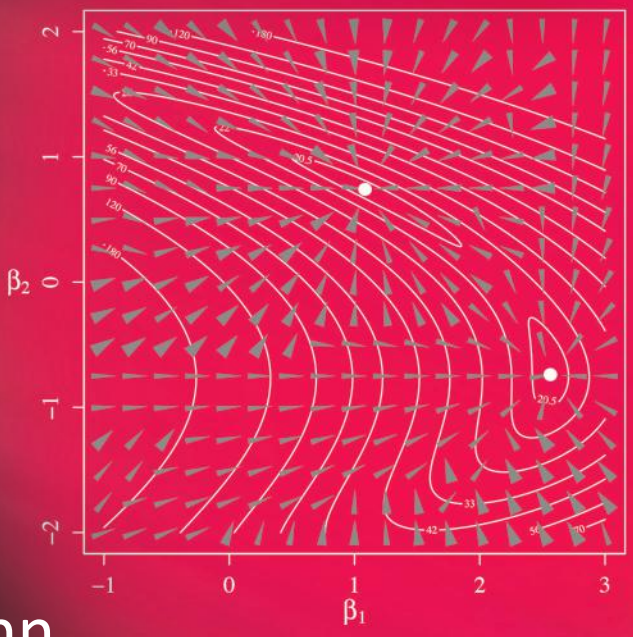
Computer Science and Data Analysis Series

Foundations of Statistical Algorithms

With References to R Packages



Claus Weihs
Olaf Mersmann
Uwe Ligges

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Foundations of Statistical Algorithms

With References to R Packages

Chapman & Hall/CRC

Computer Science and Data Analysis Series

The interface between the computer and statistical sciences is increasing, as each discipline seeks to harness the power and resources of the other. This series aims to foster the integration between the computer sciences and statistical, numerical, and probabilistic methods by publishing a broad range of reference works, textbooks, and handbooks.

Published Titles

Semisupervised Learning for Computational Linguistics
Steven Abney

Design and Modeling for Computer Experiments
Kai-Tai Fang, Runze Li, and Agus Sudjianto

Microarray Image Analysis: An Algorithmic Approach
Karl Fraser, Zidong Wang, and Xiaohui Liu

R Programming for Bioinformatics
Robert Gentleman

Exploratory Multivariate Analysis byExample Using R
François Husson, Sébastien Lê, andJérôme Pagès

Bayesian Artificial Intelligence, Second Edition
Kevin B. Korb and Ann E. Nicholson

Computational Statistics Handbook with MATLAB®, Second Edition
Wendy L. Martinez and Angel R. Martinez

Published Titles cont.

Exploratory Data Analysis with MATLAB®, Second Edition
Wendy L. Martinez, Angel R. Martinez, and Jeffrey L. Solka

Clustering for Data Mining: A Data Recovery Approach, Second Edition
Boris Mirkin

Introduction to Machine Learning and Bioinformatics
Sushmita Mitra, Sujay Datta, Theodore Perkins, and George Michailidis

Introduction to Data Technologies
Paul Murrell

R Graphics
Paul Murrell

Correspondence Analysis and Data Coding with Java and R
Fionn Murtagh

Pattern Recognition Algorithms for Data Mining
Sankar K. Pal and Pabitra Mitra

Statistical Computing with R
Maria L. Rizzo

Statistical Learning and Data Science
Mireille Gettler Summa, Léon Bottou, Bernard Goldfarb, Fionn Murtagh, Catherine Pardoux, and Myriam Touati

Foundations of Statistical Algorithms: With References to R Packages
Claus Weihs, Olaf Mersmann, and Uwe Ligges

Computer Science and Data Analysis Series

Foundations of Statistical Algorithms

With References to R Packages

Claus Weihs
Olaf Mersmann
Uwe Ligges

TU Dortmund University
Germany

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20131104

International Standard Book Number-13: 978-1-4398-7887-3 (eBook - PDF)

To
Heidrun and Max,
Sabine,
and Sandra

Contents

Preface

This book is largely based on a yearly lecture, "Computer Supported Statistics" (Computergestützte Statistik), for statistics students regularly held and improved by the authors since winter 1999/2000 at the TU University Dortmund (Germany). The exercises are based on those prepared for this lecture and other related lectures. For the book, the material of this lecture was thoroughly revised, extended, and modernized. This is particularly true for Chapters 4 and 8.

This book is not "yet another treatise on computational statistics". In fact, there is, as of this writing, no other book on the market that has a similar emphasis, for at least three reasons.

1. All the textbooks on computational statistics we know of present concise introductions to a multitude of state-of-the-art statistical algorithms without covering the historical aspect of their development, which we think is instructive in understanding the evolution of ever more powerful statistical algorithms. Many of the older algorithms are still building blocks or inspiration for current techniques. It is therefore instructive to cover these as well and present the material from a historical perspective before explaining the current best-of-breed algorithms, which naturally makes up the main body of the book.
2. With the chosen chapter titles, we try to emphasize certain recurring themes in all statistical algorithms: Computation, assessment and verification, iteration, deduction of theoretical properties, randomization, repetition and parallelization and scalability. Students should not only understand current algorithms after reading this book, but also gain a deeper understanding of how algorithms are constructed, how to evaluate new algorithms, which recurring principles are used to tackle some of the tough problems statistical programmers face, and how to take an idea for a new method and turn it into something practically useful.
3. The book contains two chapters on topics neglected in other books. One chapter is dedicated to systematic verification, a topic that is not covered in any other statistical computing book we know of. Instead of focusing on

contrived test examples, we show how to derive general classes of worst case inputs and why it is important to systematically test an algorithm over a large number of different inputs. And another chapter covers the upcoming challenge of scaling many of the established techniques to very large data sets and how the availability of many CPU cores will change the way we think about statistical computing.

To summarize, this book is based on a new and refreshingly different approach to presenting the foundations of statistical algorithms. Therefore, this book provides a great resource for both students and lecturers teaching a course in computational statistics.

Acknowledgments

We thank Daniel Horn, Sarah Schnackenberg, and Sebastian Szugat for their tireless critical proof reading, Pascal Kerschke for investigating historical literature, John Kimmel for his powerful realization of the review process, the unknown reviewers and the copy-editor for their valuable comments on draft copies of the manuscript, and last but not least Marcus Fontaine for LaTeX support in preparing the final manuscript.

The Authors

Prof. Dr. Claus Weihs studied mathematics in Bonn (Germany). After his studies, he developed a large software system at the economics department in Bonn (Germany). He received his PhD (Dr. rer. nat.) in numerical mathematics from the University of Trier (Germany) in 1986. He then practiced statistics and mathematics for 9 years as a consultant for Ciba-Geigy (Basel, Switzerland). He took on his current position as the chair of Computational Statistics in Dortmund (Germany) in 1995. So far, among more than 150 publications, Prof. Weihs has published two monographs and edited four other books. Moreover, he has served as president of the German Classification Society (GfKl) from 2004 to 2013, and he is one of the editors of the journal *Advances in Data Analysis and Classification (ADAC)*.

Olaf Mersmann has studied physics and statistics at the TU Ilmenau, University of Marburg, and the TU Dortmund University. He finished his MA in data analysis in 2011 at the department of statistics of the TU Dortmund University. Starting with his BA, Mr. Mersmann has been researching new and innovative ways to objectively test benchmark computer algorithms. He has contributed eight packages to CRAN, the R software package repository, and worked on several more.

Dr. Uwe Ligges is junior-professor for Data Analysis and Statistical Algorithms at the department of statistics, TU Dortmund University. He is author of the (German) textbook *Programmieren mit R (Programming in R)* (Springer Verlag, Heidelberg), which was first published in 2004 and is currently available in its third edition. A Japanese translation of this book was published in 2006. Uwe Ligges is also known as a member of the R Core Team and the CRAN maintainer for Windows binaries of contributed packages. Additionally, he acts as one of the editors for the *Journal of Statistical Software* as well as a column editor for *The R Journal*.

List of Algorithms