A Course in High-dimensional Statistics

Johannes Lederer
Professor of Mathematical Statistics
Ruhr-University Bochum

www.johanneslederer.com
www.github.com/LedererLab

November 26, 2018

A Course in High-dimensional Statistics, Edition 1. *Johannes Lederer.*

Published in 2019 by John Wiley & Sons Ltd.

Preface

These are my lecture notes for the course on high-dimensional statistics at Ruhr-University Bochum in Winter 2018. Some parts of the notes are based on my script for a course at UC Louvain in Spring 2018.

Please feel encouraged to give feedback and to point out mistakes.

A format for citing the script is:

> Lederer, J. 2018. A Course in High-dimensional Statistics. *Lecture Notes at Ruhr-University Bochum, Winter 2018.*

Johannes Lederer
Bochum, 2018

Pointwise Versus Probabilistic Properties

The book largely disentangles the algebraic properties of high-dimensional methods from the probabilistic idiosyncrasies of specific data generating mechanisms. In particular, many parts of this book do without concrete models. We believe that this approach avoids confusion that can be caused by artifacts from specific probability distributions, and therefore, we belive that the approach can lead to a deep understanding of the working principles of high-dimensional statistics.

Prerequisites

The following prerequisites are suggested.

Calculus and Linear Algebra Mastering of vector spaces, norms, inner products, matrix calculus, differentiation, and integration. This could include one year calculus on the level of [Spi06] and one semester linear algebra on the level of [Jän94]. Free online resources for analysis and linear algebra are [Daw] and [Hef], respectively. Test questions: *What is a vector space? What is a norm? What is an eigenvalue? How to integrate by parts?*

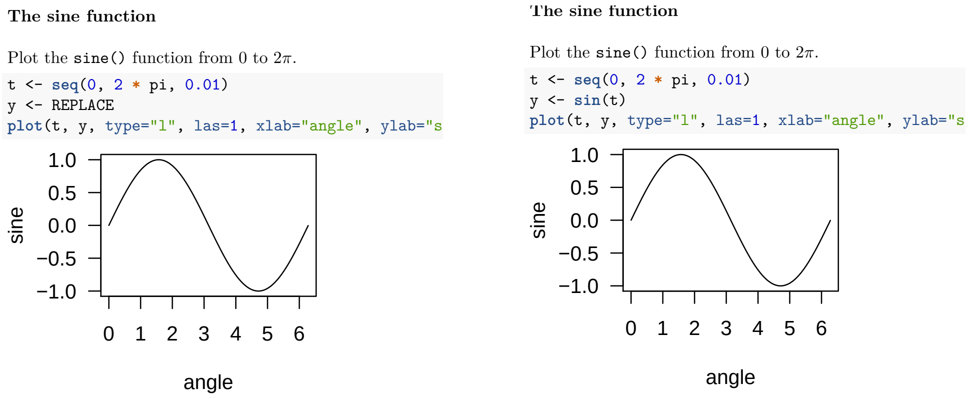
Probability Theory Mastering of probabilities and expectations. Test questions: *How are probabilities and expectations related?*

Statistics Mastering of basic models and estimators at the level of [JWHT13]. Better one additional semester course on mathematical statistics at the level of [LC98]. Test questions: *What is the form of a multivariate normal distribution? What is the linear regression model in vector formulation? What is the corresponding least-squares estimator?*

Labs and Exercises

In addition to the main text, the notes contain exercises and labs. The exercises are equipped with a diamond rating between $^{\diamond}$ and $^{\diamond\diamond\diamond}$: the more diamonds, the harder or longer the solution of an exercise. For some exercises, marked by a bullet $^{\bullet}$, solutions are in the back of the script. However, we urge the reader to attempt the exercises seriously without looking up the solutions first.

The labs are written in `R`. We propose the use of the `Rstudio` IDE, which is available freely on the web. Make sure to have downloaded the packages that are included with the `library()` command; this can be done conveniently within `Rstudio` via the `Packages` panel. To access to the manuals of the various functions, you can use the `Help` panel.



The sine function

Plot the `sine()` function from 0 to 2π.

```
t <- seq(0, 2 * pi, 0.01)
y <- REPLACE
plot(t, y, type="l", las=1, xlab="angle", ylab="s
```

The sine function

Plot the `sine()` function from 0 to 2π.

```
t <- seq(0, 2 * pi, 0.01)
y <- sin(t)
plot(t, y, type="l", las=1, xlab="angle", ylab="s
```

Figure 1: Example `R` lab (left panel) and corresponding solution (right panel). The reader is supposed to replace the keyword “REPLACE” by the correct code.

Notation

We introduce here some notation that we will use throughout the book.

Basic quantities: Lower-case letters a denote numbers and real-valued functions, boldface lower-case letters $\boldsymbol{a}$ vectors, capital letters A matrices, capital calligraphic letters $\mathcal{A}$ sets, greek letters λ real-valued parameters, boldface greek letters $\boldsymbol{\lambda}$ vector-valued parameters, capital greek letters Λ matrix-valued parameters, and additional hats $\widehat{\lambda}, \widehat{\boldsymbol{\lambda}}, \widehat{\Lambda}$ parameter estimates.

Basic functions: The logarithm is taken with respect to the basis e, that is, $\log e = 1$. The smallest integer larger or equal to a given $a \in \mathbb{R}$ is denoted by $\lceil a \rceil$. The support of a vector $\boldsymbol{a} \in \mathbb{R}^p$ is denoted by $\operatorname{supp}[\boldsymbol{a}] := \{j \in \{1, \dots, p\} : a_j \neq 0\}$. The number of elements in a set $\mathcal{A}$ is denoted by $|\mathcal{A}| \in \{0, 1, \dots, \infty\}$. The rank of a matrix $A \in \mathbb{R}^{n \times p}$ is the maximal number of linearly independent columns in A, that is, the dimension of the space spanned by the columns of A; it is denoted by $\operatorname{rank}[A]$.

ℓ_q-functions: The ℓ_q-functions on $\mathbb{R}^p$, where $q \in [0, \infty]$ and $p \in \{1, 2, \dots\}$, are defined for $q \in (0, \infty)$ as

$$\begin{aligned} \ell_q \;:\; \mathbb{R}^p \;&\to\; [0, \infty) \\ \boldsymbol{a} \;&\mapsto\; \|\boldsymbol{a}\|_q := \Biggl(\sum_{j=1}^{p} |a_j|^q\Biggr)^{1/q}, \end{aligned}$$

for $q=0$ as

$$\begin{aligned}\ell_0 \;:\; \mathbb{R}^p &\to \{0,1,\dots\}\\ \boldsymbol{a} &\mapsto \|\boldsymbol{a}\|_0 := \bigl|\bigl\{ j\in\{1,\dots,p\} \;:\; a_j\neq 0\bigr\}\bigr|\,,\end{aligned}$$

and for $q=\infty$ as

$$\begin{aligned}\ell_\infty \;:\; \mathbb{R}^p &\to [0,\infty)\\ \boldsymbol{a} &\mapsto \|\boldsymbol{a}\|_\infty := \max_{j\in\{1,\dots,p\}} |a_j|\,.\end{aligned}$$

The ℓ_q-functions are norms if and only if $q\geq 1$; accordingly, we often refer to those functions as ℓ_q-norms.

The extended real line: The real line extended by $\{-\infty,+\infty\}$ is denoted by $[-\infty,\infty] := \mathbb{R}\cup\{-\infty,+\infty\}$. Similarly, $[0,\infty] := [0,\infty)\cup\{\infty\}$. We use the conventions $0\cdot(\pm\infty) := 0$ and $a/(\pm\infty) := 0$ for $a\in\mathbb{R}$, which are continuous extentions of the rules on $\mathbb{R}$, and the convention $0/0 := 0$, which renders our expressions most concise (note that $0/0$ cannot be obtained by extending the rules on $\mathbb{R}$ continuously: if it were, then $0/0 = \lim_{a\to 0}(a/a) = 1$ and at the same time $0/0 = (2\cdot 0)/0 = 2\cdot(0/0) = 2$, which is a contradiction).

Duals: The *dual* of a function $h \;:\; \mathbb{R}^p \to [-\infty,\infty]$ with respect to the standard inner product on $\mathbb{R}^p$ is defined as

$$\overline{h}[\boldsymbol{a}] \;:=\; \sup\bigl\{\,\langle\boldsymbol{a},\,\boldsymbol{k}\rangle \;:\; \boldsymbol{k}\in\mathbb{R}^p,\, h[\boldsymbol{k}]\leq 1\,\bigr\}\,, \qquad (\boldsymbol{a}\in\mathbb{R}^p)\,.$$

As a convention, we set the supremum over the empty set equal to $-\infty$. Duals play a crucial role in Hölder's inequality B.1.3, which is a key ingredient of many proofs in this book. The third part of Exercise 6.1 shows that the dual functions of ℓ_q-norms are $\overline{\|\cdot\|_q} = \|\cdot\|_p$, where $1/q+1/p=1$. Exercise 6.3 provides further properties of dual functions.

Index sets: The complement of a set $\mathcal{A}$ is denoted by $\mathcal{A}^\complement$. Consider two positive integers l,m and two sets of indexes $\mathcal{A}\subset\{1,\dots,l\}$, $\mathcal{B}\subset\{1,\dots,m\}$ with sizes $a:=|\mathcal{A}|$ and $b:=|\mathcal{B}|$, respectively. For any vector $\boldsymbol{a}\in\mathbb{R}^l$, we denote $\boldsymbol{a}_\mathcal{A}\in\mathbb{R}^a$ as the subvector of $\boldsymbol{a}$ with indexes in $\mathcal{A}$. Similarly, we denote $\boldsymbol{a}_{\mathcal{A}^\complement}\in\mathbb{R}^{l-a}$ as the subvector of $\boldsymbol{a}$ with indexes in $\mathcal{A}^\complement$. The special case $\mathcal{A}=\varnothing$ is taken into account by setting $\boldsymbol{a}_\varnothing := 0$. We typically assume implicitly that the ordering is such that $\boldsymbol{a} = (\boldsymbol{a}_\mathcal{A}^\top, \boldsymbol{a}_{\mathcal{A}^\complement}^\top)^\top$ without loss of generality. For matrices $A\in\mathbb{R}^{l\times m}$, the matrix $A_\mathcal{A}\in\mathbb{R}^{l\times a}$ (and analogously $A_{\mathcal{A}^\complement}\in\mathbb{R}^{l\times(m-a)}$) consists of the columns of A with index in $\mathcal{A}$ ($\mathcal{A}^\complement$). Again, without loss of generality, we typically assume $A=(A_\mathcal{A}, A_{\mathcal{A}^\complement})$. Finally, we use the convention $A_\mathcal{A}^\top := (A_\mathcal{A})^\top$. Similarly, the matrix $A_{\mathcal{A}\mathcal{B}}$ is the submatrix of A with rows in $\mathcal{A}$ and columns in $\mathcal{B}$. Finally, we implicitly understand $A_{\mathcal{A}\mathcal{B}}^\top$ as $(A^\top)_{\mathcal{A}\mathcal{B}}$.

Miscellaneous: The expression $\boldsymbol{x}\sim\mathcal{N}_p[\boldsymbol{a},A]$, $\boldsymbol{a}\in\mathbb{R}^p$, $A\in\mathbb{R}^{p\times p}$, states that $\boldsymbol{x}$ is a random vector that follows a normal distribution in p dimensions with mean $\boldsymbol{a}$ and covariance matrix A. Given a positive integer $p\in\{1,2,\dots\}$, we define $\mathbf{0}_p := (0,\dots,0)^\top\in\mathbb{R}^p$.

Acknowledgements

I sincerely thank Yannick Düren, Shih-Ting Huang, Mike Laszkiewicz, Dr. Marco Rossini, Dr. Fang Xie, and the participants of my 2018 short course at UC Louvain, my 2018 seminar course at RUB, and my 2018/2019 lecture course at RUB for their valuable input.

Contents

Chapter 1

Introduction

1.1 Embracing High-Dimensionality

The technological revolution in genomics of the late 90's enabled researchers to decipher whole genomes for the very first time. It started with simple bacteria and went up all the way to the most complex eukaryotes such as mice, humans, and *Polychaos dubium*, a freshwater amoeboid whose genome consists of a record 670 billion base pairs.[1] The sections of the genome that quickly come under the spotlight are the *genes*, which are functional units of up to several million base pairs. Biologists want to understand how these genes interact among each other and how they affect the characteristics of organisms.

The human genome contains about $20'000$ genes, and it is believed that they, either individually or in small groups, determine traits such as the predisposition to certain diseases. A statistical model for this is linear regression. Say, $y \in \mathbb{R}$ is a subject's C4 blood level, a biomarker for hereditary angioedema (a disease that causes severe swellings), and $x_1, \dots, x_p \in \{0, 1, 2, \dots\}$ the copy numbers of p genes under consideration (that is, the number of repeats of the gene sequences), where $p \approx 20'000$ if the entire genome is considered.[2] The linear regression model

$$y \;=\; \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + u \tag{1.1}$$

connects the *outcome* y with the *predictors* $x_1, \dots, x_p$. The *parameters* $\beta_1, \dots, \beta_p \in \mathbb{R}$ quantify the influence of the corresponding genes on the biomarker: the larger $|\beta_i|$, the more important is the role of gene i. The parameter β_0 is an intercept. Measurement errors, non-genomic influences, etc., are meanwhile summarized in the noise $u \in \mathbb{R}$. The goal of a statistical analysis is to estimate the unknown quantities of interest $\beta_1, \dots, \beta_p$ from the observations of y and $x_1, \dots, x_p$ in n subjects.

A classical estimator for regression parameters is least-squares: what keeps us from using it here? The first suspect is the *data*: in genome data, the number of genes under consideration is often much larger than the number of study subjects. To see if this causes problems, we first focus on a single gene, that is, $y = \beta_0 + \beta_1 x_1$ and $p = 1$. Estimating the parameter β_1 by using the toy data set A on the left-hand side in Table 1.1, which consist of $n = 7$ observations of the biomarker y and of the gene's copy number x_1, is straightforward: a standard least-squares (see 1. in Exercise 1.1) yields $\widehat{\beta}_1 = 1.61$, and the corresponding model fits the data reasonably well in terms of the adjusted R^2—see the bottom of Table 1.1. Estimating β_1 by using the data set B on the right-hand side in Table 1.1, which contains the copy

A	
y	x_1
0	0
2	1
2	0
4	3
8	4
0	0
2	1

B								
y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
0	0	2	0	0	1	0	0	1
2	1	0	2	1	2	0	3	0
2	0	0	0	1	2	1	1	1
4	3	0	1	0	0	0	3	2
8	4	4	0	0	1	1	0	0
0	0	2	0	0	3	0	0	0
2	1	0	2	1	0	0	1	0

Model	**Fit measured in R^2**
$\widehat{y} = 0.50 + 1.61x_1$	0.85
$\widehat{y} = 0.17 + 1.43x_1 + 0x_2 + 0x_3 + 0x_4 + 0x_5 + 1.97x_6 + 0x_7 + 0x_8$	0.98

Table 1.1: Two regression-type data sets (top panel) and estimated models for y in terms of the x'es (bottom panel). Although the data set on the right is more comprehensive than the data set on the left, both data sets are equally suited for regressing y on x_1: one can just ignore the values for $x_2, \ldots, x_8$. However, having the values for $x_2, \ldots, x_8$ allows for estimating more refined models that are based not only on x_1. For example, the lasso model that is based on all predictors has a better fit that the simple model (the larger the coefficient of determination R^2, the better the fit; $R^2 = 1$ means that the model predictions and the observations match perfectly); refer to Exercise 1.1 for the calculations.

numbers of 7 additional genes, seems more difficult. However, we can just neglect the measurements of $x_2, \ldots, x_8$ (and similarly data from additional subjects if available) and proceed as before[3]: we can run a least-squares that uses only the realizations of y and x_1, which yields again $\widehat{\beta}_1 = 1.61$. Hence, data per se does not seem to be the problem.

What causes new challenges is instead how such data is *used.* In addition to the above analysis of the first gene, data set B can also be used to fit a model of the form $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_8 x_8 + u$, which incorporates all measured genes simultaneously. However, the number of parameters $p = 8$ now exeeds the number of samples $n = 7$, which renders the classical least-squares estimator unreliable, even ambiguous (see Section 1.2). And this failure is not specific to the least-squares: there are just not enough samples for estimating all the parameters.

There is no way around it: we need enough information to calibrate models that are rich in parameters. However, information can come in many shapes and forms, not only as observations of y's and x'es. For example, biologists know that a hereditary disease is typically associated with only a very small number of genes. A lasso approach (see 2. in Exercise 1.1) complements data set B with this information and yields the model $\widehat{y} = 0.17 + 1.43x_1 + 0x_2 + 0x_3 + 0x_4 + 0x_5 + 1.97x_6 + 0x_7 + 0x_8$, which improves the fit of the one-gene-model substantially—see the bottom of Table 1.1. The model accounts for *all* 8 genes, but it focusses the attention on two of them. This is the general trick: *high-dimensional statistics* leverages information beyond the bare data to find meaningful patterns (here, an accurate model based on only 2 genes) in complex data (here, the copy numbers of 8 genes).[4]

high-dimensional statistics

High-dimensional statistics comes into play whenever one fits complex models (p large) and is even indispensable when additionally sample sizes are comparably

	Classical statistics	**High-dimensional statistics**
$n \not\gg 1$	—	—
$n \gg 1$, $p \not\gg 1$	✓	—
$n, p \gg 1$, $p \ll n$	✓	✓
$n, p \gg 1$, $p \approx n$ or $p \gg n$	—	✓

Table 1.2: Typical scopes of classical statistics (such as least-squares estimation) and high-dimensional statistics (such as lasso estimation). The larger the number of parameters p as compared to the number of samples n, the more essential is high-dimensional statistics.

small ($p \approx n$ or even $p \gg n$)—see Table 1.2. Complex models arise naturally in "Big Data," where the total number of measurements is large. Applications include predicting shopping habits based on extensive customer profiles, finding weaknesses in opposing soccer teams by tracking every action of their players, and optimizing the fertilization of crops by using fine-grained geospacial data. Nevertheless, high-dimensionality is not limited to large data sets: in our toy example, the total number of measurements across y and $x_1, \dots, x_8$ is only $n \times (1+p) = 7 \times (1+8) = 63$. Also, high-dimensionality can be an issue in basically any model class. Our workhorse in this book is linear regression, but the insights can be transfered readily to other types of models such as logistic regression, tensor regression, and graphical modeling (see Chapter 3 for the latter). Altogether, high-dimensional statistics is useful in a very wide variety of applications.

1.2 Statistical Limitations of Classical Estimators

Parameter spaces that are high-dimensional relative to the data cannot be explored by using classical estimators. We illustrate this fact by showing that the least-squares, one of the most classical estimators, yields poor results if the number of predictors is comparable to the number of samples ($p \approx n$) or even larger ($p \gg n$).

Consider n data points $(y_1, (x_1)_1, \dots, (x_p)_1), \dots, (y_n, (x_1)_n, \dots, (x_p)_n)$ from a linear regression model of the form (1.1). We summarize the outcomes in the vector $\boldsymbol{y} := (y_1, \dots, y_n)^\top$, the predictors in the $n \times p$-*design matrix* defined through $X_{ij} := (x_j)_i$, the parameters (which remain the same across all data points) in the *regression vector* $\boldsymbol{\beta} := (\beta_1, \dots, \beta_p)^\top$ (we drop the intercept without loss of generality[5]), and the noise in the vector $\boldsymbol{u} := (u_1, \dots, u_n)^\top$. These definitions allow us to coalesce the relationships specified by (1.1) into a single vector-valued equation:

$$\boldsymbol{y} \;=\; X\boldsymbol{\beta} + \boldsymbol{u}\,. \tag{1.2}$$

The goal in linear regression is to estimate the unknown regression vector $\boldsymbol{\beta}$ from data $(\boldsymbol{y}, X)$. The classical approach to this is the *least-squares estimator*

least-squares estimator

$$\widehat{\boldsymbol{\beta}}_{\mathrm{ls}} \;\in\; \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2\,. \tag{1.3}$$

By design, the least-squares estimator provides an optimal fit to the data: no other linear combination of predictors model $\boldsymbol{y}$ more acurately than $X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$. However, a good fit to the data is not the same as a good estimation of $\boldsymbol{\beta}$. In the following, we study the least-square's performance at this latter task in terms of the *prediction error*

prediction error

$\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2$. The prediction error measures how well the estimator disentangles the data generating part $X\boldsymbol{\beta}$ from the noise $\boldsymbol{u}$.

Since the least-squares objective function $\boldsymbol{\alpha}\mapsto\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2$ is convex and differentiable, we can characterize $\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$ through derivatives of the objective function:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{ls}} \in \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2 \quad\Leftrightarrow\quad \frac{\partial}{\partial\alpha_k}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2\Big|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}} = 0 \quad \forall\, k\in\{1,\dots,p\}\,.$$

The derivatives on the right-hand side can be computed explicitly:

$$\begin{aligned}
&\frac{\partial}{\partial\alpha_k}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2\\
&= \frac{\partial}{\partial\alpha_k}\sum_{i=1}^n\Bigl(y_i-\Bigl(\sum_{j=1}^p X_{ij}\alpha_j\Bigr)\Bigr)^2 && \text{“definition of the ℓ_2-norm”}\\
&= \sum_{i=1}^n 2\Bigl(y_i-\Bigl(\sum_{j=1}^p X_{ij}\alpha_j\Bigr)\Bigr)\cdot(-X_{ik}) && \text{“sum and chain rules”}\\
&= -2\sum_{i=1}^n\Bigl(X_{ki}^\top y_i-\Bigl(\sum_{j=1}^p X_{ki}^\top X_{ij}\alpha_j\Bigr)\Bigr) && \text{“rearranging factors; $X_{ik}=X_{ki}^\top$”}\\
&= -2\bigl(X^\top\boldsymbol{y}-X^\top X\boldsymbol{\alpha}\bigr)_k\,. && \text{“evaluating the sums”}
\end{aligned}$$

Hence, setting $\boldsymbol{\alpha}=\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$ and using the right-hand side of the least-squares' characterization in the penultimate display, we find that a vector $\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$ is a solution of the least-squares estimator if and only if

$$-2\bigl(X^\top\boldsymbol{y}-X^\top X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\bigr)_k = 0 \quad \forall\, k\in\{1,\dots,p\}\,,$$

which is equivalent to the vector equality

$$X^\top X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}} = X^\top\boldsymbol{y}\,.$$

For simplicity, we assume that $X^\top X$ is invertible (see Exercise 1.2 for a generalization). We can then multiply both sides of the display by $(X^\top X)^{-1}$, which identifies $\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}=(X^\top X)^{-1}X^\top\boldsymbol{y}$ as the unique least-squares estimator. The prediction error is then

$$\begin{aligned}
&\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2\\
&= \|X\boldsymbol{\beta}-X(X^\top X)^{-1}X^\top\boldsymbol{y}\|_2^2 && \text{“mentioned form of $\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$”}\\
&= \|X\boldsymbol{\beta}-X(X^\top X)^{-1}X^\top(X\boldsymbol{\beta}+\boldsymbol{u})\|_2^2\\
& && \text{“linear regression model introduced in the beginning of the section”}\\
&= \|X(X^\top X)^{-1}X^\top\boldsymbol{u}\|_2^2\,. && \text{“simplifying the terms”}
\end{aligned}$$

We now replace the design matrix X by a singular value decomposition of it (see Appendix starting Page 171 for detailed definitions of the terms). For this, consider orthogonal (and therefore, invertible) square matrices $U\in\mathbb{R}^{n\times n}$, $V\in\mathbb{R}^{p\times p}$ and a diagonal matrix $D\in\mathbb{R}^{n\times p}$ (that is, $D_{ij}=0$ if $i\neq j$) such that

$$X = UDV^\top.$$

With these definitions, the previous display gives

$$\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2$$

$$\begin{aligned}
&= \|UDV^\top\big((UDV^\top)^\top UDV^\top\big)^{-1}(UDV^\top)^\top \boldsymbol{u}\|_2^2 \quad \text{“combining the two previous displays”}\\
&= \|UDV^\top\big(VD^\top U^\top UDV^\top\big)^{-1}VD^\top U^\top \boldsymbol{u}\|_2^2\\
&\qquad\qquad \text{“using that } (AB)^\top = B^\top A^\top \text{ for any } A, B\text{”}\\
&= \|UDV^\top\big(V^\top\big)^{-1}\big(D^\top U^\top UD\big)^{-1}V^{-1}VD^\top U^\top \boldsymbol{u}\|_2^2\\
&\qquad\qquad \text{“using that } (AB)^{-1} = B^{-1}A^{-1} \text{ for any invertible matrices } A, B\text{”}\\
&= \|UD(D^\top D)^{-1}D^\top U^\top \boldsymbol{u}\|_2^2\\
&\qquad\qquad \text{“}U^\top U = \mathrm{I}_{n\times n} \text{ since } U \text{ orthogonal by assumption; } V^\top(V^\top)^{-1} = V^{-1}V = \mathrm{I}_{p\times p} \text{ invertible”}\\
&= \big(UD(D^\top D)^{-1}D^\top U^\top \boldsymbol{u}\big)^\top UD(D^\top D)^{-1}D^\top U^\top \boldsymbol{u} \quad \text{“definition of the } \ell_2\text{-norm”}\\
&= \big(D(D^\top D)^{-1}D^\top U^\top \boldsymbol{u}\big)^\top U^\top UD(D^\top D)^{-1}D^\top U^\top \boldsymbol{u} \quad \text{“}(AB)^\top = B^\top A^\top \text{ for any } A, B\text{”}\\
&= \|D(D^\top D)^{-1}D^\top U^\top \boldsymbol{u}\|_2^2\,. \quad \text{“}U^\top U = \mathrm{I}_{n\times n}\text{; again definition of the } \ell_2\text{-norm”}
\end{aligned}$$

Next, since D is diagonal, we get for any $k, l \in \{1, \dots, p\}$

$$(D^\top D)_{kl} \;=\; \sum_{m=1}^{n} D_{mk}D_{ml} \;=\; \begin{cases} D_{kk}^2 & \text{if } k = l\\ 0 & \text{otherwise}\end{cases}\,.$$

Since $X^\top X$ is invertible by assumption, it holds that $D_{ij} \neq 0$, and therefore, $D^\top D$ is invertible. Together with the result above, this implies that $(D^\top D)^{-1}$ is also diagonal with diagonal elements $1/D_{11}^2, \dots, 1/D_{pp}^2$. Using this, and also $p \leq n$ since $X^\top X$ is invertible by assumption (see 1. of Exercise 1.2), we find for any $i, j \in \{1, \dots, n\}$

$$\begin{aligned}
&(D(D^\top D)^{-1}D^\top)_{ij}\\
&= \sum_{k,l=1}^{p} D_{ik}(D^\top D)^{-1}_{kl}D_{lj} && \text{“matrix algebra”}\\
&= \sum_{k=1}^{p} D_{ik}(1/D_{kk}^2)D_{kj} && \text{“previous observation”}\\
&= \begin{cases} D_{ii}(1/D_{ii}^2)D_{ii} = 1 & \text{if } i = j \text{ and } i \leq p\\ 0 & \text{otherwise}\end{cases}\,, && \text{“}D \text{ diagonal; } p \leq n\text{”}
\end{aligned}$$

so that

$$D(D^\top D)^{-1}D^\top \;=\; \begin{pmatrix} \mathrm{I}_{p\times p} & \\ & \mathbf{0}_{(n-p)\times(n-p)}\end{pmatrix}\,.$$

We can plug this back into the earlier display to find

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2 \;=\; \|(U^\top \boldsymbol{u})_{\{1,\dots,p\}}\|_2^2\,.$$

The term $\|(U^\top \boldsymbol{u})_{\{1,\dots,p\}}\|_2$ is the *effective noise* for the least-squares estimator: the larger this term, the larger the prediction error of the least-squares.

effective noise for the least-squares estimator

As a concrete example, consider X fixed and $\boldsymbol{u} \sim \mathcal{N}_n[\mathbf{0}_n, \sigma^2\,\mathrm{I}_{n\times n}]$ for a $\sigma \in (0, \infty)$. Then, $(U^\top \boldsymbol{u})_{\{1,\dots,p\}} \sim \mathcal{N}_p[\mathbf{0}_p, \sigma^2\,\mathrm{I}_{p\times p}]$, and the effective noise has a Chi-squared distribution with p degrees of freedom and a scaling of σ^2; in particular, taking expectations over the noise, we get the average prediction risk

$$\mathbb{E}\left[\frac{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2}{n}\right] \;=\; \frac{\sigma^2 p}{n}\,. \tag{1.4}$$

This result demonstrates that least-squares provides accurate prediction only in traditional settings where $p/n \ll 1$, that is, $p \ll n$—or if the variance of the noise σ^2 is very small. (This observation is related to the the overfitting phenomenon, which we discuss in Section 2.1.) Thus, for the high-dimensional settings that are the focus of this book, we need to introduce different approaches.

1.3 Incorporating Prior Knowledge

We have illustrated in the previous section that classical estimators are uninformative if the dimension of the parameter space p is large as compared to the number of observations n. High-dimensional statistics tackles such limitations by complementing classical estimators with terms that formulate how "likely" or "favorable" models are. For example, it is known that hereditary diseases are typically caused by only a small number of genes. This means that even though the total number of genes p is often large compared to the number of study subjects n, the number of actually relevant genes $s := |\{j : \beta_j \neq 0\}|$ can be assumed small: $s \ll n, p$. High-dimensional estimators leverage this information through the inclusion of a corresponding prior term such as $\|\boldsymbol{\alpha}\|_0 := |\{j : \alpha_j \neq 0\}|$ or $\|\boldsymbol{\alpha}\|_1 := \sum_j |\alpha_j|$.

Mathematically speaking, we want to use data $Z \in \mathcal{Z}$ to estimate unknown target parameters $\boldsymbol{\beta} \in \mathcal{B}$. In linear regression, the data is $Z = (\boldsymbol{y}, X)$ and the target parameter $\boldsymbol{\beta}$ the regression vector. More generally, the targets $\boldsymbol{\beta}$ denote "true" models if there is such a notion or just models that capture the data generating process accurately. Classical estimators such as maximum likelihood or least-squares can be written as minimizers over data-fitting functions $(\mathcal{Z}, \mathcal{B}) \to [0, \infty]$ that measure how well a parameter agrees with the observed data. *Regularized estimators* in high-dimensional statistics generalize those classical methods through the inclusion of prior functions $\mathcal{B} \to [0, \infty]$ that capture sparsity, smoothness, or other structural information or assumptions:

regularized estimator

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathcal{B}} \bigl\{ \mathrm{DataFitting}[Z, \boldsymbol{\alpha}] + r\, \mathrm{Prior}[\boldsymbol{\alpha}] \bigr\} . \tag{1.5}$$

The tuning parameters $r \in [0, \infty]$ weight the prior information: setting $r = 0$ recovers the classical estimators, in which no prior information is included, while increasing r pushes the estimates in the direction specified by the prior function.

Traditional names for the prior term are *penalty* and *regularizer;* the first name speaks to the idea of "penalizing" unrealistic or unfavorable models, the second name originates in non-parametric statistics, where one targets functions that are sufficiently "regular" (which typically means smooth). By introducing the word "prior," we want to circumvent the negative connotations that "penalty" and "regularizer" might have (do we impose a penalty on unruly parameters?) and establish a connection to Bayesian statistics: similarly as the prior distribution in the Bayesian literature, our prior term does not incorporate any data, and it formulates our knowledge or believes about the parameter space. Still, we will use the different names interchangeably in the following.

The most important feature of regularized estimators is that they can provide accurate inference even if the parameter space is high-dimensional relative to the number of observations. For example, we will show in Chapter 6 that the lasso[6], one of the most prominent regularized estimators, satisfies the risk bound[7]

$$\mathbb{E}\left[\frac{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}_{\text{lasso}}\|_2^2}{n}\right] \leq \frac{a\sigma^2 \log p}{n}$$

y	x_1	x_2
1	1	d
2	2	2

Table 1.3: Regression-type data indexed by $d \in \mathbb{R}$. The least-squares estimator on these data is not unique when $d = 1$ and not continuous in d at $d = 1$, while the ridge estimator is always unique and continuous in d. This illustrates that regularization can have numerical benefits even for seemingly simple data.

for a reasonably small constant $a \in (0, \infty)$ if $\boldsymbol{u} \sim \mathcal{N}_n[\mathbf{0}_n, \sigma^2 \mathrm{I}_{n\times n}]$ and the predictors are not too much correlated. The important difference to the earlier risk bound for the least-squares is that the dependence on p is only logarithmic, which can allow us to handle dramatically more parameters for given data: instead of $n \gg \sigma^2 p$, we now only require $n \gg \sigma^2 \log p$ for accurate prediction. The mathematical reason is that through the inclusion of prior knowledge, the least-squares' effective noise $\|(U^\top \boldsymbol{u})_{\{1,\dots,p\}}\|_2$ can be replaced by the typically smaller quantity $2\|X^\top \boldsymbol{u}\|_\infty$.

1.4 Regularization for Increasing the Numerical Stability*

The idea of prior functions predates high-dimensional statistics considerably. For example, it has been understood already in the middle of the last century that prior functions can increase the numerical stability in inverse problems.[8] We connect this early research with the modern statistical topics of this book by comparing least-squares to the ridge estimator, an estimator that has its roots in the literature on inverse problems but is also used in contemporary high-dimension statistics. We will find that least-squares estimation is like a dog that calmly walks next to its keeper at most times but immediately goes on a wild chase all around the park when a rabbit comes into sight. The additional prior function in the ridge estimator is the leash for keeping the dog under control, and the corresponding tuning parameter balances between the safety of the wildlife and the dogs freedom: the larger the tuning parameter, the tighter the leash.

To illustrate the stabilizing effect of prior functions, we study least-squares and ridge estimation on the data family in Table 1.3. These data specify the outcome $\boldsymbol{y} = (1, 2)^\top \in \mathbb{R}^2$ and the design matrix $X = ((1, 2)^\top, (d, 2)^\top) \in \mathbb{R}^{2\times 2}$, where d takes values in $\mathbb{R}$. We will vary d to evaluate the estimators' robustness against small changes in the data.

We first consider least-squares estimation. For $d \neq 1$, the least-squares estimator has the unique solution $\widehat{\boldsymbol{\beta}}_{\mathrm{ls}} = (1, 0)^\top$. To see this, note first that $\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 \geq 0$ for all $\boldsymbol{\alpha} \in \mathbb{R}^2$. Since $\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 = 0$ for $\boldsymbol{\alpha} = (1, 0)^\top$, the vector $\widehat{\boldsymbol{\beta}}_{\mathrm{ls}} = (1, 0)^\top$ must indeed be a solution of the least-squares estimator. Now, one can also check that $X^\top X$ is invertible if $d \neq 1$. Since $\boldsymbol{\alpha} \mapsto \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2$ is strictly convex for $X^\top X$ invertible (see Exercise 1.4), and since strictly convex functions have unique minima, it follows that $\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$ is the unique least-squares solution.

We can also calculate the least-squares estimator for $d \neq 1$ by using our previously established formula:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{ls}} = \bigl(X^\top X\bigr)^{-1} X^\top \boldsymbol{y} \qquad \text{“Section 1.2”}$$

$$
\begin{aligned}
&= \left(\begin{pmatrix}1 & 2\\ d & 2\end{pmatrix}\begin{pmatrix}1 & d\\ 2 & 2\end{pmatrix}\right)^{-1}\begin{pmatrix}1 & 2\\ d & 2\end{pmatrix}\begin{pmatrix}1\\ 2\end{pmatrix} && \text{“plugging in the data from Table 1.3”}\\
&= \begin{pmatrix}5 & 4+d\\ 4+d & 4+d^2\end{pmatrix}^{-1}\begin{pmatrix}1 & 2\\ d & 2\end{pmatrix}\begin{pmatrix}1\\ 2\end{pmatrix} && \text{“matrix multiplication”}\\
&= \frac{1}{4-8d+4d^2}\begin{pmatrix}4+d^2 & -4-d\\ -4-d & 5\end{pmatrix}\begin{pmatrix}1 & 2\\ d & 2\end{pmatrix}\begin{pmatrix}1\\ 2\end{pmatrix}\\
& && \text{“}\begin{pmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix}^{-1} = \begin{pmatrix}a_{22} & -a_{12}\\ -a_{21} & a_{11}\end{pmatrix}/(a_{11}a_{22}-a_{12}a_{21})\text{”}\\
&= \frac{1}{4-8d+4d^2}\begin{pmatrix}4-4d & -2d+2d^2\\ -4+4d & 2-2d\end{pmatrix}\begin{pmatrix}1\\ 2\end{pmatrix} && \text{“matrix multiplication”}\\
&= \frac{1}{4-8d+4d^2}\begin{pmatrix}4-8d+4d^2\\ 0\end{pmatrix} && \text{“matrix-vector multiplication”}\\
&= \begin{pmatrix}1\\ 0\end{pmatrix}. && \text{“vector-constant multiplication”}
\end{aligned}
$$

One can verify readily that $\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$ is a solution also for $d = 1$, but it is then not the only solution any more. For $d = 1$, it holds that $X(-a, a)^\top = \mathbf{0}_2$ for all $a \in \mathbb{R}$. Therefore, all vectors of the form $\widehat{\boldsymbol{\beta}}^a_{\mathrm{ls}} := \widehat{\boldsymbol{\beta}}_{\mathrm{ls}} + (-a, a)^\top = (1 - a, a)^\top$ satisfy $X\widehat{\boldsymbol{\beta}}^a_{\mathrm{ls}} = X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$, and consequently, $\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}^a_{\mathrm{ls}}\|_2^2 = \|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2$. Hence, in the case $d = 1$, *all* vectors $\widehat{\boldsymbol{\beta}}^a_{\mathrm{ls}}$, $a \in \mathbb{R}$, are least-squares solutions.

This ambiguity of the least-squares estimator leads to a numerical instability. The ℓ_2-difference between a response for $d = 1$ (which can be any vector of the form $\widehat{\boldsymbol{\beta}}^a_{\mathrm{ls}}$) and the response for $d \neq 1$ (which must be $\widehat{\boldsymbol{\beta}}_{\mathrm{ls}} = \widehat{\boldsymbol{\beta}}^0_{\mathrm{ls}}$) is

$$
\|\widehat{\boldsymbol{\beta}}^a_{\mathrm{ls}} - \widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2 \;=\; \|(1-a, a)^\top - (1, 0)^\top\|_2 \;=\; \sqrt{2}|a|\,.
$$

Since a can be arbitrarily large, the response for $d = 1$ can differ arbitrarily much from the response for $d \neq 1$—however close d is to 1. Hence, the least-squares procedure is discontinuous in the data: an ever so small change in the observations can lead to a dramatically different least-squares response.

We now aim at removing this numerical instability. For this, we replace the least-squares estimator by the *ridge estimator*[9] ridge

$$
\widehat{\boldsymbol{\beta}}_{\mathrm{ridge}}[r] \;\in\; \underset{\boldsymbol{\alpha}\in\mathbb{R}^p}{\operatorname{argmin}}\bigl\{\,\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_2^2\,\bigr\}\,, \tag{1.6}
$$

where $r \in (0, \infty)$ is a tuning parameter. We can proceed as in Section 1.2 to find that $\widehat{\boldsymbol{\beta}}_{\mathrm{ridge}}[r]$ is a solution of the ridge estimator if and only if

$$
\bigl(X^\top X + r\,\mathrm{I}_{p\times p}\bigr)\widehat{\boldsymbol{\beta}}_{\mathrm{ridge}}[r] \;=\; X^\top\boldsymbol{y}\,.
$$

The crux is now that the matrix $X^\top X + r\,\mathrm{I}_{p\times p}$ is *always* invertible. Indeed, for any $\boldsymbol{a} \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$,

$$
\begin{aligned}
&\boldsymbol{a}^\top\bigl(X^\top X + r\,\mathrm{I}_{p\times p}\bigr)\boldsymbol{a}\\
&= \boldsymbol{a}^\top X^\top X\boldsymbol{a} + r\boldsymbol{a}^\top\,\mathrm{I}_{p\times p}\,\boldsymbol{a} && \text{“linearity of matrices”}\\
&= (X\boldsymbol{a})^\top X\boldsymbol{a} + r\boldsymbol{a}^\top\boldsymbol{a} && \text{“properties of transposes and of the identity matrix”}\\
&= \|X\boldsymbol{a}\|_2^2 + r\|\boldsymbol{a}\|_2^2 && \text{“definition of } \ell_2\text{-norms”}\\
&> 0\,, && \text{“norms are positive definite; } r > 0, \boldsymbol{a} \neq \mathbf{0}_2\text{”}
\end{aligned}
$$

which implies that $X^\top X + r\,\mathrm{I}_{p\times p}$ is invertible. Hence, we can write

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} \;=\; \bigl(X^\top X + r\,\mathrm{I}_{p\times p}\bigr)^{-1} X^\top \boldsymbol{y}\,,$$

irrespective of whether $X^\top X$ itself is invertible or not. This shows in particular that the ridge estimator is always unique.

We now use the formula to calculate the ridge estimator on the data in Table 1.3:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\text{ridge}}[r] \;&=\; \bigl(X^\top X + r\,\mathrm{I}_{p\times p}\bigr)^{-1} X^\top \boldsymbol{y}\\
&=\; \left(\begin{pmatrix}1 & 2\\ d & 2\end{pmatrix}\begin{pmatrix}1 & d\\ 2 & 2\end{pmatrix} + \begin{pmatrix}r & 0\\ 0 & r\end{pmatrix}\right)^{-1}\begin{pmatrix}1 & 2\\ d & 2\end{pmatrix}\begin{pmatrix}1\\ 2\end{pmatrix}\\
&=\; \begin{pmatrix}5+r & 4+d\\ 4+d & 4+d^2+r\end{pmatrix}^{-1}\begin{pmatrix}1 & 2\\ d & 2\end{pmatrix}\begin{pmatrix}1\\ 2\end{pmatrix}\\
&=\; \frac{1}{4-8d+4d^2+(9+d^2)r+r^2}\begin{pmatrix}4+d^2+r & -4-d\\ -4-d & 5+r\end{pmatrix}\begin{pmatrix}1 & 2\\ d & 2\end{pmatrix}\begin{pmatrix}1\\ 2\end{pmatrix}\\
&=\; \frac{1}{4-8d+4d^2+(9+d^2)r+r^2}\begin{pmatrix}4-4d+r & -2d+2d^2+2r\\ -4+4d+dr & 2-2d+2r\end{pmatrix}\begin{pmatrix}1\\ 2\end{pmatrix}\\
&=\; \frac{1}{4-8d+4d^2+(9+d^2)r+r^2}\begin{pmatrix}4-8d+4d^2+5r\\ (4+d)r\end{pmatrix}\\
&=\; \frac{1}{4(1-d)^2+(9+d^2)r+r^2}\begin{pmatrix}4(1-d)^2+5r\\ (4+d)r\end{pmatrix}.
\end{aligned}$$

The individual steps mirror those for the least-squares estimator for $d \neq 0$. Since the denominator of the factor is strictly positive for any $r > 0$, this result shows in particular that the ridge estimator is continuous in d (see also Exercise 1.5). Thus, the ridge estimator is stable with respect to small changes in the observations.

The continuity is due to the additional regularization term in the ridge objective function. We would thus expect that the stronger the regularization, that is, the larger the tuning parameter, the more stable the estimates are with respect to changes in the data. To support this, we consider two limits. First, we derive from the above result that the limit of $\widehat{\boldsymbol{\beta}}_{\text{ridge}}[r]$ in $r \to \infty$ exists for all $d \in \mathbb{R}$:

$$\lim_{r\to\infty}\widehat{\boldsymbol{\beta}}_{\text{ridge}}[r] \;=\; \lim_{r\to\infty}\frac{1}{r^2}\begin{pmatrix}5r\\ (4+d)r\end{pmatrix} \;=\; \begin{pmatrix}0\\ 0\end{pmatrix}, \qquad (d\in\mathbb{R}).$$

In this sense, the ridge estimator is “perfectly stable” when $r \to \infty$. On the other hand, the above display also entails that for each d, the limit in $r \to 0^+$ exists and yields a least-squares solution:

$$\lim_{r\to 0^+}\widehat{\boldsymbol{\beta}}_{\text{ridge}}[r] \;=\; \begin{cases}\lim\limits_{r\to 0^+}\frac{1}{10r+r^2}\begin{pmatrix}5r\\ 5r\end{pmatrix} \;=\; \lim\limits_{r\to 0^+}\frac{1}{1+r/10}\begin{pmatrix}1/2\\ 1/2\end{pmatrix} \;=\; \begin{pmatrix}1/2\\ 1/2\end{pmatrix}, & (d=1)\\[2ex] \frac{1}{4(1-d)^2}\begin{pmatrix}4(1-d)^2\\ 0\end{pmatrix} \;=\; \begin{pmatrix}1\\ 0\end{pmatrix}, & (d\neq 1)\end{cases},$$

which are the least-squares estimators $\widehat{\boldsymbol{\beta}}_{\text{ls}}^{1/2}$ and $\widehat{\boldsymbol{\beta}}_{\text{ls}} = \widehat{\boldsymbol{\beta}}_{\text{ls}}^{0}$, respectively. That the two limits for $d = 1$ and $d \neq 1$ differ is another display of the instability of the least-squares estimator; a graphical explanation of where the two different solutions come from is given in Figure 1.1. More generally, one can derive that the smaller the regularization,

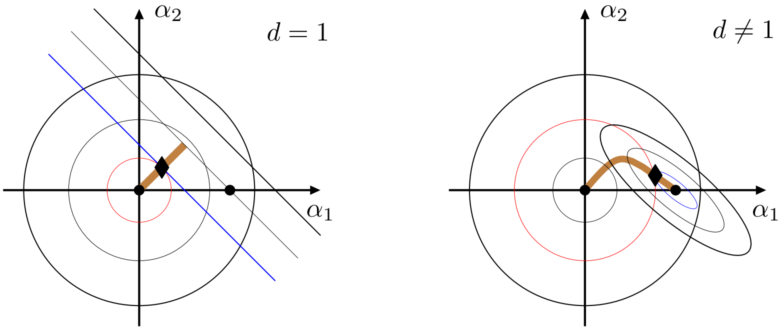


Figure 1.1: An illustration of why the limiting solutions of the ridge estimator on the data of Table 1.3 differ for $d = 1$ and $d \neq 1$. The round black dots denote $(1, 0)^\top$, which is a minimum of $\boldsymbol{\alpha} \mapsto \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2$ for any $d \in \mathbb{R}$, and $(0, 0)^\top$, which is the minimum of $\boldsymbol{\alpha} \mapsto r\|\boldsymbol{\alpha}\|_2^2$ for any $r \in (0, \infty)$. The lines (left plot) and ellipses (right plot) indicate points that have the same values in $\boldsymbol{\alpha} \mapsto \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2$. The three concentric circles indicate points that have the same values in $\boldsymbol{\alpha} \mapsto r\|\boldsymbol{\alpha}\|_2^2$. The thick brown lines indicate coordinates of the ridge estimators; each point on the brown lines corresponds to an estimator with a given value of the tuning parameter r: the larger r, the closer the corresponding ridge estimator is to the origin. Since the ridge estimators are minimizers of objective functions of the form $\boldsymbol{\alpha} \mapsto \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_2^2$, they correspond to points (indicated by black diamonds) where a line/ellipse (in blue) generated by the least-squares objective “touch” a circle (in red) generated by the prior function. See also Exercise 1.7 for details.

the better the approximation of least-squares solutions (see Exercise 1.6 for such a calculation).

In summary, the prior function stablelizes the least-squares estimator and removes potential ambiguity. The corresponding tuning parameter allows one to balance stability and closeness to the unregularized least-squares estimator. The dimensionality of the data ($n = p = 2$) shows that these effects are not tied to high dimensions, but the numerical stability is an extra benefit when using regularized estimators in high-dimensional statistics.

1.5 References and Further Reading

[1] References for the sequencing of the human genome are [KCD$^+$08] and citations therein. An example for early research on regression-type relationships in this context is [DB04], for gene-gene regulatory networks [DHJ$^+$04].

[2] Relationships between copy number variation (CNV) and certain diseases are reviewed in [AP12].

[3] Actually, dropping data is something common: for example, data sets usually come with metadata such as the lab the measurements where taken, the machines used in the experiments, and so on—still, such information is rarely included in statistical analyses.

[4] High-dimensional statistics is often associated with the phrase “curse of dimensionality,” insinuating that high-dimensionality is an unwanted but inevitable side effect of modern data. This view probably roots in the wealth of mathematical and algorithmic challenges high-dimensional spaces are known to bring about. An archetypical example

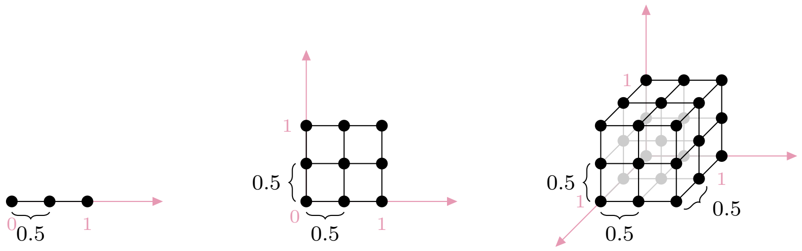


Figure 1.2: It takes $\lceil 1/d+1 \rceil$ intersection points to cover a one-dimensional unit "cube" with an equally-spaced, rectangular lattice that has distance $d \in (0,\infty)$ between neighboring points ($d=0.5$ in the left panel). In two dimensions, it takes $\lceil 1/d+1 \rceil^2$ points (middle panel), and in three dimensions, it already takes $\lceil 1/d+1 \rceil^3$ points (right panel). In general, it takes $\lceil 1/d+1 \rceil^p$ points to cover a p dimensional unit cube, which is an exponential increase in p.

for those challenges is that volumes are very difficult to estimate in $\mathbb{R}^p$ with p large, because the number of intersection points needed to form lattices with fixed distances between adjacent points increases exponentially in the number of dimensions of the ambient space (see Figure 1.2). But from a statistical perspective, we should follow Section 1.1 in viewing high-dimensionality as an *opportunity* generated by modern data: only rich enough data can bear the large parameter spaces that unravel phenomena in fine detail. Accordingly, we could associate high-dimensional statistics with the phrase "blessing of dimensionality."

[5] We just assume that one of the predictors is reserved for modeling the intercept: for example, if $x_1 = 1$ for all observations, then $\beta_1 x_1 = \beta_1$ is the intercept.

[6] The lasso has been introduced in [Tib96]. Its fast penetration into the sciences was also due to the rapid development of corresponding algorithms, such as in [OPT00] and [EHJT04].

[7] Some of the earliest theoretical results for the lasso (albeit not in the form of oracle inequalities) are derived [GR04].

[8] An early example is Tikhonov's 1943 paper [Tik43].

[9] An early paper on the Ridge estimator in regression is [HK70].

General books on high-dimensional statistics and lasso-type estimators include [vdGB11, Gir14, HTW15, vdG16].

1.6 Exercises

Exercises for Section 1.1

□ **Exercise 1.1** ◇◇● In this exercise, we confirm the models of Table 1.1. If you are not sufficiently familiar with the estimators yet, come back to this exercise after reading through Chapters 2–4.

1. Use `R` to build the least-squares model based on data set A and to check the model's adjusted R^2 value. You can use the `lm()` function.

2. Use `R` to build the lasso model based on data set B and to check the model's adjusted R^2 value. For this, load the `glmnet` package for the lasso estimator and run `coef(glmnet(y=y, x=x, lambda=1))` with the appropriate `y` and `x`, which, in our case, provides the first two parameters that enter the lasso path (as a

bonus exercise, you can confirm this). Then, run the `lm()` function on the two selected parameters to obtain the corresponding least-squares model.

Exercises for Section 1.2

□ **Exercise 1.2** $^{\diamond\diamond\diamond\bullet}$ In this exercise, we generalize the results of Section 1.2 to cases where $X^\top X$ is not necessarily invertible. We show in particular that least-squares prediction is expected to be accurate only if $\operatorname{rank}[X] \ll n$.

1. Show that if $X^\top X$ is invertible, then $p \leq n$. Conclude that the treatment in Section 1.2 applies only to settings where the number of parameters is at most as large as the number of samples.

2. Show that $\operatorname{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2$ contains only one point if an only if $X\boldsymbol{\gamma} \neq \mathbf{0}_n$ for all $\boldsymbol{\gamma} \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$. Conclude that the least-squares estimator is not unique if $X^\top X$ is not invertible.

3. Show that any $\widehat{\boldsymbol{\gamma}}_{\mathrm{ls}}, \widehat{\boldsymbol{\beta}}_{\mathrm{ls}} \in \operatorname{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2$ provide the same prediction: $X\widehat{\boldsymbol{\gamma}}_{\mathrm{ls}} = X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$.

4. Show that $\widehat{\boldsymbol{\beta}}_{\mathrm{ls}} = X^+\boldsymbol{y}$ is always a solution of the least-squares estimator, where X^+ is a Moore-Penrose inverse[1] of X. Consult Definition B.2.2 of Moore-Penrose inverses on Page 171 in the Appendix if needed.

5. Consider the singular value decomposition $X = UDV^\top$ used in the main text. (i) Show that D^+ defined as diagonal matrix with diagonal elements $D^+_{ii} := 1/D_{ii}$ if $D_{ii} \neq 0$ and $D^+_{ii} := 0$ otherwise is a Moore-Penrose inverse of D.
   (ii) Show that DD^+ is diagonal with $\operatorname{rank}[X]$ ones on its diagonal and zeros everywhere else.
   (iii) Show that $X^+ = VD^+U^\top$ is a Moore-Penrose inverse of X.

6. Show that for any least-squares solution $\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$, it holds that
$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2 \;=\; \|UDD^+U^\top\boldsymbol{u}\|_2^2\,.$$

7. Conclude that if $\boldsymbol{u} \sim \mathcal{N}_n[\mathbf{0}_n, \sigma^2\,\mathrm{I}_{n\times n}]$ for a $\sigma \in (0,\infty)$, it holds that
$$\mathbb{E}\left[\frac{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2}{n}\right] \;=\; \frac{\sigma^2\operatorname{rank}[X]}{n}\,.$$
   Here, the expectation is taken over the noise $\boldsymbol{u}$, while the design is assumed fix.

8. BONUS: Show that $(X^\top X)^+X^\top$ is a Moore-Penrose inverse of X, where $(X^\top X)^+$ is a Moore-Penrose inverse of $X^\top X$. Conclude that $\widehat{\boldsymbol{\beta}}_{\mathrm{ls}} = (X^\top X)^+X^\top\boldsymbol{y}$ is always a least-square estimator.

Exercises for Section 1.3

□ **Exercise 1.3** $^{\diamond\diamond\bullet}$ In this exercise, we motivate the ridge and lasso estimators from a Bayesian perspective. For this, we consider a linear regression model
$$\boldsymbol{y} \;=\; X\boldsymbol{\beta} + \boldsymbol{u}$$

[1] Moore-Penrose inverses are unique, so that we could also say “the” Moore-Penrose inverse. However, this is irrelevant for the question.

with outcome $\boldsymbol{y} \in \mathbb{R}^n$, design matrix $X \in \mathbb{R}^{n\times p}$, regression vector $\boldsymbol{\beta} \in \mathbb{R}^p$, and noise $\boldsymbol{u} \sim \mathcal{N}_n[\mathbf{0}_n, \sigma^2\,\mathrm{I}_{n\times n}]$. We assume the design X to be fixed. In a frequentist framework, $\boldsymbol{\beta}$ is also fixed, which means in particular that the outcome is distributed according to $\boldsymbol{y} \sim \mathcal{N}_n[X\boldsymbol{\beta}, \sigma^2\,\mathrm{I}_{n\times n}]$. In a hierarchical Bayes framework, on the other hand, $\boldsymbol{\beta}$ follows some prior distribution, and, instead of the outcome itself, the outcome *given* the parameter vector is distributed according to $\boldsymbol{y}|\boldsymbol{\beta} \sim \mathcal{N}_n[X\boldsymbol{\beta}, \sigma^2\,\mathrm{I}_{n\times n}]$. Adopting the Bayesian viewpoint, we study the *map estimator* (maximum a posterior estimator)

$$\widehat{\boldsymbol{\beta}}_{\text{map}} \in \operatorname*{argmax}_{\boldsymbol{\alpha}\in\mathbb{R}^p} l[\boldsymbol{\alpha}|\boldsymbol{y}]\,,$$

where the posterior likelihood $l[\boldsymbol{\alpha}|\boldsymbol{y}]$ is the density of $\boldsymbol{\alpha}|\boldsymbol{y}$ as a function of $\boldsymbol{\alpha}$.

Assume that the parameter vector is distributed independently of the noise and has strictly positive density[2] g, which implies that $l[\boldsymbol{\alpha}|\boldsymbol{y}]$ is proportional to $f_{\boldsymbol{\alpha}}[\boldsymbol{y}]g[\boldsymbol{\alpha}]$, where $f_{\boldsymbol{\alpha}}$ is the density of $\mathcal{N}_n[X\boldsymbol{\alpha}, \sigma^2\,\mathrm{I}_{n\times n}]$. Establish the following two relationships between the map estimator and the "frequentist estimators" lasso and ridge.

1. Show that in case of a multivariate normal prior distribution $\boldsymbol{\beta} \sim \mathcal{N}_p[\mathbf{0}_p, \tau^2\,\mathrm{I}_{p\times p}]$, the map estimator coincides with the ridge

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} \in \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\bigl\{\,\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_2^2\,\bigr\}$$

   with tuning parameter $r = \sigma^2/\tau^2$. Hint: use the identity $\max_{\boldsymbol{\alpha}\in\mathbb{R}^p} l[\boldsymbol{\alpha}|\boldsymbol{y}] = \min_{\boldsymbol{\alpha}\in\mathbb{R}^p} -\log[f_{\boldsymbol{\alpha}}[\boldsymbol{y}]g[\boldsymbol{\alpha}]]$.

2. Show that in case of a multivariate Laplace (double-exponential) prior distribution with density $e^{-\|\boldsymbol{\beta}\|_1/\tau}/(2\tau)^p$, the map estimator coincides with the lasso

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}} \in \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\bigl\{\,\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_1\,\bigr\}$$

   with tuning parameter $r = 2\sigma^2/\tau$.

We conclude that the ridge and lasso estimators can be formulated as map estimators of hierachical Bayesian frameworks.

3. Show that in general,

$$\widehat{\boldsymbol{\beta}}_{\text{map}} \in \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\bigl\{\,\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 - 2\sigma^2\log g[\boldsymbol{\alpha}]\,\bigr\}\,.$$

We conclude that there is a general correspondance between map estimators and regularized least-squares estimators.[3]

Exercises for Section 1.4

□ **Exercise 1.4** $^{\diamond\bullet}$ In this exercise, we verify and extend the claims about convexity in Section 1.4.

1. Show that $\boldsymbol{\alpha} \mapsto \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2$ is strictly convex if and only $X^\top X$ invertible.
2. Give an example of a convex but not strictly convex function that has multiple minima and one that has only one minimum. BONUS: is the function $\boldsymbol{a} \mapsto a_1^2a_2^2$ convex?

[2] We consider densities with respect to the appropirate Lebesgue measures unless stated otherwise.

[3] This connection has been highlighted through the introduction of the the Bayes lasso [PC08] and discussed further in the context of tuning parameter calibration [BL17, Section 2.3].

3. Show that in contrast, the minima of *strictly* convex functions, say $f\ :\ \mathbb{R}^p \to \mathbb{R}$, are always unique.

□ **Exercise 1.5** $^{\diamond\diamond}$ Use the data discussed in Section 1.4 to study the continuity of the ridge estimator. Specifically, show that the ridge estimator with any fixed tuning parameter is a continuous function of the parameter d of the data in that section, that is: for all $r, u > 0$, there is a $v > 0$ such that for all $d, d' \in \mathbb{R}$, $|d - d'| < v$, it holds that

$$\|\widehat{\boldsymbol{\beta}}_{\text{ridge}}[r, \boldsymbol{y}, X^d] - \widehat{\boldsymbol{\beta}}_{\text{ridge}}[r, \boldsymbol{y}, X^{d'}]\|_1 < u\,,$$

where $X^d = ((1,2)^\top, (d,2)^\top)$ (and analogeously $X^{d'}$) is the design from Table 1.3 for given d, and $\widehat{\boldsymbol{\beta}}_{\text{ridge}}[r, \boldsymbol{y}, X^d]$ (and analogeously $\widehat{\boldsymbol{\beta}}_{\text{ridge}}[r, \boldsymbol{y}, X^{d'}]$) is the ridge estimator with tuning parameter r on the data $(\boldsymbol{y}, X^d)$.

□ **Exercise 1.6** $^{\diamond\bullet}$ This exercise corroborates our statement that the smaller the tuning parameter, the better the ridge estimator approximates a least-squares solution. For this, consider the data discussed in Section 1.4 with $d = 1$. Show that

$$\|\widehat{\boldsymbol{\beta}}_{\text{ridge}}[r] - (1/2, 1/2)^\top\|_2 \ = \ \frac{r}{\sqrt{2}(10 + r)}$$

and conclude that the smaller the tuning parameter, the closer is the corresponding ridge estimator to a least-squares solution of the form $(1 - a, a)^\top$.

□ **Exercise 1.7** $^{\diamond\diamond\bullet}$ In this exercise, we provide mathematical details on the plots of Figure 1.1.

Consider again the data in Table 1.3.

1. Show that for any $d \in \mathbb{R}$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^\top$, it holds that

$$\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 \ = \ 5(\alpha_1 - 1)^2 + (4 + d^2)\alpha_2^2 + 2(4 + d)(\alpha_1 - 1)\alpha_2\,.$$

2. Verify that for $d = 1$, the equation simplifies to

$$\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 \ = \ 5(\alpha_1 - 1 + \alpha_2)^2\,.$$

3. Conclude that for $d = 1$ and any $c \in [0, \infty)$, it holds that

$$\bigl\{\boldsymbol{\alpha} \in \mathbb{R}^2\ :\ \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 = c\bigr\} \ = \ \left\{\boldsymbol{\alpha} \in \mathbb{R}^2\ :\ \alpha_2 = 1 - \alpha_1 \pm \sqrt{\frac{c}{5}}\right\}.$$

   This means that the function $\boldsymbol{\alpha} \mapsto \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2$ has level sets (a level set is a complete set of function arguments that have the same function value) specified by linear relationships between the coordinates. These level sets are indicated by the straight lines in the left panel of Figure 1.1.

4. We consider now the level sets of ellipses centered at $(1,0)^\top$ and parameterized by $\omega \in [0, 180°)$ and $a, b \in (0, \infty)$:

$$\left\{\boldsymbol{\alpha} \in \mathbb{R}^2\ :\ \frac{\bigl((\cos\omega)(\alpha_1 - 1) + (\sin\omega)\alpha_2\bigr)^2}{a^2} + \frac{\bigl(-(\sin\omega)(\alpha_1 - 1) + (\cos\omega)\alpha_2\bigr)^2}{b^2} = c\right\}.$$

   The parameter c indexes the different level sets. Show that for $\omega \to 45°$, $a \to 1/\sqrt{10}$, and $b \to \infty$, these level sets converge to

$$\bigl\{\boldsymbol{\alpha} \in \mathbb{R}^2\ :\ 5(\alpha_1 - 1 + \alpha_2)^2 = c\bigr\}\,.$$

For this, recall that a sequence of sets $\mathcal{A}_1, \mathcal{A}_2, \dots$ converges to $\mathcal{A}$ if for any point $a \in ((\cup_j \mathcal{A}_j) \cup \mathcal{A})$, there is an $n \in \{1, 2, \dots\}$ such that either $a \in (\mathcal{A}_m \cap \mathcal{A})$ for all $m \geq n$ or $a \notin \mathcal{A}_m, a \notin \mathcal{A}$ for all $m \geq n$. In the case here, it suffices to show that $((\cos\omega)(\alpha_1 - 1) + (\sin\omega)\alpha_2)^2/a^2 + (-(\sin\omega)(\alpha_1 - 1) + (\cos\omega)\alpha_2)^2/b^2 \to 5(\alpha_1 - 1 + \alpha_2)^2$.

Conclude with 2. that for $d = 1$, we can think of the least-square's level sets as 45°-rotated ellipses that are streched infinitely in one direction.

5. We switch to $d \neq 1$ and write

$$\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 \;=\; \frac{\big((\cos\omega)(\alpha_1 - 1) + (\sin\omega)\alpha_2\big)^2}{a^2} + \frac{\big(-(\sin\omega)(\alpha_1 - 1) + (\cos\omega)\alpha_2\big)^2}{b^2}$$

for $\omega \in [0°, 180°)$, $a, b \in (0, \infty)$. In the following, we compute the parameters as functions of d and show that the least-squares objective function can indeed be written in that elliptic form. For this, verify first that

$$\begin{aligned}
&\frac{\big((\cos\omega)(\alpha_1 - 1) + (\sin\omega)\alpha_2\big)^2}{a^2} + \frac{\big(-(\sin\omega)(\alpha_1 - 1) + (\cos\omega)\alpha_2\big)^2}{b^2}\\
&\quad= \left(\frac{(\cos\omega)^2}{a^2} + \frac{(\sin\omega)^2}{b^2}\right)(\alpha_1 - 1)^2 + \left(\frac{(\sin\omega)^2}{a^2} + \frac{(\cos\omega)^2}{b^2}\right)\alpha_2^2\\
&\qquad\qquad + 2\left(\frac{1}{a^2} - \frac{1}{b^2}\right)(\cos\omega)(\sin\omega)(\alpha_1 - 1)\alpha_2\,.
\end{aligned}$$

6. Compare the $(\alpha_1 - 1)$-terms of the formulations in 5. and 1. to show that if $a \neq b$, it holds that

$$(\cos\omega)^2 \;=\; \frac{5 - \frac{1}{b^2}}{\frac{1}{a^2} - \frac{1}{b^2}}\,.$$

7. Compare the α_2-terms of the formulations in 5. and 1. to show that if $a \neq b$, it holds that

$$\frac{1}{a^2} + \frac{1}{b^2} \;=\; 9 + d^2\,.$$

8. Use the previous two steps to show that if $a \neq b$, it holds that

$$(\cos\omega)^2(\sin\omega)^2 \;=\; \frac{5(4 + d^2)b^4 - (9 + d^2)b^2 + 1}{(9 + d^2)^2 b^4 - 4(9 + d^2)b^2 + 4}\,.$$

9. Show similarly as in the previous steps that $a \neq b$ for any $d \neq 1$. Conclude from 5.–7. that if $d \neq 1$, the least-squares objective function can indeed be written in the mentioned elliptic form with the parameters derived in 6.–7.

10. Compare the $(\alpha_1 - 1)\alpha_2$-terms of the formulations in 1. and 5. to show that

$$\frac{5(4 + d^2) - \frac{9 + d^2}{b^2} + \frac{1}{b^4}}{(9 + d^2)^2 - \frac{4(9 + d^2)}{b^2} + \frac{4}{b^2}} - \frac{(4 + d)^2}{(9 + d^2 - \frac{2}{b^2})^2} \;=\; 0\,.$$

11. Use the above insights to draw level sets of the least-squares loss for $d = 1$ and $d \neq 1$.

R Lab Chapter 1

1 Least-squares vs. ridge estimation

In this lab, we compare the least-squares estimator with the ridge estimator.

Your task is to replace the keyword `REPLACE` by suitable code and to answer the questions posed in the text.

1.1 Generating toy data

Generate data according to $\mathbf{y} = X\boldsymbol{\beta} + \mathbf{u}$ with $X \in \mathbb{R}^{2,2}$ as in Table 1.2 with $d = 2$, $\boldsymbol{\beta} = (1, 1/2)^\top$, and $\mathbf{u} \sim \mathcal{N}_2[\mathbf{0}_2, (0.1)^2\, \mathrm{I}_{2,2}]$. You might want to use the `rnorm()` function.

```
set.seed(11)
DesignFamily <- function(d)
{
  return(matrix(c(1, 2, d, 2), nrow=2, ncol=2))
}
design <- DesignFamily(2)
beta <- c(1, 1/2)
outcome <- REPLACE
```

1.2 Implementing the estimators

Implement the least-squares estimator (for $d \neq 1$). You might want to use the `solve()` function.

```
LsEstimator <- function(y, X)
{
  REPLACE
}
LsEstimator(outcome, design)
```

```
##           [,1]
## [1,] 1.0617625
## [2,] 0.4395672
```

Similarly, implement the ridge estimator.

```
RidgeEstimator <- function(y, X, r)
{
  REPLACE
}
RidgeEstimator(outcome, design, 1)
```

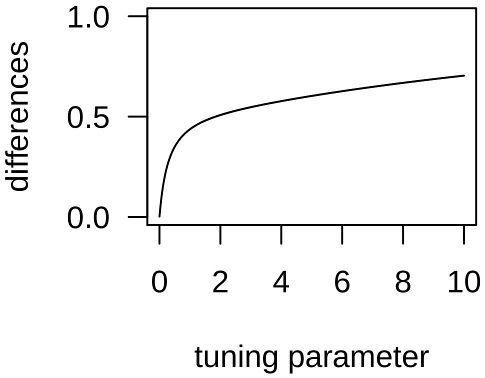
```
##           [,1]
## [1,] 0.6774037
## [2,] 0.6469656
```

1.3 Showing that the ridge estimator approximates a least-squares solution

Show that the ridge estimator approaches in ℓ_2-norm the least-squares solution when r goes to zero. For this, establish a function that compares the ℓ_2-differences between the ridge estimator and the least-squares:

$\|\widehat{\beta}_{\text{ridge}} - \widehat{\beta}_{\text{ls}}\|_2$.

```
RidgeLsDifference <- function(r, y, X)
{
  REPLACE
}
tuning.parameter <- 0.001 * c(1:10000)
difference <- apply(as.matrix(tuning.parameter), 1, RidgeLsDifference,
                             y=outcome, design)
plot(x    = tuning.parameter,
     y    = difference,
     type = "l",
     lty  = 1,
     ylim = c(0, 1),
     yaxp = c(0, 1, 2),
     las  = 1,
     xlab = "tuning parameter",
     ylab = "differences")
```

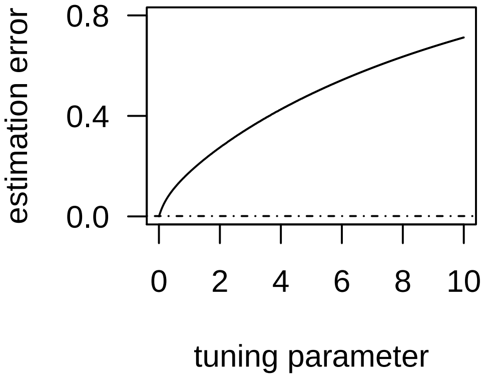


We find that the smaller the tuning parameter, the closer the ridge estimator is to the least-squares estimator. Eventually, for $r \to 0$, they coincide.

1.4 Comparing estimation errors

Compare the ℓ_2-estimator errors of the ridge estimator and the LS estimator: $\|\widehat{\boldsymbol{\beta}}^r_{\text{ridge}} - \boldsymbol{\beta}\|_2$ vs. $\|\widehat{\boldsymbol{\beta}}_{\text{ls}} - \boldsymbol{\beta}\|_2$.

```
RidgeError <- function(r, y, X)
{
  REPLACE
}
LsError <- function(y, X)
{
  REPLACE
}
tuning.parameter <- 0.001 * c(1:10000)
estimation.error <- apply(as.matrix(tuning.parameter), 1, RidgeError,
                          outcome, design)
plot(x    = tuning.parameter,
     y    = estimation.error,
     type = "l",
     lty  = 1,
     ylim = c(0, 0.8),
     yaxp = c(0, 0.8, 2),
     las  = 1,
     xlab = "tuning parameter",
     ylab = "estimation error")
abline(h=LsError(outcome, design), lty=4)
```

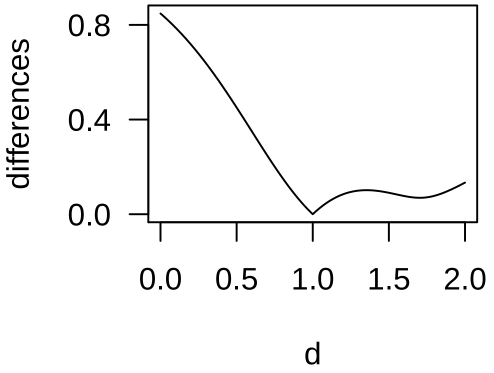


We observe in particular that the ℓ_2-error of the ridge estimator becomes large if the tuning parameter is large. What happens for $r \to \infty$?

1.5 Showing that the ridge estimator is continuous in the data

Show that the ridge estimator is continuous in the variation of the design parameter d around the critical point $d_{\text{critical}} = 1$. For this, compute the ℓ_2-distances of ridge estimators on data with different design parameters d to the ridge estimator on data with fixed design parameter $d_{\text{critical}} = 1$. Set the tuning parameter to $r = 0.5$ throughout.

```
RidgeDifference <- function(d, r, y, d.critical)
{
  REPLACE
}
tuning_parameter_fixed <- 0.5
d.critical <- 1
d <- 0.01 * c(0:200)
differences <- apply(as.matrix(d), 1,  RidgeDifference,tuning_parameter_fixed, outcome,
                     d.critical)
plot(d, differences, type="l", lty=1, yaxp=c(0, 0.8, 2), las=1)
```



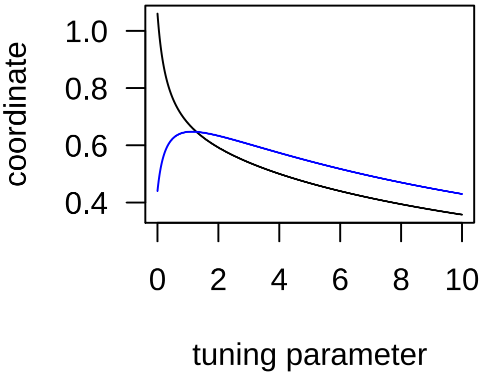
Since the curve is continuous at $d = 1$, we conclude that $\|\widehat{\beta}_{\text{ridge}}[d] - \widehat{\beta}_{\text{ridge}}[d_{\text{critical}}]\|_2 \to 0$ for $d \to d_{\text{critical}} = 1$, where $\widehat{\beta}_{\text{ridge}}[d]$ and $\widehat{\beta}_{\text{ridge}}[d_{\text{critical}}]$ denote the ridge estimators (with tuning parameter $r = 0.5$) on data indexed by d and d_{critical}, respectively. This means that the ridge estimator is continuous in the data as d varies around the critical point $d_{\text{critical}} = 1$.

Is the ridge estimator continuous as a function of d more generally?

1.6 Showing that the coordinates of the ridge estimator are not necessarily monotone in the tuning parameter

Use the toy data to show that the individual coordinates of the ridge estimator are not necessarily monotone in the tuning parameter.

```
RidgeCoordinate <- function(r, y, d, c)
{
  REPLACE
}
d <- 2
tuning.parameter <- 0.001 * c(1:10000)
coordinate_1 <- apply(as.matrix(tuning.parameter), 1, RidgeCoordinate, outcome, d, 1)
coordinate_2 <- apply(as.matrix(tuning.parameter), 1, RidgeCoordinate, outcome, d, 2)
plot(x    = tuning.parameter,
     y    = coordinate_1,
     type = "l",
     col  = "black",
     las  = 1,
     yaxp = c(0.4, 1, 3),
     xlab = "tuning parameter",
     ylab = "coordinate")
lines(tuning.parameter, coordinate_2, col="blue")
```



The curve of the second coordinate increases until about $r = 1.5$ and then decreases, which shows that the coordinates are not necessarily monotone in the tuning parameter. Nevertheless, both curves decrease monotoneously approximately as c/r for r large, where $c > 0$ can depend on the coordinate (and the data, of course), which corroborates our theoretical findings in the main text.

1.7 Comparing ridge and least-squares on economic data

We apply the ridge estimator and the least-squares on the `longley` data that contains seven economic variables and is included in the standard `R` distributions. Take as outcome the `Employed` variable, and as predictors `GNP.deflator`, `GNP`, `Unemployed`, `Armed.Forces`, and `Population`. Also add a column of ones to design matrix to account for the intercept. (We penalize the intercept as any other variable. BONUS: write a version of the ridge estimator that does not penalize the intercept.) Hint: for easier manipulations later, use the `as.matrix()` function to convert all quantities to matrices.

```
outcome <- as.matrix(longley["Employed"])
design <- REPLACE
cbind(outcome, design)[1:5, ]
```

```
##      Employed Intercept GNP.deflator     GNP Unemployed Armed.Forces
## 1947   60.323         1         83.0 234.289      235.6        159.0
## 1948   61.122         1         88.5 259.426      232.5        145.6
## 1949   60.171         1         88.2 258.054      368.2        161.6
## 1950   61.187         1         89.5 284.599      335.1        165.0
## 1951   63.221         1         96.2 328.975      209.9        309.9
##      Population
## 1947    107.608
## 1948    108.632
## 1949    109.773
## 1950    110.929
## 1951    112.075
```

```
LsEstimator(outcome, design)
```

```
##                  Employed
## Intercept    92.461307830
## GNP.deflator -0.048462828
## GNP           0.072003849
## Unemployed   -0.004038711
## Armed.Forces -0.005604956
## Population   -0.403508682
```

```
RidgeEstimator(outcome, design, 1)
```

```
##                 Employed
## Intercept     0.02227192
## GNP.deflator  0.21920075
## GNP          -0.01035088
## Unemployed   -0.01394317
## Armed.Forces -0.00579588
## Population    0.45119435
```

```
RidgeEstimator(outcome, design, 100)
```

```
##                  Employed
## Intercept     0.004790496
## GNP.deflator  0.252814814
## GNP          -0.012145722
## Unemployed   -0.012151700
## Armed.Forces -0.003817691
## Population    0.418655390
```

```
RidgeEstimator(outcome, design, 10000)
```

```
##                 Employed
## Intercept    0.001724652
## GNP.deflator 0.101214227
## GNP          0.012424157
## Unemployed   0.043045876
## Armed.Forces 0.065315837
## Population   0.156333129
```

Are the ridge estimator's coordinates monotone in the tuning parameter here? Is the ridge estimator's magnitude in ℓ_2 monotone in the tuning parameter here?

1.8 BONUS: checking the theoretical formulae of Section 1.4

We show that the formula in Section 1.4 are correct. We first compare the least-squares estimator to the version in the script for three values of d:

```
LsEstimatorTheory <- c(1, 0)
outcome <- c(1, 2)
LsEstimator(outcome, DesignFamily(0)) - LsEstimatorTheory
```

```
##                [,1]
## [1,]  2.220446e-16
## [2,] -2.220446e-16
```

```
LsEstimator(outcome, DesignFamily(2)) - LsEstimatorTheory
```

```
##                [,1]
## [1,] -4.440892e-16
## [2,]  0.000000e+00
```

```
LsEstimator(outcome, DesignFamily(1.1)) - LsEstimatorTheory
```

```
##                [,1]
## [1,] -1.705303e-13
## [2,]  5.684342e-14
```

We find equivalence within numerical precision.

We now compare the ridge estimator to the version in the script for three values of d and r:

```
RidgeEstimatorTheory <- function(d, r)
{
  return(1 / (4 - 8 * d + 4 * d^2 + (9 + d^2) * r + r^2) *
         c(4 - 8 * d + 4 * d^2 + 5 * r, (4 + d) * r))
}
RidgeEstimator(outcome, DesignFamily(0), 1) - RidgeEstimatorTheory(0, 1)
```

```
##               [,1]
## [1,] 2.220446e-16
## [2,] 0.000000e+00
```

```
RidgeEstimator(outcome,DesignFamily(2), 0.1) - RidgeEstimatorTheory(2, 0.1)
```

```
##                [,1]
## [1,]  1.443290e-15
## [2,] -6.245005e-16
```

```
RidgeEstimator(outcome, DesignFamily(1.1), 0.5) - RidgeEstimatorTheory(1.1, 0.5)
```

```
##                [,1]
## [1,] -2.220446e-16
## [2,]  5.551115e-17
```

We find again equivalence within numerical precision.

1.9 BONUS: a technical note

Consider the following two functions:

```
Function_1 <- function(y, X)
{
  return(solve(t(X) %*% X) %*% t(X) %*% y)
}
Function_2 <- function(y, X)
{
  return(solve(t(X) %*% X, t(X) %*% y))
}
y <- c(1, 2)
X <- matrix(c(1, 2, 0, 2), nrow=2, ncol=2)
Function_1(y, X)
```

```
##                [,1]
## [1,]  1.000000e+00
## [2,] -2.220446e-16
```

```
Function_2(y, X)
```

```
##      [,1]
## [1,]    1
## [2,]    0
```

What is the difference between the two functions?

To keep the code as close as possible to the script, we have generated all labs' results by using the solve function as in `Function_1`. However, the faster and numerically more stable version is `Function_2`, which solves a system of equations instead of inverting a matrix and subsequently multiplying that inverse with a vector.

Implement the least-squares and the ridge as suggested by both `Function_1` and `Function_2` and compare the results for this lab. Are there any differences?

Chapter 2

Linear Regression

In this chapter, we focus on linear regression models of the form

$$\boldsymbol{y} \;=\; X\boldsymbol{\beta} + \boldsymbol{u}\,,$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the outcome, $X \in \mathbb{R}^{n\times p}$ the design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ the regression vector, and $\boldsymbol{u} \in \mathbb{R}^n$ the noise. Our goal is to estimate the unknown model parameter $\boldsymbol{\beta}$ from the data $(\boldsymbol{y}, X)$. The statistical challenge is that the relationship between the outcome and the design is obscured by the in practice unknown noise $\boldsymbol{u}$. Since we are particularly interested in high-dimensional settings, where the number of parameters p can rival or even exceed the number of samples n, we cannot use the least-squares estimator as is. We instead consider slightly more complex estimators that complement least-squares estimation with prior terms that leverage additional information about the data generating process at hand.

2.1 Sparsity Inducing Prior Functions

In the previous chapter, we have shown that classical estimators can deteriorate rapidly with increasing number of parameters. This deterioration can show itself in a phenomenon called *overfitting*: the estimators lose themselves in the pecularities of the data at hand, thereby missing the essential structure of the data generating process (see Figure 2.1). Overfitting can be avoided by complementing the bare measurements with additional information. Such information can stem from previous studies, the experimental design, physical laws, and other sources. Prior functions formulate this information mathematically and funnel it into the statistical analysis. overfitting

The type of information that is most frequently encountered in high-dimensional statistics is *sparsity*. Sparsity states that the data generating process can be modelled accurately by using only a small number of predictors. The first attempt to leverage such information is (the Langrangian version of) *best subset selection* sparsity best subset selection

$$\widehat{\boldsymbol{\beta}}_{\text{subset}} \;\in\; \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\bigl\{\,\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_0\,\bigr\}\,,$$

where $r \in [0, \infty]$ is a tuning parameter and $\|\boldsymbol{\alpha}\|_0 := \#\{j \in \{1,\ldots,p\} : \alpha_j \neq 0\}$ counts the number of non-zero-valued coordinates in $\boldsymbol{\alpha}$. The least-squares term $\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2$

```
    mgp = c(3, 1, 0),  # location of the labels
    xpd = NA)  # allows the plots to overlap
PlotLsLasso(2, TRUE)
PlotLsLasso(5, FALSE)
PlotLsLasso(20, FALSE)
```

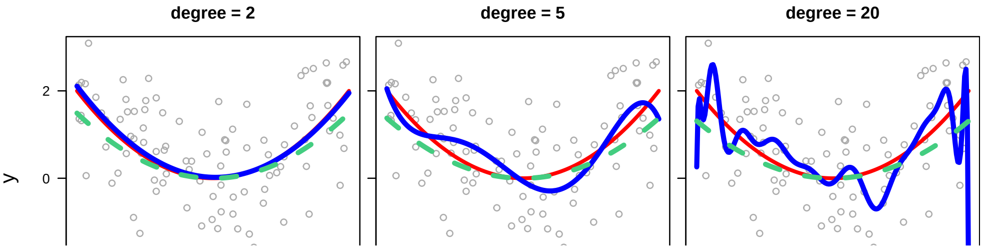


Figure 2.1: This illustration of overfitting involves noisy observations (gray circles) of a simple quadratic function (red, solid curves) for fitting polynomials with the least-squares (blue, solid lines) and the lasso (green, dashed lines). The polynomials used in the estimations have degrees $2, 5, 20$ (panels from left to right). The larger the degree of these polynomials, that is, the larger the number of irrelevant predictors (the degree of the true data generating polynomial is 2), the less the least-squares estimates capture the shape of the underlying function; we call this overfitting. The lasso (with cross-validated tuning parameter), on the other hand, simply disregards those additional predictors, and therefore, does not suffer from overfitting in this example. This is the case because the lasso assumes correctly that the underlying model is simple. See the lab for details.

ensures a good fit to the data, while the ℓ_0-prior term favors sparse parameters, that is, parameter vectors with a small number of non-zero coordinates.

The ℓ_0-prior is a non-linear, "combinatorial" function, which makes it hard to optimize over—especially for large p. One approach to lessen this computational burden is to mimic $\|\boldsymbol{\alpha}\|_0$ with $\|\boldsymbol{\alpha}\|_q := (\sum_{j=1}^{p} |\alpha_j|^q)^{1/q}$ for some $q > 0$.[1] We have already seen a corresponding estimator in the last chapter: the ridge estimator, where $q = 2$. However, while the ridge is particularly easy to compute (it even has an explicit solution), it does not preserve sparsity. A better mimic of best subset selection is the *lasso estimator*, where $q = 1$:

lasso

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}} \in \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\Bigl\{\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_1\Bigr\}. \tag{2.1}$$

The lasso has the best of both worlds: it has a convex objective function, which makes it amenable to the rich literature on convex optimization, and yet inherits the best subset selection's sparsity—see Figure 2.3.

Refinements of the simple notion of sparsity above are called *structured sparsity*. Structured sparsity in the form of group sparsity, for example, describes group-wise behaviors: Consider a partition of the index set $\{1, \dots, p\}$, that is, a collection of disjoint sets $\mathcal{A}^1, \dots, \mathcal{A}^d$ that satisfy $\cup_{j=1}^{d}\mathcal{A}^j = \{1, \dots, p\}$. While sparsity means that most of the individual predictors are irrelevant ($|\{j : \beta_j \neq 0\}| \ll \min\{n, p\}$), group sparsity means that most index sets $\mathcal{A}^j$ denote irrelevant parts of the regression vector ($|\{j : \boldsymbol{\beta}_{\mathcal{A}^j} \neq \mathbf{0}_{|\mathcal{A}^j|}\}| \ll \min\{n, p\}$). An estimator that leverages this type of structured sparsity is the *group lasso*[2]:

group lasso

$$\widehat{\boldsymbol{\beta}}_{\text{group}} \in \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\Bigl\{\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + r\sum_{j=1}^{d}\|\boldsymbol{\alpha}_{\mathcal{A}^j}\|_2\Bigr\}.$$

The group lasso can be thought of as an intermediate between the lasso and ridge estimators: The group lasso coincides with the lasso when $d = p$ and $\mathcal{A}^j = \{j\}$, and more generally, inherits the sparsity inducing property of the lasso in the sense that it sets entire subvectors $\boldsymbol{\alpha}_{\mathcal{A}^j}$ to zero. The group lasso coincides with the ridge estimator[3] when $d = 1$ and $\mathcal{A}^1 = \{1, \dots, p\}$, and more generally, acts like the ridge within the groups in the sense that the objective function is invariant under orthogonal transformations within groups and that if $\|\boldsymbol{\alpha}_{\mathcal{A}^j}\|_2 \neq 0$, then typically *all* coordinates of that subvector are non-zero,

Estimators can also comprise combinations of several prior functions. A popular example is the *elastic net*[4]

elastic net

$$\widehat{\boldsymbol{\beta}}_{\text{elastic}} \in \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\Bigl\{\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + r\bigl(w\|\boldsymbol{\alpha}\|_1 + (1-w)\|\boldsymbol{\alpha}\|_2^2\bigr)\Bigr\},$$

which interpolates between the lasso ($w = 1$) and the ridge ($w = 0$). The elastic net with an intermediate value $w \in (0, 1)$ inherits properties from both of the two base estimators: on the one hand, the resulting estimator is similar to the lasso in the sense that it tends to be sparse; on the other hand, the resulting estimator is much unlike the lasso but instead similar to the ridge in the sense that it is always unique and tends to select whole groups of highly correlated variables rather than just one representative among them.[5] The value of w is often fixed based on subjective reasonings beforehand, because a fully data-driven calibration of both tuning parameters $r \in [0, \infty]$ and $w \in [0, 1]$ is conceptually and computationally challenging. In general, combining prior functions always bears the problem of having to deal with multiple tuning parameters.

The ℓ_0-function favors solutions with many zero-valued coordinates, but it completely disregards parameter values otherwise: for example, $\|\boldsymbol{\alpha}\|_0 = 1$ whether $\boldsymbol{\alpha} = (1, 0, \dots, 0)^\top$ or $\boldsymbol{\alpha} = (100, 0, \dots, 0)^\top$. In contrast, the discussed replacements of that function keep increasing in the magnitudes of the parameter values: for example, $\|\boldsymbol{\alpha}\|_1 = 1$ for $\boldsymbol{\alpha} = (1, 0, \dots, 0)^\top$ but $\|\boldsymbol{\alpha}\|_1 = 100$ for $\boldsymbol{\alpha} = (100, 0, \dots, 0)^\top$. In addition to setting a fraction of the parameters exactly to zero as intended, this general favoring of small parameters can introduce an unwanted *overall* shrinkage of the estimates. To remove such biases, *least-squares refitting*[6] complements regularized estimators $\widehat{\boldsymbol{\beta}}$ with subsequent least-squares estimation on the support $\widehat{\mathcal{S}} := \operatorname{supp}[\widehat{\boldsymbol{\beta}}]$:

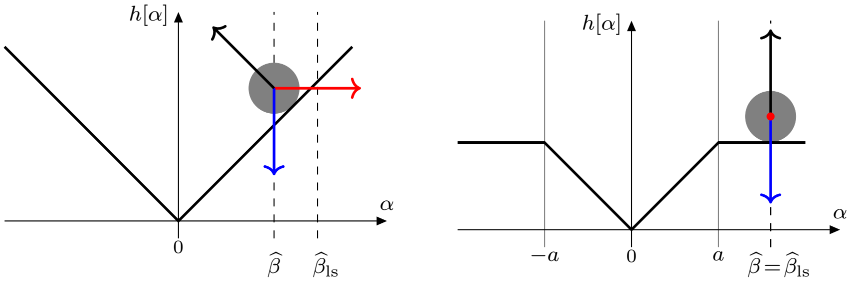
refitting

$$\widehat{\boldsymbol{\beta}}_{\text{refitting}} \in \operatorname*{argmin}_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^p\\ \operatorname{supp}[\boldsymbol{\alpha}]\subset\widehat{\mathcal{S}}}} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2\,.$$

In other words, the initial estimator is used only for a screening for non-zero coordinates, while the corresponding parameter values are determined entirely by the subsequent least-squares. The rationale is that the least-squares is an accurate and unbiased estimator in low-dimensional and correctly specified models; hence, the two-stage approach presumes that the initial estimator is sparse, that is, $|\widehat{\mathcal{S}}| \ll n$, and that it provides accurate variable selection, that is, $\widehat{\mathcal{S}}$ is a good approximation of the true support $\mathcal{S} := \operatorname{supp}[\boldsymbol{\beta}]$. In the ideal case $\widehat{\mathcal{S}} = \mathcal{S}$, the least-squares refitted estimator possesses the *strong oracle property* that it equals the oracle "estimator"

$$\widehat{\boldsymbol{\beta}}_{\text{oracle}} \in \operatorname*{argmin}_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^p\\ \operatorname{supp}[\boldsymbol{\alpha}]\subset\mathcal{S}}} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2$$

that knows the true support beforehand. But in practice, exact support recovery $\widehat{\mathcal{S}} = \mathcal{S}$ is often unrealistic and, in any case, unverifiable.[7,8]



$h[\alpha]$
α
0
$\widehat{\beta}$
$\widehat{\beta}_{\text{ls}}$
$h[\alpha]$
α
$-a$
0
a
$\widehat{\beta}=\widehat{\beta}_{\text{ls}}$

Figure 2.2: A physics analogy for how capping a prior function can avoid bias in large parameter estimates. The toy estimator is $\widehat{\beta} \in \operatorname{argmin}_{\alpha\in\mathbb{R}}\{\|\boldsymbol{y} - \alpha\boldsymbol{x}\|_2^2 + rh[\alpha]\}$, where $\boldsymbol{x} \in \mathbb{R}^n$ is the only predictor and $r \in (0,\infty)$ a fixed tuning parameter. On the left, $h[\alpha] := |\alpha|$ is the usual ℓ_1-prior in $\mathbb{R}$; on the right, $h[\alpha] := \min\{|\alpha|, a\}$ is a capped version with $a \in (0,\infty)$ defining the transition point. One can think of the estimator's objective function as a physical potential for an iron ball at α that is subject to three forces: a "magnetic force" (red) that is generated by the least-squares part of the objective function and pulls the ball horizontally to the least-squares solution $\widehat{\beta}_{\text{ls}} = \boldsymbol{x}^\top\boldsymbol{y}/\|\boldsymbol{x}\|_2^2$; a "gravitational force" (blue) that is generated by the prior function and pulls the ball vertically downwards; and a "normal force" (black) that is generated by the shape of the prior function and pushes the ball upwards perpendicular to the surface. The three forces are in equilibrium at $\widehat{\beta}$. For the ℓ_1-prior on the left, the inclined surface allows the gravitational force, more precisely, its component parallel to the surface, to pull the estimate away from the (typically unbiased) least-squares solution. For the capped prior on the right, this is still the case if $|\widehat{\beta}_{\text{ls}}| < a$. If $|\widehat{\beta}_{\text{ls}}| \geq a$, on the other hand, the graviational force is perpendicular to the surface and, therefore, cannot work against the magnetic force any more. The magnetic force then moves the ball all the way to the least-squares solution, which is the minimum of the magnetic potential (and consequently, the magnetic force is zero there). This means that our toy estimator can coincide with a least-squares solution even though regularization is imposed.

To a varying extend, least-squares refitting is already integrated in many estimators. For example, least-squares refitting leaves best subset selection completely unchanged, that is, $\widehat{\boldsymbol{\beta}}_{\text{refitting}} = \widehat{\boldsymbol{\beta}}_{\text{subset}}$ if $\widehat{\mathcal{S}} = \operatorname{supp}[\widehat{\boldsymbol{\beta}}_{\text{subset}}]$ in the refitting program. This means that adding least-squares refitting to best subset selection would be redundant. Other examples are least-squares with variants of the capped ℓ_1-function as prior. The a-capped ℓ_1-function, $a \in [0,\infty]$, is the map $\boldsymbol{\alpha} \mapsto \sum_{j=1}^{p} \min\{|\alpha_j|, a\}$. The boundary cases for least-squares with this prior are classical least-squares (for $a = 0$, the prior is zero) and lasso (for $a = \infty$, the prior is the ℓ_1-function); more generally, a is the value below which parameters are pulled towards zero—see Figure 2.2. Smooth versions of the capped ℓ_1-function are used in the *scad* and *mcp* estimators, for example.[9] Such estimators have been shown to satisfy the strong oracle property if $\widehat{\mathcal{S}} = \mathcal{S}$.[10] However, (i) exact support recovery remains a strict and unverifiable assumption; (ii) capped ℓ_1-functions render the objective functions non-convex, which can make computations challenging; and (iii) the prior functions involve one or more additional tuning parameters, such as a in the vanilla case of capped ℓ_1-regularization, that need to be calibrated.[11] In practice, these difficulties have to be weighted against a potential gain in accuracy.

scad and mcp

2.2 Optimality Conditions*

KKT conditions for the lasso:

$$-2X^\top(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\text{lasso}})+r\widehat{\boldsymbol{\kappa}} \;=\; \boldsymbol{0}_p \tag{2.2}$$

for a vector $\widehat{\boldsymbol{\kappa}}\in\boldsymbol{\partial}\|\widehat{\boldsymbol{\beta}}_{\text{lasso}}\|_1$.

2.3 References and Further Reading

[1][FF93, Equation (54) on Page 124] introduces least-squares regression with general ℓ_q-prior functions, $q\in[0,\infty]$; these estimators where later termed *bridge estimators*. The paper also motivates these priors from a Bayesian perspective; we have taken up this viewpoint in Exercise 1.3. Bridge estimators with $q\in[1,2]$ can be seen as interpolations between the lasso and the ridge; however, in contrast to the elastic net, which is another such interpolation introduced below in the main text, these estimators are sparse only if $q=1$—see Figure 2.3.

[2]Least-squares regression with a group lasso prior was first considered in [Bak99, Equation (2.7) on Page 22]. [YL06, Equation (2.1) on Page 51] invokes the more general prior function $\sum_{j=1}^{d}\sqrt{\boldsymbol{\alpha}_{\mathcal{A}^j}^\top K^j\boldsymbol{\alpha}_{\mathcal{A}^j}}$, $K^j\in\mathbb{R}^{|\mathcal{A}^j|\times|\mathcal{A}^j|}$, and those authors coin the term "group lasso" for least-squares estimators with such priors. To facilitate the calibration of the tuning parameter r, [BLS14b, Equation (4) on Page 1314] modifies the group lasso to the *square-root group lasso*

$$\widehat{\boldsymbol{\beta}}_{\sqrt{\text{group}}}\;\in\;\operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\Bigl\{\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2+r\sum_{j=1}^{d}\sqrt{\boldsymbol{\alpha}_{\mathcal{A}^j}^\top K^j\boldsymbol{\alpha}_{\mathcal{A}^j}}\Bigr\}.$$

Both [YL06] and [BLS14b] suggest $K^j:=|\mathcal{A}^j|\,\mathrm{I}_{|\mathcal{A}^j|\times|\mathcal{A}^j|}$ if the predictors are orthogonal within groups (recall that the predictors can be orthogonalized within groups without loss of generality). The factor $|\mathcal{A}^j|$ balances the selection of small and large groups and can be motivated by empirical process theory: if the noise is i.i.d. Gaussian, [vdGB11, Lemma 8.1 on Page 254] for the original case and [BLS14b, Lemma 2.1 on Page 1316] for the square-root case demonstrate that the metioned choice of K^j can allow for a suitable regularization of all groups simulateneously. The *sparse group lasso* of [SFHT13, Equation (3) on Page 232] complements the group regularization with a standard ℓ_1-regularization to obtain also within group sparsity. Overlapping groups have been studied in [OJV11] and others.

[3]One can prove that for any $r\in[0,\infty]$, there is a (data-dependent) tuning parameter r' such that $r\|\boldsymbol{\alpha}\|_2^2$- and $r'\|\boldsymbol{\alpha}\|_2$-regularization are equivalent. Similarly, for any $r\in[0,\infty]$, there is a tuning parameter r' such that $r\|\boldsymbol{\alpha}\|_2$- and $r'\|\boldsymbol{\alpha}\|_2^2$-regularization are equivalent. Hence, swapping ℓ_2- and ℓ_2^2-functions just means changing the tuning parameter.

[4]The elastic net was introduced in [ZH05, Equation (3) on Page 303].

[5]Lemma 2 on Page 306 in [ZH05] supports our earlier statement about the different selection tendencies of the ridge estimator/elastic net and the lasso: on the one hand, every least-squares estimator with *strictly* convex prior term (such as the elastic net with $w\in[0,1)$) assignes the same coordinate estimates $\widehat{\beta}_i=\widehat{\beta}_j$ to two equal predictors $\boldsymbol{x}_i=\boldsymbol{x}_j$; on the other hand, there is always one solution of the lasso estimator that sets one of those coordinate estimates to zero. Thus, the elastic net selects or disregards perfectly correlated predictors as a group, while the lasso selects at most

one representative among such predictors. More generally, Theorem 1 in that paper shows that two coordinate estimates of the elastic net converge as (i) the corresponding predictors become more correlated or (ii) the tuning parameter $(1-w)r$ increases. Further observations in this direction are made in [BW].

[6] Least-squares refitting for the lars, a sibling of the lasso, has been proposed in [EHJT04, Page 421] under the name *lars-ols hybrid*. [Mei07, Definition 1 on Page 376] introduces the *relaxed lasso*: a lasso on the support of an initial lasso, where the tuning parameter for the second stage lasso is chosen smaller than the one for the first stage.

[7] Settings where the standard least-squares refitting described in our text increases the lasso's accuracy are described in [BC$^+$13]; settings where least-square refitting decreases the lasso's accuracy are described in [Led13]. As a rule of thumb, least-squares refitting can render estimations even more accurate in "easy" settings, while in "difficult" settings, it can add considerable error.

[8] Weaker versions of strong oracle properties are *weak oracle properties*, which state that an estimator has (approximately) the same accuracy as the oracle estimator. For linear regression with Gaussian noise, these properties basically refer to oracle inequalities that do not involve the lasso's $\log p$-terms. Still, the conditions for those inequalities to hold can still be strict and are unverifiable in any case.

[9] The scad (smoothly clipped absolute deviation) regularizer was introduced in [FL01, Display (2.7) on Page 1350]; the mcp (minimax concave penalty) in [Zha10, Display (2.2) on Page 897].

[10] As two examples among many such results, [ZZ$^+$12, Theorem 1 on Page 584] and [KCO08, Theorem 3 on Page 1668] provide sufficient conditions for such estimators satisfying weak and strong oracle properties, respectively. However, these conditions are strict; for example, they imply $p \leq n$ in the latter result (see their subsequent remark).

[11] Note that ℓ_q-regularization, $q \in (0,1)$, also renders least-squares estimation non-convex but is otherwise very different from capped ℓ_1-regularization: First, ℓ_q-regularization changes the regularization landscape of the entire parameter *vector*, while the capped ℓ_1-regularizer consists of sums over regularizers for each individual *coordinate*. Second, capped ℓ_1-functions are bounded, while ℓ_q-functions are absolute homogenous, that is, linearly increasing in their function argument; this means in particular that ℓ_q-regularization (without refitting) does not have a strong oracle property. Third, the (sub-)gradients of the capped ℓ_1-function are bounded everywhere, while the gradients of ℓ_q-functions go to infinity as the function argument goes to zero; this divergence at zero can lead to numerical instabilities.

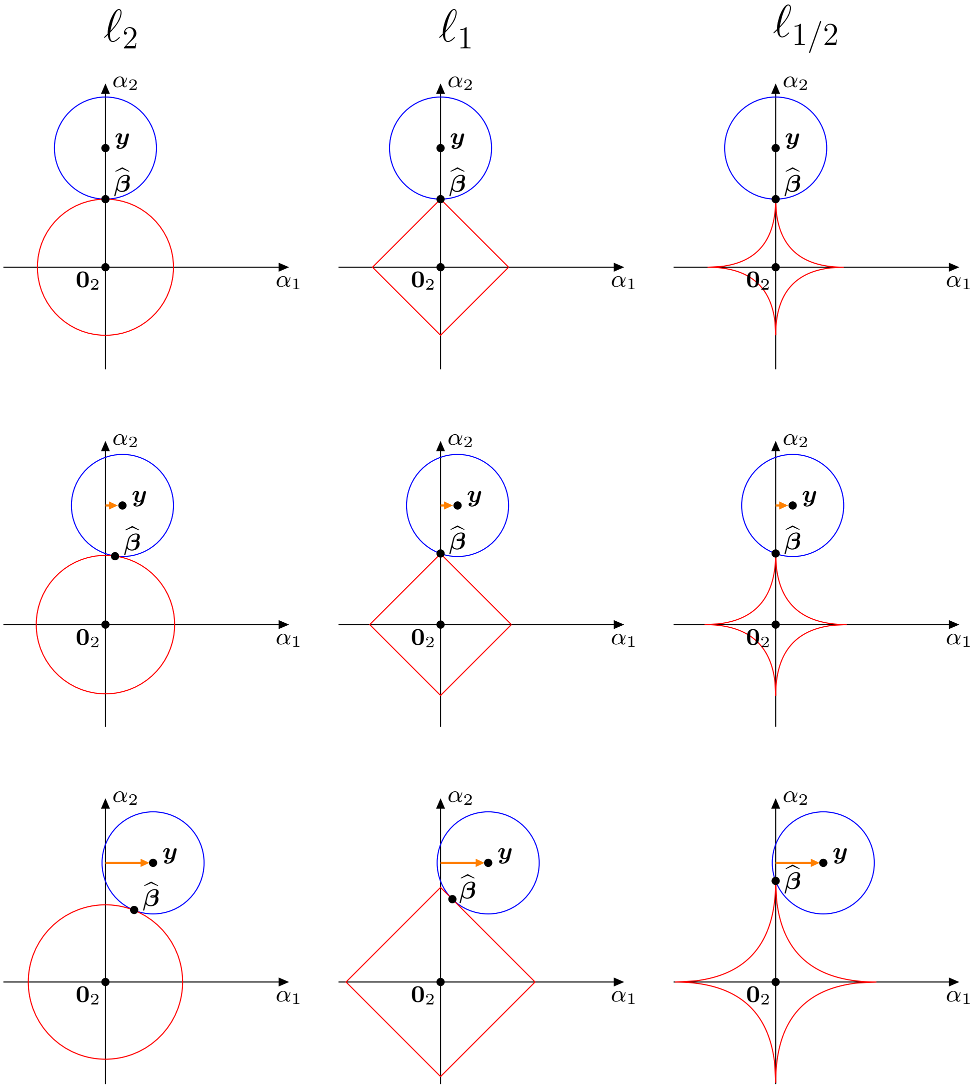


Figure 2.3: Parameter estimation in two dimensions with least-squares data fitting term $\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2$ and ℓ_2- (ridge), ℓ_1- (lasso), and $\ell_{1/2}$-prior function. The design is $X = \mathrm{I}_{2\times 2}$ throughout, while the outcome $\boldsymbol{y} \in \mathbb{R}^2$ is altered from top to bottom. The tuning parameters are chosen such that all estimators' fits are equal: $\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\|_2^2 = a$ for some constant $a \in (0, \infty)$; the blue circles denote the corresponding level sets $\{\boldsymbol{\alpha} \in \mathbb{R}^2 : \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 = a\}$. Since the estimators minimize a weighted sum of the data fitting term and the prior function, they are found where a blue level set of the least-squares function "touches" a level set of the prior function; those latter level sets are drawn in red.
The plots illustrate that the ridge estimator yields zero-valued coordinates only in very special cases (in the first column, $\widehat{\beta}_1 = 0$ only on the top), while zero-valued coordinates are more common for the lasso estimator and especially for the estimator with the $\ell_{1/2}$-prior function ($\widehat{\beta}_1 = 0$ on the top and middle of the second column and across all of the third column). The ℓ_1-regularizer is a sweet spot in that it typically yields sparse solutions (in contrast to ℓ_q with $q > 1$) and, at the same time, makes the objective function amenable to convex optimization (in contrast to ℓ_q with $q < 1$).

2.4 Exercises

Exercises for Section 2.1

□ **Exercise 2.1** $^{\diamond\diamond\bullet}$ In this exercise, we illustrate that additional predictors typically increase the complexity of least-squares estimates. We consider data $(\boldsymbol{y}, X)$ from the linear regression model in Section 2.1 and an augmented version of these data $(\boldsymbol{y}, X')$ with $X' := (X, \boldsymbol{x}) \in \mathbb{R}^{n\times(p+1)}$ for a vector $\boldsymbol{x} \in \mathbb{R}^n$. We denote the corresponding least-squares estimators by $\widehat{\boldsymbol{\beta}}_{\text{ls}}$ and $(\widehat{\boldsymbol{\beta}}_{\text{ls}})'$, respectively. Note that $\widehat{\boldsymbol{\beta}}_{\text{ls}} \in \mathbb{R}^p$, whereas $(\widehat{\boldsymbol{\beta}}_{\text{ls}})' \in \mathbb{R}^{p+1}$.

1. Show that
$$\min_{\boldsymbol{\alpha}'\in\mathbb{R}^{p+1}} \|\boldsymbol{y} - X'\boldsymbol{\alpha}'\|_2^2 \;\leq\; \min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2\,.$$

2. Show that if the additional predictor $\boldsymbol{x}$ and the residual $\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{\text{ls}}$ are not orthogonal, that is, $\langle \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{\text{ls}}, \boldsymbol{x}\rangle \neq 0$, it holds that
$$\min_{\boldsymbol{\alpha}'\in\mathbb{R}^{p+1}} \|\boldsymbol{y} - X'\boldsymbol{\alpha}'\|_2^2 \;<\; \min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2\,.$$

3. Show that the strict inequality in the above display is a sufficient condition for $(\widehat{\boldsymbol{\beta}}_{\text{ls}})'_{p+1} \neq 0$.

4. Show that if $\operatorname{rank}[X] \geq n$, it holds that
$$\min_{\boldsymbol{\alpha}'\in\mathbb{R}^{p+1}} \|\boldsymbol{y} - X'\boldsymbol{\alpha}'\|_2^2 \;=\; \min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2\,.$$

5. Show that $\operatorname{rank}[X] \geq n$ is nevertheless a sufficient condition for the existence of a least-squares solution with $(\widehat{\boldsymbol{\beta}}_{\text{ls}})'_{p+1} \neq 0$.

This proves that both $\langle \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{\text{ls}}, \boldsymbol{x}\rangle \neq 0$ (cf. Claims 2 and 3) and $\operatorname{rank}[X] \geq n$ (cf. Claim 5) are sufficient conditions for a least-squares estimator assigning non-zero weight to the additional predictor $\boldsymbol{x}$. Since these conditions seem likely to be satisfied in high-dimensional settings, our results suggest that least-squares estimation typically uses every additionally provided predictor—irrespective of whether that predictor is actually relevant or not.

Exercises for Section 2.2

□ **Exercise 2.2** $^{\diamond\diamond\bullet}$ In this exercise, we compute the lasso for orthonormal design, that is, we compute the solutions to (2.1) for data $(\boldsymbol{y}, X)$ that satisfies $X^\top X = \mathrm{I}_{p\times p}$.

Prove the following three results under the assumption of orthonormality.

1. Show that
$$\widehat{\boldsymbol{\beta}}_{\text{lasso}} \;=\; X^\top\boldsymbol{y} - \frac{r}{2}\widehat{\boldsymbol{\kappa}}$$
for a vector $\widehat{\boldsymbol{\kappa}} \in \boldsymbol{\partial}\|\widehat{\boldsymbol{\beta}}_{\text{lasso}}\|_1$.

2. Use 1. to show that for all $j \in \{1,\dots,p\}$, it holds that
$$(\widehat{\beta}_{\text{lasso}})_j \;=\; 0 \quad\Leftrightarrow\quad \bigl|(X^\top\boldsymbol{y})_j\bigr| \;\leq\; \frac{r}{2}\,.$$

3. Use 1. and 2. to show that for all $j \in \{1, \dots, p\}$, it holds that

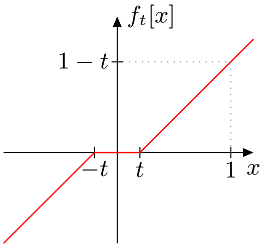
$$(\widehat{\beta}_{\text{lasso}})_j \;=\; \operatorname{sign}\bigl[(X^\top \boldsymbol{y})_j\bigr]\Bigl(\bigl|(X^\top \boldsymbol{y})_j\bigr| - \frac{r}{2}\Bigr)_+ ,$$

   where the *positive part* $(a)_+$ of a number $a \in \mathbb{R}$ is defined as $(a)_+ := a$ if $a > 0$ and $(a)_+ := 0$ otherwise.

We conclude that for orthonormal design, the coordinates of the lasso estimator are $(\widehat{\beta}_{\text{lasso}})_j = f_{r/2}[(X^\top \boldsymbol{y})_j]$, $j \in \{1, \dots, p\}$, where f_t is the *soft-thresholding operator*

$$\begin{aligned} f_t \;\to\; \mathbb{R} &\;\mapsto\; \mathbb{R}\\ x \;:\; &\operatorname{sign}[x]\bigl(|x| - t\bigr)_+ \end{aligned}$$

for a given threshold $t \in [0, \infty]$. Hence, in the orthonormal case, the lasso estimator is a soft-thresholded version of the least-squares estimator $\widehat{\boldsymbol{\beta}}_{\text{ls}} = (X^\top X)^{-1} X^\top \boldsymbol{y} = X^\top \boldsymbol{y}$.



4. *Bonus:* Can you generalize the results to the orthogonal case, where $X^\top X = D$ for a positive definite, diagonal matrix $D \in \mathbb{R}^{p \times p}$?

□ **Exercise 2.3** $^{\diamond\diamond\bullet}$ In this exercise, we derive further properties of the lasso estimator

$$\widehat{\boldsymbol{\beta}}[r] \;\in\; \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^p} \bigl\{ \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_1 \bigr\}$$

for certain tuning parameters r.

First, we establish another relationship between the lasso and the least-squares estimator.

1. Show that if the Gram matrix $X^\top X$ is invertible and $r \in [0, \infty)$, it holds that

$$\widehat{\boldsymbol{\beta}}[r] \;=\; \widehat{\boldsymbol{\beta}}_{\text{ls}} - \frac{r}{2}(X^\top X)^{-1}\widehat{\boldsymbol{\kappa}}$$

   for a $\widehat{\boldsymbol{\kappa}} \in \boldsymbol{\partial}\|\widehat{\boldsymbol{\beta}}[r]\|_1$.

2. Show that then

$$\|X\widehat{\boldsymbol{\beta}}[r] - X\widehat{\boldsymbol{\beta}}_{\text{ls}}\|_2^2 \;\le\; \frac{r^2 p}{4e_1},$$

   where e_1 is the smallest eigenvalue of the Gram matrix $X^\top X$.

In view of Equation (1.4), this means that if the lasso's tuning parameter is sufficiently small, say $r \ll 2\sqrt{e_1}\sigma$, the lasso and the least-squares estimator predict about equally well.

Second, we identify the range of tuning parameters that set the lasso to zero.

3. Show that $\widehat{\boldsymbol{\beta}}[r] = \mathbf{0}_p$ is a lasso solution if and only if $r \ge 2\|X^\top \boldsymbol{y}\|_\infty$.

4. Show that $\widehat{\boldsymbol{\beta}}[r] = \mathbf{0}_p$ is the *unique* solution if $r \ge 2\|X^\top \boldsymbol{y}\|_\infty$ *and* $r > 0$.

5. If $r = 2\|X^\top \boldsymbol{y}\|_\infty = 0$, the vector $\widehat{\boldsymbol{\beta}}[r]$ is a solution if and only if $X\widehat{\boldsymbol{\beta}}[r] = \mathbf{0}_n$.

This implies in particular that in practice, it is usually sufficient to consider tuning parameters smaller or equal to $2\|X^\top \boldsymbol{y}\|_\infty$.

R Lab Chapter 2

2 Overfitting

In this lab, we illustrate the overfitting phenomenon. For this, we generate data according to a linear regression model with standard polynomial basis functions as predictors. The parameters are set to zero except for the quadratic term. We then compare the least-squares and the lasso in estimating the corresponding curve from the noisy data. We do not assume known that the underlying function is quadratic beforehand; instead, we only assume known that the function is a polynomial, and we assume given an upper bound for the degree of that polynomial. In addition, we assume known that the model is sparse, that is, that only a small (but otherwise unknown) number of polynomial terms is relevant, which motivates the application of ℓ_1-regularization. We find that an increase in the number of such predictors can get least-squares estimation completely off course, while the lasso remains largely unimpressed.

As always, your task is to replace the keyword `REPLACE` with suitable code and to answer the questions posed in the text.

2.1 Generating data

We consider data from the linear regression model

$$y_i \;=\; \beta_0 + \beta_1 x_i + \beta_2 (x_i)^2 + \cdots + \beta_d (x_i)^d + u_i \qquad\qquad (\, i \in \{1, \ldots, n\} \,)$$

with regression vector $\boldsymbol{\beta} = (0, 0, 2, 0, \ldots, 0)^\top \in \mathbb{R}^{d+1}$ and independently distributed measurements $x_i \sim \mathrm{Unif}[-1, 1]$ and noise $u_i \sim \mathcal{N}[0, 1]$. This means that the outcomes y_i are noisy observations of the quadratic function $x \mapsto 2x^2$, but since $\boldsymbol{\beta}$ is unknown, the data analyst does not know this relationship beforehand.

Set the degree of the polynomial to $d = 20$ and generate $n = 100$ independent samples according to the above model. The function `runif()` might be helpful to sample from the uniform distribution, and the function `rnorm()` to sample from the normal distribution. Summarize the outcomes in a vector $y := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ and the predictors in a matrix $X \in \mathbb{R}^{n \times (d+1)}$ with coordinates $X_{ij} = (x_i)^j$. Including the intercept, the number of predictors is $p = d + 1$.

```r
set.seed(87)
PolynomialDesign <- function(x.vector, d)
{
  REPLACE
}
n <- 100; d <- 20
X <- PolynomialDesign(REPLACE, d)
beta <- c(0, 0, 2, rep(0, d - 2))
y <- REPLACE
cbind(y, X)[1:4, 1:5]
```

```
##          [,1] [,2]       [,3]      [,4]        [,5]
## [1,] 1.319628    1 -0.9678621 0.9367571 -0.90665165
## [2,] 1.839104    1 -0.4171889 0.1740466 -0.07261032
## [3,] 3.089433    1 -0.9150774 0.8373667 -0.76625536
## [4,] 1.657062    1  0.7153142 0.5116745  0.36600803
```

2.2 Implementing the estimators

Implement a least-squares estimator and a lasso estimator. For the latter, use the `cv.glmnet()` function from the `glmnet` package with the flag `intercept=FALSE`. The `coef()` function might be helpful in extracting the estimated regression vector from the `glmnet` object (note that the first coordinate of the resulting vector needs to be removed in our case, since that coordinate contains a placeholder for an additional intercept).

```
library(glmnet)
set.seed(98)  # glmnet uses randomized cross-validation routines
LsEstimator <- function(y, X)
{
  return(REPLACE)
}
LassoEstimator <- function(y, X)
{
  return(REPLACE)
}
cbind(LsEstimator(y, X[, 1:3]), LassoEstimator(y, X[, 1:3])) # a quick check
```

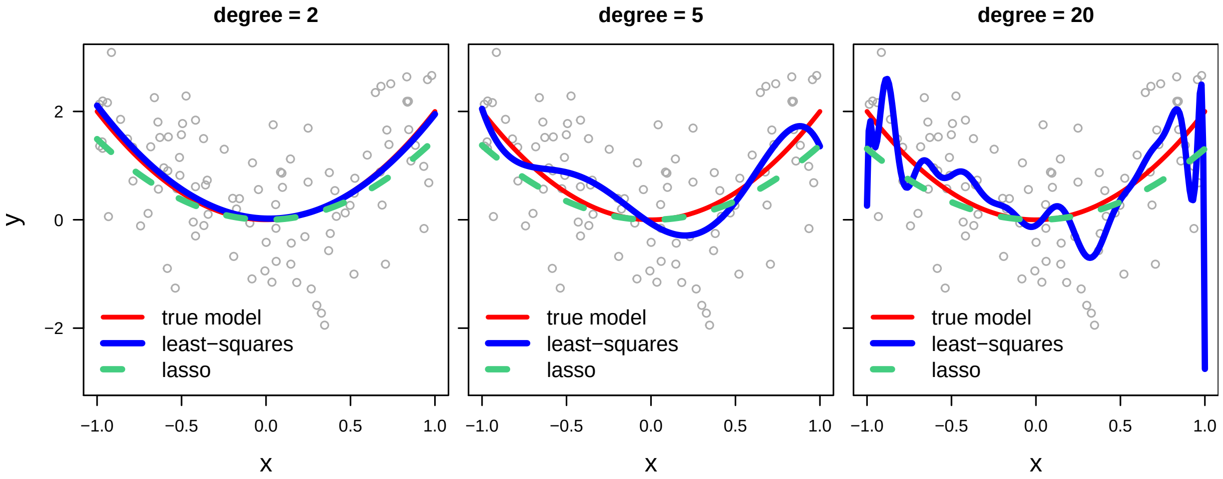
```
##             [,1]     [,2]
## [1,]  0.02221062 0.000000
## [2,] -0.07859756 0.000000
## [3,]  2.00685799 1.376057
```

A technical detail: We set `intercept=FALSE`, because otherwise, `glmnet` would estimate an additional *unregularized* intercept. However, `glmnet` also disregards the first column of X (cf. Line 2156 in the file `glmnet5.f90` on https://github.com/cran/glmnet; this line is executed even if `standardize=FALSE`). Hence, strictly speaking, our lasso implementation has a tiny advantage over the least-squares, as it is not tempted to fit an intercept. As a BONUS, you can explore the subleties.

2.3 Computing and visualizing the results

We now compute and visualize the results for varying degree $d \in \{2, 5, 20\}$. The estimators are fed with (subsets of) the above generated data: the outputs y are the values stored in `y`, and the designs X are the first $d+1$ columns of the values stored in `X`. The estimated functions are then evaluated on a fine grid called `x.plot` to obtain smooth graphs.

```
FunctionOutputs <- function(FUN, d, x.plot)
{
  return(REPLACE)
}
PlotLsLasso <- function(d, first.plot)
{
  x.plot <- seq(from=-1, to=1, by=0.01)
  if (first.plot == TRUE) ylab <- "y" else ylab <- ""  # y label only for first plot
  plot(x       = X[, 2],
       y       = y,  # samples
       xlim    = c(-1, 1),
       ylim    = c(-3, 3),
       col     = "gray68",
       xlab    = "x",
       cex.lab = 1.5,
       yaxt    = "n",
       ylab    = ylab,
       main    = paste0("degree = ", d))
  axis(side   = 2,
       labels = first.plot,
       las    = 1,
       yaxp   = c(-2, 2, 2))
  lines(x   = x.plot,
        y   = 2*x.plot^2,  # true model
        lwd = 3,
        col = "red")
  lines(x   = x.plot,
        y   = FunctionOutputs(FUN=LsEstimator, d, x.plot),  # least-squares estimates
        lwd = 4,
        col = "blue")
  lines(x   = x.plot,
        y   = FunctionOutputs(FUN=LassoEstimator, d, x.plot),  # lasso estimates
        lty = 2,
        lwd = 4,
        col = "seagreen3")
  legend("bottomleft",
         legend = c("true model", "least-squares", "lasso"),
         lty    = c(1, 1, 2),
         lwd    = c(3, 4, 4),
         col    = c("red", "blue", "seagreen3"),
         bty    = "n",  # no box
         cex    = 1.3)
}
par(mfrow=c(1, 3),  # three plots side-by-side
    oma = c(3, 3, 0, 0),  # outer margins
    mar = c(3, 1, 3, 0),  # inner margins
    mgp = c(3, 1, 0),  # location of the labels
    xpd = NA)  # allows the plots to overlap
PlotLsLasso(2, TRUE)
PlotLsLasso(5, FALSE)
PlotLsLasso(20, FALSE)
```

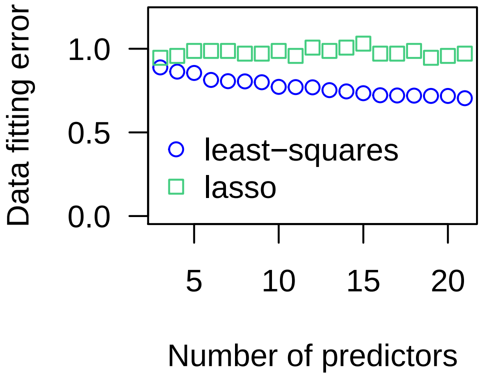


What can you conclude from the plots?

2.4 Further illustrations

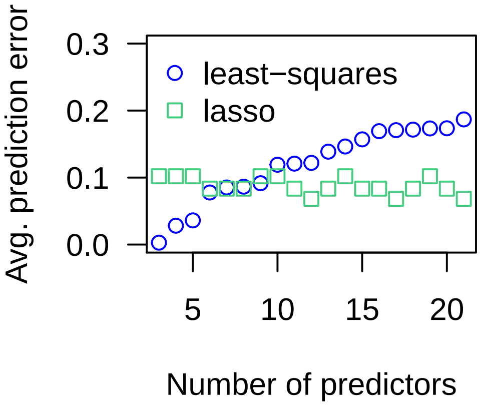
We now provide further illustrations of the above overfitting phenomenon. For this, we compare the least-squares and the lasso in four types of metrics: (i) data fitting error $\|\mathbf{y} - X\widehat{\boldsymbol{\beta}}\|_2^2/n$, (ii) average prediction error $\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2/n$, (iii) estimation error $\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_2$, and (iv) number of non-zero coordinates $\|\widehat{\boldsymbol{\beta}}\|_0$—all as functions of the number of predictors/the degree of the assumed polynomial.

```
set.seed(68)
Metrics <- function(FUN, p.max, type)
{
  REPLACE
}
plot(x       = 3:(d + 1),  # starting from degree two = three predictors
     y       = Metrics(LsEstimator, d + 1, "fitting"),
     pch     = 1,
     ylim    = c(0, 1.2),
     yaxp    = c(0, 1, 2),
     col     = "blue",
     xlab    = "Number of predictors",
     ylab    = "Data fitting error",
     las     = 1)
points(x   = 3:(d + 1),
       y   = Metrics(LassoEstimator, d + 1, "fitting"),
       pch = 0,
       col = "seagreen3")
legend("bottomleft",
       legend = c("least-squares", "lasso"),
       pch    = c(1, 0),
       lty    = 0,
       col    = c("blue", "seagreen3"),
       bty    = "n")  # no box
```



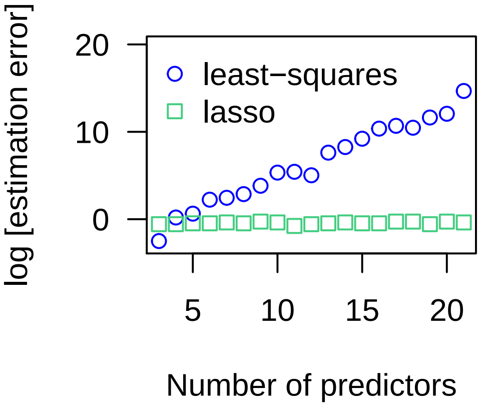
While for the least-squares, more predictors mean a better fit, the cross-validated lasso does not follow such a trend.

```
REPLACE  # generate the second plot
```



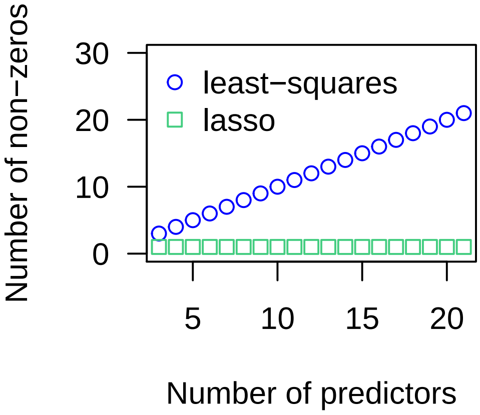
The prediction error of the least-squares increases as more predictors become available, while again, the lasso remains relatively stable. This commensurates with the results of Section 1.2, which state a linear increase for the least-squares and only a logarithmic increase for the lasso. The data fitting error above and the prediction error here are closely related to training and validation errors, which we will discuss in the context of cross-validation in Chapter 4.

```
REPLACE  # generate the third plot
```



The estimation errors have even stronger trends than the prediction errors (note the log-scaling of the y-axis).

```
REPLACE  # generate the fourth plot
```



The lasso estimator always picks only one predictor, while the least-squares estimator picks all available ones. Ambiguities among those predictors can make the least-squares numerically and statistically instable. Since the polynomial layout generates strong dependencies among the predictors, these instabilities are expressed particularly strongly in the estimation errors.

Chapter 3

Graphical Models

3.1 Overview

Data-generating mechanisms often involve multiple interdependent components. For example, neural response to sensory stimuli is produced by neuronal networks in the brain; phenotypic modulations are generated by mutations and methylations of intercorrelated groups of genes; and microbial communities in the gut and other human habitats are shaped by the rich interplay among different species. Scientists believe that mathematical models for such network structures can help to unravel the underpinning physical, mechanical, and biological principles.

A range of mathematical frameworks are known, among them *differential equations, Boolean networks,* and *graphical models*[1]. In recent years, the latter has become the dominant approach in a number of fields, because it summarized dependence structures in comprehensive, yet concise graphical representations that lend themselves naturally to scientific interpretation.[2] In this section, we focus on *undirected*, *probabilistic* graphical models, also called *Markov random fields.* The word "probabilistic" indicates that probability and measure theory are our mathematical achorage. The specific notion we base our considerations on is *conditional dependence,* which can allow for very finely detailed descriptions of the underlying relationships. As an illustration for this granularity, consider the two events `bike`="you bike to work" and `dog`="neighbor walks dog." The events `bike` and `dog` are not independent, as sunshine might make you feel tempted to use your bike as well as it might motivate the neighbor to go outside and walk the dog:

$$\texttt{bike} \not\perp \texttt{dog} \qquad \text{“}\texttt{bike}\text{ is not independent of }\texttt{dog}\text{”}\,.$$

A graphical representation is given in the left panel of Figure 3.1. However, the events `bike` and `dog` are not congruent either, as you might ride the bike even in hard rain in guilt over a big meal the night before, which on the other hand, has no influence on the neighbor's inclination to walk her dog. Stating the above dependence relation alone thus conceils that there is one common trigger for both events (sunshine) as well as independent triggers for each of the two events (your big meal the night before, ...). We can unravel this structure by adding the third event `sun`="sun is shining:" the events `bike` and `dog` are dependent, but they are independent *after adjusting* for the influence of `sun`. We call the latter statement *conditional independence*, indicated in

conditional independence

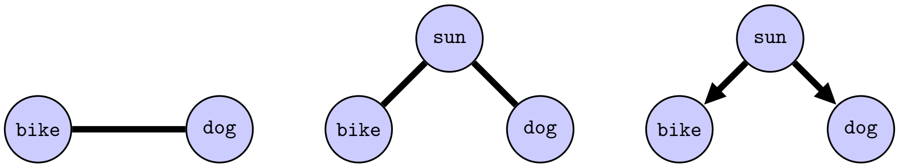


Figure 3.1: The graph on the left-hand side depicts that `bike` and `dog` are dependent. The graph in the middle captures the more granular statement that `bike` and `dog` are dependent but (conditionally) independent given `sun`. The graph on the right-hand side illustrates that `sun` causes changes in `bike` and `dog`.

speach by adding the word "given" together with the variable we adjust for:

$$\texttt{bike} \perp \texttt{dog} \,|\, \texttt{sun} \qquad \text{"}\texttt{bike} \text{ is independent of } \texttt{dog} \text{ given } \texttt{sun}\text{"}\,.$$

A graphical representation that summarizes both types of dependence is given in the middle panel of Figure 3.1.

There is a clear hierarchy in the example: sunshine makes us taking the bike more likely and also increases the chance that the neighbor walks his dog; on the contrary, as much as we might like that, neither taking the bike nor walking the dog influences the weather. There is, therefore, a *causal* relationship between `bike`, `dog`, and `sun`. Such relationships can be represented by *directed* graphs such as the one in right panel of Figure 3.1. In general, however, causal relationships are much less clear, and they are associated with a number of statistical and philosophical challenges. To avoid those difficulties, we restrict ourselves to undirected graphical models that formalize cooccurrence and avoid causal statements.

We are interested in conditional dependence relations as well as the magnitudes of the corresponding dependencies. While in some applications, (some of) the dependence relations might be provided by domain experts, previous studies, or common sense, we are interested in the more typical case in which both the dependence relations and the magnitudes need to be inferred from data.

The word "probabilistic" used earlier means that we think of the data as representations of random quantities, and therefore, base our mathematical theories on measure theory. Precisely, we model observations as independent and identically distributed copies $\boldsymbol{z}^1, \dots, \boldsymbol{z}^n$ of a random vector $\boldsymbol{z} \in \mathbb{R}^p$. Each coordinate of $\boldsymbol{z}$ could correspond to the activity of a neuron, to the expression of a gene, to the abundance of a microbial species, or simply

$$\boldsymbol{z} \;=\; \begin{pmatrix} \texttt{bike} \\ \texttt{dog} \\ \texttt{sun} \end{pmatrix}.$$

Each observation is a snapshot of the network's state. The data is no longer of regression-type: the observations consist only of the vector $\boldsymbol{z}$ in place of the pair $(\boldsymbol{y}, X)$ in regression. When discussing the neighborhood selection scheme later, we will see that $\boldsymbol{z}$ plays the roles of both $\boldsymbol{y}$ and X simultenously. Note that here, superscripts denote the indexes of an observation ($z^i \in \mathbb{R}^p$), subscripts denote the coordinate of a vector ($z_i \in \mathbb{R}$). Of course, we also account for high dimensionality, allowing the number of dependence relations under investigation (usually of the order of p^2) to be comparable or even larger than the number of observations n.

The conditional dependence relations among the coordinates of $\boldsymbol{z}$ are captured by a graph $\mathcal{G} = (\mathcal{I}, \mathcal{E})$. The nodes $\mathcal{I} = \{1, \dots, p\}$ correspond to the index set of the random vector $\boldsymbol{z}$ and represent the coordinates of the random vector under consideration; in our toy example, the nodes $\mathcal{I} = \{1,2,3\}$ can be identified with $\texttt{bike}, \texttt{dog}, \texttt{sun}$. The edge set $\mathcal{E} \subset \{(i,j) \in \mathcal{I} \times \mathcal{I}\}$ describes the conditional relations: $(i,j) \in \mathcal{E}$ if and only if $i \neq j$ and z_i and z_j are not conditionally independent given all other coordinates of $\boldsymbol{z}$. By construction, the edge set is symmetric, that is, $(i,j) \in \mathcal{E}$ if and only if $(j,i) \in \mathcal{E}$. The pair $(\boldsymbol{z}, \mathcal{G})$ of (i) the random vector $\boldsymbol{z} = (\mathbb{1}\{\texttt{bike}\}, \mathbb{1}\{\texttt{dog}\}, \mathbb{1}\{\texttt{sun}\})^\top$ that captures the observations and (ii) the corresponding graph $\mathcal{G}$ of the conditional dependencies is finally called a *graphical model*.

graphical model

3.2 A Measure-Theoretical Definition*

For those readers that are experienced in measure theory, we provide a precise mathematical definition of graphical models. To this end, consider a probability space $(\mathcal{A}, \mathfrak{A}, \mathbb{P})$, a measurable space $(\mathcal{B}, \mathfrak{B})$, and measurable functions (random variables/vectors)

$$U, V, W \quad : \quad (\mathcal{A}, \mathfrak{A}, \mathbb{P}) \;\to\; (\mathcal{B}, \mathfrak{B})$$

between the two spaces. Let $\mathfrak{A}_U$, $\mathfrak{A}_V$, and $\mathfrak{A}_W$ be the generating σ-algebras of U, V, W, respectively, that is, the smallest σ-algebras over $\mathcal{A}$ that contain $\{U^{-1}[B] : B \in \mathfrak{B}\}$, $\{V^{-1}[B] : B \in \mathfrak{B}\}$, $\{W^{-1}[B] : B \in \mathfrak{B}\}$, respectively. We say that U and V are independent if $\mathfrak{A}_U$ and $\mathfrak{A}_V$ are *independent*. In mathematical notation,

$$U \perp V \quad \Leftrightarrow \quad \mathbb{P}\{A \cap A'\} \;=\; \mathbb{P}\{A\} \cdot \mathbb{P}\{A'\} \;\text{ for all } A \in \mathfrak{A}_U, A' \in \mathfrak{A}_V\,.$$

Of course, we say that U and V are *dependent* if they are not independent, and we then write $U \not\perp V$.

Now, for conditional dependence, the third function comes into play. We say that U, V are *conditionally independent* given W if $\mathfrak{A}_U$ and $\mathfrak{A}_V$ are independent given $\mathfrak{A}_W$. In mathematical notation,

$$U \perp V \,|\, W \quad \Leftrightarrow \quad \mathbb{P}\{A \cap A' | \mathfrak{A}_W\} \;=\; \mathbb{P}\{A | \mathfrak{A}_W\} \cdot \mathbb{P}\{A' | \mathfrak{A}_W\} \;\text{ for all } A \in \mathfrak{A}_U, A' \in \mathfrak{A}_V\,.$$

Recall that $\mathbb{P}\{A|\mathfrak{A}_W\}, \mathbb{P}\{A'|\mathfrak{A}_W\}$ are *random variables* that can be defined via conditional expectations[3] over indicator functions. Again, we call U, V *conditionally dependent* given W if $\mathfrak{A}_U$ and $\mathfrak{A}_V$ are not independent given $\mathfrak{A}_W$, and we write $U \not\perp V \mid W$.

Independence and conditional independence do not imply each other. As a simple counter example, consider two independent Bernoulli random variables $U, V \in \{0,1\}$, $\mathbb{P}\{U = 0\} = \mathbb{P}\{U = 1\} = \mathbb{P}\{V\} = 0\} = \mathbb{P}\{V = 1\} = 1/2$ and the indicator function $W := \mathbb{1}\{U = V\}$. The variables U and V are independent by definition; however,

$$\begin{aligned}
&\mathbb{P}\{\{U = 0\} \cap \{V = 0\} \,|\, \{W = 1\}\} \\
&= \frac{\mathbb{P}\{\{U = 0\} \cap \{V = 0\} \cap \{W = 1\}\}}{\mathbb{P}\{W = 1\}} && \text{“definition of conditional probabilities”} \\
&= \frac{1/4}{1/2} \;=\; \frac{1}{2}\,, && \text{“U, V Bernoulli random variables”}
\end{aligned}$$

while

$$\begin{aligned}
&\mathbb{P}\{\{U=0\}\,|\,\{W=1\}\}\cdot\mathbb{P}\{\{V=0\}\,|\,\{W=1\}\}\\
&=\frac{\mathbb{P}\{\{U=0\}\cap\{W=1\}\}}{\mathbb{P}\{W=1\}}\cdot\frac{\mathbb{P}\{\{V=0\}\cap\{W=1\}\}}{\mathbb{P}\{W=1\}} && \text{“definition of conditional probabilities”}\\
&=\frac{1/4}{1/2}\cdot\frac{1/4}{1/2}\;=\;\frac{1}{4}\,. && \text{“U,V Bernoulli random variables”}
\end{aligned}$$

Hence, independence does not imply conditional independence.

Similarly, consider a Bernoulli random variable U and set $V := W := U$. Then, one can check that U is conditionally independent of V given W, while of course, U and V are not independent. Hence, also conditional independence does not imply independence.

This means that we have to decide whether to proceed with regular or conditional dependences. Motivated by the previous section, we opt for the latter.

Definition 3.2.1 (Graphical Models)

Let $\boldsymbol{z}\in\mathbb{R}^p$ be a random vector and $\mathcal{G}=(\mathcal{I},\mathcal{E})$ a graph with node set $\mathcal{I}=\{1,\dots,p\}$ and edget set $\mathcal{E}\subset\mathcal{I}\times\mathcal{I}$. We call the pair $(\boldsymbol{z},\mathcal{G})$ an (undirected) graphical model if for each $i,j\in\{1,\dots,p\}$, $i\neq j$,

$$z_i\;\perp\;z_j\mid\boldsymbol{z}_{\{1,\dots,p\}\setminus\{i,j\}}\qquad\Leftrightarrow\qquad(i,j)\notin\mathcal{E}\,.$$

Graphical models combine notions of probability theory and graph theory: the conditional dependence structure of a random variable is described by the edge set of a graph. Specifically, two coordinates z_i and z_j are conditionally independent given all other coordinates if and only if (i,j) is not an edge in the graph.

3.3 Gaussian Graphical Models

Gaussian graphical models are the most popular class of graphical models. By definition, their random vectors follow a centered, multivariate normal distribution, and it turns out that their conditional and unconditional dependence structures are determined by the inverse of the covariance matrix of that distribution. A main feature of Gaussian graphical models is that their dependence graphs are amenable to established methodology such as maximum (regularized) likelihood—we will show this in the subsequent sections.

In mathematical terms, a Gaussian graphical model is a pair $(\boldsymbol{z},\mathcal{G})$ that consists of a normal vector $\boldsymbol{z}\sim\mathcal{N}_p[\mathbf{0}_p,\Sigma]$ for a symmetric and invertible *covariance matrix* $\Sigma\in\mathbb{R}^{p\times p}$ and of a corresponding conditional independence graph $\mathcal{G}$. In such models, the graph $\mathcal{G}$ and the inverse of the covariance matrix, the *precision matrix* $\Theta:=\Sigma^{-1}$, are intimately related:

covariance and precision matrices

Theorem 3.3.1 (Hammersley-Clifford, Part I)

For all indexes $i,j\in\{1,\dots,p\}$, $i\neq j$, it holds that

$$z_i\perp z_j\mid\boldsymbol{z}_{\{1,\dots,p\}\setminus\{i,j\}}\qquad\Leftrightarrow\qquad\Theta_{ij}\;=\;0\,,$$

$$\Theta \;=\; \begin{bmatrix} 1.2 & -0.2 & 0 & 0 \\ -0.2 & 1.5 & 0 & 0.3 \\ 0 & 0 & 1 & 0 \\ 0 & 0.3 & 0 & 0.5 \end{bmatrix}$$

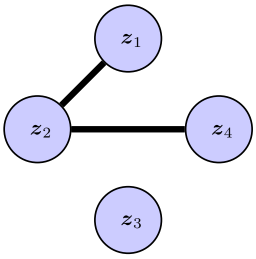


Figure 3.2: The right panel contains an example precision Θ for a Gaussian vector $\boldsymbol{z}$. Since $\Theta_{14} = 0$ but $\Theta_{12}\Theta_{24} \neq 0$, the coordinates z_1 and z_4 are conditionally independent given the other coordinates but not unconditionally independent. This is illustrated also in the corresponding dependence graph $\mathcal{G}$ on the right: $(1,4) \notin \mathcal{E}$ but $(1,2),(2,4) \in \mathcal{E}$.

that is, $\mathcal{G} = (\mathcal{I},\mathcal{E})$ with node set $\mathcal{I} = \{1,\dots,p\}$ and edge set $\mathcal{E} = \{(i,j) \in \mathcal{I}\times\mathcal{I} : i \neq j, \Theta_{ij} \neq 0\}$.

The same graph $\mathcal{G}$ also describes the unconditional dependencies among the coordinates of $\boldsymbol{z}$:

Theorem 3.3.2 (Hammersley-Clifford, Part II)

For all indexes $i,j \in \{1,\dots,p\}$, $i \neq j$, it holds that

$$z_i \;\perp\; z_j \quad\Leftrightarrow \\ \nexists\, k \in \{1,2,\dots\},\, l_1,\dots,l_k \in \{1,\dots,p\} \;:\; \Theta_{il_1}\Theta_{l_1l_2}\dots\Theta_{l_kj} \;\neq\; 0\,,$$

that is, two coordinates z_i and z_j are independent if and only if there is no path of edges $(i,l_1),(l_1,l_2),\dots,(l_k,j) \in \mathcal{E}$.

(We omit the proofs.[4]) For example, if $\Theta_{ij} \neq 0$, that is, $(i,j) \in \mathcal{E}$, then z_i and z_j are neither conditional independent (given any set of coordinates) nor unconditionally independent.

In summary, the non-zero pattern of the inverse covariance matrix captures the entire dependence structure of the Gaussian vector $\boldsymbol{z}$; Figure 3.2 contains an illustration. This makes this matrix our target for estimation in the following sections.

3.4 Maximum Regularized Likelihood Estimation

We now turn to parameter estimation for Gaussian graphical models. The parameters of interest are the inverse covariance matrix and the underpinning conditional dependence graph. As data, we assume independent realizations of the random vector in question. A natural approach to parameter estimation is then maximum likelihood. Its basic version is used for the estimation of low-dimensional parameters; for high-dimensional settings, the log-likelihood is complemented with a prior function.

The density of a Gaussian vector $\boldsymbol{z} \sim \mathcal{N}_p[\mathbf{0}_p, \Omega^{-1}]$ is

$$f[\boldsymbol{z}] \;=\; \frac{1}{\sqrt{(2\pi)^p \det[\Omega^{-1}]}}\, e^{-\boldsymbol{z}^\top\Omega\boldsymbol{z}/2} \qquad (\boldsymbol{z} \in \mathbb{R}^p)\,.$$

The maximum likelihood estimator of Θ over $\mathcal{S}_p^+$, the set of symmetric and positive definite matrices in $\mathbb{R}^{p\times p}$, based on n independent realizations $\boldsymbol{z}^1,\dots,\boldsymbol{z}^n$ of $\boldsymbol{z}$ is therefore

$$
\begin{aligned}
\widehat{\Theta}_{\mathrm{ml}} &\in \operatorname*{argmax}_{\Omega\in\mathcal{S}_p^+}\prod_{i=1}^n\frac{1}{\sqrt{(2\pi)^p\det[\Omega^{-1}]}}\;e^{-{\boldsymbol{z}^i}^\top\Omega\boldsymbol{z}^i/2}\\
&= \operatorname*{argmax}_{\Omega\in\mathcal{S}_p^+}\log\left[\prod_{i=1}^n\frac{1}{\sqrt{(2\pi)^p\det[\Omega^{-1}]}}\;e^{-{\boldsymbol{z}^i}^\top\Omega\boldsymbol{z}^i/2}\right] && \text{“log is strictly increasing”}\\
&= \operatorname*{argmax}_{\Omega\in\mathcal{S}_p^+}\sum_{i=1}^n\bigl(-p\log[2\pi]/2+\log\bigl[\det[\Omega]\bigr]/2-{\boldsymbol{z}^i}^\top\Omega\boldsymbol{z}^i/2\bigr)\\
&\qquad\qquad\text{“}\log[ab]=\log[a]+\log[b];\ \log[a^b]=b\log[a];\ \det[A]=1/\det[A^{-1}]\text{”}\\
&= \operatorname*{argmax}_{\Omega\in\mathcal{S}_p^+}\frac{1}{n}\sum_{i=1}^n\bigl(\log\bigl[\det[\Omega]\bigr]-{\boldsymbol{z}^i}^\top\Omega\boldsymbol{z}^i\bigr)\\
&\qquad\qquad\text{“}\max_x a(b+f[x])=\max_x f[x]\text{ for }a\in(0,\infty),b\in\mathbb{R}\text{ constant”}\\
&= \operatorname*{argmin}_{\Omega\in\mathcal{S}_p^+}\frac{1}{n}\sum_{i=1}^n\bigl({\boldsymbol{z}^i}^\top\Omega\boldsymbol{z}^i-\log\bigl[\det[\Omega]\bigr]\bigr) && \text{“}\operatorname{argmax}_x -f[x]=\operatorname{argmin}_x f[x]\text{”}\\
&= \operatorname*{argmin}_{\Omega\in\mathcal{S}_p^+}\left\{\frac{1}{n}\sum_{i=1}^n{\boldsymbol{z}^i}^\top\Omega\boldsymbol{z}^i-\log\bigl[\det[\Omega]\bigr]\right\} && \text{“linearity of sums”}\\
&= \operatorname*{argmin}_{\Omega\in\mathcal{S}_p^+}\left\{\frac{1}{n}\sum_{i=1}^n\operatorname{tr}[{\boldsymbol{z}^i}^\top\Omega\boldsymbol{z}^i]-\log\bigl[\det[\Omega]\bigr]\right\} && \text{“}a=\operatorname{tr}[a]\text{ for }a\in\mathbb{R}\text{”}\\
&= \operatorname*{argmin}_{\Omega\in\mathcal{S}_p^+}\left\{\frac{1}{n}\sum_{i=1}^n\operatorname{tr}[\boldsymbol{z}^i{\boldsymbol{z}^i}^\top\Omega]-\log\bigl[\det[\Omega]\bigr]\right\} && \text{“}\operatorname{tr}[AB]=\operatorname{tr}[BA]\text{”}\\
&= \operatorname*{argmin}_{\Omega\in\mathcal{S}_p^+}\left\{\operatorname{tr}\left[\frac{1}{n}\sum_{i=1}^n\boldsymbol{z}^i{\boldsymbol{z}^i}^\top\Omega\right]-\log\bigl[\det[\Omega]\bigr]\right\}. && \text{“linearity of trace”}
\end{aligned}
$$

The maximum likelihood estimator can be effective in estimating the inverse covariance matrix in low-dimensional settings, where $n\gg p$. In particular, if $n\geq p$, the maximum likelihood estimator exists, is unique, and has the closed form

$$
\widehat{\Theta}_{\mathrm{ml}} \;=\; \left(\frac{1}{n}\sum_{i=1}^n\boldsymbol{z}^i{\boldsymbol{z}^i}^\top\right)^{-1}, \tag{3.1}
$$

all with probability one—see Exercise 3.3.

In high-dimensional settings, the likelihood is complemented with a prior function. Given such a function $h\ :\ \mathcal{S}_p^+\to\mathbb{R}$ and a tuning parameter $r\in[0,\infty]$, this yields the *maximum regularized likelihood estimator*

$$
\widehat{\Theta}_{\mathrm{mrl}}\in\operatorname*{argmin}_{\Omega\in\mathcal{S}_p^+}\left\{\operatorname{tr}\left[\frac{1}{n}\sum_{i=1}^n\boldsymbol{z}^i{\boldsymbol{z}^i}^\top\Omega\right]-\log\bigl[\det[\Omega]\bigr]+rh[\Omega]\right\}. \tag{3.2}
$$

The most popular example is the *graphical lasso*[5], where $h[\Omega]=\sum_{j,k=1}^p|\Omega_{ij}|$ or $h[\Omega]=\sum_{\substack{j,k=1\\ j\neq k}}^p|\Omega_{ij}|$. The graphical lasso does not have a general closed-form solution, but its objective function is amenable to fast proximal descent algorithms—cf. Exercise 3.3.

graphical lasso

Graph estimates can finally be obtained through thresholding the estimated precision matrix. We define $\widehat{\mathcal{G}} = (\mathcal{I}, \widehat{\mathcal{E}})$ from estimates $\widehat{\Theta} \in \mathbb{R}^{p\times p}$ of the precision matrix, $\widehat{\Theta} \in \{\widehat{\Theta}_{\mathrm{ml}}, \widehat{\Theta}_{\mathrm{mrl}}\}$, through

$$\widehat{\mathcal{E}} \;:=\; \bigl\{(i,j) \in \mathcal{I}\times\mathcal{I} \;:\; i \neq j,\, |\widehat{\Theta}_{ij}| > t\bigr\}\,, \tag{3.3}$$

where $t \in [0,\infty]$ is a threshold. The discussed estimators $\widehat{\Theta}_{\mathrm{ml}}, \widehat{\Theta}_{\mathrm{mrl}}$ are symmetric by construction and, therefore, generate symmetric edge sets: $(i,j) \in \widehat{\mathcal{E}}$ if and only if $(j,i) \in \widehat{\mathcal{E}}$. The threshold t regulates how conservative the graph estimate is: the larger t, the smaller the risk of false positives but the higher the risk of false negatives. In the case of sparsity inducing estimators such as the graphical lasso, a suggested threshold is $t = 0$, that is, the estimated edges are the ones that have non-zero entries in the estimated precision matrix. This choice avoids the introduction of yet another tuning parameter. In the case of unregularized maximum likelihood estimation, however, $t = 0$ almost always leads to a full graph. To remove the spurious dependencies, a suggested threshold is then $t = \sqrt{\log[p]/n}$. The size of this threshold is related to the size of tuning parameters r in front of ℓ_1-prior terms—cf. Section 4.2.

3.5 Neighborhood Selection

Neighborhood selection is an alternative approach to estimating the parameters of interest. It exploits that the coordinates of Gaussian random vectors are connected via standard linear regressions with the entries of the inverse covariance matrix as parameters. This transforms the estimation of the inverse covariance matrix into multiple regressions, which are amenable to the techniques of the previous chapter.

The following result for Gaussian vectors $\boldsymbol{z} \sim \mathcal{N}_p[\mathbf{0}_p, \Theta^{-1}]$ is the basis of neighborhood selection.

Lemma 3.5.1 (Neighborhood Selection)

Each coordinate z_j, $j \in \{1,\dots,p\}$, is the output of a linear regression model

$$z_j \;=\; \boldsymbol{z}_{\{j\}^{\complement}}^\top \boldsymbol{\beta}^j + u_j\,,$$

where $\{j\}^{\complement} := \{1,\dots,p\}\setminus\{j\}$, $\boldsymbol{\beta}^j := -(\Theta_{j\{j\}^{\complement}})^\top/\Theta_{jj} \in \mathbb{R}^{p-1}$, and $u_j \sim \mathcal{N}_1[0, 1/\Theta_{jj}]$ is independent of $\boldsymbol{z}_{\{j\}^{\complement}}$.

The lemma states that the entries of the precision matrix form the parameters in regressions of one coordinate of a Gaussian vector onto the others.

Proof of Lemma 3.5.1. We prove a more general statement: Consider a Gaussian vector $\boldsymbol{z} \sim \mathcal{N}_p[\mathbf{0}_p, \Theta^{-1}]$ with a precision matrix that we write as

$$\Theta \;=\; \begin{bmatrix} \Theta_{\mathcal{A}\mathcal{A}} & \Theta_{\mathcal{A}\mathcal{B}} \\ \Theta_{\mathcal{B}\mathcal{A}} & \Theta_{\mathcal{B}\mathcal{B}} \end{bmatrix}\,,$$

where $\mathcal{A} := \{1,\dots,k\}$, $\mathcal{B} := \{k+1,\dots,p\}$ for a fixed integer $k \in \{1,\dots,p-1\}$. Then, the conditional distribution of $\boldsymbol{z}_{\mathcal{A}}$ given $\boldsymbol{z}_{\mathcal{B}}$ is $\mathcal{N}_k[-(\Theta_{\mathcal{A}\mathcal{A}})^{-1}\Theta_{\mathcal{A}\mathcal{B}}\boldsymbol{z}_{\mathcal{B}}, (\Theta_{\mathcal{A}\mathcal{A}})^{-1}]$, that is,

$$\boldsymbol{z}_{\mathcal{A}} \;=\; -(\Theta_{\mathcal{A}\mathcal{A}})^{-1}\Theta_{\mathcal{A}\mathcal{B}}\boldsymbol{z}_{\mathcal{B}} + \boldsymbol{u}_{\mathcal{A}}\,,$$

where $\boldsymbol{u}_{\mathcal{A}} \sim \mathcal{N}_k[\mathbf{0}_k, (\Theta_{\mathcal{A}\mathcal{A}})^{-1}]$ is independent of $\boldsymbol{z}_{\mathcal{B}}$.

The key ingredient of the proof is Lemma B.2.6, which relates submatrices and inverses of matrices. We also use properties that are specific to Gaussian data; for example, we invoke that marginals of Gaussian distributions are again Gaussian.

Lemma 3.5.1 follows by setting $k = 1$ and (reshuffling the parameters if $j \neq 1$).

Writing

$$\Sigma \;=\; \begin{bmatrix} \Sigma_{\mathcal{A}\mathcal{A}} & \Sigma_{\mathcal{A}\mathcal{B}} \\ \Sigma_{\mathcal{B}\mathcal{A}} & \Sigma_{\mathcal{B}\mathcal{B}} \end{bmatrix},$$

Lemma B.2.6 (with $M = \Sigma$ and $M^{-1} = \Theta$) yields

$$\Sigma_{\mathcal{B}\mathcal{B}}^{-1} \;=\; \Theta_{\mathcal{B}\mathcal{B}} - \Theta_{\mathcal{B}\mathcal{A}}(\Theta_{\mathcal{A}\mathcal{A}})^{-1}\Theta_{\mathcal{A}\mathcal{B}}$$

and

$$\frac{\det[\Sigma_{\mathcal{B}\mathcal{B}}]}{\det[\Sigma]} \;=\; \frac{1}{\det[(\Theta_{\mathcal{A}\mathcal{A}})^{-1}]}\,.$$

We denote the conditional density of $\boldsymbol{z}_{\mathcal{A}}$ given $\boldsymbol{z}_{\mathcal{B}}$ by $f_{\mathcal{A}|\mathcal{B}}[\boldsymbol{z}_{\mathcal{A}}|\boldsymbol{z}_{\mathcal{B}}]$ and the marginal densities of $\boldsymbol{z}$ and $\boldsymbol{z}_{\mathcal{B}}$ by $f[\boldsymbol{z}]$ and $f_{\mathcal{B}}[\boldsymbol{z}_{\mathcal{B}}]$, respectively. Using the two insights above then allows us to derive the following:

$$\begin{aligned}
&f_{\mathcal{A}|\mathcal{B}}[\boldsymbol{z}_{\mathcal{A}}|\boldsymbol{z}_{\mathcal{B}}]\\
&=\; f[\boldsymbol{z}]/f_{\mathcal{B}}[\boldsymbol{z}_{\mathcal{B}}] && \text{“Bayes' Rule”}\\
&=\; \frac{1}{\sqrt{(2\pi)^p\det[\Sigma]}}\;e^{-\boldsymbol{z}^\top\Theta\boldsymbol{z}/2}\Big/\frac{1}{\sqrt{(2\pi)^{p-k}\det[\Sigma_{\mathcal{B}\mathcal{B}}]}}e^{-\boldsymbol{z}_{\mathcal{B}}^\top\Sigma_{\mathcal{B}\mathcal{B}}^{-1}\boldsymbol{z}_{\mathcal{B}}/2}\\
& && \text{“Lemma B.3.2 about Gaussian marginals”}\\
&=\; \sqrt{\frac{\det[\Sigma_{\mathcal{B}\mathcal{B}}]}{(2\pi)^k\det[\Sigma]}}\;e^{-\boldsymbol{z}_{\mathcal{A}}^\top\Theta_{\mathcal{A}\mathcal{A}}\boldsymbol{z}_{\mathcal{A}}/2-\boldsymbol{z}_{\mathcal{B}}^\top\Theta_{\mathcal{B}\mathcal{B}}\boldsymbol{z}_{\mathcal{B}}/2-\boldsymbol{z}_{\mathcal{A}}^\top\Theta_{\mathcal{A}\mathcal{B}}\boldsymbol{z}_{\mathcal{B}}/2-\boldsymbol{z}_{\mathcal{B}}^\top\Theta_{\mathcal{B}\mathcal{A}}\boldsymbol{z}_{\mathcal{A}}/2+\boldsymbol{z}_{\mathcal{B}}^\top\Sigma_{\mathcal{B}\mathcal{B}}^{-1}\boldsymbol{z}_{\mathcal{B}}/2}\\
& && \text{“simplifying; writing out } \boldsymbol{z}^\top\Theta\boldsymbol{z}\text{”}\\
&=\; \sqrt{\frac{\det[\Sigma_{\mathcal{B}\mathcal{B}}]}{(2\pi)^k\det[\Sigma]}}\;e^{-\boldsymbol{z}_{\mathcal{A}}^\top\Theta_{\mathcal{A}\mathcal{A}}\boldsymbol{z}_{\mathcal{A}}/2-\boldsymbol{z}_{\mathcal{A}}^\top\Theta_{\mathcal{A}\mathcal{B}}\boldsymbol{z}_{\mathcal{B}}/2-\boldsymbol{z}_{\mathcal{B}}^\top\Theta_{\mathcal{B}\mathcal{A}}\boldsymbol{z}_{\mathcal{A}}/2-\boldsymbol{z}_{\mathcal{B}}^\top\Theta_{\mathcal{B}\mathcal{A}}(\Theta_{\mathcal{A}\mathcal{A}})^{-1}\Theta_{\mathcal{A}\mathcal{B}}\boldsymbol{z}_{\mathcal{B}}/2}\\
& && \text{“first equality above”}\\
&=\; \frac{1}{\sqrt{(2\pi)^k\det\big[(\Theta_{\mathcal{A}\mathcal{A}})^{-1}\big]}}\;e^{-(\boldsymbol{z}_{\mathcal{A}}+\Theta_{\mathcal{A}\mathcal{A}}^{-1}\Theta_{\mathcal{A}\mathcal{B}}\boldsymbol{z}_{\mathcal{B}})^\top\Theta_{\mathcal{A}\mathcal{A}}(\boldsymbol{z}_{\mathcal{A}}+\Theta_{\mathcal{A}\mathcal{A}}^{-1}\Theta_{\mathcal{A}\mathcal{B}}\boldsymbol{z}_{\mathcal{B}})/2}\,.\\
& && \text{“second equality above; summarizing terms in the exponent recalling that } \Theta_{\mathcal{B}\mathcal{A}} = \Theta_{\mathcal{A}\mathcal{B}} \text{ due to the symmetry of } \Theta\text{”}
\end{aligned}$$

This is the density of a normal distribution with mean $-(\Theta_{\mathcal{A}\mathcal{A}})^{-1}\Theta_{\mathcal{A}\mathcal{B}}\boldsymbol{z}_{\mathcal{B}}$ and covariance matrix $(\Theta_{\mathcal{A}\mathcal{A}})^{-1}$. In other words, the conditional distribution of $\boldsymbol{z}_{\mathcal{A}}$ given $\boldsymbol{z}_{\mathcal{B}}$ is $\mathcal{N}_k[-(\Theta_{\mathcal{A}\mathcal{A}})^{-1}\Theta_{\mathcal{A}\mathcal{B}}\boldsymbol{z}_{\mathcal{B}}, (\Theta_{\mathcal{A}\mathcal{A}})^{-1}]$, as desired. □

We now leverage Lemma 3.5.1 to estimate the parameters of interest. For each $j \in \{1, \dots, p\}$, we augment the $p-1$-dimensional regression vector $\boldsymbol{\beta}^j$ defined in the lemma by inserting a -1-valued coordinate at the jth position, and we denote the resulting p-dimensional vector by $\underline{\boldsymbol{\beta}}^j$; for example, $\underline{\boldsymbol{\beta}}^1 = (-1, (\boldsymbol{\beta}^1)^\top)^\top \in \mathbb{R}^p$. Some algebra yields (see Lemma B.2.7 in the Appendix)

$$\Theta_{ij} \;=\; -\frac{\Theta_{ii}(\underline{\boldsymbol{\beta}}^i)_j + \Theta_{jj}(\underline{\boldsymbol{\beta}}^j)_i}{2} \qquad (i, j \in \{1, \dots, p\})\,.$$

Given independent realizations $\boldsymbol{z}^1, \dots, \boldsymbol{z}^n$ of the random vector $\boldsymbol{z}$, we then estimate Θ_{ij} by estimating the four parameters on the right-hand side of the display. We can do this by regressing for each node $j \in \{1, \dots, p\}$ the vector $\boldsymbol{y}^j := ((z^1)_j, \dots, (z^n)_j)^\top \in \mathbb{R}^n$

on the matrix $X^j := ((\boldsymbol{z}^1)_{\{j\}^\mathtt{C}}, \dots, (\boldsymbol{z}^n)_{\{j\}^\mathtt{C}})^\top \in \mathbb{R}^{n\times(p-1)}$. We denote the outputs of these regressions by $\widehat{\boldsymbol{\beta}}^j \equiv \widehat{\boldsymbol{\beta}}^j[\boldsymbol{y}^j, X^j] \in \mathbb{R}^{p-1}$ and their augmented versions (again with a -1-valued coordinate inserted at the jth position) by $\underline{\widehat{\boldsymbol{\beta}}}^j \in \mathbb{R}^p$. These outputs yield the estimates $\widehat{\Theta_{jj}} := n/\|\boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j\|_2^2$ for the diagonal entries and $\widehat{(\underline{\boldsymbol{\beta}}^j)_i} := (\underline{\widehat{\boldsymbol{\beta}}}^j)_i$ for the coordinates of the regression vectors. Indeed, the above lemma ensures that $n/\|\boldsymbol{y}^j - X^j\boldsymbol{\beta}^j\|_2^2 \sim n/\|\boldsymbol{u}\|_2^2$ with $\boldsymbol{u} \sim \mathcal{N}_n[\mathbf{0}_n, \mathrm{I}_{n\times n}/\Theta_{jj}]$, and the law of large numbers that $\|\boldsymbol{u}\|_2^2$ converges to n/Θ_{jj} as $n \to \infty$. Hence, if the $\widehat{\boldsymbol{\beta}}^j$'s are consistently estimating $\boldsymbol{\beta}^j$, also $\widehat{\Theta_{jj}}$ and $\widehat{(\underline{\boldsymbol{\beta}}^j)_i}$ are consistently estimating their population counterparts.

In summary, the lemma motivates an estimate $\widehat{\Theta}_{\mathrm{ns}}$ of the precision matrix Θ with elements

$$(\widehat{\Theta}_{\mathrm{ns}})_{ij} \;:=\; -\frac{\frac{n}{\|\boldsymbol{y}^i - X^i\widehat{\boldsymbol{\beta}}^i\|_2^2}(\underline{\widehat{\boldsymbol{\beta}}}^i)_j + \frac{n}{\|\boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j\|_2^2}(\underline{\widehat{\boldsymbol{\beta}}}^j)_i}{2} \qquad (i,j \in \{1,\dots,p\})\,. \quad (3.4)$$

In the case $p \ll n$, least-squares can be chosen for the initial estimators $\widehat{\boldsymbol{\beta}}^j$; then, in fact, $\widehat{\Theta}_{\mathrm{ns}} = \widehat{\Theta}_{\mathrm{ml}}$—see Exercise 3.4. Otherwise, regularized regression methods can be applied. A particular popular choice is the lasso; we call the corresponding scheme the *neighborhood lasso*.[6]

neighborhood lasso

Finally, graph estimates are obtained again according to the rule (3.3).

Because neighborhood selection can draw on the abundant algorithmic research and software packages for linear regression, it is often faster and easier to implement than the corresponding maximum likelihood approaches. However, regularized neighborhood methods require tuning parameter calibration for p problems; then again, tuning parameter calibration is better understood for regression than for other likelihood methods. Overall, there is no clear winner between the two approaches—neither theoretically nor empirically.

A limitation of both approaches is their fundamental dependence on the data being Gaussian. Maximum likelihood invokes normality when formulating the likelihoods; neighborhood selection invokes normality when splitting the problem into multiple regressions. In contrast, because least-squares—while not necessarily optimal beyond Gaussian data—still works quite generally, the methods for linear regression in the previous chapter do not rely on normality to such an extend.

It is often recommended to scale the data before applying maximum likelihood or neighborhood selection. A standard scaling transformation is

scaling

$$z_j^i \;\mapsto\; \frac{z_j^i - \sum_{k=1}^n z_j^k/n}{\sqrt{\sum_{l=1}^n \bigl(z_j^l - \sum_{k=1}^n z_j^k/n\bigr)^2/(n-1)}} \qquad (i \in \{1,\dots,n\},\, j \in \{1,\dots,p\})\,,$$

which sets 1. the predictors' sample means to zero and 2. their Euclidean norms uniformly to $\sqrt{n}$. The rationale is that 1. circumvents intercepts and 2. homogenizes regularization. When interpreting the results, however, one has to bear in mind that, conceptionally, scaling presumes that the scaled data—rather than the original data—follows a multivariate normal distribution.

3.6 References and Further Reading

[1] Books specialized on (low-dimensional) graphical modeling include [BK, Edw12, Lau96].

[2] Applications of graphical models include Statistical Mechanics [Gal13], Quantum Field Theory [ZI77], Sociology [DC05], and Biology [KMM+15].

[3] Background on conditional expectations can be found in [Dur10, Chapter 5.1, Pages 221ff]. The same book also introduces the basic notions of measure theory.

[4] A more comprehensive version of the Hammersley-Clifford Theorem and a proof can be found in [Lau96, Theorem 3.9 on Page 36].

[5] [YL07] introduces ℓ_1-regularized maximum likelihood estimation for Gaussian data and discusses it connections to neighborhood selection with the lasso. Further algorithms for the maximum likelihood approach in the Gaussian case followed quickly in [BGd08, FHT08]. The latter paper also coined the term "graphical lasso."

[6] Neighborhood selection with the lasso was introduced in [MB06].

3.7 Exercises

Exercises for Section 3.1

□ **Exercise 3.1** $^{\diamond\diamond\bullet}$ Give further examples that show that conditional independence does not necessarily imply independence and vice versa.

Exercises for Section 3.4

□ **Exercise 3.2** $^{\diamond\diamond\bullet}$ In this exercise, we show that the function $\Omega \mapsto \operatorname{tr}[A\Omega]-\log\det\Omega$ is strictly convex on $\mathcal{S}_p^+ = \{\Omega \in \mathbb{R}^{p\times p} : \Omega \text{ symmetric and positive definite}\}$ for any $A \in \mathbb{R}^{p\times p}$. This implies in particular that the objective function of the maximum likelihood estimator for the precision matrix Θ in Gaussian graphical models (see Page 50) is strictly convex.

We first make sure that speaking about convexity makes sense here.

1. Show that the set $\mathcal{S}_p^+$ is convex.

We then consider the trace function $\Omega \mapsto \operatorname{tr}[A\Omega]$.

2. Show that for any $A \in \mathbb{R}^{p\times p}$, this function is convex on the entire space $\mathbb{R}^{p\times p}$ (which is trivially convex), that is, $\operatorname{tr}[A(v\Omega' + (1-v)\Omega'')] \leq v\operatorname{tr}[A\Omega'] + (1-v)\operatorname{tr}[A\Omega'']$ for all $\Omega', \Omega'' \in \mathbb{R}^{p\times p}$ and $v \in [0,1]$. Assure yourself that this implies that the function is convex also on $\mathcal{S}_p^+$.

We now consider the determinant function $\Omega \mapsto -\log\det\Omega$.

3. Show that for any matrix $\Omega \in \mathcal{S}_p^+$, it holds that $\log\det\Omega = \sum_{j=1}^p \log a_j$, where $a_1, \dots, a_p > 0$ are the eigenvalues of Ω.
4. Show by using 3. that for any $\Omega', \Omega'' \in \mathcal{S}_p^+$, $\Omega' \neq \Omega''$, the function $t \mapsto -\log\det[\Omega'' + t(\Omega' - \Omega'')]$ is *strictly* convex on $[0,1]$.
5. Show by using 4. that $\Omega \mapsto -\log\det\Omega$ is *strictly* convex on $\mathcal{S}_p^+$.

We can finally derive the desired claim.

6. Conclude from 2. and 5. that for any $A \in \mathbb{R}^{p\times p}$, the function $\Omega \mapsto \operatorname{tr}[A\Omega] - \log\det\Omega$ is strictly convex.

□ **Exercise 3.3** $^{\diamond\diamond\bullet}$ In this exercise, we confirm the closed-form solution of the unregularized maximum likelihood estimator stated in Section 3.4.

1. Show that for any given matrix $A \in \mathbb{R}^{p\times p}$, the (matrix-valued) gradient of the trace function $\Omega \mapsto \operatorname{tr}[A\Omega]$ on $\mathbb{R}^{p\times p}$ is $A^\top$, that is,
$$\frac{\partial}{\partial\Omega}\operatorname{tr}[A\Omega] \;=\; A^\top\,.$$
2. Show that the (matrix-valued) gradient of the log-determinant function $\Omega \mapsto \log[\det[\Omega]]$ on the invertible matrices in $\mathbb{R}^{p\times p}$ is $(\Omega^{-1})^\top$, that is,
$$\frac{\partial}{\partial\Omega}\log\bigl[\det[\Omega]\bigr] \;=\; (\Omega^{-1})^\top\,.$$
Hint: you can do a Laplace expansion of the determinant and use that an invertible matrix Ω and its cofactor matrix C are related as $\Omega^{-1} = C^\top/\det[\Omega]$.
3. Conclude that for any given matrix $A \in \mathcal{S}_p^+$, where $\mathcal{S}_p^+$ are the symmetric and invertible matrices in $\mathbb{R}^{p\times p}$, the minimization program
$$\widehat{\Theta} \;\in\; \operatorname*{argmin}_{\Omega\in\mathcal{S}_p^+}\bigl\{\operatorname{tr}[A\Omega] - \log\bigl[\det[\Omega]\bigr]\bigr\}$$
has the unique solution $\widehat{\Theta} = A^{-1}$. Hint: start from Claim 6 in Exercise 3.2.
4. Show that if $n \geq p$, the scaled Gram matrix $\sum_{i=1}^n \boldsymbol{z}^i{\boldsymbol{z}^i}^\top/n$ of n independent realizations $\boldsymbol{z}^1,\dots,\boldsymbol{z}^n$ of $\boldsymbol{z} \sim \mathcal{N}_p[\mathbf{0}_p,\Sigma]$, Σ symmetric and invertible, is symmetric and invertible with probability one: $\mathbb{P}\{\sum_{i=1}^n \boldsymbol{z}^i{\boldsymbol{z}^i}^\top/n \in \mathcal{S}_p^+\} = 1$.
5. Conclude that if $n \geq p$, Identity (3.1) on Page 50 holds true with probability one:
$$\mathbb{P}\Biggl\{\widehat{\Theta}_{\mathrm{ml}} \;=\; \Biggl(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}^i{\boldsymbol{z}^i}^\top\Biggr)^{-1}\Biggr\} \;=\; 1\,.$$

Exercises for Section 3.5

□ **Exercise 3.4** ◇◇◇• In this exercise, we show that the unregularized maximum likelihood estimator and neighborhood selection with the least-squares coincide.

Assume that the Gram matrix $\sum_{i=1}^n \boldsymbol{z}^i{\boldsymbol{z}^i}^\top/n$ is invertible and denote by $\widehat{\boldsymbol{\beta}}^j \equiv \widehat{\boldsymbol{\beta}}^j[\boldsymbol{y}^j, X^j] \in \mathbb{R}^{p-1}$, $j \in \{1,\dots,p\}$, the least-squares estimators for regressing $\boldsymbol{y}^j := ((z^1)_j,\dots,(z^n)_j)^\top \in \mathbb{R}^n$ on $X^j := ((\boldsymbol{z}^1)_{\{j\}^\complement},\dots,(\boldsymbol{z}^n)_{\{j\}^\complement})^\top \in \mathbb{R}^{n\times(p-1)}$. The augmented versions (with a -1-valued coordinate inserted at the jth position) of the least-squares estimators are denoted by $\underline{\widehat{\boldsymbol{\beta}}}^j \in \mathbb{R}^p$.

1. Show that for all $j \in \{1,\dots,p\}$, it holds that
$$\Biggl(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}^i{\boldsymbol{z}^i}^\top\Biggr)_{\{j\}^\complement\{1,\dots,p\}}\underline{\widehat{\boldsymbol{\beta}}}^j \;=\; \mathbf{0}_{p-1}\,.$$
This means that the augmented least-squares estimators are in the kernel of a Gram sub-matrix.
2. Show that for all $j \in \{1,\dots,p\}$, it holds that
$$\Biggl(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}^i{\boldsymbol{z}^i}^\top\Biggr)_{j\{1,\dots,p\}}\underline{\widehat{\boldsymbol{\beta}}}^j \;=\; -\frac{\|\boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j\|_2^2}{n}\,.$$
This connects the remaining row of the Gram matrix with the augmented least-squares estimators and the least-squares' prediction loss.

3. Show finally that

$$\widehat{\Theta}_{\mathrm{ns}} \;=\; \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}^i{\boldsymbol{z}^i}^\top\right)^{-1}.$$

We conclude that the estimates in (3.1) and (3.4) coincide under the stated assumptions.

R Lab Chapter 3

3 Estimating a gene-gene coactivation network

In this lab, we fit gene-gene coactivation networks to gene expression data. The data consists of vector-valued samples, each of them describing the expression levels of genes in a given tissue specimen. Our assumption is that these vectors are independent and identically distributed according to a multivariate Gaussian distribution. We describe the dependence structures of the expression levels by using graphs, and since gene expressions measure gene activities, we can interpret these graphs as gene-gene coactivation networks.

As always, your task is to replace the keyword `REPLACE` with suitable code and to answer the questions posed in the text.

3.1 Tests on synthetic data

We first test maximum likelihood and neighborhood selection on synthetic data generated from the model in Figure 3.2.

3.1.1 Generating data

Generate $n = 200$ independent samples from the Gaussian graphical model in Figure 3.2 and summarize these samples in a matrix $Z \in \mathbb{R}^{n \times p}$. You might want to use the `solve()` function to compute the covariance matrix and the `mvrnorm()` function from the `MASS` package to generate the data.

```
library(MASS)
set.seed(3)
invcovariance <- matrix(data=c(1.2, -0.2, 0, 0, -0.2, 1.5, 0, 0.3, 0, 0, 1, 0,
                               0, 0.3, 0, 0.5), nrow=4, ncol=4)
invcovariance
```

```
##      [,1] [,2] [,3] [,4]
## [1,]  1.2 -0.2    0  0.0
## [2,] -0.2  1.5    0  0.3
## [3,]  0.0  0.0    1  0.0
## [4,]  0.0  0.3    0  0.5
```

```
Z <- REPLACE
head(Z)  # displays the first couple of rows of Z
```

```
##             [,1]       [,2]       [,3]       [,4]
## [1,]  1.85932683  1.4249675  1.3323523 -1.0124874
## [2,]  0.84552878  0.3055378 -0.2773236 -0.3242895
## [3,] -0.05897421  1.1439170 -1.0855338  0.7471096
## [4,]  0.66070932 -0.4818663  1.6427185 -1.9616956
## [5,] -0.72279811  0.8543313  0.4571803  0.5120927
## [6,]  0.32338537  1.0411431 -1.4611718  0.3741942
```

Note that $n \gg p$, so that we can apply unregularized estimators for analyzing these data.

3.1.2 Parameter estimation via maximum likelihood

We now implement and apply the maximum likelihood estimator. We do this in two steps.

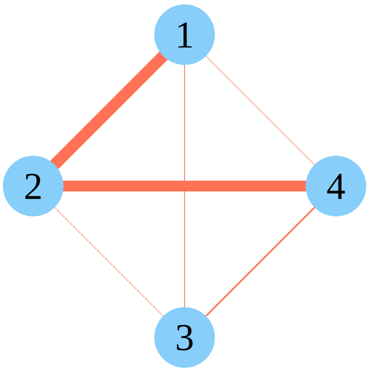
3.1.2.1 Estimating the inverse covariance matrix Recall that the (unregularized) maximum likelihood estimator (3.1) is the inverse of the scaled Gram matrix: in matrix notation, $\widehat{\Theta}_{\mathrm{ml}} = (Z^\top Z/n)^{-1}$. Compute this estimator for the above data and study the stated visualization pipeline.

```
invcovariance.ml <- REPLACE
invcovariance.ml
```

```
##             [,1]        [,2]        [,3]        [,4]
## [1,]  1.23981631 -0.29082882  0.02037997 -0.01386750
## [2,] -0.29082882  1.58948298 -0.01016519  0.27151900
## [3,]  0.02037997 -0.01016519  0.94533132 -0.04109434
## [4,] -0.01386750  0.27151900 -0.04109434  0.49903322
```

Comparing to the true inverse covariance, we find that maximum likelihood estimation is reasonably accurate in our test case.

```
library(igraph)
network.initial <- graph.adjacency(abs(invcovariance.ml), weighted=TRUE,
                                   mode="undirected", diag=FALSE)
network.layout <- layout_in_circle(network.initial, order=c(4, 1, 2, 3))
igraph_options(vertex.size        = 40,
               vertex.color       = "lightskyblue",
               vertex.frame.color = NA,
               vertex.label.cex   = 3,
               vertex.label.color = "black",
               edge.width         = 50 * E(network.initial)$weight,
               edge.color         = "coral1")
plot.igraph(network.initial, layout=network.layout)
```



What does this plot visualize?

3.1.2.2 Estimating the graph Compute the graph estimate (3.3) with threshold $t = \sqrt{\log[p]/n}$. This estimate should be written in terms of a so-called *adjacency matrix* $\widehat{A} \in \mathbb{R}^{p\times p}$ defined through $\widehat{A}_{ij} := 1$ if $(i,j) \in \widehat{\mathcal{E}}$ and $\widehat{A}_{ij} := 0$ otherwise.

Then, compare the matrix with the adjacency matrix A that captures the true graph: $A_{ij} := 1$ if $(i,j) \in \mathcal{E}$ and $A_{ij} := 0$ otherwise.

Finally, visualize the estimated adjacency matrix as above.

```
adjacencymatrix.ml <- REPLACE  # estimated graph
adjacencymatrix.ml
```

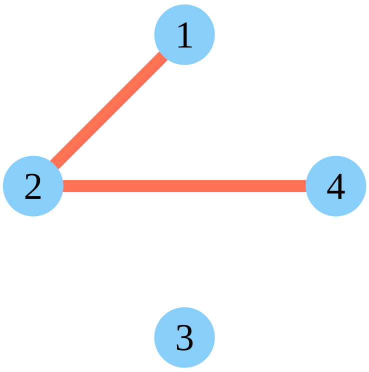
```
##      [,1] [,2] [,3] [,4]
## [1,]    0    1    0    0
## [2,]    1    0    0    1
## [3,]    0    0    0    0
## [4,]    0    1    0    0
```

```
adjacencymatrix <- REPLACE  # true graph
adjacencymatrix
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    0    1    0    0
## [2,]    1    0    0    1
## [3,]    0    0    0    0
## [4,]    0    1    0    0
```

Comparing the adjacency matrices, we find that maximum likelihood with the standard threshold recovers the graph correctly.

```
REPLACE  # draw the graph
```



3.1.3 Parameter estimation via neighborhood selection

We now implement and apply a neighborhood selection scheme. We do this in four steps.

3.1.3.1 Linear regressions Implement first a least-squares estimator (without intercept) for usual regression data of the form $y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times(p-1)}$. The output of this function is a $p-1$-dimensional vector. Then, study the function `RegressionVectors`: what does it do?

```
LsEstimator <- function(y, X)
{
  return(REPLACE)
}
RegressionVectors <- function(Z, FUN=LsEstimator)
{
  betas <- NULL
  for (j in 1:dim(Z)[2])  # passing through all nodes
  {
    betas <- cbind(betas, FUN(Z[, j], Z[, -j]))
  }
  return(betas)
}
RegressionVectors(Z)
```

```
##            [,1]        [,2]        [,3]        [,4]
## [1,] 0.23457412 0.182970704 -0.02155855  0.02778872
## [2,] -0.01643790  0.006395282  0.01075305 -0.54409004
## [3,]  0.01118512 -0.170822214  0.04347084  0.08234791
```

3.1.3.2 Calculating the individual parts of the estimator in Display (3.4) Implement a function that augments the $(p-1)$-dimensional columns of the above matrix as described in Section 3.4. Then, implement a function that calculates the estimated diagonal entries $\widehat{\Theta_{jj}}$ as described in Section 3.5.

```
AugmentMatrix <- function(betas)
{
  REPLACE
}
AugmentMatrix(RegressionVectors(Z))
```

```
##             [,1]         [,2]        [,3]        [,4]
## [1,] -1.00000000  0.182970704 -0.02155855  0.02778872
## [2,]  0.23457412 -1.000000000  0.01075305 -0.54409004
## [3,] -0.01643790  0.006395282 -1.00000000  0.08234791
## [4,]  0.01118512 -0.170822214  0.04347084 -1.00000000
```

```
DiagonalEntries <- function(Z, betas)
{
  REPLACE
}
DiagonalEntries(Z, RegressionVectors(Z))
```

```
## [1] 1.2398163 1.5894830 0.9453313 0.4990332
```

3.1.3.3 Estimating the inverse covariance matrix We now put the pieces together to estimate the inverse covariance matrix.

```
NeighborhoodSelection <- function(Z, FUN=LsEstimator)
{
  betas <- RegressionVectors(Z, FUN)
  betas.augmented <- AugmentMatrix(betas)
  diagonal.entries <- DiagonalEntries(Z, betas)
  invcovariance.ns <- matrix(data=0, nrow=dim(Z)[2], ncol=dim(Z)[2])
  for (i in 1:dim(Z)[2])
  {
    for (j in 1:dim(Z)[2])
    {
      invcovariance.ns[i,j] <- REPLACE
    }
  }
  return(invcovariance.ns)
}
invcovariance.ns <- NeighborhoodSelection(Z)
invcovariance.ns
```

```
##             [,1]        [,2]        [,3]        [,4]
## [1,]  1.23981631 -0.29082882  0.02037997 -0.01386750
## [2,] -0.29082882  1.58948298 -0.01016519  0.27151900
## [3,]  0.02037997 -0.01016519  0.94533132 -0.04109434
## [4,] -0.01386750  0.27151900 -0.04109434  0.49903322
```

We find that neighborhood selection yields the same estimate for the inverse covariance matrix as maximum likelihood–which commensurates with the theoretical finding in Exercise 3.7.4.

3.1.3.4 Estimating the graph Return the graph estimate (3.3) with threshold $t = \sqrt{\log[p]/n}$ in terms of an adjacency matrix.

```
adjacencymatrix.ns <- REPLACE
adjacencymatrix.ns
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    0    1    0    0
## [2,]    1    0    0    1
## [3,]    0    0    0    0
## [4,]    0    1    0    0
```

In line with the preceeding result, we find that the neighborhood selection scheme provides perfect graph recovery in the test case.

3.2 A low-dimensional gene network

We first estimate a network of only a small number of genes. Throughout the real data analysis, we use neighborhood selection rather than maximum likelihood. A major advantage of neighborhood selection is that once having set up the above pipeline, it is extremly easy to account for high-dimensional data: it suffices to replace the least-squares with a high-dimensional regression method.

3.2.1 Loading and preprocessing the data

Download the file `GraphicalModels_Lab_Data.rda` from the book's homepage to your `R` working directory. Loading this file into `R` populates a matrix-valued variable `data` with the measurements. The ith row of this matrix corresponds to the ith sample; the jth column corresponds to the jth gene.

Store scaled versions of the first $p = 10$ genes' expressions in a matrix $Z \in \mathbb{R}^{n \times p}$, where n is the number of samples and p the number of genes under consideration. Use the scaling described at the end of Section 3.5; the `scale()` function could be helpful for this.

```
# make sure that the file is in the current working directory
load("GraphicalModels_Lab_Data.rda")
Z <- REPLACE
head(Z)
```

```
##         MIMAT0004571 MIMAT0000318 MIMAT0000682 MIMAT0001536 MIMAT0000255
## JB5011V   -0.3539915 -0.009955891   0.04969289    0.1694809   -0.2973696
## JB5143V    0.6556383  0.470776739   0.69327724    0.7188795   -1.4154755
## JB4870E    1.3373404  1.216808945   1.46075045    1.4006642    0.5316791
## GL4660E    1.2837458  1.164566489   1.43165213    1.5955177    0.9204437
## JB2851E    2.5373463  1.922426060   2.22943935    2.2654534    2.2155121
## GL3907E    0.2346137  0.554327478   0.74559381    0.6122908    0.8785116
##         MIMAT0005880 MIMAT0016884 MIMAT0000692 MIMAT0000693 MIMAT0000244
## JB5011V    0.1055220   -0.6005179    0.9915713    0.8598650    0.8149090
## JB5143V    2.2199264    1.8321555    1.4154591    1.2689738    1.2899616
## JB4870E   -0.3707105    0.2089269    1.0052872    0.9276796    0.7200827
## GL4660E    0.5041432   -1.1400383    1.3922733    1.3402264    1.2714853
## JB2851E    1.9305868   -1.0046173    1.9828855    1.6461652    1.4039293
## GL3907E    0.2259458    0.4654818    0.6880983    0.5986850    0.6003284
```

Verify that the number of samples in `Z` is $n = 192$ and the number of parameters $p = 10$. This means in particular that we $n \gg p$, so that we can apply unregularized estimators.

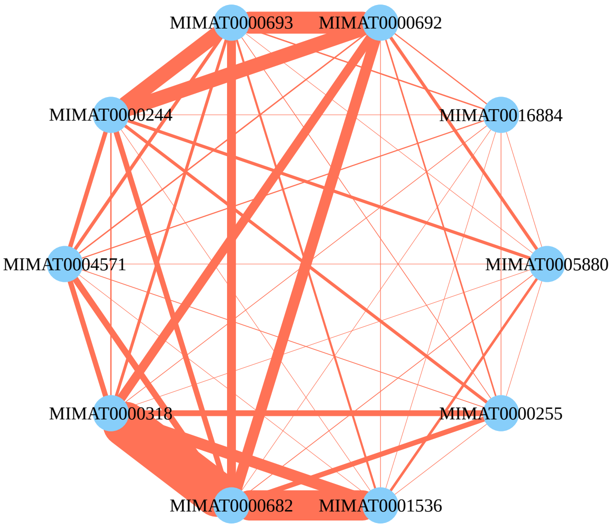
3.2.2 Estimating the inverse covariance matrix

Estimate the inverse covariance matrix with the above neighborhood selection scheme. Visualize the result.

```
invcovariance.ls <- REPLACE
# adding back row and column names
colnames(invcovariance.ls) <- rownames(invcovariance.ls) <- colnames(Z)
invcovariance.ls[1:4, 1:4]
```

```
##               MIMAT0004571 MIMAT0000318 MIMAT0000682 MIMAT0001536
## MIMAT0004571   3.99643433    -1.529433    -1.863957   0.03075732
## MIMAT0000318  -1.52943335    18.040407   -13.215965  -3.34469452
## MIMAT0000682  -1.86395690   -13.215965    24.285814  -8.90267821
## MIMAT0001536   0.03075732    -3.344695    -8.902678  13.26990150
```

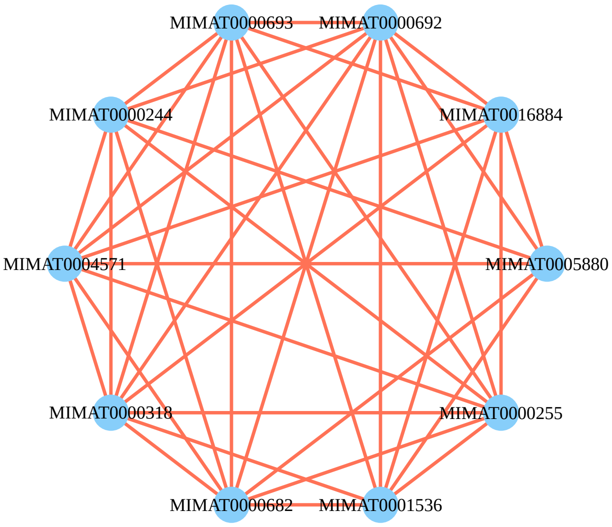
```
REPLACE  # draw the graph
```



3.2.3 Estimating the graph

Estimate the graph according to Display (3.3) with the standard threshold $t = \sqrt{\log[p]/n}$ and visualize the result.

```
REPLACE  # draw the graph
```



Once more, explain how the meanings of this plot and the preceeding one differ.

3.3 A high-dimensional gene network

We now increase the number of genes.

3.3.1 Loading and preprocessing the data

Proceed as before except for increasing the number of genes to $p = 50$.

```
Z <- REPLACE
Z[1:4, 1:4]
```

```
##         MIMAT0004571 MIMAT0000318 MIMAT0000682 MIMAT0001536
## JB5011V   -0.3539915 -0.009955891   0.04969289    0.1694809
## JB5143V    0.6556383  0.470776739   0.69327724    0.7188795
## JB4870E    1.3373404  1.216808945   1.46075045    1.4006642
## GL4660E    1.2837458  1.164566489   1.43165213    1.5955177
```

The number of samples is still $n = 192$, but the number of parameters is now $p = 50$. This means in particular that we should replace the least-squares estimator in the abve neighborhood selection pipeline with a high-dimensional estimator such as the lasso.

3.3.2 Estimating the inverse covariance matrix

Implement a lasso estimator for usual regression data of the form $y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times(p-1)}$. For this, use the `cv.glmnet()` function from the `glmnet` package. Because the data is centered, intercepts can be disregarded: set the flag `intercept=FALSE`, and make sure that the outputted regression vector has the correct dimensions.

We then estimate the inverse covariance matrix as in the neighborhood selection pipeline above except for replacing the least-squares estimator with the lasso estimator.

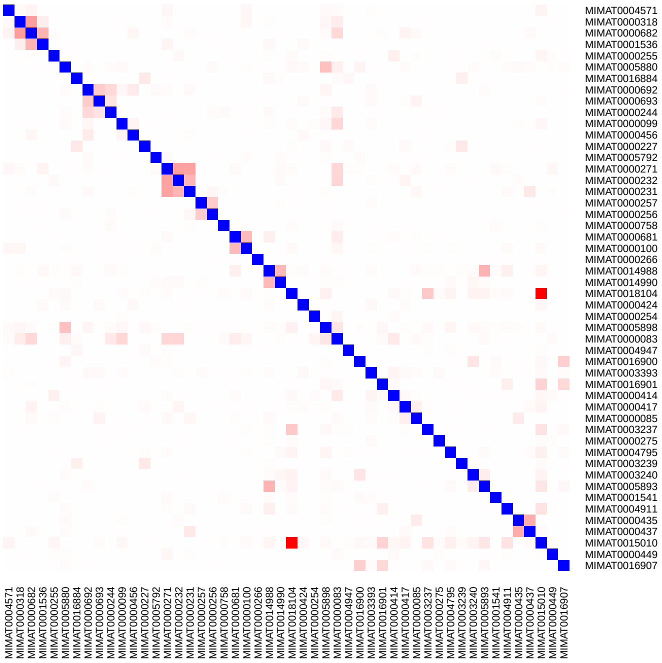
```
library(glmnet)
set.seed(84)  # the glmnet cross-validation routine is randomized
LassoEstimator <- function(y, Z)
{
  return(REPLACE)
}
invcovariance.lasso <- NeighborhoodSelection(Z, FUN=LassoEstimator)
colnames(invcovariance.lasso) <- rownames(invcovariance.lasso) <- colnames(Z)
invcovariance.lasso[1:4, 1:4]
```

```
##              MIMAT0004571 MIMAT0000318 MIMAT0000682 MIMAT0001536
## MIMAT0004571    3.8535550   -0.8242736    -1.606796     0.000000
## MIMAT0000318   -0.8242736   10.7335732   -12.988148    -2.091517
## MIMAT0000682   -1.6067958  -12.9881485    34.626414    -9.859611
## MIMAT0001536    0.0000000   -2.0915175    -9.859611    12.509681
```

3.3.3 Visualizing the inverse covariance matrix

We finally visualize the absolute values of the off-diagonals in the estimated inverse covariance matrix. In view of the large number of genes under consideration, we use a heatmap instead of an adjacency matrix or graph.

```
invcovariance.lasso.vis <- invcovariance.lasso - diag(diag(invcovariance.lasso))
invcovariance.lasso.vis <- abs(invcovariance.lasso.vis) /
                                max(abs(invcovariance.lasso.vis))  # mapping into [0,1]
diag(invcovariance.lasso.vis) <- -1 # to make the diagonal stand out visually
heatmap(x       = invcovariance.lasso.vis,
        scale   = "none",
        Rowv    = NA,
        Colv    = "Rowv",
        margins = c(10, 10),
        cexRow  = 0.45,
        cexCol  = 0.45,
        col     = colorRampPalette(c("blue", "white", "red"))(n = 1000))
```



Which genes appear to the have the largest pairwise dependence?

Chapter 4

Tuning Parameter Calibration

High-dimensional estimators consist of two terms, one for comparing parameters to the data and one for including prior information. The tuning parameters assign weights to these terms: a small tuning parameters emphasizes the data, while a large tuning parameter emphasizes the prior information. An optimally calibrated tuning parameter strikes a balance between the data and the prior information such the estimator's loss for a given task is minimized. Since this loss cannot be computed in practice, we are interested in data-driven schemes that mimic the optimal calibration as good as possible.

In this chapter, we introduce two such calibration schemes: cross-validation and adaptive validation. Cross-validation is recommended for prediction tasks, given that it is easy to implement, computationally feasible for most applications, and highly competitive in many simulation studies about prediction. Adaptive validation is recommended for tasks that can be based on suitable oracle inequalities, such as ℓ_∞-estimation and variable selection, given that it is provably optimal for such tasks, easy to implement, and computationally efficient. There are also many other calibration schemes, each one with its specific advantages and disandvantages in terms of how easy they are to use, how much computational resources they require, and how well they are backed up by mathematical and empirical support for the specific task at hand.

4.1 Overview

Data fitting terms measure the parameters' suitability for explaining the data at hand. To not distort this measuring, prior terms must be "small enough." On the other hand, data is random, and this randomness becomes more and more incalculable with increasing dimensionality of the problem. It turns out that the randomness can be contained if the prior terms are "large enough." In the following, we study these two constraints for the lasso.

We consider again the linear regression model

$$\boldsymbol{y} \;=\; X\boldsymbol{\beta}+\boldsymbol{u}$$

with outcome $\boldsymbol{y}\in\mathbb{R}^n$, design matrix $X\in\mathbb{R}^{n\times p}$, regression vector $\boldsymbol{\beta}\in\mathbb{R}^p$, and noise $\boldsymbol{u}\in\mathbb{R}^n$. If the prediction target $X\boldsymbol{\beta}$ were known, we could "estimate" a sparse surrogate of $\boldsymbol{\beta}$ via

$$\boldsymbol{\beta}^* \;\in\; \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\bigl\{\,\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+r\|\boldsymbol{\alpha}\|_1\,\bigr\}\,.$$

We can also see this as a special case of the lasso when the noise $\boldsymbol{u}$ is zero, and therefore, $\boldsymbol{y} = X\boldsymbol{\beta}$. The first term in the objective function ensures closeness to the target in prediction, and the second term induces sparsity. The tuning parameter r can be chosen at will: the smaller r, the better the results fit in prediction; the larger r, the more sparse the results. In particular, one can produce surrogates that have a perfect fit (set $r = 0$) and surrogates that are perfectly sparse (set $r \to \infty$).

In practice, where $X\boldsymbol{\beta}$ is unknown and the noise $\boldsymbol{u}$ is not zero, tuning parameter calibration has an additional facet. In place of the idealistic estimator above, consider the lasso

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\bigl\{ \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_1 \bigr\},$$

where the unknown $X\boldsymbol{\beta}$ is substituted by the known $\boldsymbol{y}$. The tuning parameter now balances sparsity and the fit to the outcome $\boldsymbol{y}$ instead of to the prediction target $X\boldsymbol{\beta}$. This change brings to bear the noise $\boldsymbol{u}$, which confines the range of tuning parameters that are feasible for finding good surrogates of the underlying model parameter $\boldsymbol{\beta}$. To show this, we write the objective function of the lasso as

$$\begin{aligned}
&\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_1 \\
&= \|\boldsymbol{y} - X\boldsymbol{\beta} + X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_1 && \text{“adding a zero-valued term”}\\
&= \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + 2\langle \boldsymbol{y} - X\boldsymbol{\beta},\, X\boldsymbol{\beta} - X\boldsymbol{\alpha}\rangle + \|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_1\\
& && \text{“expanding the data term”}\\
&= \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + 2\langle X^\top(\boldsymbol{y} - X\boldsymbol{\beta}),\, \boldsymbol{\beta}\rangle - 2\langle X^\top(\boldsymbol{y} - X\boldsymbol{\beta}),\, \boldsymbol{\alpha}\rangle\\
&\qquad + \|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_1 && \text{“properties of inner products”}\\
&= \|\boldsymbol{u}\|_2^2 + 2\langle X^\top\boldsymbol{u},\, \boldsymbol{\beta}\rangle - 2\langle X^\top\boldsymbol{u},\, \boldsymbol{\alpha}\rangle + \|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_1\,. && \text{“invoking the model”}
\end{aligned}$$

Since the first two terms are independent of the parameter $\boldsymbol{\alpha}$ we optimize over, the display shows that the lasso estimator can be written as

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\bigl\{ \|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_1 - 2\langle X^\top\boldsymbol{u},\, \boldsymbol{\alpha}\rangle \bigr\}.$$

The question is now for which tuning parameters the lasso mimics the ideal estimator $\boldsymbol{\beta}^*$ despite of the extra term in the objective function.

It turns out that a sufficient[1] condition for accurate mimicry of $\boldsymbol{\beta}^*$ is $r \geq 2\|X^\top\boldsymbol{u}\|_\infty$. Heuristically, we can see this as follows: since

$$-2\langle X^\top\boldsymbol{u},\, \boldsymbol{\alpha}\rangle \;\leq\; 2\|X^\top\boldsymbol{u}\|_\infty\|\boldsymbol{\alpha}\|_1$$

by Hölder's inequality, the mentioned condition guarantees that the prior term $r\|\boldsymbol{\alpha}\|_1$ dominates the empirical process term $2\langle X^\top\boldsymbol{u},\, \boldsymbol{\alpha}\rangle$, and consequently, that the objective function of $\widehat{\boldsymbol{\beta}}$ resembles the one of $\boldsymbol{\beta}^*$. We will substantiate this argument later in the theory chapters. Nevertheless, since the effective noise $2\|X^\top\boldsymbol{u}\|_\infty$ is inaccessible in applications, these theoretical insights cannot be directly translated into practical calibration, that is, the development of calibration schemes needs additional thought.

The outline of the chapter is now as follows: In Section 4.2, we study the condition $r \geq 2\|X^\top\boldsymbol{u}\|_\infty$ from a probabilistic point of view. In Sections 4.3 and 4.4, we discuss cross-validation, the arguably most popular calibration scheme. In Section 4.5, we discuss adaptive validation, an alternative scheme based on oracle inequalities.

4.2 The Lasso Noise Term*

The lasso noise terms $2\|X^\top \boldsymbol{u}\|_\infty$ connect our purely deterministic proofs of lasso guarantees (see especially Sections 6 and 7) with the probabilistic nature of the linear regression model. Even though the random vector $X^\top \boldsymbol{u}$ is unknown in practice, this connection is of outmost practical importance: through the condition $r \geq 2v\|X^\top \boldsymbol{u}\|_\infty$, it brings us to the question of how to calibrate the tuning parameter r for given data. This section will show what to expect from a successful calibration, while actual schemes will then be introduced in the subsequent parts of this chapter.

The key concepts of this section are *deviation inequalities*, which are one-sided tail bounds for random variables. In our treatment, we are interested in bounding the random variable $2\|X^\top \boldsymbol{u}\|_\infty$ (or similarly $2v\|X^\top \boldsymbol{u}\|_\infty$ with a constant $v \in (0,\infty)$) from above, thereby getting a sense for what range of tuning parameters r the condition $r \geq 2\|X^\top \boldsymbol{u}\|_\infty$ is satisfied. Such bounds, of course, must depend on the distribution of the noise. Consequently, also the ranges of suitable tuning parameters must depend on the noise; on a high-level, the "stronger" the noise, the larger the tuning parameters need to be.

We illustrate the main aspects with the help of a simple result for Gaussian noise.[2]

Lemma 4.2.1 (Gaussian Noise)

Consider a fixed design matrix $X \in \mathbb{R}^{n\times p}$ and Gaussian noise $\boldsymbol{u} \sim \mathcal{N}_n[\mathbf{0}_n, \sigma^2\,\mathrm{I}_{n\times n}]$. Then, for any $t \in (0,1]$ and $r_t := \sigma\sqrt{8c_{\max}\log[p/t]}$, where $c_{\max} := \max_{j\in\{1,\dots,p\}}(X^\top X)_{jj}$ is a normalization constant, it holds that

$$\mathbb{P}\bigl\{2\|X^\top \boldsymbol{u}\|_\infty \geq r_t\bigr\} \;\leq\; t\,.$$

This deviation inequality bounds the upper tails of the noise term $2\|X^\top \boldsymbol{u}\|_\infty$. The most important consequence for us is that the condition $r \geq 2\|X^\top \boldsymbol{u}\|_\infty$ is satisfied with probability at least $1-t$ for any tuning parameter that is sufficiently large, namely for any (random or deterministic) tuning parameter that is at least as large as the (deterministic) bound r_t.

The normalization constant $c_{\max}$ sets the scale of the noise with respect to the predictors. The standard deviation σ characterizes the strength of the noise: the larger σ, the stronger the noise, and consequently, the larger the bound r_t.

The parameter t enters the bound in two ways. On the one hand, it specifies to what probability the bound holds: the smaller t, the higher the probability. On the other hand, it also specifies the lower bound for the noise term: the smaller t, the larger r_t. The parameter t is, therefore, closely related to the level of a hypothesis test; the smaller the level of a hypothesis test, the more likely a correct null hypothesis is accepted, but also the smaller the power of the test.

The result looks like a manual for calibrating the lasso tuning parameter. Indeed, setting $r = r_t$ with a parameter t that suits a practitioner's needs seems like a good idea. However, such an approach is not practical for two main reasons: First, the deviation inequality presumes i.i.d. Gaussian noise with variance σ^2, but the exact type of distribution—and even more so its variance—are rarely known in practice. Second, even in the unlikely case where all such information is available, deviation inequalities do not necessarily lead to optimal results, as the inequalities are typically too loose.

Instead of addressing the practical calibration of r, deviation inequalities as the one above instead demonstrate that there usually *exist* reasonably small, deterministic tuning parameters that satisfy the condition $r \geq 2\|X^\top \boldsymbol{u}\|_\infty$ (and similarly $r \geq 2v\|X^\top \boldsymbol{u}\|_\infty$, $v \geq 1$—see 1. in Exercise 4.3) on an event that has high probability. Such deviation inequalities then allow one to understand the conditions and results of the bounds derived in the sequel of this book as functions of n, p, and σ. One example is as follows:

Corollary 4.2.1 (High-level Estimation Bound for the Lasso)

Assume that the conditions of Lemma 4.2.1 hold with some fixed normalization constant $c_{\max} \in [0,\infty)$ and that the lasso satisfies in a linear regression model for all $r \geq 2\|X^\top \boldsymbol{u}\|_\infty$ and a $a \in (0,\infty)$ the ℓ_1-estimation bound $\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}[r]\|_1 \leq ar$—cf. Theorem 7.2.1, for example. Then, for any $n \in \{1,2,\dots\}$, $p \in \{2,3,\dots\}$, and $\sigma \in (0,\infty)$, the deterministic tuning parameter r_t meets the bound $r_t \geq 2\|X^\top \boldsymbol{u}\|_\infty$ with probability at least $1-t$, and consequently, the lasso with this tuning parameter meets the bound

$$\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_1 \;\leq\; a\sigma\sqrt{\frac{\log p}{n}}$$

with probability at least $1-t$.

This exemplary result shows what precision can be expected from the lasso if the tuning parameters are calibrated well. Not surprisingly, the precision deteriorates if the standard deviation of the noise σ increases: the stronger the noise, the more it obscures the fundamental relationships in the data. Not surprising either is that the precision increases with increasing sample sizes; the rate $1/\sqrt{n}$ is the classical parametric rate. Finally, also as expected, the precision deteriorates as the number of parameters increases: the more parameters, the more difficult it is to distinguish between relevant and irrelevant ones. Importantly, however, the number of parameters enters the bound only *logarithmically*, reflecting the fact that accurate inference with the lasso is possible even for high-dimensional settings where $p \gg n$.

The bound is guaranteed only with probability $1-t<1$. Probability one cannot be achieved because of potentially adverserial instances of $\boldsymbol{u}$. However, such instances are uncommon: for example, setting $t = 0.01$ leads to a bound that holds in at least 99% of all cases. Abbreviated, we often say that such bounds hold *with high probability.*

with high probability

We conclude this section by proving the lemma.

Proof of Lemma 4.2.1. The key tool in the proof is the *union bound*; the rest is simple algebra.

By definition of the sup-norm, it holds that $2\|X^\top \boldsymbol{u}\|_\infty = \max_{j\in\{1,\dots,p\}} 2|(X^\top \boldsymbol{u})_j|$. We can, therefore, write the event in question as

$$\bigl\{2\|X^\top \boldsymbol{u}\|_\infty \geq r_t\bigr\} \;=\; \bigcup_{j=1}^{p}\bigl\{2|(X^\top \boldsymbol{u})_j| \geq r_t\bigr\}\,.$$

Applying the union bound (see Exercise 4.1) to the events on the right-hand side then yields

$$\mathbb{P}\bigl\{2\|X^\top \boldsymbol{u}\|_\infty \geq r_t\bigr\} \;\leq\; \sum_{j=1}^{p}\mathbb{P}\bigl\{2|(X^\top \boldsymbol{u})_j| \geq r_t\bigr\}\,.$$

We can now use the basic fact $\sum_{j=1}^{k} a_j \leq k \max_{j\in\{1,\dots,k\}} a_j$ for any numbers $a_1,\dots,a_k \in \mathbb{R}$, $k \in \{1,2,\dots\}$, to deduce

$$\mathbb{P}\bigl\{\,2\|X^\top\boldsymbol{u}\|_\infty \geq r_t\,\bigr\} \;\leq\; p \max_{j\in\{1,\dots,p\}} \mathbb{P}\bigl\{\,2|(X^\top\boldsymbol{u})_j| \geq r_t\,\bigr\}\,.$$

With some elementary calculations, we can massage this further as follows (verify that we can assume without loss of generality $(X^\top X)_{jj} > 0$, $j \in \{1,\dots,p\}$):

$$\begin{aligned}
&\mathbb{P}\bigl\{\,2\|X^\top\boldsymbol{u}\|_\infty \geq r_t\,\bigr\}\\
&\leq\; p \max_{j\in\{1,\dots,p\}} \mathbb{P}\biggl\{\,\frac{|(X^\top\boldsymbol{u})_j|}{\sigma\sqrt{c_{\max}}} \geq \frac{r_t}{2\sigma\sqrt{c_{\max}}}\,\biggr\}\\
&\qquad\text{“dividing inside the probability in the previous display by } 2\sigma\sqrt{c_{\max}} > 0\text{”}\\
&\leq\; p \max_{j\in\{1,\dots,p\}} \mathbb{P}\biggl\{\,\frac{|(X^\top\boldsymbol{u})_j|}{\sigma\sqrt{(X^\top X)_{jj}}} \geq \frac{r_t}{2\sigma\sqrt{c_{\max}}}\,\biggr\}\,.\\
&\qquad\text{“definition of } c_{\max} \text{ in the above Lemma 4.2.1; } \mathbb{P}\{\mathcal{A}\} \leq \mathbb{P}\{\mathcal{B}\} \text{ for } \mathcal{A} \subset \mathcal{B}\text{”}
\end{aligned}$$

Recalling that $\boldsymbol{u} \sim \mathcal{N}_n[0, \sigma^2\,\mathrm{I}_{n\times n}]$ by assumption, we observe that the random variables $(X^\top\boldsymbol{u})_j/(\sigma\sqrt{(X^\top X)_{jj}}) = \langle\boldsymbol{x}_j,\,\boldsymbol{u}\rangle/(\|\boldsymbol{x}_j\|_2\sqrt{\sigma^2})$ for $j \in \{1,\dots,p\}$, $X = (\boldsymbol{x}_1,\dots,\boldsymbol{x}_p)$, follow a standard normal distribution, and any standard normal variable $z \sim \mathcal{N}_1[0,1]$ fulfills the deviation inequality (see Exercise 4.2)

$$\mathbb{P}\{\,|z| \geq a\,\} \;\leq\; e^{-\frac{a^2}{2}} \qquad\qquad (\,a \geq 0\,)\,.$$

Combining this deviation inequality at $a = r_t/(2\sigma\sqrt{c_{\max}/n})$ and the above display yields

$$\mathbb{P}\bigl\{\,2\|X^\top\boldsymbol{u}\|_\infty \geq r_t\,\bigr\} \;\leq\; pe^{-\left(\frac{r_t}{2\sigma\sqrt{c_{\max}}}\right)^2/2}\,.$$

We finally plug in the definition of r_t in the above Lemma 4.2.1 to derive

$$\mathbb{P}\bigl\{\,2\|X^\top\boldsymbol{u}\|_\infty \geq r_t\,\bigr\} \;\leq\; pe^{-\left(\frac{\sigma\sqrt{8c_{\max}\log[p/t]}}{2\sigma\sqrt{c_{\max}}}\right)^2/2} \;=\; t\,,$$

as desired. □

4.3 Cross-Validation

Cross-validation calibrates tuning parameters for prediction. For this, it repeatedly splits the samples into two disjoint sets: a training set for fitting parameters and a validation set for estimating the prediction errors of those parameters. It then selects a tuning parameter that minimizes the average of those estimated errors, trusting that (i) the fitted parameters resemble the original estimators and (ii) the estimated errors resemble the true ones.

Within a family of linear regression estimators $\{\widehat{\boldsymbol{\beta}}[r] : r \in \mathcal{R}\} \subset \mathbb{R}^p$ indexed by a set of tuning parameters $\mathcal{R}$, an optimal choice for prediction is $\widehat{\boldsymbol{\beta}}[r^*]$ with

$$r^* \;\in\; \operatorname*{argmin}_{r\in\mathcal{R}} f^*[r]\,, \quad \text{where } f^*[r] \;:=\; \frac{1}{n}\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[r]\|_2^2\,. \tag{4.1}$$

Here, we set f^* equal to the *average* prediction error—see the scaling $1/n$—which has no influence on the minimization but facilitates the comparison between true and estimated prediction errors later on. Of course in practice, the target $\boldsymbol{\beta}$ is unknown, which means that the optimal tuning parameter r^* is unknown as well. Our goal is to leverage data $(\boldsymbol{y}, X)$ for finding a tuning parameter $\widehat{r} \equiv \widehat{r}[\boldsymbol{y}, X]$ that has a performance similar to r^*, that is, $\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[\widehat{r}]\|_2^2/n \approx \|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[r^*]\|_2^2/n$. A sufficient condition for this is $\widehat{r} \approx r^*$, which can be achieved by

$$\widehat{r} \;\in\; \operatorname*{argmin}_{r \in \mathcal{R}} \widehat{f}[r] \tag{4.2}$$

if $\widehat{f}[r] \approx f^*[r]$ for all $r \in \mathcal{R}$. The agenda is now to establish such $\widehat{f}[r]$.

The seemingly most natural estimate of the error is

$$\widehat{f}[r] \;:=\; \frac{1}{n}\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}[r]\|_2^2\,,$$

but since estimators typically try to minimize $\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}[r]\|_2^2$ already, this function can give a very distorted picture of the true prediction error. For example, least-squares estimators with norm regularizers

$$\widehat{\boldsymbol{\beta}}[r] \;\in\; \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^p}\bigl\{\,\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}[r]\|_2^2 + r\|\boldsymbol{\alpha}\|\,\bigr\}$$

always satisfy $\widehat{f}[0] = 0$ if $\boldsymbol{y}$ is in the column space of X; hence $\widehat{r} = 0$, which means that one would always select the plain least-squares estimator.

One way to improve on the described choice of $\widehat{f}$ is to add a term that takes the complexity of the selected model into account:

$$\widehat{f}[r] \;:=\; \frac{1}{n}\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}[r]\|_2^2 + \text{complexity}\bigl[\widehat{\boldsymbol{\beta}}[r]\bigr]\,.$$

The additional term should favor simple models, therefore yielding a trade-off between a good model fit (typically achieved with small r) and low model complexity (typically achieved with large r). Similarly as as the complexity measures in oracle inequalities, see Section 6.6, the complexity term here is not subject to the computational constraints that limit the choices of prior terms h in the objective functions of estimators $\widehat{\boldsymbol{\beta}}$. This is because $\widehat{f}$ is a function of the real-valued r, while the objective functions of regularized estimators are functions of the vector-valued parameters $\boldsymbol{\alpha}$. Examples for corresponding approaches include *AIC*, *BIC*, and *Mallow's* C_p.

However, if we aim purely at prediction—and not at a trade-off between prediction accuracy and model complexity (cf. our discussion on Pages 65ff), such additional terms seem unsuitable. Therefore, we focus in this section on a different way to improve the initial choice of $\widehat{f}$. Recall that the key issue with the seemingly natural estimate of the error is that the same data is used for both estimating the the target and for estimating the error. This can create a dependency between the estimates of the target and the error that leads to an inept comparison of tuning parameters; Often, the selected tuning parameters are too small, such as for the brown curve in Figure **??** and seen in the above example of norm-regularized estimators. The *holdout method* tries to alleviate the dependency through data splittling: instead of using all data for both prediction and for measuring the prediction accuracy, each data point is used only for either one of these tasks. Given data $Z^1, \dots, Z^n$, the estimators get to see only a *training set* $\{Z^i : i \in \mathcal{T}\}$, $\mathcal{T} \subset \{1, \dots, n\}$ with size $n^{\mathfrak{t}}$, and the assessment of the accuracy is performed on the remaining *validation set* (or holdout set) $\{Z^i : i \in \mathcal{V}\}$, $\mathcal{V} := \{1, \dots, n\} \setminus \mathcal{T}$ with size $n^{\mathfrak{v}} = n - n^{\mathfrak{t}}$. The assigments of the points to the training

holdout method

training and validation/holdout sets

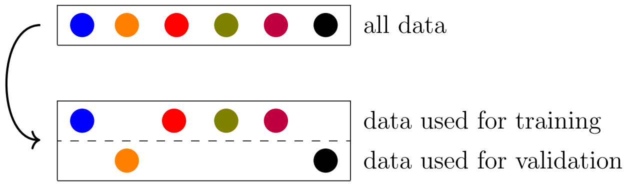


Figure 4.1: In the holdout method, the data is partioned randomly or non-randomly into two parts: one part is used only for training the estimators; one part is used only for validating the estimators.

and validation sets can be random or fixed. An illustration for $n = 6$ data points is given in Figure 4.1. In linear regression, where $Z^i = (y^i, \boldsymbol{x}^i)$, we find

$$\widehat{r} \;\in\; \operatorname*{argmin}_{r \in \mathcal{R}} \widehat{f}[r]$$

with

$$\widehat{f}[r] \;:=\; \frac{1}{n^{\mathfrak{v}}} \|\boldsymbol{y}_{\mathcal{V}} - X_{\mathcal{V}} \widehat{\boldsymbol{\beta}}[r, \boldsymbol{y}_{\mathcal{T}}, X_{\mathcal{T}}]\|_2^2\,,$$

where $\boldsymbol{y}_{\mathcal{T}} \in \mathbb{R}^{n^{\mathfrak{t}}}$ and $X_{\mathcal{T}} \in \mathbb{R}^{n^{\mathfrak{t}} \times p}$ are the vector/matrix-valued training data, $\boldsymbol{y}_{\mathcal{V}} \in \mathbb{R}^{n^{\mathfrak{v}}}$ and $X_{\mathcal{V}} \in \mathbb{R}^{n^{\mathfrak{v}} \times p}$ are the vector/matrix-valued validation data, and $\widehat{\boldsymbol{\beta}}[r, \boldsymbol{y}_{\mathcal{T}}, X_{\mathcal{T}}] \in \mathbb{R}^p$ is the estimator evaluated on the training data.

Monte Carlo cross-validation

The holdout method can be sensitive to the specific separation into training and testing sets. Cross-validation schemes attempt to alleviate this issue through the use of multiple splits. The simplest representative is *Monte Carlo cross-validation* (also called repeated random sub-sampling validation). It generates $k \in \{1, 2, \dots\}$ splits uniformly at random into training sets $\mathcal{T}_1, \dots, \mathcal{T}_k$ of size $n^{\mathfrak{t}} \in \{1, \dots, n-1\}$ and validation sets $\mathcal{V}_1, \dots, \mathcal{V}_k$ of size $n^{\mathfrak{v}} := n - n^{\mathfrak{t}}$ and then applies the holdout method on each pair of training and validation sets. An illustration of the data splitting approach for $k = 3$ and $n^{\mathfrak{t}} = 4$ is given in Figure 4.2. The tuning parameter is then selected as a minimizer of the average (or median) error. In linear regression, one gets

$$\widehat{r} \;\in\; \operatorname*{argmin}_{r \in \mathcal{R}} \widehat{f}[r]$$

with

$$\widehat{f}[r] \;:=\; \frac{1}{kn^{\mathfrak{v}}} \sum_{j=1}^{k} \|\boldsymbol{y}_{\mathcal{V}_j} - X_{\mathcal{V}_j} \widehat{\boldsymbol{\beta}}[r, \boldsymbol{y}_{\mathcal{T}_j}, X_{\mathcal{T}_j}]\|_2^2\,.$$

Empirically, Monte Carlo cross-validation is often more stable than the holdout method, because the averaging alleviates the dependence on how the training and validation sets are configured.

k-fold cross-validation

Another variant is *k-fold cross-validation*. In this approach, the data are partitioned at random or not at random into $k \in \{2, 3, \dots\}$ sets $\mathcal{A}_1, \dots, \mathcal{A}_k$ of equal size n/k (or approximately of that size if n is not divisible by k). Then, k training and validation sets are defined by $\mathcal{T}_j := \{1, \dots, n\} \setminus \mathcal{A}_j$ and $\mathcal{V}_j := \mathcal{A}_j$, $j \in \{1, \dots, k\}$. An illustration for $k = 3$ is provided in Figure 4.3. Two differences to Monte Carlo cross-validation are that the size of the training and validation sets are fixed for a given number of folds k and that each point is used for validation exactly once. Otherwise, the procedure is as

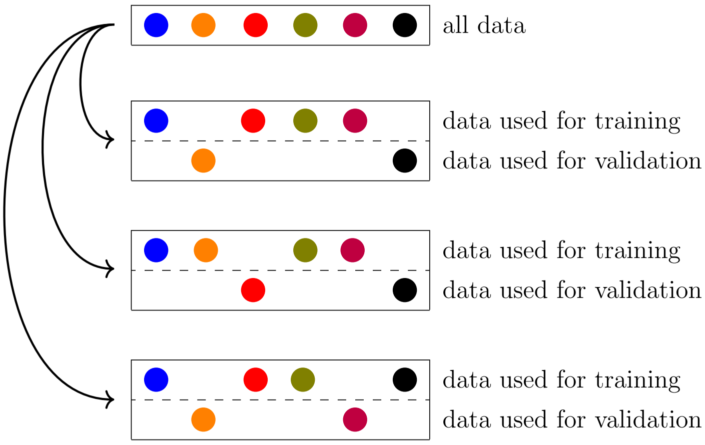


Figure 4.2: In Monte Carlo cross-validation, the data is split randomly multiple times.

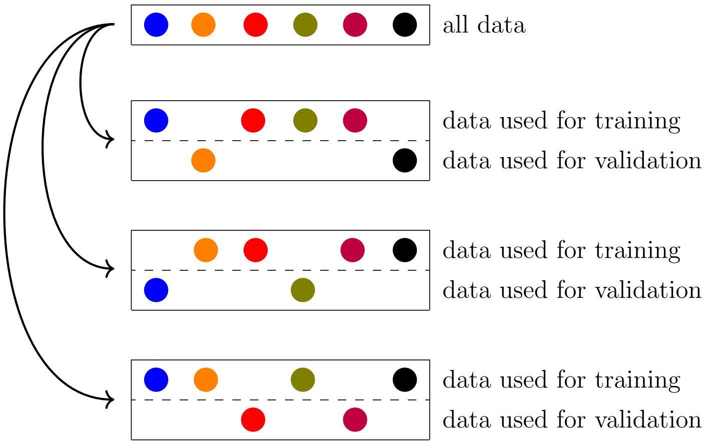


Figure 4.3: In k-fold cross-validation, the data is split randomly or non-randomly k times such that the size of each validation set is n/k and each data point is used for validation exactly once.

before. The special case $k = n$, which leads to n-validations that each use one different data point, is referred to as *leave-one-out cross-validation.*

leave-one-out cross-validation

Cross-validation schemes often yield good empirical prediction results. Not suprisingly, the selected models are typically complex: cross-validation aims purely at prediction accuracy, rather than at a trade-off between prediction performance and low model complexity. This tendency to “overfit” needs to be taken into account when interpreting the resulting models, and more generally, one has to bear in mind that cross-validation is not designed for anything else than prediction.[3]

Cross-validation schemes are also subject to a number of other limitations. First, the holdout approach implicitly assumes that the training and validation data are independent. In practice, this assumption is often violated: for example, time series data is highly dependent by design.

Next, cross-validation uses data and computing resources inefficently. By design of the holdout scheme, the training and the validation steps use only subsets of the data—

which means that the effective samples sizes are much smaller than the actual sample size n. Also, each application of one of the discussed cross-validation schemes requires k training and validations in addition to the computation of the actual estimator.

Then, there is no consistent theory for cross-validation in high-dimensional statistics. In particular, even for prediction, there is no comprehensive finite sample guarantee for lasso-type estimators calibrated by cross-validation.[4]

Moreover, the splitting procedures often comprise random elements: in k-fold cross-validation, for example, the partitioning of the data points is typically performed uniformly at random. Unless the randomness is "fixed" (by using the same random seed in all users' programs, for example), different users can then get different outputs $\widehat{\boldsymbol{\beta}}[r]$ on the exact same data.

Finally, cross-validation methods trade the original tuning parameters for new tuning parameters. For example, Monte Carlo cross-validation contains two free parameters, the number of data splits and the size of the training set; k-fold cross-validation contains one free parameter, the number of folds/data splits. Just as the original tuning parameters, these parameters are subject to trade-offs. Heuristically, the larger the number of data splits, the more stable but computationally demanding the methods are (because the specific compositions of the training and validation sets are averaged out but the training and validation results have to be computed for each split). The larger the training set, the smaller the bias but the larger the variance are (because the approximation of $\widehat{\boldsymbol{\beta}}[r]$ is more accurate but the estimation of the error more volatile). For k-fold cross-validation, $k \in \{5, 10\}$ have become standard parameters in this trade-off, but there are few theoretical justifications for any specific choice.

Still, empirical evidence shows that cross-validation schemes can be relatively stable with respect to the choice of their parameters. Therefore, if data and computational resources are enough, and if the data points are fairly independent, cross-validation can be a practical tool for prediction. One can also replace the prediction loss by the missclassification rate to use cross-validation in classification ($\boldsymbol{y}$ takes only finitely many values).

4.4 Further Insights Into Cross-Validation*

One take-away from the previous section is that we can expect cross-validation to select suitable tuning parameters if the original estimators' prediction errors are estimated well. In this section, we discuss this estimation further. We show in particular that under some conditions, the cross-validation error is indeed a good approximation of the unknown prediction error. We then discuss the roles of the number of folds and of the sample sizes of the training and validation sets.

We start by formalizing the repeated data splitting. For this, we need to assume that the data $Z \in \mathcal{Z}$ consists of n individual points $\boldsymbol{z}^1, \dots, \boldsymbol{z}^n$. In each of the splits $j \in \{1, \dots, k\}$, these data points are separated into a training set $Z^{\mathfrak{t}}_{o_j} := (\boldsymbol{z}^{o_j[1]}, \dots, \boldsymbol{z}^{o_j[n^{\mathfrak{t}}]})$ and a validation set $Z^{\mathfrak{v}}_{o_j} := (\boldsymbol{z}^{o_j[n^{\mathfrak{t}}+1]}, \dots, \boldsymbol{z}^{o_j[n^{\mathfrak{t}}+n^{\mathfrak{v}}]})$, where $n^{\mathfrak{t}} \in \{1, \dots, n-1\}$ is the size of the training set, $n^{\mathfrak{v}} := n - n^{\mathfrak{t}}$ the size of the validation set, and $o_1, \dots, o_k$ are permutations on $\{1, \dots, n\}$. In k-fold cross-validation, for example, k is the number of folds, $n^{\mathfrak{t}} = n - n/k$, $n^{\mathfrak{v}} = n/k$, and the permutations are such that each index finds itself permuted exactly once to among the last $n^{\mathfrak{v}}$ positions.

By leveraging the holdout approach on these data splits, cross-validation aims to select an optimal estimator among the options $\{\widehat{\boldsymbol{\beta}}[r] : r \in \mathcal{R}\}$. Recall that the first step is to approximate the estimators by surrogates $\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}]$ that use only the training data. The second step is to approximate the error $f^*[r]$ by an estimate $\widehat{f}[r]$ that aggregates the errors of the surrogates $\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r]$ on the holdout sets in a loss ℓ:

$$\widehat{f}[r] \;=\; \frac{1}{k}\sum_{j=1}^{k}\ell\Bigl[\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}], Z^{\mathfrak{v}}_{o_j}\Bigr]\,. \tag{4.3}$$

Losses of the form $\ell[\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}], Z^{\mathfrak{v}}_{o_j}]$ measure *out-of-sample errors.* If the observations are identically distributed, cross-validation provides an unbiased view on the ℓ-performance of the estimators trained on $n^{\mathfrak{t}}$ data points on a set of $n^{\mathfrak{v}}$ new data points. Since the accuracy of estimators generally increases with the sample size, cross-validation errors typically overestimate the out-of-sample errors of the original estimators, which are trained on all n observations.

In linear regression with i.i.d. data, we can connect the cross-validation error also with the initial (in-sample) prediction error. Consider regression-type data $\boldsymbol{z}^1 = (y^1, \boldsymbol{x}^1), \dots, \boldsymbol{z}^n = (y^n, \boldsymbol{x}^n)$. According to the definition of the (average) prediction loss, the cross-validation error is

$$\widehat{f}[r] \;=\; \frac{1}{k}\sum_{j=1}^{k}\ell\Bigl[\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}], Z^{\mathfrak{v}}_{o_j}\Bigr] \;=\; \frac{1}{kn^{\mathfrak{v}}}\sum_{j=1}^{k}\sum_{i\in\mathcal{V}_j}\bigl(y^i - \boldsymbol{x}^{i\top}\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}]\bigr)^2\,.$$

If the validation sample size is large enough and the data are i.i.d., we could hope that

$$\frac{1}{n^{\mathfrak{v}}}\sum_{i\in\mathcal{V}^j}\bigl(y^i - \boldsymbol{x}^{i\top}\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}]\bigr)^2 \;\approx\; \frac{1}{n}\sum_{i=1}^{n}\bigl(y^i - \boldsymbol{x}^{i\top}\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}]\bigr)^2 \;=\; \frac{1}{n}\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}]\|_2^2\,,$$

that is, the holdout losses averaged over the validation set are close to the naive prediction error averaged over all samples. Invoking the model, the right-hand side can be written as

$$\frac{1}{n}\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}]\|_2^2 \;=\; \frac{1}{n}\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}]\|_2^2 + \frac{2}{n}\langle\boldsymbol{u},\, X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}]\rangle + \frac{1}{n}\|\boldsymbol{u}\|_2^2\,.$$

If the noise is centered, one might further hope that the second term is neglectible, that is,

$$\widehat{f}[r] \;\approx\; \frac{1}{kn}\sum_{j=1}^{k}\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}]\|_2^2 + \frac{1}{n}\|\boldsymbol{u}\|_2^2\,.$$

If the training sample size is large and the data are i.i.d., there should be little variance among the k splits. Therefore, since the second term is independent of the tuning parameter, we could expect that

$$\operatorname*{argmin}_{r\in\mathcal{R}}\widehat{f}[r] \;\approx\; \operatorname*{argmin}_{r\in\mathcal{R}}\frac{1}{n}\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_1}]\|_2^2\,.$$

If additionally the training set is large, one might hope that $\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_1}]$ is a good approximation of the actual estimator $\widehat{\boldsymbol{\beta}}[r]$, so that

$$\operatorname*{argmin}_{r\in\mathcal{R}}\widehat{f}[r] \;\approx\; \operatorname*{argmin}_{r\in\mathcal{R}}f^*[r]\,,$$

as desired. These considerations provide further support for cross-validation. However, making the handwavy arguments more precise has proved challenging, especially in the case of finite samples.

We finally discuss the effects of the parameters k, $n^{\mathfrak{t}}$, and $n^{\mathfrak{v}}$ on the variance of the cross-validation error. First, according to Exercise 4.6 and the lab, the variance of the cross-validation error is typically decreasing in the number of splits k if everything else is fixed. This means that in Monte Carlo cross-validation, where the number of splits is disentangled from the other parameters, more splits typically lead to smaller variance.

To discuss the two other parameters, we apply Lemma B.3.1 (law of total variance) in the Appendix to separate the variance of each summand in the validation error into two terms:

$$\begin{aligned}&\mathbb{V}\mathrm{ar}\Bigl[\ell\Bigl[\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}], Z^{\mathfrak{v}}_{o_j}\Bigr]\Bigr]\\ &\quad= \mathbb{E}\Bigl[\mathbb{V}\mathrm{ar}\Bigl[\ell\Bigl[\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}], Z^{\mathfrak{v}}_{o_j}\Bigr]\Big|Z^{\mathfrak{t}}_{o_j}\Bigr]\Bigr] + \mathbb{V}\mathrm{ar}\Bigl[\mathbb{E}\Bigl[\ell\Bigl[\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}], Z^{\mathfrak{v}}_{o_j}\Bigr]\Big|Z^{\mathfrak{t}}_{o_j}\Bigr]\Bigr]\,.\end{aligned}$$

The first term measures the variance in estimating the prediction error of a given surrogate $\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}$ of the initial estimator $\widehat{\boldsymbol{\beta}}$; the second term measures the variance in estimating the surrogate itself. For illustration, we study these two terms in the context of linear regression. With ℓ the average prediction loss, the first term becomes

$$\mathbb{E}\Bigl[\mathbb{V}\mathrm{ar}\Bigl[\ell\Bigl[\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}], Z^{\mathfrak{v}}_{o_j}\Bigr]\Big|Z^{\mathfrak{t}}_{o_j}\Bigr]\Bigr] \;=\; \mathbb{E}\Bigl[\mathbb{V}\mathrm{ar}\Bigl[\frac{1}{n^{\mathfrak{v}}}\sum_{i\in\mathcal{V}_j}\bigl(y^i - \boldsymbol{x}^{i\top}\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}]\bigr)^2\Big|Z^{\mathfrak{t}}_{o_j}\Bigr]\Bigr]\,.$$

If the data is i.i.d., we can simplify this to

$$\mathbb{E}\Bigl[\mathbb{V}\mathrm{ar}\Bigl[\ell\Bigl[\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}], Z^{\mathfrak{v}}_{o_j}\Bigr]\Big|Z^{\mathfrak{t}}_{o_j}\Bigr]\Bigr] \;=\; \frac{1}{n^{\mathfrak{v}}}\mathbb{E}\Bigl[\mathbb{V}\mathrm{ar}\Bigl[\bigl(y^1 - \boldsymbol{x}^{1\top}\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}]\bigr)^2\Big|Z^{\mathfrak{t}}_{o_j}\Bigr]\Bigr]\,.$$

This shows that, everything else fixed, the first term is proportional to $1/n^{\mathfrak{v}}$. Thus, the larger the validation size, the smaller the first part of the variance.

The second term becomes

$$\mathbb{V}\mathrm{ar}\Bigl[\mathbb{E}\Bigl[\ell\Bigl[\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}], Z^{\mathfrak{v}}_{o_j}\Bigr]\Big|Z^{\mathfrak{t}}_{o_j}\Bigr]\Bigr] \;=\; \mathbb{V}\mathrm{ar}\Bigl[\mathbb{E}\Bigl[\frac{1}{n^{\mathfrak{v}}}\sum_{i\in\mathcal{V}_j}\bigl(y^i - \boldsymbol{x}^{i\top}\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}]\bigr)^2\Big|Z^{\mathfrak{t}}_{o_j}\Bigr]\Bigr]\,,$$

which is equivalent—assuming again i.i.d. data—to

$$\mathbb{V}\mathrm{ar}\Bigl[\mathbb{E}\Bigl[\ell\Bigl[\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}], Z^{\mathfrak{v}}_{o_j}\Bigr]\Big|Z^{\mathfrak{t}}_{o_j}\Bigr]\Bigr] \;=\; \mathbb{V}\mathrm{ar}\Bigl[\mathbb{E}\Bigl[\bigl(y^1 - \boldsymbol{x}^{1\top}\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}]\bigr)^2\Big|Z^{\mathfrak{t}}_{o_j}\Bigr]\Bigr]\,.$$

The estimates $\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}]$ can be expected to become more stable with increasing size of the training set, which means that—everything else fixed—the right-hand side in the penultimate display generally decreases with $n^{\mathfrak{t}}$ increasing. Thus, the larger the training size, the smaller the second part of the variance. In summary, the variance of the cross-validation error generally decreases in all of k, $n^{\mathfrak{t}}$, and $n^{\mathfrak{v}}$. However, besides the variance being only one among many possible aspects of a method, it remains difficult to understand what these insights mean in practice, where data and computatational resources are finite.

4.5 Adaptive Validation

4.6 References and Further Reading

[1]That $r \geq 2v\|X^\top\boldsymbol{u}\|_\infty$ is not a sufficient *and necessary* condition for the lasso to perform well has been shown first in [vL13]. They suggest in particular that in

correlated settings, tuning parameters should be smaller than what is suggested by the mentioned condition—see their Corollary 4.2 on Page 308, for example. Further results in this direction have derived in [HL13] and [DHL17]. Hence, although the effective noise $2\|X^\top \boldsymbol{u}\|_\infty$ is intimately connected with the lasso's tuning parameter and with the lasso more generally, it does not account for the entire intricacy of tuning parameter calibration.

[2] A refinement of Lemma 4.2.1 for correlated designs is developed in [HL13, Proof of Theorem 3.2, Pages 16-18]. Related considerations can also be found in [vL13]. Techniques for developing such concentration bounds beyond Gaussian noise can be found in [BLM13].

[3] References for incommensurable goals in estimating $\boldsymbol{\beta}$, including the so-called "AIC-BIC dilemma," can be found in [AC10, Section 2.4 and Section 2.5 on Page 48].

[4] An overview about the existing theory for Cross-Validation and about the corresponding literature is provided in [AC10]. Non-asymptotic expressions for the variance of the (again out-of-sample) risk of projection estimators in regression can be found in [Cel08, Proposition 3.4.3 on Page 66]. Some theoretical bounds for cross-validated lasso can be found in [CJ15, HM13b, HM13a, HM14]. Much more is known about cross-validated ridge regression, see [GHW79] for example.

4.7 Exercises

Exercises for Section 4.2

□ **Exercise 4.1** $^{\diamond\diamond\bullet}$ In this exercise, we establish the union bound (also called Boole's inequality), which we have used in the proof of Lemma 4.2.1.

Show that for any events $\mathcal{A}_1,\dots,\mathcal{A}_k$, $k\in\{1,2,\dots\}$, it holds that

$$\mathbb{P}\Biggl\{\bigcup_{j=1}^{k}\mathcal{A}_j\Biggr\} \;\le\; \sum_{j=1}^{k}\mathbb{P}\{\mathcal{A}_j\}\,.$$

Use only the following two properties of probability measures: (i) finite additivity, that is, $\mathbb{P}\{\mathcal{A}\cup\mathcal{B}\}=\mathbb{P}\{\mathcal{A}\}+\mathbb{P}\{\mathcal{B}\}$ for any disjoint events $\mathcal{A},\mathcal{B}$; (ii) positivity, that is, $\mathbb{P}\{\mathcal{A}\}\ge 0$ for any event $\mathcal{A}$. (That these two properties suffice indicates that the union bounds also holds for functions other than probability measures.)

□ **Exercise 4.2** $^{\diamond\diamond\bullet}$ In this exercise, we derive the deviation inequality for Gaussian random variables that we have used in the proof of Lemma 4.2.1.

Show that any standard normal variable $z\sim\mathcal{N}_1[0,1]$ fulfills

$$\mathbb{P}\bigl\{\,|z| \;\ge\; a\bigr\} \;\le\; e^{-\frac{a^2}{2}}$$

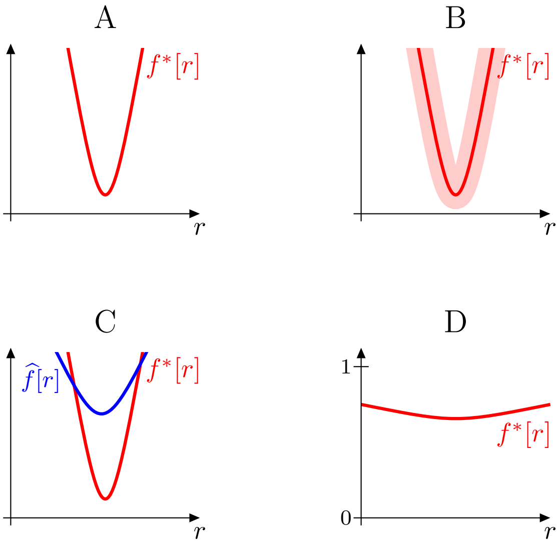
for every $a\in[0,\infty)$.

□ **Exercise 4.3** $^{\diamond}$ In this exercise, we generalize the bound in Lemma 4.2.1.

1. Prove a deviation inequality as in Lemma 4.2.1 for general $2v\|X^\top\boldsymbol{u}\|_\infty$, $v\ge 1$.
2. Prove a deviation inequality as in Lemma 4.2.1 for $2\|X^\top\boldsymbol{u}\|_\infty$ with noise $\boldsymbol{u}$ distributed according to a double-exponential distribution.

Exercises for Section 4.3

□ **Exercise 4.4** $^{\diamond\bullet}$ In this exercise, we study the optimal tuning parameter defined in the minimization (4.1) and its estimation via the minimization (4.2). For this, we consider some general settings in plots A–D:



1. Consider Plot A: Mark the optimal tuning parameter r^*. Motivate that in the depicted case, $\widehat{r} \approx r^*$ yields $f^*[\widehat{r}] \approx f^*[r^*]$.

2. Consider Plot B: Assume that the estimated prediction loss $\widehat{f}$ is close to the true loss f^* in the sense that the values $\widehat{f}[r]$ are in the red shaded area around $f^*[r]$ for all $r \in \mathcal{R}$. Can you restrict the possible values of $\widehat{r}$? Motivate that in the depicted case, $\widehat{f}$ being close to the true loss f^* across all tuning parameters $r \in \mathcal{R}$ yields $\widehat{r} \approx r^*$.

3. Consider Plot C: Mark r^* and $\widehat{r}$. Conclude that $\widehat{f}$ being close to the true loss f^* is not a necessary condition for $\widehat{r}$ being close to r^*.

4. Consider Plot D: Assume that the x-axis covers the entire set of possible tuning parameters $\mathcal{R}$. Argue that then in the depicted setting, any calibration scheme would do fine in the sense that $f^*[\widehat{r}] \approx f^*[r^*]$. Does this require $\widehat{r} \approx r^*$?

Cross-validation aims at finding a tuning parameter $\widehat{r}$ such that $f^*[\widehat{r}] \approx f^*[r^*]$, where f^* is the (scaled) prediction loss: $f^*[r] := \|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[r]\|_2^2/n$, $r \in \mathcal{R}$. Question 1 indicates that a typically case where this holds is $\widehat{r} \approx r^*$, and Question 2 indicates that a typical case where the latter holds is $\widehat{f}[r] \approx f^*[r]$ for all $r \in \mathcal{R}$.

On the other hand, Question 3 shows that $\widehat{f} \approx f^*$ is *not a necessary* condition for $\widehat{r} \approx r^*$, and Question 4 shows that $\widehat{r} \approx r^*$ is *not a necessary* condition for $f^*[\widehat{r}] \approx f^*[r^*]$ either. Hence, the only measure of success is how well $f^*[\widehat{r}]$ approximates $f^*[r^*]$.

□ **Exercise 4.5** $^{\diamond\bullet}$ In this exercise, we highlight the limitations of cross-validation outside prediction.

Assume that we want to find a data-driven version of

$$r^* \in \operatorname*{argmin}_{r \in \mathcal{R}} f^*[r]\,, \quad \text{where } f^*[r] := \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}[r]\|_2^2\,.$$

Is cross-validation suited for this?

Exercises for Section 4.4

□ **Exercise 4.6** $^{\diamond\diamond\bullet}$ In this exercise, we motivate that the variances of cross-validation errors are typically decreasing in the number of splits. For this, we consider random variables $Z^1, \dots, Z^k$ with values on a common measurable space.

1. Show that
$$\mathbb{V}\mathrm{ar}\biggl[\frac{1}{k}\sum_{j=1}^{k} Z^j\biggr] \;\leq\; \frac{1}{k}\sum_{j=1}^{k}\mathbb{V}\mathrm{ar}\bigl[Z^j\bigr]\,,$$
assuming all integrals that are involved exist and are finite. Hint: The Cauchy-Schwarz inequality in Lemma B.1.4 on Page 166 might be helpful.

2. Conclude that if the variances of the individual estimates in (4.3) do not depend on the splitting, that is, there is a constant $\gamma \in [0, \infty)$ such that
$$\mathbb{V}\mathrm{ar}\Bigl[\ell\Bigl[\widehat{\boldsymbol{\beta}}^{\mathfrak{t}}[r, Z^{\mathfrak{t}}_{o_j}], Z^{\mathfrak{v}}_{o_j}\Bigr]\Bigr] \;=\; \gamma^2 \qquad\qquad (\,j\in\{\,1,2,\dots\}\,)\,,$$
then the variances of the average errors are bounded as $\mathbb{V}\mathrm{ar}[\widehat{f}[r]] \leq \gamma^2$ irrespective of the number of splits k.

3. Show that if the random variables $Z^1, \dots, Z^k$ are also mutually independent, it holds that
$$\mathbb{V}\mathrm{ar}\biggl[\frac{1}{k}\sum_{j=1}^{k} Z^j\biggr] \;=\; \frac{1}{k^2}\sum_{j=1}^{k}\mathbb{V}\mathrm{ar}\bigl[Z^j\bigr]\,.$$

4. Conclude that if the variances of the individual errors in (4.3) do not depend on the splitting and are mutually independent, the variances of the average errors are decreasing in the number of splits.

Claim 2 states a setting where the variances of the average errors are bounded by the variance of an individual error, that is, by the variance of the cross-validation error with one single split. Claim 4 states a setting where the variances of the average errors decrease monotonously in the number of splits.

R Lab Chapter 4

4 Cross-validation

In this lab, we study cross-validation. As always, replace the keyword REPLACE with suitable code and answer the questions posed in the text.

4.1 Data generation

Generate data from a linear regression model $\mathbf{y} = X\boldsymbol{\beta} + \mathbf{u}$, where the entries of the design $X \in \mathbb{R}^{100\times 500}$ are sampled independently according to $X_{ij} \sim \mathcal{N}[0,1]$, the target is $\boldsymbol{\beta} = (1,1,1,0,0,\dots,0)^\top \in \mathbb{R}^{500}$, and the noise is sampled independently from the design according to $\mathbf{u} \sim \mathcal{N}_{100}[\mathbf{0}_{100}, \mathrm{I}_{100,100}]$.

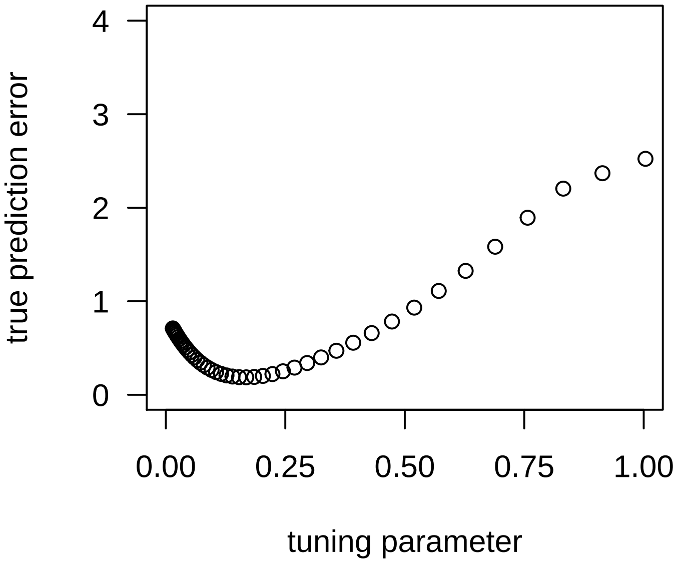
```
set.seed(1)
n <- 100; p <- 500
design <- REPLACE
target <- REPLACE
outcome <- REPLACE
cbind(outcome[1:4], design[1:4, 1:4])
```

```
##            [,1]       [,2]        [,3]       [,4]       [,5]
## [1,] -0.3115278 -0.6264538 -0.62036668  0.4094018  0.8936737
## [2,]  1.4270881  0.1836433  0.04211587  1.6888733 -1.0472981
## [3,]  0.9782890 -0.8356286 -0.91092165  1.5865884  1.9713374
## [4,]  2.6375362  1.5952808  0.15802877 -0.3309078 -0.3836321
```

4.2 Computing a set of estimators

Compute lasso estimates via the `glmnet` package. Let the `glmnet` function generate 50 tuning parameters, and store these tuning parameters for later use. Set the flag `intercept=FALSE` throughout. Use `glmnet`'s standard options otherwise. Plot the average prediction error $\|X\beta - X\widehat{\beta}\|_2^2/n$ as a function of the tuning parameter.

```
set.seed(2)
library(glmnet)
AvgPredictionLoss <- function(estimator, design, target)
{
  return(REPLACE)
}
lasso.fit <- REPLACE
estimators <- lasso.fit$beta
tuning.parameters <- lasso.fit$lambda
estimators.prediction.loss <- apply(estimators, 2, AvgPredictionLoss, design, target)
plot(x    = tuning.parameters,
     y    = estimators.prediction.loss,
     xlim = c(0, 1),
     ylim = c(0, 4),
     xaxp = c(0, 1, 4),
     yaxp = c(0, 4, 4),
     las  = 1,
     xlab = "tuning parameter",
     ylab = "true prediction error")
```

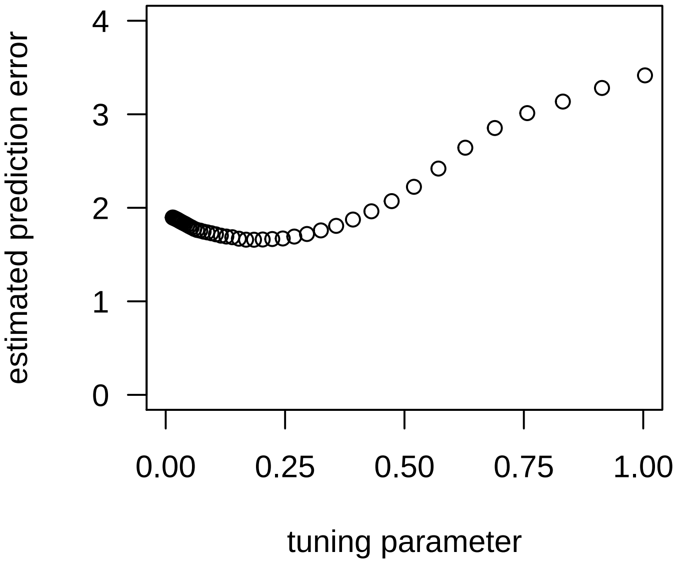


Which tuning parameters are the most favorable ones according to this plot?

4.3 Implementing a Monte Carlo cross-validation

Implement a Monte Carlo cross-validation scheme for selecting among the tuning parameters stored above. Set the training sample size to half the sample size and the number of splits to 3. Plot the estimated errors $\widehat{f}[r]$ as a function of the tuning parameter. The function `setdiff()` could be convinient in determining the validation set.

```
set.seed(3)
IndividualValidationLoss <- function(estimator, outcome, design)
{
  return(REPLACE)
}
AvgPredictionLossEstimated <- function(outcome, design, tuning.parameters, size.training,
                                       number.splits)
{
  estimators.loss <- rep(0, length(tuning.parameters))
  for (j in 1:number.splits)
  {
    indexes.training <- sample(c(1:dim(design)[1]), size.training)
    estimators <- REPLACE
    indexes.validation <- REPLACE
    estimators.loss <- estimators.loss + apply(estimators, 2, IndividualValidationLoss,
                                               outcome[indexes.validation],
                                               design[indexes.validation, ])
  }
  return(REPLACE)
}
plot(x    = tuning.parameters,
     y    = AvgPredictionLossEstimated(outcome, design, tuning.parameters,
                                       dim(design)[1] / 2, 3),
     xlim = c(0, 1),
     ylim = c(0, 4),
     xaxp = c(0, 1, 4),
     yaxp = c(0, 4, 4),
     las  = 1,
     xlab = "tuning parameter",
     ylab = "estimated prediction error")
```

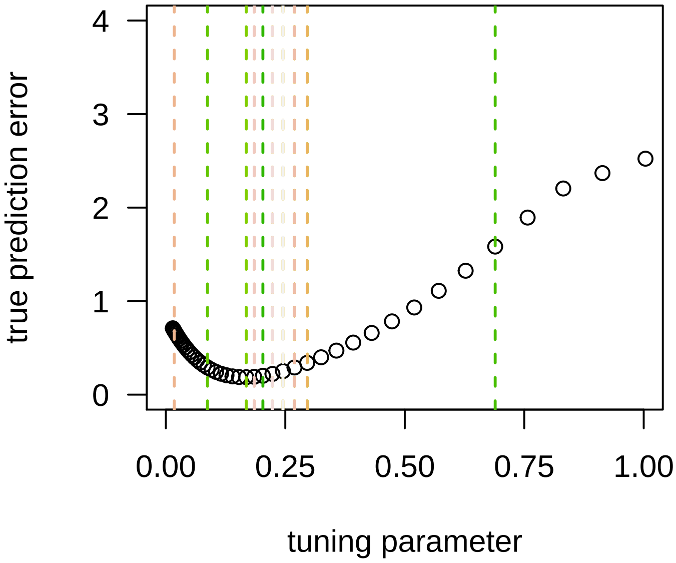


The curve looks a bit like an upward shifted version of the earlier one, which means especially that cross-validation overestimates the true errors in our example. Can you find a reason for this? Can you corroborate your reasoning by running the code with one of the model parameters changed? For the selection of tuning parameters, is an over- or undestimation of the true errors necessarily a problem?

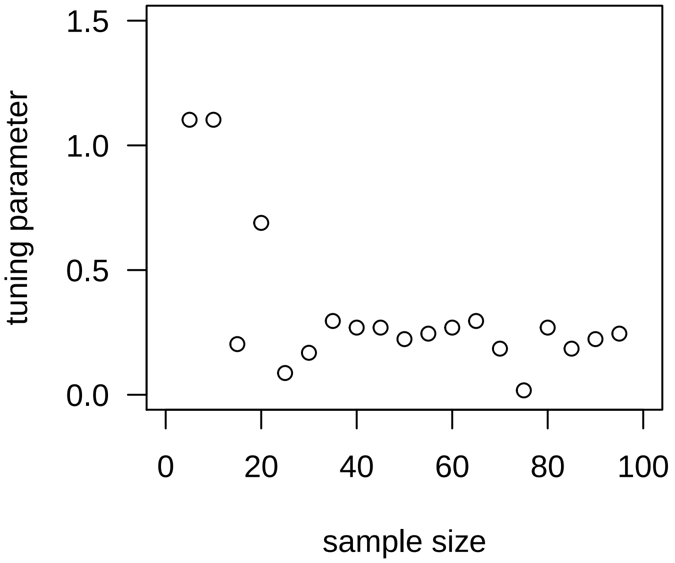
4.4 The double role of the training sample size

Complement the first plot with lines that indicate tuning parameters selected by Monte Carlo cross-validation for different training sample sizes.

```
set.seed(4)
plot(x    = tuning.parameters,
     y    = estimators.prediction.loss,
     xlim = c(0, 1),
     ylim = c(0, 4),
     xaxp = c(0, 1, 4),
     yaxp = c(0, 4, 4),
     las  = 1,
     xlab = "tuning parameter",
     ylab = "true prediction error")
nbr.lines <- 19
tuning.parameter.selected <- rep(0, nbr.lines)
colors <- terrain.colors(nbr.lines)
n.values <- seq(1:19) * n / (nbr.lines + 1)
for (i in 1:nbr.lines)
{
  tuning.parameter.selected[i] <- tuning.parameters[which.min(AvgPredictionLossEstimated(
                                    outcome, design, tuning.parameters, n.values[i], 3))]
  abline(v=tuning.parameter.selected[i], col=colors[i], lty=2, lwd=1.5)
}
```



```
plot(x    = n.values,
     y    = tuning.parameter.selected,
     xlim = c(0, 100),
     ylim = c(0, 1.5),
     xaxp = c(0, 100, 5),
     yaxp = c(0, 1.5, 3),
     las  = 1,
     xlab = "sample size",
     ylab = "tuning parameter")
```

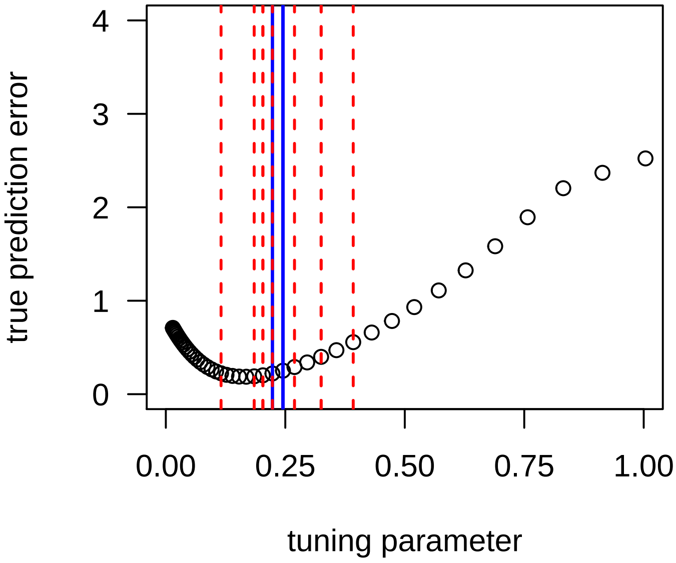


Can you identify a *clear* trend in the number of samples? Which two effects are competing?

4.5 The role of the number of splits

Complement the first plot with lines that indicate tuning parameters selected by (i) Monte Carlo cross-validation with 50 splits and (ii) a simple holdout method. Set the training sample size to 50 for both methods.

```
set.seed(1)
plot(x    = tuning.parameters,
     y    = estimators.prediction.loss,
     xlim = c(0, 1),
     ylim = c(0, 4),
     xaxp = c(0, 1, 4),
     yaxp = c(0, 4, 4),
     las  = 1,
     xlab = "tuning parameter",
     ylab = "true prediction error")
for (i in 1:10)
{
  abline(v=tuning.parameters[which.min(AvgPredictionLossEstimated(outcome, design,
                                                                  tuning.parameters,
                                                                  REPLACE, REPLACE))],
         col="blue", lty=1, lwd=1.5)  # MC with 50 splits
  abline(v=tuning.parameters[which.min(AvgPredictionLossEstimated(outcome, design,
                                                                  tuning.parameters,
                                                                  REPLACE, REPLACE))],
         col="red", lty=2, lwd=1.5)  # holdout
}
```



Which method has smaller variance in the selection of the tuning parameter? Why is this expected?

Chapter 5

Inference

5.1 Overview

5.2 Asymptotic Confidence Intervals Through One-step Updates

Many estimators can be written as roots of a system of equations

$$\widehat{\boldsymbol{\beta}} \text{ such that } \underline{\boldsymbol{h}}(Z,\widehat{\boldsymbol{\beta}}) = \mathbf{0}_h \tag{5.1}$$

with a function

$$\begin{aligned}\underline{\boldsymbol{h}} \;:\; \mathcal{Z}\times\mathcal{B} &\to \mathbb{R}^h\\ (Z,\boldsymbol{\alpha}) &\mapsto \underline{\boldsymbol{h}}(Z,\boldsymbol{\alpha})\,.\end{aligned}$$

We call such estimators *Z-estimators*. Given data Z, we write $\boldsymbol{h}(\boldsymbol{\alpha}) := \underline{\boldsymbol{h}}(Z,\boldsymbol{\alpha})$.

Example 5.2.1 (Maximum Likelihood Estimator) Given data and a class of models indexed by $\boldsymbol{\alpha}$, let $f(\boldsymbol{\alpha})$ be the likelihood function (that is, the log-density of the data as a function of the model parameter $\boldsymbol{\alpha} \in \mathbb{R}^p$). Then, under some regularity conditions, we can write the maximum likelihood estimator

$$\widehat{\boldsymbol{\beta}} \text{ such that } \widehat{\boldsymbol{\beta}} \in \operatorname*{argmax}_{\boldsymbol{\alpha}\in\mathbb{R}^p} f(\boldsymbol{\alpha})$$

as

$$\widehat{\boldsymbol{\beta}} \text{ such that } \boldsymbol{h}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}_p$$

with $\boldsymbol{h}$ the *score function*

$$\boldsymbol{h}(\boldsymbol{\alpha}) := \frac{\partial f(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\,.$$

One-Step Updates

It is not always straightforward to find a root of the system (5.1): $\boldsymbol{h}$ might depend on quantities that are not accessible in practice, solving for $\widehat{\boldsymbol{\beta}}$ might be computational challenging, or a solution of the system might not exist altogether. Nevertheless, we can then still attempt to find an approximate root. For example, we could apply a

Newton-Raphson step to an initial estimator $\widetilde{\boldsymbol{\alpha}}$, that is, we could define a one-step update of $\widetilde{\boldsymbol{\alpha}}$ by

$$\overline{\boldsymbol{\alpha}} := \widetilde{\boldsymbol{\alpha}} - \Bigl(\frac{\partial \boldsymbol{h}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\widetilde{\boldsymbol{\alpha}}}\Bigr)^{-1}\boldsymbol{h}(\widetilde{\boldsymbol{\alpha}})\,.$$

Here, we assume that $h = h$ (so that the Jacobian on the right-hand side is a square-matrix) and that the inverse Jacobian

$$\Bigl(\frac{\partial \boldsymbol{h}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\widetilde{\boldsymbol{\alpha}}}\Bigr)^{-1} \in \mathbb{R}^{p\times p}\,, \quad \text{where} \qquad \Bigl(\frac{\partial \boldsymbol{h}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Bigr)_{ij} := \frac{\partial \boldsymbol{h}_i(\boldsymbol{\alpha})}{\partial \alpha_j}\Big|_{\boldsymbol{\alpha}=\widetilde{\boldsymbol{\alpha}}}\,,$$

is well-defined. The idea of the Newton-Raphson approach (for $h = p$) is as follows: We aim at finding a root of $\boldsymbol{h}(\boldsymbol{\alpha}) = \boldsymbol{0}_p$. Given an point $\widetilde{\boldsymbol{\alpha}}$, we want to take a step $\boldsymbol{\delta}$ in towards this goal, that is, optimally, we would want to have for $\overline{\boldsymbol{\alpha}} := \widetilde{\boldsymbol{\alpha}} + \boldsymbol{\delta}$

$$\boldsymbol{h}(\overline{\boldsymbol{\alpha}}) \;=\; \boldsymbol{0}_p\,.$$

Assuming that $\boldsymbol{h}$ is smooth, we can do a Taylor expansion around $\widetilde{\boldsymbol{\alpha}}$ to find

$$\boldsymbol{h}(\overline{\boldsymbol{\alpha}}) \;=\; \boldsymbol{h}(\widetilde{\boldsymbol{\alpha}} + \boldsymbol{\delta}) \;\approx\; \boldsymbol{h}(\widetilde{\boldsymbol{\alpha}}) + \Bigl(\frac{\partial \boldsymbol{h}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\widetilde{\boldsymbol{\alpha}}}\Bigr)\boldsymbol{\delta}\,.$$

Combining these two displays and solving for $\boldsymbol{\delta}$ motivates the rule

$$\boldsymbol{\delta} \;:=\; -\Bigl(\frac{\partial \boldsymbol{h}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\widetilde{\boldsymbol{\alpha}}}\Bigr)^{-1}\boldsymbol{h}(\widetilde{\boldsymbol{\alpha}})$$

and therefore,

$$\overline{\boldsymbol{\alpha}} \;:=\; \widetilde{\boldsymbol{\alpha}} + \boldsymbol{\delta} \;=\; \widetilde{\boldsymbol{\alpha}} - \Bigl(\frac{\partial \boldsymbol{h}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\widetilde{\boldsymbol{\alpha}}}\Bigr)^{-1}\boldsymbol{h}(\widetilde{\boldsymbol{\alpha}})\,.$$

However, unlike in the Newton-Raphson scheme, we are satisfied with only one step.

It has been shown in the classical literature that the one-step updated estimator $\overline{\boldsymbol{\alpha}}$ can have good properties if $\widetilde{\boldsymbol{\alpha}}$ is $\sqrt{n}$-consistent.

Example 5.2.2 (Linear Regression) Consider a linear regression model

$$\boldsymbol{y} \;=\; X\boldsymbol{\beta} + \boldsymbol{u}$$

as usual but with the additional assumption that the design X and the noise $\boldsymbol{u}$ are independent of each other and that $\boldsymbol{u}$ has mean zero. We observe that in this case,

$$\mathbb{E}[-2X^\top(\boldsymbol{y} - X\boldsymbol{\beta})] \;=\; \mathbb{E}[-2X^\top\boldsymbol{u}] \;=\; \boldsymbol{0}_p\,.$$

This motivates a Z-estimator of $\boldsymbol{\beta}$ defined by

$$\widehat{\boldsymbol{\beta}} \;\text{ such that }\; \boldsymbol{h}(\widehat{\boldsymbol{\beta}}) \;=\; \boldsymbol{0}_p$$

with

$$\begin{aligned}\boldsymbol{h} \;:\; \mathbb{R}^p &\to \mathbb{R}^p\\ \boldsymbol{\alpha} &\mapsto -2X^\top(\boldsymbol{y} - X\boldsymbol{\alpha})\end{aligned}$$

the empirical version of $\boldsymbol{\alpha} \mapsto \mathbb{E}[-2X^\top(\boldsymbol{y} - X\boldsymbol{\alpha})]$. Assuming that $X^\top X$ is invertible, one can solve for $\widehat{\boldsymbol{\beta}}$ in the defining equation to find that the estimator is simply a least-squares:

$$\widehat{\boldsymbol{\beta}} \;=\; (X^\top X)^{-1}X^\top\boldsymbol{y}\,.$$

Instead, one can also invoke the described Newton-Raphson scheme to determine an approximate root $\overline{\boldsymbol{\alpha}}$ of the equation. Using that

$$\frac{\partial \boldsymbol{h}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = 2X^\top X\,,$$

one finds for *any* initial estimator $\widetilde{\boldsymbol{\alpha}} \in \mathbb{R}^p$

$$\begin{aligned}
\overline{\boldsymbol{\alpha}} &= \widetilde{\boldsymbol{\alpha}} - \Bigl(\frac{\partial \boldsymbol{h}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\widetilde{\boldsymbol{\alpha}}}\Bigr)^{-1} \boldsymbol{h}(\widetilde{\boldsymbol{\alpha}}) && \text{“definition of } \overline{\boldsymbol{\alpha}}\text{”}\\
&= \widetilde{\boldsymbol{\alpha}} - \frac{(X^\top X)^{-1}}{2}\bigl(-2X^\top(\boldsymbol{y} - X\widetilde{\boldsymbol{\alpha}})\bigr) && \text{“previous equality and definition of } \boldsymbol{h}\text{”}\\
&= \widetilde{\boldsymbol{\alpha}} + (X^\top X)^{-1}X^\top \boldsymbol{y} - \widetilde{\boldsymbol{\alpha}} && \text{“algebra”}\\
&= (X^\top X)^{-1}X^\top \boldsymbol{y}\,, && \text{“algebra”}
\end{aligned}$$

which is again the least-squares estimator. Thus, the one-step approximation equals the exact solution in this simple case—irrespective of what initial estimator is used.

One can check readily that multiplying the function $\boldsymbol{h}$ with any non-zero factor does not change the results. However, one can also check that $\boldsymbol{h}$ is equal to the score in linear regression with Gaussian errors only with the above choice of factors.

Now in high-dimensional settings, the matrix $\frac{\partial \boldsymbol{h}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\big|_{\boldsymbol{\alpha}=\widetilde{\boldsymbol{\alpha}}}$ is typically singular. We thus define also estimators with an approximate inverse matrix $A_{\boldsymbol{h}} \in \mathbb{R}^{p\times p}$:

$$\overline{\boldsymbol{\alpha}}_{A_{\boldsymbol{h}}} := \widetilde{\boldsymbol{\alpha}} - A_{\boldsymbol{h}}\boldsymbol{h}(\widetilde{\boldsymbol{\alpha}})\,.$$

Example 5.2.3 (Linear Regression cont.) Going back to the above example about linear regression, we find

$$\begin{aligned}
\overline{\boldsymbol{\alpha}}_{A_{\boldsymbol{h}}} &= \widetilde{\boldsymbol{\alpha}} - A_{\boldsymbol{h}}\boldsymbol{h}(\widetilde{\boldsymbol{\alpha}}) && \text{“definition of } \overline{\boldsymbol{\alpha}}_{A_{\boldsymbol{h}}}\text{”}\\
&= \widetilde{\boldsymbol{\alpha}} + 2A_{\boldsymbol{h}}X^\top(\boldsymbol{y} - X\widetilde{\boldsymbol{\alpha}}) && \text{“definition of } \boldsymbol{h}\text{”}\\
&= \widetilde{\boldsymbol{\alpha}} + 2A_{\boldsymbol{h}}X^\top X(\boldsymbol{\beta} - \widetilde{\boldsymbol{\alpha}}) + 2A_{\boldsymbol{h}}X^\top \boldsymbol{u} && \text{“invoking the model”}\\
&= 2A_{\boldsymbol{h}}X^\top \boldsymbol{u} + (2A_{\boldsymbol{h}}X^\top X - \mathrm{I}_{p\times p})(\boldsymbol{\beta} - \widetilde{\boldsymbol{\alpha}}) + \boldsymbol{\beta}\,. && \text{“rearranging”}
\end{aligned}$$

The first term can be interpreted as a noise term, the second term as an approximation or remainder term, and the third term is the desired target vector.

Asymptotic Confidence Intervals

In high-dimensions, finding roots is not so much of a problem. Instead, the challenge is rather that the asymptotic distributions of those roots are often intractable. As it turns out, the one-step corrected estimators can have much better asymptotic properties.

To illustrate this, we focus on linear regression. Example 5.2.3 yields that

$$\overline{\boldsymbol{\alpha}}_{A_{\boldsymbol{h}}} - \boldsymbol{\beta} = 2A_{\boldsymbol{h}}X^\top \boldsymbol{u} + (2A_{\boldsymbol{h}}X^\top X - \mathrm{I}_{p\times p})(\boldsymbol{\beta} - \widetilde{\boldsymbol{\alpha}})\,.$$

The idea is now that remainder term $(2A_{\boldsymbol{h}}X^\top X - \mathrm{I}_{p\times p})(\boldsymbol{\beta} - \widetilde{\boldsymbol{\alpha}})$ is small as long as $2A_{\boldsymbol{h}}$ is a good approximate inverse of the gram matrix $X^\top X$ and $\widehat{\boldsymbol{\beta}}$ a reasonable estimator of $\boldsymbol{\beta}$. We would thus hope that the asymptotic distribution of $\overline{\boldsymbol{\alpha}}_{A_{\boldsymbol{h}}} - \boldsymbol{\beta}$ is simply the one of $2A_{\boldsymbol{h}}X^\top \boldsymbol{u}$, and that the latter distributions is easy to manage.

Formally, the above equality implies for any fixed $j \in \{1, \ldots, p\}$

$$(\overline{\boldsymbol{\alpha}}_{A_{\boldsymbol{h}}})_j - \boldsymbol{\beta}_j = v\boldsymbol{v} + \boldsymbol{g}\,,$$

where $\boldsymbol{v} := (A_{\boldsymbol{h}}X^\top\boldsymbol{u})_j/\sqrt{\mathbb{E}[(A_{\boldsymbol{h}}X^\top\boldsymbol{u})_j^2]}$ and $v := 2\sqrt{\mathbb{E}[(A_{\boldsymbol{h}}X^\top\boldsymbol{u})_j^2]}$ (we assume v to be strictly positive) play the roles of the limiting random variable and its standard deviation, respectively, and $\boldsymbol{g} := \big((2A_{\boldsymbol{h}}X^\top X - \mathrm{I}_{p\times p})(\boldsymbol{\beta} - \widetilde{\boldsymbol{\alpha}})\big)_j$ plays the role of a remainder term. Hence, for any $z_\alpha, m_n > 0$,

$$\begin{aligned}
&\mathbb{P}\big(\,(\overline{\boldsymbol{\alpha}}_{A_{\boldsymbol{h}}})_j - z_\alpha v \;\leq\; \boldsymbol{\beta}_j \;\leq\; \overline{\boldsymbol{\alpha}}_{A_{\boldsymbol{h}}} + z_\alpha v\,\big) \\
&= \mathbb{P}\big(\,|(\overline{\boldsymbol{\alpha}}_{A_{\boldsymbol{h}}})_j - \boldsymbol{\beta}_j| \;\leq\; z_\alpha v\,\big) && \text{“algebra”} \\
&= \mathbb{P}\big(\,|v\boldsymbol{v} + \boldsymbol{g}| \;\leq z_\alpha v\,\big) && \text{“above equation”} \\
&\geq \mathbb{P}\big(\,|v\boldsymbol{v}| \;\leq z_\alpha v - |\boldsymbol{g}|\,\big) && \text{“triangle inequality and } \mathbb{P}(A) \leq \mathbb{P}(B) \text{ for } A \subset B\text{”} \\
&= \mathbb{P}\big(\,|\boldsymbol{v}| \;\leq z_\alpha - |\boldsymbol{g}|/v\,\big) && \text{“}v > 0 \text{ by assumption”} \\
&\geq \mathbb{P}\big(\,|\boldsymbol{v}| \;\leq z_\alpha - m_n; |\boldsymbol{g}| \leq m_n v\,\big) && \text{“}\mathbb{P}(A) \leq \mathbb{P}(B) \text{ for } A \subset B\text{”} \\
&= \mathbb{P}\big(\,|\boldsymbol{v}| \;\leq\; z_\alpha - m_n\,\big) - \mathbb{P}\big(\,|\boldsymbol{v}| \;\leq z_\alpha - m_n; |\boldsymbol{g}| > m_n v\,\big) && \text{“}\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^{\complement})\text{”} \\
&\geq \mathbb{P}\big(\,|\boldsymbol{v}| \;\leq\; z_\alpha - m_n\,\big) - \mathbb{P}\big(\,|\boldsymbol{g}| > m_n v\,\big)\,. && \text{“}\mathbb{P}(A \cap B) \leq \mathbb{P}(A)\text{”}
\end{aligned}$$

We can leverage this into the following result:

Lemma 5.2.1 (Asymptotic Distribution of One-Step Corrected Estimator) Now, the idea is that if $\boldsymbol{v} \equiv \boldsymbol{v}(n)$ is asymptotically standard normal, z_α is an α-zscore of the normal distribution, and the second term vanishes with a sequence $m_n \to 0$, we find that

$$\mathbb{P}\big(\,(\overline{\boldsymbol{\alpha}}_{A_{\boldsymbol{h}}})_j - z_\alpha v \;\leq\; \boldsymbol{\beta}_j \;\leq\; (\overline{\boldsymbol{\alpha}}_{A_{\boldsymbol{h}}})_j + z_\alpha v\,\big) \;\leq\; \alpha + o(1)\,,$$

This means that we have formed asymptotically valid α-confidence intervals for $\boldsymbol{\beta}$. Of course in practice, v is often unknown, so that is has to estimated consistently.

Note also that importantly, we have not assumed $\sqrt{n}$-concistency: all we needed is that $\boldsymbol{g}$ converges to zero in probability.

Let us conclude with a concrete example.

Example 5.2.4 (One-step Updated Lasso) Consider the linear regression case discussed in the previous examples, but explicitly with a high-dimensional, sparse setting. One could then write the lasso as a Z-estimator and discuss appropriate Newton-Raphson-type steps for it. For this, one could use that the KKT-conditions for the lasso $\widehat{\boldsymbol{\beta}}$ with tuning parameter r imply

$$-2X^\top(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}) + r\widehat{\boldsymbol{\kappa}} \;=\; \mathbf{0}_p$$

for a $\widehat{\boldsymbol{\kappa}} \in \boldsymbol{\partial}\|\widehat{\boldsymbol{\beta}}\|_1$ and set, for example, $\underline{\boldsymbol{h}} : \boldsymbol{\alpha} \mapsto \min_{\boldsymbol{\kappa}\in\boldsymbol{\partial}\|\boldsymbol{\alpha}\|_1} \|{-2X^\top(\boldsymbol{y} - X\boldsymbol{\alpha}) + r\boldsymbol{\kappa}}\|_2$. Instead, the above treatment suggests to use the (approximate version) of the least-squares as described in Example 5.2.3 for the Newton-Raphson step and the lasso as *initial estimator*. However, we can then still make use of the KKT conditions above to establish a concise form for the one-step corrected estimator: using those conditions for a lasso primal-dual pair $(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\kappa}})$, and taking any matrix $A_{\boldsymbol{h}} \in \mathbb{R}^{p\times p}$ as approximation of the Jacobien matrix, the one-step updated estimator reads

$$\overline{\boldsymbol{\alpha}}_{A_{\boldsymbol{h}}} \;=\; \widetilde{\boldsymbol{\alpha}} - A_{\boldsymbol{h}}\boldsymbol{h}(\widetilde{\boldsymbol{\alpha}}) \;=\; \widetilde{\boldsymbol{\alpha}} + 2A_{\boldsymbol{h}}X^\top(\boldsymbol{y} - X\widetilde{\boldsymbol{\alpha}}) \;=\; \widetilde{\boldsymbol{\alpha}} + rA_{\boldsymbol{h}}\widetilde{\boldsymbol{\kappa}}\,.$$

References for how to select the matrix $A_{\boldsymbol{h}}$ are stated in the literature section.

5.3 Knock-off Procedures

5.4 References and Further Reading

A reference on one-step updated estimators in classical statistics is [BKRW93, Sections 2.5 and 7.3]. The case for singular Jacobian matrices is described in [GLT, Section 2]; the corresponding regression example in [GLT, Section 4].

Suggestions for how to select the matrix $A_{\boldsymbol{h}}$ can be found in [GLT, JM14, VdGBR$^+$14].

5.5 Exercises

Exercises for Section 5.2

□ **Exercise 5.1** $^\diamond$ With the notation of Example 5.2.3, show that

$$|\boldsymbol{g}| \;\leq\; \|2A_{\boldsymbol{h}}X^\top X - \mathrm{I}_{p\times p}\|_\infty \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\alpha}}\|_1\,,$$

where $\|A\|_\infty := \max_{k,l\in\{1,\dots,p\}} |A_{kl}|$ for any matrix $A \in \mathbb{R}^{p\times p}$.

R Lab Chapter 5

5 Establishing Confidence Intervals by Using A Desparsified Lasso

In this lab, we show how to use the desparsified lasso to establish confidence intervals for regression parameters.

As always, your task is to replace the keyword `REPLACE` with suitable code and to answer the questions posed in the text.

5.1 Data loading

Load the `Fertility` data set from the `Stat2Data` package. Take the column `Embryos` as the outcome vector and the remaining data as the design matrix, and then add a column to the design matrix that takes the intercept into account. Hint: for easier manipulations later, transform the data into matrix format by applying the `as.matrix()` function.

```
library(Stat2Data)
data("Fertility")
y <- REPLACE
X <- REPLACE
cbind(y, X)[1:5, ]
```

```
##   Embryos Intercept Age LowAFC MeanAFC FSH E2 MaxE2 MaxDailyGn TotalGn
## 1      13         1  40     40    51.5 5.3 45  1427        300    2700
## 2       6         1  37     41    41.0 7.1 53   802        225    1800
## 3      15         1  40     38    41.0 4.9 40  4533        450    4850
## 4       4         1  40     36    37.5 3.9 26  1804        300    2700
## 5      12         1  30     36    36.0 4.0 49  2526        150    1500
##   Oocytes
## 1      25
## 2       7
## 3      27
## 4       9
## 5      19
```

5.2 Lasso estimation

Use the `cv.glmnet()` function from the `glmnet` package to compute a cross-validated lasso estimator. Use the standard settings of `glmnet`. Hint: the `coef()` function provides easy access to the parameters in the `cv.glmnet` object.

```
library(glmnet)
set.seed(71) # recall that cross-validation is random
lasso <- REPLACE
lasso
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                     1
## (Intercept) 2.5603339
## Age         .
## LowAFC      .
## MeanAFC     .
## FSH          .
## E2           .
## MaxE2        .
## MaxDailyGn   .
## TotalGn      .
## Oocytes      0.3519556
```

We find that the lasso estimate is sparse: only the last coordinate is non-zero.

5.3 Tuning parameter

Compute the tuning parameter that `cv.glmnet()` applied. Hint: use the KKT conditions.

```
tuningparameter <- REPLACE
```

The tuning parameter is $r = 3.7402364 \times 10^5$ in our framework. Note that we should not proceed the tuning parameters used internally in `glmnet`, as the internal normalizations of `glmnet` are different from the ones in the book; in particular, `cv.glmnet(as.matrix(subset(X, select=-Intercept)), y)}`$= 0.0820567 \neq 3.7402364 \times 10^5$.

If you cannot solve this question, take the mentioned value for the tuning parameter and continue.

5.4 Desparsifying the lasso

Compute the desparsified version of the above lasso. Verify first that the gram matrix $X^\top X$ is invertible, and thus, that its inverse can be used directly in the formulae (that is, $(X^\top X)^{-1}$ can be used as its own approximation).

```
eigen(t(X) %*% X)$value
```

```
##  [1] 3.961676e+09 3.665837e+08 1.105422e+06 1.217737e+05 4.107821e+04
##  [6] 1.073797e+04 7.252189e+03 1.025267e+03 7.446458e+02 3.328247e+00
```

```
lasso.desparsified <- REPLACE
lasso.desparsified
```

```
## 10 x 1 Matrix of class "dgeMatrix"
##                         1
## (Intercept) -0.6414146312
## Age         -0.1109954058
## LowAFC       0.2780913526
## MeanAFC     -0.2007179360
## FSH          0.0423810339
## E2           0.0034946145
## MaxE2        0.0007398015
## MaxDailyGn   0.0042646624
## TotalGn      0.0000254415
## Oocytes      0.6463559364
```

The eigenvalues of $X^\top X$ are all strictly larger than zero, that is, $X^\top X$ is indeed invertible. The desparsified version of the lasso is not sparse any more: all coordinates are non-zero.

5.5 Confidence intervals for `Oocytes`

Use the desparsified lasso to determine a 99% confidence interval for the `Oocytes` parameter.

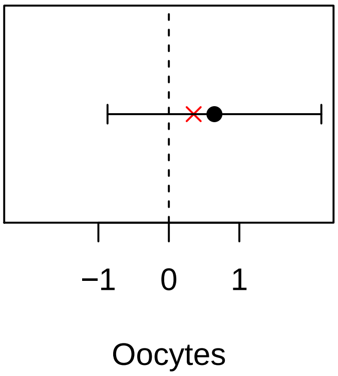
```
coverage <- 0.99
zscore <- qnorm(1 - (1 - coverage) / 2)
variance <- REPLACE
confidence.interval <- c(lasso.desparsified["Oocytes", ] - zscore * variance,
                         lasso.desparsified["Oocytes", ] + zscore * variance)
```

A 99% confidence interval for `Oocytes` is $[-0.8702943, 2.1630062]$. This confidence interval does not contain zero, which provides us with evidence that `Oocytes` is a relevant predictor for `Fertility`.

5.6 Plotting the results

We finally make a plot to illustrate our results.

```
plot(x    = lasso.desparsified["Oocytes", ],
     y    = 1,
     pch  = 19,
     xlim = c(min(confidence.interval[1], -confidence.interval[2]),
              max(confidence.interval[2], -confidence.interval[1])),
     xlab = "Oocytes",
     ylab = "",
     xaxt = "n",
     yaxt = "n")
points(lasso["Oocytes", ], y=1, pch=4, col="red")
axis(side=1, at=seq(from=-1, to=1, by=1))
arrows(confidence.interval[1], 1, confidence.interval[2], 1,
       length=0.05, angle=90, code=3)
abline(v=0, lty=2)
```



The red cross represents the initial lasso estimate of the `Oocytes` parameter; the black circle represents the desparsified estimate; the wiskers represent the 99% confidence interval. We conclude once more that `Oocytes` (number of egg cells) seems related to `Embryos` (number of embryos)—which makes sense biologically.

Chapter 6

Theory I: Prediction

In the remainder of this book, we equip high-dimensional estimators with further mathematical backing. To make our derivations as accessible as possible, we restrict ourselves to linear regression[1]

$$\boldsymbol{y} \;=\; X\boldsymbol{\beta}+\boldsymbol{u} \tag{6.1}$$

with outcome $\boldsymbol{y}\in\mathbb{R}^n$, design matrix $X\in\mathbb{R}^{n\times p}$, regression vector $\boldsymbol{\beta}\in\mathbb{R}^p$, and random noise $\boldsymbol{u}\in\mathbb{R}^n$. Nevertheless, the main concepts transfer to other high-dimensional frameworks as well, so that these two chapters should prepare the reader for mathematical research on high-dimensional statistics in general.

Our ultimate goals are mathematical guarantees for estimating the regression vector $\boldsymbol{\beta}$ from the data $(\boldsymbol{y},X)$. However, to get there, we first consider a more basic task: disentangling the true signal $X\boldsymbol{\beta}$ from the noise $\boldsymbol{u}$. A measure of how well an estimator $\widehat{\boldsymbol{\beta}}$ performs at this task is its *prediction loss*[2] $\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2$. In this chapter, we establish theoretical bounds for this loss.

prediction loss

6.1 Overview

Prediction guarantees are usually derived in two steps: first, the prediction loss is separated from the other terms of the estimator's objective function; then, those other terms are bounded such that they lose their dependence on the estimator itself. That second step is where the regularization becomes important: only if the regularization is "sufficiently strong," the randomness the problem can be controlled.

For an illustration, consider the lasso estimator

$$\widehat{\boldsymbol{\beta}} \;\in\; \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\bigl\{\,\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2+r\|\boldsymbol{\alpha}\|_1\,\bigr\}\,.$$

Since $\widehat{\boldsymbol{\beta}}$ minimizes the objective function $\boldsymbol{\alpha}\mapsto\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2+r\|\boldsymbol{\alpha}\|_1$, we obtain for any $\boldsymbol{\alpha}\in\mathbb{R}^p$ the inequality

$$\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2+r\|\widehat{\boldsymbol{\beta}}\|_1 \;\le\; \|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2+r\|\boldsymbol{\alpha}\|_1\,.$$

[1] More formally: a linear regression is a combination of a probability space $(\mathcal{A},\mathfrak{A},\mathbb{P})$, random quantities $\boldsymbol{y}\,:\,(\mathcal{A},\mathfrak{A})\to(\mathbb{R}^n,\mathcal{B}^n)$, $X\,:\,(\mathcal{A},\mathfrak{A})\to(\mathbb{R}^{n\times p},\mathcal{B}^{n\times p})$, and $\boldsymbol{u}\,:\,(\mathcal{A},\mathfrak{A})\to(\mathbb{R}^n,\mathcal{B}^n)$, where $\mathcal{B}^n,\mathcal{B}^{n\times p}$ denote the appropriate Borel σ-algebras, and a fixed vector $\boldsymbol{\beta}\in\mathbb{R}^p$ such that $\mathbb{P}\{\boldsymbol{y}-X\boldsymbol{\beta}-\boldsymbol{u}=\mathbf{0}_n\}=1$.

[2] Note that the term "prediction" can have distinctly different meanings in the literature: for example, it can also refer to the model fit $\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2$ or the prediction of a new outcome $y^{n+1}\in\mathbb{R}$ for a given vector of predictors $\boldsymbol{x}^{n+1}\in\mathbb{R}^p$.

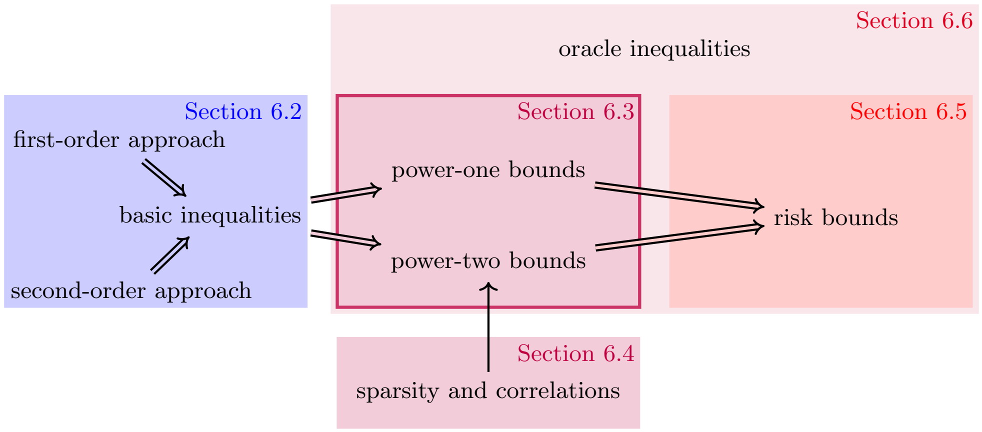


Figure 6.1: Dependencies among the different topics of this chapter. In Section 6.2, we introduce two approaches for deriving basic inequalities. In Section 6.3, we then use these basic inequalities to derive two types of probability bounds, which are the main results of this chapter. In Section 6.4, we provide further background on the assumptions needed for power-two bounds. Finally, in Section 6.5, we show how the probability bounds can be transformed into risk bounds, and in Section 6.6, we give a general description of oracle inequalities.

Adding a zero-valued term in the ℓ_2-norms on both sides then gives us

$$\|\boldsymbol{y}-X\boldsymbol{\beta}+X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2+r\|\widehat{\boldsymbol{\beta}}\|_1 \;\leq\; \|\boldsymbol{y}-X\boldsymbol{\beta}+X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+r\|\boldsymbol{\alpha}\|_1\,,$$

and expanding

$$\begin{aligned}&\|\boldsymbol{y}-X\boldsymbol{\beta}\|_2^2+2\langle\boldsymbol{y}-X\boldsymbol{\beta},X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\rangle+\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2+r\|\widehat{\boldsymbol{\beta}}\|_1\\&\qquad\leq\;\|\boldsymbol{y}-X\boldsymbol{\beta}\|_2^2+2\langle\boldsymbol{y}-X\boldsymbol{\beta},X\boldsymbol{\beta}-X\boldsymbol{\alpha}\rangle+\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+r\|\boldsymbol{\alpha}\|_1\,.\end{aligned}$$

Invoking the model and consolidating the display, we can then derive the *basic inequality*

basic inequality

$$\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; \|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+2\langle\boldsymbol{u},\,X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\alpha}\rangle+r\|\boldsymbol{\alpha}\|_1-r\|\widehat{\boldsymbol{\beta}}\|_1\,.$$

The basic inequality separates the prediction loss of the estimator from other parts of the problem.

Now, we know two ways to proceed from this. One the one hand, Hölder's inequality ($\langle\boldsymbol{a},\,\boldsymbol{b}\rangle\leq\|\boldsymbol{a}\|_\infty\|\boldsymbol{b}\|_1$ for any $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^p$) and the triangle inequality ($\|\boldsymbol{a}+\boldsymbol{b}\|_1\leq\|\boldsymbol{a}\|_1+\|\boldsymbol{b}\|_1$ for any $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^p$) lead to

$$\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; \|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2+2\|X^\top\boldsymbol{u}\|_\infty\bigl(\|\widehat{\boldsymbol{\beta}}\|_1+\|\boldsymbol{\alpha}\|_1\bigr)+r\|\boldsymbol{\alpha}\|_1-r\|\widehat{\boldsymbol{\beta}}\|_1\,.$$

Hence, since $\boldsymbol{\alpha}$ was arbitary, we eventually find for $r\geq 2\|X^\top\boldsymbol{u}\|_\infty$ the *power-one bound*

power-one bound

$$\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; \min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\bigl\{\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+2r\|\boldsymbol{\alpha}\|_1\bigr\}\,.$$

This means that up to the complexity term $2r\|\boldsymbol{\alpha}\|_1$, the lasso with sufficiently large tuning parameter minimizes the prediction loss.

On the other hand, we can start again from the basic inequality, apply Hölder's inequality as before, but use the triangle inequality slightly differently ($\|\boldsymbol{a}\|_1 - \|\boldsymbol{b}\|_1 \leq \|\boldsymbol{a}-\boldsymbol{b}\|_1$ for any $\boldsymbol{a},\boldsymbol{b} \in \mathbb{R}^p$) to find

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; \|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + 2\|X^\top\boldsymbol{u}\|_\infty\|\boldsymbol{\alpha} - \widehat{\boldsymbol{\beta}}\|_1 + r\|\boldsymbol{\alpha} - \widehat{\boldsymbol{\beta}}\|_1\,.$$

If again $r \geq 2\|X^\top\boldsymbol{u}\|_\infty$, this yields

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; \|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + 2r\|\boldsymbol{\alpha} - \widehat{\boldsymbol{\beta}}\|_1\,.$$

This result ensures that if a good approximation $\boldsymbol{\alpha}$ of the target is estimated accurately in ℓ_1-loss, the target $\boldsymbol{\beta}$ (and hence, also its approximation $\boldsymbol{\alpha}$) are recovered well in terms of prediction. We now make the assumption that the reverse is also true: we assume that the estimator $\widehat{\boldsymbol{\beta}}$ is close to the parameter $\boldsymbol{\alpha}$ in ℓ_1-loss if $\widehat{\boldsymbol{\beta}}$ is close to $\boldsymbol{\alpha}$ in prediction loss. We call this the *structural condition* for $\boldsymbol{\alpha}$ and define it formally as $\boldsymbol{\alpha} \in \widehat{\mathcal{B}}_m$, where

structural condition

$$\widehat{\mathcal{B}}_m \;:=\; \bigl\{\boldsymbol{\gamma} \in \mathbb{R}^p \;:\; \|\boldsymbol{\gamma} - \widehat{\boldsymbol{\beta}}\|_1 \leq m\|X\boldsymbol{\gamma} - X\widehat{\boldsymbol{\beta}}\|_2\bigr\}$$

for a fixed $m \in [0,\infty]$. The previous display then entails

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; \|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + 2mr\|X\boldsymbol{\alpha} - X\widehat{\boldsymbol{\beta}}\|_2\,.$$

Using the inequality $2ab \leq 4a^2 + b^2/4$ for all $a,b \in \mathbb{R}$ (see Lemma B.1.2 on Page 164 in the Appendix), we further find

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; \|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + 4m^2r^2 + \|X\boldsymbol{\alpha} - X\widehat{\boldsymbol{\beta}}\|_2^2/4\,,$$

which becomes with $\|\boldsymbol{a}+\boldsymbol{b}\|_2^2 \leq 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$ for all $\boldsymbol{a},\boldsymbol{b} \in \mathbb{R}^p$ (see Lemma B.1.1 on Page 164 in the Appendix)

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; \|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + 4m^2r^2 + \|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2/2 + \|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2/2\,.$$

In summary, if $r \geq 2\|X^\top\boldsymbol{u}\|_\infty$, we find *power-two bound*

power-two bound

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; \min_{\boldsymbol{\alpha}\in\widehat{\mathcal{B}}_m}\bigl\{3\|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + 8m^2r^2\bigr\}\,.$$

This prediction guarantee differs from the one above in three main aspects: the leading constant is strictly larger than one, that is, the oracle inequality is not "sharp"; the complexity term is quadratic in the tuning parameter r (hence the names power-one and power-two bounds) and does not involve $\|\boldsymbol{\alpha}\|_1$; and the minimum is taken over $\widehat{\mathcal{B}}_m$ rather than over the entire $\mathbb{R}^p$. Power-two bounds improve on power-one bounds if both the target is sparse and the predictors weakly correlated but not necessarily otherwise—see Section 6.4.

In both bounds, the role of the least-squares' effective noise $\|(U^\top\boldsymbol{u})_{\{1,\dots,p\}}\|_2$ (see Page 12) is assumed by the lasso's effective noise $2\|X^\top\boldsymbol{u}\|_\infty$. If, for example, $\boldsymbol{u} \sim \mathcal{N}_n[\mathbf{0}_n, \sigma^2\,\mathrm{I}_{n\times n}/n]$ and $(X^\top X)_{jj} = 1$, we have seen in Lemma 4.2.1 that the noise term $2\|X^\top\boldsymbol{u}\|_\infty$ is bounded with high probability by $a\sigma\sqrt{\log[p]/n}$, where $a \in (0,\infty)$ is some constant. For example the first result above then implies that for $r \geq a\sigma\sqrt{\log[p]/n}$, we have with high probability

effective noise for the lasso

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; \min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\Biggl\{\|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + 2a\sigma\sqrt{\frac{\log p}{n}}\|\boldsymbol{\alpha}\|_1\Biggr\}\,.$$

For applications, this means the following: if one is confident in the prior assumptions, namely that there is a vector that both approximates the target well in prediction and is not too large in ℓ_1-norm, and if $n \gg \sigma^2 \log p$, one can expect the lasso to provide accurate prediction. This means that the lasso can provide accurate prediction even if the number of parameters is much larger than the number of samples.

The goal of this chapter is to generalize and sharpen such prediction bounds. We disentangle the algebraic and probabilistic aspects of the topic as much as possible: We first derive bounds without making any assumptions on the noise $\boldsymbol{u}$ (such as being normally distributed, for example) other than satisfying (6.1); we stress this generality by typically writing $\boldsymbol{y} - X\boldsymbol{\beta}$ instead of $\boldsymbol{u}$. Guarantees for concrete models can then be obtained by plugging in specific noise distributions and, if needed, using appropriate results from empirical process theory. This separation of the two aspects yields the most general results and, at the same time, allows us to illuminate most effectively the underpinning working principles of high-dimensional theories.

6.2 Basic Inequalities

Our overall goal in this chapter is to estimate prediction targets $X\boldsymbol{\beta} \in \mathbb{R}^n$ by means of $X\widehat{\boldsymbol{\beta}}$, where

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\bigl\{ g\bigl[\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2\bigr] + rh[\boldsymbol{\alpha}]\bigr\} \tag{6.2}$$

for a link function $g : [0,\infty) \to [-\infty,\infty]$, a tuning parameter $r \in [0,\infty)$, and a prior function $h : \mathbb{R}^p \to [-\infty,\infty]$. In this section, we derive two basic inequalities as a first step towards this goal.

The two results will be based on two different proof techniques: in the first proof, we will use that the objective function is smallest at $\widehat{\boldsymbol{\beta}}$ directly; in the second proof, we will set derivatives to zero.

The first bound is valid for $g : x \mapsto x$.

Lemma 6.2.1 (First-order Basic Inequality)

For any $\boldsymbol{\alpha} \in \mathbb{R}^p$, it holds that

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \le \|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + 2\langle \boldsymbol{y} - X\boldsymbol{\beta},\, X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\alpha}\rangle + rh[\boldsymbol{\alpha}] - rh[\widehat{\boldsymbol{\beta}}]\,.$$

This result allows for any prior function h at the price of setting the link function g equal to the identity. As a special case, the bound comprises the basic inequality of the lasso in the linear regression model derived above.

Proof of Lemma 6.2.1. The proof closely follows the strategy outlined in Section 6.1 for the lasso case.

By definition of $\widehat{\boldsymbol{\beta}}$ as a minimizer of the objective function in (6.2), it holds for any $\boldsymbol{\alpha} \in \mathbb{R}^p$ that (recall that here $g : x \mapsto x$)

$$\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\|_2^2 + rh[\widehat{\boldsymbol{\beta}}] \le \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + rh[\boldsymbol{\alpha}]\,.$$

We add a zero-valued term insight the ℓ_2-norm on the left-hand side and then open up that term:

$$\begin{aligned}\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}\|_2^2 &= \|\boldsymbol{y} - X\boldsymbol{\beta} + X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2\\ &= \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + 2\langle \boldsymbol{y} - X\boldsymbol{\beta},\, X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\rangle + \|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2\,.\end{aligned}$$

We also perform a similar attack on the right-hand side of the initial inequality:

$$\begin{aligned}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2 &= \|\boldsymbol{y}-X\boldsymbol{\beta}+X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2\\ &= \|\boldsymbol{y}-X\boldsymbol{\beta}\|_2^2+2\langle\boldsymbol{y}-X\boldsymbol{\beta},\,X\boldsymbol{\beta}-X\boldsymbol{\alpha}\rangle+\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2\,.\end{aligned}$$

Plugging these two observations into the initial inequality yields

$$\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2+rh[\widehat{\boldsymbol{\beta}}]\;\le\;\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+rh[\boldsymbol{\alpha}]+2\langle\boldsymbol{y}-X\boldsymbol{\beta},\,X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\alpha}\rangle\,,$$

which can be rearranged to the desired form. □

If the prior function h is convex, we can sharpen the bounds and incorporate other link functions g. For this, we introduce an "additional version" of g:

$$\begin{aligned}\tilde{g}\;:\;\mathbb{R}^p &\to \mathbb{R}\\ \boldsymbol{\alpha} &\mapsto g\big[\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2\big]\,.\end{aligned}$$

While the arguments of the initial link function g are real-valued, the arguments of the induced function $\tilde{g}$ are vector-valued. We assume that g is differentiable at $\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2$ with strictly positive derivative, that is, $g'\big[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\big]>0$, and that $\tilde{g}$ is convex.

One function that satisfies these assumptions is the above discussed $g\,:\,x\mapsto x$: it is differentiable with derivative equal to 1 everywhere, and its induced version $\tilde{g}\,:\,\boldsymbol{\alpha}\mapsto\tilde{g}[\boldsymbol{\alpha}]=\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2$ is convex. Another function that satisfies the assumptions is $g\,:\,x\mapsto\sqrt{x}$ (assuming $\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\neq 0$): it is differentiable (assuming $x\neq 0$) with derivative equal to $1/(2\sqrt{x})>0$, and its induced version $\tilde{g}\,:\,\boldsymbol{\alpha}\mapsto\tilde{g}[\boldsymbol{\alpha}]=\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2$ is convex. The square-root function can be used to write the *square-root lasso*

$$\widehat{\boldsymbol{\beta}}\;\in\;\operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\big\{\,\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2+r\|\boldsymbol{\alpha}\|_1\,\big\}$$

in the form

$$\widehat{\boldsymbol{\beta}}\;\in\;\operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\big\{\,g\big[\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2\big]+r\|\boldsymbol{\alpha}\|_1\,\big\}\,,$$

which is of the assumed form (6.2). Thus, in contrast to the lemma derived above, the lemma to follow applies in particular to the lasso *and* the square-root lasso.

We find the following bound.

Lemma 6.2.2 (Second-order Basic Inequality)

Under the aforementioned assumptions on g and h, it holds for any $\boldsymbol{\alpha}\in\mathbb{R}^p$ that

$$\begin{aligned}\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2\le\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2&+2\langle\boldsymbol{y}-X\boldsymbol{\beta},\,X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\alpha}\rangle\\ &+\frac{r}{g'\big[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\big]}h[\boldsymbol{\alpha}]-\frac{r}{g'\big[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\big]}h[\widehat{\boldsymbol{\beta}}]\,.\end{aligned}$$

In contrast to the previous result, which allows for non-convex prior functions, this result here relies on the assumed convexity of the estimator's objective function. One can check that if both the assumptions of Lemma 6.2.1 and the assumptions of Lemma 6.2.2 are satisfied, the two bounds coincide—see 1. in Exercise 6.2.

The term "first-order" indicates that the proof of that result starts directly from the estimators' defining property

$$\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2+rh[\widehat{\boldsymbol{\beta}}]\;\le\;\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2+rh[\boldsymbol{\alpha}]$$

for all $\boldsymbol{\alpha} \in \mathbb{R}^p$. The term "second-order" indicates that the proof of the corresponding result insteads starts from setting generalized derivatives of the objective function at the estimator to zero.

Proof of Lemma 6.2.2. The individual steps of the proof are more involved than those in the previous proof; in particular, the steps here invoke convex analysis, see Section B.1 on Pages 164ff. The overall strategy is simple nevertheless: (i) we first set generalized derivatives of the estimator's objective function to zero to characterize the estimator through an equality that involves a subgradient; (ii) we then replace this subgradient by a simple term that consists only of the tuning parameter and the primal function, at the price of transforming the initial equation into an inequality; (iii) we finally separate the estimator's prediction loss from all other quantities in this inequality.

(i) We first show that the estimators $\widehat{\boldsymbol{\beta}}$ in (6.2) can be characterized by

$$\mathbf{0}_p \;=\; -2g'\big[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\big]\big(X^\top(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}})\big)+r\widehat{\boldsymbol{\kappa}}\,.$$

For this, we rewrite (6.2) as

$$\widehat{\boldsymbol{\beta}} \;\in\; \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p} f[\boldsymbol{\alpha}]$$

with the objective function f defined through

$$\begin{aligned}\mathbb{R}^p &\;\to\; \mathbb{R}\\ \boldsymbol{\alpha} &\;\mapsto\; f[\boldsymbol{\alpha}] \;:=\; \tilde{g}[\boldsymbol{\alpha}]+rh[\boldsymbol{\alpha}]\,.\end{aligned}$$

The objective function f is convex, because by assumption, both the induced link function $\tilde{g}\,:\,\boldsymbol{\alpha}\mapsto g\big[\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2\big]$ and the penalty function h are convex and $r\geq 0$.

The optimality conditions for the estimator $\widehat{\boldsymbol{\beta}}$ are $\mathbf{0}_p \in \boldsymbol{\partial} f[\boldsymbol{\alpha}]\big|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\beta}}}$. Since subdifferentials are additive, we can write $\boldsymbol{\partial} f[\boldsymbol{\alpha}]\big|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\beta}}}$ as a sum of subdifferentials of the two parts of f. In particular, using that g is differentiable (and thus, $\tilde{g}$ is differentiable), we can decompose $\mathbf{0}_p \in \boldsymbol{\partial} f[\boldsymbol{\alpha}]\big|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\beta}}}$ as

$$\mathbf{0}_p \;=\; \frac{\boldsymbol{\partial}}{\boldsymbol{\partial\alpha}}\tilde{g}[\boldsymbol{\alpha}]\big|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\beta}}}+r\widehat{\boldsymbol{\kappa}}$$

for a $\widehat{\boldsymbol{\kappa}}\in\boldsymbol{\partial} h[\boldsymbol{\alpha}]\big|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\beta}}}$. The chain rule for differentials gives us

$$\frac{\boldsymbol{\partial}}{\boldsymbol{\partial\alpha}}\tilde{g}[\boldsymbol{\alpha}]\big|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\beta}}} \;=\; -2g'\big[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\big]\big(X^\top(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}})\big)\,.$$

Combining the two displays finally yields

$$\mathbf{0}_p \;=\; -2g'\big[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\big]\big(X^\top(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}})\big)+r\widehat{\boldsymbol{\kappa}}\,,$$

as desired.

(ii) We now prove that the following inequality holds for any $\boldsymbol{\alpha}\in\mathbb{R}^p$:

$$-2g'\big[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\big]\big(X^\top(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}})\big)^\top(\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}})+r\big(h[\boldsymbol{\alpha}]-h[\widehat{\boldsymbol{\beta}}]\big)\;\geq\;0\,.$$

For this, we first conclude from (i) that for any $\boldsymbol{\alpha}\in\mathbb{R}^p$,

$$0 \;=\; \mathbf{0}_p^\top(\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}}) \;=\; \big(-2g'\big[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\big](X^\top\big(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}})\big)+r\widehat{\boldsymbol{\kappa}}\big)^\top(\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}})\,.$$

Indeed, the first equality is trivial, and the second equality follows directly from replacing $\mathbf{0}_p$ with $-2g'\bigl[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\bigr](X^\top(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}))+r\widehat{\boldsymbol{\kappa}}$ as suggested by (i). Since $\widehat{\boldsymbol{\kappa}}\in\partial h[\boldsymbol{\alpha}]\big|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\beta}}}$, the definition of subdifferentials implies that for all $\boldsymbol{\alpha}\in\mathbb{R}^p$,

$$h[\boldsymbol{\alpha}]\ \geq\ h[\widehat{\boldsymbol{\beta}}]+\langle\widehat{\boldsymbol{\kappa}},\,\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}}\rangle\,,$$

which (recall again that $r\geq 0$) can be rearranged to

$$r\bigl(h[\boldsymbol{\alpha}]-h[\widehat{\boldsymbol{\beta}}]\bigr)\ \geq\ r\widehat{\boldsymbol{\kappa}}^\top(\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}})\,.$$

Combining this with the first equality yields

$$-2g'\bigl[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\bigr]\bigl(X^\top(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}})\bigr)^\top(\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}})+r\bigl(h[\boldsymbol{\alpha}]-h[\widehat{\boldsymbol{\beta}}]\bigr)\ \geq\ 0\,,$$

as desired.

(iii) We finally derive the desired bound

$$\begin{aligned}\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2\leq\ &\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+\langle\boldsymbol{y}-X\boldsymbol{\beta},\,X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\alpha}\rangle\\&+\frac{r}{g'\bigl[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\bigr]}h[\boldsymbol{\alpha}]-\frac{r}{g'\bigl[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\bigr]}h[\widehat{\boldsymbol{\beta}}]\,.\end{aligned}$$

For this, we first observe that

$$\begin{aligned}&\bigl(X^\top(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}})\bigr)^\top(\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}})\\&=\bigl(X^\top(X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}+\boldsymbol{y}-X\boldsymbol{\beta})\bigr)^\top(\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}})&&\text{“adding a zero-valued term”}\\&=(X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}})^\top(X\boldsymbol{\alpha}-X\widehat{\boldsymbol{\beta}})+(\boldsymbol{y}-X\boldsymbol{\beta})^\top(X\boldsymbol{\alpha}-X\widehat{\boldsymbol{\beta}})\\&&&\text{“splitting the term into two parts”}\\&=(X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}})^\top(X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}+X\boldsymbol{\alpha}-X\boldsymbol{\beta})+(\boldsymbol{y}-X\boldsymbol{\beta})^\top(X\boldsymbol{\alpha}-X\widehat{\boldsymbol{\beta}})\\&&&\text{“adding another zero-valued term in the first summand”}\\&=\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2+(X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}})^\top(X\boldsymbol{\alpha}-X\boldsymbol{\beta})+(\boldsymbol{y}-X\boldsymbol{\beta})^\top(X\boldsymbol{\alpha}-X\widehat{\boldsymbol{\beta}})\,.\\&&&\text{“splitting the first term into two parts”}\end{aligned}$$

The second summand in the last line can be bounded from below according to Lemma B.1.2: setting in the first part of that lemma $\boldsymbol{a}=X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}$, $\boldsymbol{b}=X\boldsymbol{\alpha}-X\boldsymbol{\beta}$, and $v=1/a$ for any fixed $a\in(0,\infty)$, we find

$$-a\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2-\frac{\|X\boldsymbol{\alpha}-X\boldsymbol{\beta}\|_2^2}{4a}\ \leq\ (X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}})^\top(X\boldsymbol{\alpha}-X\boldsymbol{\beta})\,.$$

Using this bound in the foregoing display and consolidating gives us

$$\begin{aligned}&\bigl(X^\top(\boldsymbol{y}-X\widehat{\boldsymbol{\beta}})\bigr)^\top(\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}})\\&\qquad\geq\ (1-a)\,\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2-\frac{\|X\boldsymbol{\alpha}-X\boldsymbol{\beta}\|_2^2}{4a}+(\boldsymbol{y}-X\boldsymbol{\beta})^\top(X\boldsymbol{\alpha}-X\widehat{\boldsymbol{\beta}})\,.\end{aligned}$$

Plugging this back into the result of (ii), noting that $g'\bigl[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\bigr]$ is positive by assumption,

$$\begin{aligned}-\,2g'\bigl[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\bigr]\Bigl((1-a)\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2-\frac{\|X\boldsymbol{\alpha}-X\boldsymbol{\beta}\|_2^2}{4a}+\langle\boldsymbol{y}-X\boldsymbol{\beta},\,X\boldsymbol{\alpha}-X\widehat{\boldsymbol{\beta}}\rangle\Bigr)&\\+\,r\bigl(h[\boldsymbol{\alpha}]-h[\widehat{\boldsymbol{\beta}}]\bigr)\ \geq\ 0\,.&\end{aligned}$$

Dividing by $-2(1-a)g'\bigl[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\bigr]$ (recall again that $g'\bigl[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\bigr]>0$ by assumption) gives us

$$\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2-\frac{\|X\boldsymbol{\alpha}-X\boldsymbol{\beta}\|_2^2}{4(1-a)a}+\frac{\langle\boldsymbol{y}-X\boldsymbol{\beta},X\boldsymbol{\alpha}-X\widehat{\boldsymbol{\beta}}\rangle}{1-a}\\+\frac{r}{2(1-a)g'\bigl[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\bigr]}\bigl(h[\widehat{\boldsymbol{\beta}}]-h[\boldsymbol{\alpha}]\bigr)\ \leq\ 0\,.$$

Rearranging this yields

$$\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2\leq\frac{\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2}{4(1-a)a}+\frac{\langle\boldsymbol{y}-X\boldsymbol{\beta},\,X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\alpha}\rangle}{1-a}\\+\frac{r}{2(1-a)g'\bigl[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\bigr]}h[\boldsymbol{\alpha}]-\frac{r}{2(1-a)g'\bigl[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\bigr]}h[\widehat{\boldsymbol{\beta}}]\,,$$

from which the desired inequality follows by setting $a=1/2$.

Details: When $\boldsymbol{\alpha}=\boldsymbol{\beta}$, one can also let $a\to 0$. In the case $g'\bigl[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\bigr]=1$, this leads to an inner product term and prior terms that are reduced by a factor 2 as compared to the corresponding terms of Lemma 6.2.1—see 2. in Exercise 6.2. More generally, because one can optimize over a, the power-two approach is slightly sharper than the power-one approach.

□

6.3 Probability Bounds

We now derive probability bounds for prediction. As outlined in the overview section, we start from a basic inequality and then apply Hölder's inequality and the triangle inequality to consolidate terms. We use the triangle inequality in two slightly different ways, leading to two different types of bounds called power-one bounds and power-two bounds, respectively. These two types differ mainly in their assumptions and in how they depend on the tuning parameters: roughly speaking, power-one bounds hold more generally, while power-two bounds are more responsive to changes in the tuning parameters.

We do not restrict ourselves to any specific regression model. In particular, we do not assume any specific distribution for the noise $\boldsymbol{u}=\boldsymbol{y}-X\boldsymbol{\beta}$.

We also do not restrict ourselves to any specific estimator. Instead, we consider all estimators $\widehat{\boldsymbol{\beta}}\in\mathbb{R}^p$ of a target $\boldsymbol{\beta}\in\mathbb{R}^p$ that satisfy a generic basic inequality with respect to some function $h\ :\ \mathbb{R}^p\ \to\ [-\infty,\infty]$.

Assumption 6.3.1 (Generic Basic Inequality) Fix a function $h\ :\ \mathbb{R}^p\ \to\ [-\infty,\infty]$ and positive, finite, and potentially random quantities $\widehat{w},r\in(0,\infty)$. We assume that the estimator $\widehat{\boldsymbol{\beta}}$ satisfies for all $\boldsymbol{\alpha}\in\mathbb{R}^p$ the basic inequality

$$\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2\ \leq\ \|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+2\langle\boldsymbol{y}-X\boldsymbol{\beta},\,X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\alpha}\rangle+\widehat{w}rh[\boldsymbol{\alpha}]-\widehat{w}rh[\widehat{\boldsymbol{\beta}}]\,.$$

For regularized estimators of the form (6.2) with prior function h and tuning parameter $r\in(0,\infty)$, Lemma 6.2.1 yields such a bound with $\widehat{w}=1$, and Lemma 6.2.2 yields such a bound with $\widehat{w}=1/g'\bigl[\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}\|_2^2\bigr]$.

In our first probablity bound, we make use of three assumptions on the function h:

1. h meets the conditions of Lemma B.1.3 (Hölder's inequality);

2. h is non-negative;

3. h is symmetric.

These assumptions are weak: for example, we do not need assume h to be convex or to satisfy the triangle inequality. Moreover, 2. and 3. are used only to simplify the bounds. For example, the first two parts below do not invoke symmetry. A case where this is relevant is when variables are constraint to be non-negative, for example, see also Exercise 6.3.

We now find the following bound.

Theorem 6.3.1 (Power-one Bounds)

Consider an estimator $\widehat{\boldsymbol{\beta}}$ that complies with Assumption 6.3.1. Now, if h satisfies Assumption 1., it holds that

$$\begin{aligned}&\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2\\ &\qquad\le \min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\Big\{\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+2\overline{h}[X^\top(\boldsymbol{y}-X\boldsymbol{\beta})]h[-\boldsymbol{\alpha}]+\widehat{w}rh[\boldsymbol{\alpha}]\\ &\qquad\qquad\qquad\qquad +\big(2\overline{h}[X^\top(\boldsymbol{y}-X\boldsymbol{\beta})]-\widehat{w}r\big)h[\widehat{\boldsymbol{\beta}}]\Big\}.\end{aligned}$$

If additionally $\widehat{w}r\ge 2v\overline{h}[X^\top(\boldsymbol{y}-X\boldsymbol{\beta})]$ for a constant $v\in[1,\infty]$ and h satisfies Assumptions 2. & 3., the bound implies (consolidate the h terms)

$$\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2 \;\le\; \min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\Big\{\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+(1+\frac{1}{v})\widehat{w}rh[\boldsymbol{\alpha}]\Big\},$$

and if further $v>1$, the bound also implies (set $\boldsymbol{\alpha}=\boldsymbol{\beta}$, consolidate the h terms, and bound the prediction term from below by 0)

$$h[\widehat{\boldsymbol{\beta}}]\;\le\;\frac{v+1}{v-1}h[\boldsymbol{\beta}]\,.$$

Recall here that $\overline{h}$ is the dual function of h; for example, if h is the ℓ_1-norm, then $\overline{h}$ is the ℓ_∞-norm—see Page 4. Because in high-dimensional theories, inner product terms like $2\langle\boldsymbol{y}-X\boldsymbol{\beta},\,X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\alpha}\rangle$ in (6.3.1) are bounded via Hölder's inequality by default, dual functions are ubiquitous in these two last chapters of the book.

The second bound in the theorem is an oracle inequality for prediction as discussed in the first section. More specifically, since the bound contains the tuning parameter to the first power, it is a power-one bound. The third bound illustrates that the complexity of the estimated model is controlled as long as the tuning parameter is chosen sufficiently large. (The complexity of the target itself can be large, but one can derive refined bounds from the first result to confirm the statement more broadly.)

effective noise for regularized estimators

The specific distribution of the data enters the prediction guarantees via the effective noise $2\overline{h}[X^\top(\boldsymbol{y}-X\boldsymbol{\beta})]$ in the lower bounds for the tuning parameter. On a high level, the smaller the noise $\boldsymbol{u}=\boldsymbol{y}-X\boldsymbol{\beta}$, the smaller the effective noise $2\overline{h}[X^\top(\boldsymbol{y}-X\boldsymbol{\beta})]$, and thus, the smaller the tuning parameters r can be. Since the prediction bounds are increasing in the tuning parameter, this confirms our naive expectation that the smaller the noise, the more accurate predictions can be achieved.

Lemmas 6.2.1 and 6.2.2 imply precise values for 2 and $\widehat{w}$, and v can be set to any arbitrary number in the indicated range. The quantities that are unknown in practice are $\overline{h}[X^\top(\boldsymbol{y}-X\boldsymbol{\beta})]$ and $\boldsymbol{\beta}$. Although these quantities render the bounds impossible to be evaluated in practice (in particular, Theorem 6.3.1 does not yield any confidence intervals), the results of this section are relevant in applications: as mentioned already in Section 6.6, we learn from such oracle inequalities about different estimators' potentials in function of the model parameters, that is, we gain insights about how different estimators perform in different scenarios.

Note that the results so far do not inflict assumptions on the model or the structure of X. Moreover, the theorem per se does not presume that h is convex, albeit convexity can still sneak in if the generic basic inequality in Assumption 6.3.1 is underpinned with results, such as the second-order basic inequality in Lemma 6.2.2 rather than the first-order basic inequality in Lemma 6.2.1, that require convexity themselves.

Proof of Theorem 6.3.1. The proof is a straightforward consequence of Assumption 6.3.1.

To see this, note that the inner product on the right-hand side of the inequality in that assumption can be bounded by

$$\begin{aligned}&2\langle \boldsymbol{y}-X\boldsymbol{\beta},\, X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\alpha}\rangle\\ =\;& 2\langle X^\top(\boldsymbol{y}-X\boldsymbol{\beta}),\,\widehat{\boldsymbol{\beta}}\rangle+2\langle X^\top(\boldsymbol{y}-X\boldsymbol{\beta}),\,-\boldsymbol{\alpha}\rangle\\ &\hfill\text{“splitting the inner product into two parts”}\\ \le\;& 2\overline{h}[X^\top(\boldsymbol{y}-X\boldsymbol{\beta})]h[\widehat{\boldsymbol{\beta}}]+2\overline{h}[X^\top(\boldsymbol{y}-X\boldsymbol{\beta})]h[-\boldsymbol{\alpha}]\,. \qquad \text{“1. Assumption on } h\text{”}\end{aligned}$$

Plugging this back into the assumed bound yields for any $\boldsymbol{\alpha}\in\mathbb{R}^p$

$$\begin{aligned}\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2 \;\le\;& \|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2\\ &+2\overline{h}[X^\top(\boldsymbol{y}-X\boldsymbol{\beta})]h[\widehat{\boldsymbol{\beta}}]+2\overline{h}[X^\top(\boldsymbol{y}-X\boldsymbol{\beta})]h[-\boldsymbol{\alpha}]+\widehat{w}rh[\boldsymbol{\alpha}]-\widehat{w}rh[\widehat{\boldsymbol{\beta}}]\,,\end{aligned}$$

and therefore, rearranging,

$$\begin{aligned}&\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2\\ \le\;& \|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+2\overline{h}[X^\top(\boldsymbol{y}-X\boldsymbol{\beta})]h[-\boldsymbol{\alpha}]+\widehat{w}rh[\boldsymbol{\alpha}]+\bigl(2\overline{h}[X^\top(\boldsymbol{y}-X\boldsymbol{\beta})]-\widehat{w}r\bigr)h[\widehat{\boldsymbol{\beta}}]\,,\end{aligned}$$

which concludes the proof of the first claim since $\boldsymbol{\alpha}$ was arbitrary. The other two claims follow as indicated in the theorem. □

In our second probability bound, we make use of one additional assumption on the prior function:

4. h satisfies the triangle inequality.

Theorem 6.3.2 (Power-two Bounds)

We consider an estimator $\widehat{\boldsymbol{\beta}}$ that complies with Assumption 6.3.1. Now, if h satisfies Assumptions 1., 2. & 4., it holds for any $t\in(1,\infty]$, $m\in(0,\infty)$ that

$$\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2 \;\le\; \min_{\boldsymbol{\alpha}\in\widehat{\mathcal{B}}_m}\Bigl\{t\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+\frac{(t+1)^2m^2}{4(t-1)}\bigl(2\overline{h}[X^\top(X\boldsymbol{\beta}-\boldsymbol{y})]+\widehat{w}r\bigr)^2\Bigr\}\,,$$

where $\widehat{\mathcal{B}}_m:=\{\boldsymbol{\gamma}\in\mathbb{R}^p : h[\boldsymbol{\gamma}-\widehat{\boldsymbol{\beta}}]\le m\|X\boldsymbol{\gamma}-X\widehat{\boldsymbol{\beta}}\|_2\}$.
If additionally $\widehat{w}r\ge 2v\overline{h}[X^\top(X\boldsymbol{\beta}-\boldsymbol{y})]$ for a positive and finite constant

$v \in (0,\infty)$, the bound implies directly

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\le\; \min_{\boldsymbol{\alpha}\in\widehat{\mathcal{B}}_m}\Big\{t\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2 + \frac{(t+1)^2m^2(v+1)^2\widehat{w}^2r^2}{4(t-1)v^2}\Big\}.$$

The results again do not require that h is convex, but they *do* inflict strict assumptions on the model and the structure of X. These assumptions are somewhat hidden in the definition of the set $\widehat{\mathcal{B}}_m$, which needs to be sufficiently rich for reasonably small m in order to make the bounds useful. We discuss conditions under which this holds in the following section. In any case, the second bound in the theorem is an oracle inequality for prediction as seen in the first section. More specifically, since the bound contains the tuning parameter to the second power, it is a power-two bound. It is a general fact that power-two bounds always hinge on strict conditions, while power-one bounds can be derived under minimal assumptions.

Proof of Theorem 6.3.2. The proof follows again from Assumption 6.3.1.

To see this, note that the inner product on the right-hand side of the inequality in that assumption can be bounded by

$$\begin{aligned}
&2\langle \boldsymbol{y} - X\boldsymbol{\beta},\, X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\alpha}\rangle \\
&= 2\langle X^\top(X\boldsymbol{\beta}-\boldsymbol{y}),\, \boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}}\rangle && \text{“properties of inner products”}\\
&\le 2\overline{h}[X^\top(X\boldsymbol{\beta}-\boldsymbol{y})]h[\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}}]\,. && \text{“1. Assumption on } h\text{”}
\end{aligned}$$

Plugging this back into the assumed bound yields for any $\boldsymbol{\alpha}\in\mathbb{R}^p$

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\le\; \|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2 + 2\overline{h}[X^\top(X\boldsymbol{\beta}-\boldsymbol{y})]h[\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}}] + \widehat{w}rh[\boldsymbol{\alpha}] - \widehat{w}rh[\widehat{\boldsymbol{\beta}}]\,,$$

and further, using the triangle inequality of Assumption 4. on h and rearranging terms,

$$\begin{aligned}
&\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2\\
&\le \|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2 + 2\overline{h}[X^\top(X\boldsymbol{\beta}-\boldsymbol{y})]h[\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}}] + \widehat{w}rh[\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}}]\\
&= \|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2 + \big(2\overline{h}[X^\top(X\boldsymbol{\beta}-\boldsymbol{y})] + \widehat{w}r\big)h[\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}}]\,.
\end{aligned}$$

Using now the first part of Lemma B.1.2 with $v = 2(t-1)/((t+1)b^2)$ and $\mathbb{R}^p = \mathbb{R}$, $t > 1$ to the last term in the above display yields

$$\begin{aligned}
&\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2\\
&\le \|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2 + \frac{(t+1)m^2}{2(t-1)}\big(2\overline{h}[X^\top(X\boldsymbol{\beta}-\boldsymbol{y})]+\widehat{w}r\big)^2 + \frac{t-1}{2(t+1)m^2}\big(h[\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}}]\big)^2\,.
\end{aligned}$$

Next, we have by definition of $\widehat{\mathcal{B}}_m$ that $h[\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}}] \le m\|X\boldsymbol{\alpha}-X\widehat{\boldsymbol{\beta}}\|_2$ for any $\boldsymbol{\alpha}\in\widehat{\mathcal{B}}_m$. Using this in above display implies for such $\boldsymbol{\alpha}$

$$\begin{aligned}
&\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2\\
&\le \|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2 + \frac{(t+1)m^2}{2(t-1)}\big(2\overline{h}[X^\top(X\boldsymbol{\beta}-\boldsymbol{y})]+\widehat{w}r\big)^2 + \frac{t-1}{2(t+1)}\|X\boldsymbol{\alpha}-X\widehat{\boldsymbol{\beta}}\|_2^2\,.
\end{aligned}$$

Now, adding a zero-valued term and using $(k+k')^2 \le 2k^2+2k'^2$ for all $k,k'\in\mathbb{R}$, and thus, $\|\mathbf{k}+\mathbf{k}'\|_2^2 \le 2\|\mathbf{k}\|_2^2 + 2\|\mathbf{k}'\|_2^2$ for all $\mathbf{k},\mathbf{k}'\in\mathbb{R}^n$, we find

$$\|X\boldsymbol{\alpha}-X\widehat{\boldsymbol{\beta}}\|_2^2 \;=\; \|\underbrace{X\boldsymbol{\alpha}-X\boldsymbol{\beta}}_{\mathbf{k}}+\underbrace{X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}}_{\mathbf{k}'}\|_2^2 \le 2\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2 + 2\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2\,.$$

Plugging this into the penultimate display yields.

$$\begin{aligned}&\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2\\ &\leq\ \|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+\frac{(t+1)m^2}{2(t-1)}\big(2\overline{h}\big[\big(X^\top(X\boldsymbol{\beta}-\boldsymbol{y})\big]+\widehat{w}r\big)^2+\frac{t-1}{t+1}\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2\\ &+\frac{t-1}{t+1}\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2\,.\end{aligned}$$

Therefore, by consolidating terms,

$$\frac{2}{t+1}\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2\ \leq\ \frac{2t}{t+1}\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+\frac{(t+1)m^2}{2(t-1)}\big(2\overline{h}[X^\top(X\boldsymbol{\beta}-\boldsymbol{y})]+\widehat{w}r\big)^2$$

and then

$$\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}\|_2^2\ \leq\ t\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+\frac{(t+1)^2m^2}{4(t-1)}\big(2\overline{h}[X^\top(X\boldsymbol{\beta}-\boldsymbol{y})]+\widehat{w}r\big)^2\,.$$

This concludes the proof. □

6.4 The Power-two Bounds in Sparse and Weakly Correlated Models

The minima in power-one bounds are taken over the entire parameter space $\mathbb{R}^p$, while the minima in power-two bounds are taken over only subsets of the parameter space. These subsets can well be "too small" to contain good minimizers, and then, power-two bounds are less informative than power-one bounds—or even completely void. On the other hand, there are also cases where power-one bounds do provide very useful guarantees. In this section, we identify such latter cases. The corresponding criteria are sparsity and correlations: the regression vector must allow for a sparse approximation such that the correlations both among the predictors that correspond to the support of that surrogate vector and between those predictors and the remaining ones are low. These criteria essentially ensure correct parameter estimation, which connects this section to the following chapter on estimation and variable selection theory.

The central quantity is the random set

$$\widehat{\mathcal{B}}_m\ =\ \big\{\boldsymbol{\gamma}\in\mathbb{R}^p\ :\ h[\boldsymbol{\gamma}-\widehat{\boldsymbol{\beta}}]\ \leq\ m\|X\boldsymbol{\gamma}-X\widehat{\boldsymbol{\beta}}\|_2\big\}\,.$$

The following example demonstrates that this set is closely connected with the correlations among the predictors.

Example 6.4.1 (A Link Between $\widehat{\mathcal{B}}_m$ and the Correlations in the Design)

In this example, we show that the structural condition is closely linked to the correlations in the design: the larger the correlations, the more restrictive the condition. Assuming that the columns of X have fixed and equal lengths, a measure of the correlations is the smallest singular value of X (that is, the square-root of the smallest eigenvalue of the gram matrix $X^\top X$):

$$w \;:=\; \min_{\boldsymbol{\delta}\in\mathbb{R}^p\setminus\mathbf{0}_p}\sqrt{\frac{\boldsymbol{\delta}^\top X^\top X\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2^2}}\,.$$

The larger w, the weaker the correlations: for example, if all columns have unit length, $w\approx 0$ indicates high correlations (in particular, $w=0$ if any columns are linearly dependent), while $w\approx 1$ indicates small correlations (in particular, $w=1$ if all columns are orthogonal).

For the function $h\;:\;\boldsymbol{\alpha}\mapsto\|\boldsymbol{\alpha}\|_2$, for example, we then find (set $\boldsymbol{\delta}=\boldsymbol{\gamma}-\widehat{\boldsymbol{\beta}}$ in the above inequality)

$$h[\boldsymbol{\gamma}-\widehat{\boldsymbol{\beta}}]\;=\;\|\boldsymbol{\gamma}-\widehat{\boldsymbol{\beta}}\|_2\;\le\;\frac{1}{w}\|X\boldsymbol{\gamma}-X\widehat{\boldsymbol{\beta}}\|_2$$

for all $\boldsymbol{\gamma}\in\mathbb{R}^p$, and similarly that there exists an $\boldsymbol{\gamma}\in\mathbb{R}^p$ such that $h[\boldsymbol{\gamma}-\widehat{\boldsymbol{\beta}}]>(1/w-u)\|X\boldsymbol{\gamma}-X\widehat{\boldsymbol{\beta}}\|_2$ for all $u>0$. Therefore, $\widehat{\mathcal{B}}_m=\mathbb{R}^p$ if and only if $m\ge 1/w$. This means that the larger the correlations, the larger m needs to be to ensure that the structural condition holds for all possible parameters.

The power-two bounds of Theorem 6.3.2 are useful only if both $\widehat{\mathcal{B}}_m$ is rich enough to contain a good minimizer of the righ-hand sides in those bounds and the corresponding m is small enough to keep the second terms in those bounds at bay. The preceeding example relates this to the correlations among the predictors. In particular, $\widehat{\mathcal{B}}_m$ is even the full $\mathbb{R}^p$ for reasonably small m if the singular values of the gram matrix are sufficiently large. However, this is not a realistic scenario in high-dimensional statistics: the more predictors there are, the more likely there are strong correlations, maybe even collinearities. If $p>n$, collinearities are unavoidable altogether, and then $w=0$ (see 1. in Exercise 1.2), which means that $\widehat{\mathcal{B}}_m=\mathbb{R}^p$ if and only if $m=\infty$. Thus, power-two bounds become increasingly iffy with increasing dimensionality of the model.

Nevertheless, there is still margin for informative power-two bounds in high dimensions. In the remainder of this section, we identify this margin by showing that rather than *all* correlations among the predictors to be small, it is sufficient to ask for small correlations among predictive variables and between these variables and all other variables, that is, unpredictive variables can be correlated arbitrarily.

The set $\widehat{\mathcal{B}}_m$ intertwines prediction and estimation: it formulates that accurate prediction of a $\boldsymbol{\alpha}\in\widehat{\mathcal{B}}_m$ (that is, $m\|X\boldsymbol{\alpha}-X\widehat{\boldsymbol{\beta}}\|_2$ small) also implies accurate h-estimation of $\boldsymbol{\alpha}$ (that is, $h[\boldsymbol{\alpha}-\widehat{\boldsymbol{\beta}}]$ small). Let us, therefore, look at estimation of $\boldsymbol{\beta}$ for a moment. For simplicity, consider the noiseless case $\boldsymbol{u}=\mathbf{0}_n$, that is,

$$\boldsymbol{y}\;=\;X\boldsymbol{\beta}\,.$$

If all singular values of X are strictly positive as assumed in the above example, the equality can be directly solved for $\boldsymbol{\beta}$, and $\boldsymbol{\beta}$ is the unique vector that satisfies the equality. If X is singular instead, which is always the case for $p>n$, it seems that there is no hope to recover $\boldsymbol{\beta}$ accurately; indeed, $X\boldsymbol{\beta}=X(\boldsymbol{\beta}+\boldsymbol{\gamma})$ for any $\boldsymbol{\gamma}$ in the kernel of X, so that both $\boldsymbol{\beta}$ and $\boldsymbol{\beta}+\boldsymbol{\gamma}$ seem to be equally good candiates for parameterizing the model of $\boldsymbol{y}$. However, if $\boldsymbol{\beta}$ is sparse, that is, that $\operatorname{supp}[\boldsymbol{\beta}]=\mathcal{S}$ with $|\mathcal{S}|\ll n,p$ say,

$$\boldsymbol{y}\;=\;X_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\,,$$

and we only need invertibility of the restricted matrix $X_{\mathcal{S}}^\top X_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$, or stated differently,

$$h[\boldsymbol{\delta}] \;\leq\; m\|X\boldsymbol{\delta}\|_2 \qquad\qquad \boldsymbol{\delta} \;\in\; \mathbb{R}^p \;:\; \boldsymbol{\delta}_{\mathcal{S}^\complement} \;=\; \mathbf{0}_{p-|\mathcal{S}|}$$

for a $m > 0$. To actually do the inverse, one would need to know or estimate $\mathcal{S}$, but the point is that recovering $\boldsymbol{\beta}$ is in principle possible.

In our framework, we allow for non-zero noise $\boldsymbol{u}$ and for $\boldsymbol{\beta}$ that are only approximately sparse, that is, $\boldsymbol{\beta}$ that are not necessarily sparse themselves but that can be approximated by sparse vectors $\boldsymbol{\alpha}$. In other words, we do not necessarily talk about $\boldsymbol{\beta} \in \widehat{\mathcal{B}}_m$, but rather ask for $\boldsymbol{\alpha} \in \widehat{\mathcal{B}}_m$ for a sparse approximation of $\boldsymbol{\beta}$. Therefore, we need to ask slightly more than invertibility of the restricted matrix $X_{\mathcal{S}}^\top X_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$: we will assume that $X^\top X$ is invertible on the space of vectors that have much of their mass on their $\mathcal{S}$-coordinates. Assuming that h can be decomposed into two functions $h_{\mathcal{S}}, h_{\mathcal{S}^\complement}$ such that $h[\boldsymbol{\alpha}] = h_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}}] + h_{\mathcal{S}^\complement}[\boldsymbol{\alpha}_{\mathcal{S}^\complement}]$ for all $\boldsymbol{\alpha} \in \mathbb{R}^p$, we state the following condition:

compatiblity condition

Assumption 6.4.1 (Compatibility Condition) Given a function $h : \mathbb{R}^p \to \mathbb{R}$, a non-empty set $\mathcal{S} \subset \{1, \dots, p\}$, and positive and finite constants $m \in (0, \infty)$, $v \in (1, \infty)$, we say that the *compatibility condition* holds if

$$h[\boldsymbol{\delta}] \;\leq\; m\|X\boldsymbol{\delta}\|_2 \qquad\qquad \Big(\boldsymbol{\delta} \;\in\; \mathbb{R}^p \;:\; h_{\mathcal{S}^\complement}[\boldsymbol{\delta}_{\mathcal{S}^\complement}] \;\leq\; \frac{v+1}{v-1} h_{\mathcal{S}}[\boldsymbol{\delta}_{\mathcal{S}}]\Big)\,,$$

We call m the *compatibility constant.*

Similarly as the assumption on the minimal eigenvalues, the compatibility condition restricts the correlations among the columns of the design matrix X. However, the crux is that the inequality needs to hold only for certain vectors $\boldsymbol{\delta}$'s; for example, correlations *only* among columns with indexes in $\mathcal{S}$ and among such columns with any other columns are concerned; in contrast, the columns with indexes in $\mathcal{S}^\complement$ can basically be arbitrarily correlated among each other. The following example supports this interpretation (see also Exercise 6.7 for an example with concrete numbers).

Example 6.4.2 (Prototypical Cases Where the Compatibility Condition Holds)

In this example, we highlight a class of settings where the compatiblity condition of Assumption 6.4.1 is met. We consider fixed constants $s \in \{1, \dots, p\}$ and $m \in (0, \infty)$, the set $\mathcal{S} := \{1, \dots, s\}$, and a design with gram matrix

$$X^\top X \;=\; \begin{pmatrix} A & \\ & B \end{pmatrix}\,,$$

where the smallest eigenvalue of $A \in \mathbb{R}^{s\times s}$ is assumed to be larger or equal to $(2sv/(m(v-1)))^2$ and $B \in \mathbb{R}^{(p-s)\times(p-s)}$ is positive semi-definite by construction (see Exercise **??**). The lower bound on the eigenvalues of A ensures that the columns of X with indexes in $\mathcal{S}$ are only weakly correlated with each other, and the block structure of the gram matrix ensures that the mentioned columns are orthogonal to all remaining columns. We then find for all $\boldsymbol{\delta} \in \mathbb{R}^p$

$$\begin{aligned} m\|X\boldsymbol{\delta}\|_2 \;&=\; m\sqrt{\boldsymbol{\delta}^\top X^\top X\boldsymbol{\delta}} && \text{“definition of the } \ell_2\text{-norm”} \\ &=\; m\sqrt{\boldsymbol{\delta}_{\mathcal{S}}^\top A\boldsymbol{\delta}_{\mathcal{S}} + \boldsymbol{\delta}_{\mathcal{S}^\complement}^\top B\boldsymbol{\delta}_{\mathcal{S}^\complement}} && \text{“assumed form of } X^\top X\text{”} \end{aligned}$$

$$
\begin{aligned}
&\geq\; m\sqrt{\left(\frac{2sv}{m(1-v)}\right)^{2}\|\boldsymbol{\delta}_{\mathcal{S}}\|_2^2+0}\\
&\qquad\text{“smallest eigenvalue of } A \text{ is at least } (2sv/(m(v-1)))^2;\ B \text{ is positive semi-definite”}\\
&\geq\; \frac{2sv}{v-1}\|\boldsymbol{\delta}_{\mathcal{S}}\|_2 &&\text{“consolidating”}\\
&=\; \frac{2sv}{v-1}\sqrt{\sum_{j=1}^{s}(\boldsymbol{\delta}_{\mathcal{S}})_j^2} &&\text{“definition of the } \ell_2\text{-norm”}\\
&\geq\; \frac{2sv}{v-1}\max_{j\in\{1,\dots,s\}}|(\boldsymbol{\delta}_{\mathcal{S}})_j|\\
&\qquad\text{“}\sqrt{\textstyle\sum_{j=1}^{s}a_j^2}\geq\sqrt{\max_{j\in\{1,\dots,s\}}a_j^2}=\max_{j\in\{1,\dots,s\}}|a_j| \text{ for all } a_1,\dots,a_s\in\mathbb{R}\text{”}\\
&\geq\; \frac{2v}{v-1}\sum_{j=1}^{s}|(\boldsymbol{\delta}_{\mathcal{S}})_j|\\
&\qquad\text{“}s\cdot\max_{j\in\{1,\dots,s\}}|a_j|\geq\textstyle\sum_{j=1}^{s}|a_j| \text{ for all } a_1,\dots,a_s\in\mathbb{R}\text{”}\\
&=\; \frac{2v}{v-1}\|\boldsymbol{\delta}_{\mathcal{S}}\|_1\,. &&\text{“definition of the } \ell_1\text{-norm”}
\end{aligned}
$$

We now consider the function $h\ :\ \boldsymbol{\alpha}\ \mapsto\ \|\boldsymbol{\alpha}\|_1$ with decomposition $h_{\mathcal{S}}\ :\ \mathbb{R}^s\ \to\ \mathbb{R};\ \boldsymbol{\alpha}_{\mathcal{S}}\ \mapsto\ \|\boldsymbol{\alpha}_{\mathcal{S}}\|_1$ and $h_{\mathcal{S}^\complement}\ :\ \mathbb{R}^{p-s}\ \to\ \mathbb{R};\ \boldsymbol{\alpha}_{\mathcal{S}^\complement}\ \mapsto\ \|\boldsymbol{\alpha}_{\mathcal{S}^\complement}\|_1$. For any $\boldsymbol{\delta}\in\mathbb{R}^p$ that satisfies $h_{\mathcal{S}^\complement}[\boldsymbol{\delta}_{\mathcal{S}^\complement}]\leq(v+1)h_{\mathcal{S}}[\boldsymbol{\delta}_{\mathcal{S}}]/(v-1)$, we then find

$$
\begin{aligned}
h[\boldsymbol{\delta}]\;&=\;\|\boldsymbol{\delta}\|_1 &&\text{“choice of } h\text{”}\\
&=\;\|\boldsymbol{\delta}_{\mathcal{S}}\|_1+\|\boldsymbol{\delta}_{\mathcal{S}^\complement}\|_1 &&\text{“definition of the } \ell_1\text{-norm”}\\
&\leq\;\|\boldsymbol{\delta}_{\mathcal{S}}\|_1+\frac{v+1}{v-1}\|\boldsymbol{\delta}_{\mathcal{S}}\|_1 &&\text{“}h_{\mathcal{S}^\complement}[\boldsymbol{\delta}_{\mathcal{S}^\complement}]\leq(v+1)h_{\mathcal{S}}[\boldsymbol{\delta}_{\mathcal{S}}]/(v-1)\text{ by assumption”}\\
&=\;\frac{2v}{v-1}\|\boldsymbol{\delta}_{\mathcal{S}}\|_1\,. &&\text{“consolidating”}
\end{aligned}
$$

Combing the two displays yields

$$
h[\boldsymbol{\delta}]\;\leq\;m\|X\boldsymbol{\delta}\|_2\qquad\Bigl(\boldsymbol{\delta}\;\in\;\mathbb{R}^p\;:\;h_{\mathcal{S}^\complement}[\boldsymbol{\delta}_{\mathcal{S}^\complement}]\;\leq\;\frac{v+1}{v-1}h_{\mathcal{S}}[\boldsymbol{\delta}_{\mathcal{S}}]\Bigr)\,,
$$

as required by Assumption 6.4.1.

We now show that Assumption 6.4.1 can indeed lead to sharp prediction bounds. As a first step, we show that regularized estimators are usually concentrated on the support of sparse approximations of the target. For this, we again invoke a basic inequality.

Assumption 6.4.2 (Generic Basic Inequality Revisited) Fix a function $h\ :\ \mathbb{R}^p\ \to\ [-\infty,\infty]$ and positive, finite, and potentially random quantities $\widehat{w},r\in(0,\infty)$. We assume that the estimator $\widehat{\boldsymbol{\beta}}$ satisfies for all $\boldsymbol{\alpha}\in\mathbb{R}^p$ the basic inequality

$$
\|X\boldsymbol{\alpha}-X\widehat{\boldsymbol{\beta}}\|_2^2\;\leq\;2\langle\boldsymbol{y}-X\boldsymbol{\alpha},\,X\widehat{\boldsymbol{\beta}}-X\boldsymbol{\alpha}\rangle+\widehat{w}rh[\boldsymbol{\alpha}]-\widehat{w}rh[\widehat{\boldsymbol{\beta}}]\,.
$$

This is a version of the generic basic inequality in Assumption 6.3.1 with $\boldsymbol{\beta}$ replaced by $\boldsymbol{\alpha}$. Because the target does not appear in the set $\widehat{\mathcal{B}}_m$ altogether, we use this version rather than the earlier one. Sufficient conditions for the assumption to hold can be derived exactly as in Section 6.2 with $\boldsymbol{\beta}$ replaced by $\boldsymbol{\alpha}$; in fact, since we have

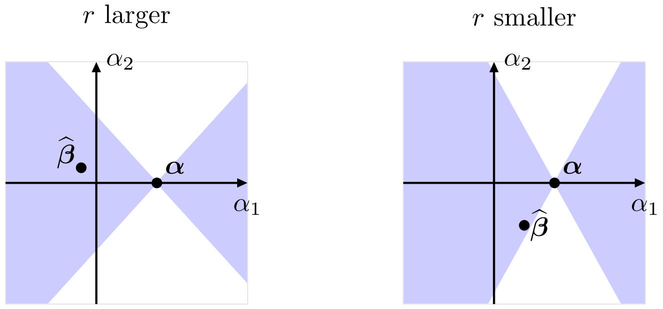


Figure 6.2: An illustration of the second part of Lemma 6.4.1 for $p = 2$, $\mathcal{S} = \{1\}$, and h the ℓ_1-norm. The estimator $\widehat{\boldsymbol{\beta}}$ is in an infinitely extended double-cone (blue area, *not* white area) with apex $\boldsymbol{\alpha}$. This guarantees that the vector of differences $\boldsymbol{\alpha} - \widehat{\boldsymbol{\beta}}$ has a considerable fraction of its mass in the first coordinate. Given an apex $\boldsymbol{\alpha}$, the smaller the tuning parameter r (yet, r needs to be sufficiently large for the second part of the lemma to hold in the first place), the smaller values for c need to be chosen, and the larger the cone. As a function of $\mathcal{S}$, the cone constraint is the more forceful the smaller $\mathcal{S}$ is relative to p, that is, the sparser the vector $\boldsymbol{\alpha}$ is.

not imposed any concrete model, $\boldsymbol{\beta}$ differs from arbitrary vectors $\boldsymbol{\alpha}$ only in that is considered the “target.”

We will also make use of five assumptions on the function h:

1. h satisfies the assumptions of Lemma B.1.3 (Hölder’s inequality);
2. h can be decomposed into two functions $h_{\mathcal{S}}, h_{\mathcal{S}^\complement}$ such that $h[\boldsymbol{\alpha}] = h_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}}] + h_{\mathcal{S}^\complement}[\boldsymbol{\alpha}_{\mathcal{S}^\complement}]$ for all $\boldsymbol{\alpha} \in \mathbb{R}^p$;
3. $h_{\mathcal{S}^\complement}$ obeys the triangle inequality;
4. $h_{\mathcal{S}}, h_{\mathcal{S}^\complement}$ are symmetric;
5. $h_{\mathcal{S}}, h_{\mathcal{S}^\complement}$ are non-negative.

We now find the following.

Lemma 6.4.1 (Double-cone)

Consider an estimator $\widehat{\boldsymbol{\beta}}$ that complies with Assumption 6.4.2, where the function h satisfies the first four conditions above. Then, it holds for any vector $\boldsymbol{\alpha} \in \mathbb{R}^p$ with $\operatorname{supp}[\boldsymbol{\alpha}] \subset \mathcal{S}$, $h_{\mathcal{S}^\complement}[\boldsymbol{\alpha}_{\mathcal{S}^\complement}] = 0$ that

$$\begin{aligned}\|X\boldsymbol{\alpha} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\le\; & \bigl(2\overline{h}[X^\top(X\boldsymbol{\alpha} - \boldsymbol{y})] + \widehat{w}r\bigr)h_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}]\\ & + \bigl(2\overline{h}[X^\top(X\boldsymbol{\alpha} - \boldsymbol{y})] - \widehat{w}r\bigr)h_{\mathcal{S}^\complement}[\boldsymbol{\alpha}_{\mathcal{S}^\complement} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}^\complement}]\,.\end{aligned}$$

If additionally $\widehat{w}r \ge 2v\overline{h}[X^\top(X\boldsymbol{\alpha} - \boldsymbol{y})]$ given a positive and finite constant $v \in (1, \infty)$, and if h satisfies the fifth condition above, the display entails in particular

$$h_{\mathcal{S}^\complement}[\boldsymbol{\alpha}_{\mathcal{S}^\complement} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}^\complement}] \;\le\; \frac{v+1}{v-1}h_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}]\,.$$

A set $\mathcal{V} \subset \mathbb{R}^p$ is called an (infinite) *double-cone* with apex $\mathbf{v} \in \mathbb{R}^p$ if $a(\mathbf{v} + \mathbf{w}) \in \mathcal{V}$ for all $a \in \mathbb{R}$ and all $\mathbf{w} \in \mathbb{R}^p$ that satisfy $\mathbf{v} + \mathbf{w} \in \mathcal{V}$. If the functions $h_{\mathcal{S}}, h_{\mathcal{S}^\complement}$ are absolutely homogenenous, that is, $h_{\mathcal{S}}[a\boldsymbol{\gamma}_{\mathcal{S}}] = |a|h_{\mathcal{S}}[\boldsymbol{\gamma}_{\mathcal{S}}]$, $h_{\mathcal{S}^{\complement}}[a\boldsymbol{\gamma}_{\mathcal{S}^{\complement}}] = |a|h_{\mathcal{S}^{\complement}}[\boldsymbol{\gamma}_{\mathcal{S}^{\complement}}]$ for all $a \in \mathbb{R}$, $\boldsymbol{\gamma} \in \mathbb{R}^p$, the set

double-cone

$$\mathcal{C}_{\boldsymbol{\alpha},c} \;:=\; \Bigl\{\boldsymbol{\gamma} \in \mathbb{R}^p \;:\; h_{\mathcal{S}^{\complement}}[\boldsymbol{\alpha}_{\mathcal{S}^{\complement}} - \boldsymbol{\gamma}_{\mathcal{S}^{\complement}}] \;\leq\; \frac{1+c}{1-c} h_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}} - \boldsymbol{\gamma}_{\mathcal{S}}]\Bigr\}$$

is a double-cone with apex $\boldsymbol{\alpha}$, and the lemma ensures that $\widehat{\boldsymbol{\beta}} \in \mathcal{C}_{\boldsymbol{\alpha},v}$. The constant v is connected to r and $\boldsymbol{\alpha}$: the larger the tuning parameter r and the better the fit of the vector $\boldsymbol{\alpha}$ to the data in terms of $\overline{h}[X^\top(X\boldsymbol{\alpha} - \boldsymbol{y})]$, the larger v can be chosen in the lemma. Large v mean small cones, which in turn means that the $h_{\mathcal{S}^{\complement}}$-size of $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\alpha})_{\mathcal{S}^{\complement}}$ is sharply constraint by the $h_{\mathcal{S}}$-size of $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\alpha})_{\mathcal{S}}$. The second part of the lemma thus states that if the tuning parameters are sufficiently large, the estimator's mass is concentrated on the support of any vector that fits the data well. Of course, the hope is that the data can be fitted by *sparse* vectors, that is, vectors with small support $\mathcal{S}$; then, the stated inequality restricts the elements off $\mathcal{S}$ by the small number of elements in $\mathcal{S}$, and therefore, is most meaningful. An illustration is provided in Figure 6.2.

There are no explicit restrictions on the design matrix X, but in view of the factor $\overline{h}[X^\top(X\boldsymbol{\alpha} - \boldsymbol{y})]$, the result is most informative if $\boldsymbol{y}$ can be approximated well by columns with indexes in $\mathcal{S}$.

Most importantly, the lemma ensures the following:

Corollary 6.4.1 ($\widehat{\mathcal{B}}_m$ Under the Compatibility Condition)

Assume that the conditions of the second part of Lemma 6.4.1 (double-cone) are met and that the Compatibility Condition 6.4.1 (compatibility condition) holds. Then,

$$\widehat{\mathcal{B}}_m \;\supset\; \bigl\{\boldsymbol{\gamma} \in \mathbb{R}^p \;:\; \boldsymbol{\gamma}_{\mathcal{S}^{\complement}} = \mathbf{0}_{|\mathcal{S}^{\complement}|}, h_{\mathcal{S}^{\complement}}[\boldsymbol{\gamma}_{\mathcal{S}^{\complement}}] = 0, \overline{h}[X^\top(X\boldsymbol{\gamma} - \boldsymbol{y})] \leq \widehat{w}r/(2(1+v))\bigr\}\,.$$

This means that under the stated assumptions, $\widehat{\mathcal{B}}_m$ contains all sparse vectors that describe the data well.

Proof of Lemma 6.4.1. The proof consists of manipulating Assumption 6.4.2 with techniques similar to those used in the previous section.

The inner product in that assumption can be bounded by virtue of Hölder's inequality as mentioned in the first condition on h, Page 107:

$$\begin{aligned} &2\langle \boldsymbol{y} - X\boldsymbol{\alpha},\, X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\alpha}\rangle \\ &\;=\; 2\langle X^\top(X\boldsymbol{\alpha} - \boldsymbol{y}),\, \boldsymbol{\alpha} - \widehat{\boldsymbol{\beta}}\rangle && \text{“properties of inner products”} \\ &\;\leq\; 2\overline{h}[X^\top(X\boldsymbol{\alpha} - \boldsymbol{y})]h[\boldsymbol{\alpha} - \widehat{\boldsymbol{\beta}}]\,. && \text{“1. Assumption on } h\text{”} \end{aligned}$$

Plugging this into the basic inequality yields

$$\|X\boldsymbol{\alpha} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; 2\overline{h}[X^\top(X\boldsymbol{\alpha} - \boldsymbol{y})]h[\boldsymbol{\alpha} - \widehat{\boldsymbol{\beta}}] + \widehat{w}rh[\boldsymbol{\alpha}] - \widehat{w}rh[\widehat{\boldsymbol{\beta}}]\,.$$

Next, with the decomposabilty of h assumed in 2., we find

$$\begin{aligned} &\|X\boldsymbol{\alpha} - X\widehat{\boldsymbol{\beta}}\|_2^2 \\ &\qquad\leq\; 2\overline{h}[\boldsymbol{y} - X^\top(X\boldsymbol{\alpha})]h_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}] + 2\overline{h}[X^\top(X\boldsymbol{\alpha} - \boldsymbol{y})]h_{\mathcal{S}^{\complement}}[\boldsymbol{\alpha}_{\mathcal{S}^{\complement}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}^{\complement}}] \\ &\qquad\qquad + \widehat{w}rh_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}}] + \widehat{w}rh_{\mathcal{S}^{\complement}}[\boldsymbol{\alpha}_{\mathcal{S}^{\complement}}] - \widehat{w}rh_{\mathcal{S}}[\widehat{\boldsymbol{\beta}}_{\mathcal{S}}] - \widehat{w}rh_{\mathcal{S}^{\complement}}[\widehat{\boldsymbol{\beta}}_{\mathcal{S}^{\complement}}]\,. \end{aligned}$$

Further, we can use $\boldsymbol{\alpha}_{\mathcal{S}^\complement} = \mathbf{0}_{|\mathcal{S}^\complement|}$ to modify the last term on the right-hand side and $h_{\mathcal{S}^\complement}(\boldsymbol{\alpha}_{\mathcal{S}^\complement}) = 0$ to remove the second term on the last line:

$$\begin{aligned}
&\|X\boldsymbol{\alpha} - X\widehat{\boldsymbol{\beta}}\|_2^2\\
&\qquad\le\; 2\overline{h}[X^\top(X\boldsymbol{\alpha}-\boldsymbol{y})]h_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}] + 2\overline{h}[X^\top(X\boldsymbol{\alpha}-\boldsymbol{y})]h_{\mathcal{S}^\complement}[\boldsymbol{\alpha}_{\mathcal{S}^\complement} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}^\complement}]\\
&\qquad\qquad + \widehat{w}rh_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}}] - \widehat{w}rh_{\mathcal{S}}[\widehat{\boldsymbol{\beta}}_{\mathcal{S}}] - \widehat{w}rh_{\mathcal{S}^\complement}[\widehat{\boldsymbol{\beta}}_{\mathcal{S}^\complement} - \boldsymbol{\alpha}_{\mathcal{S}^\complement}]\,.
\end{aligned}$$

Moreover, we can use the triangle inequality assumed in 3. to combine the first and third term on the bottom line to derive

$$\begin{aligned}
&\|X\boldsymbol{\alpha} - X\widehat{\boldsymbol{\beta}}\|_2^2\\
&\qquad\le\; 2\overline{h}[X^\top(X\boldsymbol{\alpha}-\boldsymbol{y})]h_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}] + 2\overline{h}[X^\top(X\boldsymbol{\alpha}-\boldsymbol{y})]h_{\mathcal{S}^\complement}[\boldsymbol{\alpha}_{\mathcal{S}^\complement} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}^\complement}]\\
&\qquad\qquad + \widehat{w}rh_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}] - \widehat{w}rh_{\mathcal{S}^\complement}[\widehat{\boldsymbol{\beta}}_{\mathcal{S}^\complement} - \boldsymbol{\alpha}_{\mathcal{S}^\complement}]\,.
\end{aligned}$$

We can finally use the symmetry assumed in 4. and summarize the terms to

$$\begin{aligned}
\|X\boldsymbol{\alpha} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\le\; &\big(2\overline{h}[X^\top(X\boldsymbol{\alpha}-\boldsymbol{y})] + \widehat{w}r\big)h_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}]\\
&+ \big(2\overline{h}[X^\top(X\boldsymbol{\alpha}-\boldsymbol{y})] - \widehat{w}r\big)h_{\mathcal{S}^\complement}[\boldsymbol{\alpha}_{\mathcal{S}^\complement} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}^\complement}]
\end{aligned}$$

as claimed in the first part.

For the second part, we rearrange the above result and use that $\|X\boldsymbol{\alpha} - X\widehat{\boldsymbol{\beta}}\|_2^2 \ge 0$ by definition of ℓ_2-norms:

$$\big(\widehat{w}r - 2\overline{h}[X^\top(X\boldsymbol{\alpha}-\boldsymbol{y})]\big)h_{\mathcal{S}^\complement}[\boldsymbol{\alpha}_{\mathcal{S}^\complement} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}^\complement}] \;\le\; \big(2\overline{h}[X^\top(X\boldsymbol{\alpha}-\boldsymbol{y})] + \widehat{w}r\big)h_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}]\,.$$

Now, if $\widehat{w}r \ge 2v\overline{h}[X^\top(X\boldsymbol{\alpha}-\boldsymbol{y})]$, we obtain

$$\Big(1 - \frac{1}{v}\Big)\widehat{w}rh_{\mathcal{S}^\complement}[\boldsymbol{\alpha}_{\mathcal{S}^\complement} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}^\complement}] \;\le\; \Big(1 + \frac{1}{v}\Big)\widehat{w}rh_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}]\,.$$

Here, we invoked the positivity in the 4. condition on h. Since $1/v < 1$ by assumption on v, we then find

$$\widehat{w}rh_{\mathcal{S}^\complement}[\boldsymbol{\alpha}_{\mathcal{S}^\complement} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}^\complement}] \;\le\; \frac{1+\frac{1}{v}}{1-\frac{1}{v}}\widehat{w}rh_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}] \;=\; \frac{v+1}{v-1}\widehat{w}rh_{\mathcal{S}}[\boldsymbol{\alpha}_{\mathcal{S}} - \widehat{\boldsymbol{\beta}}_{\mathcal{S}}]\,,$$

from which the second claim can be derived by dividing through $\widehat{w}r > 0$. □

We conclude this section by collecting the pieces.

Theorem 6.4.1 (Power-two Bounds Revisited)

Assume that the conditions of the second part of Theorem 6.3.2 (power-two bounds) and the first part of Lemma 6.4.1 (double-cone) are met. Now, assume that the design is weakly correlated in the sense that

1. Assumption 6.4.1 (compatibility condition) holds.

Assume also that the parameter $\boldsymbol{\alpha}$ is a sparse fit to the data in the sense that

2. $\operatorname{supp}[\boldsymbol{\alpha}] \subset \mathcal{S}$ and $\widehat{w}r \ge 2v\overline{h}[X^\top(X\boldsymbol{\alpha}-\boldsymbol{y})]$;

and that the “true” regression vector $\boldsymbol{\beta}$ is close to $\boldsymbol{\alpha}$ in the sense that

3. $\|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 \leq dr^2$

for a $d \in [0, \infty]$. Then,

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; \left(2d + \frac{9m^2(v+1)\widehat{w}^2}{2v}\right)r^2\,.$$

This result can be proved in three steps: First, it can be shown that $\boldsymbol{\alpha} \in \widehat{\mathcal{B}}_m$ by using Assumption 6.4.1 and the first part of Lemma 6.4.1. Then, a probability bound can be obtained by using the second part of Theorem 6.3.2 with $t = 2$ and $\boldsymbol{\beta}$ replaced by $\boldsymbol{\alpha}$ (cf. the comments after Assumption 6.4.2). Finally, this bound can be transferred into a bound for $\boldsymbol{\beta}$ by applying the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ for all $a, b \in \mathbb{R}$. We leave the details to the reader.

The theorem presumes that the design is not too much correlated, that there is a sparse vector that fits the data, and that this vector is a good approximation of the true regression vector. The first and second assumptions could be verified at least in principle (barring computational challenges), while the third assumption cannot. We will find the same assumptions in the discussion of parameter estimation in the following chapter, which basically means that power-two bounds relate to both prediction and parameter estimation. In contrast, power-one bounds focus sharply on prediction and, therefore, manage with much less assumptions.

6.5 Risk Bounds*

In the preceeding sections, we have discussed prediction bounds that hold point-wise for the data $Z \in \mathcal{Z}$; for example, we have shown in Section 6.1 that the lasso with tuning parameter $r \geq 2\|X^\top \boldsymbol{u}\|_\infty$ satisfies

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\left\{\|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + 2r\|\boldsymbol{\alpha}\|_1\right\}$$

for each $Z = (\boldsymbol{y}, X) \in \mathbb{R}^n \times \mathbb{R}^{n\times p}$. We have called those oracle inequalities *probability bounds.*

In this section, we transform such probability bounds into oracle inequalities for the expected prediction loss $\mathbb{E}\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2$. Unfortunately, because of the conditions on the tuning parameters, we cannot integrate the probability bounds directly. Instead, we will show that $\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2$ is Lipschitz in the noise. For Lipschitz functions of random vectors, there are sharp bounds for their deviation around their median, and we can then integrate those deviation bounds. The resulting oracle inequalities are called *risk bounds.*

Probabilistic Properties of Lipschitz Functions

We first give inequalities that bound the deviation of Lipschitz functions of Gaussian vectors around their medians. For this, we recall the concepts of Lipschitz functions and medians.

Definition 6.5.1 (Lipschitz Functions)

A function $f:\mathbb{R}^n\to\mathbb{R}$ is called *c-Lipschitz*, $c\geq 0$, if

$$|f[\mathbf{a}]-f[\mathbf{b}]|\ \leq\ c\|\mathbf{a}-\mathbf{b}\|_2\qquad(\mathbf{a},\mathbf{b}\in\mathbb{R}^n)\,.$$

A sufficient condition is that the gradient of f exists everywhere and is bounded in Euclidean norm by c (one can prove this by using the mean value theorem). A function is a *contraction* if it is c-Lipschitz with $c\leq 1$.

Definition 6.5.2 (Median of a Distribution)

Consider a random variable $\boldsymbol{v}\in\mathbb{R}^n$ and a (Borel-) measurable function $f:\mathbb{R}^n\to\mathbb{R}$. A number $m\in\mathbb{R}$ is called a *median* of (the distribution of) $f[\boldsymbol{v}]$ if

$$\mathbb{P}\bigl\{f[\boldsymbol{v}]\leq m\bigr\}\ \geq\ 1/2\quad\text{and}\quad\mathbb{P}\bigl\{f[\boldsymbol{v}]\geq m\bigr\}\ \geq\ 1/2\,.$$

The result is now as follows.

Lemma 6.5.1 (Gaussian Concentration Around the Median)

Assume that $\boldsymbol{v}\in\mathbb{R}^n$ is a standard normally distributed random vector, $f:\mathbb{R}^n\to\mathbb{R}$ deterministic and 1-Lipschitz, and m a median of $f[\boldsymbol{v}]$. Then,

$$\max\Bigl\{\{\mathbb{P}\{f[\boldsymbol{v}]\geq m+t\},\mathbb{P}\bigl\{f[\boldsymbol{v}]\leq m-t\}\Bigr\}\ \leq\ 1-\breve{g}[t]\qquad(t>0)\,,$$

where $\breve{g}[t]$ is the cumulative distribution function of a standard normal random variable. We have equality in the bound if f is linear.
Under the same assumptions, it also holds that

$$\mathbb{V}\mathrm{ar}\bigl[f[\boldsymbol{v}]\bigr]\ \leq\ 1\,.$$

We have equality in the bound if f is linear and a median is equal to the mean.

This means that Lipschitz functionals of Gaussian random vectors sharply concentrate around the median.

Example 6.5.1 (A Simple Transformation of a Gaussian Random Vector)

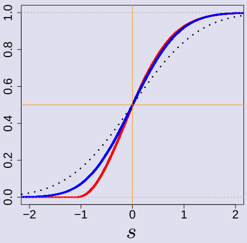
Consider a standard normal random vector $\boldsymbol{v}\in\mathbb{R}^n$ and the function $f:\mathbf{a}\mapsto\sqrt{\|\mathbf{a}\|_2^2+1}$. One can check that f is 1-Lipschitz and that the median of $f[\boldsymbol{v}]$ is $m=\sqrt{m_{\chi^2_n}+1}$, where $m_{\chi^2_n}$ is the median of a Chi-squared distribution with n degrees of freedom. Hence, Lemma 6.5.1 implies

$$\mathbb{P}\Bigl\{\sqrt{\|\boldsymbol{v}\|_2^2+1}\geq\sqrt{m_{\chi^2_n}+1}+t\Bigr\}\ \leq\ 1-\breve{g}[t]\qquad(t>0)$$
$$\mathbb{P}\Bigl\{\sqrt{\|\boldsymbol{v}\|_2^2+1}\leq\sqrt{m_{\chi^2_n}+1}-t\Bigr\}\ \leq\ 1-\breve{g}[t]\qquad(t>0)\,.$$

To validate this empirically, we rewrite the display in terms of the cumulative distribution function of $f[\boldsymbol{v}]-m$:

$$\mathbb{P}\Bigl\{\sqrt{\|\boldsymbol{v}\|_2^2+1}-\sqrt{m_{\chi_n^2}+1}\leq s\Bigr\}\begin{cases}\leq\breve{g}[s] & (s\leq 0)\\ \geq\breve{g}[s] & (s\geq 0)\end{cases}\,.$$

In the plot on the right, the solid lines depict $s\mapsto\mathbb{P}\{\sqrt{\|\boldsymbol{v}\|_2^2+1}-\sqrt{m_{\chi_n^2}+1}\leq s\}$ (red: $n=4$; blue: $n=100$); the dotted, black line depicts $s\mapsto\breve{g}[s]$. We confirm that the solid lines are never above the dotted line for $s\leq 0$ and never below the dotted line for $s\geq 0$—in accordance with the theoretical result above.



Proof of Lemma 6.5.1. The first part follows from standard isoperimetric inequalities; see [Lif12, Theorem 6.2] for a proof.

Our proof of the second part first exploits standard properties of variances and expectations and then transforms into an integration exercise.

We first derive

$$\begin{aligned}\mathbb{V}\mathrm{ar}\bigl[f[\boldsymbol{v}]\bigr]&\leq\mathbb{E}(f[\boldsymbol{v}]-m)^2 && \text{“Exercise 6.8”}\\ &=\int_0^\infty\mathbb{P}\bigl\{|f[\boldsymbol{v}]-m|^2\geq s\bigr\}\,ds && \text{“Exercise 6.9”}\\ &=\int_0^\infty\mathbb{P}\bigl\{|f[\boldsymbol{v}]-m|\geq\sqrt{s}\bigr\}\,ds && \text{“taking square-roots inside the probability”}\\ &=\int_0^\infty\mathbb{P}\bigl\{f[\boldsymbol{v}]-m\geq\sqrt{s}\bigr\}\,ds+\int_0^\infty\mathbb{P}\bigl\{m-f[\boldsymbol{v}]\geq\sqrt{s}\bigr\}\,ds\\ & && \text{“}\mathbb{P}\mathcal{A}\cup\mathcal{B}=\mathbb{P}\mathcal{A}+\mathbb{P}\mathcal{B}\text{ if }\mathcal{A}\cap\mathcal{B}=\varnothing\text{”}\\ &=\int_0^\infty\mathbb{P}\bigl\{f[\boldsymbol{v}]\geq m+\sqrt{s}\bigr\}\,ds+\int_0^\infty\mathbb{P}\bigl\{f[\boldsymbol{v}]\leq m-\sqrt{s}\bigr\}\,ds\\ & && \text{“rearranging terms inside the probabilities”}\\ &\leq 2\int_0^\infty 1-\breve{g}\bigl[\sqrt{s}\bigr]\,ds\,. && \text{“first part with }t=\sqrt{s}\text{”}\end{aligned}$$

The integral in the last line can now be evaluated by integrating by parts twice. First,

$$\begin{aligned}2\int_0^\infty 1-\breve{g}\bigl[\sqrt{s}\bigr]\,ds&=2\int_0^\infty\Bigl(1-\int_{-\infty}^{\sqrt{s}}\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx\Bigr)\,ds && \text{“definition of }\breve{g}\text{”}\\ &=2\int_0^\infty 2u\Bigl(1-\int_{-\infty}^{u}\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx\Bigr)\,du\\ & && \text{“change of variables: }u=\sqrt{s};\ ds/du=2u\text{”}\\ &=2u^2\Bigl(1-\int_{-\infty}^{u}\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx\Bigr)\Bigr|_0^\infty-2\int_0^\infty u^2\Bigl(-\frac{e^{-u^2/2}}{\sqrt{2\pi}}\Bigr)\,du\\ & && \text{“integrating by parts”}\\ &=2\int_0^\infty u^2\frac{e^{-u^2/2}}{\sqrt{2\pi}}\,du\,.\\ & && \text{“evaluating the first term; simplifying the second term”}\end{aligned}$$

Note that one can apply l'Hôpital's rule for evaluating the first term of the second to the last line at $u=\infty$.

Second,

$$\begin{aligned}
2\int_0^\infty u^2\frac{e^{-u^2/2}}{\sqrt{2\pi}}\,du \;&=\; -2\int_0^\infty u\Bigl(\frac{d}{du}\frac{e^{-u^2/2}}{\sqrt{2\pi}}\Bigr)\,du \\
&\qquad\text{“writing the exponential function as a derivative”}\\
&=\; -2u\frac{e^{-u^2/2}}{\sqrt{2\pi}}\Big|_0^\infty+2\int_0^\infty\frac{e^{-u^2/2}}{\sqrt{2\pi}}\,du && \text{“integrating by parts”}\\
&= 2\int_0^\infty\frac{e^{-u^2/2}}{\sqrt{2\pi}}\,du && \text{“evaluating first term”}\\
&= \int_{-\infty}^\infty\frac{e^{-u^2/2}}{\sqrt{2\pi}}\,du && \text{“symmetry of integrand”}\\
&= 1\,. && \text{“integrating Gaussian density”}
\end{aligned}$$

Again, one can use l'Hôpital's rule for the evaluation after integrating by parts.

Collecting the terms, we conclude

$$\mathbb{V}\mathrm{ar}\bigl[f[\boldsymbol{v}]\bigr]\;\le\;1\,,$$

as desired.

□

Regularized Least-squares Estimators Are Lipschitz

In standard least-squares, the outcome is projected on the column space of the design matrix. Since projections are contractions, least-squares prediction is 1-Lipschitz in the outcome. The question of this section is whether prediction remains Lipschitz if we complement the least-squares objective with a regularizer.

We consider

$$\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\;\in\;\operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\bigl\{\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2+rh[\boldsymbol{\alpha}]\bigr\}\,,$$

where $r\ge 0$ is a tuning parameter and $h:\mathbb{R}^p\to\mathbb{R}$ a prior function that (i) is convex, (ii) satisfies the conditions of Hölder's inequality, and (iii) is positive homogenous: $h[a\mathbf{b}]=ah[\mathbf{b}]$ for all $a>0$ and $\mathbf{b}\in\mathbb{R}^p$. The explicit inclusion of $\boldsymbol{y}$ in the notation $\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]$ manifests our interest in the estimator as a function of $\boldsymbol{y}$; meanwhile, X is assumed fix.

The answer to the initial question is positive.

Lemma 6.5.2 (Lipschitz Property of Regularized Least-squares)

The prediction function $\boldsymbol{y}\mapsto X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]$ is 1-Lipschitz.

The result holds irrespective of r, and the special case $r=0$ corresponds to the traditional least-squares estimator.

The convexity assumed in (i) allows us to use the KKT conditions of Lemma B.1.5; Hölder's inequality of Lemma B.1.3 assumed in (ii) to bound remainder terms as usual; and the homogeneity assumed in (iii) to manipulate arguments of the prior function. However, at the core of the proof are again—as for the standard least-squares—projection arguments that revolve around Lemma B.1.6.

Proof of Lemma 6.5.2. The proof relies on the optimality conditions of the regularized least-squares estimator and on Lemma B.1.6 about ℓ_2-projections.

Step 1: We first set $\mathcal{A} := \{\boldsymbol{a} \in \mathbb{R}^n \,:\, \overline{h}[X^\top \boldsymbol{a}] \leq r/2\}$ and $P_{\mathcal{A}}[\boldsymbol{y}] := \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]$ and show that $P_{\mathcal{A}}[\boldsymbol{y}] \in \mathcal{A}$ for all $\boldsymbol{y} \in \mathbb{R}^n$.

For illustration, one can check that for given X with rank n, the least-squares estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}[\boldsymbol{y}]$ satisfies $P_{\mathcal{A}}[\boldsymbol{y}] = \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}[\boldsymbol{y}] = \mathbf{0}_n$ for all $\boldsymbol{y} \in \mathbb{R}^n$, and hence $P_{\mathcal{A}}[\boldsymbol{y}] \in \mathcal{A}$.

To prove the property more generally, we use the KKT conditions for the estimator $\widehat{\boldsymbol{\beta}}$:

$$-2X^\top P_{\mathcal{A}}[\boldsymbol{y}] + r\widehat{\boldsymbol{\kappa}} \;=\; \mathbf{0}_p$$

for a $\widehat{\boldsymbol{\kappa}} \in \partial h\big[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\big]$. We rearrange this and apply the dual function of h to find

$$\overline{h}[-X^\top P_{\mathcal{A}}[\boldsymbol{y}]] \;=\; \overline{h}[r\widehat{\boldsymbol{\kappa}}/2]\,.$$

So, we are left with showing that $\overline{h}[r\widehat{\boldsymbol{\kappa}}/2] \leq r/2$. The case $r = 0$ has been discussed above. If $r > 0$, our assumptions on h and 1. in Exercise 6.3 imply that $\overline{h}[r\widehat{\boldsymbol{\kappa}}/2] \leq r/2$ is equivalent to $\overline{h}[\widehat{\boldsymbol{\kappa}}] \leq 1$. To show this latter inequality, we use again that by definition of subdifferentials,

$$h[\boldsymbol{\gamma}] \;\geq\; h\big[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\big] + \langle\widehat{\boldsymbol{\kappa}},\, \boldsymbol{\gamma} - \widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\rangle$$

for all $\boldsymbol{\gamma} \in \mathbb{R}^p$. Rearranging yields

$$\langle\widehat{\boldsymbol{\kappa}},\, \boldsymbol{\gamma}\rangle \;\leq\; h[\boldsymbol{\gamma}] - h\big[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\big] + \langle\widehat{\boldsymbol{\kappa}},\, \widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\rangle\,,$$

and with Hölder's inequality Lemma B.1.3,

$$\langle\widehat{\boldsymbol{\kappa}},\, \boldsymbol{\gamma}\rangle \;\leq\; h[\boldsymbol{\gamma}] - h\big[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\big] + \overline{h}[\widehat{\boldsymbol{\kappa}}]h\big[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\big]\,.$$

In the case $h\big[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\big] = 0$, Assumption 3. in Lemma B.1.3 (definiteness) implies $\widehat{\boldsymbol{\beta}}[\boldsymbol{y}] = \mathbf{0}_p$, and the display yields (recall the convention $0 \cdot \pm\infty = 0$)

$$\langle\widehat{\boldsymbol{\kappa}},\, \boldsymbol{\gamma}\rangle \;\leq\; h[\boldsymbol{\gamma}]\,.$$

Since this is true for any $\boldsymbol{\gamma} \in \mathbb{R}^p$, we can use the definition of duals to deduce

$$\overline{h}[\widehat{\boldsymbol{\kappa}}] \;=\; \sup\big\{\langle\widehat{\boldsymbol{\kappa}},\, \boldsymbol{\gamma}\rangle \,:\, h[\boldsymbol{\gamma}] \leq 1\big\} \;\leq\; \sup\big\{h[\boldsymbol{\gamma}] \,:\, h[\boldsymbol{\gamma}] \leq 1\big\} \;\leq\; 1\,,$$

as desired. We can thus assume $h\big[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\big] \neq 0$ in the following.

Then, we use again the earlier display and find for any $a>0$

$$\begin{aligned}
\overline{h}[\widehat{\boldsymbol{\kappa}}] &= \sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^p\\ h[\boldsymbol{\gamma}]\le 1}} \langle\widehat{\boldsymbol{\kappa}},\,\boldsymbol{\gamma}\rangle && \text{“definition of dual functions”}\\
&= \sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^p\\ h[\boldsymbol{\gamma}/(1+a)h[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]]]\le 1}} \langle\widehat{\boldsymbol{\kappa}},\,\boldsymbol{\gamma}\rangle/\bigl((1+a)h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]\bigr) && \text{“}h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]\neq 0\text{ and }a>0\text{”}\\
&\le \sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^p\\ h[\boldsymbol{\gamma}/(1+a)h[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]]]\le 1}} \bigl(h[\gamma]-h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]+\overline{h}[\widehat{\boldsymbol{\kappa}}]h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]\bigr)/\bigl((1+a)h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]\bigr)\\
&&& \text{“earlier display”}\\
&= \sup_{\substack{\boldsymbol{\gamma}\in\mathbb{R}^p\\ h[\boldsymbol{\gamma}]\le(1+a)h[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]]}} \bigl(h[\gamma]-h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]+\overline{h}[\widehat{\boldsymbol{\kappa}}]h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]\bigr)/\bigl((1+a)h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]\bigr)\\
&&& \text{“}h\text{ assumed to be positive homogenous”}\\
&\le \bigl((1+a)h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]-h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]+\overline{h}[\widehat{\boldsymbol{\kappa}}]h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]\bigr)/\bigl((1+a)h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]\bigr)\\
&&& \text{“bounding the supremum by using }h[\boldsymbol{\gamma}]\le(1+a)h[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]]\text{”}\\
&= (a+\overline{h}[\widehat{\boldsymbol{\kappa}}])/(1+a)\,.\\
&&& \text{“dividing numerator and denominator by }h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]\neq 0\text{ and simplifying”}
\end{aligned}$$

Solving for $\overline{h}[\widehat{\boldsymbol{\kappa}}]$ then gives

$$\bigl(1-1/(1+a)\bigr)\,\overline{h}[\widehat{\boldsymbol{\kappa}}]\;\le\;a/(1+a)\,,$$

and consequently, since $1-1/(a+1)>0$ for $a>0$,

$$\overline{h}[\widehat{\boldsymbol{\kappa}}]\;\le\;\frac{a/(1+a)}{1-1/(1+a)}\;=\;\frac{a}{1+a-1}\;=1\,.$$

Thus, $P_{\mathcal{A}}[\boldsymbol{y}]\in\mathcal{A}$, as desired.

Step 2: We now show that

$$\langle\boldsymbol{a}-P_{\mathcal{A}}[\boldsymbol{y}],\,P_{\mathcal{A}}[\boldsymbol{y}]-\boldsymbol{y}\rangle\;\ge\;0\qquad(\boldsymbol{y}\in\mathbb{R}^n,\boldsymbol{a}\in\mathcal{A})\,.$$

For this, we only need that h is amenable to Hölder's inequality.

To get some insights into this inequality, assume that $X^\top X$ is invertible and consider the least-squares estimator $\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$ discussed at the beginning of this section. We then find for any $\boldsymbol{a}\in\mathbb{R}^n$ with $X^\top\boldsymbol{a}=\boldsymbol{0}_p$

$$\begin{aligned}
&\langle\boldsymbol{a}-P_{\mathcal{A}}[\boldsymbol{y}],\,P_{\mathcal{A}}[\boldsymbol{y}]-\boldsymbol{y}\rangle\\
&=\langle\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}[\boldsymbol{y}]-\boldsymbol{a},\,X\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}[\boldsymbol{y}]\rangle && \text{“specifications of }P_{\mathcal{A}}[\boldsymbol{y}]\text{ and }\widehat{\boldsymbol{\beta}}\text{”}\\
&=\langle\boldsymbol{y}-X(X^\top X)^{-1}X^\top\boldsymbol{y}-\boldsymbol{a},\,X(X^\top X)^{-1}X^\top\boldsymbol{y}\rangle && \text{“definition of }\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\text{”}\\
&=\langle\boldsymbol{y}-X(X^\top X)^{-1}X^\top\boldsymbol{y}-\boldsymbol{a},\,X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top\boldsymbol{y}\rangle\\
&&& \text{“including a one-valued factor in the first argument of the inner product”}\\
&=\langle X(X^\top X)^{-1}X^\top\boldsymbol{y}-X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top\boldsymbol{y}-X(X^\top X)^{-1}X^\top\boldsymbol{a},\\
&\qquad\qquad X(X^\top X)^{-1}X^\top\boldsymbol{y}\rangle && \text{“properties of inner products”}\\
&=\langle\boldsymbol{0}_n,\,X(X^\top X)^{-1}X^\top\boldsymbol{y}\rangle && \text{“evaluating the first argument of inner product”}\\
&=0\,. && \text{“linearity of the inner product”}
\end{aligned}$$

We can think of the least-squares estimator as a egularized least-squares estimator with $r=0$ and h the ℓ_2-norm, say. One can then check that $\mathcal{A}=\{\boldsymbol{a}\in\mathbb{R}^n:X^\top\boldsymbol{a}=\mathbf{0}_p\}$, so that the above display entails

$$\langle\boldsymbol{a}-P_{\mathcal{A}}[\boldsymbol{y}],\,P_{\mathcal{A}}[\boldsymbol{y}]-\boldsymbol{y}\rangle\;=\;0\qquad(\boldsymbol{y}\in\mathbb{R}^n,\boldsymbol{a}\in\mathcal{A})\,.$$

Thus, we can interpret the desired inequality as a relaxed version of this inner product equality.

We start the proof of the inequality by writing out again the KKT conditions for the estimator $\widehat{\boldsymbol{\beta}}$:

$$-2X^\top P_{\mathcal{A}}[\boldsymbol{y}]+r\widehat{\boldsymbol{\kappa}}\;=\;\mathbf{0}_p$$

for a $\widehat{\boldsymbol{\kappa}}\in\partial h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]$. Of course, also $\widehat{\boldsymbol{\kappa}}$ depends on $\boldsymbol{y}$—but there should not be confusion about this in the following, so that we supress this dependence in the notation.

We multiply the display from the left with $\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]$ and rearrange the terms to find

$$\langle X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}],\,P_{\mathcal{A}}[\boldsymbol{y}]\rangle-\frac{r}{2}\langle\widehat{\boldsymbol{\beta}}[\boldsymbol{y}],\,\widehat{\boldsymbol{\kappa}}\rangle\;=\;0\,.$$

Now, observe that by definition of subdifferentials,

$$h[\boldsymbol{\gamma}]\;\geq\;h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]+\langle\widehat{\boldsymbol{\kappa}},\,\boldsymbol{\gamma}-\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\rangle$$

for all $\boldsymbol{\gamma}\in\mathbb{R}^p$. Rearranging yields

$$h[\boldsymbol{\gamma}]+\langle\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]-\boldsymbol{\gamma},\,\widehat{\boldsymbol{\kappa}}\rangle\;\geq\;h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]\,,$$

and in particular for $\boldsymbol{\gamma}=\mathbf{0}_p$, invoking 3. (definiteness of h) in the assumptions of Hölder's inequality in Lemma B.1.3,

$$\langle\widehat{\boldsymbol{\beta}}[\boldsymbol{y}],\,\widehat{\boldsymbol{\kappa}}\rangle\;\geq\;h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]\,.$$

We plug this back into the above display to deduce

$$\langle X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}],\,P_{\mathcal{A}}[\boldsymbol{y}]\rangle-\frac{r}{2}h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]\;\geq\;0\,.$$

Hölder's inequality allows us to bound the right-hand side according to

$$\begin{aligned}\frac{r}{2}h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr]\;&\geq\;\max_{\boldsymbol{a}\in\mathcal{A}}\overline{h}[X^\top\boldsymbol{a}]h\bigl[\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\bigr] && \text{“definition of }\mathcal{A}\text{”}\\ &\geq\;\max_{\boldsymbol{a}\in\mathcal{A}}\langle X^\top\boldsymbol{a},\,\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\rangle && \text{“Hölder's Inequality Lemma B.1.3”}\\ &=\;\max_{\boldsymbol{a}\in\mathcal{A}}\langle X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}],\,\boldsymbol{a}\rangle\,. && \text{“properties of inner products”}\end{aligned}$$

Combining the two forgoing displays yields for any vector $\boldsymbol{a}\in\mathcal{A}$,

$$\langle X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}],\,P_{\mathcal{A}}[\boldsymbol{y}]\rangle-\langle X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}],\,\boldsymbol{a}\rangle\;\geq\;0\,,$$

and thus (recall that $P_{\mathcal{A}}[\boldsymbol{y}]=\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]$),

$$\langle\boldsymbol{a}-P_{\mathcal{A}}[\boldsymbol{y}],\,P_{\mathcal{A}}[\boldsymbol{y}]-\boldsymbol{y}\rangle\;\geq\;0\,,$$

as desired.

Step 3: We show next that for any $\boldsymbol{y}\in\mathbb{R}^n$, it holds that $P_{\mathcal{A}}[\boldsymbol{y}]$ is a projection of $\boldsymbol{y}$ on the set $\mathcal{A}$.

For this, we first observe that $\mathcal{A} = \{\boldsymbol{a} \in \mathbb{R}^n : \overline{h}[X^\top \boldsymbol{a}] \le r/2\}$ is convex (since h is convex according to 5. in Exercise 6.3). Moreover, $P_{\mathcal{A}}[\boldsymbol{y}] \in \mathcal{A}$ by Step 1 and $\langle \boldsymbol{a} - P_{\mathcal{A}}[\boldsymbol{y}], P_{\mathcal{A}}[\boldsymbol{y}] - \boldsymbol{y}\rangle$ by Step 2 for all $\boldsymbol{a} \in \mathcal{A}, \boldsymbol{y} \in \mathbb{R}^n$. Hence, we can apply the second part of Lemma B.1.6 to deduce that $P_{\mathcal{A}}[\boldsymbol{y}]$ is an ℓ_2-projection of $\boldsymbol{y}$ on $\mathcal{A}$. This concludes the second step.

Step 4: Finally, we show that

$$\|X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}] - X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}']\|_2^2 \le \|\boldsymbol{y} - \boldsymbol{y}'\|_2^2 - 2\|P_{\mathcal{A}}[\boldsymbol{y}] - P_{\mathcal{A}}[\boldsymbol{y}']\|_2^2 \,,$$

which is a sharper version of the desired property.

With the preceeding steps and the third part of Lemma B.1.6, we find for any $\boldsymbol{y}, \boldsymbol{y}'$,

$$\begin{aligned}
&\|X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}] - X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}']\|_2^2\\
&= \|X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}] + P_{\mathcal{A}}[\boldsymbol{y}] - X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}'] - P_{\mathcal{A}}[\boldsymbol{y}'] - P_{\mathcal{A}}[\boldsymbol{y}] + P_{\mathcal{A}}[\boldsymbol{y}']\|_2^2 && \text{“adding zero-valued terms”}\\
&= \|\boldsymbol{y} - \boldsymbol{y}' - P_{\mathcal{A}}[\boldsymbol{y}] + P_{\mathcal{A}}[\boldsymbol{y}']\|_2^2 && \text{“}P_{\mathcal{A}}[\boldsymbol{y}] = \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\text{”}\\
&= \|\boldsymbol{y} - \boldsymbol{y}'\|_2^2 + \|P_{\mathcal{A}}[\boldsymbol{y}] - P_{\mathcal{A}}[\boldsymbol{y}']\|_2^2 - 2\langle P_{\mathcal{A}}[\boldsymbol{y}] - P_{\mathcal{A}}[\boldsymbol{y}'], \boldsymbol{y} - \boldsymbol{y}'\rangle && \text{“expanding”}\\
&\le \|\boldsymbol{y} - \boldsymbol{y}'\|_2^2 - 2\|P_{\mathcal{A}}[\boldsymbol{y}] - P_{\mathcal{A}}[\boldsymbol{y}']\|_2^2 \,. && \text{“previous steps and third part of Lemma B.1.6”}
\end{aligned}$$

This concludes the proof of Step 4 and thus of the lemma. □

As a consequence, many functionals of regularized least-squares are also Lipschitz. The following one is the most important one for us.

Lemma 6.5.3 (Lipschitz Prediction Error)

Consider a linear regression model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u}$ and an estimator $\widehat{\boldsymbol{\beta}}$ for which the prediction function $\boldsymbol{y} \mapsto X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]$ is 1-Lipschitz. Then, the prediction error

$$\boldsymbol{u} \mapsto \|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}]\|_2$$

is also 1-Lipschitz.

Proof. The proof is based only on the triangle inequality of the ℓ_2-norm.

Note first that one can assume without loss of generality $\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}]\|_2 \ge \|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}']\|_2$ (swap $\boldsymbol{u}$ and $\boldsymbol{u}'$ otherwise). Then,

$$\begin{aligned}
&\big|\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}]\|_2 - \|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}']\|_2\big|\\
&= \|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}]\|_2 - \|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}']\|_2 && \text{“without loss of generality”}\\
&= \|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}'] + X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}'] - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}]\|_2\\
&\qquad\qquad - \|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}']\|_2 && \text{“adding a zero-valued term in the first norm”}\\
&\le \|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}']\|_2 + \|X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}] - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}']\|_2\\
&\qquad\qquad - \|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}']\|_2 && \text{“triangle inequality”}\\
&= \|X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}] - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}']\|_2 && \text{“consolidating”}\\
&\le \|(X\boldsymbol{\beta} + \boldsymbol{u}) - (X\boldsymbol{\beta} + \boldsymbol{u}')\|_2 && \text{“}\boldsymbol{y} \mapsto X\widehat{\boldsymbol{\beta}}[\boldsymbol{y}]\text{ 1-Lipschitz by assumption”}\\
&= \|\boldsymbol{u} - \boldsymbol{u}'\|_2 \,, && \text{“consolidating”}
\end{aligned}$$

as desired. □

Resulting Bounds

We have learned that 1. Lipschitz functions of Gaussian random variables are sharply concentrated around their medians (Lemma 6.5.1) and 2. that the prediction losses of regularized least-squares estimators are such functions (Lemma 6.5.3). We conclude immediately that in Gaussian linear regression, the prediction losses of regularized least-squares estimators are concentrated around their medians. Relating the medians to the results derived earlier in this chapter, we find that also the *expected* prediction losses are controlled by those previous bounds.

The main result of this section is as follows.

Theorem 6.5.1 (Risk Bounds for Regularized Least-squares)

Consider regularized least-squares estimation in a linear regression model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u}$ with fixed design matrix $X \in \mathbb{R}^{n\times p}$ and normal noise vector $\boldsymbol{u} \sim \mathcal{N}_n[\mathbf{0}_n, \sigma^2\,\mathrm{I}_{n\times n}/n[$. Assume that $\mathbb{P}\{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 > b^*\} \leq 1/2$ for a non-negative constant $b^* \geq 0$. Then,

$$\mathbb{E}\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \leq b^* + \frac{\sigma^2}{\sqrt{2n}}\,.$$

The theorem ensures that any regularized least-squares estimator in linear regression with Gaussian noise that satisfies a probability bound with probability at least 1/2 also satisfies a corresponding risk bound. The additional term $\sigma^2/(\sqrt{2}n)$ is up to constants the risk of the standard least-squares estimator for fixed p, see Equation 1.4 on Page 12.

Example 6.5.2 (Power-one Risk Bound for the Lasso)

We have derived on Page 93 that the lasso $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}[r]$ with a tuning parameter that satisfies $r \geq 2\|X^\top \boldsymbol{u}\|_\infty$ obeys the *probability bound*

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \leq \min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \bigl\{\|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + 2r\|\boldsymbol{\alpha}\|_1\bigr\}\,.$$

Theorem 6.5.1 then implies (set $b^* := \min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\bigl\{\|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + 2r\|\boldsymbol{\alpha}\|_1\bigr\}$) that the lasso with a tuning parameter that satisfies $\mathbb{P}\{r \geq 2\|X^\top\boldsymbol{u}\|_\infty\} \geq 1/2$ obeys the *risk bound*

$$\mathbb{E}\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \leq \min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \bigl\{\|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + 2r\|\boldsymbol{\alpha}\|_1\bigr\} + \frac{\sigma^2}{\sqrt{2n}}\,.$$

Probability bounds are “local” in the sense that it bounds the loss for each data point $(\boldsymbol{y}, X)$ individually; the risk bound are “global” in the sense that they bound the loss in expectation. Accordingly, the conditions on the tuning parameter for the probability bounds involve a fixed data-point, while the corresponding condition for the risk bounds involve a probability over all data points. Risk bounds do not superseed probability bounds, nor the other way around: the different types of results simply answer different questions.
If $\max_{j\in\{1,\dots,p\}}(X^\top X)_{jj} \leq 1$, Lemma 4.2.1 identifies $r = \sigma\sqrt{8\log[4p]/n}$ as a suitable tuning parameter, and we find, for example,

$$\mathbb{E}\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \leq \sigma\sqrt{\frac{8\log[4p]}{n}}\|\boldsymbol{\beta}\|_1 + \frac{\sigma^2}{\sqrt{2n}}\,.$$

Proof of Lemma 6.5.1. We proceed in three steps: In the first step, we show that the condition on the tuning parameter can be slightly tightened without loss of generality. In the second step, we then apply Lemma 6.5.1 on the Gaussian concentration around medians to derive a deviation inequality for the loss. In the third step, we finally integrate this inequality to obtain the desired risk bound.

Step 1: We first show that in the rest of the proof, we can replace $\mathbb{P}\{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 > b^*\} \leq 1/2$ by the slightly stronger condition $\mathbb{P}\{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \geq b^*\} \leq 1/2$.

To this end, assume that the theorem is proved under the stronger condition: if $\mathbb{P}\{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \geq b^*\} \leq 1/2$ for a non-negative constant $b^* \geq 0$, then

$$\mathbb{E}\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; b^* + \frac{\sigma^2}{\sqrt{2n}}\,.$$

We will do this in Steps 2 and 3. However, consider now an estimator $\widehat{\boldsymbol{\beta}}$ that satisfies the weaker condition $\mathbb{P}\{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 > b^*\} \leq 1/2$ for a non-negative constant $b^* \geq 0$. Then, $\mathbb{P}\{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \geq b^* + a\} \leq 1/2$ for any $a > 0$. The theorem proved with the stronger condition applied to the (again non-negative) constant $b^* + a \geq 0$ yields

$$\mathbb{E}\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; b^* + a + \frac{\sigma^2}{\sqrt{2n}}\,.$$

Taking the limit $a \to 0$ on the right-hand side then shows that the theorem also holds under the weaker condition.

Step 2: We now show that

$$\mathbb{P}\Big\{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2 \geq \sqrt{b^*} + \sqrt{\sigma^2/n}\,t\Big\} \;\leq\; 1 - \breve{g}[t] \qquad (t \geq 0)\,.$$

We start by defining $\boldsymbol{v} := \sqrt{n/\sigma^2}\boldsymbol{u}$. This random vector is standard Gaussian. Moreover, since X is fixed, and since $\boldsymbol{u} \mapsto \|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \boldsymbol{u}]\|_2$ is 1-Lipschitz by Lemma 6.5.3, the function $f : \mathbb{R}^n \to \mathbb{R}$, $f[\mathbf{a}] := \sqrt{n/\sigma^2}\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}[X\boldsymbol{\beta} + \sqrt{\sigma^2/n}\mathbf{a}]\|_2$, is deterministic and 1-Lipschitz. We can, therefore, aim at the application of Lemma 6.5.1 with these $\boldsymbol{v}$ and f.

To this end, we need to relate $\sqrt{b^*}$ with medians m of $f[\boldsymbol{v}]$. We approach this by considering first small medians and then large medians. Since a median always exists according to Exercise 6.11, one of these cases must apply.

Case 1: Consider medians $\widetilde{m}$ of $f[\boldsymbol{v}]$ that satisfy

$$\widetilde{m} \;\leq\; \sqrt{nb^*/\sigma^2}\,.$$

In this case, the claimed inequality follows directly from Lemma 6.5.1 with $m = \widetilde{m}$.

Case 2: Consider now medians $\widetilde{m}$ of $f[\boldsymbol{v}]$ that satisfy

$$\widetilde{m} \;\geq\; \sqrt{nb^*/\sigma^2}\,.$$

In this case, we first observe that

$$\begin{aligned}
&\mathbb{P}\Big\{f[\boldsymbol{v}] \leq \sqrt{nb^*/\sigma^2}\Big\}\\
&\leq \mathbb{P}\Big\{f[\boldsymbol{v}] \leq \widetilde{m}\Big\} && \text{“assumption on } \widetilde{m}\text{; } \mathbb{P}\mathcal{A} \leq \mathbb{P}\mathcal{B} \text{ for } \mathcal{A} \subset \mathcal{B}\text{”}\\
&\leq 1/2\,. && \text{“by Definition 6.5.2 of medians with } m = \widetilde{m}\text{”}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathbb{P}\Bigl\{f[\boldsymbol{v}] \geq \sqrt{nb^*/\sigma^2}\Bigr\}\\
&= \mathbb{P}\Bigl\{\sqrt{n/\sigma^2}\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2 \geq \sqrt{nb^*/\sigma^2}\Bigr\} && \text{“definition of } f\text{”}\\
&\leq \mathbb{P}\bigl\{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \geq b^*\bigr\}\\
&\qquad \text{“multiplying both sides by } \sqrt{\sigma^2/n} \text{ and then squaring them; } \mathbb{P}\mathcal{A} \leq \mathbb{P}\mathcal{B} \text{ for } \mathcal{A} \subset \mathcal{B}\text{”}\\
&\leq 1 - 1/2 \;\;=\;\; 1/2\,. && \text{“by the assumption on } \widehat{\boldsymbol{\beta}} \text{ mentioned in Step 1”}
\end{aligned}$$

Hence, also $\sqrt{nb^*/\sigma^2}$ is a median of $f[\boldsymbol{v}]$ by Definition 6.5.2. The desired inequality then follows directly from Lemma 6.5.1 with $m = \sqrt{nb^*/\sigma^2}$.

Step 3: We now integrate the result of the second step to derive the desired risk bound:

$$\begin{aligned}
&\mathbb{E}\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 - b^*\\
&= \mathbb{E}\bigl[\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 - b^*\bigr] && \text{“linearity of integrals”}\\
&\leq \int_0^\infty \mathbb{P}\bigl\{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 - b^* \geq s\bigr\}\,ds && \text{“Exercise 6.9”}\\
&= \int_0^\infty \mathbb{P}\Bigl\{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2 \leq \sqrt{b^* + s}\Bigr\}\,ds\\
&\qquad\qquad \text{“rearranging and taking square-roots inside the probability”}\\
&\leq \int_0^\infty \mathbb{P}\Bigl\{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2 \leq \sqrt{b^*} + \sqrt{s}\Bigr\}\,ds && \text{“}\sqrt{a+b} \leq \sqrt{a} + \sqrt{b} \text{ for } a, b \geq 0\text{”}\\
&\leq \int_0^\infty 1 - \breve{g}\bigl[\sqrt{ns/\sigma^2}\bigr]\,ds && \text{“Step 2 with } \sqrt{s} = \sqrt{\sigma^2/n}\,t \Rightarrow t = \sqrt{ns/\sigma^2}\text{”}\\
&= \frac{\sigma^2}{n}\int_0^\infty 1 - \breve{g}[u]\,du && \text{“}u = \sigma^2 s/n;\ ds/du = \sigma^2 n/u\text{”}\\
&\leq \frac{\sigma^2}{n}\int_0^\infty \frac{e^{-u^2/2}}{\sqrt{\pi}}\,du && \text{“Lemma B.3.3”}\\
&= \frac{\sigma^2}{\sqrt{2}n}\int_{-\infty}^\infty \frac{e^{-u^2/2}}{\sqrt{2\pi}}\,du && \text{“symmetry and positivity of integrand”}\\
&= \frac{\sigma^2}{\sqrt{2}n}\,, && \text{“integration of Gaussian density”}
\end{aligned}$$

which yields the desired bound after rearranging the terms. □

6.6 Oracle Inequalities*

A main focus of this book is to establish theoretical guarantees for regularized estimators. A fundamental concept for formulating such guarantees are *oracle inequalities.* An oracle inequality ensures that an estimator does nearly as well as an optimal parameter-valued function. This function may depend on the data and, opposed to the estimator itself, on model properties that are unknown in practice. In this sense, we compare the estimator with an *oracle* that already knows the solution of the statistical problem. In the following, we establish a general definition of oracle inequalities that comprises the notions of classical risk minimization and model selection as special cases.

oracle inequality

The first, and arguably most important ingredient of an oracle inequality is the *loss function* $\ell : \mathcal{B} \times \mathcal{Z} \to [0, \infty]$, which contrasts any pairs of parameter and data. The loss function formalizes our notions of how well a parameter $\boldsymbol{\alpha} \in \mathcal{B}$ fits the data $Z \in \mathcal{Z}$ or a fixed target $\boldsymbol{\beta} \in \mathcal{B}$. It is important to understand that the very same estimator can minimize one loss and yet perform very poorly in another loss; there might not even exist an estimator that performs well simulateneously two given losses.

loss function

Typical examples of loss functions concern *prediction*, *estimation*, and *support recovery*. For illustration, consider a linear regression model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u}$ of the data $Z = (\boldsymbol{y}, X)$, where $\boldsymbol{y} \in \mathbb{R}^n$ is the outcome, $X \in \mathbb{R}^{n \times p}$ the fixed or random design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ the target vector, and $\boldsymbol{u} \in \mathbb{R}^n$ random noise. Then, (find the mathematical notations such as $\|\cdot\|_2$, $\operatorname{supp}[\cdot]$, etc. defined on Page 3) $\ell[\boldsymbol{\alpha}, Z] = \|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2$ measures fit in terms of prediction, $\ell[\boldsymbol{\alpha}, Z] = \|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2$ measures fit in terms of estimation, and $\ell[\boldsymbol{\alpha}, Z] = \#\{j \in \{1, \dots, p\} : \boldsymbol{\beta}_j = 0, \widehat{\boldsymbol{\beta}}_j \neq 0 \text{ or } \boldsymbol{\beta}_j \neq 0, \widehat{\boldsymbol{\beta}}_j = 0\}$ measures fit in terms of support recovery or model selection. Such losses lead to what we call *probability bounds*.

Another common class of loss functions is of the form $\ell[\boldsymbol{\alpha}, Z] = \mathbb{E}[\tilde{\ell}[\boldsymbol{\alpha}, Z]]$, where the expectation is taken over (aspects of) the data space $\mathcal{Z}$, and $\tilde{\ell} : \mathcal{B} \times \mathcal{Z} \to [0, \infty]$ is a loss function itself. Such losses lead to what we call *risk bounds*.

If the data Z is clear from the context, or if ℓ does not depend on the data altogether, we drop the second argument of ℓ, replacing $\ell[\boldsymbol{\alpha}, Z]$ by $\ell[\boldsymbol{\alpha}]$.

The second ingredient is the *complexity function* $m : \mathcal{B} \to [0, \infty]$, which assigns a number to any parameter. We first note that the role of the target $\boldsymbol{\beta}$ is to designate a model that captures the data generating process accurately. The loss is typically minimized $\ell[\boldsymbol{\beta}]$ to signify $\boldsymbol{\beta}$ as the gold standard in terms of accuracy. However, the data generating process can require that such a perfectly accurate model is complex. On the other hand, the assumption underpinning regularized methods is that there is a simple model *close* (as measured by a loss as discussed in the preceding paragraphs) to the target model. The complexity measure formalizes our notions of how simple or complex a given model is. In this sense, the specific forms of both the function m and the prior function in the estimator root in our assumptions (or knowledge if we are lucky) about the model. The key difference between the two functions is that m can have any desired shape, while the prior function is part of the estimation process and, therefore, is subject to computational constraints. However, in any case, neither of this functions need to be small at the target—they should be small at the simple surrogate of the target instead.

complexity function

Example 6.6.1 (Sparsity)

In regression-type settings, a typical loss function is $\ell : \boldsymbol{\alpha} \mapsto \|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2$ and a typical complexity function is $m : \boldsymbol{\alpha} \mapsto \#\{j : \boldsymbol{\alpha}_j \neq 0\}$. These functions reflect a desire for accurate and interpretable models, that is, modesl that both fit the data well (as measured by ℓ) and are *sparse* (as measured by m). The lasso

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^p} \Bigl\{ \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_1 \Bigr\}$$

is a standard approach to constructing such models. It estimates the original loss function ℓ by $\boldsymbol{\alpha} \mapsto \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2$, which no longer depends on the unknown target $\boldsymbol{\beta}$, and mimiks the computationally challenging complexity function m by $\boldsymbol{\alpha} \mapsto \|\boldsymbol{\alpha}\|_1$, which makes the objective function amenable to convex optimization.

The complexity measure and the prior function usually also differ, even if they have the same functional form, in terms of data-dependent factors. These factors, which we call *tuning parameters*, need to be calibrated carefully to ensure satisfactory performance in theory and practice.

The third and last ingredient is the *oracle* $\boldsymbol{\alpha}^{\bullet} : \mathcal{Z} \to \mathcal{B}$, which maps data on the parameter space. The oracle is defined as the function that optimally combines a good fit (as measured by ℓ) and low complexity (as measured by m) in that it minimizes the sum over loss function and complexity function. Therefore, the oracle is closely linked to the underpinning notion of the existence of a small model with a good fit. oracle

In probability bounds, we can treat $\boldsymbol{\alpha}^{\bullet}$ as a parameter—neglecting the functional aspect—as those bounds provide guarantees for fixed data points. In risk bounds, on the other hand, it can be necessary to work with $\boldsymbol{\alpha}^{\bullet}$ as a function indeed, as those bounds are formulated by means of expectations over a range of data points.

The oracle can depend on any model parameter and is, therefore, unknown in practice, which it has in common with the target. We can think of the oracle as a refinement of the target more generally: both the oracle and the target are supposed to fit the data well, but the oracle is additionally supposed to have low model complexity while the target is not.

Now, having all ingredients introduced, we can define oracle inequalities. An oracle inequality states that an estimator $\widehat{\boldsymbol{\beta}}$ satisfies a bound for all $Z \in \mathcal{E}$ in a set $\mathcal{E} \subset \mathcal{Z}$:

$$\ell[\widehat{\boldsymbol{\beta}}, Z] \;\leq\; c\bigl\{\ell[\boldsymbol{\alpha}^{\bullet}, Z] + m[\boldsymbol{\alpha}^{\bullet}]\bigr\} \;=\; c\min_{\boldsymbol{\alpha}\in\mathcal{H}}\bigl\{\ell[\boldsymbol{\alpha}, Z] + m[\boldsymbol{\alpha}]\bigr\}$$

for a loss function ℓ, a complexity function m, a set of parameter-valued functions $\mathcal{H}$ on the data, and a constant $c \geq 1$. We have implicitly assumed that the minimum exists. An oracle inequality is called *sharp* if $c = 1$.The term $\ell[\boldsymbol{\alpha}^{\bullet}]$ is called the *approximation error*, reflecting our objective to approximate the target; the term $m[\boldsymbol{\alpha}^{\bullet}]$ is called *complexity error*, indicating our objective to obtain simple models. The set $\mathcal{H}$ equips the oracle with leverage in minimizing the loss and complexity functions: usually the larger the set, the smaller the bound. Finally, the set $\mathcal{E}$ provides flexibility to exclude adverse data settings from consideration (in strong contrast, minimax concepts concern adverse *parameter* settings).

A probability bound holds pointwise on the event $\mathcal{E}$. The naming of this type of bound indicates that the event $\mathcal{E}$ can be associated with some—hopefully large—probability. A risk bound, on the other hand, holds in expectation. The naming refers to the standard usage of “risk” for an expected loss.

Example 6.6.2 (Risk Bounds for Standard and Regularized Least-squares)

Consider a linear regression model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u}$ with a fixed design matrix $X \in \mathbb{R}^{n\times p}$ and Gaussian noise $\boldsymbol{u} \sim \mathcal{N}_n[\mathbf{0}_n, \sigma^2\,\mathrm{I}_{n\times n}]$. For the standard least-squares estimator

$$\widehat{\boldsymbol{\beta}}_{\mathrm{ls}} \;\in\; \mathop{\mathrm{argmin}}_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2\,,$$

Exercise 1.2 (see also Equation 1.4 on Page 12) directly implies an oracle inequality (which is actually an equality):

$$\underbrace{\mathbb{E}\left[\frac{\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2}{n}\right]}_{\ell[\widehat{\boldsymbol{\beta}}]} \leq \underbrace{\mathbb{E}\left[\frac{\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}^\bullet\|_2^2}{n}\right]}_{\ell[\boldsymbol{\alpha}^\bullet]} + \underbrace{\frac{\sigma^2\operatorname{rank}[X]}{n}}_{m[\boldsymbol{\alpha}^\bullet]}\,.$$

This is a risk bound for prediction (since $\ell\,:\,\boldsymbol{\alpha}\mapsto\mathbb{E}\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2$) with oracle $\boldsymbol{\alpha}^\bullet\,:\,(\boldsymbol{y},X)\mapsto\boldsymbol{\beta}$. The oracle inequality is sharp (the leading constant is equal to 1) and valid on the entire data space ($\mathcal{E}=\mathcal{Z}=\mathbb{R}^n\times\mathbb{R}^{n\times p}$). The complexity function $m\,:\,\boldsymbol{\alpha}\mapsto\sigma^2\operatorname{rank}[X]/n$ is constant in $\boldsymbol{\alpha}$, which reflects that least-squares estimation does not invoke any prior information about the parameter space.
To exploit prior information, the estimator's objective function is complemented with a regularizer:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{rLS}} \;\in\; \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\bigl\{\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2+rh[\boldsymbol{\alpha}]\bigr\}\,,$$

where $r>0$ is a tuning parameter and $h\,:\,\mathbb{R}^p\to[0,\infty]$ formulates the prior information. Under the assumptions of Theorem 6.5.1 on Page 118 with $b^*=\min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\{\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+arh[\boldsymbol{\alpha}]\}$ for an $a\in(0,\infty)$ (cf. Theorem 6.3.1), we then find a pendant to the above oracle inequality:

$$\underbrace{\mathbb{E}\left[\frac{\|X\boldsymbol{\beta}-X\widehat{\boldsymbol{\beta}}_{\mathrm{rLS}}\|_2^2}{n}\right]}_{\ell[\widehat{\boldsymbol{\beta}}]} \leq \underbrace{\mathbb{E}\left[\frac{\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}^\bullet\|_2^2}{n}\right]}_{\ell[\boldsymbol{\alpha}^\bullet]} + \underbrace{\frac{arh[\boldsymbol{\alpha}^\bullet]}{n}+\frac{\sigma^2}{\sqrt{2}n}}_{m[\boldsymbol{\alpha}^\bullet]}\,.$$

The oracle is now $\boldsymbol{\alpha}^\bullet\,:\,(\boldsymbol{y},X)\mapsto\operatorname{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\{\|X\boldsymbol{\beta}-X\boldsymbol{\alpha}\|_2^2+arh[\boldsymbol{\alpha}]\}$ and, hence, balances prediction accuracy with the magnitude of the parameter in terms of the regularizer h. The condition for this second oracle inequality outperforming the first one is $arh[\boldsymbol{\alpha}^\bullet]+\sigma^2/\sqrt{2}\leq\sigma^2\operatorname{rank}[X]$, that is, the underlying motivation for regularization, namely that $h[\boldsymbol{\alpha}^\bullet]$ is small, is satisfied.

Oracle inequalities compare estimators with their optimal counterparts, which means that the bounds naturally scale with the difficulty of the statistical problem at hand. Indeed, the bounds are small if there are parameters that have both a good fit and small complexity, which marks a simple setting; analogously, the bounds are large if there is no parameter that has both a good fit and small complexity, which marks a difficult setting. This smoothly integrates the level of adverseness of the data generating process without limiting the scope of the bounds.

The bounds can involve model parameters such as $\boldsymbol{\beta}$ that are unknown in practice. This highlights again the spirit of oracle inequalities: oracle inequalities do not aim at computable guarantees but rather at an understanding of how model parameters influence the performance of estimators.

The bounds provided by oracle inequalities are *finite sample* gurantees. Theories in traditional fields of statistics are often formulated in terms of asymptotics in the sample size: $n\to\infty$. In high-dimensional statistics, however, sample sizes are not large enough—in view of the relatively large complexities of the parameter spaces—to render such asymptotic results sufficiently accurate.

6.7 References and Further Reading

The assumption that there is a simple model close to the target $\boldsymbol{\beta}$ can be understood in terms of projections: For example, consider $\mathcal{B} = \mathbb{R}^p$, $\ell[\boldsymbol{\alpha}] := \|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2$, and a subset of models $\mathcal{B}_s \subset \mathcal{B}$ that are simple in regards to a given complexity function. Then, the assumption means that $\min\{\|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2 : \boldsymbol{\alpha} \in \mathcal{B}_s\}$ is small, that is the ℓ_2-projection of the target onto the set $\mathcal{B}_s$ is close to the target itself. Such thoughts have been formulated in [vL13, Section 3.2], for example.

A book specialized on oracle inequalities for high-dimensional estimators is [Kol11].

The square-root lasso has been introduced and studied in [BCW11]; a version with grouped variables and further theoretical guarantees and algorithms have been established in [BLS14a]. The closely related scaled lasso has been introduced and studied in [SZ12]. The analysis of the square-root lasso in the general form 6.2, in particular the proof of Lemma 6.2.2, stems from [LYG18].

Signed-constrained regression in high-dimensions has been discussed in [Mei13] and in references therein.

Decomposability has been discussed in [NYWR12, Section 2.2] and [Wai14, Section 3.2]. Our Assumption 2. on Page 107 is a special case of [NYWR12, Definition 1] and [Wai14, Equation (22)] for $\mathcal{M} := \{\boldsymbol{\gamma} \in \mathbb{R}^p : \operatorname{supp}(\boldsymbol{\gamma}) \subset \mathcal{S}\}$.

Connections of solving $\boldsymbol{y} = X\boldsymbol{\beta}$ for sparse $\boldsymbol{\beta}$ to ℓ_1-minimization problems are discussed in [CT$^+$07, Section 1.1] and [Che95, CDS01, DET06]. Our compatibility condition is (a slightly modified and extended version of) the one introduced in [van07, Section 2.1]. The closely related *restricted eigenvalue condition* was introduced in [BRT09]. Comparisons among different conditions for the lasso can be found in [Kol09, vB09]. A geometric interpretation of compatibility constants and relationships to entropy can be found in [vL13]. A generalization of the restricted eigenvalue beyond regression-type data is *restricted strong convexity*, introduced and discussed in [NYWR12, Section 2.4] and [Wai14, Section 3.1]. Computations of compatibility constants are discussed in [DHL17, Appendix on Pages 578ff].

Derivations of risk bounds for lasso-type estimators based on Lipschitz properties of the prediction function have been established first in [BT16].

Properties of Gaussian distributions, including isoperometric inequalities, can be found in the text book [Lif12].

6.8 Exercises

Exercises for Section 6.6

□ **Exercise 6.1** $^{\diamond\diamond}$ In this exercise, we review properties of ℓ_q-norms. For this, recall the definition of the ℓ_q-“norms:”

$$\|\boldsymbol{\alpha}\|_q := \Bigl(\sum_{j=1}^p |\alpha_j|^q\Bigr)^{1/q} \qquad (\,\boldsymbol{\alpha} \in \mathbb{R}^p,\, q \in (0,\infty)\,)\,,$$

$$\|\boldsymbol{\alpha}\|_0 := \#\bigl\{j \in \{1,\dots,p\} : \alpha_j \neq 0\bigr\} \qquad (\,\boldsymbol{\alpha} \in \mathbb{R}^p\,)$$

and

$$\|\boldsymbol{\alpha}\|_\infty := \max_{j \in \{1,\dots,p\}} |\alpha_j| \qquad (\,\boldsymbol{\alpha} \in \mathbb{R}^p\,)\,.$$

1. Show that for any $q \in [1,\infty]$, the display indeed defines a norm on $\mathbb{R}^p$.

2. Show that for $q \in [0,1)$, the display does not define a norm.

3. Show that for any $q, k \in [1, \infty)$ such that $1/q + 1/k = 1$, the dual of $\mathbf{a} \mapsto \|\mathbf{a}\|_q$ is $\mathbf{a} \mapsto \|\mathbf{a}\|_k$, that is, $\overline{\|\cdot\|_q} = \|\cdot\|_k$. Conclude that $\overline{\|\cdot\|_q}$ is a norm on $\mathbb{R}^p$.

4. Show that for any $q, k \in (0, \infty]$, $q > k$, it holds that
$$\|\boldsymbol{\alpha}\|_q \;\leq\; \|\boldsymbol{\alpha}\|_k \;\leq\; p^{1/k-1/q}\|\boldsymbol{\alpha}\|_q \qquad (\,\boldsymbol{\alpha} \in \mathbb{R}^p\,)\,.$$

Exercises for Section 6.2

□ **Exercise 6.2** $^{\diamond\bullet}$ In this exercise, we compare Lemmas 6.2.1 and 6.2.2 and their corresponding proofs. For this, we assume that the conditions of both lemmas are satisfied; in particular, we consider the identity link $g \;:\; x \;\mapsto\; x$ (cf. Lemma 6.2.1) and a convex prior function h (cf. Lemma 6.2.2).

1. Show that Lemma 6.2.2 implies for any $\boldsymbol{\alpha} \in \mathbb{R}^p$ that
$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; \|X\boldsymbol{\beta} - X\boldsymbol{\alpha}\|_2^2 + 2\langle \boldsymbol{y} - X\boldsymbol{\beta},\, X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\alpha}\rangle + rh[\boldsymbol{\alpha}] - rh[\widehat{\boldsymbol{\beta}}]\,,$$
which coincides with the bound of Lemma 6.2.1.

2. Show that Lemma 6.2.1 implies
$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; 2\langle \boldsymbol{y} - X\boldsymbol{\beta},\, X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}\rangle + rh[\boldsymbol{\beta}] - rh[\widehat{\boldsymbol{\beta}}]\,,$$
while a slight modification of the proof of Lemma 6.2.1 yields
$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}\|_2^2 \;\leq\; \langle \boldsymbol{y} - X\boldsymbol{\beta},\, X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}\rangle + \frac{r}{2}h[\boldsymbol{\beta}] - \frac{r}{2}h[\widehat{\boldsymbol{\beta}}]\,.$$

Exercises for Section 6.3

□ **Exercise 6.3** (Dual Functions) $^{\diamond\bullet}$ In this exercise, we study the properties of dual functions. Recall that we define the dual function $\overline{h} : \mathbb{R}^p \to [-\infty, \infty]$ of $h : \mathbb{R}^p \to [-\infty, \infty]$ with respect to the standard inner product on $\mathbb{R}^p$ via
$$\overline{h}(\boldsymbol{a}) \;=\; \sup\bigl\{\,\langle \boldsymbol{a}, \boldsymbol{k}\rangle \;:\; \boldsymbol{k} \in \mathbb{R}^p,\, h(\boldsymbol{k}) \leq 1\bigr\} \qquad (\,\boldsymbol{a} \in \mathbb{R}^p\,)\,.$$

Recall also that $[-\infty, \infty] = \mathbb{R} \cup \{-\infty, +\infty\}$ is the extended real line, and that the supremum over the empty set is defined as $-\infty$.

1. (Positive Semi-Homogeneity) Show that for any $c \in \mathbb{R}$, $\boldsymbol{a} \in \mathbb{R}^p$, it holds that
$$\overline{h}(c\boldsymbol{a}) \;=\; |c|\overline{h}(\operatorname{sign}(c)\boldsymbol{a})\,.$$
This implies in particular that $\overline{h}(\mathbf{0}_p) = 0$ (set $c = 0$ and use the convention $0 \cdot \pm\infty = 0$).

2. (Positivity) Show that if $h(\mathbf{0}_p) \leq 1$, then $\overline{h}(\boldsymbol{a}) \geq 0$ for all $\boldsymbol{a} \in \mathbb{R}^p$.

3. (Triangle Inequality) Show that for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$, it holds that
$$\overline{h}(\boldsymbol{a} + \boldsymbol{b}) \;\leq\; \overline{h}(\boldsymbol{a}) + \overline{h}(\boldsymbol{b})\,.$$

4. (Symmetry) Show that if $h(\boldsymbol{a}) = h(-\boldsymbol{a})$ for all $\boldsymbol{a} \in \mathbb{R}^p$, then $\overline{h}(\boldsymbol{a}) = \overline{h}(-\boldsymbol{a})$ for all $\boldsymbol{a} \in \mathbb{R}^p$.

5. (Convexity) Show that for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$, $c \in [0,1]$,
$$\overline{h}[c\boldsymbol{a} + (1-c)\boldsymbol{b}] \;\leq\; c\overline{h}[\boldsymbol{a}] + (1-c)\overline{h}[\boldsymbol{b}]\,.$$

As an example, we consider $\mathbb{R}^p = \mathbb{R}$ and
$$h \;:\; a \mapsto \begin{cases} a & \text{if } a \geq 0 \\ \infty & \text{if } a < 0 \end{cases}.$$

Note that the function h is positive definite and satisfies the triangle inequality, but h is unbounded and not symmetric. In particular, h is not a norm on $\mathbb{R}$.

5. Show that $\overline{h}(a) = a$ if $a \geq 0$ and $\overline{h}(a) = 0$ otherwise.
6. Show that h satisfies the conditions of Lemma B.1.3 (Hölder's inequality) on $\mathbb{R}$. Use the conventions $0/0 = 0$ and $a/\infty = 0$ for $a \in \mathbb{R}$.

This example illustrates that functions do not need to be symmetric or bounded for our general version of Hölder's inequality.

□ **Exercise 6.4** $^{\diamond\diamond}$ The signs of $X^\top(\boldsymbol{y} - X\beta^*)$ in the bounds on the tuning parameter r in Lemmas 6.3.1 and 6.3.2 are such that the results are most concise, but other choices could be made. Develop an equivalent of Lemma 6.3.1 with the condition $\widehat{xxxr} \geq v\overline{h}[X^\top(X\beta^* - \boldsymbol{y})]$; develop an equivalent of Lemma 6.3.2 with the condition $\widehat{xxxr} \geq v\overline{h}[X^\top(\boldsymbol{y} - X\beta^*)]$;

□ **Exercise 6.5** $^{\diamond}$ In this exercise, we use the third part of Lemma 6.3.1 to relate the Lagrange versions of the estimators to their constraint versions. Consider an estimator in Lagrange formulation
$$\widehat{\boldsymbol{\beta}}_L \;\in\; \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^p} \bigl\{\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + rh[\boldsymbol{\alpha}]\bigr\}$$
with tuning parameter $r > 0$, and a corresponding constraint estimator
$$\widehat{\boldsymbol{\beta}}_C \;\in\; \operatorname{argmin}\bigl\{\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 \;:\; \boldsymbol{\alpha} \in \mathbb{R}^p, h[\boldsymbol{\alpha}] \leq t\bigr\}$$
with tuning parameter $t > 0$. Now assume for a $t < 1$, it holds that
$$rh[\widehat{\boldsymbol{\beta}}] \;\leq\; \frac{1+t}{1-t} rh[\boldsymbol{\beta}]$$
and $t \geq (1+t)/(1-t)h[\boldsymbol{\beta}]$.

1. Show that $h[\widehat{\boldsymbol{\beta}}_L] \leq t$ and
$$\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_C\|_2^2 \;\leq\; \|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_L\|_2^2\,.$$
2. Give further assumptions under which
$$\operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^p} \bigl\{\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + rh[\boldsymbol{\alpha}]\bigr\} \;=\; \operatorname{argmin}\bigl\{\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 \;:\; \boldsymbol{\alpha} \in \mathbb{R}^p, h[\boldsymbol{\alpha}] \leq t\bigr\}\,.$$

□ **Exercise 6.6** $^{\diamond\diamond}$ Specialize the results of Sections 6.2 and 6.3 to retrieve the bounds for the lasso stated in the motivational section.

Exercises for Section 6.4

□ **Exercise 6.7** $^{\diamond\diamond\bullet}$ In this exercise, we study the Compatibility Condition 6.4.1 in a simple example. We consider the following design matrix and corresponding gram matrix:

$$X \;=\; \begin{pmatrix} 1 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{\sqrt{3}}{2} & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \qquad \Rightarrow \qquad X^\top X \;=\; \begin{pmatrix} 1 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

For convinience, we define

$$\mathcal{C} \;:=\; \Bigl\{\boldsymbol{\delta} \in \mathbb{R}^4 \;:\; \|\boldsymbol{\delta}_{\mathcal{S}^\complement}\|_1 \leq \frac{v+1}{v-1}\|\boldsymbol{\delta}_{\mathcal{S}}\|_1\Bigr\} \setminus \{\mathbf{0}_4\}\,,$$

which is the set of vectors for which the inequality in the compatibility condition needs to be satisfied when h is the ℓ_1-norm (the condition trivially holds for $\boldsymbol{\delta} = \mathbf{0}_4$).

1. Show that irrespective of $\mathcal{S}$ (while we still assume that $\mathcal{S}$ is non-empty), it holds that
$$\min_{\boldsymbol{\delta}\in\mathcal{C}} \frac{\|X\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}\|_1^2} \;=\; \min_{\boldsymbol{\delta}\in\mathcal{C}} \frac{\|\boldsymbol{\delta}_{\{1,2\}}\|_1^2}{4\|\boldsymbol{\delta}\|_1^2}\,.$$
This identity is the basis for the three following tasks.

2. Conclude from 1. that for $\mathcal{S} = \{1\}$ and $\mathcal{S} = \{2\}$, it holds that
$$\min_{\boldsymbol{\delta}\in\mathcal{C}} \frac{\|X\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}\|_1^2} \;=\; \frac{(v-1)^2}{16v^2}\,. \qquad \text{“simplifying”}$$
This means that given v, the Compatibility Condition 6.4.1 is satisfied for the ℓ_1-norm if and only if $m \geq 4v/(v-1)$.

3. Conclude from 1. that for $\mathcal{S} = \{1,2\}$, it also holds that
$$\min_{\boldsymbol{\delta}\in\mathcal{C}} \frac{\|X\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}\|_1^2} \;=\; \frac{(v-1)^2}{16v^2}\,. \qquad \text{“simplifying”}$$
This means that given $v > 0$, the Compatibility Condition 6.4.1 is satisfied if and only if $m \geq 4v/(v-1)$.

4. Conclude from 1. that if $\{3\} \in \mathcal{S}$ or $\{4\} \in \mathcal{S}$, it holds that
$$\min_{\boldsymbol{\delta}\in\mathcal{C}} \frac{\|X\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}\|_1^2} \;=\; 0\,.$$
This means that for any $\mathcal{S}$ with $\mathcal{S}\cap\{3,4\} \neq \varnothing$, the Compatibility Condition 6.4.1 cannot be satisfied.

Exercises for Section 6.5*

□ **Exercise 6.8** $^{\diamond\bullet}$ Show that for any random variable $v \in \mathbb{R}$ and any constant $a \in \mathbb{R}$, it holds that
$$\mathbb{V}\mathrm{ar}[v] \;\leq\; \mathbb{E}(v-a)^2\,.$$

□ **Exercise 6.9** $^{\diamond\diamond\bullet}$ In this exercise, we derive properties of the expectations.

1. Show that for any non-negative random variable v, it holds that
$$\mathbb{E}v \;=\; \int_0^\infty \mathbb{P}\{v \geq t\}\,dt\,.$$

2. Show that for an arbitrary random variable v, it holds that
$$\mathbb{E}v \,\leq\, \int_0^\infty \mathbb{P}\{v \geq t\}\,dt\,.$$

□ **Exercise 6.10** $^{\diamond}$ Give examples of prior functions h that satisfy the assumptions in Section 6.5.

□ **Exercise 6.11** $^{\diamond\diamond}$ Show that a median of a (one-dimensional) distribution that is finite with probability one always exists.

Chapter 7

Theory II: Estimation and Support Recovery

7.1 Overview

In the previous chapter, we have aimed at prediction targets $X\boldsymbol{\beta} \in \mathbb{R}^n$. In this chapter, we drop the Xlense and aim at $\boldsymbol{\beta}$ directly. Prediction and estimation/support recovery differ in two main aspects. First, while the prediction target $X\boldsymbol{\beta}$ is always well-defined, the estimation target $\boldsymbol{\beta}$ is typically ambiguous in high-dimensional settings. For example, the collinearity of x_1 and x_3 in Table **??** on Page **??** makes it impossible to distinguish certain models (such as $\boldsymbol{\beta} = (1/2, 0, \dots, 0)^\top$ and $\boldsymbol{\beta} = (0, 0, 1, 0, \dots, 0)^\top$, for example) only on basis of the data. Second, the two types of objective concern different sets of questions. In genomics for example, prediction can provide models that relate phenotypes $\boldsymbol{y}$ with the genome expressions X, while support recovery studies *which* genes correlated with the phenotype and estimation measures the strengths of these dependencies.

These observations suggest that estimation and support recovery are more challenging than prediction. And indeed, while we have established a widely applicable power-one bound for prediction, all guarantees in this chapter hinge on strict conditions similar to those of the power-two bound.

Except for the different loss functions, the setting is the same as in the previous chapter: we consider again regression-type data $\boldsymbol{y} \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ (without specific model assumptions) and corresponding regularized estimators of the form (1.5).

7.2 Prior Function Loss

In this part, we derive a bound where the loss functions equals the prior function itself. This will be a simple task, because as it turns out, surprisingly, we have already encoutered all necessary concepts in the previous chapter. Specifically, we will invoke the power-two bound of Section 6.3 and the discussion of this bound of Section 6.4.

We find the following oracle inequality.

Theorem 7.2.1 (Prior Function Bound) Consider a function $h : \mathbb{R}^p \to [-\infty, \infty]$, a set $\mathcal{S} \subset \{1, \dots, p\}$, and positive constants $m, \tilde{t} > 0$. Assume that the conditions of Theorem 6.3.2 and Lemma 6.4.1 are met and that Assumption 6.4.1 holds regarding

$h, \mathcal{S}, m, \tilde{t}$. Assume that there is a parameter $\boldsymbol{\alpha} \in \mathbb{R}^p$ that is is a good approximation of $\boldsymbol{\beta}$ on the set $\mathcal{S}$ in the sense that

2. $\operatorname{supp}(\boldsymbol{\alpha}) = \mathcal{S}$;
3. $\overline{h}[X^\top(X\boldsymbol{\alpha} - \boldsymbol{y})] \leq (1 + c_2)\widehat{xxx}r/v$ and $h(\boldsymbol{\beta} - \boldsymbol{\alpha}) \leq dr$.

Then,

$$h(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \;\leq\; \bigl(d + m^2\sqrt{64 + 32\tilde{t}}\,\widehat{xxx}\bigr)r\,.$$

The loss function in this oracle inequality is $\ell = h$. The assumptions and the bound itself are very similar to the power-two bound on Page 101. This similarity is no coincidence, as the following proof relates the two results through the compatibility condition stated in Section 6.4.

Proof of Theorem 7.2.1. We proceed similarly as described in Section 6.4.

We first invoke the assumed triangle inequality on h (see 4. on Page 101) and then *3.* in the description of the theorem to find

$$h(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \;\leq\; h(\boldsymbol{\beta} - \boldsymbol{\alpha}) + h(\boldsymbol{\alpha} - \widehat{\boldsymbol{\beta}}) \;\leq\; dr + h(\boldsymbol{\alpha} - \widehat{\boldsymbol{\beta}})\,.$$

This means that we are left with bounding $h(\boldsymbol{\beta} - \boldsymbol{\alpha})$. For this, we combine the second part of Lemma 6.4.1 ($c = \tilde{t}$) and Assumption 6.4.1 ($\boldsymbol{\delta} = \boldsymbol{\alpha} - \widehat{\boldsymbol{\beta}}$) to find

$$h(\boldsymbol{\alpha} - \widehat{\boldsymbol{\beta}}) \;\leq\; m\|X\boldsymbol{\alpha} - X\widehat{\boldsymbol{\beta}}\|_2^2\,.$$

We can then invoke the second part power-two bound in Theorem 6.3.2 ($\boldsymbol{\beta} = \boldsymbol{\alpha}$, $t = 1 + \tilde{t}$, $a = 3$) to derive

$$h(\boldsymbol{\alpha} - \widehat{\boldsymbol{\beta}}) \leq \frac{16m^2\sqrt{2 + \tilde{t}}\,\widehat{xxx}r}{\sqrt{8}} \;=\; m^2\sqrt{64 + 32\tilde{t}}\,\widehat{xxx}r\,.$$

Plugging this into the first display yields

$$h(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \;\leq\; h(\boldsymbol{\beta} - \boldsymbol{\alpha}) + h(\boldsymbol{\alpha} - \widehat{\boldsymbol{\beta}}) \;\leq\; \bigl(d + m^2\sqrt{64 + 32\tilde{t}}\,\widehat{xxx}\bigr)r$$

as desired □

7.3 Primal-dual Witness Construction*

In this section, we derive bounds for $\overline{h}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})$ by using the *primal-dual witness technique*. This technique constructs sparse vectors through solving a reduced optimization problem, and then gives conditions for these vectors also being a solution of the full optimization problem of the estimator. Given that any of the 2^p subsets of $\{1, \dots, p\}$ is a priori a potential support of such a vector, this approach is not meant for actually computing an estimator. Instead, the approach aims at establishes theoretical "witnesses" for sparse solutions of the inital optimization problem to exist.

We consider once more estimators of the form

$$\widehat{\boldsymbol{\beta}} \;\in\; \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^p} \bigl\{\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + rh[\boldsymbol{\alpha}]\bigr\}\,, \tag{7.1}$$

where $r > 0$ is a tuning parameter and h a convex prior function that satisfies

$$h(\boldsymbol{\alpha}) \;=\; h_{\mathcal{S}}(\boldsymbol{\alpha}_{\mathcal{S}}) + h_{\mathcal{S}^\complement}(\boldsymbol{\alpha}_{\mathcal{S}^\complement}) \qquad\quad (\boldsymbol{\alpha} \in \mathbb{R}^p)$$

for (convex, cf. Exercise 7.3) functions $h_{\mathcal{S}} : \mathbb{R}^s \to [-\infty, \infty]$ and $h_{\mathcal{S}^\complement} : \mathbb{R}^{p-s} \to [-\infty, \infty]$ and an index set $\mathcal{S} \subset \{1, \dots, p\}$ with size $s := |\mathcal{S}|$.

Successful Construction

We first construct sparse primal vectors together with matching (non-sparse) dual vectors.

primal-dual witness construction

Definition 7.3.1 (Primal-Dual Witness Construction) Construct vectors $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}^\top, \widehat{\boldsymbol{\gamma}}_{\mathcal{S}^\complement}^\top)^\top \in \mathbb{R}^p$ (primal vector) and $\widehat{\boldsymbol{\nu}} = (\widehat{\boldsymbol{\nu}}_{\mathcal{S}}^\top, \widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement}^\top)^\top \in \mathbb{R}^p$ (dual vector) as follows:
(Primal 1) Define the vector $\widehat{\boldsymbol{\gamma}}_{\mathcal{S}} \in \mathbb{R}^s$ such that

$$\widehat{\boldsymbol{\gamma}}_{\mathcal{S}} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^s} \left\{\|\boldsymbol{y} - X_{\mathcal{S}}\boldsymbol{\theta}\|_2^2 + rh_{\mathcal{S}}(\boldsymbol{\theta})\right\};$$

(Primal 2) Define $\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^\complement} := \mathbf{0}_{p-s}$;
(Dual 1) Define the vector $\widehat{\boldsymbol{\nu}}_{\mathcal{S}} \in \mathbb{R}^s$ such that

$$-2X_{\mathcal{S}}^\top(\boldsymbol{y} - X_{\mathcal{S}}\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}) + r\widehat{\boldsymbol{\nu}}_{\mathcal{S}} = \mathbf{0}_s;$$

(Dual 2) Define the vector $\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement} \in \mathbb{R}^{p-s}$ such that

$$-2X_{\mathcal{S}^\complement}^\top(\boldsymbol{y} - X_{\mathcal{S}}\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}) + r\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement} = \mathbf{0}_{p-s}.$$

The hats emph that the vectors are functions of the data. Given $\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}$, all other vectors are uniquely defined (recall that $r \neq 0$ by assumption). The vector $\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}$ is a solution of a variant of the problem (7.1) that is restricted to $\mathcal{S}$; the vector $\widehat{\boldsymbol{\nu}}_{\mathcal{S}}$ is supposed to be a subgradient vector of $h_{\mathcal{S}}$ at $\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}$. Therefore, we could call the approach also "primal-subgradient witness technique." We use the terms subgradient and dual interchangeably, since subgradients and solutions of the dual problem are directly related anyways.

The KKT conditions for minima of convex functions together with the additivity of subdifferentials imply directly that if $\widehat{\boldsymbol{\nu}} \in \boldsymbol{\partial}h(\widehat{\boldsymbol{\gamma}})$, it holds that $\widehat{\boldsymbol{\gamma}}$ is a solution of (7.1). We formalize this in the following lemma.

Lemma 7.3.1 (Success of the Primal-Dual Witness Construction) Let $\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\nu}}$ be constructed according to Definition 7.3.1. Then, if $\widehat{\boldsymbol{\nu}} \in \boldsymbol{\partial}h(\widehat{\boldsymbol{\gamma}})$, it holds that $\widehat{\boldsymbol{\gamma}}$ is a solution of (7.1), that is,

$$\widehat{\boldsymbol{\gamma}} \in \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p} \left\{\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + rh[\boldsymbol{\alpha}]\right\}.$$

The only condition for this result to hold is convexity of the objective function. The lemma then gives a condition for $\widehat{\boldsymbol{\gamma}}$, which is constructed via a solving the *restricted* problem, being a solution of the *full* problem 7.1 also.

Resulting Dual Bounds

Irrespective of whether the primal-dual construction is successful according to Lemma 7.3.1 or not, any vector $\widehat{\boldsymbol{\gamma}}$ that is constructed as in Definition 7.3.1 satisfies a bound under minimal assumptions.

Theorem 7.3.1 (Dual Bound) If $X_{\mathcal{S}}^\top X_{\mathcal{S}}$ is invertible and $\overline{h}_{\mathcal{S}}$ satisfies the triangle inequality, it holds for any target $\boldsymbol{\beta} \in \mathbb{R}^p$ that

$$\overline{h}(\boldsymbol{\beta} - \widehat{\boldsymbol{\gamma}}) \le \overline{h}_{\mathcal{S}}(r(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}\widehat{\boldsymbol{\nu}}_{\mathcal{S}}/2) + \overline{h}_{\mathcal{S}}((X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}X_{\mathcal{S}}^\top(X_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{y})) + \overline{h}_{\mathcal{S}^\complement}(\boldsymbol{\beta}_{\mathcal{S}^\complement}).$$

The assumptions in Theorem 7.3.1 are minimal: the invertibility of the restricted gram matrix $X_{\mathcal{S}}^\top X_{\mathcal{S}}$ simply ensures identifiability on $\mathcal{S}$; the triangle inequality is satisfied, for example, by any norm. The main restriction is actually the decomposability of h, which disallows $h(\cdot) = \|\cdot\|_2$ (unless $\mathcal{S} = \varnothing$ or $\mathcal{S} = \{1, \dots, p\}$), for example. Popular h that satisfy all requirements are ℓ_1-norms and grouped ℓ_1/ℓ_2 variants of it (where one needs to ensure that $\mathcal{S}$ complies with the group structure).

While the result applies to any primal vector $\widehat{\boldsymbol{\gamma}}$ constructed on the set $\mathcal{S}$, the bound is only interesting if (i) the columns of $X_{\mathcal{S}}$ is not too correlated and r is reasonably small (such that the first term in the bound is small), (ii) $X_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}$ approximates $\boldsymbol{y}$ well (such that the second term is small), and (iii) $\boldsymbol{\beta}_{\mathcal{S}^{\complement}}$ is small (such that the third term is small). Because we cannot verify the latter two conditions in practice, the theorem does not motivate a strategy for how to construct a good estimator by using the primal-dual witness technique. Instead, its value is to prove that if the right-hand side is small for a set $\mathcal{S}$ and if additionally $\widehat{\boldsymbol{\nu}} \in \boldsymbol{\partial} h(\widehat{\boldsymbol{\gamma}})$, then there is a sparse solution of (7.1) that is close to the target in terms of $\overline{h}$.

Let us illustrate the bound in a simple setting.

Corollary 7.3.1 (Orthogonal Design and Sufficiently Large $\mathcal{S}$) If $X_{\mathcal{S}}^\top X_{\mathcal{S}} = n\,\mathrm{I}_{s\times s}$ and $\operatorname{supp}(\boldsymbol{\beta}) \subset \mathcal{S}$, the above bound reads

$$\overline{h}(\boldsymbol{\beta} - \widehat{\boldsymbol{\gamma}}) \le \overline{h}_{\mathcal{S}}(r\widehat{\boldsymbol{\nu}}_{\mathcal{S}}/2) + \overline{h}_{\mathcal{S}}(X_{\mathcal{S}}^\top(X_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{y}))\,.$$

The first term is typically of the order r; the second term is the usual noise term restricted to $\mathcal{S}$.

Remark 7.3.1 (On the Set $\mathcal{S}$) Previous descriptions of the primal-dual witness construction assume that (i) $\boldsymbol{\beta}$ is the true regression vector and (ii) $\mathcal{S} = \operatorname{supp}(\boldsymbol{\beta})$. Our general treatment does away these restriction, and instead suggests that the best choice of $\mathcal{S}$ for any given target $\boldsymbol{\beta}$ is

$$\begin{aligned}\operatorname{argmin}\Big\{&\overline{h}_{\mathcal{S}}(r(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}\widehat{\boldsymbol{\nu}}_{\mathcal{S}}/2) + \overline{h}_{\mathcal{S}}((X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}X_{\mathcal{S}}^\top(X_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{y})) + \overline{h}_{\mathcal{S}^{\complement}}(\boldsymbol{\beta}_{\mathcal{S}^{\complement}}) \;:\\ &X_{\mathcal{S}}^\top X_{\mathcal{S}} \text{ invertible, } \overline{h}_{\mathcal{S}} \text{ satisfies the triangle inequality, and } \widehat{\boldsymbol{\nu}} \in \boldsymbol{\partial} h(\widehat{\boldsymbol{\gamma}})\Big\}\,,\end{aligned}$$

where the argmin is taken over subsets of $\{1, \dots, p\}$. This allows one to balance aspects such as the correlations (large $\mathcal{S}$ lead to large correlations) and tuning ($\mathcal{S}$ needs to be sufficiently large to allow for effective tuning, see Lemma 7.3.3). Again, this does not mean that $\mathcal{S}$ has a say in the practical estimation, it instead sharpens the theoretical guarantees of $\widehat{\boldsymbol{\beta}}$.

We now turn to the proof of Theorem 7.3.1.

Proof of Theorem 7.3.1. Recall that the definition (Dual 1) ensures

$$-2X_{\mathcal{S}}^\top(\boldsymbol{y} - X_{\mathcal{S}}\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}) + r\widehat{\boldsymbol{\nu}}_{\mathcal{S}} = \mathbf{0}_s\,.$$

Adding a zero-valued term and rearranging yields

$$-2X_{\mathcal{S}}^\top(X_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} - X_{\mathcal{S}}\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}) + 2X_{\mathcal{S}}^\top(X_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{y}) + r\widehat{\boldsymbol{\nu}}_{\mathcal{S}} = \mathbf{0}_s$$

and

$$2X_{\mathcal{S}}^\top X_{\mathcal{S}}(\boldsymbol{\beta}_{\mathcal{S}} - \widehat{\boldsymbol{\gamma}}_{\mathcal{S}}) = r\widehat{\boldsymbol{\nu}}_{\mathcal{S}} + 2X_{\mathcal{S}}^\top(X_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{y})\,.$$

Since the restricted gram matrix $X_{\mathcal{S}}^\top X_{\mathcal{S}}$ is invertible by assumption, we find

$$\boldsymbol{\beta}_{\mathcal{S}} - \widehat{\boldsymbol{\gamma}}_{\mathcal{S}} = r(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}\widehat{\boldsymbol{\nu}}_{\mathcal{S}}/2 + (X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}X_{\mathcal{S}}^\top(X_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{y})\,.$$

We now invoke the assumed triangle inequality for $\overline{h}_{\mathcal{S}}$ to derive

$$\overline{h}_{\mathcal{S}}(\boldsymbol{\beta}_{\mathcal{S}} - \widehat{\boldsymbol{\gamma}}_{\mathcal{S}}) \le \overline{h}_{\mathcal{S}}(r(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}\widehat{\boldsymbol{\nu}}_{\mathcal{S}}/2) + \overline{h}_{\mathcal{S}}((X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}X_{\mathcal{S}}^\top(X_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{y}))\,.$$

We can now conclude

$$\begin{aligned}
&\overline{h}(\boldsymbol{\beta}-\widehat{\boldsymbol{\gamma}})\\
&= \overline{h}_{\mathcal{S}}(\boldsymbol{\beta}_{\mathcal{S}}-\widehat{\boldsymbol{\gamma}}_{\mathcal{S}})+\overline{h}_{\mathcal{S}^\complement}(\boldsymbol{\beta}_{\mathcal{S}^\complement}-\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^\complement}) && \text{(assumed form of } h)\\
&= \overline{h}_{\mathcal{S}}(\boldsymbol{\beta}_{\mathcal{S}}-\widehat{\boldsymbol{\gamma}}_{\mathcal{S}})+\overline{h}_{\mathcal{S}^\complement}(\boldsymbol{\beta}_{\mathcal{S}^\complement}) && \text{(Primal 2)}\\
&\le \overline{h}_{\mathcal{S}}(r(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}\widehat{\boldsymbol{\nu}}_{\mathcal{S}}/2)+\overline{h}_{\mathcal{S}}((X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}X_{\mathcal{S}}^\top(X_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}-\boldsymbol{y}))+\overline{h}_{\mathcal{S}^\complement}(\boldsymbol{\beta}_{\mathcal{S}^\complement})\\
& && \text{(previous display)}
\end{aligned}$$

as desired. □

Sufficient Conditions for a Successful Construction

We now give sufficient conditions for $\widehat{\boldsymbol{\nu}}\in\boldsymbol{\partial}h(\widehat{\boldsymbol{\gamma}})$.

Lemma 7.3.2 (Sufficient Condition on the Dual Vector) Assume that $h_{\mathcal{S}^\complement}$ is positive definite and satisfies Hölder's inequality in Lemma B.1.3. Then, $\overline{h}_{\mathcal{S}^\complement}(\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement})\le 1$ implies $\widehat{\boldsymbol{\nu}}\in\boldsymbol{\partial}h(\widehat{\boldsymbol{\gamma}})$.

Proof of Lemma 7.3.2. We show that $\widehat{\boldsymbol{\nu}}_{\mathcal{S}}\in\boldsymbol{\partial}h_{\mathcal{S}}(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}})$ and $\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement}\in\boldsymbol{\partial}h_{\mathcal{S}^\complement}(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^\complement})$ separately, which then implies $\widehat{\boldsymbol{\nu}}\in\boldsymbol{\partial}h(\widehat{\boldsymbol{\gamma}})$ by the additivity of subdifferentials.

Note first that $\widehat{\boldsymbol{\nu}}_{\mathcal{S}}\in\boldsymbol{\partial}h_{\mathcal{S}}(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}})$ by the KKT conditions for the objective function in (Primal 1) and by (Dual 1). Indeed, let $\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}$ be a solution of the objective function in (Primal 1), then by the KKT conditions, there is a $\widehat{\boldsymbol{\kappa}}\in\boldsymbol{\partial}h_{\mathcal{S}}(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}})$ such that

$$-2X_{\mathcal{S}}^\top(\boldsymbol{y}-X_{\mathcal{S}}\widehat{\boldsymbol{\gamma}}_{\mathcal{S}})+r\widehat{\boldsymbol{\kappa}}\;=\;\mathbf{0}_s\,.$$

By definition of $\widehat{\boldsymbol{\nu}}_{\mathcal{S}}$ in (Dual 1), we find $\widehat{\boldsymbol{\kappa}}=\widehat{\boldsymbol{\nu}}_{\mathcal{S}}$, that is, $\widehat{\boldsymbol{\nu}}_{\mathcal{S}}\in\boldsymbol{\partial}h_{\mathcal{S}}(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}})$.

The situation for $\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement}$ is more involved, because (Dual 2) does not relate directly to a minimizer of an objective function. We first observe

$$\begin{aligned}
&\langle\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement},\,\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^\complement}\rangle\\
&=\langle\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement},\,\boldsymbol{\gamma}\rangle && \text{(Primal 2)}\\
&\le \overline{h}_{\mathcal{S}^\complement}(\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement})h_{\mathcal{S}^\complement}(\boldsymbol{\gamma}) && \text{(Hölder's inequality)}\\
&\le h_{\mathcal{S}^\complement}(\boldsymbol{\gamma})\,. && (h_{\mathcal{S}^\complement}\text{ non-negative and }\overline{h}_{\mathcal{S}^\complement}(\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement})\le 1\text{ by assumption})
\end{aligned}$$

Now, rearranging and using that $h_{\mathcal{S}^\complement}(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^\complement})=0$ by the assumed positive definiteness of $h_{\mathcal{S}^\complement}$ and (Primal 2),

$$h_{\mathcal{S}^\complement}(\boldsymbol{\gamma})\;\ge\;h_{\mathcal{S}^\complement}(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^\complement})+\langle\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement},\,\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^\complement}\rangle$$

for all $\boldsymbol{\gamma}\in\mathbb{R}^{p-s}$, which means that $\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement}\in\boldsymbol{\partial}h_{\mathcal{S}^\complement}(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}^\complement})$.

In conclusion, since subdifferentials are additive, $\overline{h}_{\mathcal{S}^\complement}(\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement})\le 1$ is a sufficient condition for $\widehat{\boldsymbol{\nu}}\in\boldsymbol{\partial}h(\widehat{\boldsymbol{\gamma}})$ under the stated conditions. □

irrepresentability condition

Assumption 7.3.1 (Irrepresentability Condition) Given a function $h:\mathbb{R}^p\to\mathbb{R}$, a non-empty set $\mathcal{S}\subset\{1,\dots,p\}$, and a positive and finite constant $c\in(0,\infty)$, we say that the *irrepresentability condition* holds if

$$\sup_{\mathbf{m}\in\boldsymbol{\partial}h_{\mathcal{S}}[\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}]}\overline{h}_{\mathcal{S}^\complement}[X_{\mathcal{S}^\complement}^\top X_{\mathcal{S}}(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}\mathbf{m}]\;\le\;1-c\,.$$

We call c the *irrepresentability constant.*

This condition ensures that there is not too much correlation among the columns of X. In the orthogonal design, where $X^\top X = \mathrm{I}_{p\times p}$, the assumption is $\overline{h}_{\mathcal{S}^\complement}(\mathbf{0}_{p-s}) \le 1-c$, which is satisfied by any positive definite function, for example. One can also interpret the assumption in terms of regression coefficients in some cases—see Exercise 7.5. Typically, the Irrespresentability Condition imposes strong restrictions on the underlying model; therefore, in practice, the relevance of corresponding theoretical bounds need to be evaluated with care.

We conclude this section with a concrete condition for a successful primal-dual construction.

Lemma 7.3.3 (Sufficient Condition) Assume that the Irrepresentability Condition 7.3.1 is satisfied with constant $c > 0$, and that

$$r \;>\; \inf\{\overline{h}_{\mathcal{S}^\complement}\big(2X_{\mathcal{S}^\complement}^\top(\mathrm{I}_n - X_{\mathcal{S}}(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}X_{\mathcal{S}}^\top)(\boldsymbol{y}-X\boldsymbol{\alpha})\big)/b \;:\; \boldsymbol{\alpha}\in\mathbb{R}^p,\, \mathrm{supp}(\boldsymbol{\alpha})\subset\mathcal{S}\}\,.$$

Assume also that $h_{\mathcal{S}^\complement}$ satisfies the triangle inequality. Then,

$$\overline{h}_{\mathcal{S}^\complement}(\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement}) \;<\; 1\,.$$

A successful construction as described in Lemma 7.3.1 requires only an inequality in the bound on $\overline{h}_{\mathcal{S}^\complement}(\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement})$; neverthless, the strict inequality can be useful for showing uniqueness of the estimator $\widehat{\boldsymbol{\beta}}$, cf. Exercise 7.7. Generally speaking, the lemma ensures that if (i) the columns of the design matrix X are not too much correlated among each other and (ii) if the tuning parameter is large enough, the primal-dual construction is successful.

Proof of Lemma 7.3.3. We rewrite (Dual 1&2) in Definition 7.3.1, solve for $r\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement}$, and then show that the desired function of it is smaller than r.

For this, we recall that (Dual 1&2) yield after adding a zero-valued term in the middle

$$-2X^\top\left(X\boldsymbol{\alpha}+\boldsymbol{y}-X\boldsymbol{\alpha}-X_{\mathcal{S}}\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}\right)+r\begin{pmatrix}\widehat{\boldsymbol{\nu}}_{\mathcal{S}}\\ \widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement}\end{pmatrix} \;=\; \mathbf{0}_p\,.$$

If now $\mathrm{supp}(\boldsymbol{\alpha})\subset\mathcal{S}$, we can rearrange this to

$$-2X^\top\left(X\begin{pmatrix}\boldsymbol{\alpha}_{\mathcal{S}}\\ \mathbf{0}_{p-s}\end{pmatrix}-X\begin{pmatrix}\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}\\ \mathbf{0}_{p-s}\end{pmatrix}\right)-2X^\top(\boldsymbol{y}-X\boldsymbol{\alpha})+r\begin{pmatrix}\widehat{\boldsymbol{\nu}}_{\mathcal{S}}\\ \widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement}\end{pmatrix} \;=\; \mathbf{0}_p\,.$$

We can write this in block matrix form according to

$$-2\begin{pmatrix}X_{\mathcal{S}}^\top X_{\mathcal{S}} & X_{\mathcal{S}}^\top X_{\mathcal{S}^\complement}\\ X_{\mathcal{S}^\complement}^\top X_{\mathcal{S}} & X_{\mathcal{S}^\complement}^\top X_{\mathcal{S}^\complement}\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}_{\mathcal{S}}-\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}\\ \mathbf{0}_{p-s}\end{pmatrix}-2\begin{pmatrix}X_{\mathcal{S}}^\top(\boldsymbol{y}-X\boldsymbol{\alpha})\\ X_{\mathcal{S}^\complement}^\top(\boldsymbol{y}-X\boldsymbol{\alpha})\end{pmatrix}+r\begin{pmatrix}\widehat{\boldsymbol{\nu}}_{\mathcal{S}}\\ \widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement}\end{pmatrix} \;=\; \mathbf{0}_p\,.$$

Recall our convention $X_{\mathcal{S}}^\top = (X_{\mathcal{S}})^\top$. We now solve the second line of the display for $r\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement}$ via

$$-2X_{\mathcal{S}^\complement}^\top X_{\mathcal{S}}(\boldsymbol{\alpha}_{\mathcal{S}}-\widehat{\boldsymbol{\gamma}}_{\mathcal{S}})-2X_{\mathcal{S}^\complement}^\top(\boldsymbol{y}-X\boldsymbol{\alpha})+r\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement} \;=\; \mathbf{0}_{p-s}\,,$$

and hence, by rearranging,

$$r\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement} \;=\; 2X_{\mathcal{S}^\complement}^\top X_{\mathcal{S}}(\boldsymbol{\alpha}_{\mathcal{S}}-\widehat{\boldsymbol{\gamma}}_{\mathcal{S}})+2X_{\mathcal{S}^\complement}^\top(\boldsymbol{y}-X\boldsymbol{\alpha})\,.$$

We now also want to solve the above equation for $\boldsymbol{\alpha}_{\mathcal{S}}-\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}$ and plug this in. We find

$$-2X_{\mathcal{S}}^\top X_{\mathcal{S}}(\boldsymbol{\alpha}_{\mathcal{S}}-\widehat{\boldsymbol{\gamma}}_{\mathcal{S}})-2X_{\mathcal{S}}^\top(\boldsymbol{y}-X\boldsymbol{\alpha})+r\widehat{\boldsymbol{\nu}}_{\mathcal{S}} \;=\; \mathbf{0}_s\,,$$

and therefore, by rearranging,

$$2X_{\mathcal{S}}^\top X_{\mathcal{S}}(\boldsymbol{\alpha}_{\mathcal{S}}-\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}) \;=\; -2X_{\mathcal{S}}^\top(\boldsymbol{y}-X\boldsymbol{\alpha})+r\widehat{\boldsymbol{\nu}}_{\mathcal{S}}\,.$$

Since the matrix $X_{\mathcal{S}}^\top X_{\mathcal{S}}$ is invertible by the Irrepresentable Condition 7.3.1, we can solve this equation for $\boldsymbol{\alpha}_{\mathcal{S}}-\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}$ and find

$$\boldsymbol{\alpha}_{\mathcal{S}}-\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}=-(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}X_{\mathcal{S}}^\top(\boldsymbol{y}-X\boldsymbol{\alpha})+r(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}\widehat{\boldsymbol{\nu}}_{\mathcal{S}}/2.$$

Combining this equation with the earlier equation yields

$$\begin{aligned}r\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement} \;&=\; -2X_{\mathcal{S}^\complement}^\top X_{\mathcal{S}}(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}X_{\mathcal{S}}^\top(\boldsymbol{y}-X\boldsymbol{\alpha})+rX_{\mathcal{S}^\complement}^\top X_{\mathcal{S}}(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}\widehat{\boldsymbol{\nu}}_{\mathcal{S}}+2X_{\mathcal{S}^\complement}^\top(\boldsymbol{y}-X\boldsymbol{\alpha})\\ &=2X_{\mathcal{S}^\complement}^\top(\mathrm{I}_n-X_{\mathcal{S}}(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}X_{\mathcal{S}}^\top)(\boldsymbol{y}-X\boldsymbol{\alpha})+rX_{\mathcal{S}^\complement}^\top X_{\mathcal{S}}(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}\widehat{\boldsymbol{\nu}}_{\mathcal{S}}\,.\end{aligned}$$

So far, $\boldsymbol{\alpha}$ was abitrary. We can thus "optimize" the first term on the right-hand side over $\boldsymbol{\alpha}$: by assumption on r, there is $\boldsymbol{\alpha}$ such that

$$\overline{h}_{\mathcal{S}^\complement}(2X_{\mathcal{S}^\complement}^\top(\mathrm{I}_n-X_{\mathcal{S}}(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}X_{\mathcal{S}}^\top)(\boldsymbol{y}-X\boldsymbol{\alpha}))\;<\;rb\,.$$

For the second term, the Irrepresentable Conditions yields

$$\overline{h}_{\mathcal{S}^\complement}(X_{\mathcal{S}^\complement}^\top X_{\mathcal{S}}(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}\widehat{\boldsymbol{\nu}}_{\mathcal{S}})\;\leq\;r(1-c)\,.$$

Collecting terms and using the assumed triangle inequality gives

$$r\overline{h}_{\mathcal{S}^\complement}(\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement})\;\leq\;rb+r(1-b)\;=\;r$$

which implies, since $r>0$,

$$\overline{h}_{\mathcal{S}^\complement}(\widehat{\boldsymbol{\nu}}_{\mathcal{S}^\complement})\;\leq\;1$$

as desired. □

7.4 Sign-consistent Support Recovery

In this section, we show that the bounds developed in the previous sections can be used to establish guarantees in support recovery. The general form of the earlier bounds is as follows: if $r\geq r^*$, it holds that

$$\ell(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\;\leq\;f(r)\,,$$

where r^* is a potentially random oracle tuning parameter, ℓ some loss, $\boldsymbol{\beta}$ the target of interest, and f a real valued function of the tuning parameters. For example, if the prior function equals $h(\cdot)=\|\cdot\|_q$ for some $q\in[1,\infty)$, the dual bounds in the preceeding two sections concern the loss $\|\cdot\|_m$, $m\in[1,\infty)$ such that $1/m+1/q=1$, and directly imply entrywise bounds:

$$|\boldsymbol{\beta}_j-\widehat{\boldsymbol{\beta}}_j|\;\leq\;\Bigl(\sum_{j=1}^p|\boldsymbol{\beta}_j-\widehat{\boldsymbol{\beta}}_j|^m\Bigr)^{1/m}\;=\;\|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\|_m\;\leq\;f(r)\,.$$

The larger m, the sharper the first inequality; with equality for $m=\infty$. Motivated by this observations, we work with estimators that satisfy the following assumption.

Assumption 7.4.1 (Entrywise Bound) Given an estimator $\widehat{\boldsymbol{\beta}}$ of a target $\boldsymbol{\beta}\in\mathbb{R}^p$, and given a positive constant $c>0$, we assume that

$$|\boldsymbol{\beta}_j-\widehat{\boldsymbol{\beta}}_j|\;\leq\;c\qquad(j\in\{1,\ldots,p\})\,.$$

The following treatment is agnostic to the form or type of the estimators—as long as the entrywise bound is satisfied; however, our main motivations are regularized estimators for a given tuning parameter, noting again that such estimators can satisfy the assumption by virtue of the earlier results in this chapter.

We can now discuss support recovery, the estimation of the set $\operatorname{supp}(\boldsymbol{\beta})$ for a given target $\boldsymbol{\beta}$. Our first lemma concerns variable screening: the result ensures that all sufficiently large elements of $\operatorname{supp}(\boldsymbol{\beta})$ are detected with the correct sign.

Lemma 7.4.1 (Screening) If the entrywise bound in Assumption 7.4.1 holds, we find

$$\operatorname{sign}(\boldsymbol{\beta}_j) = \operatorname{sign}(\widehat{\boldsymbol{\beta}}_j) \qquad (j \ :\ |\boldsymbol{\beta}_j| > c)\,.$$

The result states that on each coordinate where the target has a sufficiently large entry, the estimator has the correct sign. The result ensures in particular that there are no false negatives as long as all non-zero entries of $\boldsymbol{\beta}$ are sufficiently large. The type of assumption on the non-zero entries of the target is called beta-min condition: it summarizes the simple truth that small coordinates are difficult to detect.

The proof of the result is short.

Proof of Lemma 7.4.1. Consider $j \in \{1,\dots,p\}$ in the support of $\boldsymbol{\beta}$ such that $|\boldsymbol{\beta}_j| > c$. Assume first that $\boldsymbol{\beta}_j > 0$. Then,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_j &= \boldsymbol{\beta}_j - \boldsymbol{\beta}_j + \widehat{\boldsymbol{\beta}}_j && \text{“adding a zero-valued term”}\\
&\geq |\boldsymbol{\beta}_j| - |\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j| && \text{“}\boldsymbol{\beta}_j > 0 \text{ by assumption and } -|-a| \leq a \text{ for all } a \in \mathbb{R}\text{”}\\
&> c - c = 0\,. && \text{“beta-min and assumed entrywise bound”}
\end{aligned}$$

Assume now that $\boldsymbol{\beta}_j < 0$. Then,

$$\begin{aligned}
-\widehat{\boldsymbol{\beta}}_j &= -\boldsymbol{\beta}_j + \boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j && \text{“adding a zero-valued term”}\\
&\geq |\boldsymbol{\beta}_j| - |\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j| && \text{“}\boldsymbol{\beta}_j < 0 \text{ by assumption and } -|-a| \leq a \text{ for all } a \in \mathbb{R}\text{”}\\
&> c - c = 0\,. && \text{“beta-min and assumed entrywise bound”}
\end{aligned}$$

This concludes the proof. □

Our next goal is to avoid false positives. For this, we define for $t \geq 0$ the thresholded version $\widehat{\boldsymbol{\beta}}^t$ of the estimator $\widehat{\boldsymbol{\beta}}$ via

$$\widehat{\boldsymbol{\beta}}^t_j := \begin{cases} 0 & \text{if } |\widehat{\boldsymbol{\beta}}_j| \leq t \\ \widehat{\boldsymbol{\beta}}_j & \text{otherwise} \end{cases} \qquad (j \in \{1,\dots,p\})\,.$$

The thresholding removes small entries of the estimator with the hope that the most relevant information is stored in the remaining parts.

Under some assumptions, we then get indeed a guarantee in support recovery.

Lemma 7.4.2 (Support Recovery) If the entrywise bound in Assumption 7.4.1 holds, and if $t \geq c$, we find

$$\operatorname{sign}(\boldsymbol{\beta}_j) = \operatorname{sign}(\widehat{\boldsymbol{\beta}}_j) \qquad (j \ :\ |\boldsymbol{\beta}_j| > c + t \ \text{ or } \ \boldsymbol{\beta}_j = 0)\,.$$

This bound now additionally ensures that there are no false positives. A price to pay is that the beta-min condition is slightly stronger than before. Moreover, while the $\widehat{\boldsymbol{\beta}}$'s typically satisfy favorable bounds also in vector-based losses (this is how we motivated the entrywise bounds), the thresholded versions $\widehat{\boldsymbol{\beta}}^t$ do not necessarily have good properties beyond entrywise bounds and support recovery.

Proof of Lemma 7.4.2. Consider first $j \in \{1, \dots, p\}$ in the support of $\boldsymbol{\beta}$ such that $|\boldsymbol{\beta}_j| > c + t$ and, without loss of generality, $\boldsymbol{\beta}_j > 0$. Then, as before,

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_j &= \boldsymbol{\beta}_j - \boldsymbol{\beta}_j + \widehat{\boldsymbol{\beta}}_j && \text{“adding a zero-valued term”}\\ &\geq |\boldsymbol{\beta}_j| - |\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j| && \text{“}\boldsymbol{\beta}_j > 0 \text{ by assumption”}\\ &> c + t - c = t\,. && \text{“beta-min and assumed entrywise bound”} \end{aligned}$$

Thus, by definition of $\widehat{\boldsymbol{\beta}}^t$, it holds that $\widehat{\boldsymbol{\beta}}^t_j > 0$ as desired.

Assume now that $j \notin \operatorname{supp}(\boldsymbol{\beta})$. Then,

$$\begin{aligned} |\widehat{\boldsymbol{\beta}}_j| &= |\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j| && \text{“}\boldsymbol{\beta}_j = 0 \text{ by assumption”}\\ &\leq c && \text{“assumed entrywise bound”}\\ &\leq t\,, && \text{“}t \geq c \text{ by assumption”} \end{aligned}$$

which means by definition of $\widehat{\boldsymbol{\beta}}^t$ that $\widehat{\boldsymbol{\beta}}^t_j = 0$ as desired. □

Remark 7.4.1 (Sparsity Inducing Prior Functions) For estimators with sparsity inducing prior functions, one can show that $\operatorname{supp}(\widehat{\boldsymbol{\beta}}) \subset \mathcal{S}$ under some assumptions—cf. the solution of Exercise 7.7. In that case, the conditions in Lemma 7.4.1 are sufficient to ensure correct support recovery, that is, no thresholding is required.

7.5 References and Further Reading

A *subspace compability constant* has been formulated in [NYWR12, Definition 3] and [Wai14, Equations (21)]. For ℓ_2-estimation, the main idea relates to our Assumption 6.4.1 on Page 105 as follows: For every $\boldsymbol{\delta} \in \mathcal{C}$,

$$\frac{h(\boldsymbol{\delta})}{\|X\boldsymbol{\delta}\|_2} = \frac{h(\boldsymbol{\delta})}{\|\boldsymbol{\delta}\|_2}\frac{\|\boldsymbol{\delta}\|_2}{\|X\boldsymbol{\delta}\|_2} \leq \Bigl(\sup_{\boldsymbol{\delta}\in\mathcal{C}} \frac{h(\boldsymbol{\delta})}{\|\boldsymbol{\delta}\|_2}\Bigr)\Bigl(\sup_{\boldsymbol{\delta}\in\mathcal{C}} \frac{\|\boldsymbol{\delta}\|_2}{\|X\boldsymbol{\delta}\|_2}\Bigr)\,.$$

In this sense, the assumption on the compatibility constant m can be separated in two parts: the first term, the subspace compatibility constant, measures the compatibility of the prior function and the ℓ_2-loss; the second term, the restricted eigenvalue, measures the compatibilty of ℓ_2-estimation and the prediction loss.

The primal-dual construction (with an assumed true linear regression model) is discussed in [HTW15, Chapter 5.2, Pages 95ff]. References to its origins and to further results on that topic are given in the same book on Page 312. An extension of the approach to some classes of estimators with non-convex objective functions is provided in [LW17].

7.6 Exercises

Exercises for Section 7.2

□ **Exercise 7.1** $^{\diamond\diamond\bullet}$ Theorem 7.2.1 directly entails bounds for loss functions ℓ that are dominated by the prior function h. We study an example where the loss function is the ℓ_2-norm and the penalty function the ℓ_1-norm. We proceed in three steps.
(i) Let $f : \mathbb{R}^p \to \mathbb{R}$ be a convex function. Show that for all $a_1, \dots, a_k \geq 0$ with $a_1 + \dots + a_k = 1$ and for all $\mathbf{b}^1, \dots, \mathbf{b}^k \in \mathbb{R}^p$ *Jensen's inequality* holds:

$$f\Bigl(\sum_{i=1}^k a_i \mathbf{b}^i\Bigr) \leq \sum_{i=1}^k a_i f\bigl(\mathbf{b}^i\bigr)\,.$$

(ii) Use the above inequality to show that $\|\cdot\|_s \le \|\cdot\|_t$ for all $s, t \ge 0$, $s \ge t$.
(iii) Conclude that if the conditions of Theorem 7.2.1 with $h(\cdot) := \|\cdot\|_1$ are satisfied, it holds that

$$\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_2 \;\le\; \bigl(d + m^2\sqrt{64 + 32\tilde{t}\,\widehat{xxx}}\bigr)r\,.$$

Remark: An important aspect here is the condition $\overline{h}[X^\top(X\boldsymbol{\alpha} - \boldsymbol{y})] \le (1 + c_2)\widehat{xxx}r/v$, which involves $\overline{h}(\cdot) = \|\cdot\|_\infty$ for ℓ_1-regularization. Instead, ℓ_2-regularization would involve $\overline{h}(\cdot) = \|\cdot\|_2$, which can lead to a need for larger tuning parameters, and thus, less favorable bounds.

Exercises for Section 7.3

□ **Exercise 7.2** $^\diamond$ Show that the following function from $\mathbb{R}^s$ to $\mathbb{R}$

$$\theta \;\mapsto\; \|\boldsymbol{y} - X_{\mathcal{S}}\theta\|_2^2$$

1. is convex and
2. is strictly convex if and only if $X_{\mathcal{S}}^\top X_{\mathcal{S}}$ is invertible.

□ **Exercise 7.3** $^\diamond$ Consider the function h on Page 130. Show that h is convex if and only if both $h_{\mathcal{S}}$ and $h_{\mathcal{S}^\complement}$ are convex.

□ **Exercise 7.4** $^{\diamond\bullet}$ Consider the group lasso penalty $h : \boldsymbol{\alpha} \mapsto \sum_{i=1}^k \|\boldsymbol{\alpha}_{\mathcal{A}_i}\|_2$ for a number $k \in \{1, \dots, p\}$ and index sets $\mathcal{A}_1, \dots, \mathcal{A}_k \subset \{1, \dots, p\}$.

1. Show that the lasso penalty $\boldsymbol{\alpha} \mapsto \|\boldsymbol{\alpha}\|_1$ is a special case of the group lasso penalty.
2. Show that we can write h in the form

$$h(\boldsymbol{\alpha}) \;=\; h_{\mathcal{S}}(\boldsymbol{\alpha}_{\mathcal{S}}) + h_{\mathcal{S}^\complement}(\boldsymbol{\alpha}_{\mathcal{S}^\complement})$$

if an only if for all $i \in \{1, \dots, p\}$, it holds that $\mathcal{A}_i \subset \mathcal{S}$ or $\mathcal{A}_i \subset \mathcal{S}^\complement$.

□ **Exercise 7.5** $^{\diamond\diamond\bullet}$ We study Assumption 7.3.1 on Page 133 with h the ℓ_1-norm.

1. Assume that $g : \mathbb{R}^k \to [0, \infty)$ is convex and satisfies the conditions of Hölder's inequality B.1.3. Then, $\overline{g}(\mathbf{a}) \le 1$ for all $\mathbf{a} \in \partial g(\mathbf{b})$, $\mathbf{b} \in \mathbb{R}^k$.
2. Use 1. to show that for $h_{\mathcal{S}^\complement}(\cdot) \equiv \|\cdot\|_1$,

$$\sup_{\mathbf{m} \in \partial h[\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}]} \overline{h}_{\mathcal{S}^\complement}(X_{\mathcal{S}^\complement}^\top X_{\mathcal{S}}(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}\mathbf{m}) \;\le\; \max_{j \in \mathcal{S}^\complement} \|(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1} X_{\mathcal{S}}^\top x_j\|_1\,,$$

where we write $X = (x_1 \dots x_p)$.

Note: $(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1} X_{\mathcal{S}}^\top x_j$ is the least-squares estimator for linear regression with outcome x_j and design matrix $X_{\mathcal{S}}$. So we measure the correlation between the columns outside $\mathcal{S}$ with the columns inside $\mathcal{S}$.

□ **Exercise 7.6** $^{\diamond\diamond\diamond}$ Let $\mathcal{B} := \{\boldsymbol{\alpha} \in \mathbb{R}^p \;:\; X\boldsymbol{\alpha} = \mu\}$ for given $X \in \mathbb{R}^p$ and $\mu \in \mathbb{R}^n$. Let $\boldsymbol{\beta}$ such that $|\operatorname{supp}(\boldsymbol{\beta})| = \min\{\operatorname{supp}(\boldsymbol{\alpha}) \;:\; \boldsymbol{\alpha} \in \mathcal{B}\}$. Prove or find a counter example to $\|\boldsymbol{\beta}\|_1 = \min\{\|\boldsymbol{\alpha}\|_1 \;:\; \boldsymbol{\alpha} \in \mathcal{B}\}$.

□ **Exercise 7.7** $^{\diamond\diamond\diamond\bullet}$ We study the uniqueness of the vectors $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\beta}}$ of Section 7.3.

1. Show that the vector $\widehat{\boldsymbol{\gamma}}$ is unique if the restricted Gram matrix $X_{\mathcal{S}}^\top X_{\mathcal{S}}$ is invertible.

2. Does this mean that also $\widehat{\boldsymbol{\beta}}$ is unique if the restricted Gram matrix $X_{\mathcal{S}}^{\top}X_{\mathcal{S}}$ is invertible? What if additionally, the primal-dual witness construction is successful?

3. Assume finally that $X_{\mathcal{S}}^{\top}X_{\mathcal{S}}$ is invertible, $h(\cdot) = \|\cdot\|_1$, and $\|\widehat{\boldsymbol{\nu}}_{\mathcal{S}^{\complement}}\|_\infty < 1$. Is the estimator $\widehat{\boldsymbol{\beta}}$ now unique? *Hint:* Use Lemma 7.3.2 or prove a special case of it.

Appendix A

Solutions

A.1 Solutions for Chapter 1

Solutions for Section 1.1

Solution 1.1 See the R file on the book homepage.

Solutions for Section 1.2

Solution 1.2 We prove the claims in order.

1. The proof of the first claim follows from basic linear algebra.

Write $X = (\boldsymbol{x}^1, \dots, \boldsymbol{x}^p)$ with columns $\boldsymbol{x}^1, \dots, \boldsymbol{x}^p \in \mathbb{R}^n$. If $p > n$, we know by linear algebra that the columns must be linearly dependent. For example, there are $a_1, \dots, a_{p-1} \in \mathbb{R}$ such that $\boldsymbol{x}^p = \sum_{j=1}^{p-1} a_j \boldsymbol{x}^j$. Then, for $\mathbf{w} := (a_1, \dots, a_{p-1}, -1)^\top \in \mathbb{R}^p$, it holds that $\mathbf{w} \neq \mathbf{0}_n$ and

$$X\mathbf{w} \;=\; \sum_{j=1}^{p} w_j \boldsymbol{x}^j \;=\; \sum_{j=1}^{p-1} w_j \boldsymbol{x}^j + w_p \boldsymbol{x}^p \;=\; \sum_{j=1}^{p-1} a_j \boldsymbol{x}^j - \boldsymbol{x}^p \;=\; \mathbf{0}_n\,.$$

Hence,

$$\begin{aligned}
\mathbf{w}^\top X^\top X\mathbf{w} &= (X\mathbf{w})^\top X\mathbf{w} && \text{“}(AB)^\top = B^\top A^\top\text{”}\\
&= \|X\mathbf{w}\|_2^2 && \text{“definition of the } \ell_2\text{-norm”}\\
&= \|\mathbf{0}_n\|_2^2 && \text{“above display”}\\
&= 0 && \text{“positive definitness of norms”}
\end{aligned}$$

for $\mathbf{w} \neq \mathbf{0}_n$, which means that $X^\top X$ is not invertible.

2. We again resort to linear algebra.

$\Rightarrow$: Assume that $X\boldsymbol{\gamma} = \mathbf{0}_n$ for a $\boldsymbol{\gamma} \neq \mathbf{0}_p$. Then, for any $\widehat{\boldsymbol{\beta}}_{\text{ls}} \in \operatorname{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2$, also $\widehat{\boldsymbol{\beta}}_{\text{ls}} + \boldsymbol{\gamma} \in \operatorname{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2$. Hence, $\widehat{\boldsymbol{\beta}}_{\text{ls}}$ is not unique.

$\Leftarrow$: Assume that $\operatorname{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2$ is not unique. Then, there are least-squares estimators $\widehat{\boldsymbol{\gamma}}_{\text{ls}}, \widehat{\boldsymbol{\beta}}_{\text{ls}}$ such that $\widehat{\boldsymbol{\gamma}}_{\text{ls}} \neq \widehat{\boldsymbol{\beta}}_{\text{ls}}$ but

$$\|\boldsymbol{y} - X\widehat{\boldsymbol{\gamma}}_{\text{ls}}\|_2^2 \;=\; \|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{\text{ls}}\|_2^2\,.$$

However, since $\boldsymbol{a} \mapsto \|\boldsymbol{y} - \boldsymbol{a}\|_2^2$ is *strictly* convex, this means that $X\widehat{\boldsymbol{\gamma}}_{\text{ls}} = X\widehat{\boldsymbol{\beta}}_{\text{ls}}$, which in turn means that $X\boldsymbol{\gamma} = \mathbf{0}_n$ for $\boldsymbol{\gamma} := \widehat{\boldsymbol{\gamma}}_{\text{ls}} - \widehat{\boldsymbol{\beta}}_{\text{ls}} \neq \mathbf{0}_p$.

The stated conclusion now follows from the fact that $X\boldsymbol{\gamma} \neq \mathbf{0}_n$ for all $\boldsymbol{\gamma} \in \mathbb{R}^p \setminus \{\mathbf{0}_p\}$ if and only if $X^\top X$ invertible.
3. We use again that the function $\mathbf{a} \mapsto \|\boldsymbol{y} - \mathbf{a}\|_2^2$ is *strictly* convex.

Assume that $X\widehat{\boldsymbol{\gamma}}_{\mathrm{ls}} \neq X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$. Then, for any $a \in (0,1)$,

$$\begin{aligned}
&\|\boldsymbol{y} - X\big(a\widehat{\boldsymbol{\gamma}}_{\mathrm{ls}} + (1-a)\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\big)\|_2^2\\
&= \|\boldsymbol{y} - aX\widehat{\boldsymbol{\gamma}}_{\mathrm{ls}} + (1-a)X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2 && \text{“multiplying out the design-related part”}\\
&< a\|\boldsymbol{y} - X\widehat{\boldsymbol{\gamma}}_{\mathrm{ls}}\|_2^2 + (1-a)\|\boldsymbol{y} + X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2\\
&\qquad \text{“strict convexity of the mentioned function; use the function values } \boldsymbol{a} = aX\widehat{\boldsymbol{\gamma}}_{\mathrm{ls}} \text{ and } \boldsymbol{a} = (1-a)X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\text{, respecitvely”}\\
&= a\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2 + (1-a)\|\boldsymbol{y} + X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2 && \text{“both } \widehat{\boldsymbol{\gamma}}_{\mathrm{ls}} \text{ and } \widehat{\boldsymbol{\beta}}_{\mathrm{ls}} \text{ are least-squares solutions”}\\
&= \|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2\,, && \text{“consolidating terms”}
\end{aligned}$$

which means that $a\widehat{\boldsymbol{\gamma}}_{\mathrm{ls}} + (1-a)\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$ has strictly smaller value in objective function than $\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$. This contradicts that $\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}$ is a least-squares solution.

4. The proof is short: we just need to use the properties of the Moore-Penrose inverse.

According to Section 1.2, we have to show that $X^\top X(X^+\boldsymbol{y}) = X^\top\boldsymbol{y}$. For this, we compute

$$\begin{aligned}
X^\top X(X^+\boldsymbol{y}) &= X^\top(XX^+)^\top\boldsymbol{y} && \text{“Part 3 in Definition B.2.2 (Moore-Penrose)”}\\
&= (XX^+X)^\top\boldsymbol{y} && \text{“properties of transposes”}\\
&= X^\top\boldsymbol{y}\,, && \text{“Part 1 in Definition B.2.2 (Moore-Penrose)”}
\end{aligned}$$

as desired.

Alternatively, we could show that for all $\boldsymbol{\alpha} \in \mathbb{R}^p$,

$$\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 \;\geq\; \|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2\,.$$

For this, we first observe that

$$\begin{aligned}
&2\langle(\mathrm{I}_{n\times n} - XX^+)\boldsymbol{y},\, X(X^+\boldsymbol{y} - \boldsymbol{\alpha})\rangle\\
&= 2\langle\boldsymbol{y},\, (\mathrm{I}_{n\times n} - XX^+)^\top X(X^+\boldsymbol{y} - \boldsymbol{\alpha})\rangle && \text{“properties of transposes”}\\
&= 2\langle\boldsymbol{y},\, (\mathrm{I}_{n\times n} - XX^+)X(X^+\boldsymbol{y} - \boldsymbol{\alpha})\rangle && \text{“3. in Definition B.2.2 (Moore-Penrose)”}\\
&= 2\langle\boldsymbol{y},\, (X - XX^+X)(X^+\boldsymbol{y} - \boldsymbol{\alpha})\rangle && \text{“reorganizing”}\\
&= 0\,. && \text{“1. in Definition B.2.2 (Moore-Penrose)”}
\end{aligned}$$

Using this, we find for any $\boldsymbol{\alpha} \in \mathbb{R}^p$

$$\begin{aligned}
&\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2\\
&= \|(\mathrm{I}_{n\times n} - XX^+)\boldsymbol{y} + XX^+\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 && \text{“adding a zero-valued term”}\\
&= \|(\mathrm{I}_{n\times n} - XX^+)\boldsymbol{y} + X(X^+\boldsymbol{y} - \boldsymbol{\alpha})\|_2^2 && \text{“factorizing } X\text{”}\\
&= \|(\mathrm{I}_{n\times n} - XX^+)\boldsymbol{y}\|_2^2 + \|X(X^+\boldsymbol{y} - \boldsymbol{\alpha})\|_2^2 + 2\langle(\mathrm{I}_{n\times n} - XX^+)\boldsymbol{y},\, X(X^+\boldsymbol{y} - \boldsymbol{\alpha})\rangle\\
&&& \text{“expanding the square”}\\
&= \|(\mathrm{I}_{n\times n} - XX^+)\boldsymbol{y}\|_2^2 + \|X(X^+\boldsymbol{y} - \boldsymbol{\alpha})\|_2^2 && \text{“previous display”}\\
&= \|\boldsymbol{y} - XX^+\boldsymbol{y}\|_2^2 + \|XX^+\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 && \text{“expanding the bracket in the second term”}\\
&= \|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2 + \|X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}} - X\boldsymbol{\alpha}\|_2^2 && \text{“definition of } \widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\text{”}\\
&\geq \|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2\,, && \text{“positive definitness of norms”}
\end{aligned}$$

as desired.
5. Each of the three claims is a brief exercise in linear algebra.

(i) One needs to verify that X^+ satisfies the four properties of Definition B.2.2 (Moore-Penrose). The first property can be derived as follows:

$$\begin{aligned}
&(DD^+D)_{ij}\\
&= \sum_{k=1}^{p} D_{ik} \sum_{l=1}^{n} D^+_{kl} D_{lj} && \text{“explicit form of matrix-matrix multiplications”}\\
&= D_{ii} D^+_{ij} D_{jj} && \text{“}D\text{ diagonal by assumption”}\\
&= \begin{cases} D_{ii} D^+_{ii} D_{ii} = D_{ii} & \text{for } i = j\\ 0 & \text{for } i \neq j \end{cases} && \text{“definition of }D^+\text{”}\\
&= D_{ij}\,. && \text{“}D\text{ diagonal by assumption”}
\end{aligned}$$

The remaining three properties are proved similarly.
(ii) Observe first that

$$\begin{aligned}
&(DD^+)_{ij}\\
&= \sum_{k=1}^{p} D_{ik} D^+_{kj} && \text{“explicit form of matrix-matrix multiplication”}\\
&= D_{ii} D^+_{ij} && \text{“}D\text{ diagonal”}\\
&= \begin{cases} D_{ii} D^+_{ii} = 1 & \text{for } i = j,\ D_{ii} \neq 0\\ D_{ii} D^+_{ii} = 0 & \text{for } i = j,\ D_{ii} = 0\\ 0 & \text{for } i \neq j \end{cases}. && \text{“definition of }D^+\text{”}
\end{aligned}$$

Therefore, DD^+ is diagonal with zero and ones as entries, and the number of ones is $|\{i \in \{1, \dots, \min\{n, p\}\} : D_{ii} \neq 0\}| = \operatorname{rank}[D]$. Since U, V orthogonal, $\operatorname{rank}[D] = \operatorname{rank}[X]$, from which we can conclude the claim.
(iii) We need to verify that X^+ satisfies the four properties of Definition B.2.2 (Moore-Penrose). The first property can be derived as follows:

$$\begin{aligned}
&XX^+X\\
&= UDV^\top(VD^+U^\top)UDV && \text{“SVD and definition of }X^+\text{”}\\
&= UDD^+DV && \text{“}U, V\text{ orthogonal by assumption”}\\
&= UDV && \text{“}D^+\text{ is a Moore-Penrose inverse of }D\text{ by assumption”}\\
&= X\,. && \text{“SVD”}
\end{aligned}$$

The other three properties can be verified similarly.
6. The short prove is based on the previous findings.

We first note that in view of Claim 3, it is sufficient to show the equality for *any* least-squares solution. We opt for $\widehat{\boldsymbol{\beta}}_{\text{ls}} = X^+\boldsymbol{y}$ motivated by Claim 4. We then find

$$\begin{aligned}
&\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}_{\text{ls}}\|_2^2\\
&= \|X\boldsymbol{\beta} - XX^+\boldsymbol{y}\|_2^2 && \text{“our choice of }\widehat{\boldsymbol{\beta}}_{\text{ls}}\text{”}\\
&= \|X\boldsymbol{\beta} - XX^+(X\boldsymbol{\beta} + \boldsymbol{u})\|_2^2 && \text{“model assumptions: }\boldsymbol{y} = X\boldsymbol{\alpha} + \boldsymbol{u}\text{”}\\
&= \|XX^+\boldsymbol{u}\|_2^2 && \text{“1. in Definition B.2.2 (Moore-Penrose) and consolidating”}\\
&= \|UDV^\top VD^+U^\top\boldsymbol{u}\|_2^2 && \text{“SVD and 5.(iii) above”}\\
&= \|UDD^+U^\top\boldsymbol{u}\|_2^2\,, && \text{“}V\text{ orthogonal”}
\end{aligned}$$

as desired.
7. After some simple calculations, the claim follows from the above results and the properties of Chi-squared distributions.

We first rewrite the right-hand side of the display in 6. as

$$
\begin{aligned}
&\|UDD^{+}U^{\top}\boldsymbol{u}\|_2^2\\
=\ &(UDD^{+}U^{\top}\boldsymbol{u})^{\top}UDD^{+}U^{\top}\boldsymbol{u} && \text{“definition of the } \ell_2\text{-norm”}\\
=\ &(U^{\top}\boldsymbol{u})^{\top}(DD^{+})^{\top}U^{\top}UDD^{+}U^{\top}\boldsymbol{u} && \text{“properties of transposes”}\\
=\ &(U^{\top}\boldsymbol{u})^{\top}(DD^{+})^{\top}DD^{+}U^{\top}\boldsymbol{u} && \text{“}U\text{ orthogonal”}\\
=\ &(U^{\top}\boldsymbol{u})^{\top}DD^{+}DD^{+}U^{\top}\boldsymbol{u} && \text{“5.(i) and 3. in Definition B.2.2 (Moore-Penrose)”}\\
=\ &(U^{\top}\boldsymbol{u})^{\top}DD^{+}U^{\top}\boldsymbol{u}\,. && \text{“4.(i) and 1. or 2. in Definition B.2.2 (Moore-Penrose)”}
\end{aligned}
$$

Setting $\boldsymbol{\gamma} := U^{\top}\boldsymbol{u}/\sigma$ and combining with 6. yields

$$\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2 \sim \sigma^2\boldsymbol{\gamma}^{\top}DD^{+}\boldsymbol{\gamma}\,,$$

and thus, dividing both sides by n,

$$\frac{\|X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2}{n} \sim \frac{\sigma^2\boldsymbol{\gamma}^{\top}DD^{+}\boldsymbol{\gamma}}{n}\,.$$

Since U is orthogonal, $\boldsymbol{\gamma} \sim \mathcal{N}_n[\mathbf{0}_n, \mathrm{I}_{n\times n}]$. Moreover, according to 4.(ii), DD^{+} has $\operatorname{rank}[X]$ ones on its diagonal and zeros everywhere else. These two observations imply that $\boldsymbol{\gamma}^{\top}DD^{+}\boldsymbol{\gamma} \sim \chi^2_{\operatorname{rank}[X]}$. The claim now follows from the fact that the expectation of a (standard) Chi-squared random distribution equals its degrees of freedom.
8. We confirm the properties of Moore-Penrose matrices.

All four properties in Definition B.2.2 can be approached similarly, so that we focus only on 1.:

$$X(X^{\top}X)^{+}X^{\top}X \;=\; X\,.$$

For this, we consider first general $A \in \mathbb{R}^{p\times p}, B \in \mathbb{R}^{n\times p}$ that satisfy $B^{\top}BA = \mathbf{0}_{p\times p}$. Then,

$$
\begin{aligned}
&& B^{\top}BA &= \mathbf{0}_{p\times p}\\
&\Rightarrow & A^{\top}B^{\top}BA &= \mathbf{0}_{p\times p}\\
&&&\quad \text{“multiplying both sides of the assumed equality by } A^{\top} \text{ from the left”}\\
&\Rightarrow & (BA)^{\top}(BA) &= \mathbf{0}_{p\times p} && \text{“properties of transpose”}\\
&\Rightarrow & \big((BA)^{\top}(BA)\big)_{jj} &= 0 \quad \text{for all } j\in\{1,\dots,p\}\\
&&&\quad \text{“definition of matrix-matrix multiplication”}\\
&\Rightarrow & \sum_{i=1}^{n}((BA)^{\top})_{ji}(BA)_{ij} &= 0 \quad \text{for all } j\in\{1,\dots,p\}\\
&&&\quad \text{“definition of matrix-matrix multiplication”}\\
&\Rightarrow & \sum_{i=1}^{n}((BA)_{ij})^2 &= 0 \quad \text{for all } j\in\{1,\dots,p\} && \text{“properties of transposes”}\\
&\Rightarrow & (BA)_{ij} &= 0 \quad \text{for all } i\in\{1,\dots,n\}, j\in\{1,\dots,p\}\\
&&&\quad \text{“squared-terms are non-negative”}\\
&\Rightarrow & BA &= \mathbf{0}_{n\times p}\,. && \text{“definition of matrices”}
\end{aligned}
$$

Now, we use the definition of Moore-Penrose inverses for $(X^\top X)^+$ and the above display with $A = (XX^\top)^+X^\top - \mathrm{I}_{p\times p}$ and $B = X^\top$ to find

$$\begin{aligned} X^\top X(X^\top X)^+X^\top X &= X^\top X \\ &\text{“1. in Definition B.2.2 (Moore-Penrose) applied to } (X^\top X)^+\text{”} \\ \Rightarrow\quad X^\top X\bigl((X^\top X)^+X^\top X - \mathrm{I}_{p\times p}\bigr) &= \mathbf{0}_{n\times n} \qquad \text{“factorizing } X^\top X\text{”} \\ \Rightarrow\quad X\bigl((X^\top X)^+X^\top X - \mathrm{I}_{p\times p}\bigr) &= \mathbf{0}_{n\times p} \qquad \text{“previous display”} \\ \Rightarrow\quad X(X^\top X)^+X^\top X &= X\,, \\ &\text{“multiplying the bracket out and rearranging terms”} \end{aligned}$$

as desired.

Solutions for Section 1.3

Solution 1.3 The proofs are straightforward calculations.

1. By assumption, $f_{\boldsymbol{\alpha}}$ is the density of $\mathcal{N}_n[X\boldsymbol{\alpha}, \sigma^2\,\mathrm{I}_{n\times n}]$ and g the density of $\mathcal{N}_p[\mathbf{0}_p, \tau^2\,\mathrm{I}_{p\times p}]$, that is,

$$f_{\boldsymbol{\alpha}}[\boldsymbol{y}] \;=\; \frac{1}{(2\pi\sigma^2)^{n/2}}e^{-\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2/(2\sigma^2)}$$

and

$$g[\boldsymbol{\alpha}] \;=\; \frac{1}{(2\pi\tau^2)^{p/2}}e^{-\|\boldsymbol{\alpha}\|_2^2/(2\tau^2)}\,.$$

Using these formulae and the properties of the logarithm yields

$$\begin{aligned} \log\bigl[f_{\boldsymbol{\alpha}}[\boldsymbol{y}]g[\boldsymbol{\alpha}]\bigr] &= \log\Biggl[\frac{1}{(2\pi\sigma^2)^{n/2}}e^{-\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2/(2\sigma^2)}\frac{1}{(2\pi\tau^2)^{p/2}}e^{-\|\boldsymbol{\alpha}\|_2^2/(2\tau^2)}\Biggr] \\ &= -\frac{n}{2}\log[2\pi\sigma^2] - \frac{1}{2\sigma^2}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2 - \frac{p}{2}\log[2\pi\tau^2] - \frac{1}{2\tau^2}\|\boldsymbol{\alpha}\|_2^2\,. \end{aligned}$$

Then,

$$\begin{aligned} &\operatorname*{argmax}_{\boldsymbol{\alpha}\in\mathbb{R}^p} l[\boldsymbol{\alpha}|\boldsymbol{y}] \\ &= \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\Bigl\{\frac{n}{2}\log[2\pi\sigma^2] + \frac{1}{2\sigma^2}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2 + \frac{p}{2}\log[2\pi\tau^2] + \frac{1}{2\tau^2}\|\boldsymbol{\alpha}\|_2^2\Bigr\} \\ &\hspace{20em}\text{“hint and above display”} \\ &= \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\Bigl\{\frac{1}{2\sigma^2}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2 + \frac{1}{2\tau^2}\|\boldsymbol{\alpha}\|_2^2\Bigr\} \qquad \text{“omitting terms that do not depend on } \boldsymbol{\alpha}\text{”} \\ &= \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\Bigl\{\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2 + \frac{\sigma^2}{\tau^2}\|\boldsymbol{\alpha}\|_2^2\Bigr\}\,. \\ &\qquad\text{“multiplying through by } 2\sigma^2 \text{ and noting that this does not affect any minimizer”} \end{aligned}$$

Hence, $\widehat{\boldsymbol{\beta}}_{\mathrm{map}} = \widehat{\boldsymbol{\beta}}_{\mathrm{ridge}}$ for $r = \sigma^2/\tau^2$, as desired.

2. Now,

$$g[\boldsymbol{\alpha}] \;=\; \frac{1}{(2\tau)^p}e^{-\|\boldsymbol{\alpha}\|_1/\tau}\,,$$

but we can proceed as above otherwise. In particular, we find

$$\begin{aligned} \log\bigl[f_{\boldsymbol{\alpha}}[\boldsymbol{y}]g[\boldsymbol{\alpha}]\bigr] &= \log\Biggl[\frac{1}{(2\pi\sigma^2)^{n/2}}e^{-\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2/(2\sigma^2)}\frac{1}{(2\tau)^p}e^{-\|\boldsymbol{\alpha}\|_1/\tau}\Biggr] \\ &= -\frac{n}{2}\log[2\pi\sigma^2] - \frac{1}{2\sigma^2}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2 - p\log[2\tau] - \frac{1}{\tau}\|\boldsymbol{\alpha}\|_1 \end{aligned}$$

and

$$\begin{aligned}
&\operatorname*{argmax}_{\boldsymbol{\alpha}\in\mathbb{R}^p} l[\boldsymbol{\alpha}|\boldsymbol{y}]\\
&= \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\Bigl\{\frac{n}{2}\log[2\pi\sigma^2]+\frac{1}{2\sigma^2}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2+p\log[2\tau]+\frac{1}{\tau}\|\boldsymbol{\alpha}\|_1\Bigr\}\\
&= \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\Bigl\{\frac{1}{2\sigma^2}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2+\frac{1}{\tau}\|\boldsymbol{\alpha}\|_1\Bigr\}\\
&= \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\Bigl\{\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2+\frac{2\sigma^2}{\tau}\|\boldsymbol{\alpha}\|_1\Bigr\}
\end{aligned}$$

by following the same lines as above. Hence, $\widehat{\boldsymbol{\beta}}_{\text{map}}=\widehat{\boldsymbol{\beta}}_{\text{lasso}}$ for $r=2\sigma^2/\tau$, as desired.

3. The proof follows the exact same scheme as above.

Solutions for Section 1.4

Solution 1.4 The solutions are simple applications of the definitions of convexity and strict convexity.

1. This follows readily from the properties of smooth, strictly convex functions.

A smooth function $\mathbb{R}^p\to\mathbb{R}$ is strictly convex if and only if its Hessian matrix is positive definite everywhere. In our case, $\boldsymbol{\alpha}\mapsto\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2$ is indeed smooth, and its Hessian matrix is

$$\nabla^2\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2 \;=\; 2X^\top X\,.$$

Hence, the function in question is strictly convex if and only if $X^\top X$ is positive definite. Since $X^\top X$ is symmetric, and positive definiteness and invertibility are equivalent for symmetric matrices, this yields the desired claim.

2. There is a plethora of such functions, we just look at two simple examples.

Considering functions $\mathbb{R}^p\to\mathbb{R}$, a simple example for the first case is the constant function $\boldsymbol{\alpha}\mapsto 0$, for which every point is a minimum; a simple example for the second case is $\boldsymbol{\alpha}\mapsto\|\boldsymbol{\alpha}\|_1$, for which $\mathbf{0}_p$ is the only minimum.

For the bonus question, the answer is "no." Indeed, naming the function in question f and $\boldsymbol{a}:=(1,0)^\top$, $\boldsymbol{b}:=(0,1)^\top$ and assuming that f is convex, we find

$$\begin{aligned}
1 &= f\bigl[(1,1)^\top\bigr] && \text{"definition of } f\text{"}\\
&= f\bigl[(2,0)^\top/2+(0,2)^\top/2\bigr] && \text{"basic vector algebra"}\\
&\le f\bigl[(2,0)^\top\bigr]/2+f\bigl[(0,2)^\top\bigr]/2 && \text{"assumed convexity"}\\
&= 0/2+0/2 && \text{"definition of } f\text{"}\\
&= 0\,, && \text{"consolidating"}
\end{aligned}$$

which is a contradiction.

3. We prove this by a simple contradiction.

Assume that $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ are two distinct minima of f. The presumed strict convexity implies for every intermediate point $a\boldsymbol{\alpha}+(1-a)\boldsymbol{\alpha}'$, $a\in(0,1)$, that

$$f\bigl[a\boldsymbol{\alpha}+(1-a)\boldsymbol{\alpha}'\bigr] \;<\; af[\boldsymbol{\alpha}]+(1-a)f[\boldsymbol{\alpha}']\,.$$

Since both $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ are minima of f, it holds that $f[\boldsymbol{\alpha}]=f[\boldsymbol{\alpha}']$. Plugging this into the display and collecting terms yields

$$f\bigl[a\boldsymbol{\alpha}+(1-a)\boldsymbol{\alpha}'\bigr] \;<\; af[\boldsymbol{\alpha}]+(1-a)f[\boldsymbol{\alpha}] \;=\; f[\boldsymbol{\alpha}]\,,$$

which means that $\boldsymbol{\alpha}$ is not a minimum of f as assumed initially. This is the desired contradiction.

Solution 1.6 The proof follows readily from the explicit form of the ridge estimator.

Section 1.4 provides us with

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}}[r] \;=\; \frac{1}{4-8d+4d^2+(9+d^2)r+r^2}\begin{pmatrix}4-8d+4d^2+5r\\(4+d)r\end{pmatrix}$$

for all $r\in(0,\infty)$ and $d\in\mathbb{R}$, and therefore, for $d=1$,

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}}[r] \;=\; \frac{1}{10r+r^2}\begin{pmatrix}5r\\5r\end{pmatrix} \;=\; \frac{1}{1+r/10}\begin{pmatrix}1/2\\1/2\end{pmatrix}\,.$$

From here, a straightforward calculation gives

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}_{\text{ridge}}[r]-(1/2,1/2)^\top\|_2 \;&=\; \sqrt{\Bigl(\frac{1/2}{1+r/10}-\frac{1}{2}\Bigr)^2+\Bigl(\frac{1/2}{1+r/10}-\frac{1}{2}\Bigr)^2}\\ &=\; \sqrt{\frac{1}{2}\Bigl(\frac{1}{1+r/10}-1\Bigr)^2}\\ &=\; \sqrt{\frac{1}{2}\Bigl(\frac{r/10}{1+r/10}\Bigr)^2}\\ &=\; \sqrt{\frac{1}{2}\Bigl(\frac{r}{10+r}\Bigr)^2}\\ &=\; \frac{r}{\sqrt{2}(10+r)}\,,\end{aligned}$$

as desired.

This result and the fact that $r/(\sqrt{2}(10+r))$ continuously approaches 0 as $r\to 0$ ensure that the ridge estimator continuously approaches the least-squares solution $(1-a,a)^\top$ with $a=1/2$ as $r\to 0$.

Solution 1.7 We prove the eleven claims in order.

1. The proof of the first claim simply consists of plugging in the values for $\boldsymbol{y}$ and X and some reformulations.

We first derive

$$\begin{aligned}&\boldsymbol{y}-X\boldsymbol{\alpha}\\ &=\begin{pmatrix}1\\2\end{pmatrix}-\begin{pmatrix}1&d\\2&2\end{pmatrix}\begin{pmatrix}\alpha_1\\\alpha_2\end{pmatrix} && \text{“definitions of } \boldsymbol{y} \text{ and } X \text{ in Table 1.3”}\\ &=\begin{pmatrix}1\\2\end{pmatrix}-\begin{pmatrix}\alpha_1+d\alpha_2\\2\alpha_1+2\alpha_2\end{pmatrix} && \text{“matrix-vector multiplication”}\\ &=\begin{pmatrix}1-\alpha_1-d\alpha_2\\2-2\alpha_1-2\alpha_2\end{pmatrix} && \text{“vector-vector addition”}\\ &=-\begin{pmatrix}\alpha_1-1+d\alpha_2\\2(\alpha_1-1)+2\alpha_2\end{pmatrix}\,. && \text{“rearraning the quantities”}\end{aligned}$$

Hence,

$$\begin{aligned}&\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2\\ &=\;\bigl(\alpha_1-1+d\alpha_2\bigr)^2+\bigl(2(\alpha_1-1)+2\alpha_2\bigr)^2\\ &\qquad\qquad \text{“invoking the above display and the definition of } \ell_2\text{-norms”}\\ &=\;5(\alpha_1-1)^2+(4+d^2)\alpha_2^2+2(4+d)(\alpha_1-1)\alpha_2\,,\\ &\qquad\qquad \text{“expanding and summarizing terms”}\end{aligned}$$

as desired.

2. This proof follows from 1. by plugging in d.
   Indeed, we find

$$\begin{aligned}
&\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2\\
&= 5(\alpha_1-1)^2+(4+d^2)\alpha_2^2+2(4+d)(\alpha_1-1)\alpha_2 && \text{“1. just above”}\\
&= 5\bigl((\alpha_1-1)^2+\alpha_2^2+2(\alpha_1-1)\alpha_2\bigr) && \text{“using that } d=1 \text{ and summarizing terms”}\\
&= 5(\alpha_1-1+\alpha_2)^2\,, && \text{“binomial theorem”}
\end{aligned}$$

as desired.

3. The proof follows from rewriting the terms in 2.
   We find

$$\begin{aligned}
\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2 &= c\\
\Leftrightarrow\quad 5(\alpha_1-1+\alpha_2)^2 &= c && \text{“2. just above”}\\
\Leftrightarrow\quad \alpha_1-1+\alpha_2 &= \pm\sqrt{\frac{c}{5}} && \text{“dividing both sides by 5 and taking square-roots”}\\
\Leftrightarrow\quad \alpha_2 &= 1-\alpha_1\pm\sqrt{\frac{c}{5}}\,, && \text{“rearranging terms”}
\end{aligned}$$

as desired.

4. The proof consists of plugging in the values and taking limits.
   We find

$$\begin{aligned}
&\frac{\bigl((\cos\omega)(\alpha_1-1)+(\sin\omega)\alpha_2\bigr)^2}{a^2}+\frac{\bigl(-(\sin\omega)(\alpha_1-1)+(\cos\omega)\alpha_2\bigr)^2}{b^2}\\
&\to 10\bigl((\cos 45^\circ)(\alpha_1-1)+(\sin 45^\circ)\alpha_2\bigr)^2+\frac{\bigl(-(\sin 45^\circ)(\alpha_1-1)+(\cos 45^\circ)\alpha_2\bigr)^2}{b^2}\\
&&& \text{“plugging in the values for } \omega \text{ and } a\text{”}\\
&= \frac{10(\alpha_1-1+\alpha_2)^2}{2}+\frac{(\alpha_1-1+\alpha_2)^2}{2b^2} && \text{“}\sin 45^\circ=\cos 45^\circ=1/\sqrt{2}\text{”}\\
&\to 5(\alpha_1-1+\alpha_2)^2\,. && \text{“consolidating and taking the limit } b\to\infty\text{”}
\end{aligned}$$

By the comment in the exercise, this means that the level sets converge to the specified target. We conclude with 2. that the straight lines that form the level sets of the least-squares can be interpreted as degenerate ellipses.

5. Just expand the square-terms and summarize the factors.
6. This is a simple algebra exercise.
   We find

$$\begin{aligned}
\frac{(\cos\omega)^2}{a^2}+\frac{(\sin\omega)^2}{b^2} &= 5 && \text{“comparing the } (\alpha_1-1)\text{-terms”}\\
\Rightarrow\quad \frac{(\cos\omega)^2}{a^2}+\frac{1-(\cos\omega)^2}{b^2} &= 5 && \text{“}(\sin\omega)^2+(\cos\omega)^2=1\text{”}\\
\Rightarrow\quad (\cos\omega)^2\Bigl(\frac{1}{a^2}-\frac{1}{b^2}\Bigr) &= 5-\frac{1}{b^2} && \text{“factorizing and rearranging terms”}\\
\Rightarrow\quad (\cos\omega)^2 &= \frac{5-\frac{1}{b^2}}{\frac{1}{a^2}-\frac{1}{b^2}}\,,\\
&&& \text{“dividing both sides by } 1/a^2-1/b^2 \text{ under the assumption } a\neq b\text{”}
\end{aligned}$$

as desired.

7. This works as above.

We find

$$\begin{aligned}
&& \frac{(\sin\omega)^2}{a^2}+\frac{(\cos\omega)^2}{b^2} &= 4+d^2 && \text{“comparing the α_2-terms”}\\
\Rightarrow && \frac{1-(\cos\omega)^2}{a^2}+\frac{(\cos\omega)^2}{b^2} &= 4+d^2 && \text{“$(\sin\omega)^2+(\cos\omega)^2=1$”}\\
\Rightarrow && (\cos\omega)^2\Bigl(\frac{1}{b^2}-\frac{1}{a^2}\Bigr)+\frac{1}{a^2} &= 4+d^2 && \text{“factorizing terms”}\\
\Rightarrow && \frac{5-\frac{1}{b^2}}{\frac{1}{a^2}-\frac{1}{b^2}}\Bigl(\frac{1}{b^2}-\frac{1}{a^2}\Bigr)+\frac{1}{a^2} &= 4+d^2 && \text{“Claim 6”}\\
\Rightarrow && \frac{1}{b^2}-5+\frac{1}{a^2} &= 4+d^2 && \text{“simplifying the left-hand side”}\\
\Rightarrow && \frac{1}{a^2}+\frac{1}{b^2} &= 9+d^2\,, && \text{“consolidating”}
\end{aligned}$$

as desired.

8. The proof combines the results from the two previous proof steps.

We first observe that

$$\begin{aligned}
(\cos\omega)^2 &= \frac{5-\frac{1}{b^2}}{\frac{1}{a^2}-\frac{1}{b^2}} && \text{“Claim 6”}\\
&= \frac{5-\frac{1}{b^2}}{\frac{1}{a^2}+\frac{1}{b^2}-\frac{2}{b^2}} && \text{“adding a zero-valued term”}\\
&= \frac{5-\frac{1}{b^2}}{9+d^2-\frac{2}{b^2}} && \text{“Claim 7”}\\
&= \frac{5b^2-1}{(9+d^2)b^2-2}\,, && \text{“multiplying through by b^2”}
\end{aligned}$$

and similarly,

$$\begin{aligned}
1-(\cos\omega)^2 &= 1-\frac{5b^2-1}{(9+d^2)b^2-2} && \text{“previous display”}\\
&= \frac{(9+d^2)b^2-2-5b^2+1}{(9+d^2)b^2-2}\\
& && \text{“putting the two terms on a common denominator”}\\
&= \frac{(4+d^2)b^2-1}{(9+d^2)b^2-2}\,. && \text{“consolidating”}
\end{aligned}$$

Using that $(\sin\omega)^2+(\cos\omega)^2$ and combining the two displays yields

$$\begin{aligned}
(\cos\omega)^2(\sin\omega)^2 &= (\cos\omega)^2(1-(\cos\omega)^2)\\
&= \frac{(5b^2-1)\bigl((4+d^2)b^2-1\bigr)}{\bigl((9+d^2)b^2-2\bigr)^2}\\
&= \frac{5(4+d^2)b^4-(9+d^2)b^2+1}{(9+d^2)^2b^4-4(9+d^2)b^2+4}\,,
\end{aligned}$$

as desired.

9. We do a proof by contradiction.

Assume that $a=b$. Similarly as in 6., we then find

$$\begin{aligned}
\frac{(\cos\omega)^2}{a^2}+\frac{(\sin\omega)^2}{b^2} &= 5 && \text{“comparing the }(\alpha_1-1)\text{-terms as in 6.”}\\
\Rightarrow\quad \frac{1}{a^2} &= 5 && \text{“}a=b;\ (\sin\omega)^2+(\cos\omega)^2=1\text{”}\\
\Rightarrow\quad a &= b=\frac{1}{\sqrt{5}}\,. && \text{“solving for }a\text{; using again that }a=b\text{”}
\end{aligned}$$

Then, similarly as in 7.,

$$\begin{aligned}
\frac{(\sin\omega)^2}{a^2}+\frac{(\cos\omega)^2}{b^2} &= 4+d^2 && \text{“comparing the }\alpha_2\text{-terms as in 7.”}\\
\Rightarrow\quad 5 &= 4+d^2\\
&\text{“}a=1/\sqrt{5}\text{ according to the previous display; }(\sin\omega)^2+(\cos\omega)^2=1\text{”}\\
\Rightarrow\quad d &= \pm 1\,. && \text{“consolidating”}
\end{aligned}$$

The case $d=1$ is excluded by assumption. For $d=-1$, we find

$$\begin{aligned}
2\biggl(\frac{1}{a^2}-\frac{1}{b^2}\biggr)(\cos\omega)(\sin\omega) &= 2(4+d) && \text{“comparing the }(1-\alpha_1)\alpha_2\text{-terms”}\\
\Rightarrow\quad 0 &= 6\,, && \text{“}a=b\text{ by assumption; }d=-1\text{”}
\end{aligned}$$

which means that $a=b$ cannot be true if $d=-1$ either. Thus, we have contradicted $a=b$, as desired.

10. We proceed similarly as before.

Observe that

$$\begin{aligned}
2\biggl(\frac{1}{a^2}-\frac{1}{b^2}\biggr)(\cos\omega)(\sin\omega) &= 2(4+d)\\
&\text{“comparing the }(1-\alpha_1)\alpha_2\text{-terms of 5. and 1.”}\\
\Rightarrow\quad 2\biggl(\frac{1}{a^2}+\frac{1}{b^2}-\frac{2}{b^2}\biggr)(\cos\omega)(\sin\omega) &= 2(4+d)\\
&\text{“adding a zero-valued term inside the first bracketed term”}\\
\Rightarrow\quad 2\biggl(9+d^2-\frac{2}{b^2}\biggr)(\cos\omega)(\sin\omega) &= 2(4+d) && \text{“Claim 7”}\\
\Rightarrow\quad (\cos\omega)(\sin\omega) &= \frac{4+d}{9+d^2-\frac{2}{b^2}}\\
&\text{“rearranging terms”}\\
\Rightarrow\quad (\cos\omega)^2(\sin\omega)^2 &= \frac{(4+d)^2}{(9+d^2-\frac{2}{b^2})^2}\\
&\text{“squaring both sides”}\\
\Rightarrow\quad \frac{5(4+d^2)b^4-(9+d^2)b^2+1}{(9+d^2)^2b^4-4(9+d^2)b^2+4} &= \frac{(4+d)^2}{(9+d^2-\frac{2}{b^2})^2} && \text{“Claim 8”}\\
\Rightarrow\quad \frac{5(4+d^2)-\frac{9+d^2}{b^2}+\frac{1}{b^4}}{(9+d^2)^2-\frac{4(9+d^2)}{b^2}+\frac{4}{b^2}}-\frac{(4+d)^2}{(9+d^2-\frac{2}{b^2})^2} &= 0\,,\\
&\text{“reducing the first fraction by }b^4\text{ and rearranging terms”}
\end{aligned}$$

as desired.

11. Find `R` code on the book homepage.

A.2 Solutions for Chapter 2

Solutions for Section 2.1

Solution 2.1 We prove the five claims in turn.

1. The first claim follows readily from the properties of minima. Indeed, we find

$$\begin{aligned}\min_{\boldsymbol{\alpha}'\in\mathbb{R}^{p+1}}\|\boldsymbol{y}-X'\boldsymbol{\alpha}'\|_2^2 &= \min_{\alpha_{p+1}\in\mathbb{R}}\min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\|\boldsymbol{y}-X\boldsymbol{\alpha}-\alpha\boldsymbol{x}\|_2^2 && \text{“}X'=(X,\boldsymbol{x})\text{”}\\ &\le \min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2\,, && \text{“consider }\alpha_{p+1}=0\text{”}\end{aligned}$$

as desired.

2. The proof of this claim consists of simple but slightly tedious derivations. Note first that the proof is complete if we can find a vector $\boldsymbol{\alpha}'\in\mathbb{R}^{p+1}$ such that

$$\|\boldsymbol{y}-X'\boldsymbol{\alpha}'\|_2^2 \;<\; \min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2\,.$$

To this end, we define $\boldsymbol{\alpha}':=(\widehat{\boldsymbol{\beta}}_{\text{ls}}^\top,\alpha_{p+1})^\top$ with $\alpha_{p+1}\in\mathbb{R}$ specified later. Now,

$$\begin{aligned}&\|\boldsymbol{y}-X'\boldsymbol{\alpha}'\|_2^2-\min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2\\ &= \|\boldsymbol{y}-X'\boldsymbol{\alpha}'\|_2^2-\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\text{ls}}\|_2^2 && \text{“definition of }\widehat{\boldsymbol{\beta}}_{\text{ls}}\text{”}\\ &= \|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\text{ls}}-\alpha_{p+1}\boldsymbol{x}\|_2^2-\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\text{ls}}\|_2^2\\ & && \text{“substituting }X'=(X,\boldsymbol{x})\text{ and }\boldsymbol{\alpha}'=(\widehat{\boldsymbol{\beta}}_{\text{ls}}^\top,\alpha_{p+1})^\top\text{”}\\ &= \|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\text{ls}}\|_2^2-2\langle\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\text{ls}},\,\alpha_{p+1}\boldsymbol{x}\rangle+\|\alpha_{p+1}\boldsymbol{x}\|_2^2-\|\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\text{ls}}\|_2^2\\ & && \text{“expanding the first term”}\\ &= -2\alpha_{p+1}\langle\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\text{ls}},\,\boldsymbol{x}\rangle+(\alpha_{p+1})^2\|\boldsymbol{x}\|_2^2\,.\\ & \text{“consolidating terms and using the linearity/positive homogeneity of inner products/norms”}\end{aligned}$$

In view of the stated condition $\langle\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\text{ls}},\,\boldsymbol{x}\rangle\neq0$, it must hold that $\boldsymbol{x}\neq\boldsymbol{0}_n$. Therefore, we can massage the display further into

$$\begin{aligned}&\|\boldsymbol{y}-X'\boldsymbol{\alpha}'\|_2^2-\min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2\\ &\qquad\qquad = \|\boldsymbol{x}\|_2^2\biggl(\frac{\langle\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\text{ls}},\,\boldsymbol{x}\rangle}{\|\boldsymbol{x}\|_2^2}-\alpha_{p+1}\biggr)^2-\frac{\bigl(\langle\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\text{ls}},\,\boldsymbol{x}\rangle\bigr)^2}{\|\boldsymbol{x}\|_2^2}\,.\end{aligned}$$

Setting $\alpha_{p+1}:=\langle\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\text{ls}},\,\boldsymbol{x}\rangle/\|\boldsymbol{x}\|_2^2$, this yields

$$\|\boldsymbol{y}-X'\boldsymbol{\alpha}'\|_2^2-\min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2 \;=\; -\frac{\bigl(\langle\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\text{ls}},\,\boldsymbol{x}\rangle\bigr)^2}{\|\boldsymbol{x}\|_2^2}\,,$$

which implies—again in view of the stated condition $\langle\boldsymbol{y}-X\widehat{\boldsymbol{\beta}}_{\text{ls}},\,\boldsymbol{x}\rangle\neq0$—that

$$\|\boldsymbol{y}-X'\boldsymbol{\alpha}'\|_2^2-\min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2 \;<\; 0\,,$$

and thus, rearranging,

$$\|\boldsymbol{y}-X'\boldsymbol{\alpha}'\|_2^2 \;<\; \min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2\,,$$

as desired.

3. We establish a proof by contradiction.

Consider a least-squares estimator of the form $(\widehat{\boldsymbol{\beta}}_{\mathrm{ls}})' = ((\widehat{\boldsymbol{\beta}}_{\mathrm{ls}})'_{-}{}^{\top}, 0)^{\top}$ for some $(\widehat{\boldsymbol{\beta}}_{\mathrm{ls}})'_{-} \in \mathbb{R}^p$. It then holds that

$$\begin{aligned}\min_{\boldsymbol{\alpha}' \in \mathbb{R}^{p+1}} \|\boldsymbol{y} - X'\boldsymbol{\alpha}'\|_2^2 &= \|\boldsymbol{y} - X'(\widehat{\boldsymbol{\beta}}_{\mathrm{ls}})'\|_2^2 && \text{“definition of } (\widehat{\boldsymbol{\beta}}_{\mathrm{ls}})'\text{”}\\ &= \|\boldsymbol{y} - X(\widehat{\boldsymbol{\beta}}_{\mathrm{ls}})'_{-}\|_2^2 && \text{“}(\widehat{\boldsymbol{\beta}}_{\mathrm{ls}})'_{p+1} = 0 \text{ by assumption; } X' = (X, \boldsymbol{x})\text{”}\\ &\geq \min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2\,. && \text{“taking the minimum”}\end{aligned}$$

This contradicts the strict inequality of 2., and thus, proves the claim.

4. For proving this claim, we argue with linear dependence.

Recall that if $\mathrm{rank}[X] \geq n$, the columns of X form a basis for the $\mathbb{R}^n$. Applied to the columns of $X' \in \mathbb{R}^{n\times(p+1)}$, this means that for any $\alpha_{p+1} \in \mathbb{R}$, there exists a $\boldsymbol{\gamma} \in \mathbb{R}^p$ such that $\alpha_{p+1}\boldsymbol{x} = X\boldsymbol{\gamma}$. We can use this in the following derivation:

$$\begin{aligned}&\min_{\boldsymbol{\alpha}' \in \mathbb{R}^{p+1}} \|\boldsymbol{y} - X'\boldsymbol{\alpha}'\|_2^2\\ &= \min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \min_{\alpha_{p+1} \in \mathbb{R}} \|\boldsymbol{y} - X\boldsymbol{\alpha} - \alpha_{p+1}\boldsymbol{x}\|_2^2 && \text{“properties of minima; } X' = (X, \boldsymbol{x})\text{”}\\ &= \min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \min_{\boldsymbol{\gamma} \in \mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\alpha} - X\boldsymbol{\gamma}\|_2^2 && \text{“above insight: } \alpha_{p+1}\boldsymbol{x} = X\boldsymbol{\gamma}\text{”}\\ &= \min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2\,, && \text{“}\boldsymbol{\alpha} \text{ taking the role of } \boldsymbol{\alpha} + \boldsymbol{\gamma}\text{”}\end{aligned}$$

as desired.

5. We prove this by constructing and verifying a specific solution.

Set $\boldsymbol{\gamma}' := (\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}^{\top}, 0)^{\top} + (\boldsymbol{\gamma}^{\top}, -\alpha_{p+1})^{\top} \in \mathbb{R}^p$, where $\alpha_{p+1} \in \mathbb{R} \setminus \{0\}$ and $\boldsymbol{\gamma} \in \mathbb{R}^p$ such that $\alpha_{p+1}\boldsymbol{x} = X\boldsymbol{\gamma}$ (cf. the proof of 4.). Then,

$$\begin{aligned}&\|\boldsymbol{y} - X'\boldsymbol{\gamma}'\|_2^2\\ &= \|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}} - X\boldsymbol{\gamma} + \alpha'_{p+1}\boldsymbol{x}\|_2^2 && \text{“choice of } \boldsymbol{\gamma}'\text{; } X' = (X, \boldsymbol{x})\text{”}\\ &= \|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2 && \text{“}\alpha_{p+1}\boldsymbol{x} = X\boldsymbol{\gamma} \text{ by construction”}\\ &= \min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 && \text{“definition of } \widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\text{”}\\ &= \min_{\boldsymbol{\alpha}' \in \mathbb{R}^{p+1}} \|\boldsymbol{y} - X'\boldsymbol{\alpha}'\|_2^2\,, && \text{“4.”}\end{aligned}$$

Hence, $(\widehat{\boldsymbol{\beta}}_{\mathrm{ls}})' := \boldsymbol{\gamma}'$ is a least-squares solution with a non-zero-valued last coordinate, as desired.

Solutions for Section 2.2

Solution 2.2 The proof is a relatively straightforward reformulation of the lasso's KKT equation. We show this in three steps, as indicated by the layout of the exercise.

1. The first result follows almost directly from the lasso's KKT conditions (2.2).

Recall that these conditions are

$$-2X^{\top}(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{\mathrm{lasso}}) + r\widehat{\boldsymbol{\kappa}} = \mathbf{0}_p$$

for a vector $\widehat{\boldsymbol{\kappa}} \in \boldsymbol{\partial}\|\widehat{\boldsymbol{\beta}}_{\mathrm{lasso}}\|_1$. Using $X^{\top}X = \mathrm{I}_{p\times p}$ and rearranging the terms of the display yields

$$\widehat{\boldsymbol{\beta}}_{\mathrm{lasso}} = X^{\top}\boldsymbol{y} - \frac{r}{2}\widehat{\boldsymbol{\kappa}}\,,$$

as desired.

2. The second result can be derived from the equation in 1. by using only basic algebra. We show this in two steps, proving first ⇒ and then ⇐.

⇒: Assume that $(\widehat{\beta}_{\text{lasso}})_j = 0$. We then find from 1. that

$$0 \;=\; (X^\top \boldsymbol{y})_j - \frac{r}{2}\widehat{\kappa}_j\,,$$

which implies—rearrange the terms and take absolute values on both sides—

$$\bigl|(X^\top \boldsymbol{y})_j\bigr| \;=\; \frac{r}{2}|\widehat{\kappa}_j|\,.$$

Now, $|\widehat{\kappa}_j| \leq 1$ due to $\widehat{\boldsymbol{\kappa}} \in \boldsymbol{\partial}\|\widehat{\boldsymbol{\beta}}_{\text{lasso}}\|_1$—see Display (2.2). Plugging this into the preceeding equality yields

$$\bigl|(X^\top \boldsymbol{y})_j\bigr| \;\leq\; \frac{r}{2}\,,$$

as desired.

⇐: Assume first $(\widehat{\beta}_{\text{lasso}})_j > 0$. Then, 1. yields

$$(X^\top \boldsymbol{y})_j - \frac{r}{2}\widehat{\kappa}_j \;>\; 0\,,$$

which can be reformulated as

$$(X^\top \boldsymbol{y})_j \;>\; \frac{r}{2}\widehat{\kappa}_j\,.$$

Hence, since $\widehat{\kappa}_j = 1$ for $(\widehat{\beta}_{\text{lasso}})_j > 0$,

$$(X^\top \boldsymbol{y})_j \;>\; \frac{r}{2}\,,$$

which implies (recall that $r \geq 0$)

$$\bigl|(X^\top \boldsymbol{y})_j\bigr| \;>\; \frac{r}{2}\,.$$

Similarly, if $(\widehat{\beta}_{\text{lasso}})_j < 0$, 1. yields

$$(X^\top \boldsymbol{y})_j - \frac{r}{2}\widehat{\kappa}_j \;<\; 0\,,$$

which can be reformulated as

$$-(X^\top \boldsymbol{y})_j \;>\; -\frac{r}{2}\widehat{\kappa}_j\,.$$

Hence, since $\widehat{\kappa}_j = -1$ for $(\widehat{\beta}_{\text{lasso}})_j < 0$,

$$-(X^\top \boldsymbol{y})_j \;>\; \frac{r}{2}\,,$$

which implies

$$\bigl|(X^\top \boldsymbol{y})_j\bigr| \;>\; \frac{r}{2}\,,$$

as desired.

3. The proof of this part uses 1. and 2. and the properties of the sign function and of the positive part.

Note first that for $|(X^\top \boldsymbol{y})_j| \leq r/2$, the right-hand side of the desired equality is equal to zero, which is correct in view of 2. We can thus assume $|(X^\top \boldsymbol{y})_j| > r/2$ in the following.

We consider first $(X^\top \boldsymbol{y})_j > r/2$. Then,

$$\begin{aligned}(\widehat{\beta}_{\text{lasso}})_j &= (X^\top \boldsymbol{y})_j - \frac{r}{2}\widehat{\kappa}_j && \text{“from 1.”}\\ &\geq (X^\top \boldsymbol{y})_j - \frac{r}{2} && \text{“}\widehat{\kappa}_j \leq 1 \text{ since } \widehat{\kappa}_j \in \boldsymbol{\partial}\|\widehat{\boldsymbol{\beta}}_{\text{lasso}}\|_1\text{”}\\ &> 0\,. && \text{“since } (X^\top \boldsymbol{y})_j > r/2 \text{ by assumption”}\end{aligned}$$

Similarly for $-(X^\top \boldsymbol{y})_j > r/2$,

$$\begin{aligned}(\widehat{\beta}_{\text{lasso}})_j &= (X^\top \boldsymbol{y})_j - \frac{r}{2}\widehat{\kappa}_j && \text{“from 1.”}\\ &\leq (X^\top \boldsymbol{y})_j + \frac{r}{2} && \text{“}\widehat{\kappa}_j \geq -1 \text{ since } \widehat{\kappa}_j \in \boldsymbol{\partial}\|\widehat{\boldsymbol{\beta}}_{\text{lasso}}\|_1\text{”}\\ &< 0\,. && \text{“since } (X^\top \boldsymbol{y})_j < -r/2 \text{ by assumption”}\end{aligned}$$

Hence,

$$\begin{aligned}(\widehat{\beta}_{\text{lasso}})_j &= (X^\top \boldsymbol{y})_j - \frac{r}{2}\widehat{\kappa}_j && \text{“again from 1.”}\\ &= \begin{cases} |(X^\top \boldsymbol{y})_j| - \frac{r}{2} & \text{if } (X^\top \boldsymbol{y})_j > r/2\\ -|(X^\top \boldsymbol{y})_j| + \frac{r}{2} & \text{if } -(X^\top \boldsymbol{y})_j > r/2\end{cases}\end{aligned}$$

“using the two preceeding displays and $\widehat{\kappa}_j = 1$ and $\widehat{\kappa}_j = -1$ for $(\widehat{\beta}_{\text{lasso}})_j > 0$ and $(\widehat{\beta}_{\text{lasso}})_j < 0$, respectively ”

$$\begin{aligned}&= \operatorname{sign}\big[(X^\top \boldsymbol{y})_j\big]\Big(\big|(X^\top \boldsymbol{y})_j\big| - \frac{r}{2}\Big) && \text{“definition of the sign function”}\\ &= \operatorname{sign}\big[(X^\top \boldsymbol{y})_j\big]\Big(\big|(X^\top \boldsymbol{y})_j\big| - \frac{r}{2}\Big)_+\,, && \text{“}|(X^\top \boldsymbol{y})_j| \geq \tfrac{r}{2} \text{ by assumption”}\end{aligned}$$

as desired.

Alternatively, we could approach in the case $|(X^\top \boldsymbol{y})_j| > r/2$ “the other way round:”

$$\begin{aligned}&\operatorname{sign}\big[(X^\top \boldsymbol{y})_j\big]\Big(\big|(X^\top \boldsymbol{y})_j\big| - \frac{r}{2}\Big)_+\\ &= \operatorname{sign}\big[(X^\top \boldsymbol{y})_j\big]\Big(\big|(X^\top \boldsymbol{y})_j\big| - \frac{r}{2}\Big) && \text{“}|(X^\top \boldsymbol{y})_j| > r/2 \text{ by assumption”}\\ &= (X^\top \boldsymbol{y})_j - \frac{r}{2}\operatorname{sign}\big[(X^\top \boldsymbol{y})_j\big] && \text{“multiplying out”}\\ &= (X^\top \boldsymbol{y})_j - \frac{r}{2}\operatorname{sign}\Big[(X^\top \boldsymbol{y})_j - \frac{r}{2}\widehat{\kappa}_j\Big]\\ &\qquad \text{“since } |(X^\top \boldsymbol{y})_j| > r/2 \text{ by assumption and } |\widehat{\kappa}_j| \leq 1\text{, the sign cannot change”}\\ &= (X^\top \boldsymbol{y})_j - \frac{r}{2}\operatorname{sign}\big[(\widehat{\beta}_{\text{lasso}})_j\big] && \text{“1.”}\\ &= (X^\top \boldsymbol{y})_j - \frac{r}{2}\widehat{\kappa}_j && \text{“}(\widehat{\beta}_{\text{lasso}})_j > 0 \text{ by 2.; } \widehat{\kappa}_j \in \partial|(\widehat{\beta}_{\text{lasso}})_j|\text{”}\\ &= (\widehat{\beta}_{\text{lasso}})_j\,. && \text{“1.”}\end{aligned}$$

This also provides a complete proof.

Solution 2.3 We prove the five parts in order.

1. This equation is a simple reformulation of the lasso's KKT condition.

Recall that Equation 2.2 states

$$-2X^\top\big(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}[r]\big) + r\widehat{\boldsymbol{\kappa}} = \mathbf{0}_p$$

for a vector $\widehat{\boldsymbol{\kappa}} \in \boldsymbol{\partial}\|\widehat{\boldsymbol{\beta}}[r]\|_1$. Rearranging the terms yields

$$2X^\top X\widehat{\boldsymbol{\beta}}[r] = 2X^\top \boldsymbol{y} - r\widehat{\boldsymbol{\kappa}}\,,$$

and multiplying then both sides by $(X^\top X)^{-1}/2$

$$\widehat{\boldsymbol{\beta}}[r] \;=\; \underbrace{(X^\top X)^{-1}X^\top \boldsymbol{y}}_{\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}} - \frac{r}{2}(X^\top X)^{-1}\widehat{\boldsymbol{\kappa}}\,,$$

as desired.

2. This inequality follows readily from 1.

Verify first that if $A \in \mathbb{R}^{p\times p}$ is a positive definite matrix with ordered eigenvalues $e_p \geq \cdots \geq e_1 > 0$, the ordered eigenvalues of A^{-1} are $1/e_1 \geq \cdots \geq 1/e_p > 0$ and $\|A^{-1}\boldsymbol{a}\|_2 \leq \|\boldsymbol{a}\|_2/e_1$.

We then find

$$\begin{aligned}
&\|X\widehat{\boldsymbol{\beta}}[r] - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2\\
&= \|-\frac{r}{2}X(X^\top X)^{-1}\widehat{\boldsymbol{\kappa}}\|_2^2 && \text{“1.”}\\
&= \frac{r^2\|X(X^\top X)^{-1}\widehat{\boldsymbol{\kappa}}\|_2^2}{4} && \text{“absolute homogeneity of norms”}\\
&= \frac{r^2\big(X(X^\top X)^{-1}\widehat{\boldsymbol{\kappa}}\big)^\top X(X^\top X)^{-1}\widehat{\boldsymbol{\kappa}}}{4} && \text{“definition of the } \ell_2\text{-norm”}\\
&= \frac{r^2\widehat{\boldsymbol{\kappa}}^\top\big((X^\top X)^{-1}\big)^\top X^\top X(X^\top X)^{-1}\widehat{\boldsymbol{\kappa}}}{4} && \text{“properties of the transpose”}\\
&= \frac{r^2\widehat{\boldsymbol{\kappa}}^\top\big((X^\top X)^{-1}\big)^\top\widehat{\boldsymbol{\kappa}}}{4} && \text{“simpifying”}\\
&= \frac{r^2\widehat{\boldsymbol{\kappa}}^\top(X^\top X)^{-1}\widehat{\boldsymbol{\kappa}}}{4} && \text{“using 3. in Lemma B.2.4 and that } X^\top X \text{ is symmetric”}\\
&\leq \frac{r^2\|\widehat{\boldsymbol{\kappa}}\|_2\|(X^\top X)^{-1}\widehat{\boldsymbol{\kappa}}\|_2}{4} && \text{“Hölder's inequality”}\\
&\leq \frac{r^2\|\widehat{\boldsymbol{\kappa}}\|_2^2}{4e_1}\,. && \text{“previous comments applied to } A = X^\top X\text{”}
\end{aligned}$$

Now, recall that $|\widehat{\kappa}_j| \leq 1$ since $\widehat{\boldsymbol{\kappa}} \in \boldsymbol{\partial}\|\widehat{\boldsymbol{\beta}}[r]\|_1$, so that the definition of the ℓ_2-norm and simple algebra yield

$$\|\widehat{\boldsymbol{\kappa}}\|_2^2 \;=\; \sum_{j=1}^p \widehat{\kappa}_j^2 \;\leq\; p\,.$$

This combined with the previous display then implies

$$\|X\widehat{\boldsymbol{\beta}}[r] - X\widehat{\boldsymbol{\beta}}_{\mathrm{ls}}\|_2^2 \;\leq\; \frac{r^2 p}{4e_1}\,,$$

as desired.

3. The claim follows again readily from the lasso's KKT conditions.

Once more, recall that Equation (2.2) states that $\widehat{\boldsymbol{\beta}}[r]$ is a solution of the lasso if and only if

$$-2X^\top\big(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}[r]\big) + r\widehat{\boldsymbol{\kappa}} \;=\; \mathbf{0}_p$$

for a $\widehat{\boldsymbol{\kappa}} \in \boldsymbol{\partial}\|\widehat{\boldsymbol{\beta}}[r]\|_1$. Setting $\widehat{\boldsymbol{\beta}}[r] = \mathbf{0}_p$ and rearranging yield that $\mathbf{0}_p$ is a solution if and only if

$$2X^\top\boldsymbol{y} \;=\; r\widehat{\boldsymbol{\kappa}}$$

for a $\widehat{\boldsymbol{\kappa}} \in \boldsymbol{\partial}\|\mathbf{0}_p\|_1$.

Next, recall that $\widehat{\boldsymbol{\kappa}} \in \boldsymbol{\partial}\|\mathbf{0}_p\|_1$ is equivalent to $\widehat{\kappa}_j \in [-1,1]$ for all $j \in \{1,\dots,p\}$. Consequently, formulated in terms of the individual coordinates of the vectors, $\mathbf{0}_p$ is a solution if and only if for all $j \in \{1,\dots,p\}$, there is a $\widehat{\kappa}_j \in [-1,1]$ such that

$$2(X^\top \boldsymbol{y})_j \;=\; r\widehat{\kappa}_j\,.$$

The latter is true if and only if $2|(X^\top \boldsymbol{y})_j| \leq r$ for all $j \in \{1,\dots,p\}$, that is, $2\|X^\top \boldsymbol{y}\|_\infty \leq r$, as desired.

4. For this question, we leverage the estimators definition directly.

Assume that $\mathbf{0}_p$ is a lasso solution, that is, $\mathbf{0}_p$ is a minimizer of the lasso's objective function. Then, the value of the lasso's objective function at any solution $\boldsymbol{\gamma} \in \mathbb{R}^p$ must equal the value at $\mathbf{0}_p$:

$$\|\boldsymbol{y} - X\boldsymbol{\gamma}\|_2^2 + r\|\boldsymbol{\gamma}\|_1 \;=\; \|\boldsymbol{y} - X\mathbf{0}_p\|_2^2 + r\|\mathbf{0}_p\|_1 \;=\; \|\boldsymbol{y}\|_2^2\,.$$

Expanding the first term on the left-hand side of this equality yields

$$\|\boldsymbol{y}\|_2^2 - 2\langle \boldsymbol{y},\, X\boldsymbol{\gamma}\rangle + \|X\boldsymbol{\gamma}\|_2^2 + r\|\boldsymbol{\gamma}\|_1 \;=\; \|\boldsymbol{y}\|_2^2$$

and simplifying and rearranging then

$$\|X\boldsymbol{\gamma}\|_2^2 + r\|\boldsymbol{\gamma}\|_1 \;=\; 2\langle \boldsymbol{y},\, X\boldsymbol{\gamma}\rangle\,.$$

Using the properties of inner products, Hölder's inequality, and that $r \geq \|X^\top \boldsymbol{y}\|_\infty$, we can bound the right-hand side of this equality according to

$$2\langle \boldsymbol{y},\, X\boldsymbol{\gamma}\rangle \;=\; 2\langle X^\top\boldsymbol{y},\, \boldsymbol{\gamma}\rangle \;\leq\; 2\|X^\top\boldsymbol{y}\|_\infty\|\boldsymbol{\gamma}\|_1 \;\leq\; r\|\boldsymbol{\gamma}\|_1\,.$$

Plugging this back into the penultimate display gives

$$\|X\boldsymbol{\gamma}\|_2^2 + r\|\boldsymbol{\gamma}\|_1 \;\leq\; r\|\boldsymbol{\gamma}\|_1$$

and, therefore,

$$\|X\boldsymbol{\gamma}\|_2^2 \;\leq\; 0\,,$$

which means that $X\boldsymbol{\gamma} = \mathbf{0}_n$ by the positive definiteness of norms.

We can plug this into the previously derived equality $\|X\boldsymbol{\gamma}\|_2^2 + r\|\boldsymbol{\gamma}\|_1 = 2\langle \boldsymbol{y},\, X\boldsymbol{\gamma}\rangle$ to find

$$r\|\boldsymbol{\gamma}\|_1 \;=\; 0\,.$$

Since $r > 0$ by assumption, we can again use the positive definitness of norms to derive from this equality that $\boldsymbol{\gamma} = \mathbf{0}_p$, as desired.

5. This is a simple algebra exercise.

Observe first that the positive definiteness of norms yields

$$2\|X^\top\boldsymbol{y}\|_\infty \;=\; 0 \quad\Rightarrow\quad X^\top\boldsymbol{y} \;=\; \mathbf{0}_p\,.$$

Therefore, the lasso's objective function becomes

$$\begin{aligned}
&\|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 + r\|\boldsymbol{\alpha}\|_1 \\
&= \|\boldsymbol{y} - X\boldsymbol{\alpha}\|_2^2 && \text{“}r = 0\text{ by assumption”}\\
&= \|\boldsymbol{y}\|_2^2 - 2\langle \boldsymbol{y},\, X\boldsymbol{\alpha}\rangle + \|X\boldsymbol{\alpha}\|_2^2 && \text{“expanding the squared-norm”}\\
&= \|\boldsymbol{y}\|_2^2 - 2\langle X^\top\boldsymbol{y},\, \boldsymbol{\alpha}\rangle + \|X\boldsymbol{\alpha}\|_2^2 && \text{“properties of the inner product”}\\
&= \|\boldsymbol{y}\|_2^2 - 2\langle \mathbf{0}_p,\, \boldsymbol{\alpha}\rangle + \|X\boldsymbol{\alpha}\|_2^2 && \text{“previous display”}\\
&= \|\boldsymbol{y}\|_2^2 + \|X\boldsymbol{\alpha}\|_2^2\,. && \text{“linearity of the inner product”}
\end{aligned}$$

Hence, since $\|\boldsymbol{y}\|_2^2$ does not depend on $\boldsymbol{\alpha}$,

$$\operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\bigl\{\,\|\boldsymbol{y}-X\boldsymbol{\alpha}\|_2^2+r\|\boldsymbol{\alpha}\|_1\,\bigr\} \;=\; \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^p}\|X\boldsymbol{\alpha}\|_2^2 \;=\; \bigl\{\boldsymbol{\alpha}\in\mathbb{R}^p \;:\; X\boldsymbol{\alpha}=\mathbf{0}_n\bigr\}\,,$$

as desired.

Alternatively, one could verify that the KKT conditions imply in our case (where $2X^\top\boldsymbol{y}=r\widehat{\boldsymbol{\kappa}}/2=\mathbf{0}_p$) that $X^\top X\widehat{\boldsymbol{\beta}}_{\text{lasso}}=\mathbf{0}_p$. Multiplying both sides of this equation by $\widehat{\boldsymbol{\beta}}_{\text{lasso}}$ yields $\|X\widehat{\boldsymbol{\beta}}_{\text{lasso}}\|_2=0$, from which the proof follows readily.

A.3 Solutions for Chapter 3

Solutions for Section 3.1

Solution 3.1: Refer to online resources such as Wikipedia.

Solutions for Section 3.4

Solution 3.2 We prove the six claims in order.

1. The proof of this claim is a simple algebra exercise.

Consider two matrices $\Omega',\Omega''\in\mathcal{S}_p^+$ and a constant $v\in[0,1]$. We have to show that $v\Omega'+(1-v)\Omega''\in\mathcal{S}_p^+$.

The symmetry of the matrix in question is proved as follows:

$$\begin{aligned}&\bigl(v\Omega'+(1-v)\Omega''\bigr)^\top\\ &=\; v\Omega'^\top+(1-v)\Omega''^\top && \text{“linearity of the transpose”}\\ &=\; v\Omega'+(1-v)\Omega''\,, && \text{“Ω',Ω'' are symmetric by assumption”}\end{aligned}$$

as desired.

The positive definiteness of the matrix is proved as follows: for any $\boldsymbol{b}\in\mathbb{R}^p\setminus\{\mathbf{0}_p\}$,

$$\begin{aligned}&\boldsymbol{b}^\top\bigl(v\Omega'+(1-v)\Omega''\bigr)\boldsymbol{b}\\ &=\; v\boldsymbol{b}^\top\Omega'\boldsymbol{b}+(1-v)\boldsymbol{b}^\top\Omega''\boldsymbol{b} && \text{“linearity of matrix-vector multiplications”}\\ &>\; 0\,, && \text{“Ω',Ω'' positive definite by assumption; $v,1-v\geq 0$ by assumption”}\end{aligned}$$

as desired.

This concludes the proof of Claim 1.

2. This claim follows directly from the linearity of matrix multiplications and of the trace function.

Indeed, these linearity properties applied in sequence yield for any $\Omega',\Omega''\in\mathbb{R}^{p\times p}$ and $v\in[0,1]$ that

$$\operatorname{tr}\bigl[A\bigl(v\Omega'+(1-v)\Omega''\bigr)\bigr] \;=\; \operatorname{tr}\bigl[vA\Omega'+(1-v)A\Omega''\bigr] \;=\; v\operatorname{tr}[A\Omega']+(1-v)\operatorname{tr}[A\Omega'']\,,$$

which implies the desired claim.

3. This proof is again a simple algebra exercise.

Since Ω is symmetric by assumption, there exists a diagonalization $\Omega=ODO^\top$ with $O\in\mathbb{R}^{p\times p}$ an orthogonal matrix (in particular, $\det O=\det O^\top=\pm1$) and $D\in\mathbb{R}^{p\times p}$

a diagonal matrix with diagonal $\operatorname{diag} D = (a_1, \dots, a_p)^\top$. Using this decomposition, we then find

$$\begin{aligned}
&\log\det\Omega && \text{“well-defined since } \det\Omega > 0 \text{ by assumption”}\\
&= \log\det[ODO^\top] && \text{“}\Omega = ODO^\top \text{ by construction”}\\
&= \log\bigl[\det[O]\det[D]\det[O^\top]\bigr] && \\
& && \text{“1. in Lemma B.2.1 (multiplicativeness of the determinant)”}\\
&= \log\det D && \text{“}\det O = \det O^\top = \pm 1 \text{ by construction”}\\
&= \log\prod_{j=1}^{p} D_{jj} && \text{“Leipniz form of determinant; } D \text{ diagonal”}\\
&= \sum_{j=1}^{p}\log D_{jj} && \text{“rules for the logarithm: } \log[ab] = \log[a] + \log[b]\text{”}\\
&= \sum_{j=1}^{p}\log a_j\,, && \text{“diag } D_{jj} = a_j \text{ by assumption”}
\end{aligned}$$

as desired.

4. This follows from 3. and the strict concavity of the logarithm.

We first fix a point $t \in [0,1]$ and show that $\mathrm{I}_{p\times p} + t\Omega''^{-1}(\Omega' - \Omega'') \in \mathcal{S}_p^+$.

First, one can verify readily that the matrix is symmetric.

Second, for any $\boldsymbol{b} \in \mathbb{R}^p \setminus \{\boldsymbol{0}_p\}$ and $t < 1$

$$\begin{aligned}
&\boldsymbol{b}^\top\bigl(\mathrm{I}_{p\times p} + t\Omega''^{-1}(\Omega' - \Omega'')\bigr)\boldsymbol{b} &&\\
&= \boldsymbol{b}^\top\,\mathrm{I}_{p\times p}\,\boldsymbol{b} + t\boldsymbol{b}^\top\bigl(\Omega''^{-1}(\Omega' - \Omega'')\bigr)\boldsymbol{b} && \text{“linearity of matrix-vector multiplications”}\\
&> \|\boldsymbol{b}\|_2^2 - t\boldsymbol{b}^\top\bigl(\Omega''^{-1}\Omega''\bigr)\boldsymbol{b} &&\\
& && \text{“}\boldsymbol{b}^\top\Omega''^{-1}\Omega'\boldsymbol{b} > 0 \text{ since } \Omega', \Omega'' \text{ are positive definite by assumption”}\\
&= \|\boldsymbol{b}\|_2^2 - t\|\boldsymbol{b}\|_2^2 && \text{“}\Omega''^{-1}\Omega'' = \mathrm{I}_{p\times p}\text{”}\\
&> 0\,; && \text{“}t < 1 \text{ by assumption”}
\end{aligned}$$

and for $t = 1$, it holds similarly that $\boldsymbol{b}^\top(\mathrm{I}_{p\times p} + t\Omega''^{-1}(\Omega' - \Omega''))\boldsymbol{b} = \boldsymbol{b}^\top\Omega''^{-1}\Omega'\boldsymbol{b} > 0$. Hence, $\mathrm{I}_{p\times p} + t\Omega''^{-1}(\Omega' - \Omega'') \in \mathcal{S}_p^+$.

Then, using 3.,

$$\begin{aligned}
&-\log\det\bigl[\Omega'' + t(\Omega' - \Omega'')\bigr] &&\\
&= -\log\det\Bigl[\Omega''\bigl(\mathrm{I}_{p\times p} + t\Omega''^{-1}(\Omega' - \Omega'')\bigr)\Bigr] && \text{“factoring out } \Omega''\text{”}\\
&= -\log\Bigl[\det[\Omega'']\cdot\det\bigl[\mathrm{I}_{p\times p} + t\Omega''^{-1}(\Omega' - \Omega'')\bigr]\Bigr] &&\\
& && \text{“1. in Lemma B.2.1 (multiplicativeness of the determinant)”}\\
&= -\log\det[\Omega''] - \log\det\bigl[\mathrm{I}_{p\times p} + t\Omega''^{-1}(\Omega' - \Omega'')\bigr] &&\\
& && \text{“rules for the logarithm: } \log[ab] = \log a + \log b\text{”}\\
&= -\log\det[\Omega''] - \sum_{j=1}^{p}\log[1 + ta_j]\,, && \text{“Claim 3”}
\end{aligned}$$

where $a_1, \dots, a_p$ are the eigenvalues of $\Omega''^{-1}(\Omega' - \Omega'')$. The first term is constant in t; the second term is strictly convex due to the strict concavity of the logarithm (at least one a_j must be different from zero since $\Omega' \neq \Omega''$ by assumption). This demonstrates that the function in question is strictly convex on $[0,1]$, as desired.

5. This claim follows from 4. indeed.
Fix any two $\Omega', \Omega'' \in \mathcal{S}_p^+$, $\Omega' \neq \Omega''$, and $v \in (0,1)$. Then,

$$\begin{aligned}
&-\log\det\bigl[v\Omega' + (1-v)\Omega''\bigr]\\
&= \;-\log\det\bigl[\Omega'' + v(\Omega' - \Omega'')\bigr] && \text{“rearranging terms”}\\
&= \;-\log\det\Bigl[\Omega'' + \bigl(v\cdot 1 + (1-v)\cdot 0\bigr)(\Omega' - \Omega'')\Bigr]\\
& && \text{“adding a zero-valued term and multiplying by one”}\\
&< \;-v\log\det\bigl[\Omega'' + 1\cdot(\Omega' - \Omega'')\bigr] - (1-v)\log\det\bigl[\Omega'' + 0\cdot(\Omega' - \Omega'')\bigr] && \text{“Claim 4”}\\
&= \;-v\log\det[\Omega'] - (1-v)\log\det[\Omega''] \,, && \text{“consolidating”}
\end{aligned}$$

as desired.

6. This claim follows directly from 2., 5., and that the sum of a convex function and a strictly convex function is again a strictly convex function.

Solution 3.3 We prove the five claims in order.

1. The proof of the first claim is based on the linearity of the trace function.
By the definition of the trace, we find

$$\operatorname{tr}[A\Omega] \;=\; \sum_{j=1}^{p}(A\Omega)_{jj} \;=\; \sum_{j=1}^{p}\sum_{i=1}^{p} A_{ji}\Omega_{ij}\,.$$

Therefore, taking derivatives yields for any $i, j \in \{1, \dots, p\}$

$$\frac{\partial}{\partial\Omega_{ij}}\operatorname{tr}[A\Omega] \;=\; A_{ji} \;=\; (A^\top)_{ij}\,,$$

which is in matrix form

$$\frac{\partial}{\partial\Omega}\operatorname{tr}[A\Omega] \;=\; A^\top\,,$$

as desired.

2. For this proof, we use the Laplace expansion of the determinant.
Given an invertible matrix $\Omega \in \mathbb{R}^{p\times p}$, let $C \equiv C[\Omega] \in \mathbb{R}^{p\times p}$ be its cofactor matrix defined as follows: $C_{ij} := (-1)^{i+j}m_{ij}$ with m_{ij} the determinant of the $(p-1)\times(p-1)$-matrix that results from deleting the ith row and the jth column of Ω (C_{ij} is called the (i,j)th cofactor of Ω and m_{ij} the (i,j)th minor of Ω). It holds that (i) m_{ij} does not depend on Ω_{ij} nor does any m_{kj}, $k \in \{1, \dots, p\}$, (because the jth column of Ω is not considered in them) and (ii) $\Omega^{-1} = C^\top/\det[\Omega]$ (cf. Cramer's rule—we omit the details). The Laplace expansion of the Ω's determinant for the jth column, $j \in \{1, \dots, p\}$, is then (again, we omit the details)

$$\det[\Omega] \;=\; \sum_{k=1}^{p} C_{kj}\Omega_{kj}\,.$$

Hence, taking derivatives yields for any $i, j \in \{1, \dots, p\}$

$$\begin{aligned}
\frac{\partial}{\partial\Omega_{ij}}\log\bigl[\det[\Omega]\bigr] \;&=\; \frac{1}{\det[\Omega]}\frac{\partial}{\partial\Omega_{ij}}\det[\Omega] && \text{“chain rule; } \tfrac{d}{da}\log[a] \;=\; 1/a\text{”}\\
&=\; \frac{1}{\det[\Omega]}\frac{\partial}{\partial\Omega_{ij}}\sum_{k=1}^{p} C_{kj}\Omega_{kj} && \text{“penultimate display”}\\
&=\; \frac{C_{ij}}{\det[\Omega]} && \text{“sum rule and (i) above”}\\
&=\; \bigl((\Omega^{-1})^\top\bigr)_{ij}\,, && \text{“(ii) above”}
\end{aligned}$$

which can be written in the desired matrix form.

3. This part follows readily from the hint, the KKT conditions, and Claims 1 and 2 above.

Exercise 3.2.5 ensures that the objective function is strictly convex. Hence, if a minimizer exists, it must be unique.

To check the existence and the form of the minimizer, we just need to set derivatives to zero (KKT conditions):

$$\begin{aligned}\frac{\partial}{\partial\Omega}\Big|_{\Omega=\widehat{\Theta}}\big(\operatorname{tr}[A\Omega]-\log\big[\det[\Omega]\big]\big) &= \mathbf{0}_{p\times p} && \text{“KKT conditions for } \widehat{\Theta}\text{”}\\ \Rightarrow\qquad A^\top-(\widehat{\Theta}^{-1})^\top &= \mathbf{0}_{p\times p} && \text{“sum rule and 1. and 2.”}\\ \Rightarrow\qquad \widehat{\Theta} &= A^{-1}\,,\,. && \text{“}A\text{ and }\widehat{\Theta}\text{ are invertible by assumption”}\end{aligned}$$

This shows that the minimizer indeed exists and is equal to A^{-1}, as desired.

4. For this part, we use that sets that live in spaces that are of a dimension smaller than the ambient spaces' dimension are nullsets with respect to any Gaussian measure.

First, symmetry is straightforward:

$$\begin{aligned}\Big(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}^i\boldsymbol{z}^{i\top}\Big)^\top &= \frac{1}{n}\sum_{i=1}^n\big(\boldsymbol{z}^i\boldsymbol{z}^{i\top}\big)^\top && \text{“linearity of the transpose”}\\ &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}^i\boldsymbol{z}^{i\top}\,, && \text{“}(AB)^\top=B^\top A^\top\text{”}\end{aligned}$$

as desired.

For the invertibility, consider first the case $p=1$. Then,

$$\begin{aligned}&\mathbb{P}\Big\{\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}^i\boldsymbol{z}^{i\top}\text{ not invertible}\Big\}\\ &=\ \mathbb{P}\Big\{\frac{1}{n}\sum_{i=1}^n (z^i)^2\ =\ 0\Big\} && \text{“}p=1\text{ by assumption”}\\ &=\ \mathbb{P}\big\{z^i\ =\ 0\ \forall i\in\{1,\dots,p\}\big\} && \text{“all summands are non-negative”}\\ &\le\ \mathbb{P}\big\{z^1\ =\ 0\big\} && \text{“probability functions are increasing: } \mathbb{P}\{\mathcal{A}\}\le\mathbb{P}\{\mathcal{A}'\}\text{ if }\mathcal{A}\subset\mathcal{A}'\text{”}\\ &=\ 0\,,\end{aligned}$$

“affine subspaces with codimension larger or equal to one are nullsets of Gaussian distributions”

as desired.

Consider now the case $p>1$. Take the first k outcomes $\boldsymbol{z}^1,\dots,\boldsymbol{z}^k$, $k\in\{1,\dots,p-1\}$. These vectors span an affine subspace of dimension at most k, which makes a nullset with respect to any (non-degenerate) Gaussian distribution on the $\mathbb{R}^p$. Hence, $\boldsymbol{z}^{k+1}$ is linear independent of the first k outcomes with probability one. Since unions of finitely many nullsets are still nullsets, we conclude by induction that the first $\boldsymbol{z}^1,\dots,\boldsymbol{z}^p$ form a basis of the $\mathbb{R}^p$ with probability one.

Now,

$$\begin{aligned}
&\mathbb{P}\biggl\{\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}^i{\boldsymbol{z}^i}^\top \text{ not invertible}\biggr\}\\
&= \mathbb{P}\biggl\{\min_{\boldsymbol{a}\in\mathbb{R}^p\setminus\{\mathbf{0}_p\}} \boldsymbol{a}^\top\Bigl(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}^i{\boldsymbol{z}^i}^\top\Bigr)\boldsymbol{a} \;=\; 0\biggr\} && \text{“A invertible $\Leftrightarrow \boldsymbol{a}^\top A\boldsymbol{a}\neq 0$ for all $\boldsymbol{a}\neq\mathbf{0}$”}\\
&= \mathbb{P}\biggl\{\min_{\boldsymbol{a}\in\mathbb{R}^p\setminus\{\mathbf{0}_p\}} \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{a}^\top\boldsymbol{z}^i{\boldsymbol{z}^i}^\top\boldsymbol{a} \;=\; 0\biggr\} && \text{“linearity of sums”}\\
&= \mathbb{P}\biggl\{\min_{\boldsymbol{a}\in\mathbb{R}^p\setminus\{\mathbf{0}_p\}} \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{a}^\top\boldsymbol{z}^i)^2 \;=\; 0\biggr\} && \text{“symmetry of inner products”}\\
&= \mathbb{P}\bigl\{\exists\boldsymbol{a}\in\mathbb{R}^p\setminus\{\mathbf{0}_p\}\;:\;\boldsymbol{a}^\top\boldsymbol{z}^i \;=\; 0\;\forall i\in\{1,\dots,n\}\bigr\}\\
& && \text{“all summands are non-negative”}\\
&= \mathbb{P}\bigl\{\exists\boldsymbol{a}\in\mathbb{R}^p\setminus\{\mathbf{0}_p\}\;:\;\boldsymbol{a}\perp\boldsymbol{z}^i \;=\; 0\;\forall i\in\{1,\dots,n\}\bigr\} && \text{“reformulation”}\\
&\leq \mathbb{P}\bigl\{\exists\boldsymbol{a}\in\mathbb{R}^p\setminus\{\mathbf{0}_p\}\;:\;\boldsymbol{a}\perp\boldsymbol{z}^i \;=\; 0\;\forall i\in\{1,\dots,p\}\bigr\}\\
& && \text{“$n\geq p$ by assumption and $\mathbb{P}\{\mathcal{A}\}\leq\mathbb{P}\{\mathcal{A}'\}$ if $\mathcal{A}\subset\mathcal{A}'$”}\\
&\leq \mathbb{P}\bigl\{\boldsymbol{z}^1,\dots,\boldsymbol{z}^p \text{ is not a basis of the } \mathbb{R}^p\bigr\}\\
& && \text{“by definition, bases span the entire ambient space”}\\
&= 0\,, && \text{“above insights”}
\end{aligned}$$

as desired.

Details: The fact that affine subspaces with codimension larger or equal to one have Gaussian measure zero can be proved as follows (an affine subspace that has dimension d has codimension $p-d$ with respect to the $\mathbb{R}^p$): Consider an affine subspace $\mathcal{A}$ of dimension k in the $\mathbb{R}^p$, $k<p$. There is a rotation R such that for any element $\boldsymbol{a}\in R\mathcal{A}$, it holds that $a_p=0$. Denoting the Lebesgue measure on $\mathbb{R}^p$ by $\mathbb{P}$, we then find

$$\begin{aligned}
\mathbb{P}\mathcal{A} &= \mathbb{P}R\mathcal{A} && \text{“the Lebesgue measure on $\mathbb{R}^p$ is rotation invariant (we omit that proof)”}\\
&\leq \mathbb{P}\{\boldsymbol{a}\in\mathbb{R}^p\;:\;a_p \;=\; 0\} && \text{“definition of the rotation R; $\mathbb{P}\{\mathcal{A}\}\leq\mathbb{P}\{\mathcal{A}'\}$ if $\mathcal{A}\subset\mathcal{A}'$”}\\
&= \mathbb{P}\bigcup_{i=1}^{\infty}\bigl\{\boldsymbol{a}\in\mathbb{R}^p\;:\;a_1,\dots,a_{p-1}\in[-i,i],a_p=0\bigr\} && \text{“reformulation of the event”}\\
&= \lim_{i\to\infty}\mathbb{P}\bigl\{\boldsymbol{a}\in\mathbb{R}^p\;:\;a_1,\dots,a_{p-1}\in[-i,i],a_p=0\bigr\} && \text{“measures are countably additive”}\\
&= \lim_{i\to\infty}\bigl([-i,i]^{p-1}\cdot 0\bigr)\\
& \text{“definition of the Lebesgue measure in [Dud02, Theorem 3.2.6 on Page 98 and Chapter 4.4 on Pages 134ff]”}\\
&= 0\,. && \text{“taking the limit”}
\end{aligned}$$

(Note that one cannot take boxes $\{\boldsymbol{a}\in\mathbb{R}^p : |a_p|\leq 1/i\}$ instead, because then, the stated definition of the Lebesgue measure does not apply.) Since the Lebesgue measure dominates all (non-degenerate) Gaussian distributions, this proves that affine subspace of dimension at most $p-1$ are Gaussian nullsets in $\mathbb{R}^p$.

5. In view of the formulation of the maximum likelihood estimator on Page 50, this claim follows directly from 3. and 4. when taking $A=\sum_{i=1}^{n}\boldsymbol{z}^i{\boldsymbol{z}^i}^\top/n$.

Solutions for Section 3.5

Solution 3.4 In essence, we reformulate basic properties of the least-squares solutions.

1. The proof of the first claim is a reorganization of the least-squares' KKT condition.

From the least-squares definition

$$\widehat{\boldsymbol{\beta}}^j \;\in\; \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^{p-1}} \|\boldsymbol{y}^j - X^j\boldsymbol{\alpha}\|_2^2\,,$$

we find by setting derivatives to zero (KKT conditions) that

$$\begin{aligned}
&\frac{\boldsymbol{\partial}}{\boldsymbol{\partial}\boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\beta}}^j}\|\boldsymbol{y}^j - X^j\boldsymbol{\alpha}\|_2^2 \;=\; \mathbf{0}_{p-1}\\
&\qquad\text{“KKT conditions”}\\
\Rightarrow\quad & -2(X^j)^\top\big(\boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j\big) \;=\; \mathbf{0}_{p-1}\\
&\qquad\text{“computing the derivative”}\\
\Rightarrow\quad & -2\big((\boldsymbol{z}^1)_{\{j\}^\complement},\cdots,(\boldsymbol{z}^n)_{\{j\}^\complement}\big)\left(\begin{pmatrix}(z^1)_j\\ \vdots\\ (z^n)_j\end{pmatrix} - \begin{pmatrix}\big((\boldsymbol{z}^1)_{\{j\}^\complement}\big)^\top\\ \vdots\\ \big((\boldsymbol{z}^n)_{\{j\}^\complement}\big)^\top\end{pmatrix}\cdot\widehat{\boldsymbol{\beta}}^j\right) \;=\; \mathbf{0}_{p-1}\\
&\qquad\text{“definition of the } \boldsymbol{y}^j\text{'s and } X^j\text{'s”}\\
\Rightarrow\quad & -2\big((\boldsymbol{z}^1)_{\{j\}^\complement},\cdots,(\boldsymbol{z}^n)_{\{j\}^\complement}\big)\left(-\begin{pmatrix}(\boldsymbol{z}^1)^\top\\ \vdots\\ (\boldsymbol{z}^n)^\top\end{pmatrix}\cdot\underline{\widehat{\boldsymbol{\beta}}}^j\right) \;=\; \mathbf{0}_{p-1}\\
&\qquad\text{“}(\underline{\widehat{\boldsymbol{\beta}}}^j)_j = -1\text{ by definition”}\\
\Rightarrow\quad & \frac{1}{n}\big((\boldsymbol{z}^1)_{\{j\}^\complement},\cdots,(\boldsymbol{z}^n)_{\{j\}^\complement}\big)\begin{pmatrix}(\boldsymbol{z}^1)^\top\\ \vdots\\ (\boldsymbol{z}^n)^\top\end{pmatrix}\cdot\underline{\widehat{\boldsymbol{\beta}}}^j \;=\; \mathbf{0}_{p-1}\,.\\
&\qquad\text{“consolidating and dividing both sides by } 2n\text{”}
\end{aligned}$$

Now, we find for any $k,l\in\{1,\dots,p\}$, $k\neq j$,

$$\begin{aligned}
\frac{1}{n}\left(\big((\boldsymbol{z}^1)_{\{j\}^\complement},\cdots,(\boldsymbol{z}^n)_{\{j\}^\complement}\big)\begin{pmatrix}(\boldsymbol{z}^1)^\top\\ \vdots\\ (\boldsymbol{z}^n)^\top\end{pmatrix}\right)_{kl} \;&=\; \frac{1}{n}\sum_{i=1}^n (z^i)_k(z^i)_l\\
&\qquad\text{“computing the matrix-matrix multiplication”}\\
&=\; \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}^i{\boldsymbol{z}^i}^\top\right)_{kl}\,.\\
&\qquad\text{“rewriting in terms of the observation vectors”}
\end{aligned}$$

Combining this with the penultimate display yields the desired claim.

2. This is a straightforward calculation.

We find

$$\begin{aligned}
\|\boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j\|_2^2 &= \langle \boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j, \boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j\rangle && \text{“definition of the } \ell_2\text{-norm”}\\
&= \langle \boldsymbol{y}^j, \boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j\rangle - \langle X^j\widehat{\boldsymbol{\beta}}^j, \boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j\rangle\\
& && \text{“linearity of the inner product”}\\
&= \langle \boldsymbol{y}^j, \boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j\rangle - \langle \frac{\widehat{\boldsymbol{\beta}}^j}{2}, 2(X^j)^\top(\boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j)\rangle\\
& && \text{“linearity of inner product; properties of the transpose”}\\
&= \langle \boldsymbol{y}^j, \boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j\rangle - \langle \frac{\widehat{\boldsymbol{\beta}}^j}{2}, \mathbf{0}_{p-1}\rangle\\
& && \text{“}2(X^j)^\top(\boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j) = \mathbf{0}_{p-1}\text{ by the KKT conditions for the least-squares—see 1. above”}\\
&= \langle \boldsymbol{y}^j, \boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j\rangle && \text{“linearity of the inner product”}\\
&= \sum_{i=1}^n (z^i)_j\Bigl((z^i)_j - \sum_{k\neq j}(z^i)_k(\underline{\widehat{\boldsymbol{\beta}}}^j)_k\Bigr)\\
& && \text{“definition of the } \boldsymbol{y}^j\text{'s, } X^j\text{'s, and } \underline{\widehat{\boldsymbol{\beta}}}^j\text{'s”}\\
&= \sum_{i=1}^n (z^i)_j\Bigl(-\sum_{k=1}^p(z^i)_k(\underline{\widehat{\boldsymbol{\beta}}}^j)_k\Bigr) && \text{“}(\underline{\widehat{\boldsymbol{\beta}}}^j)_j = -1\text{ by definition”}\\
&= -n\sum_{k=1}^p\Bigl(\frac{1}{n}\sum_{i=1}^n(z^i)_j(z^i)_k\Bigr)(\underline{\widehat{\boldsymbol{\beta}}}^j)_k && \text{“linearity of sums”}\\
&= -n\sum_{k=1}^p\Bigl(\frac{1}{n}\sum_{i=1}^n\boldsymbol{z}^i\boldsymbol{z}^{i\top}\Bigr)_{jk}(\underline{\widehat{\boldsymbol{\beta}}}^j)_k && \text{“as above”}\\
&= -n\Bigl(\frac{1}{n}\sum_{i=1}^n\boldsymbol{z}^i\boldsymbol{z}^{i\top}\Bigr)_{j\{1,\dots,p\}}\underline{\widehat{\boldsymbol{\beta}}}^j\,.\\
& && \text{“formulating in terms of matrix-vector multiplications”}
\end{aligned}$$

Dividing both sides by $-n$ yields the desired result.

3. We plug the results from above into the definition of the neighborhood selection estimator.

The two claims above can be combined into

$$\Bigl(\frac{1}{n}\sum_{i=1}^n\boldsymbol{z}^i\boldsymbol{z}^{i\top}\Bigr)\underline{\widehat{\boldsymbol{\beta}}}^j = -\frac{\|\boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j\|_2^2}{n}\boldsymbol{e}^j\,,$$

where $\boldsymbol{e}^j \in \mathbb{R}^p$ is the jth standard unit vector: $(e^j)_k = 1$ if $k = j$ and $(e^j)_k = 0$ otherwise. By assumption, we can invert the Gram matrix, so that the equation can be written as

$$\underline{\widehat{\boldsymbol{\beta}}}^j = -\frac{\|\boldsymbol{y}^j - X^j\widehat{\boldsymbol{\beta}}^j\|_2^2}{n}\Bigl(\frac{1}{n}\sum_{i=1}^n\boldsymbol{z}^i\boldsymbol{z}^{i\top}\Bigr)^{-1}\boldsymbol{e}^j\,.$$

With this in mind, we find for any $k,l \in \{1,\dots,p\}$ that

$$
\begin{aligned}
(\widehat{\Theta}_{\mathrm{ns}})_{kl} &= -\frac{\frac{n}{\|\boldsymbol{y}^k - X^k\widehat{\boldsymbol{\beta}}^k\|_2^2}(\underline{\widehat{\boldsymbol{\beta}}}^k)_l + \frac{n}{\|\boldsymbol{y}^l - X^l\widehat{\boldsymbol{\beta}}^l\|_2^2}(\underline{\widehat{\boldsymbol{\beta}}}^l)_k}{2} && \text{“Definition (3.4)”}\\
&= \frac{\left(\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}^i\boldsymbol{z}^{i\top}\right)^{-1}\boldsymbol{e}^k\right)_l + \left(\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}^i\boldsymbol{z}^{i\top}\right)^{-1}\boldsymbol{e}^l\right)_k}{2}\\
& && \text{“penultimate display”}\\
&= \frac{\left(\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}^i\boldsymbol{z}^{i\top}\right)^{-1}\right)_{lk} + \left(\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}^i\boldsymbol{z}^{i\top}\right)^{-1}\right)_{kl}}{2}\\
& && \text{“}\boldsymbol{e}^k,\boldsymbol{e}^l \text{ are the } k\text{th and } l\text{th unit vectors, respectively”}\\
&= \left(\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}^i\boldsymbol{z}^{i\top}\right)^{-1}\right)_{kl},\\
& && \text{“the Gram matrix (and, therefore, also it's inverse) are symmetric”}
\end{aligned}
$$

as desired.

Appendix B

Mathematical Background

B.1 Convex Analysis

Lemma B.1.1 (Exponents) For any $a, b \in \mathbb{R}$ and $t \geq 1$, it holds that

$$|a+b|^t \;\leq\; 2|a|^t + 2|b|^t\,.$$

This directly implies, for example, $\|\mathbf{v}+\mathbf{w}\|_2^2 \leq 2\|\mathbf{v}\|_2^2 + 2\|\mathbf{w}\|_2^2$ for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^p$.

Proof of Lemma B.1.1. The inequality is straightforward if $a = 0$ or $b = 0$; therefore, we can assume $a, b \neq 0$ without loss of generality, We can then also assume that $b = \max\{|a|, |b|\}$ (flip signs and/or interchange variables otherwise) and $|a| = 1/2$ (divide both sides by $2|a|$ otherwise) without loss of generality.

Now we find by virtue of the triangle inequality in the case $t = 1$

$$|a+b|^t \;\;=\;\; |1+b| \;\leq\; |a| + |b| \;\leq\; 2|a|^t + 2|b|^t\,.$$

Therefore, it suffices to show that the function $t \mapsto |a+b|^t - 2|a|^t - 2|b|^t$ is decreasing. For this, we take derivatives and find

$$\begin{aligned}
&\frac{\partial}{\partial t}\bigl(|a+b|^t - 2|a|^t - 2|b|^t\bigr)\\
&= \frac{\partial}{\partial t}\bigl(e^{t\log|a+b|} - e^{t\log(2|a|)} - e^{t\log(2|b|)}\bigr) && \text{“properties of log”}\\
&= e^{t\log|a+b|}\log|a+b| - e^{t\log(2|a|)}\log|a| - e^{t\log(2|b|)}\log(2|b|) && \text{“product rule”}\\
&= |a+b|^t\log|a+b| - (2|a|)^t\log(2|a|) - (2|b|)^t\log(2|b|) && \text{“properties of log”}\\
&= |a+b|^t\log|a+b| - (2|b|)^t\log(2|b|)\,. && \text{“$|a| = 1/2$, hence $\log(2|a|) = 0$”}
\end{aligned}$$

We can conclude by noting that by the above assumptions on a, b

$$|a+b| \;\leq\; |a| + |b| \;\leq\; 2|b|$$

and that $x\log x$ is an increasing function. □

Lemma B.1.2 (Decomposition of the Inner Product) For any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$ and $v > 0$, it holds that

$$-\frac{\|\boldsymbol{a}\|_2^2}{v} - \frac{v\|\boldsymbol{b}\|_2^2}{4} \;\leq\; \langle \boldsymbol{a}, \boldsymbol{b}\rangle \;\leq\; \frac{\|\boldsymbol{a}\|_2^2}{v} + \frac{v\|\boldsymbol{b}\|_2^2}{4}\,,$$

and similarly, for any $u > 0$,

$$-\frac{\|\boldsymbol{a}\|_2^2}{v} - \frac{u^2v\|\boldsymbol{b}\|_2^2}{4} \;\leq\; u\langle \boldsymbol{a}, \boldsymbol{b}\rangle \;\leq\; \frac{\|\boldsymbol{a}\|_2^2}{v} + \frac{u^2v\|\boldsymbol{b}\|_2^2}{4}\,.$$

Proof of Lemma B.1.2. We derive

$$\begin{aligned}
0 &\le \|\sqrt{2/v}\,\boldsymbol{a} - \sqrt{v/2}\,\boldsymbol{b}\|_2^2 && \text{“non-negativity of norms”}\\
&= \sum_{j=1}^{p}\bigl(\sqrt{2/v}\,a_j - \sqrt{v/2}\,b_j\bigr)^2 && \text{“definition of two-norm”}\\
&= \sum_{j=1}^{p}\bigl(2/v\cdot a_j^2 + v/2\cdot b_j^2 - 2a_jb_j\bigr) && \text{“expanding”}\\
&= 2/v\cdot\|\boldsymbol{a}\|_2^2 + v/2\cdot\|\boldsymbol{b}\|_2^2 - 2\langle\boldsymbol{a},\,\boldsymbol{b}\rangle\,.\\
& && \text{“definition of two-norms and standard inner product”}
\end{aligned}$$

Rearranging then yields

$$\langle\boldsymbol{a},\,\boldsymbol{b}\rangle \le \frac{\|\boldsymbol{a}\|_2^2}{v} + \frac{v\|\boldsymbol{b}\|_2^2}{4}$$

as desired for the first part.

For the second part, we find for any $v, v', u > 0$, $v' = uv$,

$$\begin{aligned}
u\langle\boldsymbol{a},\,\boldsymbol{b}\rangle &\le u\Bigl(\frac{\|\boldsymbol{a}\|_2^2}{v'} + \frac{v'\|\boldsymbol{b}\|_2^2}{4}\Bigr) && \text{“first part applied to } v'\text{”}\\
&\le \frac{\|\boldsymbol{a}\|_2^2}{v} + \frac{u^2v\|\boldsymbol{b}\|_2^2}{4}\,, && \text{“}v' := uv\text{”}
\end{aligned}$$

as desired.

The right-hand inequalities finally follow from setting $\boldsymbol{a} \to -\boldsymbol{a}$. □

Lemma B.1.3 (A General Hölder's Inequality) Consider a function $h : \mathbb{R}^p \to [-\infty, \infty]$ that satisfies

1. (positive definiteness) $h[\mathbf{a}] \ge 0$ for all $\mathbf{a} \in \mathbb{R}^p$ with equality if and only if $\mathbf{a} = \mathbf{0}_p$;
2. (homogeneity) $h[\mathbf{a}/h[\mathbf{a}]] \le 1$ for all $\mathbf{a} \in \mathbb{R}^p$.

Recall that the dual function $\overline{h} : \mathbb{R}^p \to [-\infty, \infty]$ of h is defined according to Page 4 via

$$\overline{h}[\mathbf{a}] := \sup\bigl\{\langle\mathbf{a},\,\mathbf{k}\rangle \;:\; \mathbf{k} \in \mathbb{R}^p,\, h[\mathbf{k}] \le 1\bigr\} \qquad (\,\mathbf{a} \in \mathbb{R}^p\,)\,.$$

Then, also $\overline{h}$ is positive definite and

$$\langle\mathbf{a},\,\mathbf{b}\rangle \le \overline{h}[\mathbf{a}]h[\mathbf{b}] \qquad (\,\mathbf{a}, \mathbf{b} \in \mathbb{R}^p\,)\,. \tag{B.1}$$

In particular, any norm on $\mathbb{R}^p$ is a valid function h. In contrast, a function that violates 1. is $h : \mathbf{a} \mapsto \mathbb{1}\{a_1, \dots, a_p \ge 0\}$; a function that violates 2. is $h : \mathbf{a} \mapsto \|\mathbf{a}\|_2^2$.

Proof of Lemma B.1.3. The proof is based on the definition of dual functions and on the linearity of inner products. The only subtlety is the calculus with infinite values.

We first prove that $\overline{h}$ is positive definite. For this, we observe that

$$\begin{aligned}
\overline{h}[\mathbf{0}_p] &= \sup\bigl\{\langle\mathbf{0}_p,\,\mathbf{k}\rangle \;:\; \mathbf{k} \in \mathbb{R}^p,\, h[\mathbf{k}] \le 1\bigr\} && \text{“definition of the dual function } \overline{h}\text{”}\\
&= \sup\bigl\{0 \;:\; \mathbf{k} \in \mathbb{R}^p,\, h[\mathbf{k}] \le 1\bigr\} && \text{“linearity of inner products”}\\
&= 0 && \text{“evaluating the supremum”}
\end{aligned}$$

and that for all $\mathbf{a} \neq \mathbf{0}_p$

$$\begin{aligned}\overline{h}[\mathbf{a}] &= \sup\bigl\{\langle \mathbf{a},\, \mathbf{k}\rangle \;:\; \mathbf{k} \in \mathbb{R}^p,\, h[\mathbf{k}] \leq 1\bigr\} && \text{“definition of the dual function } \overline{h}\text{”}\\ &\geq \langle \mathbf{a},\, \mathbf{a}/h[\mathbf{a}]\rangle\\ & && \text{“considering } \mathbf{k} = \mathbf{a}/h[\mathbf{a}];\ h[\mathbf{a}] > 0 \text{ by 1. (positive definitness of } h) \text{ and } h[\mathbf{a}/h[\mathbf{a}]] \leq 1 \text{ by 2. (homogeneity of } h)\text{”}\\ &= \frac{\langle \mathbf{a},\, \mathbf{a}\rangle}{h[\mathbf{a}]} && \text{“linearity of inner products”}\\ &> 0\,. && \text{“}\langle \mathbf{a},\, \mathbf{a}\rangle > 0 \text{ for all } \mathbf{a} \neq 0;\ h[\mathbf{a}] > 0 \text{ by 1. (positive definiteness of } h)\text{”}\end{aligned}$$

Hence, $\overline{h}[\mathbf{a}] = 0$ if and only if $\mathbf{a} \neq \mathbf{0}_p$, which means that h is positive definite.

We now prove Inequality (B.1). For this, we consider three cases.

Case 1: $h[\mathbf{b}] = 0$. In this case, the left-hand side of the desired inequality satisisfies

$$\begin{aligned}\langle \mathbf{a},\, \mathbf{b}\rangle &= \langle \mathbf{a},\, \mathbf{0}_p\rangle && \text{“}\mathbf{b} = \mathbf{0}_p \text{ by 1. (positive definitness of } h)\text{”}\\ &= 0\,. && \text{“linearity of inner products”}\end{aligned}$$

The right-hand side of the desired inequality satisfies

$$\begin{aligned}\overline{h}[\mathbf{a}]h[\mathbf{b}] &= \overline{h}[\mathbf{a}] \cdot 0 && \text{“}\mathbf{b} = \mathbf{0}_p \text{ by 1. (positive definitness of } h)\text{”}\\ &= 0\,. && \text{“note that } \infty \cdot 0 = 0 \text{ by the conventions on Page 4”}\end{aligned}$$

Hence, $\langle \mathbf{a},\, \mathbf{b}\rangle = \overline{h}[\mathbf{a}]h[\mathbf{b}]$, which implies the desired inequality (B.1).

Case 2: $h[\mathbf{b}] \neq 0$, $h[\mathbf{b}] \neq \pm\infty$. In this case, we use the linearity of inner products to write

$$\langle \mathbf{a},\, \mathbf{b}\rangle = \langle \mathbf{a},\, \mathbf{b}\rangle \cdot \frac{h[\mathbf{b}]}{h[\mathbf{b}]} = \langle \mathbf{a},\, \mathbf{b}/h[\mathbf{b}]\rangle h[\mathbf{b}]\,.$$

In view of 1. (which implies $h \geq 0$), we are then left with showing that $\langle \mathbf{a},\, \mathbf{b}/h[\mathbf{b}]\rangle \leq \overline{h}[\mathbf{a}]$. To this end, we note that the homogeneity of h assumed in 2. ensures that

$$h\left[\frac{\mathbf{b}}{h[\mathbf{b}]}\right] \leq 1\,.$$

Hence,

$$\begin{aligned}&\langle \mathbf{a},\, \mathbf{b}/h[\mathbf{b}]\rangle\\ &\leq \sup\bigl\{\langle \mathbf{a},\, \mathbf{k}\rangle \;:\; \mathbf{k} \in \mathbb{R}^p,\, h[\mathbf{k}] \leq 1\bigr\}\\ & && \text{“considering } \mathbf{k} = \mathbf{b}/h[\mathbf{b}];\ h[\mathbf{b}/h[\mathbf{b}]] \leq 1 \text{ by the previous display”}\\ &= \overline{h}[\mathbf{a}]\,, && \text{“definition of the dual function } \overline{h}\text{”}\end{aligned}$$

as desired.

Case 3: $h[\mathbf{b}] = \infty$. Recall that $h \geq 0$ by 1., which especially implies that $h[\mathbf{b}] \neq -\infty$; hence, Case 3 is the only missing piece in the proof.

If $\overline{h}[\mathbf{a}] \neq 0$ in this case, the right-hand side of the desired inequality is infinite (cf. the conventions on Page 4), which means that the inequality holds irrespective of the left-hand side. If $\overline{h}[\mathbf{a}] = 0$, the right-hand side of the desired inequality is zero (cf. the conventions on Page 4). However, since $\overline{h}$ is positive definite as shown above, it then also holds that $\mathbf{a} = \mathbf{0}_p$, and therefore, using the linearity of inner products, $\langle \mathbf{a},\, \mathbf{b}\rangle = 0$. Thus, $\langle \mathbf{a},\, \mathbf{b}\rangle = \overline{h}[\mathbf{a}]h[\mathbf{b}]$, which implies Inequality (B.1). This concludes the proof of Lemma B.1.3. □

Lemma B.1.4 (A General Cauchy-Schwarz Inequality) Let $\mathcal{V}$ be a vector space over the reals equipped with an inner product $(\mathbf{a}, \mathbf{b}) \mapsto \langle \mathbf{a},\, \mathbf{b}\rangle$ and the associated norm $\mathbf{a} \mapsto \|\mathbf{a}\| := \sqrt{\langle \mathbf{a},\, \mathbf{a}\rangle}$. Then,

$$\langle \mathbf{a},\, \mathbf{b}\rangle \leq \|\mathbf{a}\|\|\mathbf{b}\| \qquad (\mathbf{a}, \mathbf{b} \in \mathcal{V})\,.$$

Proof of Lemma B.1.4. The proof leverages the basic properties of inner products and norms.

In the case $\mathbf{b} = \mathbf{0}$, we have

$$\langle \mathbf{a},\, \mathbf{b}\rangle \;=\; \|\mathbf{a}\|\|\mathbf{b}\| \;=\; 0$$

by the linearity of inner products and the positive definitness of norms. Therefore, can assume without loss of generality that $\mathbf{b} \neq \mathbf{0}$, which is equivalent to $\|\mathbf{b}\| \neq 0$ by the positive definiteness of norms.

We can now apply the properties of inner products and norms:

$$\begin{aligned}
0 &\leq \|\mathbf{a} - \frac{\langle \mathbf{a},\, \mathbf{b}\rangle}{\|\mathbf{b}\|^2}\mathbf{b}\|^2 && \text{“algebra”}\\
&= \langle \mathbf{a} - \frac{\langle \mathbf{a},\, \mathbf{b}\rangle}{\|\mathbf{b}\|^2}\mathbf{b},\, \mathbf{a} - \frac{\langle \mathbf{a},\, \mathbf{b}\rangle}{\|\mathbf{b}\|^2}\mathbf{b}\rangle && \text{“definition of the norm”}\\
&= \langle \mathbf{a} - \frac{\langle \mathbf{a},\, \mathbf{b}\rangle}{\|\mathbf{b}\|^2}\mathbf{b},\, \mathbf{a}\rangle - \frac{\langle \mathbf{a},\, \mathbf{b}\rangle}{\|\mathbf{b}\|^2}\langle \mathbf{a} - \frac{\langle \mathbf{a},\, \mathbf{b}\rangle}{\|\mathbf{b}\|^2}\mathbf{b},\, \mathbf{b}\rangle && \text{“linearity of inner products”}\\
&= \langle \mathbf{a},\, \mathbf{a}\rangle - \frac{\langle \mathbf{a},\, \mathbf{b}\rangle}{\|\mathbf{b}\|^2}\langle \mathbf{b},\, \mathbf{a}\rangle - \frac{\langle \mathbf{a},\, \mathbf{b}\rangle}{\|\mathbf{b}\|^2}\langle \mathbf{a},\, \mathbf{b}\rangle + \frac{\langle \mathbf{a},\, \mathbf{b}\rangle^2}{\|\mathbf{b}\|^4}\langle \mathbf{b},\, \mathbf{b}\rangle && \text{“linearity of inner products”}\\
&= \langle \mathbf{a},\, \mathbf{a}\rangle - 2\frac{\langle \mathbf{a},\, \mathbf{b}\rangle^2}{\|\mathbf{b}\|^2} + \frac{\langle \mathbf{a},\, \mathbf{b}\rangle^2}{\|\mathbf{b}\|^4}\langle \mathbf{b},\, \mathbf{b}\rangle && \text{“symmetry of inner products”}\\
&= \|\mathbf{a}\|^2 - 2\frac{\langle \mathbf{a},\, \mathbf{b}\rangle^2}{\|\mathbf{b}\|^2} + \frac{\langle \mathbf{a},\, \mathbf{b}\rangle^2}{\|\mathbf{b}\|^2} && \text{“definition of the norm”}\\
&= \|\mathbf{a}\|^2 - \frac{\langle \mathbf{a},\, \mathbf{b}\rangle^2}{\|\mathbf{b}\|^2}\,. && \text{“consolidation”}
\end{aligned}$$

Rearraning the inequality and taking square-roots complete the proof. □

Definition B.1.1 (Subgradient) [1] Consider a convex function $f : \mathbb{R}^p \to \mathbb{R}$ and a vector $\beta \in \mathbb{R}^p$. We call a vector $z \in \mathbb{R}^p$ a subgradient of f at β if

$$f(\beta') \;\geq\; f(\beta) + \langle z,\, \beta' - \beta\rangle$$

for all $\beta' \in \mathbb{R}^p$. We call the set of all subgradients of f at β the subdifferential of f at β and denote this set by $\partial f(\beta)$.

Example B.1.1 (Differentiable) If f is convex and differentiable at β, it holds that $\partial f(\beta) = f'(\beta)$.

Example B.1.2 (ℓ_1-norm) If $p = 1$ and $f : \beta \mapsto |\beta|$, it holds that

$$\partial f(\beta) \;=\; \begin{cases} \{+1\} & \text{if } \beta \;>\; 0\\ [-1, 1] & \text{if } \beta \;=\; 0\\ \{-1\} & \text{if } \beta \;<\; 0 \end{cases}.$$

Similarly, for general $p \in \{1, 2, \dots\}$ and $f : \beta \mapsto \|\beta\|_1$,

$$\partial f(\beta) \;=\; \{z \;\in\; \mathbb{R}^p \;: \|z\|_\infty \;\leq\; 1 \;\text{ and }\; z_j \;=\; \operatorname{sign}(\beta_j) \text{ if } \beta_j \;\neq\; 0\}\,.$$

Example B.1.3 (Subdifferential of the Lasso) Consider a fixed $r \geq 0$ and

$$\begin{aligned} f \;:\; &\mathbb{R}^p \to \mathbb{R}\\ &\beta \mapsto \|\boldsymbol{y} - X\beta\|_2^2 + r\|\beta\|_1\,. \end{aligned}$$

[1] cf. [HTW15, Chapter 5.2.2, Pages 98-99]

Then,

$$\partial f(\beta) \;=\; \bigl\{-2X^\top(\boldsymbol{y}-X\beta)+\kappa \;:\; \kappa\in\partial\|\beta\|_1\bigr\}\;.$$

Lemma B.1.5 (Minima of Convex Functions/KKT Conditions) Consider a convex function $f:\mathbb{R}^p\to\mathbb{R}$ and a vector $\beta\in\mathbb{R}^p$. The vector β is a minimizer of f if and only if zero is in the subdifferential of f at β, that is,

$$f(\beta') \;\geq\; f(\beta)\;\;\forall\,\beta'\;\in\;\mathbb{R}^p \qquad\Leftrightarrow\qquad \mathbf{0}_p\;\in\;\partial f(\beta)\,.$$

Proof of Lemma B.1.5. We prove the two parts of the statement in turn.
Step 1: sufficiency. We first prove that

$$\mathbf{0}_p\;\in\;\partial f(\beta)\qquad\Rightarrow\qquad f(\beta')\;\geq\;f(\beta)\;\;\forall\,\beta'\;\in\;\mathbb{R}^p\,.$$

For this, we assume $\mathbf{0}_p\in\partial f(\beta)$, which leads by the Definition B.1.1 of the subdifferential to

$$f(\beta')\;\geq\;f(\beta)+\langle 0,\,\beta'-\beta\rangle\;=\;f(\beta)$$

for all $\beta'\in\mathbb{R}^p$. This concludes the sufficiency part of the proof.
Step 2: necessity. We now prove that

$$f(\beta')\;\geq\;f(\beta)\;\;\forall\,\beta'\;\in\;\mathbb{R}^p\qquad\Rightarrow\qquad\mathbf{0}_p\;\in\;\partial f(\beta)\,.$$

For this, assume $f(\beta')\geq f(\beta)$ for all $\beta'\in\mathbb{R}^p$. Then, by linearity of the inner product,

$$f(\beta')\;\geq\;f(\beta)\;=\;f(\beta)+\langle 0,\,\beta'-\beta\rangle\,,$$

so that $\mathbf{0}_p\in\partial f(\beta)$ by Definition B.1.1 of the subdifferential. This concludes the necessity part of the proof. □

Example B.1.4 (Minima of the Lasso) For the lasso as above,

$$\hat\beta\;\in\;\operatorname*{argmin}_{\beta\in\mathbb{R}^p}\bigl\{\|\boldsymbol{y}-X\beta\|_2^2+r\|\beta\|_1\bigr\}$$

is equivalent to

$$-2X^\top(\boldsymbol{y}-X\hat\beta)+r\hat\kappa\;=\;\mathbf{0}_p$$

for a $\hat\kappa$ with $\|\hat\kappa\|_\infty\leq 1$ and $\hat\kappa_j\;=\;\operatorname{sign}(\hat\beta_j)$ if $\hat\beta_j\;\neq\;0$. We call this the KKT conditions for the lasso.

Lemma B.1.6 (ℓ_2-Projections) Let $\mathcal{A}\subset\mathbb{R}^n$ be an arbitrary set. Then, $P_{\mathcal{A}}(\boldsymbol{a})\in\mathcal{A}$ is an ℓ_2-projection of $\boldsymbol{a}\in\mathbb{R}^n$ on $\mathcal{A}$, that is,

$$\|P_{\mathcal{A}}(\boldsymbol{a})-\boldsymbol{a}\|_2^2\;=\;\min_{\boldsymbol{b}\in\mathcal{A}}\|\boldsymbol{b}-\boldsymbol{a}\|_2^2\,,$$

if and only if

$$\langle\boldsymbol{b}-P_{\mathcal{A}}(\boldsymbol{a}),\,P_{\mathcal{A}}(\boldsymbol{a})-\boldsymbol{a}\rangle\;\geq\;-\|\boldsymbol{b}-P_{\mathcal{A}}(\boldsymbol{a})\|_2^2/2$$

for all $\boldsymbol{b}\in\mathcal{A}$.

[2]Moreover, if $\mathcal{A}$ is convex, $P_{\mathcal{A}}(\boldsymbol{a})\in\mathcal{A}$ is an ℓ_2-projection of $\boldsymbol{a}\in\mathbb{R}^n$ on $\mathcal{A}$ if and only if

$$\langle\boldsymbol{b}-P_{\mathcal{A}}(\boldsymbol{a}),\,P_{\mathcal{A}}(\boldsymbol{a})-\boldsymbol{a}\rangle\;\geq\;0$$

for all $\boldsymbol{b}\in\mathcal{A}$.

[3]Finally, if $\mathcal{A}$ is convex and $P_{\mathcal{A}}(\boldsymbol{a}),P_{\mathcal{A}}(\boldsymbol{b})\in\mathcal{A}$ are ℓ_2-projections of $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^n$, respectively, on $\mathcal{A}$, it holds that

$$\|P_{\mathcal{A}}(\boldsymbol{a})-P_{\mathcal{A}}(\boldsymbol{b})\|_2^2\leq\langle P_{\mathcal{A}}(\boldsymbol{a})-P_{\mathcal{A}}(\boldsymbol{b}),\,\boldsymbol{a}-\boldsymbol{b}\rangle\,.$$

[2] cf. [HUL13, Theorem 3.1.1. on Page 117]
[3] cf. [HUL13, Proposition 3.1.3. on Page 118]

Note that in this sort of results, it is typically assumed that $\mathcal{A}$ is closed, non-empty, convex to ensure that the projections are well defined; we instead treated this issue by our definition of projections (first display of the lemma) that implicitly assumes existence.

Proof of Lemma B.1.6. The proof is a simple algebra exercise. However, we only proof the first part and refer to [HUL13] for the other two parts.

We start by noting that if $P_{\mathcal{A}}(\boldsymbol{a})$ is an ℓ_2-projection of $\boldsymbol{a}$ on $\mathcal{A}$, it holds by assumption that

$$\|P_{\mathcal{A}}(\boldsymbol{a})-\boldsymbol{a}\|_2^2 \leq \|\boldsymbol{b}-\boldsymbol{a}\|_2^2$$

for all $\boldsymbol{b}\in\mathbb{R}^n$. Adding a zero-valued term on the right-hand side yields

$$\|P_{\mathcal{A}}(\boldsymbol{a})-\boldsymbol{a}\|_2^2 \leq \|\boldsymbol{b}-P_{\mathcal{A}}(\boldsymbol{a})+P_{\mathcal{A}}(\boldsymbol{a})-\boldsymbol{a}\|_2^2$$

and expanding subsequently

$$\|P_{\mathcal{A}}(\boldsymbol{a})-\boldsymbol{a}\|_2^2 \leq \|\boldsymbol{b}-P_{\mathcal{A}}(\boldsymbol{a})\|_2^2+2\langle\boldsymbol{b}-P_{\mathcal{A}}(\boldsymbol{a}),\,P_{\mathcal{A}}(\boldsymbol{a})-\boldsymbol{a}\rangle+\|P_{\mathcal{A}}(\boldsymbol{a})-\boldsymbol{a}\|_2^2\,.$$

We can rewrite this as

$$\langle\boldsymbol{b}-P_{\mathcal{A}}(\boldsymbol{a}),\,P_{\mathcal{A}}(\boldsymbol{a})-\boldsymbol{a}\rangle \geq -\|\boldsymbol{b}-P_{\mathcal{A}}(\boldsymbol{a})\|_2^2/2\,,$$

which coincides with the desired inequality. □

B.1.1 Exercises

Exercise B.1.1 Proof the remaining parts of Lemma B.1.6.

B.2 Matrix Algebra

B.2.1 Basics

Lemma B.2.1 (Determinants) Let $M,N\in\mathbb{R}^{p\times p}$ be two square matrices. Then, 1. $\det[MN]=\det[M]\det[N]$, 2. $\det[M^\top]=\det[M]$, 3. $\det[\mathrm{I}_{p\times p}]=1$, and 4. $\det[M]\neq 0$ if and only if M is invertible.

The proof of Lemma B.2.1 is omitted; we refer to standard textbooks on linear algebra.

Lemma B.2.2 (Determinant of a Special Block Matrix) For any square matrix $M\in\mathbb{R}^{(p+q)\times(p+q)}$ of the form

$$M=\begin{bmatrix}\mathrm{I}_{p\times p} & U\\ \boldsymbol{0}_{q\times p} & V\end{bmatrix}$$

with $U\in\mathbb{R}^{p\times q}$, $V\in\mathbb{R}^{q\times q}$, it holds that $\det[M]=\det[V]$.

Similarly, for any square matrix $M\in\mathbb{R}^{(q+p)\times(q+p)}$ of the form

$$M=\begin{bmatrix}U & V\\ \boldsymbol{0}_{q\times p} & \mathrm{I}_{q\times q}\end{bmatrix}$$

with $U\in\mathbb{R}^{p\times p}$, $V\in\mathbb{R}^{p\times q}$, it holds that $\det[M]=\det[U]$.

Proof of Lemma B.2.2. The proof consists of rewriting Leipniz' formula by exploiting the special form of M.

Recall that the Leipniz formula for the determinant of a square matrix $M \in \mathbb{R}^{(p+q)\times(p+q)}$ reads

$$\det[M] \;=\; \sum_{\sigma\in\mathcal{S}^{p+q}} \operatorname{sign}[\sigma] \prod_{i=1}^{p+q} M_{i\sigma[i]}\,,$$

where $\mathcal{S}^{p+q}$ is the set of permutations of $\{1,\dots,p+q\}$, and where the sign function returns $+1$ for even and -1 for odd permutations.

The upper left part of M is an identity matrix; more precisely, $M_{i\sigma[i]}=1$ if $\sigma[i]=i$ and $M_{i\sigma[i]}=0$ otherwise for all $i\in\{1,\dots,p\}$ and $\sigma\in\mathcal{S}^{p+q}$. For the previous display, this means that the sum can be restricted to permutations that keep the first p indexes unchanged, or stated differently, it is sufficient to consider permutations of the last q indexes. In mathematical terms,

$$\det[M] \;=\; \sum_{\sigma\in\mathcal{S}^{q}} \operatorname{sign}[\sigma] \prod_{i=1}^{q} M_{(p+i)(p+\sigma[i])}\,,$$

where $\mathcal{S}^q$ is the set of permutations of $\{1,\dots,q\}$.

The lower right part of M is V; more precisely, $M_{(p+i)(p+\sigma[i])} = V_{i\sigma[i]}$ for all $i\in\{1,\dots,q\}$ and $\sigma\in\mathcal{S}^q$. For the previous display, this means that

$$\det[M] \;=\; \sum_{\sigma\in\mathcal{S}^{q}} \operatorname{sign}[\sigma] \prod_{i=1}^{q} V_{i\sigma[i]}\,.$$

We finally invoke again Leipniz' formula, but this time for the matrix V, to deduce from the display that $\det[M]=\det[V]$, as desired.

The second statement follows in the same way. □

Lemma B.2.3 (Submatrices) Let $M\in\mathbb{R}^{p\times p}$ be an invertible matrix and $A\in\mathbb{R}^{q\times q}$ a submatrix of M, $q\in\{1,\dots,p\}$. (This means in particular that that there is a permutation π of $\{1,\dots,p\}$ such that $A_{ij}=M_{\pi[i]\pi[j]}$.) Then, A is invertible.

Lemma B.2.4 (Inverse Matrices) Let $M\in\mathbb{R}^{p\times p}$ be an invertible matrix. Then, 1. if M is symmetric, then M^{-1} is also symmetric, 2. $\det[M^{-1}]=1/\det[M]$, and 3. $(M^\top)^{-1}=(M^{-1})^\top$.

Proof of Lemma B.2.4. For 1., note first that for any two square matrices $U,V\in\mathbb{R}^{p\times p}$, it holds that $(UV)^\top=V^\top U^\top$. Indeed, by direct calculation, we find

$$(UV)^\top_{ij} \;=\; (UV)_{ji} \;=\; \sum_{k=1}^{p} U_{jk}V_{ki} \;=\; \sum_{k=1}^{p} V_{ki}U_{jk} \;=\; \sum_{k=1}^{p} V^\top_{ik}U^\top_{kj} \;=\; (V^\top U^\top)_{ij}$$

for each $i,j\in\{1,\dots,p\}$. We now calculate

$$\begin{aligned}
M^{-1} &= M^{-1}\,\mathrm{I}_{p\times p} && \text{(property of identity)}\\
&= M^{-1}\,\mathrm{I}^\top_{p\times p} && \text{(property of identity)}\\
&= M^{-1}(M^{-1}M)^\top && (M \text{ invertible})\\
&= M^{-1}M^\top(M^{-1})^\top && \text{(Step 1)}\\
&= M^{-1}M(M^{-1})^\top && (M \text{ symmetric})\\
&= \mathrm{I}_{p\times p}(M^{-1})^\top && (M \text{ invertible})\\
&= (M^{-1})^\top && \text{(property of identity)}
\end{aligned}$$

as desired.

For 2., we invoke 1. and 2. in Lemma B.2.1 to derive

$$\det(M)\det(M^{-1}) \;=\; \det(MM^{-1}) \;=\; \det(\mathrm{I}_{p\times p}) \;=\; 1$$

and then 2. in Lemma B.2.1 to conclude the proof. □

Definition B.2.1 (Diagonal Matrices and Orthogonal Matrices) We call a matrix $D \in \mathbb{R}^{n\times p}$ diagonal if $D_{ij} = 0$ for all $i \in \{1,\dots,n\}$, $j \in \{1,\dots,p\}$, $i \neq j$. We call an invertible matrix $U \in \mathbb{R}^{n\times n}$ orthogonal if $U^{-1} = U^\top$.

Lemma B.2.5 (Singular Value Decomposition) We can write any matrix $A \in \mathbb{R}^{n\times p}$ as

$$A \;=\; UDV^\top$$

with $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{p\times p}$ orthogonal and $D \in \mathbb{R}^{n\times p}$ diagonal. We call such an expansion a *singular value decomposition* of A.

The proof can be found in any advanced book on linear algebra.

Definition B.2.2 (Moore-Penrose Inverse) A matrix $A^+ \in \mathbb{R}^{p\times n}$ is called a *Moore-Penrose inverse* of $A \in \mathbb{R}^{n\times p}$ if the following four conditions are met:

1. $AA^+A = A$;
2. $A^+AA^+ = A^+$;
3. $(AA^+)^\top = AA^+$;
4. $(A^+A)^\top = A^+A$.

B.2.2 Results for Gaussian Graphical Modeling

Lemma B.2.6 (Matrix Algebra) Consider a symmetric, invertible matrix square matrix $M \in \mathbb{R}^{(p+q)\times(p+q)}$ of the form

$$M \;=\; \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}$$

with $A \in \mathbb{R}^{p\times p}, B \in \mathbb{R}^{p\times q}, C \in \mathbb{R}^{q\times q}$. We write its inverse M^{-1} (which is also symmetric by 1. in Lemma B.2.4) in the form

$$M^{-1} \;=\; \begin{bmatrix} \tilde{A} & \tilde{B} \\ \tilde{B}^\top & \tilde{C} \end{bmatrix}$$

with $\tilde{A} \in \mathbb{R}^{p\times p}, \tilde{B} \in \mathbb{R}^{p\times q}, \tilde{C} \in \mathbb{R}^{q\times q}$. It then holds that (i)

$$C^{-1} \;=\; \tilde{C} - \tilde{B}^\top \tilde{A}^{-1}\tilde{B}$$

and (ii)

$$\frac{\det[C]}{\det[M]} \;=\; \frac{1}{\det[\tilde{A}^{-1}]}\,.$$

Proof of Lemma B.2.6. We only need elementary matrix calculus and some results from Section B.2.1 about basic matrix algebra.

For proving (i), we first calculate

$$MM^{-1} \;=\; \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}\begin{bmatrix} \tilde{A} & \tilde{B} \\ \tilde{B}^\top & \tilde{C} \end{bmatrix} \;=\; \begin{bmatrix} A\tilde{A} + B\tilde{B}^\top & A\tilde{B} + B\tilde{C} \\ B^\top\tilde{A} + C\tilde{B}^\top & B^\top\tilde{B} + C\tilde{C} \end{bmatrix}.$$

Since $MM^{-1} = \mathrm{I}_{(p+q)\times(p+q)}$, the displays yields in particular

$$B^\top \tilde{A} + C\tilde{B}^\top \;=\; \mathbf{0}_{q\times p}$$

and

$$B^\top \tilde{B} + C\tilde{C} \;=\; \mathrm{I}_{q\times q}\;.$$

Since $\tilde{A}$ and C are submatrices of M^{-1} and M, respectively, they are also invertible according to Lemma B.2.3. We can, therefore, multiply the first equation by $\tilde{A}^{-1}$ from the right and the second equation by C^{-1} from the left to find

$$B^\top + C\tilde{B}^\top \tilde{A}^{-1} \;=\; \mathbf{0}_{q\times p}$$

and

$$C^{-1}B^\top \tilde{B} + \tilde{C} \;=\; C^{-1}\,,$$

respectively. The rest is then a simple calculation:

$$\begin{aligned}
C^{-1} &= \tilde{C} + C^{-1}B^\top \tilde{B} && \text{“previous equation”}\\
&= \tilde{C} + C^{-1}(-C\tilde{B}^\top \tilde{A}^{-1})\tilde{B} && \text{“substituting } B^\top \text{ via the penultimate equation”}\\
&= \tilde{C} - \tilde{B}^\top \tilde{A}^{-1}\tilde{B}\,, && \text{“}C^{-1}(-C) = -\mathrm{I}_{q\times q}\text{”}
\end{aligned}$$

as desired.

For proving (ii), we first calculate

$$\begin{aligned}
\det[M^{-1}] &= \det\begin{bmatrix}\tilde{A} & \tilde{B}\\ \tilde{B}^\top & \tilde{C}\end{bmatrix} && \text{“writing } M^{-1} \text{ as stated in the lemma”}\\
&= \det\left[\begin{bmatrix}\tilde{A} & \mathbf{0}_{p\times q}\\ \mathbf{0}_{q\times p} & \mathrm{I}_{q\times q}\end{bmatrix}\begin{bmatrix}\mathrm{I}_{p\times p} & \tilde{A}^{-1}\tilde{B}\\ \tilde{B}^\top & \tilde{C}\end{bmatrix}\right]\\
&&& \text{“separating the entire matrix into two factors”}\\
&= \det\left[\begin{bmatrix}\tilde{A} & \mathbf{0}_{p\times q}\\ \mathbf{0}_{q\times p} & \mathrm{I}_{q\times q}\end{bmatrix}\begin{bmatrix}\mathrm{I}_{p\times p} & \mathbf{0}_{p\times q}\\ \tilde{B}^\top & \mathrm{I}_{q\times q}\end{bmatrix}\begin{bmatrix}\mathrm{I}_{p\times p} & \tilde{A}^{-1}\tilde{B}\\ \mathbf{0}_{q\times p} & \tilde{C} - \tilde{B}^\top \tilde{A}^{-1}\tilde{B}\end{bmatrix}\right]\\
&&& \text{“separating the second argument further into two factors”}\\
&= \det\begin{bmatrix}\tilde{A} & \mathbf{0}_{p\times q}\\ \mathbf{0}_{q\times p} & \mathrm{I}_{q\times q}\end{bmatrix}\det\begin{bmatrix}\mathrm{I}_{p\times p} & \mathbf{0}_{p\times q}\\ \tilde{B}^\top & \mathrm{I}_{q\times q}\end{bmatrix}\det\begin{bmatrix}\mathrm{I}_{p\times p} & \tilde{A}^{-1}\tilde{B}\\ \mathbf{0}_{q\times p} & \tilde{C} - \tilde{B}^\top \tilde{A}^{-1}\tilde{B}\end{bmatrix}\\
&&& \text{“using 1. in Lemma B.2.1 twice”}\\
&= \det\begin{bmatrix}\tilde{A} & \mathbf{0}_{p\times q}\\ \mathbf{0}_{q\times p} & \mathrm{I}_{q\times q}\end{bmatrix}\det\begin{bmatrix}\mathrm{I}_{p\times p} & \tilde{B}^\top\\ \mathbf{0}_{p\times q} & \mathrm{I}_{q\times q}\end{bmatrix}\det\begin{bmatrix}\mathrm{I}_{p\times p} & \tilde{A}^{-1}\tilde{B}\\ \mathbf{0}_{q\times p} & \tilde{C} - \tilde{B}^\top \tilde{A}^{-1}\tilde{B}\end{bmatrix}\\
&&& \text{“using 2. in Lemma B.2.1 for the second determinant”}\\
&= \det[\tilde{A}]\det[\mathrm{I}_{q\times q}]\det[\tilde{C} - \tilde{B}^\top \tilde{A}^{-1}\tilde{B}]\\
&\text{“using the second part of Lemma B.2.2 for the first term and the first part for the other two terms”}\\
&= \det[\tilde{A}]\cdot 1\cdot \det[\tilde{C} - \tilde{B}^\top \tilde{A}^{-1}\tilde{B}] && \text{“using 3. in Lemma B.2.1”}\\
&= \det[\tilde{A}]\det[C^{-1}]\,. && \text{“Part (i)”}
\end{aligned}$$

Hence, using 2. in Lemma B.2.4, we find

$$\frac{1}{\det[M]} \;=\; \frac{1}{\det[\tilde{A}^{-1}]\det[C]}\,,$$

which can be rearranged to

$$\frac{\det[C]}{\det[M]} \;=\; \frac{1}{\det[\tilde{A}^{-1}]}\,,$$

as desired □

Lemma B.2.7 (Parameters in Neighborhood Selection) With the notation on Page 52, it holds that

$$\Theta_{ij} \;=\; -\frac{\Theta_{ii}(\underline{\boldsymbol{\beta}}^i)_j + \Theta_{jj}(\underline{\boldsymbol{\beta}}^j)_i}{2}\,.$$

Proof of Lemma B.2.7. The proof is simple yet slightly tedious algebra.

We first derive

$$\begin{aligned}
&& \boldsymbol{\beta}^j \;&=\; -\,(\Theta_{j\{j\}^\complement})^\top/\Theta_{jj} && \text{“by definition in Lemma 3.5.1”}\\
\Rightarrow\quad && (\Theta_{j\{j\}^\complement})^\top \;&=\; -\,\Theta_{jj}\boldsymbol{\beta}^j &&\\
&&&& \text{“multiplying through by } \Theta_{jj}\text{; } \Theta_{jj}>0 \text{ since } \Theta\in\mathcal{S}_p^+ \text{ by assumption”}\\
\Rightarrow\quad && \begin{pmatrix}\Theta_{j1}\\ \vdots\\ \Theta_{j(j-1)}\\ \Theta_{j(j+1)}\\ \vdots\\ \Theta_{jp}\end{pmatrix} \;&=\; -\,\Theta_{jj}\begin{pmatrix}(\boldsymbol{\beta}^j)_1\\ \vdots\\ (\boldsymbol{\beta}^j)_{j-1}\\ (\boldsymbol{\beta}^j)_j\\ \vdots\\ (\boldsymbol{\beta}^j)_{p-1}\end{pmatrix} && \text{“writing previous equation more explicit”}\\
\Rightarrow\quad && \begin{pmatrix}\Theta_{j1}\\ \vdots\\ \Theta_{j(j-1)}\\ \Theta_{j(j+1)}\\ \vdots\\ \Theta_{jp}\end{pmatrix} \;&=\; -\,\Theta_{jj}\begin{pmatrix}(\underline{\boldsymbol{\beta}}^j)_1\\ \vdots\\ (\underline{\boldsymbol{\beta}}^j)_{j-1}\\ (\underline{\boldsymbol{\beta}}^j)_{j+1}\\ \vdots\\ (\underline{\boldsymbol{\beta}}^j)_p\end{pmatrix} && \text{“definition of } \underline{\boldsymbol{\beta}}^j\text{”}\\
\Rightarrow\quad && \begin{pmatrix}\Theta_{j1}\\ \vdots\\ \Theta_{j(j-1)}\\ \Theta_{jj}\\ \Theta_{j(j+1)}\\ \vdots\\ \Theta_{jp}\end{pmatrix} \;&=\; -\,\Theta_{jj}\begin{pmatrix}(\underline{\boldsymbol{\beta}}^j)_1\\ \vdots\\ (\underline{\boldsymbol{\beta}}^j)_{j-1}\\ (\underline{\boldsymbol{\beta}}^j)_j\\ (\underline{\boldsymbol{\beta}}^j)_{j+1}\\ \vdots\\ (\underline{\boldsymbol{\beta}}^j)_p\end{pmatrix} && \text{“}(\underline{\boldsymbol{\beta}}^j)_j=-1\text{ by definition”}\\
\Rightarrow\quad && \Theta_{ji} \;&=\; -\,\Theta_{jj}(\underline{\boldsymbol{\beta}}^j)_i\,. && \text{“taking the } i\text{th coordinate on either side”}
\end{aligned}$$

Similarly, $\Theta_{ji} = -\Theta_{ii}(\underline{\boldsymbol{\beta}}^i)_j$, and since Θ is symmetric by assumption, $\Theta_{ij} = \Theta_{ji}$. Combining these three observations yields

$$\Theta_{ij} \;=\; -\frac{\Theta_{ii}(\underline{\boldsymbol{\beta}}^i)_j + \Theta_{jj}(\underline{\boldsymbol{\beta}}^j)_i}{2}\,,$$

as desired. □

B.3 Probability and Measure Theory

Lemma B.3.1 (Law of Total Variance) Consider two random variables X, Y that satisfy (i) $\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y|X]\big]$, (ii) $\mathbb{E}[Y^2] = \mathbb{E}\big[\mathbb{E}[Y^2|X]\big]$, and (iii) $\mathbb{E}\Big[\big(\mathbb{E}[Y|X]\big)^2\Big] = \mathbb{E}\Big[\mathbb{E}\big[\big(\mathbb{E}[Y|X]\big)^2\big|X\big]\Big]$ (iterated expectations). Then,

$$\mathbb{V}\mathrm{ar}[Y] \;=\; \mathbb{E}\big[\mathbb{V}\mathrm{ar}[Y|X]\big] + \mathbb{V}\mathrm{ar}\big[\mathbb{E}[Y|X]\big]\,.$$

Proof of Lemma B.3.1. The proof is a simple algebra exercise. We find

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}[Y] &= \mathbb{E}\Big[\big(Y-\mathbb{E}[Y]\big)^2\Big] && \text{“definition of variance”}\\
&= \mathbb{E}\Big[Y^2-2Y\mathbb{E}[Y]+\big(\mathbb{E}[Y]\big)^2\Big] && \text{“expanding”}\\
&= \mathbb{E}\big[Y^2\big]-\big(\mathbb{E}[Y]\big)^2 && \text{“linearity of integrals”}\\
&= \mathbb{E}\Big[\mathbb{E}\big[Y^2|X\big]\Big]-\big(\mathbb{E}\big[\mathbb{E}[Y|X]\big]\big)^2 && \text{“iterated expectations (i) and (ii)”}\\
&= \mathbb{E}\Big[\mathbb{E}\big[Y^2|X\big]\Big]-\mathbb{E}\Big[\big(\mathbb{E}[Y|X]\big)^2\Big]+\mathbb{E}\Big[\big(\mathbb{E}[Y|X]\big)^2\Big]-\big(\mathbb{E}\big[\mathbb{E}[Y|X]\big]\big)^2\\
& && \text{“adding a zero-valued term”}\\
&= \mathbb{E}\Big[\mathbb{E}\big[Y^2|X\big]\Big]-\mathbb{E}\Big[\mathbb{E}\big[\big(\mathbb{E}[Y|X]\big)^2|X\big]\Big]+\mathbb{E}\Big[\big(\mathbb{E}[Y|X]\big)^2\Big]-\big(\mathbb{E}\big[\mathbb{E}[Y|X]\big]\big)^2\\
& && \text{“iterated expectations (iii)”}\\
&= \mathbb{E}\Big[\mathbb{E}\big[Y^2-\big([\mathbb{E}[Y|X]\big)^2|X\big]\Big]+\mathbb{E}\Big[\big(\mathbb{E}[Y|X]\big)^2-\big(\mathbb{E}\big[\mathbb{E}[Y|X]\big]\big)^2\Big]\\
& && \text{“linearity of conditional expectations and expectations”}\\
&= \mathbb{E}\Big[\mathbb{E}\big[\big(Y-\mathbb{E}[Y|X]\big)^2|X\big]\Big]+\mathbb{E}\Big[\big(\mathbb{E}[Y|X]-\mathbb{E}\big[\mathbb{E}[Y|X]\big]\big)^2\Big]\\
& && \text{“linearity of conditional expectations”}\\
&= \mathbb{E}\big[\mathbb{V}\mathrm{ar}[Y|X]\big]+\mathbb{V}\mathrm{ar}\big[\mathbb{E}[Y|X]\big] && \text{“definition of variance and conditional variance”}
\end{aligned}
$$

as desired. □

Lemma B.3.2 (Gaussian Marginals) Consider $\boldsymbol{z}\sim\mathcal{N}_p(\mathbf{a},\Psi)$ and the index set $\mathcal{B}:=\{k+1,\dots,p\}$ for given $k\in\{1,\dots,p-1\}$. Then, $\boldsymbol{z}_{\mathcal{B}}\sim\mathcal{N}_{p-k}(\mathbf{a}_{\mathcal{B}},\Psi_{\mathcal{B}\mathcal{B}})$.

The proof is left as an exercise.

Lemma B.3.3 (Gaussian Tails) For all $t\geq 0$, it holds that

$$1-\breve{g}(t)\;\leq\;\frac{e^{-t^2/2}}{\sqrt{\pi}}\,.$$

Sharper bounds are known, but the bound above is sufficient for most purposes.

Proof of Lemma B.3.3. The proof is an elementary algebra exercise.

Part 1: We first show that for all $t\in[0,1/\sqrt{2}]$

$$1-\breve{g}(t)-\frac{e^{-t^2/2}}{\sqrt{\pi}}\;\leq\;0\,.$$

We first observe that by definition of $\breve{g}$,

$$1-\breve{g}(0)-\frac{e^{-0^2/2}}{\sqrt{\pi}}\;=\;1-\frac{1}{2}-\frac{1}{\sqrt{\pi}}\;\approx\;-0.06\;\leq\;0\,.$$

Now, for the derivatives, we find

$$\begin{aligned}
&\frac{d}{dt}\big(1-\breve{g}(t)-\frac{e^{-t^2/2}}{\sqrt{\pi}}\big)\\
&=\frac{d}{dt}\big(1-\int_{-\infty}^{t}\frac{e^{-x^2/2}}{\sqrt{\pi}}\,dx-\frac{e^{-t^2/2}}{\sqrt{\pi}}\big) && \text{“definition of } \breve{g}\text{”}\\
&=-\frac{e^{-t^2/2}}{\sqrt{2\pi}}+\frac{te^{-t^2/2}}{\sqrt{\pi}} && \text{“fundamental theorem of differentiation and integration”}\\
&=\big(t-\frac{1}{\sqrt{2}}\big)\frac{e^{-t^2/2}}{\sqrt{\pi}} && \text{“algebra”}\\
&\leq 0\,. && \text{“}t\leq 1/\sqrt{2}\text{”}
\end{aligned}$$

Combining the two observations concludes the proof of Part 1.

Part 2: We now show that also for all $t\in[1/\sqrt{2},\infty)$

$$1-\breve{g}(t)-\frac{e^{-t^2/2}}{\sqrt{\pi}}\;\leq\;0\,.$$

We find

$$\begin{aligned}
&1-\breve{g}(t)-\frac{e^{-t^2/2}}{\sqrt{\pi}}\\
&\leq 1-\breve{g}(t)-\frac{e^{-t^2/2}}{t\sqrt{2\pi}} && \text{“}t\geq 1/\sqrt{2}\text{ by assumption”}\\
&=\int_{t}^{\infty}\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx-\frac{e^{-t^2/2}}{t\sqrt{2\pi}} && \text{“definition of } \breve{g}\text{ and integration rules”}\\
&\leq\int_{t}^{\infty}\frac{xe^{-x^2/2}}{t\sqrt{2\pi}}\,dx-\frac{e^{-t^2/2}}{t\sqrt{2\pi}} && \text{“}x\geq t\text{ under the integral”}\\
&\leq\int_{t}^{\infty}\frac{d}{dx}\Big(-\frac{e^{-x^2/2}}{t\sqrt{2\pi}}\Big)\,dx-\frac{e^{-t^2/2}}{t\sqrt{2\pi}} && \text{“differentiation laws”}\\
&=\frac{e^{-t^2/2}}{t\sqrt{2\pi}}-\frac{e^{-t^2/2}}{t\sqrt{2\pi}} && \text{“fundamental law of integration and differentiation”}\\
&=0 && \text{“algebra”}
\end{aligned}$$

as desired. □

B.3.1 Exercises

Exercise B.3.1 Formulate sufficient conditions for the assumptions of Lemma B.3.1 to hold.

Exercise B.3.2 Prove Lemma B.3.2.

Index

Bibliography

[AC10] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[AP12] Suhani H Almal and Harish Padh. Implications of gene copy-number variation in health and diseases. *Journal of human genetics*, 57(1):6, 2012.

[Bak99] S. Bakin. Adaptive regression and model selection in data mining problems. 1999.

[BC$^+$13] Alexandre Belloni, Victor Chernozhukov, et al. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

[BCW11] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[BGd08] Onureena Banerjee, Laurent El Ghaoui, and Alexandre dAspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

[BK] C. Borgelt and R. Kruse. *Graphical Models – Methods for Data Analysis and Mining*. Wiley.

[BKRW93] P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, 1993.

[BL17] Y. Bu and J. Lederer. Integrating additional knowledge into estimation of graphical models. *arXiv:1704.02739*, 2017.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[BLS14a] F. Bunea, J. Lederer, and Y. She. The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms. *IEEE Trans. Inform. Theory*, 60(2):1313–1325, 2014.

[BLS14b] Florentina Bunea, Johannes Lederer, and Yiyuan She. The group square-root lasso: Theoretical properties and fast algorithms. *IEEE Transactions on Information Theory*, 60(2):1313–1325, 2014.

[BRT09] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[BT16] P. Bellec and A. Tsybakov. Bounds on the prediction error of penalized least squares estimators with convex penalty. *arXiv:1609.06675*, 2016.

[BW] Jacob Bien and Marten Wegkamp. Discussion of correlated variables in regression: Clustering and sparse estimation.

[CDS01] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[Cel08] A. Celisse. *Model selection via cross-validation in density estimation, regression, and change-points detection.* PhD thesis, Université Paris Sud-Paris XI, 2008.

[Che95] Scott Shaobing Chen. Basis pursuit, 1995.

[CJ15] S. Chatterjee and J. Jafarov. Prediction error of cross-validated lasso. *arXiv:1502.06291*, 2015.

[CT$^+$07] Emmanuel Candes, Terence Tao, et al. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[Daw] P. Dawkins. Pauls online math notes.

[DB04] Marcel Dettling and Peter Bühlmann. Finding predictive gene groups from microarray data. *J. Multivariate Anal.*, 90(1):106–131, 2004.

[DC05] Jana Diesner and Kathleen Carley. Exploration of communication networks from the Enron email corpus. In *SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA*, pages 3–14, 2005.

[DET06] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2006.

[DHJ$^+$04] Adrian Dobra, Chris Hans, Beatrix Jones, Joseph R Nevins, Guang Yao, and Mike West. Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.*, 90(1):196–212, 2004.

[DHL17] Arnak S Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581, 2017.

[Dud02] R. Dudley. *Real Analysis and Probability*, volume 74. Cambridge University Press, 2002.

[Dur10] R. Durrett. *Probability: Theory and Examples.* Cambridge University Press, 2010.

[Edw12] D. Edwards. *Introduction to Graphical Modelling.* Springer, 2012.

[EHJT04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[FF93] LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[FHT08] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[FL01] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[Gal13] Giovanni Gallavotti. *Statistical mechanics: A short treatise.* Springer, 2013.

[GHW79] G. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[Gir14] C. Giraud. *Introduction to high-dimensional statistics.* CRC Press, 2014.

[GLT] D. Gold, J. Lederer, and J. Tau. Inference for high-dimensional nested regression.

[GR04] E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.

[Hef] J. Hefferon. Linear algebra.

[HK70] A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[HL13] M. Hebiri and J. Lederer. How Correlations Influence Lasso Prediction. *IEEE Trans. Inform. Theory*, 59(3):1846–1854, 2013.

[HM13a] Darren Homrighausen and Daniel McDonald. The lasso, persistence, and cross-validation. In *International Conference on Machine Learning*, pages 1031–1039, 2013.

[HM13b] Darren Homrighausen and Daniel J McDonald. Risk-consistency of cross-validation with lasso-type procedures. *arXiv:1308.0810*, 2013.

[HM14] Darren Homrighausen and Daniel J McDonald. Leave-one-out cross-validation is risk consistent for lasso. *Machine learning*, 97(1-2):65–78, 2014.

[HTW15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations.* Chapman and Hall, 2015.

[HUL13] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms*, volume 1. Springer, 2013.

[Jän94] K. Jänich. Linear algebra. *Springer*, 1994.

[JM14] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[JWHT13] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning.* Springer, 2013.

[KCD$^+$08] Jeffrey Kidd, Gregory Cooper, William Donahue, Hillary Hayden, Nick Sampas, Tina Graves, Nancy Hansen, Brian Teague, Can Alkan, Francesca Antonacci, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64, 2008.

[KCO08] Yongdai Kim, Hosik Choi, and Hee-Seok Oh. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673, 2008.

[KMM$^+$15] Z. Kurtz, C. Müller, E. Miraldi, D. Littman, M. Blaser, and R. Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.*, 11(5):e1004226, 2015.

[Kol09] V. Koltchinskii. Sparsity in penalized empirical risk minimization. 45(1):7–57, 2009.

[Kol11] V. Koltchinskii. *Introduction.* Springer, 2011.

[Lau96] S. Lauritzen. *Graphical Models*, volume 17 of *Oxford Statistical Science Series.* The Clarendon Press, Oxford University Press, New York, 1996. Oxford Science Publications.

[LC98] E. Lehmann and G. Casella. *Theory of point estimation*, volume 31. Springer, 1998. second edition.

[Led13] Johannes Lederer. Trust, but verify: benefits and pitfalls of least-squares refitting in high dimensions. *arXiv:1306.0113*, 2013.

[Lif12] M. Lifshits. Lectures on gaussian processes. In *Lectures on Gaussian Processes*, pages 1–117. Springer, 2012.

[LW17] P.-L. Loh and M. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

[LYG18] J. Lederer, L. Yu, and I. Gaynanova. Oracle inequalities for high-dimensional prediction. *Bernoulli*, 2018.

[MB06] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

[Mei07] Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.

[Mei13] Nicolai Meinshausen. Sign-constrained least squares estimation for high-dimensional regression. *Electronic Journal of Statistics*, 7:1607–1631, 2013.

[NYWR12] S. Negahban, B. Yu, M. Wainwright, and P. Ravikumar. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27:538–557, 2012.

[OJV11] Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv:1110.0413*, 2011.

[OPT00] M. Osborne, B. Presnell, and B. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

[PC08] T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[SFHT13] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[Spi06] M. Spivak. Calculus, 2006.

[SZ12] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 58(1):267–288, 1996.

[Tik43] A. Tikhonov. On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198, 1943.

[van07] S. van de Geer. The deterministic lasso. JSM Proceedings, 2007.

[vB09] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[vdG16] S. van de Geer. *Estimation and testing under sparsity.* Springer, 2016.

[vdGB11] S. van der Geer and P. Bühlmann. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, 2011.

[VdGBR$^+$14] Sara Van de Geer, Peter Bühlmann, Yaacov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[vL13] S. van de Geer and J Lederer. The lasso, correlated design, and improved oracle inequalities. pages 303–316. 2013.

[Wai14] M. Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and Its Application*, 1:233–253, 2014.

[YL06] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[YL07] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[ZH05] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[Zha10] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[ZI77] Jean-Bernard Zuber and Claude Itzykson. Quantum field theory and the two-dimensional Ising model. *Phys. Rev. D*, 15(10):2875, 1977.

[ZZ$^+$12] Cun-Hui Zhang, Tong Zhang, et al. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.