Springer Texts in Statistics

Johannes Lederer

Fundamentals of High-Dimensional Statistics

With Exercises and R Labs

Springer

Springer Texts in Statistics

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

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

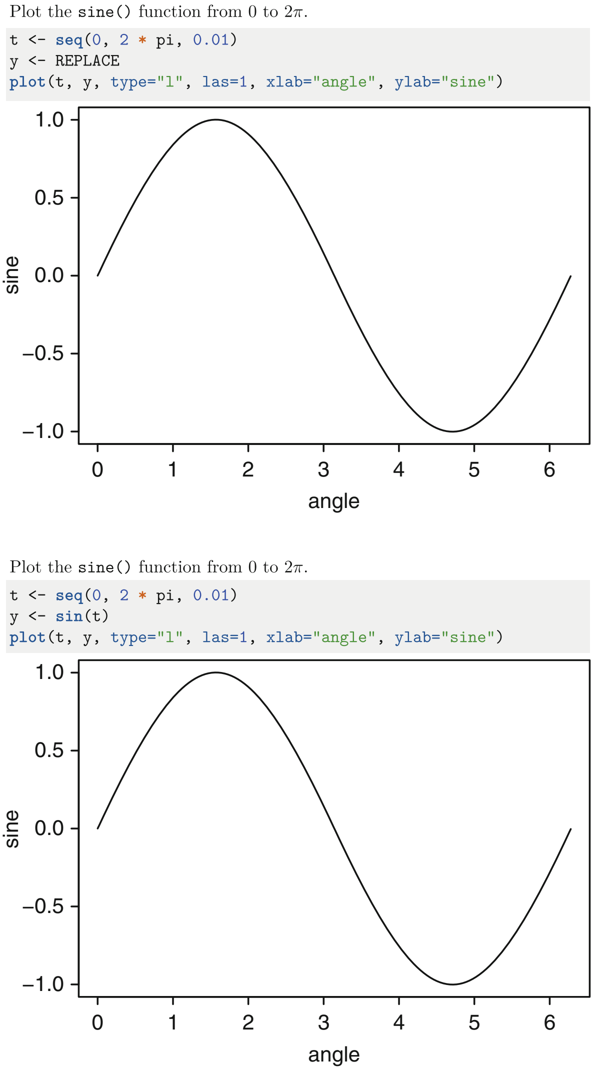
In addition to the main text, the book contains exercises, labs, and literature notes. The diamond ratings ◊/♦, ◊◊/♦♦, ◊◊◊/♦♦♦ next to an exercise number indicate the difficulty of the problem and if solutions are provided: the more diamonds, the harder or longer the solution of an exercise, and filled 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 $\mathscr{f}$ real-valued functions; boldface lowercase letters $\boldsymbol{a}$ (column) vectors; boldface, calligraphic lowercase letters $\boldsymbol{\mathscr{f}}$ vector-valued functions; capital letters A matrices; calligraphic capital letters $\mathscr{A}$ 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 $\operatorname{supp}[\boldsymbol{a}] := \{j \in \{1, \ldots, p\} : a_j \neq 0\}$. Minima over the empty set are set to infinity: $\min_{\boldsymbol{a}\in\varnothing} \mathscr{f}[\boldsymbol{a}] := \infty$ for every function $\mathscr{f}$. The *signum function* sign : $\mathbb{R} \to \mathbb{R}$ is defined via $\operatorname{sign}[a] := \mathbb{1}\{a \geq 0\} - \mathbb{1}\{a \leq 0\}$. The *cardinality of a set* $\mathscr{A}$, that is, the number of elements in $\mathscr{A}$, is denoted by $|\mathscr{A}| \in \{0, 1, \ldots, \infty\}$. The sum of two sets $\mathscr{A}$, $\mathscr{B}$ that are defined over the same vector space is $\mathscr{A} + \mathscr{B} := \{\boldsymbol{a} + \boldsymbol{b} : \boldsymbol{a} \in \mathscr{A},\ \boldsymbol{b} \in \mathscr{B}\}$, and $a\mathscr{B} := \{a\boldsymbol{b} : \boldsymbol{b} \in \mathscr{B}\}$.

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

Assumptions 1–3 imply that norms are non-negative: $0 = \|\mathbf{0}_p\| = \|\boldsymbol{a} - \boldsymbol{a}\| \leq \|\boldsymbol{a}\| + \|-\boldsymbol{a}\| = \|\boldsymbol{a}\| + |-1|\|\boldsymbol{a}\| = 2\|\boldsymbol{a}\|$, that is, $\|\boldsymbol{a}\| \geq 0$ for all $\boldsymbol{a} \in \mathbb{R}^p$; Assumption 2 implies that norms are symmetric: $\|-\boldsymbol{a}\| = \|(-1) \cdot \boldsymbol{a}\| = |-1|\|\boldsymbol{a}\| = \|\boldsymbol{a}\|$; Assumption 2 also implies that norms are scalable: $\|\boldsymbol{b}/\|\boldsymbol{b}\|\| = \|\boldsymbol{b}\|/\|\boldsymbol{b}\| = 1$ for all $\boldsymbol{b} \neq \mathbf{0}_p$. The ℓ_q-functions on $\mathbb{R}^p$, where $q \in [0, \infty]$ and $p \in \{1, 2, \ldots\}$, are defined for $q \in (0, \infty)$ as

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

for $q = 0$ as

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

and for $q = \infty$ as

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

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

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 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). 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 $\mathscr{A}$ with respect to an ambient set $\mathscr{B}$ is denoted by $\mathscr{A}^{\complement} := \mathscr{B} \setminus \mathscr{A}$. For example, the complement of $\{1, 2\}$ with respect to $\{1, \ldots, p\}$, where $p \in \{3, 4, \ldots\}$, is $\{3, \ldots, p\}$. Typically, it is clear what the ambient set is, so that there is no further mention of it.

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

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

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

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

$$C = \begin{pmatrix} C_{\mathscr{B}\mathscr{A}} & C_{\mathscr{B}\mathscr{A}^\complement} \\ C_{\mathscr{B}^\complement\mathscr{A}} & C_{\mathscr{B}^\complement\mathscr{A}^\complement} \end{pmatrix}.$$

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

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

Miscellaneous The expression $\boldsymbol{z} \sim \mathscr{N}_p[\boldsymbol{\mu}, \Sigma]$, $\boldsymbol{\mu} \in \mathbb{R}^p$, $\Sigma \in \mathbb{R}^{p\times p}$, states that $\boldsymbol{x}$ 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 $z \sim \mathscr{N}[\mu, \sigma^2]$, where $\mu \in \mathbb{R}$ is the mean and $\sigma^2 \in (0, \infty)$ the variance.

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

Contents