Cambridge University Press
978-0-521-89619-1 - Algebraic and Geometric Methods in Statistics
Edited by Paolo Gibilisco, Eva Riccomagno, Maria Piera Rogantin and Henry P. Wynn
Frontmatter
More information

ALGEBRAIC AND GEOMETRIC METHODS IN STATISTICS

This up-to-date account of algebraic statistics and information geometry explores the emerging connections between the two disciplines, demonstrating how they can be used in design of experiments and how they benefit our understanding of statistical models and in particular, exponential models. This book presents a new way of approaching classical statistical problems and raises scientific questions that would never have been considered without the interaction of these two disciplines.

Beginning with a brief introduction to each area, using simple illustrative examples, the book then proceeds with a collection of reviews and some new results by leading researchers in their respective fields. Parts I and II are mainly on contingency table analysis and design of experiments. Part III dwells on both classical and quantum information geometry. Finally, Part IV provides examples of the interplay between algebraic statistics and information geometry. Computer code and some proofs are also available on-line, where key examples are also developed in further detail.

ALGEBRAIC AND GEOMETRIC METHODS IN STATISTICS

Edited by

PAOLO GIBILISCO
EVA RICCOMAGNO
MARIA PIERA ROGANTIN
HENRY P. WYNN

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521896191

First published 2010

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-89619-1 Hardback

This volume is dedicated to
Professor Giovanni Pistone
on the occasion of
his sixty-fifth birthday

Contents

Contributors

Satoshi Aoki
Department of Mathematics and Computer Science, Kagoshima University, 1-21-35, Korimoto, Kagoshima 890-0065, Japan

Enrico Carlini
Department of Mathematics, Polytechnic of Turin, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

Yuguo Chen
Department of Statistics, University of Illinois at Urbana-Champaign, 725, S. Wright Street Champaign, IL 61820 USA

Ian H. Dinwoodie
214 Old Chemistry Building, Box 90251, Duke University, Durham, NC 27708-0251 USA

Adrian Dobra
Department of Statistics, University of Washington, Seattle WA 98195-4322 USA

Stephen E. Fienberg
Department of Statistics, Machine Learning Dept. and Cylab, Carnegie Mellon University, Pittsburgh, PA 15213-3890 USA

Roberto Fontana
Department of Mathematics, Polytechnic of Turin, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

Kenji Fukumizu
Institute of Statistical Mathematics, 4-6-7 Minamiazabu, Minatoku, Tokyo 106-8569, Japan

Paolo Gibilisco
Department S.E.F.E.M.E.Q., University of Roma Tor Vergata, Via Columbia 2, 00133 Rome, Italy

Frank Hansen
Department of Economics, University of Copenhagen, Studiestrde 6, 1455 Kbenhavn K, Denmark

Patricia Hersh
Department of Mathematics, Indiana University, Bloomington, IN 47405-7000 USA

Serkan Hoşten
Department of Mathematics, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA, 94132 USA

Daniele Imparato
Department of Mathematics, Polytechnic of Turin, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

Anne Krampe
Fakultät Statistik, Technische Universität Dortmund, 44221 Dortmund, Germany

Sonja Kuhnt
Technische Universiteit Eindhoven, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Reinhard Laubenbacher
Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Washington Street, MC 0477, USA

Guy Lebanon
Colleges of Science and Engineering, Purdue University, 250 N. University Street, West Lafayette, IN, 47907-2066, USA

Hugo Maruri-Aguilar
Department of Statistics, London School of Economics, London WC2A 2AE, United Kingdom

Roberto Notari
Department of Mathematics, Polytechnic of Milan, Via Bonardi 9, 20133 Milano, Italia

Giovanni Pistone
Department of Mathematics, Polytechnic of Turin, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

Fabio Rapallo
Department DISTA, University of Eastern Piedmont, Via Bellini, 25/G, 15100 Alessandria, Italy

Eva Riccomagno
Department of Mathematics, Genoa University, Via Dodecaneso, 35, 16146 Genova, Italia

Alessandro Rinaldo
Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890 USA

Maria Piera Rogantin
Department of Mathematics, Genoa University, Via Dodecaneso, 35, 16146 Genova, Italia

Aleksandra B. Slavković
Department of Statistics, Pennsylvania State University, State College, PA USA

Brandilyn Stigler
Mathematical Biosciences Institute, The Ohio State University, 231 West 18th Avenue, Columbus, OH 43210, USA

Raymond F. Streater
Department of Mathematics, Kings College London, The Strand, London WC2R 2LS, United Kingdom

Seth Sullivant
Department of Mathematics, Harvard University, One Oxford Street, Cambridge, MA 02138 USA

Akimichi Takemura
Department of Mathematical Informatics, University of Tokyo, Bunkyo, Tokyo 113-0033, Japan

Barbara Trivellato
Department of Mathematics, Polytechnic of Turin, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

Henry P. Wynn
Department of Statistics, London School of Economics, London WC2A 2AE, United Kingdom

Anna Jenčová
Mathematical Institute, Slovak Academy of Sciences, Stefanikova 49, SK-84173 Bratislava, Slovakia

Ruriko Yoshida
Department of Statistics, University of Kentucky, 805A Patterson Office Tower, Lexington, KY 40506-0027, USA

Yi Zhou
Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213-3890 USA

Preface

Information Geometry and Algebraic Statistics are brought together in this volume to suggest that the interaction between them is possible and auspicious.

To meet this aim, we couple expository material with more advanced research topics sometimes within the same chapter, cross-reference the various chapters, and include many examples both in the printed volume and in the on-line supplement, held at the Cambridge University Press web site at www.cambridge.org/9780521896191. The on-line part includes proofs that are instructive but long or repetitive, computer codes and detailed development of special cases.

Chapter 1 gives a brief introduction to both Algebraic Statistics and Information Geometry based on the simplest possible examples and on selected topics that, to the editors, seem most promising for the interlacing between them. Then, the volume splits naturally in two lines. Part I, on contingency tables, and Part II, on designed experiments, are authored by researchers active mainly within Algebraic Statistics, while Part III includes chapters on both classical and quantum Information Geometry. This material comes together in Part IV which consists of only one chapter by Giovanni Pistone, to whom the volume is dedicated, and provides examples of the interplay between Information Geometry and Algebraic Statistics.

The editors imagine various entry points into the volume according to the reader's own interests. These are indicated with squared boxes in Figure 0.1. Maximum likelihood estimation in models with hidden variables is revisited in an algebraic framework in Chapter 2 (S. E. Fienberg *et al.*) which is supported by a substantial on-line section, including Chapter 22 (Y. Zhou) where the role of secant varieties for graphical models is detailed. Chapter 3 (A. Slavkovich and S. E. Fienberg) gives old and new geometric characterizations of the joint distribution on $I \times J$ contingency tables and can be used to gain familiarity with algebraic geometric jargon and ideas common in Algebraic Statistics. The next two chapters present fast algorithms for the computation of Markov bases in model selection (Chapter 4 by A. Krampe and S. Kuhnt) and under strictly positive margins (Chapter 5 by Y. Chen *et al.*), while Chapter 6 (E. Carlini and F. Rapallo) defines a class of algebraic statistical models for category distinguishability in rater agreement problems. The algebraic notion of index of complexity of maximum likelihood equations is used in Chapter 7 (S. Hoşten and S. Sullivant) for bivariate data missing at random. This part of the volume ends with Chapter 8 by S. E. Fienberg and A. Dobra.

Part II considers the two technologies of Algebraic Statistics most employed in design and analysis of experiments. Chapter 12 (R. Fontana and M. P. Rogantin) uses the game of sudoku to review polynomial indicator functions and links to Part I via the notion of Markov bases. This link is developed for a special case in Chapter 13 (S. Aoki and A. Takemura). This chapter should appeal to a reader acquainted with the classical theory of experimental design. Chapters 9, 10 and 11 develop in different settings the ideas and techniques outlined in the first part of Chapter 1: Chapter 9 (H. Maruri-Aguilar and H. P. Wynn) argues that algebraic sets can be used as repositories of experimental designs; Chapter 10 (R. Laubenbacher and B. Stigler) presents an application to the identification of biochemical networks from experimental data; and Chapter 11 (E. Riccomagno and R. Notari) considers designs with replicated points.

The Information Geometry part of the volume starts with Chapter 14 (R. F. Streater) which provides a gentle and short, though comprehensive, introduction to Information Geometry and its link to the theory of estimation according to Fisher. It keeps as far as possible the analogy between the classical and the quantum case. It extends to the purely quantum case in Chapter 15 (R. F. Streater) which, together with Chapter 16 (A. Jenčová), provides an extension to the quantum case of the statistical manifolds modelled on an Orlicz space. Also, Chapter 20 (F. Hansen) deals with quantum Information Geometry. A construction of a statistical manifold modelled on a Reproducing Kernel Hilbert Space is presented in Chapter 18 (K. Fukumizu), where the application to the theory of estimation is based on a suitable class of likelihood functions defined point-wise. Chapter 19 (D. Imparato and B. Trivellato) extends the standard non-parametric exponential model by considering its limit, developing ideas in Chapter 21. An application of classical information geometry for text analysis is developed by G. Lebanon in Chapter 17.

Chapter 1 includes a glossary of terms from Algebraic Geometry that are recurrent in the volume.

The editors thank the authors for providing interesting papers, the many referees who helped with the peer-reviewing, our publisher CUP and the ever patient and capable Diana Gillooly. Some chapters in this volume were first presented to the conference ‘Mathematical explorations in contemporary statistics’ held in Sestri Levante on 19–20 May 2008. Some chapters were also presented at the opening workshop of the 2008–09 SAMSI Program on Algebraic Methods in Systems Biology and Statistics, 14–17 September 2008.

This volume is dedicated to Giovanni Pistone on the occasion of his sixty-fifth birthday. We are grateful for his discreet and constant support.

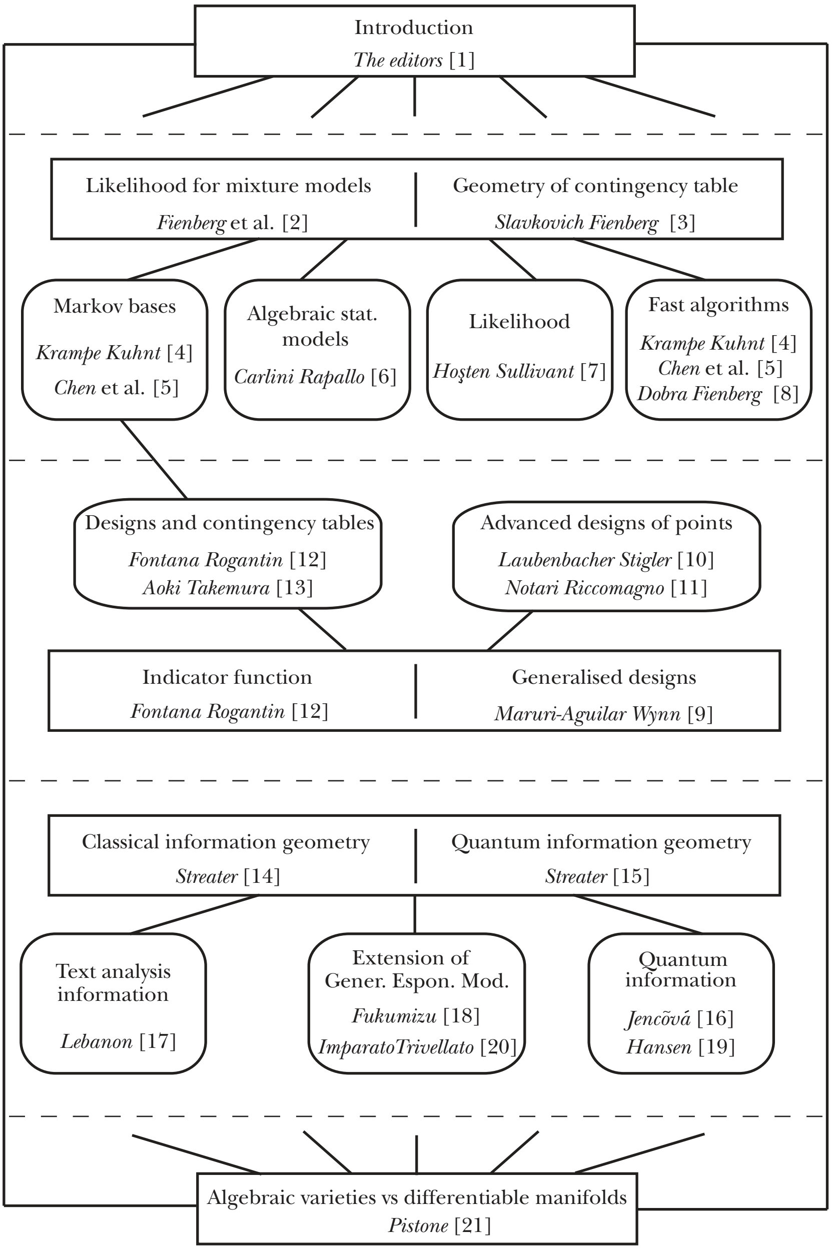

Fig. 1 Layout of the volume.

Frequently used notations and symbols

$\mathbb{N}$	natural numbers
$\mathbb{Z}$	integer numbers
$\mathbb{Q}$	rational numbers
$\mathbb{R}$	real numbers
$\mathbb{C}$	complex numbers
$\mathbb{R}_{>0}$	strictly positive real numbers
$\mathbb{R}_{\geq 0}$	non-negative real numbers

$\mathrm{E}_p[X]$	expectation of a random variable X w.r.t. the probability measure p
$\mathrm{Cov}_p(X, Y)$	covariance of X, Y w.r.t. p
$\mathrm{Var}_p(X)$	variance of X w.r.t. p
M_n	space of the $n \times n$ matrices with complex entries
$\mathrm{K}(p, q)$ or $\mathrm{KL}(q \Vert p)$	Kullback–Leibler relative entropy
I_X or $\mathcal{I}_X$ or G (resp. I_f or $\mathcal{I}_f$ or G)	Fisher information of X (resp. the density f)
$(\Omega, \mathcal{F}, \mu)$	measure space
$\mathcal{M}_>$, $\mathcal{M}_>(\mu)$ (resp. $\mathcal{M}_\geq$, $\mathcal{M}_\geq(\mu)$)	space of strictly positive (resp. non-negative) densities w.r.t the measure μ

$\vert . \vert$	cardinality of a set
k	number of factors
n	number of observations
$\mathcal{D}$	design
$\mathcal{K}[x_1, \ldots, x_k]$	set of polynomials in $x_1, \ldots, x_k$ with coefficients in $\mathcal{K}$
$\mathrm{I}(f_1, \ldots, f_l)$ or $\langle f_1, \ldots, f_l \rangle$	ideal generated by the polynomials $f_1, \ldots, f_l$
$\mathrm{I}(\mathcal{D})$	ideal of the points in the design
$\mathbb{R}[x_1, \ldots, x_k]/\mathrm{I}(f_1, \ldots, f_l)$	quotient space modulo $\mathrm{I}(f_1, \ldots, f_l)$
$\mathrm{NF}(f, I)$	normal form of f w.r.t. I
A or A_T	constraint matrix