

Springer Texts in Statistics

Johannes Lederer

Fundamentals of High-Dimensional Statistics

With Exercises and R Labs



Springer

Springer Texts in Statistics

Series Editors

G. Allen, Department of Statistics, Houston, USA

R. De Veaux, Department of Mathematics and Statistics,
Williams College, Williamstown, USA

R. Nugent, Department of Statistics,
Carnegie Mellon University, Pittsburgh, USA

Springer Texts in Statistics (STS) includes advanced textbooks from 3rd- to 4th-year undergraduate courses to 1st- to 2nd-year graduate courses. Exercise sets should be included. The series editors are currently Genevera I. Allen, Richard D. De Veaux, and Rebecca Nugent. Stephen Fienberg, George Casella, and Ingram Olkin were editors of the series for many years.

More information about this series at <http://www.springer.com/series/417>

Johannes Lederer

Fundamentals of High-Dimensional Statistics

With Exercises and R Labs



Springer

Johannes Lederer
Statistics, Machine Learning
& Data Science
Ruhr-University Bochum
Bochum, Germany

ISSN 1431-875X

ISSN 2197-4136 (electronic)

Springer Texts in Statistics

ISBN 978-3-030-73791-7

ISBN 978-3-030-73792-4 (eBook)

<https://doi.org/10.1007/978-3-030-73792-4>

© The Editor(s) (if applicable) and The Author(s), under exclusive licence to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my family

Preface

This textbook is designed for beginning graduate and advanced undergraduate students in statistics, biostatistics, and bioinformatics, but it may also be useful to a broader audience. Particular emphasis is put on

- Step-by-step introductions to the mathematical tools and principles
- Exercises that complement the main text, many of them with detailed solutions
- Computer labs that convey practical insights and experience
- Suggestions for further reading

This approach should give the reader a smooth start in the field.

I am grateful to Dr. Marco Rossini for the inspiring discussions about drafts of this book and about statistics in general. I also thank Yannick Düren, Shih-Ting Huang, Dr. Tobias Kaufmann, Janosch Kellermann, Mike Laszkiewicz, Mahsa Taheri, and Dr. Fang Xie for their valuable suggestions and corrections.

Johannes Lederer
Bochum, Germany
January 2020

Exercises, Labs, and Literature

In addition to the main text, the book contains exercises, labs, and literature notes. The diamond ratings \diamond/\blacklozenge , $\diamond\diamond/\blacklozenge\blacklozenge$, $\diamond\diamond\diamond/\blacklozenge\blacklozenge\blacklozenge$ next to an exercise number indicate the difficulty of the problem and if solutions are provided: the more diamonds, the harder or longer the solution of an exercise, and filled diamonds mean that there are solutions at the back of the book (however, I still recommend strongly to attempt all 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 for free 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 the manuals of the various functions, you can use the Help panel. ■ Fig. 1 shows how the lab exercises look like (top panel) and how to solve them (bottom panel). The labs are interpreted with R version 3.5.2; your outputs might slightly differ if you use a different version of R (especially the `set.seed()` function changed with version 3.6.0).

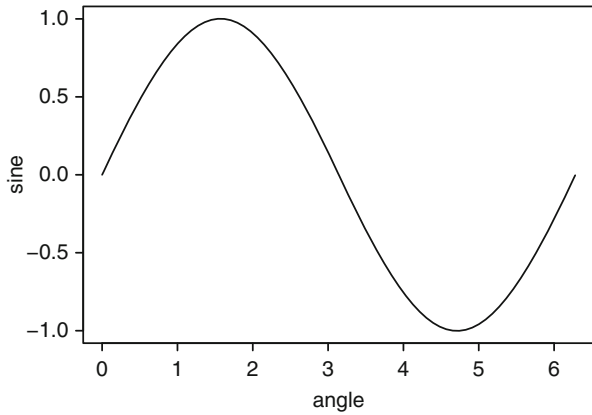
Further notes and references are indicated by numbered superscripts (such as “[...] copy-number variation (CNV).²”) in the main text and stated in the Notes and References sections toward the end of each chapter.

Finally, we denote sections that can be skipped at first reading with an asterisk (such as “2.4 Hölder Inequality*”).

■ **Fig. 1** Example R lab (top panel) and corresponding solution (bottom panel). The reader is supposed to replace the keyword `REPLACE` by the correct code

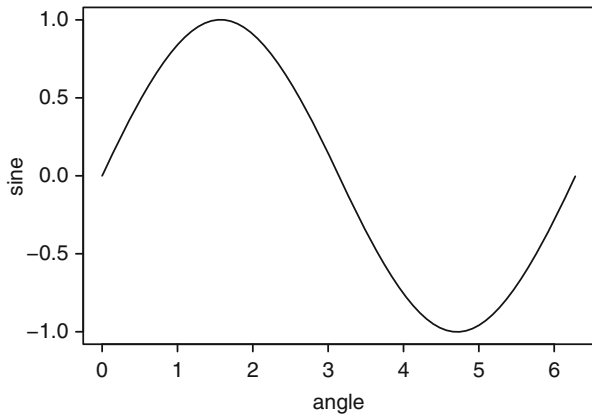
Plot the `sine()` function from 0 to 2π .

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



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="sine")
```



Notation

Here, we introduce some notation that we will use throughout the book.

Basic Quantities Lowercase letters a denote numbers; calligraphic lowercase letters \mathcal{f} real-valued functions; boldface lowercase letters \mathbf{a} (column) vectors; boldface, calligraphic lowercase letters \boldsymbol{f} vector-valued functions; capital letters A matrices; calligraphic capital 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* $\mathbf{a} \in \mathbb{R}^p$ is denoted by $\text{supp}[\mathbf{a}] := \{j \in \{1, \dots, p\} : a_j \neq 0\}$. Minima over the empty set are set to infinity: $\min_{\mathbf{a} \in \emptyset} \mathcal{f}[\mathbf{a}] := \infty$ for every function \mathcal{f} . The *signum function* $\text{sign} : \mathbb{R} \rightarrow \mathbb{R}$ is defined via $\text{sign}[a] := \mathbb{1}\{a \geq 0\} - \mathbb{1}\{a \leq 0\}$. The *cardinality of a set* \mathcal{A} , that is, the number of elements in \mathcal{A} , is denoted by $|\mathcal{A}| \in \{0, 1, \dots, \infty\}$. The sum of two sets \mathcal{A}, \mathcal{B} that are defined over the same vector space is $\mathcal{A} + \mathcal{B} := \{\mathbf{a} + \mathbf{b} : \mathbf{a} \in \mathcal{A}, \mathbf{b} \in \mathcal{B}\}$, and $\mathbf{a}\mathcal{B} := \{\mathbf{a}\mathbf{b} : \mathbf{b} \in \mathcal{B}\}$.

Norms and the Standard Inner Product A function $\|\cdot\| : \mathbb{R}^p \rightarrow \mathbb{R}$ is called a *norm* on \mathbb{R}^p if it 1. satisfies the *triangle inequality* ($\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$ for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$), 2. is *absolutely homogenous* ($\|a\mathbf{b}\| = |a|\|\mathbf{b}\|$ for all $a \in \mathbb{R}, \mathbf{b} \in \mathbb{R}^p$), and 3. is *positive definite* ($\|\mathbf{a}\| = 0$ if and only if $\mathbf{a} = \mathbf{0}_p$).

Assumptions 1–3 imply that norms are non-negative: $0 = \|\mathbf{0}_p\| = \|\mathbf{a} - \mathbf{a}\| \leq \|\mathbf{a}\| + \|-\mathbf{a}\| = \|\mathbf{a}\| + |-1|\|\mathbf{a}\| = 2\|\mathbf{a}\|$, that is, $\|\mathbf{a}\| \geq 0$ for all $\mathbf{a} \in \mathbb{R}^p$; Assumption 2 implies that norms are symmetric: $\|-\mathbf{a}\| = \|(-1) \cdot \mathbf{a}\| = |-1|\|\mathbf{a}\| = \|\mathbf{a}\|$; Assumption 2 also implies that norms are scalable: $\|\mathbf{b}/\|\mathbf{b}\|\| = \|\mathbf{b}\|/\|\mathbf{b}\| = 1$ for all $\mathbf{b} \neq \mathbf{0}_p$. 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

$$\ell_q : \mathbb{R}^p \rightarrow [0, \infty);$$

$$\mathbf{a} \mapsto \|\mathbf{a}\|_q := \left(\sum_{j=1}^p |a_j|^q \right)^{1/q},$$

for $q = 0$ as

$$\ell_0 : \mathbb{R}^p \rightarrow \{0, 1, \dots\};$$

$$\mathbf{a} \mapsto \|\mathbf{a}\|_0 := |\{j \in \{1, \dots, p\} : a_j \neq 0\}|,$$

and for $q = \infty$ as

$$\ell_\infty : \mathbb{R}^p \rightarrow [0, \infty);$$

$$\mathbf{a} \mapsto \|\mathbf{a}\|_\infty := \max_{j \in \{1, \dots, p\}} |a_j|.$$

The ℓ_q -functions are norms if and only if $q \geq 1$ (see 1. in Exercise 2.2); accordingly, we often refer to those functions as ℓ_q -norms. The ℓ_2 -norm is also called *Euclidean norm*; the ℓ_∞ -norm is also called *sup-norm* or *max-norm*.

The (standard) *inner product* on \mathbb{R}^p is the function $\langle \cdot, \cdot \rangle : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ defined through $\langle \mathbf{a}, \mathbf{b} \rangle := \mathbf{a}^\top \mathbf{b} = \sum_{j=1}^p a_j b_j$ for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$. The inner product is 1. *symmetric* ($\langle \mathbf{a}, \mathbf{b} \rangle = \langle \mathbf{b}, \mathbf{a} \rangle$ for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$), 2. *linear* ($\langle \mathbf{a}b, \mathbf{c} \rangle = a \langle \mathbf{b}, \mathbf{c} \rangle$ and $\langle \mathbf{a} + \mathbf{b}, \mathbf{c} \rangle = \langle \mathbf{a}, \mathbf{c} \rangle + \langle \mathbf{b}, \mathbf{c} \rangle$ for all $a \in \mathbb{R}, \mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^p$), and 3. *positive definite* ($\langle \mathbf{a}, \mathbf{a} \rangle \geq 0$ for all $\mathbf{a} \in \mathbb{R}^p$ and $\langle \mathbf{a}, \mathbf{a} \rangle = 0$ if and only if $\mathbf{a} = \mathbf{0}_p$). Two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ are *orthogonal* if $\langle \mathbf{a}, \mathbf{b} \rangle = 0$.

Intervals and the Extended Real Line We denote by $[a, b]$ the interval between $a \in \mathbb{R}$ and $b \in \mathbb{R}$ that contains the endpoints ($a, b \in [a, b]$), by $[a, b)$ and $(a, b]$ the intervals between a and b that contain the left end right endpoint, respectively ($a \in [a, b)$, $b \notin [a, b)$, $a \notin (a, b]$, $b \in (a, b]$), and by (a, b) the interval between a and b that does not contain the endpoints ($a, b \notin (a, b)$).

The real line extended by $\{-\infty, +\infty\}$ is denoted by $[-\infty, \infty] := \mathbb{R} \cup \{-\infty, +\infty\}$. Similarly, $[0, \infty] := [0, \infty) \cup \{\infty\}$, $(0, \infty] := (0, \infty) \cup \{\infty\}$, and so forth. We use the conventions $0 \cdot (\pm\infty) := (\pm\infty) \cdot 0 := 0$, $a/(\pm\infty) := 0$ for $a \in \mathbb{R}$, $a \cdot (\pm\infty) := (\pm\infty) \cdot a := \pm\infty$ for $a \in (0, \infty]$, and $a \cdot (\pm\infty) := (\pm\infty) \cdot a := \mp\infty$ for $a \in [-\infty, 0)$, which are all continuous extensions 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 $0/0 = \lim_{a \rightarrow 0} (a/a) = 1$ and at the same time $0/0 = (2 \cdot 0)/0 = 2 \cdot (0/0) = 2$, which is a contradiction). The ordering of the values in $[-\infty, \infty]$ is as expected: for example, $a < \infty$ for all $a \in [-\infty, \infty)$.

Index Sets and Matrices The complement of a set \mathcal{A} with respect to an ambient set \mathcal{B} is denoted by $\mathcal{A}^c := \mathcal{B} \setminus \mathcal{A}$. For example, the complement of $\{1, 2\}$ with respect to $\{1, \dots, p\}$, where $p \in \{3, 4, \dots\}$, is $\{3, \dots, p\}$. Typically, it is clear what the ambient set is, so that there is no further mention of it.

Consider a vector $\mathbf{c} \in \mathbb{R}^l$ and a corresponding index set $\mathcal{A} \subset \{1, \dots, l\}$ with size $a := |\mathcal{A}|$. We denote $\mathbf{c}_{\mathcal{A}} \in \mathbb{R}^a$ as the vector that consists of the coordinates of \mathbf{c} with indexes in \mathcal{A} . For example, for $\mathbf{c} = (3, 4, 5)^\top$ and $\mathcal{A} = \{1, 3\}$, it holds that $\mathbf{c}_{\mathcal{A}} = (3, 5)^\top$. The special case $\mathcal{A} = \emptyset$ is taken into account by setting $\mathbf{c}_{\emptyset} := 0$.

Consider a matrix $C \in \mathbb{R}^{l \times m}$ and corresponding index sets $\mathcal{A} \subset \{1, \dots, l\}$ and $\mathcal{B} \subset \{1, \dots, m\}$ with sizes $a := |\mathcal{A}|$ and $b := |\mathcal{B}|$, respectively. We denote $C_{\mathcal{A}} \in \mathbb{R}^{l \times a}$ as the matrix that consists of the columns of C with indexes in \mathcal{A} , and we denote $C_{\mathcal{B}\mathcal{A}} \in \mathbb{R}^{l \times m}$ as the matrix that consists of the rows and columns of C with indexes in \mathcal{B} and \mathcal{A} , respectively. For example,

$$C = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \mathcal{A} = \{2, 3\}, \mathcal{B} = \{1\} \Rightarrow C_{\mathcal{A}} = \begin{pmatrix} 2 & 3 \\ 5 & 6 \end{pmatrix}, C_{\mathcal{B}\mathcal{A}} = (2 \ 3).$$

However, we typically assume that the coordinates of the vectors/the rows and columns of the matrices are shuffled such that $\mathcal{A} = \{1, \dots, a\}$ and $\mathcal{B} = \{1, \dots, b\}$. This allows us to write, for example, $\mathbf{c} = (\mathbf{c}_{\mathcal{A}}^{\top}, \mathbf{c}_{\mathcal{A}^c}^{\top})^{\top}$ and

$$C = \begin{pmatrix} C_{\mathcal{B}\mathcal{A}} & C_{\mathcal{B}\mathcal{A}^c} \\ C_{\mathcal{B}^c\mathcal{A}} & C_{\mathcal{B}^c\mathcal{A}^c} \end{pmatrix}.$$

We finally use the convention $C_{\mathcal{A}}^{\top} := (C_{\mathcal{A}})^{\top}$.

A brief review of matrix algebra can be found in ► Sect. B.2.

Miscellaneous The expression $\mathbf{z} \sim \mathcal{N}_p[\boldsymbol{\mu}, \Sigma]$, $\boldsymbol{\mu} \in \mathbb{R}^p$, $\Sigma \in \mathbb{R}^{p \times p}$, states that \mathbf{x} is a random vector that follows a Gauss distribution in p dimensions with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ . In $p = 1$ dimensions, we write $z \sim \mathcal{M}[\mu, \sigma^2]$, where $\mu \in \mathbb{R}$ is the mean and $\sigma^2 \in (0, \infty)$ the variance.

Given a positive integer $p \in \{1, 2, \dots\}$, we define $\mathbf{0}_p := (0, \dots, 0)^{\top} \in \mathbb{R}^p$.

Contents

1	Introduction	1
1.1	Embracing High-Dimensionality	2
1.2	Statistical Limitations of Classical Estimators	5
1.3	Incorporating Prior Information	10
1.4	Regularization for Increasing the Numerical Stability*	13
1.5	Outlook	18
1.6	Exercises	19
1.7	R Lab: Least-Squares vs. Ridge Estimation	25
1.8	Notes and References	34
2	Linear Regression	37
2.1	Overview	38
2.2	Sparsity-Inducing Prior Functions	41
2.3	Post-Processing Methods	45
2.4	Hölder Inequality*	48
2.5	Optimality Conditions*	53
2.6	Exercises	63
2.7	R Lab: Overfitting	71
2.8	Notes and References	77
3	Graphical Models	81
3.1	Overview	82
3.2	Gaussian Graphical Models	85
3.3	Maximum Regularized Likelihood Estimation	87
3.4	Neighborhood Selection	91
3.5	Exercises	95
3.6	R Lab: Estimating a Gene–Gene Coactivation Network	98
3.7	Notes and References	107
4	Tuning-Parameter Calibration	109
4.1	Overview	110
4.2	Bounds on the Lasso’s Effective Noise	113
4.3	Cross-Validation	116
4.4	Adaptive Validation	121
4.5	Exercises	127

4.6	R Lab: Cross-Validation	131
4.7	Notes and References	135
5	Inference	139
5.1	One-Step Estimators	140
5.2	Confidence Intervals	146
5.3	Exercises	153
5.4	R Lab: Confidence Intervals in Low and High Dimensions	157
5.5	Notes and References	166
6	Theory I: Prediction	169
6.1	Overview	170
6.2	Basic Inequalities	175
6.3	Prediction Guarantees	182
6.4	Prediction Guarantees for Sparse and Weakly Correlated Models	192
6.5	Exercises	204
6.6	Notes and References	208
7	Theory II: Estimation and Support Recovery	211
7.1	Overview	212
7.2	Estimation Guarantees in ℓ_1-loss	213
7.3	Estimation Guarantees in $\overline{\ell}_1$-loss	218
7.4	Support Recovery Guarantees	231
7.5	Exercises	235
7.6	Notes and References	237
	Supplementary Information	239
A	Solutions	240
A.1	Solutions for ► Chap. 1	240
A.2	Solutions for ► Chap. 2	257
A.3	Solutions for ► Chap. 3	271
A.4	Solutions for ► Chap. 4	285
A.5	Solutions for ► Chap. 5	288
A.6	Solutions for ► Chap. 6	295
A.7	Solutions for ► Chap. 7	300
B	Mathematical Background	309
B.1	Analysis	309
B.2	Matrix Algebra	312
	Bibliography	338
	Index	345