Springer Series in Statistics

Advisors:
P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin, S. Zeger

Springer Series in Statistics

Athreya/Lahiri: Measure Theory and Probability Theory
Bilodeau/Brenner: Theory of Multivariate Statistics
Brockwell/Davis: An Introduction to Time Series and Forecasting
Carmona: Statistical Analysis of Financial Data in S-PLUS
Chow/Teicher: Probability Theory: Independence, Interchangeability,Martingales, 3rd ed.
Christensen: Advanced Linear Modeling: Multivariate, Time Series, and Spatial Data; Nonparametric Regression and Response Surface Maximization, 2nd ed.
Christensen: Log-Linear Models and Logistic Regression, 2nd ed.
Christensen: Plane Answers to Complex Questions: The Theory of LinearModels, 2nded.
Cryer/Chan: Time Series Analysis, Second Edition
DasGupta: Asymptotic Theory of Statistics and Probability
Davis: Statistical Methods for the Analysis of Repeated Measurements
Dean/Voss: Design and Analysis of Experiments
Dekking/Kraaikamp/Lopuhaä/Meester: A Modern Introduction to Probability and Statistics
Durrett: Essential of Stochastic Processes
Edwards: Introduction to Graphical Modeling, 2nd ed.
Everitt: An R and S-PLUS Companion to Multivariate Analysis
Gentle: Matrix Algebra: Theory, Computations, and Applications in Statistics
Ghosh/Delampady/Samanta: An Introduction to Bayesian Analysis
Gut: Probability: A Graduate Course
Heiberger/Holland: Statistical Analysis and Data Display; An Intermediate Course with Examples in S-PLUS, R, and SAS
Jobson: Applied Multivariate Data Analysis, Volume I: Regression andExperimental Design
Jobson: Applied Multivariate Data Analysis, Volume II Categorical and Multivariate Methods
Karr: Probability
Kulkarni: Modeling, Analysis, Design, and Control of Stochastic Systems
Kolaczyk: Statistical Analysis of Network Data
Lange: Applied Probability
Lange: Optimization
Lehmann: Elements of Large Sample Theory
Lehmann/Romano: Testing Statistical Hypotheses, 3rd ed.
Lehmann/Casella: Theory of Point Estimation, 2nd ed.
Longford: Studying Human Populations: An Advanced Course in Statistics
Marin/Robert: Bayesian Core: A Practical Approach to Computational Bayesian Statistics
Nolan/Speed: Stat Labs: Mathematical Statistics Through Applications
Pitman: Probability
Rawlings/Pantula/Dickey: Applied Regression Analysis
Robert: The Bayesian Choice: From Decision-Theoretic Foundations toComputational Implementation, 2nd ed.
Robert/Casella: Monte Carlo Statistical Methods, 2nd ed.
Rose/Smith: Mathematical Statistics with *Mathematica*
Ruppert: Statistics and Finance: An Introduction

(continued after index)

Eric D. Kolaczyk

Statistical Analysis of Network Data

Methods and Models

Springer

Eric D. Kolaczyk
Department of Mathematics & Statistics
Boston University
111 Cummington St.
Boston MA 02215
USA

ISBN 978-0-387-88145-4 e-ISBN 978-0-387-88146-1
DOI 10.1007/978-0-387-88146-1

Library of Congress Control Number: 2009921812

Printed on acid-free paper

springer.com

To the ‘network’ that is my family . . .
. . . those both near and far.

In Memorium

Br. Robert C. Ruhl, CSV
March 18, 1947 – Feb 2, 2008

“A teacher affects eternity; he can never tell where his influence stops.”
(Henry Brooks Adams)

Preface

In recent years there has been an explosion of *network data* – that is, measurements that are either of or from a system conceptualized as a network – from seemingly all corners of science. The combination of an increasingly pervasive interest in scientific analysis at a systems level and the ever-growing capabilities for high-throughput data collection in various fields has fueled this trend. Researchers from biology and bioinformatics to physics, from computer science to the information sciences, and from economics to sociology are more and more engaged in the collection and statistical analysis of data from a network-centric perspective.

Accordingly, the contributions to statistical methods and modeling in this area have come from a similarly broad spectrum of areas, often independently of each other. Many books already have been written addressing network data and network problems in specific individual disciplines. However, there is at present no single book that provides a modern treatment of a core body of knowledge for statistical analysis of network data that cuts across the various disciplines and is organized rather according to a statistical taxonomy of tasks and techniques. This book seeks to fill that gap and, as such, it aims to contribute to a growing trend in recent years to facilitate the exchange of knowledge across the pre-existing boundaries between those disciplines that play a role in what is coming to be called 'network science.'

The book is written for students and researchers with a 'mature' knowledge of statistics and hence is intended not only for statisticians but also for people involved with network data in various other areas, like those mentioned above. Background in calculus and linear algebra and some reasonable foundation in statistics and probability are expected. Beyond that, I have attempted to build all necessary material as needed.

In an effort to reach this admittedly diverse audience successfully, I have aimed in each chapter to communicate the material in a manner that strikes an appropriate balance between concepts, on the one hand, and technical depth and rigor, on the other. It is expected that the interested reader will want – and, indeed, is encouraged – to pursue the relevant primary sources for details I may have chosen to omit. The book is in this sense intended to serve as an *entrée* to the larger literature. Copious use of references has been made throughout the book for this very purpose. In addition, the exercises at the end of each chapter provide further opportunities to explore some of the topics in greater depth. There are both analytical and computational exercises to be found, with the latter frequently designed to be fairly open-ended in nature, so as to encourage exploration. Finally, the methods and models presented herein are illustrated throughout the book with examples from a wide range of disciplines. I have found this overall approach to the pedagogy of the material to work well when I taught classes of precisely the diversity that I envision for the readership of this book.

The book itself would not have been possible without the help, feedback, and support of many. I thank my editor, John Kimmel, and the other folks at Springer, for their help and guidance throughout the publication process. The idea for the writing of this book arose while I was on sabbatical during the 2004-2005 academic year, in the Laboratoire d'Informatique Algorithmique, Fondements, et Applications (LIAFA), at l'Université Paris 7, courtesy of CNRS, and in the Department of Statistics, at Harvard University. My own work in this area has been generously supported by grants from the United States National Institutes of Health (NIH), National Science Foundation (NSF), and Office of Naval Research (ONR). Students and colleagues attending the courses I taught on this material, at Harvard University in Spring 2005 and at Boston University in Fall 2005, are gratefully acknowledged for their interest, questions, and hard work. Many of the datasets used in this book were generously shared by colleagues from around the world, including Ignacio Alvarez-Hamelin, Mark Coates, Mark Crovella, Tim Gardner, Sucharita Gopal, Boris Hayete, Mark Kramer, Emmanuel Lazergas, Naoki Nariai, Robert Nowak, Xiaoyu Jiang, and Fabien Viger. In addition, Ignacio Alvarez-Hamelin, Kevin Boyack, Ulrik Brandes, Rui Castro, Sucharita Gopal, and Mark Kramer were kind enough to produce and share some of the more stunning figures found herein. For graciously responding to my unabashed solicitations for feedback on material in various chapters, I am indebted to Ignacio Alvarez-Hamelin, Alain Barrat, Ulrik Brandes, Tom Britton, Hugh Chipman, Mark Crovella, Tim Gardner, Boris Hayete, Peter Hoff, David Hunter, Xiaoyu Jiang, Simon Kasif, Naoki Nariai, Robert Nowak, Pip Pattison, Mike Rabbat, Garry Robbins, Martin Steffen, Shu Yang, and Ji Zhu. Special thanks in this regard are due to Joe Whittaker, whose comments were extensive and invariably helpful. Notwithstanding all of this feedback, however, any typos, mistakes, and other similar errors are of course my own. Andrej Mrvar, of the `Pajek` software team, is also to be thanked for his help in responding to my numerous questions. Finally, although surely inadequate in measure, I offer my heartfelt thanks to my wife, Josée Dupuis, not only for her love, patience, and support over the many months that I was submersed in the writing of this book, and for her reading of the final manuscript, but also for serving as my **R** guru, fielding even my many late-night queries when she would surely have rather preferred to continue uninterrupted with her own work!

Eric D. Kolaczyk
Boston, Massachusetts
March, 2009

Contents