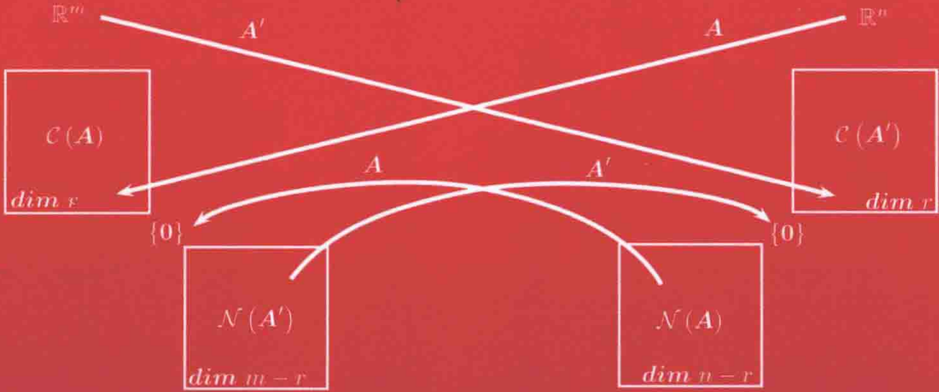
Texts in Statistical Science

Linear Algebra and Matrix Analysis for Statistics



Sudipto Banerjee

Anindya Roy

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Texts in Statistical Science

Linear Algebra and Matrix Analysis for Statistics

Sudipto Banerjee
Professor of Biostatistics
School of Public Health
University of Minnesota, U.S.A.

Anindya Roy
Professor of Statistics
Department of Mathematics and Statistics
University of Maryland, Baltimore County, U.S.A.

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

Printed on acid-free paper
Version Date: 20140407

International Standard Book Number-13: 978-1-4200-9538-8 (Hardback)

Library of Congress Cataloging-in-Publication Data

Banerjee, Sudipto, author.
Linear algebra and matrix analysis for statistics / Sudipto Banerjee, professor of biostatistics, School of Public Health, University of Minnesota, U.S.A., Anindya Roy, professor of statistics, Department of Mathematics and Statistics, University of Maryland, Baltimore County, U.S.A.
pages cm. -- (Chapman & Hall/CRC texts in statistical science)
Summary: "Linear algebra and the study of matrix algorithms have become fundamental to the development of statistical models. Using a vector-space approach, this book provides an understanding of the major concepts that underlie linear algebra and matrix analysis. Each chapter introduces a key topic, such as infinite-dimensional spaces, and provides illustrative examples. The authors examine recent developments in diverse fields such as spatial statistics, machine learning, data mining, and social network analysis. Complete in its coverage and accessible to students without prior knowledge of linear algebra, the text also includes results that are useful for traditional statistical applications."-- Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-4200-9538-8 (hardback)
1. Algebras, Linear. 2. Matrices. 3. Mathematical statistics. I. Roy, Anindya, 1970- author. II. Title.

QA184.2.B36 2014
512'.5--dc23 2014004884

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Linear Algebra and Matrix Analysis for Statistics

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Analysis of Failure and Survival Data
P. J. Smith

The Analysis of Time Series: An Introduction, Sixth Edition
C. Chatfield

Applied Bayesian Forecasting and Time Series Analysis
A. Pole, M. West, and J. Harrison

Applied Categorical and Count Data Analysis
W. Tang, H. He, and X.M. Tu

Applied Nonparametric Statistical Methods, Fourth Edition
P. Sprent and N.C. Smeeton

Applied Statistics: Handbook of GENSTAT Analyses
E.J. Snell and H. Simpson

Applied Statistics: Principles and Examples
D.R. Cox and E.J. Snell

Applied Stochastic Modelling, Second Edition
B.J.T. Morgan

Bayesian Data Analysis, Third Edition
A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin

Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians
R. Christensen, W. Johnson, A. Branscum, and T.E. Hanson

Bayesian Methods for Data Analysis, Third Edition
B.P. Carlin and T.A. Louis

Beyond ANOVA: Basics of Applied Statistics
R.G. Miller, Jr.

The BUGS Book: A Practical Introduction to Bayesian Analysis
D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter

A Course in Categorical Data Analysis
T. Leonard

A Course in Large Sample Theory
T.S. Ferguson

Data Driven Statistical Methods
P. Sprent

Decision Analysis: A Bayesian Approach
J.Q. Smith

Design and Analysis of Experiments with SAS
J. Lawson

Elementary Applications of Probability Theory, Second Edition
H.C. Tuckwell

Elements of Simulation
B.J.T. Morgan

Epidemiology: Study Design and Data Analysis, Third Edition
M. Woodward

Essential Statistics, Fourth Edition
D.A.G. Rees

Exercises and Solutions in Statistical Theory
L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Exercises and Solutions in Biostatistical Theory
L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models
J.J. Faraway

A First Course in Linear Model Theory
N. Ravishanker and D.K. Dey

Generalized Additive Models: An Introduction with R
S. Wood

Generalized Linear Mixed Models: Modern Concepts, Methods and Applications
W. W. Stroup

Graphics for Statistics and Data Analysis with R
K.J. Keen

Interpreting Data: A First Course in Statistics
A.J.B. Anderson

Introduction to General and Generalized Linear Models
H. Madsen and P. Thyregod

An Introduction to Generalized Linear Models, Third Edition
A.J. Dobson and A.G. Barnett

Introduction to Multivariate Analysis
C. Chatfield and A.J. Collins

Introduction to Optimization Methods and Their Applications in Statistics
B.S. Everitt

Introduction to Probability with R
K. Baclawski

Introduction to Randomized Controlled Clinical Trials, Second Edition
J.N.S. Matthews

Introduction to Statistical Inference and Its Applications with R
M.W. Trosset

Introduction to Statistical Limit Theory
A.M. Polansky

Introduction to Statistical Methods for Clinical Trials
T.D. Cook and D.L. DeMets

Introduction to Statistical Process Control
P. Qiu

Introduction to the Theory of Statistical Inference
H. Liero and S. Zwanzig

Large Sample Methods in Statistics
P.K. Sen and J. da Motta Singer

Linear Algebra and Matrix Analysis for Statistics
S. Banerjee and A. Roy

Logistic Regression Models
J.M. Hilbe

Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition
D. Gamerman and H.F. Lopes

Mathematical Statistics
K. Knight

Modeling and Analysis of Stochastic Systems, Second Edition
V.G. Kulkarni

Modelling Binary Data, Second Edition
D. Collett

Modelling Survival Data in Medical Research, Second Edition
D. Collett

Multivariate Analysis of Variance and Repeated Measures: A Practical Approach for Behavioural Scientists
D.J. Hand and C.C. Taylor

Multivariate Statistics: A Practical Approach
B. Flury and H. Riedwyl

Multivariate Survival Analysis and Competing Risks
M. Crowder

Nonparametric Methods in Statistics with SAS Applications
O. Korosteleva

Pólya Urn Models
H. Mahmoud

Practical Data Analysis for Designed Experiments
B.S. Yandell

Practical Longitudinal Data Analysis
D.J. Hand and M. Crowder

Practical Multivariate Analysis, Fifth Edition
A. Afifi, S. May, and V.A. Clark

Practical Statistics for Medical Research
D.G. Altman

A Primer on Linear Models
J.F. Monahan

Principles of Uncertainty
J.B. Kadane

Probability: Methods and Measurement
A. O'Hagan

Problem Solving: A Statistician's Guide, Second Edition
C. Chatfield

Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition
B.F.J. Manly

Readings in Decision Analysis
S. French

Sampling Methodologies with Applications
P.S.R.S. Rao

Stationary Stochastic Processes: Theory and Applications
G. Lindgren

Statistical Analysis of Reliability Data
M.J. Crowder, A.C. Kimber, T.J. Sweeting, and R.L. Smith

Statistical Methods for Spatial Data Analysis
O. Schabenberger and C.A. Gotway

Statistical Methods for SPC and TQM
D. Bissell

Statistical Methods in Agriculture and Experimental Biology, Second Edition
R. Mead, R.N. Curnow, and A.M. Hasted

Statistical Process Control: Theory and Practice, Third Edition
G.B. Wetherill and D.W. Brown

Statistical Theory: A Concise Introduction
F. Abramovich and Y. Ritov

Statistical Theory, Fourth Edition
B.W. Lindgren

Statistics for Accountants
S. Letchford

Statistics for Epidemiology
N.P. Jewell

Statistics for Technology: A Course in Applied Statistics, Third Edition
C. Chatfield

Statistics in Engineering: A Practical Approach
A.V. Metcalfe

Statistics in Research and Development, Second Edition
R. Caulcutt

Stochastic Processes: An Introduction, Second Edition
P.W. Jones and P. Smith

Survival Analysis Using S: Analysis of Time-to-Event Data
M. Tableman and J.S. Kim

The Theory of Linear Models
B. Jørgensen

Time Series Analysis
H. Madsen

Time Series: Modeling, Computation, and Inference
R. Prado and M. West

Understanding Advanced Statistical Methods
P.H. Westfall and K.S.S. Henning

To my parents, Shyamali and Sunit, my wife, Sharbani, and my son, Shubho.

—Sudipto Banerjee

To my wife, Ishani.

—Anindya Roy

Preface

Linear algebra constitutes one of the core mathematical components in any modern curriculum involving statistics. Usually students studying statistics are expected to have seen at least one semester of linear algebra (or applied linear algebra) at the undergraduate level. In particular, students pursuing graduate studies in statistics or biostatistics are expected to have a sound conceptual grasp of vector spaces and subspaces associated with matrices, orthogonality, projections, quadratic forms and so on.

As the relevance and attraction of statistics as a discipline for graduate studies continues to increase for students with more diverse academic preparations, the need to accommodate their mathematical needs also keeps growing. In particular, many students find their undergraduate preparation in linear algebra rather different from what is required in graduate school. There are several excellent texts on the subject that provide as comprehensive a coverage of the subject as possible at the undergraduate level. However, some of these texts cater to a broader audience (e.g., scientists and engineers) and several formal concepts that are important in theoretical statistics are not emphasized.

There are several excellent texts on linear algebra. For example, there are classics by Halmos (1974), Hoffman and Kunze (1984) and Axler (1997) that make heavy use of vector spaces and linear transformations to provide a coordinate-free approach. A remarkable feature of the latter is that it develops the subject without using determinants at all. Then, there are the books by Strang (2005, 2009) and Meyer (2001) that make heavy use of echelon forms and canonical forms to reveal the properties of subspaces associated with a matrix. This approach is tangible, but may not turn out to be the most convenient to derive and prove results often encountered in statistical modeling. Among texts geared toward statistics, Rao and Bhimsankaram (2000), Searle (1982) and Graybill (2001) have stood the test of time. The book by Harville (1997) stands out in its breadth of coverage and is already considered a modern classic. Several other excellent texts exist for statisticians including Healy (2000), Abadir and Magnus (2005), Schott (2005) and Gentle (2010). The concise text by Bapat (2012) is a delightful blend of linear algebra and statistical linear models.

While the above texts offer excellent coverage, some expect substantial mathematical maturity from the reader. Our attempt here has been to offer a more gradual exposition to linear algebra without really dumbing down the subject. The book tries to be as self-contained as possible and does not assume any prior knowledge of linear algebra. However, those who have seen some elementary linear algebra will be able to move more quickly through the early chapters. We have attempted to present both the vector space approach as well as the canonical forms in matrix theory.

Although we adopt the vector space approach for much of the later development, the book does not begin with vector spaces. Instead, it addresses the rudimentary mechanics of linear systems using Gaussian elimination and the resultant decompositions (Chapters 1–3). Chapter 4 introduces Euclidean vector spaces using less abstract concepts and makes connections to systems of linear equations wherever possible. Chapter 5 is on the rank of a matrix. Why devote an entire chapter to rank? We believe that the concept of rank is *that* important for a thorough understanding. In several cases we show how the same result may be derived using multiple techniques, which, we hope, will offer insight into the subject and ensure a better conceptual grasp of the material. Chapter 6 introduces complementary subspaces and oblique projectors. Chapter 7 introduces orthogonality and orthogonal projections and leads us to the Fundamental Theorem of Linear Algebra, which connects the four fundamental subspaces associated with a matrix. Chapter 8 builds upon the previous chapter and focuses on orthogonal projectors, which is fundamental to linear statistical models, and also introduces several computational techniques for orthogonal reduction. Chapter 9 revisits linear equations from a more mature perspective and shows how the theoretical concepts developed thus far can be handy in analyzing solutions for linear systems. The reader, at this point, will have realized that there is much more to linear equations than Gaussian elimination and echelon forms. Chapter 10 discusses determinants. Unlike some classical texts, we introduce determinants a bit late in the game and present it as a useful tool for characterizing and obtaining certain useful results. Chapter 11 introduces eigenvalues and eigenvectors and is the first time complex numbers make an appearance. Results on general real matrices are followed by those of real symmetric matrices. The popular algorithms for eigenvalues and eigenvectors are outlined both for symmetric and unsymmetric matrices. Chapter 12 derives the Singular value decomposition and the Jordan Canonical Form and presents an accessible proof of the latter. Chapter 13 is devoted to quadratic forms, another topic of fundamental importance to statistical theory and methods. Chapter 14 presents Kronecker and Hadamard products and other related materials that have become conspicuous in multivariate statistics and econometrics. Chapters 15 and 16 provide a taste of some more advanced topics but, hopefully, in a more accessible manner than more advanced texts. The former presents some aspects of linear iterative systems and convergence of matrices, while the latter introduces more general vector spaces, linear transformations and Hilbert spaces.

We remark that this is not a book on matrix computations, although we describe several numerical procedures in some detail. We have refrained from undertaking a thorough exploration of the most numerically stable algorithms as they would require a lot more theory and be too much of a digression. However, readers who grasp the material provided here should find it easier to study more specialized texts on matrix computations (e.g., Golub and Van Loan, 2013; Trefthen and Bau III, 1997). Also, while we have included many exercises that can be solved using languages such as

MATLAB and R, we decided not to marry the text to a specific language or platform. We have also not included statistical theory and applications here. This decision was taken neither in haste nor without deliberation. There are plenty of excellent texts on the theory of linear models, regression and modeling that make abundant use of linear algebra. Our hope is that readers of this text will find it easier to grasp the material in such texts. In fact, we believe that this book can be used as a companion text in the more theoretical courses on linear regression or, perhaps, stand alone as a one-semester course devoted to linear algebra for statistics and econometrics.

Finally, we have plenty of people to thank for this book. We have been greatly influenced by our teachers at the Indian Statistical Institute, Kolkata. The book by Rao and Bhimsankaram (2000), written by two of our former teachers whose lectures and notes are still vivid in our minds, certainly shaped our preparation. Sudipto Banerjee would also like to acknowledge Professor Alan Gelfand of Duke University with whom he has had several discussions regarding the role of linear algebra in Bayesian hierarchical models and spatial statistics. The first author also thanks Dr. Govindan Rangarajan of the Indian Institute of Science, Bangalore, India, Dr. Anjana Narayan of California Polytechnic State University, Pomona, and Dr. Mohan Delampady of Indian Statistical Institute, Bangalore, for allowing the author to work on this manuscript as a visitor in their respective institutes. We thank the Division of Biostatistics at the University of Minnesota, Twin Cities, and the Department of Statistics at the University of Maryland, Baltimore, for providing us with an ambience most conducive to this project. Special mention must be made of Dr. Rajarshi Guhaniyogi, Dr. Joao Monteiro and Dr. Qian Ren, former graduate students at the University of Minnesota, who have painstakingly helped with proof-reading and typesetting parts of the text. This book would also not have happened without the incredible patience and cooperation of Rob Calver, Rachel Holt, Sarah Gelson, Kate Gallo, Charlotte Byrnes and Shashi Kumar at CRC Press/Chapman and Hall. Finally, we thank our families, whose ongoing love and support made all of this possible.

SUDIPTO BANERJEE — Minneapolis, Minnesota
ANINDYA ROY — Baltimore, Maryland
August, 2013.

Contents