ_1-penalty:*
If $\lambda = \mathcal{O}\left(\sqrt{\frac{\log(n)}{n}}\right)$ *and the conditions on the design required for point 1. of Theorem 3.10 are met, then*
$$\lim_{n\to\infty}\mathbb{P}\left(\hat{I}^L = I^*\right) = 1.$$
2. *$\ell_1+\ell_2$-penalty:*
If $\lambda_1 = \mathcal{O}\left(\sqrt{\frac{\log(n)}{n}}\right)$ *and the conditions on the design required for point 2. of Theorem 3.10 are met, then*
$$\lim_{n\to\infty}\mathbb{P}\left(\hat{I}^{EN} = I^*\right) = 1.$$

All in all, the results collected in the previous section for logistic regression do not differ much from those summarized in Chapter 1 for linear models. Also within the penalized logistic regression framework, correct variable selection both asymptotically and in finite samples can be guaranteed under relatively strict conditions on the correlations of the design. It remains unclear, whether the Lasso and Elastic Net for logistic regression are able to select the true model in presence of high correlations among the explanatory variables such that those conditions are violated. The aim of the following section is to get further insight on both penalisation methods' variable selection properties within the present setting.

3.2 A simulation study on logistic regression

As mentioned earlier in this chapter, theoretical results on variable selection in presence of multicollinearity such as the grouping effect have to the best of our knowledge not been derived for generalized linear models. In order to nonetheless further investigate variable selection properties of the Lasso and the Elastic Net in a GLM setting, we resort to simple simulation studies. Inspired by the numerical studies in Meinshausen and Bühlmann (2010), we like to determine the probability that a certain share of the (known) relevant variables can be recovered *without error*. For $\ell_1 + \ell_2$-penalized generalized linear regression, this means that there is no non-zero estimated coefficient corresponding to a superfluous predictor. As before, we choose to consider the frequently used logistic regression for the simulation studies. Below, we give further details on the setting that serves as basis for the subsequent analyses.

3.2.1 Data generating process

For the ensuing studies we generate data as follows: we simulate $n = 100$ realisations of p-dimensional explanatory variables with $p = 10$ according to $X_i \sim \mathcal{N}_p(0, \Sigma)$, $i = 1, \dots, n$. Three different scenarios are considered concerning the correlation structure among the regressors:

a) Independent setting: $\Sigma = I_{10}$.

b) Block structure:

$$\Sigma = \begin{pmatrix} B_1 & 0 & 0 \\ 0 & B_2 & 0 \\ 0 & 0 & I_4 \end{pmatrix}, \quad \text{with} \quad B_1 = \begin{pmatrix} 1 & 0.99 & 0.99 \\ 0.99 & 1 & 0.99 \\ 0.99 & 0.99 & 1 \end{pmatrix} \quad \text{and} \quad B_2 = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}.$$

c) Toeplitz design: $\Sigma_{i,j} = \rho^{|i-j|}$, $1 \leq i, j \leq 10$, with $\rho = 0.99$.

Simulated values for the target variable are then determined based on a logistic regression model with true coefficient vector $\beta = (1, 1, 0, 1, 0, 0, 1, 0, 0, 1)^T$, i.e.

$$Y_i \overset{id}{\sim} Ber(p_i), \quad \ln\left(\frac{p_i}{1 - p_i}\right) = \beta^T X_i, \qquad i = 1, \dots, n.$$

3.2.2 Probability of success for the Elastic Net

With the simulated data at hand, we then assess the probability that $\gamma \cdot q$ of the $q = 5$ relevant variables can be recovered without error, with $\gamma \in \{0.2, 0.4, 0.6, 0.8, 1\}$, by the Elastic Net solving

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(- Y_i X_i^T \beta + \log\left(1 + \exp\left(X_i^T \beta\right)\right) \right) + \lambda \left((1 - \alpha) \|\beta\|_1 + \alpha \|\beta\|_2^2 \right), \tag{3.3}$$

with varying parameter α. In other words, for each setting a)-c) and certain choices of α, we are interested in estimating the probability of the EN to give one to five true positives without generating any false positives.

The study is conducted in R as follows: for each scenario, a dataset is generated according to Section 3.2.1, points a)-c) respectively. Then, for $\alpha \in \{0, 0.1, 0.2, 0.5, 0.7, 0.9\}$, a logistic regression model is fitted using the function *glmnet()* for a sequence of λ-values suggested by the algorithm

itself. Successful recovery is then said to be in place for some $\gamma \in \{0.2, 0.4, 0.6, 0.8, 1\}$, if there exists a value of the penalty parameter λ among the proposed choices, such that the solution to (3.3) $\hat{\beta}$ satisfies $\hat{\beta}_i = 0$ for all $i \in I \backslash I^*$ and $\hat{\beta}_i \neq 0$, for at least $\gamma \cdot q$ many of the true variables I^*. For each scenario and each choice of γ and α, this procedure is repeated 100 times and the proportion of resulting successes is taken as an estimate for the desired probability. For the sake of comparability, within each repetition the Elastic Net is fitted using the λ-sequence suggested by the first *cv.glmnet()*-fit.

Figure 3.1 shows the results of the simulation study outlined above. For each the orthogonal, block and Toeplitz design and a range of α-values, the estimated probability of recovery without error is depicted as a function of the desired proportion of true positives. First, we observe that for any choice of penalty parameter α the Elastic Net is able (i.e. there is at least one value among the proposed ones for the penalty parameter λ) to recover one true positive while at the same time avoiding any false positives. However, selecting a larger number of the true variables comes at the cost of generating false positives as well. In all three settings and for almost any choice of α, the probability of correctly selecting two of the relevant variables without error already falls below 0.5. The number of times the EN was able to recover three or more true variables without error even ranks between zero to ten out of a hundred.

Overall, the more weight is put on the ℓ_1-norm, that is the smaller the α, the higher the estimated probability of success. Surprisingly, for $\alpha \leq 0.2$ the EN was able to more frequently select true coefficients without error in a scenario with block correlations than in the orthogonal setting. With the Toeplitz design as well, the EN with $\alpha \leq 0.2$ exhibits slightly improved selection properties at least for $\gamma \geq 0.6$.

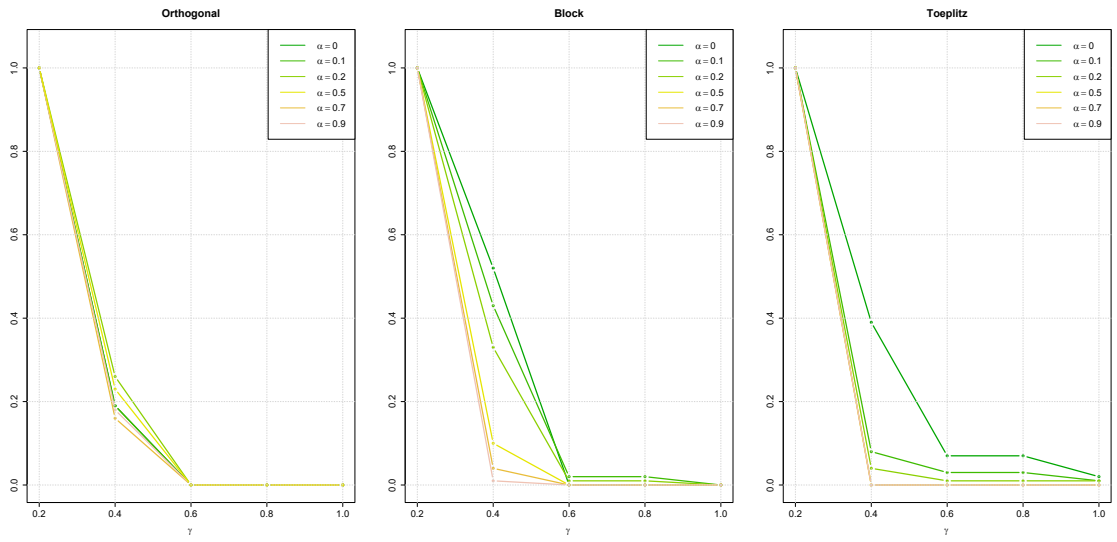

Figure 3.1: Probability of recovering $\gamma \cdot 5$ of the relevant variables without error, as a function of $\gamma \in (0, 1]$, for an orthogonal, block and Toeplitz design and for each $\alpha \in \{0, 0.1, 0.2, 0.5, 0.7, 0.9\}$.

By the above, the Elastic Net tends to include noise variables when choosing a penalty parameter λ with the aim to select more of the relevant variables. As Meinshausen and Bühlmann (2010) by the example of the ℓ_1-penalized linear model empirically show, Stability Selection can lead to large improvements in this regard. In a similar manner, we will in what follows investigate, whether EN-penalized logistic regression joint with Stability Selection leads to higher probabilities of successful

recovery as introduced above.

3.2.3 Probability of success for the Elastic Net with Stability Selection

We once more consider data generated as in Section 3.2.1. In what follows, we seek to compare the probability that $\gamma \cdot q$ of the $q = 5$ relevant variables can be successfully recovered by the EN and the EN with Stability Selection, for $\gamma \in \{0.2, 0.4, 0.6, 0.8, 1\}$. To that end, we choose $\alpha = 0.5$ fixed and determine the probability of success for ordinary $\ell_1 + \ell_2$-penalized logistic regression as shown in Section 3.2.2. For the EN joint with Stability Selection, successful recovery is in place if variables with the highest selection probability (according to a specified threshold) are all relevant. More precisely, the analysis is conducted as follows: first, we construct 100 subsamples of size $n/2 = 50$ without replacement of the dataset obtained by 3.2.1. For each subset, an EN-penalized logistic regression with $\alpha = 0.5$ is fit using *glmnet()*. The values for λ are obtained from the sequence previously specified by the ordinary EN fit on the entire dataset. Thereafter, the value $\lambda_{\min}$ is chosen among said sequence such that for all $\lambda \geq \lambda_{\min}$ at most $\sqrt{0.8p}$ variables are selected. Then, for each $k \in \{1, \ldots, p\}$ and every $\lambda \geq \lambda_{\min}$ the probability of selecting the k^{th} variable when optimizing with penalty parameter λ across samples Π_k^λ is determined. In other words, we have

$$\Pi_k^\lambda = \frac{\sum_{s \in S} \mathbb{1}\left\{\hat{\beta}_k^\lambda(s) \neq 0\right\}}{|S|},$$

with $\hat{\beta}_k^\lambda(s)$ the estimated coefficient vector corresponding to subsample s, S the set of subsamples and $|S| = 100$ its size. Thereafter, for every variable $k \in \{1, \ldots, p\}$, the highest selection probability over all $\lambda \geq \lambda_{\min}$ is assessed, i.e. we determine

$$\Pi_k^{\max} = \max_{\lambda \geq \lambda_{\min}} \Pi_k^\lambda, \quad k = 1, \ldots, p.$$

Those p probabilities $\Pi_k^{\max}$ are then ranked and the $\gamma \cdot q$ variables corresponding to the $\gamma \cdot q$ highest such probabilities are selected. Finally, we speak of successful recovery (or recovery without error) at the rate γ, if all of these ultimately selected variables are relevant. As for the ordinary Elastic Net, repeating this procedure a hundred times ultimately yields the probabilities of success.
As stated in Section 2.4, Meinshausen and Bühlmann (2010) argue by the example of the Lasso in a linear model framework, that Stability Selection works best when adding some extra randomness. Similarly to the *randomised Lasso* proposed in Meinshausen and Bühlmann (2010) (see Definition 2.6), we introduce what we call the *randomised Elastic Net*.

Definition 3.12 (Randomised Elastic Net) *The* randomised Elastic Net *estimator with weakness $w \in (0, 1]$ is obtained as solution to*

$$\min_{\beta \in \mathbb{R}^p} \ell(\beta; X, Y) + \lambda \left\{ \sum_{k=1}^{p} \frac{(1-\alpha)|\beta_k| + \alpha\beta_k^2}{W_k} \right\}, \tag{3.4}$$

with $W_k \overset{iid}{\sim} Unif(w, 1)$, $k = 1, \ldots, p$ and penalisation parameters $\lambda > 0$ and $0 < \alpha < 1$.
Here, $\ell(\cdot; X, Y)$ may be the negative log-likelihood function corresponding to some generalized linear model for explanatory variables $X \in \mathbb{R}^{n \times p}$ and target $Y \in \mathbb{R}^n$.

To extend our analysis, we additionally perform Stability Selection jointly with randomised Elastic Net for $\alpha = 0.5$ on the given dataset. The idea being that possibly, the extra randomness

introduced to the penalty improves variable selection in a similar way as randomised Lasso did for linear regression.

To incorporate the random component, the procedure introduced for Stability Selection with ordinary EN barely has to be altered. The function *glmnet()* allows for a certain penalty factor to be specified, which, if set as the randomly generated weights $(W_k)_{k=1,\dots,p}$, precisely works as desired for the randomised Elastic Net (3.4).

Figure 3.2 visualizes the resulting probabilities of successful recovery for each the ordinary EN, the Elastic Net with Stability Selection and randomised EN with Stability Selection for $\alpha = 0.5$ fixed as a function of the proportion of true variables that need be selected. We observe that considerable improvements could only be achieved in the independent setting. Nonetheless, for $\gamma \geq 0.4$, i.e. if more than one true variable is to be recovered without error, both Stability Selection with ordinary and with randomised EN yielded slightly higher probabilities of success for the scenarios with more evolved correlations in the design as well. However, the methods involving Stability Selection were not able to always recover one true variable without selecting any noise variables, as is the case for ordinary EN.

Furthermore, we note that for all three settings, randomised EN performs best for $\gamma \geq 0.4$. However, the difference to Stability Selection joint with ordinary Elastic Net often is quite small.

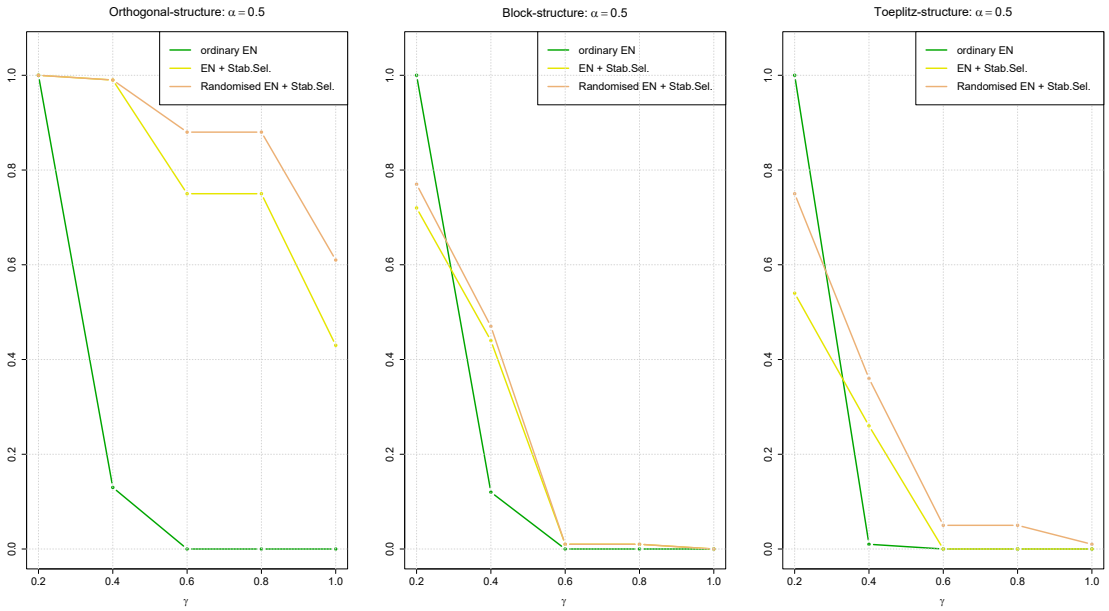

Figure 3.2: Comparison of ordinary EN, Elastic Net with Stability Selection and randomised EN with Stability Selection with regards to their probability of recovering $\gamma \cdot 5$ of the relevant variables without error, as a function of $\gamma \in (0, 1]$. With $\alpha = 0.5$ the methods' selection properties are compared within the orthogonal, block and Toeplitz settings introduced in Section 3.2.1.

We conclude, that in a small-p setting such as the one considered within this simulation, Stability Selection only leads to improved variable selection properties, if it is of interest to recover more than one of the true variables without error. Adding extra randomness in the sense of 3.12 only leads to slightly higher chances of success. The true virtue of Stability Selection however, may lie in selecting true variables while conservatively controlling for false positives in a large dimensional setting. Such a "large-p small-n" scenario forms the basis for an empirical study in (Meinshausen

and Bühlmann, 2010, Section 4), where, as already mentioned in chapter 2, Stability Selection proves to be highly beneficial.

Appendix A

Subgradients for convex optimisation

In the following we will outline some results about subgradients and their application to optimization of convex functions without proofs that are used throughout this thesis. These subsequent findings are both obtained from Nesterov (2003) and the lecture notes Boyd and Vandenberghe (2008).
Throughout this chapter, let $n \in \mathbb{N}$ and $f : \mathbb{R}^n \to \mathbb{R}$ be a convex, possibly non-differentiable function.

Definition A.1 *A vector $g \in \mathbb{R}^n$ is called the subgradient of f at $x_0 \in dom(f)$, if for any $x \in dom(f)$ it holds that*

$$f(x) \geq f(x_0) + g^T(x - x_0). \tag{A.1}$$

The set of all subgradients of f at x_0, termed $\partial f(x_0)$, is called the subdifferential of f at x_0.

Remark A.2 *If the convex function f and differentiable at some point $x \in dom(f)$, then $\partial f(x) = \{\nabla f(x)\}$, i.e. its subdifferential at x consists of its gradient at x only.*

Lemma A.3 *Let $f = \sum_{i=1}^m f_i$ with $f_1, \ldots, f_m$ convex functions. Then we have*

$$\partial f(x) = \sum_{i=1}^{m} \partial f_i(x).$$

Example A.4 *The subdifferential of $|\cdot| : \mathbb{R} \to \mathbb{R}$ is given by*

$$g_{|\cdot|}(x) = \begin{cases} sign(x), & x \neq 0 \\ \{c \in \mathbb{R} : |c| < 1\}, & x = 0 \end{cases}, \quad x \in \mathbb{R}.$$

Example A.5 *The subdifferential of the ℓ_1-norm $\|\cdot\|_1$ at some point $x \in \mathbb{R}^n$ is given by*

$$g_{\ell_1}(x) = \{g \in \mathbb{R}^n : \quad g_i = sign(x_i), x_i \neq 0 \quad \wedge \quad g_i \in \{c \in \mathbb{R} : |c| < 1\}, x = 0\}.$$

Theorem A.6 *(Optimality condition) The function f attains its minimum at $x^* \in dom(f)$ if and only if $g = 0$ is a subgradient of f at x^*, i.e. $0 \in \partial f(x^*)$.*

Bibliography

S. Boyd and L. Vandenberghe. Subgradients. 04 2008. URL https://see.stanford.edu/materials/lsocoee364b/01-subgradients_notes.pdfhttps://doi.org/10.3150/07-BEJ6011.

Florentina Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electron. J. Statist.*, 2:1153–1194, 2008. doi: 10.1214/08-EJS287. URL https://doi.org/10.1214/08-EJS287.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. doi: 10.1214/009053604000000067. URL https://doi.org/10.1214/009053604000000067.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL http://www.jstatsoft.org/v33/i01/.

Edouard Grave, Guillaume R Obozinski, and Francis R. Bach. Trace lasso: a trace norm regularization for correlated designs. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2187–2195. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4277-trace-lasso-a-trace-norm-regularization-for-correlated-designs.pdf.

Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5):1356–1378, 10 2000. doi: 10.1214/aos/1015957397. URL https://doi.org/10.1214/aos/1015957397.

Chenlei Leng, Yi Lin, and Grace Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, pages 1273–1284, 2006.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2010.00740.x. URL http://dx.doi.org/10.1111/j.1467-9868.2010.00740.x.

Y Nesterov. Introductory lectures on convex optimization: A basic course. 01 2003.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, December 2006. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1248547.1248637.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x. URL http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.

List of Figures

1.1 Two-dimensional estimation picture for the Lasso and Elastic Net. For $t = 1$, the constraint region of the Lasso is depicted in grey, while that of the EN is indicated by the dashed black line. The contours of the objective are given in blue. 14
1.2 Simulation 1: Estimated coefficients in dependence of the correlation in the explanatory variables. 22
1.3 Simulation 2: Estimated coefficients in dependence of the correlation in the explanatory variables. 23

3.1 Comparison of the Elastic Net with differing penalty parameter α, with regards to its probability of recovery without error. 41
3.2 Comparison of ordinary EN, Elastic Net with Stability Selection and randomised EN with Stability Selection with regards to their probability of recovery without error. 43