universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Linear Regression and the LASSO“

verfasst von / submitted by

Lorenz Bazant, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 821
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Mathematik
Betreut von / Supervisor:	Univ.-Prof. Dr. Jiří Černý

Contents

1 Preface

Research fields such as economics, medicine or other natural sciences cannot be imagined without statistical investigations. Nevertheless mathematical statistics form an important sub-discipline that unfortunately have been covered only rudimentarily in my studies at the university. This circumstance has been my central motivation in choosing the topic which will be elaborated in this thesis.

Linear regression is an outstanding statistical technique to analyse and predict data. Already in the beginning of the 19th century Gauß and Legendre published papers about the least squares method. At that time their pioneering investigations were successfully applied in the orbit determination of the dwarf planet Ceres for instance. From that point onwards linear regression became more and more important for all science fields.

Having at least as many samples as parameters that are to be estimated is a necessary condition in getting a unique solution in least squares estimation.

Imagine for example that we want to study the changes in climate in the geological past. Suppose that we were given the mean global surface temperature in each year from 1850 to 1998, as well as 1209 indirect indicators for climate, such as pollen, tree-ring data or ice cores in order to reconstruct the mean global surface temperature in the last thousand years. Thus, if we try putting the temperatures with these indicators into a linear relation, we then have $n = 1998 - 1850 + 1 = 149$ samples with the great number of 1209 parameters for this palaeontological problem. This is an example of high-dimensional data that recently led to scientific controversies. Here the least squares fitting of the model is ill-posed.

In 1996 Robert Tibshirani, a statistician at Stanford University, introduced the LASSO, which is an acronym for least absolute shrinkage and selection operator. It will be the main objective of this text, being an approach for the case where one has too less information.

However, even if there are enough samples available, we are often not satisfied with the ordinary least squares method. Therefore it is sometimes helpful to use the LASSO instead. On the one hand the greater the size of the data is, i.e. the more parameters are to be estimated, interpretation of the model becomes more complicated. On the other hand correlation of some variables in the data may lead to high variance in the least squares method. As it will be discussed and further explained later on, LASSO facilitates interpretation by selecting a subset of the data and it improves the prediction accuracy.

This thesis is organized as follows:

Section 3, entitled Preliminaries, gives an overall review of statistical methods primarily based on [14], [13], [20], [4] and [15]. Section 3.1 then introduces the Bayesian approach. Here we follow [7] and [4].

Section 4 deals with linear regression. In Section 4.1 we will define the least squares estimator and will prove some of its important properties. We will use the conventions and notations of [16] and [11].

In Section 5 we discuss the general process of statistical learning. Then we establish penalty methods in Section 5.1 as it is explained in [19]. This is necessary to be able to define the LASSO, which will be carried out in the end of that part. Furthermore, in Section 5.2, we will present ridge regression. This is a method of linear regularization older than the LASSO. Later on, in Section 5.3, we will compare these two methods.

In the last two sections we follow basically [2]. Section 6 leads to error estimations for the LASSO. There, we firstly introduce a so-called "basic inequality" for the LASSO. We will show that we can get rid of its error term with high probability. This will yield consistency. In the end we will derive the so-called compatibility condition.

Finally we will conclude some refinements of the model for the LASSO in Section 7. In the first subsection we will define the regression function and will show what happens if it is by any chance not a sparse linear combination of the vectors $X^{(j)}$. This leads to the definition of some oracle whose refinement will be topic of the second subsection.

We expect the reader of this text to have basic knowledge of probability theory as well as to have fundamental skills in linear algebra and mathematical analysis of course.

2 Acknowledgements

Before immersing ourselves into the material, I first of all want to thank professor Jiří Černý, who supervised my master's thesis. I am utterly grateful for his everlasting reachability to help me with arisen problems, although statistics are not his main area of expertise.

My further thanks are dedicated to my parents and my family who supported me during my studies not only financially but also for standing behind me at all times.

Special thanks also go to my sister-in-law Marina Bazant who always helped me in drafting some complicated English phrases and even read parts of this text.

Finally I want to thank all my friends and my colleagues at the university. The latter I owe great success in extending my skills in mathematics by preparing exercises or studying for exams together.

3 Preliminaries

In this section we introduce the basic statistical framework we are going to work with. First of all note that for mathematical statistics, modelling an experiment is essential:

Definition 3.1. *One calls* $\mathcal{E} = (\mathcal{X}, \mathcal{A}, \mathcal{P})$ statistical model *or* statistical experiment*, if* $(\mathcal{X}, \mathcal{A})$ *is a measurable space and* $\mathcal{P}$ *is a set of probablility distributions on the sample space* $(\mathcal{X}, \mathcal{A})$*. Briefly one says that* $\mathcal{P}$ *is a* statistical model *on* $(\mathcal{X}, \mathcal{A})$*.*

Very often it is $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$*, where* Θ *is called* parameter set*. For* $\Theta \subset \mathbb{R}^k$ *the model is called* parametric*.*

Fundamental in statistical inference is that the experiment is described by a probability distribution that is known except for one parameter $\vartheta \in \Theta$. The goal is to estimate this parameter, or, even more generally, to estimate some measurable and usually real-valued function $g(\vartheta)$ of this parameter. Based on an observation $x \in \mathcal{X}$ we make a decision for ϑ or $g(\vartheta)$ respectively. This concept leads now to the next definitions.

Definition 3.2.

- *A measurable space* $(\Delta, \mathcal{A}_\Delta)$ *with* $\{a\} \in \mathcal{A}_\Delta$ *for every* $a \in \Delta$ *is called* decision space*.*

- *A measurable function* $d : (\mathcal{X}, \mathcal{A}) \to (\Delta, \mathcal{A}_\Delta)$ *is called* non-randomized decision function*, whereas a* randomized decision function δ *is a stochastic kernel from* $\mathcal{X}$ *to* Δ*, i.e. a mapping* $\delta : \mathcal{X} \times \mathcal{A}_\Delta \to [0, 1]$ *sucht that*

  1. $\delta(\cdot, A)$ *is measurable for all* $A \in \mathcal{A}_\Delta$*, and*
  2. $\delta(x, \cdot)$ *is a probability measure for all* $x \in \mathcal{X}$*.*

  We denote the set containing all non-randomized or randomized decision functions with D *and* $\mathcal{D}$ *respectively.*

- *A measurable mapping* $L : \Theta \times \Delta \to \overline{\mathbb{R}}_+$ *is called* loss function*, if for all* $\vartheta \in \Theta$ *:*

$$L(\vartheta, .) : (\Delta, \mathcal{A}_\Delta) \to (\mathbb{R}_+, \overline{\mathcal{B}}_+).$$

  It measures the "amount of loss" caused by choosing $a \in \Delta$ *instead of the actual value* $\vartheta \in \Theta$*, where either* $a = d(x)$ *or in the case of a randomized decision* a *is* $\delta(x, \cdot)$*-distributed.*

- $(\mathcal{E}, \Delta, L)$ *is called* statistical decision problem*, if*
  - $\mathcal{E}$ *is a statistical experiment,*
  - $(\Delta, \mathcal{A}_\Delta)$ *is a decision space and*

– *L is a loss function.*

Note, that if we define for each non-randomized decision function d the function

$$\delta_d(x, A) := \begin{cases} 1 & \text{if } d(x) \in A \\ 0 & \text{if } d(x) \notin A \end{cases},$$

we achieve an injective embedding $D \hookrightarrow \mathcal{D}$ by $d \mapsto \delta_d$. Hence it actually suffices to observe randomized decision problems.

If the amount one loses choosing $a \in \Delta$ is the squared distance between a and the unknown ϑ, then *quadratic loss* is at hand. That's the only situation we will consider throughout this text.

As already mentioned before, we will try to estimate the unknown parameter ϑ in the probability distribution corresponding to the statistical experiment. For that purpose we will observe realisations x_i of X_i having this distribution and merge them into a function that we will call *estimator*. In order to define this exactly, we let $X_1, \ldots, X_n$ be i.i.d. real valued random variables for simplicity, each having the density function $f(\cdot, \vartheta)$. Using the standard convention, let us denote its observed values in small letters: $x_1, \ldots, x_n$.

Definition 3.3.

- *For $j = 1, \ldots, m$, let $T_j : \mathbb{R}^n \to \mathbb{R}$ be measurable functions that do not depend on ϑ or any other unknown quantities, and set $T = (T_1, \ldots, T_m)^t$. Then*

$$T(X_1, \ldots, X_n) = (T_1(X_1, \ldots, X_n), \ldots, T_m(X_1, \ldots, X_n))^t$$

is called an m-dimensional statistic.

- *Any statistic $T = T(X_1, \ldots, X_n)$ that is used for estimating the unknown quantity $g(\vartheta)$ is called an* estimator *of $g(\vartheta)$. The value $T(x_1, \ldots, x_n)$ of T for the observed values of the $X's$ is called an* estimate *of $g(\vartheta)$.*

Note that the terms estimator and estimate are often used interchangeably. Further it shall be mentioned that we normally write $\hat{\vartheta}$ for an estimator estimating ϑ. The "hat" shall simply indicate that we are estimating the parameter immediately beneath it.

Having already established the main statistical setting, we can introduce some estimation methods and give a few examples keeping always the theoretical background in mind.

Definition 3.4. *The function L_x that maps every ϑ to the value $L_x(\vartheta) := P_\vartheta(x)$ is called* Likelihood function. *If it holds $L_x(\hat{\vartheta}) := \sup\{L_x(\vartheta) : \vartheta \in \Theta\}$, then one calls $\hat{\vartheta}(x)$ a* maximum likelihood estimation *of ϑ and $g(\hat{\vartheta}(x))$ a maximum likelihood estimation of $g(\vartheta)$.*

That is, the maximum likelihood method chooses that parameter for which the likelihood function attains its supremum. Now let us have a look at some examples involving the maximum likelihood estimator.

Example 3.5. *Let $X_1, \ldots, X_n$ be a $\mathcal{N}(\mu, \sigma^2)$-distributed random sample with parameter $\vartheta = (\mu, \sigma^2)^t$. Then*

$$L_{X_1,\ldots,X_n}(\vartheta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{2}\right\}.$$

We maximize this function now. We first apply the logarithm, which leads to

$$\log L_{X_1,\ldots,X_n}(\vartheta) = -n\log\sqrt{2\pi} - n\log\sqrt{\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Next, we differentiate with respect to μ and σ^2 and equate them to zero afterwards:

$$\frac{\partial}{\partial\mu}\log L_{X_1,\ldots,X_n}(\vartheta) = \frac{n}{\sigma^2}(\bar{x}-\mu) = 0, \text{ and}$$

$$\frac{\partial}{\partial\sigma^2}\log L_{X_1,\ldots,X_n}(\vartheta) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0,$$

where $\bar{x}$ stands for the mean $\frac{1}{n}\sum_{i=1}^{n} x_i$.

We see immediately that

$$\mu = \bar{x}$$

solves the first equation and plugging this into the second one leads to

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Finally it remains to show that this extremum is a maximum and we will get $\hat{\vartheta} = (\bar{x}, \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2)$. This is straightforward and we omit it.

The use of the logarithm in Example 3.5 legitimates the following definition.

Definition 3.6 (Log-Likelihood-function)**.** *The Log-Likelihood-function is given by*

$$l(\vartheta) = \log L(\vartheta).$$

So one can clearly define the maximum-likelihood-estimator as well by

$$\hat{\vartheta}_{ML} = \arg\max_{\vartheta\in\Theta} l(\vartheta).$$

In the above example the estimator is quite natural. In order to illustrate that there are also other ways to construct a good estimator we observe a further example that is a discrete and quite simple one. It is well and further described in [9]. Allegedly, variations of it were important in the Second World War. With the help of the serial numbers of knocked out tanks the allied powers purportedly tried to estimate the number of tanks of the German Armed Forces.

Example 3.7 (Taxi problem)**.** *Imagine a big city that has N cabs labelled with the numbers $1, \dots, N$. A pedestrian now observes n cabs with the numbers $x_1, \dots, x_n$ without repetitions. Without loss of generality we assume that $x_1 < x_2 < \cdots < x_n$. Then we try to estimate N with this information.*

Obviously it is $N \geq x_n$ and $P_N(x) = \binom{N}{n}^{-1}$, the probability that the pedestrian observes exactly the n cabs $x_1, \dots, x_n$. We see that the smaller N gets, the bigger becomes this probability. So $\hat{N}(x) = \max_i x_i = x_n$ is the maximum likelihood estimator. Since $\hat{N}(x) \leq N$ this method never gives an estimation higher than the true value N.

Another way to argue would be that due to symmetrie reasons the numbers of unobserved cabs $x_1 - 1$ and $N - x_n$ should have in average of many samples the same size. This idea establishes another estimator $\hat{N}_1(x) = x_n + x_1 - 1$.

A third way to construct an estimator is to replace the length of the above gap of unobserved cabs $\{x_n + 1, \dots, N\}$ by the mean length of gaps between the observations, i.e by

$$\frac{1}{n}((x_1 - 1) + (x_2 - x_1 - 1) + \cdots + (x_n - x_{n_1} - 1)) = \frac{x_n - n}{n}.$$

This approach gives another estimator: $\hat{N}_2(x) = x_n + \frac{x_n - n}{n}$.

As we have seen, for one and the same problem often exist various estimators. We are now interested in finding a way to decide whether an estimator is a good one or not. We achieve one possible classification by looking at its expectation value.

Definition 3.8.

- *The* bias *of an estimator $\hat{\vartheta}$ is defined by*

$$B\left[\hat{\vartheta}\right] = \mathrm{E}_\vartheta\left[\hat{\vartheta}\right] - \vartheta.$$

- *An estimator $\hat{\vartheta}$ is called* unbiased *if its bias equals zero, i.e. if $\mathrm{E}_\vartheta\left[\hat{\vartheta}\right] = \vartheta$ for all $\vartheta \in \Theta$. If $\mathrm{E}_\vartheta\left[\hat{\vartheta}\right] \neq \vartheta$ we say that $\hat{\vartheta}$ is biased.*

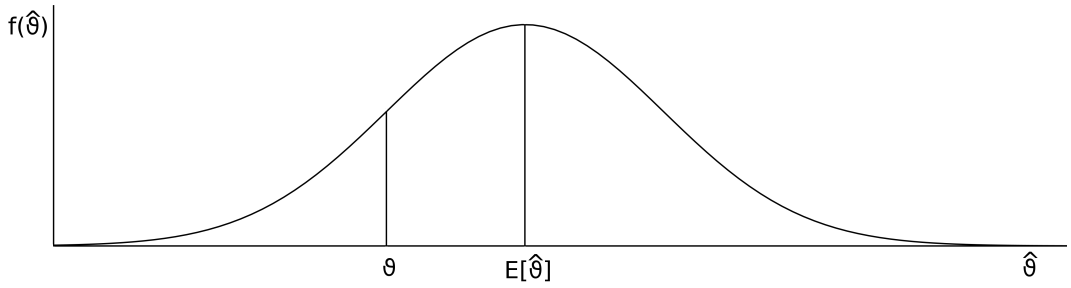

Figure 1: positively biased estimator

It is not difficult to show that in Example 3.7, the maximum likelihood estimator is biased and the estimators $\hat{N}_1$ and $\hat{N}_2$ are unbiased. We rather omit this proof and give instead another example.

Example 3.9. *Let $X_1, X_2 \ldots, X_n$ be a random sample and let $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$. Then it holds*

- *$S'^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ is a biased estimator for σ^2 and*
- *$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ is a unbiased estimator for σ^2.*

Proof. By an elementary computation,

$$\begin{aligned}
\mathrm{E}\left[S'^2\right] &= \frac{1}{n}\,\mathrm{E}\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = \frac{1}{n}\,\mathrm{E}\left[\sum_{i=1}^n X_i^2 - 2\sum_{i=1}^n X_i\bar{X} + \sum_{i=1}^n \bar{X}^2\right] \\
&= \frac{1}{n}\left(\sum_{i=1}^n \mathrm{E}\left[X_i^2\right] + \mathrm{E}\Big[-2\bar{X}\underbrace{\sum_{i=1}^n X_i}_{n\bar{X}} + n\bar{X}^2\Big]\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^n (\sigma^2 - \mu^2) - n\,\mathrm{E}\left[\bar{X}^2\right]\right) \\
&= (\sigma^2 - \mu^2) - \mathrm{E}\left[\bar{X}^2\right] = (\sigma^2 - \mu^2) + \mathrm{E}\left[\bar{X}\right]^2 - \mathrm{Var}\left[\bar{X}\right] = \frac{(n-1)\sigma^2}{n}.
\end{aligned}$$

For the last equality we used that obviously it holds $\mathrm{E}\left[\bar{X}\right] = \mu$ and $\mathrm{Var}\left[\bar{X}\right] = \frac{\sigma^2}{n}$ for the mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. So S'^2 is (negatively) biased. From the above calculation follows that

$$\mathrm{E}\left[S^2\right] = \frac{1}{n-1}\,\mathrm{E}\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = \frac{1}{n-1}(n(\sigma^2 - \mu^2) + n\mu^2 - \sigma^2) = \sigma^2,$$

and that's why S^2 is biased. □

Clearly we prefer an estimator having also a small variance. That's why we introduce as well the *mean square error*.

Definition 3.10. *The* mean square error *of an estimator $\hat{\vartheta}$ is defined as the expectation of its squared distance to the "real" parameter.*

$$\mathrm{MSE}\left[\hat{\vartheta}\right] = \mathrm{E}_\vartheta\left[(\hat{\vartheta}-\vartheta)^2\right].$$

A simple calculation given below shows that the mean square error in fact invokes the variance. First of all, observe that it holds

$$\left(\hat{\vartheta}-\vartheta\right) = \left(\hat{\vartheta}-\mathrm{E}_\vartheta\left[\hat{\vartheta}\right]\right) + \left(\mathrm{E}_\vartheta\left[\hat{\vartheta}\right]-\vartheta\right) = \left(\hat{\vartheta}-\mathrm{E}_\vartheta\left[\hat{\vartheta}\right]\right) + B\left(\hat{\vartheta}\right).$$

By squaring this equation and taking the expectation afterwards, this leads to

$$\begin{aligned}\mathrm{MSE}\left[\hat{\vartheta}\right] &= \mathrm{E}_\vartheta\left[\left(\hat{\vartheta}-\vartheta\right)^2\right] \\ &= \mathrm{E}_\vartheta\left[\left(\hat{\vartheta}-\mathrm{E}\left[\hat{\vartheta}\right]\right)^2\right] + 2\,\mathrm{E}_\vartheta\left[\left(\hat{\vartheta}-\mathrm{E}_\vartheta\left[\hat{\vartheta}\right]\right)\left(\mathrm{E}_\vartheta\left[\hat{\vartheta}\right]-\vartheta\right)\right] \\ &\quad + \mathrm{E}_\vartheta\left[\left(\mathrm{E}\left[\hat{\vartheta}\right]-\vartheta\right)^2\right] \\ &= \mathrm{Var}_\vartheta\left[\hat{\vartheta}\right] + 0 + \left(\mathrm{E}_\vartheta\left[\hat{\vartheta}\right]-\vartheta\right)^2 = \mathrm{Var}_\vartheta\left[\hat{\vartheta}\right] + B\left[\hat{\vartheta}\right]^2.\end{aligned} \tag{1}$$

To bring this initial subsection to an end we introduce now another important statistical requirement on estimators making them viable.

Definition 3.11. *The estimator $\hat{\vartheta}_n$, which is calculated using a sample of size n, is said to be a* consistent estimator *of ϑ, if it converges to ϑ in probability, i.e. if for all $\varepsilon > 0$ and all $\vartheta \in \Theta$*

$$\lim_{n\to\infty} \mathrm{P}_\vartheta\left[|\hat{\vartheta}_n - \vartheta| > \varepsilon\right] = 0.$$

Consistency and unbiasedness are of course not equivalent. However, as the following lemma shows, with a further condition an unbiased estimator becomes consistent.

Lemma 3.12. *An unbiased estimator $\hat{\vartheta}_n$ for ϑ is a consistent estimator of ϑ if*

$$\lim_{n\to\infty} \mathrm{Var}_\vartheta\left[\hat{\vartheta}_n\right] = 0.$$

Proof. Set $\sigma_{\hat{\vartheta}_n} = \sqrt{\mathrm{Var}\left[\hat{\vartheta}_n\right]}$, let $\varepsilon > 0$ and assume that the sample size n is fixed. Then, by the Chebyshev inequality, using that $\mathrm{E}_\vartheta\left[\hat{\vartheta}\right] = \vartheta$,

$$0 \leq \mathrm{P}\left[\left|\hat{\vartheta}_n - \vartheta\right| > \varepsilon\right] \leq \frac{\mathrm{Var}_\vartheta\left[\hat{\vartheta}_n\right]}{\varepsilon^2}.$$

By taking $n \to \infty$ the claim follows. □

3.1 Bayesian Inference

The usual procedure in statistical estimation as described above is to regard the value of ϑ as fixed but unknown. One then uses some observations to draw appropriate conclusions.

Now we will introduce an important different approach called Bayesian inference, where ϑ is a random variable with a unknown distribution called the *prior distribution* $f(\vartheta)$ that incorporates all available information about it. After observing some data one then considers the *posteriori distribution* $f(\vartheta|x)$ based on which we can construct a *Bayesian estimator*.

Readers who are already familiar with this topic can skip this section and go directly to the linear regression model on page 17. We will use the methods of bayesian inference only very superficially later on but nevertheless it is something that cannot be absent in an introduction to statistics. Basically, we follow here [7].

The following theorem, on which the method described in this section is based on, is a fundamental result of probability theory. The proof is known from elementary lecture.

Theorem 3.13 (Bayes' Rule)**.** *Let A and B be two events $A, B \subset \Omega$, with $0 < \mathrm{P}[A] \leq 1$ and $\mathrm{P}[B] > 0$. Then it holds*

$$\mathrm{P}[A|B] = \frac{\mathrm{P}[B|A]\,\mathrm{P}[A]}{\mathrm{P}[B]}.$$

Similarly, for continuous random variables X and Y

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int f_{(X,Y)}(x,y)\mathrm{d}y},$$

where f_X and f_Y denote the densities of X and Y respectively and $f_{Y|X}$ stands for the conditional density of Y given X, i.e. $f_{Y|X}(y|x) := \frac{f_{(X,Y)}(x,y)}{f_X(x)}$ with $f_{(X,Y)}(x,y)$ being the joint density of X and Y. One defines $f_{X|Y}$ analogously.

Proof. See [12], for example. □

Even before any data are collected one has to determine the prior distribution $f(\vartheta)$ firstly. In order to abbreviate, note that we will often call this distribution just *prior*. Figure 2 illustrates that lack of information leads to a wide, flat probability density function of the prior while lots of information give a peaked prior that is highly concentrated about some value. After observing

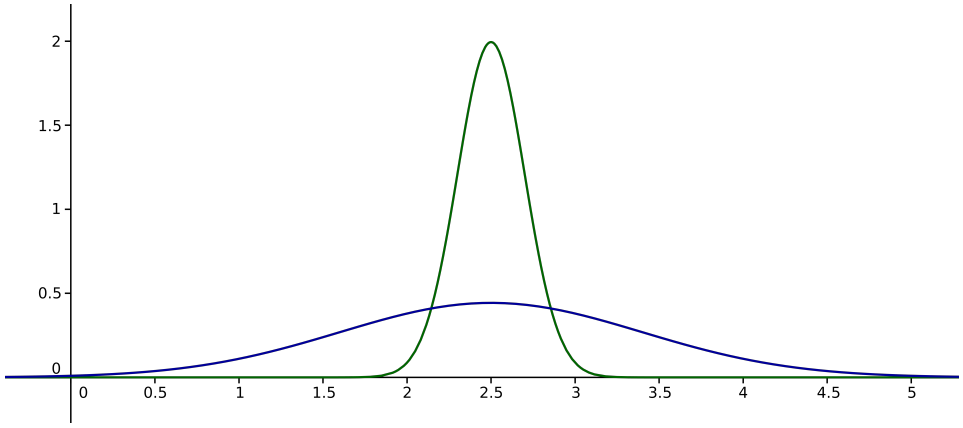

Figure 2: A narrow, concentrated prior (the green one) and a wider, less informative prior (the blue one)

some data, one constructs the conditional distribution of ϑ given $X = x$. This then is called *posterior distribution*. Here comes its exact definition:

Definition 3.14 (posterior distribution)**.** *Let $X = x$ be the observed realization of a random variable X with density function $f(x|\vartheta)$. After the prior distribution with density function $f(\vartheta)$ is determined, we obtain the density function of the* posterior distribution *by Bayes' rule (see Theorem 3.13):*

$$f(\vartheta|x) = \frac{f(x|\vartheta)f(\vartheta)}{\int f(x|\vartheta')f(\vartheta')\mathrm{d}\vartheta'}. \tag{2}$$

For discrete parameter spaces one has just to replace the integral in the denominator by a sum.

The denominator in (2) obviously does not depend on ϑ, so the posterior distribution is proportional to the product of likelihood and prior distribution. One often writes

$$f(\vartheta|x) \propto f(x|\vartheta)f(\vartheta) \text{ or } f(\vartheta|x) \propto L(\vartheta)f(\vartheta). \tag{3}$$

In the Bayesian model we use the posterior distribution for inferences about ϑ. The following definitons show how this can be done.

Definition 3.15. *The* mode *of a continuous distribution f is the value x^* that maximizes $f(x)$.*

Note that in statistics the expression "mode" refers to other situations as well. On the one hand, the *mode* of a discrete random variable X with probability distribution $p(x)$ is that value x^* for which $p(x)$ is largest, i.e. it is the most probable x value. On the other hand the *mode* of a numerical data set is the value that occurs most frequently in the set.

Now the following definition makes sense:

Definition 3.16 (Estimators in Bayesian Inference)**.** *The* posterior expectation value $\mathrm{E}[\vartheta|x]$ *is the expectation value of the posterior distribution $f(\vartheta|x)$:*

$$\mathrm{E}[\vartheta|x] = \int \vartheta f(\vartheta|x)\mathrm{d}\vartheta.$$

The posterior mode $\mathrm{Mod}(\vartheta|x)$ *is the mode of the posterior distribution $f(\vartheta|x)$:*

$$\mathrm{Mod}(\vartheta|x) = \arg\max_{\vartheta} f(\vartheta|x).$$

The posterior median $\mathrm{Med}(\vartheta|x)$ *is the median of the posterior distribution $f(\vartheta|x)$, i.e. the value a for which*

$$\int_{-\infty}^{a} f(\vartheta|x)\mathrm{d}\vartheta = 0.5 \text{ and } \int_{a}^{\infty} f(\vartheta|x)\mathrm{d}\vartheta = 0.5$$

holds.

In order to support the developed theory about Bayesian inference, we close this section with an example that is further described in [4].

Example 3.17. *In this example we make an inference about a population proportion p. Since $p \in [0,1]$ we chose a particular beta distribution for a prior on p. Note that*

$$f(x) = \begin{cases} \frac{1}{\mathrm{B}(p,q)} x^{p-1}(1-x)^{q-1} & \text{if } x \in [0,1] \\ 0 & \text{else} \end{cases}$$

is the density function of the standard beta distribution, where $\mathrm{B}(p,q)$ is given by

$$\mathrm{B}(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

Further recall that a $\mathrm{B}(p,q)$*-distributed random variable* X *has expectation* $\frac{p}{p+q}$*. According to the data from an American survey of 1574 reported people, 803 of them incorrectly said that antibiotics kill viruses. It suggests itself that the data can be considered being binomial* $\mathrm{Bin}(n=1574,p)$ *distributed.*

So it is not difficult to calculate the posterior distribution:

$$\begin{aligned} f(p|x) &= \frac{f(x|p)f(p)}{\int_{-\infty}^{\infty} f(x|s)f(s)\mathrm{d}s} = \frac{\binom{n}{x}p^x(1-p)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1}}{\int_0^1 \binom{n}{x}s^x(1-s)^{(n-x)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}s^{a-1}(1-s)^{b-1}\mathrm{d}s} \\ &= \frac{\binom{n}{x}p^{x+a-1}(1-p)^{n-x+b-1}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}{\binom{n}{x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 s^{x+a-1}(1-s)^{n-x+b-1}\mathrm{d}s} = \frac{p^{x+a-1}(1-p)^{n-x+b-1}}{\int_0^1 s^{x+a-1}(1-s)^{n-x+b-1}\mathrm{d}s} \\ &= \frac{p^{x+a-1}(1-p)^{n-x+b-1}}{\frac{\Gamma(x+a)\Gamma(n-x+b)}{\Gamma(n+a+b)}\int_0^1 \frac{\Gamma(n+a+b)}{\Gamma(x+a)\Gamma(n-x+b)}s^{x+a-1}(1-s)^{n-x+b-1}\mathrm{d}s}. \end{aligned}$$

The integral in the denominator is 1, because the integrand has the form of a standard beta probability density. Consequently it is

$$f(p|x) = \frac{\Gamma(n+a+b)}{\Gamma(x+a)\Gamma(n-x+b)}p^{x+a-1}(1-p)^{n-x+b-1}.$$

We see that the posterior distribution of p *is* $\mathrm{B}(x+a,n-x+b)$*-distributed.*

By taking the posterior mean for example, we obtain the Bayesian estimate

$$\mathrm{E}\left[p|x\right] = \frac{x+a}{n+a+b} = \frac{803+a}{1574+a+b}$$

for p*.*

Table 1: data corresponding to Figure 3

Year	**1951**	**1956**	**1961**	**1966**	**1971**	**1976**	**1981**
Rate of Divorces	17.70	14.40	13.80	14.80	17.68	20.83	26.50

Year	**1986**	**1991**	**1996**	**2001**	**2006**	**2011**
Rate of Divorces	29.50	33.50	38.30	45.97	48.86	43.02

Source: [1]

4 Linear Regression

In this section we will introduce the linear regression model. It is a method to put a *dependent variable* in a linear context with *independent variables*. However, before defining the exact model we are going to work with, let us have a look at an introducting example of linear regression to some statistical data that will give us an idea of how the procedure works.

Example 4.1 (Divorce Rates in Austria)**.** *In the year 2011 a total number of 20582 divorces were filed in Austria. So far this is the highest absolute number of divorces. In the eighties and nineties the total divorce number was between 16000 and 18000 each year.*

In this example we want to analyse the total divorce rates. They indicate the magnitude of the percentage of marriages that end by a divorce. The total divorce rates are calculated based on the observed divorces in a year related to the year of the corresponding marriage.

In the diagram below, each grey point corresponds to the total divorce rate in Austria in some five years step from 1951 to 2011. Our data corresponds to a time line which forms a special case of regression analysis. The line represents the linear regression and is in this example given by $y = 0.6195x - 1199.0872$*. It is kind of an equilization of the point cloud. Aim of this section is to expplain the method how we can obtain this.*

The line here is increasing. Therefore an increase of the divorce rates in the future is expected. However, obviously linear regression is not appropriate in this example for long time predictions. In the year 2100 for instance, we would expect the impossible divorce rate of 101.78*. Nevertheless we get meaningful results for the near future or for a year between 1951 and 2011 that is not contained in the data. For the year 1994 for instance, the linear regression gives a divorce rate of* 36.12*, which means a difference of* 0.74 *to the lower true value that is not contained in the data. For this year, 2016, we expect a divorce rate of* 49.75*.*

As the above example illustrates, given n pairs of observations $(x_i, y_i), (i =$

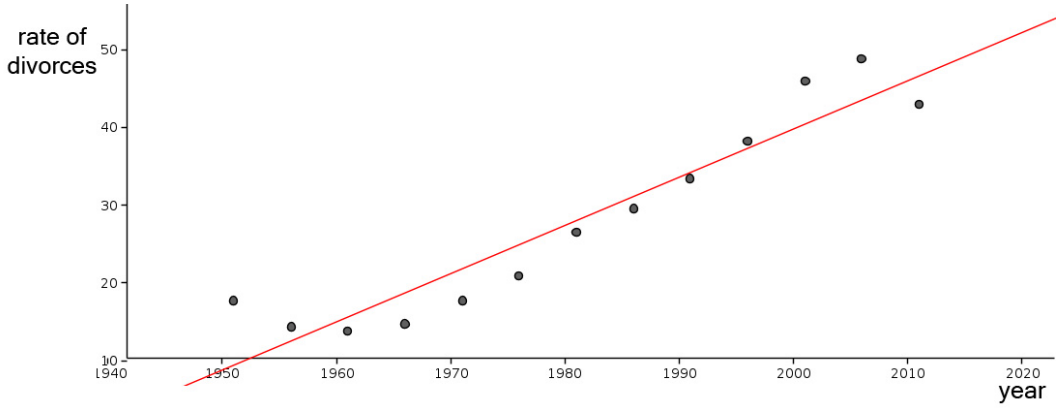

Figure 3: rates of divorces in Austria in the years 1951-2011

$1, 2, \ldots, n$) we can plot them to a scatter diagram. Regression analysis then tries to fit a smooth curve through the points such that they are as close as possible to the curve. In our case, as the title of this section already suggests, the curve will be a straight line.

In contrast to Example 4.1 there likewise exist pairs of variables that have an "exact" (e.g. physical) relationship between each other. Nevertheless we have to take fluctuations into account that are caused by measurement errors for instance.

Statistical relationships do not imply causal correlations. However, one can use them for further investigations in prediction.

Aside from prediction, another reason for regression analysis is testing scientific hypotheses. Imagine for example that we are examining Ohm's law. It states that $U = R \cdot I$, where I denotes the current through a resistor of R ohms and U is the voltage across the resistor. Now, we could measure these quantities in some way. Then, after plotting them into a scatter diagram, it will give support to the law if they are close to a straight line through the origin.

As we have already enough motivation, let us define the linear regression model exactly.

Definition 4.2 (Linear Regression Model)**.** *Let* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *correspond to the observed data, where* $X_i = (X_i^{(1)}, \ldots, X_i^{(p)}) \in \mathbb{R}^p$ *and* $Y_i \in \mathbb{R}$ *for all* $i = 1, \ldots, n$*. Then the relation*

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i,\ i = 1, \ldots, n \tag{4}$$

is called a linear regression model*. One calls the variables* X_i independent

(also predictor *or* regressor*)* variables *and* Y_i *the* dependent *(or* response*)* variables.

For simplicity we assume the errors ε_i *to be* $\mathcal{N}(0, \sigma^2)$*-distributed and independent of the regressor variables.*

Sometimes it is more convenient to write this model in matrix form

$$Y = X\beta + \varepsilon,$$

where $Y_{n\times 1}$ *is the vector of responses,* $X_{n\times p}$ *the design matrix and* $\varepsilon_{n\times 1}$ *the vector of measurement errors.*

Furthermore it is assumed that the linear regression model holds exactly with the parameter β^0*.*

Note that in some slightly more general models in the literature a single β_0 called the *intercept*, which we assume to be zero throughout this text, is added in (4). In fact, with the intercept being non-zero one could just include β_0 in β and replace $X_i = \left(X_i^{(1)}, \ldots, X_i^{(p)}\right) \in \mathbb{R}^p$ by $X_i' = \left(1, X_i^{(1)}, \ldots, X_i^{(p)}\right) \in \mathbb{R}^{p+1}$ for all $i \in \{1, \ldots, n\}$.

4.1 The Least Squares Estimator

First of all let us consider the case where the number of unknown parameters is at most equal to the sample size, i.e. $p \leq n$, and that the matrix X has full rank p. One method to estimate β is the *least squares estimation* which we will derive right now in a geometrical way. It consists of minimizing $\sum_{i=1}^n \varepsilon_i^2$ with respect to β. So by setting $\vartheta = X\beta$, we minimize $\varepsilon^t\varepsilon = \|Y - \vartheta\|_2^2$ subject to $\vartheta \in \{z : z = Xy \text{ for any } y\} =: \Omega$, the column space of X. Therefore we set $\hat{\vartheta} = PY$, where P represents the orthogonal projection onto Ω.

In the following we just use properties of orthogonal projections. We want to show now first that $\|Y - \vartheta\|_2^2 \geq \left\|Y - \hat{\vartheta}\right\|_2^2$ for all $\vartheta \in \Omega$.

$$\begin{aligned}(Y - \hat{\vartheta})^t(\hat{\vartheta} - \vartheta) &= (Y - PY)^t(PY - \underbrace{\vartheta}_{=P\vartheta}) \\ &= (Y - PY)^t P(Y - \vartheta) \\ &= Y^t \underbrace{(\mathbb{I}_n - P)P}_{=P-P^2}(Y - \vartheta) = Y^t(P - \underbrace{P^2}_{=P})(Y - \vartheta) = 0,\end{aligned}$$

and thus trivially also

$$(\hat{\vartheta} - \vartheta)^t(Y - \hat{\vartheta}) = ((Y - \hat{\vartheta})^t(\hat{\vartheta} - \vartheta))^t = 0.$$

If we write now

$$Y - \vartheta = (Y - \hat{\vartheta}) + (\hat{\vartheta} - \vartheta),$$

this leads us to

$$\begin{aligned}
\|Y - \vartheta\|_2^2 &= (Y - \vartheta)^t (Y - \vartheta) \\
&= ((Y - \hat{\vartheta}) + (\hat{\vartheta} - \vartheta))^t ((Y - \hat{\vartheta}) + (\hat{\vartheta} - \vartheta)) \\
&= ((Y - \hat{\vartheta})^t + (\hat{\vartheta} - \vartheta)^t)((Y - \hat{\vartheta}) + (\hat{\vartheta} - \vartheta)) \\
&= (Y - \hat{\vartheta})^t (Y - \hat{\vartheta}) + \underbrace{(Y - \hat{\vartheta})^t (\hat{\vartheta} - \vartheta)}_{=0} + \underbrace{(\hat{\vartheta} - \vartheta)^t (Y - \hat{\vartheta})}_{=0} + (\hat{\vartheta} - \vartheta)^t (\hat{\vartheta} - \vartheta) \\
&= \left\|Y - \hat{\vartheta}\right\|_2^2 + \left\|\hat{\vartheta} - \vartheta\right\|_2^2 \geq \left\|Y - \hat{\vartheta}\right\|_2^2.
\end{aligned}$$

Here equality holds clearly if and only if $\vartheta = \hat{\vartheta}$. Since $Y - \hat{\vartheta} \perp \Omega$, it is

$$X^t (Y - \hat{\vartheta}) = 0 \Leftrightarrow X^t \hat{\vartheta} = X^t Y.$$

X has full rank, so $X^t X$ is positive definite and therefore invertible. It exists a unique $\hat{b}$ such that $X\hat{b} = \hat{\vartheta}$. Therefore we get the so called *normal equations*

$$X^t X \hat{b} = X^t Y, \tag{5}$$

which lead immediately to the following definition.

Definition 4.3. *The* least squares estimator *in the linear model (4) is of the form*

$$\hat{b} := (X^t X)^{-1} X^t Y.$$

Note that we intentionally write $\hat{b}$ here and in the following for the least squares estimator instead of $\hat{\beta}$ in order not to mix up with other estimators.

Next we analyse the prediction error one has to put up with least squares estimation. To this end we recall two important definitions:

Definition 4.4. *Let $X_1, \ldots, X_n$ be identical $\mathcal{N}(0,1)$-distributed random variables. Then the sum $X_1^2 + \cdots + X_n^2$ is χ_n^2 distributed. It has the desity function*

$$p(x) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} e^{-\frac{x}{2}} x^{\frac{n}{2}-1},$$

where Γ is the usual Gamma function. The parameter n is called the number of degrees of freedom.

Definition 4.5. *The random vector X with expectation vector μ and covariance matrix Σ has multivariate normal distribution if its density function is*

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)}.$$

One writes $X \sim \mathcal{N}_n(\mu, \Sigma)$ as usual. Note that for $n = 1$ this is the common one dimensional normal distribution.

Lemma 4.6. *Let Y be $\mathcal{N}_n(\mu, \Sigma)$-distributed with Σ being positive definite. Then*

$$Q := (Y-\mu)^t \Sigma^{-1} (Y-\mu) \text{ has } \chi^2_n \text{ distribution.}$$

Proof. We can write $Y = \Sigma^{\frac{1}{2}} Z + \mu$ with $Z \sim \mathcal{N}_n(0, \mathbb{I}_n)$. Hence,

$$Q = \underbrace{Z^t \Sigma^{\frac{1}{2}}}_{(Y-\mu)^t} \Sigma^{-1} \underbrace{\Sigma^{\frac{1}{2}} Z}_{Y-\mu} = Z^t Z = \sum_{i=1}^n Z_i^2 \sim \chi^2_n$$

because of independence. □

Theorem 4.7. *If Y is $\mathcal{N}(X\beta, \sigma^2 \mathbb{I}_n)$-distributed, where X is a $n \times p$ matrix of rank p, then*

$$\frac{\left\| X(\hat{b} - \beta^0) \right\|_2^2}{\sigma^2} \text{ has } \chi^2_p\text{-distribution}$$

Proof. One can write $\hat{b} = (X^t X)^{-1} X^t Y = CY$ for some $p \times n$ matrix C such that $\operatorname{rank} C = \operatorname{rank} X = p$. Therefore it has a mulivariate normal distribution.

A simple calculation gives the following equations:

$$\mathrm{E}\left[\hat{b}\right] = \mathrm{E}\left[(X^t X)^{-1} X^t Y\right] = (X^t X)^{-1} X^t \underbrace{\mathrm{E}\left[Y\right]}_{X\beta} = \beta^0 \tag{6}$$

and

$$\begin{aligned} \operatorname{Var}\left[\hat{b}\right] &= \operatorname{Var}\left[(X^t X)^{-1} X^t Y\right] = (X^t X)^{-1} X^t \operatorname{Var}\left[Y\right] X (X^t X)^{-1} \\ &= \sigma^2 (X^t X)^{-1} (X^t X)(X^t X)^{-1} = \sigma^2 (X^t X)^{-1}. \end{aligned} \tag{7}$$

In both equations one uses that $\varepsilon \sim \mathcal{N}_n(0, \mathbb{I}_n \sigma^2)$. In the second equation we have used $\operatorname{Var}[AX] = A \operatorname{Var}[X] A^t$ which can be derived easily.

Therefore $\hat{b} \sim \mathcal{N}_n(\beta, \sigma^2 (X^t X)^{-1})$. Now finally observe that

$$\frac{\left\| X(\hat{b} - \beta^0) \right\|_2^2}{\sigma^2} = \frac{(\hat{b} - \beta^0)^t X^t X (\hat{b} - \beta^0)}{\sigma^2} = (\hat{b} - \beta^0)^t \left(\operatorname{Var}\left[\hat{b}\right] \right)^{-1} (\hat{b} - \beta^0) \sim \chi^2_p,$$

by Lemma 4.6. □

The theorem means in particular that

$$\frac{\mathrm{E}\left[\left\|X(\hat{b}-\beta^0)\right\|_2^2\right]}{n} = \frac{\sigma^2}{n}p.$$

So each paramter β_j^0 is estimated with an error of order $\frac{\sigma^2}{n}$ after "reparametrizing to orthonormal design", which gives an overall squared accuracy of $\frac{\sigma^2}{n}p$.

Note that in the case of non linearly independent columns of X, which always happens for example if $p < n$, one gets similar results working with any generalized inverse of X^tX.

Definition 4.8.

- *The values $X\hat{b}$ are called* fitted values *and we use the notation $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)$.*
- *The elements of the vector $Y - \hat{Y}$ are called* residuals *which we denote by e.*
- *The minimum value of $\varepsilon^t\varepsilon$, namely*

$$\begin{aligned} e^te &= (Y - X\hat{b})^t(Y - X\hat{b}) \\ &= Y^tY - 2\hat{b}^tX^tY + \hat{b}^tX^tX\hat{b} \\ &= Y^tY - \hat{b}^tX^tY + \hat{b}^t(X^tX\hat{b} - X^tY) \\ &\overset{(5)}{=} Y^tY - \hat{b}^tX^tY \\ &= Y^tY - \hat{b}^tX^tX\hat{b}, \end{aligned}$$

is called the residual sum of squares. *One writes* $\mathrm{RSS}\,(\beta)$

Observe that by the uniqueness of $\hat{\vartheta} = X\hat{b}$ it follows that $\hat{Y}$, e and the residual sum of squares are unique, no matter which rank X has. Further note that the least squares method consists in nothing else than minimizing $\mathrm{RSS}\,(\beta)$.

In [5] there is a proof given to the important and famous Gauß-Markov-theorem.

Theorem 4.9 (Gauß-Markov)**.** *In the linear regression model with regressor matrix X, the least squares estimator $\hat{b}$ is the minimum variance linear unbiased estimator of β. For any vector of constants w, the minimum variance linear unbiased estimator of $w'\beta$ in the regression model is $w'\hat{b}$, where $\hat{b}$ is the least squares estimator.*

Proof. We already know, that the least squares estimator $\hat{b}$ is a linear unbiased estimator (see proof of Theorem 4.7). We show now, that any other linear unbiased estimator of β has a larger variance.

Let $\hat{a} = CY$ be another linear unbiased estimator of β, where C is a $K \times n$ matrix. By $\hat{a}$ being unbiased it is

$$\mathrm{E}\left[CY|X\right] = \mathrm{E}\left[(CX\beta + C\varepsilon)|X\right] = \beta,$$

which implies that $CX = I$.

Now set $D = C - (X^tX)^{-1}X^t$, so that $DY = \hat{a} - \hat{b}$. Then it is

$$\begin{aligned}
\mathrm{Var}\left[\hat{a}|X\right] &= \sigma^2((D + (X^tX)^{-1}X^t)(D + (X^tX)^{-1}X^t)^t) \\
&= \sigma^2 DD^t + \sigma^2 D((X^tX)^{-1}X^t)^t \\
&\quad + \sigma^2(X^tX)^{-1}X^tD + \sigma^2(X^tX)^{-1}X^t((X^tX)^{-1}X^t)^t \\
&= \sigma^2 DD^t + \sigma^2 DX(X^tX)^{-1} + \sigma^2(X^tX)^{-1}X^tD + \sigma^2(X^tX)^{-1}.
\end{aligned}$$

Furthermore we see that DX must equal 0, because $CX = I = DX + (X^tX)^{-1}(X^tX)$ and thus

$$\mathrm{Var}\left[\hat{a}|X\right] = \sigma^2(X^tX)^{-1} + \sigma^2 DD^t = \mathrm{Var}\left[\hat{b}|X\right] + \sigma^2 DD^t.$$

Since $\sigma^2 DD^t$ is nonnegative definite, every quadratic form in $\mathrm{Var}\left[\hat{a}|X\right]$ is larger than the corresponding quadratic form in $\mathrm{Var}\left[\hat{b}|X\right]$.

As well, the second claim now follows immediately since the variance of $w'b$ is a quadratic form in $\mathrm{Var}\left[b|X\right]$, and likewise for any $\hat{a}$ and proves that each individual slope estimator b k is the best linear unbiased estimator of β_k. To see this let w be the zero vector that has just a single 1 in the kth coordinate. □

The Gauss-Markov theorem together with the derivation in (1) implies that the least squares estimator has the smallest mean square error of all linear estimators with no bias.

Of course, this does not say that there do not exist a biased estimator (which of course are also commonly used) with smaller mean square error. As we will discuss in section 5, such an estimator would trade a little bias while reducing its variance. Every estimator that shrinks or sets to zero some of the least squares coefficients may be a biased one.

5 Methods in Supervised Statistical Learning

We discuss how to handle statistical data in order to learn from it. The general process works as follows: Usually one tries to predict a quantitative or categorical *outcome measurement* based on a set of *features*, which are called in the regression model predictor variables, independent or regressor variables (see Definition 4.2). Analogously the outcome measurement corresponds to the dependent or response variables. With a training set of data we can observe the outcome and feature measurements for a set of objects. Then we construct a so called *learner*. This is a prediction model with the aid of which we predict the outcome for new objects. A learner is said to be *good* if it accurately predicts an outcome. In this text we only treat supervised problems which means that we also deal with outcome measurements. In unsupervised learning problems just the features are observed and the task is rather to describe how the data is organized or clustered. Statisticians speak about *regression* when quantitative outputs shall be predicted, and *classification* refers to the prediction of qualitative outputs.In this section we mostly follow [6] and [8].

Aim of the previous section was to discuss the linear regression model (4) which is usually fitted by using least squares. For some reasons we now try to extend the linear model framework:

The least squares estimates often have low bias (in our situation even no bias) but large variance. So on the one hand, there is the *prediction accurracy* that we want to improve by shrinking or setting some of the coefficients to zero. Thereby we sacrifice a little bias but we can reduce the variance of the predicted values at the same time. This results in an improvement of the overall prediction accurracy.

On the other hand we seek a better or easier way of *interpretation*. In many cases some predictor variables play a less important role than others. With a large number of variables we often want to determine a smaller subset that exhibits the strongest effects. For that purpose we sacrifice some details in order to get a better overview.

For some investigations it might happen that there are more parameters in the linear regression model in (4) to be estimated than there are available samples, i.e. $p > n$. We already know that in other cases the least squares method has a unique solution, but in this scenario that is no longer true. One should either try somehow to get more observations or use some regularization penalty.

Here we discuss two classes of methods using least squares:

- **Subset selection:** The name of this class of methods is already self-

explanatory. One tries to reduce the set of variables on which one then fits a model.

- **Shrinkage:** As an alternative to the discrete process of subset selection we can use all p predictor variables. This model is "more continous": It shrinks the coefficient estimates towards zero and does not suffer much from high variability.

One method that shall be mentioned here is the *best subset selection*. This approach finds for each $k \in \{0, 1, \ldots, p\}$ the subset of size k that gives the smallest RSS.

But instead of going more into detail here we will explain the theory about regularization next in order to be able to define the LASSO and ridge regression.

5.1 Penalty Methods for the Case $p > n$ and Definition of the LASSO

Now let us analyse the case where in the linear regression model it is $p > n$, i.e. there are more parameters to be estimated than available samples. Then there is no longer a unique least squares coefficient estimate. The idea is now constraining to a subset of coefficients:

Definition 5.1. *The set of all the indices for which β_j^0 is not equal to zero,*

$$S_0 := \{j : \beta_j^0 \neq 0\},$$

is called active set. *Its cardinal number $s_0 := |S_0|$ is the* sparsity index *of β^0.*

"Believing" that in fact only s_0 of the β_j^0 are non-zero, we get similarly to before (see page 22) the overall squared accuracy $\frac{\sigma^2}{n} s_0$, if we only take those variables $X^{(j)}$ into account with $j \in S_0$. The problem is that in general we don't know the active set.

Recall that we obtained the least squares estimator by minimizing $\mathrm{RSS}(\beta)$. The LASSO does the same subject to $\sum_{j=1}^{p} |\beta_j| \leq t$, where $t \geq 0$ is a so called *tuning parameter*.

For being able to define the LASSO estimator as a penalized least squares procedure, one needs some basic knowledge about optimization. More precisely it is important to understand the aim of penalty methods that form an helpful

concept in constraint optimization, i.e. problems of the form:

$$
\begin{array}{ll}
\text{minimize} & f(x) \\
\text{with the constraints} & g_1(x) \leq 0 \\
 & g_2(x) \leq 0 \\
 & \vdots \\
 & g_p(x) \leq 0,
\end{array}
\tag{8}
$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, p$.

As described in [19], their idea consists of solving a sequence of unconstrained optimization problems, that are formed by adding penalty terms to the objective function. These penalty terms are weighted by a positive parameter - the *penalty parameter*. The penalty problems have the form

$$\min_x f(x) + \alpha\pi(x)$$

with the penalty parameter $\alpha > 0$ and the penalty-function $\pi : \mathbb{R}^n \to \mathbb{R}$, $\pi(x) = 0$ on X and $\pi(x) > 0$ on $\mathbb{R}^n \setminus X$. Here one calls the set X containing all $x \in \mathbb{R}^n$ for which the constraints are fulfilled *permissible range*.

It was not mentioned yet what a penalty-function exactly is.

Definition 5.2. *A map $\pi : \mathbb{R} \to \mathbb{R}$ is a* penalty function *for the optimization problem as in (8) if these three conditions are fulfilled:*

1. *π is continuous*
2. *$\pi(x) \geq 0 \quad \forall x \in \mathbb{R}^n$*
3. *$\pi(x) = 0$ if and only if x is feasible, i.e. if $g_i(x) \leq 0$ for all $i \in \{1, \ldots, p\}$.*

In [3] there is a good example given that introduces the ℓ^1-penalty:

Example 5.3. *Let the functions $h_1(x) := x - 2$, $h_1 : \mathbb{R} \to \mathbb{R}$ and $h_2(x) := -(x+1)^3$, $h_2 : \mathbb{R} \to \mathbb{R}$ be given. We want that $h_1(x) \leq 0$ and $h_2(x) \leq 0$ hold. That is the case for all $x \in [-1, 2]$*

$$h_1^+(x) := \max(0, h_1(x)) = \begin{cases} 0 & \text{if } x \leq 2 \\ x - 2 & \text{otherwise} \end{cases}$$

$$h_2^+(x) := \max(0, h_2(x)) = \begin{cases} 0 & \text{if } x \geq -1 \\ -(x+1)^3 & \text{otherwise} \end{cases}$$

Here the penalty-function becomes

$$\pi(x) = h_1^+(x) + h_2^+(x) = \begin{cases} x-2 & \text{if } x > 2 \\ 0 & \text{if } -1 \le x \le 2 \\ -(x+1)^3 & \text{if } x < -1 \end{cases}$$

So ℓ^1-penalty refers to a *absolute value penalty function* of the form $\sum_{i=1}^{p} |h_i(x)|$, where the the summation is taken over all constraints that are violated at x. This is exactly the ℓ^1-norm. Analogously one defines ℓ^p-penalty for $p \ge 2$.

Back to our situation we now define the LASSO estimator.

Definition 5.4 (The LASSO estimator)**.** *A good choice for the needed regularization penalty is ℓ^1-penalty, i.e. the* LASSO *is defined by*

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \left(\frac{\|Y - X\beta\|_2^2}{n} + \lambda \|\beta\|_1 \right). \tag{9}$$

with $\|Y - X\beta\|_2^2 = \sum_{i=1}^{n} (Y_i - (X\beta)_i)^2$, $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ and where λ is a penalty parameter.

The acronym LASSO stands for *"Least absolute shrinkage and selection operator"*. Namely we will see that the LASSO combines both shrinkage and variable selection.

5.2 The Ridge Regression

Ridge regression is a shrinkage method that imposes a penalty on the size of the regression coefficient which shrinks them. Let us have a look at its exact definition.

Definition 5.5. *The ridge regression estimator is defined as*

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \}.$$

So in the ridge regression, instead of ℓ^1-penalty, ℓ^2 penalty is used.

Similarly as in the model for the LASSO, we have here a penalty term $\lambda \ge 0$. The larger the value of λ is, the greater becomes the amount of shrinkage towards zero. Alternatively one can write

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\}, \quad \text{subject to } \sum_{j=1}^{p} \beta_j^2 \le t. \tag{10}$$

The case of correlation of some predictor variables leads to high variance in the "ordinary" linear regression model. For example, a very large positive coefficient on one variable can be cancelled by a similarly large negative coefficient on another predictor variable that is correlated. The size constraint in the regression model as in (10) can solve this problem.

Another way to define the ridge regression is to write it in matrix form. With

$$\mathrm{RSS}(\lambda) = (y - X\beta)^t(y - X\beta) + \lambda\beta^t\beta,$$

the ridge regression solutions are easily seen to be

$$\hat{\beta}^{ridge} = (X^tX + \lambda\mathbb{I})^{-1}X^ty,$$

where $\mathbb{I}$ is here the $p \times p$ identity matrix. The addition of a positive constant to the diagonal elements of X^tX makes the problem nonsingular. This was the main motivation for introducing the ridge regression in statistics in 1970 by Hoerl and Kennard.

In statistics it is often very useful to apply the singular value decomposition, which we abbreviate in the following by SVD. The reader is recommended at this point to read the relevant pages of [17] (p. 364 ff.). A better insight of what is happening in ridge regression is attained using the SVD. The SVD of the $n \times p$ matrix X reads

$$X = UDV^t,$$

where U and V are $n \times p$ and $p \times p$ orthogonal matrices respectively. The columns of U span the column space of X and the columns of V span the row space. D is a $p \times p$ diagonal matrix. Its diagonal entries $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ are called singular values of X. If there is a value $d_j = 0$, then X is singular.

The least squares fitted values have now the form

$$\begin{aligned} X\hat{b} &= X(X^tX)^{-1}X^tY = UDV^t((UDV^t)^tUDV^t)^{-1}(UDV^t)^tY = \\ &= UDV^t(VDU^tUDV^t)^{-1}VDU^tY = UDV^t(VD^2V^t)^{-1}VDU^tY \\ &= UU^tY \end{aligned}$$

We see that the least squares method computes the coordinates of Y with respect to the orthonormal basis U.

The ridge solutions are

$$\begin{aligned} X\hat{\beta}^{ridge} &= X(X^tX + \lambda I)^{-1}X^tY = UDV^t((UDV^t)^tUDV^t + \lambda I)^{-1}(UDV^t)^tY \\ &= UDV^t(VD^2V^t + \lambda VV^t)^{-1}VDU^tY = UD(D^2 + \lambda I)^{-1}DU^tY. \end{aligned}$$

D is a diagonal matrix and thus also $D(D^2 + \lambda I)D$. That is why we can write the ridge regression fitted values as

$$X\hat{\beta}^{ridge} = \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^t Y,$$

where u_j stands for the j-th column of U.

It holds $\frac{d_j^2}{d_j^2+\lambda} \leq 1$, because $\lambda \geq 0$. Ridge regression computes the coordinates of Y with respect to the orthonormal basis U and shrinks then these coordinates by the factors $\frac{d_j^2}{d_j^2+\lambda}$ aferwards. Therefore, as d_j gets small, the amount of shrinkage becomes greater. The small singular values d_j correspond to directions in the column space of X, which have small variance. Ridge regression shrinks these directions the most. Indeed, if we analyse the *principal components* of the variables in a centred matrix X, we will see that. The SVD of X gives us

$$X^t X = V D^2 V^t.$$

The eigenvectors v_j that are the columns of V are called *principal components directions* of X. That is why above expression is called eigen decomposition of $X^t X$. The first principal component direction v_1 has the largest variance among all normalized linear combinations of the columns of X. It is

$$\mathrm{Var}\left[Xv_1\right] = \mathrm{Var}\left[u_1 d_1\right] = \frac{d_1^2}{N}.$$

The variable Xv_1 is called *first principal component* wherefore one calls u_1 normalized *first principal component*. Subsequent principal components Xv_j are orthogonal to the previous ones and have therefore maximum variance $\frac{d_j}{N}$. The last principal component has minimum variance.

Definition 5.6. *The monotone increasing function*

$$\mathrm{df}(\lambda) = \mathrm{tr}\left[(X^t X + \lambda I)^{-1} X^t\right] = \sum_{i=1}^{p} \frac{d_i^2}{d_i^2 + \lambda}$$

is called effective degrees of freedom *of the ridge regression fit.*

Observe, that if $\lambda = 0$, which means that no regularization happens, it holds $\mathrm{df}(\lambda) = p$. Further it is obviously $df(\lambda) \to 0$ as $\lambda \to \infty$.

Finally let us now interpret the ridge regression model geometrically. Figure 4 shows the case where $p = 2$. Here the constraints correspond to the blue

circle and the ellipses to the contours of residual sum of squares. The RSS of the inner ellipse is smaller than the RSS of the outer one. It is minimized at the ordinary least squares (OLS) estimate. The ridge regression method intends to minimize the ellipse size and circle simultanously. Its estimate is the point that has the circle in common with the ellipse.

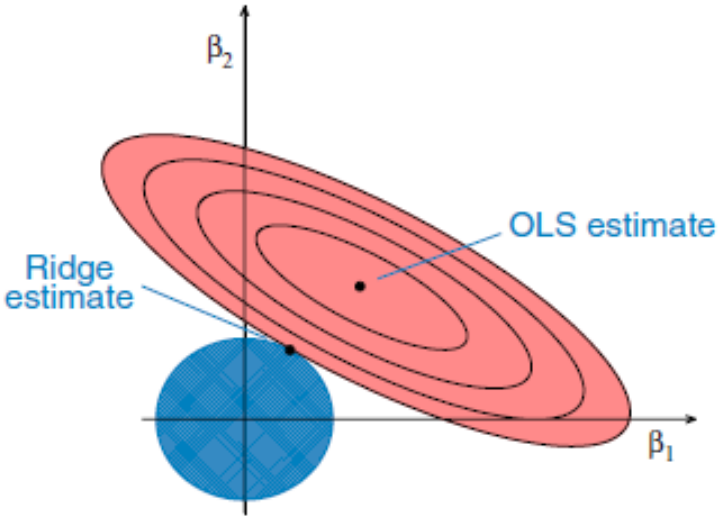

Figure 4: Geometric Interpretation of Ridge Regression
Source: `https://onlinecourses.science.psu.edu/stat857/node/155`

5.3 Ridge Regression versus LASSO

Even though the orthonormal case will never occur in practise, we consider this very unlikely scenario anyway in order to get an insight about the nature of shrinkage. So if the matrix X is orthonormal, both procedures have explicit solutions and apply a simple transformation to the least squares estimate $\hat{b}$. In the previous subsection we have seen that both methods apply a simple transformation to the least squares estimate $\hat{\beta}_j$. Ridge regression does a proportional shrinkage, whereas LASSO translates each coefficient by a constant factor λ and truncates at zero.

The use of ℓ^1-penalty makes the solutions nonlinear in the Y_i. That is why for the LASSO there do not exist any closed form expression as for ridge regression.

The LASSO in fact does some kind of continuous subset selection. If one chooses t to be sufficiently small then some of the coefficients will be exactly

zero. In cotrast, if t is chosen larger than the ℓ^1-norm of the least squares estimate $\hat{b}$, then the LASSO-estimates are exactly the $\hat{b}_j$'s.

Let us now have a look at the illustration of the LASSO. In Figure 5 we see that in contrast to the case of ridge regression, the constraint now forms a rhombus. The LASSO-estimate corresponds to the point where the ellipse hits the diamond. We have to distinguish between two cases, because compared to the circle in the ridge regression model, the rhombus has corners. If the solution occurs at a corner, then it has one parameter β_j equal to zero. Assuming an higher dimensional model, the equilateral quadrilateral becomes a rhomboid. It has a lot of corners, faces and flat edges. So there are much more possibilities for the ellipse to intersect at a corner, i.e. there are more parameters that possibly equal zero.

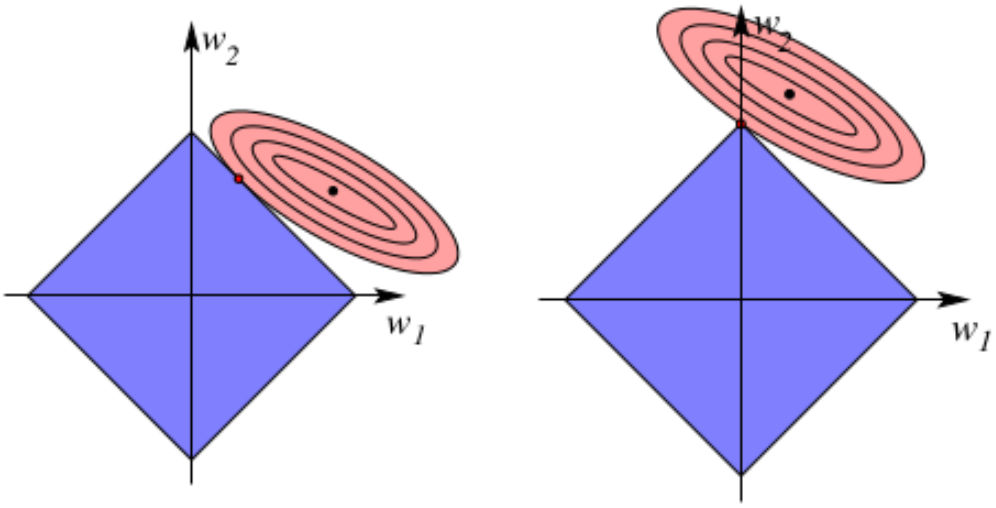

Figure 5: Geometric Interpretation of the LASSO. In this diagram w is used instead of β. Apart from that it is convenient for our situation.
Source:
`http://www.cmat.edu.uy/~mordecki/modelos/pdf_files/cours_Montevideo_1.pdf`

As the observations above already let us guess, we can generalize LASSO and ridge regression with the following criterion

$$\tilde{\beta} = \arg\min_{\beta} \{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q\},$$

with $q \geq 0$. Clearly, $q = 1$ corresponds to the LASSO and the case $q = 2$ to ridge regression. With $q = 0$ variable subset selection is at hand. Because

then just parameters that are non-zero are counted.

Now, we want to give a Bayesian interpretation of the above criterion. To this end, let us have a look at the case of ridge regression for instance and suppose that the regression coefficients β_j are independent and normally distributed with mean 0 and variance τ^2. Furthermore we work with a Gaussian sampling model $Y \sim \mathcal{N}(X\beta, \sigma^2\mathbb{I})$ as usual. We go on now with the Bayes'inference procedure as it is described in Section 3.1. As in equation (3) there, it holds

$$f(\beta|Y) \propto f(Y|\beta)f(\beta),$$

for the prior density function f and the corresponding posterior density.

For appropriate constants C_1 and C_2 then it is

$$f(\beta) = C_1 \exp\left\{-\frac{\|\beta\|_2^2}{2\tau^2}\right\}, \text{ and } f(Y|\beta) = C_2 \exp\left\{-\frac{\|Y - X\beta\|_2^2}{2\sigma^2}\right\}.$$

In consequence, the posterior distribution has the form

$$f(\beta|Y) = C_3 \exp\left\{-\frac{\|Y - X\beta\|_2^2 + \frac{\sigma^2}{\tau^2}\|\beta\|_2^2}{2\sigma^2}\right\}$$

for a suitable constant C_3. Now we set $\lambda = \frac{\sigma^2}{\tau^2}$, take the negative logarithm of the above posterior distribution to obtain

$$-\log f(\beta|Y) = -\log C_3 + \frac{\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2}{2\sigma^2}.$$

Thus, for the constant $C_4 = -2\sigma^2 \log C_3$, this can be written as

$$\sum_{i=1}^{N}\left(Y_i - \sum_{j=1}^{p}\beta_j X_j\right)^2 + \lambda\sum_{i=1}^{p}\beta_i^2 + C_4.$$

Taking the mode of the posterior density means to minimize the negative log-posterior density. Therefore the ridge regression estimate is in fact the mode of the posterior distribution.

Similarly, the LASSO estimate can be shown to be the Bayes posterior mode for independent double exponentially distributed (i.e. Laplace distributed) priors with parameters $\mu = 0$ and $\gamma = \frac{1}{\lambda}$. Recall that this means that the priors have the density

$$f(\beta_j) = \lambda\frac{1}{2}\exp\{-\lambda|\beta_j|\}.$$

It is to mention that as well subset selection can be seen as a Bayesian estimate, also derived as some posterior mode.

Figure 6 shows the contours of ℓ^q-norms in three dimensions for various values of q. We see that as the value of q gets smaller, the size of the corresponding ℓ^q-ball decreases as well.

Figure 6: ℓ^p ball in three dimensions
Source: `http://www.stat.ucla.edu/~ybzhao/teaching/stat101c/`, modified graphic

Now let us consider the case where q has a value different to $0, 1$ or 2. For a value $q \in (1, 2)$ we expect some kind of compromise between the LASSO and ridge regression. The case $q > 1$ makes $|\beta_j|$ differentiable at 0, which is why no coefficient will be set exactly to zero as LASSO does. In 2005 Zou and Hastie introduced a possible way out of this problem.

Definition 5.7 (Elastic Net Penalty)**.** *The estimator defined via "elastic net penalty" is given by*

$$\hat{\beta}^{\text{en}} := \lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1 - \alpha) \left| \beta_j \right| \right).$$

The elastic net penalty is a good alternative to LASSO and ridge regression. It selects variables like the LASSO, while it shrinks the coefficients of correlated predictor variables like the ridge regression model. Furthermore it has considerable computational advantages. Figure 7 compares ℓ^1-, ℓ^2- and the Elastic Net Penalty. The latter one has sharp (i.e. non-differentiable) corners, while penalty for $q > 1$ does not.

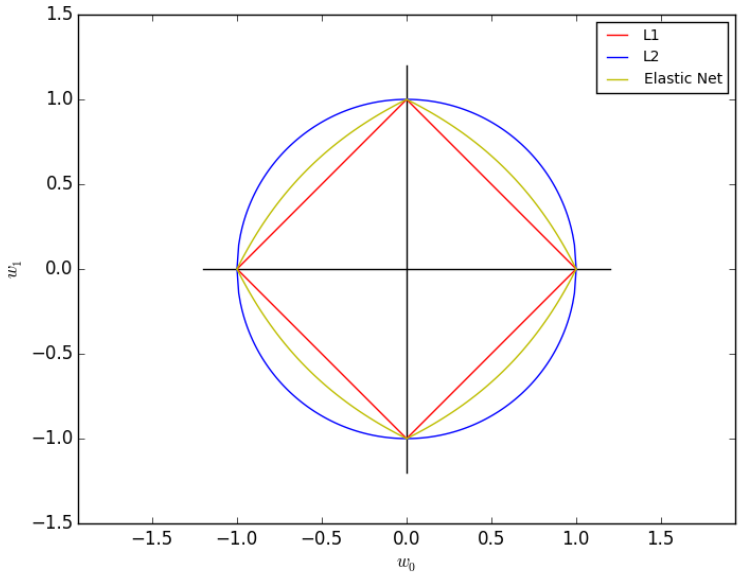

Figure 7: Comparison of three penalty types
Source: `http://scikit-learn.org/stable/modules/sgd.html`

6 Error Estimations for the LASSO

In this section we first introduce the so called "basic inequality" based on which we obtain a consistency result and will be able to estimate the error of the LASSO. For that purpose we will establish the commonly named "compatibility condition", which we will use to prove a certain bound for the ℓ^1-error as well as an estimation for the prediction error.

Recall that in the following it is assumed that the linear model given in Definition 4.2 holds exactly with the parameter β^0.

Lemma 6.1 (Basic Inequality)**.** *We have*

$$\frac{\left\|X(\hat{\beta}-\beta^0)\right\|_2^2}{n}+\lambda\left\|\hat{\beta}\right\|_1 \leq \frac{2\varepsilon^t X(\hat{\beta}-\beta^0)}{n}+\lambda\left\|\beta^0\right\|_1$$

Proof. By the definition of $\hat{\beta}$, it clearly holds

$$\frac{\left\|Y-X\hat{\beta}\right\|_2^2}{n}+\lambda\left\|\hat{\beta}\right\|_1 \leq \frac{\|Y-X\beta^0\|_2^2}{n}+\lambda\left\|\beta^0\right\|_1.$$

By plugging in for $Y = X\beta^0 + \varepsilon$, this is equivalent to

$$\frac{\left\|-X(-\beta^0+\hat{\beta})+\varepsilon\right\|_2^2}{n} + \lambda\left\|\hat{\beta}\right\|_1 \leq \frac{\|\varepsilon\|_2^2}{n} + \lambda\left\|\beta^0\right\|_1.$$

We expand the term $\left\|-X(-\beta^0+\hat{\beta})+\varepsilon\right\|_2^2$ and obtain

$$\frac{\left\|-X(\hat{\beta}-\beta^0)\right\|_2^2}{n} \underbrace{-\frac{1}{n}(X(\hat{\beta}-\beta^0))^t\varepsilon - \frac{1}{n}\varepsilon^t(X(\hat{\beta}-\beta^0))}_{-\frac{2\varepsilon^t(X(\hat{\beta}-\beta^0))}{n}} + \frac{\|\varepsilon\|_2^2}{n} + \lambda\left\|\hat{\beta}\right\|_1$$

$$\leq \frac{\|\varepsilon\|_2^2}{n} + \lambda\left\|\beta^0\right\|_1.$$

Finally, this is equivalent to

$$\frac{\left\|X(\hat{\beta}-\beta^0)\right\|_2^2}{n} + \lambda\left\|\hat{\beta}\right\|_1 \leq \frac{2\varepsilon^t X(\hat{\beta}-\beta^0)}{n} + \lambda\left\|\beta^0\right\|_1,$$

which is the required inequality. □

Now let us have a look back again at the basic inequality given in Lemma 6.1 above. The term that contains the measurement error, that is

$$\frac{2\varepsilon^t X(\hat{\beta}-\beta^0)}{n},$$

is called "empirical process" part in the case of quadratic loss. It clearly holds

$$2|\varepsilon^t X(\hat{\beta}-\beta^0)| \leq (\max_{1\leq j\leq p} 2|\varepsilon^t X^{(j)}|)\left\|\hat{\beta}-\beta^0\right\|_1.$$

In order to get rid of this random part somehow, we now define the set

$$\mathcal{J} := \left\{\max_{1\leq j\leq p}\frac{2\varepsilon^t X^{(j)}}{n} \leq \lambda_0\right\}$$

and assume that $2\lambda_0 \leq \lambda$. On $\mathcal{J}$ the basic inequality now reads

$$\frac{\left\|X(\hat{\beta}-\beta^0)\right\|_2^2}{n} + \lambda\left\|\hat{\beta}\right\|_1 \leq \lambda_0\left\|\hat{\beta}-\beta^0\right\|_1 + \lambda\left\|\beta^0\right\|_1 \tag{11}$$

Before we are able to prove the next Lemma, we need a probabilistic estimation. Readers that are familiar with Markov inequality arguments can skip the following Proposition.

Proposition 6.2 (Gaussian Tail Estimate)**.** *Let X be $\mathcal{N}(0,1)$ distributed. Then for $a > 0$ it holds*

$$P[X \geq a] \leq e^{-\frac{a^2}{2}}.$$

Proof. By Markov's inequality it is

$$\mathrm{P}\left[X \geq a\right] = \mathrm{P}\left[e^{tX} \geq e^{ta}\right] \leq \mathrm{E}\left[e^{tX}\right] e^{-ta} \quad \forall t > 0.$$

X is a standard normal random variable, so $\mathrm{E}\left[e^{tX}\right] = e^{\frac{t^2}{2}}$. Thus the above inequality now reads

$$\mathrm{P}\left[X \geq a\right] \leq e^{\frac{t^2}{2}} e^{-ta}.$$

Since this holds $\forall t > 0$ we can minimize $e^{\frac{t^2}{2} - ta}$ which means to minimize $\frac{t^2}{2} - ta$. This expression attains its minimum at $t = a$. Therefore

$$\mathrm{P}\left[X \geq a\right] \leq e^{\frac{a^2}{2} - a^2} = e^{-\frac{a^2}{2}}.$$

This completes the proof. □

Dealing with Gaussian errors, the target is to prove the following lemma which states that $\mathcal{J}$ has large probability.

Lemma 6.3. *Let $\hat{\Sigma}_{jj}$, $j = 1, \ldots, p$ denote the diagonal elements of the scaled Gramian matrix $\hat{\Sigma} := X^t X \frac{1}{n}$. Suppose that they all equal 1. Then for all $t > 0$ and for $\lambda_0 := 2\sigma\sqrt{\frac{t^2 + 2\log p}{n}}$ it holds*

$$P\left[\mathcal{J}\right] \geq 1 - 2e^{-\frac{t^2}{2}}$$

Proof. The diagonal entries of $\hat{\Sigma}$ have the form $\frac{1}{n}\sum_{i=1}^n (X_i^{(j)})^2 = 1$ for all $1 \leq j \leq p$, which obviously means $\sum_{i=1}^n (X_i^{(j)})^2 = n$ for all $1 \leq j \leq p$. Define $V_j := \frac{\varepsilon^t X^{(j)}}{\sqrt{n\sigma^2}}$. By a simple calculation one can see that this is standard normal distributed. Indeed, it is

$$\mathrm{E}\left[V_j\right] = \mathrm{E}\left[\frac{\varepsilon^t X^{(j)}}{\sqrt{n\sigma^2}}\right] = \mathrm{E}\left[\frac{\sum_{i=1}^n \varepsilon_i X_i^{(j)}}{\sqrt{n\sigma^2}}\right] = \frac{1}{\sqrt{n\sigma^2}} \sum_{i=1}^n X_i^{(j)} \underbrace{\mathrm{E}\left[\varepsilon_i\right]}_{=0} = 0,$$

and

$$\begin{aligned}
\operatorname{Var}[V_j] &= \operatorname{E}\left[V_j^2\right] - \underbrace{\operatorname{E}[V_j]}_{=0}{}^2 = \operatorname{E}\left[\frac{\sum_{i=1}^n \varepsilon_i X_i^{(j)} \sum_{i=1}^n \varepsilon_i X_i^{(j)}}{\sqrt{n\sigma^2}\sqrt{n\sigma^2}}\right] \\
&= \frac{1}{n\sigma^2} \operatorname{E}\left[\sum_{i=1}^n \varepsilon_i^2 (X_i^{(j)})^2 + \sum_{1\le i\ne k\le n} \varepsilon_i \varepsilon_k X_i^{(j)} X_k^{(j)}\right] \\
&= \frac{1}{n\sigma^2} \sum_{i=1}^n \operatorname{Var}[\varepsilon_i] (X_i^{(j)})^2 + \frac{1}{n\sigma^2} \sum_{1\le i\ne k\le n} \underbrace{\operatorname{E}[\varepsilon_i]}_{=0} \underbrace{\operatorname{E}[\varepsilon_k]}_{=0} X_i^{(j)} X_k^{(j)} \\
&= \frac{1}{n\sigma^2}\sigma^2 \sum_{i=1}^n (X_i^{(j)})^2 = \frac{1}{n} n = 1.
\end{aligned}$$

So it is

$$\begin{aligned}
\operatorname{P}[\mathcal{J}^c] &= \operatorname{P}\left[\max_{1\le j\le p} \frac{2|\varepsilon^t X^{(j)}|}{n} > \lambda_0\right] = \operatorname{P}\left[\max_{1\le j\le p} 2|\varepsilon^t X^{(j)}| > \underbrace{2\sqrt{n}\sigma\sqrt{t^2+2\log p}}_{=n\lambda_0}\right] \\
&= \operatorname{P}\left[\max_{1\le j\le p} \frac{|\varepsilon^t X^{(j)}|}{\sqrt{n\sigma^2}} > \sqrt{t^2+2\log p}\right] = \operatorname{P}\left[\max_{1\le j\le p} |V_j| > \sqrt{t^2+2\log p}\right] \\
&\le p \operatorname{P}\left[|V_j| > \sqrt{t^2+2\log p}\right] = 2p \operatorname{P}\left[V_j > \sqrt{t^2+2\log p}\right] \\
&\le 2p \exp\left[-\frac{t^2+2\log p}{2}\right] = 2p \exp\left[-\frac{t^2}{2}\right] \underbrace{\exp[-\log p]}_{=\frac{1}{p}} = 2e^{-\frac{t^2}{2}},
\end{aligned}$$

which is equivalent to

$$P[\mathcal{J}] \ge 1 - 2e^{-\frac{t^2}{2}}.$$

This completes the proof. □

We now obtain a consistency result for the LASSO:

Corollary 6.4 (Consistency of the LASSO)**.** *Assume that* $\hat{\Sigma}_{jj} = 1$ *for all* j*. For some* $t > 0$ *define* $\lambda := 4\hat{\sigma}\sqrt{\frac{t^2+2\log p}{n}}$*, where* $\hat{\sigma}$ *is some estimator of* σ*. Further define* $\alpha := 2e^{-\frac{t^2}{2}} + \operatorname{P}[\hat{\sigma} \le \sigma]$*. Then we have*

$$\frac{2\left\|X(\hat{\beta} - \beta^0)\right\|_2^2}{n} \le 3\lambda \left\|\beta^0\right\|_1$$

with probability at least $1 - \alpha$*.*

Proof. Clearly on $\mathcal{J}$ with $2\lambda_0 \leq \lambda$ it holds

$$\frac{\left\|X(\hat{\beta}-\beta^0)\right\|_2^2}{n} + \lambda \left\|\hat{\beta}\right\|_1 \leq \underbrace{\lambda_0 \left\|\hat{\beta}-\beta^0\right\|_1}_{\leq \frac{1}{2}\lambda \left\|\hat{\beta}-\beta^0\right\|_1} + \lambda \left\|\beta^0\right\|_1 .$$

This is equivalent to

$$\begin{aligned}\frac{2\left\|X(\hat{\beta}-\beta^0)\right\|_2^2}{n} + 2\lambda \left\|\hat{\beta}\right\|_1 &\leq \lambda \left\|\hat{\beta}-\beta^0\right\|_1 + 2\lambda \left\|\beta^0\right\|_1 \\ &\leq \lambda \left\|\hat{\beta}\right\|_1 + \underbrace{\lambda \left\|\beta^0\right\|_1 + 2\lambda \left\|\beta^0\right\|_1}_{=3\lambda\|\beta^0\|_1}.\end{aligned}$$

So in the end we now get

$$\underbrace{\frac{2\left\|X(\hat{\beta}-\beta^0)\right\|_2^2}{n} + \lambda \left\|\hat{\beta}\right\|_1}_{\geq 2\frac{\|X(\hat{\beta}-\beta^0)\|_2^2}{n}} \leq 3\lambda \left\|\beta^0\right\|_1 .$$

□

For the following, recall that on page 25 we defined S_0 to be the active set. It is the set of indices for which β_j^0 is nonzero. We call its cardinal number s_0 the corresponding sparsity index.

Let us introduce now some notation. To exploit sparsity of β^0, we write $\beta_{j,S} := \beta_j \mathbb{1}_{j\in S}$ for an index $S \subset \{1,\ldots,p\}$ and hence $\beta_{j,S^c} := \beta_j \mathbb{1}_{j\notin S}$. Then obviously $\beta = \beta_S + \beta_{S^c}$.

Lemma 6.5. *On $\mathcal{J}$ it holds with $\lambda \geq 2\lambda_0$*

$$\frac{2\left\|X(\hat{\beta}-\beta^0)\right\|_2^2}{n} + \lambda \left\|\hat{\beta}_{S_0^c}\right\|_1 \leq 3\lambda \left\|\hat{\beta}_{S_0} - \beta^0_{S_0}\right\|_1 .$$

Proof. On $\mathcal{J}$ the basic inequality with $\lambda \geq 2\lambda_0$ reads

$$\frac{2\left\|X(\hat{\beta}-\beta^0)\right\|_2^2}{n} + 2\lambda \left\|\hat{\beta}\right\|_1 \leq 2\frac{1}{2}\lambda \left\|\hat{\beta}-\beta^0\right\|_1 + 2\lambda \left\|\beta^0\right\|_1 .$$

By the above definition of β_{j,S^c} it is

$$\left\|\hat{\beta}\right\|_1 = \left\|\hat{\beta}_{S_0}\right\|_1 + \left\|\hat{\beta}_{S_0^c}\right\|_1 .$$

Here we now apply the reverse triangle inequality to get

$$\left\|\hat{\beta}\right\|_1 \geq \left\|\beta^0_{S_0}\right\|_1 - \left\|\hat{\beta}_{S_0} - \beta^0_{S_0}\right\|_1 + \left\|\hat{\beta}_{S_0^c}\right\|_1 .$$

On the right side of the basic inequality one uses

$$\left\|\hat{\beta} - \beta^0\right\|_1 = \left\|\hat{\beta}_{S_0} - \beta^0_{S_0}\right\|_1 + \left\|\hat{\beta}_{S_0^c}\right\|_1 .$$

Obviously it is

$$\left\|\beta^0\right\|_1 = \left\|\beta^0_{S_0}\right\|_1 ,$$

so the required inequality follows. □

We now want to derive the so called "compatibility conditions" on the design matrix X. The idea behind it is to discard the ℓ^1-term on the right hand side of the above equation incorporating it into the ℓ^2-term on the left side.

In general it is well known that for p-norms on n-dimensional vector spaces with $1 \leq p \leq r \leq \infty$ it holds

$$\|x\|_r \leq \|x\|_p \leq n^{\frac{1}{p}-\frac{1}{r}} \|x\|_r$$

by the Cauchy-Schwarz-inequality.

That is why in our case we can write

$$\left\|\hat{\beta}_{S_0} - \beta^0_{S_0}\right\|_1 \leq \sqrt{s_0} \left\|\hat{\beta}_{S_0} - \beta^0_{S_0}\right\|_2$$

being just interested in the s_0 non-vanishing coordinates.

Again let $\hat{\Sigma} = \frac{1}{n} X^t X$ be the scaled Gramian matrix, then

$$\frac{\left\|X(\hat{\beta} - \beta^0)\right\|_2^2}{n} = (\hat{\beta} - \beta^0)^t \hat{\Sigma} (\hat{\beta} - \beta^0).$$

Let now hold

$$\left\|\hat{\beta}_{S_0} - \beta^0_{S_0}\right\|_2^2 \leq \frac{(\hat{\beta} - \beta^0)^t \hat{\Sigma} (\hat{\beta} - \beta^0)}{\Phi_0^2} \tag{12}$$

for some constant and positive Φ_0. Then it is possible to appropriately proceed with the chain of inequalities. We then may restrict to $\mathcal{J}$ where by the basic inequality from Lemma 6.1 it is

$$\left\|\hat{\beta}_{S_0^c}\right\|_1 \leq 3 \left\|\hat{\beta}_{S_0} - \beta^0_{S_0}\right\|_1 ,$$

because requiring (12) for all β would need $\hat{\Sigma}$ to be non-singular.

Definition 6.6 (Compatibility Condition)**.** *We say that the* compatibility condition *is met for the set S_0, if for some $\Phi_0 > 0$ and for all β satisfying $\left\|\beta_{S_0^c}\right\|_1 \leq 3 \left\|\beta_{S_0}\right\|_1$ it holds that*

$$\left\|\beta_{S_0}\right\|_1^2 \leq (\beta^t \hat{\Sigma} \beta) \frac{s_0}{\Phi_0^2}.$$

One calls Φ_0^2 compatibility constant *of the matrix $\hat{\Sigma}$.*

Theorem 6.7. *Suppose that the compatibility condition holds for S_0. Then on $\mathcal{J}$ we have for $\lambda \geq 2\lambda_0$,*

$$\frac{\left\|X(\hat{\beta} - \beta^0)\right\|_2^2}{n} + \lambda \left\|\hat{\beta} - \beta^0\right\|_1 \leq \frac{4\lambda^2 s_0}{\Phi_0^2}.$$

Proof. First observe that it holds

$$\begin{aligned} &\frac{2\left\|X(\hat{\beta} - \beta^0)\right\|_2^2}{n} + \lambda \left\|\hat{\beta} - \beta^0\right\|_1 \\ &= \frac{2\left\|X(\hat{\beta} - \beta^0)\right\|_2^2}{n} + \lambda \left\|\hat{\beta}_{S_0} - \beta^0_{S_0}\right\|_1 + \lambda \left\|\hat{\beta}_{S_0^c}\right\|_1. \end{aligned}$$

By Lemma 6.5 this is bounded from above by

$$\lambda \left\|\hat{\beta}_{S_0} - \beta^0_{S_0}\right\|_1 + 3\lambda \left\|\hat{\beta}_{S_0} - \beta^0_{S_0}\right\|_1 = 4\lambda \left\|\hat{\beta}_{S_0} - \beta^0_{S_0}\right\|_1.$$

By the compatibility condition, Definition 6.6, this is again bounded from above by

$$4\lambda \sqrt{\frac{(\hat{\beta} - \beta^0)^t \hat{\Sigma} (\hat{\beta} - \beta^0) s_0}{\Phi_0^2}} = \frac{4\lambda \sqrt{s_0} \left\|X(\hat{\beta} - \beta^0)\right\|_2}{\sqrt{n} \Phi_0}.$$

Now we use that $4uv \leq u^2 + v^2$ (with $u = \frac{\left\|X(\hat{\beta} - \beta^0)\right\|_2}{\sqrt{n}}$ and $v = \frac{\lambda\sqrt{s_0}}{\Phi_0}$), which holds by being equivalent to $0 \leq u^2 - 4uv + 4v^2 = (u - 2v)^2$. So the above expression is once more bounded from above by

$$\frac{\left\|X(\hat{\beta} - \beta^0)\right\|_2}{n} + \frac{4\lambda^2 s_0}{\Phi_0^2}.$$

This yields

$$\frac{\left\|X(\hat{\beta}-\beta^0)\right\|_2^2}{n}+\lambda\left\|\hat{\beta}-\beta^0\right\|_1\leq\frac{4\lambda^2 s_0}{\Phi_0^2},$$

which is exactly the inequality that was to be proven.

□

The above theorem is yielding for two reasons. On the one hand it gives us the bound

$$\left\|\hat{\beta}-\beta^0\right\|_1\leq\frac{4\lambda s_0}{\Phi_0^2}$$

for the ℓ^1-error, but it also leads to this estimation for the prediction error:

$$\frac{\left\|X(\hat{\beta}-\beta^0)\right\|_2}{n}\leq\frac{4\lambda^2 s_0}{\Phi_0^2}.$$

A further result that follows immediately from this theorem is the following:

Corollary 6.8. *Assume that $\hat{\sigma}_j^2=1$ for all j and that the compatibility condition holds for S_0 with $\hat{\Sigma}$ normalized in this way. For some $t>0$ define $\lambda:=4\hat{\sigma}\sqrt{\frac{t^2+2\log p}{n}}$, where $\hat{\sigma}^2$ is some estimator of σ^2. Furthermore let $\alpha:=2e^{-\frac{t^2}{2}}+\mathrm{P}\left[\hat{\sigma}\leq\sigma\right]$. Then with probability at least $1-\alpha$ we have*

$$\frac{\left\|X(\hat{\beta}-\beta^0)\right\|_2^2}{n}+\lambda\left\|\hat{\beta}-\beta^0\right\|_1\leq\frac{4\lambda^2 s_0}{\Phi_0^2}$$

7 Refinements of the Model

7.1 Linear Approximation of the Truth

We search now a function $f(X)$ for predicting Y given values of the regressor X. Therefore we assume that X and Y have the joint distribution $\Pr(X, Y)$. The expected squared prediciton error then reads

$$\mathrm{EPE}(f) = \mathrm{E}\left[Y - f(X)\right]^2.$$

By conditioning on X it is possible to rewrite this to

$$\mathrm{EPE}(f) = \mathbb{E}_X \mathbb{E}_{Y|X}([Y - f(X)]^2|X).$$

It suffices to minimize EPE pointwise:

$$\arg\min_c \ \mathbb{E}_{Y|X}([Y - c]^2|X = x),$$

where the solution is given by

$$f(x) = \mathrm{E}\left[Y|X = x\right].$$

This is called the *regression function*. In this section we briefly discuss the case where we assume the regression function $\mathrm{E}[Y] := f^0$ possibly not to be a sparse linear combination of the vectors $X^{(j)}$. The next lemmas will show, that there are very similar results to the theory established in the previous section.

Lemma 7.1 (General Basic Inequality)**.** *For any vector β^* it is*

$$\frac{\left\|X\hat{\beta} - f^0\right\|_2^2}{n} + \lambda \left\|\hat{\beta}\right\|_1 \leq \frac{2\varepsilon^t X(\hat{\beta} - \beta^*)}{n} + \lambda \left\|\beta^*\right\|_1 + \frac{\|X\beta^* - f^0\|_2^2}{n}.$$

Proof. By the definition of $\hat{\beta}$ it clearly holds

$$\frac{\left\|Y - X\hat{\beta}\right\|_2^2}{n} + \lambda \left\|\hat{\beta}\right\|_1 \leq \frac{\|Y - X\beta^*\|_2^2}{n} + \lambda \left\|\beta^*\right\|_1$$

for any vector β^*. Now by plugging in for $Y = f^0 + \varepsilon$ this inequality reads

$$\frac{\left\|f^0 + \varepsilon - X\hat{\beta}\right\|_2^2}{n} + \lambda \left\|\hat{\beta}\right\|_1 \leq \frac{\|f^0 + \varepsilon - X\beta^*\|_2^2}{n} + \lambda \left\|\beta^*\right\|_1.$$

This is equivalent to

$$\frac{\left\|-(X\hat{\beta}-f^0)\right\|_2^2}{n}\underbrace{-\frac{(X\hat{\beta}-f^0)^t\varepsilon}{n}-\frac{\varepsilon^t(X\hat{\beta}-f^0)}{n}}_{-2\varepsilon^t(X\hat{\beta}-f^0)\frac{1}{n}}+\frac{\|\varepsilon\|_2^2}{n}+\lambda\left\|\hat{\beta}\right\|_1$$
$$\leq\frac{\|X\beta^*-f^0\|_2^2}{n}+\lambda\|\beta^*\|_1\underbrace{+\frac{(X\beta^*-f^0)^t\varepsilon}{n}+\frac{\varepsilon^t(X\beta^*-f^0)}{n}}_{2\varepsilon^t(X\beta^*-f^0)\frac{1}{n}}+\frac{\|\varepsilon\|_2^2}{n}+\lambda\|\beta^*\|_1.$$

And this now becomes

$$\frac{\left\|X\hat{\beta}-f^0\right\|_2^2}{n}+\lambda\left\|\hat{\beta}\right\|_1$$
$$\leq\frac{\|X\beta^*-f^0\|_2^2}{n}\underbrace{+2\varepsilon^t(X\beta^*-f^0)\frac{1}{n}+2\varepsilon^t(X\hat{\beta}-f^0)\frac{1}{n}}_{\frac{2\varepsilon^t(X\hat{\beta}-X\beta^*)}{n}}+\lambda\|\beta^*\|_1.$$

□

The above Lemma shows that in our further derivations we have to take into account the approximation error

$$\frac{\|X\beta^*-f^0\|_2^2}{n}$$

as for instance in the next Lemma.

Therefore we firstly define $S_*:=\{j:\beta_j^*\neq 0\}$. Bear in mind that this set depends on the arbitrary vector β^*.

Lemma 7.2. *On $\mathcal{J}$ it holds for any vector β^**

$$\frac{4\left\|X\hat{\beta}-f^0\right\|_2^2}{n}+3\lambda\left\|\hat{\beta}_{S_*^c}\right\|_1\leq 5\lambda\left\|\hat{\beta}_{S_*}-\beta^*_{S_*}\right\|_1+\frac{4\|X\beta^*-f^0\|_2^2}{n}$$

with $\lambda\geq 4\lambda_0$.

Proof. The proof of this lemma works analogously to the proof of Lemma 6.5. On $\mathcal{J}$ the general basic inequality reads with $\lambda\geq 4\lambda_0$

$$\frac{4\left\|X\hat{\beta}-f^0\right\|_2^2}{n}+4\lambda\left\|\hat{\beta}\right\|_1\leq 4\frac{1}{4}\lambda\left\|\hat{\beta}-\beta^*\right\|_1+4\lambda\|\beta^*\|_1+\frac{4\|X\beta^*-f^0\|_2^2}{n}.$$

By the same argument as in the proof of Lemma 6.5 it is

$$\left\|\hat{\beta}\right\|_1 \geq \left\|\beta^*_{S_*}\right\|_1 - \left\|\hat{\beta}_{S_*} - \beta^*_{S_*}\right\|_1 + \left\|\hat{\beta}_{S_*^c}\right\|_1$$

and on the right hand side one gets

$$\left\|\hat{\beta} - \beta^*\right\|_1 = \left\|\hat{\beta}_{S_*} - \beta^*_{S_*}\right\|_1 + \left\|\hat{\beta}_{S_*^c}\right\|_1.$$

Plugging this two observations into the above inequality and observing that $\left\|\beta^*_{S_*}\right\|_1 = \|\beta^*\|_1$ leads to the inequality that was to be proven. □

In order to get similar results as before we have to get rid of this error term somehow. Therefore let us now distinguish between the following two cases:

$$\begin{aligned} &\textit{I)}\ \lambda\left\|\hat{\beta}_{S_*} - \beta^*_{S_*}\right\|_1 \geq \frac{\|X\beta^* - f^0\|_2^2}{n}, \text{ and} \\ &\textit{II)}\ \lambda\left\|\hat{\beta}_{S_*} - \beta^*_{S_*}\right\|_1 < \frac{\|X\beta^* - f^0\|_2^2}{n}. \end{aligned} \tag{13}$$

So one can derive from the inequality of the previous lemma, that either in case I it holds

$$\frac{4\left\|X\hat{\beta} - f^0\right\|_2^2}{n} + 3\lambda\left\|\hat{\beta}_{S_*^c}\right\|_1 \leq 9\lambda\left\|\hat{\beta}_{S_*} - \beta^*_{S_*}\right\|_1,$$

or in case II it holds

$$\frac{4\left\|X\hat{\beta} - f^0\right\|_2^2}{n} + 3\lambda\left\|\hat{\beta}_{S_*^c}\right\|_1 \leq \frac{9\|X\beta^* - f^0\|_2^2}{n},$$

or both hold.

In the first case we find again $\left\|\hat{\beta}_{S_*^c}\right\|_1 \leq 3\left\|\hat{\beta}_{S_*} - \beta^*_{S_*}\right\|_1$ and we can go on similarly as before with the

Definition 7.3 (Compatibility Condition for general sets)**.** *We say that the* compatibility condition *holds for the set S of the form $\{j : \beta^*_j \neq 0\}$ for general β^*, if for some constant $\Phi(S) > 0$ and for all β with $\|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1$ one has*

$$\|\beta_S\|_1^2 \leq \frac{(\beta^t\hat{\Sigma}\beta)|S|}{\Phi^2(S)}.$$

We will denote a collection of sets S for which the compatibility condition is fulfilled by $\mathcal{S}$.

In the second case we already reached a quite good inequality for a well-chosen β^*:

Definition 7.4. *The* oracle β^* *is defined as*

$$\beta^* = \underset{\beta: S_\beta \in \mathcal{S}}{\arg\min} \{ \frac{\|X\beta - f^0\|_2^2}{n} + \frac{4\lambda^2 s_\beta}{\Phi^2(S_\beta)} \},$$

where $S_\beta := \{j : \beta_j \neq 0\}$ *and* $s_\beta := |S_\beta|$.

Note, that it would also be possible to minimize over all β with the convention that $\Phi(S) = 0$ if S infringes the compatibility conditions. If the "true regression function" $f^0 = f^0_\beta$ is linear we take $\mathcal{S} = \{S_0\}$.

For a given set S we are interested in the best approximation of f^0 using only non-zero coefficients inside the set S:

$$b^S := \underset{\beta = \beta_S}{\arg\min} \left\| X\beta - f^0 \right\|_2 .$$

Theorem 7.5. *Assume that* $\hat{\sigma}_j^2 = 1$ *for all* j *and that the compatibility condition is fulfilled for all* $S \in \mathcal{S}$, *with* $\hat{\Sigma}$ *normalized in this way. For* $t > 0$ *let* $\lambda = 8\hat{\sigma}\sqrt{\frac{t^2 + 2\log p}{n}}$ *where* $\hat{\sigma}^2$ *is some estimator of* σ^2 *and* $\alpha := 2e^{-\frac{t^2}{2}} + \mathrm{P}\left[\hat{\sigma} \leq \sigma\right]$. *Then with probability at least* $1 - \alpha$, *we have*

$$\frac{2\left\| X\hat{\beta} - f^0 \right\|_2^2}{n} + \lambda \left\| \hat{\beta} - \beta^* \right\|_1 \leq \frac{6\left\| X\beta^* - f^0 \right\|_2^2}{n} + \frac{24\lambda^2 s_*}{\Phi_*^2}$$

Proof. For the proof of this theorem, we have to analyse each of the two cases in (13) separately.

Let us now start with **Case** I**:** On $\mathcal{J}$, whenever $\lambda \left\| \hat{\beta}_{S_*} - \beta^*_{S_*} \right\|_1 \geq \frac{\left\| X\beta^* - f^0 \right\|_2^2}{n}$, we have

$$\frac{4\left\| X\hat{\beta} - f^0 \right\|_2^2}{n} + 3\lambda \left\| \hat{\beta}_{S_*^c} \right\|_1 \leq 9\lambda \left\| \hat{\beta}_{S_*} - \beta^*_{S_*} \right\|_1 ,$$

as already observed. So by using

$$\left\| \hat{\beta} - \beta^* \right\|_1 = \left\| \hat{\beta}_{S_*} \right\|_1 = \left\| \hat{\beta}_{S_*} - \beta^*_{S_*} \right\|_1 + \left\| \hat{\beta}_{S_*^c} \right\|_1 , \tag{14}$$

this is equivalent to

$$\begin{aligned}
\frac{4\left\|X\hat{\beta}-f^0\right\|_2^2}{n}+3\lambda\left\|\hat{\beta}-\beta^*\right\|_1 &\leq 12\lambda\left\|\hat{\beta}_{S_*}-\beta^*_{S_*}\right\|_1 \\
&\overset{Def.7.3}{\leq} \frac{12\lambda\sqrt{s_*}\left\|X(\hat{\beta}-\beta^*)\right\|_2}{\sqrt{n}\Phi_*} \\
&\overset{\triangle-inequ.}{\leq} \frac{12\lambda\sqrt{s_*}\left\|X(\hat{\beta}-f^0)\right\|_2}{\sqrt{n}\Phi_*}+\frac{12\lambda\sqrt{s_*}\left\|X(\beta^*-f^0)\right\|_2}{\sqrt{n}\Phi_*} \\
&\overset{(\star)}{\leq} \frac{18\lambda^2 s_*}{\Phi_*^2}+\frac{2\left\|X\hat{\beta}-f^0\right\|_2^2}{n}+\frac{6\lambda^2 s_*}{\Phi_*^2}+\frac{6\left\|X\hat{\beta}-f^0\right\|_2^2}{n} \\
&= \frac{24\lambda^2 s_*}{\Phi_*^2}+\frac{2\left\|X\hat{\beta}-f^0\right\|_2^2}{n}+\frac{6\left\|X\beta^*-f^0\right\|_2^2}{n}
\end{aligned}$$

By subtraction of the term $\frac{2\left\|X(\hat{\beta}-f^0)\right\|_2^2}{n}$ in this inequality we hence get

$$\frac{2\left\|X\hat{\beta}-f^0\right\|_2^2}{n}+3\lambda\left\|\hat{\beta}-\beta^*\right\|_1 \leq \frac{6\left\|X\beta^*-f^0\right\|_2^2}{n}+\frac{24\lambda^2 s_*}{\Phi_*^2}.$$

In $(\star)$ we just used for each term of the sum the inequalities $12uv \leq 18u^2+2v^2$ and $12uv \leq 6u^2+6v^2$ which hold by being equivalent to $0 \leq 18u^2-12uv+2v^2 = 2(9u^2-6uv+v^2) = 2(3u-v)^2$ and $0 \leq 6u^2-12uv+6v^2 = (\sqrt{6}u-\sqrt{6}v)^2$, respectively.

Otherwise, if **Case** II holds, then on $\mathcal{J}$, whenever $\lambda\left\|\hat{\beta}_{S_*}-\beta^*_{S_*}\right\|_1 < \frac{\left\|X\beta^*-f^0\right\|_2^2}{n}$, we have

$$\frac{4\left\|X\hat{\beta}-f^0\right\|_2^2}{n}+3\lambda\left\|\hat{\beta}_{S_*^c}\right\|_1 \leq \frac{9\left\|X\beta^*-f^0\right\|_2^2}{n},$$

as already observed and by the equation in (14), also that

$$\left\|\hat{\beta}_{S_*^c}\right\|_1 = \left\|\hat{\beta}-\beta^*\right\|_1-\left\|\hat{\beta}_{S_*}-\beta^*_{S_*}\right\|_1 > \left\|\hat{\beta}-\beta^*\right\|_1-\frac{1}{\lambda}\frac{\left\|X\beta^*-f^0\right\|_2^2}{n}.$$

Together this yields

$$\frac{4\left\|X\hat{\beta}-f^0\right\|_2^2}{n}+3\lambda\left\|\hat{\beta}-\beta^*\right\|_1 < 12\frac{\left\|X\beta^*-f^0\right\|_2^2}{n}.$$

Bear in mind that in both cases we have proven some slightly stronger results.

□

7.2 Handling Smallish Coefficients

Target of this subsection will be to refine the oracle that we defined in the previous subsection. There the oracle contains the approximation error and some ℓ^0-penalty as well as the compatibility constant. Here we now want the refined oracle to combine ℓ^0 and ℓ^1 penalties.

With this new oracle we will prove Theorem 7.8. This will lead to a result that gives us both consistency as well as the oracle result like the previous subsection.

To begin with, we start to define the trade off that includes both ℓ^0 and ℓ^1 penalties.

Definition 7.6. *For each set S, we define*

$$S^{\text{sub}} := \operatorname*{arg\,min}_{S^\circ \subset S} \left\{ \frac{3\lambda^2 \left|S^\circ\right|}{\Phi^2(S^\circ)} + \lambda \left\| (b^S)_{S \setminus S^\circ} \right\|_1 \right\}.$$

This means that smaller coefficients b_j^S go into the ℓ^1-penalty, and the larger ones in the ℓ_0-penalty. One can show that putting fewer coefficients into the ℓ^0-penalty will increase the value of the compatibility constant $\Phi(S^\circ)$.

Now let us define the refined oracle.

Definition 7.7 (Definition of the Oracle)**.** *Let*

$$S_* := \operatorname*{arg\,min}_{S \in \mathcal{S}} \{ \frac{3 \left\| f_S - f^0 \right\|_2^2}{n} + \frac{12\lambda^2 \left|S^{\text{sub}}\right|}{\Phi^2(S^{\text{sub}})} + 4\lambda \left\| (b^S)_{S \setminus S^{\text{sub}}} \right\|_1 \}.$$

The oracle is defined as $\beta^* := b^{S_*}$ and we use the notation $s_*^{\text{sub}} := \left|S_*^{\text{sub}}\right|$ as well as $\Phi_*^{\text{sub}} := \Phi(S_*^{\text{sub}})$.

Theorem 7.8. *For $\lambda \geq 4\lambda_0$ it holds on the set $\mathcal{J} := \left\{ \max_{1 \leq j \leq p} \frac{2\varepsilon^t X^{(j)}}{n} \leq \lambda_0 \right\}$ that*

$$\frac{2 \left\| X\hat{\beta} - f^0 \right\|_2^2}{n} + \lambda \left\| \hat{\beta} - \beta^* \right\|_1 \leq \frac{6 \left\| X\beta^* - f^0 \right\|_2^2}{n} + \frac{24\lambda^2 s_*^{\text{sub}}}{(\Phi_*^{\text{sub}})^2} + 8\lambda \left\| \beta^*_{S_* \setminus S_*^{\text{sub}}} \right\|_1.$$

Proof. Throughout the proof of this theorem, we assume operating just on $\mathcal{J}$. By the basic inequality it is

$$\frac{\left\| X\hat{\beta} - f^0 \right\|_2^2}{n} + \lambda \left\| \hat{\beta} \right\|_1 \leq \lambda_0 \left\| \hat{\beta} - \beta^* \right\|_1 + \lambda \left\| \beta^* \right\|_1 + \frac{\left\| X\beta^* - f^0 \right\|_2^2}{n}.$$

Therefore it is

$$
\begin{aligned}
&\frac{\left\|X\hat{\beta}-f^0\right\|_2^2}{n}+\lambda\left\|\hat{\beta}_{S_*^c}\right\|_1+\lambda\left\|\hat{\beta}_{S_*^{sub}}\right\|_1+\lambda\left\|\hat{\beta}_{S_*\setminus S_*^{sub}}\right\|_1 \\
&\leq \lambda_0\left\|\hat{\beta}_{S_*^c}\right\|_1+\lambda_0\left\|\hat{\beta}_{S_*^{sub}}-\beta^*_{S_*^{sub}}\right\|_1+\lambda_0\left\|\hat{\beta}_{S_*\setminus S_*^{sub}}\right\|_1 \\
&\quad+\lambda\left\|\beta^*_{S_*^{sub}}\right\|_1+(\lambda+\lambda_0)\left\|\beta^*_{S_*\setminus S_*^{sub}}\right\|_1+\frac{\|X\beta^*-f^0\|_2^2}{n}.
\end{aligned}
\tag{15}
$$

After taking the terms $\lambda_0\left\|\hat{\beta}_{S_*^c}\right\|_1$ and $\lambda_0\left\|\hat{\beta}_{S_*\setminus S_*^{sub}}\right\|_1$ to the left side of the above inequality and taking $\lambda\left\|\hat{\beta}_{S_*^{sub}}\right\|_1$ to the right side, we apply the reverse triangle inequality as follows

$$
\lambda\left\|\beta^*_{S_*^{sub}}\right\|_1-\lambda\left\|\hat{\beta}_{S_*^{sub}}\right\|_1\leq\lambda\left\|\hat{\beta}_{S_*^{sub}}-\beta_{S_*^{sub}}\right\|_1,
$$

the inequality (15) now reads

$$
\begin{aligned}
&\frac{\left\|X\hat{\beta}-f^0\right\|_2^2}{n}+(\lambda-\lambda_0)\left\|\hat{\beta}_{S_*^c}\right\|_1+(\lambda-\lambda_0)\left\|\hat{\beta}_{S_*\setminus S_*^{sub}}\right\|_1 \\
&\leq(\lambda+\lambda_0)\left\|\hat{\beta}_{S_*^{sub}}-\beta^*_{S_*^{sub}}\right\|_1+(\lambda+\lambda_0)\left\|\beta^*_{S_*\setminus S_*^{sub}}\right\|_1+\frac{\|X\beta^*-f^0\|_2^2}{n}.
\end{aligned}
\tag{16}
$$

Using the inequality

$$
\left\|\hat{\beta}_{S_*\setminus S_*^{sub}}-\beta^*_{S_*\setminus S_*^{sub}}\right\|_1\leq\left\|\hat{\beta}_{S_*\setminus S_*^{sub}}\right\|_1+\left\|\beta^*_{S_*\setminus S_*^{sub}}\right\|_1
$$

on the left hand side, we then get

$$
\frac{\left\|X\hat{\beta}-f^0\right\|_2}{n}+(\lambda-\lambda_0)\left\|\hat{\beta}_{(S_*^{sub})^c}-\beta^*_{(S_*^{sub})^c}\right\|_1
$$

$$
\leq(\lambda+\lambda_0)\left\|\hat{\beta}_{S_*^{sub}}-\beta^*_{S_*^{sub}}\right\|_1+2\lambda\left\|\beta^*_{S_*\setminus S_*^{sub}}\right\|_1+\frac{\|X\beta^*-f^0\|_2^2}{n}.
$$

And now we use the assumption that $\lambda\geq 4\lambda_0$. So the inequality reads now

$$
\frac{4\left\|X\hat{\beta}-f^0\right\|_2^2}{n}+3\lambda\left\|\hat{\beta}_{(S_*^{sub})^c}-\beta^*_{(S_*^{sub})^c}\right\|_1
$$

$$
\leq 5\lambda\left\|\hat{\beta}_{S_*^{sub}}-\beta^*_{S_*^{sub}}\right\|_1+8\lambda\left\|\beta^*_{S_*\setminus S_*^{sub}}\right\|_1+\frac{4\left\|X\beta^*-f^0\right\|_2}{n}.
$$

Let us now analyse the following two cases:

It either holds **case I**:

$$\lambda \left\| \hat{\beta}_{S_*^{sub}} - \beta^*_{S_*^{sub}} \right\|_1 \geq 2\lambda \left\| \beta^*_{S_* \setminus S_*^{sub}} \right\|_1 + \frac{\|X\beta^* - f^0\|_2}{n},$$

or **case II**:

$$\lambda \left\| \hat{\beta}_{S_*^{sub}} - \beta^*_{S_*^{sub}} \right\|_1 < 2\lambda \left\| \beta^*_{S_* \setminus S_*^{sub}} \right\|_1 + \frac{\|X\beta^* - f^0\|_2}{n}.$$

If **case I** is true, then it holds

$$\frac{4 \left\| X\hat{\beta} - f^0 \right\|_2^2}{n} + 3\lambda \left\| \hat{\beta}_{(S_*^{sub})^c} - \beta^*_{(S_*^{sub})^c} \right\|_1 \leq 9\lambda \left\| \hat{\beta}_{S_*^{sub}} - \beta^*_{S_*^{sub}} \right\|_1.$$

Otherwise, if **case II** is true, then it holds

$$\frac{4 \left\| X\hat{\beta} - f^0 \right\|_2^2}{n} + 3\lambda \left\| \hat{\beta}_{(S_*^{sub})^c} - \beta^*_{(S_*^{sub})^c} \right\|_1 \leq 10\lambda \left\| \beta^*_{S_* \setminus S_*^{sub}} \right\|_1 + \frac{9 \|X\beta^* - f^0\|_2^2}{n},$$

or both. In **case I** we can use the same argument as in Proof of Theorem 7.5. In **case 2**, we have

$$\frac{4 \left\| X\hat{\beta} - f^0 \right\|_2^2}{n} + 3\lambda \left\| \hat{\beta} - \beta^* \right\|_1 \leq 16\lambda \left\| \beta^*_{S_* \setminus S_*^{sub}} \right\|_1 + \frac{12 \|X\beta^* - f^0\|_2^2}{n}.$$

□

So as already stated, by replacing S^{sub} for some S by the suboptimal choice $S^\circ := \emptyset$, this theorem gives

$$\frac{2 \left\| X\hat{\beta} - f^0 \right\|_2^2}{n} + \lambda \left\| \hat{\beta} - b^S \right\|_1 \leq \frac{6 \left\| Xb^S - f^0 \right\|_2^2}{n} + 8\lambda \left\| b^S \right\|_1 \text{ for all } S,$$

which implies consistency.

On the other hand, replacing S^{sub} for some S by the suboptimal choice $S^\circ := S$ leads to

$$\frac{2 \left\| X\hat{\beta} - f^0 \right\|_2^2}{n} + \lambda \left\| \hat{\beta} - b^S \right\|_1 \leq \frac{6 \left\| Xb^S - f^0 \right\|_2}{n} + \frac{24\lambda^2 |S|}{\Phi^2(S)} \quad \text{for all } S.$$

In other words, Theorem 7.8 cobines a consistency result with an oracle result.

8 References

[1] Statistik Austria. Ehescheidungen, Scheidungsrate und Gesamtscheidungsrate seit 1946. `http://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/bevoelkerung/ehescheidungen/index.html`, 2016.

[2] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data*. Springer, 2011.

[3] Edwin K. P. Chong and Stanislaw H. Żak. *An Introduction to Optimization*. John Wiley & Sons, Inc, 2001.

[4] Jay L. Devore and Kenneth N. Berk. *Modern mathematical statistics with applications*. Thomson, 2007.

[5] William H. Greene. *Econometric Analysis*. Pearson, 2012.

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[7] Leonhard Held. *Methoden der statistischen Inferenz – Likelihood und Bayes*. Spektrum, 2008.

[8] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

[9] Ulrich Krengel. *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. vieweg, 2005.

[10] Hansruedi Künsch. Daten und modelle – die zwei Säulen der Statistik. `ftp://ftp.stat.math.ethz.ch/U/hkuensch/hsgym-feb12.pdf`, 2012.

[11] Douglas C. Montgomery, Elizabeth A. Peck, and G.Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley, 2001.

[12] Sheldon Ross. *A first course in probability*. Pearson, 2010.

[13] George G. Roussas. *A Course in Mathematical Statistics*. Academic Press, 1997.

[14] Ludger Rüschendorf. *Mathematische Statistik*. Springer Spektrum, 2014.

[15] Mark J. Schervish. *Theory of Statistics*. Springer, 1995.

[16] George A.F. Seber and Alan J. Lee. *Linear Regression Analysis*. Wiley, 2003.

[17] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley – Cambridge Press, 2009.

[18] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267 – 288, 1996.

[19] Stefan Ulbrich. *Nichtlineare Optimierung*. Birkhäuser, 2012.

[20] Dennis D. Wackerly, William Mendenhall III, and Richard L. Scheaffer. *Mathematical Statistics with Applications*. Thomson, 2008.

[21] John L. Weatherwax and David Epstein. A Solution Manual and Notes for: The Elements of Statistical Learning. `http://waxworksmath.com/Authors/G_M/Hastie/WriteUp/weatherwax_epstein_hastie_solutions_manual.pdf`, 2013.

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

Abstract

This master's thesis gives an overview of statistical methods related to linear regression.

Its main objective is to analyse the LASSO. This is a method that uses an ℓ^1-penalty for the case where one has more parameters that are to be estimated than samples.

LASSO is an acronym that stands for "least absolute shrinkage and selection operator". This means that it both shrinks the coefficient estimates towards zero as well as by the nature of the ℓ^1-penalty some coefficients are shrunken exactly to zero. That is it performs variable selection.

In this text the LASSO is compared to ridge regression, which is another, slightly older, shrinkage method that relates to an ℓ^2-penalisation. Moreover several of the properties of the LASSO such as consistency are proven.

The LASSO was firstly introduced by Robert Tibshirani in 1996.

Zusammenfassung

Diese Masterarbeit gibt einen Überblick über statistische Methoden in Bezug auf lineare Regression.

Das Hauptziel ist es, den LASSO zu untersuchen. Dieser ist eine Methode, die ein ℓ^1-Strafverfahren verwendet für den Fall, in dem mehr zu schätzende Parameter zur Verfügung stehen als Stichproben.

LASSO ist eine Abkürzung die für "least absolute shrinkage and selection operator" steht. Das bedeutet, dass Koeffizienten sowohl gegen Null geschrumpft werden als auch einige exakt zu Null. Also leistet er auch Variablen-Selektion.

In diesem Text wird der LASSO mit Ridge Regression verglichen, was eine weitere, ein wenig ältere Schrumpfmethode ist, bei der es sich um ein ℓ^2-Strafverfahren handelt. Außerdem werden einige Eigenschaften des LASSO bewiesen wie zum Beispiel Konsistenz.

Der LASSO wurde erstmals von Robert Tibshirani im Jahr 1996 veröffentlicht.