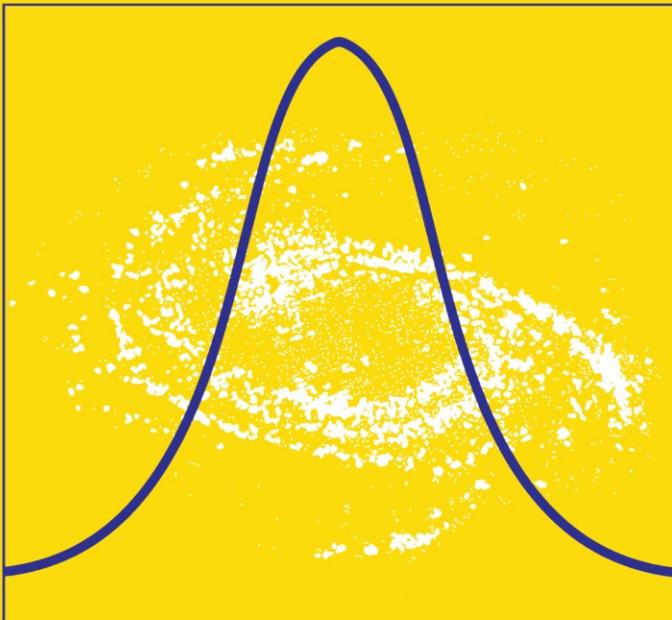


G. Jogesh Babu / Eric D. Feigelson
Editors

Statistical Challenges in Modern Astronomy II



Springer

Statistical Challenges in Modern Astronomy II

Springer-Science+Business Media, LLC

G. Jogesh Babu Eric D. Feigelson
Editors

Statistical Challenges in Modern Astronomy II

With 72 Illustrations



Springer

G. Jogesh Babu
Department of Statistics
Pennsylvania State University
University Park, PA 16802
USA

Eric D. Feigelson
Department of Astronomy
and Astrophysics
Pennsylvania State University
University Park, PA 16802
USA

Cover art: Conference logo of the cross-disciplinary conference, Statistical Challenges in Modern Astronomy," held on August 11-14, 1991, at the University Park campus of the Pennsylvania State University.

Library of Congress Cataloging-in-Publication Data
Statistical challenges in modern astronomy II / G. Jogesh Babu and
Eric D. Feigelson, editors
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4612-7360-8 ISBN 978-1-4612-1968-2 (eBook)
DOI 10.1007/978-1-4612-1968-2
1. Statistical astronomy—Congresses. I. Babu, Gutti Jogesh,
1949— . II. Feigelson, Eric D.
QB149.S74 1997
520'.72—dc21 97-5785

Printed on acid-free paper.

© 1997 Springer Science+Business Media New York
Originally published by Springer-Verlag New York, Inc. in 1997
Softcover reprint of the hardcover 1st edition 1997

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Science+Business Media, LLC), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

Production managed by Steven Pisano; manufacturing supervised by Jeffrey Taub.
Photocomposed pages prepared from the authors' LaTeX files.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4612-7360-8

Preface

Over the several centuries, astronomers provided basic concepts and laid foundations to the subsequent developments of mathematical statistics. But interaction between the two fields declined since the late 19th century as statisticians moved towards applications in biology, social sciences and industry, while astronomers developed the powerful links between physics and celestial phenomena. However, astronomy now faces a vast range of statistical issues: terabyte multiwavelength databases, nonlinear modeling using astrophysical theory, variable objects with unevenly spaced observations, complex clustering of galaxies in space and photons in images. Such problems have spawned a revival in methodological activity among astronomers during the past decade, evidenced by frequent papers and conferences on data analysis techniques. Over 200 articles on statistical techniques are published annually in the astronomical literature.

Although the needs for an active field of astrostatistics have grown, few statisticians work on astronomical problems. Some of the methodological problems arising in modern astronomical research have established solutions known among statisticians (though often not among astronomers), while others require development of new advanced statistical methods. The historical divergence of astronomy and statistics created a language and knowledge barrier between scholars in the two fields, and reestablishment of the longstanding relationship has not been rapid. Our recent monograph *Astrostatistics* (G. J. Babu and E. D. Feigelson, Chapman and Hall 1996) provides a basis for further cross-disciplinary interactions.

This situation led to the organization of the first cross-disciplinary meeting for astronomers and statisticians in the U.S. entitled *Statistical Challenges in Modern Astronomy* (SCMA). Held at Penn State in August 1991, the conference brought together researchers and students from observational astronomy and mathematical statistics to discuss methodological issues of common interest. Topics included galaxy clustering and large-scale structure of the universe, time series analysis of variable stars and galactic

nuclei, censoring and truncation from astronomical surveys with nondetections, image analysis, multivariate classification, and Bayesian approaches to many problems. Review talks (mainly by astronomers) were followed by commentary (mainly by statisticians) and discussions. Its proceedings were published in 1992 by Springer-Verlag with Feigelson and Babu as editors.

The present volume is the product of a second conference, SCMA II held in June 1996. About 125 researchers and students from 13 countries attended. The format was similar to SCMA I with an increased emphasis on talks by statisticians and introducing new topics. Some astronomical presentations focussed on statistical issues arising in large astronomical projects and space observatories. These include: Laser Interferometry Gravitational-Wave Observatory (LIGO), where very brief or faint signals must be acquired from long correlated time series; star/galaxy discrimination, a complex problem in multivariate classification arising in digitized optical sky surveys; Advanced X-ray Astrophysics Facility (AXAF), where complex spectral models must be fit to sparse spectroscopic data; Hipparcos, where accurate calibration of both positional and photometric data is required; X-ray Timing Explorer (XTE), where mixtures of aperiodic, quasi-periodic and periodic behaviors are studied in X-ray binary systems; gravitational lensing search for dark matter (MACHO project), where rare achromatic events must be extracted from a terabyte-sized database of noisy unevenly spaced time series.

It is difficult to summarize the statistical advice provided by the statisticians for such problems. The difficulties of treating flux limits and measurement errors were discussed. In some cases, standard methods such as ‘survival analysis’ for censored data and weighting for known measurement errors suffice. But statisticians noted that common methods can be incorrect under many circumstances. Both astronomers and statisticians expressed enthusiasm for techniques associated with the wavelet transform, and astronomers are particularly advanced in wavelet image analysis methods.

Considerable interest was expressed in the potential of Bayesian methods in addressing astronomical problems. Bayesian statistics are philosophically attractive to physical scientists (measured data are held fixed while hypotheses are considered uncertain), but require knowledge of prior distributions to be implemented. Recent progress in Markov Chain Monte Carlo and related computational tools were discussed.

The issue of statistical software is thorny. The statistical community, and their client researchers in social and biological sciences, rely almost exclusively on commercial companies to produce statistical software. Astronomers, on the other hand, develop public (or very low cost) software. Due to this incompatibility, astronomy is often decades behind other research fields in implementing useful advanced methods. One outgrowth of the SCMA II meeting has been the establishment of a Web site for statistical codes useful to astronomers (<http://www.astro.psu.edu/statcodes>).

Acknowledgments

We would like to thank many people for their assistance in producing the conference and this book. The Scientific Organizing Committee – Peter Bickel, Bradley Efron, Robert Hanisch, William Jefferys, Fionn Murtagh, William Press, Brian Ripley, Grace Yang – contributed to the scientific program. At Penn State University, the conference received support from Gregory Geoffroy, Dean of the Eberly College of Sciences; James Rosenberger, Head of the Department of Statistics; Peter Mészárós, Head of the Department of Astronomy & Astrophysics. Debby Noyes and Amy Spangler of Continuing Distance & Education helped with logistic arrangements.

The conference would not have been possible without the support of the leading international organizations in the two fields and U.S. funding agencies. The conference was co-sponsored by the International Astronomical Union (as an IAU Technical Workshop), International Statistical Institute and Institute for Mathematical Statistics. Travel grants were provided by the National Science Foundation, National Aeronautical and Space Administration, International Astronomical Union and International Science Federation. Penn State University provided support for conference organization. Our astrostatistical research at Penn State is supported by NSF grant DMS-9626189 and NASA grant NAGW-2120.

Finally, we would like to thank the invited speakers, commentators and panel discussants for their fine presentations and papers.

G. Jogesh Babu
Eric D. Feigelson

Contents

Preface	v
Acknowledgments	vii
List of Participants	xiii

General Methods in Astrostatistics

1 Pre and Post Least Squares: The Emergence of Robust Estimation	3
C. Radhakrishna Rao	
2 Some Recent Developments in Bayesian Analysis, with Astronomical Illustrations	15
James O. Berger	
Discussion by Alanna Connors	39
3 Bayesian Analysis of Lunar Laser Ranging Data	49
William H. Jefferys and Judit Györgyey Ries	
Discussion by Steven F. Arnold	63
4 Modern Statistical Methods for Cosmological Testing	67
I. E. Segal	
5 Comparing Censoring and Random Truncation via Nonparametric Estimation of a Distribution Function	83
Grace L. Yang	
Discussion by David M. Caditz	100
Response by Grace L. Yang	103
6 Astronomical (Heteroscedastic) Measurement Errors: Statistical Issues and Problems	105
Michael G. Akritas	
Discussion by William H. Jefferys	118
7 New Problems and Approaches Related to Large Databases in Astronomy	123
Fionn Murtagh and Alex Aussem	

8 Object Classification in Astronomical Images	135
Richard L. White	
Discussion by Francisco G. Valdes	149
9 Recent Advances in Large-scale Structure Statistics	153
Vicent J. Martínez	
Discussion by Michael L. Stein	166
10 Wavelet Transform and Multiscale Vision Models	173
Albert Bijaoui, Frédéric Rué and Renaud Savalle	
11 Statistical Software, Software and Astronomy	185
Edward J. Wegman, Daniel B. Carr, R. Duane King, John J. Miller, Wendy L. Poston, Jeffrey L. Solka, and John Wallin	
Discussion by John Nousek	203

Major Astronomical Projects

12 Statistical Issues in the MACHO Project	209
T. S. Axelrod, C. Alcock, R. A. Allsman, D. Alves, A. C. Becker, D. P. Bennett, K. H. Cook, K. C. Freeman, K. Griest, J. Guern, M. J. Lehner, S. L. Marshall, B. A. Peterson, M. R. Pratt, P. J. Quinn, A. W. Rodgers, C. W. Stubbs, W. Sutherland, and D. L. Welch	
13 LIGO: Identifying Gravitational Waves	225
Bernard F. Schutz and David Nicholson	
Discussion by Curt Cutler	237
14 AXAF Data Analysis Challenges	241
Aneta Siemiginowska, Martin Elvis, Alanna Connors, Peter Freeman, Vinay Kashyap, and Eric Feigelson	
Discussion by Joseph Horowitz	254
15 Statistical Aspects of the Hipparcos Photometric Data	259
F. van Leeuwen, D. W. Evans and M. B. van Leeuwen-Toczek	
Discussion by P. J. Bickel	275

Time Series Analysis

16 Application of Wavelet Analysis to the Study of Time-dependent Spectra	283
M. B. Priestley	
17 Nonparametric Methods for Time Series and Dynamical Systems	303
D. Guégan	
Discussion by Jeffrey D. Scargle	317
18 Quantifying Rapid Variability in Accreting Compact Objects	321
M. van der Klis	

19 Wavelet and Other Multi-resolution Methods for Time Series Analysis	333
Jeffrey D. Scargle	

Summaries

20 An Overview of “SCMA II”	351
P. J. Bickel	
21 Late-Night Thoughts of a Classical Astronomer	365
Virginia Trimble	

Contributed Papers

22 Algorithms for the Detection of Monochromatic and Stochastic Gravitational Waves	389
Pia Astone	
23 Statistical Tests for Changing Periods in Sparsely Sampled Data	391
Paul Hertz	
24 Analyzing X-ray Variability by Linear State Space Models	393
Michael König and Jens Timmer	
25 The Time Interferometer: Spectral Analysis of the Gapped Time Series from the Stand Point of Interferometry	395
V. V. Vityazev	
26 Structures in Random Fields	397
Juan E. Betancort-Rijo	
27 The New γ-CFAR Detector For Astronomical Image Processing	401
A. D. Nair, Jose C. Principe and Munchurl Kim	
28 Bayesian Image Reconstruction with Noise Suppression	403
Jorge Núñez and Jorge Llacer	
29 Astronomical Images Restoration by the Multiscale Maximum Entropy Method	405
Jean-Luc Starck and Eric Pantin	
30 Nested Test for Point Sources	407
James Theiler and Jeff Bloch	
31 Segmenting Chromospheric Images with Markov Random Fields	409
Michael J. Turmon and Judit M. Pap	
32 Analysis of Hipparcos Data in the Orthonormal Wavelet Representation	413
E. Chereul, M. Crézé, and O. Bienaymé	
33 Statistical Properties of Wavelet Transforms Applied to X-Ray Source Detection	417
F. Damiani, A. Maggio, G. Micela and S. Sciortino	
34 Wavelet Based X-Ray Spatial Analysis: Statistical Issues	419
V. Kashyap and P. Freeman	

35 Wavelet and Multifractal Analyses of Spatial and Temporal Solar Activity Variations	421
J. K. Lawrence, A. C. Cadavid and A. A. Ruzmaikin	
36 Wavelets in Unevenly Spaced Data: OJ 287 light curve	423
Harry J. Lehto	
37 Wavelet Based Analysis of Cosmic Gamma-Ray Burst Time Series	427
C. Alex Young, Dawn C. Meredith, and James M. Ryan	
38 Smoothed Nonparametric Density Estimation for Censored or Truncated Samples	429
David M. Caditz	
39 Luminosity and Kinematics: A Maximum Likelihood Algorithm for Exploitation of the Hipparcos Data	433
X. Luri, M. O. Mennessier, F. Figueras, J. Torra, and A. E. Gómez	
40 Assessing Statistical Accuracy to the Orbital Elements of Visual Double Stars by Means of Bootstrap	437
G. Ruymakers and J. Cuypers	
41 A Poisson Parable: Bias in Linear Least Squares Estimation	441
Wm. A. Wheaton	
42 Neural Network Classification of Stellar Spectra	445
Coryn Bailer-Jones, Mike Irwin, and Ted von Hippel	
43 Multidimensional Index for Highly Clustered Data with Large Density Contrasts	447
I. Csabai, A. Szalay, R. Brunner, and K. Ramaiyer	
44 Quantitative Morphology of Moderate Redshift Peculiar Galaxies	449
Avi Naim, Kavan U. Ratnatunga and Richard E. Griffiths	
45 Bayesian Inference on Mixed Luminosity Functions	451
David Walshaw	
46 Stochastic Solutions for the Hipparcos Astrometric Data Merging	455
F. Arenou, L. Lindegren, and R. Wielen	
47 Identification of Nonlinear Factors in Cosmic Gamma-ray Bursts	457
Anton M. Chernenko	
48 Statistical Challenges in Asteroid Dynamics	459
S. Dikova	
49 Brief Annotated Bibliography on Point Processes	461
Joseph Horowitz	
50 Testing the Hubble Law from Magnitude-Redshift Data of Field Galaxies: The Effect of the Shape of the Luminosity Function	463
O. Ullmann	
Index	465

List of Participants

- Mark A. Abney** Department of Astronomy and Astrophysics, University of Chicago, 5640 S. Ellis Ave., Chicago, IL 60637
- Carmen O. Acuna** Department of Mathematics, Lewisburg, PA 17837
- Michael G. Akritas** Department of Statistics, Pennsylvania State University, 326 Thomas Building, University Park, PA 16802
- Charles H. Anderson** Washington University Medical School, Box 8108, S. Euclid Ave., Saint Louis, MO 63110
- Frederic Arenou** Observatoire de Paris-Meudon, 5 Place J. Janssen, Meudon 82195, France
- Steven F. Arnold** Department of Statistics, Pennsylvania State University, 313 Thomas Building, University Park, PA 16802
- Pia Astone** University La Sapienze, P. Aldo Moro 2, Rome 00185, Italy
- Tim S. Axelrod** Mount Stromlo Observatory, Weston Creek, ACT 2611, Australia
- G. J. Babu** Department of Statistics, Pennsylvania State University, 326 Thomas Building, University Park, PA 16802
- Sandy D. Balkin** Department of Statistics, Pennsylvania State University, 326 Thomas Building, University Park, PA 16802
- James O. Berger** Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251
- Juan E. Betancort-Rijo** Instituto de Astrofisica de Canarias, Via Lactea S/M, La Laguna, E-38200 Tenerife, Canary Islands, Spain
- Peter J. Bickel** Department of Statistics, University of California, 367 Evans Hall, Berkeley, CA 94720
- Albert Bijaoui** Observatoire de la Cote d'Azur, B.P. 229, Nice Cedex 06304, France
- Jerome J. Brainard** Space Sciences Laboratory, ES-84, NASA Marshall Space Flight Center, Huntsville, AL 35812
- Michael S. Briggs** Department of Physics, University of Alabama - Huntsville, OB 201-C, Huntsville, AL 35899

- Robert F. Burns** 400 West 58th St., Apt. 38, New York, NY 10019
- David N. Burrows** Department of Astronomy & Astrophysics, Pennsylvania State University, 525 Davey Laboratory, University Park, PA 16802
- David M. Caditz** Department of Physics, Montana State University, Bozeman, MT 59717
- George Chartas** Department of Astronomy & Astrophysics, Pennsylvania State University, 525 Davey Laboratory, University Park, PA 16802
- Emmanuel Chereul** Observatoire de Strasbourg, 11 rue de l'Université, Strasbourg 67000, France
- Anton M. Chernenko** Space Research Institute, Profsojuznaja 84/32, Moscow 117810, Russia
- Mary C. Christman** Department of Mathematics and Statistics, American University, 4400 Massachusetts Ave. NW, Washington, DC 20016-8050
- Leon Cohen** City University, 695 Park Avenue, New York, NY 10047
- Alanna Connors** Space Science, University of New Hampshire, Morse Hall, Durham, NH 03824
- Istvan Csabai** Department of Physics & Astronomy, Johns Hopkins University, Charles and 34th Street, Baltimore, MD 21218
- Curt J. Cutler** Department of Physics, Pennsylvania State University, 104 Davey Laboratory, University Park, PA 16802
- Francesco Damiani** Observatorio Astronomico di Palermo, Piazza del Parlamento 1, Palermo 90134, Italy
- Victor de Oliveira** Department of Mathematics, University of Maryland, College Park, MD 20742
- Smiliana D. Dikova** Institute of Astronomy, Bulgarian Academy of Sciences, Tzarigradsko Schosses 72, Sofia 1784, Bulgaria
- Debiprosad Duari** Theoretical Astrophysics Group, Tata Institute of Fundamental Research, Homi Bhabha Road, Bombay 400005, India
- Suzanne R. Dubnicka** Department of Statistics, Pennsylvania State University, 330A Thomas Building, University Park, PA 16802
- Bradley Efron** Department of Statistics, Stanford University, Stanford, CA 94305
- Eric D. Feigelson** Department of Astronomy & Astrophysics, Pennsylvania State University, 525 Davey Laboratory, University Park, PA 16802
- Grant Foster** American Association of Variable Star Observers, 25 Birch Street, Cambridge, MA 02138
- Peter E. Freeman** Department of Astronomy, University of Chicago, 55640 S. Ellis Ave., Chicago, IL 60637
- Audrey B. Garmire** Department of Astronomy & Astrophysics, Pennsylvania State University, 525 Davey Laboratory, University Park, PA 16802

- Gordon P. Garmire** Department of Astronomy & Astrophysics, Pennsylvania State University, 525 Davey Laboratory, University Park, PA 16802
- Madhumita Ghosh-Destidar** Department of Statistics, Pennsylvania State University, 330A Thomas Building, University Park, PA 16802
- Richard E. Griffiths** Department of Physics & Astronomy, Johns Hopkins University, 3400 Charles Street North, Baltimore, MD 21218
- Dominique Guégan** Department of Statistics, ENSAE, Timbre J120, 3 Avenue Pierre Larousse, 92245 Malakoff Cedex, France
- Hasan N. Hamdan** Department of Mathematics & Statistics, American University, 4400 Massachusetts Ave. NW, Washington, DC 20016
- Paul L. Hertz** Space Science Division, Naval Research Laboratory, Code 7621, Washington, DC 20375-5352
- Tim C. Hesterberg** Mathematics Department, Franklin & Marshall College, Lancaster, PA 17604-3003
- Joseph Horowitz** Department of Mathematics & Statistics, University of Massachusetts, Amherst, MA 01003
- Nan-Jung Hsu** Department of Statistics, Iowa State University, Snedecor Hall, Ames, IA 50014
- Hsin-cheng Huang** Department of Statistics, Iowa State University, 204 Snedecor Hall, Ames, IA 50014-3571
- William H. Jefferys** Department of Astronomy, University of Texas, Austin, TX 78703
- Jiming Jiang** Department of Statistics, Case Western Reserve University, Cleveland, OH 44106
- Coryn A. L. Jones** Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, United Kingdom
- M. Chris Jones** Department of Statistics, Open University, Walton Hall, Milton Keynes MK7 6AA, United Kingdom
- Vinay L. Kashyap** AXAF Science Center, Department of Astronomy, University of Chicago, 5640 S. Ellis Ave., Chicago, IL 60637
- Kristin E. Kearns** Department of Astronomy, Wesleyan University, Middletown, CT 06459
- Do J. M. Kester** Space Research Organization of the Netherlands, P.O. Box 800, Groningen 9700, Netherlands
- Michael König** Institute of Astronomy, Tübingen University, Waldhäuserstrasse 64, Tübinben D-72076, Germany
- Soundar Kumara** Department of Industrial Engineering, The Pennsylvania State University, University Park, PA 16802
- William D. Langer** Jet Propulsion Laboratory, California Institute of Technology, MS 169-506, 4800 Oak Grove Dr., Pasadena, CA 91109
- John K. Lawrence** Department of Physics & Astronomy, California State University, 1811 Nordhoff St., Northridge, CA 91330-8268
- Harry J. Lehto** Tuorla Observatory, University of Turku, Väisäläntie 20, FIN-21500, Piikkiö, Finland

- Bing Li** Department of Statistics, Pennsylvania State University, 326 Thomas Building, University Park, PA 16802
- Haihong Li** Department of Statistics, Pennsylvania State University, 326 Thomas Building, University Park, PA 16802
- Suzanne Linder** Department of Astronomy & Astrophysics, Pennsylvania State University, 525 Davey Laboratory, University Park, PA 16802
- Bruce G. Lindsay** Department of Statistics, Pennsylvania State University, 422 Thomas Building, University Park, PA 16802
- Marc A. Loizeaux** Department of Statistics, Florida State University, Tallahassee, FL 32308
- Thomas Loredo** Department of Astronomy, Cornell University, Space Sciences Building, Ithaca, NY 14853-6801
- Xavier Luri** Departament d'Astronomia i Meteorologia, Universitat de Barcelona, Avda. Diagonal 647, Barcelona 08028, Spain
- Fernando Marques** Villafranca Satellite Tracking Station, European Space Agency, P.O. Box 50727, Madrid 28080, Spain
- Herman L. Marshall** Center for Space Research, Massachusetts Institute of Technology, 37-667A, Cambridge, MA 02139
- Vicent J. Martínez** Department of Astronomy, University of Valencia, Av. Dr. Moliner 50, Burjassot E-46100, Spain
- Mary Sara McPeek** Department of Statistics, University of Chicago, 5734 S. University Ave., Chicago, IL 60637
- Soma Mukherjee** Physics & Applied Mathematics Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India
- Fionn D. Murtagh** Faculty of Informatics, University of Ulster, Londonderry BT48 7JL, United Kingdom
- Avi Naim** Department of Physics & Astronomy, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218
- Achotham D. Nair** Department of Astronomy, University of Florida, 217 S.S.R.B., Gainesville, FL 33611
- John A. Nosek** Department of Astronomy & Astrophysics, Pennsylvania State University, 525 Davey Laboratory, University Park, PA 16802
- Jorge C. Núñez** Department D' Astronomia, Universitat De Barcelona, Av. Diagonal 647, Barcelona E-08028, Spain
- John T. O'Gorman** Department of Statistics, Pennsylvania State University, 326 Thomas Building, University Park, PA 16802
- Vahé Petrosian** Stanford University, Varian Building, Rm. 302C, Stanford, CA 94305-4060
- Harri T. Pietila** Tuorla Observatory, University of Turku, Vaisalantie 20, Piikkio 21500, Finland
- Jennifer L. Pittman** Department of Statistics, The Pennsylvania State University, 301 Thomas Building, University Park, PA 16802
- Maurice B. Priestley** Department of Mathematics, Institute of Science and Technology, University of Manchester, Manchester M60 10D, United Kingdom

- Francis A. Primini** Smithsonian Astrophysical Observatory, 60 Garden St., Cambridge, MA 02138
- Lianfen Qian** Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824
- Calyampudi R. Rao** Statistics Department, The Pennsylvania State University, 417 Thomas Building, University Park, PA 16802
- Makarand V. Ratnaparkhi** Wright State University, 4821 Silver Oak St., Dayton, OH 45424
- Kavan U. Ratnatunga** Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218
- Frederick A. Ringwald** Department of Astronomy and Astrophysics, The Pennsylvania State University, 525 Davey Lab, University Park, PA 16802
- Goedele Ruymakers** Royal Observatory of Belgium, Ringlaan 3, Brussels B-1180, Belgium
- Jeffrey D. Scargle** NASA-Ames, Mail Stop 245-3, Moffett Field, CA 94035-1000
- Bernard F. Schutz** Albert Einstein Institute, Schlaatzweg 1, Potsdam D-14473, Germany
- Karen G. Shaefer** Space Telescope Science Institute, 3700 San Martin Drive, Baltimore MD 21218
- Irving E. Segal** Mass. Institute of Tech., Rm. 2-244, 77 Mass. Ave., Cambridge, MA 02139
- Richard A. Shaw** Space Telescope Science Institute, 3700 San Martin Dr., Baltimore, MD 21218
- Leon H. Sibul** ARL, P.O. Box 30, The Pennsylvania State University, State College, PA 16804
- Aneta Siemiginowska** AXAF Science Center, 60 Garden ST., MS 70, Cambridge, MA 02138
- Jean-Luc Starck** CEA, 91 191 Gif sur Yvette, Cedex, France
- Michael L. Stein** Department of Statistics Department of Statistics, University of Chicago, 5734 University Ave., Chicago, IL 60637.
- Mitchell F. Struble** Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104-6394
- Edmund C. Sutton** University of Illinois, 1002 W. Green ST., Urbana, IL 61801
- Jean H. Swank** NASA/GSFC, Code 662 NASA/GSFC, Greenbelt, MD 20740
- Alex Szalay** Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, MD 21218
- HenSiong Tan** Department of Statistics, The Pennsylvania State University, 326 Thomas Building, University Park, PA 16802
- James P. Theiler** Los Alamos National Laboratory, MS-D436, Los Alamos, NM 87545
- David J. Thomson** Bell Labs, 2C-360 Bell Labs, Murray Hill, NJ 079974

- Kip S. Thorne** CALTECH, 130-33 Caltech, Pasadena, CA 91125
- Jordi Torra** Departament d'Astronomia i Meteorologia, Universitat de Barcelona, Avda. Diagonal 647, Barcelona 08028, Spain
- Leisa K. Townsley** Astronomy and Astrophysics, The Pennsylvania State University, 525 Davey Lab, University Park, PA 16802-6305
- Virginia L. Trimble** Physics Department, University of California Irvine, Irvine, CA 92717-4575
- Michael Turmon** JPL/Caltech, M/S 525-3550, 4800 Oak Grove Drive, Pasadena, CA 91109
- Oskar Ullmann** MPA Garching, Karl-Schwarzschild-Strasse 7, D-Garching D-8046, Germany
- Esa I. Uusipaikka** University of Turku, Porvarinkuja 1 i 34, Helsinki 00750, Finland
- Francisco G. Valdes** NOAO/IRAF Group, Box 25732, Tucson, AZ 85726
- Michiel van der Klis** Sterrekundig Instituut 'Anton Pennekkoek', Universiteit van Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
- Floor van Leeuwen** Royal Greenwich Observatory, Madingley Road, Cambridge CB30E2, United Kingdom
- Veniamin Vityazev** St. Petersburg University, Astronomy Department, ST Petersburg 198904, Russia
- David Walshaw** Speech Department, University of Newcastle, King George VI Bldg, Queen Victoria Road, Newcastle Upon Tyne, United Kingdom
- Yazhen Wang** Department of Statistics, University Of Missouri-Columbia, Columbia, MO 65211
- Edward J. Wegman** Center for Computational Statistics, George Mason University, 157 Sci-Tech II, Fairfax, VA 22030
- William A. Wheaton** JPL 169-327, 4800 Oak Grove Dr., Pasadena, CA 91109
- Richard L. White** Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218
- Carolee J. Winstein** University of Southern California, 1540 East Alcazar St., CHP 155, Los Angeles, CA 90033
- Grace L. Yang** Department of Mathematics, University of Maryland, College Park, MD 20742
- Alex Young** University of New Hampshire, 311 Morse Hall-UNH, Durham, NH 03824
- Saleem Zaroubi** Department of Astronomy, Campbell Hall, University of California, Berkeley, CA 94720

General Methods in Astrostatistics

A wide variety of statistical issues are addressed in the first portion of the book. RAO discusses the history of least-squares and robust estimation methods. BERGER, JEFFERYS and commentators CONNORS and ARNOLD present powerful Bayesian ideas and techniques. Nonparametric treatments of truncated and censored datasets arise in flux-limited astronomical surveys. Their applications to luminosity functions and cosmological modeling are considered by SEGAL, YANG and commentator CADITZ. AKRITAS and commentator JEFFERYS consider the common but tricky problems caused by heteroscedastic measurement errors in astronomical data.

Several authors discuss a wide range of multivariate problems. WHITE and commentator VALDES review multivariate techniques to classify stars and galaxies in digitized sky surveys. MURTAGH and BIJAOUI present innovative wavelet and multiscale approaches to database and image analysis. MARTÍNEZ and commentator STEIN update galaxy clustering statistics, which were extensively discussed in the first SCMA conference. Finally, WEGMAN and commentator NOUSEK discuss the wide range of statistical software and their applicability to astronomical problems.

Pre and Post Least Squares: The Emergence of Robust Estimation

C. Radhakrishna Rao¹

ABSTRACT The paper traces the history of estimation of unknown parameters when measurements are subject to error from the time of Ptolemy to Gauss and Laplace, the inventors of the method of least squares estimation (LSE). The modern theory of LSE started with the papers by Markoff and Aitken and later contributions by Bose and the author. The LSE's have some nice properties. However, they are found to be sensitive to outliers and contamination in the data. To overcome this defect, robust methods are introduced using measures of discrepancy between a measurement and its expected value which have a slower rate of growth than the squared value. A unified theory of robust estimation is described using the difference of two convex functions as the measure of discrepancy.

1.1 Introduction

Almost all significant developments in statistical methodology were motivated by problems in biological and social sciences. The names of journals in statistics like *Biometrika*, *Biometrics*, *Psychometrics*, *Econometrics* and *Technometrics* signify such a link between statistics and problems in real life. These developments, however, took place in the twentieth century based on the advances in probability theory and codification of inductive reasoning by which decisions could be made under uncertainty. Astronomers from Kepler (1571-1630) and Newton (1642-1727) of the earlier centuries to Newcomb (1835-1909) and Poincare (1854-1912) of the last century tried to apply probabilistic reasoning to astronomical phenomena, but much progress could not be made because of the inadequacy of the available analytical tools. In one area, however, probabilistic reasoning to

¹Center for Multivariate Analysis, Statistics Department, 417 Thomas Building, Penn State University, University Park, PA 16802

Research supported by the Army Research Office under Grant DAAHO4-96-1-0082.

astronomy led to a result of signal and lasting importance: the theory of observational errors. From the time of Ptolemy, astronomers were faced with the problem of obtaining best estimates of unknown parameters from measurements subject to error. Various attempts, partly objective and partly subjective, were made over the last five centuries, which finally led to the discovery of the method of least squares estimation (LSE) in the beginning of the last century, in which Gauss (1777-1855) and Laplace (1749-1827) played major roles.

LSE's have nice properties when errors are normally distributed. However, they are sensitive to contamination in the data and the presence of outliers. Attempts have been made during the last thirty years to introduce methods of estimation by which the influence of outliers and departure from the normality of the distribution of the errors could be minimized. The object of this paper is to provide a historical account of the evolution of LSE and the recent emergence of robust methods as alternatives to LSE.

1.2 Pre least squares

While estimating an unknown parameter when several measurements were made, some of the earlier astronomers selected among the available observations the one they believed to be the best (Tycho Brahe (1546-1601), Johannes Hevelius (1611-1687), John Flamsteed (1646-1719, England's first Astronomer Royal), while others used the arithmetic mean or the one observation closest to the arithmetic mean (Johannes Kepler (1571-1630), James Bradley, Nicolas-Louis de Lacaille), or the weighted mean (Roger Cotes (1682-1716)). When more than one parameter, say $\theta_1, \dots, \theta_p$, is involved and the measurements, m_1, \dots, m_n , refer to known linear combinations of the parameters, so that

$$a_{1i}\theta_1 + \dots + a_{pi}\theta_p = m_i, \quad i = 1, \dots, n \quad (1.1)$$

omitting the error term, the problem is to estimate $\theta_1, \dots, \theta_p$ when $n > p$. Leonard Euler (1707-1783) and Tobias Mayer (1723-1762) used some intuitive methods of combining some of the equations to reduce the number of n “conditions”, as the equations (1.1) were called, to p , the number of parameters, and then solving the p equations in p unknowns.

The first objective method of estimation was developed by Rudjer, J. Bošković (1711-1787). Denoting the difference, $m_i - a_{1i}\theta_1 - \dots - a_{ip}\theta_p$ by d_i , Bošković suggested the estimation of $\theta_1, \dots, \theta_p$ by minimizing

$$\sum |d_i| \quad \text{subject to} \quad \sum d_i = 0. \quad (1.2)$$

He provided a geometric solution to the problem. Later Pierre-Simon Laplace (1749-1827) provided the algebraic steps by which Bošković's problem could be solved.

An interesting contribution, a forerunner of the method of maximum likelihood, was due to Johann Heinrich Lambert (1728-1777), who suggested estimation of a parameter θ by minimizing

$$\Phi(x_1 - \theta)\Phi(x_2 - \theta) \dots \Phi(x_n - \theta) \quad (1.3)$$

choosing $\Phi(x - \theta)$ as the frequency curve for the error $x - \theta$ (observed value minus the parameter value). In most cases, he stated that $\hat{\theta}$ the minimizer of (1.3) would not differ very much from the arithmetic mean.

Thomas Simpson (1710-1761) and later Joseph Louis Lagrange (1736-1813) tried to justify the use of the arithmetic mean in the case of one parameter by determining its distribution under simple models for the error distribution. They showed that the arithmetic mean is a reasonable estimator in all the cases they studied, and it is better than a single observation.

1.3 Invention of the method of least squares

In the last decade of the eighteenth century and the first decade of the nineteenth century, a number of mathematicians independently hit upon the idea of determining the unknowns of a set of equations of condition (1.1) by minimizing the sum of squares of the residuals, which was later to be called the method of “least squares” by Legendre (moindre carrés). It was assumed that such a method would yield the most probable values of the unknowns.

Peter Merian wrote in an obituary note that Daniel Huber (1768-1830), a Swiss mathematician, invented the method of least squares estimation (LSE), but published nothing about it. Adrien-Marie Legendre (1752-1833) mentioned in an appendix to one of his books, dated March 6, 1805, the LSE as the most general and most accurate way of determining several unknowns from a larger number of equations of condition. He derived the rule of the arithmetic mean using LSE in the case of a single unknown, but did not supply any justification. In 1808, Robert Adrian (1775-1843) published a paper in which he derived the normal law of errors and used it to establish the principle of least squares. The first published work on LSE was in 1809 by Carl Friedrich Gauss (1777-1855), where he used probabilistic arguments. Assuming that $e_i = y_i - \xi_i(\theta)$, where ξ_i is a known function of the unknown vector parameter $\theta' = (\theta_1, \dots, \theta_k)$, are independent random variables with distribution $f(e_i)$ for e_i , Gauss suggested the most probable value of θ as the one which maximizes $\prod f(e_i)$. This is obtained as a root of

$$\sum \frac{\partial \log f(y_i - \xi_i)}{\partial \xi_i} \frac{\partial \xi_i}{\partial \theta_j} = 0, \quad j = 1, \dots, k. \quad (1.4)$$

To proceed further, the mathematical form of f must be known. Gauss took the special case $y_i = \theta_1 + e_i$ for all i and showed that his method will yield the arithmetic mean of y_i 's as the desired estimate of θ_1 , when f has normal distribution. He then recommended the use of the normal distribution for f in the general equation (1.4). Gauss also mentioned in his 1809 memoir that he had used this method as early as in 1795.

Gauss further asserted that the alternative principle of Bošković and Laplace - the method of seeking a minimum of the sum of absolute values of the errors - is unsatisfactory because the estimates come out as the solution of a subset of equations of condition equal in number to that of the unknown parameters, although the remaining equations have an influence on the selection of equations to be exactly satisfied. In return Laplace criticized Gauss for his using a circular argument in deriving the method of least squares assuming that the arithmetic mean is the ideal one in the case of a single unknown.

In a memoir read to the Paris Academy in April 1810, Laplace established the central limit theorem stating the normal law as the limiting distribution of a linear function of random variables as the number of variables tend to infinity, and gave a new rationale, different from Gauss's, for the choice of a normal distribution of errors, hence a different argument for the method of least squares. Starting with observational equations, $y_i = x_i\beta + e_i$, $i = 1, \dots, n$, for one parameter β , Laplace considered a linear estimate $\hat{\beta} = \sum a_i y_i / \sum a_i x_i$ which is consistent for β in the sense that $\hat{\beta} = \beta$ if $\sum a_i e_i = 0$; the actual error $(\hat{\beta} - \beta) = \sum a_i e_i / \sum a_i x_i$. When n is large $\hat{\beta}$ will have a limiting normal distribution with variance proportional to $\sum a_i^2 / (\sum a_i x_i)^2$. Then the probability of $\hat{\beta}$ lying in any symmetric interval about β increases as variance decreases. Hence the optimum choice of a_i of the coefficients for which there is greater concentration of probability is when $\sum a_i^2 / (\sum a_i x_i)^2$ is a minimum, i.e., when a_i is proportional to x_i , giving the optimum $\hat{\beta}$ as $\sum x_i y_i / \sum x_i^2$, which is the LSE. Laplace went on to generalize his conclusions to problems with several unknowns in an argument that effectively derived the multivariate limiting distribution of two or more LSE's.

Gauss's new justification for LSE appeared in 1823. Using the linear model $Y = X\beta + \epsilon$, with $\text{cov}(\epsilon) = \sigma^2 I$, Gauss considered a linear estimate $\hat{\beta} = CY$, where C is such that $CX = I$, so that $\hat{\beta} = \beta$ when the errors are all zero (some kind of consistency condition). Then the estimator of $g'\beta$, a linear function of β , is $g'CY$ which has the mean square error $\sigma^2 g'CC'g$. Minimizing $g'CC'g$ subject to $CX = I$ yields the solution $C = (X'X)^{-1}X'$, which leads to the LSE of β . Thus was born the theory of LSE without any distributional assumptions, based only on the properties of unbiasedness (or consistency) and minimum variance.

The historical account of this section is based on a book by Stigler (1986).

1.4 Consolidation of least squares

1.4.1 Gauss-Markoff Theorem

In 1900, A.A. Markoff (1856-1922) introduced the linear model

$$Y = X\beta + \epsilon, \quad \text{Cov}(\epsilon) = \sigma^2 I \quad (1.5)$$

where $\rho(X)$, the rank of X , is equal to m , the size of the vector parameter β to be estimated and showed, exactly as Gauss did, that the best linear estimate (i.e., unbiased with minimum variance) of $p'\beta$, a linear function of β , is $p'\hat{\beta}$ where

$$\hat{\beta} = \arg \min_{\beta} (Y - X\beta)'(Y - X\beta). \quad (1.6)$$

However, in statistical literature the statistical inference associated with the model (1.5) is known as Gauss-Markoff theory. The $\hat{\beta}$ in (1.6) is called the LSE (least squares estimate) of β .

1.4.2 Aitken LSE

Gauss also considered the model (1.5) with $\text{Cov}(\epsilon) = \Delta$, a nonsingular diagonal matrix, and showed that the best estimate of β (in the above sense) is

$$\hat{\beta} = \arg \min_{\beta} (Y - X\beta)' \Delta^{-1} (Y - X\beta). \quad (1.7)$$

The next step, when $\text{Cov}(\epsilon) = \sigma^2 \Sigma$, a general positive definite matrix, was taken by Aitken (1935), who showed that the best estimate of β is

$$\hat{\beta} = \arg \min_{\beta} (Y - X\beta)' \Sigma^{-1} (Y - X\beta) \quad (1.8)$$

which is referred to as Aitken LSE.

1.4.3 Contribution of Bose

In the Gauss-Markoff-Aitken method it is assumed that $\rho(X) = m$, the number of unknown parameters to be estimated. But in many practical applications, the linear model is formulated in such a way that $\rho(X) \leq m$. During the early forties R.C. Bose (1901-1987) observed that when $\rho(X) < m$ (i.e., when there does not exist a matrix C such that $CX = I$ as Gauss assumed), there are some linear functions of β that are not unbiasedly estimable. He derived the condition for unbiased estimability of $p'\beta$, a given linear parametric function, and developed a method of estimating it when estimable.

Rao (1946) showed that even when $\rho(X) < m$, the normal equation

$$(X'X)\beta = X'Y \quad (1.9)$$

obtained by minimizing $(Y - X\beta)'(Y - X\beta)$ in the usual way, admits a solution $\hat{\beta}$ which may not be unique, but $p'\hat{\beta}$, for any solution $\hat{\beta}$ of (1.9), is unique, unbiased for $p'\beta$ and has minimum variance among linear estimates provided $p'\beta$ is estimable. In later papers, Rao (1962, 1973) showed that a solution of (1.9) can be expressed as $\hat{\beta} = (X'X)^{-}X'Y$, where $(X'X)^{-}$ is any given g -inverse of $X'X$. (A^{-} is said to be a g -inverse, generalized inverse, of A if $AA^{-}A = A$). Also for estimable parametric functions $p'\beta$ and $q'\beta$,

$$\begin{aligned} \text{var}(p'\hat{\beta}) &= \sigma^2 p'(X'X)^{-}p, \quad \text{var}(q'\hat{\beta}) = \sigma^2 q'(X'X)^{-}q \\ \text{Cov}(p'\hat{\beta}, q'\hat{\beta}) &= \sigma^2 p'(X'X)^{-}q \end{aligned} \quad (1.10)$$

so that $\sigma^2(X'X)^{-}$ can be formally considered as $\text{Cov}(\hat{\beta})$. Thus the Gauss-Markoff method holds for estimable functions.

1.4.4 Unified Theory of Linear Estimation

If $\text{Cov}(\epsilon) = \sigma^2\Sigma$ is positive definite, the LSE of β is obtained by minimizing

$$(Y - X\beta)' \Sigma^{-1} (Y - X\beta) \quad (1.11)$$

a solution of which can be expressed as

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-}X'\Sigma^{-1}Y \quad (1.12)$$

where $(X'\Sigma^{-1}X)^{-}$ is any g -inverse of $X'\Sigma^{-1}X$, with the formal covariance matrix of $\hat{\beta}$ as

$$\text{Cov}(\hat{\beta}) = \sigma^2(X'\Sigma^{-1}X)^{-}. \quad (1.13)$$

The expressions (1.12) and (1.13) can be used as in (1.10) to estimate estimable parametric functions and to compute their variances and covariances. Thus Aitken's method works even when X is deficient in rank.

However, if Σ is singular, the formulas (1.11)-(1.13) do not hold by just replacing Σ by any kind of g -inverse of Σ . Rao (1973) showed that the expression to be minimized to obtain least squares estimates when X and Σ are possibly deficient in rank is the quadratic form

$$(Y - X\beta)'G^{-}(Y - X\beta) \quad (1.14)$$

where $G = \Sigma + XUX'$ with U chosen such that $\text{rank } G = \text{rank}(\Sigma : X)$ and G^{-} is any g -inverse of G , thus generalizing the theorem of Aitken. The most general theorem of least squares is then as follows.

Theorem (Rao (1973)). Let $Y = X\beta + \epsilon$ be a linear model with $E(\epsilon) = \sigma^2\Sigma$ where X and/or Σ may be deficient in rank and G, G^- and U be as defined above. Further let $\hat{\beta}$ be a minimizer of (1.14). Then

- 1) $p'\hat{\beta}$ is the minimum variance unbiased estimate of $p'\beta$ if estimable.
- 2)

$$V(p'\hat{\beta}) = \sigma^2 p'[(X'G^-X)^- - U]p,$$

$$\text{Cov}(p'\hat{\beta}, q'\hat{\beta}) = \sigma^2 p'[(X'G^-X)^- - U]q.$$

- 3) An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = f^{-1}(Y - X\hat{\beta})'G^-(Y - X\hat{\beta})$$

where $f = \rho(\Sigma : X) - \rho(X)$. [Note that when $\rho(\Sigma)$ is full, U can be chosen to be zero in which case $G^- = \Sigma^{-1}$ as in (1.11). Even if $\rho(\Sigma)$ is full, we can have any U with the condition $\rho(\Sigma : X) = \rho(\Sigma + XUX')$.]. For further details on g -inverses and their applications to linear models, reference may be made to Rao (1971) and Rao and Mitra (1971).

1.4.5 Linear Models without Moments

In the discussion of linear models in Sections 4.1-4.4, it is assumed that the error variables have second order moments. What properties does the LSE, $\hat{\beta} = (X'X)^{-1}X'Y$, have if the first and second order moments do not exist? The question is answered by Jensen (1979) when ϵ has a spherical distribution with the density of the form

$$\mathcal{L}(Y) = \sigma^{-n}\Psi_n\{(Y - X\beta)'(Y - X\beta)/\sigma^2\}. \quad (1.15)$$

We represent this class by $S_k(X\beta, \sigma^2 I)$, where k represents the integral order of moments which ϵ admits. If $k = 0$, no moments exist. Jensen (1979) proved among other results the following.

Theorem (Jensen, 1979). Consider $\hat{\beta} = (X'X)^{-1}X'Y$ as an estimator β in the model $Y = X\beta + \epsilon$. Then

- (i) If $\mathcal{L}(Y) \in S_0(X\beta, \sigma^2 I)$, then $\hat{\beta}$ is median unbiased for β and $\hat{\beta}$ is at least as concentrated about β as any other median unbiased estimator of β .
- (ii) If $\mathcal{L}(Y) \in S_1(X\beta, \sigma^2 I)$, then $\hat{\beta}$ is unbiased for β and is at least as concentrated around β as any other unbiased linear estimator.
- (iii) If $\mathcal{L}(Y) \in S_0(X\beta, \sigma^2 I)$ and in addition unimodal, then $\hat{\beta}$ is modal unbiased for β .

[Note that an s -vector $T \in R^s$ is said to be median unbiased for $\theta \in R^s$ if $a'T$ is median unbiased for $a'\theta$ for all $a \in R^s$; T is modal unbiased for θ if $\mathcal{L}(T)$ is unimodel and its mode coincides with θ . A measure $\mu(\cdot)$ on R^n is said to be more concentrated about $\theta \in R^n$ than $\nu(\cdot)$, if for each compact convex set A symmetric about θ under reflection, $\mu(A) \geq \nu(A)$.]

1.4.6 A Characterization of the LSE

Consider the model $Y = X\beta + \epsilon$ with $\text{Cov}(\epsilon) = \sigma^2 I$, $\rho(X) = m$, the size of vector β , and a submodel $Y_{(i)} = X_{(i)}\beta + \epsilon_{(i)}$ obtained by choosing $k \geq m$ rows of the original model. Further let

$$\hat{\beta} = (X'X)^{-1}X'Y, \quad \hat{\beta}_{(i)} = (X'_{(i)}X_{(i)})^{-1}X'_{(i)}Y_{(i)} \quad (1.16)$$

be the LSE's from the original and the submodel respectively. Subramaniam (1972) and Rao and Precht (1985) proved the following result.

Theorem. Denoting $d_{(i)} = |X'_{(i)}X_{(i)}|$, we have

$$\hat{\beta} = \sum_{i=1}^c d_{(i)}\hat{\beta}_{(i)} / \sum_{i=1}^c d_{(i)} \quad (1.17)$$

where c is the number of all possible subsets of size k from n .

The result (1.17), which expresses $\hat{\beta}$ as a weighted average of $\hat{\beta}_{(i)}$ is useful in regression diagnostics. We may calculate all possible $\hat{\beta}_{(i)}$, and look for consistency among them. If some appear to be much different from others, then we may examine the data for outliers or existence of clusters and consider the possibility of combining them with a different set of weights (some may be zero) than the those in (1.17). Further results of interest in this direction are contained in Wu (1986).

1.4.7 Asymmetric Least Squares (ALS)

Efron (1991) considered estimation of β as

$$\hat{\beta}_w = \arg \min_{\beta} (Y - X\beta)' \Delta_{\beta} (Y - X\beta) \quad (1.18)$$

where Δ_{β} is a diagonal matrix with 1 or $w (> 0)$ as the i -th diagonal element according as the i -th component of $Y - X\beta$ is ≤ 0 or > 0 . The plane (or the regression line) $y = x'\beta_w$ where y is the dependent variable and x is the concomitant is called the estimated $100p(w)$ th regression percentile, where $p(w)$ is the proportion of the components of $Y - X\hat{\beta}_w$ which are ≤ 0 . If we need the 100α th regression percentile for given α , w is so adjusted as to satisfy the relation $p(w) = \alpha$. Efron discusses the computation of such ALS estimates and their use in regression diagnostics and statistical inference.

1.5 Emergence of robust estimation

Consider the model $Y = X\beta + \epsilon$, $\text{Cov}(\epsilon) = \sigma^2 I$ and denote the i -th component of $Y - X\beta$ by $r_i = y_i - x'_i\beta$. The LSE of β is obtained by minimizing

$\sum r_i^2$. It is seen that the use of the discrepancy function r^2 of the residual gives undue influence to large residuals if there are outliers in the data. In order to downplay such influence, other functions of r which do not increase as rapidly as r^2 are chosen. Any such function may be denoted by $M(r)$ and the method of estimating β by minimizing $\sum M(r_i)$ is called M -estimation. Reference may be made to books by Huber (1981) and Hampel, Ronchetti, Rousseeuw and Stahel (1986), and a forthcoming *Handbook of Statistics*, Vol 15, edited by Maddala and Rao (1996), for details of such estimation and the choice of M functions.

It is observed by Windham (1994) that a general method of constructing an M function is to first choose a discrepancy function $f(r)$ which is, in some sense, appropriate when there are no outliers and contamination in the data and then robustize it by considering $c[f(r)]$, where c is an increasing concave function. The function to be minimized is $\sum c[f(r_i)]$. Windham (1994) shows that most of the M functions appearing in the literature on robust estimation are concave functions of $f(r) = r^2$, the basic discrepancy function used in the least square theory.

1.6 Unified theory of M -estimation

The theory of M -estimation using a convex discrepancy function under a minimal set of assumptions was developed in a series of papers (Bai, Rao and Wu (1991), Rao and Liu (1991), Rao and Zhao (1992) and Bai, Rao and Zhao (1993)). Recently the theory is extended to a general discrepancy function M , which is the difference of two convex functions under the same set of minimal conditions as in the case of a single convex function. Details are given in Bai, Rao and Wu (1996) where it is also shown that any discrepancy function can be approximated by the difference of two convex functions. Thus the theory based on the difference of two convex functions provides a unified treatment of M -estimation under a minimal set of conditions. We give two examples of well known discrepancy functions used in robust estimation.

For example consider the discrepancy function

$$f(r) = \begin{cases} r^2, & |r| \leq c \\ c^2, & |r| > c \end{cases}$$

where c is a constant. This can be expressed as the difference of the convex functions

$$\rho_1(r) = r^2 + c^2, \quad \rho_2(r) = \begin{cases} c^2, & |r| \leq c, \\ r^2, & |r| > c. \end{cases}$$

Another example is

$$f(r) = \begin{cases} r^2, & |r| \leq c \\ |r|, & |r| > c \end{cases}$$

where c is a constant. This can be expressed as the difference of the convex functions

$$\rho_1(r) = r^2 + |r|, \quad \rho_2(r) = \begin{cases} |r|, & |r| \leq c, \\ r^2, & |r| > c. \end{cases}$$

For details, reference may be made to Bai, Rao and Wu (1996).

REFERENCES

- [1] Aitken, A. C. (1935). On least squares and linear combination of observations. *Proc. Roy. Soc. Edin. A*, **55**, 42-48.
- [2] Bai, Z. D., Rao, C. R. and Wu, Y. (1991). Recent contributions to robust estimation. In *Probability, Statistics and Design of Experiments*, R. C. Bose Symposium Volume, Wiley Eastern, Ed. R. R. Bahadur, 30-50.
- [3] Bai, Z. D., Rao, C. R. and Zhao, L. C. (1993). Manova tests under a convex discrepancy function for the standard multivariate normal distribution. *J. Statist. Plann. Inference* **36**, 77-90.
- [4] Bai, Z. D., Rao, C. R. and Wu, Y. (1996). Robust inference in multivariate linear regression using difference of two convex functions as discrepancy measure. In *Handbook of Statistics*, Vol 15 (Eds G. S. Maddala and C. R. Rao), to appear.
- [5] Bose, R. C. (1950-51). Least Squares Aspects of the Analysis of Variance. *Mimeographed Series*, No.9, North Carolina University.
- [6] Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica* **1**, 93-126.
- [7] Gauss, K. F. (1809). *Werke* **4**, 1-93, Gottingen.
- [8] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics*. Wiley.
- [9] Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.
- [10] Jensen, D. R. (1979). Linear models without moments. *Biometrika* **66**, 611-617.
- [11] Maddala, G. S. and Rao, C. R. (1996). *Handbook of Statistics*, Vol 15, (in press).
- [12] Markoff, A. A. (1900). *Wrscheinlichkeitrechnung*, Tebner, Leipzig.
- [13] Rao, C. R. (1946). On the linear combination of observations and the general theory of least squares. *Sankhyā* **7**, 237-256.
- [14] Rao, C. R. (1962). A note on a generalized inverse of a matrix with applications to problems in mathematical statistics. *J. Roy. Statist. Soc. B* **24**, 152-158.
- [15] Rao, C. R. (1971). Unified theory of linear estimation. *Sankhyā A*, **33**, 371-374.
- [16] Rao, C. R. (1973). Unified theory of least squares. *Communications in Statistics* **1**, 1-18.
- [17] Rao, C. R. and Liu, Z. J. (1991). Multivariate analysis under M -estimation theory using a convex discrepancy function. *Biometric Letters* **28**, 89-95.
- [18] Rao, C. R. and Zhao, L. C. (1992). Linear representation of M -estimates in linear models. *Canadian J. Statistics* **20**, 359-368.
- [19] Rao, C. R. and Mitra, S. K. (1971). *Generalized Inverse of Matrices and its Applications*. Wiley, New York.

- [20] Rao, P. S. S. N. V. P. and Precht, M. (1985). On a conjecture of Hoel and Kennard on a property of least squares estimators of regression coefficients. *Linear Algebra and its Applications* **67**, 99-101.
- [21] Stigler, S. M. (1986). *The History of Statistics*. Harvard University Press.
- [22] Subramanyam, M. (1972). A property of simple least square estimates. *Sankhyā B*, **34**, 355-356.
- [23] Windham, M. P. (1994). Robust parameter estimation. *Tech. Report*.
- [24] Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261-1350.

Some Recent Developments in Bayesian Analysis, with Astronomical Illustrations

James O. Berger

ABSTRACT New developments in default Bayesian hypothesis testing and model selection are reviewed. As motivation, the surprising differences between Bayesian and classical answers in hypothesis testing are discussed, using a simple example. Next, an example of model selection is considered, and used to illustrate a new default Bayesian technique called the “intrinsic Bayes factor”. The example involves selection of the order of an autoregressive time series model of sunspot data. Classification and clustering is next considered, with the default Bayesian approach being illustrated on two astronomical data sets. In part, Bayesian analysis is experiencing major growth because of the development of powerful new computational tools, typically called Markov Chain Monte Carlo methods. A brief review of these developments is given. Finally, some philosophical comments about reconciliation of Bayesian and classical schools of statistics are presented.

2.1 Introduction

Bayesian Analysis and Astronomy have a long history together. Although Bayesian analysis is named after Thomas Bayes, who wrote the first paper on the subject (Bayes, 1783), it was Laplace who, in the late eighteenth and early nineteenth centuries, extensively developed the Bayesian approach to statistics, culminating in his revolutionary statistics book, Laplace (1812). Much of Laplace’s motivation in this development was the solution of problems in celestial mechanics.

The Bayesian approach to statistics, then called “inverse probability,” dominated the statistical scene for most of the nineteenth century, but fell into disfavor in the first half of the twentieth century. It has recently staged a major revival, and is indeed today the dominant approach in several statistical areas of interest to astronomy, such as image processing. Use of maximum entropy methods, which is a type of Bayesian analysis, has also become very popular in some astronomical circles. Additional insight into the use of Bayesian analysis in the astronomical community is available in

the excellent papers Loredo (1992) and Ripley (1992), from the previous conference.

Virtually any problem involving uncertainty can be approached from a Bayesian perspective, and the quantity of Bayesian methodological development being undertaken today is enormous. Hence a review of recent Bayesian developments or assessment of its future impact is virtually impossible. The goal of this paper is thus considerably more modest. We will simply try to illustrate some of the features of Bayesian analysis, with an eye towards helping astronomers to better judge whether or not they should consider using Bayesian methodology. There is a substantial learning curve involved in becoming adept at use of Bayesian methodology, and so some indications of the rewards and potential uses of the methodology can be helpful in deciding whether or not to invest effort in this direction.

A personal caveat is in order before proceeding. I have had little direct involvement in the analysis of astronomical problems, and so the illustrations I will consider are rather sterile, containing rather minimal astronomical content; indeed, the illustrations might even be “bad science” in terms of astronomical understanding. I hope readers will not be too distracted by this, and that at least some of the potential of Bayesian analysis is nevertheless apparent.

Related to this is the caveat that I will be focusing primarily on ‘default’ Bayesian methodology. This can best be described as a toolkit of Bayesian procedures which can be used automatically, much the same as many standard classical statistical procedures. In a sense, limiting the paper to this subject is somewhat unfortunate, because the Bayesian paradigm is much more; indeed, many argue that its greatest strength is to allow completely general interactive modelling between science, data, and opinions of the investigator, to reach a holistic end. My own experiences in this regard are not within astronomy, however, and hence I cannot provide an astronomical illustration of this possibility.

Even within default Bayesian analysis, I will primarily focus on one issue: use of Bayesian methods in hypothesis testing and model selection. This area is of considerable interest because it is an area in which Bayesian answers systematically differ from classical answers. (In contrast, when dealing with estimation or prediction problems for reasonably large data sets, there are typically only modest differences between classical and Bayesian answers.) Section 2 introduces the subject through an elementary, but surprising, example. Section 3 considers a more involved application to time series analysis. Section 4 deals with analysis of mixture problems, with applications to clustering.

One of the reasons for the recent upsurge in use of Bayesian methods is the advent of powerful computational engines based on Markov Chain Monte Carlo procedures. In Section 5 we give a brief introduction to these techniques. Section 6 concludes with some philosophical perspective.

2.2 Bayesian hypothesis testing and model selection

2.2.1 Motivation and a simple example

The Bayesian approach to hypothesis testing and model selection is conceptually straightforward. One assigns *prior* probabilities to all unknown hypotheses or models, and uses probability theory to compute the *posterior* probabilities of the hypotheses or models, given observed data. The key advantage of this approach is that the answers can be interpreted as the actual probabilities that the hypotheses or models are true, in contrast to standard significance testing which lacks any such interpretation. While there is a common intellectual appreciation of the difference, the impact of this intellectual appreciation upon practice seems to be quite limited; the following illustration of the difference is useful in helping to convince oneself that it is a serious matter.

Example 1. Suppose we observe independent $N(\theta, 1)$ data (i.e., data following the normal or Gaussian distribution with mean θ and variance 1), and that it is desired to test $H_1 : \theta = 0$ versus $H_2 : \theta \neq 0$. The typical classical procedure that is used is to compute the P -value or observed significance level, and to consider there to be significant evidence against H_1 if this P -value is small enough.

The following interesting simulation is easy to perform: repeatedly generate random data from H_1 and H_2 , corresponding to a series of tests of H_1 versus H_2 , and see where the series of P -values happens to fall. For data generated from H_1 , this is no mystery; by definition, the fraction of P -values that would fall in a given interval, say $0.04 < P\text{-value} < 0.05$, would be the size of the interval (here 0.01). The *nonshaded* vertical bars in Figure 1 reflect this, showing the fraction of P -values that would fall in each of many intervals, were the data generated from H_1 . (Three separate graphs are presented, because it is the smaller ranges of P -values that are typically of more interest.)

Now let us see where the series of P -values happen to fall if they arise from H_2 . There is a certain ambiguity here, because what does it mean to generate data from H_2 ? One possibility is to simply pick some value of θ , say 2, and generate data from the $N(2, 1)$ distribution. Another possibility is to randomly select θ from some distribution $\pi(\theta)$, generate data from the selected θ , and then repeat the process with new θ arising from $\pi(\theta)$. This latter possibility better mirrors the actual daily use of significance testing, and so Figure presents the results of one such simulation, with $\pi(\theta)$ chosen to be a $N(0, 2)$ distribution and the sample size (for each P -value computation) chosen to be one. (The simulation itself repeats the generation of θ and data many times.) The *shaded* columns in Figure 1 present the fraction of P -values that fell in each interval. (Note that, while

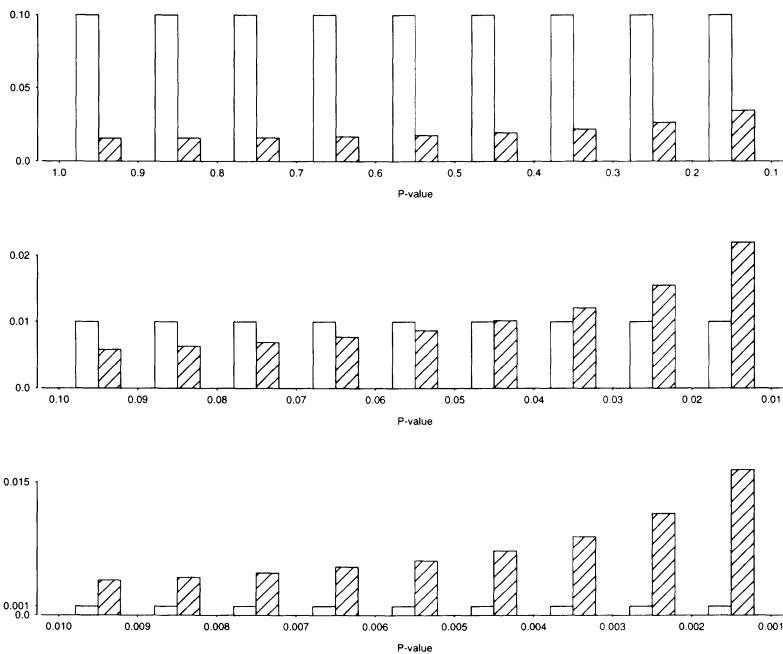


FIGURE 1. The fraction of normal observations that will fall in each given interval of p-values under $H_1 : \theta = 0$ (fraction given by unshaded column) and under $H_2 : \theta \neq 0$ (fraction given by shaded column).

the data for the shaded columns were generated from H_2 , we still just compute the P -value relative to H_1 .)

The qualitative nature of the results is no surprise: the larger the P -value, the more likely that it arose from H_2 than from H_1 . But the quantitative implications come as a considerable surprise to those who see this for the first time. For instance, suppose you observe a P -value in the interval $(0.04, 0.05)$. This is a rare event under H_1 , happening only 1% of the time, but it is nearly as rare under H_2 , happening only about 2% of the time. Thus, if one had to bet whether this P -value arose from H_1 or H_2 , it would be prudent to bet on H_2 at no more than 2 to 1 odds (assuming the two hypotheses were judged to be equally likely a priori). This last is essentially the Bayesian conclusion from the problem, that a P -value of 0.05 provides at most weak evidence in favor of H_2 .

Consider next the case in which the P -value is in the interval $(0.009, 0.01)$. In classical language this is typically termed “highly significant evidence against H_1 ,” and yet we see that the likelihood that the data came from H_2 is only 5 times the likelihood that it came from H_1 . Odds of 5 to 1

are meaningful, but hardly carry the level of conclusiveness that is usually accorded the phrase “highly significant evidence.”

A final comment before leaving this example: one might well be suspicious that the startling nature of Figure 1 is due to the particular way in which we generated data from H_2 . This is not the case. One can make *any* choices of the parameters under H_2 , and use *any* sample size (for the P -value computation), and the story remains roughly the same or worse. For instance, the fraction of P -values that will fall in the interval (0.04, 0.05), when the data is generated under H_2 , can be shown to be *at most* 0.034, and so the odds of H_2 to H_1 can be *at most* 3.4 to 1 when the P -value is in this interval. Interestingly, the fraction of P -values in the interval has no lower bound, and decreases rapidly as the sample size increases; thus if one performed this simulation with a large sample size (for computing the P -value), it would typically indicate that a P -value in the interval (0.04, 0.05) is evidence *in favor* of H_1 .

The above example illustrates what is, perhaps, the most attractive feature of Bayesian analysis, namely that its conclusions carry a clear interpretation. While there is nothing “wrong” with the classical P -value here, in that its definition is precise and it does convey information of interest, learning how to interpret a P -value as a measure of the evidence for H_1 relative to H_2 is extremely difficult, requiring major adjustments for the sample size and the type of testing that is done (among other things).

The phenomenon in the above example is very general, applying to most testing problems in which the hypotheses are of differing dimensions, including typical chi-squared testing of fit. For more extensive discussion, see Edwards, Lindeman, and Savage (1963), Berger and Sellke (1987), Berger and Delampady (1987), and Delampady and Berger (1990).

2.2.2 Notation

Hypothesis testing and model selection are essentially the same from a Bayesian perspective; we will henceforth use notation that is more designed for model selection. Also, we will only consider analysis of parametric models.

The data, X , is assumed to have arisen from one of several possible models M_1, \dots, M_m . Under M_i , the density of X is $f_i(x|\theta_i)$, where θ_i is an unknown vector of parameters of f_i .

The Bayesian approach to model selection begins by assigning prior probabilities, $P(M_i)$, to each model; often, equal prior probabilities are used, i.e. $P(M_i) = 1/m$. It is also necessary to choose prior distributions $\pi(\theta_i)$ for the unknown parameters of each model; sometimes these can also be chosen in a “default” manner, as will be illustrated later.

The analysis then proceeds by computing the posterior probabilities of each model, which elementary probability theory (Bayes theorem) shows to be equal to

$$P(M_i|\tilde{x}) = \frac{P(M_i)m_i(\tilde{x})}{\sum_{j=1}^m P(M_j)m_j(\tilde{x})}, \quad (2.1)$$

where $m_j(x) = \int f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j$. Typically one selects the model (or models) with largest posterior probability. Note that, when the prior probabilities, $P(M_i)$, are equal, then the $P(M_i|\tilde{x})$ equal

$$P_i \equiv m_i(\tilde{x}) / \sum_{j=1}^m m_j(\tilde{x}). \quad (2.2)$$

It is common to simply report the P_i in the summary of an investigation, since someone with unequal $P(M_i)$ can easily recover *their* $P(M_i|\tilde{x})$ via the alternate expression

$$P(M_i|\tilde{x}) = P(M_i) \cdot P_i / \sum_{j=1}^m P(M_j) \cdot P_j.$$

Example 1 (continued). Here, the data is $X = (X_1, \dots, X_n)$, where the X_i are independent $N(\theta, 1)$ observations. Model M_1 corresponds to assuming $\theta = 0$, in which case $f_1(\tilde{x}|0)$ is just the standard normal density. Model M_2 corresponds to assuming $\theta \neq 0$. Since M_1 has no unspecified parameters, $\pi_1(\theta_1)$ is not needed. For M_2 , we assumed in the earlier simulation that θ has a $N(0, 2)$ distribution, which would thus be $\pi_2(\theta)$. Computation then yields

$$\begin{aligned} m_1(\tilde{x}) &= f_1(\tilde{x}|0) = \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} \right), \\ m_2(\tilde{x}) &= \int_{-\infty}^{\infty} f_2(\tilde{x}|\theta)\pi_2(\theta)d\theta \\ &= \int_{-\infty}^{\infty} \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i - \theta)^2/2} \right) \cdot \frac{1}{\sqrt{4\pi}} e^{-\theta^2/4} d\theta. \end{aligned}$$

Evaluation of the above integral, and use of (2.2) yields

$$P_1 = 1 - P_2 = \left[1 + (1 + 2n)^{-1/2} \exp\{n\bar{x}^2/(2 + \frac{1}{n})\} \right]^{-1}. \quad (2.3)$$

Thus if $n = 1$ and $|\bar{x}| = 1.96$ (so that the P -value would be exactly 0.05), then $P_1 = 0.325$ and $P_2 = 0.675$. These mirror the conclusion from Figure 1 that, even though the P -value is 0.05, there is almost a (1/3) chance that $\theta = 0$ is true (assuming that we initially give M_1 and M_2 equal prior probabilities of being true).

2.2.3 Default implementation

The two difficulties in implementing Bayesian model selection are (i) choosing the prior distributions $\pi_i(\theta_i)$, and (ii) computing the $m_i(x)$. A variety of strategies exist for carrying out the integrations necessary to compute the $m_i(x)$; see Kass and Raftery (1995) for discussion. Choosing the $\pi_i(\theta_i)$ is more of a problem.

It may well be the case that subjective knowledge about the θ_i is available, and can be incorporated into subjective proper priors for the θ_i . This is clearly desirable if it can be done. Thus, in Example 1, one might feel that, if $\theta \neq 0$, then θ will be near 5 with an uncertainty of ± 1 . (Perhaps the alternative theory, M_2 , would predict this.) Choosing $\pi_2(\theta)$ to then be, say, a $N(5, 1)$ distribution would be reasonable. If one had no specific alternatives in mind, one might at least want to specify a guess, τ , as to the likely amount of departure of θ from 0 under H_2 . Interpreting τ as a prior “standard error”, it would then be reasonable to use a $N(0, \tau^2)$ prior distribution for $\pi_2(\theta)$. An alternative to guessing τ is to simply specify a plausible range for τ , and see if the desired conclusion holds for the entire range. This is called *robust Bayesian analysis*; see Jeffreys and Berger (1992) for a simple but interesting astronomical example, and Berger (1994) for a recent review.

Subjective Bayesian analysis is often avoided, for a variety of reasons. The most common argument against subjective Bayesian analysis is that “scientific discourse demands objectivity, and hence one cannot use a subjective Bayesian analysis.” The merits of this argument are highly debatable (cf, Berger and Berry, 1988), but one cannot deny that at least the appearance of objectivity can be helpful.

The other primary objection to subjective Bayesian analysis is that it is often too difficult. This is especially true of model selection problems, in which there may be several high-dimensional models and eliciting all the needed high-dimensional prior distributions would be a truly formidable undertaking. (We do not mean to imply that there is anything wrong with subjective Bayesian analysis - indeed it can be argued that one should always first try to implement such an analysis - but the difficulties are real.)

For these and other reasons, the most popular Bayesian methods tend to be default methods, which operate with default prior distributions. For estimation and prediction problems, the default Bayesian theories are well developed, and use prior distributions that are designed to be “noninformative” in some sense. The most famous of these is the *Jeffreys prior*, named after the famous geophysicist who, through his exemplary book Jeffreys (1961), was most responsible for the modern Bayesian revival. *Maximum entropy* priors are another well-known type of noninformative prior (although they often have certain features specified). The more recent statistical literature emphasizes what are called *reference priors*, which prove

remarkably successful even in higher dimensional problems (cf., Berger and Bernardo, 1992, and Yang and Berger, 1995).

Testing and model selection has proved much more resistant to the development of default Bayesian methods. This is because the “noninformative” priors discussed above are typically improper distributions, meaning they do not have mass equal to one. This does not typically pose a problem in estimation and prediction, but it does for testing and model selection. The expressions in (2.1) and (2.2) really make sense only if the $\pi_i(\theta_i)$ (and hence the $m_i(x_i)$) are proper distributions.

Jeffreys faced this problem squarely, and considered the various (unappealing) choices. One choice is to use classical measures, such as the P -value; but their very misleading nature made that choice especially unattractive. Another choice is to demand subjective Bayesian analysis, but Jeffreys felt that this was too restrictive a requirement. The third possible choice is to simply pick some proper prior distributions that seem reasonable (for the given models), and *conventionally* use them for default hypothesis testing or model selection. This was the choice that Jeffreys adopted. Thus, in Example 1, he gave extensive arguments supporting use of a standard Cauchy distribution as the default prior. (The choice we made, of a $N(0, 2)$ distribution, gives almost the same answers and is easier to handle computationally.)

In principle, this approach of Jeffreys is our favorite approach. Its major disadvantage is that there is no well-developed theory for determining default priors for hypothesis testing and model selection. Hence progress in this direction has been very sporadic, with only certain special models being treated on a case-by-case basis. Because of this limitation, Bayesian model selection has, instead, typically been performed using an asymptotic approach which is known as the BIC (Bayesian Information Criterion, see Kass and Raftery, 1995, for discussion). This approach, however, has the usual disadvantages of asymptotics, including the need for regular models and large sample sizes (though see Kass and Wasserman, 1995).

Recently, two very general default Bayesian methods have appeared: the *fractional Bayes factor* approach of O’Hagan (1995), and the *intrinsic Bayes factor* approach of Berger and Pericchi (1996a). This last approach appears to be applicable to virtually any hypothesis testing and model selection problem (even for nonregular models), and seems to closely correspond to the recommended analyses of Jeffreys for the cases he considered. Hence we feel that it holds great promise for solving the hypothesis testing and model selection problem.

There are several versions of intrinsic Bayes factors, with different versions argued to be particularly useful in certain settings. There is, however, one quite simple version that seems to work well across all problems, and that is what we will describe here.

The idea behind intrinsic Bayes factors is simple, and is based on an ad hoc method that has long been used to address the model selection

problem. The ad hoc method is to use part of the data (typically as small a part as possible) as “training data”, to convert the standard noninformative priors (used in estimation problems) to proper distributions. Then one uses these proper distributions, with the remainder of the data, to compute the posterior probabilities in (2.1) or (2.2).

Besides being an ad hoc approach with no clear justification, the method was rather arbitrary in that the choice of the training data is typically arbitrary. The simple idea in Berger and Pericchi (1996a) was to eliminate this arbitrariness by doing the computation for all possible choices of the training data (or some reasonably large random subset of such choices), and then picking an “average” of the ensuing answers. As a general purpose method, picking the median of the ensuing answers seems to work extremely well, and defines what we call the *median intrinsic Bayes factor*. (This is strictly defined only for pairwise comparisons of models, but pairwise answers can easily be adapted to deal with multiple models or hypotheses.)

Example 1 (continued). The common noninformative prior for θ , under M_2 , is $\pi(\theta) = 1$. This is improper, but if we use just one of the observations, say x_1 , as a training sample, then we can convert $\pi(\theta)$ to the *posterior distribution*

$$\pi(\theta|x_1) = \frac{f_2(x_1|\theta) \cdot \pi(\theta)}{\int f_2(x_1|\theta) \cdot \pi(\theta) d\theta} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta-x_1)^2}.$$

Using this as the (now proper) distribution for θ , and computing (2.2) for the remaining data x_2, \dots, x_n , yields

$$P_1 = 1 - P_2 = \left[1 + \sqrt{n} e^{-n\bar{x}^2/2} \cdot e^{x_1^2/2} \right]^{-1}.$$

(Here, \bar{x} is the average of all the data.) Since choice of x_1 as the training sample was arbitrary, one can do the computation for all possible choices, and then take the median of the resulting answers. The result is clearly

$$P_1^* = 1 - P_2^* = \left[1 + \sqrt{n} e^{-n\bar{x}^2/2} \cdot e^{x^{*2}/2} \right]^{-1}, \quad (2.4)$$

where x^{*2} is defined to be the median of x_1^2, \dots, x_n^2 .

While simple to implement (modulo possible computational difficulties), the intrinsic Bayes factor approach at first appeared to be just another ad hoc approach. Interest grew enormously, however, when it was shown that the answers resulting from this approach correspond very closely to answers from actual proper default prior analyses, of the type recommended by Jeffreys. One could now obtain reasonable answers without going through the involved arguments needed for the Jeffreys-type implementation.

Detailed discussion of this approach, its limitations, and advice for its computational implementation, can be found in Berger and Pericchi (1996a,

1996b). Note, however, that the median intrinsic Bayes factor is not emphasized therein; we have only recently come to appreciate it, for its general applicability and stability; except for extremely small sample sizes, we have not found any serious contraindications to its use.

2.3 An application to time series analysis

A situation in which one typically is considering a multitude of models is in time series analysis. Consider the time series in Figure 2, for instance. This

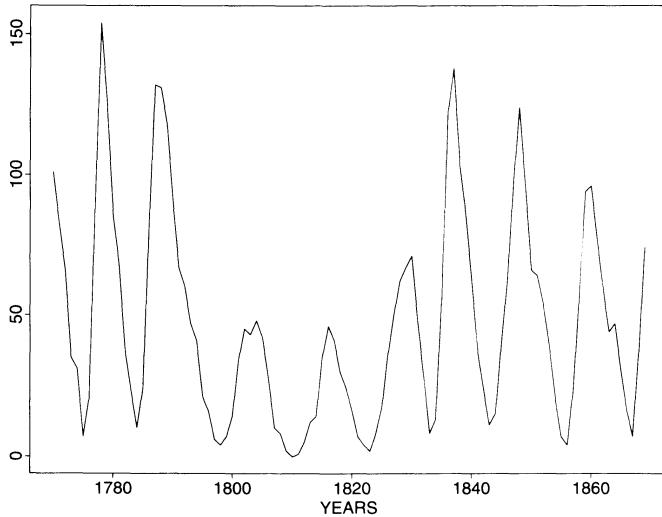


FIGURE 2. Wolf sunspot numbers.

is the famous Wolf sunspot series data, consisting of the number of sunspots observed each year from 1770 to 1869. A possibly reasonable model for the series is a stationary autoregressive process with drift. For instance, the AR(1) model with a linear drift would be described as

$$Y_t = \beta_1 t + \phi_1(Y_{t-1} - \beta_1(t-1)) + \epsilon_t. \quad (2.5)$$

where Y_t is the observation at time t , β_1 is the unknown linear coefficient, ϕ_1 is the unknown autocorrelation, and the ϵ_t are i.i.d. $\mathcal{N}(0, \sigma^2)$ errors, σ^2 also unknown.

It is decided to consider autoregressive models of order 1, 2, 3 and 4, and also to consider constant (C), linear (L), and quadratic (Q) drift. Thus the

TABLE 2.1. Posterior probabilities of models, assuming equal prior probabilities.

Model	$P(M_i x)$	Model	$P(M_i x)$
AR(1), C	~ 0	AR(3), C	0.740
AR(1), L	~ 0	AR(3), L	0.001
AR(1), Q	~ 0	AR(3), Q	0.001
AR(2), C	0.161	AR(4), C	0.076
AR(2), L	0.011	AR(4), L	0.006
AR(2), Q	0.001	AR(4), Q	0.001

AR(j) model with drift of polynomial order k ($k = 0, 1, 2$) can be written

$$Y_t = \sum_{\ell=0}^k \beta_\ell t^\ell + \sum_{r=1}^j \phi_r (Y_{t-r} - \sum_{\ell=0}^k \beta_\ell (t-r)^\ell) + \epsilon_t. \quad (2.6)$$

We are thus considering twelve models (any of the four AR models together with any of the three polynomial drifts).

The “intrinsic Bayes factor” approach applies directly to this problem. It utilizes only standard noninformative priors for the parameters (constant priors for the β_i and ϕ_i , and $1/\sigma^2$ for the variance, σ^2). Recall that one cannot use these noninformative priors directly, but must use them through the “intrinsic Bayes factor” algorithm. There is also a computational complication: because of the stationarity assumption, $\phi = (\phi_1, \phi_2, \dots, \phi_j)$ is restricted to the “stationarity region,” and so the integration in computation of the $m_i(x)$ must be carried out over this region. Methods of doing this, as well as the relevant intrinsic Bayes factor algorithm, can be found in Varshavsky (1996). The results are summarized in Table 2.1.

The AR(3) model with no drift is clearly the preferred model, although the AR(2) model with no drift receives some support, and should not be ruled out on the basis of this data. Higher or lower autoregressive structures receive little support.

Note that the analysis very strongly discourages including any drift term in the model; none of the models with a drift term has posterior probability greater than 0.011. Indeed, there was no scientific reason to include drift terms in the model, but we did so as a pedagogical illustration. One of the highly attractive features of the Bayesian approach to model selection is that it acts as a natural “Ockham’s razor”, in the sense of favoring simpler models over more complex models, if the data provides roughly comparable fits for the models. Here, the models with drift will certainly fit the data slightly better than will the models without drift, but the Bayesian analysis automatically prefers the simpler non-drift models in the absence of a clearly superior fit. And this is without having to introduce any

explicit penalty, such as reduced prior probabilities for the more complex models (although this is often also recommended - cf, Jeffreys, 1961).

In classical statistics, overfitting is avoided by introducing an ad hoc penalty term (as in AIC, Akaike Information Criterion), which increases as the complexity (i.e., the number of unknown parameters) of the model increases. Not only are such corrections ad hoc, but the standard ones do not sufficiently penalize complex models. For instance, a recent bias corrected version of AIC, designed for time series (see Hurvich and Tsai, 1989), when applied to the above data selects as the top four models the (AR(4), Q), (AR(4), C), (AR(4), L), and finally (AR(3), C) models. For an interesting historical example of Ockham's razor, and general discussion and references, see Jeffreys and Berger (1992).

We have limited the discussion of this example to the model selection question. There is, of course, considerably more of interest to the problem. One might also want to estimate the parameters of the selected model (and provide associated standard errors), and/or use the selected model for optimal prediction. Any such inferences are readily available in the Bayesian approach; for instance, the noninformative priors mentioned earlier in the example could be used to compute the posterior distribution of the parameters, from which any inferences follow directly.

One item of special interest here is that the analysis would automatically incorporate the stationarity constraint, since the prior distribution is supported only on the stationarity region for the autocorrelation parameters. Incorporating this constraint in classical analysis is problematic; maximum likelihood estimates will often be on the boundary of the stationarity region, in which case the standard errors produced from likelihood asymptotics will usually be considerably too small.

Prediction also deserves special mention. It is a well-known problem that predictions based on selected models typically turn out to be much less accurate than the model would have suggested. This is especially so if the selected model is used in raw form, with estimated values of the parameters inserted. The obvious reason for this overly optimistic prediction is that one is pretending to know the model and the parameter values, when one really does not. To obtain reasonable estimates of predictive accuracy, one must not only incorporate the uncertainty in parameter values (as reflected by their posterior distributions), but also must incorporate the uncertainty in the model. In the above example, for instance, it would be reasonable to base the predictions on the appropriate weighted mixture of predictions from the (AR(3), C) and (AR(2), C) models. For discussion on this general issue, see Draper (1995).

2.4 An application to classification and mixture models

Figures 3 and 4 show two different sets of bivariate observations that arose in an astronomical setting. It was desired to (i) identify how many clusters are present in each set of data; (ii) determine which data set exhibits stronger clustering; (iii) provide a characterization of the clusters in each figure; and (iv) classify each observation in terms of cluster membership.

The natural Bayesian approach to this problem is to model the data as arising from a mixture distribution. We will talk as if each cluster can be identified with a separate distribution from this mixture, and that each distribution in the mixture corresponds to a separate population of observations. These interpretations are not strictly necessary, but they are useful conceptually and will often be reasonable as an explanation of the underlying process.

We will assume the observations arise from an additive mixture of bivariate normal populations; thus the overall density of an observation $\tilde{x} = (\tilde{x}_1, \tilde{x}_2)$ is

$$f(\tilde{x}) = \sum_{j=1}^m \gamma_j f_j(\tilde{x}|\tilde{u}_j, \tilde{\Sigma}_j),$$

where m is the number of populations (clusters) in the mixture; γ_j is the probability that an observation arises from population j (i.e., γ_j is the proportion of elements in cluster j); and f_j is a bivariate normal distribution with mean vector \tilde{u}_j and covariance matrix $\tilde{\Sigma}_j$. Here we will assume that m and all the γ_j , \tilde{u}_j , and $\tilde{\Sigma}_j$ are completely unknown. Frequently, one might want to make further assumptions (such as assuming that the $\tilde{\Sigma}_j$ have a specified form or are all equal), but for analysis of the data in Figures 3 and 4 we make no such assumptions.

The mixture model is a difficult model for classical statistical analysis because standard asymptotics fails, and because maximum likelihood methods are often very unstable due to the presence of multiple modes in the likelihood. For Bayesian analysis there are also difficulties, primarily because standard (improper) noninformative priors cannot be used. (Subjective Bayesian analysis of mixture models is relatively straightforward, if one is willing to make the investment in prior specification; cf, Lavine and West, 1992).

We have recently modified the intrinsic Bayes factor algorithm to also enable analysis of mixture models. The idea is to, again, use (small) parts of the data as training samples; however, since we do not know which populations gave rise to which data, this has to be done as an iterative simulation involving the classification probabilities of the data. Details of the algorithm can be found in Shui (1996). The output of the algorithm is the posterior distribution of all unknown parameters. For m , only the

TABLE 2.2. The posterior probability of $m = 1, 2$, or 3 clusters, in each of the data sets in Figures 3 and 4.

	m		
	1	2	3
Data Set 1	8.1×10^{-14}	5.7×10^{-10}	~ 1
Data Set 2	1.3×10^{-6}	0.918	0.082

values 1, 2, and 3 were considered, and the posterior probabilities of each (assuming equal prior probabilities of $1/3$) are given in Table 2.2 (for both the data set in Figure 2 and that in Figure 3).

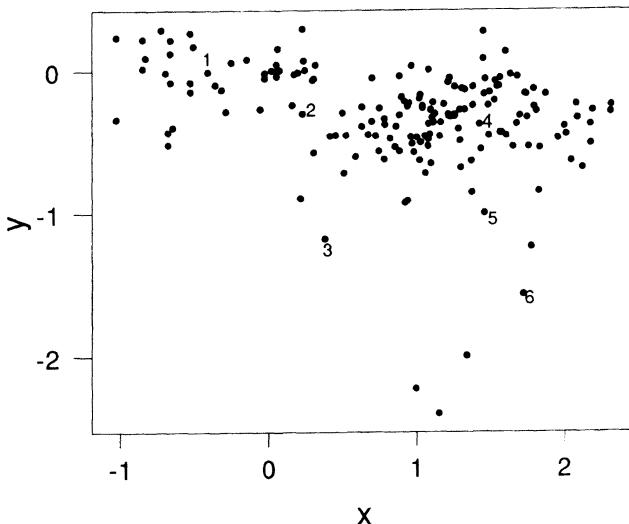


FIGURE 3. Scatter plot for data set 1

For Data Set 2, there is overwhelming evidence that at least two clusters are present, and little evidence to support more than two clusters. This last is another illustration of the automatic Ockham's razor effect of Bayesian analysis; the simpler two-cluster model is preferred, because three clusters do not provide a markedly better fit.

The situation with Data Set 1 is more interesting. The two-cluster model is much preferred over the one-cluster model (the “odds” would be the ratio of the posterior probabilities, here 703), but three clusters is the overwhelmingly preferred model. The reason is that there are a number of obvious

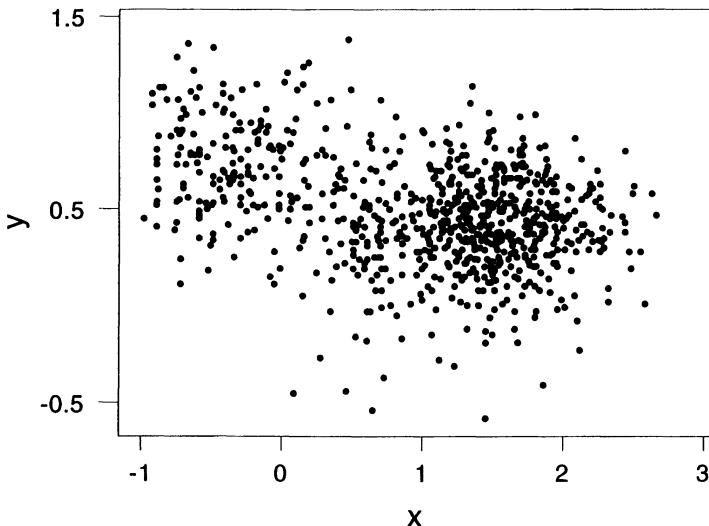


FIGURE 4. Scatter plot for data set 2

“outliers” in the lower right quadrant of Figure 2, and these outliers become their own “cluster”. This effect is typical of normal mixture models, as normal distributions do not admit outliers. In a sense, this is not bad, since outliers often deserve special identification and treatment.

Table 2.3 gives the Bayesian estimates (here, posterior medians) of all parameters in the various populations for Data Set 1. (Standard errors of these estimates are also available, but are not reported here.) The populations, or clusters, have been labelled in terms of their size, with “1” being the largest.

There are no surprises in these results. For the preferred 3-cluster model, the third component indeed appears to be the outliers, and has the very small $\hat{\gamma}_3 = 0.025$. (There were 174 observations in all, so $\hat{\gamma}_3$ intuitively reflects about 4 outliers.) Note that the covariance matrices seem to be quite different, in both the 2-component and the 3-component models.

The final output available from the algorithm is the posterior probability that each observation arises from each cluster. This is given in Table 2.4 for the labeled observations from Figure 3. (As before, the clusters are labeled according to their size, with “1” being the largest and “3” the smallest.) There are no surprises in the table, although note that there is considerable uncertainty attached to the classification of some of the observations. Ordinary classification and clustering algorithms do not typically allow for reflection of such uncertainty.

TABLE 2.3. Estimates (posterior medians) of parameters in the mixture model for Data Set 1.

Parameter	1-Cluster	Model 2-Cluster	3-Cluster
γ_1		0.774	0.749
u_1	(0.892, -0.325)	(1.236, -0.408)	(1.266, -0.363)
Σ_1	$\begin{pmatrix} 0.628 & -0.095 \\ -0.095 & 0.147 \end{pmatrix}$	$\begin{pmatrix} 0.223 & 0.010 \\ 0.010 & 0.157 \end{pmatrix}$	$\begin{pmatrix} 0.201 & 0.001 \\ 0.001 & 0.057 \end{pmatrix}$
γ_2		0.226	0.226
u_2		(-0.309, -0.075)	(-0.290, -0.051)
Σ_2		$\begin{pmatrix} 0.193 & 0.001 \\ 0.001 & 0.040 \end{pmatrix}$	$\begin{pmatrix} 0.166 & 0.000 \\ 0.000 & 0.037 \end{pmatrix}$
γ_3			0.025
u_3			(1.195, -1.729)
Σ_3			$\begin{pmatrix} 0.203 & -0.011 \\ -0.011 & 0.265 \end{pmatrix}$

TABLE 2.4. Posterior probabilities that the labeled observations from Figure 3 belong to cluster 1, 2, or 3, in the favored three-cluster model.

Observation	Belongs to Cluster		
	“1”	“2”	“3”
1	0.002	0.998	~ 0
2	0.429	0.571	~ 0
3	0.654	~ 0	0.346
4	~ 1	~ 0	~ 0
5	0.906	~ 0	0.094
6	~ 0	~ 0	~ 1

2.5 Advances in Bayesian computation

2.5.1 Introduction

Recent computational tools have allowed application of Bayesian methods to highly complex and nonstandard models. Indeed, for many complicated

models, Bayesian analysis has now become the simplest (and often only possible) method of analysis.

Although other goals are possible, most Bayesian computation is focused on calculation of posterior expectations $E^*[g(\theta)]$, where E^* represents expectation with respect to the posterior distribution and $g(\theta)$ is some function of interest. For instance, if $g(\theta) = \theta$, then $E^*[g(\theta)] = E^*[\theta] \equiv \mu$, the posterior mean; if $g(\theta) = (\theta - \mu)^2$, then $E^*[g(\theta)]$ is the posterior variance of θ ; and, if $g(\theta)$ is 1 if $\theta > C$ and 0 otherwise, then $E^*[g(\theta)]$ is the posterior probability that θ is greater than C . Another common type of Bayesian computation is calculation of the posterior mode (as in computation of MAP estimates in image processing); we do not formally discuss this here, although a number of techniques discussed below can also be useful in this regard.

2.5.2 Traditional numerical methods

The ‘traditional’ numerical methods for computing $E^*[g(\theta)]$ are numerical integration, Laplace approximation, and Monte Carlo Importance Sampling. Brief introductions to these methods can be found in Berger (1985). Here we say only a few words, to place the methods in context and provide references.

A successful general approach to **numerical integration** in Bayesian problems, using adaptive quadrature methods, was developed in Naylor and Smith (1982). This was very effective in moderate (e.g., 10) dimensional problems.

Extension of the *Laplace approximation* method of analytically approximating $E^*[g(\theta)]$, leading to a reasonably accurate general technique, was carried out in Tierney et al. (1989). The chief limitations of the method are the need for analytic derivatives, the need to redo parts of the analysis for each different $g(\theta)$, and the lack of an estimate of the error of the approximation. For many problems, however, the technique is remarkably successful.

Monte Carlo importance sampling [see Geweke (1989) and Wolpert (1991) for discussion] has been the most commonly used traditional method of computing $E^*[g(\theta)]$. The method can work in very large dimensions, and carries with it a fairly reliable accuracy measure. Although one of the oldest computational devices, it is still one of the best, being nearly ‘optimal’ in many problems. It does require determination of a good ‘importance function’, however, and this can be a difficult task. Current research continues to address the problem of choosing a good importance function; for instance, Oh and Berger (1993) developed a method of selecting an importance function for a multimodal posterior.

2.5.3 Markov chain simulation techniques

The newest techniques to be extensively utilized for numerical Bayesian computations are Markov chain simulation techniques, including the popular Gibbs Sampling. [Certain of these techniques are actually quite old — see, e.g., Hastings (1970); it is their application and adoption to Bayesian problems that is new.] A brief generic description of these methods is as follows:

- Step 1.* Select a ‘suitable’ Markov chain on the parameter space Θ , with $p(\cdot, \cdot)$ being the transition probability density (i.e., $p(\theta, \theta^*)$ gives the transition density for movement of the chain from θ to θ^*). Here ‘suitable’ means primarily that the posterior distribution of θ given the data x , $\pi(\theta|x)$, is a stationary distribution of the Markov chain, which can be assured in a number of ways.
- Step 2.* Starting at a point $\theta^{(0)} \in \Theta$, generate a sequence of points $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(m)}$ from the chain.
- Step 3.* Then, for large m , $\theta^{(m)}$ is (approximately) distributed as $\pi(\theta|x)$ and

$$\frac{1}{m} \sum_{i=1}^m g(\theta^{(i)}) \cong E^*[g(\theta)]. \quad (2.7)$$

The main strengths of Markov chain methods for computing $E^*[g(\theta)]$ are:

- (1) Many different g can simultaneously be handled via Step 3, once the sequence $\theta^{(1)}, \dots, \theta^{(m)}$ has been generated.
- (2) Programming tends to be comparatively simple.
- (3) Methods of assessing convergence and accuracy exist and/or are being developed.

The main weaknesses of the Markov chain methods are:

- (1) They can be quite slow. It is not uncommon in complicated problems to need m to be in the hundreds of thousands, requiring millions of random variable generations if the dimension of θ is appreciable.
- (2) One can be misled into prematurely judging that convergence has obtained.

The more common Markov chain methods, corresponding to different choices of $p(\cdot, \cdot)$, will briefly be discussed. A recent general guide to these methods, and their use in practice, is Gelman, et. al. (1995). See also Smith (1991) and Besag, et. al. (1995).

Metropolis-Hastings algorithm: One generates a new θ^* based on a ‘probing’ distribution, and then moves to the new θ^* or stays at the old θ according to a certain ‘accept-reject’ probabilities, see Hastings (1970).

Gibbs sampling: The Markov chain moves from $\theta^{(i)}$ to $\theta^{(i+1)}$ one coordinate at a time (or one group of coordinates at a time), the transition density being the conditional posterior density of the coordinate(s) being moved given the other coordinates. This is a particularly attractive procedure in many Bayesian scenarios, such as analysis of hierarchical models, because the conditional posterior density of one parameter given the others is often relatively simple (or can be made so with the introduction of auxiliary variables). Extensive discussion and illustration of Gibbs sampling can be found in Gelfand and Smith (1990), Gelman and Rubin (1992), Raftery (1992), and Smith and Gelfand (1992). We confine ourselves here to a very elementary example, but one which illustrates the basic technique.

Example 3. The following posterior density is a very simplified version of posterior densities which occur commonly in Bayesian analysis, and which are particularly amenable to Gibbs sampling. Suppose the posterior density is

$$\pi(\theta_1, \theta_2 | \text{data}) = \frac{1}{\pi} \exp\{-\theta_1(1 + \theta_2^2)\} \quad (2.8)$$

on the domain $\theta_1 > 0$, $-\infty < \theta_2 < \infty$. Many posterior expectations involving this density cannot be done in closed form. Gibbs sampling, however, can easily be applied to this distribution to compute all integrals of interest.

Note, first, that the conditional distribution of θ_2 , given θ_1 , is Normal with mean zero and variance $1/2\theta_1$; and, given θ_2 , θ_1 has an exponential distribution with mean $1/(1 + \theta_2^2)$. Hence the Gibbs sampling algorithm can be given as follows:

- Step 0.* Choose an initial value for θ_2 ; for instance, the maximizer of the posterior, $\theta_2^{(0)} = 0$.
- Step i(a).* Generate $\theta_1^{(i)} = \mathcal{E}/(1 + [\theta_2^{(i-1)}]^2)^{1/2}$, where \mathcal{E} is a standard exponential random variable.
- Step i(b).* Generate $\theta_2^{(i)} = Z/\sqrt{2\theta_1^{(i)}}$, where Z is a standard normal random variable.
- Repeat* Steps i(a) and i(b) for $i = 1, 2, \dots, m$.
- Final Step.* Approximate the posterior expectation of $g(\theta_1, \theta_2)$ by

$$\begin{aligned} E[g(\theta_1, \theta_2)] &= \int_{-\infty}^{\infty} \int_1^{\infty} g(\theta_1, \theta_2) \pi(\theta_1, \theta_2 | \text{data}) d\theta_1 d\theta_2 \\ &\cong \frac{1}{m} \sum_{i=1}^m g(\theta_1^{(i)}, \theta_2^{(i)}). \end{aligned} \quad (2.9)$$

For instance, the typical estimate of θ_1 would be its posterior mean, approximated by $\hat{\theta}_1 = (1/m) \sum_{i=1}^m \theta_1^{(i)}$. Table 2.5 presents the results of this computation for various values of m . Note that the true posterior mean here is 0.5.

TABLE 2.5. Approximate values of posterior mean of θ_1 from Gibbs Sampling.

m	100	500	1,000	10,000	50,000
$\hat{\theta}_1$	0.43761	0.53243	0.48690	0.49857	0.50002

Hit and run sampling: The idea here is roughly that one moves from $\theta^{(i)}$ to $\theta^{(i+1)}$ by choosing a random direction and then moving in that direction according to the appropriate conditional posterior distribution. This method is particularly useful when Θ is a sharply constrained parameter space. Extensive discussion and illustration can be found in Belisle et al. (1993), and Chen and Schmeiser (1993).

Hybrid methods: Complex problems will typically require a mixture of the above (and other) methods. Here is an example, from Müller (1991), the purpose of which is to do Gibbs sampling when the posterior conditionals [e.g., $\pi(\theta_i|x, \text{other } \theta_k)$] are not ‘nice’.

Step 1. Each step of the Markov chain will either

- generate $\theta_j^{(i)}$ from $\pi(\theta_j|x, \text{other } \theta_k^{(i)})$ if the conditional posterior is ‘nice’ or
- generate $\theta_j^{(i)}$ by employing one or several steps of the Metropolis-Hastings algorithm if the conditional is not nice.

Step 2. For the probing function in the Metropolis-Hastings algorithm, use the relevant conditional distribution from a global multivariate normal (or t) importance function, as typically developed in Monte Carlo importance sampling.

Step 3. Adaptively update the importance function periodically, using estimated posterior means and covariance matrices.

Other discussions or instances of use of hybrid methods include Geyer (1992, 1995), Gilks and Wild (1992), Tanner (1991), Smith and Roberts (1993), Berger and Chen (1993), and Tierney (1994).

2.5.4 Software existence and development

Availability of general user-friendly Bayesian software would rapidly advance use of Bayesian methods. A number of software packages do exist, and are very useful for particular scenarios. An example is BATS (cf., Pole, West, and Harrison, 1994, and West and Harrison, 1989), which is designed for Bayesian time series analysis. A listing and description of pre-1990 Bayesian software can be found in Goel (1988) and Press (1989).

Four recent software developments are BAIES, a Bayesian expert system (see Cowell, 1992); [B/D], an ‘expectation based’ subjective Bayesian system (see Wooff 1992); BUGS, designed to analyze general hierarchical models via Gibbs sampling (see Thomas et. al. 1992); and XLISP-STAT,

a general system with excellent interactive and graphics facilities, but limited computational power (see Tierney 1990). Numerous other Bayesian software developments are currently underway.

Two of the major strengths of the Bayesian approach create certain difficulties in developing generic software. One is the extreme flexibility of Bayesian analysis, with virtually any constructed model being amenable to analysis. Classical packages need contend with only a few well-defined models or scenarios for which a classical procedure has been determined. Another strength of Bayesian analysis is the possibility of extensive utilization of subjective prior information, and Bayesians tend to feel that software should include an elaborate expert system for prior elicitation. This is hard, in part because much remains to be done empirically to determine optimal ways to elicit priors. Note that such an expert system is not, by any means, a strict need for Bayesian software; it is possible to base a system on use of noninformative priors.

2.6 Conclusions

Papers on Bayesian analysis frequently tout the advantages of Bayesian over classical methods, and this paper has been no exception. In a sense, this is unavoidable since, for a scientist to try Bayesian methods, a considerable retooling and investment of effort may be required, and the case must be made that this effort is worthwhile. At the same time, criticism of classical statistics is rather unfortunate, because Bayesian statistics and classical statistics share a great deal in common, and have much the same aims. Indeed, the two schools of statistics have been drawing closer together of late, so much so that one can envisage at least a philosophical unification sometime in the near future.

As an example of this, consider the situation of Bayesian testing, as illustrated by Example 1. We were quite critical of the use of P -values in that example, but P -values are also criticized by classical frequentist statisticians, in part because they are not true frequentist procedures having an interpretation in terms of a long-run error rate. In a recent surprising development (based on ideas of Kiefer, 1977), Berger, Brown, and Wolpert (1995) and Berger, Boukai, and Wang (1994) show for simple versus simple testing and for testing a precise hypothesis, respectively, that Bayesian tests (with, say, equal prior probabilities of the hypotheses) yield posterior probabilities which have direct interpretations as conditional frequentist error probabilities. In Example 1, for instance, P_1 in (2.3) can be interpreted as the conditional Type I frequentist error probability, and P_2 can be interpreted as an average conditional Type II error probability. Note that the reported error probabilities thus vary with the data, in contrast with the usual Neyman-Pearson error probabilities. Also, use of these con-

ditional error probabilities is arguably greatly superior to use of the usual Neyman-Pearson error probabilities, even from the frequentist perspective.

The necessary technical detail to make this work is the defining of suitable conditioning sets upon which to compute the conditional error probabilities. These sets necessarily include data in both the acceptance and the rejection regions, and can roughly be described as the sets which include data points providing equivalent strength of evidence for and against H_1 . Note that computation of these sets is not necessary for practical implementation of the procedures.

The primary limitation of this Bayesian - frequentist equivalence is that there will typically be a region, which is called the no-decision region, in which frequentist and Bayesian interpretations are incompatible. Hence this region is excluded from the decision space. In Example 1, for instance, and if $n = 20$, then the no-decision region is the set of all points where the usual z -statistic ($\sqrt{n}|\bar{x}|$) is between 1.18 and 1.95. In all examples we have studied, the no-decision region is similarly a region where both frequentists and Bayesian would feel indecisive, and hence its presence in the procedure is not detrimental from a practical perspective.

While a philosophical reconciliation of the statistical schools appears to be within the realm of possibility, the ease of interpretation of Bayesian answers and the comparative simplicity in implementing the (default) Bayesian techniques will still argue in favor of their use.

Acknowledgments: This research was supported by the National Science Foundation, under Grant DMS - 9303556. Chimei Shui and Dejun Tang were instrumental in carrying out the computations.

REFERENCES

- [1] Bayes, T. (1783). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.*, **53**, 370-418.
- [2] Belisle, C., Romeijn, H. E. and Smith, R. (1993). Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operation Research*, **18**, 255-266.
- [3] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd edition). Springer-Verlag, NY.
- [4] Berger, J. (1994). An overview of robust Bayesian analysis. *Test*, **3**, 5-124.
- [5] Berger, J. and Bernardo, J. (1992). On the development of the reference prior method. In J. Bernardo, J. Berger, A. Dawid and A. F. M. Smith (editors), *Bayesian Statistics*, **4**, Oxford University Press, London.
- [6] Berger, J. and Berry, D. (1988). Analyzing data: Is objectivity possible? *American Scientist*, **76**, 159-165.
- [7] Berger, J., Brown, L. and Wolpert, R. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential hypothesis testing. *Annals of Statistics*, **22**, 1787-1807.

- [8] Berger, J., Boukai, B., and Wang, Y. (1994). Unified frequentist and Bayesian testing of a precise hypothesis. Technical Report 94-25C, Purdue University, West Lafayette.
- [9] Berger, J. and Chen, M. H. (1993). Determining retirement patterns: prediction for a multinomial distribution with constrained parameter space. *The Statistician*, **42**, 427–443.
- [10] Berger, J. and Delampady, M. (1987). Testing precise hypotheses (with discussion). *Statist. Science*, **2**, 317–352.
- [11] Berger, J., and Pericchi, L. R. (1996a). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109–122.
- [12] Berger, J., and Pericchi, L. R. (1996b). The intrinsic Bayes factor for linear models. *Bayesian Statistics*, **5**. J. M. Bernardo, et. al. (eds.), pp. 23–42, Oxford University Press, London.
- [13] Berger, J. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P values and evidence. *J. Amer. Statist. Assoc.*, **82**, 112–122.
- [14] Besag, J., Green, P., Higdon, D., and Mengerson, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, **10**, 1–58.
- [15] Chen, M. H. and Schmeiser, B. (1993). Performance of the Gibbs, hit-and-run, and Metropolis samplers. *Journal of Computational and Graphical Statistics*, **2**, 1–22.
- [16] Delampady, M. and Berger, J. (1990). Lower bounds on posterior probabilities for multinomial and chi-squared tests. *Annals of Statistics*, **18**, 1295–1316.
- [17] Draper, D. (1995). Assessment and propagation of model uncertainty. *J. Roy. Statist. Soc. B*, **57**, 45–98.
- [18] Cowell, R. G. (1992). BAIES: A probabilistic expert system shell with qualitative and quantitative learning. In: *Bayesian Statistics*, **4** (J. Bernardo, J. Berger, A. Dawid and A. F. M. Smith, Eds.). Oxford University Press, Oxford.
- [19] Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242.
- [20] Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398–409.
- [21] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.
- [22] Gelman, A. and Rubin, D. B. (1992). On the routine use of Markov Chains for simulation. In J. Bernardo, J. Berger, A. Dawid, and A. F. M. Smith (editors), *Bayesian Statistics*, **4**, Oxford University Press, London.
- [23] Geweke, J. (1989). Bayesian inference in econometrics models using Monte Carlo integration. *Econometrica*, **57**, 1317–1340.
- [24] Geyer, C. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, **7**, 473–483.
- [25] Geyer, C. (1995). Conditioning in Markov Chain Monte Carlo. *J. Comput. Graph. Statist.*, **4**, 148–154.
- [26] Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. In J. Bernardo, J. Berger, A. Dawid, and A. F. M. Smith (editors), *Bayesian Statistics*, **4**, Oxford University Press, London.

- [27] Goel, P. (1988). Software for Bayesian analysis: current status and additional needs. In: *Bayesian Statistics*, **3**, J. M. Bernardo, M. DeGroot, D. Lindley and A. Smith, (Eds.). Oxford University Press, Oxford.
- [28] Hastings, W. K. (1970). Monte-Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.
- [29] Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.
- [30] Jeffreys, H. (1961). *Theory of Probability* (3rd edition), Oxford University Press, London.
- [31] Jeffreys, W. and Berger, J. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, **80**, 64–72.
- [32] Kass, R. and Raftery, A. (1995). Bayes factors and model uncertainty. *J. Amer. Statist. Assoc.*, **90**, 773–795.
- [33] Kass, R. E., and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**, 928–934.
- [34] Kiefer, J. (1977). Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association*, **72**, 789–827.
- [35] Laplace, P. S. (1812). *Theorie Analytique des Probabilités*. Courcier, Paris.
- [36] Lavine, M. and West, M. (1992). A Bayesian method for classification and discrimination. *Canadian J. of Statistics*, **20**, 421–461.
- [37] Loredo, T. (1992). Promise of Bayesian inference for astrophysics. In: *Statistical Challenges in Modern Astronomy*, E. Feigelson and G. J. Babu (Eds.). Springer-Verlag, New York.
- [38] Naylor, J. and Smith, A. F. M. (1982). Application of a method for the efficient computation of posterior distributions. *Appl. Statist.*, **31**, 214–225.
- [39] O'Hagan, A. (1995). Fractional Bayes factors for model comparisons. *J. Roy. Statist. Soc. B*, **57**, 99–138.
- [40] Oh, M. S. and Berger, J. (1993). Integration of multimodal functions by Monte Carlo importance sampling. *J. Amer. Statist. Assoc.*, **88**, 450–456.
- [41] Raftery, A. (1992). How many iterations in the Gibbs sampler? In J. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith (editors), *Bayesian Statistics 4*, Oxford University Press.
- [42] Ripley, B. D. (1992). Bayesian methods of deconvolution and shape classification. In: *Statistical Challenges in Modern Astronomy*, E. Feigelson and G. J. Babu (Eds.). Springer-Verlag, New York.
- [43] Shui, C. (1996). Default Bayesian Analysis of Mixture Models. Ph.D. Thesis, Purdue University.
- [44] Smith, A. (1991). Bayesian computational methods. *Phil. Trans. Roy. Soc.*, **337**, 369–386.
- [45] Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling-resampling perspective. *American Statistician*, **46**, 84–88.
- [46] Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. B*, **55**, 3–23.
- [47] Tanner, M. A. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*, Lecture Notes in Statistics **67**, Springer Verlag, New York.

- [48] Thomas, A., Spiegelhalter, D. J. and Gilks, W. (1992). BUGS: A program to perform Bayesian inference using Gibbs sampling. In: *Bayesian Statistics*, 4 (J. Bernardo, J. Berger, A. Dawid and A. F. M. Smith, Eds.). Oxford University Press, Oxford.
- [49] Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.*, **22**, 1701–1762.
- [50] Tierney, L. (1990). *Lisp-Stat, an Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. Wiley, New York.
- [51] Tierney, L., Kass, R. and Kadane, J. (1989). Fully exponential Laplace approximations to expectations and variances of non-positive functions. *J. Amer. Statist. Assoc.*, **84**, 710–716.
- [52] Varshavsky, J. (1996). Intrinsic Bayes factors for model selection with autoregressive data. To appear in J. Bernardo et. al. (editors), *Bayesian Statistics*, 5, Oxford University Press, London.
- [53] Wolpert, R. L. (1991). Monte Carlo importance sampling in Bayesian statistics. In: *Statistical Multiple Integration* (N. Flournoy and R. Tsutakawa, Eds.). *Contemporary Mathematics*, Vol. 115.
- [54] Wooff, D. A. (1992). [B/D] works. In: *Bayesian Statistics*, 4 (J. Bernardo, J. Berger, A. Dawid and A. F. M. Smith, Eds.). Oxford University Press, Oxford.
- [55] Yang, R. and Berger, J. (1996). A catalogue of noninformative priors. Technical Report, Purdue University.

Discussion by Alanna Connors

Putting the Newest of Bayes into Context for Astronomers

Goals/Context

Introduction

Why might recent developments in Bayesian analysis, or even standard Bayesian procedures, be of interest to astronomers and physicists? J. Berger, in [BE96] presents some examples, from the point of view of a statistician. In this paper, I try to translate these concepts to a point of view more familiar to astronomers and physicists. [BE96] focuses on hypothesis testing and model selection. I try to start more slowly. I first highlight terms that may be unfamiliar; and then very briefly sketch out standard Bayes parameter estimation and likelihood ratios for two examples from γ -ray astrophysics. With these in mind, one can see where [BE96] presents classic examples of Bayesian hypothesis testing; plus both some intriguing new ideas on the difficult area of priors; and new developments in computer techniques. I hope this might also briefly give statisticians some of the flavor of trying to eke out inferences about physical conditions of objects in the distant

sky; and where Bayesian methods might be more practical. I close with a few personal thoughts on moving towards the use of likelihood ratios.

What is it?

Bayesian inference is a clear procedure for building measurement tools (probabilities and their ratios) for: 1) parameter estimation; 2) model selection and hypothesis testing; 3) robustness and sensitivity of results to model choice, and prior information; and 4) prediction. Many astrophysicists are more familiar with *sampling statistics*: the probability of the data \mathcal{X} , given a model or hypothesis \mathcal{M} and parameters Θ , $p(\mathcal{X}|\Theta\mathcal{M}I)$ (or $p(\mathcal{X}|\mathcal{M}I)$). With Bayesian inference one works with the inverse: the probability of a model or hypothesis \mathcal{M} and parameters Θ given the data, $p(\Theta|\mathcal{M}\mathcal{X})$ (or $p(\mathcal{M}|\mathcal{X})$). One gets from one (*data-space, on the right*) to the other (*parameter- or hypothesis-space, on the left*) via *Bayes's Theorem*:

$$p(\Theta|\mathcal{X}I) = \frac{p(\Theta|I)}{p(\mathcal{X}|I)} p(\mathcal{X}|\Theta I), \text{ or } p(\mathcal{M}|\mathcal{X}I) = \frac{p(\mathcal{M}|I)}{p(\mathcal{X}|I)} p(\mathcal{X}|\mathcal{M}I).$$

Here “ I ” represents prior measurements and information; $p(\Theta|I)$ (or $p(\mathcal{M}|I)$) is called the *prior probability*; $p(\mathcal{X}|\Theta I)$ the *direct probability* or *sampling statistic*; $p(\Theta|\mathcal{X}I)$ (or $p(\mathcal{M}|\mathcal{X}I)$) is the *posterior probability*; and $p(\mathcal{X}|I)$ serves as a normalization term.

The references cited by [BE96] give fine overviews and bibliographies. I would like to highlight two: [JA78] contains a classic historical account from the perspective of a physicist. Perhaps the earliest modern use of Bayesian inference in astronomy is [BI71].

How is it different from what I'm used to doing?

Sampling statistics is based on the long-term (asymptotic) frequency of occurrence of a particular pattern of data, assuming the model is true. Many astronomers use the recipes for likelihood ratios in [LM79],[CA78] to generate confidence intervals, which are based on the Central Limit Theorem asymptotically holding. (Also, some astrophysicists might be more comfortable with the applied math term “inverse problems” [CB86]. Or, they may not have realized that “forward-fitting”, using χ^2 , is a maximum-likelihood method that assumes a Gauss–Normal form for the sampling statistic.) By contrast, Bayesian inference calculates the probability of the parameters (or model) given any prior information, plus just the data one has.

The concept of *priors*, of assigning probability distributions to parameters before making inferences from the data, may also be new to astrophysicists. [BE96] lists many standard options then spends some time discussing new “one-size-fits-all” priors: usually they are “custom-built”. I want to highlight two distinctions: *informative* versus *uninformative* priors; and *proper* versus *improper* priors. When one has significant prior

information (such as a previous background measurement, or knowledge of atomic line strengths), one can use an *informative prior*. Without such knowledge, one uses an *uninformative prior*. In the latter case, a physicist or astronomer can often constrain the form of the prior from knowledge of the geometry of the physical system, or physics theory, or invariance arguments (see also [JA78]). A *proper prior* is one that is normalized to one; while an *improper prior* is a handy analytic form (such as a constant or log distribution) that, when integrated over all parameter space, tends to ∞ and so is not normalizable. [BE96] notes that the latter can work well for parameter estimation, but has drawbacks for model comparison and hypothesis testing. This drives his “intrinsic Bayes factor” approach.

When working in parameter-space one can integrate over uninteresting (or “nuisance”) parameters; or indeed over all parameters. This is called *marginalization*; another potentially unfamiliar term. Note that (by marginalizing over all parameter space) one can directly calculate and compare the global probabilities of two hypotheses with differing numbers of parameters. There is no need to add an extra factor for each degree of freedom (e.g. in sampling statistics one might require the difference in χ^2 , equivalent to $-2 \log[p(\mathcal{X}|\Theta I)]$, to be more than 1). As [BE96] illustrates, integrating over each extra dimension intrinsically takes this into account.

Benefits / Objections

Benefits

It gives a clear mechanism to build a tool to get the best measure of distance between two clearly stated hypotheses. It is always a *sufficient statistic*; that is, it incorporates all the information about the hypotheses that is available in the data; and it includes a mechanism to optimally incorporate prior information. For example, [BH93] suggests an appealing but “ad hoc” statistic for incorporating imaging information when searching for periodic γ -ray emission from a known radio pulsar. Each γ -ray photon is weighted by its angular distance from the source according to a telescope point-spread function, before the data are binned at the pulsar period into a phase histogram, and a χ^2 test for a flat light-curve is performed. This seems intuitive, but how does one know whether it incorporates all of the information available in the data, and in one’s prior information?

One can tackle any problem where the hypotheses are clearly stated. For example, many image processing applications have very large numbers of parameters, comparable to the number of data points. This can be a numerically intractable “inverse problem”, until one notices that with Bayesian methods one has a prior that can act as a regularizer.

It is valid for moderate and small data sets (no asymptotics required). The familiar recipes used by astronomers to generate confidence intervals are based on the Central Limit Theorem [CA78], [LM79]. Often this does not strictly hold. For example there may be multiple peaks in the probability space. Or, the sample size may be very small and the measurement not repeatable: [LO92] points out there was only one chance to measure neutrinos from SN 1987A; there were roughly two dozen neutrinos, and approximately 8 parameters.

One can reduce dimension of problems by integrating over uninteresting parameters. A common example: an interesting source energy spectrum might have $\sim 10^2$ energy bins, low Poisson counts per channel, plus measurements of the $\sim 10^2$ background rates in each bin. One does not subtract the background rate from each energy channel in the source spectrum, but instead *marginalizes* over the imperfectly known background rates [LO92]. It also clears up what to do with the “number of trials” question: one integrates over a range of trial parameters.

One can compare the likelihoods of non-nested models with different numbers of parameters. [BE96]. Also, by definition, *one can handle uncertainties in the model or in prior information.* Examples include uncertainties in stellar coronal models; or in energy response matrices.

Objections

Learning the language, retooling. “It’s not in Bevington.” [BR92] No, it’s not; but neither are most of the techniques discussed in these proceedings. Becoming familiar with the language of priors, posteriors, marginalizations, and credible regions requires a significant effort.

Getting practical, reliable priors. This is an active area of research, as [BE96] makes clear. One approach is to report one’s results in a form where the effect of using different priors is easy to calculate.

Computation time. “Rev. Thomas Bayes started his calculation in 1783, and they’re just now finishing.” – D. J. Forrest on the recent rise in interest in Bayesian methods. Although marginalization is a Bayesian technique of great power, it requires integrating over parameter space. Numerical integration in high dimensions is one of the classic high-CPU problems. [BE96] touches on some new techniques. However, when the integration can be done analytically, marginalizing can actually speed up a calculation [LO92].

No general “goodness of fit” like χ^2 . “That’s an objection?” – standard Bayesian response. Standard significance tests use the tail of the distribution. [BE96] works through an example showing this is often not a very good discriminator between two hypotheses. Instead a Bayesian analysis

specifically calculates the probability or likelihood of two (or more) hypotheses.

Simple example: Astrophysicists have it easier than statisticians

Specifying the problem

Periodic Time Series Analysis. Suppose one is searching for γ -ray emission from a known pulsar, with position, period, and all period derivatives known from radio data. Given a set of γ -ray data, what is the likelihood that a periodic signal has been detected? This is a quick sketch. For more details, [GL92] carefully treat a problem that is similar but has a different shape function.

Data. The data are in the form of time-tagged events (point Poisson process): a list of photon arrival times with a 3° window around the source position, and standard data quality cuts on the other parameters [MU95]. The two sets I show here are 1–3 MeV and 10–30 COMPTEL data on the well-known 33 ms Crab pulsar. It is a 14 day observation. There are 54626 photons in this 1–3 MeV dataset (about 1 every 20 seconds); and 1981 in the 10–30 MeV data (about 1 every 10 minutes). There is known to be a significant ($> 80\%$ of the events) background component. For this example we look for the total pulsed fraction of the source + background rate.

Null hypothesis, \mathcal{M}_0 . The photon arrival times are completely random, and can be described by a Poisson process with a constant rate $\mu_0(t) = B$.

Interesting hypothesis, \mathcal{M}_1 . The photon arrival times are periodic, with a shape described by $\rho(t)$, with $\langle \rho(t) \rangle \equiv 1$ when averaged over one cycle; and total normalization described by B : $\mu_1(t) = B\rho(t)$.

Shape function for interesting hypothesis. Since this is a Poisson process, it is convenient to describe the periodic shape by an exponentiated Fourier series, or generalized von Mises distribution. For one component, $\rho(t) \propto \exp[-\kappa \cos(\Theta(t) + \phi)]$, with $\Theta(t)$ the pulsar phase from radio data; and ϕ the unknown phase difference between the radio and gamma-ray energies. The parameter κ is known as the shape or concentration parameter, with pulsed fraction $f = \tanh(\kappa)$. The normalization condition $\langle \rho(t) \rangle = 1$ requires $\rho(t) = \exp[-\kappa \cos(\Theta(t) + \phi)] / I_0(\kappa)$, where I_0 is the modified Bessel function of order zero.

Assigning probabilities

Priors. Knowing the physical meaning of the pulsed fraction $f \in [0, 1]$ and relative phase $\phi \in [0, 2\pi]$ allows one to assign unambiguous properly

normalized prior probabilities, even when one has no previous measurements. From symmetry, one argues that the prior for the phase ϕ should be $p(\phi|I)d\phi = d\phi/(2\pi)$. Likewise, the prior on the pulsed fraction can be given by $p(f|I)df = df$. The prior on $B \in [0, B_0]$ is the only ambiguous assignment. Should it be a uniform prior, $p(B|I)dB = dB/B_0$? A log-uniform prior, $p(B|I)dB = dB/(B \log(B_0))$? However, whatever the choice, all dependence on B will be exactly the same for the null and interesting hypotheses, and so will cancel when a likelihood ratio is taken. For this example, I chose the former, and let $B_0 \rightarrow \infty$ at the end of the calculation.

Direct probability. For both null and interesting hypotheses, one uses the Poisson probability, given a model rate $\mu(t)$, and detection of N photons at times $\{t_k\}$, in a total live-time T_L , in (very small) time bins δt [GL92]:

$$p(\{t_k\}|\mu(t), I) = \exp \left[- \int_{T_L} \mu(t) dt \right] \prod_{k=1}^N \mu(t_k) \delta t.$$

Turning the crank. For each hypothesis, one applies Bayes's Theorem, integrates analytically over the amplitude and phase parameters B and ϕ , and then takes the ratio. (The normalization term $p(\mathcal{X}|I)$ cancels, and so is not calculated.) This gives $\lambda(f)$, the log likelihood for parameter estimation:

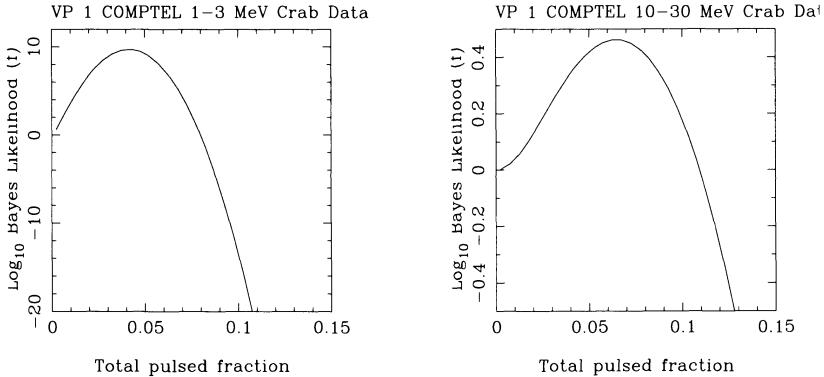
$$\lambda(f) = \log_{10} \left[I_0(\kappa \mathcal{S}_N) / I_0(\kappa)^N \right],$$

where \mathcal{S}_N is defined as $\mathcal{S}_N \equiv \frac{1}{N} \sum_{k=1}^N \cos^2 \Theta(t_k) + \sin^2 \Theta(t_k)$. For $\kappa = 1$, this is analogous to a frequentist Rayleigh statistic.

For hypothesis testing, one obtains the Bayes factor, or ratio of the total probabilities of the interesting to null hypotheses:

$$\mathcal{L} = \int_0^1 df \frac{I_0(\kappa \mathcal{S}_N)}{I_0(\kappa)^N}; \quad f = \tanh(\kappa).$$

Application to data



Here we plot $\lambda(f)$ for two different datasets. Both are from a two week CGRO–COMPTEL observation of the Crab pulsar. The first shows the 1–3 MeV band, where it was detected very significantly (total pulsed fraction $f = 0.042 \pm 0.006$; Bayes factor $\mathcal{L} = 10^{7.8}$). The second shows the 10–30 MeV Crab data. The total pulsed fraction $f = 0.063 \pm 0.03$ is suggestive, but not a formally significant detection (Bayes factor $\mathcal{L} = 10^{-0.6} < 1$).

Adding a complication: astrophysicists need help from statisticians

Joint imaging and timing analysis. With Bayesian inference, it is straightforward to add more information. Since these data were from an imaging telescope, why not use the imaging response on the full dataset, rather than just an angular window around the source? One should be able to derive a likelihood ratio for joint imaging and timing analysis, and at once obtain credible regions for both the source flux and pulsed fraction. The data are the same, save that a much wider angular window was used. There are 157175 photons in this 1–3 MeV dataset (about 1 every 8 seconds); and 7096 in the 10–30 MeV data (about 1 every 3 minutes). The models are a little more complicated. Let j be the index for the spatial imaging bins; β_j the shape of the background as a function of bin position, with $\sum_j \beta_j \equiv 1$; \mathcal{R}_j the instrument response (or point-spread function) in bin j , given the known pulsar position; and A the source flux ($\text{photons}\cdot\text{cm}^{-2}\cdot\text{s}^{-1}$). Note that the shape of the instrument background β_j and the response \mathcal{R}_j are both known a priori. The rate for the null hypothesis, \mathcal{M}_0 , is still one component: $\mu_{0j}(t) = B\beta_j(t)$. However, the rate for the interesting hypothesis, \mathcal{M}_1 , is now two (background + source): $\mu_{1j}(t) = B\beta_j(t) + A\mathcal{R}_j\rho(t)$.

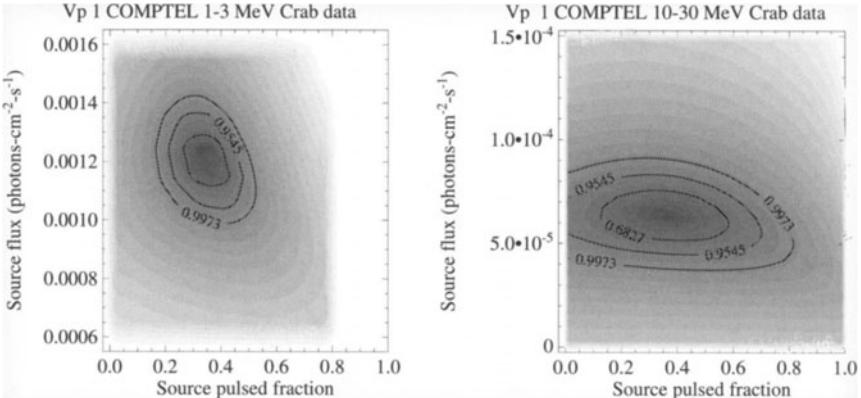
Assigning probabilities. One assigns the same priors for ϕ , f , and B one did previously, but how does one assign a prior for A ? There is no one unambiguous choice, and dependence on A will not cancel when the likelihood ratio is taken. For this calculation, I used a uniform prior on $A \in [0, A_0]$, with $A = 10^4$ photons-cm $^{-2}$ -s $^{-1}$. Once the μ are given, the direct probabilities have the same form as before.

Turning the crank. This gives $\lambda(f)$, the log likelihood for parameter estimation:

$$\lambda(f, A) = \log_{10} \left[p(A|I) \int_0^{B_0} dB \frac{T_L^{N+1}}{N!} \times \exp \left[-T_L \sum_j (B\beta_j + AR_j) \Delta V_j \right] \prod_{k=1}^N (B\beta_{j_k} + AR_{j_k} \rho(t_k)) \right],$$

and global Bayes factor $\mathcal{L} = \int_0^1 df \int_0^{A_0} dA 10^{\lambda(f,A)}$, where the integrations over B , f and A are performed numerically.

Application to data. The results (68.27, 95.45, and 99.73% posterior probability credible regions) are displayed for the same CGRO-COMPTEL Crab observations as before.



The detections appear more significant. For the 1–3 MeV data, one finds a source flux $A = 1.2 \times 10^{-3} \pm 6 \times 10^{-5}$ photons-cm $^{-2}$ -s $^{-1}$; a source pulsed fraction $f = 0.35 \pm 0.05$; and a global Bayes factor $\mathcal{L} = 10^{75.8}$. For the 10–30 MeV data, one finds a source flux $A = 6.1 \times 10^{-5} \pm 7.6 \times 10^{-6}$ photons-cm $^{-2}$ -s $^{-1}$; a source pulsed fraction $f = 0.36 \pm 0.15$; and a global Bayes factor $\mathcal{L} = 10^{5.99}$. However, without a prior for A with an unambiguous normalization, it is hard to interpret the total likelihood of the hypothesis that there is a pulsed γ -ray source. A different choice of prior and A_0 would have given about the same parameter constraints, but different global Bayes factors. This was the problem addressed by J. Berger's “intrinsic Bayes factor” method.

Future thoughts

For the future. Clearly thoughtful priors are an active area of concern for the future. For many problems, an astrophysicist may be able to use physical knowledge of a system to assign reasonable, proper priors; for others, the choice may be ambiguous, so much remains to be worked out. We are aided by both increases in computation speed, and by new numerical integration techniques such as MCMC (Markov Chain Monte Carlo). This allows a greater flexibility in the kinds of problems one can tackle in a reasonable amount of time.

Personal thoughts. I often find that, once having derived a Bayesian likelihood ratio, I later see a relation to a standard maximum likelihood statistic. I find the Bayes prescription clearer, especially when exploring the problem. [TA93] coined term “likelihoodist” to describe those basing their inference on the shape of a likelihood, Bayesian or otherwise. Astronomers are clever people, and come up with many ingenious, intuitive, and speedy ad-hoc statistics. I am coming to consider these as methods of data exploration and visualization; but for the final calculations of probabilities and uncertainties, I encourage astrophysicists to make more use of a “likelihoodist” perspective.

Acknowledgments: I thank T. Loredo, E. Linder and D. Sinha for pivotal discussions. T. Loredo provided software for calculating the Bayesian credible regions shown in the figures. AC is supported through the CGRO-COMPTEL project, which is supported in part through NASA grant NAS 5-26646, DARA grant 50 QV 90968, and the Netherlands Organization for Scientific Research (NWO).

REFERENCES

- [BE96] J. Berger, these proceedings
- [BR92] P. R. Bevington and D. K. Robinson. *Data Reduction and Error Analysis for the Physical Sciences*. Second Edition. McGraw-Hill, 1992.
- [BH93] L. E. Brown and D. H. Hartman. *Astrophys. & Space Science*, **209**, 285, 1993.
- [BI71] A. B. Bijaoui. *Astron. & Astrophys.*, **13**, 226, 1971.
- [CB86] I. J. D. Craig and J. C. Brown. *Inverse Problems in Astronomy*. Hilger, 1986.
- [CA78] W. Cash. *Astrophys. J.*, **228**, 939, 1978.
- [GL92] P. Gregory and T. Loredo. *Astrophys. J.*, **398**, 146, 1992.
- [JA78] E. T. Jaynes. Where do we stand on Maximum Entropy?. *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*. Kluwer, 1978.
- [LM79] M. Lampton, B. Margon and S. Bowyer. *Astrophys. J.*, **208**, 177, 1976.
- [LO92] T. Loredo. In *Statistical Challenges in Modern Astronomy*, Springer-Verlag, 1992.

- [MU95] R. Much *et al.* In *Proceedings of the Compton Symposium, Munich 1995*.
- [TA93] M. A. Tanner. *Tools for Statistical Inference*. Kluwer, 1993.

3

Bayesian Analysis of Lunar Laser Ranging Data

William H. Jefferys and Judit Györgyey Ries

ABSTRACT In 1969, astronauts first placed a retroreflector on the moon for laser ranging of the moon, and since then the McDonald Observatory of the University of Texas has been ranging to these and other later-placed retroreflectors. By determining the round-trip time of a very short but powerful laser pulse, important and extremely precise information about lunar motion and earth rotation can be obtained. The problem is an interesting one from the point of view of signal-to-noise, for in unfavorable circumstances, nearly all of the detected photons are not laser returns but simply background photons. Other interesting features of this problem are the fact that the data are censored; and that it is necessary to take into account the Poisson nature of the data. Determining which photons are actual returns is critical to the initial data analysis. In this paper we describe how a Bayesian analysis of the return data can be used to improve the results.

3.1 Introduction

The only experiment of the Apollo lunar missions still in progress is the lunar laser ranging (LLR) experiment. The arrays of reflecting cornercubes that the Apollo astronauts left on the Moon, along with two other arrays delivered by Soviet spacecraft, do not require power, and their surfaces have not shown measurable degradation since they were deployed. Improvements in laser technology and timing devices have increased the accuracy of the range measurements in the intervening quarter century, but lunar laser ranging remains a technically and scientifically challenging measurement. Lunar laser ranging provides a wide range of scientific results as well as a three orders of magnitude improvement in the lunar ephemeris and lunar rotation variations over earlier techniques. These include, for example, measurements of the Earth's precession, of the moon's tidal deceleration, of the relativistic precession of the lunar orbit, and a test of the Strong Equivalence Principle [DBF⁺94].

Lunar laser ranging is the measurement of the round-trip travel time of a photon emitted from an Earth-based laser. Changes in travel time,

which indicate changes in the separation between the transmitter and the reflector, contain a great deal of information about the Earth-Moon system, which can be retrieved by estimating model parameters. An important signal in the difference between the observed and predicted range is the error in the predicted Earth rotation parameters. The motivation for better identification of photons returning from the moon is to better determine these Earth Orientation Parameters (EOP). The rotation of the Earth is far from constant at the millisecond level, and even after removing variations due to tidal and seasonal periodic variations, there are still signals which at this point are best described as random. Predicted Earth Orientation Parameters provided by the US Naval Observatory are important in navigation and in artificial satellite data acquisition. The more accurately the Earth orientation can be measured in near real time, the better the short-term predictions become. Artificial satellite observations provide very reliable information on the polar motion components of the EOP series. However, model deficiencies of the nodal precession of the satellite orbit (i.e., changes in the orientation of the orbital plane in space) cannot be separated from the rotational component of the EOP, a major problem. Although Very Long Baseline Interferometry can provide accurate measurement of all three components of Earth orientation using distant radio sources, the time-consuming data reduction gives results only about three weeks after the actual measurement. The precision of EOP determination from lunar laser ranging measurements is not as high as with the other two techniques, but it has the advantage of a quick turn-around time of about 12 hours. Combined with the satellite EOP determination, it can correct for the orientation of the satellite's orbital plane, and improve the predictions. However, the quality and quantity of lunar data depends strongly on atmospheric conditions and on the lunar phase. High humidity, atmospheric turbulence and high background noise can make the detection of lunar returns quite difficult.

3.2 Data acquisition

The basic elements of a lunar laser ranging station are a laser, timing equipment, a telescope through which the laser is beamed to the Moon, and which in case of the French and American stations also collects the returning photons, and a computer. The McDonald LLR has a mirror of 0.76 meter diameter, and a Nd-YAG laser which fires 10 pulses in a second, with 200 picosecond pulse width (full width at half maximum), delivering 120 mJoule of energy per pulse. The laser beam contains approximately 3×10^{19} photons. There is a substantial loss of signal due to transmission through the optical elements and the atmosphere, beam divergence, and the distance to the retroreflector. When the laser beam reaches the Moon,

it is spread over an area of about 7 km in diameter. The area covered by the retroreflector is 10^{-9} times smaller than the beam itself, and the cornercubes spread the beam further, producing a 20 km spot on the Earth. This results in a factor of 10^{21} decrease in the signal; for every 30 laser firings, on average one photon arrives back at the detector, making the lunar laser ranging a single photon detection experiment.

An essential requirement in collecting LLR data is an adequate initial model of lunar dynamics, atmospheric refraction, station coordinates and Earth orientation. Based on this model, the telescope can be accurately pointed to the retroreflector on the Moon, which can be tracked during the observation. The interval timer is started as the laser fires, and it stops when the first appropriate photon reaches the detector. The precision of the epoch timing system is approximately 25 picoseconds. To eliminate most of the non-lunar photons, a filter centered on the wavelength of the laser is placed in the path of the returning beam. A temporal filter, a *range gate*, is also implemented. From the initial model, the predicted round trip travel time can be estimated for a given shot, and the detector allowed to observe only while the gate is open for a limited time interval centered on the predicted arrival time. The most commonly used range gate width at the McDonald Observatory LLR station is 400 nanoseconds. A typical lunar run is about 45 minutes long. During this time, depending on the lunar phase (i.e., on the illumination of the lunar disk), and on the atmospheric conditions, ten to a thousand photons reach the detector. When the Moon is in first and third quarter and the reflectors are in the dark, most of the detected photons are laser returns, but in the intermediate phases from first to last quarter, background photons can overpower the lunar returns because the location of the retroreflector is illuminated by the Sun. The individual returns are later analyzed and compressed into normal points of 1 cm precision. Data accuracy is at about the 2-3 cm level.

3.3 Present filtering method

In addition to the spectral and temporal filter the data are run through a software filtering program. The difference between the predicted and measured round-trip times, that is, the residuals, are calculated using the adopted model. For a perfect measurement and model the residuals should be zero. Different model errors introduce different signatures into the residuals, and the noise photons do not follow any pattern. We assume that the mathematical model we are using is correct except for UT0 (timing) errors in the predicted EOP. During the 45 minute run this error would cause the residuals of the returning photons from the laser shot to lie on a straight line of small but unknown slope. The background photons are still randomly distributed in the gate. Assuming that the background noise is uniformly

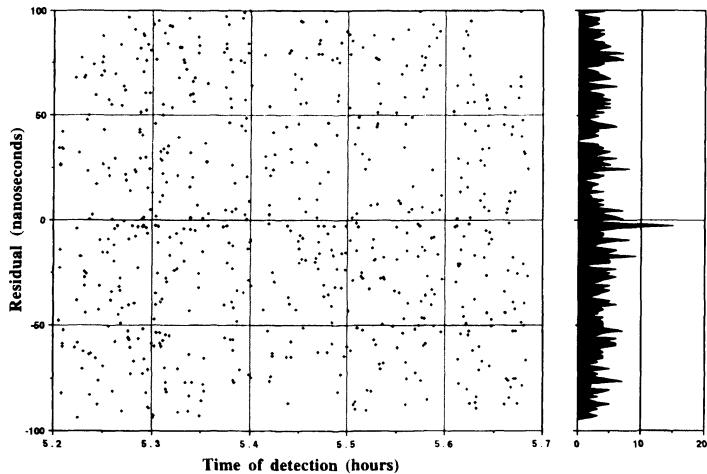


FIGURE 1. Fair-to-average data: The difference between the observed and predicted range is plotted as a function of time for an actual lunar run. The clumping is obvious to the eye on the histogram. The traditional filtering method can identify the laser returns.

distributed, to identify the lunar return the analyst looks for clumping, that is for significant deviation from a uniform distribution [RS92]. The residuals are binned (usually into 1 nanosecond bins) and the maximum number of photons expected in a bin is calculated from the total number of detections. The program looks at the bins in pairs, and looks for a significant deviation from this expected number (The slope and the width of the bins can be adjusted in this process). When it finds such pairs all photons in the two bins are identified as lunar returns and compressed into normal points. The EOP are recovered through another step using nightly corrections based on the normal points.

When the signal is strong, the laser returns can easily be identified even by the eye (Figure 1). This approach breaks down if the number of the total returns is small, or if the noise level is high (Figure 2); in such cases the program cannot decide based only on the maximum expected number of returns whether the detected photons are from the retroreflector or not. In this case no photons are so identified, and some data can be lost. The returning laser pulse is wider than the outgoing pulse. The width of the returning pulse, if we can estimate it, could provide additional information about the precision of the normal point, but at present is not taken into account, except that the bins are chosen to be wide enough to contain the smeared pulse. The EOP determination can be improved by increasing the

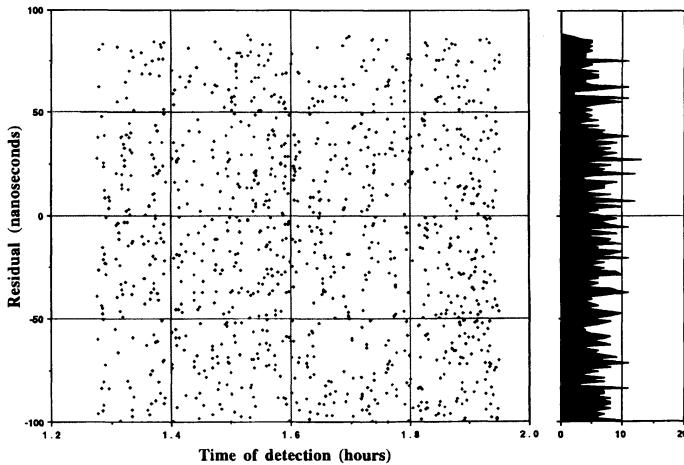


FIGURE 2. Poor data. The data are marginal, and the clumping is not obvious. The traditional filtering method breaks down. However, the ranging crew, with years of experience at lunar laser ranging, indicated that they think they obtained real laser returns in this run.

amount of reliable data going into the calculations, and by recovering the parameters close to real time. The Bayesian approach could help in two ways: It could recover data deemed unusable before, and it eliminates the extra step of forming the normal point for EOP determination. In addition to determining EOP, the data are distributed to the LLR community for scientific analysis, and we hope that the Bayesian approach will help provide a better estimate of the uncertainty of the normal points.

3.4 Bayesian analysis

Loredo [Lor90] has given a nice introduction to Bayesian analysis for astronomers. Rather than try to duplicate his excellent discussion, we will summarize the results.

Bayesian analysis requires the scientist to provide two things: A prior distribution $p(\text{model})$ which is a function of all of the model parameters in the problem, and a likelihood function $L(\text{model}; \text{data})$ which describes in a probabilistic sense what sort of data we expect to see given any particular choice of model parameters. To give a simple example, we may be interested in estimating some quantity, such as the distance D to the Moon. The prior distribution represents our knowledge and/or opinions about this parameter prior to taking some data set. It is a probability density, say, that

tells us (on prior information) that we believe that it is more likely that the parameter lies in certain subsets of the possible values than in others. The prior distribution is a measure of our own ignorance: if we are very sure of the value of the parameter, it will be sharply peaked, and if we are less sure, it will be more broadly spread out.

The prior distribution can vary from individual to individual for various reasons, including the fact that different individuals usually have different prior information. Under many circumstances the result of a Bayesian analysis may not depend critically on the choice of prior within a fairly wide class of priors. In other cases, care must be taken and if the prior is uncertain in such a way that the results do depend sensitively on the prior, one should carefully investigate the way the results depend on the prior.

The likelihood function $L(\text{model}; \text{data})$ is any function that is proportional to the probability of obtaining certain data, given a model; but it is considered as a function of the model (i.e., the model parameters), since the data are considered fixed. Frequently the particular model is specified by a particular parameter set. In our simple example, each possible value of D corresponds to a particular model, and the probability of obtaining a particular set of data—return timings, for example—depends upon the value of D . Data that are consistent with a high value of D are more likely to be obtained if the value of D is high than if it is low, and vice versa. Expressed in the language of conditional probability,

$$L(\text{model}; \text{data}) \propto p(\text{data}|\text{model}) \quad (3.1)$$

The Bayesian prescription tells us that, by Bayes' theorem, the posterior distribution of the parameter given the data, $p(\text{model}|\text{data})$, is proportional to the prior distribution times the likelihood, with a normalization factor that is just the reciprocal of this product integrated over all models (i.e., sets of parameters). Thus

$$p(\text{model}|\text{data}) \propto L(\text{model}; \text{data}) \times p(\text{model}) \quad (3.2)$$

with proportionality factor C given by

$$C^{-1} = \int_{\text{all models}} L(\text{model}; \text{data}) \times p(\text{model}) \quad (3.3)$$

A key characteristic of the Bayesian paradigm is that the results are *conditional* on the data that have been *actually observed*. In other words, the Bayesian analysis does not consider data that might have been observed but were not. This is displayed by the conditional nature of the posterior probability distribution.

All results of interest can be derived from the posterior distribution. For example, by integrating out (marginalizing with respect to) parameters that are not of interest, we can obtain a posterior distribution which is a

function of just those parameters that we are interested in. Such calculations are often the most difficult part of a Bayesian analysis, since it is relatively straightforward in many cases to write down the likelihood function and even the prior, but the integration of the posterior may be difficult because it may be complex and not integrable in closed form. To handle this problem, a number of Monte Carlo methods have been developed in recent years that have proved rather effective [Tan93]. We will utilize this strategy in our discussion.

3.5 The Likelihood Function

The lunar laser ranging problem is an interesting one, not only because the signal is very weak but also because the observations are censored owing to the closing of the range gate whenever the first photon is detected. When the reflector is in the Sun, typically 95% of detections are in fact not genuine returns from the laser pulse, and most laser shots do not result in a detection.

Because we are counting discrete events which are for all practical purposes independent, the statistics governing this problem are Poisson. Often it is possible in large-signal situations to approximate Poisson statistics by an appropriate normal approximation, but that route is not available here. We must deal with the Poisson nature of the data at the outset, without fudging.

We approach the problem of writing down the likelihood function similarly to the way Loredo did in another small-signal problem involving a comparable number of neutrino detections from Supernova 1987A [Lor90]. Loredo observed that in a very short interval of time Δt , the probability that we detect no photons is

$$(r\Delta t)^0 \exp(-r\Delta t)/0! = \exp(-r\Delta t), \quad (3.4)$$

and the probability that we detect a single photon is

$$(r\Delta t)^1 \exp(-r\Delta t)/1! = r\Delta t \exp(-r\Delta t), \quad (3.5)$$

where r is the expected rate per unit time of a detection. By making Δt very small, we can ignore the probability of two detections in the interval. In our case, the rate varies with time, $r = r(t)$, because the gate is automatically opened and closed for each shot (closing can be at the end of the window or when a photon arrives), and also because the probability of the arrival of a photon is significantly enhanced during the very short interval that the range coincides with the actual light-time to the moon.

If the intervals Δt are disjoint, then the detected events are independent and the likelihood function is just the product of (3.4) and (3.5) over all

intervals:

$$L(\text{model}; \text{data}) \propto \exp(-\sum r(t)\Delta t) \prod_N r(t_i)(\Delta t)^N, \quad (3.6)$$

where the data t_i are the times when a photon was detected and N is the total number of detections. Since we only need the likelihood function up to a constant factor, we drop the last term $(\Delta t)^N$ then take the limit as $\Delta t \rightarrow 0$ to obtain

$$L(\text{model}; \text{data}) \propto \exp(-\int r(t)dt) \prod_N r(t_i), \quad (3.7)$$

where the integral is taken over the entire time that the range gate is open (or, equivalently, over the entire duration of the observation set, noting that the probability of detection $r = 0$ when the gate is closed).

The same result is obtained when we analyze the problem using the approach suggested in [Tan93, §2.1].

For our problem, we presume the following form for the rate $r(t)$:

$$r(t) = \begin{cases} 0 & \text{if the gate is closed} \\ r_{bg} & \text{if the gate is open but the return is not expected} \\ r_{bg} + r_s & \text{if the pulse return is expected} \end{cases}$$

Here, r_{bg} is the background detection rate per unit time and r_s is the signal detection rate per unit time.

What do we mean by saying that return is expected? We presume that the width of the returning laser pulse is very brief, a few hundred picoseconds. The pulse travels to the moon and is reflected back to the Earth, arriving some time later. We do not know the time of arrival, since that is what we have to determine. We have a model for return time that is dependent on certain parameters. We predict (on the basis of our model) that at some time t the center of the pulse will arrive; that is the expected time for the return of the pulse.

By using the “box” function $\Pi(x)$, which is 1 when $-1/2 \leq x \leq 1/2$ and zero otherwise, we can express a simple model for $r(t)$ as follows:

$$r(t) = \sum_i \Pi((t - t_{gi})/a_i)(r_{bg} + r_s \Pi((t - t_{ri})/a_{pw})) \quad (3.8)$$

where in the i th shot, a_i is the length of time the range gate is open (normally 400 nanoseconds, but shorter if a detection occurs), t_{gi} is the mean of the opening and closing times of the range gate, a_{pw} is the pulse width, and t_{ri} is the predicted time of the pulse return. It is assumed that there is an unknown bias in the ephemeris of the Moon, so that there is an unknown offset in the Moon’s distance: and that furthermore, the offset varies

in time, linearly to first order, by an amount that is also unknown. Thus we can write

$$t_{ri} = b + c(t - \bar{t}), \quad (3.9)$$

where b is the expected pulse return time at time $t = \bar{t}$, \bar{t} is the midpoint of the data take, and c is the slope of the unknown pulse return time. The unknown parameters b and c of the model are to be determined.

The expression for the likelihood function can be simplified by carrying out the integration and product. We see immediately that

$$\int r(t)dt = Tr_{bg} + (m - k/2)r_s\Delta t, \quad (3.10)$$

where T is the total time that the range gate was open, Δt is the assumed width of the returning pulse, k is the number of detections that occurred within $\Delta t/2$ of the expected pulse arrival time t_{ri} , and m is the number of times that the range gate was open when the pulse return was expected at time t_{ri} . The factor $1/2$ in the last term takes into account the fact that the pulse is detected, on average, halfway through the pulse width, an approximation that is convenient but inessential. We adopt it for the purpose of this calculation.

The product can be written up to a constant factor as

$$r_{bg}^N \left(1 + \frac{r_s}{r_{bg}} \right)^k,$$

where as before N is the total number of detections.

3.6 Prior distribution

The formal unknown parameters in this problem are the detection rates r_{bg} and r_s , and the parameters b and c that describe the expected time of pulse arrival. Usually one would have a pretty good idea of the errors of the ephemeris, and can bound b and c reasonably well. For this investigation we usually adopted a simple prior that is uniform within a range typical of what might be expected for the parameter and zero outside that range. In some cases we adopted a normal prior. For some runs we made life as difficult as we could by assuming that the pulse could return at any time that the range gate was open and setting the width of the prior accordingly. In real life, that is far too pessimistic, but it allowed us to find out just how well we could pin down the actual return time from the data. We restricted the slope c so that over the entire run the expected time of arrival would not vary by more than 10 nanoseconds. The width of the priors on b and c were variable, i.e., we allowed ourselves to be very sure or quite ignorant, depending on the run.

3.7 Gibbs sampler

Once we have determined the prior and the likelihood, we can write down the posterior distribution (up to a constant factor). Now the fun begins. We consider some of the parameters, in particular r_{bg} and r_s , to be “nuisance parameters.” That is, we are not much interested in their actual values. We are most interested in the marginal distributions of b and c , which provide the desired information about the Moon’s orbital motion. These are obtained by integrating over all of the other parameters to obtain marginal posterior distributions $p(b|\text{data})$ and $p(c|\text{data})$ from the complete posterior $p(b, c, r_{bg}, r_s|\text{data})$. We might also be interested in the marginal distributions of r_{bg} and r_s ; during the initial runs, we assumed that we knew r_{bg} and r_s ; this information is actually pretty well known since the characteristics of the laser pulse and the reflection process on the Moon under ideal conditions are well understood after over 25 years of ranging, but can be influenced by weather and other conditions. In later runs we allowed these too to be uncertain.

Our integration method was the Metropolis subchain Gibbs sampler, as suggested in [Mül91] and described in [Tan93, §6.5.3]. The idea behind the Gibbs sampler is to generate a Markov chain using the posterior distribution to generate each next step in the chain. The transition probabilities at each step are prescribed by the posterior distribution, in such a way that one is more likely to make a transition from a region of lower posterior probability to one of higher posterior probability than vice versa. Thus, the Markov chain tends to spend more time in regions of high posterior probability than in regions of low posterior probability. The Markov chain is defined so that a step is taken first in b , then in c , then in r_{bg} , then in r_s , say, to constitute one iteration. Then the process is repeated indefinitely, always starting the new step where the old one left off. After each iteration, the current values of the parameters obtained are tallied separately, to obtain marginal distributions for each parameter. It can be shown that under reasonable conditions that are usually met in practice, the resulting Markov chain yields marginal distributions that approach the actual marginals in the limit.

The difficulty in carrying out this prescription is that the one-dimensional distributions of b , c , r_{bg} and r_s are themselves difficult to sample from, since in general (and in our case) they do not belong to the narrow class of distributions for which analytical sampling schemes exist. However, many probabilistic schemes exist for such sampling. We applied Müller’s ideas to generate trial steps from the one-dimensional distributions at each iteration using a Metropolis-Hastings approach. The details are described in [Tan93], but the basic procedure is as follows. Suppose we are ready to generate the next step in a parameter, say b . From a symmetric, but otherwise *arbitrary* distribution $q(\Delta b)$, generate a delta step Δb . The new trial value of the parameter is $b^* = b + \Delta b$. We accept or reject this trial value

probabilistically based on a simple function of the posterior distribution at the two points b and b^* . In particular, with probability

$$\alpha(b^*, b) = \min \left\{ \frac{p(b^*, c, r_{bg}, r_s | \text{data})}{p(b, c, r_{bg}, r_s | \text{data})}, 1 \right\} \quad (3.11)$$

accept b^* as the new step; otherwise keep b . Then proceed in the same way with c , r_{bg} and r_s to complete the iteration.

A key advantage of this method is that it is unnecessary to work with the normalized posterior probability, since the normalization factor cancels out of the expression for α .

We chose $q(\Delta b)$ to be *uniform* over an interval that was typically 10% of the allowed variation in the particular parameter in question (i.e., the range of the prior for that parameter). This is admittedly crude, and other choices are to be investigated, but it is remarkable how effective this choice turned out to be. Having done this, we then repeated the sample/accept/reject procedure for the second parameter, then the third, and so on, until all parameters had been sampled once. One trip through this procedure for all of the parameters constitutes one iteration of the Gibbs sampler. At this point the current position of the point in parameter space is noted, and the process repeated. Typically we would iterate on the order of 10,000 times to allow the calculation to “burn in,” and then start recording the data for another 10,000 iterates. Finally, histograms of the later iterates’ marginal distributions were tallied and plotted.

3.8 Results

The data used were a set of simulated data from a simulator that has been used previously in the lunar laser ranging project ([RJ95]). The data consisted of 14,400 shots, of which 2313 generated detections. Most of those detections were of noise; only 8 were actual returns, or 0.34%. This particular set of data was intended to represent data of poor quality. Figure 3 shows the actual data as simulated.

The slope c generated by the simulator was $+1.4$; we used various starting values to see if the Gibbs sampler would find the actual slope. Also, we sometimes started the procedure well away from the actual bias to see if we would converge on the actual bias (which for these data was $b = +2.35$ nanoseconds; we would typically start at positions like ± 10 nanoseconds from the true value; the range gate width was 400 nanoseconds for these data).

Our biggest question was whether we would be able to pin down b and c sufficiently well to tell when the returns came. The likelihood function can be expected to have a very narrow peak near the return; would the Metropolis subchain Gibbs sampler find it? We tried a number of runs, first

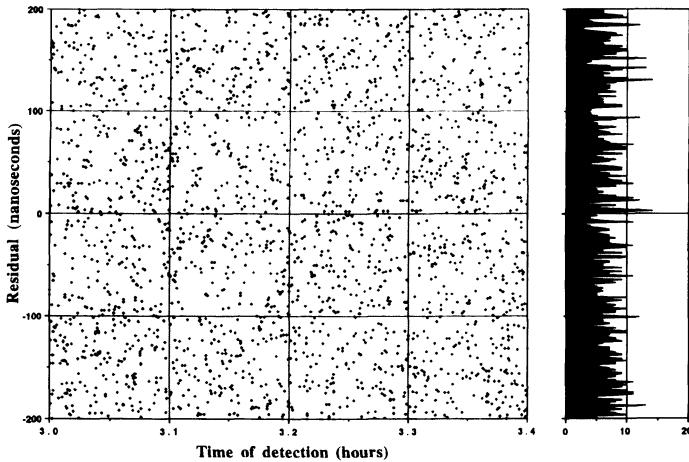


FIGURE 3. Simulated poor data: Simulated returns are plotted as a function of time. The 8 actual returns can barely be seen on the histogram and are difficult or impossible to see on the scatter diagram.

assuming that we knew the detection probabilities and later on computing their marginals as well. The results were really quite gratifying, as can be seen in the figures for one of the later runs.

In the run shown in Figures 4 and 5, we adopted a normal prior for b with mean 0 and standard deviation 15 nanoseconds, cut off at ± 30 nanoseconds and started at $b = 0$ nanoseconds. This prior represents the typical state of knowledge for actual runs. The prior on c was still relatively pessimistic: uniform for $-10 \leq c \leq 10$ and 0 outside that range. The priors on the rates were uniform for rates ≥ 0 , and both the signal and background rates were estimated. The results were that the median of the smallest posterior distribution of b was +2.27 nanoseconds (true value +2.35 nanoseconds) with the smallest 80% Bayesian confidence interval (+2.25, 2.35) nanoseconds. The smallest 95% Bayesian confidence interval was (+1.89, 2.44) nanoseconds. Thus, the tails of the distribution are quite heavy relative to a normal distribution, and the center is very strongly peaked near the true value. This run is actually quite typical, and it shows that despite the fact that the posterior probability is strongly peaked, the Metropolis subchain Gibbs sampler was able to handle it well and find the peak with no apparent difficulty.

The posterior distributions of the slope and of the rates are roughly normal, as expected, and not very interesting, so will not be discussed here.

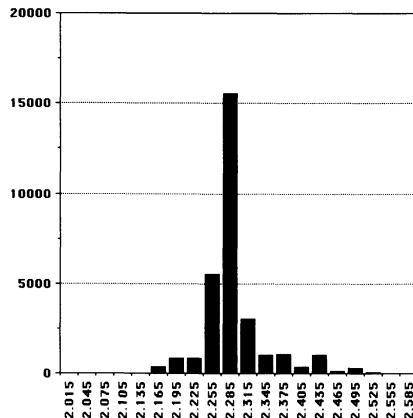


FIGURE 4. Posterior marginal distribution for the bias b that represents the expected return of the photons from the laser at time $t = \bar{t}$. It is strongly peaked near the true value; 80% of the posterior probability is contained within an interval of 0.1 nanoseconds, and 95% within an interval of 0.55 nanoseconds.

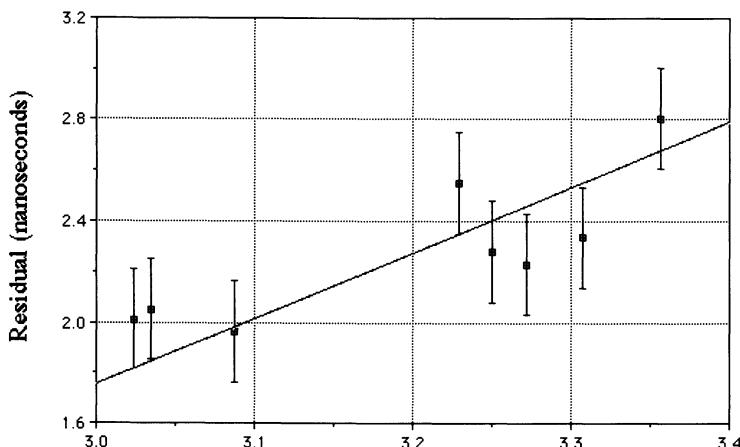


FIGURE 5. The eight actual returns are plotted together with the median line from the marginal distributions of b and c . The error bars indicate the uncertainty in return time.

Figure 5 shows the 8 actual lunar returns, together with the median line from the Gibbs sampler calculation. The vertical scale has been blown up substantially; the error bars in the vertical scale are ± 200 picoseconds (0.2 nanoseconds). The fit is very satisfactory. Indeed, we have been delighted with the way the Gibbs sampler homes in on the input answer even when the signal is so small that one can't pick it out by eye. This is very promising for practical application.

3.9 Conclusions and Future Research

This paper describes a demonstration in principle that a Bayesian approach can be used to analyze the results of lunar laser ranging experiments. With poor data and somewhat pessimistic priors, one clearly and unambiguously picks out the return signal, even though over 99% of the photons detected were noise.

Much work remains to be done. We have merely scratched the surface of this interesting project. Extensive simulations need to be run, for all kinds of data. Real data have not yet been considered and must be analyzed. It will be particularly interesting to see if and how well this method can recognize signal where the older method cannot. We also need to investigate methods of improving the Gibbs sampler, e.g., convergence criteria and methods of deciding ideal step sizes for trial steps.

REFERENCES

- [DBF⁺94] J.O. Dickey, P.L. Bender, J.E. Faller, X.X. Newhall, R.L. Ricklefs, J.G. Ries, P.J. Shelus, A.L. Whipple, J.R. Wiant, J.G. Williams, and C.F. Yoder. Lunar laser ranging: A continuing legacy of the apollo program. *Science*, 265:482–490, 1994.
- [Lor90] T.J. Loredo. From Laplace to Supernova 1987a: Bayesian inference in astrophysics. In P. Fogère, editor, *Maximum Entropy and Bayesian Methods*, pages 81–142. Kluwer Academic Publishers, Dordrecht, 1990.
- [Mül91] P. Müller. A generic approach to posterior integration and bayesian sampling. Technical report 91-09, statistics department, Purdue University, 1991.
- [RJ95] J.G. Ries and W.H. Jefferys. Application of bayesian statistics to lunar data analysis. *Bulletin of the American Astronomical Society*, 27(3):1200, 1995.
- [RS92] R. L. Ricklefs and P.J. Shelus. Poisson filtering of laser ranging data. In *Proceeding of the Eighth International Workshop on Laser Ranging Instrumentation*. Annnapolis: NASA conference Publication 3214, 1992.
- [Tan93] Martin A. Tanner. *Tools for Statistical Inference*. Springer-Verlag, New York, 1993.

Discussion by Steven F. Arnold

This paper presents a very interesting application of Bayesian statistics to an astronomical problem. In fact, the authors have performed essentially the same analysis I would have performed, which does not leave me much to comment on. For this reason, I thought I would take this opportunity to describe why I (a confirmed classical statistician) often do Bayesian analyses on complicated problems such as the one in this paper.

As Jeffreys and Ries describe in Section 4, a Bayesian analysis for a problem starts with the classical formulation of a statistical model for a data set (consisting of the observations, the unobserved parameters and the likelihood function) and adds a probability distribution on the parameters called the prior distribution. There are several advantages to a Bayesian analysis over a classical analysis, particularly the following:

1. Questions are often more easily formulated from a Bayesian perspective;
2. Bayesian procedures are conditional on the observed data and do not involve averages over unobserved possible outcomes for the experiment as classical procedures often are;
3. Bayesian procedures do not have to be adjusted to account for the method used to collect the data as classical procedures often do;
4. Bayesian procedures are "coherent". Optimal classical procedures may be inadmissible even by classical standards.

In spite of these nice qualities of Bayesian procedures, until recently, they were rarely used for several reasons including the following:

1. It is often difficult to think of the unknown parameter as random;
2. It is often difficult to determine an appropriate prior distribution;
3. The Bayesian analysis is only appropriate for the prior the particular prior chosen. Therefore Bayesians with different priors have incompatible solutions to the same problem;
4. Because of the dependence on the prior distribution, Bayesian procedures are not "objective".
5. It is nearly impossible to solve Bayesian problems analytically unless they are very simple;

These disadvantages to Bayesian analysis have always convinced many statisticians such as myself not to use Bayesian analyses except when the parameter is random and I have a natural prior distribution to use

(which hardly ever happens in practice). In recent years I find myself using Bayesian procedures more frequently, especially for complicated problems for reasons which are illustrated in the paper by Jeffreys and Ries.

The first reason I often find a Bayesian analysis appealing is that I can often formulate the problem I want to solve much more easily from a Bayesian perspective than from a classical perspective. This property of Bayesian analysis is primarily due to its being conditional on the data observed as discussed in Section 4 of this paper. In fact I can often formulate problems from a Bayesian perspective for which a classical formulation is essentially impossible.

When the prior distribution is not too peaked and there is a large quantity of data, the data swamp the prior and the Bayesian solution is very close to the classical solution (if I could find the classical solution or even formulate it). In other words, with fairly flat priors and large data sets, the Bayesian solution is not too sensitive to the actual prior chosen. In this case, I often consider the Bayesian solution as an approximation to the classical solution. (In Jeffreys and Ries, the priors chosen were normal and uniform priors which are not too sharply peaked). In fact, in the example in Jeffreys and Ries, the data is simulated, so that the true values of the various parameters are known. Section 8 indicates how accurately the Bayesian procedure approximates the true values.

The most important reason for the increased use of Bayesian procedures is the development of Markov Chain Monte Carlo (MCMC) methods (such as the Gibbs sampler and Metropolis-Hastings algorithm described in Section 7 of Jeffreys and Ries). These algorithms make it possible to do the calculations for many complicated Bayesian problems, calculations which had previously not been possible. In fact for many of these problems it does not appear practical to do the calculations for a classical approach even yet. Another advantage of these MCMC methods is the possibility is the ease of computing histograms for the posterior distributions of the parameters (i.e., the conditional distributions of the parameters given the data) which can be used in an obvious way to find Bayesian "confidence intervals" for the parameters as given in Section 8.

One difficulty with the MCMC procedures is determining when the Markov chain has converged. It is often quite difficult to tell from the process when we have had a long enough burn in period, especially in an experiment with real data in which the true values of the parameters are unknown. (Note that even after the Markov chain has converged to the limiting distribution, it may still jump around a lot, so that convergence of a Markov chain is very different from convergence of a numerical algorithm.)

In using Gibbs sampling, there are two methods which are often employed. Jeffreys and Ries appear to use the single path Gibbs sampler in which a single path of the Markov chain is run with a long burn-in period. Various aspects of the posterior distribution are then inferred from the later occurrences in the chain. Many authors prefer the multiple path

Gibbs sampler in which the Markov chain is run many times and only the last observation is chosen from each path after a long burn-in. Using a multiple path Gibbs sampler, we get something which is very close to a sample from the posterior distribution which makes it more straightforward to estimate variances of posterior distributions and find confidence intervals. It also seems a little easier to look at convergence issues with a multiple path Gibbs sampler. However, the single path method makes more efficient use of the simulated data. Therefore, today many people run multiple paths but also take many observations form each path, thereby perhaps getting the best of both worlds.

Jeffreys and Ries are to be congratulated for giving a very good example of how Bayesian methods and MCMC methods can be used to analyze complicated astronomical data.

Modern Statistical Methods for Cosmological Testing

I. E. Segal¹

ABSTRACT Statistically efficient and equitable methods for cosmological analysis and testing on the basis of objective samples of extragalactic sources are presented. Nontrivial such methods inevitably assume that the population of sources has well-defined uniformity features, or that departures from uniformity follow a designated pattern (parametric or otherwise). The most basic assumption, almost universally made in analysis at lower redshifts, is that of ‘luminosity uniformity’ (LU); i.e., that the intrinsic brightnesses of the sources form a well defined statistical population that is independent of the positions of the sources, in a designated redshift range. A secondary assumption, which can be utilized only in conjunction with LU, is that of ‘spatial uniformity’ (SU).

Assuming only LU, the problem of the maximum likelihood estimation of the luminosity function (L ; i.e., the distribution of the intrinsic brightnesses) on the basis of a sample that is selected without discrimination on the basis of flux, down to a given limit, is soluble in closed form when the L is a step function, in which form LFs are commonly reported in observational studies. The solution, known as ROBUST, is applied to a number of well-known flux limited samples of galaxies and quasars. At low redshifts, the corresponding directly observable predictions of Friedman-Lemaître cosmology (FLC) are extremely deviant, in fact more so than those of any redshift-distance power law up to the cubic. The second power law appears optimal, and this is predicted by the chronometric cosmology (CC) proposed by Segal. At high redshifts, CC remains consistent with observation, without the hypothetical ‘evolution’ required by FLC for consistency. At all redshifts, the deviations of the FLC predictions for cosmology independent directly observable quantities are as predicted by CC for the results of analysis predicated on FLC.

¹Massachusetts Institute of Technology, Room 2-244, Cambridge MA 02139

4.1 Introduction

Is cosmology truly a science, in the sense of modern quantitative physics? Pauli is said to have once rejected a physics paper with the words ‘it isn’t even good enough to be wrong’; is cosmology? In the more sedate language of Popper, is it ‘falsifiable’? Many eminent astronomers, such as Rees and Zel’dovich, have argued for a view of cosmology as a kind of high-tech scenario, whose function is to provide a unified framework into which cosmic observations may be coherently fitted. As realists, they accept the apparent impracticability of rigorous quantitative confirmation of astrophysical theories in the traditional sense of physics. Some, however, such as Kippenhahn and Zwicky, have worried that current theory might be totally illusionary.

My thesis here is that the combination of advances in statistical technology with increased discipline in astronomical observation now make possible rigorous quantitative testing of cosmological theories. When this is applied to many of the best known complete samples, in all four observed wave bands, the results are strongly contraindicative of FLC, and suggestive of CC.

After Einstein (1917) founded modern theoretical cosmology on the basis of gravitational and stability considerations, cosmology became focused on the galaxy redshift phenomenon that had been discovered by Slipher (1917). This culminated in the announcement by Hubble (1929) of the linear redshift-distance relation that bears his name. The linear law was the origin of the acceptance of the expanding universe model that was suggested by the work of Friedman and Lemaître, and remains basic for empirical cosmology. However, Hubble himself remained skeptical of what he called the ‘motion’ theory of the redshift. Hubble and Tolman (1935) suggested rather a space curvature effect as its possible origin; such an effect was later embodied in CC (1972 *et seq.*). In the meantime, the announcement of the linear law inspired Eddington’s influential book, *The Expanding Universe* (1933), and reversed Einstein’s strong opposition to the theories of Friedman and Lemaître, although both men had earlier ridiculed the notion that the universe might be ‘expanding’.

Today there are objective and well-documented samples inclusive of hundreds, thousands, indeed tens of thousands of redshift observations of sources, accompanied by measurements of the flux from and/or apparent angular size of these sources. These are directly observed, theory-independent quantities, in contrast to the ‘distance’ to an extragalactic source, which is remote from direct measurement. The observable scientific content of the linear law lies in its implied relations between such directly observed quantities. These quantities are, to be sure, subject to statistical noise, a variety of perturbations, such as ‘peculiar’ (random) motions, absorption and aperture effects, instrumental deficiencies, variable spectra, etc. This provides a discretionary range for observers, who naturally tend to

resolve ambiguities in accordance with the dominant theoretical paradigm. However, in modern objectively observed samples, such effects should be quantitatively small, only marginally affect the underlying flux-redshift relation, and certainly have no apparent reason to favor an unconventional cosmology. The marginal character of such perturbative effects is in large part testable in conjunction with computer simulations (e.g. Nicoll & Segal, 1982).

In the early days, the difficulties and limitations of observation together with the theoretical dominance of the ‘expanding universe’ model, produced samples whose unknown or subjective specification make reasonably rigorous statistical analysis impossible. Hubble’s derivation of the linear law was based on what he himself described as rough estimates of the distances. These estimates were in turn based on assumptions about stars that had not been, and perhaps could not be, independently validated. Moreover, no attempt was made to correct for the strong magnitude cutoff (*i.e.*, truncation) involved in the sample. The directly observed magnitude and redshift data Hubble reported are in fact indicative of a rough cubic law, if no allowance is made for truncation. It is not clear why the redshift-distance relation he reported shows no corresponding effect. But the truncation, known as the observational magnitude or flux cutoff, or less precisely as Malmquist bias, became recognized as the quintessential statistical problem in extragalactic astronomy. Fortunately, this problem is soluble in an efficient and equitable manner by the procedure known as ROBUST, which is described below.

4.2 Directly observed galaxy statistics

As noted above, the *directly observable*, and thus *scientifically falsifiable*, content of a redshift-distance law resides in its implied relation between flux (and/or angular diameter) and the redshift. This relation derives from geometry, according to which flux varies as the inverse of the surface area on which the light is spread, or locally, as the inverse square of the distance, and the angular diameter inversely with its first power. If the redshift z varies locally as the power p of the distance – a law or ‘cosmology’ that we will denote at C_p – then the flux F will vary locally as $z^{-2/p}$, apart from secondary effects such as those indicated. Traditionally, optical astronomers have measured the flux F on a logarithmic scale, and generally use the magnitude $m = -2.5 \log F + C$ (C being a presently immaterial constant). The corresponding theoretical magnitude-redshift relation is then $m = (5/p) \log z + C'$ (C' another constant).

The first clear-cut determination from a directly observed magnitude-redshift relation that $p = 1$ was based on a sample of 10 objects of the type known as bright cluster galaxies (BCGs, henceforth), of unknown selection

criterion by Hubble & Humason (1931). Over a period of four decades, this sample was enlarged by a succession of observers, notably Mayall, Sandage, and Gunn. BCGs were claimed to be ‘standard candles’, and on this basis no attempt was made to correct for cutoff bias. The graph of the magnitude-redshift data for 41 BCGs due to Sandage (1972) was often presented in textbooks as establishing the linear law beyond serious question.

Unfortunately, this sample was selected and corrected in a thoroughly subjective manner, with the result that its confirmation of the linear law appears essentially circular, if not the product of a self-fulfilling prophecy. Thus, the sample was largely taken from the cluster catalog of Abell (1958), whose selection criterion explicitly assumes the linear law. Sandage sought also to reject a certain class of galaxies known as cDs, which were thought to be excessively luminous and identifiable by subtle morphological characteristics. But occasion the morphology was not clear, and the candidate galaxy was rejected because it appeared excessively luminous under the assumption of the linear law! A variety of procedures, e.g. for the subtraction of the background cluster light from the flux for the individual BCG, or for aperture, or for cluster richness, etc., appear *ad hoc*. The probably most definitive BCG sample, due to Gunn & Oke (1975) and Hoessel *et al.* (1980), and inclusive of some hundred-odd galaxies, noted the uncertain objectivity of some of these corrections and avoided them. But it also granted the cosmology-dependence of its aperture corrections and was taken entirely from the Abell catalog.

The statistical validity of the BCG samples, which was questioned by Zwicky in his classic book on clusters (1959), was argued simply from the closeness of the fit of the linear law, $m = 5 \log z + C$, to the corrected data. The rationalization for this was put succinctly in a letter I received from an exponent of the ‘standard candle’ school as follows: “If $\sigma(M) = 0$, you don’t need completeness”. Here “ $\sigma(M)$ ” refers to the standard deviation of the absolute magnitudes M , which of course are theory-dependent quantities, and “completeness” refers to the sample property of objective definition by a specific flux limit. The simple fallacy implicit here is evident to a statistician, and otherwise requires only a modicum of thought.

Today there is a new wave of claims for the validation of the Hubble law, on the basis of observations of another quite non-generic type of object, namely supernovae. Bold, if not somewhat disingenuous, claims for ‘measurement’ of the distances to supernovae are made, notwithstanding that the crucial difficulty in extragalactic astronomy is that the distance to a source can never be measured in a truly model-independent way. The *directly observable* content of the Hubble law in no way involves putative distances, but is rather is to the effect that the apparent magnitude m and redshift z are related by the equation $m = 5 \log z + M$, where M is a random variable that is independent of z . The ‘distances’ of supernovae are, like the ‘standard candle’ character of the BCGs, *theorized* rather than *observed*. Because of their transience, irregularity, scarcity, and difficulty

of classification into appropriate types, the use of supernovae as primary sample objects for cosmological testing would probably serve to moot the redshift-distance relation indefinitely.

4.3 Studies of complete samples

In contrast to both BCG and supernova samples, ‘complete’ galaxy samples are objective, endlessly observable, and in the range of several hundreds to tens of thousands. The general idea of these samples originated with the work of Shapley and Ames (1932), which began the formation of an objectively specified galaxy sample much larger than the bright cluster galaxy samples used to support the linear law. The sample was to be ‘complete’ in the sense that it included all galaxies whose *apparent* brightness exceeded a specified limit. This is an objective and cosmology-independent criterion that defines a reproducible and otherwise statistically viable sample. But much observational work, notably the observation of the redshifts, was necessary to complete the Shapley-Ames initiative. It was not until the 1970s that large complete galaxy samples inclusive of redshift observations became available.

Modern statistical methodology is actually quite easy to apply to cosmological testing, but (for whatever reason) seems not to have been done prior to our approach in the early seventies. There was no earlier attempt to test complete samples with due allowance for the cutoff, with the exception of the Schmidt V/V_m method (1968), which required SU in addition to LU (and was presented without clear statistical justification).

Prospective researchers in the application of statistics to cosmology should perhaps be made aware of the difficulty of publishing papers whose conclusions will be disliked by referees who know and love FLC, irrespective of their statistical legitimacy. Thus, that LU suffices for LF estimation was noted and applied to the study of the redshift-distance relation by Nicoll & Segal in the early 1970s, but like most of our papers on comparative cosmology (up to the present day in fact), the work was rejected by leading journals. It was published only in an unrefereed summary form (1975a).

After some years we succeeded in getting a more detailed study published (Nicoll & Segal 1978). The procedure for maximum likelihood estimation (MLE) of the LF from a complete sample, without any assumption as to SU, was spelled out and applied. In this paper we parametrized the bright end of the LF by a truncated normal law, and estimated p together with the normal law parameters by MLE. The MLE of p was not significantly different from 2, and the normal law parametrization was validated *a posteriori* with appropriate allowance for the degrees of freedom lost by the MLE. C_1 was strongly contraindicated, but paradoxically, like any poorly fitting theory in the cosmology context, could seek to escape rejection by arguing

that the assumed parametrization was inapplicable. The parametric MLE method assuming only LU was later applied by Sandage, Tammann & Yahil (1979), who referred to our work to assert (without even an attempt at justification) that we reached the conclusion $p \simeq 2$ rather than 1 by neglect of the cutoff bias whose proper treatment was our main methodological contribution!

The rigorous answer to the theoretical exculpation of the linear law by the argument that there is no unexceptional *a priori* parametrization for the LF is that a hypothesis that does not provide a basis for objective testing is not a truly scientific one. A more specific answer derived from an encounter with Michael Woodroffe, as follows.

4.4 Efficient nonparametric LF estimation

Woodroffe suggested a nonparametric approach via representation of the LF by division of the range into some number of bins, and approximation of the LF by a function constant in each bin, in conjunction with the application of MLE to the determination of these constants. At the time, almost two decades ago, computer power was a problem, and even today there is hardly any sure way to globally maximize a highly nonlinear function of a large number of variables by brute force. But a natural closed form nonparametric method for LF estimation eventually occurred to me, and was brilliantly programmed by J. F. Nicoll for use on a small computer. Herman Chernoff suggested that the method, which we dubbed ROBUST because of its universality, might be a MLE. We verified this, and found also that the estimate was a sufficient statistic for the LF.

Following a year or two of rejections from journals in the U.S., a letter summarizing it was accepted by *Astronomy & Astrophysics*, subject to the referee's stipulation that no comparative cosmological results be included! After publication, I received a note from Lynden-Bell to the effect that ROBUST seemed reminiscent of a method he had proposed for quasar studies. The estimator proposed by Lynden-Bell (1971) is similar to ROBUST in being nonparametric and interpretable as MLE, albeit in the rigorously murky case where the number of parameters is infinite. It is however fundamentally different in that it assumes SU as well as LU, and in the potential instability that derives from estimating as many parameters as the sample size. The Lynden-Bell estimate was greatly clarified by Woodroffe (1985), who showed its statistical consistency and placed it in a general context. He also noted that, as we suggested (1983), the Lynden-Bell estimate coincides with ROBUST in the limit of infinitesimal bins, but a reference to ROBUST in this connection seems to have been misinterpreted and given rise to the totally incorrect belief that ROBUST represents an adaptation of the Lynden-Bell estimate to grouped data.

A further and statistically somewhat subtle but important difference is that the Lynden-Bell estimate is automatically self-consistent, by virtue of the law of large numbers. The consistency of ROBUST, however, depends on the correctness of the underlying cosmology, and provides thereby a non-trivial cosmological test. To put it another way, the Lynden-Bell estimate derives from the *joint* magnitude-redshift distribution, while ROBUST derives from the *conditional* distribution of magnitude at given redshifts. The mutual compatibility of the observed conditional distributions of magnitudes at different redshifts imposes nontrivial cosmological constraints.

From a practical position, ROBUST is quite advantageous in being applicable even without completeness in redshift. This is much more difficult to attain and assure than the absence of discrimination on the basis of flux down to the given limit, which is all that ROBUST needs. Moreover, the effectively infinitesimal bins of the Lynden-Bell estimate makes it extremely sensitive to otherwise rather inconsequential (e.g. Segal & Nicoll 1982) and largely uncontrollable astronomical noise, such as absorption,, aperture, peculiar motion effects, etc., which can not be objectively separated from the observed quantities.

In contrast, ROBUST requires only a small number of bins (or parameters) and is correspondingly quite sable. Moreover, it can tolerate missing redshifts, spectroscopic selection, possible departures from SU, etc. It requires only the absence of bias in selection on flux, down to the prescribed limited flux. In comparative cosmological studies, in the interest of equitability, each cosmology (or at low redshifts, value of p) is allotted the same number n of equal-sized bins to cover the *variably truncated* part of the LF. In the part of the LF that is sufficiently bright as to be unaffected by the truncation, the usual simple frequency estimate is applicable. In all of our studies in the past decade, n has been taken as 10. The resulting bins are comparable in size to those used in the astronomical literature, and the use of the fixed number 10 provides a simple systematic baseline for comparisons of studies of different surveys. As n is allowed to increase, the *comparative ranking* of the prediction accuracy of different cosmologies remains largely unaffected, but as was to be expected, the *absolute differences* in the prediction accuracy tend to decrease slowly.

The ROBUST closed form for the MLE of the LF distribution is a quotient of polynomials in the occupation numbers of the bins. The denominators may on occasion vanish, signifying the non-existence of the MLE. In practice, this happens only if the sample is extremely small, or if the cosmology fits very poorly, as in the case of the original form of FCC as applied to quasars. To see how this comes about, it suffices to apply a little thought to the case when just two redshifts are involved. In the brute force implementation of the goal of ROBUST, involving computations of the likelihood, which has been used by Saunders et al. (1988), such nonexistence may well be overlooked, whereas it is immediately apparent when the exact formula is used.

4.5 Prediction of observable statistics

The LF of a population uniquely determines the prediction of any given cosmology for samples drawn from the population. The ROBUST estimate thus provides a basis for predictions of any observable sample statistics. Among the statistics, the first and second order moments usually play a fundamental role. When the redshifts are given, these moments consist of the following: the mean magnitude; the dispersion in magnitude; the slope of the magnitude-redshift relation. Here, ‘magnitude’ may mean either the apparent or the absolute magnitude, but those that derive from the apparent magnitude are cosmology independent, and so are the fundamental statistics for scientific validation of a given cosmology, or for comparative cosmological studies. As an illustration of how the LF, and more specifically ROBUST, may be used, I summarize a phenomenological study of the redshift-distance exponent.

It is particularly easy to predict the slope of the magnitude-redshift relation and the rms of the deviations between the observed magnitudes and their predictions as the conditional expectation of magnitude given the observed redshift. For this reason and for brevity, I restrict consideration at this point to these two statistics. The predictions will be derived from the application of ROBUST to each of several well-known and documented complete samples, in optical, infrared, and X-ray wave bands. The samples are denoted as FIRAS (Fisher et al. 1995); RSA (Sandage & Tammann 1981); RC3-ALL (de Vaucouleurs et al. 1992, D25 diameters); RC3-ELLIPTICALS (elliptical subsample); VISVANATHAN (Visvanathan 1979); QDOT (Saunders et al. 1988); EMSS [AGNs] (Gioia et al. 1990 and Stocke et al. 1991).

Fig. 1 shows the slope prediction errors, and Fig. 2 the rms errors, as a function of cosmology, under the assumption *Cap*. Because the slope of the basic relation varies as $1/p$, this parameter is more appropriate than p itself, and is plotted along the x-axis. In Fig. 1, the absolute difference in slope between the upper and lower redshift halves of the sample, divided by the log of the quotient of the geometric mean redshifts in these two half-samples is plotted along the y-axis. In Fig. 2 the rms in magnitudes between the observed magnitude and the expected magnitude at the observed redshift is plotted along the y-axis, except for the subtraction of a constant for each sample in order to simplify comparison of the results for different samples.

The figures show that the linear law prediction is less accurate than that of any higher power law at least up to $p = 3$, and that $p = 2$ appears to be approximately optimal. The absolute differences between cosmologies shown in Fig. 2 are small, but the resulting comparative rankings are substantially identical to those obtained by replacement of individual redshifts by redshift bins of any given number of sources; the absolute difference between the rms for different values of p increases to large values as the bin size is increased (cf. e.g. Segal 1986a). For the standpoint of conventional

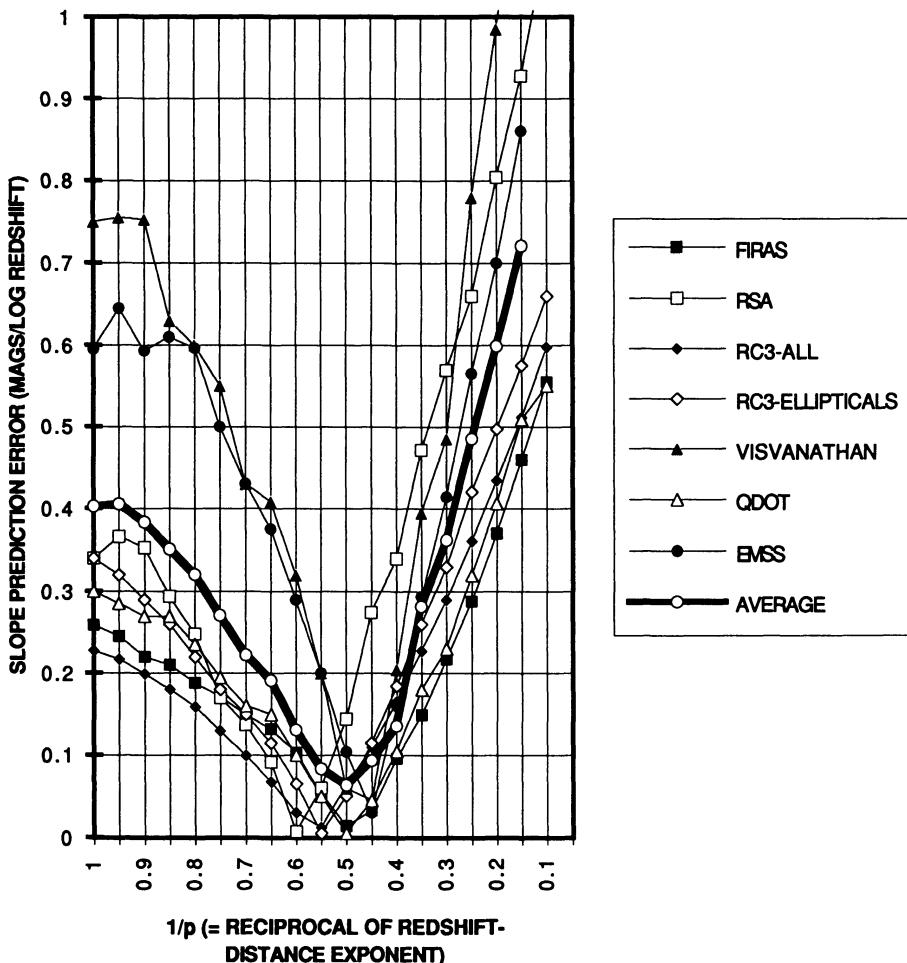


FIGURE 1. Slope prediction error as a function of cosmology in complete optical, infrared, and X-ray samples in the pre-evolutionary range.

cosmology, it may appear surprising that the linear law predictions are not only deviant, but among the least accurate. This, however, is what would be expected if C_2 were correct. For the form of the theoretical magnitude-redshift relation indicates that in lowest order the errors of C_p should vary roughly as $|1/p - 1/2|$. in contrast, the accuracy of the C_2 predictions, including those for the results of analysis predicated on C_1 , are striking.

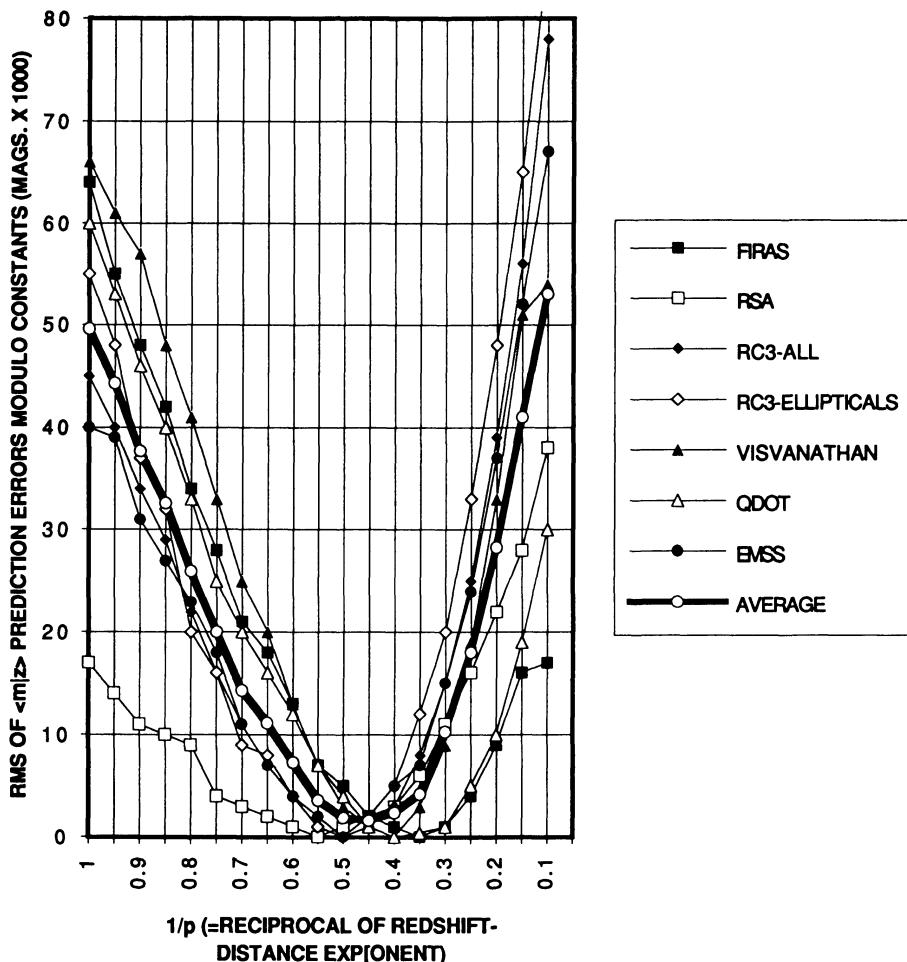


FIGURE 2. Apparent magnitude prediction error as a function of cosmology in complete optical, infrared, and X-ray samples in the pre-evolutionary range.

4.6 Objective estimation of probabilistic significance

The figures indicate that the linear law is seriously flawed, and that the square law is highly tenable. To estimate probabilistic significance requires a more detailed study of the statistics, which will be limited to the cases $p = 1$ or 2 , which are the only values of p associated with a general cosmological theory. To rigorously validate or invalidate in the most rigorous way one

needs to estimate the probability of a deviation between prediction and observation as large as that actually observed. It is also desirable to test the predictions of each cosmology for the results of assuming the alternative cosmology. If one fits poorly and the other fits well, the burden of the poorly-fitting one is to explain the good fit of the other; and conversely, can the well-fitting one fully explain the deviations of the poorly-fitting one?

This detailed study can be done in an efficient and physically intelligent way by a Monte Carlo analysis. Using a large number of random samples drawn from the LF subject to the observational constraints on apparent magnitude and redshift, the distribution of any given statistic can be objectively and reproducibly estimated. Such studies have been made in all four wave bands in which substantial complete samples are well-documented; see, for instance, Segal & Nicoll (1992, 1996b), Segal et al. (1993, 1994), and references therein regarding C_1 and C_2 in lower redshift samples, and Segal & Nicoll (1986, 1996b) and references therein regarding CC and FLC in quasar samples. The cosmologically most sensitive statistic appears as the dispersion in apparent magnitude. Not only are the C_1 predictions for this statistic deviant by as much as 10σ , but the prediction is in *excess* of the observed value, whereas if unobserved factors were affecting the magnitudes, it would be expected that the predicted value would be less than the observed dispersion.

Moreover, the C_2 predictions for the results of analysis predicated on C_1 agree with the actual deviations of the C_1 deviations from observation. C_1 , however, is quite unable to explain the excellent fit of C_2 . A large repertoire of statistics was determined and tested in the cited samples. Among cosmology dependent statistics, the correlation of absolute magnitude with redshift was the most sensitive. C_2 correctly predicted both the C_1 and C_2 correlations, whereas the C_1 prediction for its own (as well as the C_2) correlation was highly deviant.

4.7 Quasars and other high-redshift sources

The largest redshift in Hubble's original paper was about 0.0035, but quasars have been observed with redshifts more than 1000 times greater. It would be absurd to neglect these high-redshift observations in considerations of the nature of the redshifts: luminosities appear to be monotone increasing with redshifts, rather than from a fixed population. The quasar observations gave rise on the one hand to the suggestions by H. Arp and G. Burbidge (cf. Burbidge et al. 1990 and references therein) that some redshifts are not of Doppler origin. But the more orthodox explanation was that the source population was *evolving* in its luminosity and/or spatial distribution. This however is not an *explanation* but a *description*, which

involved whole adjustable functions, and substantially total loss of predictive capacity for FLC at large redshifts regarding directly observable quantities.

In contrast, CC predicts the main observed features of quasars, such as the relative paucity of large-redshift quasars, and the surprising smallness of the decrease in apparent brightness for quasars at very high redshifts. It does this entirely without adjustable cosmological parameters such as q_0 or Λ , and is devoid of ‘evolution’, or any comparable adjustable features. Its predictions for the collective statistics of high redshift complete samples have been validated by the same Monte Carlo procedure described in the low-redshift regime.

The high-redshift predictions of FLC can of course not be proved wrong when an unlimited number of adjustable functions are provided under the rubric ‘evolution’. What can be done, however, is to show that the deviations of the original nonevolutionary theory from observation are identical to what is predicted by the non-parametric CC for the results of analysis that assumes FLC. This has been done for many samples. Studies at high-redshifts of quasars, active galactic nuclei in the X-ray band, and sources in the radio band (e.g. Nicoll & Segal 1975b, 1985; Segal 1986b, 1987, 1990; Segal & Nicoll 1986, 1996; Segal & Segal 1980; Segal et al. 1994a, 1994b) confirm the empirical consistency of CC, and its ability to explain the deviations of the nonevolutionary FLC predictions from direct observation. In the absence of any direct observational or model-independent means to substantiate the evolution hypothesis, its primary scientific role would appear to be that of an exculpatory theoretical artifact rather than an actual physical phenomenon.

Objective cosmology independent definition of quasars or other high-redshift sources is an intrinsically difficult problem, but one of the most carefully selected sample, at the highest redshifts at which carefully selected samples, at the highest redshifts at which complete samples are available, due to Warren et al. (1994) exemplifies these results. A recent much larger and broader sample, the ‘Large Bright Quasar Survey’ of Foltz et al., confirms these results; e.g. (i) the CC predictions are accurate within $\sim 2\sigma$; (ii) the nonevolutionary FLC predictions are deviant by as much as $\sim 10\sigma$; (iii) the CC predictions of the FLC errors are correct within $\sim 2\sigma$ (Segal & Nicoll 1996c).

4.8 Spatial distribution considerations

Traditionally, SU has been the focus of many investigations, such as those of Schmidt regarding quasars, and it is interesting to explore its validity in various contexts, notwithstanding that only LU is required for the more conservative analysis above. This was done by Segal and Nicoll both for

low-redshift galaxy samples, and high-redshift quasar samples, using V/V_m tests, the $N(< z)$ relation, and variants such as the predictions of $< z >$. The technique is the same as that described above, except that in Monte Carlo studies the redshifts must be chosen at random in the frame of the cosmology under consideration, rather than taken as the observed redshifts, in order for the test to be nontrivial. CC was found to be not only consistent with SU, but to explain the deviations of the FLC predictions from observation. Of course, no statistical analysis is immune to sample ‘gerrymandering’. E.g. by arbitrarily deleting a part of the sample of Schmidt & Green (1983), merging it with a sample having a different selection criterion, and oversimplifying the spectral index relation (important for SU considerations at the highest redshifts), a nominal departure from SU for quasars was set up for CC by a guardian of the conventional faith.

4.9 Discussion

By normal statistical standards, the extensive and consistent pattern of deviance of the FLC predictions, in all large complete samples that have been available to us, in all observed redshift ranges, and in all observed wave bands, implies that FLC is scientifically unacceptable. If the deviance of FLC predictions at low redshifts be ascribed to ‘evolution’, as has been done at higher redshifts, there is then no physically falsifiable content to the linear law, on which the FLC is based.

To be sure, there is more to the evaluation of a scientific theory than statistical considerations. But FLC gives up the greatest achievement of 19th century physics, the global law of the conservation of energy, in return for a *deus ex machina* that provides an intuitive explanation for the redshift. Unfortunately, this explanation appears somewhat sophistical from a mathematical or philosophical perspective. In contrast, CC restores global conservation of energy, and explains why the cosmic spectral shift is to the red rather than blue. It provides (Segal 1990) an effective and self-consistent means of determination of the cosmic distance scale, in contrast to the persistently contradictory estimates derived from FLC. It redeems the natural global space-time structure proposed by Einstein, and the suggestion by Hubble and Tolman for the space curvature origin of the redshift, while at the same time directly predicting, without any contrived scenario, the cosmic microwave background (CMB) radiation and its isotropy (e.g. Segal & Zhou 1995). The replacement of FLC by CC as the working model of choice would therefore appear to be highly beneficial to observational astronomy, and long overdue.

REFERENCES

- [1] Abell, G. O. 1958, *Astrophysical J.*, S 3, 211.

- [2] Burbidge, G., A. Hewitt, J. V. Narlikar & P. Das Gupta 1990, *Astrophysical J.*, S 74, 675.
- [3] de Vaucouleurs, G., A. de Vaucouleurs, R. J. Buta, H. G. Corwin Jr., P. Fouqué & G. Paturel 1992, *Third Reference Catalog of Bright Galaxies* (3 vols.), Springer-Verlag.
- [4] Eddington, A.S. 1933, *The Expanding Universe*, Cambridge Univ. Press.
- [5] Einstein, A. 1917, *Sizer. Preuss. Akad. Wiss.*, 142.
- [6] Fisher, K. B., J. P. Huchra, M. A. Strauss, M. Davis, A. Yahil & D. Schlegel, 1995, *Astrophysical J.*, S 100, 69.
- [7] Gioia, I. M., T. Maccacaro, R. E. Schild, A. Wolter, J. T. Stocke, S. L. Morris & J. B. Henry 1990, *Astrophysical J.*, S 72, 567.
- [8] Gunn, J. E. & J. B. Oke 1975, *Astrophysical J.*, 195, 255.
- [9] Hoessel, J. G., J. E. Gunn & T. X. Thuan 1980, *Astrophysical J.*, 241, 486.
- [10] Hubble, E. 1929, *Proc. Nat. Acad. Sci., USA*, 15, 168.
- [11] Hubble, E. & Humason, M. L. 1931, *Astrophysical J.*, 74, 43.
- [12] Hubble, E. & R. C. Tolman 1935, *Astrophysical J.*, 82, 302.
- [13] Lynden-Bell, D. 1971, *Mon. Not. Roy. Astron. Soc.*, 155, 95.
- [14] Nicoll, J. F. & I. E. Segal 1975a, *Proc. Nat. Acad. Sci., USA*, 72, 2473.
- [15] Nicoll, J. F. & I. E. Segal 1975b, *Proc. Nat. Acad. Sci., USA*, 72, 4691.
- [16] Nicoll, J. F. & I. E. Segal 1978a, *Ann. Phys.*, 113, 1.
- [17] Nicoll, J. F. & I. E. Segal 1978b, *Proc. Nat. Acad. Sci., USA*, 75, 535.
- [18] Nicoll, J. F. & I. E. Segal 1980, *Astron. & Astrophys.*, 82, L3.
- [19] Nicoll, J. F., D. Johnson, I. E. Segal & W. Segal 1980, *Proc. Nat. Acad. Sci., USA*, 77, 6275.
- [20] Nicoll, J. F. & I. E. Segal 1982, *Astron. & Astrophys.*, 115, 225.
- [21] Nicoll, J. F. & I. E. Segal 1983, *Astron. & Astrophys.*, 118, 180.
- [22] Nicoll, J. F. & I. E. Segal 1985, *Astron. & Astrophys.*, 144, L23.
- [23] Sandage, A. R. 1972, *Astrophysical J.*, 178 1 & 25.
- [24] Sandage, A. R. & G. A. Tammann 1981, *A Revised Shapley-Ames Catalog of Bright Galaxies*, Carnegie Inst. of Washington.
- [25] Saunders, W., M. Rowan-Robinson, A. Lawrence, G. Efstathiou, N., Kaiser, R. S. Ellis & C. S. Frenk 1990, *Mon. Not. Roy. Astron. Soc.*, 242, 318.
- [26] Schmidt, M. 1968, *Astrophysical J.*, 151, 393.
- [27] Schmidt, M. & R. F. Green 1983, *Astrophysical J.*, 269, 352.
- [28] Segal, I. E. 1972 *Astron. & Astrophys.*, 18, 143.
- [29] Segal, I. E. 1976, *Mathematical Cosmology and Extragalactic Astronomy*, Academic Press, New York.
- [30] Segal, I. E. 1983, *Astron. & Astrophys.*, 123, 151.
- [31] Segal, I. E. 1983, in *Quasars and Gravitational Lenses*. Proc. 24th Liege Astrophys. Coll., 293.
- [32] Segal, I. E. 1986a, *Proc. Nat. Acad. Sci., USA*, 83, 7129.
- [33] Segal, I. E. 1986b, *Publ. Astron. Soc. Japan*, 38, 611
- [34] Segal, I. E. 1987, *Astrophysical J.*, 316, L5.
- [35] Segal, I. E. 1990, *Mon. Not. Roy. Astron. Soc.*, 242, 423.
- [36] Segal, I. E. 1993, *Proc. Nat. Acad. Sci., USA*, 90, 4798.
- [37] Segal, I. E. & J. F. Nicoll, 1986, *Astrophysical J.*, 300, 224.
- [38] Segal, I. E. & J. F. Nicoll, 1992, *Proc. Nat. Acad. Sci., USA*, 89, 11669.
- [39] Segal, I. E. & J. F. Nicoll, 1996a, *Astrophysical J.*, 459, 496.
- [40] Segal, I. E. & J. F. Nicoll, 1996b, *Astrophysical J.*, 465, 578.

- [41] Segal, I. E. & J. F. Nicoll, 1996c, preprint.
- [42] Segal, I. E. & W. Segal 1980, *Proc. Nat. Acad. Sci., USA*, 77, 3080.
- [43] Segal, I. E., J. F. Nicoll & E. Blackman 1994a, *Astrophysical J.*, 430, 63.
- [44] Segal, I. E., J. F. Nicoll & P. Wu 1994b, *Astrophysical J.*, 431, 52.
- [45] Segal, I. E., J. F. Nicoll, P. Wu & Z. Zhou 1991, *Naturwiss.*, 78, 289.
- [46] Segal, I. E., J. F. Nicoll, P. Wu & Z. Zhou 1993, *Astrophysical J.*, 411, 465.
- [47] Segal, I. E. & Z. Zhou 1995, *Astrophysical J.*, S 100, 307.
- [48] Shapley, H. & A. Ames 1932, *Harvard Ann.*, 988, 2.
- [49] Slipher, V. M. 1917, *Proc. Amer. Phil. Soc.* 56, 503.
- [50] Spinrad, H., S. Djorgovski, J. Marr & L. Aguilar 1985, *Publ. Astron. Soc. Pacific*, 97, 932
- [51] Stocke, J. T., S. L. Morris, I. M. Gioia, T. Maccacaro, R. Schild, A. Wolter & J. P. Henry 1991, *Astrophysical J.*, S 76, 813.
- [52] Visvanathan, N. 1979, *Astrophysical J.*, 228, 81.
- [53] Warren, S. J., P. C. Hewett & P. S. Osmer 1994, *Astrophysical J.*, 421, 412.
- [54] Woodroofe, M. E. 1985, *Ann. Stat.* 13, 163.
- [55] Zwicky, F. 1959, *Clusters of Galaxies: Handbuch der Physik LIII*, Springer, Berlin.

5

Comparing Censoring and Random Truncation via Nonparametric Estimation of a Distribution Function

Grace L. Yang ¹

ABSTRACT The Kaplan-Meier estimator and the Lynden-Bell estimator of a distribution function play pivotal roles in the nonparametric analysis of incomplete data. The former is constructed with a right-censored sample of lifetimes and the latter with a randomly truncated sample. Although both estimators look similar in their product-limit forms, they are quite different in distributional properties, especially the variances.

We use these two estimators to compare censoring and truncation. We first consider four models for incomplete data: the right-censoring model, the random truncation model, and two of their generalizations, the censoring-truncation model and the double censoring model. The generalizations are introduced to contrast the first two which are our focus. By way of comparison, we discuss model identifiability, hazard functions, a unified way of constructing nonparametric estimator of the distribution function by using an inversion formula, some of the difficulties in the application of the method, and some recent results particularly on random truncation.

5.1 Introduction

Over the last twenty some years there has been intensive research activities on the analysis of censored lifetimes data. The research was particularly stimulated by a publication of Kaplan-Meier (1958) who proposed a nonparametric estimator (KM) of the distribution function with right-censored data. The estimator is a complicated product of stochastically dependent terms (see equation (5.24) below) which makes the study of its distributional properties very difficult. Thanks to the effort of many, there are now a variety of probabilistic tools to study the KM estimator and more compli-

¹Department of Mathematics, University of Maryland, College Park, MD 20742

cated censoring problems. Results are now routinely used in biostatistics, industrial reliability studies, and many other fields. They have also been applied to the analysis of astronomical data, see e.g., Feigelson and Nelson (1985), Schmitt (1985) and the proceedings of the last conference on Statistical Challenges in Modern Astronomy (Feigelson & Babu 1992).

As regard to truncated data, it is abundant in astronomy. However, the nonparametric analysis of truncated data appeared in the statistical literature much later than that of censored data. Woodrooffe(1985)'s paper seems to be the first systematic study of the nonparametric estimator of the distribution function with randomly truncated data (see equation (5.26) below) constructed by Lynden-Bell(1971).

Since both estimators play pivotal roles in the nonparametric analysis of incomplete data, in this presentation we shall use these two estimators to compare censoring and truncation. By way of comparison we discuss probability models for incomplete data that arise from different sampling and censoring-truncation mechanisms, point out some of the difficulties in the application of the method, and also identify some statistical problems for further investigation.

This is an expository talk and only some of the recent results will be cited with selection bias toward random truncation.

The paper is organized as follows. Section 2 contains four models for incomplete data: the right-censoring model, the random truncation model and two of their generalizations, namely the censoring-truncation model and double censoring model. The generalizations are presented for the purpose of contrasting the first two which will be our focus. Section 3 discusses very briefly the analysis of luminosity data. The reader is referred to Feigelson and Nelson (1985), Schmitt (1985) and Feigelson and Babu (1992) for thorough discussions and further references.

Section 4 discusses model identifiability and the general definition of hazard or intensity functions. The result in Section 4 leads to a unified way of constructing estimate of the distribution function with either right-censored or randomly truncated data or both. This is contained in Section 5. The construction demonstrates the importance of the hazard function and cumulative hazard function in estimation and model identifiability with right censored or randomly truncated data. However, we also point out that these functions are not easy to use for doubly censored data for which explicit formulas for nonparametric estimates of the distribution function are not available. Asymptotic optimal estimates have to be calculated recursively and numerically. Section 6 contains some distributional results of the estimates, particularly some recent ones. Conditions for finite variance are discussed. Concluding remarks are in Section 7.

5.2 Stochastic models for censored and randomly truncated data

Consider a pair of independent random variables X and Y . Let $F(t) = P[X \leq t]$ and $G(t) = P[Y \leq t]$ denote their distribution functions. Right-censoring refers to the situation that one cannot simultaneously observe both X and Y but the minimum of X and Y (to be denoted by $Z = X \wedge Y$) and the indicator $\delta = I[X \leq Y]$ where $\delta = 1$ if $X \leq Y$ and $\delta = 0$ if $X \geq Y$. We say X is subject to right-censoring by Y .

The joint distribution of the observable pair (Z, δ) is called the right-censoring model which is given by

$$Q(t, 1) = P[Z \leq t, \delta = 1] = P[X \leq t, X \leq Y] = \int_0^t P[Y \geq u]dF(u) \quad (5.1)$$

$$Q(t, 0) = P[Z \leq t, \delta = 0] = P[Y \leq t, X > Y] = \int_0^t P[X > u]dG(u) \quad (5.2)$$

The problem is to estimate the distribution function F of X from a random sample of n pairs of (Z, δ) to be denoted by $\{(Z_j, \delta_j), j = 1, \dots, n\}$. For simplicity, X and Y are assumed to be nonnegative random variables.

The duality of X and Y is apparent. Calling Q a right-censoring model is a matter of emphasis. The same data set can be used for estimation of either G or F or both. So if the problem is to estimate G , then Q should be called a left-censoring model. To fix the idea, we shall use the problem of estimating F in our discussion.

In some other situations, especially in astronomy, the experimenter can observe both X and Y provided that $X \geq Y$. In this case, it is convenient to denote the observable X and Y by U and V respectively, for U and V have distributions different from that of X and Y . Unlike Z and δ , it is not possible to write U and V as functions of X and Y ; U and V are related to X and Y in distribution only. The observable ranges of U and V are affected by the random truncation. The ranges will be specified in section 4.2. The random truncation model refers to the joint distribution of U and V given by

$$\begin{aligned} H(x, y) &= P[U \leq x, V \leq y] = \frac{P[X \leq x, Y \leq y, X \geq Y]}{P[X \geq Y]} \\ &= (1/\alpha) \int_0^x G(y \wedge u)dF(u) \quad \text{for } x \geq y, \end{aligned} \quad (5.3)$$

where

$$\alpha = P[X \geq Y] \quad (5.4)$$

which is assumed to be positive. Then the marginal distributions of U and V are given by

$$F^*(x) = P[U \leq x] = H(x, \infty) = \frac{1}{\alpha} \int_0^x P[Y \leq u] dF(u), \quad (5.5)$$

$$G^*(x) = P[V \leq x] = H(\infty, y) = \frac{1}{\alpha} \int_0^x P[X \geq u] dG(u), \quad (5.6)$$

An important quantity in estimation is the probability $R(x)$ of the random interval $[V, U]$ that covers an arbitrarily chosen value say x . Since $V \leq U$, we have

$$R(x) = P[V \leq x \leq U] = G^*(x) - F^*(x-). \quad (5.7)$$

The right-censoring model Q of (5.1)-(5.2) is an unconditional distribution whereas the random truncation model H of (5.3) is a conditional distribution.

The above models can be generalized. Data collected from a truncated population may still subject to censoring. For instance, under the condition $X \geq Y$, X may be subject to right censoring as $Z = X \wedge C$ and $\delta = I[X \leq C]$ where C is another independent random variable. This gives rise to a truncation-censoring model

$$\frac{P[X \leq x, Y \leq y, X \geq Y, Z \leq t, \delta = i]}{P[X \geq Y]} \quad \text{for } i = 0, 1, \quad (5.8)$$

where x, y , and t are subject to constraints induced by either censoring or truncation.

Another generalization would be to subject X to double censoring from left and right as follows (without truncation). The variable X is observable if and only if it lies in the random interval $(Z, Y]$ where Z and Y are random variables with $Z \leq Y$. If X is not in $(Z, Y]$, the exact value of X cannot be determined and we only know whether X is less than Z or greater than Y . The observable part of X can be expressed by a pair of random variables W and δ , where

$$\begin{aligned} W &= \max(\min(X, Y), Z) \\ \delta &= 1, \quad \text{if } Z < X \leq Y, \\ &= 2, \quad \text{if } X > Y, \\ &= 3, \quad \text{if } X < Z. \end{aligned} \quad (5.9)$$

The bivariate distribution of W and δ (under the assumption that X is independent of Z and Y) given below is called the double censoring model,

$$Q(t, 1) = P[W \leq t, \delta = 1] = - \int_t^\infty (S_Y(u) - S_Z(u)) dS_X(u),$$

$$\begin{aligned} Q(t, 2) &= P[W \leq t, \delta = 2] = - \int_t^\infty S_X(u) dS_Y(u), \\ Q(t, 3) &= P[W \leq t, \delta = 3] = - \int_t^\infty (1 - S_X(u)) dS_Z(u) \end{aligned} \quad (5.10)$$

where $S_X(u) = P[X > u]$ is the survival function of X , S_Y and S_Z are similarly defined.

The assumption of stochastic independence stated in each of these models must be imposed. Without it the distribution of F of X cannot be determined uniquely or estimated consistently by nonparametric methods.

5.3 Analysis of luminosity data

In the statistics journals, not many papers addressed the nonparametric analysis of luminosity data. Notable exceptions are papers by Bhattacharya (1983), Bhattacharya, Chernoff and Yang (1983), and Woodroffe (1985). They are concerned with nonparametric estimation and linear regression analysis of the luminosity function using random truncation models. Since then, the random truncation model has received increasing attention. I will mention some recent development in these two areas. To proceed, it is necessary to cast the model into an astronomical setting.

The luminosity function is the distribution of the absolute luminosities of objects of certain type in a specified region of the sky. Each object in the defined population has an absolute luminosity L and a distance d . The problem is to determine the distribution of L . However L is not directly observable by the instrument. The observable quantity is the apparent luminosity l of the object and L is deduced from l by the relation

$$L = 4\pi d^2 l.$$

The distance d is not directly measurable either. Its value has to be inferred from models. The determination of d also varies with the defined survey population. For instance, for the population of quasars, d is a function of the redshift and other cosmological parameters, whereas for the population of stars in the vicinity of our sun it is something else.

The apparent luminosity has a lower limit l_0 such that any object with absolute luminosity smaller than $L_0 = 4\pi d^2 l_0$ will not be detected by the instrument.

The luminosity is often measured in a log scale called magnitude. The absolute magnitude of an object is defined by $M = -2.5 \log L$ and the apparent magnitude $m = -2.5 \log l$. The functional relation between L and l is translated into that of M and m as

$$m = M + f$$

where f is a function of some other cosmological parameters. For our discussion it is not necessary to specify the form of f . Similarly, put $m_0 = -2.5 \log l_0$, so that the observable range of the apparent magnitude is $-\infty < m \leq m_o$, truncated at m_o .

Woodroffe (1985) formulated the estimation of the luminosity function as a problem of estimating the distribution function G of Y in the random truncation model. With galaxies in mind, the f he used is a function of the redshift z and the relation

$$m = f(z) + M \quad (5.11)$$

is converted to that of X and Y in the random truncation model (5.3) by the transformation:

$$X = \exp[-f(z)], \quad \text{and} \quad Y = \exp[M - m_0]. \quad (5.12)$$

where z and M are assumed to be independent random variables (according to the Cosmological Principle). The estimation problem will be discussed in the next section.

Bhattacharya, Chernoff and Yang (1983) considered the following truncated regression model for the apparent magnitude which is denoted by Y :

$$Y = \beta_0 X + M \quad (5.13)$$

where X and M are independent random variables. X is a function of the redshift, and β_0 is an unknown positive parameter. Both X and M are observable only if $Y \leq m_0$. The problem is to estimate the regression slope β_0 and the distribution of M based on n independent observations $(X_i, Y_i), i = 1, \dots, n$. We shall not dwell on this topic and refer the reader to Lai and Ying (1991) for some recent development.

There are of course other types of truncations. For instance Trumpler and Weaver (1953, page 253) discussed truncation problems in which X and Y are observable only if some known function $g(X, Y) \leq \text{constant}$. If we take $g(X, Y) = Y - X$, the random truncation model (5.3) becomes a special case. It would be interesting to investigate this problem nonparametrically.

The censoring-truncation model (5.8) seems to fit the description of a problem discussed in Feigelson (1992). The problem is to compare the luminosity function of a sample selected from an early survey of the objects using different measurement instruments. If the original sample is randomly truncated, then the observations of the second sample, collected from the first survey, are subject to further censoring by a variable C like the one given in equation (5.8). The censoring-truncation model has been used to analyze cross-sectional lifetime data, see, e.g. Tsai, et al. (1987).

5.4 Intensity functions and model identifiability

For right-censoring and random truncation models, it is especially convenient to use hazard function and cumulative hazard functions of X and Y for statistical inference. We shall illustrate their usefulness in model identification and construction of estimates.

The name hazard function comes from the analysis of survival data in biostatistics. For computing luminosity functions, it would be more suggestive to call it the intensity function (for light intensity). In fact, intensity function is a commonly used term in probability theory.

Since the nonparametric estimators considered in this paper are functions of the empirical distributions of the data which are discrete distributions, we need to work primarily with the intensity function of a discrete random variable. Thus suppose X is a discrete random variable that takes nonnegative values x_1, x_2, x_3, \dots . The intensity function (hazard rate) of X at t is defined by the conditional probability

$$\begin{aligned} P[X = t | X \geq t] &= \frac{\Delta F(t)}{S_X(t_-)} = \frac{P[X=x_k]}{P[X \geq x_k]} \quad \text{if } t = x_k, k = 1, 2, \dots \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Here we have used the notation $S_X(t_-) = P[X \geq t]$ and $\Delta F(t) = F(t) - F(t_-) = P[X = t]$. The difference notation Δ is very convenient for calculations. It is a trivial fact but important to note that the conditioning event is $[X \geq x_k]$ and not $[X > x_k]$. The latter would make the conditional probability zero and useless. Thus the clumsy notation $S_X(t_-)$ serves its useful purpose.

The cumulative intensity function $\Lambda(t)$ of X at t is the sum of the intensities up to t , i.e.,

$$\Lambda(t) = \sum_{u \leq t} \frac{\Delta F(u)}{S_X(u_-)} \quad \text{for } t \geq 0, \tag{5.14}$$

with the convention that $\frac{0}{0} = 0$. In terms of the difference notation Δ , the intensity function or the hazard rate is

$$\Delta\Lambda(t) = \frac{\Delta F(t)}{S_X(t_-)} \quad \text{and} \quad \Lambda(t) = \sum_{u \leq t} \Delta\Lambda(u) \quad \text{for } t \geq 0. \tag{5.15}$$

Given S we can compute Λ . Conversely, given Λ or $\Delta\Lambda$, it is easy to deduce S_X from (5.15),

$$S_X(t) = \prod_{0 \leq u \leq t} (1 - \Delta\Lambda(u)) = \prod_{x_j \leq t} (1 - \Delta\Lambda(x_j)) \quad \text{for } t \geq 0, \tag{5.16}$$

The general definition of the cumulative intensity function for an arbitrary distribution function F (continuous or discrete or a mixture of both)

is given by the (Lebesgue-Stieltjes) integral:

$$\Lambda(t) = \int_0^t \frac{1}{S_X(u_-)} dF(u), t \geq 0. \quad (5.17)$$

According to the Doléans-Dade exponential formula (1970), any survival function S_X has the following representation (5.18). Due to its importance, it will be called the inversion formula.

5.4.1 The inversion formula

Given $\Lambda(u)$, the unique solution of (5.17) is

$$S_X(t) = \exp\{-\Lambda^c(t)\} \prod_{u \leq t} [1 - \Delta\Lambda(u)], t \geq 0 \quad (5.18)$$

where $\Lambda^c(t) = \Lambda(t) - \sum_{u \leq t} \Delta\Lambda(u)$ is the continuous part of $\Lambda(t)$. See, e.g., Shorack and Wellner (1986).

The formula (5.18) is easy to deduce if S_X is either a step function or a continuous function. The proof is more involved if S_X has both continuous and discrete components. If $S_X(t)$ is a step function, (5.18) reduces to (5.16). If $S_X(t)$ is continuous, $\Lambda(t) = -\log S_X(t)$. Furthermore, if $S_X(t)$ is differentiable, the derivative $\lambda(t) = \frac{d}{dt} \Lambda(t)$ is the hazard rate which, of course, is no longer a conditional probability.

5.4.2 Calculation of the cumulative intensity function of X from the right censoring or random truncation model

The calculation of the cumulative intensity function is for estimation purposes. There are various methods of estimating F . The particular approach we take unifies the construction of the estimates and proof of consistency for both right censoring and random truncation models. The approach is to first solve the model identification problem. In the case of right censoring, it is to solve eqs (5.1) and (5.2) for F or G with the given $Q(t, 1)$ and $Q(t, 0)$. In the case of random truncation, it is to solve eq. (5.3) for F or G with the given $H(x, y)$.

It turns out that it is easy to solve these equations for the cumulative intensity function $\Lambda_X(t)$ instead of F or the survival function $S_X(t)$. To avoid ambiguity, here and in the sequel a subscript X is attached to Λ to clearly indicate it is the cumulative intensity of X .

For the right censored model (5.1)-(5.2), it can be easily deduced that

$$\Lambda_X(t) = \int_0^t S_Z^{-1}(u_-) dQ(u, 1), \quad \text{for } 0 \leq t < \beta_z \quad (5.19)$$

where β_z is the upper limit of the distribution of Z , the minimum of X and Y . The observation range of X is the interval $(0, \beta_z)$.

Once $\Lambda_X(t)$ is identified, the survival function $S_X(t)$ for t in the interval $(0, \beta_z)$ is obtained immediately by invoking the inversion formula (5.18).

In exactly the same way, $\Lambda_X(t)$ can be calculated from the random truncation model (5.3). If $[a_F, b_F]$ and $[a_G, b_G]$ denote the original ranges of X and Y , under random truncation the observation of X is limited to the range $[a_G, b_F]$. Thus, only the conditional distributions $F_0(x) = P[X \leq x | X \geq a_G]$ and $G_0(x) = P[Y \leq x | Y \leq b_F]$ can be estimated. For simplicity, we assume that $a_G \leq a_F, b_G \leq b_F$, then $F_0 = F, G_0 = G$. Furthermore, to simplify the notation we set $a_F = 0$. Then,

$$\Lambda_X(t) = \int_0^t \frac{dF^*(s)}{R(s)} \quad \text{for } 0 \leq t < b_F, \quad (5.20)$$

where F^* and R are defined by (5.5) and (5.7). See Woodrooffe (1985).

It is interesting to note that although the fraction α defined by (5.4) is unknown, it gets canceled out in the integral because both F^* and R have α as denominator. Therefore Λ_X and hence F is still identifiable in the random truncation model. Likewise we can identify G . It follows that we can also determine α ,

$$\alpha = P[X \geq Y] = \int G(s)dF(s).$$

The above procedure can also be used to calculate the cumulative intensity from the censoring-truncation model (5.8). However, to our knowledge, no explicit formulas for $\Lambda_X(t)$ and F have been calculated for the double censoring model. The identifiability of F (under some mild conditions) was proved by Chang and Yang (1987) using a different method.

5.5 Construction of estimators for F

The identifiability result can be utilized to construct estimates for F (and G). For this we need the empirical distributions. If Z_j, δ_j for $j = 1, \dots, n$ is our right-censored sample, then the empirical distribution $Q_n(t, 1)$ of $Q(t, 1)$ and the empirical survival function of Z , $S_{Z,n}$ are:

$$Q_n(t, 1) = \frac{\sum_{j=1}^n I[Z_j \leq t, \delta_j = 1]}{n} \quad (5.21)$$

$$S_{Z,n} = \frac{\sum_{j=1}^n I[Z_j \leq t]}{n}, \quad (5.22)$$

where $I[A]$ denotes the indicator of the event A . Replacing $Q(t, 1), S_Z$ in eq. (5.19) by the corresponding empirical distributions yields immediately

a nonparametric estimate $\Lambda_{X,n}$ of the cumulative hazard function Λ_X :

$$\Lambda_{X,n}(t) = \int_0^t S_{Z,n}^{-1}(u_-) dQ_n(u, 1). \quad (5.23)$$

By invoking the inversion formula an estimate \hat{F}_n for F can be obtained. Since the empirical distributions, (5.21) and (5.22), are step functions, the estimate \hat{F}_n is a step function having only the “product-limit” part as given by

$$\hat{S}_{X,n}(t) = 1 - \hat{F}_n(t) = \prod_{k=1}^n \left[1 - \frac{\delta_{[k]}}{n-k+1} \right]^{I[Z_{(k)} \leq t]} \quad \text{for } 0 \leq t < \infty \quad (5.24)$$

where $Z_{(1)} \leq Z_{(2)} \leq \dots \leq Z_{(n)}$ denote the ordered values of the Z and $\delta_{[k]}$ the concomitant of $Z_{(k)}$, i.e. $\delta_{[k]} = \delta_i$ if $Z_{(k)} = Z_i$. The ties are ordered arbitrarily among themselves. This formula is self-adjusted for tied $Z_{(j)}$. The estimator $\hat{S}_n(t)$ is a right continuous decreasing step function, and is strictly positive on $[Z_{(n)}, \infty)$ if the largest observation is censored, $\delta_{[n]} = 0$. Some authors set $\hat{S}_n(t)$ equal to 0 on $[Z_{(n)}, \infty)$ regardless of whether $\delta_{[n]} = 0$ or 1.

This estimator is well known. It was derived by Kaplan-Meier (1958) using the Maximum likelihood principle. One can also use the self-consistent criterion, see Efron (1967), Turnbull (1976).

In a similar way, we replace F^* , R in (5.20) by the empirical distributions F_n^* , R_n computed from a randomly truncated sample, (U_j, V_j) , $j = 1, \dots, n$, to arrive at an estimate $\tilde{\Lambda}_{X,n}$ of the cumulative intensity function of X as

$$\tilde{\Lambda}_{X,n}(t) = \int_0^t \frac{dF_n^*(s)}{R_n(s)} = \sum_{U_j \leq t} \frac{d_j}{n R_n(U_j)} \quad \text{for } 0 \leq t \leq b_F, \quad (5.25)$$

where d_j is the number of observations in the U -sample having value U_j and

$$R_n(s) = \frac{\sum_{j=1}^n I[V_j \leq s, U_j \geq s]}{n} = \frac{\sum_{j=1}^n I[V_j \leq s \leq U_j]}{n}. \quad (5.26)$$

Applying the inversion formula yields a product-limit estimate for F as

$$\tilde{F}_n(t) = 1 - \prod_{s \leq t}^n \left[1 - \frac{\Delta F_n^*(s)}{R_n(s)} \right]. \quad (5.27)$$

This is the estimate derived by Lynden-Bell (1971) from the Maximum likelihood principle and the formula is self-adjusted to tied observations.

All these estimates have a product-limit form, since by the inversion formula any discrete distribution can be put into that form. A major difference

between the two estimators lies in the denominators. In the truncation estimator the denominator $R_n(s)$ is random and not a monotone function of s . On the contrary, in the censoring estimator the comparable term is $(n - k + 1)$ which is nonrandom and a monotonically decreasing function of k . Evidently the analytical treatment of the two can be quite different. In particular the variances of both estimators have very different behavior as to be seen in the next section.

The same technique can be used to construct estimate for F with censored-truncated data, see e.g. He and Yang (1992). But it does not seem to work with doubly censored data for which estimates are usually obtained by the self-consistent criterion. There, a self-consistent estimate need not be the nonparametric maximum likelihood estimate. See, e.g. Tsai and Crowley (1985), Gu and Zhang (1993).

5.6 Distributional properties of the estimates

There is a rich literature on asymptotic properties of the KM estimator. These results are well known and readily available in numerous books of all level on survival analysis. The results on the LB estimator are less complete. we discuss some of them below.

5.6.1 The KM estimator

As the sample size $n \rightarrow \infty$, the error in the KM estimate tends to a normal distribution, that is, $\sqrt{n}(\hat{F}_n(t) - F(t))$ tends to the normal distribution $N(0, V(t))$ with variance

$$V(t) = S_X^2(t) \int_0^t \frac{d\Lambda(u)}{S_Z(u-)(1 - \Delta\Lambda(u))}. \quad (5.28)$$

both for a fixed t such that $S_Z(t) > 0$, and as a random function of t .

The asymptotic variance $V(t)$ may be estimated by

$$\hat{V}(t) = \left(\hat{S}_{X,n}(t) \right)^2 \sum_j \frac{\Delta\Lambda_{X,n}(Z_{(j)})}{1 - \Delta\Lambda_{X,n}(Z_{(j)})}, \quad (5.29)$$

where $\hat{S}_{X,n}(t) = 1 - \hat{F}_n(t)$. The sum extends to all $Z_{(j)} \leq t$ where $Z_{(j)}$ is the j th smallest value in the sample $\{Z_1, \dots, Z_n\}$. This is equivalent to the classical Greenwood formula (1926). The approximated $(1 - \alpha)\%$ confidence intervals have confidence limits

$$\hat{S}_{X,n}(t) \pm c_\alpha \left(\hat{V}(t) \right)^{1/2},$$

where c_α is the $(\alpha/2)$ th percentile of the standard normal distribution.

The asymptotic efficiency of the KM estimator has been established by Wellner (1982).

In theory the asymptotic normality holds for all t in the interval $(0, \beta_z)$ but the normal approximation can be very poor when t is close to the upper boundary β_z , the heavy censoring region. In that case, Poisson approximation may work better, see, Wellner (1985).

There are estimators of F and its variance that are minor modifications of the \hat{F}_n and $\hat{V}(t)$. They do not affect the asymptotic normality but might either improve finite sample behavior under special circumstances or be more convenient to compute. For instance some authors change $n - k + 1$ to $n - k + 2$ in the denominator to avoid a zero denominator.

Finite sample behavior of the KM estimator has been difficult to study. In a finite sample setting, Akritas (1986) used bootstrap methods to obtain variance estimates and confidence bands. The results applies to both discrete and continuous distributions F and G . Asymptotic confidence bands were constructed by Hall & Wellner (1980) for continuous F and G .

Recently, Chang (1991) obtained, under the continuity assumption of $F(t)$ and $G(t)$, finite sample formulas for the 2nd, 3rd and the fourth moments of the KM estimator with accuracy of order $O(n^{-2})$. Stute and Wang (1993) showed, under very weak conditions, that the bias $b(t) = E\hat{F}_n(t) - F(t) \leq 0$ for any finite sample size n and any fixed $t < \beta_z$, the upper boundary of Z ; and $E\hat{F}_n(t)$ converges upwards to $F(t)$.

5.6.2 The LB estimator and comparison of asymptotic variances

Little is known about the finite sample behavior of the LB estimator, $\tilde{F}_n(t)$.

Asymptotically, a rather stringent condition is needed to ensure the finiteness of the asymptotic variance of $\sqrt{n}(\tilde{F}_n(t) - F(t))$ for $a_F < t < b_F$. The condition is

$$\int_{a_G}^{\infty} \frac{dF(u)}{G(u)} < \infty. \quad (5.30)$$

The following are some examples:

1. If $a_G < a_F$ so that $G(a_F) > 0$, then (5.30) holds.
2. If F and G have the same continuous distribution, then (5.30) does not hold.
3. If $G = F^\theta$ and F is continuous, then

$$\int_{a_G}^{\infty} \frac{dF(u)}{G(u)} \quad \left\{ \begin{array}{ll} < \infty & \text{for } \theta < 1 \\ = \infty & \text{for } \theta \geq 1 \end{array} \right.$$

The estimation problem under the additional assumption (5.30) is no longer completely nonparametric. Furthermore, this condition is difficult to verify in applications, since F and G are unknown and that is why we

have to estimate them. However, it is interesting to note that this condition is not needed for establishing the uniform consistency property of $\tilde{F}_n(t)$ but for asymptotic normality.

Under condition (5.30) and the continuity of F and G , Woodroffe (1985) proved that $\sqrt{n}(\tilde{F}_n(t) - F(t))$ converges weakly to a zero mean Gaussian process. Its asymptotic variance is given by

$$\tilde{V}(t) = (S_X(t))^2 \int_{a_F}^t \frac{dF^*(s)}{R^2(s)}. \quad (5.31)$$

An estimate of $\tilde{V}(t)$ is

$$\tilde{V}_n(t) = (\tilde{S}_{X,n}(t))^2 \int_{a_F}^t \frac{d\tilde{F}_n(s)}{R_n^2(s)} = (\tilde{S}_{X,n}(t))^2 \sum_{U_j \leq t} \frac{d_j}{R_n^2(U_j)}, \quad (5.32)$$

where d_j is defined as in (5.25) and $\tilde{S}_{X,n}(t) = 1 - \tilde{F}_n(t)$. We could use

$$\tilde{F}_n(t) \pm c_\alpha \left(\tilde{V}_n(t) \right)^{1/2} \quad (5.33)$$

as confidence intervals for F , where c_α is the confidence limit defined earlier. We are not aware of any published results on confidence bands for F .

The asymptotic efficiency of the LB estimators of F and G has been established by van der Vaart (1991) under conditions (5.30) and (5.35) and the continuity of F and G .

The following results summarize the effect of censoring and truncation on the asymptotic variances, σ^2 ,

$$\begin{aligned} \sqrt{n}(F_n(t) - F(t)) &\approx N(0, \sigma^2 = F(t)(1 - F(t))) \quad \text{for all } t, \\ \sqrt{n}(\hat{F}_n(t) - F(t)) &\approx N(0, \sigma^2 = V(t) < \infty) \\ &\quad \text{for any } t \text{ satisfying } S_Z(t) > 0, \\ \sqrt{n}(\tilde{F}_n(t) - F(t)) &\approx N(0, \sigma^2 \text{ may be infinite}), \end{aligned}$$

where $F_n(t)$ is the empirical distribution function of a complete X-sample (i.e. no censoring or truncation), $V(t)$ is the variance of the KM estimate given by (5.28). The asymptotic variance becomes progressively worse. It is bounded by $1/4$ in a complete sample, finite in a right-censored sample, and may be infinite in a randomly truncated sample.

We have for convenience discussed exclusively only the estimation of F . However, since we have used Y as the absolute magnitude (see(5.12)) in section 3, the luminosity distribution is therefore G . For completeness, we give below the LB estimate for G ,

$$\tilde{G}_n(t) = \prod_{s>t}^n \left[1 - \frac{\Delta G_n^*(s)}{R_n(s)} \right], \quad (5.34)$$

where $G_n^*(s)$ is the empirical distribution function of the V-sample.

Parallel to (5.30), the following condition

$$\int_{-\infty}^{b_F} \frac{dG(u)}{1 - F(u)} < \infty \quad (5.35)$$

is necessary for the finiteness of the asymptotic variance of

$$\sqrt{n}(\tilde{G}_n(t) - G(t)) \quad \text{for } a_G < t < b_F. \quad (5.36)$$

Under (5.35), $\sqrt{n}(\tilde{G}_n(t) - G(t))$ is asymptotically normal with mean zero and variance

$$G^2(t) \int_t^{b_G} \frac{dG(s)}{R^2(s)}.$$

With the estimates \tilde{F}_n and \tilde{G}_n the probability of truncation $\alpha = P[X \geq Y]$ (subject to proper boundary conditions) can be obviously estimated by

$$\alpha_n = \int \tilde{G}_n(s) d\tilde{F}_n(s). \quad (5.37)$$

This estimator has a simpler representation, namely choose any t such that $R_n(t) > 0$,

$$\alpha_n = \frac{\tilde{G}_n(t)\tilde{S}_{X,n}(t-)}{R_n(t)}. \quad (5.38)$$

The errors $\sqrt{n}(\alpha_n - \alpha)$ have an asymptotic normal distribution with mean zero and variance

$$\alpha^2 \left\{ \int_{a_F}^t \frac{dF^*(s)}{R^2(s)} + \int_t^{b_G} \frac{dG^*(s)}{R^2(s)} - \frac{1}{R(t)} + 2\alpha - 1 \right\}. \quad (5.39)$$

Different methods of investigation can be found in Chao (1987), Keiding and Gill (1990), and He and Yang (1996). Again, finiteness of the asymptotic variance of $\sqrt{n}(\alpha_n - \alpha)$ requires that both (5.30) and (5.35) hold.

The continuity restriction used in the proof of asymptotic normality and consistency of \tilde{F}_n and α_n has been removed in He and Yang (1996). Thus these results are applicable to discrete distributions F and G , and to the grouped or binned data as well.

Finally we conclude this section with a remark that conditions (5.30) and (5.35) can be avoided if we estimate $F(t)$ not in its entire range but limited to, say, $t \geq \epsilon$ where ϵ is an arbitrarily selected number from (a_G, b_F) . According to Example 1, condition (5.30) is automatically satisfied if $a_G < a_F$, thus we only need to discuss the case $a_G = a_F$. In this case, we can estimate the conditional survival distribution $\frac{S_X(t)}{S_X(\epsilon)}$ by $\frac{\tilde{S}_{X,n}(t)}{\tilde{S}_{X,n}(\epsilon)}$ for $t \geq \epsilon$. Then the

errors $\sqrt{n} \left[\frac{\tilde{S}_{X,n}(t)}{\tilde{S}_{X,n}(\epsilon)} - \frac{S_X(t)}{S_X(\epsilon)} \right]$ have an asymptotic normal distribution with mean zero and variance

$$\left(\frac{S_X(t)}{S_X(\epsilon)} \right)^2 \int_{\epsilon}^t \frac{dF^*(s)}{R^2(s)}.$$

This is finite. The integral is the same as that given in (5.31) except for the lower limit. This says that a fully nonparametric inference can only be achieved on a restricted range, $t \geq \epsilon$. Similar comments apply to the estimation of G . In astronomical applications, it may well be that conditions $a_G < a_F$ and $b_G < b_F$ are justifiable. Thus (5.30) and (5.35) are satisfied.

5.6.3 Sample moments and quantiles

The convergence and distributional properties of the sample moments and sample quantiles computed with the KM estimator and the LB estimator extend the classical results of the empirical cumulative distribution function of a complete sample. However, the observable range of X matters. Under censoring, X is observable in the interval $[0, \beta_z]$. Under random truncation, X is observable in the interval $[a_G, b_F]$.

In the case of a complete sample of n iid observations, $\{X_1, \dots, X_n\}$, we can write the k th sample moment in an integral form as follows:

$$\frac{\sum_{j=1}^n X_j^k}{n} = \int_0^\infty x^k dF_n(x) \quad (5.40)$$

where $F_n(x)$ is the empirical distribution function of the sample. The classical results of the convergence of the sample moments to the corresponding population moments (the Strong Law of Large Numbers)

$$\int_0^\infty x^k dF_n(x) \longrightarrow \int_0^\infty x^k dF(x) \quad (5.41)$$

have only recently been proved for the right censored estimator \hat{F}_n (Stute and Wang (1993)) and for the truncation estimator \tilde{F}_n (He and Yang (1995)).

The limit in (5.41) will not be the k th population moment of F unless the upper limit of integration β_z equals to the upper limit of X . Similarly comments applied to \tilde{F}_n .

In another direction, both \hat{F}_n and \tilde{F}_n have been used to construct estimates for population quantiles. If t is the p th quantile of the distribution F i.e. $F(t) = p$, (for simplicity we assume t is the unique p th quantile), we denote t by the inverse $F^{-1}(p)$. Then the p th sample quantile computed from \tilde{F}_n is the inverse $(\tilde{F}_n)^{-1}(p)$. The following result have been obtained by Gürler, Stute and Wang (1993).

$$\sqrt{n}[(\tilde{F}_n)^{-1}(p) - F^{-1}(p)] \longrightarrow N(0, \sigma^2) \quad (5.42)$$

where $\sigma^2 = \tilde{V}(t)(F^{-1}(p))[f(F^{-1}(p))]^{-2}$, and $\tilde{V}(t)$ is given by (5.31).

For the sample quantile computed from the KM estimate, see the monograph by Csörgö (1983) and references therein.

5.7 Concluding remarks

We have discussed only very briefly some of the distributional results of the two fundamental estimators used in the analysis of incomplete data.

For recent developments on two-sample problems of testing hypotheses, see Neuhaus (1993) which contains a nice summary of the chronological development of the problem.

In the double censoring case, the asymptotic normality has been obtained by Chang (1990) for continuous distributions of X, Y, Z , and generalized to arbitrary distributions by Gu and Zhang (1993). The asymptotic variance formula is very complicated. Zhan and Wellner (1995) have developed a highly efficient computational algorithm which facilitates the construction of confidence bands by bootstrapping.

The two dimensional extension of the KM estimator is not without difficulties. Some of the better known bivariate estimates may assume negative values for any finite sample size and, therefore, cannot be a proper survival distribution. The reader is referred to Pruitt (1991, 1993) for further discussion.

Adding to an already large literature on survival analysis are some new books on the subject by Crowder *et al.* (1991), Andersen *et al.* (1993), Gill (1994).

REFERENCES

- [1] Akritas, M. G.(1986). Bootstrapping the Kaplan-Meier estimator. *J. Amer. Statist. Assoc.* 81 1032-1038.
- [2] Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer, New York.
- [3] Bhattacharya, P. K., Chernoff, H. and Yang. S. S. (1983), Nonparametric estimation of the slope of a truncated regression, *Ann. Statist.* v.11, 505-514.
- [4] Bhattacharya, P. K. (1983), Justification for a K -S type test for the slope of a truncated regression, *Ann. Statist.* v.11, 697-701.
- [5] Chang, M. N. and Yang, G. L. (1987). Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *Ann. Statist.* 15, 1536-1547.
- [6] Chang, M. N. (1990) Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *Ann. Statist.* v.18, 391-404.
- [7] Chang, M. N. (1991). Moments of the Kaplan-Meier estimator. *Sankhya Ser. A* 53 27-50.
- [8] Chao, M.-T.(1987). Influence curves for randomly truncated data, *Biometrika* 74, 426-429.

- [9] Crowder, M. J., Kimber, A. C. Smith, R. L. and Sweeting, T. J. (1991) *Statistical Analysis for Reliability Data*, Chapman and Hall.
- [10] Csörgő, M. (1983). *Quantile Processes with Statistical Applications*. SIAM, Philadelphia.
- [11] Doléans-Dade, C (1970). Quelques applications de la formule de changement de variables pour les semimartingales. *Z. Wahrsch. verw. Gebiete* 16 181-194.
- [12] Efron, B. (1967). The two sample problem with censored data. in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, v. 4, 831-853, eds. Le Cam L. and Neyman, J., Univ. of California at Berkeley Press.
- [13] Feigelson, E.D. and Nelson, P. I. (1985) Statistical methods for astronomical data with upper limits. I. Univariate distributions. *Astrophysical J.* v. 293, 192 - 206.
- [14] Feigelson, E.D. and Babu, G.J. eds. (1992). *Statistical Challenges in Modern Astronomy*, Springer Verlag, New York.
- [15] Gill, R (1994). Lectures on Survival Analysis. Ecole d'Eté de Probabilités de Saint Flour XXII-1992, p. 115-242, ed. P. Bernard, *Lecture Notes in Mathematics*, no. 1581, Springer, New York.
- [16] Gu, M. G. and Zhang, C. H. (1993). Asymptotic properties of self-consistent estimators based on doubly censored data. *Ann. Statist.* v. 21, 611-624.
- [17] Gürler, Ü., Stute, W. and Wang, J. L. (1993). Weak and strong quantile representations for randomly truncated data with applications. *Statistics & Probability Letters* v. 17, 139-148.
- [18] Hall, W. J. and Wellner, J. A. (1980). Confidence bands for a survival curve from censored data. *Biometrika* 67 133-143.
- [19] He, S. and Yang, G. L. (1993). Estimating a lifetime distribution under different sampling plans. In *Statistical Decision Theory and Related Topics*, V, 73-85, eds. Berger, J. and Gupta, S., Springer, New York.
- [20] He, S. and Yang, G. L. (1995). The strong law in the random truncation model. U.of Maryland, Math. Dept.TR95-05.
- [21] He, S. and Yang, G. L. (1996). Estimation of the truncation probability in the random truncation model. U.of Maryland, Math. Dept.TR96-07.
- [22] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* 53 457-481.
- [23] Lai, T. L. and Ying, Z. (1991), Rank regression methods for left-truncated and censored-data. *Ann. Statist.* v. 19, 531-554.
- [24] Pruitt, Ronald C. (1991). On negative mass assigned by the bivariate Kaplan-Meier estimator. *Ann. Statist.* 19 443-453.
- [25] Pruitt, Ronald C. (1993). Identifiability of bivariate survival curves from censored data. *J. Amer. Statist. Assoc.* 88, 573-579.
- [26] Schmitt, J. H. M. M. (1985). Statistical analysis of astronomical data containing upper bounds: General methods and examples drawn from X-Ray astronomy. *Astrophysical J.* v. 293, 178-191.
- [27] Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.
- [28] Stute, W. and Wang, J.-L. (1993). The strong law under random censorship. *Ann. Statist.* 21, 1591-1607.

- [29] Tsai, W. Y. and Crowley, J. (1985). A large sample study of generalized maximum likelihood estimators from incomplete data via self-consistency. *Ann. Statist.* v. 13, 1317-1334. Correction. *Ann. Statist.* v. 18, 470.
- [30] Tsai, W. Y., Jewell, N. and Wang, M. C. (1987). A note on the PL estimator under right censoring and left truncation. *Biometrika* 76, 883-886.
- [31] Trumpler, R. J. and Weaver, H. F. (1953). *Statistical Astronomy*. Dover publ. New York.
- [32] Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B* 38, 290-295.
- [33] van der Vaart, Aad. (1991). On differentiable functionals. *Ann. Statist.* 19, 178-204.
- [34] Wang, M.-C., Jewell, N. P. and Tsai, W.-Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *Ann. Statist.* 14, 1597-1605.
- [35] Wellner, J. A. (1982). Asymptotic optimality of the product limit estimator. *Ann. Statist.* 10, 595-602.
- [36] Wellner, J. A. (1985). A heavy censoring limit theorem for the product limit estimator. *Ann. Statist.* 13, 150-162.
- [37] Zhan, Y. and Wellner, J. A. (1995). Double censoring: Characterization and computation of the nonparametric maximum likelihood estimator. Technical Rep. no. 292, Dept. of Statistics, Univ. of Washington, Seattle.
- [38] Woodroffe, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.* 13, 163-177.

Discussion by David M. Caditz

Professor Yang has provided a very interesting and insightful talk comparing the Kaplan-Meier (KM) and Lynden-Bell (LB) estimators. These estimators are designed for censored and truncated samples, respectively, both of which are abundant in astronomical studies. In the following, I will quickly describe a few points which I believe to be important to astronomer's and I will provide some examples which illustrate concerns regarding the application of these estimators to astronomical data. These concerns are: 1) the assumption of stochastic independence of the sample distributions, 2) the so called 'large jump' problem for the LB estimator which is related to the instability of the LB variance as described by Dr. Yang, and 3) the effect of observational uncertainties and uncertain censoring and truncation boundaries.

The KM and LB estimators are both of the product limit form which is a consequence of the inversion formula and the discrete treatment of the sample data. As Dr. Yang points out, if the hazard rate, $\Delta\Lambda$, can be estimated from the sample data, this inversion formula can be applied to estimate the distribution function of interest. The difference between KM and LB comes in estimation of $\Delta\Lambda$ as described in Figure 1. For this example, the LB hazard rate at $C = .45$ counts/sec (the location of the vertical line) is determined using all detections within the box labeled LB,

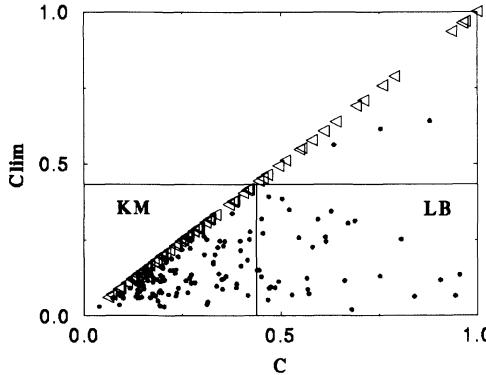


FIGURE 1. QSO detections and upper limits. Count rate, C , vs. minimum detectable count rate, C_{lim} . Hazard rates for Lynden-Bell and Kaplan-Meier distributions are estimated using the sources within the boxes labeled LB and KM, respectively.

while the KM hazard rate is determined using all detections and upper limits within the box labeled KM.

The standard estimators for $\Delta\Lambda$, which make use of the maximum amount of sample data, assume stochastic independence of the underlying (unknown) distributions $F(x)$ and $G(y)$. This is a crucial assumption which is often not true in astronomical data. The QSO luminosity vs. redshift sample shown in Figure 2 is a classic example of dependent distributions as can be seen by comparing the LB luminosity functions for subsets of the sample binned by redshift. The luminosity distribution appears to vary with redshift being, on average, brighter at higher z . It is important, therefore, to test for independence of the underlying distributions. There are good separate discussions of this point by [P92, A92, W92]. [EP94] describe a rank type test for independence. For illustrative purposes I will simply attempt to parameterize the underlying evolution by defining a new 'luminosity' $L' = L/h(z)$ such that $F(L')$ and $G(z)$ are stochastically independent. The function $h(z) = (1+z)^{3.2}$ is found empirically to give a reasonable fit based on comparing the redshift binned distributions of Figure 2, however, I know of no two sample test for LB distributions which could be used to quantify this statement.

For the LB estimator, we estimate $\Delta\Lambda_i$, the hazard rate at the i th data point, as $1/N_i$, where N_i is the number of points in the so called comparable region: $N_i = \{x \geq x_i, y \leq x_i\}$. Unlike the KM hazard rate, the LB hazard rate can be unstable because N_i is random and can be small or even zero for any x_i . An example of this is given in Figure 3 which is a combined sample of several QSO redshift surveys plotted on the L vs. L_{lim} plane where $L_{lim} = 4\pi d^2(z)f_k$, f_k being the flux limit of the k th survey. For the case shown, $N_i = 1$, causing a large step height in $\Phi(L_i)$. Should we

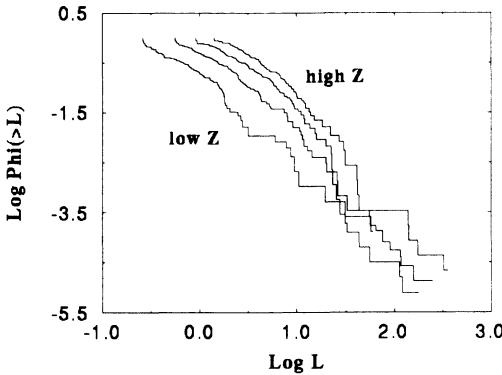


FIGURE 2. QSO luminosity functions for various redshift bins using the LB estimator. Luminosity functions appear to evolve towards lower luminosities at lower redshifts indicating failure of stochastic independence.

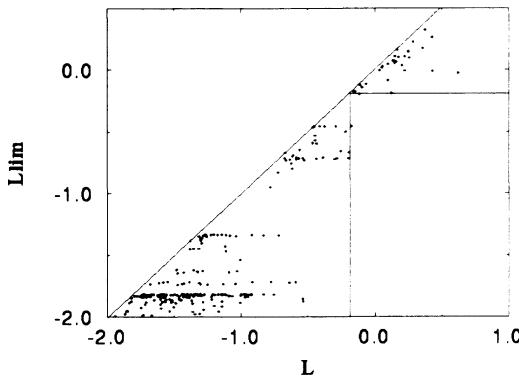


FIGURE 3. The rectangle shows the comparable region used to determine N_i for $L_i = .$ In this case $N_i = 1$ giving a large uncertainty in the LB estimator.

have confidence in the LB estimator in this case? It is easy to imagine that if several samples are drawn from the same underlying distribution, some might have a vanishing N_i 's giving an infinite variance estimate. This issue is addressed in a more rigorous manner by Dr. Yang who has shown that the variance of the LB estimator can be estimated as $\hat{V}(x_i) \propto \sum_{j=1}^i 1/(N_j)^2.$

Consider again the situation depicted in Figure 3. There are several sources very near the boundary of a comparable region. If observational uncertainties are considered, then there is some probability that these sources could have been found within the comparable region, altering the LB estimator. Three types of uncertainties may be important here: 1) uncertainties in the reported values of source fluxes and redshifts which could move ob-

jects into or out of the comparable region of the i th source, 2) uncertainties in source fluxes and redshifts which could alter the position of the i th source and hence its comparable region and associated weight, and 3) uncertainty or fuzziness of the truncation limit which may give an incomplete sampling near the limit and also would cause a fuzziness of the boundaries of the comparable region associated with each source. The first two issues have been addressed by [CP93], [C95] and poster 2 in these proceedings, using source smoothing.

Finally, I would like to remark that because of the different properties of the KM and LB hazard rates, the distributional properties of these estimators may be quite different. The asymptotic behavior of the LB estimator is apparently somewhat problematic. In particular, finiteness of the asymptotic LB variance depends upon the form of the (unknown) marginal distribution functions such that

$$\int \frac{dF(u)}{G(u)} < \infty.$$

The finite sample behaviors of the KM and LB estimators are, apparently, not well understood. Dr. Yang provides a useful summary of recent investigations into the distributional properties of these estimators.

REFERENCES

- [A92] Akritas, M. G. 1992, Statistical Challenges in Modern Astronomy, Feigelson, E.D. and Babu, G. J., eds. Springer-Verlag, New York
- [CP93] Caditz, D. & Petrosian, V. 1993, ApJ, 416, 450
- [C95] Caditz, D. 1995, ApJ, 452, 140
- [EP94] Efron, B. & Petrosian, V. 1994, JASA, 89, 426, 452
- [P92] Petrosian, V. 1992, Statistical Challenges in Modern Astronomy, Feigelson, E.D. and Babu, G. J., eds. Springer-Verlag, New York
- [W92] Woodroffe, M. 1992, Statistical Challenges in Modern Astronomy, Feigelson, E.D. and Babu, G. J., eds. Springer-Verlag, New York

Response by Grace L. Yang

I thank Professor Caditz for providing very interesting astronomical examples demonstrating the use as well as some difficulties in applying the LB estimate, and suggesting important problems for further statistical research.

My comments will be limited to the assumption of independence and Figure 2.

Following the notation used in Section 3 of my paper, let

$$X = \exp[-f(z)] \quad \text{and} \quad Y = \exp[M - m_0] = \frac{L}{t_0}, \quad (5.43)$$

where z is the redshift and M the absolute magnitude of a source. The statistical independence assumption of X and Y is based on the Cosmological

Principle (homogeneity of the Universe) which implies that the absolute M of a source is independent of its redshift z .

In my setup, the distribution function G of Y is the (normalized) luminosity function under consideration. Since the LB estimate uses the X -sample (see eqs. (5.25), (5.26) and (5.34)), it's dependence on redshift z is expected according to the definition of X in (5.1). The value of X varies with the postulated functional form of $f(z)$ where $f(z)$ is used to define the relationship between the absolute magnitude M and the apparent magnitude m as

$$m = M + f(z). \quad (5.44)$$

For instance $f(z) \doteq 5 \log z$ for Hubble's Law and $f(z) = p \log \frac{z}{(1+z)}$ in Segal's Chronometric Theory, where p is some parameter. This does not violate the independence of X and Y . Thus the shift in luminosity functions in Figure 2 need not due to the dependence of M and z but could be due to a particular selection of f .

If however, contrary to the Cosmological Principle, one assumes dependence of M and z , then X and Y would be dependent. Under this circumstance the luminosity function would not be identifiable and we would not be able to decide if the LB estimate is a correct estimate. Statistical tests for independence of X and Y have been considered by Tsai (1990), Efron and Petrosian (1994), and discussed by Woodrooffe, Akritas in the proceedings of SCMA I (1991). Woodrooffe points out that it is hardly possible to test the independence of M and z but that it is possible to test a consequence of the independence, namely the validity of formula (5.3) in my paper. Much work remains to be done in this area, such as studying the properties of the tests.

The suggestion of using $L' = L/h(z)$ is of a different nature. The random variables L' and z would typically be statistically dependent. If such an $h(z)$ can be found to ensure the independence of L' and z , the $h(z)$ would be of a very special form and need not be the one that fits the luminosity functions given in Figure 2.

Indeed, I fully agree with Professor Segal's suggestion (Ch. 4) that the LB estimate should be called the Lyden-Bell-Woodrooffe estimator. The paper by Woodrooffe (1985) is the first one in the statistical literature that systematically and rigorously studied the properties of this estimate. The paper laid the ground work for further developments of the statistical theory of random truncation.

REFERENCES

- [1] Efron and Petrosian (1994). Survival analysis of the gamma-ray burst data, *JASA*, v. 89, no. 426, 452-462.
- [2] Tsai, W-Y. (1990). Testing the assumption of independence of truncation time and failure time, *Biometrika*, v. 77, 169-177.

6

Astronomical (Heteroscedastic) Measurement Errors: Statistical Issues and Problems

Michael G. Akritas ¹

ABSTRACT

A statistical model for astronomical data with measurement errors is described and discussed. Attention is drawn to the distinction between two types of measurement errors according to whether or not the magnitude (variance) of the measurement error depends on the measurement. It is emphasized that when the magnitude of the measurement error does not depend on the measurement, more efficient procedures based on suitable weighting of the observations are possible. However, when the magnitude of the measurement error depends on the measurement, weighting biases the procedure. A method for comparing multivariate data sets, valid for both kinds of measurement error, is described and a variety of other statistical problems are considered and solutions are proposed.

6.1 Introduction

A common feature of many astronomical data sets is the presence of background noise which results in data measured with errors. Examples include color-luminosity relations for field galaxies, relations between X-ray temperatures and velocity dispersions for galaxy clusters, the Tully-Fisher relation and the Tolman test.

Typically, the magnitude (or variance) of this error differs for different observations (i.e. is *heteroscedastic*) and forms part of the data set. That is, different observations in a sample may have considerably different measurement errors due to different observing conditions, exposure times, source brightness, or other conditions. However, the variance of the errors in each observation is determined separately from the measurement of the object of

¹Department of Statistics, The Pennsylvania State University, State College PA 16802, U.S.A. This work was supported in part by NSF grant DMS-9208066.

direct interest. This rarely occurs in the social and biological sciences which have been the focus of statistical applications for many decades, and gives rise to many interesting statistical problems. Astronomers have long been aware of these problems (Eddington 1913) but the procedures currently in use are largely ad hoc and often erroneous.

This paper has two main goals. The first is to present a *statistical model* for data with astronomical (heteroscedastic) measurement errors. This is a necessary guide for developing suitable statistical methods. It is also a useful conceptual tool as it points to the distinction between two different types of measurement error according to whether or not the variance of the measurement error depends (in the statistical sense) on the observation. It will be emphasized that some of the weighted procedures that are currently commonly used in astronomical research can be biased if the variance of the measurement error depends on the observation; see §6.3.2 and §6.4.4 on multiple regression, and polynomial regression. The second goal of the paper is to highlight some of the many interesting statistical problems. In particular, we will discuss the problems of a) comparing multivariate samples, b) polynomial regression, c) estimation of the intrinsic scatter in regression models, d) goodness of fit tests and e) density estimation. A worked out solution or suggested solution is offered for each of these problems; the suggested solutions are under development by two students involved in Thesis work at Penn State.

The paper is organized as follows. In Section 2 we present the statistical model for data with astronomical measurement errors. Section 3 describes a method for comparing multivariate observations. Other approaches to the problems mentioned are described in Section 4.

6.2 A statistical model

We will present the model for bivariate data which is sufficient for describing methods for comparing two (univariate or bivariate) samples and polynomial regression with one independent variable. Multivariate generalizations of this model (which are needed to describe methods for comparing multivariate samples and general polynomial regression) are straightforward. Let the variables of interest be denoted by (X_{1i}, X_{2i}) and the observed data be denoted by

$$(Y_{1i}, Y_{2i}, \mathbf{V}_i), \quad i = 1, \dots, n, \quad (6.1)$$

where for each i , \mathbf{V}_i is a symmetric 2×2 matrix with elements denoted by $V_{11,i}$, $V_{22,i}$, and $V_{12,i}$, for the two diagonal and the common off diagonal elements, respectively. The observed data are related to the unobserved variables of interest by

$$Y_{1i} = X_{1i} + \epsilon_{1i}, \text{ and } Y_{2i} = X_{2i} + \epsilon_{2i}, \quad (6.2)$$

where the errors $(\epsilon_{1i}, \epsilon_{2i})$ have a joint bivariate distribution with zero mean and covariance matrix \mathbf{V}_i , for all i . In this model we allow \mathbf{V}_i to depend on (Y_{1i}, Y_{2i}) and thus, implicitly, on (X_{1i}, X_{2i}) . Thus we do not require that $(\epsilon_{1i}, \epsilon_{2i})$ be independent from (X_{1i}, X_{2i}) . However, \mathbf{V}_i is the only aspect of the distribution of $(\epsilon_{1i}, \epsilon_{2i})$ that depends on (Y_{1i}, Y_{2i}) . In other words, it is assumed that, given \mathbf{V}_i , $(\epsilon_{1i}, \epsilon_{2i})$ is independent from (X_{1i}, X_{2i}) .

The intuitive meaning of the technical assumption that “given \mathbf{V}_i , $(\epsilon_{1i}, \epsilon_{2i})$ is independent from (X_{1i}, X_{2i}) ” is that ϵ_{1i} , for example, is equally likely to be positive or negative no matter what the value of X_{1i} , and the size of its absolute value is governed only by the magnitude of $V_{11,i}$ which is given. All astronomical data sets that we are aware of comply to this assumption.

In most cases, the measurement errors for the two variables are independent (so $V_{12,i} = 0$ for all i), and the observed data are of the form

$$(Y_{1i}, Y_{2i}, V_{1,i}, V_{2,i}), \quad (6.3)$$

with $V_{k,i}$ denoting the variance of ϵ_{ki} , $k = 1, 2$.

Some of the methods we will describe will not require additional assumptions. However, methods based on weighted moments, methods for polynomial regression and those based on deconvolution do require that the error distribution be known. For simplicity, when discussing such methods, it will be assumed that the bivariate distribution of the errors is normal and that the measurement errors in different coordinates are independent (so the data will be as in (6.3)).

We close this section by giving the joint distribution of the observed data under the stated model and under the above assumption about the error distribution. Note first that the observed data in (6.3) are four-dimensional. (Accordingly, when there is interest in only one variable, the observed data are bivariate.) Let $F_1(y, v)$, $f_1(y, v)$, denote the joint cumulative distribution function (cdf), joint probability density function (pdf), when it exists, of $(Y_{1i}, V_{1,i})$, and $G_1(y, v)$, $g_1(y, v)$, denote the joint cdf, joint pdf (when it exists) of $(X_{1i}, V_{1,i})$. Similarly we denote by $F_2(y, v)$, $f_2(y, v)$, $G_2(y, v)$, $g_2(y, v)$, the corresponding quantities from the second sample. It can be seen that

$$F_1(y, v) = \int \int_0^v \Phi\left(\frac{y-x}{u}\right) G_1(dx, du), \quad (6.4)$$

where $\Phi(\phi)$ denotes the cdf (pdf) of the standard normal distribution. The marginal cdf's or pdf's will be denoted by dropping the other argument; thus $G_1(x)$ ($g_1(x)$) denotes the marginal cdf (pdf) of X_1 . When dealing with one sample, the subscript 1 will be omitted.

6.3 Comparing two samples

6.3.1 Methods based on moments

Univariate data

The procedures described in this section are based on the estimation of the first two moments. We consider first two univariate data sets. Thus the data are

$$(Y_{1i}, V_{1.i}), \quad i = 1, \dots, n_1, \quad \text{and} \quad (Y_{2i}, V_{2.i}), \quad i = 1, \dots, n_2, \quad (6.5)$$

where, as described in Section 2, $Y_{ki} = X_{ki} + \epsilon_{ki}$, $k = 1, 2$, and $i = 1, \dots, n_k$. We let G_1, G_2 denote the distribution functions of X_{1i} , respectively X_{2i} . Of interest is the hypothesis that $G_1 = G_2$. Instead of testing the equality of distribution functions, we propose here to test

$$H_{10} : \mu_1 = \mu_2 \quad \text{and} \quad \sigma_1^2 = \sigma_2^2, \quad (6.6)$$

where μ_k, σ_k^2 , is the mean and variance of G_k , $k = 1, 2$. While it is true that equality of the means and variances does not imply equality of the distributions, it is quite likely that the difference between two distributions will manifest itself in the first two moments.

The proposed procedure for testing H_{10} in (6.6) will be based on estimates of the parameters involved. The estimates we will consider are valid under the general model (6.1) - (6.2) which allows for the variance of the measurement error to depend on the measurement. These estimates are based on the relations

$$\mathbb{E}(Y_{ki}) = \mathbb{E}(X_{ki}), \quad \text{Var}(Y_{ki}) = \text{Var}(X_{ki}) + E(V_{k.i}). \quad (6.7)$$

The proof of these results is given in Akritas & Bershady (1996). Relations (6.7) suggest that μ_k and σ_k^2 , $k = 1, 2$, can be estimated by

$$\hat{\mu}_k = \bar{Y}_k, \quad \hat{\sigma}_k^2 = n_k^{-1} \sum_{i=1}^{n_k} (Y_{ki} - \bar{Y}_k)^2 - \bar{V}_k, \quad (6.8)$$

respectively, where $\bar{Y}_k = n_k^{-1} \sum_{i=1}^{n_k} Y_{ki}$ and $\bar{V}_k = n_k^{-1} \sum_{i=1}^{n_k} V_{k.i}$. Now let $N = n_1 + n_2$ and define

$$\mathbf{D}_{1N} = (\hat{\mu}_1 - \hat{\mu}_2, \hat{\sigma}_1^2 - \hat{\sigma}_2^2)' \quad (6.9)$$

where ' denotes the transpose of a vector (so \mathbf{D}_{1N} is a 2×1 vector). The test statistic for H_{10} will be a quadratic form based on \mathbf{D}_{1N} . The matrix of the quadratic form will be the inverse of the pooled sample covariance matrix evaluated from

$$\hat{\boldsymbol{\xi}}_{ki} = (Y_{ki}, \hat{\xi}_{2,ki})', \quad \text{where} \quad \hat{\xi}_{2,ki} = Y_{ki}^2 - 2(\bar{Y}_k)Y_{ki} - V_{k,i}, \quad (6.10)$$

for $k = 1, 2$ and $i = 1, \dots, n_k$. Thus, the test statistic is

$$T_{1N} = N\mathbf{D}_{1N}\mathbf{S}_N^{-1}\mathbf{D}'_{1N}, \quad (6.11)$$

where $S_N = (N/n_1)S_{1,N} + (N/n_2)S_{2,N}$, with $S_{k,N} = n_k^{-1} \sum_{i=1}^{n_k} (\hat{\boldsymbol{\xi}}_{ki} - \bar{\hat{\boldsymbol{\xi}}}_k)(\hat{\boldsymbol{\xi}}_{ki} - \bar{\hat{\boldsymbol{\xi}}}_k)'$, and $\bar{\hat{\boldsymbol{\xi}}}_k = n_k^{-1} \sum_{i=1}^{n_k} \hat{\boldsymbol{\xi}}_{ki}$, $k = 1, 2$. Under the null hypothesis, and as $N \rightarrow \infty$ in such a way that both n_1/N and n_2/N remain bounded away from 0 and 1, it can be shown that the statistic T_{1N} has a central chi-square distribution with 2 degrees of freedom.

Multivariate data

The above method for comparing two univariate populations extends relatively easily to multivariate populations. Recall that the data also consist of the variance-covariance matrix of the measurement error vector, so for a p -dimensional population, the data will consist of the p -dimensional observation in addition to a $p \times p$ variance-covariance matrix. For $k = 1, 2$, denote the two data sets by

$$(\mathbf{Y}_{1i}, \mathbf{V}_{1,i}), \quad i = 1, \dots, n_1, \quad \text{and} \quad (\mathbf{Y}_{2i}, \mathbf{V}_{2,i}), \quad i = 1, \dots, n_2, \quad (6.12)$$

where, as described in Section 2, $\mathbf{Y}_{ki} = \mathbf{X}_{ki} + \boldsymbol{\epsilon}_{ki}$, $k = 1, 2$, and $i = 1, \dots, n_k$. We let G_1, G_2 denote the distribution functions of the p -dimensional vectors \mathbf{X}_{1i} , respectively \mathbf{X}_{2i} . Of interest is the hypothesis that $G_1 = G_2$. As before, instead of testing the equality of distribution functions, we propose here to test

$$H_{20} : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{and} \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2, \quad (6.13)$$

where $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$, is the mean and variance-covariance matrix of G_k , $k = 1, 2$.

The proposed procedure for testing H_{20} in (6.13) will be based on estimates of the parameters involved. The estimates we will consider are valid under the general model (6.1), (6.2) which allows for the variance-covariance matrix of the measurement error to depend on the measurement. In addition to the relations (6.7), these estimates are also based on

$$\text{Cov}(Y_{kj_1i}, Y_{kj_2i}) = \text{Cov}(X_{kj_1i}, X_{kj_2i}) + E(V_{k(j_1j_2),i}), \quad (6.14)$$

where, Y_{kji} (X_{kji}) denotes the j -th element of \mathbf{Y}_{ki} (\mathbf{X}_{ki}) and $V_{k(j_1j_2),i}$ denotes the (j_1j_2) element of $\mathbf{V}_{k,i}$. The proof of this can also be found in Akritas & Bershady (1996).

Relations (6.7) and (6.14) imply that $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ can be estimated by

$$\hat{\boldsymbol{\mu}}_k = \bar{\mathbf{Y}}_k, \quad \hat{\boldsymbol{\Sigma}}_k = n_k^{-1} \sum_{i=1}^{n_k} (\mathbf{Y}_{ki} - \bar{\mathbf{Y}}_k)(\mathbf{Y}_{ki} - \bar{\mathbf{Y}}_k)' - \bar{\mathbf{V}}_k, \quad (6.15)$$

where $\bar{\mathbf{Y}}_k = n_k^{-1} \sum_{i=1}^{n_k} \mathbf{Y}_{ki}$, $\bar{\mathbf{V}}_k = n_k^{-1} \sum_{i=1}^{n_k} \mathbf{V}_{ki}$. Let $N = n_1 + n_2$ and define

$$\begin{aligned} \mathbf{D}_{2N} &= ((\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)', (\hat{\sigma}_{1,11} - \hat{\sigma}_{2,11}, \dots, \hat{\sigma}_{1,pp} - \hat{\sigma}_{2,pp}), \\ &\quad (\hat{\sigma}_{1,jj'} - \hat{\sigma}_{2,jj'}, j < j', j, j' = 1, \dots, p))' \end{aligned} \quad (6.16)$$

where $\hat{\sigma}_{k,jj'}$ is the (j, j') element of $\hat{\Sigma}_k$. Then \mathbf{D}_{2N} is a $(2p + p(p-1)/2)$ -dimensional vector. As in the univariate case, the test statistic for H_{20} will be a quadratic form based on of \mathbf{D}_{2N} . The matrix of the quadratic form will be the inverse of the pooled variance-covariance matrix evaluated from

$$\hat{\boldsymbol{\xi}}_{ki} = ((\mathbf{Y}_{ki})', (\hat{\boldsymbol{\xi}}_{2,ki})', (\hat{\boldsymbol{\xi}}_{3,ki})')'. \quad (6.17)$$

where $\hat{\boldsymbol{\xi}}_{2,ki}$ is the p -dimensional vector with components $Y_{kji}^2 - 2(\bar{Y}_{kj})Y_{kji} - v_{kjj'i}$ and $\hat{\boldsymbol{\xi}}_{3,ki}$ is the $p(p-1)/2$ dimensional vector with elements $Y_{kji}Y_{kj'i} - (\bar{Y}_{kj})Y_{kj'i} - (\bar{Y}_{kj'})Y_{kj'i} - v_{kjj'i}$ for $j < j', j, j' = 1, \dots, p$, with $v_{kjj'i}$ denoting the (jj') element of \mathbf{V}_{ki} . Note that in order to avoid complicating the notation with additional subscripts, we used the notation $\hat{\boldsymbol{\xi}}_{ki}$ which in the univariate context is defined in a different way; similarly, the matrices S_N , $S_{1,N}$ and $S_{2,N}$ below are different from the corresponding matrix in the univariate case. The test statistic for H_{20} will be

$$T_{2N} = N \mathbf{D}_{2N} S_N^{-1} \mathbf{D}'_{2N}. \quad (6.18)$$

Under the null hypothesis, and as $N \rightarrow \infty$ in such a way that both n_1/N and n_2/N remain bounded away from 0 and 1, it can be shown that the statistic T_{2N} has a central chi-square distribution with $2p + p(p-1)/2$ degrees of freedom.

6.3.2 Methods based on weighted moments

The procedures described above apply to all data sets generated according to the statistical model described in §6.2. However, when the variance of the measurement error is independent from the measurement, procedures based on weighted averages are much more efficient.

Population means are commonly estimated in astronomical research by a weighted average of the observations with weights proportional to the inverse of the variance of the measurement error (Bevington & Robinson, 1992). When the variance of the measurement error depends on the measurement, this weighted average may be biased and should be avoided. On the other hand, when the variance of the measurement error is independent from the measurement, the optimal set of weights are proportional to the inverse of the variance of each observation. Depending on the magnitude of the "intrinsic scatter", the optimal set of weights can be very different from the weights recommended in the astronomical literature. From the

second relation in (6.8) it follows that, for estimating the first moment of population k ($k = 1, 2$), the weight assigned to observation Y_{ki} should be proportional to $(\hat{\sigma}_k^2 + V_{ki})^{-1}$.

Estimation of the variance by a weighted average requires knowledge of the fourth moment of the measurement error. Such knowledge is at hand under the additional assumption that the measurement errors are normal.

Testing procedures based on weighted moment estimators will be described in detail in a forthcoming paper.

6.3.3 Other methods

It is also possible to develop extensions of the Wilcoxon-Mann-Whitney test in the present setting with measurement errors. Such extensions would be based on successful estimation of the functional $\int G_1(x)dG_2(x)$. To see the relevance of this quantity, note that, up to a centering constant, this is the same as $\int H(x)d(G_1(x) - G_2(x))$, where $H(x) = p_1G_1(x) + p_2G_2(x)$, for any p_1, p_2 such that $p_1 + p_2 = 1$. Therefore it can be estimated as in §6.4.3. From the last expression, however, it is seen that the important aspect of the Wilcoxon-Mann-Whitney statistic is that it 'contrasts' G_1 with G_2 , while the function H merely specifies the type of contrast. Meaningful contrast functions other than the H given above can also be chosen.

6.4 Deconvolution methods

6.4.1 Background

Suppose that, due to measurement error the random variable of interest, X , cannot be observed. Instead the observable random variable is $Y = X + \epsilon$, where the measurement error ϵ is assumed to be independent from X . Let G denote the distribution function of X and Φ denote the distribution function of ϵ . Thus, the distribution function of Y is the convolution between G and Φ , $F(y) = G \star \Phi(y)$. Obtaining G from F is known as the deconvolution problem, which is a special case of the mixture problem. This line of research was initiated in the context of the nonparametric empirical Bayes problem (Robbins, 1950 and 1956), while Kiefer & Wolfowitz, 1956) is the earliest work involving maximum likelihood methods. For other approaches and a description of the state of the art, see the books Lindsay (1995), McLachlan & Basford (1988), and Titterington, Smith & Makov (1985).

6.4.2 Nonparametric density estimation and regression

It is quite remarkable that Fourier methods for deconvolution have never been applied to the measurement errors problem by astronomers. In fact,

the use of Fourier methods for deconvolution has a long history in astronomy. Applications include line profile analysis to measure stellar rotation and turbulence (Carroll, 1933), image formation and restoration in radio interferometry (Cornwell & Perley 1991), speckle interferometry to reduce atmospheric and telescopic blurring of images (Labeyrie, 1970), and other problems. On the contrary, the statistical literature is rich with applications of Fourier methods in the traditional setting of homoscedastic measurement errors; see Stefanski & Carroll (1990), Carroll & Hall (1988), Liu & Taylor (1989), Zhang (1990), Fan (1991), and Diggle & Hall (1993). See §6.4.3 for further references on deconvolution methods that are not specific to density estimation.

Nonparametric Density Estimation

In this paragraph we focus attention to the approach in Stefanski & Carroll (1990), and briefly indicate how their main ideas apply in our heteroscedastic measurement error context. For simplicity we will illustrate the univariate case only.

For any density function f , let γ_f denote the characteristic function of f . Also, let f^* denote the density (if it exists) of the observable random vector $(Y/V, V)$; thus $f^*(y, v) = \int \phi(y - \frac{x}{v})g(x, v)dx$ (see notation in §6.2). Then,

$$\gamma_{f^*}(t, s) = \gamma_\phi(t) \int \int e^{i(tx/v + sv)} g(x, v) dx dv = \gamma_\phi(t) \gamma_{g^*}(t, s). \quad (6.19)$$

where $g^*(w, v)$ is the pdf of (W, V) , where $W = X/V$. Next let $\hat{f}^*(y, v) = (n\lambda^2)^{-1} \sum_{j=1}^n K(\lambda^{-1}(Y_j/V_j - y), \lambda^{-1}(V_j - v))$ be a nonparametric kernel estimator for f^* corresponding to kernel K and bandwidth λ . It then follows that

$$\hat{\gamma}_{f^*} = \gamma_K(-\lambda t, -\lambda s) \hat{\gamma}_{f^*}(t, s). \quad (6.20)$$

where $\hat{\gamma}_{f^*}$ is the empirical characteristic function of $(Y/V, V)$. Relations (6.19) and (6.20) imply

$$\hat{\gamma}_{g^*}(t, s) = \gamma_K(-\lambda t, -\lambda s) [\gamma_\phi(t)]^{-1} \hat{\gamma}_{f^*}(t, s). \quad (6.21)$$

The kernel K can be chosen so $\hat{\gamma}_{g^*}(t, s)$ is integrable over the entire plane and thus the inversion formula (Feller, 1971, p. 524) yields an estimator for the density $g^*(t, s)$. This, in turn provides the needed estimator of $g(x)$.

Nonparametric Regression

Nonparametric regression seeks to estimate a measure of conditional location (e.g. conditional mean value) of the dependent variable given a specified value of the independent variable. What distinguished it from the usual (parametric) regression is that no model (such as linear, polynomial etc.,)

is known to describe how the conditional location changes with the independent variable. Fan & Truong (1993) present an idea for nonparametric regression in the classical (homoscedastic) measurement errors problem. Using the above extension of nonparametric density estimation, Fan's approach can be extended for data with astronomical (heteroscedastic) measurement errors. A related idea is to estimate the conditional density via estimation of the corresponding bivariate and univariate densities. This will yield the desired estimator of conditional location.

6.4.3 Estimation of functionals

Here we outline an approach to the estimation of functionals which, to our knowledge, is new even in the case of known homoscedastic measurement error. For this reason (and for simplicity), we present the main ideas in the classical homoscedastic measurement error case. As a consequence, the notation will be somewhat different (but analogous) to that used up to now.

A New Method for Estimation of Functionals

The proposed estimation method is based on a method for expressing functionals $\int \xi(x)dG(x)$, as corresponding functionals of F (see also Tierney & Lambert, 1984). The main idea is to express a wide enough class of functionals of G as functionals of F . This will provide a basis for either reexpressing a given functional of G as a functional of F , or approximate the given functional of G by a functional of F at any prespecified level of accuracy.

Our approach will be to work with a separating class of functionals (c.f. Breiman, (1968), p. 165) which, in turn can also provide good approximations to other functionals of interest. Thus it is different from other applications of Fourier methods for deconvolution used in statistics (see also §6.4.2). We illustrate this idea with the class of separating functionals given in Breiman (1968), pp. 217-218. In particular, consider the class of functionals $\int h_v(x)dG(x)$ where $h_v(x) = h_0(x)e^{ivx}$, $v \in \mathcal{R}$, $h_0(x) = (\lambda_1 x)^{-2}(\sin \lambda_1 x)^2 + (\lambda_2 x)^{-2}(\sin \lambda_2 x)^2$, for λ_1, λ_2 positive and not rational multiples of each other. The important features of these functions (other than being separating) are that they are everywhere positive and $\tilde{h}_v(u) = \int e^{iux}h_v(x)dx$ vanishes outside a compact set. Thus, if we define

$$\tilde{\psi}_v(u) = \frac{\tilde{h}_v(u)}{\gamma_\phi(-u)}, \quad \psi_v(y) = (2\pi)^{-1} \int e^{-iuy}\tilde{\psi}_v(u)du, \quad (6.22)$$

we can obtain the relation $h_v(x) = \int \psi_v(y)\phi(y-x)dy$ from which it follows that,

$$\int h_v(x)dG(x) = \int \psi_v(y)dF(y). \quad (6.23)$$

At this point it should be noted that both h_v and ψ_v are complex valued functions. Thus relation (6.23) is implies the two relations

$$\int \operatorname{Re}[h_v(x)]dG(x) = \int \operatorname{Re}[\psi_v(y)]dF(y) \quad (6.24)$$

$$\int \operatorname{Im}[h_v(x)]dG(x) = \int \operatorname{Im}[\psi_v(y)]dF(y), \quad (6.25)$$

where Re , Im denote the real, imaginary part of a complex number, respectively. Simply put, this result means that for any h_v , $\int h_v(x)dG(x)$ can be expressed as an integral in terms of F and thus can be estimated from the data as an average. Clearly, this is also possible for any function ξ that can be expressed as a finite linear combination of the real or the imaginary parts of h_v 's.

Next we will estimate an approximation any given integral $\int \xi(x)dG(x)$, where ξ is such that $\int |\xi(x)|dG(x) < \infty$, in terms of a functional in F . The approximation can be achieved to any specified degree of accuracy. The first step to this is to approximate $\int \xi(x)dG(x)$ by $\int \xi(x)b(x)dG(x)$ where $b(x)$ vanishes off a finite interval. This approximation can be done to any degree of accuracy. It is then possible to express

$$\begin{aligned} \int \xi(x)b(x)dG(x) &= \sum_{k=0}^{\infty} \left[c_k \int h_0(x) \cos(2\pi kx) dG(x) \right. \\ &\quad \left. + d_k \int h_0(x) \sin(2\pi kx) dG(x) \right]. \end{aligned} \quad (6.26)$$

where

$$\begin{aligned} c_k &= \int [\xi(x)b(x)/h_0(x)] \cos(2\pi kx) dx \\ d_k &= \int [\xi(x)b(x)/h_0(x)] \sin(2\pi kx) dx. \end{aligned}$$

According to (6.24) and (6.25) it is possible to express each of the integrals in (6.26) as an integral in F which can be estimated. Since the infinite sums can be approximated to any degree of accuracy, we have thus estimated a quantity which approximates $\int \xi(x)dG(x)$ to any degree of accuracy.

6.4.4 Some applications

Multiple Regression.

This problem has received considerable attention. However, as far as we know, the pitfalls of weighting when the variance of the measurement error depends on the measurement has not been pointed out. Thus, when only the dependent variable is measured with error, weighted least squares is

recommended (Bevington & Robinson, 1992). In this case, weighted least squares is valid only when the variance of the measurement error is independent from the measurement and in this case the optimal weights are obtained in Akritas & Bershady (1996).

Until very recently the only available methods for fitting regression models when both variables are subject to measurement errors were based on the assumption that there is no *intrinsic scatter*. Intrinsic scatter refers to the variability about the regression line for the true (i.e. without the measurement errors) variables. Thus, the true points are assumed to lie exactly on a straight line; see for example, the software package ORDPACK (Boggs et al., 1990) and the review papers by Feigelson & Babu (1992) and MacDonald & Thompson (1992). However, this assumption is not met in most practical situations and therefore the procedure is erroneous. In addition, these methods require normal error distribution and, more importantly, they require that the variance of the measurement error is independent from the measurement. This last pitfall has never been pointed out.

Akritas & Bershady (1996) developed a method for simple regression which is based on realistic assumptions, namely, that intrinsic scatter about the regression line exists in addition to heteroscedastic measurement error, the variance of the measurement error is allowed to depend on the measurement, and the error distribution need not be normal. This method is based on the observation that the ordinary least squares estimator is given in terms of the first two moments, and uses the moment estimators described in §6.3. Though not yet developed, extension of this method to multiple regression is possible under the same general assumptions. In addition, when the variance of the measurement errors is independent from the measurements, the more efficient weighted moment estimators can be used instead.

Polynomial Regression.

Polynomial regression is typically dealt within the framework of multiple regression when the covariates are observed without error. However, when the covariates are measured with error, polynomial regression requires special attention. For example, application of the moments method (described above) to polynomial regression requires the estimation of higher order product moments. This is not possible without the additional assumptions that a) the variances of the measurement errors are independent from the measurements and b) the measurement errors have a joint multivariate normal distribution or that they are independent. Thus, the method that applies quite generally for multiple regression, does not accommodate both types of measurement error in polynomial regression.

However, it is possible to use the method developed in Akritas (1996) in conjunction with the nonparametric regression method described in §6.4.2. This method involves replacing the data set by n pairs consisting of spec-

ified values of the covariate and conditional location estimators for the response variable given each of the specified covariate values. The polynomial regression parameters are then estimated by ordinary least squares on these n pairs. Both point and interval estimates obtained by this method in incomplete data settings performed well in simulation studies.

Estimation of the Intrinsic Scatter.

The issue of estimating the intrinsic scatter in a regression model (and how it depends on the covariate) is of independent interest in astronomical research. When the covariate is not contaminated by measurement error, various methods for estimating the intrinsic scatter have been proposed; see Dixon & McKean (1995) and the book by Carroll & Ruppert (1988). However, when the covariate is also contaminated with measurement error, it is not obvious how to use the above methods because the residuals cannot be evaluated.

We propose here methods for estimation of the intrinsic variability. Suppose first that there is no measurement error in the independent variable (see also Akritas & Bershady, 1996). In this case the method of Akritas (1996) (see description above) can be adapted to the present context by simply replacing an estimate of conditional location by an estimate of the conditional variance of the response given the covariate value. The debiased sample variance (see §6.3)) computed from a window around the specified covariate value will be used as estimate of the conditional variance. When there is measurement error in the covariate, the conditional variance of the response given a value of the covariate will be estimated in a way similar to the estimation of conditional location described in §6.4.2.

Goodness-of-Fit Testing

Astronomers are often interested in testing the plausibility of a particular model when the observations are contaminated by measurement error. In the case where V is independent from X (see description of the astronomical measurement error model in §6.2, but for the univariate case) it is fairly simple to develop Pearson-type goodness-of-fit tests. Indeed, consider the hypothesis $H_0 : G = G_0$, where G denotes the distribution function of the (unobserved due to measurement error) variable of interest. Then, relation (6.4) implies that

$$F_0(y) = \int E[\Phi(\frac{y-x}{V})]G_0(dx). \quad (6.27)$$

and H_0 can be tested by testing $H_0^* : F = F_0$. However, it is clear that Pearson's chi-square statistic (or, for that matter, any other procedure) cannot be applied directly since F_0 needs to be estimated by estimating the expected value inside the integral in (6.27). While this estimation can be achieved by averaging over the observed V_i , the theory of Pearson's chi-

square tests covers only estimation of Euclidean parameters and does not include this case – but it can be developed.

In the general case where V depends on X we propose to treat the chi-square statistic as a functional in the sense described in §4.2.

Acknowledgments: The material presented in this paper is the outcome of many stimulating discussions with Professors M. A. Bershady and E. D. Feigelson, and questions submitted to the Statistical Consulting Center for Astronomy by D. Rabinowitz and W. Vacca.

REFERENCES

- [1] Akritas, M. G. (1996). On the use of nonparametric regression techniques for fitting parametric regression models. *Biometrics, In press*.
- [2] Akritas, M. G., & Bershady, M. A. (1996). Linear regression for astronomical data with measurement errors and intrinsic scatter. *The Astrophysical Journal, In press*.
- [3] Bevington, P. R., & Robinson, D. K. (1992). *Data Reduction and Error Analysis for the Physical Sciences*. Second edition, McGraw-Hill, New York.
- [4] Boggs, P. T., Donaldson, J. R., Byrd, R. H., & Schnabel, R. B. (1990). ODRPACK: Software for weighted orthogonal distance regression. *ACM Trans. Math. Software, 15*, 348–364.
- [5] Breiman, L. (1968). *Probability*. Addison-Wesley, Reading.
- [6] Carroll, J. A. (1933). The spectroscopic determination of stellar rotation and its effect on line profiles. *Mon. Not. Royal Astr. Soc., 93*, 478–507.
- [7] Carroll, R. J., & Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association, 83*, 1184–1186.
- [8] Carroll, R. J., & Ruppert, D. (1988). *Transformation and Weighting in Regression* (First edition). Chapman and Hall, New York.
- [9] Cornwell, T. J., & Perley, R. A. e. (1991). *Radio Interferometry, Theory, Techniques, and Applications*. Astro. Soc. Pacific, San Francisco.
- [10] Diggle, P. J., & Hall, P. (1993). A Fourier approach to nonparametric deconvolution of a density estimate. *Journal of the Royal Statistical Society, Ser. B, 55*, 523–531.
- [11] Dixon, S. L., & McKean, J. W. (1995). Rank-based analysis of the heteroscedastic linear model. *The Journal of the American Statistical Association, in press*.
- [12] Eddington, A. S. (1913). On a formula for correcting statistics for the effects of a known probable error of observation. *Mon. Not. Royal Astro. Soc., 73*, 359.
- [13] Fan, J., & Truong, Y. K. (1993). Nonparametric regression with errors in variables. *The Annals of Statistics, 21*, 1900–1925.
- [14] Fan, J. Q. (1991). On the optimal rates of convergence for nonparametric deconvolution problem. *The Annals of Statistics, 19*, 1257–1272.
- [15] Feigelson, E. D., & Babu, G. J. (1992). Linear regression in astronomy. II. *Astrophys. J., 397*, 55–67.

- [16] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Vol. II* (Second edition). Wiley, New York.
- [17] Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27, 886–906.
- [18] Labeyrie, A. (1970). Attainment of diffraction limited resolution in large telescopes by Fourier analyzing speckle patterns in star images. *Astron. Astrophys.*, 6, 85–87.
- [19] Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, California.
- [20] Liu, M. C., & Taylor, R. L. (1989). A consistent nonparametric density estimator for the deconvolution problem. *Canadian Journal of Statistics*, 17, 427–438.
- [21] MacDonald, J. R., & Thompson, W. J. (1992). Least-squares fitting when both variables contain errors: Pitfalls and possibilities. *Amer. J. Physics*, 60, 60–73.
- [22] McLachlan, G. J., & Basford, K. E. (1988). *Mixture Models*. Dekker, New York.
- [23] Robbins, H. (1950). A generalization of the method of maximum likelihood: Estimating a mixing distribution (Abstract). *Annals of Mathematical Statistics*, 21, 314–315.
- [24] Robbins, H. (1956). An empirical Bayes approach to statistics. In J., N. (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1, pp. 157–163. Prentice-Hall, New York.
- [25] Stefanski, L. A., & Carroll, R. A. (1990). Deconvoluting kernel density estimators. *Statistics*, 21, 169–184.
- [26] Tierney, L., & Lambert, D. (1984). Asymptotic efficiency of estimators of functionals of mixed distributions. *The Annals of Statistics*, 12, 1380–1387.
- [27] Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.
- [28] Zhang, C. H. (1990). Fourier methods for estimating mixing densities and distributions. *The Annals of Statistics*, 18, 806–831.

Discussion by William H. Jefferys

There is little that I can add to Prof. Akritas paper. Quoting from his paper, I would like to remind everyone of his essential point:

It is emphasized that when the magnitude of the measurement error does not depend on the observation, more efficient procedures based on suitable weighting of the observations are possible. However, *when the magnitude of the measurement error depends on the observation, weighting biases the procedure* (Emphasis added).

In other words, the *obvious* thing to do can be the *wrong* thing to do. Oftentimes we find people using particular statistical procedures as “black boxes,” without fully understanding them. It is important to understand what you are doing and why. Unthinking use of a particular procedure without understanding its properties is a recipe for disaster.

I also wish to call attention to William Wheaton’s poster paper at this conference, “A Poisson parable: Bias in linear least squares estimation,” (Ch. 41) which presents a simple example of the same phenomenon that Akritas has discussed. Wheaton considers the estimation of an unknown quantity which is measured with Poisson noise. For example, suppose we make k independent measurements of the luminosity L of a star by counting photons, obtaining for the i th observation N_i counts during the integration period t_i to yield an estimated luminosity $\hat{L}_i = N_i/t_i$. What is the correct method of combining the \hat{L}_i to estimate L ? Since the process being observed is Poisson, the variance of an individual observation is proportional to L_i , and one might be tempted to calculate a weighted average of the \hat{L}_i with weights proportional to $1/N_i$; but this would be wrong, and the resulting estimator would be biased and in fact inconsistent. Wheaton shows that an unbiased estimator is given by $\sum N_i / \sum t_i$.

The example that Akritas discusses is somewhat more complex, but related. Here, we are measuring some property of a class of objects which has a natural cosmic scatter. For example, we could be measuring the luminosities of different members of a given class of stars in a cluster or galaxy, and the luminosity L_i of a given star i could be distributed according to some probability density $P(L_i | L, \dots)$.

The individual L_i are unidentified; the best we can do is to estimate the L_i by $\hat{L}_i = N_i/t_i$, where N_i photons are counted during an integration period t_i . Again, Akritas shows that the “obvious” estimator for L obtained by computing a weighted average of the \hat{L}_i , weighting by the variance estimated in the obvious way from the Poisson nature of the photon counting process, yields a biased and even inconsistent estimator of L . Instead, the unweighted average is preferable, as it is manifestly unbiased. (In deriving his estimators, Akritas is following a moment method that has also been applied in astronomy by Deeming [Dee68].)

Since Akritas, Wheaton and I agree on these essential points, I thought it would be useful to use the remainder of my time to discuss an alternative, Bayesian approach to such problems. Thus, we observe counts (N_1, \dots, N_k) for stars $(1, \dots, k)$ in some class. Each star is characterized by its actual luminosity L_i (expected counts/second) and integration time t_i . It follows that N_i follows a Poisson distribution:

$$P(N_i | L_i, t_i) = \frac{(L_i t_i)^{N_i} \exp(-L_i t_i)}{N_i!}$$

Because of cosmic scatter, each individual star in the class may have a different luminosity L_i . Assume, therefore, that the L_i are distributed ac-

cording to some probability density. e.g., we might assume normality:

$$L_i \sim \mathcal{N}(L, \sigma^2)$$

so that

$$P(L_i | L, \sigma) \propto \frac{1}{\sigma} \exp\left(-\frac{(L_i - L)^2}{2\sigma^2}\right)$$

To approach this problem from a Bayesian viewpoint, we will also find it necessary to specify a prior distribution $P(L, \sigma)$ on L and σ . It follows from the definition of conditional probability that the prior distribution on (L_i, L, σ) is given by

$$P(L_i, L, \sigma) = P(L_i | L, \sigma)P(L, \sigma).$$

By Bayes' theorem, therefore, the posterior distribution of (L_i, L, σ) , given the data, is proportional to the prior times the likelihood:

$$\begin{aligned} P(L_i, L, \sigma | N_i) &\propto P(L_i, L, \sigma)P(N_i | L_i, t_i, L, \sigma) \\ &= P(L_i, L, \sigma)P(N_i | L_i, t_i). \end{aligned}$$

Assuming independence, the complete posterior distribution for all observations is just the product of these over i :

$$P(L_1, \dots, L_k, L, \sigma | N_1, \dots, N_k) \propto \prod_i P(L_i, L, \sigma)P(N_i | L_i, t_i, L, \sigma).$$

Everything of interest is to be inferred from the posterior distribution. For example, we are interested in making inferences about L . The standard Bayesian prescription is to marginalize (integrate) with respect to the nuisance variables $(L_1, \dots, L_k, \sigma)$, obtaining a posterior distribution in L alone. Thus

$$P(L | data) \propto \int \dots \int P(L_1, \dots, L_k, L, \sigma | data) dL_1 \dots dL_k d\sigma.$$

Once we have the posterior distribution of L in hand, we can compute Bayesian confidence intervals, posterior means, posterior medians, and so forth, for L .

In actuality, the particular case we have probably can't be integrated in closed form, so some approximate method such as Markov Chain Monte Carlo (MCMC) would have to be used. However, we can consider Wheaton's limiting case, obtained by letting $\sigma \rightarrow 0$, which results in a simplified problem that can be solved in closed form. We use the usual "automatic" (improper) prior

$$P(L) \propto \frac{1}{L},$$

which yields the posterior distribution

$$\begin{aligned} P(L \mid t_i, N_i) &\propto \frac{1}{L} \prod_i (Lt_i)^{N_i} \exp(-Lt_i) \\ &\propto L^{N-1} \exp(-LT) \end{aligned}$$

where

$$T = \sum_i t_i, \quad N = \sum_i N_i$$

are sufficient statistics. The full normalized posterior distribution is therefore

$$P(L \mid data) = \frac{T(LT)^{N-1} \exp(-LT)}{(N-1)!}$$

The procedure at this point would be to derive whatever is desired from the posterior distribution. For example, if we want an estimator for L , we can compute the posterior mean or mode. The posterior mean is

$$\hat{L}_{mean} = \int L P(L \mid data) dL = \frac{N}{T}$$

The mode would be $(N-1)/T$, which is biased but consistent.

The interesting thing about this is that the Bayesian prescription automatically tells us not to use weighted averages, and instead leads us to estimators similar to (and in this simple case even identical to) the estimator advocated by Akritas and Wheaton. In more complex cases such as those with cosmic scatter, however, it is to be expected that the Bayesian estimators would not end up being a simple unweighted average, but would in general be nonlinear and computable only by numerical means. Nonetheless, they may well turn out to be better than the simple unweighted averages advocated by Akritas.

The discussion can be extended to the problem where the background count must also be considered [Lor92]. Whereas a straightforward approach using classical estimators can run into the problem in low signal situations of yielding unphysical negative luminosities, the natural Bayesian solution to the same problem cannot result in such anomalies. At the same time, the Bayesian solution is typically simple and to set up, while at the same time handling the problem of unphysical parameters quite automatically.

Response to Professor Rao

Professor Rao asks about maximizing the likelihood function instead of calculating marginal distributions. A Bayesian would be more likely to ask

about maximizing the posterior distribution, obtaining the so-called maximum a posteriori (MAP) estimate, but in the situation at hand either method may run into difficulty. It is well-known that in situations where some variables are unidentified, like errors-in-variables problems, maximum likelihood gives the wrong answer for the variance—it is off by a factor of two [KS79]. Similarly, in nonlinear errors-in-variables problems, maximum likelihood may similarly produce inconsistent estimators for other interesting parameters [Ful87]. In contrast, the standard Bayesian prescription is to marginalize, that is, to integrate over the unidentified and nuisance variables. This does not result in an inconsistent estimator, if the prior is chosen properly. For a discussion of the Bayesian approach to this problem in the context of linear regression, see [Zel87].

REFERENCES

- [Dee68] Terence J. Deeming. The analysis of linear correlation in astronomy. In Arthur Beer, editor, *Vistas in Astronomy*, volume 10, pages 125–142. Pergamon Press, New York, 1968.
- [Ful87] Wayne A. Fuller. *Measurement Error Models*. John Wiley & Sons, Inc., New York, 1987.
- [KS79] Maurice Kendall and Alan Stuart. *The Advanced Theory of Statistics, Volume 2*. Charles Griffin & Co., Ltd., London, 1979.
- [Lor92] T.J. Loredo. The promise of Bayesian inference for astrophysics. In E.D. Feigelson and G.J. Babu, editors. *Statistical Challenges in Modern Astronomy*, pages 275–306. Springer-Verlag, New York, 1992.
- [Zel87] Arnold Zellner. *An Introduction to Bayesian Inference in Econometrics*. Krieger Publishing Co., Florida, 1987.

New Problems and Approaches Related to Large Databases in Astronomy

Fionn Murtagh and Alex Aussem¹

ABSTRACT Analyzing large image and text databases poses particular computational problems. Computational problems can sometimes be solved by using traditional analysis techniques, and by throwing more and more memory cycles at them. A more aesthetic way to tackle such scalability problems is to find new data structures and new algorithms which will more thoroughly deal with these issues. One of the most looming issues in data analysis is the laborious phase prior to the main analysis: selection of data, coding, etc. We summarize some recent results in data coding. We then look at how the incorporation of the wavelet transform into data analysis can helpfully mitigate some problems related to preliminary data processing. We look at how these same principles (but with a different wavelet transform) can be used in time series prediction.

7.1 Preliminary data processing, coding and analysis: Overview

The major part of data analysis does not go into the analysis itself, but rather into the demarcation of the problem, selection of data, preliminary processing, coding and recoding, followed by perhaps the final 5% of the investigators' time in the use of analysis methods and programs. For multivariate data analysis (which we take here as the same set of problems, and more or less the same methods and techniques, as neural networks, or data mining) data coding is of pivotal importance.

A rich tradition has been built up in conjunction with correspondence analysis and so it is there that we seek some interesting developments. Ter-

¹Fionn Murtagh, ST-ECF, Karl-Schwarzschild-Str. 2, D-85748 Garching (Germany) and Faculty of Informatics, University of Ulster, Londonderry BT48 7J (Northern Ireland).

Alex Aussem, Université René Descartes, 45, rue des Saints-Pères, 75006 Paris (France)

minology includes doubling, complete disjunctive form, double rescaling (all related to the recoding of quantitative or qualitative values into qualitative or categorical vales), and the lever principle (related to the projections in the reduced dimensional space). Such coding may be very useful for giving equal importance to different variables and parts of the population studied. (Note that – at least in correspondence analysis – a population is aimed at, rather than a sample.) This is a problem in discriminant analysis, for example, where some classes of object may be very few in number, while others are very numerous (see [17]).

A very interesting form of coding is fuzzy coding (also called piecewise linear coding, or barycentric coding). This is also a categorical coding, and takes a form such as: $x \rightarrow \{1, 0\}$ if $x \leq L$; $x \rightarrow \{0, 1\}$ if $x \geq U$; else $x \rightarrow \{a, 1 - a\}$, where a is linearly interpolated between 1 and 0; and L and U are thresholds. More enhanced versions of this form of coding can be used. It is justified by the close link between correspondence analysis and categorical data. It also can be used to balance the effect of presence or absence of particular variables, and it does this in a continuous manner. Perhaps most importantly of all, it facilitates interpretation. As an example, in a meteorological-astronomical case, values of “low” and “high” seeing (observing quality) have much importance for the astronomer, and this justifies using such fuzzy terms at all stages of the analysis (see [9], [13], [14]).

7.2 Multiscale image processing based on image noise analysis: Overview

Multiscale methods (wavelet analysis, or other such sequences of the data at progressively less resolved scales) have been used on 2-dimensional images, 1-dimensional images (spectra), and 3-dimensional images (data cubes) for noise filtering, image deconvolution and other applications ([22, 16]). Computational advantages are available when, for example, image fusion takes place at successively increasing resolution scale [8]. Noise is sought at different resolution scales, based on a model for the input image. In astronomy, knowledge of such a model is available – perhaps more so than in other fields. CCD detectors have additive Poisson noise, and Gaussian read-out noise. Digitized photographic images are contain a mixture of Poisson and Gaussian noise (with saturation effects). In order to bring such images to a common noise framework, we have made extensive use of variance stabilization (see e.g. [1], generalized by Albert Bijaoui to the Poisson and Gaussian case). Work is ongoing for cases with low numbers of counts, and for non-stationary noise (medical images of different origins).

This work motivates the use of wavelet transforms for analysis of multivariate data analysis, as described below. How can we bring the effec-

tiveness of multiscale methods more widely to bear on multivariate data analysis?

7.3 Orthogonal wavelet transforms and data analysis

Data analysis, for exploratory purposes, or prediction, is usually preceded by various data transformations and recoding. The wavelet transform offers a particularly appealing data transformation, as a preliminary to data analysis, for de-noising, smoothing, etc., in a natural and integrated way.

For an introduction to the orthogonal wavelet transform, see e.g. [24, 7]. We consider the signal's detail signal, ξ_m , at resolution levels, m . With the residual, smoothed image, x_0 , we have the wavelet transform of an input signal x as follows. Define ξ as the row-wise juxtaposition of all $\{\xi_m\}$ and x_0 , and consider W given by

$$Wx = \xi = \begin{bmatrix} \xi_{N-1} \\ \vdots \\ \xi_0 \\ x_0 \end{bmatrix}$$

with $W^T W = I$ (the identity matrix). Examples of these orthogonal wavelets are the Daubechies family, and the Haar wavelet transform (see [19], [7]). Computational time is $O(n)$ for an n -length input data set.

The basis for multivariate data analysis, based on the orthogonal wavelet transform, is now described. We consider the wavelet transform of x , Wx . Considering two vectors, x and y , we have $\|x - y\|^2 = \|Wx - Wy\|^2$. This is the fundamental result used in this part of the paper: for use of the squared Euclidean distance, the wavelet transform can replace the original data in the data analysis. This in turn allows us to directly manipulate the wavelet transform values, using any of the stratagems which have been found useful in other areas. The results based on the orthogonal wavelet transform exclusively imply use of the Euclidean distance, which nonetheless covers a considerable area of current data analysis practice. Future work will investigate extensions to other metrics.

Foremost among modifications of the wavelet transform coefficients is to approximate the data, progressing from coarse representation to fine representation, but stopping at some resolution level m' . Filtering or non-linear regression of the data can be carried out by deleting insignificant wavelet coefficients at each resolution level (noise filtering), or by “shrinking” them (data smoothing) [6]. Reconstitution of the data then provides a cleaned data set. In [22], noise suppression is discussed, based on a noise model for the input data. If linear combinations of *i.i.d.* (independent identically

distributed) Gaussian-distributed data are formed, then the successive resolution levels remain Gaussian. If the input data is Poisson, or a mixture of Gaussian and Poisson, then variance stabilization may be used to provide a new input data set with Gaussian properties.

7.4 Example: a wavelet-transform based Kohonen net

The Kohonen “self-organizing feature map” (SOFM) approach has been described in many texts (see references in [15]). We used a set of 45 astronomical spectra. These were of the complex AGN (active galactic nucleus) object, NGC 4151, and were taken with the small but very successful IUE (International Ultraviolet Explorer) satellite which is still active in 1996 after nearly two decades of operation [11]. We chose a set of 45 spectra observed with the SWP spectral camera, with wavelengths from 1191.2 Å to approximately 1794.4 Å, with values at 512 interval steps. A wavelet transform (Daubechies 4 wavelet used) of these spectra was generated. An overall 0.1σ (standard deviation, calculated on all wavelet coefficients) was used as a threshold, and coefficient values below this were set to zero. On average, 76% of the wavelet coefficients were zeroed in this way.

Fig. 1 shows an SOFM output using a 5×6 output representational grid. When a number of spectra were associated with a representational node, one of these is shown here, together with an indication of how many spectra are clustered at this node. Hatched nodes indicate no assignment of a spectrum.

We then constructed the SOFM on the wavelet coefficients (following zeroing of 76% of them). The assignments of the 45 spectra were identical to the assignments associated with Fig. 1. The values associated with output representational nodes were in this case the *wavelet transform* of SOFM neurons (cluster centers), which can be converted back to neuron (cluster center) values with linear computational cost. This approach to SOFM construction leads to the following possibilities:

1. Efficient implementation: a good approximation can be obtained by zeroing most wavelet coefficients, which opens the way to more appropriate storage (e.g. offsets of non-zero values) and distance calculations (e.g. implementation loops driven by the stored non-zero values). Similarly, compression of large datasets can be carried out. Finally, calculations in a high-dimensional space, \mathbf{R}^m , can be carried out more efficiently since, as seen above, the number of non-zero coefficients may well be small with very little loss of useful information.
2. Data “cleaning” or filtering is a much more integral part of the data analysis processing. If a noise model is available for the input data,

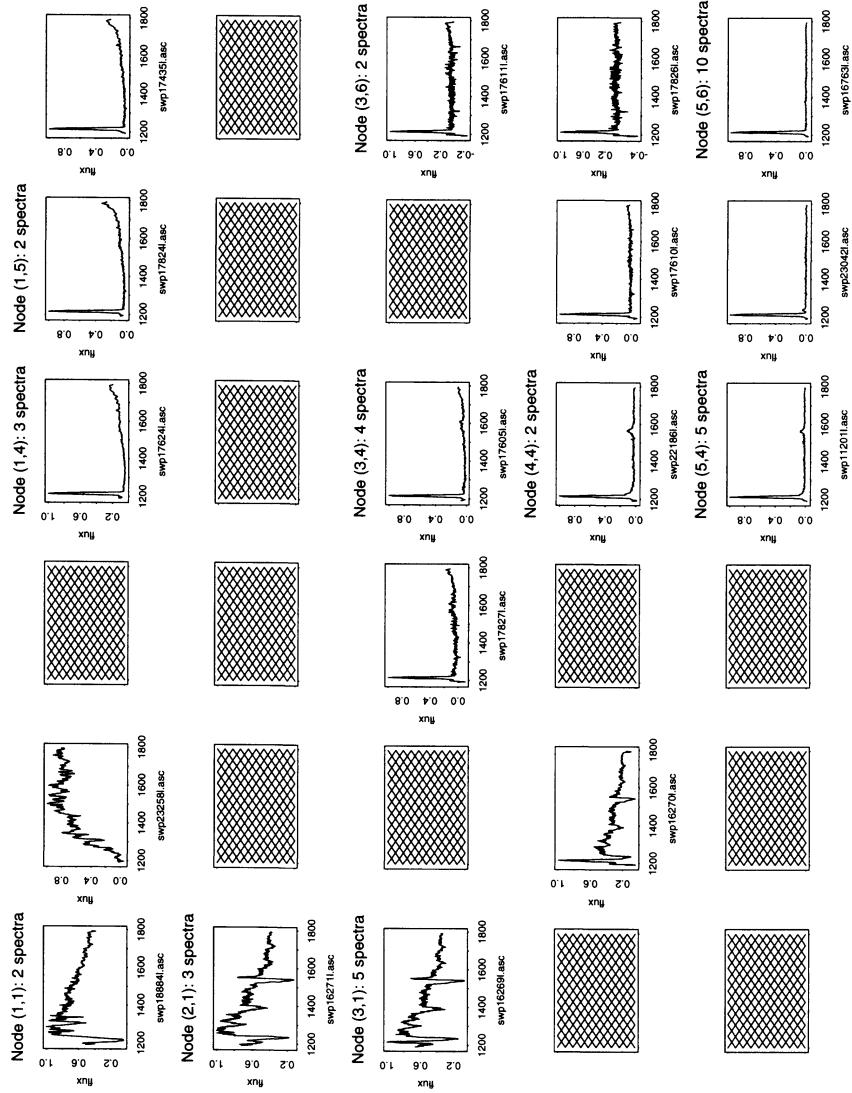


FIGURE 1. Kohonen SOFM of 45 spectra: original data normalized to maximum = 1 for each spectrum.

then the data can be de-noised at multiple scales [22]. By suppressing wavelet coefficients at certain scales, high-frequency or low-frequency information can be removed. Part of the data coding phase, prior to the analysis phase, can be dealt with more naturally in this new integrated approach.

7.5 *K*-means and principal components analysis in wavelet space

We used the set of 45 spectra which were “cleaned” in wavelet space, by putting 76% of the coefficients to zero (cf. above). This was therefore a set of cleaned wavelet-transformed vectors. We also reconstructed these cleaned spectra using the inverse wavelet transform. Due to design, therefore, we had an input data array of dimensions 45×512 , with 76% of values equal to zero; and an input data array of dimensions 45×512 , with no values exactly equal to zero.

A number of runs of the *k*-means partitioning algorithm were made. The exchange method, described in [21] was used. Four, or two, clusters were requested. Identical results were obtained for both data sets, which is not surprising given that this partitioning method is based on the Euclidean distance. For the 4-cluster, and 2-cluster, solutions we obtained respectively these assignments:

123213114441114311343133141121412222222121114

1222111111111111111111111121112222222121111

The case of principal components analysis was very interesting. We know that the basic PCA method uses Euclidean scalar products to define the new set of axes. Often PCA is used on a variance-covariance input matrix (i.e. the input vectors are centered); or on a correlation input matrix (i.e. the input vectors are rescaled to zero mean and unit variance). These two transformations destroy the Euclidean metric properties vis-à-vis the raw data. Therefore we used PCA on the unprocessed input data. We obtain identical eigenvalues and eigenvectors for the two input data sets.

The eigenvalues are similar up to numerical precision:

1911.217163	210.355377	92.042099	13.908587	7.481989
2.722113	2.304520			

1911.220703	210.355392	92.042336	13.908703	7.481917
2.722145	2.304524			

The eigenvectors are similarly identical. The actual projection values are entirely different. This is simply due to the fact that the principal components in wavelet space are themselves inverse-transformable to provide principal components of the initial data.

Various aspects of this relationship between original and wavelet space remain to be investigated. We have argued for the importance of this, in the framework of data coding and preliminary processing. We have also noted that if most values can be set to zero with limited (and maybe beneficial) effect, then there is considerable scope for computational gain also.

The processing of sparse data can be based on an “inverted file” data-structure which maps non-zero data entries to their values. The inverted file data-structure is then used to drive the distance and other calculations. Reference [12] (pp. 51–54 in particular) discusses various algorithms of this sort.

The results described here, from the multivariate data analysis perspective, are very exciting. They not only open up computational advances. They also provide a thrust into the very difficult issue of data coding and preliminary processing.

7.6 Combining scale-based forecasts in time series analysis

We discuss a simple strategy aimed at improving statistical or neural network prediction accuracy, based on the combination of predictions at varying resolution levels of the domain under investigation (here: time series). First, a wavelet transform is used to decompose the time series into varying scales of temporal resolution. The latter provide a sensible decomposition of the data so that the underlying temporal structures of the original time series become more tractable. Then, a statistical or neural network or other prediction is carried out independently at each resolution scale. The individual wavelet scale forecasts are afterwards recombined to form the current estimate. The predictive ability of this strategy is assessed with the well-known sunspot series.

An additive wavelet transform is provided by the *à trous* (with holes) algorithm [10][23]. This is a “stationary” [18] or redundant transform, i.e. decimation is not carried out.

The wavelet expansion of the multivariate time series, vector x , in terms of wavelet coefficients, is given by

$$x(t) = c_p(t) + \sum_{i=1}^p w_i(t)$$

The term c_p is the residual.

This equation provides a reconstruction formula for the original time series. It is additive, which leads us to fuse predictions also in an additive manner. We will see shortly, however, that we can adopt a hybrid strategy in regard to exactly what is combined to yield an overall prediction. That is, we can test a number of short-memory and long-memory predictions at each resolution level, and retain the method which performs best.

For a time series of size n , the wavelet decomposition used here can be determined with $O(n)$ computational cost. It therefore has more favorable computational cost than the Fast Fourier Transform (FFT).

We are thus armed with (i) a powerful modeling and forecasting method, and (ii) a versatile and tractable data decomposition scheme (which may well aim at capturing interpretable resolution levels in the data).

7.7 Example: Multiscale prediction

The sunspot series was the first time series studied with autoregressive models [20, 25, 28], and thus has served as a benchmark in the forecasting literature. The sunspots are dark blotches on the sun that can be related to other solar activities such as the magnetic field cycles, which in turn influence, by indirect and intangible means, the meteorological conditions on earth. Although the data exhibit strong regularities, attempts to understand the underlying features of the series have failed because the amount of available data was insufficient. Thus the sunspots provide an interesting data-set to test our wavelet decomposition method.

Consistent with previous appraisals [25, 26], we use range-normalized yearly averages of the sunspot data tabulated from 1720 to 1979. One-step ahead error is used as a performance criterion. The single-step prediction error is monitored on 59 withheld sunspot values ranging from 1921 to 1979, while the remaining data is used for training.

For the individual forecasts, a Dynamical Recurrent Neural Network (DRNN) was used, which was trained on each resolution scale with the temporal-recurrent backpropagation (TRBP) algorithm (see [2, 3, 4, 5]). By virtue of its internal dynamic, this general class of dynamic connectionist network approximates the underlying law governing each resolution level by a system of nonlinear difference equations.

Results are discussed in [5], where superior MSE (mean squared error) performance is found compared to a classical multilayer perceptron, or a primitive autoregressive model. In fact, the approach described here allows for a hybrid forecast, using a combination of *any* prediction engine. This powerful feature was used (the autoregressive prediction at low-resolution scales was combined with the DRNN approach – requiring more data to perform acceptably – at the higher-resolution scales). Fig. 2 shows the data used, and one-step ahead prediction obtained.

7.8 Conclusion

Wavelet-based multivariate data analysis is very powerful. It allows complete integration of different methods in a common mathematical framework. It addresses certain aspects of data transformation and coding which are vital in practice, in data analysis.

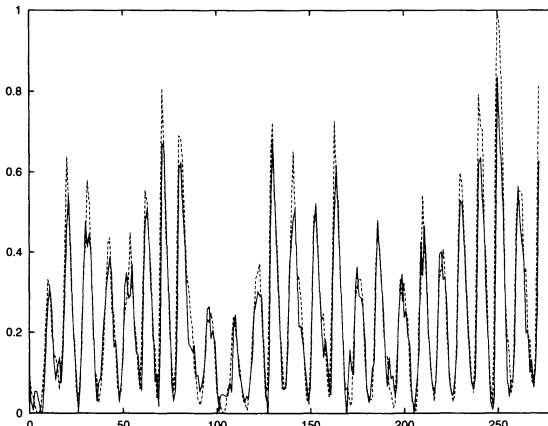


FIGURE 2. Single step prediction with several DRNNs. Original and predicted sunspot series overplotted. The dashed curve corresponds to the actual series and the plain curve is the prediction. Hybrid multiresolution neural network/autoregressive prediction used.

We have seen that computational advantages may be obtained. We have applied this new methodology to time series analysis, and to change-point detection. We have also applied it to various clustering and dimensionality-reduction studies.

This integrated wavelet/multivariate analysis methodology opens up new theoretical and practical directions in this field.

The spectral data used in this study are available at <ftp://ftp.infm.ulst.ac.uk:/pub/Image/wt-mda>

REFERENCES

- [1] F.J. Anscombe, "The transformation of Poisson, binomial and negative-binomial data", *Biometrika*, 15, 246–254, 1948.
- [2] A. Aussem, F. Murtagh and M. Sarazin, "Dynamical recurrent neural networks – towards environmental time series prediction," *International Journal on Neural Systems*, 6, 145–170, 1995.
- [3] A. Aussem, F. Murtagh and M. Sarazin, "Fuzzy astronomical seeing nowcasts with a dynamical and recurrent connectionist network," *Neurocomputing*, 1995, in press.
- [4] A. Aussem, *Theory and applications of dynamical and recurrent neural networks towards prediction, modeling and adaptive control of dynamical processes*, Ph.D. Thesis, Université René Descartes, June, 1995.
- [5] A. Aussem and F. Murtagh, "Combining neural network forecasts on wavelet-transformed time series", *Connection Science*, 1996, submitted.
- [6] A. Bruce and H.-Y. Goa, *S+Wavelets User's Manual*, Version 1.0, StatSci Division, MathSoft Inc., Seattle, WA, 1994.

- [7] I. Daubechies, *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1992.
- [8] J.P. Djambji, A. Bijaoui and R. Maniere, “Geometrical registration of images: the multiresolution approach”, *Photogrammetric Engineering and Remote Sensing*, 59, 645–653, 1993.
- [9] T.K. Gopalan and F. Murtagh, “The role of input data coding in multivariate data analysis: the example of correspondence analysis”, *International Statistical Review*, 1995, submitted.
- [10] M. Holschneider, R. Kronland-Martinet, J. Morlet and Ph. Tchamitchian, “A real-time algorithm for signal analysis with the help of the wavelet transform”, in *Wavelets: Time-Frequency Methods and Phase Space*, Berlin: Springer-Verlag, 286–297, 1989.
- [11] J.P.D. Mittaz, M.V. Penston and M.A.J. Snijders, “Ultraviolet variability of NGC 4151: a study using principal component analysis”, *Monthly Notices of the Royal Astronomical Society*, 242, 370–378, 1990.
- [12] F. Murtagh, *Clustering Algorithms*. Physica-Verlag, Würzburg, 1985.
- [13] F. Murtagh and M. Sarazin, “Nowcasting astronomical seeing: a study of ESO La Silla and Paranal”, *Publications of the Astronomical Society of the Pacific*, 105, 932–939, 1993.
- [14] F. Murtagh, A. Aussem, A. and M. Sarazin, “Nowcasting astronomical seeing: towards an operational approach”, *Publications of the Astronomical Society of the Pacific*, 107, 702–707, 1995.
- [15] F. Murtagh, and M. Hernández-Pajares, “The Kohonen self-organizing feature map method: an assessment”, *Journal of Classification*, 12, 165–190, 1995.
- [16] F. Murtagh, J.-L. Starck, and A. Bijaoui, “Multiresolution in astronomical image processing: a general framework”, *International Journal of Image Systems and Technology*, 6, 332–338, 1995.
- [17] F. Murtagh, “Application de l’analyse factorielle et de l’analyse discriminante à des données colligées pour être soumises à des réseaux de cellules”, *Les Cahiers de l’Analyse des Données*, XXI, 53–74, 1996.
- [18] G.P. Nason and B.W. Silverman, “The stationary wavelet transform and some statistical applications”, preprint, University of Bath, 1995.
- [19] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes*, 2nd ed., Chapter 13, Cambridge University Press, New York, 1992.
- [20] M.B. Priestley, *Spectral Analysis and Time Series*, New York: Academic Press, 1981.
- [21] H. Späth, *Cluster Dissection and Analysis*, Ellis Horwood, Chichester, 1985.
- [22] J.L. Starck, A. Bijaoui, F. Murtagh, “Multiresolution support applied to image filtering and deconvolution”, *Graphical Models and Image Processing*, 57, 420–431, 1995.
- [23] M.J. Shensa, “Discrete wavelet transforms: wedding the à trous and Mallat algorithms”, *IEEE Transactions on Signal Processing*, 40, 2464–2482, 1992.
- [24] G. Strang, “Wavelets and dilation equations: a brief introduction”, *SIAM Review*, 31, 1989, 614–627.
- [25] H. Tong, *Non Linear Time Series*. Oxford: Clarendon Press, 1990.
- [26] A.S. Weigend, D.E. Rumelhart and B.A. Huberman, “Predicting the future: a connectionist approach,” *International Journal of Neural Systems*, 1, 195–220, 1990.

- [27] Yansun Xu, J.B. Weaver, D.M. Healy and Jian Lu, "Wavelet transform domain filters: a spatially selective noise filtration technique", *IEEE Transactions on Image Processing*, 3, 747–758, 1994.
- [28] G.U. Yule, "On a method of investigating periodicities in disturbed series with special reference to Wolf's sunspot numbers," *Philos. Trans. Roy. Soc. London Ser. A*, 226, 267, 1927.

Object Classification in Astronomical Images

Richard L. White

ABSTRACT Automated classification methods are needed for processing the huge quantities of data generated by modern astronomical instruments. The star-galaxy classification problem and some techniques that have been applied to it are briefly reviewed. Methods for constructing training sets and selecting parameters are described.

A new method of scaling parameter values using ranks has been developed. This approach is found to be of great utility for distinguishing stars and galaxies on digitized photographic plates. It should be widely applicable to other classification problems, especially when the data being classified are not completely homogeneous.

8.1 Introduction

The combination of ever larger detector formats, powerful computers capable of processing vast amounts of data, and on-line access to large databases has made an ocean of data available to astronomers. Because of the large size of individual databases, and the large number of databases, automated techniques must be used to search a single or combined database for particular classes of interesting or rare objects. Moreover, automated methods are required to construct the databases themselves from the raw data collected at observatories.

In this paper I focus mainly on the problem of star-galaxy separation. For our purposes, the main difference between stars and galaxies is that stars look completely unresolved and so have sharp images, while galaxies look fuzzy. Even though stars are not fuzzy, images taken through telescopes have finite resolution and the stellar images do have non-zero sizes. Figure 1 shows a small section from a digitized photographic plate with galaxies marked. Distinguishing stars from galaxies becomes difficult when both are faint and the galaxies are small; it is also sometimes difficult for very bright objects, when the stars can saturate the detector and no longer look compact. All galaxies are centrally concentrated, and a significant fraction of galaxies have activity near their centers and so appear to have a stellar

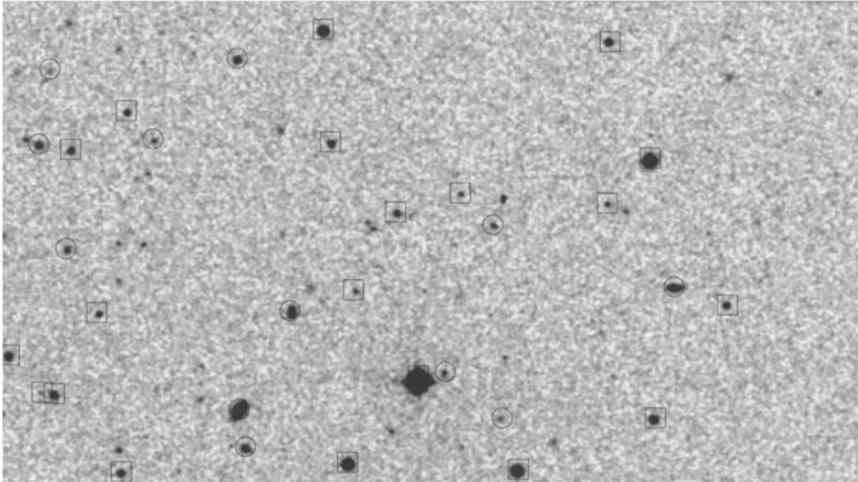


FIGURE 1. 530×300 pixel section of digitized Palomar Sky Survey II plate. Pixels are one arcsecond. The contrast has been enhanced to make visible both faint objects and the noise in the sky. Objects brighter than $V = 20.5$ from Postman's deep CCD images are marked (squares are stars, circles are galaxies.) Our goal is to develop a classification algorithm that can distinguish the stars from the galaxies. There are more than 3 million image patches this size across the entire sky.

core surrounded by faint fuzz. Such objects are especially difficult to classify correctly.

Two examples will suffice to demonstrate the scope of the problem. The Sloan Digital Sky Survey, scheduled to begin in 1997, will generate a survey of 10,000 square degrees (1/4 of the sky) in 5 colors using an array of 30 2048×2048 pixel CCD detectors [GK92]. The ~ 12 terabytes of raw pixel values generated by this survey will be processed to produce a catalog of $> 10^8$ objects with positions, brightness in 5 colors, and classifications as stars or galaxies. The classification will be based on both the objects' morphological parameters (shape, central concentration, etc.) and on their colors. Accurate classification of this vast number of objects down to the detection limit of the survey is a daunting task that will require new classification methods.

The Space Telescope Science Institute constructed a catalog of guide stars for pointing the Hubble Space Telescope [LSM⁺90]. The original catalog contained about 2×10^7 objects to a brightness of 15th magnitude and was constructed from a digitized version of photographic plates in a single color taken at the Palomar and the UK Schmidt telescopes. A second generation catalog is now under construction [LMJ⁺95]: it will have nearly one billion objects and will be based on digitized photographic data taken in a variety of colors and at several different epochs, so it will include both

color and information on proper motions of objects in the catalog. This project also needs an accurate, automatic star-galaxy classifier.

Although these projects may sound similar, they present rather different classification problems because of the non-linear response of photographic emulsions to light.

The next section (§8.2) briefly describes some of the common methods used for classification, with some attention to ways to handle noise in the data. The steps required during the development of a classifier for a particular problem are discussed (§8.3). A new method of constructing more robust features by using ranks is described (§8.4), with examples from the Guide Star Catalog II star/galaxy discrimination problem. The concluding section (§8.5) discusses promising areas for future work.

8.2 Methods for classification

Any classification method uses a set of *features* or *parameters* to characterize each object, where these features should be relevant to the task at hand. We consider here methods for *supervised* classification, meaning that a human expert both has determined into what classes an object may be categorized and also has provided a set of sample objects with known classes. This set of known objects is called the *training set* because it is used by the classification programs to learn how to classify objects. There are two phases to constructing a classifier. In the training phase, the training set is used to decide how the parameters ought to be weighted and combined in order to separate the various classes of objects. In the application phase, the weights determined in the training set are applied to a set of objects that do *not* have known classes in order to determine what their classes are likely to be.

If a problem has only a few (two or three) important parameters, then classification is usually an easy problem. For example, with two parameters one can often simply make a scatter-plot of the feature values and can determine graphically how to divide the plane into homogeneous regions where the objects are of the same classes. The classification problem becomes very hard, though, when there are many parameters to consider. Not only is the resulting high-dimensional space difficult to visualize, but there are so many different combinations of parameters that techniques based on exhaustive searches of the parameter space rapidly become computationally infeasible. Practical methods for classification always involve a heuristic approach intended to find a “good-enough” solution to the optimization problem.

8.2.1 Neural networks

There are a number of standard classification methods in use. Probably neural network methods are most widely known. Odewahn et al. [OSP⁺92] applied neural network methods to the star-galaxy classification problem on digitized photographic plates. They obtained good results for objects in a limited brightness range. The biggest advantage of neural network methods is that they are general: they can handle problems with very many parameters, and they are able to classify objects well even when the distribution of objects in the N -dimensional parameter space is very complex. The disadvantage of neural networks is that they are notoriously slow, especially in the training phase but also in the application phase. Another significant disadvantage of neural networks is that it is very difficult to determine how the net is making its decision. Consequently, it is hard to determine which of the image features being used are important and useful for classification and which are worthless. As I discuss below (§8.3), the choice of the best features is an important part of developing a good classifier, and neural nets do not give much help in this process.

8.2.2 Nearest-neighbor classifiers

A very simple classifier can be based on a nearest-neighbor approach. In this method, one simply finds in the N -dimensional feature space the closest object from the training set to an object being classified. Since the neighbor is nearby, it is likely to be similar to the object being classified and so is likely to be the same class as that object. Nearest neighbor methods have the advantage that they are easy to implement. They can also give quite good results if the features are chosen carefully (and if they are weighted carefully in the computation of the distance.) There are several serious disadvantages of the nearest-neighbor methods. First, they (like the neural networks) do not simplify the distribution of objects in parameter space to a comprehensible set of parameters. Instead, the training set is retained in its entirety as a description of the object distribution. (There are some thinning methods that can be used on the training set, but the result still does not usually constitute a compact description of the object distribution.) The method is also rather slow if the training set has many examples. The most serious shortcoming of nearest neighbor methods is that they are very sensitive to the presence of irrelevant parameters. Adding a single parameter that has a random value for all objects (so that it does not separate the classes) can cause these methods to fail miserably.

8.2.3 Decision trees

Decision tree methods have also been used for star-galaxy classification problems [FWD93, DWF94, Wei94]. In axis-parallel decision tree methods,

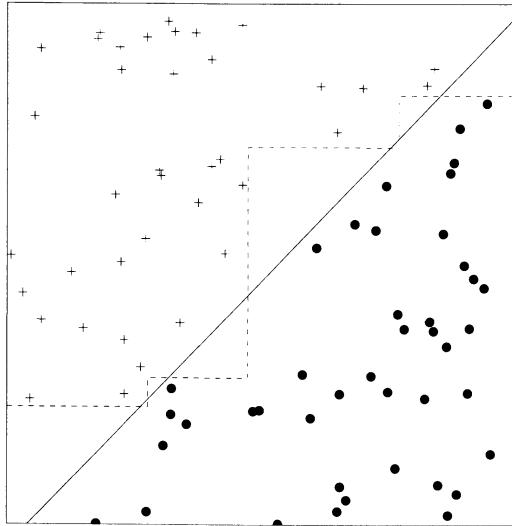


FIGURE 2. Sample classification problem with only two features. *Solid line*: oblique decision tree. *Dashed line*: axis-parallel decision tree. The true dividing line for the simulated data is the square's diagonal. The oblique decision tree is both simpler and more accurate than the axis-parallel tree.

a binary tree is constructed in which at each node a single parameter is compared to some constant. If the feature value is greater than the threshold, the right branch of the tree is taken; if the value is smaller, the left branch is followed. After a series of these tests, one reaches a leaf node of the tree where all the objects are labeled as belonging to a particular class. These are called axis-parallel trees because they correspond to partitioning the parameter space with a set of hyperplanes that are parallel to all of the feature axes except for the one being tested.

Axis-parallel decision trees are usually much faster in the construction (training) phase than neural network methods, and they also tend to be faster during the application phase. Their disadvantage is that they are not as flexible at modeling parameter space distributions having complex distributions as either neural networks or nearest neighbor methods. In fact, even simple shapes can cause these methods difficulties. For example, consider a simple 2-parameter, 2-class distribution of points with parameters x, y that are all of type 1 when $x > y$ and are of type 2 when $x < y$ (Fig. 2). To classify these objects with an axis-parallel tree, it is necessary to approximate the straight diagonal line that separates the classes with a series of stair-steps. If the density of points is high, very many steps may be required. Consequently, axis-parallel trees tend to be rather elaborate, with many nodes, for realistic problems.

8.2.4 *Oblique decision trees*

Oblique decision trees attempt to overcome the disadvantage of axis-parallel trees by allowing the hyperplanes at each node of the tree to have any orientation in parameter space [HKS93, MKS94]. Mathematically, this means that at each node a linear combination of some or all of the parameters is computed (using a set of feature weights specific to that node) and the sum is compared with a constant. The subsequent branching until a leaf node is reached is just like that used for axis-parallel trees.

Oblique decision trees are considerably more difficult to construct than axis-parallel trees because there are so many more possible planes to consider at each tree node. As a result the training process is slower. However, they are still usually much faster to construct than neural networks. They have one major advantage over all the other methods: they often produce very simple structures that use only a few parameters to classify the objects. It is straightforward through examination of an oblique decision tree to determine which parameters were most important in helping to classify the objects and which were not used.

Most of our work has concentrated on oblique decision trees using the freely available software OC1¹ [MKS94], though we have used all the methods mentioned above. We find that oblique decision trees represent a good compromise between the demands of computational efficiency, classification accuracy, and analytical value of the results.

8.2.5 *Adapting classification methods to noise in data*

Noise is common in astronomical data. In the presence of noise, classification becomes a statistical estimation problem. Objects should not be classified as either stars or galaxies, but should be assigned probabilities of being in one or the other class.

All of the above methods can be modified to give a probabilistic estimate of class rather than a simple yes or no answer. Neural nets can have an output parameter that represents probability. It is also possible to provide to a neural network (or the other methods, for that matter) the noise in parameters as additional object features. It is hard to know what the neural network actually does with this information, though — this approach asks the classifier to discover that a particular parameter measures the noise on another parameter. Why should the classifier be burdened with deducing this when we already know what the parameters mean?

The nearest-neighbor method can be generalized to use the K nearest neighbors to an object, which can vote for the object's class. If the vote is not unanimous, it gives an estimate of the uncertainty of the object's

¹OC1 is available at <http://www.cs.jhu.edu/~salzberg/announce-oc1.html> and <ftp://ftp.cs.jhu.edu/pub/oc1>.

majority classification. The same approach can be applied to decision trees. If the search algorithm used to construct the decision tree has a randomized component (which is typical for optimization searches in high dimensional spaces and is in fact the case for OC1), one can construct a number of different decision trees and let them vote on an object's class.

For decision trees, it is also possible to use known noise estimates for the parameters to determine at each node the *probability* that an object belongs on the left branch or the right branch. These probabilities, multiplied down to the leaf nodes and summed over all the leaves belonging to a particular class, give a probability that an object belongs that class. We are just beginning to explore this approach, which has great intuitive appeal and appears to be a promising avenue for further work.

8.3 Steps in developing a classifier

The choice of an algorithm for classification is in many ways the easiest part of developing a scheme for object classification. The discussion above demonstrates that there are several “off-the-shelf” approaches available (though there is obviously still room for improvement.) There are two major hurdles to be faced before these methods can be used, though: a training set must be constructed for which the true classifications of the objects are known, and a set of object parameters must be chosen that are powerful discriminators for classification. Once a possible classifier has been identified, it is necessary to measure its accuracy.

8.3.1 Training sets

A training set must contain a list of objects with known classifications. Ideally the training set should contain many examples (typically thousands of objects) so that it includes both common and rare types of objects. Creating a training set requires a source of true object classifications, which is usually difficult even for human experts to generate if it must rely on the same data being used by the classifier.

To construct a training set for the star-galaxy classification problem, the best approach we have found is to use a more sensitive, higher resolution image to get true classifications of the objects seen in the survey data. In some cases (e.g. for the Sloan Digital Sky Survey), it may be quite difficult and time-consuming to acquire images that are significantly better than the survey images; in that case it may be possible to use instead simulated data (where the true classifications are known.) A risk in using simulated images is that they may not include all the effects that make classification difficult from the real data.

8.3.2 Feature selection

Adding many irrelevant parameters makes classification harder for all methods, not just the nearest neighbor methods. Training classifiers is an optimization problem in a many-dimensional space. Increasing the dimensionality of the space by adding more parameters makes the optimization harder (and the difficulty grows exponentially with the number of parameters.) It is always better to give the algorithm only the necessary parameters rather than expecting it to learn to ignore the irrelevant parameters.

One should not ask the classifier to rediscover everything you already know about the data. Not only should irrelevant parameters be omitted, but highly correlated parameters should be combined when possible to produce a few powerful features. For example, if you expect the shapes of images of a particular class of object to be similar, include a brightness-independent shape parameter rather than simply giving the classifier raw pixel values and expecting it to figure out how to extract shape information from the pixel values.

If the training process does not require too much computation, a useful approach to identifying the best parameters is to train many times on subsets of the features. We have used this method in two ways, both starting from the complete list of features and reducing it by removing parameters, and starting from a minimal list, augmenting it by adding parameters. Both methods have proven effective at pruning unnecessary parameters [SCF⁺95]. This procedure can be very fast if axis-parallel trees are used for the exploration. (Note that OC1 can construct either axis-parallel or oblique trees.)

Another useful approach with the decision tree methods is to examine directly the weights assigned to the various features. Important features are given a high weight, while unimportant features may not be used at all. This information can be used as a guide for pruning experiments.

8.3.3 Assessing classifier accuracy

Once a potentially useful classifier has been constructed, the accuracy of the classifier must be measured. Knowledge of the accuracy is necessary both in the application of the classifier and also in comparison of different classifiers.

The accuracy can be determined by applying the classifier to an independent training set of objects with known classifications. This is sometimes trickier than it sounds. Since training sets are usually difficult to assemble, one rarely has the resources to construct yet another set of objects with known classifications purely for testing. One must avoid the temptation to train and test on the same set of objects, though. Once an object has been used for training, any test using it is necessarily biased.

We normally use five-fold cross-validation to measure the accuracy of our classifiers. The training set is divided into five randomly selected subsets having roughly equal numbers of objects. The classifier is then trained five times, excluding a single subset each time. The resulting classifier is tested on the excluded subset. Note that each training session must be completely independent of the excluded subset of objects; one cannot, for example, use the results of an earlier training session as a starting point.

The advantage of cross-validation is that all objects in the training set get used both as test objects and as training objects. This ensures that the classifier is tested on both rare and common types of objects. The cost of cross-validation is that the training process must be repeated many times, adding to the computational cost of the training. In most applications, though, the computer time necessary to repeat the training is more readily available than is the human expert time required to generate completely independent test and training sets.

8.4 Use of ranks to make parameters more robust

Inhomogeneous data sets can make developing a classifier devilishly difficult. In the Guide Star Catalog II (GSC-II) star/galaxy classification problem, for example, there is a considerable variation in the characteristics of images from different photographic plates. The sky brightness and atmospheric “seeing” (which determines the stellar image size) change with time. The telescope optical characteristics also cause a slow variation in these parameters from the center to the edges of the plates. The atmosphere causes the images of objects photographed directly overhead (at the zenith) to differ from images of the same objects photographed near the horizon. Finally, the photographic plates themselves do not all respond identically when exposed to light. All of these factors combine to make the image features of identical objects appear to change from one plate to another.

Almost all large-scale surveys have similar problems. The Sloan Digital Sky Survey does not suffer plate emulsion variations, but it will have seeing, sky, and zenith-angle variations.

The straightforward solution to this problem is to generate many training sets to cover all the possible variations in object parameters. That is almost never a viable approach, however. For the GSC-II project, we have many thousands of photographic plates. Both the human expense of generating training sets and the computational expense of retraining the classifier on each plate are prohibitive.

A better approach is to find image parameters that are independent of the plate-to-plate variations. This section describes a new approach to generating robust parameters for classification. This approach appears to

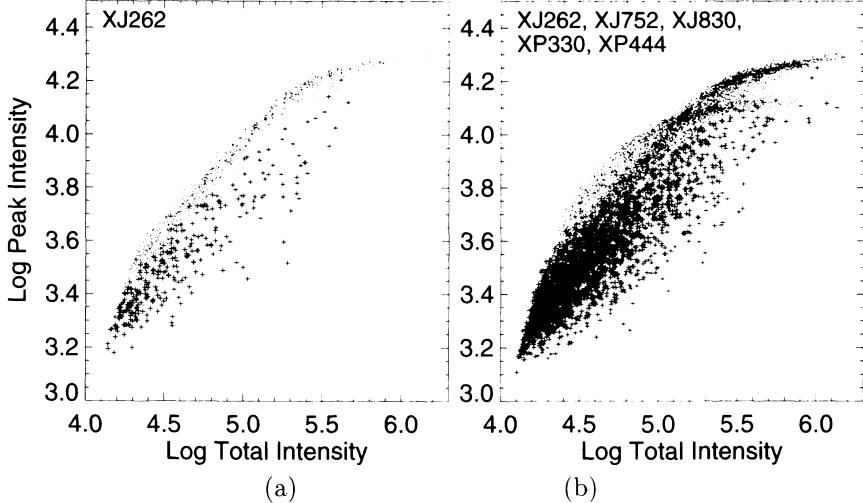


FIGURE 3. Total density versus peak density for stars (dots) and galaxies (plusses) measured on GSC-II photographic plates. (a) The distribution for a single plate (XJ262) is well-defined and could be used for classification. (b) The distribution for five plates differs for different plates and so a classifier for all plates could not be constructed from these parameters.

solve our problems for the GSC-II: it is a very general technique that will be applicable to many other problems as well.

It is not terribly difficult to find image parameters that separate stars and galaxies for a single GSC-II plate. Figure 3(a) shows a scatter-plot of the total density² (summed over all pixels in the object) versus the peak density (brightest pixel) for objects from one plate. Stars and galaxies for this training set were identified using the deep CCD catalog of Postman [PLG⁺96]. Stars are (mainly) well-separated from galaxies: stars have sharper images and so have brighter peak density values for a given total density. The distribution is non-linear as a result of the non-linear response of the photographic plate to light, but otherwise a reasonably simple and accurate classifier can be constructed using only these two parameters.

The problem gets very messy when one compares objects from different plates. Figure 3(b) shows the same plot with objects from five different plates. The well-defined distributions for individual plates can be discerned here, but it would be practically impossible to develop an accurate classifier based on these parameters.

We are currently using what appears to be a new approach for scaling these parameters so that they are plate-independent. The idea for this scal-

²Density measures the darkness of the photographic plate and is a non-linear function of the intensity of light to which the plate was exposed.

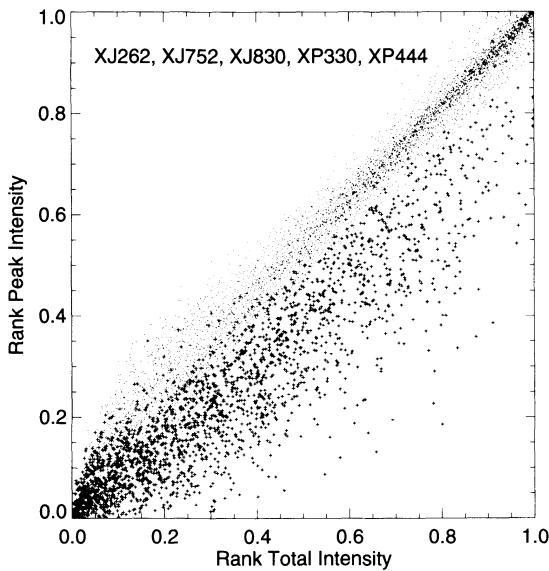


FIGURE 4. Distribution of total density versus peak density for objects from Fig. 2(b) after transformation using ranks. Stars and galaxies from different photographic plates separate very well. An accurate classifier can be constructed based on these rank parameters.

ing grew from some methods used in robust statistics. A standard “trick” in robust statistics is to use not the value of a parameter but its rank. For example, the Spearman rank-order correlation coefficient can be used to test for a correlation between two variables using not the actual values of the variables but their ranks.

The advantages of ranks in statistical applications are well-known: by using ranks, we are able to construct statistical tests that do not rely on the probability distribution of the variables. The ranks are by definition uniformly distributed. Similarly, using ranks for classification allow us to use parameters such as the total and peak intensities even though their detailed distributions vary from plate to plate.

In the current application, we proceed as follows. We compute the “raw” feature values for all objects on a plate (or on a portion of the plate if there are variations within the plate.) We then sort the raw values of each parameter and determine the rank of each raw value within the sorted list. These ranks are scaled to be in the range zero to one. Thus each of the raw features is transformed into a corresponding rank feature.

Note that the ranks are computed separately for objects from each plate. Once converted to ranks, the features for all objects can be combined and a single classifier can be used for objects from different plates.

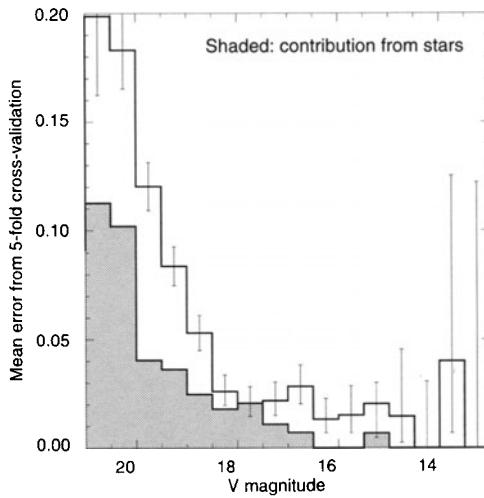


FIGURE 5. Error rate as a function of object brightness for oblique decision tree based on parameters transformed using ranks. All objects in the training set shown in Figures 3 and 4 are included. Excellent accuracy is achieved for brighter objects, and even for faint objects the accuracy is good.

Figure 4 shows the distribution of the total density rank versus the peak density rank for the same five plates shown in Figure 3(b). Using the rank transformation results in excellent separation between stars and galaxies. Transforming these parameters using their ranks has three significant benefits for our classifier. (1) The distributions become much more similar on different plates. (2) The distribution of objects with feature values becomes uniform, so for example there are equal number of objects in the intervals (0,0.1) and (0.9,1). That makes separating objects in feature space easier. (3) The line separating the stars from galaxies becomes nearly perfectly linear, which makes the decision tree classifiers even more effective than they would be for the untransformed features.

A single decision tree classifier with an accuracy of 91.5% on all five plates can be constructed using only these two parameters. Most of the classification errors are faint, noisy objects. When additional features are included, we are able to construct a classifier with an accuracy of 95–96% for all objects and 98–99% for moderately bright objects (see Fig. 5.) We apply the rank transformation to most of the additional parameters. The complete details of our classifier and the features we are using will be reported elsewhere.

This method was developed only recently, and some details are still being worked out. What is the best way to store the rank transformation so that small object lists (e.g., from small images extracted from the plate)

can be classified? How can we best adapt this method to cases where the population of objects changes substantially from one data set to another? For example, in the plane of the Milky Way, the sky images are densely crowded with many stars; galaxies are rare or non-existent. Looking straight up out of the plane of our galaxy, stars are scattered much more sparsely across the sky and a larger fraction of the objects are galaxies. It may be necessary to have independent training sets for the two extremes and to somehow interpolate between them for intermediate regions of the sky.

Despite the uncertainty about how some details should be properly handled, it is clear that this new method is a powerful tool for converting parameters to a more useful form that in most cases leads to simpler and more accurate classifiers.

8.5 Summary

In this paper I have briefly reviewed methods for object classification, focusing on the problem of distinguishing stars from galaxies. There are several methods available for classification; the oblique decision tree method is well-suited for our problem, but the other methods are also useful and may be better for other problems. Choice of parameters and the construction of high quality training/test data sets are important steps in the problem.

A new method has been described for scaling feature values based on ranks to make the features more robust. This approach should be generally applicable to many classification problems and is especially useful when the data being classified are not completely homogeneous.

There are several interesting avenues for future work. Two topics related to this conference come to mind. First, since the use of ranks (borrowed from statistical methods) looks like a powerful tool for classification, perhaps there are other transformation methods from robust or non-parametric statistics that should be explored for similar applications. Second, most of the classifiers discussed here do not view the classification as a statistical estimation problem. When the object features are noisy, though, classification really is a statistical problem. All the classification methods we have used may be adapted to utilize noise estimates on parameters (§8.2.5), but there has been relatively little work on this problem to date.

Acknowledgments: This project is an outgrowth of an on-going collaboration with Steven Salzberg, Holland Ford, Rupali Chandar, and Sreerama Murthy on applications of classification methods to astronomical problems. Thanks to all of them for many useful discussions. I am especially grateful to Murthy and Steven for freely distributing the OC1 decision tree program, which I have used extensively in this work. The development of the high

quality training sets was done with my GSC-II colleague, Andrea Zacchei, whose participation in the project is supported by the Italian Council for Research in Astronomy. Marc Postman's CCD galaxy catalog has proved invaluable in this effort.

REFERENCES

- [DWF94] S. Djorgovski, N. Weir, and U. Fayyad. Processing and analysis of the Palomar-STScI Digital Sky Survey using a novel software technology. In D. R. Crabtree, R. J. Hanisch, and J. Barnes, editors, *Astronomical Data Analysis Software and Systems III*, pages 195–204, San Francisco, 1994. ASP.
- [FWD93] U. M. Fayyad, N. Weir, and S. Djorgovski. SKICAT: A machine learning system for automated cataloging of large scale sky surveys. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 112–119, Amherst, MA, 1993. Morgan Kaufmann.
- [GK92] J. E. Gunn and G. R. Knapp. The Sloan Digital Sky Survey. *Proceedings of the Astronomical Society of the Pacific*, 43:267–279, 1992.
- [HKS93] D. Heath, S. Kasif, and S. Salzberg. Learning oblique decision trees. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1002–1007, Chambery, France, 1993. Morgan Kaufmann.
- [LMJ⁺95] B. M. Lasker, B. J. McLean, H. Jenkner, M. G. Lattanzi, and A. Spagna. Potential application of GSC-II for GAIA operations. In F. van Leeuwen and M. Perryman, editors, *Future Possibilities for Astrometry in Space*, ESA SP-379, 1995.
- [LSM⁺90] B. M. Lasker, C. R. Sturch, B. J. McLean, J. L. Russell, H. Jenkner, and M. M. Shara. The Guide Star Catalog. I. Astronomical foundations and image processing. *Astronomical Journal*, 99:2019, 1990.
- [MKS94] S. K. Murthy, S. Kasif, and S. Salzberg. Induction of oblique decision trees. *Journal of Artificial Intelligence Research*, 2:1–33, 1994.
- [OSP⁺92] S. C. Odewahn, E. B. Stockwell, R. L. Pennington, R. M. Humphreys, and W. A. Zumach. Automated star/galaxy discrimination with neural networks. *Astronomical Journal*, 103:318–331, 1992.
- [PLG⁺96] M. Postman, L. M. Lubin, J. E. Gunn, J. B. Oke, J. G. Hoessel, D. P. Schneider, and J. A. Christensen. The Palomar distant cluster survey: I. The cluster catalog. *Astronomical Journal*, 111:615, 1996.
- [SCF⁺95] S. Salzberg, R. Chandar, H. Ford, S. K. Murthy, and R. White. Decision trees for automated identification of cosmic rays in Hubble Space Telescope images. *Proceedings of the Astronomical Society of the Pacific*, 107:279–288, 1995.
- [Wei94] N. Weir. *Automated analysis of the digitized Second Palomar sky survey: system design, implementation, and initial results*. PhD thesis, California Institute of Technology, Pasadena, California, 1994.

Discussion by Francisco G. Valdes

The problem of automatic object classification in astronomical images may be summarized as follows. A luminous astronomical object is observed as a set of brightness values, with some uncertainty and noise, in a two dimensional array where the array dimensions are projected spatial positions. Observation classes are defined based on the expected brightness values for each type of astrophysical object. Assign the observation to the class with the greatest probability of having produced the measured brightnesses.

The challenges which arise are that the observation classes can be similar due to limited resolution even though the physical objects are not similar, a particular type of object can have variations in its appearance, and the fainter objects will have a significant amount of noise. It is the ability to give the most probable class in the presence of shape variations and noise that distinguishes classifiers.

The simplest set of classes which are of astronomical interest are stars, galaxies, and other. A star is essentially an unresolved point of light. However, due to the effects of the atmosphere and detector the point of light will be spread out into a characteristic shape called the point spread function or PSF. The instrumental component will normally be fixed and the atmospheric component will vary from observation to observation. The key point is that whatever the PSF all stars will have the same underlying brightness distribution.

Galaxies are distant collections of stars which appear as patches of light in various shapes. The detector and atmospheric distortions will also affect the brightnesses, but if the patch of light is big enough it will have a shape different (more extended) than that of a star. The "other" category is used for noise, instrumental defects, or moving objects such as planets and aircraft.

The problem of automated star-galaxy classification or discrimination has been the subject of considerable work over the past 20 years. There are basically three approaches that have been applied with reasonable degrees of success. These are parametric clustering, neural networks, and model fitting. The paper by Richard L. White reviews the problem of automatic star-galaxy classification using parametric clustering methods with a brief mention of neural networks. I will make few comments about his paper and add a brief description on model fitting.

I concur with Dr. White's comments about the strengths and weaknesses of neural networks. In particular, it is hard to interpret how the information in the observed image is used to reach particular classifications. The training of such classifiers is also a time consuming task. Because neural networks are an important area of research in pattern recognition there will continue to be applications of this approach to astronomical object recognition but it remains to be seen if this is the best or most general technique.

Parametric clustering methods are based on some set of parameters or features, which are generally integrated properties of the array of brightness values. Some examples of such parameters are the total brightness and a characteristic width. These are clustering methods because the idea is that the set of parameters cluster around certain points in the space of the parameters which define the various astronomical classes. The paper presents a good review of various ways to use parametric information.

I also like decision trees. They provide a good way to organize parameters and rules in an understandable way. The distinction between parallel and oblique trees is an obvious detail. The rules of the tree should be more complex than considering one parameter at time. There is some good work being done in how to optimally create and prune decision trees.

The invited paper does not discuss the important approach of model fitting. This is a powerful technique because it most closely matches the problem statement given at the beginning of my comments. It is especially well suited to the star-galaxy classification problem since the expected distribution of brightness values for a star is described by the common PSF.

The matching or fitting of observed brightness to a PSF model is the basis of two of the most widely used software packages in astronomy. DAOPHOT, and many similar programs, fit a PSF model to make accurate measurements of the total brightness of stars, particularly in conditions where the star images start to blend together. These do not classify though they detect data that do not fit the PSF model well.

For dealing with a mixture of stars and galaxies and classifying them a widely used program is the Faint Object Classification and Analysis System or FOCAS [Val89]. The proceedings that include the previous citation provide a good description of the various programs in astronomy that provide star and galaxy analysis. FOCAS provides classification of stars, galaxies, and other using an algorithm called *resolution classification* [Val82]. The algorithm makes classifications exactly as defined in the first paragraph. It uses the observed brightnesses and uncertainties and models of the expected brightnesses for different classes or templates to compute the most probable assignment using Bayesian methods.

In the specific implementation the templates are derived from the empirical PSF which is the average of a subset of objects thought to be stars. The choice of objects is analogous to providing a training set as discussed by White. The various classes have expected brightness distributions that have the same form as the PSF but are stretched and compressed in scale. Finding a particular template that maximizes the probability (i.e. best fits the observation) is translated to a classification by considering how close the template comes to a perfect PSF. In effect the templates measure how likely the observation is to be resolved (bigger than the PSF) or to be noise (too narrow to match the PSF).

Obviously the templates do not represent galaxies so this classifier only separates stars from galaxies. In principle, galaxy templates could be used.

The Bayesian classification method has advantages in addition to matching the description of the classification model most directly. These are that it directly uses the uncertainties and knowledge of the noise properties of the observation. It can also directly supply relative probabilities for the classification. (In FOCAS this is not done because the calculation is prohibitively slow in the current implementation.) As White noted these are important properties for a classifier and ones that are only now being looked at for parametric clustering and neural network methods.

My principle conclusion on the subject of star-galaxy classification are that all the various methods have been shown to work well by their practitioners. No method has been shown to be clearly superior. It is my feeling that the most robust method is one that takes results using all the methods – neural nets, model or PSF fitting, parameter clustering – and combines them together using a decision tree. Decision trees are very attractive because they are understandable and can use a variety of criteria for obtaining a classification. The derivation of a final probability in decision trees using information about the noise and uncertainties in the observation needs to be developed and demonstrated.

Finally I want to point out that the future challenges in automated classification are to move beyond separating stars and galaxies to automatic classification of various types of galaxies and morphologies. Various methods are also under investigation which fall into similar types as discussed above. These are parametric clustering, model fitting including bulge-disk decomposition, principle component analysis, and neural networks.

REFERENCES

- [Val82] Francisco Valdes. Resolution classifier. In David L. Crawford, editor, *SPIE Vol. 331 Instrumentation in Astronomy IV*, pages 465–472. SPIE, SPIE, March 1982.
- [Val89] Francisco Valdes. Faint object classification and analysis system standard test images results. In P. J. Grosbøl, F. Murtagh, and R. H. Warmels, editors, *1st ESO/ST-ECF Data Analysis Workshop (ESO Conference and Workshope Proceedings No. 31)*, pages 35–67. ESO, ESO, September 1989.

Recent Advances in Large-scale Structure Statistics

Vicent J. Martínez¹

ABSTRACT I review the most recent redshift surveys used to probe the large scale structure of the Universe. Then I provide an overview of some of the statistical tools used to describe the galaxy distribution, trying to connect these measures with some of the statistics used in the mainstream of spatial statistics. Special topics include intensity functions, topology, and second-order statistics (2-point correlation function, K -function).

9.1 Introduction

The study of the large-scale structure (LSS) of the Universe is one of the most active fields in modern astronomy. One of the most interesting aspects of this subject is the statistical description of the cosmological observations.

“Cosmography”, the mapping of the distribution of galaxies and clusters of galaxies in our local part of the Universe, has provided in the past recent years a view of the Universe where galaxies lie in knots, filaments and walls surrounding big voids almost devoid of luminous matter. This view has led to new methods of describing the statistical and geometrical properties of the clustering. Recent galaxy redshift surveys and future observations such as the SDSS and 2df projects will produce an explosion in the quantity and quality of the cosmological data. The new data will demand new methods for the statistical analysis of the spatial pattern revealed by the surveys.

There are several theoretical models for describing the formation of the large-scale structure. N -body simulations are also used to describe the observed Universe. One theory or simulation is usually rejected when the statistical analysis of the model gives results which are not compatible with the results of applying the same statistics to the real data. Therefore the statistics used to describe the large-scale structure need to be [Bo96]

¹Departament d’Astronomia i Astrofísica. Universitat de València. E-46100 – Burjassot. València. Spain. e-mail: martinez@hubble.matapl.uv.es

robust, reliable, easy to interpret and with discriminative power. Inside the field of the statistical analysis of point fields, new techniques and estimators of the clustering of spatial point patterns have been developed in the past decades.

Unfortunately, the connection between the spatial statisticians and cosmologists is not today as important as it was in the late fifties when the Berkeley statisticians Neymann and Scott carried out an intensive programme about the analysis of the Lick catalogue ([Ne52]; [Ne55]). In this review I shall summarize the present state of the art in the mapping of the large-scale distribution of galaxies. I shall then comment on some of the different statistics applied to describe the clustering, trying to break the language barrier between cosmologists and statisticians working in the field of point processes. In the following table we can see the equivalences between the terms used in both scientific communities:

Terms used in spatial statistics	Terms used in cosmology
Intensity function	Density field
Fourth Minkowski functional	Topology of the LSS
Second-order intensity function	Two-point correlation function
Second moment measure	Correlation integral
Empty space function	Void probability function

TABLE 9.1. Some statistical measures

9.2 Galaxy redshift surveys

The three-dimensional view of the distribution of matter in the Universe comes mainly from the redshift surveys. The redshift z of a galaxy is just the relative variation between the wavelength of the observed and the emitted radiation $z = (\lambda_o - \lambda_e)/\lambda_e$. The redshift is directly measured from the displacement of the emission or absorption lines of the galaxy spectrum. For small recession velocities v , this is just the Doppler shift $z = v/c$ (valid for $z < 0.1$). Hubble law states that the recession velocity is proportional to the distance ($v = H_0 r$). With this simple relation, redshifts are transformed into distances. However the Hubble constant, H_0 , is still not accurately known. Moreover its actual value is also a matter of an interesting debate [Pe96]. It is accepted that $50 < H_0 < 100 \text{ km s}^{-1} \text{ Mpc}^{-1}$. In Cosmology distances are usually expressed in $h^{-1} \text{ Mpc}$ (1 pc being 3.26 light years), where h is the Hubble constant in units of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$. Therefore $h \in [0.5, 1]$ reflects our ignorance of the distance scale.

During the past two decades systematic collections of redshifts have been compiled following different strategies. Today the possibility of using multi-

fiber spectrographs permits to increase the number of available redshifts very rapidly. Some common features which affect all the surveys are summarized here:

1. Usually the catalogs contain all the galaxies lying in a particular region of the sky down to a given flux limit or to a given diameter. This fact implies that when performing statistical analysis one has to choose between two strategies:
 - a) To extract volume-limited samples by keeping only galaxies intrinsically brighter than $M = m_{\text{lim}} - 25 - 5 \log(D_{\text{max}})$ where m_{lim} is the apparent-magnitude limit of the survey and D_{max} is the maximum depth of the sample.
 - b) To use the selection function $\varphi(x)$, which gives an estimate of the probability that a galaxy at a distance x is included in the sample. With this strategy each galaxy has a weight $w = 1/\varphi(x)$ depending on its distance x to us. To derive the selection function one has to make strong assumptions about the distribution of galaxies with different luminosities.
2. There exists a zone of avoidance, because the dust in our own Galaxy absorbs most of the light coming from extragalactic sources. Therefore catalogues are incomplete below galactic latitudes $|b| < 20^\circ - 30^\circ$.
3. Positions extracted from redshifts are distorted by the peculiar motions, because the observed radial velocity is due not only to the Hubble flow but also to the peculiar velocities. The effect of this radial distortion is clearly illustrated when dense clusters of galaxies, almost spherical in real space, appear as structures elongated along the line of sight, in redshift space. These structures are known as ‘fingers of God’.
4. In the surveys there are different kinds of galaxies. Galaxies have intrinsic properties such as luminosity, which varies in a range going from $3 \times 10^5 L_\odot$ to $2 \times 10^{10} L_\odot$, and morphology (spirals, ellipticals, irregulars, etc). There are supergiant galaxies like the big cD elliptical galaxies with $10^{13} M_\odot$ and 2 Mpc diameter and dwarf galaxies with $10^5 M_\odot$ and 1 kpc diameter. The statistical properties of the spatial distribution of different kinds of objects can be different. In fact, it is well established that elliptical galaxies are more frequent in denser regions such as rich clusters, while spirals are more often found in low-density environments.

Now I shall briefly describe some of the wide-angle redshift surveys used up to now and some of the projects already going ahead to map the Universe.

1. **Center for Astrophysics (CfA) redshift survey.** The Catalog of Galaxies and Clusters of Galaxies elaborated by F. Zwicky and co-workers ([Zw61–68]) contains 29363 galaxies down to a blue magnitude of 15.5. The angular position of each galaxy together with its apparent blue magnitude and other properties are listed in the compilation. It covers the whole northern hemisphere (equatorial declination $\delta \geq 0^\circ$). The CfA-II redshift survey is drawn from this catalog. The strategy has been to measure redshifts of galaxies in slices 6° wide in declination and 9 hours in right ascension ([dL86]; [Ge89]). Now it is complete over several contiguous slices and covers a large region of both Northern and Southern Galactic hemispheres, mapping a solid angle of 2.95 sr : $8^h < \alpha < 17^h$, $8.^{\circ}5 < \delta < 44.^{\circ}5$ in the North with 6500 galaxies and $20^h < \alpha < 4^h$, $-2.^{\circ}5 < \delta < 48^\circ$ in the South with 4283 galaxies.
2. **Southern Sky Redshift Survey (SSRS).** This catalog is the counterpart of the CfA catalog for the Southern Celestial hemisphere ($\delta < 0^\circ$). It is derived from plate scans and is also magnitude limited to $m_B \leq 15.5$ [dC94]. It is complete in the declination range $-40^\circ \leq \delta \leq -2.^{\circ}5$ and Galactic latitude $b \leq -40^\circ$. It covers a solid angle of 1.13 sr .
3. **Pisces–Perseus sample.** It is also based in the Zwicky catalog, and points into the direction of a huge super-cluster. Redshifts have been measured at 21cm with the Arecibo radiotelescope. It is complete in the region of the sky with right ascension $22^h < \alpha < 4^h$ and declination $0^\circ < \delta < 45^\circ$, comprising 5183 galaxies. The average redshift of these three catalogs is $\langle z \rangle \sim 0.02$ and redshifts have been generally measured one at a time.
4. **IRAS catalogs.** Several catalogs are based on the *IRAS* Infrared Astronomical Satellite (see [Ro96] and references therein). In the infrared part of the spectrum the Galactic extinction is rather low and therefore the sky coverage is bigger (up to 96%). However, elliptical galaxies are undersampled. There are already four redshift surveys extracted from this database:
 - **QDOT.** This is a sparse sample. It contains one in six of all *IRAS*-point sources with $60 \mu\text{m}$ flux, S_{60} , exceeding 0.6 Jy (see Fig. 1). There are 2387 sources ([Ro90]; [Ma94]).
 - **2 Jy.** Galaxies brighter than 1.94 Jy. It contains 2685 galaxies [Str90].
 - **1.2 Jy.** Galaxies brighter than 1.2 Jy. It contains 5339 galaxies [Fi92].

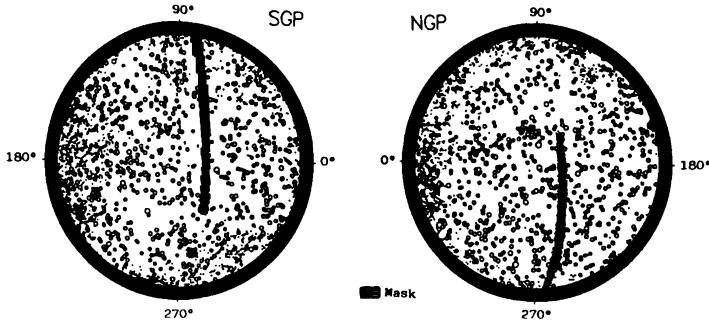


FIGURE 1. A subsample of the *IRAS-QDOT* survey. The equal-area projections are centered at the North and South Galactic Poles respectively. A band corresponding to galactic latitudes $|b| < 10^\circ$ has been excluded. Each circle corresponds to a galaxy and its size has been scaled proportionally to $1/\log(x)$, x being its distance.

- **Point Source Catalog (PSC-z).** It is the complete QDOT sample with 15000 sources exceeding 0.6 Jy. It will be soon available.

- 5. **Las Campanas redshift survey (LCRS)** This is a very deep survey with an average redshift of $\langle z \rangle \sim 0.1$. Redshifts have been measured using multi-fiber optics. The catalog consists in six slices with 1.5×80 degrees each and contains 23697 galaxies. The whole catalog is now available together with some references and scientific results in <http://manaslu.astro.utoronto.ca/~lin/lcrs.html>.

- 6. **Sloan Digital Sky Survey (SDSS)** This will be the largest galaxy survey ever compiled (see [Gu95]). The observations will be carried out in a totally dedicated to this task 2.5m telescope in New Mexico provided with two double fiber spectrographs with 320 fibers. The solid angle covered by the survey will be around π sr. Redshifts for more than one million galaxies will be measured. For details see the web page (<http://www-sdss.fnal.gov:8000/>).

- 7. **2 degree field (2df)** This survey will be operated by the Anglo-Australian telescope. Redshifts of about 250,000 galaxies brighter than $m_B = 19.5$ will be measured, covering 1700 square degrees of the sky. For details see (<http://msowww.anu.edu.au/~colless/2dF/>) In [Guz96] we can see a plot with an equal-area Aitoff projection showing the location of the deeper surveys mentioned above.

9.3 First order characteristics

9.3.1 Density field

When cosmologists estimate the density field from the discrete galaxy distribution, they are doing a kernel estimation of what spatial statisticians call first-order intensity function. Let $\{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\}$ be the positions in 3D of N galaxies in a bounded region W . The estimator of the intensity function with kernel κ_s (also called filter function) and smoothing radius (band width) $s > 0$ is

$$\hat{\lambda}_s(\vec{x}) = \sum_{i=1}^N \kappa_s(\vec{x} - \vec{x}_i), \quad \vec{x} \in W \quad (9.1)$$

where κ_s is a symmetric density probability function.

It is well known in the field of point processes that the election of the kernel function is not an important matter. However, the chosen value for the smoothing radius s has strong consequences in how the reconstructed density field looks.

The standard kernel function used in Cosmology is the Gaussian filter

$$\kappa_s(\vec{y}) = \frac{1}{(2\pi)^{d/2} s^d} \exp\left(-\frac{|\vec{y}|^2}{2s^2}\right) \quad (9.2)$$

In the next section we shall see how the estimation of the intensity might affect some conclusions which arose from the study of the topology of the large scale structure.

9.3.2 Topology of the large-scale structure

With this name cosmologists refer to the art of measuring the degree of connectivity of the large-scale structure of the Universe once the redshift survey has been smoothed with an appropriate filter function as explained above. This is done by means of the topological genus ([Go86]; [Me90]). The genus of a surface g is basically

$$g = (\text{number of wholes}) - (\text{number of isolated regions}) + 1 \quad (9.3)$$

For example, a sphere has topological genus equal to 0, a torus has genus +1, while N disjoint spheres have $g = -(N - 1)$.

A more formal definition is based in the Gauss-Bonnet theorem which gives a relationship between the curvature of the surface and its topological genus. According to this theorem, the integral of the Gauss curvature of a compact two-dimensional surface is

$$\mathcal{C} = \int K dS = 2\pi\chi = 4\pi(1 - g) \quad (9.4)$$

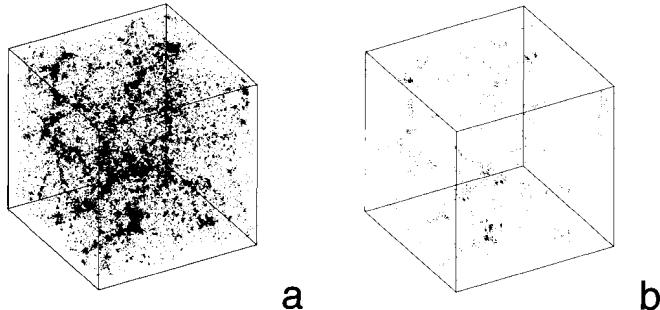


FIGURE 2. a) A 32^3 CDM simulation in a box of side $80 h^{-1}$ Mpc. b) A sample of 762 “galaxies” extracted from this simulation.

where χ is the Euler-Poincaré characteristic (related with the fourth Minkowski functional) ([St87], [Ke96]).

Gott et al. (1986) proposed an algorithm for calculating the topological genus of an isodensity surface. Isodensity surfaces are specified by means of the fraction f of the volume contained in regions with density exceeding a given threshold. We can also use the number of standard deviations ν that a given density threshold is above or below the average density.

$$f = \frac{1}{\sqrt{2\pi}} \int_{\nu}^{\infty} e^{-t^2/2} dt \quad (9.5)$$

The topological analysis consists in studying how the genus of an isodensity surface varies with f or ν . If these curves are symmetric we have Gaussian topology corresponding to a “sponge-like” surface. If the curves are biased to the left the topology corresponds to isolated clusters superimposed on a smooth background. It is referred to as “meatball” topology. Finally if the curve is biased to the right, the topology is of the “Swiss-cheese” kind with empty regions surrounded by one connected high density region.

In order to see how the estimation of the intensity with different band widths affects the results of the topological analysis, we have used a CDM (Cold Dark Matter) N -body simulation [Co96]. We have extracted galaxies from the simulation by means of a new algorithm [Ma96b]. The resulting sample contains 762 galaxies in a box of side $80 h^{-1}$ Mpc (see Fig. 2). The density of this sample is of the same order of the density of volume-limited samples extracted from the CfA-I catalog. The results of applying the topological analysis to this simulation with different smoothing radius are shown in Fig. 3. We can see how the shape of the genus curve is clearly affected by the election of the band width s . An optimal selection of s when one wants to estimate the intensity function from a point process is a well studied problem in the mainstream of spatial statistics [Di83, Di85, Cr91,

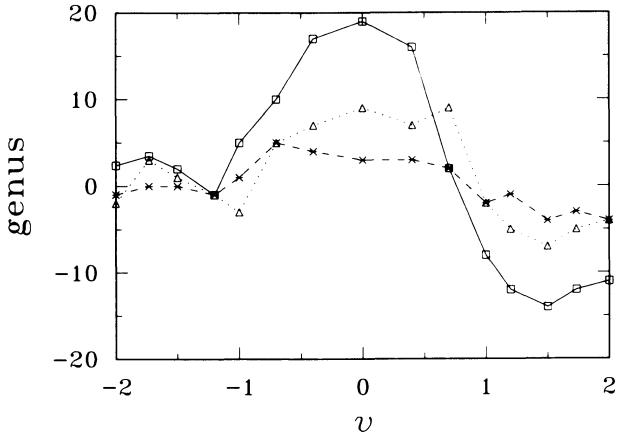


FIGURE 3. The genus function corresponding to the sample shown in Fig. 2b calculated for three different values of the band width $s \simeq 3.5 h^{-1}$ Mpc (solid line), $s \simeq 5.3 h^{-1}$ Mpc (dotted line) and $s \simeq 7.1 h^{-1}$ Mpc (dashed line).

St94] and it is a clear example where the collaboration between astronomers and statisticians will be useful.

9.4 Second order characteristics

9.4.1 Correlation function

I shall start this section by using the terminology and notation often employed by spatial statisticians and I shall relate it with that used by cosmologists.

The second order intensity function [St94, Di83] $\lambda_2(\vec{x}_1, \vec{x}_2)$ is defined by means of the probability that in two infinitesimal volumes dV_1 and dV_2 centered in \vec{x}_1 and \vec{x}_2 lies a point of the point process

$$dP = \lambda_2(\vec{x}_1, \vec{x}_2) dV_1 dV_2. \quad (9.6)$$

If the point field is stationary and isotropic, $\lambda_2(\vec{x}_1, \vec{x}_2)$ depends only on the distance $r = |\vec{x}_1 - \vec{x}_2|$. The *two-point correlation function* commonly used in Cosmology is [Pe80]

$$\xi(r) = \frac{\lambda_2(r)}{n^2} - 1. \quad (9.7)$$

where n is the mean number density of galaxies in a fair sample of the Universe.

At small distances, $0.1 < r < 10 h^{-1}$ Mpc, the galaxy two-point correlation function shows a behaviour compatible with a power-law [Da83, Da88, Ma93].

$$\xi_{gg}(r) = \left(\frac{r}{r_g} \right)^{-\gamma}, \quad (9.8)$$

where the exponent $\gamma \simeq 1.8$ and the correlation length $r_g \simeq 5 h^{-1}$ Mpc.

9.4.2 Estimators of $\xi(r)$

Different estimators may be used to evaluate $\xi(r)$. For volume-limited samples all the galaxies have the same weight $w = 1$, while when selection functions are used to account for the incompleteness of the sample, each galaxy is counted with a weight $w \geq 1$.

1. Davis & Peebles (1983) use the estimator

$$1 + \xi_{DP}(r) = \frac{DD(r)}{DR(r)} \frac{N_R}{N_D}, \quad (9.9)$$

where $DD(r)$ is the number of pairs with separation between r and $r + dr$ in the galaxy catalogue containing N_D galaxies and $DR(r)$ is the number of pairs with separation between r and $r + dr$ formed by the data and a random sample with N_R points distributed in the same volume as the data.

2. Equivalently we can use [Riv86, Ma93] the following estimator by averaging over the N galaxies of the sample

$$1 + \xi_R(r) = \frac{1}{N} \sum_{i=1}^N \frac{N_i(r)}{nV_i(r)}, \quad (9.10)$$

where $N_i(r)$ is the number of galaxies lying in a shell of thickness dr at distance r from galaxy i and $V_i(r)$ is the volume of the part of the shell lying within the sample boundaries.

3. Martínez et al. (1996) are using the following estimator proposed by [St96], which makes use of a smoothed version $\widetilde{DD}(r)$ of $DD(r)$ obtained by means of a kernel function $k(x)$. The Epanechnikov kernel will be used here

$$k(x) = \begin{cases} \frac{3}{4h} (1 - \frac{x^2}{h^2}) & \text{for } |x| \leq h \\ 0 & \text{otherwise} \end{cases}. \quad (9.11)$$

$$1 + \xi_S(r) = \frac{1}{4\pi r^2 n^2} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N \frac{k(r - |\vec{x}_i - \vec{x}_j|)}{V(W \cap W_{|\vec{x}_i - \vec{x}_j|})} \quad (9.12)$$

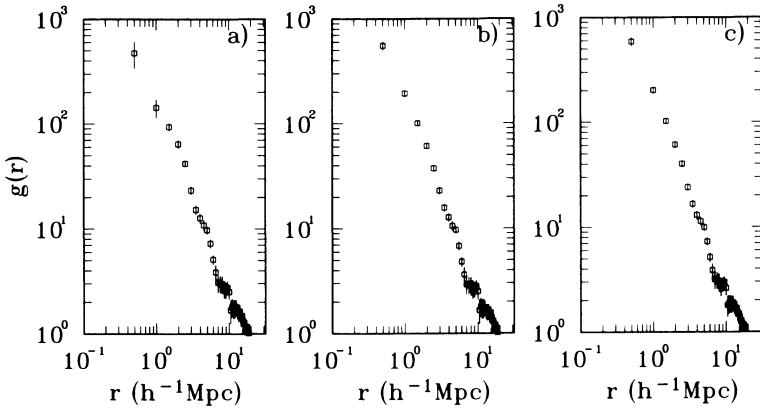


FIGURE 4. The correlation function $g(r) = 1 + \xi(r)$, calculated over the model sample by means of the 3 estimators: a) Davis & Peebles; b) Rivolet and c) Stoyan & Stoyan.

where V denotes the volume occupied by the sample and $W_{\vec{y}}$ is the window W shifted by \vec{y} , $W_{\vec{y}} = \{\vec{x} : \vec{x} = \vec{z} + \vec{y}, \vec{z} \in W\}$. This acts as an edge-correction which provides us with an unbiased estimator.

In Fig. 4 we show how these different estimators act on the galaxies of our model sample. At large scales they give approximately the same results. There are noticeable differences at small distances. The use of the random sample in the *DP* estimator produces larger errors at small scales. The plotted errors come from resampling on Cox processes having similar $\xi(r)$ [Ma96b].

9.4.3 The K function and the correlation dimension

The expected number of points within a distance r from an arbitrary given galaxy is

$$\langle N \rangle_r = \int_0^r 4\pi n s^2 (1 + \xi(s)) ds = \frac{4\pi}{n} \int_0^r s^2 \lambda_2(s) ds. \quad (9.13)$$

The last expression may also be referred to as the correlation integral $C(r)$ [Ma95]. $K(r) = C(r)/n$ is called the K -function [Ri81] and is extensively used in the literature of point fields. We can estimate $K(r)$ by means

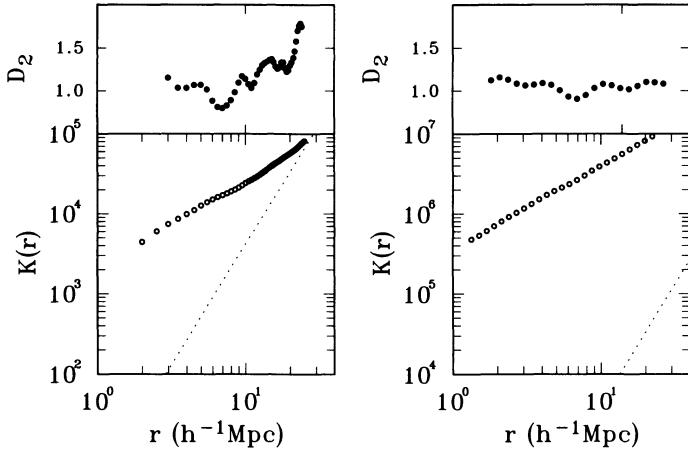


FIGURE 5. The K function and its local log-log slope D_2 as a function of the scale for the model sample (left panel). In the right panel we see the same for the Soneira & Peebles fractal model. Dotted line is the Poissonian case.

of [Ba93]):

$$\hat{K}(r) = \frac{V}{N^2} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N \frac{\theta(r - |\vec{x}_i - \vec{x}_j|)}{\omega_{ij}}, \quad (9.14)$$

where θ is Heaviside's step function, and ω_{ij} is the proportion of the area of the sphere centered at \vec{x}_i and passing through \vec{x}_j that is contained in W . Note that the sum in Eq.(9.14) is an unbiased estimator of $n^2 V K$ ([Ri76]; [St94]); however, we introduce a certain bias when we estimate n through N/V . Consequently K , being an integral quantity, does not suffer of the hindrance of splitting the possibly sparse information into disjoint bins, as the differential quantity ξ does [Ri92].

If scaling of the first moment of the count of neighbours holds, then $K(r)$ is proportional to r^{D_2} where D_2 is the correlation dimension and it tends to 3 when homogeneity is reached. Once we compute K , we obtain the local dimensions D_2 as the slope of a five-point log-log linear regression on the function $K(r)$.

We have calculated K and D_2 for our model sample and the results are shown in Fig. 5 (left panel). The expected behaviour of $K(r)$ for a uniform Poissonian distribution is shown with dotted line. For comparison we show in Fig. 5 (right panel) the same functions calculated on the Soneira–Peebles (SP) model. It is a hierarchical fractal dust with the parameters chosen to obtain $D_2 = 1$ ([So78]; [Ma90]). In the SP model, homogeneity is not reached at all, having $K(r)$ a power-law behaviour with exponent $D_2 \simeq 1$

in the whole range of scales depicted in the plot. Instead, the K function corresponding to the CDM model shows a power-law shape at small scales ($r \leq 10 h^{-1}$ Mpc) with exponent $D_2 \simeq 1$ (with some fluctuations) and then a transition to homogeneity is appreciated in both, the approach of $K(r)$ to the Poissonian line and the increasing behaviour of D_2 with the scale [Ma96a].

9.5 Conclusions

We have summarized the available redshift surveys used to study the distribution of matter in the Universe. From our point of view, these surveys are a challenge for statisticians working in the field of point processes, because despite their intrinsic shortcomings (truncation, distortion, segregation) they contain precious information about how the Universe has evolved.

We have reviewed some of the statistical tools used to probe the large-scale structure of the Universe. We have focused our attention in the methodological aspects rather than the results of applying the statistics described here to the real catalogs. This latter aspect can be easily found in the literature. We have tried to make clear the connections between the statistics and their estimators used by cosmologists and statisticians. It is our hope that future collaborations between the two scientific communities will produce soon fruitful outcome. I am sure that this meeting will encourage such collaborations.

Acknowledgments: I thank my collaborators María-Jesús Pons-Bordería, Dietrich Stoyan, Helga Stoyan, Enn Saar, Silvestre Paredes and Rana Moyeed for their contribution to the work presented here and for their permission for using part of our common unpublished results. A good part of this work was written in the relaxing atmosphere of a visit to the Kapteyn Astronomical Institute in Groningen. I want to thank their hospitality. This research has been supported by the Conselleria d'Educació i Ciència de la Generalitat Valenciana (grant number GV-2207/94) and by the EC Human Capital and Mobility Programme network (Contract ERB CHRX-CT93-0129).

REFERENCES

- [Bo96] Borgani, S., 1996, in the Enrico Fermi School Proceedings: Dark Matter in the Universe. Eds. S. Bonometto, J. Primack & A. Provenzale, IOP publishing.
- [Ba93] Baddeley, A.J., Moyeed, R.A., Howard, C.V., & Boyde, A. 1993, *Appl. Statist.*, 42, 641.

- [Co96] Couchman, H.M.P., 1996, Cosmological Simulations Using Particle-Mesh Methods, Springer-Verlag (in press).
- [Cr91] Cressie, N., 1991, Statistics for Spatial Data, New York: J. Wiley & Sons.
- [dC94] da Costa et al., 1994, ApJ 424, L1.
- [Da83] Davis, M. & Peebles, P.J.E., 1983, ApJ 267, 465.
- [Da88] Davis, M., Meiksin, A., Strauss, M.A., da Costa, L.N. & Yahil, A., 1988, ApJL 333, 9.
- [dL86] de Lapparent, V, Geller, M.J & Huchra, J.P. 1986 ApJ 302, L1.
- [Di83] Diggle, P.J., 1983, Statistical Analysis of Spatial Point Patterns, London: Academic Press.
- [Di85] Diggle, P.J. 1985, Appl. Statist., 34, 138.
- [Fi92] Fisher, K.B., Strauss, M.A., Davis, M., Yahil, A., & Huchra, J.P., 1992, ApJ 389, 188.
- [Ge89] Geller, M.J & Huchra, J.P., 1989, Science 246, 897.
- [Go86] Gott, J.R., Melott, A.L., & Dickinson, M., 1986, ApJ 306, 341.
- [Gu95] Gunn, J.E., and Weinberg, D.H., 1995 in Wide-Field Spectroscopy and the Distant Universe. Eds. S.J Maddox & A. Aragón-Salamanca, World-Scientific (Singapore).
- [Guz96] Guzzo, L., 1996, in Mapping, Measuring and Modelling the Universe. Eds. P. Coles, V.J. Martínez & M.J. Pons-Bordería, ASP conference series.
- [Ke96] Kerscher, M., Schmalzing, J., & Buchert, T., 1996 in Mapping, Measuring and Modelling the Universe. Eds. P. Coles, V.J. Martínez & M.J. Pons-Bordería, ASP conference series.
- [Ma90] Martínez, V.J., Jones, B.J.T., Dominguez-Tenreiro, R. & van de Weygaert, R., 1990 ApJ 357, 50.
- [Ma93] Martínez, V.J., Portilla, M., Jones, B.J.T. & Paredes, S., 1993 AA 280, 5.
- [Ma94] Martínez, V.J., & Coles, P., 1994 ApJ, 437, 550.
- [Ma95] Martínez, V.J., Paredes, S., Borgani, S. and Coles, P. 1995 Science 269, 1245.
- [Ma96a] Martínez, V.J., Pons-Bordería, M.J., & Moyeed, R., 1996 (submitted).
- [Ma96b] Martínez, V.J., Stoyan, D., Stoyan, H., Saar, E. & Pons-Bordería, M.J., 1996 (submitted to MNRAS).
- [Me90] Melott, A.L., 1990, Phys. Rep. 193, 1.
- [Ne52] Neymann, J., & Scott, E., 1952, ApJ 116, 144.
- [Ne55] Neymann, J., & Scott, E. 1955, AJ 60, 33.
- [Pe80] Peebles, P.J.E., 1980, The Large Scale Structure of the Universe, Princeton University Press.
- [Pe96] Peebles et al., 1996, in Critical Dialogues in Cosmology. Princeton University Press.
- [Ri76] Ripley, B.D. 1976, J. Appl. Prob. 13, 255.
- [Ri77] Ripley, B.D. 1977, J. Royal Statist. Soc. 39, 172.
- [Ri81] Ripley, B.D. 1981, Spatial Statistics, New York: John Wiley & Sons.
- [Ri92] Ripley, B.D. 1992, in Statistical Challenges in Modern Astronomy. Eds. E.D. Feigelson & J. Babu.
- [Riv86] Rivolo, A.R., 1986, ApJ 301, 70.
- [Ro90] Rowan-Robinson et al., 1990, MNRAS 247, 1.

- [Ro96] Rowan-Robinson, M., 1996 in *Mapping, Measuring and Modelling the Universe*. Eds. P. Coles, V.J. Martínez & M.J. Pons-Bordería. ASP conference series.
- [So78] Soneira, R.M. & Peebles, P.J.E. 1978 AJ 83, 845.
- [St87] Stoyan, D., Kendall, W.S. & Mecke, J., 1987, *Stochastic Geometry and its Applications*, New York: John Wiley & Sons.
- [St94] Stoyan, D., & Stoyan, H., 1994, *Fractals, Random Shapes and Point Fields*, Chichester: John Wiley & Sons.
- [St96] Stoyan, D., & Stoyan, H. 1996, Biom. J. 38, 259.
- [Str90] Strauss, M.A., Davis, M., Yahil, A., & Huchra, J.P., 1990, ApJ 361, 49.
- [Zw61–68] Zwicky et al, 1961–1968, *Catalogue of Galaxies and of Clusters of Galaxies*, Pasadena: California Institute of Technology.

Discussion by Michael L. Stein²

Introduction

Despite the fact that galaxy clustering received considerable attention at the first conference on Statistical Challenges in Modern Astronomy, the statistical issues arising in studying this problem have only barely begun to be addressed. Vicent Martínez describes the current state of statistical methodology in this area, including some ideas from the recent statistical literature. It appears that statisticians' contributions to date have mainly concerned how to better estimate quantities of interest to astronomers. While this is important, I believe it is at least as important for statisticians to learn enough astronomy so they can become engaged in questions of what should be estimated.

Density field

As an example of when interesting statistical issues of both “how” and “what” arise, consider estimating the density field or first order intensity for galaxies. Martínez notes that kernel methods are commonly used for this problem. Note that the usual assumption statisticians have in mind when using such methods is that the observations are independent with a common density $\lambda(\vec{x})$. For now, suppose this makes sense as a model

²Department of Statistics, University of Chicago, 5734 University Ave., Chicago, IL 60637. e-mail: stein@galton.uchicago.edu

for galaxy locations and consider the question of how to estimate $\lambda(\vec{x})$. Martínez gives the estimator

$$\hat{\lambda}_h(\vec{x}) = \sum_{i=1}^N \kappa_h(\vec{x} - \vec{x}_i), \quad \vec{x} \in V,$$

where, for example, in three dimensions,

$$\kappa_h(\vec{x}) = \frac{1}{(2\pi)^{3/2} h^3} \exp\left(-\frac{|\vec{x}|^2}{2h^2}\right).$$

However, such an estimator will not work well in regions where there are rapid changes in the density, which might occur near the boundary of a filament or wall. Allowing h to depend on \vec{x} , which statisticians call adaptive density estimation ([Si86], [Be92]) would help somewhat. More generally, one could consider a kernel of the form

$$\kappa_A(\vec{x}) = \frac{1}{(2\pi)^{3/2} \det(A)^{1/2}} \exp\left(-\frac{1}{2} \vec{x}^T A^{-1} \vec{x}\right),$$

where A is a nonsingular 3×3 matrix. This kernel reduces to the previous isotropic form when A equals h^2 times the identity matrix. By allowing A to depend on \vec{x} , one could allow for locally smoothing the observations more in some directions than others depending on the local structure. For example, for \vec{x} in the interior of a thin wall, it would make sense to smooth more in directions along the wall than in directions perpendicular to the wall. How one would actually do this in practice deserves further study. However, no kernel method, even an adaptive one, will do well very near a place where there is a sharp change in density. Some procedure for segmenting the universe into regions such that the density varies smoothly within each region and then estimating the density separately within each region (making appropriate adjustments when near a boundary [Si86]) might do better.

Let us next consider what we are estimating by $\hat{\lambda}_h(\vec{x})$. If, as is often done, one assumes galaxy locations form a stationary point process, then $\lambda(\vec{x})$ does not depend on \vec{x} , so that using an estimate of $\lambda(\vec{x})$ that varies in \vec{x} would appear to be silly. One possible way to reconcile this difference between the spatially varying estimate and the presumed constant true density is to imagine that galaxy locations are a stationary Cox, or doubly stochastic, process ([Di83], [Da88]). A stationary Cox process is obtained by first generating a realization $\lambda(\cdot)$ of a stationary and nonnegative random intensity function $\Lambda(\cdot)$ and then taking the point process to be an inhomogeneous Poisson process with intensity function $\lambda(\cdot)$. We see that using kernel estimation for estimating a spatially varying density corresponds to conditioning on $\lambda(\cdot)$; unconditionally, the process has constant intensity but is not Poisson. While both the conditional and unconditional properties of the process may be of interest, it seems to me that the unconditional

properties are more directly related to the dynamics of the universe. Furthermore, note that if one were interested in $\lambda(\cdot)$, standard kernel methods do not make use of the assumption that $\lambda(\cdot)$ is a realization of a stationary process. For processes in one dimension. Diggle (1985) discusses estimating $\lambda(\cdot)$ when $\Lambda(\cdot)$ is stationary, although I disagree with his assertion that stationarity implies one should use a kernel with a constant bandwidth.

The unconditional properties of a Cox process directly follow from the probability law of the random intensity $\Lambda(\cdot)$. If there is a parametric model available for this law, it may in some cases be possible to approximate the likelihood function using recent advances in Markov Chain Monte Carlo methods ([Sm93], [Bes93], [Ge94]). Møller, Syversveen and Waagepetersen (1996) use these methods to estimate $\lambda(\cdot)$, the realized value of the intensity, when the logarithm of $\Lambda(\cdot)$ is a Gaussian process. However, there may not be available a credible parametric model for the probability law of $\Lambda(\cdot)$. In these cases, one possible approach is to estimate moments of the point process and make use of the fact that the factorial moments of the point process equal the ordinary moments of $\Lambda(\cdot)$ [Da88].

To make a fairer judgement of the value of kernel density estimates for studying galaxy clustering requires examining how these estimates are actually used. One use of density estimates is as a summary of large datasets and certainly that use applies here. As Martínez describes, density estimates are often used in galaxy clustering as part of a procedure to discriminate between various models for the universe by examining the genus of surfaces given by the contour lines of the density estimate as the value of the estimate varies. While the overall procedure is too complicated for any in-depth analysis, I suspect that very different processes could lead to rather similar curves of the type given by Martínez in Figure 3. Whether or not this procedure is effective for making the relevant distinctions is not apparent to me from the literature I have seen. Furthermore, it is not clear that developments in the statistical literature mentioned by Martínez for choosing the bandwidth can be directly applied here, since the goal of obtaining genus estimates with good discrimination is different than those considered by statisticians. Incidentally, it strikes me as a mistake to draw different conclusions about two functions on V if one is just a monotonically increasing transformation of the other, which suggests, using the notation in Equation (9.5), that genus should be plotted as a function of f rather than ν .

One final “shot-in-the-dark” on density estimation: given that part of the reason for doing this is to identify structures such as walls or filaments ([Ba92], [Co92]), I wonder if some adaptation of principal curves and surfaces ([Ha89], [Le94]) might be helpful. The basic idea of these procedures is to find curves or surfaces along which a set of observations are concentrated.

Second order characteristics

For a stationary and isotropic point process, it is natural to consider estimating the correlation structure of the process as a function of distance. This structure can be described in terms of the correlation at a specific distance r , as measured by $\xi(r)$ or $\rho(r)$, or by the number of points of the process within r of a given point, as measured by $K(r)$ or $C(r)$. In either case, as Martínez describes, it is important to correct for bias due to edge effects. If I may suggest one more edge correction, Stein (1993) describes a method for estimating $K(r)$ that has essentially the same bias characteristics as do those mentioned by Martínez but has smaller variance under various circumstances, which is demonstrated via asymptotic approximations ([St93], [St95]) and by simulations ([St91], [St93]). These results show that the variance of this procedure can sometimes be much smaller than previously suggested procedures when r is a substantial fraction of the dimensions of the observation region, which may be particularly important in the present application. This estimator is quite a bit harder to calculate than those presented by Martínez (indeed, it is not so easy to describe), requiring numerical integrations even for simple observation regions. However, given the effort and expense used in collecting the data, there is no excuse for not doing the extra calculations, even if it turns out that this procedure leads to the same scientific conclusions. The same basic approach as described in Stein (1993) can also be used for edge-correcting estimates of $\xi(r)$, although in this case there is not presently any direct evidence that it works better than other edge-correction methods.

While the edge effects are the main potential source of bias for estimates of $K(r)$, for $\xi(r)$ there is an additional source of bias from smoothing caused by averaging over pairs of points close to but not exactly r apart in order to estimate $\xi(r)$. Thus, when Martínez says that $\hat{\xi}(r)$ as defined in Section 4.2 is unbiased, that is not quite correct. It is free of bias due to edge effects but not due to smoothing, nor is it possible to entirely remove this smoothing bias.

A more serious problem is providing believable standard errors for estimates of either K or ξ . For estimates of ξ , Stoyan, Bertram & Wendrock (1993) suggest using a Poisson approximation to the number of pairs of points very nearly r apart. While this approximation is plausible, in practice it can work poorly, especially for clustered processes [Stoyan, Bertram & Wendrock 1993]. I do not know of any practically useful method for obtaining the standard error for $\hat{K}(r)$. It is possible to use bootstrapping, but I am skeptical that it can be successfully applied when r is more than a small fraction of the dimensions of the observation region. With the presently available data sets, I suspect that it is not possible to obtain reliable standard errors without making strong parametric assumptions about the nature of the process.

Other issues

Martínez describes two noteworthy aspects of the process of measuring the locations of galaxies. One is the fact that certain galaxies are not observed because of some combination of their dimness, distance and the presence of intervening objects. The astronomers' practice of restricting analysis to galaxies whose intrinsic brightness is such that they would have been observed no matter where in the survey they were is sensible even though it does throw away information. Developing more reliable ways of using all of the data other than weighting by the reciprocal of the probability an object is observed is a good and challenging problem for statisticians to address. The other noteworthy aspect of locating galaxies is the effect of their peculiar motions on determining their distance. The errors in the radial direction due to the peculiar motions are not independent across galaxies but show spatial structure due to the fact that neighboring galaxies that are part of some dense cluster tend to have similar peculiar velocities, leading to the "finger of God" phenomenon. Describing this phenomenon by some simple model for errors with spatial correlation depending on just the actual distance between two galaxies is not adequate since the correlation presumably also depends on at least the density of galaxies in the area.

One last issue I would like to raise is the problem of reconstructing the dynamics of the universe from essentially a single "snapshot". While this problem may be so basic to astronomy as to not merit mention, it is interesting to note its existence in other disciplines. For example, in ecology, point process models are sometimes used to model the location of trees in a forest [Di83]. These locations often show clustering, but from observing the forest at a single time, it may be difficult or impossible to distinguish between the possibilities that the clustering is due to smooth spatial variations in soil fertility or to some aspect of seed dispersal. If the forest can be observed over a sufficiently long time interval, it should be possible to distinguish between these alternatives. Unfortunately, the universe changes too slowly for this approach to be of much use to any living astronomers. However, the fact that astronomers can see the past by looking far away does provide some information on how galaxy clustering has changed over time, even if it is not as direct as observing the same forest over many years. While astronomers have tried to exploit the fact that far away is also far back in time in studying, for example, quasars [Tytler 1992], perhaps future datasets will make it feasible to apply this approach to the dynamics of galaxy clustering. The statistical challenges of looking for differences in clustering over time in a way that appropriately accounts for the uncertainty in radial distances would be considerable and could keep both statisticians and astronomers busy for a long time.

Acknowledgments: This research was supported in part by the National Science Foundation grant DMS 95-04470. This manuscript was prepared using computer facilities supported in part by the National Science Foundation grant DMS 89-05292 awarded to the Department of Statistics at The University of Chicago, and by The University of Chicago Block Fund.

REFERENCES

- [Ba92] Barrow, J.D. 1992, in *Statistical Challenges in Modern Astronomy*, p. 21, Eds. E.D. Feigelson & G. J. Babu.
- [Be92] Beers, T.C. 1992, in *Statistical Challenges in Modern Astronomy*, p. 111, Eds. E.D. Feigelson & G. J. Babu.
- [Bes93] Besag, J. & Green, P.J. 1993, J. Roy. Statist. Soc. B, 55, 25.
- [Co92] Coles, P.C. 1992, in *Statistical Challenges in Modern Astronomy*, p. 57, Eds. E.D. Feigelson & G. J. Babu.
- [Da88] Daley, D.J. & Vere-Jones, D. 1988, An Introduction to the Theory of Point Processes, New York: Springer-Verlag.
- [Di83] Diggle, P.J. 1983, Statistical Analysis of Spatial Point Patterns, London: Academic Press.
- [Di85] Diggle, P.J. 1985, Appl. Stat., 34, 138.
- [Ha89] Hastie, T. & Stuetzle, W. 1989, J. Amer. Statist. Assoc., 84, 502.
- [Le94] LeBlanc, M. & Tibshirani, R. 1994, J. Amer. Statist. Assoc., 89, 53.
- [Ge94] Geyer, C.J. & Møller, J. 1994, Scand. J. Statist., 21, 359.
- [Mo96] Møller, J., Syversveen, A.R. & Waagepetersen, R. 1996, manuscript.
- [Si86] Silverman, B.W. 1986, Density Estimation for Statistics and Data Analysis, London: Chapman & Hall.
- [Sm93] Smith, A.F.M. & Roberts, G.O. 1993, J. Roy. Statist. Soc. B, 55, 3.
- [St91] Stein, M.L. 1991, Biometrika, 78, 281.
- [St93] Stein, M.L. 1993, Biometrika, 80, 443.
- [St95] Stein, M.L. 1995, Statistica Sinica, 5, 221.
- [Sto93] Stoyan, D., Bertram, U. & Wendrock, H. 1993, Ann. Inst. Statist. Math., 45, 211.
- [Ty92] Tytler, D. 1992, in *Statistical Challenges in Modern Astronomy*, p. 83, Eds. E.D. Feigelson & G. J. Babu.

Wavelet Transform and Multiscale Vision Models

Albert Bijaoui, Frédéric Rué and Renaud Savalle

ABSTRACT We have implemented multiscale vision models based on the wavelet transform to analyze field astronomical images. The discrete transform is performed by the *à trous* or the pyramidal algorithms. The vision models are based on the notion of the significant structures. Different kind of noises have been taken into account. We identify the pixels of the wavelet transform space (WTS) associated with the objects. At each scale a region labelling is carried out. An interscale connectivity graph is then established. In accordance with some rules that permit false detections to be removed, the objects and their sub-objects are identified. They define respectively trees and sub-trees in the graph. So, the identification of the WTS pixels of the tree related to a given object leads to the reconstruction of its image by the conjugate gradient method. The model has been tested successfully on astronomical images which shows that complex structures are better analyzed than using usual astronomical vision models.

10.1 Model visions in astronomy

The astronomical images contain typically a large set of point-like sources (the stars), some quasi point-like objects (faint galaxies, double stars,...) and some complex and diffused structures (galaxies, nebulous, planetary stars, clusters, etc.). A *vision model* is defined by the sequence of operations required for the automated image analysis. Astronomical images need specific ones which take into account the scientific purposes, the characteristic of the objects and the existence of hierarchical structures.

The classical vision model for robotic and industrial images is based on the edges detection. We have applied first this conception to the astronomical imagery [BLOM78]. We choose the Laplacian of the intensity as the edge line. The results are independent of large scale spatial variations, such the ones due to the sky background. The main disadvantage of the resulting model lies in the difficulty to get a correct object classification: astronomical sources can not be accurately recognized from their edges.

Many reduction procedures were built using a model for which the image is the sum of a slowly variable background with superimposed small scale objects [Sto86] [SMB⁺88] [Val89] [Kru89]. We build first a background mapping [Bij80]. For that purpose we need to introduce a scale: the background is defined in a given area. Each pixel with a value significantly greater than the background is considered to belong to a real object. A same label is given to each significant pixel belonging to the same connected field. For each field we determine the area, the position, the flux and some pattern parameters. Generally, this procedure leads to quite accurate measurements, with a correct detection and recognition. The model works very well for poor fields. If it is not the case, a labeled field may correspond to many objects. The background map is done at a given scale: larger objects are removed. The smoothing is only adapted to the star detection not to larger objects.

The classical vision models fail to arrive at a complete analysis of astronomical images because they are based on a single spatial scale for the adapted smoothing and background mapping. They are only adapted to stars or quasi stellar sources with a slowly varying background. A multiscale analysis allows us to get a background adapted to a given object and to optimize the detection of different size objects.

10.2 The discrete wavelet transforms and the multiscale vision

10.2.1 Continuous and discrete wavelet transforms

For performing a multiscale analysis we have to apply a specific transformation on the image. We choose a transformation which has the following properties: i/linear in order to easily control the data statistics, and for simple computations, ii/covariant under translations, the vision does not depend on the frame origin, iii/covariant under dilations. if the objects are dilated the vision model has to detect them in the same manner. The continuous wavelet transform satisfies these properties. Its definition for a 1D function $f(x) \in L^2(\mathbb{R})$ is [GKMM89]:

$$w(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(x) \psi^*(\frac{x-b}{a}) dx \quad (10.1)$$

$\psi(x)^*$ is the conjugate of the analyzing wavelet $\psi(x)$. $a (> 0)$ the scale parameter and b the position parameter.

The use of the wavelet transform with a computer can be foreseen through the sampling theorem [Bra65]. The wavelet transform is a set of convolutions, so if we process an image with a cut-off frequency, we just have to do some multiplications in the Fourier space. The complexity of the algorithm

is in $N \log^2 N$ for a 1D signal and faster algorithms have been searched to reduce this complexity.

10.2.2 The multiresolution analysis

The classical discrete wavelet transform results from the *Multiresolution Analysis* [Mal89] which is based on an increasing sequence of closed linear subspace V_i of $L^2(R)$. A function $f(x)$ is projected at each step i on the subset V_i . This projection is defined by the scalar product $c(i, k)$ of $f(x)$ with the function $\phi(x)$ which is dilated and translated:

$$c(i, k) = \frac{1}{2^i} \langle f(x), \phi\left(\frac{x}{2^i} - k\right) \rangle \quad (10.2)$$

$\phi(x)$ is named the scaling function of the analysis. Its main property lies in the following relation (dilation equation [Str89]):

$$\frac{1}{2} \phi\left(\frac{x}{2}\right) = \sum_n h(n) \phi(x - n) \quad (10.3)$$

This relation allows to compute the set $\{c(i, k)\}$ from $\{c(i - 1, k)\}$:

$$c(i, k) = \sum_n h(n - 2k) c(i - 1, n) \quad (10.4)$$

At each step, the number of scalar products is divided by 2. An information is lost, and step by step the signal is smoothed. The remaining information can be restored using the complementary subspace W_i of V_i in V_{i-1} . This subspace can be generated by a suitable wavelet function $\psi(x)$ with translation and dilation. We have:

$$\frac{1}{2} \psi\left(\frac{x}{2}\right) = \sum_n g(n) \phi(x - n) \quad (10.5)$$

We compute the scalar products $\frac{1}{2^i} \langle f(x), \psi\left(\frac{x}{2^i} - k\right) \rangle$, i.e. the discrete wavelet coefficients, with:

$$w(i, k) = \sum_n g(n - 2k) c(i - 1, n) \quad (10.6)$$

The 2D multiresolution analysis is generally performed separately in line and column. This does not lead to an isotropic vision, three wavelet functions are used and it is not easy to associate wavelet coefficients to a given pixel.

This discrete wavelet transform is not shift-invariant. At each scale the wavelet coefficients are decimated and it is impossible to restore the lost coefficients by a simple interpolation. Different ways exist to keep the shift-invariance. For the so-called *à trous* (with holes) algorithm [HKMMT89]

[Bij91] no decimation is done. The sampling is the same at each scale. Pyramidal wavelet transforms using a wavelet which has a cut-off frequency [SBLP94] are also shift invariant. We have also applied a pyramidal transform with a wavelet for which the values are negligible after the frequency 0.5. If the redundancy is not critical we prefer to avoid any decimation, using the *à trous* algorithm. Else, a pyramidal algorithm is applied.

10.2.3 The *à trous* algorithm

The sampled data, $F(k, l)$, are assumed to be the scalar product of the continuous function, \mathcal{F} with a scaling function ϕ :

$$F(k, l) = F(0, k, l) = \langle \mathcal{F}(x, y), \phi(x - k, y - l) \rangle \quad (10.7)$$

Let us consider the following scalar products that give the smoothed image of $F(k, l)$ at scale i :

$$F(i, k, l) = \frac{1}{4^i} \langle \mathcal{F}(x, y), \phi\left(\frac{x - k}{2^i}, \frac{y - l}{2^i}\right) \rangle \quad (10.8)$$

ϕ is chosen to satisfy the 2D dilation equation:

$$\frac{1}{4} \phi\left(\frac{x}{2}, \frac{y}{2}\right) = \sum_{n,m} h(n, m) \phi(x - n, y - m) \quad (10.9)$$

which permits to compute iteratively the $F(i, k, l)$ from a scale to the double one using the relation which the introduction of the low pass filter $H(i)$:

$$\begin{aligned} F(i, k, l) &= \sum_{n,m} h(n, m) F(i - 1, k + 2^{i-1}n, l + 2^{i-1}m) \quad (10.10) \\ &= H(i)(F(i - 1))(k, l) \end{aligned}$$

The wavelet coefficient at scale i and location (k, l) can be computed by a scalar product between the associated wavelet function ψ and \mathcal{F} . Using the projection of ψ on the basis formed by the ϕ functions, we get the following relation where appears the high pass filter $G(i)$:

$$\begin{aligned} W(i, k, l) &= \sum_{n,m} g(n, m) F(i - 1, k + 2^{i-1}n, l + 2^{i-1}m) \quad (10.11) \\ &= G(i)(F(i - 1))(k, l) \end{aligned}$$

If ψ results from the difference of two successive approximations, the wavelet coefficients can be computed very easily:

$$W(i, k, l) = F(i - 1, k, l) - F(i, k, l) \quad (10.12)$$

The algorithm allowing the initial image to be rebuilt is obvious: the last smoothed array $F(I, k, l)$ is added to all the I differences $W(i, k, l)$ (I is the number of scales).

We choose the following scaling function:

$$\phi(x, y) = B_3(x)B_3(y) \quad (10.13)$$

where B_3 is the B -spline function of degree 3 [UA92]. This function is very similar to a Gaussian one. The iterative relation (10.11) can be computed separately in row and column with the 1D mask $h(n) = \{\frac{1}{16}; \frac{1}{4}; \frac{3}{8}; \frac{1}{4}; \frac{1}{16}\}$:

10.2.4 The pyramidal algorithm

The approximation is decimated at each scale:

$$F(i, k, l) = \frac{1}{4^i} \langle \mathcal{F}(x, y), \phi\left(\frac{x}{2^i} - k, \frac{y}{2^i} - l\right) \rangle \quad (10.14)$$

Let us call $\tilde{F}(i, k, l)$ the approximation before decimation, we have the recursive expression:

$$\tilde{F}(i, k, l) = \sum_{n,m} h(n, m) F(i-1, k+n, l+m) \quad (10.15)$$

The wavelet coefficients are:

$$W(i, k, l) = F(i-1, k, l) - \tilde{F}(i, k, l) \quad (10.16)$$

and:

$$F(i, k, l) = \tilde{F}(i, 2k, 2l) \quad (10.17)$$

The reconstruction can not easily be done. We use an iterative algorithm based on interpolations and additions.

We have used also the cubic B -spline scaling function. The best results have been obtained with a small redundancy. At the first scale we do not decimate the approximation, leading to a redundancy of 2 in each direction, then we apply the pyramidal algorithm. This operation allows us to get a quite correct sampling of the wavelet transform at each scale.

A real time algorithm has been implemented. The pyramidal wavelet transform is computed simultaneously at all scales during a line by line reading. The computing time is largely reduced, and the needed memory allows us to process very large images in real time.

10.3 The multiscale vision models

10.3.1 Object definition in the wavelet transform space

An object has to be defined in the wavelet transform space (WTS). In the image, an object occupies a physical connected region and each pixel of the

region can be linked to the others. The connectivity in the direct space has to be transported to the WTS. All the structures form a 3D connected set which is hierarchically organized: at a given scale the structures are linked to smaller structures of the previous scale. This set gives the description of an object in the WTS. The steps of the multiscale model can now be defined:

After applying the wavelet transform on the image, a **thresholding in the WTS** is performed to identify the statistically significant pixels. These are regrouped in connected fields by a **scale by scale segmentation procedure**, in order to define the object structures. Then an **interscale connectivity graph** is established and the **object identification procedure** extracts each connected sub-graph that corresponds to 3D connected sets of pixels in the WTS and, by referring to the object definition, can be associated with the objects. Finally, from each set of pixels an image of the object can be reconstructed using **reconstruction algorithms**. Then, measurement and classification operations can be carried out.

10.3.2 Thresholding in the wavelet space

We have studied the distribution of the wavelet coefficients $W(i, k, l)$ for different statistics (Gauss [SB94], Poisson [SdLB93], Gauss + Poisson [MSB95], etc...). If the image is locally uniform we can compute the probability density function (PDF) of the wavelet coefficient $p(w)$. Then we can introduce a statistical meaning of the observed value from the classical decision theory [Har63]. \mathcal{H}_0 is the hypothesis that at the scale i the image is constant in the neighbourhood of the pixel (k, l) . For a positive coefficient w , the \mathcal{H}_0 rejection depends on the probability p :

$$P = \text{Prob}(W > w(i, k, l)) = \int_{w(i, k, l)}^{+\infty} p(W)dW \quad (10.18)$$

For a negative coefficient we examine:

$$P = \text{Prob}(W < w(i, k, l)) = \int_{-\infty}^{w(i, k, l)} p(W)dW \quad (10.19)$$

We fix a decision level ϵ . If $P > \epsilon$, \mathcal{H}_0 is not excluded at level ϵ , then the coefficient value can be due to the noise. On the other hand, if $P < \epsilon$, we can not consider that the value results only from the noise and \mathcal{H}_0 must be rejected at this decision level. We say that we have detected a significant coefficient. Our vision model is based only on the detected significant coefficients.

10.3.3 Scale by scale segmentation and the interscale relation

The region labelling is done by a classical growing technique. At each scale, neighboring significant pixels are grouped together to form a segmented

field. A label $n > 0$ is assigned to each field pixel. If a pixel is not significant, it does not belong to a field and its label is 0. We denote by $L(i, k, l)$ the label corresponding to the pixel (k, l) at scale i and $D(i, n)$ a segmented field of label n at the same scale.

Now we have to link the labelled fields from a scale to the following one, in order to construct a graph from which we can extract the objects. Let us consider the fields $D(i, n)$ at scale i and $D(i + 1, m)$ at scale $i + 1$. The pixel coordinates of the maximum coefficient $W(i, k_{i,n}, l_{i,n})$ of $D(i, n)$ are $(k_{i,n}, l_{i,n})$. $D(i, n)$ is said to be connected to $D(i + 1, m)$ if the maximum position belongs to the field $D(i + 1, m)$, i.e $L(i + 1, k_{i,n}, l_{i,n}) = m$. With this criterion of interscale neighbourhood, a field of a given scale is linked to at most one field of the upper scale. Now we have a set of fields $D(i, n)$ and a relation \mathcal{R} :

$$D(i, n) \mathcal{R} D(i + 1, m) \quad \text{if} \quad L(i + 1, k_{i,n}, l_{i,n}) = m \quad (10.20)$$

This relation leads to build the interscale connectivity graph whose summits correspond to the labelled fields. Statistically, some significant structures can be due to the noise. They contain very few pixels and are generally isolated, i.e connected to no field at upper and lower scales. So, to avoid false detection, the isolated fields are removed from the initial interscale connection graph.

10.3.4 The object identification

An object is associated with each local maximum of the image wavelet transform. For each field $D(i, n)$ of the interscale connection graph, its highest coefficient $W(i, k_{i,n}, l_{i,n})$ is compared with the corresponding coefficients of the connected fields of the upper scale, $W(i + 1, k_+, l_+)$ and lower scale, $W(i - 1, k_-, l_-)$.

If $W(i - 1, k_-, l_-) < W(i, k_{i,n}, l_{i,n}) > W(i + 1, k_+, l_+)$, $D(i, n)$ corresponds to a local maximum of the wavelet coefficients. It defines an object. No other fields of the scale i are attributed to the object; $D(i, n)$ concentrates the main information which permits the object image to be reconstructed. Only the fields of the lower scales connected to $D(i, n)$ are kept. So the object is extracted from larger objects that may contain it. On the other hand, some of these fields may define other objects. They are sub-objects of the object. To get an accurate representation of the object cleaned of its components, the fields associated with the sub-objects can not be directly removed; as experiments show, their images will have to be restored and subtracted from the reconstructed global image of the object. By construction, $D(i, n)$ is the root of a sub-graph which defines a tree noted \mathcal{T} . \mathcal{T} expresses the hierarchical overlapping of the object structures and

10.3.5 The object image reconstruction

Let us consider an object (or a sub-object) \mathcal{O} previously defined and its associated tree T . It corresponds to a set of wavelet coefficients \mathcal{V} defined on a 3D support \mathcal{S} in WTS:

$$\mathcal{O} \iff \{\mathcal{V}(i, k, l), \text{ for } (i, k, l) \in \mathcal{S}\} \quad (10.21)$$

where

$$\mathcal{S} = \{(i, k, l) \text{ such that } W(i, k, l) \in D(i, n) \text{ element of } T\} \quad (10.22)$$

F is an image and W is its corresponding wavelet transform. F can be considered as a correct restored image of the object \mathcal{O} if:

$$\mathcal{V}(i, k, l) = W(i, k, l) \quad \forall (i, k, l) \in \mathcal{S} \quad (10.23)$$

Let us denote $P_{\mathcal{S}}$ the projection operator in the subspace \mathcal{S} and WT the operator associated with the wavelet transform, we can write:

$$\mathcal{V} = (P_{\mathcal{S}} \circ WT)(F) = A(F) \quad (10.24)$$

We have to solve the inverse problem which consists of determining F knowing A and \mathcal{V} . We minimize the distance $\|\mathcal{V} - A(F)\|$ leading to:

$$\tilde{A}(\mathcal{V}) = (\tilde{A} \circ A)(F) \quad (10.25)$$

The initial equation (10.24) is modified with the introduction of \tilde{A} the joint operator associated with A . \tilde{A} is applied to a wavelet transform W and gives an image \tilde{F} :

$$\tilde{F} = \tilde{A}(W) = \sum_{i=1}^I (H(1) \cdots H(i-1)G(i))(W(i)) \quad (10.26)$$

Initially, we have implemented the reconstruction algorithm with this operator \tilde{A} , but several tests have shown the existence of spurious rings around the objects in the restored images. This phenomenon is due to the positive bumps of the filtering of the negative components at each scale of the wavelet structures processed by $G(i)$. The artifacts are removed by suppressing this operator and we simply write:

$$\tilde{F} = \tilde{A}(W) = \sum_{i=1}^I (H(1) \cdots H(i-1))(W(i)) \quad (10.27)$$

The equation (10.25) is solved either by *the gradient algorithm* [BF95] or by *the conjugate gradient algorithm* [P. 94] which improves the restoration quality and the convergence speed.

10.3.6 The pyramidal vision

The previous vision scheme has been also applied to a pyramidal wavelet transform. The interscale connectivity graph is determined taking into account the decimation from one scale to the following one. The restoration algorithm is derived from the conjugate gradient algorithm.

In a variant method, we have only take into account the significant maxima. At each scale we extract a 3×3 window centered on them. The interscale graph and the restoration are done in a same manner. This is well adapted to a real time processing.

10.4 Applications to astronomical images

We test the multiscale models on the image L384-350 (see figure 1) corresponding to the galaxy 384350 of the Lauberts catalogue.

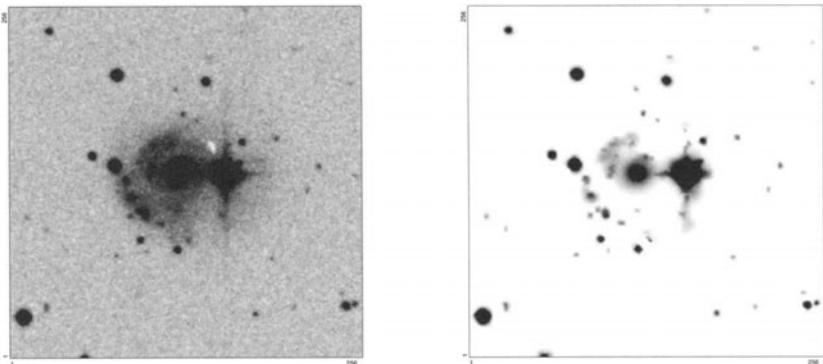


FIGURE 1. Image of L384-350 and the restored image

We performed a 7 scales wavelet transform of L384-350. 58 objects are detected. The restored image with the *à trous* algorithm, made of the reconstructed images of each object, is given in Figure 1. The tree of the central galaxy object is plotted in Figure 2. The corresponding restored image is plotted in Figure 3. A sub-object of the galaxy, which corresponds to a spiral arm, has been extracted; its image is shown in the same figure.

On left figure 4 we have plotted the reconstruction using a pyramidal algorithm. All the objects are restored, with a quality similar to the one using the *à trous* algorithm. On right figure 4 the reconstruction based on the maxima detection is displayed. The quality of the restoration is less good. This model is now not adapted to correctly describe complex structures, but it is sufficient to analyze point-like objects.

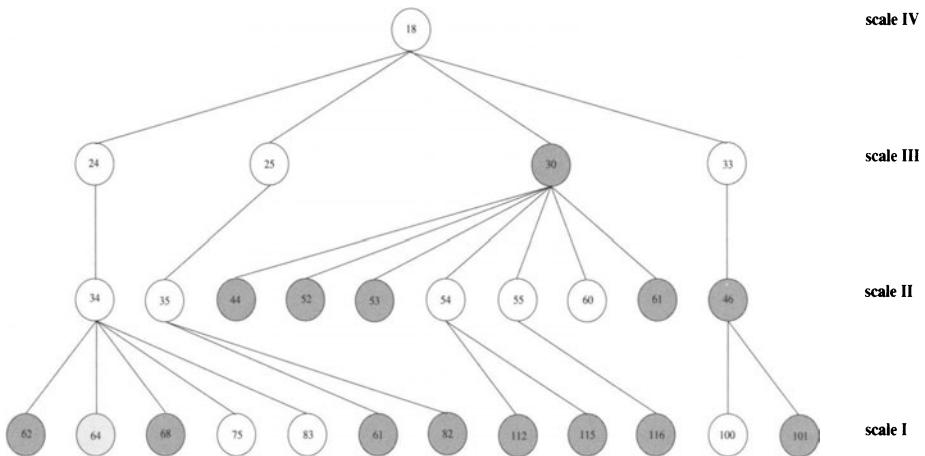


FIGURE 2. Tree of the galaxy object of L384-350

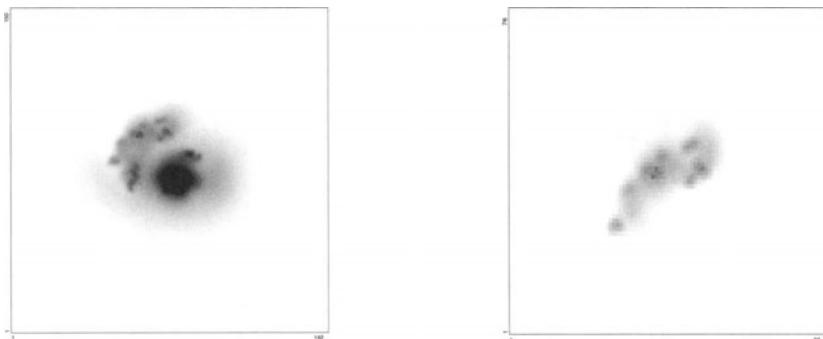


FIGURE 3. Restored images of the galaxy object and one of its sub-object

In the case of simple objects of small size, usual astronomical imagery methods and the multiscale model give very close results [RB95]. But, the multiscale model permits not only point-like objects to be identified but also objects which are much more complex (for instance the central galaxy of L384-350). Such objects with their structure hierarchy can be decomposed by our model thanks to the notion of sub-object.

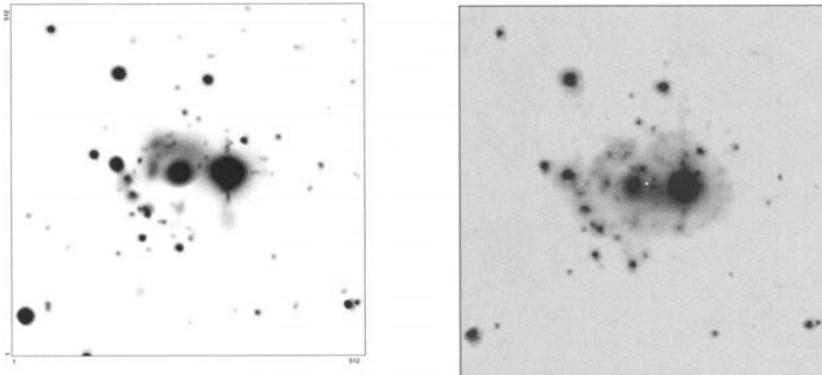


FIGURE 4. The restored images for the pyramidal algorithm with the significant positive coefficients (left) and with the maxima (right).

REFERENCES

- [BF95] A. Bijaoui and F. Rué. A multiscale vision model adapted to the astronomical images. *Signal Processing*, 46:345–362, 1995.
- [Bij80] A. Bijaoui. Skybackground estimation and applications. *Astronomy and Astrophysics*, 84:81–84, 1980.
- [Bij91] A. Bijaoui. Algorithmes de la transformation en ondelettes. applications en astronomie. In INRIA, editor, *Ondelettes et Paquet d’Ondes*, pages 115–140, 1991.
- [BLMO78] A. Bijaoui, G. Lago, J. Marchal, and C. Ounnas. Le traitement automatique des images en astronomie. In INRIA, editor, *Traitemet des Images et Reconnaissance des Formes*, pages 848–854, 1978.
- [Bra65] R.M. Bracewell. *The Fourier transform and its applications*, chapter 10, page 189. Mac-Graw-Hill New-York, 1965.
- [GKMM89] A. Grossmann, R. Kronland-Martinet, and J. Morlet. *Reading and Understanding Continuous Wavelet transform*, pages 2–20. Springer Berlin, 1989. in *Wavelets: Time-Frequency Methods and Phase-Space*, Springer-Verlag, J.M. Combes., A. Grossmann., Ph. Tchamitchian Editors.
- [Har63] W.W. Harman. *Principles of the Statistical Theory of Communication*, chapter 11, page 217. Mac-Graw Hill, New York, 1963.
- [HKMMT89] M. Holdschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian. A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform. In *Wavelets: Time-Frequency Methods and Phase-Space*, pages 286–297. Springer Berlin, 1989.
- [Kru89] A. Kruszewski. Inventory-searching, photometric and classifying package. In *1st ESO/ST-ECF Data Analysis*. Warsaw University Observatory, April 1989.
- [Mal89] S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 11(7):674–693, 1989.

- [MSB95] F. Murtagh, J.L. Starck, and A. Bijaoui. Image restauration with noise suppression using the wavelet transform ii. *AA Sup Ser*, 112:179–189, 1995.
- [P. 94] P. Lascaux and R. Théodor. *Analyse numérique matricielle appliquée à l'art de l'ingénieur*, volume 2, chapter 8, pages 405–458. Masson, 1994.
- [RB95] F. Rué and A. Bijaoui. A multiscale vision model to analyse field astronomical images. submitted to Experimental Astronomy, September 1995.
- [SB94] J.L. Starck and A. Bijaoui. Filtering and deconvolution by the wavelet transform. *Signal Processing*, 35:195–211, 1994.
- [SBLP94] J.L. Starck, A. Bijaoui, B. Lopez, and Ch. Perrier. Image reconstruction by the wavelet transform applied to the aperture synthesis. *Astronomy and Astrophysics*, 283:349–360, 1994.
- [SdLB93] E. Slezak, V. de Lapparent, and A. Bijaoui. Objective detection of voids and high density structures in the first. *Ap. J.*, 409:517–529, 1993.
- [SMB⁺88] E. Slezak, G. Mars, A. Bijaoui, C. Balkowski, and P. Fontanelli. Galaxy counts in the coma supercluster field: automated image detection and classification. *Astron. Astrophys. Sup. Ser.*, 74:83–106, 1988.
- [Sto86] R.S. Stobie. The COSMOS image analyzer. *Pattern Recognition Letters*, 4:317–324, 1986.
- [Str89] G. Strang. Wavelets and dilation equations: a brief introduction. *SIAM Review*, 31:614–627, 1989.
- [UA92] M. Unser and A. Aldroubi. Polynomial splines and wavelets - a signal processing perspective. In *Wavelets: a tutorial in theory and applications*, pages 91–122. C.K. Chui. Academic Press, New York, 1992.
- [Val89] F. Valdes. Faint object classification and analysis system standard test image. In *1st ESO/ST-ECF Data Analysis*. IRAF group, Tucson, Arizona, April 1989.

11

Statistical Software, Siftware and Astronomy

**Edward J. Wegman, Daniel B. Carr,
R. Duane King, John J. Miller,
Wendy L. Poston, Jeffrey L. Solka, and
John Wallin¹**

ABSTRACT This paper discusses statistical, data analytic and related software that is useful in the realm of astronomy and spaces sciences. The paper does not seek to be comprehensive, but rather to present a cross section of software used by practicing statisticians. The general layout is first to discuss commercially available software, then academic research software and finally some possible future directions in the evolution of data-oriented software. We specifically exclude commercial database software from the discussion, although it is relevant. The paper focuses on providing internet (world wide web) pointers for a variety of the software discussed.

11.1 Introduction

It seems somewhat presumptuous for a group of statisticians (and one astronomer) to tell a group of astronomers what manner of statistical software they need. The alternative is an attempt at an encyclopedic cataloguing of existing statistical software, an effort that would seem to have little value added. Fortunately, there are a few guides to the type of statistical methods perceived by astronomers as being required. An early work by Trumpler and Weaver (1953) is based on lectures given in 1935 and focuses on the application of then emerging statistical theory to astronomy. While traditional statistical theory is exposited, applications focus on statistical techniques for spectral distributions and spatial distribution of stars. These presage elements of time series analysis and spatial statistics, the latter being particularly a topic of considerable interest among statisticians. Interest in spatial point processes is also reflected in the much more recent collaborative work by Babu and Feigelson (1996). Slightly earlier work by Rolfe (1983) and by Murtagh and Heck (1988) carried strong elements of

¹George Mason University, Fairfax, VA 22030-4444

analysis based on large databases reflecting current statistical interest in massive data sets. Jaschek and Murtagh (1989) introduce concerns of data analysis and fitting, and of small sample issues suggesting bootstrapping and jackknifing techniques. Perhaps the most definitive articulation of the role of statistics in astronomy is Feigelson and Babu (1992). They identify work cluster analysis, truncation and censoring, Bayesian methods and image analysis, time series analysis, and multivariate methods. We would perhaps add to the themes articulated above graphical exploratory analysis and visualization. Our attempt to describe statistical and related software will be built around these statistical and related computing themes: 1) spatial statistics and spatial point process; 2) time series analysis, spectral distributions; 3) massive data sets, databases; 4) clustering methods, pattern analysis; 5) truncation and censoring; 6) image analysis, particularly Bayesian methods; 7) multivariate methods; and 8) visual exploratory analysis.

In our subsequent discussion, we take **Past** to be synonymous with **commercially available software**, **Present** to be synonymous with **academic and research software** that is not commercially supported, and **Future** to be **software based on a little speculation on the likely nature of future requirements**. Of course, commercially available software is legacy software for which there is a major investment in both people and existing databases. The companies that release commercial software have a very big infrastructure to support that software and do an excellent job in keeping up with the latest developments. However, because of their required adherence to data structures and styles of computing along which their analysis systems were originally developed, they lack much of the agility that academic and research code can exploit. Academic code, on the other hand is developed with much less discipline and is traditionally unsupported or supported comparatively poorly. Our speculation on future code may, of course, ultimately prove to be foolish. Yet it does perhaps point the way to what we should at least expect.

As turbulent as computing has been since the introduction of microprocessor-based personal computers and workstations, we are on the threshold of an even more uncertain and exciting era. Several phenomena are worth noting:

1) Within the last 18 months, the face of supercomputing has changed dramatically. Cray Research was sold to Silicon Graphics, Convex Computers became a wholly owned subsidiary of Hewlett-Packard, and Intel announced that they would cease production of their Paragon, the current holder of world speed record. A 90 megahertz Pentium PC has essentially the same computing power as the \$12,000,000 Cray 1 supercomputer did in 1975. Thus, personal computing supplants supercomputing to a large extent.

2) The ubiquity of the world wide web is a phenomenon none can escape. Having a web page is now a mode of business as crucial as having a fax

machine was three years ago. A business now seems to be at a serious disadvantage if it doesn't have a web page.

3) High performance computing which, at one time, was essentially synonymous with vector processor supercomputers solving partial differential equations has been broadened to include not only computationally-intensive applications, but also data-intensive and information-intensive applications. The high speed network becomes even more crucial in this context. The notion of a hollow computer, a personal computer or workstation for which most of the computations are done transparently by other machines on the network, will become a reality.

The software discussed in this paper is selected for discussion based on highly personalized experience. It reflects a cumulative experience of a number of people who really do statistical and scientific computing on a daily basis. We certainly do not guarantee that we are inclusive in describing all possible packages. Earlier works by Francis (1981) and by Hayes (1982) survey respectively 60 and 213 statistical packages. A summary of the latter work appeared as Wegman and Hayes (1988), but all of these are hopelessly outdated now and are only of historical interest. Rather than an all-inclusive survey, we choose to describe a few of the packages we find useful in day-to-day computing.

11.2 Past: Commercially available software

We believe there are three general classes of software available using several different user interfaces. Statistical software begins to blend in one direction with relational database software such as Oracle (software we do not discuss here) and with mathematical software such as MATLAB in the other direction. Mathematical software exhibits not only statistical capabilities flowing from code for matrix manipulation, but also optimization and symbolic manipulation useful for statistical purposes. Finally visualization software overlaps to some extent with software intended for exploratory data analysis.

The SAS System for Statistical Analysis SAS began as a statistical analysis system in the late 1960's growing out of a project in the Department of Experimental Statistics at North Carolina State University. The SAS Institute was founded in 1976. The SAS System has expanded to become a system for data management and analysis. The SAS System includes products for: management of large data bases; time series analysis; analysis of most classical statistical problems, including multivariate analysis, generalized linear models, and clustering; data visualization and plotting. A geographic information system is one of the products available in the system. The SAS System is available on PC and UNIX platforms, as well as on mainframe computers.

SAS supports simulation studies with random number generators for many different distributions. User written functions can be integrated into the system with the product, SAS/BASE. Programs may be written in a language which resembles C, however, many applications can be accomplished using simple point and click operations. For users with a need to write an applications program using a matrix language, the product SAS/IML provides the ability to program using matrices as objects. Data may be imported into and exported from SAS using the SAS/ACCESS product. In PC SAS, data may be imported from most commercial spread sheet or database software. The SAS/STAT product linear models (regression, analysis of variance and covariance), generalized linear models (including logistic and Poisson regression), multivariate methods (MANOVA, canonical correlation, discriminant analysis, factor analysis, clustering), categorical data analysis (including log-linear models), and all standard techniques for descriptive and confirmatory statistical analysis. The statistical analyses may be used with the graphical product, SAS/GRAFH, to produce relevant plots such as q-q plots, residual plots, and other relevant graphical descriptions of the data. The SAS/ETS product allows the user to accomplish sophisticated analyses of time series data. The SAS/GIS product is a geographic information system built in to the larger SAS system. Spatial data may be stored, linked, analyzed, and displayed using SAS/GIS.

SAS is to a large extent an industry standard statistical software package. We find that demand for students with SAS skills is considerably greater than for students with skills other statistical packages. Some useful URL's are <http://www.sas.com> which is the main URL for SAS and also <http://is.rice.edu/~radam/prog.html> which contains some user-developed tips on using SAS. Web search engines also can turn up many, many references to SAS. An AltaVista (<http://altavista.digital.com>) search on SAS turns up more than 40,000 hits.

Other statistical systems which are of the same general vintage as SAS are MINITAB, BMDP and SPSS. All of these systems began as main-frame systems, but have evolved to smaller scale systems as computing has evolved. The URL for MINITAB is <http://www.minitab.com>. An AltaVista search on MINITAB turns up about 1000 hits. An AltaVista search of BMDP turns up about 2000 hits. A reference URL for BMDP is <http://www.ppgsoft.com/bmdp00.html>. An AltaVista search for SPSS turns up about 10,000 hits. A reference URL for SPSS is <http://www.spss.com>. MINITAB is used extensively in the educational community. Indeed, we use it for our introductory courses. BMDP and SPSS tend to find users among the communities in which they originated, respectively the biomedical community and the social sciences community. While this may be a value judgment from our perspective, it appears that mainstream applied statisticians tend to use SAS more extensively. S-PLUS on the other hand seems to be a package that is highly regarded among the

more research-oriented statisticians, particularly those interested in computational statistics.

S-PLUS While there are many different packages for performing statistical analysis, S-PLUS offers great flexibility with regard to the implementation of user-defined functions and the customization of ones environment. S-PLUS can be thought of as a high-level programming language that has been designed for the easy implementation of statistical functions. Besides excellent support for statistical and user defined operations this language offers the user extensive graphics and hardcopy capability. S-PLUS is a supported extension of the statistical analysis language S. S was originally developed at AT&T Bell Labs by R. Becker, J. Chambers, A. Wilks, W. Cleveland and T. Hastie. The original description of the S language was written by Becker, Chambers, and Wilks (1988). A good introduction to the application of S to statistical analysis problems is contained in Chambers and Hastie (1992).

S-PLUS is manufactured and supported by the Statistical Sciences Corporation, now a division of MathSoft. There have been many researchers who have worked on the S-PLUS extension. Some of the code has been contributed by prominent individuals from the academic and industrial communities. Much of this code resides at the **Statlib** library system. The reader may obtain an index of the S Statlib software by sending email to **statlib@lib.stat.cmu.edu** with the one line message send index from S.

S-PLUS runs on both PC and UNIX based platforms. In addition the company offers easy links for the user to call S-PLUS from within C/ FORTRAN or for the user to call C/FORTRAN compiled functions within the S-PLUS environment. Statistical Sciences has made great efforts to keep the software current with regard to the needs of the statistical community. There is a recently released wavelets module and at the time of the writing of this article there is a planned release of a spatial statistics module. The S-PLUS package provides the user with a plethora of statistical capabilities. These include the ability to generate random data from 20 different distributional types and the ability to perform 12 different types of hypothesis tests including Student's t-test and the Wilcoxon test. It allows one to perform linear, nonlinear, and projection pursuit regression. There are also capabilities for multivariate analysis including graphical methods, cluster analysis and discriminant analysis. In addition there are the standard time series analysis tools for ARIMA models and seasonally adjusted data. Finally we point out that the graphical display capabilities are well developed.

The capabilities inherent in the spatial statistics module may be of particular interest to the astronomical community. Sky catalog images can be viewed as spatial point patterns on one level. Remote sensing planetary information might fall into the realm of spatially continuous data or area data depending on the nature of the collection process. Hence the S-PLUS spatial capabilities may be quite useful for the analysis of these

data types. An AltaVista search on S-PLUS turns up about 1000 hits. The S-PLUS home page can be reached at <http://www.mathsoft.com>. The URL: <http://www.gcrc.ufl.edu/gopher.documents/sas/sas.vs.splus.html> features a comparison between SAS and S-PLUS.

Other statistically oriented packages enjoying good reputations are SYSTAT, DataDesk, and JMP. SYSTAT originated as a PC-based package developed by Leland Wilkinson. SYSTAT is now owned by SPSS and more information on SYSTAT can be found at URL <http://www.spss.com>. SYSTAT has about 1000 AltaVista hits. DataDesk is a Macintosh-based product authored by Paul Velleman from Cornell University. This is a GUI-based product which contains many innovative graphical data analysis and statistical analysis features. More information about DataDesk can be found at URL: <http://www.lightlink.com/> **datadesk**. DataDesk has about 200 AltaVista hits. JMP is another SAS product that is highly visualization oriented. JMP is a stand alone product for PC and Macintosh platforms. It originated as a Macintosh product and resembles DataDesk in some ways. Information on JMP can be found at <http://www.sas.com>. An AltaVista query on JMP is indeterminate since this abbreviation appears to have many other meanings. While more could be written about these individual products (and probably should be), we leave the discussion at this stage.

The descriptions of statistical software above cover the most well-established commercially available software packages. Mathematical packages often exhibit some statistical capabilities, especially when engineering or other basic science applications have overlap with statistics. Among the most extensively used mathematical packages is MATLAB. MATLAB has many features that resemble APL, a language popular for statistical computing in the 1970s.

MATLAB is an interactive computing environment that can be used for scientific and statistical data analysis and visualization. It is similar to the data analysis software IDL that may already be familiar to many researchers in astronomy. The basic data object in MATLAB is the matrix. The user can perform numerical analysis, signal processing, image processing and statistics on matrices, thus freeing the user from programming considerations inherent in other programming languages such as C and FORTRAN. Versions of MATLAB are available for UNIX platforms, PC's running Microsoft Windows and Macintosh. Because the functions are platform independent, provides the user with maximum reusability of their work.

MATLAB comes with many functions for basic data analysis and graphics. Most of these are written as M-file functions, which are basically text files that the user can read and adapt for other uses. The user also has the ability to create their own M-file functions and script files. The recent addition of the MATLAB C-Compiler and C-Math Library allows the user to write executable code from their MATLAB library of func-

tions, yielding faster execution times and stand-alone applications. For researchers who need more specific functionality, MATLAB offers several modules or toolboxes. The toolboxes are a collection of M-file functions that implement algorithms and functions common to an area of interest. Some of the toolboxes that would be useful in astronomy are Statistics, Signal Processing, Image Processing, and Symbolics. The Statistics Toolbox performs basic hypothesis tests, regression, and statistical visualization. The Signal and Image Processing Toolboxes include functions for signal display, filtering and analysis. The Symbolics Toolbox contains the Maple Kernel and comes in a basic collection of functions or an extended version which includes the Maple programming features. There are also several third-party packages that are available, including a package called Wavbox that implements wavelet analysis algorithms. There is a considerable amount of contributed MATLAB code available on the internet. One notably useful source for astronomers is the MATLAB Astronomy Library at the Astronomy Department of the University of Western Ontario. This library has M-file functions that have been developed by the department for analyzing astronomical data. This site can be accessed at <http://phobos.astro.uwo.ca/~etittley/matlab/matlab-astrolib.html>. Another source of code is available via the home page for MATLAB at <http://www.mathworks.com>. MATLAB has more than 10,000 hits in an AltaVista search.

MATLAB is used extensively in our University and is a particular favorite of engineers, physicists and chemists. Other mathematical software worth noting is *Mathematica* and MAPLE, both of which have powerful symbolic processing capabilities. *Mathematica* also has numerical and graphical features, but is comparatively complex to learn. Information on *Mathematica* is available at URL <http://www.wolfram.com> while additional information on MAPLE is available at <http://www.maplesoft.com>. An AltaVista search on *Mathematica* turns up some 40,000 hits while a search on MAPLE turns up 7,000 hits many of which refer to trees. Another useful mathematical package is MATHCAD, a package which combines numerical, symbolic, and graphical features. MathSoft, Inc., producers of MATHCAD, have recently acquired S-PLUS; information on MATHCAD is available at <http://www.mathsoft.com>. MATHCAD turns up about 3,000 hits in an AltaVista search. A longtime standard package that spans both statistical and mathematical techniques is the IMSL scientific subroutine library. Most scientists are probably familiar with IMSL. The IMSL corporation merged several years ago with the producers of PVWave and is now known as Visual Numerics, Inc. More information on both of these products can be obtained at <http://www.vni.com>. A good source of expertise and helpful hints for IMSL users is available at http://www-c8.lanl.gov/dist_comp2/MATH/Imsl/imsl_keyword.html.

Visualization tools are becoming more powerful and hence more useful for the statistician and data analyst. S-Plus, DataDesk, JMP and more re-

cently MATLAB are incorporating advanced visualization tools. We have alluded to PVWave above as a visualization packages. Other advanced visualization tools include AVS and IDL. As indicated above, more information on PVWave is available at <http://www.vni.com>. Information on AVS is available at <http://www.avs.com>. AVS has an extensive users group and a library of software applications and other information for AVS is available at <http://testavs.ncsc.org>. Information on IDL is available at <http://www.rsinc.com/>. URL, <http://axp2.ast.man.ac.uk:8000/~dsb/visual/sg8.htm/node16.html>, is a general resource for visualization packages.

11.3 Present: Academic and research software

To state what is obvious, the down-side of academic and research software is that it tends to be less comprehensive and less reliable than commercial-grade, supported software. The upside is that it tends to be more innovative and daring in concept. Most of the commercial software discussed in Section 2 has roots in academic, research software. Because academic, research software is generally not as widely distributed, our discussion of it will be more limited in scope and personalized in perception. We discuss four academic, research packages: 1) XGobi, 2) Xlisp-Stat, 3) ExplorN, and 4) Manet.

XGobi XGobi is an interactive high-dimensional visualization package. This X-Window-based system implements many useful data exploration concepts developed or promoted by the statistical graphics community during the last two decades. The concepts include focusing via rescaling, conditioning and sectioning; point linking across and rearrangement of multiple views, direct interactive manipulation including stretching, panning, zooming, rotation, point identification, and brushing; projection and section tours, data transformations, and algorithms for finding optimized views. XGobi is the result of a continuing effort by the statistics community to make powerful methods accessible to the scientific community. The history behind XGobi can in part be found in Cleveland and McGill (1988). XGobi is free over the network. The documentation is now sold for a modest price. While XGobi is both fun and easy to use, knowledge about how seek and interpret structure is important. Furnas and Buja (1994) provide an enlightening discussion on the discovery of the dimensionality of objects embedded in high dimensional space via low dimensional views. With appropriate knowledge and procedures one can identify the dimensionality of objects through 6-dimensions using scatterplots. The dimensionality that XGobi can practically handle is limited. The words, high-dimensional, here means something like 20 variables and not 500 variables. For very high-dimensional problems, dimension reduction methods are required be-

fore XGobi is useful. As always, scientifically insightful transformations and dimension reductions can be crucial. The structure one looks for depends on the type of problem. A surprisingly common situation is that low-dimensional structure is embedded in high-dimensional data. For example satellite spectral intensity bands are one source of multivariate data. When high dimensional data is composite of fairly simple low-dimensional structures, humans can understand a lot. The key is to have the right set of clustering, projection and sectioning tools that help find and isolate understandable constituent elements.

The efforts to extend the domain of application of XGobi continue; XGobi can communicate with ARC/VIEW and ARC/INFO via remote procedure calls (Symanzik, Majure, and Cook, 1995). Brushed points in XGobi change color in the ARC/VIEW map and vice versa. XGobi provides insightful re-expression of data coming from ARC/VIEW. A reference URL for XGobi is <http://lib.stat.cmu.edu/general/XGobi/index.html>. An AltaVista search on XGobi returns more than 550 hits.

Xlisp-Stat Xlisp-Stat is an object-oriented environment for statistical computing and dynamic graphics. Written by Luke Tierney, University of Minnesota, Xlisp-Stat was motivated by the "S" system, with the basic principal that an extendible system is necessary for conducting research on new computationally based statistical methods. Xlisp-Stat provides a set of high-level tools to develop new dynamic graphics techniques. Although motivated by S, Xlisp-Stat is based on Lisp. Like S, Xlisp-Stat is an interpreted language, which is much more suited towards exploration than a compiled language such as C/C++ or Pascal. These require lengthy recompilations to fix bugs or test simple new ideas. When greater speed is required, a byte-code compiler can be run from inside Xlisp-Stat. This compiler can give increase the speed of execution by an order of magnitude. The defining reference book on Xlisp-Stat is Tierney (1990).

Xlisp-Stat is available for Unix/X Windows, Macintosh/MacOS and for IBM PC compatibles running Windows 3.1. Xlisp-Stat is freeware, available by anonymous ftp to [ftp.stat.umn.edu](ftp://ftp.stat.umn.edu). For Unix workstations, the R-code project (<http://www.stat.umn.edu/~bjm/rcode>) has information on retrieving and compiling Xlisp-Stat for Unix. One needs a workstation running X11R4 or later. There are several sources of further information. UCLA's Xlisp-Stat Archive (<http://www.stat.ucla.edu/develop/lisp>) is run by Jan Deleeuw at UCLA. This is the largest public collection of Xlisp-Stat code. Penn State University's Lisp-Stat page includes descriptions of projects around the world using Xlisp-Stat (<http://euler.bd.psu.edu/lispstat/lispstat.html>). Another forum for questions about Xlisp-Stat is available by sending email to luke@stat.umn.edu. An AltaVista search on Xlisp-Stat turns up 300 hits.

ExplorN ExplorN is a statistical graphics and visualization package designed for the exploration of high dimensional data with a practical limit of 30 or so dimensions. The software has its roots in Explor4 (see Carr

and Nicholson, 1988), but has evolved well beyond. ExplorN is authored by Qiang Luo, Edward J. Wegman, Daniel B. Carr and Ji Shen and is written in C and exploits the GL graphics library available on Silicon Graphics workstations. ExplorN combines the early work of Carr and Nicholson on stereo ray glyph plots with more recent multidimensional visualization tools such as parallel coordinates, Wegman (1990). d-dimensional grand tour, Wegman (1991) and saturation brushing, Wegman and Luo (1996). Multidimensional display is available in either a scatter plot matrix, a parallel coordinate plot, or a stereo ray glyph plot. Brushing is available in any of these displays and the brushed color becomes an attribute of the brushed data so that brushed color is linked to all other plots.

ExplorN uses a general d-dimensional grand tour rather than simply a two-dimensional grand tour. The results of the tour are available in all three forms of multidimensional display. The high interaction graphics allows one to temporarily suspend the tour, brush with color and then resume the grand tour. Two additional features are of interest. The color saturation may be varied. The idea is that the display may be brushed with very low color saturation levels, i.e. nearly black. The program uses the alpha-channel feature of Silicon Graphics workstations to add saturation levels. Thus in regions of heavy overplotting, color saturation is high, while in regions of little overplotting, color saturation is low. This feature is useful for very large data sets since the net effect is to produce a color-coded density plot. This density has the interesting feature that it does not require any smoothing (convolutions) and hence preserves edges and boundaries well. Coupled with a partial grand tour it can be used to produce tree structured decision rules. The software runs on any SG workstation supporting alpha-channel and 24 bit color. We have used ExplorN for data sets as large as 250,000 observations in 10 dimensions and so is capable of handling fairly massive data sets. ExplorN also produces three-dimensional rendered density surfaces using lighting models, described in more detail in Wegman and Carr (1993) and Wegman and Luo (1995). Some of this work is described at URL <http://www.galaxy.gmu.edu/papers/inter96.html> and some images of the rendered densities are available at URL http://www.galaxy.gmu.edu/images/gallery/research_arcade.html.

MANET MANET is software for interactive statistical graphics running on a Macintosh computer. MANET is designed by Antony Unwin, Chair of the Computer-oriented Statistics and Data Analysis Group, Institute for Mathematics at the University of Augsburg, Augsburg, Germany—with contributions from George Hawkins, Heike Hoffman, Bernd Siegel, and Martin Theus. MANET is written in C++ and provides standard interactive graphical features. It is similar in design to DataDesk or JMP. Unlike its commercial counterparts, however, the MANET software focuses on innovative methods for graphically dealing with missing values. To our knowledge, it is the only software that consistently attempts to represent missing val-

ues graphically. All graphics are fully linked and may be interacted with directly. MANET follows Macintosh conventions and is consistent with other Macintosh packages. It is an exploratory tool and is intended to be used with other more traditional software. Unlike ExplorN, MANET does not support massive data sets. The current version is Version 0.1832. The next version, we are told by Professor Unwin, will be Version 0.1848. The version numbers correspond to important dates in the life of the impressionist painter, Manet, and presumably have the advantage that there cannot be an unlimited number of versions. All of the software developed by Professor Unwin's group at Augsburg is named for impressionist painters because the software is intended to give a visual impression of the data. MANET is freeware and may be obtained by sending email to unwin@uni-augsburg.de. More information on MANET is available at <http://www1.Math.Uni-Augsburg.DE/~theus/Manet/ManetEx.html>.

11.4 Future: Massive datasets and SIFTWARE

The software described above reflects traditional thinking about data sets and data analysis. Essentially they reflect mental world views of a fairly conventional nature about the size and dimensionality of data sets. Wegman (1995) articulated some issues of computational complexity in conjunction with data set sizes and discussed the limits of computational feasibility and well as visualization feasibility. This was motivated in part by considerations of NASA's EOS-DIS project as well as by massive data sets available as a result of accumulations in financial transaction databases. Following Huber (1994), Wegman discusses a range of data set sizes ranging from tiny (10^2 bytes) to huge (10^{10} bytes) and even beyond this to the multi-terabyte data sets promised by EOS. It is clear that data sets of these latter magnitudes test the computational limits as well as the visualization limits of all the software discussed in Sections 2 and 3. Automated or semi-automated accumulation of data has been a hallmark of space and astronomical experimental science for decades. In this section, we attempt to articulate the impact of computational and electronic instrumentation advances on what we conventionally think of as data analysis.

The World Wide Web has become a nearly ubiquitous fact of daily life that provides an essentially new learning paradigm, a virtual library at the finger tips of anyone with a network connected computer. Much like Marshall McLuhan's famous adage of the 1960's, the *Medium is the Message*, the web is the message and McLuhan's global village is in reality a global cyber-village. One can easily anticipate in much the same way that text material is available and searchable by the dozen or so indexed web search engines, that in the future databases of numerical and symbolic data will be searchable and retrievable through similar mechanisms. Just as there is

a virtual text library today, there will be a virtual data library in the future. Some virtual data libraries in primitive forms are available even today; some cancer and related medical databases are privately held and, to a limited extent, fragmented pieces of genome databases are publicly available on the web. The widespread availability of easily searchable and retrievable databases would have the effect of accelerating research both for subject matter scientists who could rapidly verify or discard conjectures based on empirical evidence as well as for methodological scientists such as statisticians who could test and refine methodologies based on application of their methods to real data.

We believe that this is a direction that statistical and computational research will take. Data acquisition, sorting and refinement will become part of the data analysis process. The phrase, *siftware*, we have coined in the title has its origins in a typographical error (*o* is next to *i* on the qwerty keyboard), but in fact massive databases will generally not simply be one massive data set, but many, many somewhat smaller data sets. A terabyte database could easily be a million 10^6 data sets. A database of this magnitude is not feasible for an individual to browse in an afternoon. Thus, data analysis software must also be data siftware ... software designed to aid in isolating interesting worthwhile data sets for the researcher to examine. In this spirit, we discuss three tools we think will reflect future directions.

JAVA JAVA is a programming language which represents an extension of the world wide web capabilities. Basic documents on the web are constructed using HTML, the hypertext markup language. HTML is a simple addition to ASCII text which allows the inclusion of simple formatting commands that are interpreted by the client browser. Most web pages are static in the sense that once a server delivers the HTML text to the browser, the server has done its job and the static text, images, and multimedia content are interpreted and displayed by the client's browser. There is a possibility of interaction between client and server through so-called common gateway interface (CGI) scripts. CGI scripts allow for the client to send information back to the server, for the server to carry out some action based on data from the client, and for the server to generate a web page dynamically based on the client's data. Typically the two most used applications of CGI scripts are requests for searching some database resident on the server or registration/purchase requests used, for example, to register for a scientific meeting. With a CGI script, the text is generated on demand by the server, but is still typically a static display in the client browser.

JAVA is a fully distributed, object-oriented programming language which allows for the creation a fully interactive web-based system. Data and the tools to view it can be sent to the client browser. In object oriented programming, one creates objects instead of variables. Objects have attributes (values) and methods (subroutines). JAVA allows attributes and methods to be linked together. In particular, JAVA allows applets, small applications

or subroutines, to be created and transmitted across the web just as static HTML documents are transmitted. The applets run on the client machine rather than the server. A typical application might be to have a data set and a linked statistical routine, for example a plotting routine or, say, a time series analysis routine. The routine would typically be interactive, so, for example, the graphic might be rotated dynamically or there might be a slider bar for the dynamic adjustment of a parameter in a statistical model.

JAVA is similar to C++ is comparatively easy to learn for those familiar with C++. JAVA is distributed in the sense that JAVA applets can run on machines connected by networks. Communication between applets is possible and so in this sense there is a resemblance to a massively parallel computer. JAVA is both interpreted and compiled. The original source code is compiled to *byte code*. Byte code is a machine neutral code that is interpreted locally by each client in a *JAVA Virtual Machine* (JVM). The JVM runs inside of a client browser and is isolated from the other resources of the client machine. JAVA byte code is interpreted by the client as it is loaded. JAVA is intended to be a secure system in that running inside the JVM although security problems do exist with present implementations. However, access to local data is restricted and the JAVA is a securable environment. The JAVA environment is architecturally neutral in that it will run on any machine that has a browser supporting a JVM. Interpreted byte code is very fast, producing near machine speeds and typically byte code sizes are very small even for comparatively complex programs.

So why is JAVA related to statistical/data analysis software of the future? We declare this as software to be watched because it is a practical implementation of a new paradigm in distributed computing, just as web browsers were a new paradigm in anonymous access to non-local machines. JAVA will allow for not only the distribution of text and multimedia, but also of computer applications and data. We could imagine an applet launched to search a distant database for a particular class of data, and to return to the client small subsets of the database that fulfill the search terms. The search applet could run in the background and could use heuristic search algorithms (sometimes called by statisticians *cognostics*), sort of a data analyst's analog to the military's fire and forget weapons. Moreover, under a JAVA framework, new statistical, data analytic and other methodologies could be made available over the web, and could be tried out by practitioner's in other research fields on their own data and their own computer. The possibilities are quite extensive and we have not yet seen even the most elementary of these uses implemented. More information on JAVA can be found at <http://java.sun.com>. An AltaVista search on JAVA returns more than 10,000 websites referencing JAVA.

JAVA is a response to the enormous popularity of the world wide web. If we consider the possibility of extending the web in a natural way to acquiring numerical or symbolic data in the same way we now acquire human-consumable information, new mechanisms must be sought to pro-

vide for the distribution of that data. The next two items, VDADC and METANET, are concepts that have been proposed as methods for accessing and distributing data. The ideas amount to a substantive method for evolving our notion of software. These ideas are under development, but of course are not available for usage yet. They depend not only on technology developments, but also on political developments. The latter are much more problematic. We believe something akin to these will ultimately be developed.

VDADC NASA has created a wonderfully vulnerable concept to deal with the massive data sets anticipated from the Earth Observing System (EOS). Called the DAAC, Distributed Active Archive Centers, NASA manages to encode two oxymorons in a single name (distributed centers and active archives). The DAACs are intended as central repositories for the massive amounts of data expected from EOS and, as such, form a prototype for other application fields with massive data sets. One proposal currently under development for access data in the DAACs is the Virtual Domain Application Data Center (VDADC). A VDADC is a way of organizing massive data sets to provide an optimal search for any particular group of users. The VDADC is designed to accommodate large data sets in the sense that the VDADC contains a large number of descriptors (metadata) that characterize the data rather than containing the data itself. The VDADC views its world of data as organized into distinct trees, without any specific linkages among the trees. Each tree is called a "database". The database is the most general parameter to describe the data. The early part of the search mechanism is to determine which database is the closest fit to the user request. The trees are constructed so that most searches are resolved in the least amount of time. The trees are dynamically constructed based on user queries. The trees are conceptually viewed as three level structures. The top level is a gross organizing scheme; it also serves to define any number of parameters to be inherited by the levels below it. The middle level consists on any number of nodes which will form the search mechanism. The bottom layer consists descriptors pointing to the actual data.

It is envisioned that user queries would be generated from web pages. These web pages (forming the user queries) would be generated by the VDADC system itself to reflect the current state of the dynamic search trees. Since the forms are being generated by the system, this would allow a specific user request to be partially formulated by the system. If a user were to request a very narrow field of the data base, the query presented to the search mechanism will have actually been generated by the web page, and that query could include a number of hints to the search mechanism to allow a very fast descent through the search trees. The user query will then be broken into a number of independent queries and the appropriate result of those queries will be delivered to the user. While the VDADC concept is specifically focused on NASA's EOS data, the METANET concept

described below envisions a national and international digital data library which would be available via the Internet. We consider a heterogeneous collection of scientific databases.

METANET Automated Generation of Metadata In general, it is assumed there are metadata that describe file and variable type and organization, but that have minimal information on scientific content of the data. In the raw form, a data set and its metadata have minimal usability. Thus a strategy for making the data usable is to link the data set to digital objects that are used to index the data set. The search operation for particular scientific content in the data becomes a simple indexing operation on the digital objects linked to the data set. The idea is to link digital objects with scientific meaning to the data set. The digital objects become part of the searchable metadata associated with the data set. The goal of creating digital objects reflecting the scientific content of the data is not to replace the judgment of the scientist, but to narrow the scope of the data sets that the scientist must consider. The key element is to automate the process of creating digital objects with scientific meaning to be linked to the data set. The concept is to have a background process, launched either by the database owner or, more likely, via applet created by the virtual data center (e.g. a VDADC), examining databases available on the dataweb and searching within data sets for recognizable patterns. When a pattern is found in a particular data set, the digital object corresponding to that pattern is made part of the metadata associated with that data set. Also pointers would be added to that metadata pointing to metadata associated with other distributed databases containing the same pattern. This metadata will be located in the virtual data center and through this metadata, distributed databases will be linked. This linking is to be done on the fly as data is accumulated in the database. On existing databases, the background process would run as compute cycles are available. Because the database is dynamic, the background process would always be running adding metadata dynamically.

Query and Search The idea of the automated creation of metadata is to develop metadata that reflects the scientific content of the data sets within the database rather just data structure information. The locus of the metadata is the virtual data center. The end user would see only the virtual data center. The original metadata, resident in the actual data centers, would be reproduced in the virtual center. However, that original metadata would be augmented by metadata collected by the automated creation procedures mentioned above, by pointers used to link related data sets in distributed databases, and by metadata collected in the process of interacting with system users. Query and search would contain four major elements: 1) client browser, 2) expert system for query refinement, 3) search engine and 4) reporting mechanism.

Client Browser The client browser would be a piece of software running on the scientist's client machine. The client machine is likely to be a PC or a workstation. The client software is essentially analogous to the myriad of browsers available for the world-wide web.

Expert System for Query Refinement There are two basic scenarios for the interaction of the scientist with the server: first, the scientist knows precisely the location and type of data he desires, and second, he knows generally the type of question he liked to ask, but has little information about the nature of the databases with which he hopes to interact. The first scenario is comparatively straightforward, but the expert system would still be employed to keep a record of the nature of the query. The idea is to use the queries as a tool in the refinement of the search process. The second scenario, however, is the more complex. The approach is to match a vague query formulated by the scientist to one or more of the digital objects discovered in the automated-generation-of-metadata phase. The expert system would initially be given rules devised by discipline experts for performing this match. Given an inquiry, the expert system would attempt to match the query to one or more digital objects (patterns). It would provide the scientist with an opportunity to confirm the match or to refine the query. This interplay would continue until the scientist is satisfied with the proposed matches. The expert system would also take advantage of the interaction with the scientist to form a new rule for matching the original query to the digital objects developed in the refinement process.

Search Engine As indicated above, large scale scientific information systems will likely be distributed in nature and contain not only the basic data but both structured metadata, for example, sensor type, sensor number, measurement date and unstructured metadata, for example, a text-based description of the data. These systems will typically have multiple main repository sites that together will house a major portion of the data as well as some smaller sites, virtual data centers, containing the remainder of the data. Clearly, given the volume of the data, particularly within the main servers, high performance engines that integrate the processing of the structured and unstructured data would be required to support desired response rates for user requests.

Reporting Mechanism The basic idea is not only to retrieve data sets appropriate to the needs of the scientist, but also to scale down the potentially large databases the scientist must consider. That is, the scientist would consider megabytes instead of terabytes of data. The search and retrieval process may still result in a massive amount of data. The reporting mechanism would thus initially report the nature and magnitude of the data sets to be retrieved. If the scientist agrees that the scale is appropriate to his needs, the data will be delivered by an FTP or similar mechanism to his local client machine or to another server where he wants the synthesized data to be stored.

11.5 General comments

In this paper, we have tried to provide general assessments and pointers to a variety of statistical, data analysis and related software that would appear to address some of the needs of astronomers and space scientists. This is, of course, a highly personalized view of the statistical software world. We attempt to represent a variety of opinions by drawing in a rather larger number of authors than might be typical for a statistics paper. We have divided our discussion into commercial, academic research and as-yet-not-developed software. The intent is to provide a broader vision of software, not to merely catalog a dozen or so packages.

The use of pointers (URLs) from the world wide web extends the utility of this paper by giving references that are likely to be dynamically updated. Particularly with commercial software, the vendors have a significant stake in maintaining current information while at the same time ensuring the continuity of the web page address. We believe that these URLs may be the most valuable aspect of this summary paper. Also of note is the number of hits a particular software gets under an AltaVista search. This search engine is by far the most comprehensive of all the search engines available. We believe a reasonable inference is that a particular software's popularity (utility?)(market penetration?) is proportional to the number of web pages devoted to it. This gives the reader a gauge of the success of a piece of software.

Finally we want to note some addition URLs that may be of general interest for both the astronomy and the statistics audiences. Most statisticians know of STATLIB available at URL <http://lib.stat.cmu.edu>. STATLIB is an extraordinarily comprehensive resource for data, software, and other information for the statistics community. It is a highly recommended site to visit and browse. A visit to the URL at Cornell University, <http://www.stat.cornell.edu/compsites.html>, yields a series of pointers to a variety of software and computing resources. A guide to statistical computing resources on the net can be found at http://asa.ugl.lib.umich.edu/chdocs/statistics/stat_guide_home.html. Finally a somewhat longer version of this paper with all of the URLs mentioned here is available on-line at <http://www.galaxy.gmu.edu/papers/astr.html>.

Acknowledgments: The work of the principal author, Dr. Wegman, was supported by the Army Research Office under contract DAAH04-94-G-0267. The associate authors are listed in alphabetical order of last names. Drs. Wegman, Carr and Miller are members of the Department of Applied and Engineering Statistics at George Mason University and are also affiliated with the Center for Computational Statistics at GMU. Dr. Wallin and Mr. King have primary affiliations with the Institute for Computational Sciences and Informatics at GMU. Drs. Poston and Solka have primary af-

filiation with the Naval Surface Warfare Center in Dahlgren, VA. All of the authors have an affiliation with the Institute for Computational Sciences and Informatics. The work of Dr. Carr was supported by the Environmental Protection Agency. The work of Drs. Poston and Solka was supported by the Office of Naval Research under the ILIR program.

REFERENCES

- [1] Babu, G. J. and Feigelson, E. D. (1996) "Spatial point processes in astronomy," **Journal of Statistical Planning and Inference**, 50(3), 311-326.
- [2] Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) **The New S Language**, Wadsworth and Brooks/Cole, Pacific Grove, CA.
- [3] Carr, D. and Nicholson, W. (1988) "Explor4: A program for exploring four-dimensional data using stereo-ray glyphs, dimensional constraints, rotation and masking," In **Dynamic Graphics for Statistics**, (W. S. Cleveland and M. E. McGill, eds.) Wadsworth Inc., Belmont CA.. 309-329.
- [4] Chambers, J. M. and Hastie, T. K.. (eds.) (1992) **Statistical Models in S**, Wadsworth and Brooks/Cole, Pacific Grove, CA.
- [5] Cleveland, W. S. and McGill, M. E. (1988) **Dynamic Graphics for Statistics**, Wadsworth Inc., Belmont CA.
- [6] Feigelson, E. D. and Babu J. G. (eds.) (1993) **Statistical Challenges in Modern Astronomy**, Springer-Verlag, New York.
- [7] Francis, Ivor (1981) **Statistical Software: A Comparative Review**, North Holland: New York.
- [8] Furnas, G. W. and Buja, A. (1994) "Prosection views, dimensional inference through sections and projections" (with discussion), **Journal of Computational and Graphical Statistics**, 3, 323-385.
- [9] Hayes, Annie (1982) **Statistical Software: A Survey and Critique of its Development**, Office of Naval Research, Arlington, VA
- [10] Huber, Peter J. (1994) "Huge data sets," **COMPSTAT: Proc. in Computat. Statist.**, 11th Symp., 3-13, (Dutter, R.; Grossmann, W. eds.) Physica-Verlag, Heidelberg.
- [11] Jaschek, C. and Murtagh, F. (eds.) (1990) **Errors, Bias and Uncertainties in Astronomy**, Cambridge University Press, Cambridge.
- [12] Majure, J. J., Cook, D., Cressie, N., Kaiser, M., Lahiri, S. and Symanzik, J. (1995) "Spatial CDF estimation and visualization with application to forest health monitoring," **Computing Science and Statistics**, 27, (Rosenberger, J. and Meyer, M., editors).
- [13] Murtagh, F. and Heck, A. (eds.) (1988) **Astronomy from Large Databases-Scientific Objectives and Methodological Approaches**, ESO Conference and Workshop Proceedings, No. 28.
- [14] Rolfe, E. J. (ed.) (1983) **Statistical Methods in Astronomy**, European Space Agency Special Publication ESA SP-201.
- [15] Statistical Sciences (1993) **Statistical Analysis in S-Plus**. Version 3.1, Seattle: StatSci, a division of MathSoft, Inc.
- [16] Symanzik, J., Majure, J. J. and Cook, D. (1995) "Dynamic graphics in a GIS: A bi-directional link between ArcView 2.0 and XGobi, **Computing Science and Statistics**, 27, (Rosenberger, J. and Meyer, M., editors).

- [17] Tierney, Luke (1990) **Lisp-Stat**, John Wiley and Sons, New York.
- [18] Trumpler, R. J. and Weaver, H. F. (1953) **Statistical Astronomy**, University of California Press, republished in 1962 by Dover, New York
- [19] Venables, W. N. and Ripley, B. D. (1994) **Modern Applied Statistics with S-PLUS**, Springer-Verlag, New York.
- [20] Wegman, E. J. (1990) "Hyperdimensional data analysis using parallel coordinates," **J. American Statist. Assoc.**, 85, 664-675.
- [21] Wegman, E. J. (1991) "The grand tour in k-dimensions," **Computing Science and Statistics: Proceedings of the 22nd Symposium on the Interface**, 127-136.
- [22] Wegman, E. J. (1995) "Huge data sets and the frontiers of computational feasibility," **Journal of Computational and Graphical Statistics**, 4(4), 281-195.
- [23] Wegman, E. J. and Carr, D. B. (1993) "Statistical graphics and visualization," in **Handbook of Statistics 9: Computational Statistics**, (Rao, C. R., ed.), 857-958, North Holland, Amsterdam.
- [24] Wegman, E. J. and Hayes, A. R. (1988) "Statistical software," **Encyclopedia of Statistical Sciences**, 8, 667-674.
- [25] Wegman, E. J. and Luo, Qiang (1995) "Visualizing densities," Technical Report 100, Center for Computational Statistics, George Mason University, Fairfax, VA
- [26] Wegman, E. J. and Luo, Qiang (1996) "High dimensional clustering using parallel coordinates and the grand tour," Technical Report 124, Center for Computational Statistics, George Mason University, Fairfax, VA

Discussion by John Nousek

Prof. Wegman has given an excellent summary review of a sample of available statistical packages. I encourage him to make his listing available on the World Wide Web to serve as a useful entry point for people searching for relevant packages. (*Prof. Wegman interjected at this point to say that his talk was already posted at <http://www.stat.gmu.edu/papers/astr.html>.*)

Two aspects of his presentation are worth emphasizing. He recognized the great value of both commercial software and shareware. The former can be expected to be comprehensive and well supported, while the latter can be expected to contain the latest results of research and the newest and most innovative functions and interfaces. It is wise to consider both types, with the URLs given by Prof. Wegman as a starting point.

Another excellent point is that the use of statistical software should not be limited to a tool kit of tests, but realize that before the tests are framed the astronomer may well first need to visualize his or her data. Many of the data sets mentioned during this workshop are enormous and require visualization tools to even begin to comprehend the data. XGobi is particularly interesting to try.

This brings me to the metaquestion raised by the talk: why don't astronomers already know about these tools? I think the answer is that astronomers do not routinely use *any* statistics packages, except for those already integrated into the software environments they are using for data reduction (IRAF, AIPS, IDL, MIDAS, etc.). In the remainder of my talk I present a hypothetical dialog to understand why such packages are not in greater use, perhaps pointing the way to removing such reluctance.

Why don't astronomers use statistics packages?

"If you need statistics to claim a result it's probably not true"

Such a statement may be more characteristic of the older range of practicing astronomers, but it reveals a deep discomfort with application of advanced statistical methods. Taken to extremes this reply is patently false as it refutes the entire scientific method, but there is a grain of truth to the argument that if a critical measurement depends on the kind of statistical argument used to decide whether a theory is confirmed or rejected, then there is too great a chance that the statistics are uncertain to come to a definitive conclusion on the result.

Note that this attitude is heavily influenced by the nature of the field of astronomy. Especially in the past, observing capabilities were limited so the number of cases being examined to test a result were usually small. Many times observation of some kind of anomaly, or unique case formed the lynchpin to simulate a new theory.

Such samples are very difficult to handle properly. Calculation of *a posteriori* probabilities, with strong selection biases is very difficult. Samples had totally unknown systematics. The small sample sizes would often be characterized by non-Gaussian (especially Poisson) distributions, but normal errors were typically assumed.

Today the larger samples, more uniform detection limits, and better experimental design are removing these objections, but a healthy skepticism about results will probably remain.

"If you use a package you don't know what you are really doing"

Astronomers generally come from a physics background. This culture has a great optimism that the Universe is amenable to reduction to simple, elementary principles. The paradigm holds that anything that is properly understood can be explained on the back of an envelope. Use of large packages can often obscure what operations are actually being performed on the data and break the connection to the physical interpretation of the results.

While mere use of a package is not necessarily bad, it is critical that the package provide excellent documentation at all levels (entry-level, where the beginner starts to use the package; usage-level, where an experienced

user can add new and advanced capabilities; and reference-level, where descriptions of the algorithms and rigorous treatments of the validity and assumptions are offered).

“My case isn’t in the package”

Often particular problems have aspects which don’t seem to be addressed by the packages. Examples of common astronomical cases which don’t seem covered by the packages are: variable error bars (i.e. different sized uncertainties associated with measurements, called by statisticians heteroscedastic); errors in both independent and dependent values; censored and truncated data with errors in the truncation and censoring levels; Poisson point processes; spherical distribution effects; and quantified uncertainties in the models.

“It takes too long to find out how to do what I need using this package”

Even if the capability is in the package it may be difficult to find it, and having found it there is a learning curve to climb to figure out how to use the package beyond the effort to learn the needed statistics.

The learning curve problem is not serious if an astronomer just repeatedly attacks the same kind of questions, but often the astronomer will be involved in many different kinds of analysis efforts, each needing different tools.

“If I don’t use a ‘standard’ statistical tool no one will believe the result”

The astronomer must also confront the problem that results need to be published and defended. If a statistical technique has not been widely applied in astronomy before, then there are the additional burdens of convincing the journal referees and the community at large that the statistical methods are valid.

Certain techniques which are widespread in astronomy and seem to be accepted without special justification are: linear and non-linear regression (χ^2 analysis in general), Kolmogorov-Smirnov tests, and bootstraps. It also appears that if you find it in *Numerical Recipes* (Press et al. 1992) that it will be more likely to be accepted without comment.

Outside of these cases the astronomer has to present his own demonstration of validity, which may be difficult and is likely to be time consuming. The general demonstration of validity usually requires a Monte Carlo simulation, as the community does not ordinarily deal in proofs or transformations. Note an insidious effect of this bias, astronomers will often choose to utilize a widely accepted statistical tool, even into regimes where the tool is known to be invalid, just to avoid the problem of developing or researching appropriate tools.

"I want the statistics interactive with the rest of my data extraction"

Many astronomers spend a great deal of time interacting with their data, using powerful analysis environments specially designed for astronomy. The data may be in internal formats, and the data extraction process may be closely coupled with the results of statistical fitting. (For example an image of the sky can be used to select a region for spectral analysis. The result of the analysis can lead to making more images, and the images more spectra.) Examples of astronomical analysis environments which merge interactive graphics, image analysis, spectral analysis and time series analysis are IDL, IRAF, and XANADU.

If an external statistical package cannot accept data in standard forms from these packages, then the work involved in writing a format translation is likely to overwhelm the gain in using the statistical package.

Within astronomy this same problem exists in that users develop strong attachments to particular environments. These attachments arise from the suite of tools available, the ease of interaction with data in these environments, or simply getting comfortable with a particular way of doing things. The best interfacing solution we have come up with in astronomy is the FITS (Flexible Image Transport System) standard, which has grown to accommodate many more structures than just images. All the packages in standard use in astronomy can import and export data in FITS format, which allows the adventurous user to freely mix and match tools across environment boundaries.

Major Astronomical Projects

Four examples of ‘big science’ in modern astronomy, each requiring the investment of dozens of scientists and $\$10^7 - 10^9$ of funds. AXELROD describes the search for a few gravitational lensing events, possibly associated with the mysterious dark matter in the Galaxy, amidst a million stellar light curves. SCHUTZ and commentator CUTLER discuss the statistical difficulties encountered in the effort to discover weak signals in long time series expected from a new generation of gravitational wave observatories. SIEMIGINOWSKA and commentator HOROWITZ outline problems and possible solutions for modeling high-energy spectra to be obtained with a forthcoming X-ray astronomical satellite, while VAN LEEUWEN and commentator BICKEL discuss time series data from a recently completed stellar photometric satellite experiment.

12

Statistical Issues in the MACHO Project

**T. S. Axelrod, C. Alcock, R. A. Allsman,
D. Alves, A. C. Becker, D. P. Bennett,
K. H. Cook, K. C. Freeman, K. Griest,
J. Guern, M. J. Lehner, S. L. Marshall,
B. A. Peterson, M. R. Pratt, P. J. Quinn,
A. W. Rodgers, C. W. Stubbs,
W. Sutherland, and D. L. Welch**

ABSTRACT The MACHO project is an ongoing survey project which collects photometric timeseries data on roughly 20 million stars in the LMC, SMC, and Galactic bulge. The data is irregularly time sampled, largely due to weather interruptions, and has noise which is non-stationary and non-Gaussian. The time series data is analyzed for microlensing events, rare brightenings of a star that result from an otherwise undetected massive object that passes close to the line of sight, thereby forming a gravitational lens. In addition to microlensing events, there are a much larger number of brightenings that result from intrinsic stellar variability. This background is only partially understood, since some classes of variable stars have received little study, and there are doubtless some classes yet to be discovered. Since the background can not be reliably simulated, the experiment must aim at a false alarm rate near zero, at the cost of reduced detection efficiency. At the same time, to achieve the scientific results at which it aims, the detection efficiency must be measured reliably.

12.1 The MACHO project

The MACHO [MAssive Compact Halo Objects] Project was initiated to test the hypothesis that a significant fraction of the dark matter in the halo of our galaxy consists of normal baryonic matter in the form of compact dark objects such as brown dwarfs, white dwarfs, or other dense objects. We began our effort following a suggestion of Paczynski [Pac86] that such objects could be detected through gravitational microlensing. The basic idea is simple. One looks through the halo at a background star. If a massive dark object passes near to one's line of sight, the intensity of the background

star will be amplified by the gravitational lens formed by the dark object. Very high amplifications will be produced if the approach is near enough, and the relative motion between the background star and lens will make the amplification time dependent. So, at least in some circumstances, the effect should be readily measurable.

The principal difficulty with putting the idea into practice is that the probability that any particular star is being microlensed at a given moment is expected to be quite small. This probability is referred to as the microlensing optical depth, τ . A dark object of mass M in the halo will appreciably amplify the light of a background star only if the line of sight to the star passes within an Einstein ring radius, r_E , given by

$$r_E = \sqrt{\frac{4GmLx(1-x)}{c^2}} \quad (12.1)$$

where L is the observer-star distance and x is the ratio of the observer-lens and observer-star distances. The microlensing optical depth is the probability that a random line of sight through the halo passes within r_E of one of the lensing objects. For most models of the galactic halo this probability is of order 10^{-6} [Pac86, Gri91] if the halo dark mass is entirely made up of compact baryonic objects. This implies that we must observe at least 10^6 , and more realistically, 10^7 stars to have a realistic expectation of detecting microlensing.

Hence MACHO is designed to continuously monitor a large number of stars, looking for microlensing events. As of February 1996, we have detected over 80 microlensing events [AAA⁺95a, AAA⁺95c], and other microlensing searches [USK⁺94, ABB⁺93] have been successful as well. So the basic method has been proven. But, as will become clear in the remainder of this paper, interpreting the data is complex, and challenges remain.

12.2 Data acquisition and reduction

12.2.1 Image Acquisition

The MACHO project has full-time use of the 1.27-meter telescope at Mount Stromlo Observatory, Australia, until the year 2000. A dichroic beamsplitter and filters provide simultaneous images in two passbands, a ‘red’ band and a ‘blue’ band. Two very large CCD cameras are employed at the two foci; each contains a 2×2 mosaic of 2048×2048 pixel Loral CCD imagers. The pixel size is $15 \mu\text{m}$ which corresponds to $0.63''$ on the sky, giving a sky coverage of 0.7×0.7 degrees. Details of the telescope, camera system, and data acquisition pipeline are given by [HvHH⁺96], [SMC⁺93], [AAQ⁺95], respectively.

Observations are obtained during all clear nighttime hours, except for occasional gaps for telescope maintenance. Data taking began in July 1992,

and as of February 1996 nearly 40000 exposures have been taken with the system, resulting in approximately 3 terabytes of image data. The images are taken at standard sky positions ('fields'), of which we have defined 82 in the LMC [Large Magellanic Cloud], 21 in the SMC [Small Magellanic Cloud] and 94 in the [galactic] bulge.

For several reasons, the time sampling of our fields is quite nonuniform. Some of these, such as weather outages and seasonal limitations on which fields can be observed, are beyond our control. Others result from explicit decisions to give higher priority to some fields than to others. The major factor entering into this prioritization is the number of stars contained in a field, but the decision process is complex and has resulted in several changes in observation strategy since the project began.

12.2.2 Photometric reductions

Photometric measurements from these images are made with a special-purpose code known as SoDoPHOT, derived from DoPHOT [SMS93]. First, one image of each field with good seeing and dark sky is chosen as a 'template image'. This is processed in a manner similar to a standard DoPHOT reduction except that after one color of the image has been reduced, the coordinates of the stars found in the first color are used as starting points for the positions of stars in the second color, which improves the star matching between colors. (The final positions of the matched stars are forced to be the same in both colors, after allowing for differential refraction.) This procedure provides a 'template' catalog of stellar positions and magnitudes for each field.

All other images are processed in 'routine' mode, which proceeds as follows. First the image is divided into 120 'chunks' of $\sim 512 \times 512$ pixels, and for each chunk ~ 30 bright unsaturated stars are located and matched with the template. These stars are used to determine an analytic fit to the point spread function, a coordinate transformation, and a photometric zero point relative to the template. Then, all the template stars are subtracted from the image using the model PSF [Point Spread Function] and coordinate transformation; noise is added to the variance estimate for each pixel to allow for errors in the subtraction. Next, photometric fitting is carried out for each star in descending order of brightness, by adding the analytic model of the star back to the subtracted frame and fitting a 2-parameter fit to the star's flux and sky background, with pixels weighted by inverse variance, while the model PSF and computed position of the star are kept fixed. When a star is found to vary significantly from its template magnitude, it and its neighbors undergo a second iteration of fitting. For each star, the estimated magnitude and error are determined, along with 6 other parameters (quality flags) measuring the object 'type' (single/blended etc.); the χ^2 of the PSF fit; the 'crowding', i.e. the amount of flux contributed from nearby stars; the weighted fractions of the PSF masked due to bad

pixels and cosmic rays respectively; and the fitted sky value. The photometric error estimate is the PSF fit uncertainty (as in DoPHOT) with an empirically determined 1.4% systematic error added in quadrature.

12.2.3 Noise characteristics

The noise characteristics of the data vary greatly both between individual stars and over time. Although the major contribution to the noise level of an individual measurement is the Poisson noise resulting from the finite number of photons collected in the stellar PSF, the full story is much more complex. Atmospheric conditions, which affect the PSF width, or ‘seeing’, vary on a variety of timescales from hours to weeks. Although the photometry code compensates for this through its PSF fitting procedure, it does so imperfectly, and as a result fluctuations in seeing add noise to the resulting star intensity measurements. Similar effects occur due to changing sky brightness, and a myriad of smaller effects.

The result is noise which is strongly non-Gaussian. Figure 1 shows the noise distribution from a composite of 68 lightcurves of non-variable stars,

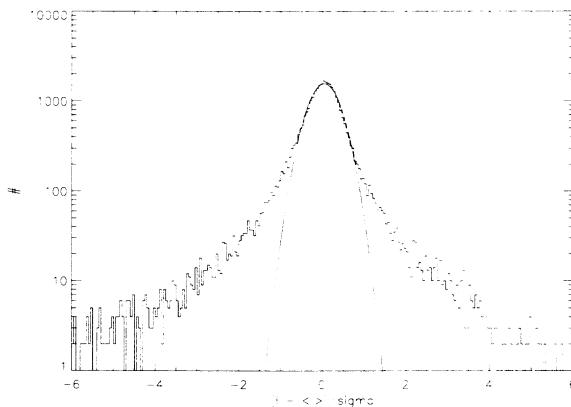


FIGURE 1. Noise distribution for a composite of 68 lightcurves of nonvariable stars. X axis is deviation from mean in units of distribution standard deviation.

and compares it to the best fit Gaussian. The noise has a Gaussian core, but strongly non-Gaussian wings. Although we have not fully investigated this, it also seems certain that the noise is correlated over a variety of timescales.

12.3 Current analysis techniques

In this section we take a simplified view of the scientific goals of the project: we consider only observations of the LMC, and desire to compare the number and durations of the microlensing events that we find to the predictions of theoretical models for the structure of the galactic dark halo. Several analysis tasks are required to accomplish this. First, microlensing events must be detected; secondly, the detection efficiency of the experiment must be measured; finally, the observations must be compared in some astrophysically meaningful way with the predictions of the models. We discuss each of these tasks in turn.

12.3.1 Event detection

Our discussion of event detection must begin with the predicted characteristics of microlensing lightcurves. In the case that both the source and the lens objects are point sources, the theory is simple (eg. [Pac86]), and predicts the following form for the amplification:

$$A(t) = \frac{u(t)^2 + 2}{u(t)\sqrt{u(t)^2 + 4}} \quad (12.2)$$

where

$$u(t) = \sqrt{u_{min}^2 + (2(t - t_0)/\hat{t})^2} \quad \text{and} \quad \hat{t} = \frac{2r_E}{v_\perp} \quad (12.3)$$

Here t_0 is the time at which maximum amplification occurs, r_E is the Einstein ring radius, and v_\perp is the transverse relative velocity between lens and source. The lightcurves thus form a two parameter family, with the parameter \hat{t} setting the overall timescale of the event, and u_{min} being a measure of how close the lensing object comes to the line of sight. Figure 2 shows this lightcurve family. When we first designed the experiment, we realized that our major source of background (i.e. false detections) would be variable stars of various kinds, and designed our detection algorithm to exploit the following differences between microlensing events and known variables:

1. Since the microlensing optical depth is so low, only one event should be seen in any given star.
2. The gravitational deflection of light is wavelength-independent, hence the star should not change color during the amplification.
3. The accelerations of galactic objects are negligible on timescales of these events, hence the events should be symmetrical in time, and have a shape given by equation 2 above.

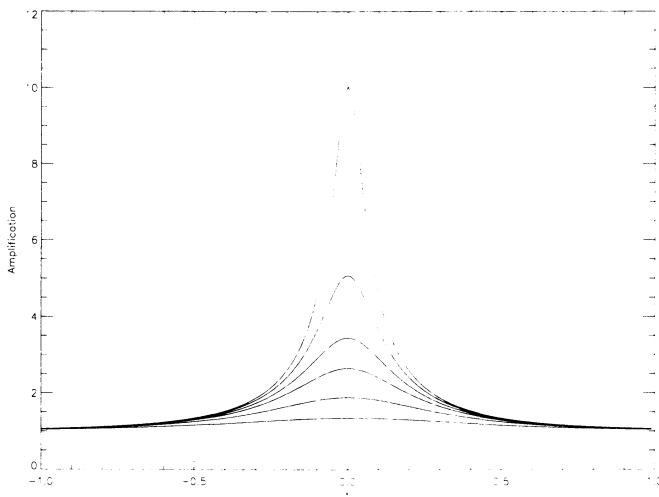


FIGURE 2. Lightcurve family generated by point-source, point-lens (PSPL) microlensing model. X axis is time in units of \hat{t} . Curves are shown for $u_{min} = 0.1, 0.2, 0.3, 0.4, 0.6, 1.0$.

All these characteristics are distinct from known types of intrinsic variable stars; most variable stars are periodic or semi-regular, and do not remain constant for long durations. Most change temperature and hence color as they vary, and usually have asymmetrical lightcurves with a rapid rise and slower fall.

The detection scheme we used for our first year of data [AAA⁺96a, AAA⁺95d] relied heavily on these differences. However, we have now detected over 80 events which we attribute to microlensing, with most of them being toward the Galactic bulge [AAA⁺95a, AAA⁺95c], and have had ample opportunity to realize that the simple microlensing lightcurve does not adequately describe a significant number of the events we see. This complicates the situation, and has forced us to revise our detection algorithm, a process which is likely to continue for some time as we gain experience.

Before outlining our current event detection algorithm, it is useful to describe some of the cases where the point source-point lens microlensing model breaks down in practice.

1. Since most of our star fields are quite crowded, a detected object is often not a single star, but a blend of two or more closely spaced stars. Although our images can not resolve the individual stars, their separations are still large compared with the Einstein ring radii, and

so only a single one will participate in a microlensing event. The presence of the unlensed stars modifies the shape of the lightcurve given in (x) and may make it chromatic.

2. A substantial fraction of stars are members of binary systems. If the lens is a binary, the lightcurve can be drastically modified, exhibiting caustics and other complex behavior [MDS95]. If the source is a binary the lightcurve is again modified, being a superposition of two microlensing curves, and can again be chromatic.
3. In some cases the finite angular size of the source star can modify the shape of the lightcurve peak.
4. The effects of the accelerated motion of the Earth can significantly modify the shapes of long duration microlensing events.

With this broader definition of microlensing events, the originally clear distinction between microlensing and variable stars becomes blurred, particularly if the signal-to-noise is low. Of the three criteria listed above, only the first, non-repetition of the event, survives unmodified. Our current event detection algorithm takes these issues into account, although imperfectly, and is performed as follows:

Before starting the microlensing search, measurements with substandard PSF chi square, crowding, missing pixel, or cosmic ray flags are flagged as ‘suspect’ measurements and removed from further consideration. We require that stars have an acceptable template measurement, and at least 7 simultaneous red-blue data pairs, to be searched for microlensing. The reddest stars, those with $V - R > 0.9$, are excluded from the microlensing search because they are often long-period variables. The event detection then proceeds in two stages; the first stage to define a ‘loose’ collection of events is very similar to that described in [AAA^{+95d}]; briefly, a set of matched filters of timescales 7, 15 and 30 days is run over each lightcurve. If after convolution, a lightcurve shows a peak above a pre-defined significance level in both colors, it is defined as a ‘level-1 candidate’, a full 5-parameter fit to microlensing is made, and many statistics describing the significance of the deviation, goodness of fit etc. are calculated. We use the standard point-source, point-lens approximation (x), which will be referred to as the ‘PSPL’ model. There are 5 free parameters: the un-lensed flux in red and blue passbands, the minimum impact parameter in units of the Einstein radius, u_{min} , the Einstein diameter crossing time \hat{t} , and the time of maximum amplification t_0 .

Lightcurves passing loose cuts on these statistics are defined as ‘level-1.5’ candidates. More stringent cuts are then applied to select final ‘level-2’ candidates.

As explained above, the criteria for ‘final’ microlensing candidates have been modified from those used in [AAA^{+95d}]. One parameter that is used

for a number of the cuts is $\Delta\chi^2 \equiv \chi_{\text{const}}^2 - \chi_{ml}^2$, where χ_{const}^2 and χ_{ml}^2 are the χ^2 values for the constant flux and microlensing fits respectively. χ_{peak}^2 refers to the χ^2 of the microlensing fit in the “peak” region where $A_{\text{fit}} > 1.1$. We use the following criteria to select candidate microlensing events:

1. The fitted time of peak amplification t_{max} must be within the time span of the observations, and the event duration $\hat{t} < 300$ days.
2. There must be at least 40 ‘baseline’ points on each light curve outside the time interval $t_{\text{max}} \pm 2\hat{t}$, and the χ^2 per degree of freedom (d.o.f.) of the microlensing fit in this region must be $\chi_{ml-out}^2(\text{d.o.f.}) < 4$.
3. We require a fit $A_{\text{max}} > 2 \times$ the mean estimated error of the data points.
4. We require at least 6 data points $> 1\sigma$ above median brightness in the peak region $t_{\text{max}} \pm 0.5\hat{t}$.
5. We exclude stars brighter than $V < 17.5$ which contain a class of bright blue variables known as “bumpers”.
6. We exclude stars in a region surrounding SN 1987A in order to avoid spurious microlensing triggers due to the supernova light echo.
7. We remove events with low signal to noise or a poor peak fit with $\Delta\chi^2/\chi_{\text{peak}}^2(\text{d.o.f.}) > 200$.
8. We remove stars that may have crowding related spurious photometry.
9. Our main signal-to-noise cut is $\Delta\chi^2/\chi_{ml}^2(\text{d.o.f.}) > 500$.
10. We require a fit $A_{\text{max}} > 1.75$.

Two examples of lightcurves that pass these cuts in the first 2 years of LMC data [AAA⁺96b] are shown in Figure 3. These lightcurves are chosen to show roughly the range of quality in the events which pass our current cuts, with one being near to our minimum amplification cut, and the other showing higher signal-to-noise.

12.3.2 Efficiency determination

The principal goal of our experiment is to quantitatively compare the predictions of galactic models with the numbers and durations of the microlensing events we observe. To do so, we need to be able to measure our detection efficiency, i.e. given a real microlensing event of amplitude A and duration \hat{t} that lenses a star with magnitude M , what is the probability

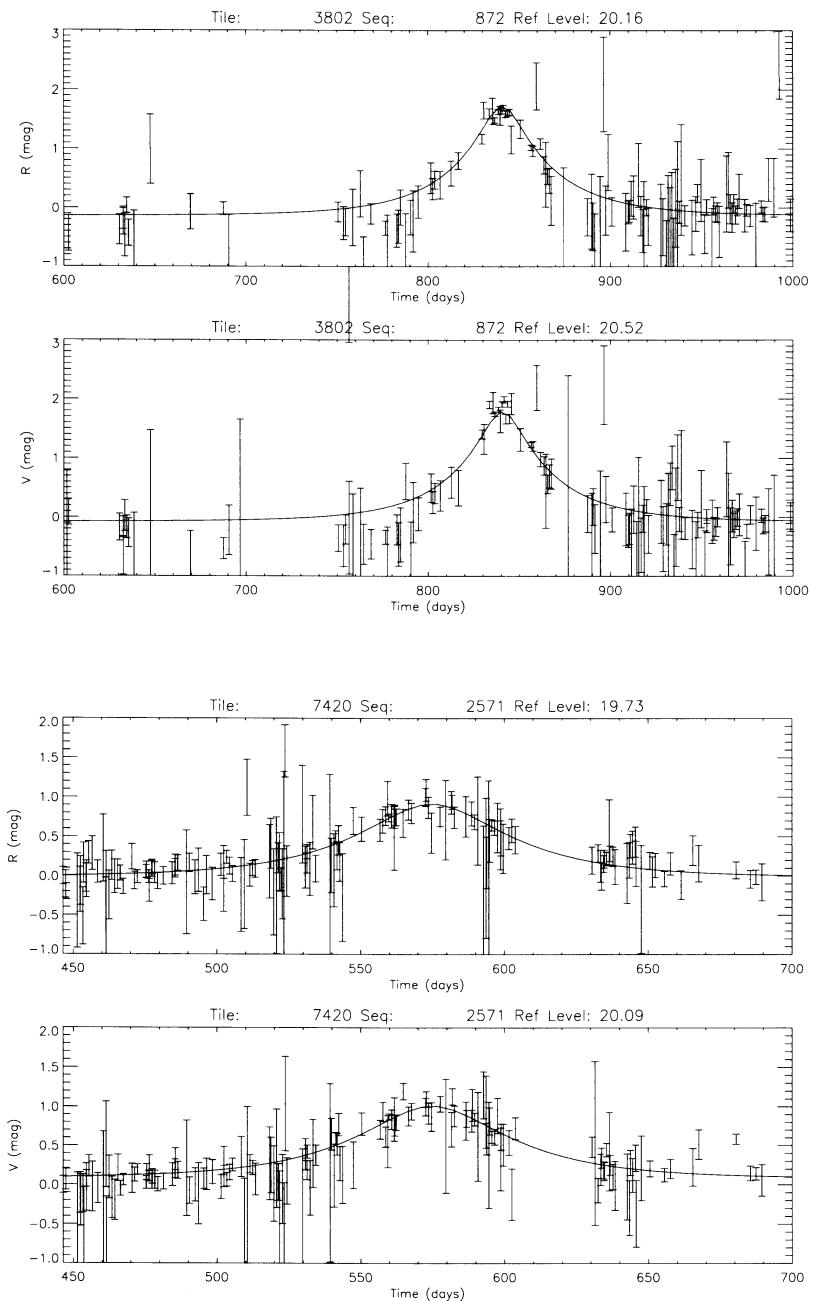


FIGURE 3. Two lightcurves of LMC microlensing events that pass the analysis cuts. Y axis shows brightness in magnitudes.

$p(A, \hat{t}, M)$ that we detect it? The problem is more difficult than appears at first glance. One major issue is that it is unclear how many parameters are required to determine this probability. For example, what about the crowding of a star? We have every reason to expect that the performance of the photometry code will degrade as a star of given magnitude is placed in more crowded environments, and that this will affect the efficiency. What about its color? Again, an argument could be made that the efficiency should be color dependent. Clearly, the list of possible parameters could be made almost arbitrarily long. It is essential to arrive at some defensible parameterization, however. Certainly it is not practical to determine the efficiency for each of twenty million individual stars!

A second issue is that in general an event will not be recovered with its true amplitude and duration. Blending effects in crowded fields result in biased estimates for both A and \hat{t} .

The approach we have taken to efficiency calculation takes account of all these issues, although imperfectly. The procedure, in outline, is the following:

First, artificial stars are added to real images taken in a wide variety of seeing conditions and sky brightness. Care is taken in this process to model the real PSF as accurately as possible, and the magnitudes and colors of the artificial stars are randomly chosen from our known distribution of detected stars. The intensities of the artificial stars are then systematically varied through a wide range of amplifications. The resulting images are then processed through our photometry code, and the resulting intensities of the artificial stars are used to construct ‘response functions’, $RF(A; M, s, b)$, which give the recovered amplification as a function of the input amplification, A , and parameterized by the star brightness, M , the seeing, s , and the sky brightness, b . Approximately 10000 such response functions are calculated.

Secondly, the response functions are used to determine the response of the entire analysis process to simulated microlensing events. To do this, a 1% random sample of all objects we track in our database is created. Each object then has a particular sequence of times at which it was actually imaged by the telescope system. The idea is then to randomly generate microlensing events for each of these objects, and at each point in the time sequence for that object, use the proper response function to predict what amplification would have been obtained by the photometry code. The resulting time series of amplifications is then fed into the microlensing analysis path. The analysis may or may not detect microlensing in the time series. If it does, the recovered A and \hat{t} will not necessarily be the same as that of the simulated event. The resulting recovered information is accumulated and used to construct the efficiencies we require.

12.3.3 Comparison with models

Having detected events and determined our detection efficiency, we are in a position to compare our results with the predictions of theoretical models. There are two classes of questions that arise. The first class asks in some form whether our event distributions (as opposed to individual events) are compatible with the microlensing hypothesis. The answers to these questions do not depend on detailed models of the galaxy, but rather on general characteristics of microlensing. Examples of these include:

1. Is the distribution of u_{min} consistent with microlensing?
2. Is the distribution of the lensed stellar magnitude and color consistent with the color-magnitude distribution for the whole stellar population?
3. Is the spatial distribution of events compatible with microlensing by halo objects?

In all of these cases, the simple predictions of microlensing theory, for example that the event distribution should be uniform in u_{min} , must be modified by the detection efficiency. This efficiency falls as u_{min} increases, and in fact, with our current cuts is set to zero when $u_{min} > 0.66$. Therefore, the validity of the tests depends on the certainty with which the efficiency is known. We currently have only rough estimates of this. The statistical tests themselves are currently being performed with the K-S test, and its two-dimensional form [FF87]. It is not clear whether more powerful tests exist.

The second class of questions compares our results with detailed galactic models, and in some sense the answers represent the ultimate scientific results of the project. Space limitations preclude covering this aspect in detail here. The interested reader is referred to [AAA^{+95d}, AAA^{+95b}]. However, it is worth mentioning some of the issues that come up, even though they are shared with many other problems in astronomy.

There are really two problems here. The first is: Given a galactic model, parameterized by some set of parameters G , what is the likelihood $p(E; G)$ that we would observe our set of event amplitudes and durations, E ? If the model is well specified, and the detection efficiency is well determined, this question is straightforward, although messy, to answer.

The second problem does not have such a clear answer: What is a reasonable space of models, and how should we assign prior probabilities in it, so that we can answer the questions of greatest interest from an astrophysical point of view: What is the best estimate of the MACHO fraction in the halo? What is the best estimate of the masses of the MACHOs? For two approaches to this, the reader is referred to [AAA^{+95d}, GGT95], but it is fair to say that a lot of work is left to be done in this area.

12.4 An unresolved issue

In the preceding discussion the determination of the detection probability, p_D , and its importance to our final results, has featured prominently. The complementary quantity, the false alarm probability, p_{FA} , has so far not been discussed. We in fact expect p_{FA} to play an equally important role. Viewed from a sufficient distance, the entire experiment can be quantified in terms of a ‘receiver operating curve’, or ROC, a function which relates p_D to p_{FA} . As always when faced with a ROC, the challenge is to pick a point along it where the false alarm rate is just tolerable, thereby maximizing the detection rate. However this choice is made, knowledge of p_{FA} is crucial when interpreting the results of the experiment in the light of some model.

Our situation is unusual in that, while the detection probability is conceptually straightforward to calculate, the false alarm probability can not at present be determined with confidence. The reason for this is the nature of our background. As discussed earlier, the major source of background is certainly variable stars. But this background is imperfectly understood, and there is even the possibility that some heretofore unknown class of variables exists that can mimic microlensing. Given the expected rarity of microlensing events, an equally rare type of variable star could plausibly remain undiscovered, and could cause a major error in interpretation if it were to be confused with microlensing. Only much larger datasets, spanning longer times, can definitively answer this last concern. Given a large enough set of microlensing events, for example, the test mentioned in Section 3.3 that checks the population of lensed stars for consistency with the entire stellar population, should be able to remove this concern.

Most astronomers will probably feel comfortable in putting aside concern about unknown types of variable stars, and perhaps we are then justified in asserting that p_{FA} is “low enough”, even though we do not know its value. But, even granted that, there is still a related problem that affects the determination of both p_D and p_{FA} . Perhaps it is one which is amenable to solution with the data that we have available now.

As we have seen above, the PSPL microlensing model is too restrictive. We have seen a number of events now which involve binary lenses and/or strong blending, as well as more subtle effects. It would seem that we should adapt to this reality simply by replacing the simple microlensing model everywhere it is used with a completely general microlensing model, which we’ll refer to as ‘GSGL’. So, in the analysis of lightcurves for events, discussed in Section 3.x, we would fit each lightcurve to the GSGL model. Similarly, in the calculation of the efficiency, instead of injecting PSPL lightcurves into the analysis pipeline, we would inject lightcurves from the GSGL model.

Two problems immediately arise. For the PSPL model, which has only two nontrivial parameters, u_{min} and \hat{t} , we know exactly what the a priori distribution of u_{min} should be, and can make reasonable guesses for the

distribution of \hat{t} . On the other hand, for the GSGL model, there are is a much larger number of nontrivial parameters. Deciding on the correct a priori probability distribution within this large parameter space is a difficult task with a very uncertain outcome. Additionally, some parameters are nearly degenerate with others, so that a lightcurve fit can not determine a unique parameter set.

The second, related, problem is that the GSGL model can generate such a wide range of behavior that we are sure that there will be false alarms from non-microlensing lightcurves which can be accurately fit by the GSGL model. Unfortunately, there is again no clear path to determining p_{FA} , and we are back to the problem that opened the discussion in this section. To give this aspect of the problem some reality, let's look at a case which actually arose in practice.

In 1994 a lightcurve was noticed that was an excellent fit to a binary source microlensing event (a subset of the GSGL model), as shown in Figure 4. Although it inhabits a region of the color-luminosity diagram in which variable stars known as ‘bumpers’ frequently turn up, it appeared that it might really be a microlensing event. In addition to the excellent lightcurve fit, its color behavior was significantly more achromatic than the bumpers that were then known, suggesting that it might be a different type of object. Opinion was divided on the nature of this event, with many of the arguments hinging on the poorly determined a-priori probabilities for observing such an event. The matter was definitively resolved only by more data: the star ‘bumped’ again while the argument was still in progress!

Although this particular event is eliminated in our current analysis by a cut in the color-magnitude diagram specifically intended to eliminate bumpers, we may not always be so fortunate. Certainly we need a better thought out procedure for applying more complex models of microlensing to our data.

12.5 Conclusions

I'd like to conclude with a philosophical point. A number of us came to the MACHO experiment with a background in high energy physics. Partly because of this, there initially appeared to be many similarities in the data analysis issues between high energy physics experiments and the sort of astronomy we were attempting to do. In both cases, one is searching for a very small number of special events in a large sea of events. The challenge was initially seen mainly in terms of selecting a set of “cuts” in parameter space that would optimally separate the signal from the background, and to do so with known detection efficiency and false alarm rates. As in high energy experiments, we hoped to determine the cuts through Monte Carlo simulations, with the idea that the cuts would be set once and for all, without being affected by what we found in the data.

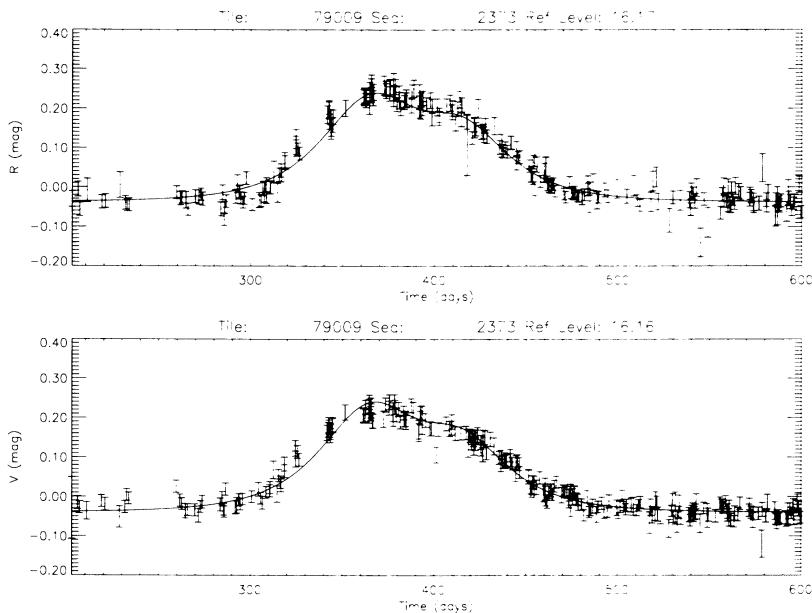


FIGURE 4. Lightcurve of a variable star with fit to a binary source microlens.

One lesson from the last two years of analysis is that this analogy, initially so attractive, is flawed. In spite of our best intentions, the cuts have evolved continuously in response to the data. Sometimes we have come uncomfortably close to adjusting a cut solely to place a particular event on one side or the other of the line. Although I think we have ultimately avoided that temptation, we have all been a little uncomfortable with the whole process.

Our behavior in this regard, however, has reflected the true nature of the situation: the background is imperfectly known, and our knowledge of it continues to grow as we accumulate data. To freeze our cuts before data taking began would simply have been foolish, no matter how elaborate a Monte Carlo simulation was undertaken.

But this is not to say that we cannot do a better, more systematic, job than we have done so far. This is a situation which calls out for a careful mathematical treatment, and if such is possible, a number of astronomy experiments will be greatly improved.

REFERENCES

- [AAA⁺95a] C. Alcock, R. A. Allsman, T. S. Axelrod, D. P. Bennett, S. Chan, K. H. Cook, K. C. Freeman, K. Griest, S. L. Marshall, S. Perlmuter, B. A. Peterson, M. R. Pratt, P. J. Quinn, A. W. Rodgers, C. W.

- Stubbs, and W. Sutherland. Probable gravitational microlensing towards the galactic bulge. *ApJ*, 445(1):133–139, 1995.
- [AAA⁺95b] C. Alcock, R. A. Allsman, T. S. Axelrod, D. P. Bennett, K. H. Cook, N. W. Evans, K. C. Freemann, K. Griest, J. Jijina, S. L. Marshall, B. A. Peterson, M. R. Pratt, P. J. Quinn, A. W. Rodgers, C. W. Stubbs, and W. Sutherland. Theory of exploring the dark halo with microlensing 1: Power-law models. *ApJ*, 449(1):28–41, 1995.
- [AAA⁺95c] C. Alcock, R. A. Allsman, T. S. Axelrod, D. P. Bennett, K. H. Cook, K. C. Freemann, K. Griest, S. L. Marshall, B. A. Peterson, M. R. Pratt, P. J. Quinn, A. W. Rodgers, C. W. Stubbs, W. Sutherland, and D. Welch. Gravitational microlensing results toward the galactic bulge. *ApJ*, 1995. submitted.
- [AAA⁺95d] C. Alcock, R. A. Allsman, T. S. Axelrod, D. P. Bennett, K. H. Cook, K. C. Freemann, K. Griest, J. A. Guern, M. J. Lehner, S. L. Marshall, H.-S. Park, S. Perlmutter, B. A. Peterson, M. R. Pratt, P. J. Quinn, A. W. Rodgers, C. W. Stubbs, and W. Sutherland. Experimental limits on the dark matter halo of the galaxy from gravitational microlensing. *Phys. Rev. Lett.*, 74(15):2867–2871, 1995.
- [AAA⁺96a] C. Alcock, R. A. Allsman, T. S. Axelrod, D. P. Bennett, K. H. Cook, K. C. Freemann, K. Griest, J. A. Guern, M. J. Lehner, S. L. Marshall, H.-S. Park, S. Perlmutter, B. A. Peterson, M. R. Pratt, P. J. Quinn, A. W. Rodgers, C. W. Stubbs, and W. Sutherland. The macho project first year lmc results: The microlensing rate and the nature of the galactic dark halo. *ApJ*, 1996. in press.
- [AAA⁺96b] C. Alcock, R. A. Allsman, T. S. Axelrod, D. P. Bennett, K. H. Cook, K. C. Freemann, K. Griest, J. A. Guern, M. J. Lehner, S. L. Marshall, H.-S. Park, S. Perlmutter, B. A. Peterson, M. R. Pratt, P. J. Quinn, A. W. Rodgers, C. W. Stubbs, and W. Sutherland. Microlensing of lmc stars in the macho 2-year data. *ApJ*, 1996. in preparation.
- [AAQ⁺95] T. S. Axelrod, R. A. Allsman, P. J. Quinn, D. P. Bennett, K. C. Freemann, B. A. Peterson, A. W. Rodgers, C. Alcock, K. H. Cook, K. Griest, S. L. Marshall, M. R. Pratt, C. W. Stubbs, and W. Sutherland. The macho data pipeline. *PASP*, submitted, 1995.
- [ABB⁺93] E. Aubourg, P. Bareyre, S. Brehin, M. Gros, M. Lachieze-Rey, B. Laurent, E. Lesquoy, C. Magneville, A. Milsztajn, L. Moscosco, F. Queinnec, J. Rich, M. Spiro, L. Vigroux, S. Zylberajch, R. Ansari, F. Cavalier, M. Moniez, J.-P. Beaulieu, R. Ferlet, Ph. Grison, A. Vidal-Madjar, J. Guibert, O. Moreau, F. Tajahmady, E. Maurice, L. Prevot, and C. Gry. Evidence for gravitational microlensing by dark objects in the galactic halo. *Nature*, 365(6447):623–625, 1993. EROS collaboration.
- [FF87] G. Fasano and A. Franceschini. A multidimensional version of the kolmogorov-smirnov test. *MNRAS*, 225:155–170, 1987.
- [GGT95] E. I. Gates, G. Gyuk, and M. S. Turner. Microlensing and halo cold dark matter. *Phys. Rev. Lett.*, 74(19):3724–3727, 1995.
- [Gri91] K. Griest. Galactic microlensing as a method of detecting massive compact halo objects. *ApJ*, 366(2):412–421, 1991.

- [HvHH⁺96] J. Hart, J van Harmelen, G. Hovey, K. C. Freemann, B. A. Peterson, T. S. Axelrod, P. J. Quinn, A. W. Rodgers, R. A. Allsman, C. Alcock, D. P. Bennett, K. H. Cook, K. Griest, S. L. Marshall, M. R. Pratt, C. W. Stubbs, and W. Sutherland. The telescope system of the macho program. *PASP*, 108:220, 1996.
- [MDS95] S. Mao and R. Di Stefano. Interpretation of gravitational microlensing by binary systems. *ApJ*, 440:22, 1995.
- [Pac86] B Paczyński. Gravitational microlensing by the galactic halo. *ApJ*, 304:1, 1986.
- [SMC⁺93] C. W. Stubbs, S. L. Marshall, K. H. Cook, R. Hills, J. Noonan, C. W. Akerlof, C. Alcock, T. S. Axelrod, D. P. Bennett, K. Dagley, K. C. Freemann, K. Griest, H.-S. Park, S. Perlmutter, B. A. Peterson, P. J. Quinn, A. W. Rodgers, C. Sosin, and W. Sutherland. A 32 megapixel dual color CCD imaging system. In M Blouke, editor, *Charge Coupled Devices and Solid State Optical Sensors III*, volume 1900 of *Proceedings of the SPIE*, 1993.
- [SMS93] P. L. Schechter, M. Mateo, and A. Saha. Dophot, a CCD photometry program - description and tests. *PASP*, 105(693):1342–1353, 1993.
- [USK⁺94] A. Udalski, M. Szymanski, J. Kaluzny, M. Kubiak, W. Krzeminski, M. Mateo, G. W. Preston, and B. Paczyński. The optical gravitational lensing experiment - the optical depth to gravitational microlensing in the direction of the galactic bulge. *Acta Astronomica*, 44(3):165–189, 1994.

13

LIGO: Identifying Gravitational Waves

Bernard F. Schutz¹ and David Nicholson²

ABSTRACT We discuss the statistical challenges presented by data from large-scale gravitational-wave interferometers now under construction. Extracting information from signals requires measuring parameters, and sophisticated algorithms may be required to attain maximal possible accuracy. Detecting continuous waves from neutron stars presents many problems, particularly for a survey of the sky, where optimum algorithms are not known.

13.1 Introduction

Identification of gravitational waves is an issue well deserved of its place on the agenda at this meeting. As a scientific endeavour, it poses a host of novel *statistical challenges* in the analysis of large time-series data sets for weak signals. Success promises rich scientific rewards and should pave the way to a *modern* new branch of *astronomy*.

Our intention here is to survey two of the key statistical challenges in gravitational wave data analysis. The first is information extraction from noisy gravitational wave signals. The second is the detection of signals of a periodic origin in coherent data taken over a year or more.

Information extraction for weak signals in noise depends upon estimating parameters of the family of filters used to detect the signal. This is clearly a vital step towards opening up a new gravitational window on the universe: it is not enough simply to detect signals, because we really want to use them to study their sources. As a concrete example we focus our attention upon gravitational waves from compact coalescing binary systems comprising of neutron stars and black holes. These systems are deemed by theorists to be, in the long run, the most promising sources of gravitational radiation for the

¹Department of Physics and Astronomy, University of Wales College of Cardiff, Cardiff, U.K. and Max Planck Institute for Gravitational Physics, The Albert Einstein Institute, Potsdam, Germany

²Department of Physics and Astronomy, University of Wales College of Cardiff, Cardiff, U.K.

very sensitive network of laser interferometric gravitational wave detectors that are now being assembled at sites in the United States (the LIGO project [1]) and in Europe (the French/Italian VIRGO [2] and UK/German GEO600 [3] projects). First observations may take place as early as 1999, but the detectors will continue to improve their sensitivity for many years after that.

Gravitational waves from coalescing binaries are encoded with a rich suite of information about their sources. Its extraction would lead to new insights about strong-field gravity, measurements of neutron star masses and black hole masses and spins, and new, independent, determinations of the cosmological parameters [4].

In order to gauge the usefulness of coalescing binaries as astrophysical probes of the universe, statisticians and data analysts have to meet two important challenges. The first is to compute the theoretical accuracy with which the information conveyed by the waves can in principle be extracted. The second is to devise practical measurement algorithms that attain this accuracy or at least come very close to it. We will describe progress towards meeting these challenges in Section 2.

The second key statistical challenge in gravitational wave data analysis that we shall examine here concerns continuous sources of gravitational radiation from systems involving neutron stars, such as pulsars. Our goal is to develop a practical algorithm for searching very long stretches of data for weak periodic signals subject to amplitude and frequency modulation due to the motion of the detector. We distinguish between targeted searches for waves arriving from a specific direction in the sky, such as the site of a known radio pulsar, and wideband searches across all of the sky for hitherto unknown periodic sources. Continuous wave sources and issues in search algorithm development are reviewed in Section 3. Our main conclusions are drawn in Section 4.

13.2 Information extraction from gravitational waves

The extraction of information from signals that are immersed in noise has a long and venerable history [5]. Let us consider an observation, for example of a waveform. A simple model is,

$$h(t) = s(t; \boldsymbol{\theta}) + n(t); \quad -T/2 < t < T/2 \quad (13.1)$$

where s is assumed to be known for all possible choices of the parameter vector $\boldsymbol{\theta} \equiv \theta_1, \dots, \theta_m$, and $n(t)$ is a realization of a zero-mean Gaussian random noise process. Suppose some measurement process on $h(t)$ designed to extract the parameter vector yields an estimate $\hat{\boldsymbol{\theta}}$ with error $\boldsymbol{\epsilon} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$. A complete statistical summary of the process is embodied in the error

covariance matrix, $\mathcal{R} \equiv \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T$. The diagonal elements of this matrix give the errors on each separate parameter while the off-diagonal terms are covariances of the errors. In the general case when the parameters enter the signal s nonlinearly, it is not possible to evaluate this matrix exactly. Simpler, computable, lower bounds on \mathcal{R} must instead be sought. There are two theoretical classes of lower bounds on parameter estimation accuracy: *local* bounds and *global*, or Bayesian, bounds. We briefly review their main features next, citing a common example drawn from each of the classes. Our intention is to point out that astronomers commonly use local bounds to gauge the potential accuracy of their experiments. However, global bounds are much more versatile for this purpose. Information theorists have recently developed tight global bounds that are relatively straightforward to compute. We describe one of these bounds here — the Ziv-Zakai bound — and apply it to compute measurement accuracy for the parameters of compact coalescing binary gravitational wave signals.

13.2.1 Local bounds

Local bounds treat the desired parameters as unknown deterministic quantities. They bound the minimum mean-square error (MSE) on the parameters. Their locality stems from the fact that they generally depend upon the actual values of the parameters themselves. Local bounds have two main drawbacks. The first is that they apply only to a restricted set of estimators, namely unbiased estimators. Such estimators do not exist in the important case when the parameter space is finite [14]. A further drawback of local bounds is that they cannot incorporate prior information that one might have about the parameters, such as their support. A familiar example of a local bound is the Cramer-Rao bound (CRB) [6]. This is a simple bound to compute and we have used it almost exclusively to calculate the theoretical accuracy with which gravitational wave parameters can be measured [10, 11]. The CRB has also found widespread application elsewhere in astronomy [7, 8]. Lending to the bound's appeal is the fact that it can be approached asymptotically, with increasing signal-to-noise ratio (SNR) by the well-known maximum-likelihood method of parameter estimation. However, astronomical observations, and in particular those we anticipate in gravitational wave astronomy, can rarely afford high SNR. In operating conditions of low-to-moderate SNR the CRB can often be a very weak lower bound. One should thus look for bounds that have a more general domain of validity. Barankin's bound is one such local bound that has found application in radar and sonar problems [9], but it is rather difficult to compute in the general case. A better alternative is to turn to global bounds which are much more versatile than their local counterparts.

13.2.2 Global bounds

Global bounds treat the unknown parameters of the signal as random variables with known prior distributions. They bound the global MSE averaged over these prior distributions. Weighing heavily in their favour, global bounds are not restricted to a certain class of estimator. Indeed, they are completely independent of the estimation process. In addition, global bounds are easily able to incorporate any prior information one may have about the desired parameters. An example of a global bound is the conditional-mean estimation bound (CMB) [5]. In fact, this is not really a lower bound since it can in principle be attained by the conditional-mean estimator. The latter has been briefly discussed in the context of gravitational wave parameter estimation, where it is referred to as the method of nonlinear filtering [11, 19]. Unfortunately the bound is generally impossible to compute, involving multidimensional integrals over the parameter space. Recently, however, information theorists have devised tight global bounds that are much less difficult to compute [14, 15]. One of these is the Ziv-Zakai bound (ZZB) [14]

13.2.3 Ziv-Zakai bound

This bound has been restricted in application to parameter estimation (*e.g.* direction-of-arrival, Doppler modulation) problems in radar and sonar [16]. We feel, however, that it may also be useful to astronomers. A simple derivation of the bound was pointed out to us by Kristine Bell [16]. Space limits its inclusion here. Instead, we present only the key points and a general expression for the bound.

Consider a K -dimensional vector random parameter $\boldsymbol{\theta}$ for which an estimate is sought, based on noisy observations \mathbf{X} . An estimator $\hat{\boldsymbol{\theta}}(\mathbf{X})$ yields an error $\boldsymbol{\epsilon} = \hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}$ and error covariance matrix \mathcal{R} . We seek a lower bound on $\mathbf{a}^T \mathcal{R} \mathbf{a}$, where \mathbf{a} is any K -dimensional vector.

The start point for the ZZB is an identity drawn from the theory of stochastic processes [17],

$$\mathbf{a}^T \mathcal{R} \mathbf{a} = \langle |\mathbf{a}^T \boldsymbol{\epsilon}|^2 \rangle = \int_0^\infty \frac{\Delta}{2} \Pr \left(|\mathbf{a}^T \boldsymbol{\epsilon}| \geq \frac{\Delta}{2} \right) d\Delta \quad (13.2)$$

Now the probability term under the integral can be arranged into a form such that it can be identified as the probability of error, P_e , in a detection problem defined by the hypotheses,

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\phi}; \quad \Pr(H_0) = \frac{p(\boldsymbol{\phi})}{p(\boldsymbol{\phi}) + p(\boldsymbol{\phi} + \boldsymbol{\delta})}; \quad \mathbf{X} \sim p(\mathbf{X} | \boldsymbol{\theta} = \boldsymbol{\phi}) \quad (13.3)$$

$$H_1 : \boldsymbol{\theta} = \boldsymbol{\phi} + \boldsymbol{\delta}; \quad \Pr(H_1) = 1 - \Pr(H_0); \quad \mathbf{X} \sim p(\mathbf{X} | \boldsymbol{\theta} = \boldsymbol{\phi} + \boldsymbol{\delta}) \quad (13.4)$$

and subject to a suboptimal decision rule,

$$\text{Decide } H_0 : \boldsymbol{\theta} = \boldsymbol{\phi} \quad \text{if} \quad \mathbf{a}^T \hat{\boldsymbol{\theta}}(\mathbf{x}) > \mathbf{a}^T \boldsymbol{\phi} + \frac{\Delta}{2} \quad (13.5)$$

$$\text{Decide } H_1 : \boldsymbol{\theta} = \boldsymbol{\phi} + \boldsymbol{\delta} \quad \text{if} \quad \mathbf{a}^T \hat{\boldsymbol{\theta}}(\mathbf{x}) \leq \mathbf{a}^T \boldsymbol{\phi} + \frac{\Delta}{2} \quad (13.6)$$

The ZZB is obtained by bounding P_e with the minimum probability of error, $P_{\min}(\boldsymbol{\theta}, \boldsymbol{\theta} + \boldsymbol{\delta})$, based on an optimal likelihood-ratio test of the two hypotheses.

A closed form expression for $P_{\min}(\boldsymbol{\theta}, \boldsymbol{\theta} + \boldsymbol{\delta})$ generally cannot be found, however it can be tightly bounded from below for equally likely hypotheses [5]. Furthermore, the computation of the bound is often simplified because in real problems (like the one we consider next), the probability of error is often only a function of $\boldsymbol{\delta}$ and not of $\boldsymbol{\theta}$. In this case the ZZB takes the following form

$$\mathbf{a}^T \mathcal{R} \mathbf{a} \geq \int_0^\infty \Delta \cdot V \left\{ \max_{[\boldsymbol{\delta}: \mathbf{a}^T \boldsymbol{\delta} = 1]} P_{\min}^{\text{e.l.}}(\boldsymbol{\delta}) A(\boldsymbol{\delta}) \right\} d\Delta \quad (13.7)$$

where

$$A(\boldsymbol{\delta}) = \int \min[p(\boldsymbol{\theta}), p(\boldsymbol{\theta} + \boldsymbol{\delta})] d\boldsymbol{\theta} \quad (13.8)$$

and $V\{\cdot\}$ is a valley-filling function [16]. Calculation of the minimum probability of error for equally likely hypotheses, $P_{\min}^{\text{e.l.}}$, depends on the ambiguity function of the signal and $A(\boldsymbol{\delta})$ depends on the prior distribution of $\boldsymbol{\theta}$. The tightest bound is given by maximizing over the vector $\boldsymbol{\delta}$, subject to the constraint $\mathbf{a}^T \boldsymbol{\delta} = \Delta$. In practise, this means defining a path of integration that captures structure in the signal's ambiguity function.

13.2.4 Application

Let us now apply the ZZB (13.7) to estimate the accuracy with which the parameters of a noisy Newtonian compact coalescing binary waveform can be measured. This waveform has been written down many times and in a certain parametrization can be described by: an arrival time at the detector t_a , a phase upon arrival ϕ_a , and a ‘chirp’ time τ that depends on the two component masses of the system [11, 12]. We shall assume that these parameters have uniform priors. However this does not impact on our results because the parameters regions of support are much larger than their typical estimation errors. A strong correlation exists between t_a and τ which shows up as a ridge in the signal’s ambiguity function [12]. This defines a suitable path along which to integrate to obtain the tightest possible ZZB for this problem. Assuming additive Gaussian noise with a spectral shape characteristic of the the first stage LIGO detectors, we compute the bound using a standard numerical integrator for various SNR’s. Our results are presented below.

13.2.5 Result

The conventional approach to parameter estimation for gravitational wave signals is the method of maximum likelihood (ML) [12]. A bank of filters for the assumed known waveform shape are correlated against the noisy data and the parameters of the filter that trigger the peak correlation are identified as the parameters of the gravitational wave signal.

Monte-Carlo simulations of this method have recently been applied to coalescing binary signals embedded in noise with the same statistical properties as the noise we adopted for our theoretical calculations above. The numerical results (filled circles) for errors on the time-of-arrival parameter as a function of SNR are compared against the ZZB (unfilled circles) and CRB (bold line) in Fig. 1. This parameter is an important one for gravitational wave astronomy: it governs how accurately we are able to locate the positions of gravitational wave sources on the sky. The ZZB on time-of-arrival accuracy deviates from the CRB at an SNR of ~ 15 , but only by a few percent. We can understand this in terms of the shape of the signal ambiguity function. The CRB essentially probes the curvature of this function. In the presence example, the ambiguity function has only one, rather extended, lobe. In test problems, we have seen that the ZZB deviatea markedly from the CRB when the signal ambiguity function has a ‘choppy’ structure, comprising of a main lobe with a string of large secondary sidelobes. As the CRB probes the curvature of the main lobe only it is effectively ‘blind’ to the presence of the sidelobes and a poor predictor of real measurement errors. Although the ZZB results in Fig. 1 are a little tighter than thos of the CRB, they still lie substantially below the errors obtained in the numerical experiment. As we suspect the ZZB to be a very tight bound, this suggests that a better parameter estimation algorithm must be found.

We are currently giving some thought to this problem [18]. The conditional mean estimator should give more accurate estimates than the ML method at these moderate SNR’s. However, as we have noted, it is very difficult to implement this estimator in practise. Building on unpublished results [13], we are studying an approximate iterative implementation of the conditional mean estimator which reduces the amount of computation by replacing a multidimensional integral over the parameter space with a sequence of one-dimensional integrals.

13.3 Searching for periodic gravitational wave sources

Gravitational wave sources that emit continuous, periodic or nearly-periodic signals in their own frame of rest are a prime target for the new inter-

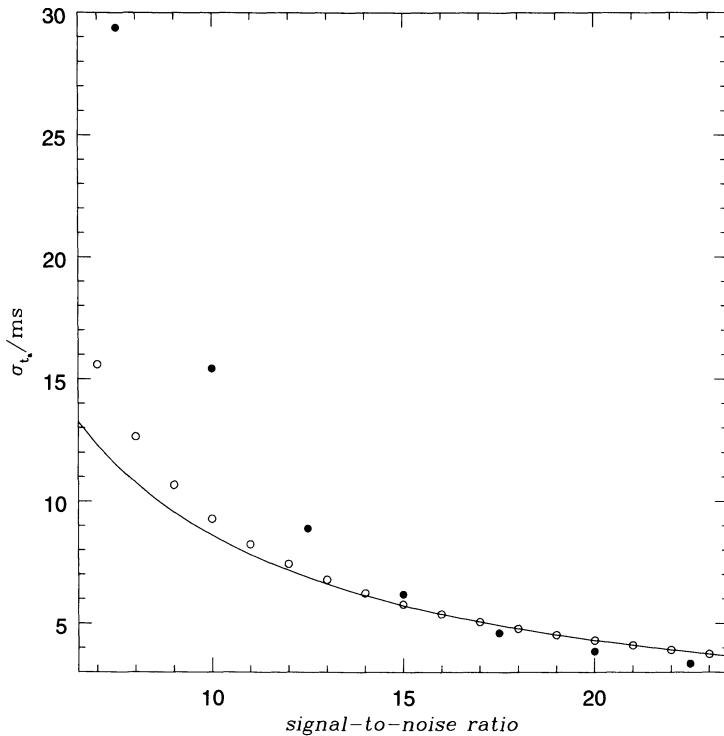


FIGURE 1. Accuracy of estimation of the time-of-arrival parameter (in milliseconds) against SNR for a coalescing binary gravitational wave signal. Results of a Monte-Carlo simulation of the ML method (filled circles) are compared with the ZZB (unfilled circles) and CRB (bold line).

ferometric gravitational wave detectors. This arises for two reasons, one astrophysical and the other practical.

The astrophysical interest in these sources comes from the fact that they will almost certainly be neutron stars, since nothing else is likely to be able to emit strong gravitational waves at frequencies above 20 Hz. Neutron stars are among astronomy's most interesting objects. As the endpoint of the evolution of many stars (those that die in supernova explosions), they tell us about stellar evolution and statistics. As places where physics is extreme (they are as dense as a uranium nucleus and more massive than the Sun, with a radius of only 10 km or so), they are laboratories for

nuclear physics in conditions not achievable in Earth-based experiments. But neutron stars are hard to observe. We see them as pulsars, which present interesting statistical problems for detection in radio waves. We see them sometimes in X-ray binaries, which have also been discussed at this meeting. But most neutron stars are invisible, no longer emitting detectable radiation. Gravitational waves present an opportunity to find a different subset of them.

The practical reason for wanting to search for them is that they are just about the only thing that a single gravitational wave detector, working in isolation, can reliably identify. They are therefore important targets for whichever detector is the first to come on-line (which may be only 3 years away), and later for whichever detectors achieve the best duty cycles. Short bursts of gravitational radiation are too likely to resemble local, poorly-modelled detector noise, so they can only be identified in coincidence searches among two or more detectors. But long wavetrains, lasting months or years, have a distinct signature that comes from the amplitude and phase modulation induced in the signal by the motion of the detector, and it seems less likely that local sources of noise will preferentially mimic such a pattern. Moreover, a detection of such a source is repeatable, both by the same detector and by others, so once it is found it can be studied again to raise the level of confidence.

Detectors are essentially omni-directional: they have a quadrupolar antenna pattern, but cannot be pointed like telescopes. Searching for sources in specific directions must therefore be done in software. We believe that, with relatively modest computing power, we shall be able to perform a targeted search of several hundred locations on the sky with data sets as long as a year or more. But the problem of searching the whole sky for a completely unknown source is many orders of magnitude harder. The central signal-analysis problem is to devise an efficient algorithm that permits us to search, say, a 1-year data set with a dedicated teraflop computer.

In the next section I will describe the nature of the data and what it is we want (intend) to do with it for the two problems, a targeted search and an all-sky all-frequency survey. In the final section I will pose the questions we face today and suggest partial answers. There will be more questions than answers, and many of them pose interesting statistics problems in their own right. The problem is discussed in more detail in [21, 22, 23].

13.3.1 Why the data-analysis problem is difficult

The interferometers under construction will be wide-band detectors that operate reliably almost all the time. Each will produce a signal data stream sampled at up to 20 kHz. Since we do not expect neutron stars to radiate continuously at frequencies much above 2 kHz, we can probably expect to work with a data stream that is filtered and resampled down to about 5 kHz. In one year the data set consists of 1.5×10^{11} 2-byte samples, or

300 GB. Interesting searches can be performed on subsets of this data: most neutron stars will be emitting radiation below 100 Hz, giving us a data set of only 15 GB, although it is not at all clear that the strongest gravitational-wave sources will be at these low frequencies.

If we were on an observing platform that had a fixed velocity relative to the stars, and therefore to any pulsar we might be looking for, then finding the signal would be just a matter of taking the Fourier transform of the data and looking for a peak at the known frequency. The signal-to-noise ratio d for an observation that lasts a time T_{obs} would increase as $T_{obs}^{1/2}$.

However, the Earth rotates on its axis and moves about the Sun and Moon, and these motions would Doppler-spread the frequency and reduce its visibility against the noise. The rotation of the Earth, after a time

$$T_{max} \approx 70 \left(\frac{f}{1\text{kHz}} \right)^{-1/2} \text{ min}, \quad (13.9)$$

smears out a pulsar signal of frequency f by more than the observational frequency resolution of $1/T_{max}$. The effect of the Earth's orbit around the Sun dominates after a day or so. The Earth's motion about the Earth-Moon barycentre also has a significant effect. Since any serious observation is likely to last days or longer, the Doppler effects of all these motions must be removed, even in searches for very low-frequency signals (10 Hz).

The Doppler corrections one has to apply depend on the location of the source in the sky. Since the spin axis of the Earth is not parallel to orbital angular momentum vectors of its motion about the Sun or Moon, there is no symmetry in the Doppler problem, and every location on the sky needs its own correction. The orbital motion of the Earth sets the most stringent limit for observation times T_{obs} greater than a couple of days:

$$\Delta\theta_{orbit} = 1 \times 10^{-6} \left(\frac{f}{1\text{kHz}} \right)^{-1} \left(\frac{T_{obs}}{10^7 \text{s}} \right)^{-2} \text{ rad}, \quad \text{for } T_{obs} < 1 \times 10^7 \text{s}. \quad (13.10)$$

This reaches a minimum of about 0.2 arcsec for a millisecond pulsar observed for 4 months.

Observations can therefore determine the position of a pulsar very accurately, but of course there is a compensating problem. Uncertainties in the position of the pulsar being searched for, orbital motion of the pulsar in a binary system, proper motion of the pulsar (*e.g.*, a transverse velocity of 150 km/s at 100 pc), or unpredicted changes in the period [anything larger than an accumulated fractional change $\Delta f/f$ of $10^{-10}(f/1\text{kHz})^{-1}$] will all require special techniques to compensate for the way they spread the frequency out over more than the frequency resolution of the observation.

If the source is in a known position, then one can apply the appropriate corrections to the data either in the time domain or the frequency domain and obtain in the end a Fourier transform of the data as if it had been taken

by a detector at rest at the solar system barycenter. This can then be used for further filtering for orbital motion, spindown, etc. Data analysis groups will be targeting certain sources: known pulsars in an accessible frequency range, X-ray binaries that may harbor emitting pulsars, and giant stars that may hide neutron stars inside their envelopes, a situation that arises in binary evolution.

To perform a survey, one needs in principle to search all positions on the sky. If this were done naively, using a separate Fourier transform to perform essentially a matched filter for each of the approximately 10^{13} locations on the sky that an observation could resolve, then we would require a computer capable of performing 10^{17} floating-point computations per second if we wanted to analyze a 1-year data set in one year. Clearly, more efficient methods are needed.

13.3.2 Problems and possible approaches

The targeted search does not seem to pose significant problems in terms of reducing data to a barycentric data set, from which further filtering can be done. But the further filtering does pose a number of problems.

1. In the case where the neutron star could be hidden inside a stellar envelope, it would be executing an orbital motion with an unknown period. As in the case of radio searches for pulsars in binary systems, this complicates the search enormously. However, unlike the radio case, here the binary will complete hundreds of orbits during a single observation, so the radio technique of searching for an acceleration parameter would not help. Instead, we need to search for the particular frequency modulation produced by this orbital motion.

What is the most efficient way of finding, with optimum signal-to-noise ratio, the signal from an orbiting neutron star that completes hundreds of orbits during an observation?

A possible solution is to apply a family of “picket-fence” filters to the power spectrum, choosing the spacing to reflect astrophysical constraints on the orbital periods.

2. Pulsar spin-down and spin-up pose significant problems. Accreting neutron stars in X-ray systems and inside stellar envelopes are likely to be spinning up at a rate which is not always known. Field neutron-stars, that might turn up in an all-sky survey, are likely to spin down. A typical pulsar may change its frequency by several milliHertz in one year, spreading its power over millions of frequency bins in a one-year Fourier transform. Second-, third-, and fourth-derivatives of the frequency may be important.

How can one find a signal from a spinning-up or spinning-down neutron star most efficiently?

Possible solutions include the acceleration-search techniques developed for pulsar radio astronomy, or fast search methods such as those discussed in [24]. Other possible methods include wavelets or fractional Fourier transforms. Nonlinear methods like adaptive filtering may also work, depending on signal-to-noise ratios.

3. Many known pulsars move across the sky fairly rapidly, up to 100 arcseconds per year. This spreads their signals out across hundreds of sky resolution elements.

How does one search for a signal that crosses many sky resolution elements, given data that have been reduced for a particular location?

If the proper motion of the target is known, then this can presumably be done rather efficiently. But if it is not known (this will be the normal case in survey mode) then one seems to face a considerable filtering problem.

The all-sky survey is a more difficult problem, as described in the previous section. It is ameliorated somewhat by realizing that we can only reliably detect signals stronger than perhaps 10σ , because with 10^{13} possible locations on the sky, there are many chances for false alarms. But even this observation raises a question, which is the first in our list for the survey problem:

1. The noise in data sets for different locations on the sky all comes from the same original data set. The statistics of this noise is important.

Given a data set of 10^{11} samples, say of pure Gaussian noise, what are the statistics of the Fourier transforms in each of the 10^{13} patches on the sky? Can they be treated simply as 10^{13} independent samples of 10^{11} Gaussian data, or are there important correlations among them?

This is essential for setting appropriate thresholds and deciding the significance of a detection.

2. Given that we are looking for relatively strong signals, it seems that some kind of hierarchical search technique would simplify the search and reduce its computational demands.

What is the best hierarchical search techniques, and what are its statistics?

There are many possible styles of search. One could break the sky up into larger patches or zones, in which strong signals might still appear at, say, 3σ . Each such candidate would then be looked at in more detail. Alternatively, one could analyze shorter stretches of data, in which the effective patch size on the sky might be much bigger, and then either add successive power spectra or use pattern-matching techniques to find candidates that repeat from one time-slice to another. With the technique of adding power spectra, we have been able to show that a 1 teraflop computer could do an all-sky survey that lasted a few months. We suspect that we can improve on this, but not enough work is being done on this problem yet.

3. Finally, a simple hierarchical search would not seem to work if one is trying to find a signal from a binary system, or any other signal that needs filtering to raise its amplitude. Until the extra filter has been applied, the signal won't stand out enough to pass through the hierarchical cutoffs.

How does one incorporate filtering for spindown, binaries, and proper motion into a hierarchical search?

13.4 Conclusions

Searching for weak gravitational wave signals presents interesting challenges. Parameter estimation and detection of modulated signals are among the most pressing. Despite the difficulty of constructing detectors that will actually accumulate data of sufficient sensitivity, it seems likely today that the actual performance of detectors for some sources, like continuous signals, will be limited by available computer resources and the efficiency of data analysis algorithms, rather than by the sensitivity of the detectors.

REFERENCES

- [1] Abramovici, A. *et al.*, *Science*, **256**, 325–333 (1992).
- [2] Bradaschia, C. *et al.*, *Nucl. Instr. Meth. Phys. A*, **289**, 518–525 (1990).
- [3] Danzmann, K., in Coccia, E., Pizzella, G., Ronga, F., eds., *Gravitational Wave Experiments*, (World Scientific, Singapore, 1995), 100–111.
- [4] Schutz, B.F., “The Detection of Gravitational Waves”, in Marck, J.-A., Lassota, J.-P., eds., *Astrophysical Sources of Gravitational Radiation*, (Springer, Paris, 1996).
- [5] Van Trees, H. L., *Detection Estimation and Modulation Theory, Part I* (John Wiley and Sons Inc., New York, 1968).
- [6] Cramér, H., *Mathematical Methods of Statistics*, (Princeton University Press, Princeton NJ, 1946).
- [7] Strong, A.W., *Astron. Astrophys.*, **150**, 273 (1985).

- [8] Tegmark, M., Taylor, A.N., and Heavens, A.F., preprint, astro-ph/9603021 (1996).
- [9] Barankin, E.W., *Ann. Math. Stat.*, **20**, 477 (1949).
- [10] Finn, L.S., *Phys. Rev. D*, **46**, 5236 (1992).
- [11] Cutler, C. and Flanagan, E.E., *Phys. Rev. D.*, **49**, 2658 (1994).
- [12] Balasubramanian, R., Sathyaprakash, B.S., and Dhurandhar, S.V., *Phys. Rev. D.*, **53**, 3033 (1996).
- [13] Pasetti, A., M.Sc Thesis, University of London (1987).
- [14] Ziv, J and Zakai, M., *IEEE Trans. Info. Theory*, **IT-15**, 386 (1969).
- [15] Weinstein, E. and Weiss, A.J., *IEEE Trans. Info. Theory*, **34**, 338 (1988).
- [16] Bell, K.L., “*Performance Bounds in Parameter Estimation with Application to Bearing Estimation*”, Ph.D. Thesis, George Mason University, Fairfax, VA. (1995).
- [17] Cinlar, E., *Introduction to Stochastic Processes*, (Prentice Hall, Englewood Cliffs, 1975).
- [18] Nicholson, D. and Vecchio, A., in preparation (1996).
- [19] Davis, M.H.A., “A Review of the Statistical Theory of Signal Detection”, in Schutz, B.F., ed., *Gravitational Wave Data Analysis*, (Kluwer, Dordrecht, 1989), p. 73–94.
- [20] B.F. Schutz, “Cosmic sources of gravitational radiation”, *Class. Quantum Grav.* **10** (1993) S135–S145.
- [21] B.F. Schutz, “Data Processing Analysis and Storage for Interferometric Antennas”, in Blair, D.G. (ed) *The Detection of Gravitational Waves*, Cambridge University Press, Cambridge United Kingdom (1991) pp. 406–452.
- [22] Schutz, B.F., “Searching for Gravitational Waves” in Vandoni, C.E., and Verkerk, C., eds., *1993 CERN School of Computing* (CERN, Geneva, 1994), 274.
- [23] Jones, G.S., PhD thesis (University of Wales Cardiff, 1996).
- [24] Bailey, D.H., and Swarztrauber, P.N., *Siam Journal on Scientific Computing*, **16**, 1239 (1995).

Discussion by Curt Cutler

General remarks

I just want to make a few general remarks about gravitational wave data analysis, followed by a short list of statistics-related questions that we are groping with in this field. For those interested in learning more about this field, a good place to start would be *The Detection of Gravitational Waves*, D. Blair, ed. (Cambridge Univ. Press, 1991), and especially B. Schutz’s article in that volume. For simplicity I confine my remarks to the largest laser-interferometer detectors that are currently under construction: the two LIGO detectors and VIRGO. In essence, each of these detectors measures one polarization of the gravitational wave; each outputs a time-series

$h(t) \equiv (\delta L_1(t) - \delta L_2(t))/L$, the fractional change in length of the first arm of the interferometer, minus the fractional change in length of the second arm. These detectors are relatively broad-band, with best sensitivity roughly in the range 10 – 1000 Hz. In this band, they will be capable of measuring $h(t)$ as small as $\sim 10^{-22}$ for short-lived, burst sources and $\sim 10^{-25}$ for long-lived, periodic sources. Despite this remarkable sensitivity, it is not really clear that there are any types of sources sufficiently energetic or numerous (so that some are close to us) that LIGO/VIRGO can be assured of seeing them in a few years of operation. However there are a number of very good possibilities: coalescing neutron star or black hole binaries, supernovae, gravitational wave pulsars (i.e., rapidly rotating, non-axisymmetric neutron stars), and stochastic gravitational waves produced in the very early universe. Any detections are expected to be at relatively low signal-to-noise: typically $S/N < 10$, even after matched filtering. Therefore, given the complexity of the detectors and the possibility of unmodelled noise posing as signal, one relies crucially on coincidence of detection (the three detectors are separated by thousands of miles) to establish confidence that gravitational waves have really been detected. Also, since the detectors are effectively “all-sky monitors” with very broad, quadrupole beam patterns, at least three detectors are necessary for determining the position of the source on the sky, by a standard time-of-flight method.

Some questions

With that as a bare-bones introduction, I list below some of the questions that have arisen as we try to prepare for the actual data analysis process, which is still a few years away. I will admit upfront that these questions are not well-posed, nor will I give sufficient detail that the reader could be expected to solve them, or even pose them properly! However, I imagine each might be the beginning of a useful conversation, or prompt a suggestion to try looking at reference X or Y. Questions:

- 1) Some expected signals are known quite accurately, and matched filtering will be used. However for others, only qualitative features are known. For the latter, one possibility of course is just to parametrize one's ignorance, and filter for a correspondingly enlarged class of signals. Is it likely that non-linear, adaptive filtering methods would be more useful? How does one establish confidence levels with such methods?
- 2) We would like to know how accurately these detectors will determine various quantities characterizing the detected sources. For instance, when two black holes collide, we'd like to know how accurately we will extract from the data stream the masses of the two black holes. We have been using the standard, Fisher-matrix approach to estimate the size of the errors, but this approach is only valid for sufficiently high signal-to-noise. There are

indications that this approach simply will not be valid for colliding black holes detected with $S/N = 10$. In this case, the likelihood function may well have several “competing” local maxima. In this case, is there a formalism for estimating errors that is more accurate than the the simple Fisher-matrix approach, but easier to implement than a full-blown Monte Carlo analysis?

3) This question is really about efficient search algorithms, not statistics, but I'll include it anyway. It seems likely that the search for gravitational wave pulsars (the rapidly rotating, deformed neutron stars mention above) will be limited primarily by computer speed. To a first approximation, these pulsars just emit a sinusoidal signal with fixed frequency. If that were really the whole story, we could find them all very efficiently just using an FFT. However the signals are expected to be weak, so integration times of order a year may well be necessary to build up sufficient signal-to-noise for detection. On that timescale, the signal is periodically Doppler-shifted by the Earth's motion, and the frequency may also change by a couple percent because the pulsar is slowing down (slowing down in a way that's not understood theoretically, but which can probably be fit with a few parameters). Are there efficient generalizations of the FFT one could use in this case, where the signals one is filtering for are not quite sine waves? If not, are there any known tricks for doing the filtering in a computationally efficient way? In a computationally-limited search, there's clearly a premium on finding/developing efficient methods.

AXAF Data Analysis Challenges

**Aneta Siemiginowska¹, Martin Elvis¹,
 Alanna Connors², Peter Freeman³,
 Vinay Kashyap³, and Eric Feigelson⁴**

ABSTRACT The high quality of the AXAF X-ray data provides new challenges for the X-ray data analysis. It is clear that an “old” approach is not enough to fully exploit the capabilities of the AXAF instruments. We describe a few of the statistical and computational problems that we have so far identified. Some of them appear to be theoretically solvable but computationally challenging, while others state problems for theoretical statistics which, so far as we know, are unsolved. The problems divide, from an astronomical point of view, into: Modeling the Data (e.g. non-linear parameter estimation, uncertainties in the model, weighting the data, correlated residuals), Source Detection (events in N-space, use of wavelets, significance of detected structures) and Instrument Related Issues (pile-up in AXAF ACIS, overlapping orders in grating spectra).

14.1 Introduction

Study of X-ray emission from stars and galaxies requires placing highly specialized telescopes and detectors on space-based satellites, because X-rays do not penetrate the Earth’s atmosphere. Following the initial discoveries of cosmic X-ray sources in the early 1960s, 28 satellite-borne X-ray missions have been launched by several nations (Bradt et al. 1992). NASA’s forthcoming “Advanced X-ray Astrophysics Facility” (AXAF) mission will provide the highest spatial and spectral resolution yet achieved in X-ray astronomy (see Zombeck 1996, 1982, 1979).

For the first time X-ray astronomy will obtain comparable resolution to that commonly available in the other regions of the spectrum. Detec-

¹AXAF Science Center, 60 Garden Street, Cambridge, MA 02138

²University of New Hampshire, Durham, NH 03824

³University of Chicago, IL 60637

⁴Department of Astronomy & Astrophysics, Penn State University, University Park PA 16802

tion of very faint point sources (\sim 10 times fainter than ROSAT) becomes possible because of the reduced background per beam, and accurate locations of sources facilitate their identification with counterparts at other wavelengths. Moreover one of the imaging detectors records a spectrum in each spatial pixel allowing spatially resolved spectroscopy. Transmission gratings provide wavelength resolution which improves linearly with the reductions in a beam size, so that high resolution spectroscopy ($E/\Delta E \sim$ several hundred), particularly at low energies, becomes feasible. However, these instrumental advances will generate new computational and conceptual challenges for X-ray data analysis. The statistical methodology traditionally used in X-ray astronomy may not prove adequate for AXAF.

Table 14.1 compares the basic characteristics of AXAF and the two X-ray astronomy satellites now operating: the German Röntgen Satellite (ROSAT) and the Japanese Advanced Satellite for Cosmology and Astrophysics (ASCA). Both carry US instrumentation. ROSAT has high spatial resolution and low spectral resolution, while the reverse applies to ASCA. AXAF will outperform these satellites in both respects although its field of view is more limited. Though the X-ray data is transmitted by the satellite to ground stations in a linear telemetry stream, each observation can be considered to be a four-dimensional multivariate database where each photon is characterized by its position in the detector (representing two-dimensional location in the sky or wavelength along a grating spectrum); its energy in units of kilo-electron Volts (keV); and its arrival time. Different analysis problems can thus be viewed as challenges in image restoration, interpretation of spectra, and time series analysis.

TABLE 14.1. Instrument characteristic

Instrument	Mirror PSF ^a [arcsec]	ΔE^b [keV]
ROSAT PSPC	5	0.4
ASCA SIS	150	0.1
AXAF ACIS	0.5	0.1
AXAF ACIS/HETG	0.5 (1-D)	0.005

^a Width of the Point Spread Function.

^b Energy resolution at 1 keV.

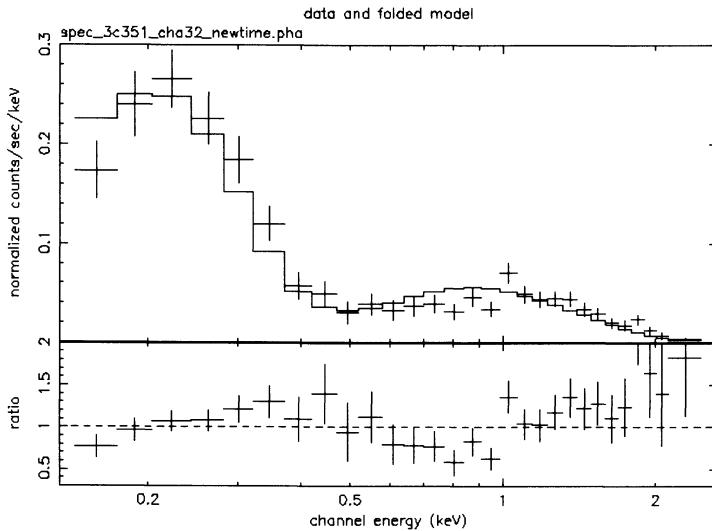


FIGURE 1. Upper panel shows the best fit power law emission model to the ROSAT PSPC quasar spectrum (3C 351). Lower panel shows the residuals. Large deviations from the model indicate more complicated structure present in the data. Thanks to Fabrizio Nicastro.

14.2 Example of X-ray data analysis problems

14.2.1 Spectral analysis of ROSAT, ASCA and AXAF data

A quasar spectrum, or plot of X-ray photon flux F_x against energy E , observed with ROSAT PSPC (Position Sensitive Proportional Counter) is shown in Figure 1.

Detection of the X-ray photons is an intrinsically Poisson process. When binned, and with high enough source counts per bin, a Gauss–Normal distribution is a reasonable approximation. Further, a “smearing” of the true energy and angular position of each photon by an instrument response function is fundamental to the X-ray measurement process. The wider the “smearing”, the lower the spatial or energy resolution. The ROSAT spectrum contains just a few independent energy channels (3–4) binned on a finer scale (32 bins) between 0.1 and 2.5 keV; this is a low resolution spectrum with resolution $\Delta E \simeq 0.5$ keV). In a simple analysis of this source, the model spectrum of a power law ($F_x \propto E^{-\alpha}$) with the galactic absorption (a complicated nonlinear function) has been assumed. The χ^2 statistic has been used to find the best-fit model parameters: spectral index α of the power law emission and a column density of the absorber. Note that

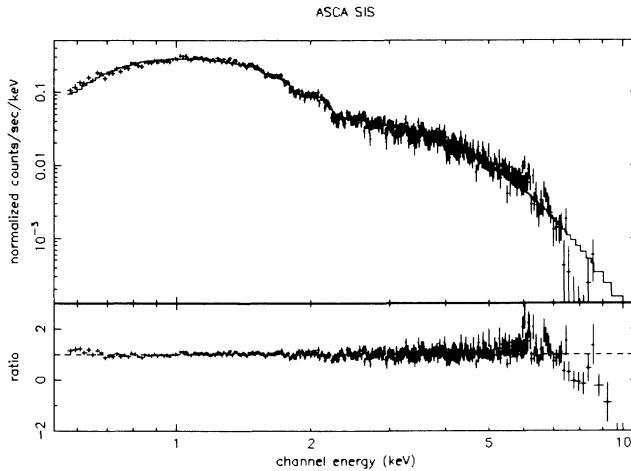


FIGURE 2. Upper panel shows the best fit power law model to the ASCA SIS spectrum of the Seyfert 1 galaxy (NGC 7469, observed in 1993). The complicated structure (emission lines and edges) above 5 keV is clearly present in this data.

the overall shape of the spectrum is not a simple power law because the nonlinear spectral reflectivity of the focussing mirrors and of the detector have been included in the model. Additional features in the spectrum are identified by comparing the model prediction to the observed data and searching through the residuals (Figure 1, lower panel). For example, absorption edges from ionized oxygen are present in this spectrum around 0.6-0.8 keV, but this low resolution spectrum does not allow us to distinguish between different ionization states (O VII and O VIII) or to decide whether both edges or only one are present.

A higher resolution quasar spectrum obtained with the ASCA CCD detectors (the "SIS", Solid-state Imaging Spectrometer) is shown in Figure 2. Compared to the ROSAT PSPC detection, the SIS spectrum contains more independent channels over a wider range of energies (0.5-10 keV) and provides higher spectral resolution ($\Delta E \sim 0.1$ keV). The data are binned into 256 energy bins. More features, emission lines and edges, can be found in such a spectrum, such as the likely iron line complex around 6-7 keV. These data were analyzed in exactly the same way as in the ROSAT PSPC example. First a power law emission model is assumed and plot of the residuals to this fit is made. Strong deviations are identified. The significance of the additional features is estimated by comparing the χ^2 values of different

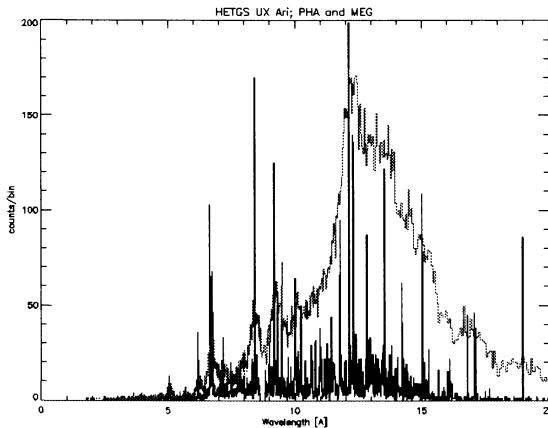


FIGURE 3. Simulated high resolution spectra of a stellar X-ray source (UX Ari) observed with the AXAF ACIS/HETG. The spectra obtained with the AXAF ACIS – dotted line, from zero order photons – and the AXAF/HETG (with medium energy grating of the HETG instrument) – solid line, are plotted on the same scale for the comparison. The assumed source model was a continuous emission measure distribution with a peak near $\log T=7.0$. Courtesy David Huenemoerder.

models (e.g. power law, power law + emission lines, blackbody emission, plasma emission). This spectrum is similar to these we will get from each pixel of ACIS, the CCD spectral imager on AXAF.

AXAF spectra at much higher resolutions can be obtained with the help of grating elements. An example of a simulated High Energy Transmission Grating (HETG) spectrum is presented in Figure 3. The energy covered with this spectrum ranges from 0.4-10 keV similar to the ASCA SIS spectrum, but with much higher spectral resolution ($\Delta E \sim 0.005$ keV) for a point source. A forest of emission lines is clearly visible in this simulation, though each may be represented by only a few photons. Using the HETG, for the first time, the quality of X-ray spectra will be comparable to those obtained in the optical band. Global statistics like χ^2 are unlikely to be effective in modeling such complex and low- signal (e.g. no longer Gauss-Normal) spectra.

14.2.2 Spatial analysis and imaging with ROSAT and ASCA

The spatial resolution of ROSAT allows us to distinguish individual point sources or extended emission regions. However, a point source always

spreads over a finite region of the detector mostly because of non-ideal optics. A “Point Spread Function” (PSF) is used to characterize the way the photons are spread around the central position of a point source. The half power diameter of the AXAF PSF will be less than 0.5 arcsec (Table 14.1). The pixel size in ACIS is 0.5 arcsec so the PSF is undersampled, although the random jitter of the spacecraft and analysis of ‘split events’ may allow sub-pixel imaging.

The high imaging resolution allows us to identify and separate sources in the image. The improved energy resolution of ACIS will simultaneously provide spatially resolved spectra of particular regions in the image, which was not possible with previous missions. Combined spectral-spatial analysis, with complicated nonlinear parametric models in both domains, will be important in studying supernova remnants, galaxies and clusters of galaxies, and in general for any extended X-ray source. The high quality of the AXAF X-ray data provides new challenges for the X-ray data analysis. It is clear that the current approaches can not fully exploit the capabilities of the AXAF instruments. Here, we describe a few of the statistical and computational problems that we have so far identified. Some of them appear to be theoretically solvable but computationally challenging, while others state problems for mathematical statistics which, so far as we know, are unsolved. From an astronomical point of view, the problems can be classified as follows: Modeling the Data, Source Detection, and Instrument Related Issues.

14.3 Modeling the data

The properties of X-ray detectors, combined with low source and background fluxes and fast read-out times, allow the position (x, y), time (t) and energy (E) of each photon to be recorded. Most traditional methods involve binning (grouping) the data so that Gaussian statistics apply and χ^2 can be calculated for each bin. But this results in loss of spatial, temporal or spectral resolution, and unbinned methods are preferable. Tests based on the empirical distribution functions (Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling) are available, but are not readily applicable to multivariate datasets. For univariate data, some astronomers use these tests repeatedly for parameter estimation (Fasano et al. 1993), but the validity of this approach has not been evaluated statistically. Bayesian methods based on the Poisson likelihood in four dimensions are possible, but are not fully developed or easy to apply. In general, we have difficulty envisioning a full data analysis system performed in the unbinned “event space”.

Another goal is to develop analysis tools that simultaneously treat spatial, temporal and/or spectral information. The complex temporal variability of some X-ray sources is discussed in the chapter by M. van der Klis in this

volume. For joint spatial-timing analysis, Giommi et al. (1995) suggest visualizing the image where the value in each pixel represents the Kolmogorov-Smirnov statistic measuring source variability. Kashyap (1996) suggests a source detection algorithm to be applied in the 3D space. More general statistical tools are needed for modeling the multi-dimensional datasets.

14.3.1 Searching a large parameter space

Complex global models have to be used to describe the new X-ray data. They are derived from astrophysical theory, can have highly nonlinear forms (e.g. with sharp discontinuities due to atomic absorption and emission lines) and include many parameters to be estimated. The main scientific goal of the investigations is to constrain these parameters with the high-quality AXAF data. When fitting complex models, such as fitting multiple plasma temperatures and elemental abundances to the grating spectrum of a star, or fitting a non-equilibrium ionization model to a spatially resolved supernova remnant, there may be dozens of model parameters. In such a complex parameter space, many ‘best fit’ solutions with similar “goodness of fit” statistics may be present. How can we efficiently search for the minima in the large parameter space? Is it possible to know when the entire parameter space has been adequately explored? Can the statistical probabilities of distant minima be evaluated? How should parameter confidence intervals be determined in cases where the goodness-of-fit statistic is unusually low or high (e.g. reduced χ^2 far from unity)? How should the confidence intervals be represented when there are many model parameters?

In some cases, a solution may be mathematically “best” or acceptable but have physically unreasonable parameters. Can “unphysicality” be included as a constraint on the fitting process in advance of obtaining solution? Perhaps physical priors can be established within a Bayesian approach. The search through a large parameter space is a serious computational challenge: effective and rapid search algorithms are needed. A Euclidean grid search is not good enough; perhaps Metropolis or Markov Chain Monte Carlo algorithms would be helpful.

14.3.2 Uncertainties in the model

The models applied during data analysis may contain some intrinsic uncertainties from the astrophysical theory. For example, many X-ray emission lines do not have fully determined atomic physics to predict their strength, or even their wavelength, and the physicist often can estimate the amplitude of these uncertainties. Modeling of the high resolution spectra would require us to include these atomic physics uncertainties with known variances in the model. Are there techniques to assign uncertainties to the predictions, knowing that each wavelength bin may contain many lines? In the Bayesian approach, uncertainties on the model can sometimes be

included directly in the priors. Is it possible to include uncertainties on the model using the frequentist (i.e. maximum likelihood) approach? In either case, how do errors propagate through the calculations when both a model and data contain uncertainties?

In cases where the count rate is sufficiently high (or one is willing to bin the data sufficiently) so that Gaussian probability distribution apply, then one might apply a modified χ^2 statistic where errors in both the data and model are used to weight the variance. A best-fit solution in a least-squares sense could then be obtained. But in general AXAF data will lie in the Poisson regime, and specialized likelihood, semiparametric or Bayesian methods must be developed for this problem.

14.3.3 Weighting data by its information content

In general a spectrum can be divided into continuum and emission/absorption lines components. In a grating spectrum, most of the counts and most of the bins will be due to continuum emission. The continuum is usually fully determined by just a few parameters (e.g. plasma temperature, density and volume).

Overwhelmingly, most of the interesting physical constraints will be made using the emission lines. Line ratios can provide information about the temperatures or ionization structures of the emitter, and line profiles can be used to study dynamics of the emitting system. However, the lines may contain only 10% of the signal. Some lines contain more information than others; for instance, the existence of some lines or a ratio of certain lines, may determine the density of the emitting gas uniquely. Are there methods for weighting the data by the astrophysical information it carries, rather than simply by its signal-to-noise? Once again, χ^2 is not an adequate statistic.

14.3.4 Correlated residuals

X-ray astronomers normally use χ^2 statistic to find a global best fit, and then examine the residuals of the fit (data-model) to find new structures or features (lines or edges) in the spectra. Often these residuals are obviously correlated (Figures 1-2). This usually indicates a localized feature (e.g. an emission line) that cannot drive χ^2 . χ^2 though is blind to the clustering of the contributors to the statistic. Nonparametric ‘run statistics’ might be used, but these do not take into account the known measurement errors for each spectral channel. Are there other statistics available that include this information?

14.4 Source detection

14.4.1 *Analysis of events in N-space*

Traditionally, X-ray astronomers find sources in their images, where all temporal and spectral information has been ignored. First the density of counts attributable to an “uninteresting” background level is evaluated, and then a window is passed across the field to locate regions where the local photon density is significantly above the background (Marshall 1992). Threshold levels for source existence are set using Poisson probabilities, the likelihood ratio test based on the Poisson distribution (Cash, W. 1979) or by Monte Carlo simulation. The window can be a simple square, or can be a filter matched to the known point spread function of the telescope (Vikhlinin et al. 1995) These procedures are reasonably successful in locating constant point sources with continuous spectra, but suffer inefficiencies for spatially diffuse structure and unusual objects that are discontinuous in spectra or time.

Methods that search for clustering of events in the 4-dimensional position-time-energy space without resorting to binning may be more sensitive than standard techniques requiring binning, and may be sensitive to different types of X-ray sources (e.g. bursts, emission-line only sources). Such methods operate directly on the event files, so the data are not manipulated and all the original information remains there during the detection process. Percolation methods such as the “friends-of-friends” algorithm (known in statistics as single linkage hierarchical clustering) are commonly used to locate galaxy clusters in studies of large-scale structure in the Universe (see Feigelson & Babu 1992). Recently percolation has been used in X-ray source detection algorithms (Ebeling et al. 1996). The main difficulty is to evaluate the statistical significance of detected structures and to reliably distinguish real physical structures from statistical noise in regions of diffuse low surface brightness.

14.4.2 *Multi-scale analysis of complex source structures*

Wavelets can be used in the source detection process as well, as described by Bijaoui; Damiani et al., Kashyap et al. (see Ch. 10, 33 and 34) in these proceedings. Binned images are correlated with wavelet functions at various scales and the resulting coefficients are compared across the scales in order to determine source parameters. Methods to extend the use of wavelets to entire fields-of-view and beyond the detection of point sources are under development (see the aforementioned contributions). Whereas methods to detect point sources are in good standing, much work is left to be done with source characterization: what is the significance of a multi-bin source (and what defines its extent and shape); and how do we combine information over multiple scales and statistically characterize the results?

Consider an ACIS image subject to a wavelet transform. Wavelet coefficients below some amplitude and/or spatial thresholds are deleted, and the image is reconstructed from the remaining coefficients. What is the statistical error of structures in the transformed image? If an extended source is present, how sensitive is its shape (size, eccentricity, etc.) to the manner of the reconstruction? And more general question how can we put the confidence limits on the source shapes?

In the Bayesian context, progress has been made recently on the development of “pixons” (Puetter 1996). The data are described by a model which is smoothed locally at the best scale for a given structure. Parts of the image with less structure (e.g. source-free regions) are sufficiently well represented at large scales, while parts of the image with more structure (e.g. many point sources in a small region) need smaller scales. As in the wavelet methods the main problems are related to characterization of detected sources.

14.5 Instrument related issues

14.5.1 *Photon pile-up in ACIS*

The time between readouts of the AXAF ACIS is 2.7 seconds. If in that time two photons land in the same pixel, the electric charge created by the two will be summed. The existence of two photons and their individual energies will be unknown. This problem is called photon ‘pile-up’. Since 50% of the photons from a point source in AXAF fall into a single pixel this will be a common occurrence for bright X-ray sources.

As Poisson statistics apply to photon arrival times (assuming a constant source), it is easy to calculate when pile-up will set in. Pile-up becomes a 10% effect at ~ 0.1 ct/s (Figure 4). It will not be clear which photons are doubly counted. The long tail of the Poisson distribution also means that triple and quadruple countings will be significant too. If the total charge exceeds that from a ~ 15 keV photon then the total charge collected will exceed the capacity of the telemetry, and the single “overscale” event is considered a background event (due to a particle not a photon), and is lost. Worse still, real events can be ‘split’ over two or more pixels, so if two counts arrive next to each other they will be considered as a single event in normal processing. This lowers the pile-up count rate limit by almost a factor 10, so it will be very common. Figure 5 shows a simulated high count rate spectrum with pile-up and the corresponding clean spectrum with no pile-up. It is clear that many soft energy photons were redistributed into the high energy band changing the slope of the spectrum significantly and smearing out the structure at high energies. Are there ways to recover the initial spectrum given an estimate of the fraction of pile-up events, and possibly given a clean spectrum involving about 10% of the events?

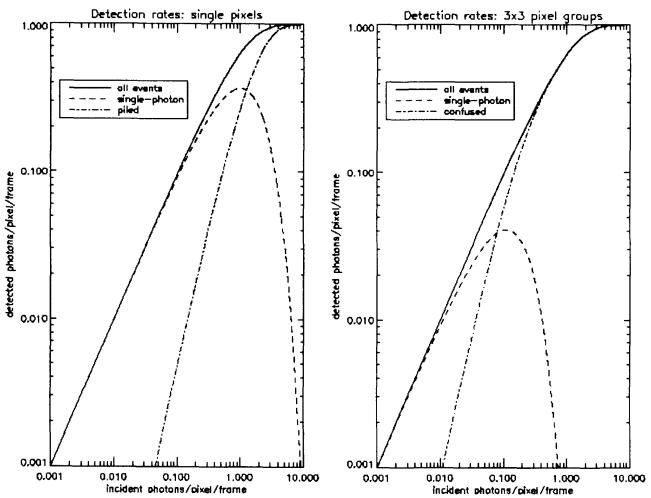


FIGURE 4. Detected event rate vs. incident photon rate for single-pixel case, left, and (more realistic) 3×3 pixel case, right. For the single-pixel case, the likelihood of multiple-photon ("piled") events becomes significant ($\sim 10\%$) at incident rates of only ~ 0.2 counts per frame. In the 3×3 pixel case, photon confusion (photons arriving in neighboring pixels during a single frame) lowers this count rate threshold by an order of magnitude. Courtesy Joel Kastner.

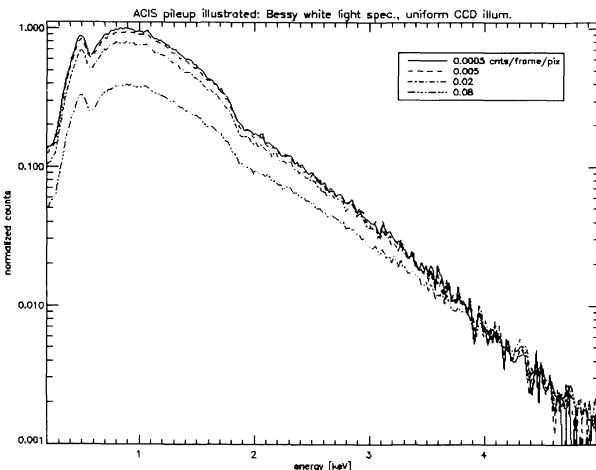


FIGURE 5. Modifications to an input continuum spectrum caused by photon pileup in the detector. Pileup results in false high-energy events, effectively flattening the spectrum. The effect grows more prominent with source strength. Based on simulated ACIS data produced by Andy Rasmussen.

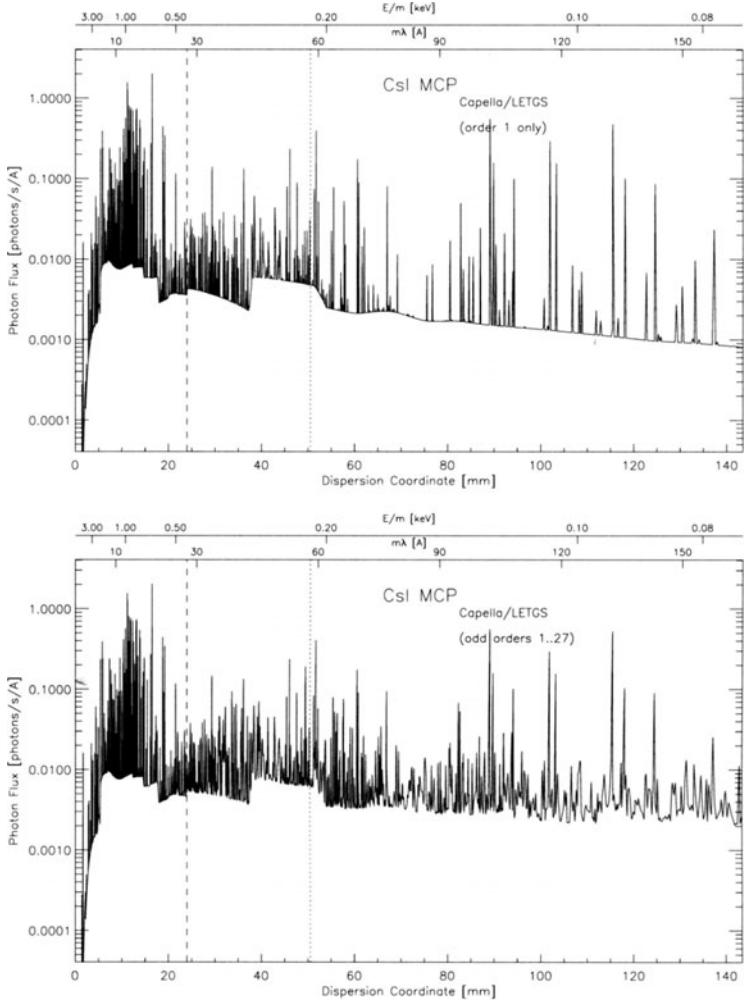


FIGURE 6. Simulated high resolution spectra of a star (Capella) observed with AXAF HRC-S/LETG. It is a grating spectrum with many diffraction orders. a) Only the first order spectrum is plotted. b) The contributions from the odd orders (1–27) are added and the resulted spectrum is plotted (from Internal ASC Memorandum by David Huenemoerder, November 1994). Courtesy David Huenemoerder.

14.5.2 Overlapping orders in low energy grating data

When a grating spectrum is projected onto the AXAF detectors, the different orders of diffraction overlap in space. With the ACIS CCD detector this is a minor problem since the orders separate quite cleanly in

pulse height space. However, when the other major instrument, the High Resolution Camera (“HRC”) is used, as it must be for low energy spectra, there is no direct way to discriminate the different orders, since the HRC has almost no inherent energy resolution (Figures 6a,6b).

An iterative deconvolution method may work for continuum points, while a pattern matching technique may be effective for lines, at least if they are not so numerous that they are heavily blended with one another. Higher orders have higher wavelength resolution, which would separate blended lines, so changing the pattern. The higher orders can dominate the counts, especially in spectral lines, so the problem is not a perturbative one, and error propagation and ‘blow-up’ is a concern. Can alternative techniques be considered?

Acknowledgments: We would like to thank Prof. Ernst Linder, Herald Ebeling, Frank Primini, Joel Kastner, David Huenemoerder, Fabrizio Nicastro for discussions. AS and ME are supported by NASA grant NAS8-39073. AC is supported by the COMPTEL project on board CGRO, which is supported in part through NASA grant NAS 5-26646, DARA grant 50 QV 90968, and the Netherlands Organization for Scientific Research (NWO). EDF is supported by NASA grants: NAGW-2120 and NAS8-38252. PF and VK acknowledge support from AXAF Sience Center.

REFERENCES

- [1] Bijaoui, A. 1996, these proceedings
- [2] Bradt et al., 1992, ARAA 30, 391
- [3] Cash, W. 1979, Ap.J. 228, 939
- [4] Damiani, F. et al. 1996, these proceedings
- [5] Ebeling et al. 1996, MNRAS, in press
- [6] Feigelson, E.D. & Babu, G.J., 1992, “Statistical Challenges in Modern Astronomy”, Springer-Verlag, New York, 1992.
- [7] Fasano et al. 1993, ApJ, 416, 546
- [8] Giommi et al. 1995, ADASS IV, Ed. Shaw, R.A., Payne, H.E. & Hayes, J.J.E. p.117
- [9] Huenemoerder, D., 1994, Internal ASC Memorandum
- [10] Kashyap, V. 1996, ADASS V Ed. Jacoby, G.H. & Barnes, J., p.25
- [11] Kashyap, V. et al. 1996, these proceedings
- [12] Marshall, H., 1992, in “Statistical Challenges in Modern Astronomy”, Ed. Feigelson, E.D., & Babu, G.J., p. 247.
- [13] Puetter 1996, in XV Intl. Workshop on Maximum Entropy and Bayesian Methods, ed. Hanson and Silver, in press
- [14] Swank, J., 1996, these proceedings
- [15] van der Klis, M., 1996, these proceedings
- [16] Vikhlinin, A., Forman, W. Jones, C. & Murray, S, 1995, Ap.J. 451, 542

- [17] Zombeck, M. 1996, "Advanced X-ray Astrophysics Facility" to be published in the Proceedings of the International School of Space Science Course on "X-ray Astronomy" L'Aquila, Italy August 29 - September 10, 1994
- [18] Zombeck, M.V., 1982, COSPAR and International Astronomical Union, Symposium on Advanced Space Instrumentation in Astronomy, 4th, Ottawa, Canada, May 20-22, 1982. Advances in Space Research, vol. 2, no. 4, p. 259-270.
- [19] Zombeck, M.V., 1979, in: Space optics: Imaging X-ray optics workshop; Proceedings of the Seminar, Huntsville, Ala., May 22-24, 1979. Bellingham, Wash., Society of Photo-Optical Instrumentation Engineers p. 50; Discussion, p. 62.

Discussion by Joseph Horowitz⁵

There are many statistical questions touched on in the paper of Siemiginowska et al. (referred to briefly as [S+]), most of which are either implicit or very general. Thus, some of my comments will also be implicit or very general, consisting, in some cases, only of pointing out some relevant references.

Chi-square

The notion of χ^2 comes up often in the astronomy literature, and in [S+] as well. Now χ^2 has several possible meanings to statisticians, and Eric Feigelson was kind enough to clue me in to what astronomers mean by " χ^2 ". (I am not the only one who has been confused on this point; see [B], p.322.)

According to him, the "bible" of statistics for astronomers is ... *Numerical Recipes* [NR]! In ch.15 of [NR], χ^2 refers to the (scaled) error sum of squares in nonlinear regression, which, under certain conditions, has a χ^2 -distribution with the appropriate degrees of freedom.

In more familiar terms (for statisticians), the *Pearson χ^2 -statistic*, designed to test hypotheses about Poisson (or multinomial) observations,

$$\chi_P^2 = \sum_{i=1}^k (N_i - \lambda_i(\theta))^2 / \lambda_i(\theta),$$

is often used in astronomy.

Here N_1, \dots, N_k are independent Poisson counts, with means $\lambda_i(\theta)$, typically representing photon counts in k distinct energy or spatial bins, and

⁵Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003

the parameter θ is some physical characteristic of, for example, an x-ray source.

When the means are large, the N_i have approximately normal distributions, but then the variances are necessarily equal to the means (also pointed out in [B]). Then the nonlinear regression material in [NR] becomes relevant, but the constraint on the variances is often ignored.

The usual strategy is to fit a model by finding the value $\hat{\theta}$ that minimizes χ_P^2 , and to search the residuals for further structure. The χ_P^2 -statistic has a good intuitive motivation, and its asymptotic distribution is known. The minimizer $\hat{\theta}$ is asymptotically equivalent, and in some cases identical to, the maximum likelihood (ML) estimator of θ . For details, see [C], chs. 30, 33.

Another version of χ^2 is the *likelihood ratio - χ^2* ,

$$\chi_{LR}^2 = 2 \sum_{i=1}^k (N_i \log(N_i/\lambda_i(\theta)) - (N_i - \lambda_i(\theta))),$$

which is asymptotically equivalent to χ_P^2 and has the same limiting distribution. But χ_{LR}^2 is the correct likelihood ratio statistic for testing the “null” hypothesis, H_0 , that the Poisson means are given by the model $\lambda_1(\theta), \dots, \lambda_k(\theta)$, for some θ , against the alternative, that the means are not of that form. Tests of H_0 against specific alternative models allow rigorous assessment of the reality of features that do not conform to the H_0 -model.

In this connection, χ_{LR}^2 has good decomposition properties, not shared by χ_P^2 , for certain sequences of nested models of successively greater complexity (see [MN]).

The “Bible”

Although it is a great read, [NR] is no more suitable as a statistical bible than Ptolemy is for astronomy. A glance at the main statistical references in [NR], ch.15, confirms this: Bevington 1969, von Mises 1964, Brownlee 1965, Martin 1971, plus two 1976 papers in *Ap. J.*. It is as though nothing had happened in statistics over the last 25 years or so. For more recent, though not necessarily astronomer- (or statistician-) friendly, expositions of nonlinear regression, see [BW], [SW], [G1]. Some recent *linear* regression books, for instance, [M], [R], also contain some nonlinear material.

Searching the parameter space

The optimization problems hinted at in [S+], §3.1, are standard, difficult ones in numerical analysis. Some statistical theory that could be applied

to grid searches for ML and similar estimators is available (e.g., [WS] and references therein) to the intrepid. The added feature is that accuracy of the grid estimator *with high probability* can sometimes be asserted.

Recently, effective Markov Chain Monte Carlo (MCMC) calculations have been developed for ML and other function optimization problems; see [G2] and [B+].

The question of “unphysicality” ([S+], §3.1) is of course not a statistical one. If it can be expressed in a reasonable mathematical form, physicality can, in principle, be added as a constraint to the model.

Which residuals?

Residuals are mentioned several times in [S+] (§§2.1, 3.4). In dealing with low counts, where the Poisson distribution must be respected, there are various types of residuals specifically tailored for Poisson data, viz., Anscombe-Cox-Snell and deviance residuals, all discussed in [MN].

To bin or not to bin

Many of the phenomena discussed in [S+] can be modeled directly, with no binning.

Let B be the energy \times time “box”, $E_1 \leq E \leq E_2$, $t_1 \leq t \leq t_2$. Astrophysical models for the photon count $N(B)$ in B often specify that, as a random variable, $N(B)$ have a Poisson distribution with mean $(t_2 - t_1) \int_{E_1}^{E_2} g(E) dE$, and that, for disjoint energy \times time regions, the counts be statistically independent. The physics is contained in the function $g(E)$. This type of model is a *space-time Poisson point process*, although “energy-time” would be a better term for this example.

If a photon of energy E is detected in the energy interval $E' \pm dE'$ with probability $k(E, E')dE'$, where $k(E, E')$ models the detector, then (theorem) $N'(B)$, the number of detections in the box B , also follows a Poisson point process model with g -function $g'(E') = \int k(E, E') g(E) dE$. The observations are the counts $N'(B)$, for all boxes B , which is equivalent to the full, unbinned data set. The inference problem is to find out information about the *original* $g(E)$.

There is an elaborate statistical literature on modeling and inference for Poisson point processes, some of which is cited elsewhere in these proceedings. For astronomers, a good source for this point is [SM].

Conclusion

Reading between the lines of [S+], many of the statistical questions sound really fascinating, but it is usually not possible to say whether there are statistical techniques for this or that purpose without knowing the details of the problem. Choosing a statistical technique is not like choosing a pair of shoes off the shelf, especially for such complex phenomena as those discussed in [S+]. Rather than “statistician as shoe clerk”, a more appropriate metaphor might be “statistician as psychotherapist”. Serious collaboration between astronomers and statisticians requires lots of conversation, and should start early in the project. Recent technological advances are generating fundamental new statistical and scientific challenges that would be best met by such collaborative efforts.

REFERENCES

- [BW] D.M. Bates and D.G. Watts, *Nonlinear Regression and its Applications*. Wiley 1988.
- [B+] J. Besag, P. Green, D. Higdon, and K. Mengersen, Bayesian computation and stochastic systems. *Stat. Sci.*, 10, 1995, 3-66.
- [B] P. Bickel, in *Statistical Challenges in Modern Astronomy*, E. Feigelson and G. Babu (eds.). Springer 1992.
- [C] H. Cramèr, *Mathematical Methods of Statistics*. Princeton Univ. Press 1946.
- [G1] A.R. Gallant, *Nonlinear Statistical Models*. Wiley 1987.
- [G2] C. Geyer, On the convergence of Monte Carlo maximum likelihood calculations. *J. Roy. Statist. Soc. Ser. B* 56, 1994, 261-274.
- [MN] P. McCullagh and J.A. Nelder, *Generalized Linear Models* (2nd ed.). Chapman and Hall 1989.
- [M] R.H. Meyers, *Classical and Modern Regression with Applications* (2nd ed.). PWS-Kent 1990.
- [NR] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes* (2nd ed.). Cambr. Univ. Press 1992.
- [R] J.O. Rawlings, *Applied Regression Analysis: A Research Tool*. Wadsworth 1988.
- [SW] G.A.F. Seber and C.J. Wild, *Nonlinear Regression*. Wiley 1989.
- [SM] D.L. Snyder and M.I. Miller, *Random Point Processes in Space and Time* (2nd ed.). Springer 1991.
- [WS] X. Shen and W.H. Wong, Convergence rate of sieve estimates. *Ann. Stat.* 22, 1994, 580-615.

Statistical Aspects of the Hipparcos Photometric Data

F. van Leeuwen, D. W. Evans and
M. B. van Leeuwen-Toczek

ABSTRACT A brief outline of the ESA Hipparcos mission is presented, with special emphasis on photometric aspects. Limitations to interpretation of the data due to constraints on the reductions and as imposed by the scanning law are described. The limited value of period searching algorithms for a data-set with many large gaps is shown and discussed in the light of planned future missions of similar kind.

15.1 Introduction

In August 1989 the European Space Agency launched the first and so far only scientific satellite dedicated to the determination of positions, proper motions and parallaxes of a sample of 118 000 stars on the sky. The mission, called Hipparcos after the Ancient Greek astronomer Hipparchus, had two main aims that could not be achieved from the ground: to determine accurate distances for a large number of stars, and establish a reliable (milli-arcsec, mas) accuracy optical reference frame, linked with the VLBI (radio) reference frame. The radio reference frame has an accuracy at the level of 1 mas, and is directly linked to extra-galactic sources, many of which are active at radio wavelengths. Through this link, an inertial optical wavelength reference frame can be obtained, which is essential for studies of the dynamics of our own galaxy. Descriptions of the Hipparcos mission can be found in [P*92], [P*95] and references therein.

Distances of stars are difficult to measure, and the only method that is fully independent of any astrophysical models is by means of the parallax: the reflection of the Earth's orbit as observed in the variation of the apparent position of a star on the sky. The largest parallaxes are of the order of a few tenths of an arcsecond. In order to get distances for a representative sample of stars (containing not only common, fairly faint stars like the Sun, but also much rarer and intrinsically brighter stars), parallaxes need to be measured to accuracies of 1 mas or better. A displacement of 1 mas is equivalent to the size of a child standing on the Moon, as seen from

the Earth. The Hipparcos mission reached this limit for many stars. Over a 3.5 year mission it collected a sufficient amount of data to present in April 1997 a catalogue with astrometric data of unprecedented accuracy: parallaxes with accuracies better than 10% will soon available for some 20 000 stars, compared to no more than about a hundred with similar accuracy as obtained from the ground. It should be realized at this point, that contrary to any other scientific mission, the Hipparcos measurements could only be fully reduced and interpreted after all data had been collected. This is the reason why all Hipparcos results only become public when all the reductions and checks have been completed.

The Hipparcos satellite was a scanning satellite, spinning around one axis, and describing with that axis a cone around the direction to the Sun. The two apertures of the instrument described great circles on the sky. Astrometric data were obtained from images passing through the focal plane over a modulating grid, behind which there was an image dissector tube (IDT) photon counting device. This IDT can be directed such that only a $30''$ diameter area of the 0.9 by 0.9 degrees field is observed at any one time.

A fundamental part of the Hipparcos instrument is the beam combiner: it allows two parts of the sky, separated by 58 degrees and describing the same great circle, to be projected on the same focal plane. This makes it possible to measure very accurately large angles on the sky, through measuring small displacements in the focal plane and calibrating the basic angle through closing the circle. This mechanism allows, for the first time ever, the measurement of absolute parallaxes, and is based on an idea first developed by P. Lacroute in the 1960s. All ground based measurements can only produce relative parallaxes, with precision at the level of 8 mas (except for a small number of stars measured in recent years at Yale Observatory, where 1 to 2 mas precisions have been obtained). Actual accuracies are often much lower.

During the development stages of the mission, it was realized that the photon counts used in the astrometry could also be used as a measurement of intensity. The modulated signal obtained behind the grid is folded with the modulation period (corrected for scan velocity variations and projection effects). The modulated signal then consists of a first and second harmonic superimposed on a zero level. The phase of this signal provides the (one-dimensional) astrometric measurement. The zero-level and the amplitude of the first harmonic are measurements of the intensity of the object. In addition, the relation between first and second harmonic, as well as between first harmonic and zero level, can indicate multiplicity or extended objects such as planetary nebula; in both cases the modulation of the signal is decreased.

The IDT assembly consisted of a large number of small glass lenses, which in space were subject to chemical changes as a result of radiation, an effect usually referred to as darkening. Over the 3.5 year mission, the total trans-

mission of these lenses decreased by 20 to 50% depending on wavelength, resulting in a slight shift of the photometric passband. The radiation effects were worse than originally foreseen, due to an orbit problem: during all of its mission the satellite has been stuck in its geostationary transfer orbit, crossing the Van Allen radiation belts every 5 hours. As an unexpected side effect, this produced also a monitoring of the radiation belts over that period, which contained two large outbursts on the Sun [D*94].

Next to the main detector was a star mapper detector assembly, equipped with a prism splitting the light into a B and a V channel. The primary aim of this detector was the monitoring and reconstruction of the pointing of the satellite. The star mapper detector consisted of two groups of four slits, at uneven distances, allowing the recognition of a signal. One slit-group was situated perpendicular to the scan direction, and used to determine the instantaneous scan phase, the other slit-group was situated at an angle of 45 degrees with the scan-direction, allowing reconstruction of the pointing direction of the spin-axis. The continuous data stream from the star mapper detectors was also examined, in a project called Tycho (after the Danish astronomer Tycho Brahe). Further information on this aspect can be found in [H*95].

The above defines a rather hostile and far from optimal, environment for collecting photometric data. So, what were the reasons to spend a lot of time and energy on handling these data? First and most importantly, the photometric data is complementary to the astrometric data. Colour information obtained with the star mapper detectors was crucial in the astrometric reductions in all cases where no ground-based colours were available. During the reduction processes, updated photometric data from the mission has been used in the astrometric reductions. Secondly, there is no ground-based telescope that can observe both hemispheres and is unaffected by seasonal changes in observing conditions. The Hipparcos photometry will be able to provide a global photometric system for calibrating ground-based photometric systems. Thirdly, and that is where all statistical problems start, the data could be used for a systematic search for variability among the 118000 stars observed. The problems encountered here can be split in two groups: problems associated with reducing the photon-counts to magnitudes, and those associated with the coverage in time of the observations of individual stars.

All reductions of Hipparcos data have been carried out in parallel and semi-independently by two reduction consortia: FAST and NDAC. The FAST consortium was led by J. Kovalevsky and contained groups from France, Italy, Germany and the Netherlands. In the NDAC consortium, led by L. Lindegren, were groups from Lund Observatory (Sweden), Copenhagen University Observatory (Denmark) and the Royal Greenwich Observatory (UK). Comparisons of results and methods have driven both groups to obtain the best possible results from the data, as became clear when final results were compared and merged.

15.2 Photon-count reductions

15.2.1 Reconstruction of the modulated signal

The Hipparcos mission photometry is obtained from the IDT photon-counts collected behind a modulating grid. The photon-count reductions had been designed and optimized for the astrometric data. All the data were collected over 2.1 second frames, during which time between 1 and 10 stars could be observed. An observing strategy programme on-board the satellite decided which stars were to be observed. The implication is that, in the basic measurement, the amount of time spent on a star can vary by a factor 20: from just over 0.1 to the full 2.1 seconds. In general, this allowed spending longer time on faint stars and less time on bright stars when possible. Thus, measurements of the same star can be of very different accuracy. The signal obtained in a 2.1 second frame formed the basis for further reductions.

The modulation period of the signal was not constant: it varied with scan velocity and, due to projection effects, with position on the grid. The scan velocity effects were modelled using the results of the reconstructed attitude: a process that through the use of Star Mapper transits of selected stars was able to reconstruct the pointings of the satellite axes as a function of time to an accuracy of around $0.05''$. This was part of the work done by the reduction consortia, so minor changes in these values do occur as a result of different models and techniques used. After correcting for scan velocity and grid distortion effects, relative phases could be assigned to all collected IDT samples.

Using the relative phases, the data were binned: the measured intensities were collected in 12 bins by one consortium, and 64 bins by the other. When collecting data in a small number of bins, distortions due to the spread of phases of the data collected in the bins have to be accounted for. Thus, the 12 bin model used also first and second order phase corrections, the 64 bin model did not use phase corrections. The model for the mean count collected in a bin, using a first and second harmonic signal with phase corrections looks like:

$$\begin{aligned} \sum N_{k|i}/n_i = & b_1 + b_2 \frac{1}{n_i} \left(\cos p_i \sum (1 - \frac{1}{2} \delta p_{k|i}^2) - \sin p_i \sum \delta p_{k|i} \right) \\ & + b_3 \frac{1}{n_i} \left(\sin p_i \sum (1 - \frac{1}{2} \delta p_{k|i}^2) + \cos p_i \sum \delta p_{k|i} \right) \\ & + b_4 \frac{1}{n_i} \left(\cos 2p_i \sum (1 - 2\delta p_{k|i}^2) - \sin 2p_i 2 \sum \delta p_{k|i} \right) \\ & + b_5 \frac{1}{n_i} \left(\sin 2p_i \sum (1 - 2\delta p_{k|i}^2) + \cos 2p_i 2 \sum \delta p_{k|i} \right), \end{aligned} \quad (15.1)$$

where p_i represents the phase at the centre of bin i , $p_{k|i}$ the relative phase assigned to a sample k falling in bin i , $\delta p_{k|i} = p_{k|i} - p_i$, $N_{k|i}$ the photon count of sample k and n_i the total number of contributors to bin i .

The parameters to be estimated are b_1 to b_5 . Without phase corrections equation (15.1) reduces to:

$$\sum N_{k|i}/n_i = b_1 + b_2 \cos p_i + b_3 \sin p_i + b_4 \cos 2p_i + b_5 \sin 2p_i. \quad (15.2)$$

From the estimated values of b_1 to b_5 were derived the mean intensity, modulation amplitudes, and phases with their estimated errors. The estimated errors and correlations between these parameters were obtained from a Jacobian transformation of the information matrix resulting from the least squares solution of equations (15.1) or (15.2). The modulation amplitude, being derived from $\sqrt{(b_2^2 + b_3^2)}$, had to be statistically corrected for the squared-up (estimated) errors to remove a well-known bias occurring when errors become comparable to amplitudes. These corrections amounted to a few percent for the faintest stars. The final signal can be expressed as:

$$E = I_b + I_s(1 + M_1 \cos g_1 + M_2 \cos(2g_1 + 2g_2)), \quad (15.3)$$

where the dc-component is given by $I_b + I_s$ and the ac-component by $I_s M_1$ or as a weighted combination of M_1 and M_2 .

15.2.2 Photometric calibrations

Two independent photometric signals were available at this stage: the zero-level or dc-component, and the modulation amplitude or ac-component. The first is the stronger and therefore more precise signal, but is affected by background contributions, especially for faint stars. It could also become affected by accidental superposition of an image from the other field of view. The ac-signal is weaker, not affected by background, but sometimes affected by other stars within the sensitivity area of the detector. These other stars can come from the same field of view (double stars) or from the other field of view (accidental super positions). The latter cases were investigated using an all-sky catalogue down to magnitude 11.5, derived from preliminary results of the Tycho mission mentioned above. The effect of the latter cases is an incidental increase in the dc-intensity, and a decrease in the ac-modulation. Double stars produce a constant dc-intensity and a variable ac-modulation.

The observed intensities were affected by the position on the grid while being measured, and the colour of the star. The removal of these effects was the task of the photometric reductions. Different calibration models, but the same set of standard stars, were used by the two reduction consortia. This applied in particular to the modelling of the background. In addition, the evolution of the passband over the mission had to be reconstructed. This work was done by M. Grenon using mission data supplemented with measurements of very red stars made specially for this purpose. The final corrections for the passband changes defined the Hipparcos H_p band

as the reconstructed passband for January 1, 1992. All other observations were reduced to this passband.

In order to optimize the use of the dc-component, extensive modelling of the background was introduced. Being an astrometric rather than a photometric mission, no instantaneous measurements of the background were available for the IDT detector. The background in the star mapper detectors, however, was possible to be measured, but the star mapper detectors reacted differently to the radiation environment. However, in every misfortune there can be some advantage: on a few occasions the combined efforts of ground-control and satellite were unable to establish the pointing of the satellite over a 10 hour period, and it is by using these cases (when measurements were made, but did not reach the stars intended), that we were able to check the background modelling. During these periods most of the measurements showed only background.

The reduction of measured intensities to magnitudes was carried out in a pseudo-magnitude scale. This avoids a very large range of values to enter a least squares solution. In addition, most of the influences on the intensities act as a factor (a certain percentage of light is blocked), thus when transformed to magnitudes they become off-sets. However, doing the reduction in magnitudes also introduces a bias. Using a Monte-Carlo simulation, differences between the weighted mean of the magnitudes (used in the least squares solutions) and the mean of counts (which would have produced the proper values) were determined and applied as off-set in the reductions. After calibration (involving amongst others position in the field of view and star-colour) magnitude estimates were obtained. Errors were assigned to these magnitudes according to the estimated errors on the intensities as obtained from the modulation fit (equation (15.1)).

In order to stabilize the less well-defined coefficients in the solution, NDAC implemented a “running solution”. The part of the solution that changed only slowly was kept, and entered as additional (down weighted) observations for the next solution. Using a Householder transformations based algorithm for the least squares solutions. this was simple to implement (see e.g. [GJB77]).

It took a star nominally 20 seconds to cross the field of view, giving up to 10 frames of 2.1 seconds with a possible observation of this star. Depending on its requirement for observing time, a star was observed during some or all of the 10 frames. In general, faint stars managed usually only 5 or 6 observations, while bright stars often reached 10. The group of observations associated with the crossing is called the field transit, and these are the data finally used in variable star research. The mean magnitude for a field transit is obtained by first converting the measured frame-transit magnitudes into a pseudo- intensity scale, then averaging, and converting back to magnitudes. The error on the field transit is estimated from the weighted residuals of the frame-transit data. As a result of the low number of frame transits used, the estimated errors on the field transits are subject

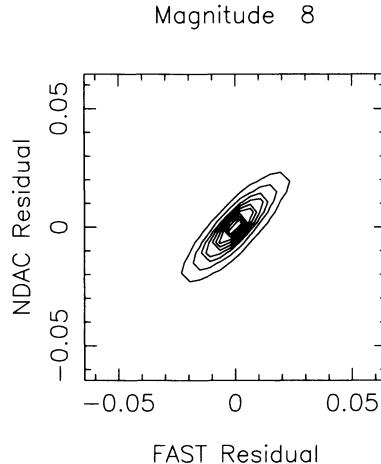


FIGURE 1. A contour plot for over 300 000 residuals for stars between 7.5 and 8.5 magn

to Student-t distributions. Collective corrections for estimated errors have been applied afterwards, and were obtained from comparisons between observed spread in the data within small ranges of estimated errors, using data from constant (standard) stars only.

Final corrections to the magnitudes were made at the end of the mission, when it was known how the passband of the system had developed. A reference system was defined for a date near the middle of the mission, to which all photometric data were finally reduced.

15.2.3 Comparison and merging of the reduced data

The last stage of the photometric reductions concerned the merging of the data from the two reduction consortia. Comparisons of methods and results had taken place all through the mission, but without compromising the independence of the reduction methods. The correlations between the two sets of reduction results as shown in the final data, made it very clear that for all but the very brightest and the very faintest stars, the reductions had been very successful. Figure 1 shows an example of the correlation between estimates of magnitudes as obtained by the two reduction processes, relative to median values of their estimates, for constant stars of 8th magnitude: the high correlation coefficient of 0.83 shows that in the final reduction results of both consortia the main noise contributor is the original photon noise, which of course was the same for both reduction processes. Figure 2 shows the correlation coefficients as a function of magnitude.

Reduced correlation coefficients for the faintest stars were the result of background modelling problems, while for the brightest stars this was due

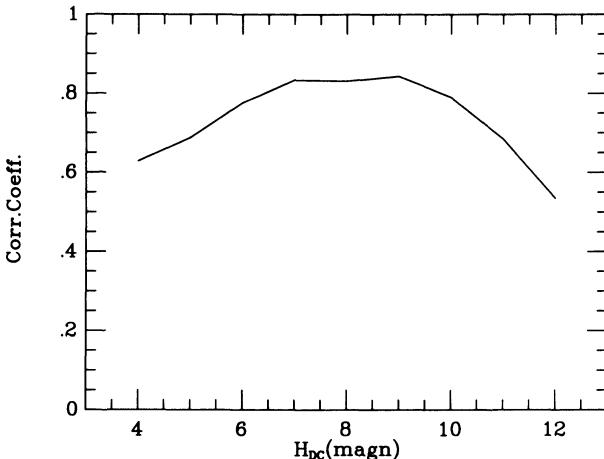


FIGURE 2. The correlation coefficients between the FAST and NDAC photometric results as a function of magnitude

to small problems in the modelling of the detector response over the grid and as a function of star colour. Insufficient very bright stars were available for calibration processes to calibrate and remove the minor influences on the brightest stars.

15.3 The scanning-law effects

The Hipparcos mission was ruled by a scanning law: a predetermined systematic coverage of the sky. The scanning law was optimized for astrometric purposes, in particular for obtaining parallaxes. It should therefore not come as a surprise that when used for photometric data in the search for periodic variability, the resulting window function is far from ideal. The effect of the scanning law, which is strongly linked to the ecliptic coordinate system as it follows the apparent mean position of the Sun, is a function of ecliptic latitude. Figure 3 shows the distribution of lengths of stretches of observations and of lengths of gaps between observations, at different latitudes. Figure 4 shows two examples of the resulting window functions.

When expressed in terms of duty cycle (as used by [WW95]), rather low values (0.01 to 0.001) are found for most stars. Similarly, the window functions were often of such low quality that only very limited automated processing in the search for periodic variability could be done (see e.g. [TD75]). A look at a typical power spectrum shows in addition that the sensitivity to detecting real periods in the range of a few days to 100 days is very low. The best detection possibility for periods is in the range between

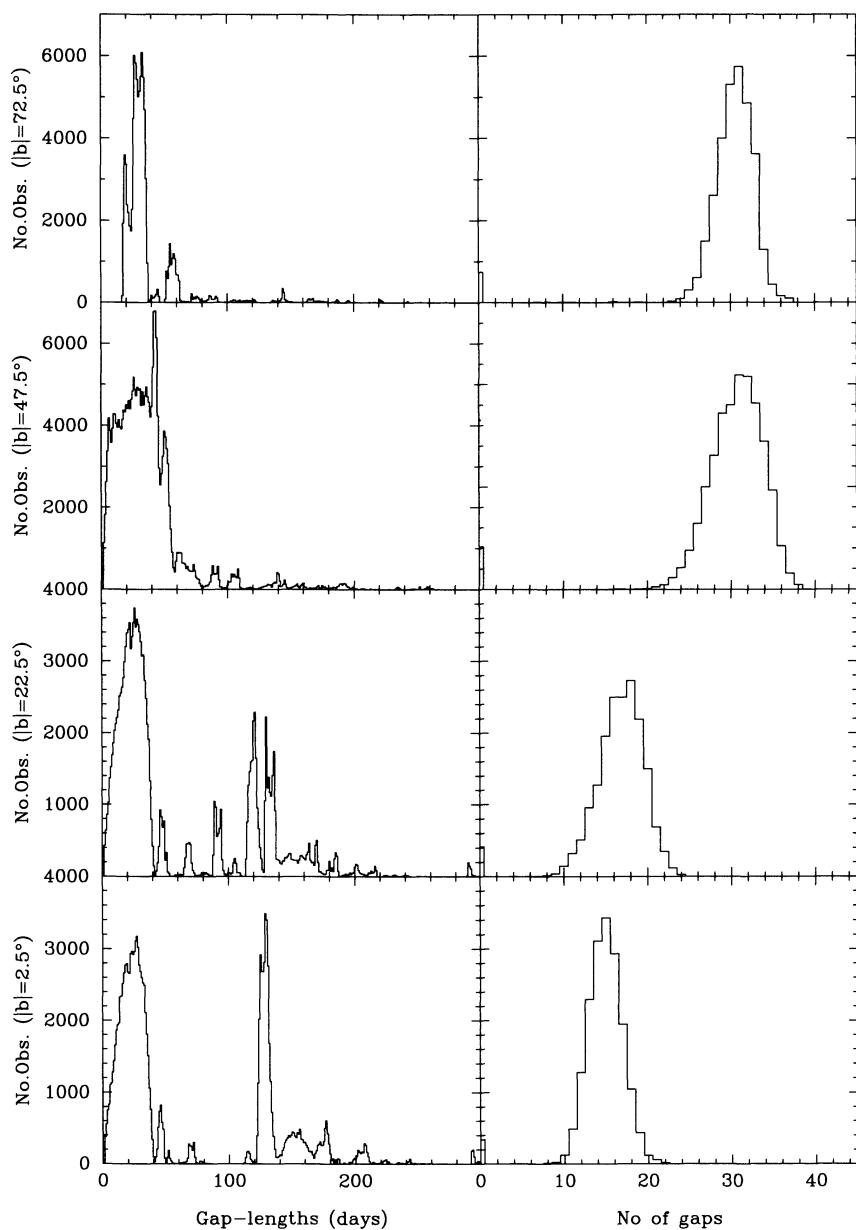


FIGURE 3. Distributions of lengths of gaps (left) and number of gaps (right) at 4 different ecliptic latitudes

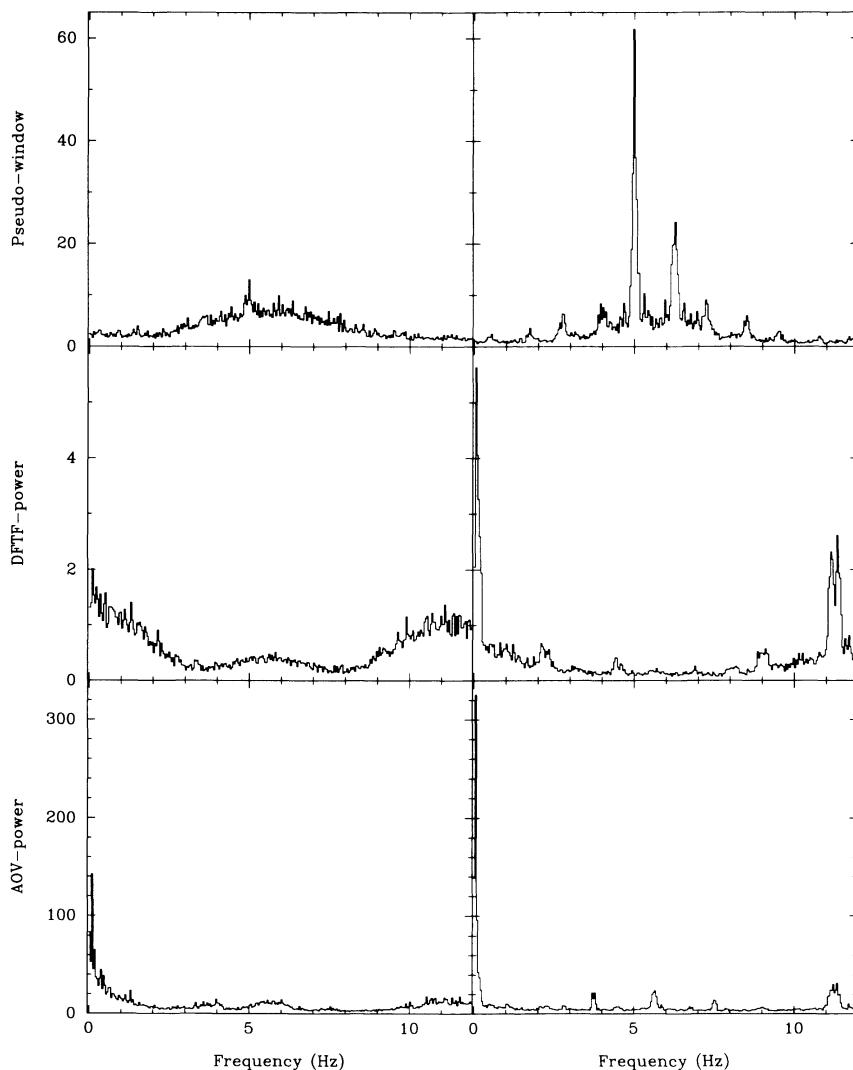


FIGURE 4. Two examples of power spectra and pseudo window functions. Instead of the full power spectra and window functions, we show here the maxima in intervals of 200 points only, to retain visibility of the main features. On the left is shown a typical case for low numbers of observations (period 11.1d), on the right a typical case for relatively high numbers of observations (period 12.34d)

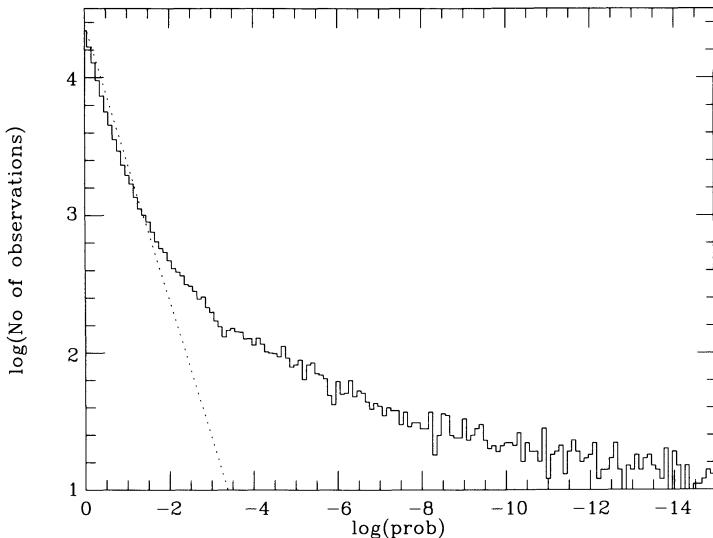


FIGURE 5. The observed distribution of χ^2 values (full line) compared with the distribution for all stars being constant (dotted line)

0.09 and 1.5 to 2 days, closely related to the average length of a stretch of data.

The processing for the variability detection was done in three steps at the Royal Greenwich Observatory, and in parallel along similar lines at Geneva Observatory (M. Grenon). Results from the two processes were compared before being accepted or rejected for inclusion in the final catalogue. The first step concerned detection of possible variability. A reference distribution of unit weight residuals for probable non-variable stars was made, with respect to which the data of each star was compared, resulting in a χ^2 value indicating the probability for a star not to be variable. When comparing these probabilities with the number of stars found within a probability interval, one should expect a fraction of non-variable stars plus a contribution from variables. Figure 5 shows that this is not entirely like the results obtained.

It appears that various corrections applied to estimated errors (which enter as weights in the accumulation of the histogram of weighted residuals) may have somewhat distorted this diagram and diminished our possibility to reliably distinguish small amplitude (micro) variability. Short of redesigning the first steps of the data reductions in order to calculate and carry along more reliable error estimates, there is very little that could be done about this.

The larger amplitude variables can be easily detected, although we do encounter various effects mimicking variability: drifts in the data due to the combined effect of the darkening of the optics and the application of the

wrong colour in the reductions; unrecognized external influences (primarily super-position of bright parts of the Milky Way in the other field of view); and duplicity. The sample of stars with sufficiently high probability to be real variables was then subjected to two period search mechanisms: the Discrete Fourier Analysis [JSc82], [JSc89] and the Analysis of Variance [ASC89]. These methods are briefly explained in the Appendix. The latter method had to be adapted to real life before it could be used: not only was the frequency searched, but at each frequency the sampling bins were examined at various offsets. Also runs with different numbers (4 to 6) of sampling bins had to be used to improve detections. Agreement between the results of the two methods was used as an indication of reliability of a detection. Special methods were used in order to resolve eclipsing binaries, in particular those of the Algol type, which defy, with their narrow minima, most automated period searches. Other binaries, with wider minima, were often detected at half the actual period. Here, the analysis of variance method was very helpful: by defining a sufficiently high number of bins (usually 8 to 10) and searching within a limited period range, most periods could still be found.

The length of the mission determined the minimum frequency step of $0.0002d^{-1}$, used in the searches. This is equivalent to the Nyquist frequency for evenly sampled data. The resulting power spectrum is severely affected by the distribution of gaps and the generally very low duty cycle. Due to the often very poor power spectrum with severe leakage, exact interpretation of peak height was mostly not possible. This resulted in detection of spurious periods for many semi-regularly pulsating M-giant stars. Problems with aliasing occurred for the very short period stars, due to the short time-scale observations rhythm: the rotation period of the satellite was 128 minutes and the time between observations of the preceding and the following field of view 20 minutes. The pattern was thus: observation, 20 min. gap, observation, 108 min. gap, observation, 20 min. gap, etc., over a period of 6 to 8 hours, sometimes restarted 2 to four hours later.

Most of the periodic variable stars observed with the Hipparcos mission were already more or less well-studied objects. It thus made sense not to start the investigation for these objects from scratch but to make use instead of what was available in the literature already, and to use the Hipparcos data only for period refinements and phase determinations. For this purpose a database of references to variability or associated observations and discussions was collected at the Royal Greenwich Observatory over the past 2.5 years. A total of over 4300 references have helped us establishing improved periods for many low amplitude stars, and prevented us from claiming new discoveries for objects that were presented in the literature over the past 10 years.

Table 15.1 shows a summary of types of variability detected within the sample of Hipparcos stars. Most remarkable is the number of eclipsing binaries (the first three types in Table 15.1), which nearly doubled from

Type	old	new
Algol	253	158
β Lyrae	82	247
W UMa	78	42
δ Cephei	269	4
RR Lyrae	169	18
β Cephei	28	31
δ Scuti	55	51
Semi-regular	88	131

TABLE 15.1. Numbers of some known and newly discovered periodic and semi-period variables in the Hipparcos catalogue

447 cases among the Hipparcos stars known from the literature to a total of 860 known among these stars now.

15.4 Fine-tuning of periods

The periods detected with either the analysis of variance [ASC89] or the discrete Fourier analysis [JSc82] were hardly ever the optimal periods. The reason for this is that both methods assume certain characteristics of the data which in most cases are not fulfilled. This situation is made worse by the poor to very poor window functions. Therefore, a simple procedure for fine-tuning the periods was developed that makes use of the features of a light curve. Here we report on the method used at the Royal Greenwich Observatory, the method used at Geneva Observatory uses similar principles.

The first step is the fit of a spline function through the data, in such a way that the spline function is continuous in zero, first and second order derivatives when going from phase 1 back to phase 0: the resulting fitted curve can be used as a periodic signal. The number of nodes and the positions of the nodes depend on type of light curve and total amplitude of the signal with respect to the noise on the data.

If $f(\phi)$ is the function describing the light curve, and p is the period used to fold the signal, then the period correction can be obtained from a least squares solution of the following observation equations:

$$O_i - f(\phi_i) = -\frac{t_i - t_0}{p} \frac{\Delta p}{p} \left(\frac{\partial f(\phi)}{\partial \phi} \right)_{\phi=\phi_i}, \quad (15.4)$$

where O_i represents observation i for which, when folded with period p the phase is given by $\phi_i = \left(\frac{t_i - t_0}{p} \bmod 1 \right)$. The time of measurement (in baricentric Julian days) is give by t_i , while t_0 is an arbitrary zero- point

somewhere halfway in the total stretch of data. The new period is then $p' = p + \Delta p$.

The solution of equation (15.4) is iterated a few times with the solution for the curve fit $f(\phi)$, until period corrections become negligible. Solving the period through such an iteration has the advantage of optimizing the period, while effectively putting most weight on those measurements that contain most information on the period: the measurements on the steepest parts of a light curve. The final solution provides an accuracy estimate for the period which in most cases appeared to be slightly on the optimistic side, and tends to depend to some extent on the actual curve fit $f(\phi)$. For automated processing, however, it was a practical solution, as it could provide from the solution a measurement of the phases and magnitudes at minimum and maximum luminosity. Equation (15.4) can also be used to fit data to a slowly varying period, by expressing the correction Δp as a function of time.

15.5 Aspects for a future mission

The European Space Agency carries out a program with so-called corner-stone missions and medium scale missions. As one of its possible future corner-stone missions (in the Horizon 2000+ program), it has opted for an interferometric satellite. Two options are considered: an infra-red interferometer for work on extra-solar planets, and a second astrometry mission. The latter project is preferred, if it can be shown that accuracies of 0.01 mas can be reached. Recent work by [HMF96] shows that this is possible. The preliminary designs for this possible mission have been named GAIA. An accuracy of 0.01 mas and a complete survey down to 16th magnitude would chart large parts of our galaxy and firmly establish distances and luminosities (provided reddening corrections can be derived) for various extra-galactic distance calibrators. Details of the GAIA concept can be found in [PvL95].

A GAIA mission would measure 50 to 100 million stars over a period of 5 to 10 years. It would contain 2 to 3 double telescopes, giving 4 to 6 fields of view. The possible orbit would put it much further away from Earth than Hipparcos, decreasing the interruption time due to Earth occultations. Photometric data will be accumulated in some 8 intermediate band-width filters or possibly as spectrophotometric information. For photometric data this has the following consequences: the coverage over short time-scales is very much improved with respect to Hipparcos. The spectroscopic or colour information is much better, giving a better interpretation possibility of the character of variability. However, the long gaps remain and will keep on damaging the window function with leaks for periods longer than a few days. Given the fact that the amount of data obtained from GAIA will

be about 20 000 times more than for Hipparcos, for which we examined 13 million data points, automated processing is the only way forward. The challenge we put forward therefore is: can a reliable period-searching mechanism be designed to work under these rather bad circumstances, or will it be necessary to partially fill in the gaps using one or two small auxiliary satellites, describing the same scanning law with a phase-lag?

Acknowledgments: Many people have contributed to the final success of the Hipparcos mission, the present paper shows only one minor aspect of its results. We would like to express our thanks in particular to Margaret Penston, Michel Grenon, François Mignard and Michael Perryman, as well as the Input Catalogue Consortium led by Catherine Turon.

REFERENCES

- [D*94] E. Daly, F. van Leeuwen, H. D. R. Evans and M. A. C. Perryman, 1994 IEEE Trans. Nucl. Science, 41, 2376-2382
- [GJB77] G. J. Bierman, 1977 Factorization Methods for Discrete Sequential Estimation, Academic Press, London
- [TD75] T. J. Deeming, 1975 Astroph. SpaceSc., 36, p137-158
- [GM78] J. R. Green & D. Margison, 1978 Statistical treatment of experimental data, Elsevier
- [H*95] E. Høg et al, 1995 Astron. Astroph., 304, p150-159
- [HMF96] E. Høg, V. Makarov, C. Fabricius, 1996 preprint
- [APp91] A. Papoulis, 1991 Probability, Random Variables, and Stochastic Processes, Third edition, McGraw-Hill
- [P*92] M. A. C. Perryman et al, 1992 Astron. Astroph., 246, p1-6
- [P*95] M. A. C. Perryman et al, 1995 Astron. Astroph., 304, p69-81
- [PvL95] M. A. C. Perryman and F. van Leeuwen (ed), 1995 ESA SP379
- [JSc82] J. D. Scargle, 1982 Astroph. Journ., 263, p835-853
- [JSc89] J.D. Scargle, 1989 Astroph. Journ., 343, p874-887
- [ASC89] A. Schwarzenberg-Czerny, 1989 Mon. Not. Roy. Astr. Soc., 241, p153-165
- [WW95] J. Z. Wilcox and T. J. Wilcox, 1995 Astron. Astroph. Suppl. Ser., 112, p395-405

15.6 Appendix: Period searching methods used

This section contains a brief outline of the methods used at the Royal Greenwich Observatory for period searches. One of us conducted a literature search aimed at finding the method best equipped to cope with the Hipparcos data. It soon became apparent, however, that the irregularly sampled (but not entirely random), sparse data imposed great limitations. The two methods chosen were selected because of their simplicity and reli-

bility, but even they could not be fully implemented: due to the complicated window functions it was decided not to determine statistical significances of the detected periods. Instead, the appearance of the light curve together with any additional information available were used as final criteria for the acceptance or rejection of a “detected” period.

15.6.1 Analysis of variance

Analysis of variance is a well-known statistical method and readers unfamiliar with it should refer to publications devoted to data analysis for a description (see e.g. [GM78]). The version and method used for the analysis at the Royal Greenwich Observatory was derived from the one described and analyzed by [ASC89].

Let $X(t)$ denote a time series, usually assumed to be of the form $X(t) = S(t) + R(t)$, where $S(t)$ is the signal and $R(t)$ the noise. $X(t)$ consists of N elements $x(t)$, individual measurements of measurement error $\sigma(t)$, like the magnitudes and their estimated errors described in Section 15.2. For every trial period P with which the data is folded, the time series $X(t)$ can be distributed over k phase bins, corresponding to k different phase intervals. The overall mean value of $x(t)$ is defined by (see e.g. [APP91], p187, example 8-4):

$$\tilde{x} = \left(\sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij} / \sigma_{ij}^2 \right) / \sum_{i,j} 1 / \sigma_{ij}^2,$$

where ij refers to the element j in bin i .

Similarly the mean value of the observations x contained in bin i :

$$\tilde{x}_i = \left(\sum_{j=1}^{n_i} x_{ij} / \sigma_{ij}^2 \right) / \sum_{j=1}^{n_i} 1 / \sigma_{ij}^2,$$

and its variance:

$$\sigma_i = \frac{1}{\sum_{j=1}^{n_i} 1 / \sigma_{ij}^2},$$

where n_i denotes the number of elements x in bin i , and $\sum_{i=1}^k n_i = N$. The following two statistics can be constructed:

$$s_1 = \sum_{i=1}^k \frac{(\tilde{x}_i - \tilde{x})^2}{\sigma_i^2}$$

and

$$s_2 = \sum_{i=1}^k \sum_{j=1}^{n_i} \frac{(x_{ij} - \tilde{x}_i)^2}{\sigma_{ij}^2}.$$

In the absence of a signal S and in the presence of white Gaussian noise, statistics s_1 and s_2 have a χ^2 distribution of respectively $k - 1$ and $N - k$ degrees of freedom, and since they are independent, the ratio $\mathbf{r} = (s_1/(k - 1))/(s_2/(N - k))$ will have a Fisher-Snedecor distribution of $k - 1$ and $N - k$ degrees of freedom.

15.6.2 Fourier analysis - Scargle's periodogram

The method defined in [JSc82] is a modification of the discrete Fourier analysis which corrects for non-orthogonality caused by uneven distribution of time-points. The periodogram is defined as:

$$\mathbf{P}_x(\omega) = \frac{1}{2} \left\{ \frac{\left[\sum_j x_j \cos \omega(t_j - \tau) \right]^2}{\sum_j \cos^2 \omega(t_j - \tau)} + \frac{\left[\sum_j x_j \sin \omega(t_j - \tau) \right]^2}{\sum_j \sin^2 \omega(t_j - \tau)} \right\},$$

where τ is defined by:

$$\tan(2\omega\tau) = \frac{\sum_j \sin 2\omega t_j}{\sum_j \cos 2\omega t_j},$$

and ω is the trial frequency.

Such modification maintains the exponential distribution of the power \mathbf{P} for unevenly spaced data if x is pure Gaussian noise (in parallel with the classical periodogram for evenly spaced data). For a full description the reader is referred to the references quoted above.

The distributions of both \mathbf{r} and \mathbf{P} change if a periodic signal is present in the data (but also if the assumptions relating to the noise are violated). The search for a periodic signal involves the construction of \mathbf{r} and \mathbf{P} for several trial periods and testing the hypothesis that no periodic signal is present against the one which assumes existence of such signal of some period.

Discussion by P. J. Bickel¹

As Dr. van Leeuwen ably described, the Hipparchos satellite gathered astrometric and photometric information on 118000 nearby stars resulting in 1.3×10^7 data points. The primary goal of the mission was to make

a) Highly precise determinations of positions, proper motions and parallaxes (with measures of precision).

To this were added photometric goals,

¹Prepared with partial support of NSF Grant DMS 9504955 and NSA Grant MDA 904-94-11-2020
University of California, Berkeley

- b) To determine with high accuracy magnitudes of these stars (with measures of precision), and
- c) To determine whether these objects were variable or not and if so estimate types of variability, periods etc.

An interesting particular feature of this study was that two groups analyzed the data from each star to a considerable extent independently. Statistical issues related to a) above were not presented and those related to b) were discussed but the statistics deemed adequate. Questions were raised partly in relation to c) and a proposed new larger mission (GAIA) as to whether there were “ways of improving methodology so that a reliable period seeking mechanism can be designed to work under rather bad circumstances”.

Dr. van Leeuwen added a good bit of detail in his presentation on b) above to the extent that I was able to try to translate at least in part what was being done into language more familiar to statisticians. I devoted part of my presentation to this translation and checking on the statistical approaches in b) for instance seeing whether the feature of two independent analyses was being used to do model checking. My main purpose was to determine whether we were on the same wavelength, an indispensable feature of any interdisciplinary discussion. Dr. van Leeuwen’s response made it clear that indeed the types of empirical models and diagnostic checks I discussed were used in their analyses and are in common use among astronomers. So I will dispense with this part of my discussion and briefly discuss,

- (i) His account of the methods used to search for variable objects and possible alternatives.
- (ii) Ways of using information from Hipparcos and other available data to perhaps improve the search in future missions.

The light curves used in the variability search can be thought of as vectors. $\{M_i(t_k) : 0 \leq k \leq K_i\}$, $1 \leq i \leq I$ where $M_i(t_k)$ is the consensus (of the two analysis groups) reduced observation of the flux of object i at time t_k . Then one can represent

$$M_i(t_k) = s_i(p^{-1}(t_k - t_0)) + e_{ik}, 0 \leq k \leq K_i \quad (1)$$

where $s_i(p^{-1}(t_k - t_0))$ is the “true” light curve and e_{ik} represents measurement error due to various systematic components introduced by the reduction process and the factors necessitating the reduction process. A nonvariable star has s_i constant. If the star is periodically variable then p is to be interpreted as its period and s_i is of period 1. If s_i is not constant but aperiodic the star is a nonperiodic variable.

The statistical character of the e_{ik} is difficult. There is a range of assumptions that can be made. It appears to me that implicitly van Leeuwen

et al. are assuming that the $\{e_{ik} : 1 \leq k \leq K_i\}$ are independent in i and possibly that $e_{ik} = \tau_i \epsilon_{ik}$ where the ϵ_{ik} are identically distributed. That is, reduced observations on stars are independent and even if the consecutive measurements on an individual star are not viewed as independent at least the standard deviation of the errors doesn't depend on time. Even the first of these assumptions is dubious since stars are not treated entirely individually during data reduction. Nevertheless, as I understand it, the HIPPARCOS approach is then

- i) Test the hypothesis $H : s_i(t) \equiv \mu_i$ (no variability) by calculating for a star i

$$T_i \equiv \sum_{j=1}^{K_i} (M_i(t_j) - \bar{M}_i)^2 / \tau_i^2 \quad (2)$$

where \bar{M}_i is the average of the $M_i(t_j)$ and τ_i^2 is variance of the e_{ij} , estimated more or less exogeneous to the $M_i(t_0)$ and

- ii) Decide to treat the star as variable and compare further if $T_i \geq t_0$ where t_0 is a threshold obtained by computing the distribution of T_i based on a large set of "known" nonvariable stars. As van Leeuwen et al. note, their thresholding method is not quite satisfactory. Their Figure 2 plots the log of the significance probabilities versus flux for the whole population of stars they have studied. Rather than, as might have been expected, showing a linear relationship initially, curving away from linearity in the tail because of the presence of variable stars the plot is distinctly nonlinear throughout. This, of course, suggests that the subpopulation of nonvariable stars of the study population is different than the reference population from which the threshold is constructed.

If the differences have to do with different proportions of star "types" with different "noise" variances τ_i^2 , $i = 1, \dots, I$ a possible remedy is to bin the reference population according to values of τ^2 and then threshold a new star i in the bin corresponding to τ_i^2 .

Evidently, further binning of the reference population according to other possibly relevant characteristics (e.g. color) may also make sense, though evidently one has to make sure that there are a reasonable number of stars in each bin. Plots such as those of Figure 2 can then be recomputed, hopefully with more satisfactory results. Here are two other minor suggestions.

- (i) It may be reasonable to try simple classifiers other than T_i , for instance,

$$MAX_i = \max\{|M_i(t_j) - \bar{M}_i|/\tau_i : 1 \leq j \leq K_i\} \quad (3)$$

- (ii) For stars in which the series of measurements is long i.e. K_i not too small, if it is believed that the e_{ik} are independent as well as identically distributed then one could use a bootstrap ([ET 93]) threshold.

That is, resample samples $e_{1b}^*, \dots, e_{K_i b}^*$, $b = 1, \dots, B$ from the residuals $M_i(t_j) - \bar{M}_i$; $1 \leq j \leq K_i$. Form $T_{i1}^*, \dots, T_{iB}^*$ where

$$T_{ib}^* = \sum_{j=1}^{K_i} [e_{jb}^*]^2 / \tau_i^*$$

and then use a threshold based on the $\{T_{ib}^* : 1 \leq b \leq B\}$. For instance if a “significance” of .001 is desired then $B = 1000$ and having the observed T_i be as large or larger than the largest T_{ib}^* would do.

Following the initial determination of variability, van Leeuwen et al. fit one period to the light curves of stars by first estimating the period crudely using an analysis of variance criterion or Scargle’s periodogram and then refine that estimate in a way they describe. I believe their method boils down to the following algorithm whose goal is to obtain the maximum likelihood solution to model (1) when

- (1a) $s_i(t)$ is assumed to be a cubic spline with continuous 0 first and second order derivatives and
- (1b) The e_{ik} are assumed to be independent and identically distributed Gaussian mean 0 variables. normal distribution

That is they, in the end, minimize

$$\sum_{j=0}^{K_i} (M_i(t_j) - s_i(\frac{(t_j - t_0)}{p}))^2$$

over all smooth cubic splines s of period 1 and all periods $p > 0$. To see this note that the likelihood equations for this model correspond to,

- (i) For $p = \hat{p}$ fixed. The normal (least squares fit) equations for the observation equations given by

$$M_i(t_j) = s_i(\frac{1}{\hat{p}}(t_j - t_0)) + e_{ij} \quad (4)$$

- (ii) A nonlinear equation in $\frac{1}{p}$ for s_i fixed. However, for s_i fixed and p close to p_0 we can replace (1) by

$$M_i(t_j) = s_i((t_j - t_0)/p_0) - (\frac{p - p_0}{p_0})(\frac{t_j - t_0}{p_0}) \frac{\partial s_i}{\partial v}(v)|_{\frac{(t_j - t_0)}{p_0}} + e_{ij} \quad (5)$$

and this is just their model (15.4).

So using a good starting point and iterating (4) and (5) to convergence we obtain the solutions to the likelihood equations for model (1a), (1b).

Of course, there may be multiple solutions and if the initial ANOVA or Scargle values are not good, there can be trouble. Still, what else can be done? One possibility, see my summary lecture, is to use robust fits, for example, minimize

$$\sum_{j=0}^{K_i} |M_i(t_j) - s\left(\frac{(t_j - t_0)}{p}\right)|$$

More interesting in preparing for the GAIA project might be to use empirical Bayes ideas (see [ML 89], [M 83], [R 83]) for example, as follows: There are from Table 15.1 a number of different types of variable stars. For each type one would expect to have available from HIPPARCOS and other sources a substantial number of light curves many one would hope taken under better conditions than available under HIPPARCOS or GAIA. I suppose this is the case.

We can think of the light curve of a variable periodic star as corresponding to a fitted cubic spline of period 1 as above and period p . Suppose then we have M types ($M = 8$ in Table 15.1 though, as I understand semiregular stars are not periodic in the usual sense.) Suppose for type m we have N_m known good light curves. Each of these light curves has a vector parameter $\boldsymbol{\theta} = (\mathbf{a}, p)$ where \mathbf{a} are the parameters of the best cubic spline fit and p is the period. So, for type m we have N_m vectors $\boldsymbol{\theta}_{m1}, \dots, \boldsymbol{\theta}_{mN_m}$ as above.

For simplicity suppose that we can, in advance, determine what type m a new variable star i whose period we are trying to determine belongs to and that we have light curve measurements (possibly very poor) $M_i(t_j)$, $1 \leq j \leq K_i$ on it. The “nonparametric” empirical Bayes point of view would now be to put a prior on the parameter $\boldsymbol{\theta}$ of the new star which is uniform on the known parameters $\boldsymbol{\theta}_{m1}, \dots, \boldsymbol{\theta}_{mN_m}$. If we couple this with model (1a), (1b) we arrive at the following fitting procedure:

I: Compute $\pi_k \equiv \alpha_k / \sum_{\ell=1}^{N_m} \alpha_\ell$, $1 \leq k \leq N_m$ where,

$$\alpha_k = \prod_{j=1}^{K_i} \varphi_i(M_i(t_j) - s_{km}\left(\frac{(t_j - t_0)}{p_{km}}\right))$$

where $\varphi_i(z) = \frac{\tau_i^{-1}}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2\tau_i^2}\right\}$ the Gaussian density with variance τ_i^2 (corresponding to measurements on the new star), the t_j refer to the light curve times on the new star, but s_{km}, p_{km} are determined by $\boldsymbol{\theta}_{km}$. (p_{km} is just the period of known star k of type m .) Now take the estimated period of the new star to be,

$$\hat{p} = \sum_{k=1}^{N_m} \pi_k p_k$$

and the corresponding light curve to be

$$\hat{s}_i(t) = \sum_{k=1}^{N_m} \pi_k s_{km}(t)$$

There is a possible logical inconsistency in this approach since one begins by assuming that only the S_{km} and p_{km} , $1 \leq k \leq N_m$, are possible and ends up by adding the new \hat{p}, \hat{s}_i , but in practice I do not believe this is troublesome. There are many other possibilities. For instance, one could assume a prior Gaussian distribution, of which $\theta_{1m}, \dots, \theta_{N_m m}$ are independent observed realizations. This would correspond to a parametric empirical Bayes approach. Whatever algorithm one produces in this fashion or another can now be tested for performance in a model free way as follows. Generate “bad” light curves from the “good” stars by taking the signals determined by $\theta_{1m}, \dots, \theta_{N_m m}$, at irregularly spaced time points. Add Gaussian noise with the expected characteristics. Run this “data” whose true character is known through the algorithm and examine performance characteristics.

It is more or less clear how to develop both algorithms and tests if the type of a star is not known for sure, but we simply have prior probabilities based on characteristics other than just fluxes or even just relative frequencies of types among known variable objects. I hope these notions, which I suspect are not new to astronomers, will prove useful if not in this study then in the similar endeavours. I have discussed them in a different unrealistic context (discussion to [N 92]) earlier. I don’t know whether the data of the type I have assumed are available to give them a chance here.

REFERENCES

- [ML 89] Maritz, J. and Lwin, T. **Empirical Bayes Methods**. 2nd Ed. Chapman and Hall London 1989.
- [M 83] Morris, C.N. Parametric empirical Bayes inference: Theory and application. *J. Amer. Statist. Assoc.* **78**, 47-59 1983.
- [R 83] Robbins, H. Some thought on empirical Bayes estimation. *Ann. Statist.* **11**, 713-723 1983.
- [N 92] Nousek, J.A. Source existence and parametric fitting when few counts are available in *Statistical Challenge, in Modern Astronomy*. E. Feigelson and G.S. Babu Eds. Springer Verlag 1992.
- [ET 93] Efron, B. and Tibshirani, R. **An Introduction to the Bootstrap**. Chapman and Hall 1993.

Time Series Analysis

Advanced problems and methodologies of time series analysis are treated in the next few chapters. PRIESTLEY and SCARGLE present the theory and application of wavelet analysis for nonstationary time series. Priestley emphasizes the relationship to Fourier analysis and Scargle elucidates tools and applications. GUÉGAN and commentator SCARGLE discuss nonparametric approaches to nonlinear and chaotic time series. VAN DER KLIS details an important application of Fourier methods to the complex variability of accreting X-ray binary star systems.

16

Application of Wavelet Analysis to the Study of Time-dependent Spectra

M. B. Priestley

ABSTRACT

In recent years wavelet analysis has attracted considerable interest as a method of constructing time/frequency decompositions of signals, and it has been suggested that wavelet transforms can be used to estimate time-varying power spectra. In this paper we examine the mathematical framework required for a physically meaningful interpretation of time-varying spectra, and then discuss the extent to which wavelet transforms can be used to estimate such spectra.

16.1 Introduction

Spectral analysis is a fundamental technique of signal processing and has an enormously wide range of applications throughout all the physical sciences. In its standard form, however, its application is limited to signals which are “stationary”, i.e. signals whose statistical properties do not change over time. Like all physical ideas the concept of “stationarity” is an idealisation which, to a certain degree of approximation, may hold for some types of signals but will be invalid for other types. When we enter the territory of “non-stationary” processes the situation becomes much more complex, and attempts to extend the ideas of spectral analysis to this more general case encounter formidable difficulties both of a mathematical and physical nature.

It is not difficult to see why the notion of stationarity is such an appealing one: it endows the signal with “statistical stability” so that quantities originally defined as ensemble averages (such as autocovariance and autocorrelation functions) can be estimated from a single run of data by computing the corresponding time domain averages. If we drop the assumption of stationarity and do not replace it with any alternative assumptions then there is very little which we can say about a given signal. In order to develop a useful theory we need to replace stationarity by a more general notion which still allows us to carry out meaningful statistical analyses. The basic

strategy is to move away “gradually” from the property of stationarity and consider signals which, although globally non-stationary, are in some sense “locally stationary”. This leads us very naturally to the idea of dividing up the time domain into a sequence of small intervals and defining a different form of power spectrum for each of the time intervals. Extending this idea leads us to the notion of a *continuously changing power spectrum*, i.e. a *time-dependent power spectrum*.

There is now a vast literature of the subject of ‘time-frequency decompositions’ of signals, and virtually all the approaches considered lead to the notion of a time-dependent spectrum, suitably defined (see, e.g., Priestley [Pri88], p158, for a review). However, although the idea of a power spectrum changing continuously over time may seem a fairly simple one, its precise mathematical formulation turns out to be an extremely difficult one: it involves, in particular, a careful re-examination of the concept of “frequency”. Moreover, since we cannot “see” a sinusoidal component unless we examine the signal over a non-zero time interval, we encounter a fundamental physical limitation, called the “*uncertainty principle*”, which tells us, in effect, that we cannot obtain simultaneously a high degree of resolution in both the time domain and the frequency domain. This principle applies to all methods of constructing time-dependent power spectra (Priestley [Pri88], p151)), and, as pointed out by Daniels [Dan65] and Tjøstheim [Tjø76], it is completely analogous to Heisenberg’s uncertainty principle in quantum mechanics.

Recently, the study of time-dependent spectral analysis has received new impetus from the explosive interest in *wavelet analysis*. The basic aim of wavelet analysis is to represent a signal as a linear superposition of “wavelets” centred on a sequence of time points. In this respect it represents a natural tool for the investigation of “local” properties of time varying signals, and many authors have attempted to interpret wavelet analysis as a form of “time-frequency” decomposition (see, e.g., Strang [Str93], Scargle [Sca94]). In fact, phases such as “local power spectra” are widely used in the wavelet literature, but the motivation for this type of terminology has so far been largely intuitive. In this paper we will try to establish a more rigorous and analytical link between wavelet analysis and time-dependent spectral analysis. We first present a brief review of the basic ideas underlying wavelet analysis.

16.2 Wavelet analysis

The essence of wavelet analysis is to expand a given function $f(t)$ as a sum of “elementary” functions called *wavelets*. These wavelets are themselves derived from a single function $\psi(t)$, called the *mother wavelet*, by translations and dilations as explained below. The mother wavelet $\psi(t)$ is

chosen so that $\int_{-\infty}^{\infty} \psi(t)dt = 0$, and has to satisfy certain other analytical conditions which effectively ensure that $\psi(t)$ is “well localised”, i.e. decays “quickly” to zero at a suitable rate.

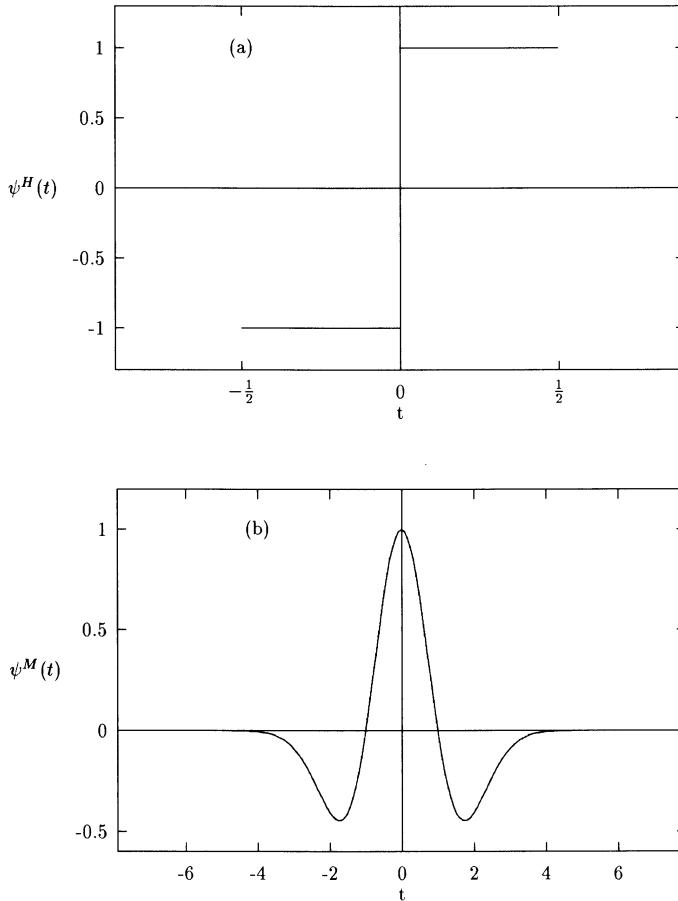


FIGURE 1. (a) The Haar function. (b) The “Mexican hat function”.

There are many different forms of $\psi(t)$ all of which satisfy the above conditions. The classical mother wavelet is the *Haar function* (fig 1a) defined by,

$$\psi^H(t) = \begin{cases} 1, & 0 \leq t \leq \frac{1}{2}, \\ -1, & -\frac{1}{2} \leq t < 0, \\ 0, & \text{otherwise.} \end{cases} \quad (16.1)$$

However, another commonly used wavelet is the “Mexican hat function” (fig 1b),

$$\psi^M(t) = (1 - t^2)e^{-t^2/2}. \quad (16.2)$$

Given a mother wavelet $\psi(t)$ we now construct a doubly infinite sequence of *wavelets* by applying varying degrees of translations and dilations to $\psi(t)$. Specifically, for all real a , b , ($a \neq 0$), write

$$\psi_{a,b}(t) = |a|^{-\frac{1}{2}} \psi\left(\frac{t-b}{a}\right) \quad (16.3)$$

so that a (usually restricted to positive values) represents the *scale parameter*, and b the *translation parameter*. (The normalising factor $|a|^{-\frac{1}{2}}$ is chosen so that $\int |\psi_{a,b}(t)|^2 dt = \int |\psi(t)|^2 dt$.) The crucial property of the set of wavelets $\{\psi_{a,b}(t)\}$ is that any (deterministic) function $f(t) \in L_2$ (the space of square integrable functions) can be expressed as a linear superposition of the $\{\psi_{a,b}(t)\}$. Thus we can write (see, e.g. Daubechies [Dau92], p25)),

$$f(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \langle f, \psi_{a,b} \rangle \psi_{a,b}(t) \frac{dadb}{a^2}. \quad (16.4)$$

where $\langle f, \psi_{a,b} \rangle$ denotes the L_2 inner product, i.e.

$$\langle f, \psi_{a,b} \rangle = \int_{-\infty}^{\infty} f(t) \psi_{a,b}(t) dt \quad (16.5)$$

and

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega$$

The expression (16.5) is called the *continuous wavelet transform* of $f(t)$. Remarkably, we can still obtain a representation of $f(t)$ similar to (16.4) even when a and b are restricted to *discrete* sets of values. Thus, if we choose fixed values a_0 , b_0 , and set $a = a_0^{-m}$, $b = nb_0a_0^{-m}$, $n, m = 0, \pm 1, \pm 2, \dots$ and set

$$\psi_{m,n} = a_0^{-m/2} \psi\left(\frac{t - nb_0a_0^{-m}}{a_0^{-m}}\right)$$

then with suitably chosen $\psi(t)$, a_0 , b_0 , we can still write any function $f(t) \in L_2$ in the form

$$f(t) = \sum_m \sum_n \tilde{\psi}_{m,n}(t) \langle f, \psi_{m,n} \rangle \quad (16.6)$$

where $\{\tilde{\psi}_{m,n}(t)\}$ form the so-called “dual frame” wavelets (see Daubechies [Dau92], p54)), and

$$\langle f, \psi_{m,n} \rangle = \int_{-\infty}^{\infty} f(t) \psi_{m,n}(t) dt \quad (16.7)$$

is called the *discrete wavelet transform*.

The most commonly used discrete set of wavelets, associated with “*multiplesolution analysis*,” is generated by setting $a_0 = \frac{1}{2}$, $b_0 = 1$, yielding for $m, n = 0, \pm 1, \pm 2, \dots$

$$\psi_{m,n}(t) = 2^{m/2} \psi(2^m t - n) \quad (16.8)$$

It may be shown that the class of wavelets given by (16.8) forms a *complete orthonormal basis* for L_2 (Mallat [Mal89], Daubechies [Dau92], p129)), and hence any $f(t) \in L_2$ admits the representation

$$f(t) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} a_{m,n} \psi_{m,n}(t) \quad (16.9)$$

where

$$a_{m,n} = \int_{-\infty}^{\infty} f(t) \psi_{m,n}(t) dt \quad (16.10)$$

Note that in (16.8) the mother wavelet ψ is *dilated* by the factor 2^{-m} and *translated* to the position $2^{-m}n$. In the context of multiresolution analysis the approximation of the function $f(t)$ at resolution $2^{-\ell}$ is given by

$$\hat{f}_{\ell}(t) = \sum_{m=-\infty}^{\ell} \sum_{n=-\infty}^{\infty} a_{m,n} \psi_{m,n}(t), \quad (16.11)$$

and the inner sum in (16.11) is called the “*detail signal at level* 2^{-m} ”. The approximation $\hat{f}_{\ell}(t)$ can be written in the alternative form,

$$\hat{f}_{\ell}(t) = \sum_{n=-\infty}^{\infty} b_{\ell,n} \phi_{\ell,n}(t) \quad (16.12)$$

where $\{\phi_{\ell,n}(t)\}$ constitute an orthonormal basis and are derived from the corresponding *scale function* $\phi(t)$ by setting $\phi_{\ell,n}(t) = 2^{\ell/2} \phi(2^{\ell}t - n)$, $n = 0, \pm 1, \pm 2 \dots$, and

$$b_{\ell,n} = \int_{-\infty}^{\infty} f(t) \phi_{\ell,n}(t) dt \quad (16.13)$$

(See, e.g., Mallat [Mal89]).

16.3 Relationship with Fourier analysis

There is an obvious analogy between wavelet analysis and Fourier analysis in the sense that both techniques aim to represent a function as a liner superposition of “basis” functions. In the case of wavelet analysis the basis functions are the wavelets $\{\psi_{a,b}(t)\}$ or $\{\psi_{m,n}(t)\}$ whereas in Fourier analysis they are the complex exponentials $\{e^{i\omega t}\}$. Thus, given a function $f(t) \in L_1$ we may represent it in the form

$$f(t) = \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega \quad (16.14)$$

where $F(\omega)$, the *Fourier transform* of $f(t)$, is given by

$$F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)e^{-i\omega t} dt \quad (16.15)$$

The analogue of the discrete wavelet transform arises when $f(t)$ is periodic with period 2π (say), in which case we can represent $f(t)$ as a *Fourier series* of the form,

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{int} \quad (16.16)$$

where

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)e^{-int} dt \quad (16.17)$$

However, although this analogy is frequently referred to in the literature (see, e.g. Strang [Str93]), it should be treated cautiously since despite the common feature mentioned above there are important differences between these two approaches. The most obvious difference is that the wavelet basis functions are indexed by two parameters (a, b or m, n) whereas the Fourier basis functions are indexed by the single parameter ω . In physical terms this means that wavelet transforms (or coefficients) are characteristics of the *local* behaviour of function whereas Fourier transforms (or coefficients) are characteristics of the *global* behaviour of the function. In the case of Fourier analysis the parameter ω has the physical interpretation of *frequency*; in wavelet analysis the second parameter (b or n) represents a *time location* and the first parameter (a or m) determines the “*width*” of the wavelet. There is thus no immediate physical connection between the wavelet parameters and the Fourier frequency parameter ω , but we may establish a tenuous relationship between ω and the wavelet dilation parameter a (or m) by noting that if the mother wavelet has “oscillatory” characteristics — as, e.g., in the case of Mexican hat function (16.2) — then as a decreases the “oscillations” become compressed in the time domain, i.e. exhibit “high frequency” behaviour, whereas as a increases the “oscillations” become drawn out, i.e. exhibit “low frequency” behaviour.

In a *loose sense*, therefore, we may tentatively identify the wavelet parameters a, b (or m, n) with “*frequency*” and “*time*”, respectively, but it should be firmly noted that at this stage the physical interpretation of the wavelet parameters is purely heuristic. In particular it should be emphasised that *the physical concept of “frequency” is related purely to the family of coupled exponential functions* and has no precise meaning when applied to other families of functions — unless, of course, we can define a more general concept of “frequency” which agrees with our physical understanding. (We return to this discussion in section 5.).

One notable feature of wavelet representations is that the wavelet transforms (or coefficients) are “localised”, i.e. are time-varying and depend only

on the *local* properties of $f(t)$ in the neighbourhood of each time point. Thus, if $f(t)$ has singularities (such as discontinuities or “spikes”) these will affect only the wavelet transforms at time points near the singularities; wavelet transforms at time points well removed from the singularities will be (essentially) unaffected. By contrast, the standard Fourier transforms (or coefficients) depend on the global properties of $f(t)$, and any singularity in $f(t)$ will affect *all* such transforms. As is well-known, if a periodic function $f(t)$ has a discontinuity then we would require a very large number of terms in its Fourier series in order to obtain an adequate approximation of $f(t)$ in the region of the discontinuity. However, if $f(t)$ is “smooth” except at the discontinuity (say, piece-wise continuous) then we may obtain quite a good approximation by using a relatively small number of wavelet coefficients (see, e.g. Nason and Silverman [NS94]).

One way in which we can strengthen the analogy between wavelet and Fourier analysis is by introducing the notion of *time-dependent* or *windowed* Fourier transforms, constructed as follows. Choose a function $g(t)$ which is concentrated around $t = 0$ and decays to zero as $t \rightarrow \pm\infty$ — rather like a mother wavelet but without the condition that it integrates to zero. We now define the windowed Fourier transform of $f(t)$ by

$$F_t(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(s)g(s-t)e^{-i\omega s} ds \quad (16.18)$$

(What we have done, in effect, is to “isolate” a portion of the function f in the neighbourhood of the time point t by giving high weight to the values of f near t and low weight to values far removed from t). The windowed Fourier transform $F_t(\omega)$ now shares with wavelet transforms the property that it is time-dependent, i.e. is a function of two variables, t and ω , but it is, of course, dependent on the arbitrary choice of the “window” $g(t)$ and does not therefore characterise purely the properties of the function $f(t)$. The physical interpretation of $F_t(\omega)$ is open to question; it is tempting (and indeed customary) to think of $F_t(\omega)$ as a “short time” Fourier transform — particularly if $g(t)$ is chosen to have a rectangular form and vanishes outside a finite interval. However, the question arises as to exactly what $F_t(\omega)$ is measuring, and what type of representation for $f(t)$ leads to a meaningful interpretation of $F_t(\omega)$.

The use of windowed Fourier transforms raises a further and quite fundamental difficulty. If we wish to obtain “high resolution” in the time domain, i.e. we wish $F_t(\omega)$ to reflect the properties of f strictly near the time point t (and not be contaminated by values of f far removed from the point t) then the window $g(t)$ must be chosen so that it decays to zero very quickly, i.e. it must have a very small time domain “width”. In this case the integral in (16.18) operates only over a very small time interval and $F_t(\omega)$ then loses resolution in the frequency domain. Conversely, if we wish $F_t(\omega)$ to have high frequency domain resolution then $g(t)$ will have to be chosen

so that it has a large “width” and we then lose resolution in the time domain. This feature is a consequence of a fundamental result known as the “*Uncertainty Principle*” (Priestley [Pri88], p151), and applies equally well to wavelet transforms. The often stated objective that (mother) wavelets should be chosen so as to achieve “good localisation” in both the time and frequency domains is, in fact, impossible to achieve in an ideal sense.

Despite the fact that wavelet and Fourier transforms are both time and frequency dependent there is an important difference between the two techniques. In the case of windowed Fourier transforms the “width” of the window $g(t)$ remains constant as the frequency variable ω changes. Thus, both high and low frequency components are evaluated over the same (effective) time interval. However, in the case of wavelet transforms a decrease in the value of the parameter a (or an increase in the value of m) *increases* the frequency of the wavelet oscillations and simultaneously *shrinks* the time domain width of the wavelet. Similarly, an increase in the value of a *decreases* the frequency of the wavelet oscillations and *stretches* the time domain width — as shown in fig 2 for the Mexican hat wavelet.

This is a crucial feature of wavelet analysis and means that in this approach high frequency components are evaluated over “short” time intervals whereas low frequency components are evaluated over “large” time intervals. This enables wavelet transforms to highlight “short time” high frequency phenomena such as short duration transients or singularities.

We could, of course, incorporate the same feature in windowed Fourier transforms by changing the width of the window $g(t)$ as the frequency variable ω changes - but the relationship between ω and the width of $g(t)$ would be quite arbitrary and would not be effected in the “automatic” way in which wavelets achieve this effect.

16.4 Spectral analysis of stationary processes

Fourier transforms are a natural tool for the frequency analysis of (suitable) deterministic functions. However, if instead of deterministic functions we now consider a *stochastic process*, $X(t)$, then in general Fourier transforms no longer exist and the appropriate tool for frequency analysis becomes *Spectral Analysis*. Thus a general stochastic process would not admit a standard Fourier integral representation of the form (16.14), but if we restrict attention to zero mean (stochastically continuous) *stationary processes* then there exists a generalised Fourier-Stieltjes integral representation of the form,

$$X(t) = \int_{-\infty}^{\infty} e^{i\omega t} dZ(\omega) \quad (16.19)$$

where $Z(\omega)$ is a complex-valued process with orthogonal increments over $(-\infty, \infty)$ — see, e.g. Priestley [Pri81], p246. The *integrated spectrum*, $H(\omega)$,

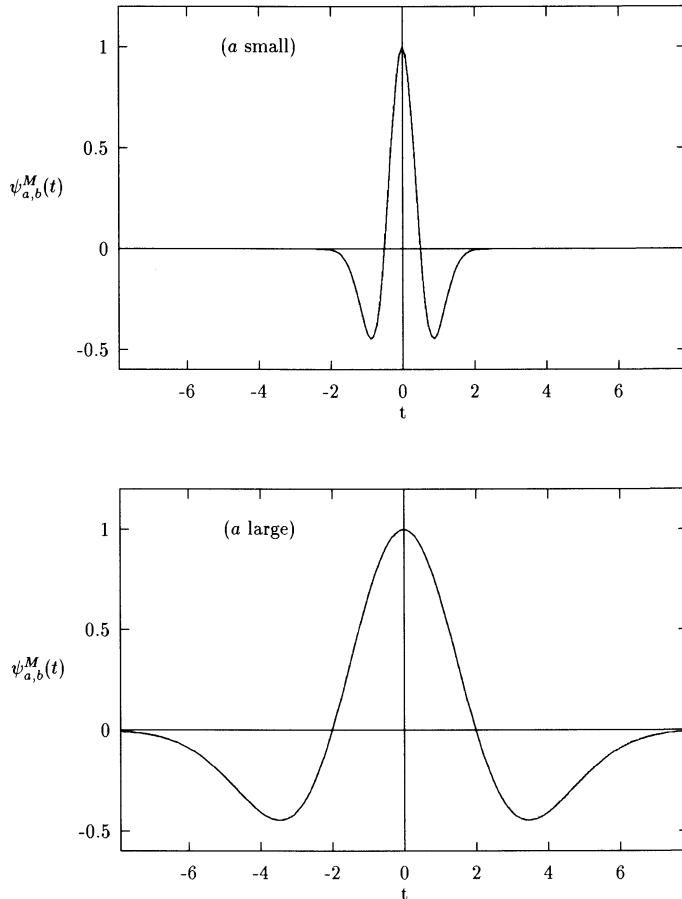


FIGURE 2. The Mexican hat wavelet.

is then defined by

$$dH(\omega) = E[|dZ(\omega)|^2] \quad (16.20)$$

When $H(\omega)$ is differentiable, $h(\omega) = H'(\omega)$ is called the (power) *spectral density function*. Since $|dZ(\omega)|$ is the (random) amplitude of the component in $X(t)$ with frequency ω , $h(\omega)d\omega = E[|dZ(\omega)|^2]$ may be interpreted as the average contribution to the total power of $X(t)$ with frequencies between $\omega, \omega + d\omega$, — hence the description of $h(\omega)$ as a power *density* function.

A basic result, known as Wiener-Khintchine theorem, allows us to evaluate $h(\omega)$ (when it exists) as the standard Fourier transform of the auto-covariance function of $X(t)$, i.e. we have,

$$h(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} R(\tau) e^{-i\omega\tau} dt \quad (16.21)$$

where $R(\tau) = E[X(t)X(t+\tau)]$ denotes the autocovariance function. Given a continuously observed finite record of $X(t)$, say for $0 \leq t \leq T$, we may construct an estimate of $h(\omega)$ by passing the record through a narrow band filter centred on frequency ω_0 , say, which (ideally) will remove all components except those with frequencies in the neighbourhood of ω_0 . The power in $X(t)$ near frequency ω_0 is then estimated by estimating the total power (i.e. variance) of the filter output. More specifically, we choose a filter with impulse response function $g(u)$ such that the corresponding transfer function

$$\Gamma(\theta) = \int_{-\infty}^{\infty} g(u)e^{-iu\theta} du, \quad (16.22)$$

is highly concentrated in the region $\theta = 0$. We then multiply $X(t)$ by $e^{i\omega_0 t}$ and pass it through this filter, giving the output

$$Y(t) = \int_{t-T}^t g(u)X(t-u)e^{-i\omega_0(t-u)} du \quad (16.23)$$

(Assuming that the impulse response function $g(u)$ decays to zero sufficiently fast, and that $t \gg 0$, the limits in the integral in (16.23) may be effectively replaced by $-\infty, \infty$). We now estimate $h(\omega_0)$ by evaluating the sample variance of $Y(t)$, i.e. we form

$$\hat{h}(\omega_0) = \frac{1}{T} \int_0^T |Y(t)|^2 dt \quad (16.24)$$

The above estimate is, in fact, equivalent to computing the (continuous time) periodogram of $X(t)$ and then smoothing this periodogram over neighbouring frequencies via the “spectral window” $|\Gamma(\theta)|^2$, i.e. if we evaluate the periodogram of $X(t)$,

$$I(\theta) = \frac{1}{2\pi T} \left| \int_0^T X(t)e^{-i\theta t} dt \right|^2 \quad (16.25)$$

then $\hat{h}(\omega)$ may be written in the equivalent form

$$\hat{h}(\omega_0) = \int_{-\infty}^{\infty} I(\theta) |\Gamma(\omega_0 - \theta)|^2 d\theta \quad (16.26)$$

— see Priestley [Pri81], p496.

16.5 Time dependent spectral analysis

The theory of spectral analysis described in the preceding section is valid only for stationary processes, i.e. processes whose second order statistical properties do not change over time. A non-stationary process does

not possess a spectrum in the conventional sense, but since its statistical properties are now changing over time it is natural to try to describe its power/frequency properties by introducing the notion of “*time dependent spectra*”, $h_t(\omega)$, constructed so that, for each t , $h_t(\omega)$ describes a *local* power/frequency distribution which is characteristic of the behaviour of the process in the neighbourhood of the time point t . From an empirical point of view the construction of time dependent spectral *estimates* is quite straightforward; given a sample record of $X(t)$ observed over the interval $0 \leq t \leq T$ we simply divide the interval $(0, T)$ into a number of sub-intervals and compute expressions of the form (16.24) separately for each sub-interval — leading to a sequence of functions $\hat{h}_t(\omega)$, say, with t corresponding (say) to the mid-part of the sub-interval. (Note, however, that in view of the “Uncertainty Principle” — referred to briefly in section 3 — the smaller we make each sub-interval the more we lose resolution in the frequency domain). This approach is echoed in the wavelet analysis literature where the squared modulus of the wavelet transform (16.5), $|\langle f, \psi_{a,b} \rangle|^2$ is often referred to as a time-dependent “*spectrogram*” — see, e.g. Daubechies [Dau92], p86, Scargle [Sca94]. However, at this stage the interpretation of any of these qualities as time dependent power spectra is merely wishful thinking. The crucial question is; what are (e.g.) the $\hat{h}_t(\omega)$ estimating? To answer this we need first to define *theoretical* time dependent spectra in order to provide a framework against which we can interpret the sample estimates described above.

Various definitions of theoretical time dependent spectra have been proposed in the literature (see Priestley [Pri88], p855, for further discussion) but most of these fail to capture the physical interpretation of “local” power/frequency distributions. The crucial difficulty concerns the interpretation of “frequency”: in the case of stationary processes we have the *spectral representation* (16.19), which tells us that a stationary process $X(t)$ can be expressed as a “sum” of sine and cosine functions with varying frequencies and (random) amplitudes. We can then identify that component in $X(t)$ which has frequency ω , and meaningfully discuss the contribution of this component to the total power of the process. In the absence of such a representation we cannot immediately talk about “power distributions over *frequency*”. Now a non-stationary process cannot be represented as a “sum” of sine and cosine functions (with orthogonal coefficients) — instead we have to represent it as a “sum” of other kinds of functions. Since, according to its conventional definition, the term “frequency” is associated specifically with sine and cosine functions we cannot talk about the “frequency components” of a non-stationary process unless we first define a more general concept of “frequency” which agrees with our physical understanding.

Consider, for example, a deterministic function which has the form of a damped sine wave, say,

$$f(t) = Ae^{-t^2/\alpha^2} \cos(\omega_0 t + \phi)$$

If we perform a conventional Fourier analysis of $f(t)$ we see that it contains Fourier components at *all* frequencies — the Fourier transform of $f(t)$ consists of two Gaussian functions, one centred on ω_0 and the other on $(-\omega_0)$, the width of these functions being inversely proportional to the parameter α . Thus, if we represent $f(t)$ as a “sum” of sine and cosine functions with constant amplitudes we need to include components at all frequencies. However, we can equally well describe $f(t)$ by saying that it consists of just two “frequency” components ($\omega = \pm\omega_0$), each component having a time varying amplitude, Ae^{-t^2/α^2} . In fact, if we were to examine the *local* behaviour of $f(t)$ in the neighbourhood of a particular time point this is precisely what we would observe, i.e. if the interval of observations was small compared with α then $f(t)$ would appear simply as a cosine function with frequency ω_0 and amplitude Ae^{-t^2/α^2} . Nevertheless, it would not be physically meaningful to attempt to assign a “frequency” to a function $f(t)$ of arbitrary form — it would make little sense to talk about the “frequency” of the function $\log \omega t$. For the term “frequency” to be physically meaningful the function $f(t)$ must possess what we can describe loosely as an “*oscillatory form*”, and we can characterise this property by saying that the Fourier transform of such a function will be concentrated around a particular point ω_0 (or around $\pm\omega_0$ in the real case). Thus if we have a non-periodic function $f(t)$ whose Fourier transform has an absolute maximum at the point ω_0 we may define ω_0 as the “frequency” of this function, the argument being that locally $f(t)$ behaves like a sine wave with (conventional) frequency ω_0 , modulated by a “smoothly varying” amplitude function. It is this type of reasoning which forms the basis of the “evolutionary spectra” approach to the analysis of non-stationary processes, and this approach rests on a spectral representation which is a direct generalisation of (16.19).

16.6 Evolutionary spectral analysis

We consider a fairly general class of stochastic processes $\{X(t)\}$ for which \exists a family of functions $\{\phi_t(\omega)\}$ such that $X(t)$ admits a representation of the form

$$X(t) = \int_{-\infty}^{\infty} \phi_t(\omega) dZ(\omega) \quad (16.27)$$

where $Z(\omega)$ is again a process with orthogonal increments. Suppose that, for each fixed ω , $\phi_t(\omega)$ (considered as a function of t) possesses a Fourier transform whose modulus has an absolute maximum at frequency $\theta(\omega)$,

say. Then $\phi_t(\omega)$ may be regarded as an amplitude modulated sine wave with frequency $\theta(\omega)$, i.e. we may write

$$\phi_t(\omega) = A_t(\omega)e^{i\theta(\omega)t}$$

where now the Fourier transform of $A_t(\omega)$ has an absolute maximum at *zero frequency*. The function $\phi_t(\omega)$ is then called an *oscillatory function*, and if the family $\{\phi_t(\omega)\}$ is such that $\theta(\omega)$ is a single valued function of ω then we may transform the variable in the integral in (16.27) from ω to $\theta(\omega)$, and by suitably redefining $A_t(\omega)$ we may write

$$X(t) = \int_{-\infty}^{\infty} A_t(\omega)e^{i\omega t} dZ(\omega) \quad (16.28)$$

When $X(t)$ admits a representation of the form (16.28) (with $A_t(\omega)$) satisfying the required condition we call it an *oscillatory process*. By virtue of the orthogonality of the $\{dZ(\omega)\}$ it follows immediately from (16.28) that

$$\text{var}\{X(t)\} = \int_{-\infty}^{\infty} |A_t(\omega)|^2 d\mu(\omega) \quad (16.29)$$

where $d\mu(\omega) = E[|dZ(\omega)|^2]$. Since $\text{var}\{X(t)\}$ may be interpreted as the “total power” of the process at time t , (16.29) gives a decomposition of total power in which the contribution from frequency ω is $|A_t(\omega)|^2 d\mu(\omega)$. This result is consistent with the interpretation of (16.28) as an expression for $X(t)$ as the “sum” of sine and cosine terms with varying frequencies and time dependent (random) amplitudes $\{A_t(\omega)dZ(\omega)\}$. We now define the *evolutionary spectrum* at time t by (cf Priestley [Pri88], p148),

$$dH_t(\omega) = |A_t(\omega)|^2 d\mu(\omega) \quad (16.30)$$

In particular, when $H_t(\omega)$ is differentiable with respect to ω (for each t), $h_t(\omega) = dH_t(\omega)/d\omega$ is called the *evolutionary spectral density function* at time t . Note that the evolutionary spectrum has the same physical interpretation as the conventional spectrum of a stationary process, namely that it describes a distribution of *power over frequency*, but whereas the latter is determined by the behaviour of the process over all time the former represents specifically the spectral content of the process in the neighbourhood of the time point t .

Evolutionary spectral density functions may be estimated from data by an extension of the method described in section 4 for the estimation of spectral density functions of stationary processes. Thus, given observations over the interval $(0, T)$ we pass the data through a linear filter centred on frequency ω_0 , say, yielding an output $U(t)$, say. We then compute a weighted average of $|U(t)|^2$ in the neighbourhood of the time point t to provide an estimate of the local power density function at frequency ω_0 .

Specifically, we set

$$U(t) = \int_{t-T}^t g(u)X(t-u)e^{-i\omega_0(t-u)}du, \quad (16.31)$$

$$\hat{h}_t(\omega_0) = \int_{t-T}^t w(v)|U(t-v)|^2dv \quad (16.32)$$

where $g(u)$ is a filter whose transfer function $\Gamma(\theta)$ (defined as in (16.22)) is peaked in the neighbourhood of $\theta = 0$ and is normalised so that $\int |\Gamma(\theta)|^2 d\theta = 1$. The filter width, $B_g = \int_{-\infty}^{\infty} |u| |g(u)| du$, is chosen so that B_g is much smaller than “characteristic width” B_X of $X(t)$, and the weight function $w(v)$ is chosen so that $\int_{-\infty}^{\infty} w(v)dv = 1$. (The characteristic width, B_X is defined so that, roughly speaking, $2\pi B_X$ may be interpreted as the maximum time interval over which the process may be treated as “approximately stationary”. For a precise definition of B_X , together with a more detailed discussion of the estimation procedure see Priestley [Pri88], Ch 6).

Assuming that both $g(u)$ and $w(v)$ decay to zero sufficiently fast so that the limits in the integrals in (16.31), (16.32) may be replaced effectively by $(-\infty, \infty)$, it may be shown that

$$E[|U(t)|^2] \sim \int_{-\infty}^{\infty} h_t(\omega + \omega_0) |\Gamma(\omega)|^2 d\omega \quad (16.33)$$

so that

$$E[\hat{h}_t(\omega_0)] \sim \int_{-\infty}^{\infty} \bar{h}_t(\omega + \omega_0) |\Gamma(\omega)|^2 d\omega \quad (16.34)$$

where

$$\bar{h}_t(\omega + \omega_0) = \int_{-\infty}^{\infty} w(v) h_{t-v}(\omega + \omega_0) dv \quad (16.35)$$

Making the usual assumption that $h_t(\omega)$ is “flat” over the bandwidth of $|\Gamma(\omega)|^2$ we may write

$$E[\hat{h}_t(\omega_0)] \sim \bar{h}_t(\omega_0),$$

ie $\hat{h}_t(\omega_0)$ is an approximately unbiased estimate of the weighted average of $h_t(\omega_0)$ in the neighbourhood of the time point t .

For discrete parameter processes (i.e. processes defined for say integer values of t), the expressions corresponding to (16.31), (16.32) becomes

$$U_t = \sum_u g_u X_{t-u} e^{-i\omega_0(t-u)}, \quad (16.36)$$

$$\hat{h}_t(\omega_0) = \sum_v w_v |U_{t-v}|^2, \quad (16.37)$$

where now $\{g_u\}$, $\{w_v\}$ are suitably chosen sequences.

16.7 Relationship with wavelet analysis

Most of the literature on wavelets is concerned with the analysis of deterministic functions and there is very little discussion of their applications to stochastic processes — although it should be noted that Cambanis and Masry [CM94] discuss the wavelet representation of both stationary and non-stationary processes but in the case of stationary processes restrict their analysis to stochastically continuous processes defined only on a finite time interval, $(0, T)$. (This condition is imposed so that, with probability 1, $\int_0^T X^2(t)dt < \infty$ and thus $X(t) \in L_2$, with probability 1.)

Let us now return to the discrete wavelet representation (16.9), viz,

$$f(t) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} a_{m,n} \psi_{m,n}(t), \quad (16.38)$$

where

$$\psi_{m,n}(t) = 2^{m/2} \psi(2^m t - n), \quad (16.39)$$

$\psi(t)$ being a given mother wavelet. As noted in section 3, it is conventional to identify the wavelet parameter m with “frequency”, although in general this interpretation is highly dubious. However, if we focus attention on the particular mother wavelet,

$$\psi(t) = \begin{cases} \sqrt{2} \sin 2\pi t, & 0 \leq t \leq 1, \\ 0, & \text{otherwise,} \end{cases} \quad (16.40)$$

then the “frequency” interpretation of m becomes much stronger. For with this class of $\psi(t)$ we see that, for fixed m , $\sum_{n=-\infty}^{\infty} a_{m,n} \psi_{m,n}(t)$ has the form of an amplitude modulated sine wave with the modulating function taking the form of a step-function as shown in fig 3.

Thus, we may write

$$\sum_{n=-\infty}^{\infty} a_{m,n} \psi_{m,n}(t) = A_t(\omega) \sin \omega t, \quad \omega = 2\pi 2^m, \quad (16.41)$$

with

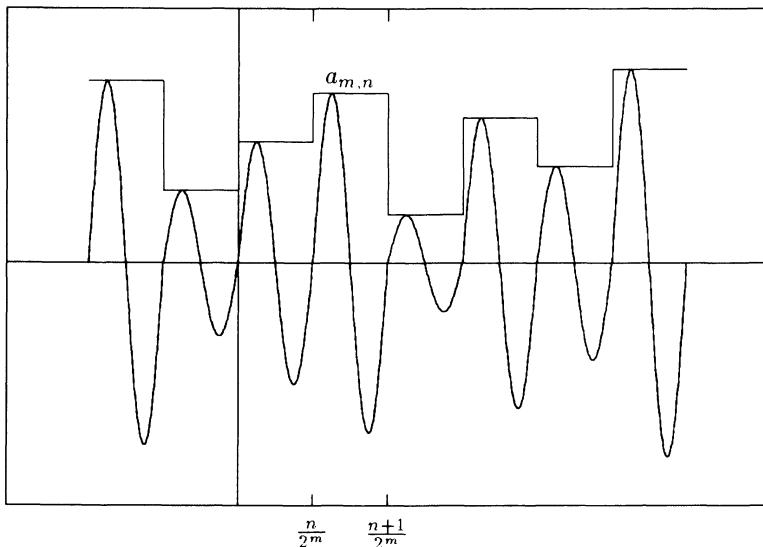
$$A_t(2\pi 2^m) = 2^{\frac{m+1}{2}} a_{m,n}, \quad \frac{n}{2^m} \leq t < \frac{n+1}{2^m} \quad (16.42)$$

Hence, we may now re-write the representation (16.38) in the form

$$f(t) = \int_{-\infty}^{\infty} A_t(\omega) \sin \omega t dZ(\omega) \quad (16.43)$$

with $A_t(\omega)$ given by (16.42) and

$$dZ(\omega) = \begin{cases} 1, & \omega = 2\pi 2^m, \quad m = 0, \pm 1, \pm 2, \dots \\ 0, & \text{otherwise.} \end{cases} \quad (16.44)$$

FIGURE 3. The form of $\sum_{n=-\infty}^{\infty} a_{m,n} \psi_{m,n}(t)$.

The similarity between (16.43) and the evolutionary spectral representation (16.28) is now quite compelling: in effect, (16.43) is a discrete approximation to (16.28), the discretisation occurring both over time (with $A_t(\omega)$ taking the form of a step-function over the time intervals $(n/2^m, (n+1)/2^m)$) and over frequency (with ω restricted to values $2\pi 2^m$). Comparing (16.43) with (16.28) and using the definition (16.30) we now have the heuristic relationship,

$$h_t(\omega) d\omega = |A_t(\omega)|^2 E[|dZ(\omega)|^2] \sim 2^{m+1} |a_{m,n}|^2, \quad n = \frac{t}{2^m}, \quad \omega = 2\pi 2^m, \quad (16.45)$$

thus justifying the time-dependent spectral interpretation of the squared modulus of the wavelet coefficients. It is important, however, to recall that in order for the function $A_t(\omega) \sin \omega t$ to be said to have "frequency" ω , the function $A_t(\omega)$ must not oscillate "too fast", i.e. its Fourier transform must satisfy the condition referred to in section 6. (Note that, with $\psi(t)$ chosen as in (16.40), $A_t(\omega)$ is constant over precisely one cycle of the function $\sin \omega t$.) Whether or not the Fourier transform of $A_t(\omega)$ satisfies the required condition depends on the form of the wavelet coefficients, $a_{m,n}$ which in turn depends on the function (or process) being analysed.

We may note further that in the representation (16.43) the frequency variable ω is restricted to discrete values at the end points of *octave bands*.

Thus, as m takes values through the range $\dots, -3, -2, -1, 0, 1, 2, 3, \dots$, ω takes values $\dots, \frac{\pi}{4}, \frac{\pi}{2}, \pi, 2\pi, 4\pi, 8\pi, 16\pi, \dots$. Also, as is characteristic of this form of wavelet analysis, *the time variable t is discretised at intervals which become progressively narrower as the frequency variable increases*; for frequency $\omega = 2\pi 2^m$ the time points are sampled at intervals of width $\frac{1}{2^m} = \frac{2\pi}{\omega}$. Thus, high frequency components are sampled at small time intervals while low frequency components are sampled at large time intervals.

We now re-examine the heuristic relationship between $h_t(\omega)$ and $|a_{m,n}|^2$ (equation 16.45) in a more detailed form. If we apply the above wavelet analysis to a zero mean stochastic process $X(t)$ then replacing $f(t)$ by $X(t)$ in (16.10) and using (16.8) we may write

$$a_{m,n} = \int_{-\infty}^{\infty} X(u)\psi_m\left(u - \frac{n}{2^m}\right) du \quad (16.46)$$

where we have written $\psi_m(t) = 2^{m/2}\psi(2^m t)$, ψ being the mother wavelet. The RHS of (16.46) has an obvious similarity with the process $U(t)$ defined by (16.31). If, for the moment, we set $\omega_0 = 0$ in (16.31) and re-write the expression for $U(t)$ in the form

$$U(t) = \int_{-\infty}^{\infty} X(u)g(u-t)du \quad (16.47)$$

then (16.46) becomes equivalent to (16.47) on identifying $g(u)$ with $\psi_m(u)$ and t with $\frac{n}{2^m}$. We now have from (16.33),

$$E[|a_{m,n}|^2] \sim \int_{-\infty}^{\infty} h_t(\omega)|\Psi_m(\omega)|^2 d\omega \quad (16.48)$$

where $\Psi_m(\omega)$ is the Fourier transform of $\psi_m(t)$. Now consider the complex form of the mother wavelet given by (16.40), namely $\psi(t) = e^{2\pi i t}$, $0 \leq t \leq 1$. Then

$$\psi_m(t) = 2^{m/2} \exp\{2\pi i(2^m t)\}, \quad 0 \leq t \leq \frac{1}{2^m}$$

and its Fourier transform is easily shown to be

$$\Psi_m(\omega) = 2^{m/2} D_{\frac{1}{2^{m+1}}}(\omega - \omega_m) e^{-i\omega/2^m}$$

where $D_M(\omega)$, the *Dirichlet kernel*, is given by $D_M(\omega) = \sin \frac{M\omega}{\omega}$, and $\omega_m = 2\pi 2^m$ (see, e.g., Priestley [Pri81], p437). Hence in this case $|\Psi_m(\omega)|^2$ is proportional to the square of a Dirichlet kernel centred on frequency ω_m , and thus $|\Psi_m(\omega)|^2$ is peaked in the neighbourhood of $\omega = \omega_m$. In fact, $|\Psi_m(\omega)|^2$ has effectively the form of a *Fejer kernel*, $F_M(\omega)$, centred on $\omega = \omega_m$ and with $M = \frac{1}{2^m}$. We may therefore conclude tentatively from (16.48) that if $|\Psi_m(\omega)|^2$ is suitably concentrated around $\omega = \omega_m$,

$$E[|a_{m,n}|^2] \sim 2\pi h_t(\omega_m), \quad t = \frac{n}{2^m} \quad (16.49)$$

(Note that $\int_{-\infty}^{\infty} |\Psi_m(\omega)|^2 d\omega = 2\pi$.) However, we have to examine the behaviour of $|\Psi_m(\omega)|^2$ for varying m very carefully. It is well-known that as $M \rightarrow \infty$ the Fejer kernel becomes highly concentrated and acts as a pseudo δ -function, while as M decreases towards zero the Fejer kernel becomes “flat”. These features are evident once it is noted that the (half-power) bandwidth of the Fejer kernel is $O(1/M)$. The band width of $|\Psi_m(\omega)|^2$ is correspondingly $O(2^m)$, and consequently $|\Psi_m(\omega)|^2$ will be concentrated around $\omega = \omega_m$ for small or negative m but for large positive m $|\Psi_m(\omega)|^2$ will be a “flat” function.

We may thus conclude that, as spectral estimates, *the discrete wavelet transforms will have good frequency resolution at low frequencies but very poor frequency resolution at high frequencies*. (This effect cannot be remedied by “smoothing” the $|a_{n,m}|^2$ over neighbouring values of n — as is done in the case of evolutionary spectral estimates — cf (16.32), (16.37). Such smoothing would simply reduce the sampling fluctuations but could not improve the resolution in the frequency domain.)

The behaviour of $|\Psi_m(\omega)|^2$ with increasing values of m is a built-in feature of wavelet transforms and holds quite generally. Indeed, it is an inevitable consequence of the fact that as m increases (or equivalently as a decreases) the wavelet $\Psi_{m,n}(t)$ “shrinks” in the time domain (cf the discussion in section 4) and correspondingly its Fourier transform becomes “flat”. It is also perfectly consistent with the often quoted property of wavelets that they have the ability to “zoom in” on local “fine detail”: high resolution in the time domain inevitably implies poor resolution in the frequency domain.

Whilst the discussion on the behaviour of $|\Psi_m(\omega)|^2$ for large m is quite general, the behaviour of $|\Psi_m(\omega)|^2$ for small m depends very much on the choice of the mother wavelet. The mother wavelet (16.40) has an “oscillatory” character, and as we have seen the Fourier transform of $\psi_m(t)$, $\Psi_m(\omega)$, is then centred on frequency ω_m and thus provides some degree of frequency domain resolution. However, if we consider instead the *Haar wavelet* (16.1), or for simplicity the slightly modified version,

$$\psi(t) = \begin{cases} 1, & 0 \leq t \leq 1, \\ 0, & \text{otherwise} \end{cases} \quad (16.50)$$

then the situation becomes quite different. (Strictly speaking, (16.50) is not a “wavelet” since it does not integrate to zero, but it is in fact the so-called “scale function” associated with the Haar wavelet.). It is now easily seen that

$$|\Psi_m(\omega)|^2 = 2^{m+2} |D_{\frac{1}{2^{m+1}}}(\omega)|^2, \quad (16.51)$$

so that, for all m , $|\Psi_m(\omega)|^2$ has a peak solely at the origin. We now have in place of (16.49),

$$E[|a_{n,m}|^2] \sim 2\pi h_t(0), \quad t = \frac{n}{2^m}, \quad (16.52)$$

provided m is sufficiently small or negative for $|\Psi_m(\omega)|^2$ to be sufficiently concentrated around $\omega = 0$. For m large,

$$\mathbb{E}[|a_{n,m}|^2] \propto \int_{-\infty}^{\infty} h_t(\omega) d\omega = \text{total power at time } t = \frac{n}{2^m} \quad (16.53)$$

(Note that those results assume that $X(t)$ is a zero mean process.). Thus in the case of the Haar wavelet *the discrete wavelet transforms can at best estimate only the spectral ordinates at zero frequency and are totally ineffective at other frequencies.*

REFERENCES

- [CM94] S. Cambanis and E. Masry. Wavelet approximation of deterministic and random signals: convergence properties and rates. *I.E.E.E. Trans. Information Theory*, 40, 1994.
- [Dan65] H. E. Daniels. Discussion on “evolutionary spectra and non-stationary processes”. *J. Roy. Statist. Soc. (B)*, 19:1–63, 1965.
- [Dau92] I. Daubechies. *Ten lectures on wavelets*. SIAM publications, Philadelphia, 1992.
- [Mal89] S. G. Mallat. Multiresolution approximation and wavelets. *Trans. Amer. Math. Soc.*, 315:69–88, 1989.
- [NS94] G. Nason and B. Silverman. The discrete wavelet transform in s. *J. Computational and Graphical Statistics*, 3:163–191, 1994.
- [Pri81] M. B. Priestley. *Spectral Analysis of Time Series (vols I, II)*. Academic Press, London, 1981.
- [Pri88] M. B. Priestley. *Non-linear and Non-stationary Time Series Analysis*. Academic Press, London, 1988.
- [Sca94] J. D. Scargle. Wavelet methods in astronomical time series analysis. In T. Subba Rao and O. Lessie, editors, *Applications of time series analysis in astronomy and meteorology*. 1994.
- [Str93] G. Strang. Wavelet transforms versus fourier transform. *Bull. Amer. Math. Soc.*, 28:288–305, 1993.
- [Tjø76] D. Tjøstheim. Spectral generating operators for non-stationary processes. *Adv. Appl. Prob.*, 8:831–846, 1976.

Nonparametric Methods for Time Series and Dynamical Systems

D. Guégan ¹

ABSTRACT We present different approaches to reconstruct a chaotic map and to identify existence of chaos. Using nonparametric techniques, we construct - from observational data - estimates of the embedding dimension, the chaotic map, the invariant measure and the Lyapunov exponent.

17.1 Introduction

The bibliographies in papers relative to chaos show that, in the fields of economics and astronomy, many have tried different techniques to detect and identify chaos from real sample data. However, as it has been remarked, most papers favor an empirical approach of the subject and few deal with a theoretical and consistent approach. The question is to find out whether consistent tools can be developed, see SCMA I (1992) for instance.

In fact the nonparametric techniques studied in stochastic context can be used and generalized to build tools permitting the identification of chaos in real data. In this talk, we present results on this approach that can be promising if worked on any further. By way of nonparametric approach, one hopes to meet the expectations of astronomers who wish to discover the mappings which generate the apparent chaotic times series.

When we observe data, the first thing we get are the trajectories: the representation of the data with respect to time (either continuous or discrete). It is well known that it is impossible to observe chaos from such a representation, even if some nonlinearities can be detected. The second thing to do, then, be it for the economist or for the astrophysicist, is to represent the data in the state space. The difficulty is to know the true dimension of this space.

¹ENSAE-CREST, UA742, Timbre J120, 3 av. Pierre Larousse 92245 Malakoff Cedex, France. e-mail: guegan@ensae.fr

In classical chaotic systems, as the Rossler or the Henon map for instance, this dimension is known. When we have a trajectory however, this dimension called the embedding dimension is *not* known, therefore we need to estimate it. Different approaches have been developed to estimate the embedding dimension, as for instance the Grassberger - Proccacia method, but none has proved entirely satisfying so far. People resort to nonparametric inference to construct the embedding dimension. In Section 2, we propose such an estimate and we show how to apply it from Monte Carlo simulations. Applications on real data are currently being made.

Then, if we assume now that the embedding dimension is known, we can use various techniques to reconstruct the orbits in state space. At that point, chaos can be identified either analytically or geometrically, depending on the approach one chooses to favor. Using the analytical approach, the Lyapunov exponents are computed. We then try to prove their positivity, from which stems the chaoticity of the data. Using the geometrical approach, we construct the attractor linked to the chaos by way of the return map, then we compute its dimension. If the latter is fractal, we know that we have a chaos. In the present paper, we shall use the analytical approach. In Section 3 we propose different ways to reconstruct the chaotic map, and in Section 4 we show how a consistent estimator of Lyapunov exponents can be built.

When we are convinced of the presence of chaos in the data, then, we generally want to make predictions. We know then that if the chaotic function is well rebuilt, we can make short term predictions. It is near to impossible to make mean term predictions. But if we know the invariant measure of the system, we can make long term predictions. In Section 5, we discuss the problem of prediction. In Section 6 we introduce unsolved problems. By the approach that we develop in this paper, we try to have means to identify existence of chaos (with the exponents of Lyapunov), and to reconstruct the chaotic map (using the embedding dimension and the different estimates of the mapping).

17.2 Estimation of the embedding dimension

As we say in the Introduction, the first information that we need when we work on chaos is the knowledge of the embedding dimension. In this section, we present a consistent estimate of the embedding dimension using the so-called zero-one explosive nonparametric method.

We assume that X_1, \dots, X_N , are real valued observed random variables generated by the transformation

$$X_t = 3D \psi(X_{t-1}, \dots, X_{t-m_o}), \quad t \in \mathbb{Z}. \quad (2.1)$$

where $\psi : [0, 1]^{m_o} \rightarrow [0, 1]$ is measurable. The $[0, 1]$ set can be replaced by a parallelepiped in \mathbb{R}^m , but for the sake of simplicity we only consider the

$[0, 1]$ case here. We suppose that m_o is unknown and we wish to construct an estimate of m_o based on X_1, \dots, X_N . Thus, we propose another approach of that defined by Takens (1981) to reconstruct the embedding dimension. Several other approaches for estimation of the embedding dimension have also been proposed in the literature, our approach is different of those: we construct a consistent estimate of m_o using a nonparametric approach. Froehling, Crutchfield, Farmer, Packard and Shaw (1981), and Broomhead, Jones and King (1981) search for the dimension for which the reconstructed attractor can be locally approximated by its tangent. The approach of Fraser (1989) is based on the mutual information between the components of the reconstructed state vectors.

Eckmann and Ruelle (1985) consider the rate of time divergencies of orbits to determine the embedding dimension. Some authors use the predictive approach to detect m_o , see e.g., Crutchfield and McNamara (1987), Castagli (1989), Aleksic (1991), Murray (1993). Numerical approaches have also been developed, with Cenys and Pyragas (1992), Kennel and Isabelle (1992).

The probabilistic frame of our problem should be justified by ergodic assumptions on ψ stated below. Setting

$$\underline{X}_t^{m_o} = 3D(X_t, \dots, X_{t+m_o-1}), \quad t \in \mathbb{Z}$$

and using (2.1) repeatedly we may construct a mapping $\varphi : [0, 1]^{m_o} \rightarrow [0, 1]^{m_o}$ such that

$$\underline{X}_t^{m_o} = 3D\varphi(\underline{X}_{t-1}^{m_o}), \quad t \in \mathbb{Z}. \quad (2.2)$$

We assume that φ is a chaotic transformation with invariant ergodic measure μ in the classical sense defined by Ruelle (1989). It is known that for classical chaos this property is verified. For instance, the invariant measure μ of the Logistic map has the density $f(x) = 3D/(\pi\sqrt{x(1-x)})$, $0 < x < 1$; it is chaotic as soon as $\lambda > 3.57$ and it verifies the ergodic conditions for $\lambda = 3D4$. For the r -adic function the invariant measure is the Uniform measure and the ergodic assumptions are verified. For more details, see Lasota, Mac Key (1987).

Our method to estimate m_0 is based on the following alternative: if $\underline{X}_t^{m_o}$ admits a density with respect to Lebesgue measure, then

- \underline{X}_t^m has a density for any $m \leq m_o$
- The distribution of \underline{X}_t^m is singular for any $m > m_o$.

In the simple example of the r -adic function, clearly, as soon as X_0 has the Uniform measure on $[0, 1]$, the variable X_t has a density and (X_t, X_{t+1}) is concentrated on the diagonal of $[0, 1] \times [0, 1]$. Then using a nonparametric density estimate one can detect m_o as a “breaking point”. The assumptions needed to obtain a consistent estimate for m_0 can now be precisely stated. The first one concerns the existence of a density for the vector $\underline{X}_t^{m_o}$. The

second one assumes that ψ satisfies a local Lipschitz condition. The last two specify the notion of ergodicity we use throughout this paper for the fonction φ .

A_1 - $\underline{X}_t^{m_o}$ possesses a strictly positive continuous density f_t with respect to Lebesgue measure over $[0, 1]^{m_o}$.

A_2 - These exists $x_o \in (0, 1)^{m_o}$ and a positive number ℓ such that, for any $x_o \in [0, 1]^{m_o}$,

$$|\psi(x) - \psi(x_o)| \leq \ell \|x - x_o\|$$

where $\|\cdot\|$ be the sup norm in \mathbb{R}^{m_o} defined by

$$\|(u_1, \dots, u_{m_o})\| = 3D \sup_{1 \leq j \leq m_o} |u_j|$$

and let $\|\cdot\|_\infty$ be the sup norm on $\mathcal{M}([0, 1]^{m_o})$ the space of all bounded real functions defined on $[0, 1]^{m_o}$.

A_3 - There exists $\gamma_o > 0$, $c_o > 0$ and $\rho \in]0, 1[$ such that

$$|\mu(B \cap \varphi^{-t}(B)) - (\mu(B))^2| \leq c_o \rho^t, \quad t \geq 1$$

for each hypercube B in $[0, 1]^{m_o}$ satisfying $\mu(B) < \gamma_o$.

A_4 - μ has a strictly positive continuous density f on $[0, 1]^{m_o}$ and

$$\|f_t - f\|_\infty \rightarrow 0 \text{ as } t \rightarrow +\infty.$$

Note that if φ is the mapping $y = 3Drx \pmod{1}$, $0 \leq x \leq 1$ where $r \geq 2$ is an integer, then the hypotheses $(A_1) - (A_4)$ hold provided f_1 has a bounded derivative. Similarly these hypotheses hold for the bidimensional "baker transformation" (see Lasota and Mackey (1985) p.48).

We now turn to the construction of $\hat{m}_{0,N}$, the estimate of m_0 . For each integer m such that $1 \leq m \leq D$, where D is some constant to be chosen. Let us consider the density estimate

$$f_N^{(m)}(x) = 3D \frac{1}{(N-m+1) h_N^m} \sum_{t=3D1}^{N-m+1} K^{\otimes m} \left(\frac{x - \underline{X}_t^m}{h_N} \right), \quad x \in [0, 1]^m, \quad (2.3)$$

where $K = 3D \mathbb{I}_{[-\frac{1}{2}, +\frac{1}{2}]}$ can be the naive kernel or the Gaussian kernel $K(x) = 3D(2\pi)^{-1/2} \exp(-\frac{x^2}{2})$, $x \in \mathbb{R}$, and h_N is a given positive number.

The estimate $\hat{m}_{0,N}$ is defined by the following

$$\hat{m}_{0,N} = 3D \max\{m : 1 \leq m \leq D, \|f_N^m\|_\infty < a\} \quad (2.4)$$

where a is a given positive number. If $\|f_N^m\|_\infty \geq a$ for each m , we set $\hat{m}_{0,N} = 3D1$.

The sequence (h_N) and the number a have to be specified. The consistency result is the following:

Proposition 2.1: If the assumptions (A_1) – (A_4) hold and if $h_N = 3DcN^{-\delta}$ where $\delta < \frac{1}{2D}$ and where c is a positive constant, then, as $N \rightarrow \infty$

(1) If $m \leq m_o$, $\|f_N^{(m)} - f^{(m)}\|_\infty \rightarrow 0$ in probability where f^m is a m -dimensional marginal density of f .

(2) If $m > m_o$, $\|f_N^{(m)}\|_\infty \rightarrow +\infty$ in probability. Furthermore, there exists $a > 0$ such that $P(\|f_N^{(m)}\|_\infty > \frac{a}{h_n}) \rightarrow 1$ as $n \rightarrow \infty$. ■

The consistency of $\hat{m}_{0,N}$ is thus a consequence of Proposition 2.1.

Corollay 2.2: If conditions in Proposition 2.1 hold and if $a = 3Dc' \frac{N^\gamma}{\log N}$ where c' is a positive constant, then

$$\lim_{N \rightarrow \infty} P(\hat{m}_{0,N} = 3Dm_o) = 3D1. \blacksquare$$

For details on the proof, see Bosq and Guégan (1994).

In order to compute $\hat{m}_{0,N}$ it is necessary to evaluate $\|f_N^{(1)}\|_\infty, \|f_N^{(2)}\|_\infty, \dots$ until obtaining $\|f_N^{(m)}\|_\infty > a$. A practical way to detect $\hat{m}_{0,N}$ should be the use of a graphical representation of $\|f_N^{(m)}\|_\infty$ with respect to m , and to try to detect the break, as we can see in the following example.

We have simulated three random variables X, Y, Z uniformly distributed on $[0, 1]$. Using $N = 3D3000$ observations for each variable, we estimate respectively the *sup* of the joint densities for the uplets (X) , (X, Y) , (X, Y, Z) , (X, Y, Z, X^*Y) , (X, Y, Z, X^*Y, X^2) . We define a constant for each window in dimension one, then we choose the width of the windows for the different values of m . We estimate $f_N^{(m)}$ for each m and we look for the maximum of the function. In our example, we expect a burst as soon as $m = 3D4$ and we are going to see what choice of window is better to detect the jump.

One of the difficulties is the choice of initial conditions. We have taken the observations which maximize the estimate $\hat{m}_{0,N}$ as initial conditions. The various choices made for h_N are:

- $h_{N,m}^1 = 3Dh_{N,1}$: constant window.
- $h_{N,m}^2 = 3DN^{1-1/m} h_{N,1}$: in dimension one, we associate the probability $p_1 = 3D\frac{1}{N}$ to each observation, in dimension m , we associate the probability $p_m = 3D\frac{1}{N^m}$ to them. Then the number of observations in windows $h_{N,1}$ and $h_{N,m}$ is respectively $N_1 = 3D\frac{h_{N,1}}{N}$ and $N_m = 3D\frac{h_{N,m}^m}{N^m}$. We equalize these two numbers to have a window which has the same quality.
- $h_{N,m}^3 = 3D(\frac{C}{m \cdot N(2\sqrt{\pi})^m})^{\frac{1}{4+m}}$: this bandwidth has been built using the optimal window with the MISE as it is defined in the classical nonparametric regression estimate. C is a constant to be chosen.

In the array below, for the four uplets that we have introduced before, we give the values that we obtain for the *sup* of the estimate $f_N^{(m)}$ defined

in (2.3). In each column, we indicate these values for the different values of the three windows we have chosen. The rows correspond to the four different uplets that we have considered. For each case, we have simulated 3000 sample size.

Uplets	$h_{N.m}^1$	$h_{N.m}^2$	$h_{N.m}^3$
(X)	1,05	1,05	1,06
(X, Y)	1,27	0,97	1,34
(X, Y, Z)	1,99	0,56	1,89
(X, Y, Z, X^*Y)	14,32	0,26	8,89
(X, Y, Z, X^*Y, X^2)	107,61	0,11	36,53

We can see that the first window $h_{N.m}^1$ and the third window $h_{N.m}^3$ detect the correlation between the variables for the uplets (X, Y, Z, X^*Y) and (X, Y, Z, X^*Y, X^2) . The jump is not so important with the second one. This example shows how the burst can be detected. Other examples can be found in Guégan and Léorat (1995a).

17.3 Estimation of the function φ

When the embedding dimension is known, another important problem concern the reconstruction of the function φ as it is defined in (2.2). Different nonparametric methods can be used to reconstruct the chaotic map. We present them here.

For sake of simplicity, we use now the following notations and we assume that we observe X_1, \dots, X_N , with $X_i \in \mathbb{R}^m, i = 3D1, \dots, N$, which are generated by a system as

$$X_t = 3D\varphi(X_{t-1}), \quad (3.1)$$

where φ is some map from $D \rightarrow D$, where D is some closed subset of \mathbb{R}^m . We assume that φ is ergodic in the sense defined previously. We present estimates for the reconstruction of φ . We propose four methods for the estimation of the chaotic function, two based on kernel nonparametric method, two based on the nearest neighbors method.

1. The zero-one “explosive” method.

The functional relationship (3.1) induces that the joint distribution $P_{t,t-1}$ of (X_t, X_{t-1}) is singular with respect to the Lebesque measure λ , thus a suitable density estimate for $P_{t,t-1}$ will explode in a neighborhood of the graph of φ and will vanish elsewhere. So to obtain an estimate $\hat{\varphi}_N^1$ of φ , we set

$$g_N(x, y) = 3D \frac{1}{Nh_N^{2m}} \sum_{t=3D1}^{N-1} K\left(\frac{x - X_t}{h_N}\right) K\left(\frac{y - X_{t-1}}{h_N}\right), \quad (3.2)$$

where $(x, y) \in \mathbb{R}^{2m}$, K is a kernel bounded density function of \mathbb{R}^m , and (h_N) is a sequence which tends to zero as $N \rightarrow \infty$. Then an estimate $\hat{\varphi}_N^1(x)$ of $\varphi(x)$ will be defined by:

$$g_N(x, \hat{\varphi}_N^1(x)) \geq \sup_{y \in \mathbb{R}^m} g_N(x, y) - \zeta_N \quad (3.3)$$

where $\zeta_N \rightarrow 0^+$ as $N \rightarrow \infty$.

Theorem 3.1: Under classical ergodicity conditions for φ , we have

$$h_N^{-p}[\hat{\varphi}_N^1(x) - \varphi(x)] \rightarrow 0, \text{ in probability, for each } p < 1. \blacksquare$$

The interest of this estimate is its robustness in particular if we add a noise to the system (3.1). We will come back on this discussion latter.

2. The regressogram method.

Another approach consists to interprete φ as a regression function, then a natural estimate should be, for $m = 3D1$,

$$\hat{\varphi}_N^2(x) = 3D \frac{\sum_{t=3D1}^{N-1} X_{t-1} K(\frac{x-X_t}{h_N})}{\sum_{t=3D1}^N K(\frac{x-X_t}{h_N})}$$

where $x \in \mathbb{R}$, K is a strictly positive kernel function, and where the sequence (h_N) tends to zero, as $N \rightarrow \infty$. We can choose for $K = 3D_J$, where $J = 3D[x - h_N, x + h_N]$ or the Gaussian Kernel, then we get,

Theorem 3.2: If φ is locally Lipschitz and if the system (3.1) is uniformly ergodic then,

$$\hat{\varphi}_N^2(x) \rightarrow \varphi(x),$$

for almost all x in \mathbb{R} , as $N \rightarrow \infty$. \blacksquare

We present now the two other estimates for φ based on the nearest neighbors method. We denote by i^n , i^ℓ and i^r , the indexes of the three elements of the observed set of data which are respectively the nearest neighbor of x , the left nearest neighbor of x and the right nearest neighbor of x . Now we assume that $m = 3D1$.

3. The nearest neighbor method:

Let be x a point of \mathbb{R} , then we define the estimate $\hat{\varphi}_N^3(x)$ for φ as:

$$\hat{\varphi}_N^3(x) = 3D X_{i^n+1}.$$

Theorem 3.3: If φ is locally Lipschitz

$$\hat{\varphi}_N^3(x) \rightarrow \varphi(x),$$

for almost all x in \mathbb{R} , as $N \rightarrow \infty$. \blacksquare

4. The two nearest neighbors method:

Let be $x \in \mathbb{R}$. We define the estimate $\hat{\varphi}_N^4(x)$ of $\varphi(x)$ by:

$$\hat{\varphi}_N^4(x) = 3D (x - X_{i^\ell}) \frac{X_{i^r+1} - X_{i^\ell+1}}{X_{i^r} - X_{i^\ell}} + X_{i^\ell+1}$$

Theorem 3.4: If φ is locally Lipschitz

$$\hat{\varphi}_N^4(x) \rightarrow \varphi(x).$$

for almost all x in \mathbb{R} , as $N \rightarrow \infty$. ■

The first two estimates considered here, $\hat{\varphi}_N^1(x)$ and $\hat{\varphi}_N^2(x)$ are based on the kernel method, and the last two, $\hat{\varphi}_N^3(x)$ and $\hat{\varphi}_N^4(x)$ on the nearest neighbours method. The first one assumes that the observations are random variables, and that X_0 has a density. We do not use this framework in the three other methods, nor do we assume that X_0, X_1, \dots are random variables, see Guégan (1994a). To prove the consistency of the four estimates considered here, note that there exists a great difference between these four methods. If we have a convergence rate in the assumptions concerning the ergodicity of φ , this permits to obtain a convergence rate for the estimate $\hat{\varphi}_N$. We use this approach when we study the estimate $\hat{\varphi}_N^1(x)$ for instance, see Bosq and Guégan (1995). This last hypothesis is verified in particular for the r-adic function. In the regressogram approach, see Delecroix, Guégan and Léorat (1995), we assume that the system is uniformly ergodic and that the function φ is locally Lipschitz. In the nearest neighbours method we assume that the process is ergodic and that φ is locally Lipschitz, see Delecroix, Guégan and Léorat (1994). In the last three approaches it is not possible to obtain the rate of convergence. The first approach is essentially useful for the theoretical computation of the embedding dimension as we have seen in the previous section. the second one permits to obtain a good estimate for the chaotic function and the last two ones to obtain easy estimates for the Lyapunov exponents in one dimensional setting.

We can see that we use the same kinds of estimates than in the stochastic approach, see Chen and Härle (1995). The great difference is the kinds of hypotheses we use. They need to be realistic for the model we want to develop. As in model (3.1) we consider there is no noise, we cannot use the same kind of proofs than in the stochastic case. It is very difficult in that case to obtain a rate of convergence and in any case convergence in law for the estimates. But, if we use strong hypotheses as in Bosq and Guégan (1995) where we explicitly define a rate of convergence in the ergodicity hypothesis, then we can obtain a rate of convergence for the estimate. The problem is to be able to test this kind of hypotheses on real data. With the other methods we cannot compute bias and variance for the estimates, as people classically do with the nonparametric approach, but these methods are more easy to implement, see Guégan (1995) for more details.

17.4 Estimation of the Lyapunov exponent

The estimation of the function φ can be used to obtain an estimate of the Lyapunov exponent. Indeed this last quantity permits to measure the sen-

sitivity of the system to initial conditions. In dimension one, the Lyapunov exponents is the real number λ defined by:

$$\lambda = 3D \lim_{n \rightarrow \infty} \frac{1}{n} \log \left| \frac{d}{dx} \varphi^{(n)}(x) \right|. \quad (4.1)$$

Here $\varphi^{(n)}$ represents the n^{th} iterate of the function φ . When the dynamical system is ergodic, then λ is well defined independently of the point $x \in \mathbb{R}$, see Ruelle (1989). If λ is strictly positive, the system is sensitive to initial conditions; that is, two trajectories starting from two initial values very closed to each other, can diverge exponentially until they are not at all similar.

The goal of this section is to estimate the Lyapunov exponent in an ergodic dynamical system. To obtain an estimate for the Lyapunov exponents, we can use the definition (4.1). This definition is equivalent to

$$\lambda = 3D \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=3D1}^{n-1} \log |\varphi'(X_i)| \quad (4.2)$$

where X_i describes all the trajectory.

In the first case, we need an estimate of $\varphi^{(n)'}(x)$ and in the second case we can use the estimate of φ' . Here we are going to use the definition (4.1) and we assume that $m = 3D1$. We proceed in the following way. Let be a sequence of integers increasing to infinity denoted $p(N)$, x be a point of D , a compact subset of \mathbb{R} . Then using the previous notations, we can consider the left and the right nearest neighbors of x among $X_1, \dots, X_{N-p(N)}$. An estimate $\hat{\varphi}^{p(N)'}$ of $\varphi^{p(N)'}$ will be, for $x \in D$

$$\hat{\varphi}^{p(N)'}(x) = 3D \frac{X_{i^r+p(N)} - X_{i^\ell+p(N)}}{X_{i^r} - X_{i^\ell}}$$

And the estimate of λ is

$$\hat{\lambda}_N^{(1)} = 3D \frac{1}{p(N)} \log |(\hat{\varphi}^{p(N)})'(x)|, \quad (4.3)$$

for all $x \in D$.

Theorem 4.1: If the assumptions (A_2) - (A_4) are satisfied, the estimate $\hat{\lambda}_N^{(1)}$ tends to λ , for every x such that the following assumption (A_5) is verified:

(A_3) : $\forall \varepsilon > 0$, $\exists V_x^\varepsilon = 3D[x - \varepsilon, x + \varepsilon]$ such that $(\varphi^{(k)})'(y) \leq M$, $\forall k \geq 1$, $\forall y \in V_x^\varepsilon$, where M is some constant. ■

In case of the logistic map, $a = 3D4$. For $n = 3D5000$ and $p(N) = 3D \log N$, we have computed $\hat{\lambda}_N^{(1)}$ for different values of x . The results are the following:

x	$\hat{\lambda}_N^{(1)}$
0.1	0.70878869
0.2	0.67450682
0.3	0.66004223
0.4	0.64949465
0.5	0.21654712
0.6	0.64224023
0.7	0.66049692
0.8	0.67537244
0.9	0.70500775

The expected value for λ is 0.69. When $x = 3D0.5$, this corresponds to a fixed point and we will expect 0 for the limit. We can also investigate the importance of the choice of $p(N)$ for the logistic map with $a = 3D4$, $\hat{\lambda}^{(1)}$ in function of $p(n)$, see Guégan, Léorat (1995b) for other applications.

17.5 Predictions

Prediction has been extensively developed for chaotic systems. As Lillekjendlie, Kugiumtzis and Christoffersen (1994) recall in their article, different groups of methods can be considered to predict series: the local, the semi-local and the global methods. The local approach has been developed and discussed in the papers of Farmer and Sidorowich (1987), Jimenez, Moreno and Ruggieri (1992), Murray (1994), Doerner, Hubinger and Martienssen (1994), Finkenstädt and Kuhbier (1995), among others. The semi-local approach, in the papers of Smith (1992), Stockbro and Umbberger (1992) and the global approach, for instance by Casdagli (1989), Lebaron and Weigend (1995), Fang and Cao (1995). Other methods have also been developed using marginal densities, see for example, El Gamal (1987), Geweke (1988). Experimental algorithms have been proposed in the works of Sugihara and May (1990) and Meyer and Packard (1991).

Then most of these nonparametric or semiparametric methods can be used to make predictions on chaos. Indeed, when the reconstruction of the function φ is correct, it is possible to obtain short term predictions. If we assume that we observe data generated by a system as

$$X_t = 3D\varphi(X_{t-1}, \dots, X_{t-m})$$

when X_{t-1}, \dots, X_{t-m} are known, we can expect to forecast X_t, \dots, X_{t+p} where the horizon of lag of prediction p is small (same order as m). We have already present consistent nonparametric estimates of the function φ in the previous section that we can use to make short term predictions.

Other approaches, which are more empirical, can also be used to make predictions. We briefly present them.

For example, we can try to develop an approximation of the chaotic function by global polynomials. For that we define:

$$\hat{\varphi}_N^5(X_t) = 3D \sum_i w_i f_i(X_t)$$

where w_i are parameters and the basis functions f_i are powers and cross products of the components X_t . The difficulty with this method concerns the choice of the functions f_i . On the other hand, the noise, if it exists, can have a great influence on the quality of the estimation.

Another class of global methods which can be considered is the multi-layer perceptron neural networks. In that case the estimate is defined as:

$$\hat{\varphi}_N^6(X_t) = 3Df\left(\sum_{i_N} w_{i,N} f\left(\sum_{i_{N-1}} w_{i_{N-1}} \varphi(X_t)\right)\right),$$

where the sum on i_N concerns the number of nodes of the network N and the sum on i_{N-1} concerns the number of nodes of the network $N - 1$, where the $w_{i,N}$ are real valued parameters called the weights, and where the function f is usually chosen as the sigmoid shaped basis function. This method is more robust than the previous one, but is time consuming because of the learning process.

We can also consider semi-global methods using spline or radial basis functions, then the estimate for φ is defined in the following way:

$$\hat{\varphi}_N^7(X_t) = 3D \sum_{i=3D1}^{\zeta} w_i f_i(\|X_t - \zeta^i\|)$$

where w_i are weights to be chosen, the functions f_i are radially symmetric around a center value ζ_i . All these methods have been compared for different known systems and apply on real data in a recent work, see Guégan and Mercier (1996). However, the future of a nonlinear system in the chaotic phase is not predictable for a mean term time interval because forecasting errors grow exponentially fast.

We can say that we make a long term prediction when the prediction horizon is bigger than several pseudo-periods (when pseudo-periods can be defined). In that case, providing the historical data is long enough, we can compute the density of the prediction if we know the invariant measure. In order to compute long term predictions, we present now a manner to estimate the invariant measure using a kernel method.

We denote by f the density of the invariant measure μ , and f_N an estimate of f defined by

$$f_N(x) = 3D \frac{1}{Nh_N^{2m}} \sum_{t=3D0}^{N-1} \mathbb{I}_{B_o}\left(\frac{x - X_t}{h_N}\right), \quad x \in \mathbb{R}^m$$

where $B_o = 3D] - \frac{1}{2}, \frac{1}{2}[^k$. Then the theorem we get is the following:

Theorem 5.1 Under the ergodic hypothesis. we get

$$E[f_N(x) - f(x)]^2 = 3D0\left(\frac{\log N}{N}\right)^{4/k+4}, \quad x \in I\!\!R^m$$

if $h_N \sim (\frac{\log N}{N})^{1/k+4}$. ■

For details on the proof, see Bosq and Guégan (1995).

17.6 Conclusion

In this paper, we have presented new results concerning tools permitting the study of chaos. These tools constructed in a nonparametric context provide estimates of the embedding dimension, the chaotic map, the Lyapunov exponents and the invariant measure. Most of the tools used here have been developed historically in a stochastic context. We adapt them in various situations that we encounter here, and prove their consistency in a deterministic framework which corresponds to the dynamical system. There exists different points of view concerning the investigation of chaos in the statistician population. Basically when we observe data we can consider the two following situations:

- The observational data set X_1, X_2, \dots, X_N is generated by

$$X_t = 3D \varphi(X_{t-1}) + \varepsilon_t. \quad (6.1)$$

where X_t and ε_t are random variables.

- The observational data set Y_1, Y_2, \dots, Y_N is generated by

$$Y_t = 3DX_t + \varepsilon_t \quad (6.2)$$

where

$$X_t = 3D \varphi(X_{t-1}). \quad (6.3)$$

Y_t and ε_t are random variables and X_t is deterministic.

In the first case, we are in a classical stochastic context and this concerns works as they are developed, for instance, in the books of Tong (1991) and Guégan (1994). In the second case, we are in a deterministic context with a kind of measure noise. All the difficulty in that latter context is to separate correctly the noise from the state. This problem is always non solved. Here, we have considered a very simple representation, (6.2)- (6.3), because there is no reason to have additive noise in this context. The situation can be more complex and we certainly have not a unique decomposition. Consequently it is important to be sure that the estimates that we propose in the previous

sections are robust in presence of noise. These ideas have begun to be developed in some empirical works, but the theory has to be finished (see Guégan 1994a; Guégan and Léorat 1995b).

Another problem is the existence of correct definitions of the different characteristics of chaos when there is noise. We point, for instance, to the notions of state space, Lyapunov exponents and dimension. Castagli, Eubank, Farmer and Gibson (1991) give an interesting discussion on the possible procedures to reconstruct the state space in presence of noise, and they give a particular example with the Ikeda map. Spiro (1993) proposes a method to determine the level of dynamical noise inherent in a system using proportions of observations within spheres of radius r , in different embedding dimensions. Some modified versions of the algorithm has been proposed in view to reduce the noise level (see Dvorak and Klaschka 1990; Fraedring and Wang 1993).

If we assume that the reconstruction of the state space is satisfactory, the problem now is to determine exactly what happens concerning the two fundamental characteristics of chaos when there is noise. In effect, most of the methods to identify chaos require a very large sample and almost noise free data set in order to produce a reliable estimate of the estimations of those previous quantities. We have to be careful when we work with a model such as (6.1). If (6.2) is a pure deterministic model, close to the “true” physical model, model (6.1) becomes more like a statistical model. In that case, the transformation φ now has the status of being the conditional mean given the past, instead of being a physical law. One may therefore question the relevance of studying the “chaotic” properties, such as dimension or Lyapunov exponents of this map. It seems that the characteristics of chaos are relevant when noise is small (Jensen 1993). But even if the notions of dimension of an attractor and the Lyapunov exponents can be generalized to random transformations, it is important to remark that they have not the same meaning as for chaotic systems. When working with noise, we have to treat systems in terms of conditional probability density function, and this is not possible with the deterministic approach (see Guégan 1996 for more detailed discussion).

REFERENCES

- [1] Z. Aleksic (1991): “Estimating the embedding dimension”, *Physica D*, 52, 362-368.
- [2] D. Bosq, D. Guégan (1994): “Estimation of the embedding dimension of a dynamical system. $n = B0\ 9451$ Preprint INSEE. Paris .
- [3] D. Bosq, D. Guégan (1995): “Nonparametric estimation of the chaotic function and the invariant measure of a dynamical system”. *Statistic and Probability Letters*, 25, 201 -212.
- [4] D. S. Broomhead, G. P. King (1986): “Extracting qualitative dynamics from experimental data”, *Physica D*, 20, 217-236.

- [5] M. Castagli (1989): "Nonlinear prediction of chaotic time series", *Physica D*, 35, 335-356.
- [6] M. Castagli, S. Eubank, J. Farmer, J. Gibson (1991): " State space reconstruction in the presence of noise", *Physica D*, 51, 52-98.
- [7] A. Cenys, K. Pyragas (1988): "Estimation of the number of degrees of freedom from chaotic time series", *Physics Letters A*, 129, 227-230.
- [8] R. Chen, W. Hardle (1995): "Nonparametric time series analysis: a selective review with examples", in *Proceedings of the 50th ISI Conference in Beijing*, IP 10, 375-394.
- [9] M. Delecroix, D. Guégan, G. Léorat (1994): "Detecting deterministic chaos from observational data. $n = B09450$ ". Preprint INSEE. Paris.
- [10] M. Delecroix, D. Guégan, G. Léorat (1996): "Detecting deterministic chaos using regressogram". In preparation.
- [11] J. P. Eckmann, D. Ruelle (1985): "Ergodic theory of chaos and strange attractors", *Reviews of Modern Physics*, 57, 617-656.
- [12] M. A. El Gamal (1987): "Simple estimationn and forecasting methods in systems characterized by deterministic chaos: the univariate case", Preprint.
- [13] H. P. Fang, L. Y. Cao (1995): "Predicting and charaterizing data sequences from structure variable systems", Preprint.
- [14] J. D. Farmer, V. Sidorovich (1988): "Exploiting chaos tro predict the future-and reduce noise", in *Evolution, Learning and Cognition*, eds. Lee, World Scientific Press.
- [15] B. Finkenstadt, P. Kuhbier 51995): "Forecasting nonlinear economic time series: a simple test to accompany the nearest neighbor method", *Empirical economics*, 20, 243-263.
- [16] A. M. Fraser, H. L. Swinney (1986): "Indépendant coordinates for strange attractors frrom mutual information", *Phys. Review. A*.33, 1134-1140.
- [17] J. Geweke (1988): "Inference and forecasting for deterministic nonlinear time series observed with measurement error", Preprint.
- [18] D. Guégan (1994a): "Deterministic versus Stochastic Systems". Preprint n°9438, INSEE.
- [19] D. Guégan (1994b): "Nonparametric estimation in chaotic deterministic system", to appear in the *Proceedings of the International Conference on "Chaos and dynamical systems"*, Tokyo, May 1994.
- [20] D. Guégan (1994 c): "Séries chronologiques non linéaires temps discret. Economica.
- [21] D. Guégan (1995): "Invariance in stochastic and Deterministic Systems", in the *Proceedings of the 50th ISI Conference*, Beijing, IP 10, 357-373..
- [22] D. Guégan (1996): "How can noise be brought out in dynamical system", in the proceedings of the *Multidisciplinary International Conference in Matra-fured*, Hungary, Sept. 1995.
- [23] D. Guégan, G. Léorat (1995a): "Consistent estimates to detect chaos in financial data", Preprint Paris XIII $n = B0 95-07$.
- [24] D. Guégan, G. Léorat (1995b): "What is the good identification theory for noisy chaos: an empirical approach", Preprint.
- [25] D. Guégan, L. Mercier (1996): "Rising and Falling Predictions in Intra-Day financial data", in the *Proceedings of the International Forecasting Financial Markets Conference organized by the Chemical Bank*, London.

- [26] J. Jimenez, J. A. Moreno, G. J. Ruggieri (1992): “Forecasting in chaotic time series: a local optimal linear reconstruction method”, Physical review A, 45, 3553-3558.
- [27] J. Lasota, O. Mac Key (1987): “Probabilistic properties of deterministic system”. Cambridge Ltd.
- [28] B. Lillekjendlie, D. Kugiumtzis, N. Christoffersen (1994): “Chaotic time series: system identification and prediction”, Preprint.
- [29] T. P. Meyer, N. H. Packard (1991): “Local forecasting of high dimensional chaotic dynamics”, Preprint.
- [30] D. B. Murray (1993): ”Forecasting a chaotic time series using an improved metric for embedding space”, Physica D, 68, 318-325.
- [31] D. Ruelle (1989): “Chaotic motion and strange attractors”, Cambridge Univ. Press.
- [32] SCMA I (1992), E. D. Feigelson, G. J. Babu, Statistical Challenges in Modern Astronomy, Springer Verlag.
- [33] G. C. Spiro (1993): “Measuring dynamical noise in dynamical systems”. Physica D, 65, 289-299.
- [34] K. Stokbro, D. K. Umberger (1992): “Forecasting with weighed maps”, in Nonlinear Modeling and Forecasting, Proc. Vol. XII, eds. M. Castagli and S. Eubank, Addison-Wesley.
- [35] F. Takens (1981): “Detecting strange attractors in turbulence, in Lecture Notes in Mathematics, Vol. 898. D. A. Rand and L. S. Young, Eds. Springer Verlag, 366.
- [36] H. Tong (1991): “Nonlinear Time Series: A Dynamical System Approach”, Oxford science Publications.

Discussion by Jeffrey D. Scargle

What is a nonlinear process?

Nonlinear and nonstationary time series have become important topics in applied data analysis – see [Pr88, To90] for example. It seems to me that the term “nonlinear time series” has been used so loosely that its true meaning has become somewhat obscured. In ordinary usage *nonlinear* is a property of physical *systems* in which there is an *input* and an *output*. The output is linear in the input: the sum of two inputs yields as output the sum of the corresponding outputs; if the input is multiplied by a constant, so is the output.

Clearly this concept simply cannot be applied to a time series, but only to a mathematical model of the underlying random process. One defines in effect two classes of models, linear and nonlinear. A nonlinear time series is then one that in some sense conforms to a model from the latter class.

Almost universally, the way in which this is done is to model the dependence of the current value of the process to its previous history as an input/output relation: past values = input; current value = output. The

autoregressive model, *e.g.*, explicitly expresses the X_n as a linear form in the previous values $X_{n-1}, X_{n-2}, X_{n-3}, \dots$

A good example of this procedure is Priestly's [Pr88] approach. He considers ordinary causal moving average models (equivalent to corresponding autoregressive models), in which X_n is linear in current and previous values of an input white noise process, $R_{n-1}, R_{n-2}, R_{n-3}, \dots$. He then defines an extension of this class with a Volterra series expansion, involving nonlinear terms in the input noise – *e.g.* $R_{n-k}R_{n-j}$, $R_{n-k}R_{n-j}R_{n-l}$, *etc.* Using a dynamical systems approach, Tong [To90] defines a nonlinear class of autoregressive models in which the current value of the process is in general a nonlinear function of previous values.

Thus the accepted concept of nonlinear time series depends on establishing at least three things:

- models with input/output structure
- linear and nonlinear classes of such models
- criteria for associating time series data with one of these classes

The point I am trying to make is that there is no universal concept of time series linearity – and that previous work has invoked special, somewhat arbitrary choices for the three items listed here. There are time series models that are not in the form of an input/output relation at all. And there are processes in which the independent variable is not time, so that the causality associated with representing the current value of the process with its previous history loses significance, and for which – at the very least – the structure of the models classes has to be reworked.

In a very interesting paper, Bickel and Bühlmann [Bi96] pour even more cold water on the prospects for a universal concept of linearity (and hence nonlinearity) for time series. They adopt essentially Priestly's classes of models, and then show that even in the context of such a clearly defined concept of nonlinearity, it is impossible to test whether a given set of observations were generated by a “linear” process.

This is not to say that one should give up. There are many situations in which it is meaningful to try to understand whether the underlying process is linear or not. One just has to be careful about one means by even this, and to assess data analysis methods that purport to be “nonlinearity meters.”

Progress in nonlinear modeling

At the previous *Statistical Challenges in Modern Astronomy* (hereafter referred to as SCMA I) I reviewed techniques for detecting and modeling chaotic processes using time series and summarized the search for chaos in astronomical data. The concluding section contained these somewhat pessimistic remarks:

“ ... (1) typical astronomical time series are relatively short and noisy compared to what is needed for a good detection and analysis of chaotic processes; (2) this problem is compounded by the fact that the most easily computed numbers – dimension in particular – unfortunately do not provide conditions sufficient for the presence of chaos.”

In an effort to be optimistic, I punctuated the above with

“Almost all astronomical discoveries are preceded by a period of mere theoretical possibilities. Neutron stars and gravitational lenses are examples, as someday black holes may be. Perhaps a future addition to this list will be chaotic processes pervading the Universe.”

Some progress has been made in the years since SCMA I. Roughly a year after that meeting, a NATO Advanced Research Workshop on Comparative Time Series Analysis was held in Santa Fe, New Mexico, and the proceedings [WG94] are a good summary of the state of the art – at that time – of prediction based on nonlinear models. Over the years, many workshops and publications at the Institute for Mathematics and its Applications (see *e.g.* [Br92]) and at the Santa Fe Institute have dealt directly or indirectly with nonlinear phenomena and data analysis (see *e.g.*, [SR91]). A group mainly at Los Alamos has published a number of papers on nonlinear modeling of chaotic time series [Ca92, Th92], and have in particular developed the idea of “surrogate data” as a way of verifying time series data-analytic conclusions.

More recently (30 November - 2 December 1995), the Eleventh Annual Florida Workshop in Nonlinear Astronomy “NONLINEAR SIGNAL AND IMAGE ANALYSIS,” took place, and the many interesting contributions will soon be published by the New York Academy of Science.

Wavelets and related methods (see my main contribution to this volume) are general purpose tools that may have application in the area of identification and characterization of nonlinear dynamical processes. There are several repositories of current work on nonlinear dynamics and related data analysis methods on the World Wide Web. These can easily be found with any of the many “search engines.” *Nonlinear Dynamics and Topological Time Series Analysis Archive* at

<http://cnls-www.lanl.gov/nbt/intro.html>

the Springer-Verlag journal *Nonlinear Science Today*, at

<http://www.springer-ny.com/nst/>

and *Neural Nets, Autonomous Agents, Fuzzy Systems, Time Series Analysis, and Computational Biology* at

<http://www-psych.nmsu.edu/linda/complex/nnaafuzz.htm>

are examples.

Importance of Guégan's work

I do not think the fundamental data-related problems lamented above have been effectively solved. Thus the importance of Dominique Guégan's contribution here is that it represents a novel approach that may well meet the pressing need for new and better methods to detect and characterize chaotic dynamics from time series data.

This context of the method is the standard state-space embedding arena. What is novel is the "zero-one explosive" method applied to a direct estimate of the probability density, or "invariant measure," in order to identify the system's chaotic map and its properties.

In particular, the fact that the approach is nonparametric should make the method applicable in a wide range of astrophysical contexts. Of course, this does not mean that there are no adjustable parameters of the method – there are in fact several: the window sequence h_N , and the sequence of "explosion" thresholds a_N . It does mean, however, that one does not need to have a full model of a postulated dynamical system.

Furthermore, Guégan's methods will hopefully shed light on the connection between nonlinear dynamics and *scaling behavior* – which is becoming increasing important in both theoretical and observational astronomy.

REFERENCES

- [Bi96] Bickel, Peter J., and Bühlmann. Peter, (1996), "What is a Linear Process," preprint, to appear in the Proceedings of the National Academy of Sciences.
- [Br92] *New Directions in Time Series Analysis, Part I*, eds. D. Brillinger, P. Caines, J. Geweke, E. Parzen, M. Rosenblatt, M. Taqqu, Springer-Verlag: New York.
- [Ca92] Casdagli, M., Des Jardins, D., Eubank, S., Farmer, J.D., Gibson, J., Hunter, N. and Theiler, J. (1992) "Nonlinear Modeling of Chaotic Time Series: Theory and Applications" *Applied Chaos*, ed. Jong Kim. (Addison-Wesley) pp. 335-383.
- [Pr88] Preistly, M.B., (1988) *Non-Linear and Non- Stationary Time Series Analysis*, Academic Press: London.
- [SR91] Subba Rao, T., (1991), "Analysis of Nonlinear Time Series (and Chaos) by Bispectral Methods," in *Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity*, Proc. Vol. XII, Eds. M. Casdagli and S. Eubank, Addison-Wesley. 1991.
- [Th92] Theiler, J. Lindsay, P. S., and Rubin, D. M. (1992): "Detecting Nonlinearity in Data with Long Coherence Times" Proc. of NATO Workshop on Comparative Time Series Analysis, Santa Fe Institute, 14-17 May, 1992.
- [To90] Tong, H. (1990): *Non-Linear Time Series, A Dynamical System Approach*, Oxford University Press: Oxford.
- [WG94] Weigend, A., and Gershenfeld, N. (1994), *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison-Wesley Publishing Co., Reading, MA

Quantifying Rapid Variability in Accreting Compact Objects

M. van der Klis

ABSTRACT I discuss some practical aspects of the analysis of millisecond time variability X-ray data obtained from accreting neutron stars and black holes. First I give an account of the statistical methods that are at present commonly applied in this field. These are mostly based on Fourier techniques. To a large extent these methods work well: they give astronomers the answers they need. Then I discuss a number of statistical questions that astronomers don't really know how to solve properly and that statisticians may have ideas about. These questions have to do with the highest and the lowest frequency ranges accessible in the Fourier analysis: how do you determine the shortest time scale present in the variability, how do you measure steep low-frequency noise. The point is stressed that in order for any method that resolves these issues to become popular, it is necessary to retain the capabilities the current methods already have in quantifying the complex, concurrent variability processes characteristic of accreting neutron stars and black holes.

18.1 Introduction

The purpose of this talk is to explain to statisticians how astrophysicists, mostly using Fourier transform techniques, go about analyzing X-ray time-series data obtained from accreting compact objects (neutron stars and black holes), and to point out a few problems with the usual approaches. The point will be made, that the conglomerate of statistical methods that is being applied in this branch of high-energy astrophysics, even though most definitely not always rigorous, on the whole serves its purpose well and is providing astronomers with the quantitative answers they require. This talk will aim, however, at a few areas where we run into problems, and where more statistical expertise might help. In thinking about how to solve the problems that I shall outline, it will be important to keep an eye on what the capabilities of the existing methods are, as those capabilities will need to be preserved in whatever new approach one would like to propose.

Mostly, accreting neutron stars and black holes occur in double star systems known as X-ray binary stars, where a normal star and the compact

object are in a close orbit around each other (Fig. 1). Matter flows from the normal star to the compact object by way of a flat, spiraling flow called an accretion disk, and finally accretes onto the compact object. A large amount of energy is released in this process (typically 10^{36} to 10^{38} erg/sec), and is emitted, mostly in the form of X-rays. The characteristic variability timescale for the X-ray emitting regions is predicted (and observed) to be very short (less than a millisecond). By studying the properties of this rapid X-ray variability it is possible to extract a great deal of information about the flow of matter onto the compact object, and, indirectly, about the object itself. See van der Klis (1995) for a recent review of the results of studies of this type.

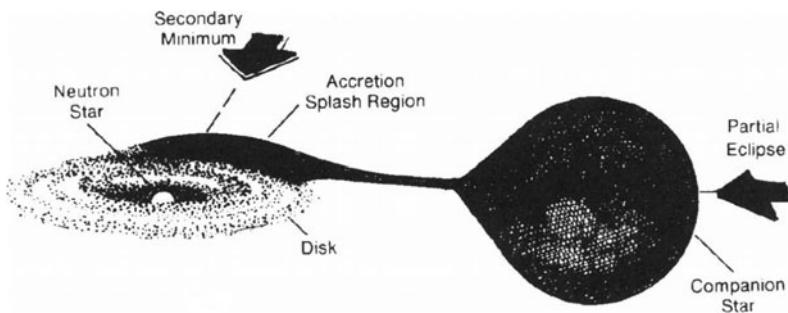


FIGURE 1. X-ray binary star.

18.2 The data

In order to understand the character of the data we are dealing with, I shall follow the flow of information from star to computer. An X-ray binary star emits X-ray photons at a very large rate (say, 10^{46} photons/sec). For all practical purposes, the X-ray photon rate produced by the star can be considered as a continuous function of time $I(t)$. These photons are emitted isotropically, or at least over a solid angle of order 4π sterad. Because the X-ray detector onboard the X-ray satellite spans only a very small solid angle as seen from the X-ray star, only a very small fraction ϵ (say, 10^{-43}) of the photons is detected in the instrument.

The time series of photon arrival times $t_i, i = 1, \dots, N_{phot}$ (with N_{phot} the total number of detected photons) is the information that, ideally, we would like to have available for analysis. However, typically, instrumental limitations (and the maximum telemetry rate) prevent the registration and transmission of all these arrival times. Therefore, onboard the satellite,

the data are *binned* into equidistant time bins. The information that is finally telemetered to the ground station consists of a sequence of numbers $x_m, m = 0, \dots, N_{tot} - 1$, where N_{tot} is the total number of time bins, x_m representing the number of photons that was detected during time bin m . In most cases, and unlike the usual case in astronomy, the time bins are equidistant and contiguous (no gaps). The statistical problem facing us can be summarized as follows: “*Given $x_m, m = 0, \dots, N_{tot} - 1$, reconstruct as much as possible about $I(t)$.*”

We shall assume that for the bright X-ray binaries that I am discussing here, the rate of background photons can be considered to be negligible, and that there are no relevant effects affecting the photon arrival time series other than the huge geometrical dilution factor just described.¹ Therefore, if the X-ray star does not vary intrinsically, the t_i will to a high degree of precision be uniformly and randomly distributed, so that the x_k follow Poisson statistics appropriate to a rate $\mu = \langle x_m \rangle$, with a standard deviation σ_{x_m} equal to $\sqrt{\mu}$.

Of course these stars *do* vary. The variability time scales of interest are ≤ 1 millisecond, and the detected photon rates 10^2 - 10^5 photons/sec, which means that the time bins must be chosen such that typically there are on average only a few (sometimes $\ll 1$) photons per time bin, and σ_{x_m} is of order μ (sometimes $\sigma_{x_m} \gg \mu$). As the intrinsic variability of the star often has a relatively small amplitude, only a few percent of the total flux, it is clear that we are in a low signal-to-noise regime (with the “signal” the intrinsic stellar variability and the “noise” the Poisson fluctuations). Fortunately, typical observations span 10^3 to 10^5 sec, so that the number of time bins N_{tot} (the number of “measurements”) is 10^6 to 10^8 , which allows us to recover the signal from the noise. The techniques used for this are described in the next section.

18.3 Standard analysis

The standard approach (see van der Klis 1989) is to divide the time series into M equal-length segments of N time bins $x_k, k = 0, \dots, N - 1$ each (so, ideally $N_{tot} = MN$), to calculate the discrete Fourier transform of each

¹This is of course not exactly true. The background does not pose large problems in practice, as it just constitutes an additional source of detections (of photons as well as charged particles) that is not strongly variable on the time scales we are interested in and uncorrelated to the fluctuations due to the star. Detector deadtime processes leading to “missed” photons *do* constitute a serious complication (e.g., van der Klis 1989, Mitsuda and Dotani 1989, Vikhlinin et al. 1994, W. Zhang 1995) that will be ignored here.

segment:

$$a_j \equiv \sum_{k=0}^{N-1} x_k e^{2\pi i j k / N} \quad j = 0, \dots, N/2,$$

to convert this into a power spectrum for each segment

$$P_j \equiv \frac{2}{a_0} |a_j|^2 \quad j = 0, \dots, N/2,$$

and then to *average* these power spectra (see below). Note that in our application $a_0 \equiv \sum_{k=0}^{N-1} x_k = N_{ph}$, the number of photons detected in the segment. With this power spectral normalization, due to Leahy et al. (1983), it is true that if the x_k are distributed according to the Poisson distribution, then the P_j follow the χ^2 distribution with 2 degrees of freedom, so $\langle P_j \rangle = 2$ and $\sigma_{P_j} = 2$. This white noise component with mean level 2 and standard deviation 2, induced in the power spectrum due to the Poisson fluctuations in the time series, is called the “Poisson level”. It can be considered as “background” in the power spectrum, against which we are trying to observe the other power spectral features, which are caused by the intrinsic variability of the X-ray binary.

The physical dimension of the thus defined powers is the same as that of the time series: $[P_j] = [a_j] = [x_k]$. Often the Y-axis of plots of power spectra in this normalization is just labeled “POWER”, which reflects the fact that the physical interpretation of the P_j in terms of properties of the star is inconvenient. For this reason, in recent years another power spectral normalization has become popular where the powers are reported as $Q_j \equiv P_j / \lambda$, with λ the “count rate”, the number of detected photons per second: $\lambda = N_{ph}/T$, where T is the duration of a segment. The Q_j are dimensionless, and can be interpreted as estimates of the power density $Q(\nu_j)$ near the frequency $\nu_j \equiv j/T$, where $Q(\nu)$ is a function of frequency whose integral gives the fractional root-mean-square amplitude r of the variability. This latter quantity is defined as

$$r \equiv \sqrt{\frac{1}{N} \sum_{k=0}^{N-1} (x_k - \bar{x})^2} \quad \text{where} \quad \bar{x} \equiv \frac{1}{N} \sum_{k=0}^{N-1} x_k.$$

(This follows directly from Parseval’s theorem.) The fractional rms amplitude r_{12} due to fluctuations in a given frequency range (ν_1, ν_2) is given by

$$r_{12} = \int_{\nu_1}^{\nu_2} Q(\nu) d\nu.$$

So, the physical interpretation of $Q(\nu)$ is easy: it is the function whose integral gives you the square of the fractional rms amplitude of the variability in the original time series. The physical unit used for $Q(\nu)$ is $(\text{rms}/\text{mean})^2/\text{Hz}$,

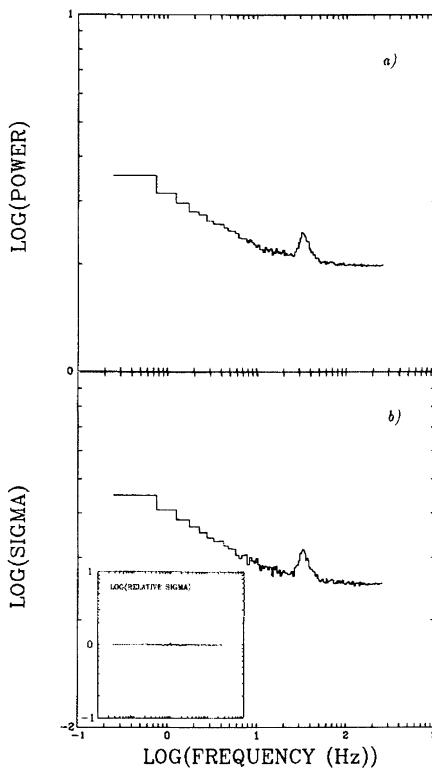


FIGURE 2. Top: An average of 6166 power spectra of EXOSAT ME data on the X-ray binary GX 5–1 showing power-law noise and QPO. Bottom: The standard deviation of the 6166 power values averaged in each frequency bin. Inset: the ratio of standard deviation to mean. The standard deviation equals the mean power as expected for χ^2 distributed powers. From Van der Klis (1989).

where “rms” and “mean” both refer to the time series; “rms/mean” is just the dimensionless quantity r .

The averaging of the power spectra mentioned above is usually performed both by averaging individual power spectra (from different segments) *together* (averaging the P_j 's with the same j from the M different segments) and by averaging powers at adjacent frequencies (P_{j+1} to P_{j+W} , say). The main purpose of this is, of course, to decrease the standard deviation of the power estimates, which in the raw spectra is equal to the mean power. The reason to calculate many power spectra of segments of the data rather than one very large power spectrum of the whole data set, apart from computational difficulties with this approach, is that in this way it is possible to study the variations in the power spectrum as a function of time.

The final step in the analysis is to fit various functional shapes $f(\nu)$ to the power spectra using the method of χ^2 minimization that is also used in X-ray spectroscopy (the Levenberg-Marquardt method described in Press et al. 1992, Chapter 15). Because many individual power estimates have been averaged in the analysis process, the central limit theorem ensures that the uncertainties of the final power estimates \bar{Q} are approximately normally distributed, as required for this method to work well.

A problem arises concerning what “uncertainties” $\sigma_{\bar{Q}}$ should be assigned to these power estimates. Usually $\sigma_{\bar{Q}} = \bar{Q}/\sqrt{MW}$ is assumed, where MW is the number of individual powers averaged to obtain \bar{Q} (W stands for the “width”, the number of adjacent powers averaged; M for the number of averaged power spectra). This is approximately correct, and was experimentally verified (Fig. 2), in the case that the dynamic range of the power spectrum is dominated by the intrinsic differences in the mean amplitude of the star’s variability as a function of frequency rather than by the stochastic fluctuations in power. If instead the stochastic variations dominate then this procedure for estimating the uncertainties can lead to severe underestimation of the power, as accidentally low powers will get high weights in the fitting procedure and vice versa. A solution to this problem that is sometimes adopted is to estimate the uncertainty in \bar{Q} as $f(\bar{\nu})/\sqrt{MW}$, where $f(\bar{\nu})$ is the fit function and $\bar{\nu}$ the frequency corresponding to \bar{Q} .

18.4 Applications and results

The method described in the previous section works. It allows astronomers to quickly characterize the variability properties of large amounts of data, to study the changes in the properties of the variability as a function of time and other source characteristics, and to measure amplitudes and characteristic time scales of the variability. The method straightforwardly extends to simultaneous multiple time series, for example time series obtained in different photon energy bands. For N_{band} simultaneous time series, it is possible to calculate $N_{band}(N_{band} - 1)$ different cross-spectra between them, and to look for systematic time delays in the variability as a function of energy. A comprehensive description of the methods used for this can be found in the paper by Vaughan et al. (1994).

A very important aspect is the possibility to identify different “power spectral components” in the variability. This is done by studying the changes in the shape of the power spectra as a function of time and other source properties, such as brightness or photon energy spectrum. It usually turns out that the simplest way to describe the changes in the power spectrum as a function of time is in terms of the sum of a number of components whose properties (strength, characteristic frequency) depend in a smooth, systematic and repeatable way on, for example, brightness. If this is the

BLACK-HOLE-CANDIDATE POWER SPECTRA

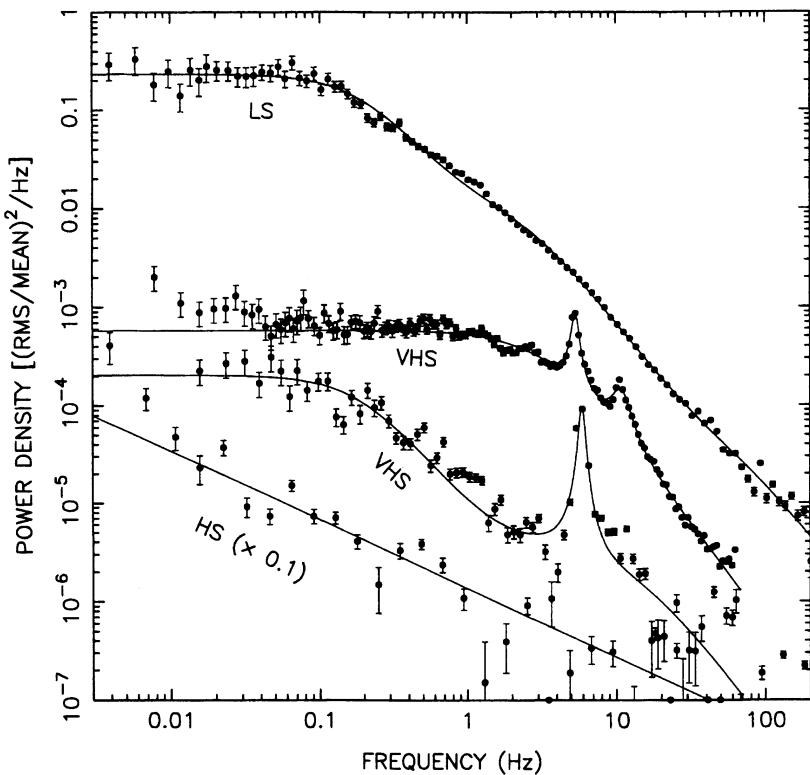


FIGURE 3. Power spectra of black hole candidates. The spectrum labeled LS (low state) is of Cygnus X-1; the other two are of GS 1124–68. From Van der Klis (1995).

case, then it is natural to interpret these power spectral components as due to different physical processes (or different aspects of the same process) that are all affecting the count rates at the same time, and that have been disentangled from each other in the analysis just described.

Examples of power spectral components that are distinguished in practice are “power law noise”, “band-limited noise” and “quasi-periodic oscillations (QPO)”. All of these are all presumed to be stochastic processes in the time series that cause, respectively, a component in the power spectra that fits a function $f_{PL}(\nu) = C_{PL}\nu^{-\alpha}$, one that fits $f_{BLN}(\nu) = C_{BLN}\nu^{-\alpha}e^{-\nu/\nu_{cut}}$ and one that fits a Lorentzian $f_{QPO}(\nu) = C_{QPO}/(\nu - \nu_{QPO})^2 + (\Gamma/2)^2$. For a component to be called QPO it is sufficient that the corresponding power spectral component has a shape approximately described by $f_{QPO}(\nu)$ with

the peak full-width-at-half- maximum (FWHM) Γ less than half its centroid frequency ν_{QPO} ; of course this definition is arbitrary.

Figure 3 shows a number of actually observed power spectra, plus the functions that were fit to them in order to describe them in terms of the power spectral components just mentioned. QPO peaks, band-limited noise and power law noise can all be seen.

Sometimes the shape of the observed power spectrum is ambiguous with respect to the decomposition into power spectral components as described here. An advantage of the χ^2 minimization method is that it allows to quantify the degree of this ambiguity by comparing the χ^2 values of the different possible combinations of fit parameters.

18.5 Problems at the low-frequency end: low-frequency leakage

At the low frequency end (typically, below 0.01 Hz) of the power spectra a number of problems occurs in an analysis along the lines described above that is usually not really dealt with in a satisfactory manner.

One problem is, that in order to reach the lowest frequencies in the first place, it is necessary to choose the length of the time segments T relatively large and therefore M relatively small. This means that at the lowest frequencies the statement made in Section 18.3, that a large number of individual powers has been averaged and therefore the average power is approximately normally distributed is no longer true. Other (e.g., maximum-likelihood) fitting procedures are required to take the true distribution of the average powers into account, but such methods are not usually applied. See Papadakis and Lawrence (1993) for a discussion of a method to remedy some of these problems.

A more serious problem is, that the true power spectrum at these frequencies often seems to be a quite steep power law. The finite time window T in those situations leads to so-called “low-frequency leakage” (see Deeter 1984 and references therein): power shows up at a higher frequency in the power spectrum than where it belongs. One way to see this is by noting that the lowest frequency accessible in the discrete Fourier transform is $1/T$. If there is a lot of variability at lower frequencies than this, then due to these slow variations the time series of an individual time segment will usually have a large overall trend. The Fourier transform of this trend produces a power law with index -2 in the power spectrum. Another way to describe the effect of low-frequency leakage is by noting that, according to the Fourier convolution theorem, the Fourier transform calculated in the finite time window T is related to the true Fourier transform by a convolution of the true transform with the transform of the window function. As the window function is a boxcar, its Fourier transform is the well-known sinc function, with a big central lobe and upper and lower sidelobes that

gradually decrease in amplitude. For true power spectra steeper than a power law of index -2 , the contribution to the convolution of the lower side lobes overwhelms that of the central lobe, and the result is a power spectrum that is a power law with index -2 (e.g., Bracewell 1965). So, for any true power spectrum steeper than -2 , the actually measured power spectrum will have slope of ~ -2 .

There are well-known solutions to this well-known problem. The most famous one is data tapering: instead of a boxcar window a tapered window is used, i.e., a window that makes the data go to zero near its end points more gradually than by a sudden drop to zero. Other methods are polynomial detrending (fitting a polynomial to the data and subtracting it) and end-matching (subtracting a linear function that passes through the first and the last data point). Deeter et al. (1982) and Deeter (1984) have explored a number of non-Fourier methods. All these methods work in the sense that to some extent they suppress the side lobes of the response function and therefore they are able to recover power laws steeper than with index -2 (the value of the power law index where the method breaks down is different in each case).

However, typically these methods have only been evaluated for the case where the time series is pure power-law noise, and in many cases even only with respect to their effectiveness in recovering the power law index, not even the noise strength. Some methods require that the index of the power law is known in advance! Nearly nothing is known about the way in which these methods affect the results of fits to power spectra with complicated shapes such as those described above. These methods may recover the correct power law index for the low-frequency part of the power spectrum, but what will be the effect on the fractional rms values and time scales of all the other components? For this reason, these methods are not usually applied.

18.6 Problems at the high frequency end: what is the shortest time scale?

The high-frequency end of the power spectra, near the Nyquist frequency, is of particular interest to astrophysicists, as it is there that we expect to find the signatures of the fastest accessible physical processes going on in the star. A question that is often asked in this context is: “what is the shortest variability time scale τ we can detect in the data?”. Generally, what is observed at high frequency is that the power spectrum more or less gradually slopes down towards the Poisson level. The problem is to decide out to which frequency intrinsic power is observed above the Poisson background, and *what this number means*.

A decidedly misleading way to answer the question about the shortest time scale, I think, is by determining the *shortest time interval Δt within*

which significant variations can be detected. It is obvious, that for a slowly varying source, by zooming in on some gradual slope in its light curve $I(t)$, this shortest time interval can be made as short as one wishes, just by improving the quality of the data (by increasing ϵ for example). Yet, most of the work on “shortest time scales” seems to aim for measuring Δt rather than τ . It seems clear that one must also take into account the amplitude of the variations, not just the time within which they occur to do something that is physically useful. Following Fabian (1987) one can for example define the variability time scale as

$$\tau(t) = \frac{I(t)}{\dot{I}(t)},$$

where the dot denotes the time derivative.

Defined this way, τ is a measure of how steeply I changes with t , and depends itself on t . Using a power spectrum, one would measure some average of this quantity by determining the fractional rms amplitude $r(\nu_{hi})$ near some high frequency ν_{hi} in the way described in Section 18.3, and then write

$$\bar{\tau} = C \times \frac{1}{\nu_{hi} r(\nu_{hi})},$$

where C is a constant of order unity depending on assumptions on exactly how the variations causing the power in the spectrum took place. However, for precise work one has to worry about low-frequency leakage, too. Exactly the same problem as described in Section 18.5 can occur here: power from lower frequencies can leak up to higher frequencies due to the sinc response associated with the power estimators. That this problem is serious is apparent from the fact that the observed power spectrum of, for example, the famous black-hole candidate Cyg X-1 has an index of ~ -2 for frequencies above ~ 10 Hz (see Fig. 3), just the value at which low-frequency leakage begins to worry us. It would be of great interest to have a foolproof way to subtract power that has leaked up from lower frequencies, or even to have a way to make a conservative estimate of (obtain a lower limit on) the true high-frequency power. It is not clear to what extent this can be accomplished. Obviously, as in any convolution problem, some information has been lost, but how much can be recovered is a problem X-ray astronomers do not know the answer to. I wonder to what extent standard deconvolution procedures might be useful here.

I note parenthetically that a method that has been applied for determining the shortest variability in a time series by Meekins et al. (1984) seems to suffer from the same problem of low-frequency leakage. In this method, the time series is divided up in very short segments of, for example, $N = 10$ bins each. In each segment a quantity called “chi-squared” is calculated as follows:

$$\chi^2_{Meekins} = \sum_{k=0}^{N-1} \frac{(x_k - N_{ph}/N)^2}{N_{ph}/N}.$$

Here x_k is the number of photons detected in bin k and N_{ph} is the total number of photons in the segment. One recognizes an “observed over expected” variance ratio for an expected Poisson distribution. The distribution of this quantity is then compared to that expected if all variability in the time series would be due to Poisson fluctuations and if a significant excess is found, this is interpreted as detection of variability on time scales between the length T of a segment and the duration of one bin.

One would expect low-frequency leakage to be as much of a problem here as in the power spectral method. It is easy to see that if there are variations in the time series on time scales much longer than the segment length, the data points in the segments will usually follow steep trends, causing large values of $\chi^2_{Meekins}$ that are not related to variability on time scales shorter than T . Indeed, Meekins et al. show that $\chi^2_{Meekins}$ is closely related to the Fourier power in the segment, so from the point of view of low-frequency leakage the Meekins et al. method is less effective than the standard power spectral approach, as it requires T to be quite short and therefore increases the probability of steep trends in the data segments.

REFERENCES

- [1] Bracewell, R., “The Fourier Transform and its Applications”, McGraw-Hill, 1965.
- [2] Deeter, J., *Astrophys. J.* 281, 482, 1984.
- [3] Deeter, J., and Boynton, P.E., *Astrophys. J.* 261, 337, 1982.
- [4] Deeter, J., *Astrophys. J.*
- [5] Fabian, A.C., in Proc. “The Physics of Accretion onto Compact Objects”, Lecture Notes in Physics 266, 229, 1987.
- [6] Leahy, D.A., Darbro, W., Elsner, R.F., Weisskopf, M.C., Sutherland, P.G., Kahn, S., Grindlay, J.E., *Astrophys. J.* 266, 160, 1983.
- [7] Meekins, J.F., Wood, K.S., Hedler, R.L., Byram, E.T., Yentis, D.J., Chubb, T.A., Friedman, H., *Astrophys. J.* 278, 288, 1984.
- [8] Mitsuda, K., Dotani, T., *Publ. Astron. Soc. Japan*, 41, 557, 1989.
- [9] Papadakis, I.E. and Lawrence, A., *Mon. Not. R. Astron. Soc.* 261, 612, 1993.
- [10] Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., “Numerical Recipes in FORTRAN”, 2nd Edition, Cambridge Univ. Press, p. 650, 1992.
- [11] Van der Klis, M., in “X-Ray Binaries”, W.H.G. Lewin et al. (eds.), Cambridge Univ. Press, p. 252, 1985.
- [12] Van der Klis, M., in “Timing Neutron Stars”, Ögelman and Van den Heuvel (eds.), NATO ASI C262, Kluwer, p. 27, 1989.
- [13] Vaughan, B.A., Van der Klis, M., Wood, K.S., Norris, J.P., Hertz, P., Michelsson, P.F., Van Paradijs, J., Lewin, W.H.G., Mitsuda, K., Penninx, W., *Astrophys. J.* 435, 362, 1994.
- [14] Vikhlinin, A., Churazov, E., Gilfanov, M., *Astron. Astrophys.* 287, 73, 1994.
- [15] Zhang, W., Jahoda, K., Morgan, E.H., Giles, A.B., *Astrophys. J.* 449, 930, 1995.

19

Wavelet and Other Multi-resolution Methods for Time Series Analysis

Jeffrey D. Scargle

ABSTRACT Wavelets are the central idea of a broad framework for thinking about, displaying, and analyzing data. These localized basis functions provide flexible and efficient methods for analysis of time series and image data – including noise removal, characterization of stochastic behavior, depiction of power spectrum time-variations, and data compression.

19.1 Wavelet analysis of digital signals

Wavelets comprise a complete set of basis functions which are localized in both time and frequency. As astronomers understand their great utility in problems of noise reduction, signal characterization and modeling, data compression, and so on, the applications [Sc96] will continue to grow. Within a few years multiresolution techniques will be familiar fare, and wavelets will be more used than Fourier techniques. This section summarizes the special properties which make wavelets so useful.

19.1.1 Basic properties

Wavelets are *localized*. That is, each member of the basis is concentrated on a subinterval of T , the total sampled time interval. (In contrast, the Fourier basis functions are all spread over T .) This feature makes them especially useful for signals containing sharply defined, localized features like jumps, jerks, and bumps. But their frequency (or, equivalently, time-scale) properties make them useful as well for representing $\frac{1}{f}$ noise, chirps, fractals, and other multi-scale or self-similar structures.

Another important feature of wavelets is that there is a close relationship between the magnitudes of the wavelet coefficients and the smoothness of the corresponding function – a relationship not shared by Fourier analysis. In simple terms, an attempt to smooth a function by diminishing the amplitudes of some of its wavelet components is guaranteed to succeed – no such guarantee applies to Fourier components. Technically, wavelets form

unconditional bases, which are in turn optimal bases for data compression and for statistical estimation [Me92a, Me92b, Do93c]. The practical importance is that smoothing algorithms with very nice properties can be implemented using wavelets: decreasing the size (or eliminating entirely) some of the wavelet coefficients results in a relatively noise-free version of the sampled data (see §19.1.4).

Wavelets elucidate not just fluctuations on a specific scale, but also relationships between those on different scales. One invents display techniques for highlighting behaviors on different scales and visualizing their relationships. One analyzes data (especially images) into the sum of a smoothed and a detailed version, and then treats them separately. One decomposes data into components on a hierarchy of scales, treats them individually, and then re-synthesizes the data; this theme pervades much of wavelet data analysis.

One thinks in terms of not just the harmonic and time-scale content of fluctuation spectra, but of the time-evolution of such spectra, and the detection of sharp (“localized”) changes in them. One seeks methods to remove random observational noise from data, without rounding off the sharp edges present in the signal – “denoising without smoothing.” And one seeks data analysis and visualization techniques that are suited to the multi-scale behaviors characteristic of $\frac{1}{f}$ -noise and chaotic dynamical systems.

This paper is not meant to introduce the reader to technical aspects of wavelets, or to their actual use in data analysis. There is now a large literature on the theory and applications of wavelets: good sources include [Da92, Me92a, Me92b, Me93, Ch92a, Ch92b]. See also [Gr95] and the World Wide Web sites given in §19.3. Much of the published research is oriented toward abstract functional analysis, but the monograph by Meyer [Me93] discusses the goals of wavelets in data analysis and provides a historical overview, a treatment of time-frequency methods, plus four chapters on applications (computer vision, fractals, turbulence, and the study of distant galaxies). The book [Wi94] is a comprehensive overview of wavelet theory with C source code for many practical algorithms. A recent review [BDG] compares over thirty commercial and public domain wavelet analysis software packages.

Here I will outline the basic definitions for wavelets in which both the data and the scale variable in the transform are discrete – to set the terminology and notation. For definiteness, assume that we have $N = 2^M$ (M an integer) evenly spaced samples of a function X ; the data are then

$$X_n, n = 1, 2, 3, \dots, N. \quad (19.1)$$

The *mother wavelet* (also called the *analyzing wavelet*) ψ , can be thought of as a localized shape function, that in turn generates the set of basis functions. For example, the simplest mother wavelet is the *Haar wavelet*,

which in discrete time is:

$$\psi(n) = \begin{cases} \frac{1}{2}, & n = 1; \\ -\frac{1}{2}, & n = 2; \\ 0, & \text{all other } n \end{cases} \quad (19.2)$$

The wavelets making up the complete, orthonormal basis are shifted and time-scaled copies of the mother wavelet:

$$\psi_{s,l}(n) = 2^{-\frac{s}{2}} \psi\left(\frac{n}{2^s} - l\right). \quad (19.3)$$

The shift and scale factors are defined by the location index l scale index s , respectively. The index ranges

$$s = 0, 1, \dots, M - 1 \quad (19.4)$$

$$l = 0, 2^{s+1}, \dots, 2^M - 2^{s+1} \quad (19.5)$$

generate $N - 1$ copies. (For Haar, the remaining degree of freedom is \bar{X} .) Since this set is complete and orthogonal, the data can be represented as:

$$X_n = \sum_s \sum_l W_{s,l} \psi_{s,l}(n), \quad (19.6)$$

the *wavelet coefficients* $W_{s,l}$ being defined as the inner product of the corresponding wavelet with the data:

$$W_{s,l} = \sum_n \psi(n)_{s,l} X_n. \quad (19.7)$$

In practice, however, this *wavelet transform* is computed with a fast recursive algorithm.

19.1.2 Analysis and visualization of temporal structure: The scalogram

The wavelet transform is a kind of quick-and-dirty time-scale distribution (see §19.1.5). It is therefore useful to display the log of the absolute value of the wavelet coefficients as a function of scale and location – this is called the *scalogram* [Ri92], or sometimes the *wavelet modulus*. For example, [Go90, Go91] contain pictures of such three-dimensional displays.

Szatmary and co-workers have used scalograms to study changes in the harmonic content of regular and semiregular variable stars [Sz92a, Sz92b]. They carried out simulations on a suite of synthetic signals [Sz93, Sz94], providing useful information on how the wavelet transform of periodic signals responds to things like additive noise, uneven sampling, data gaps, amplitude and phase modulation, phase jumps, period changes, and mode switching.

19.1.3 Detection/characterization of $1/f$ noise: The scalegram

To study a signal as a function of scale, ignoring location, wavelets offer an alternative to the power spectrum, namely the *scalegram*, defined as the mean square wavelet coefficient (averaged over the location index):

$$V(s) = \frac{2^s}{N} \sum_l (W_{s,l})^2. \quad (19.8)$$

where the l -sum is over the $2^{-s}N$ location index values allowed at scale s [Eq. (19.5)]. To avoid confusion with the scalogram, the scalegram is sometimes called the *wavelet spectrum*. For data with non-normal errors, a modification of the scalegram using the absolute value of the wavelet transform in place of its square is more robust [No94, Cl73]. [CM96] compares the scalegram with other wavelet approaches related to power spectra.

Randomly and/or chaotically fluctuating systems are of great importance in astronomy [Sc92]. The power spectrum is perforce continuous and usually steeply declining with frequency. The famous " $\frac{1}{f}$ " noise and the related *fractal Brownian motion* are examples. Flandrin and co-workers have applied wavelet methods, including the scalegram, to the study of such processes [Fl92a, Fl92b, Ab95, Ab96]. A recent monograph on self-similar processes, with a wavelet flavor is [Wo96].

It is instructive to evaluate the scalegram of a noisy signal and to relate it to the scalegram of the unknown intensity X^{true} . Assume that the observed data consist of a series of photon counts in N evenly spaced time bins. Then X_n is a Poisson-distributed random variable with expected value $E(X_n) = X_n^{true}$, and has the probability distribution

$$\text{Prob}\{X_n = k\} = \frac{(X_n^{true})^k e^{-X_n^{true}}}{k!}. \quad (19.9)$$

Let this relationship between the measured and true signals be written

$$X_n = \text{Pois}(X_n^{true}).$$

(19.10)

It should be stressed that this relation is:

- Very different in form from the additive noise model that most statisticians assume is relevant to scientific data.
- Quite accurate; in practice deviations are small correlations produced by detector dead-time effects, *etc.*, not departures from this formula.

The structure of this non-additive, non-Gaussian noise is determined by the precisely Poisson nature of photon counting.

Assuming that any other errors are in the form of an additive, stationary, normally distributed process R_n , we have for the observed data:

$$X_n = \text{Pois}(X_n^{true}) + R_n. \quad (19.11)$$

We assume that both R and the Poisson photon noise are uncorrelated (white) and uncorrelated with each other. A straightforward computation gives

$$E[V^{X_n}(s)] = V^{X_n^{true}}(s) + \sigma_R^2 + \frac{1}{N} \sum_n X_n^{true} \quad (19.12)$$

In short, the net effect of both noise processes is to add a constant to the scalegram. An identical result holds for the power spectrum.

If there is no counting noise, the last term on the right-hand side vanishes, and one needs an estimate of σ_R , *e.g.* from instrumental considerations. If, as is often the case in X-ray or γ -ray data, there is no additive noise, the constant $\frac{1}{N} \sum_n X_n^{true}$ in Eq. (19.12) can be estimated from the total counts, $\frac{1}{N} \sum_n X_n^n$. This procedure seems to work very well in practice [Sc93]. Indeed, using the corrected scalegram we identified self-similar variability in Sco X-1; this led to a model – the “dripping handrail” [Yo96] – for chaotic accretion in this low mass X-ray binary star/disk system. Fritz [FR95] has used the scalegram in a wavelet-based study of flickering in cataclysmic variables.

19.1.4 Wavelet denoising methods

In a marvelous series of papers, Donoho and co-workers [Do93a, Do93b, Do93c, Do93d, Do93e, Do93f, Do93g, DJKP93a, DJKP93b, DJ93a, DJ93b, DJ93c, DJ93d] have proposed wavelet procedures for dealing with a number of estimation and time series analysis problems.

Of great potential application in astronomy is *denoising*. These algorithms yield accurate estimates of signals embedded in noise, without losing sharp features such as steps, bumps, or spikes. This methodology should not be confused with the usual approach to noise reduction by *smoothing*, which invariably blurs such features.

For a practical introduction to these methods, I recommend consulting [Do93f] or the documentation in the **WaveLab** system (see §19.3) – even if the reader does not use the software itself. For details on estimating unknown functions embedded in noise see [Do93a, Do93b, Do93d]; more technical matters are discussed in [DJ93c]. The use of wavelets to estimate probability distributions is discussed in [DJKP93a]. For an introduction to the theory of wavelet and wavelet packet based de-noising, inverse methods, segmented multiresolutions, and nonlinear multi-resolutions, consult [Do93g, DJKP93b].

Among the mathematical results in these papers are proofs that the basic thresholding technique is amazingly good both in a practical sense (as depicted with synthetic examples) and in a theoretical sense (in that the rate at which the variance of the estimator improves with sample size is greater than for conventional methods – nearly as great as for an ideal estimator [Do93a]).

Kolaczyk [Ko96] has developed denoising theory for the case of Poisson noise. His simulation studies and applications to γ -ray bursts are promising.

Bendjoya, Petit and Spahn [Be93] have developed a procedure that attempts to recognize significant patterns in the wavelet coefficients of noisy data. Their approach begins with the isolation of significant data structures by retaining only wavelet coefficients above a relatively high threshold. On a second pass smaller coefficients are added if they relate in a specified way to those retained in the first pass. For example, coefficients can be added if they exceed a second, lower, threshold and are contiguous in scale-location space to the first-pass coefficients.

19.1.5 Wavelet time-frequency methods

The topic of time-frequency [Coh89] (and time-scale) representations is wide-ranging and complicated, primarily because its goal (the portrayal of the time-evolution of a quantity that is defined as a time-integral) is somewhat self contradictory. See [Da90] for a wavelet-flavored discussion.

From their definition as localized basis functions, oscillating about zero, wavelets clearly have a potential role in the representation of time-varying spectra [Ri92, Fl92c]. Donoho and co-workers have developed and included in **WaveLab** (see §19.3) a number of algorithms for carrying out a generalized multi-scale analysis of time series, with the goal of producing time-frequency distributions. The basic idea involves displaying the sum of the time-frequency distributions of the atoms in an atomic representation [see Eq. (19.13) below].

Wavelet methods appear to be able to avoid some unpleasant side-effects of the classical approach to time-frequency distributions, such as the infamous interference terms between multiple periodic components [Coh89, Ka92]. Donoho and von Sachs are developing wavelet based tools for time-frequency and time-scale displays [VS96].

Maurice Priestley, in his contribution to this volume, discusses some problems with the use of wavelets to study time-dependent spectra.

19.1.6 Data compression

Most of the “action” in many signals arising in the real world is contained in a few wavelet coefficients. That is to say, only a very few of the N coefficients are significantly different from zero. So if the many small coefficients are discarded, one obtains a very parsimonious but accurate representation of the signal – image or time series. Indeed, §19.1.4 above indicates that such a procedure will often discard more noise than signal. With wavelet compression one gets denoising free. Donoho [Do93c] describes the further compression obtained by quantizing wavelet coefficients.

19.1.7 Detection/characterization of discontinuities: The CWT

The *continuous wavelet transform* is evaluated at a continuum of scales, not just discrete scales related to each other by factors of 2. It is useful in locating and measuring the order of discontinuities in time series. On a 3D map of the wavelet coefficients, the ridge lines (loci of local maxima) locate discontinuities, and the logarithmic slopes of the ridge lines give the orders [Al95]. This procedure works surprisingly well even in the presence of substantial noise.

19.1.8 Spectral estimation

It is perhaps natural that wavelet denoising be used to estimate the notoriously noisy power spectrum. The distribution of the power spectrum is so pathological that the statisticians [Ga93a, Ga93b, Mo92, Mo93, Mo94] have attacked the better-behaved logarithm of the spectrum.

A variant of the STF transform, called the Gabor transform, is being used by a group at NASA-Goddard and elsewhere [Bo94a, Bo94b] to investigate astronomical systems, both observational and theoretical, containing harmonic signals with time-varying frequencies. The goal of their techniques is to provide insight into non-stationary evolution of chaotic and other non-linear physical systems. Boyd *et al.* [Bo94b] studied numerical data from computations of a single star scattering gravitationally off of a binary system. The Gabor transform was used to generate time-frequency plots exhibiting the dual-frequency behavior of this 3-body system, and to show the evolution of the frequencies over time. In addition, they analyzed Hubble telescope photometric data of the star HD 605435 and found evidence that changes in the frequencies are too abrupt to be due to a beat phenomenon, as proposed by other workers.

Stark (personal communication) is experimenting with wavelet denoising of spectra from the complicated multiply periodic power spectra of solar oscillations obtained by the Global Oscillation Network Group (GONG).

19.1.9 Fixing up the wavelet transform

Here are three practical problems, and their fixes, with the use of the standard wavelet transform.

Edge Effects The simplest implementations of the wavelet transform assume that the data are periodically replicated outside of the sampled interval. As with Fourier methods, it is desirable to avoid this “wraparound” by zero-padding the data or otherwise. This problem of “wavelet bases for the interval” has been solved by Cohen, Daubechies, Jawerth and Vial [CDJV], by constructing special sets of wavelets such that the inner product in Eq. (19.7) never extends beyond the edges of the sampled data.

The **WaveLab** implementation of these left-edge, middle, and right-edge wavelets is called the *CDJV Boundary-Corrected Transform*.

Wavelet Transform of Unevenly Spaced Data For well-known reasons, astronomical time series are often unevenly spaced. Techniques have been developed to compute – for data with arbitrary sampling in time – Fourier transforms [Sc89], power spectra [Sc82], and correlation functions [Sc89]. The goals are computational techniques for estimating these quantities, and ways to correct for the unevenness of the sampling to the extent possible.

Since the Haar wavelet is piecewise constant, wavelet coefficients can be computed by simply counting the number of points in the corresponding intervals – no matter what the sampling. But it is not obvious which levels in the scale hierarchy to include, since the *minimum scale length* is not well defined. Taking it to be somewhat less than the mean sampling interval is probably reasonable in practice. The only other problem is the treatment of any dyadic intervals that contain no samples – as will sometimes be the case due to pure chance if the sampling is random. In the computation of the scalogram, one should average the squared wavelet coefficients corresponding only to intervals which contain samples.

Lehto [Le96a, Le96b] has used this approach to the scalogram to study unevenly sampled optical observations of OJ 287. He introduced the interesting idea of studying the effects of the sampling by computing the scalogram of white noise sampled at the same times as the real data (the *noisegram*).

Translation-invariant Wavelet Transforms A third practical problem with wavelets results from their localized nature. Suppose the signal has one or more localized features. Consider those wavelets whose characteristic scale is on the same order as that of the feature – and whose location overlaps the feature. The coefficients of such wavelets will depend strongly on the relative locations of the wavelet and the feature. This renders the analysis sensitive to the whims of how the wavelets fall.

Coifman and Donoho [Co95] have solved this problem with a special *translation invariant wavelet transform* that, in essence but not directly, considers all possible rotations of the data by one sample (with wraparound) – thus in effect averaging over all phases. The **WaveLab** software contains tools for implementing this extension of the ordinary wavelet transform.

19.2 Beyond wavelets

There is a class of signal processing methods in which overcomplete sets of functions, some of which are wavelets or wavelet-like, are used in place of complete, orthogonal bases. The general idea is that we want to represent

the observed signal as a linear superposition of the form

$$X_n = \sum_k c_k \psi_k(n), \quad n = 1, 2, \dots, N. \quad (19.13)$$

The c_k are constant coefficients, and the ψ_k belong to the set

$$D = \{\psi_k(n)\}, k = 1, 2, 3, \dots, \quad (19.14)$$

chosen so that any function of interest can be accurately, but not necessarily uniquely, so represented.

The relation in (19.13) is called an *atomic representation* of the data, the functions ψ_k are called *atoms*, and D is a *dictionary*. Behind these colorful terms is the idea that the data are like a body of information that can be described by a sequence of words (atoms) listed in a dictionary. The same information can be translated to another language, using words from its dictionary. Even more generally, one can construct multi-lingual libraries, the shelves of which house many complete dictionaries. An expression (read “signal”) can be represented in many ways in any of these languages. It is nevertheless understandable that the data may find its most efficient representation in one language. Thus to express a given idea, in *matching pursuit* we seek the best wording, in *basis pursuit* the best language. As with human languages, the best choice depends on the data – French is the language of philosophy, Italian the language of love, etc.

Atoms are intended to be elementary building blocks, well suited to compose the structures known or likely to be present in the signal of interest. Often the atoms are simple in form; atomic representation then amounts to *assembling a complicated structure out of simple building blocks*. An example is classical Fourier analysis, where the building blocks – sine and cosine functions – comprise a complete, orthonormal set which provides a unique expansion in the form in Equation (19.13) for any sufficiently well-behaved function.

The Fourier basis functions – sines and cosines – are precisely localized in frequency, but completely unlocalized in time. Atoms that, on the contrary, are well localized in both time and frequency are important in theory and applications, and are called *time-frequency atoms*. An example is the windowed or short-time Fourier transform (STF), in which the dictionary consists of sines and cosines modulated by a (sliding) localizing window function (see [Da92] for the relation of STF to wavelets, and [Coh89] for its shortcomings).

One general way to make a dictionary is to pick a function $g(t)$, and then construct copies of it shifted, $g(t-l)$, and scaled, $g(\frac{t}{S})$, in time. If g is localized in time, and its Fourier transform is localized in frequency, then these copies – $g(\frac{t-l}{S})$, for all allowed values of S and l – comprise a dictionary of *time-frequency atoms*. Wavelets are an example of this construction. In some applications it makes sense to also frequency-modulate the function.

But in the most general setting atoms are arbitrary functions, considered in some way to be suitable for piecing together functions of interest.

If the dictionary D is a complete, orthogonal set (*i.e.*, a basis), then each f has a unique atomic representation, and we have an invertible transformation between the data and the coefficients c_k . Fourier and wavelet transforms are examples. But we allow D to be *overcomplete*. Since each f can be constructed in a number of ways, one wants to find the linear combination of basis functions that best represents f . There are many choices for what “best” means in this context – for example, the one with the fewest number of atoms.

Further, one can define a set of dictionaries, called a *library*; then the goal is to find the best dictionary, and in turn its best representation of the data. Wavelet packets and cosine packets [Co92] are examples of such libraries.

An important case is that in which the functions $\psi_k(t)$ can be grouped into subsets S_k of D that correspond to different “scales of variation.” Start with a basic function $\psi(t)$ confined to the unit interval $0 \leq t \leq 1$. Define S_0 to be the set consisting of this function and all copies of it translated in time by integer multiples of one unit – *i.e.*, $\psi(t-l)$, $l = 0, \pm 1, \pm 2, \dots$

Then consider the functions $\psi(\frac{t}{2^s})$ – stretched out or compressed versions of $\psi(t)$, according as s is > 0 or < 0 . Define D_s as the set of this scaled function translated by all integer multiples of 2^s . This generates a *hierarchy of scales*. The set of all linear combinations of functions in D_s is completely contained in the set of all linear combinations of functions in D_{s-1} . Any linear combination of functions in D_s is said to *vary on the time scale* 2^s . This abstraction, known as *multiresolution analysis* (MRA), was invented and rigorously developed in 1986 by Mallat and Meyer, as described in [Da92, Me92b]. Wavelets are an example.

19.3 WaveLab – wavelet and MRA software tools

There is a tremendous amount of information about wavelets available on the Internet and the World Wide Web (WWW). TeX/LaTeX and other versions of most of the papers by Donoho and co-workers are readily available on the Internet. Do an ftp to the node **playfair.stanford.edu** as “anonymous,” enter your internet address in lieu of a password, change to the directory `./pub/donoho`, (type “`cd pub/donoho`”), and then “get” followed by the name of the file you want. This can be done for any of the papers for which a file name in the form `—.tex` appears in the bibliography below; the WWW address is <http://playfair.stanford.edu/~donoho/>. In addition, a software package called **WaveLab**, implementing all of the wavelet operations discussed here and many more, can be obtained at no cost from this WWW site: <http://playfair.stanford.EDU:80/~wavelab/>.

Wavelet Digest is a good general source, with pointers to papers and meetings, plus a Q & A forum: <http://www.math.sc.edu/~wavelet/>. Fionn Murtagh has an extensive bibliography on astronomical applications of wavelets, multiresolution, noise suppression, filtering, image restoration and compression at <http://http.hq.eso.org/~fmurtagh/wavelets.html>.

Acknowledgments: I wish to thank Dave Donoho and Ian Johnstone, of Stanford University, who have made many valuable suggestions. This paper is based in part on work supported by grants from NASA's Astrophysics Data Program.

REFERENCES

- [Ab95] Abry, Patrice, Goncalves, Paulo, and Flandrin, Patrick, (1995), "Wavelets, spectrum analysis and 1/f processes," in *Lecturm Notes in Statistics*, No. 103, Wavelets and Statistics, eds. A. Antoniadis and G. Oppenheim, Springer-Verlag, 15-30.
- [Ab96] Abry, Patrice, and Flandrin, Patrick, (1996), "Point Processes, Long-Range Dependence and Wavelets," in *Wavelets in Medecine and Biology*, eds. A. Aldroubi and M. Unser, CRC Press, Boca Raton, 413-437.
- [Al95] Alexandrescu, M., Gilbert, D., Hulot, G., Le Mouël, J.-L., and Saracco, G., (1995), "Detection of Geomagnetic jerks using wavelet analysis," *J. Geophys. Res.*, **100**, No. B7, 12,557 - 12,572.
- [Be93] Bendjoya, Ph., Petit, J.-M., E., and Spahn, F. (1993), "Wavelet Analysis of the Voyager Data on Planetary Rings. I. Description of the Method," *Icarus*, **105**, pp. 385-399.
- [Bo94a] Boyd, P. T., Carter, P.H., Gilmore, R., and Dolan, J. F. (1995), "Nonperiodic Variations in Astrophysical Systems: Investigating Frequency Evolution," *Ap. J.*, **445**, pp. 861-871.
- [Bo94b] Boyd, P. T., McMillan, S. L. W. (1994), "Chaotic scattering in the gravitational three-body problem," *Chaos*, **3**, pp. 507-523.
- [BDG] Bruce, A., Donoho, D., and Gao, H-Y., (1996), "Wavelet Analysis," *IEEE Spectrum*, **33**, No. 10, October 1996, pp. 26-35.
- [CM96] Chiann, Chang, and Morettin, Pedro A., (1996), "A Wavelet Analysis for Time Series," available at <ftp://ftp.ime.usp.br/pub/morettin>, in directory ./pub/morettin
- [Ch92a] Chui, C. K. (1992) *An Introduction to Wavelets*, (Acamedic Press: Boston).
- [Ch92b] Chui, C. K. (1992) *Wavelets: A Tutorial in Theory and Applications*, (Acamedic Press: Boston).
- [Cl73] Claerbout, J.F., and Muir, F., (1973), "Robust Modeling with Erratic Data," *Geophysics*, **38**, 826-844.
- [Coh89] Cohen, L., (1989), "Time-Frequency Distributions – A Review," *Proc. IEEE*, **77**, pp. 941-981.

- [CDJV] Cohen, A., Daubechies, I., Jawerth, B., and Vial, P., (1993), "Multi-resolution analysis, wavelets and fast algorithms on an interval," *Comptes Rendus Acad. Sc. Paris*, **316** (série I), pp. 417-421.
- [Co92] Coifman, R. and Wickerhauser, M. V. (1992), "Entropy-based algorithms for best basis selection.", *IEEE Trans Information Theory*, **38**, pp. 713-718.
- [Co95] Coifman, R. R., and Donoho, D. L.. "Translation- Invariant De-Noising," (1995), to appear in *Wavelets and Statistics*, Anestis Antoniadis, ed. Springer-Verlag Lecture Notes.
- [Da90] Daubechies, I. (1990), "The Wavelet Transform, Time-Frequency Localization and Signal Analysis," *IEEE Transactions on Information Theory*, **36**, pp. 961-1005.
- [Da92] Daubechies, I. (1992), *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics: Philadelphia.
- [Do93a] Donoho, D.L., (1993), "De-Noising by Soft- Thresholding," preprint (denoiserelease3.tex).
- [Do93b] Donoho, D.L., (1993), "Interpolating Wavelet Transforms," preprint (interpol.tex).
- [Do93c] Donoho, D.L., (1993), "Unconditional Bases are Optimal Bases for Data Compression and for Statistical Estimation," preprint (UBRelease.tex).
- [Do93d] Donoho, D.L., (1993), "Smooth Wavelet Decompositions with Blocky Coefficient Kernels," preprint (blocky.txt) to appear in *Recent Advances in Wavelet Analysis*, L Schumaker and G. Webb, eds., Academic Press, pp. 1-43.
- [Do93e] Donoho, D.L., (1993), "Nonlinear Solution of Linear Inverse Problems by Wavelet-Vaguelette Decomposition." preprint (nslip.tex).
- [Do93f] Donoho, D.L., (1993), "Wavelet Shrinkage and W.V.D.: A 10-minute tour," preprint (toulouse.tex).
- [Do93g] Donoho, D.L., (1993). "Nonlinear Wavelet Methods for Recovery of Signals, Densities, and Spectra from Indirect and Noisy Data," soon to appear in workshop book edited by I. Daubechies.
- [DJ93a] Donoho, D.L., and Johnstone, I. M., (1993), "Ideal Spatial Adaptation by Wavelet Shrinkage," preprint (isaws.tex).
- [DJ93b] Donoho, D.L., and Johnstone, I. M., (1993). "Adapting to Unknown Smoothness via Wavelet Shrinkage," preprint (ausws.tex).
- [DJ93c] Donoho, D.L., and Johnstone, I. M., (1993), "Minimax Risk over l_p -Balls for l_q -error," preprint (mrlp.tex).
- [DJ93d] Donoho, D.L., and Johnstone, I. M.. (1993), "Minimax Estimation via Wavelet Shrinkage," preprint (mews.tex).
- [DJKP93a] Donoho, D.L., Johnstone, I. M., Kerkyacharian G., and Picard, D. (1993), "Density estimation by wavelet thresholding," preprint (dens.tex).
- [DJKP93b] Donoho, D.L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1993), "Wavelet Shrinkage: Asymptopia?," preprint (asymp.tex).
- [Fl92a] Flandrin, P., (1992), "Wavelet Analysis and Synthesis of Fractional Brownian Motion," *IEEE Transactions on Information Theory*, **38**, pp 910-917.

- [Fl92b] Flandrin, P., (1992), "On the Spectrum of Fractional Brownian Motion," *IEEE Transactions on Information Theory*, **35**, pp. 197-199.
- [Fl92c] Flandrin, P., Vidalie, B., and Rioul, O., (1992), "Fourier and Wavelet Spectrograms seen as smoothed Wigner-Ville Distributions," in [Me92a].
- [FR95] Fritz, Thomas, (1995), "Perspektiven der Analyse des Flickering in CV's durch Wavelet-Transformation," WWW document at <http://aquila.uni-muenster.de/~fritz/seminar/semin.html>
- [Ga93a] Gao, Hong-Yee, (1993), "Wavelet estimation of spectral densities in time series analysis," Ph.D. dissertation, University of California, Berkeley.
- [Ga93b] Gao, Hong-Yee (1993), "Choice of Thresholds for Wavelet Shrinkage Estimate of the Spectrum," *J. of Time Series Analysis*, in press.
- [Go90] Goupil, M. J., Auvergne, M., and Baglin, A. (1990) "A Wavelet Analysis of the ZZ Ceti Star G191 16" Proc. of the 7th European Workshop on White Dwarfs, NATA, Toulouse, France, September 1990, ed. G. Vauclair.
- [Go91] Goupil, M. J., Auvergne, M., and Baglin, A. (1991) "Wavelet Analysis of Pulsating White Dwarfs," *Astron. Astrophys.*, **250**, pp. 89-98.
- [Gr95] Graps, A., (1995), "An Introduction to Wavelets," *Computational Science and Engineering*, **2**, No. 2, Summer 1995, IEEE Computer Society. Also see <http://www.best.com/~agraphs/agraphs.html>
- [Ka92] Kadambe, S., and Boudreaux-Bartels, G. F., (1992), "A Comparison of the Existence of 'Cross Terms' in the Wigner Distribution and the Squared Magnitude of the Wavelet Transform and the Short Time Fourier Transform," *IEEE Transactions on Signal Processing*, **40**, pp. 2498-2517.
- [Ko96] Kolaczyk, Eric D., (1996), "Estimation of Intensities of Inhomogeneous Poisson Processes Using Haar Wavelets," Technical Report 436, Department of Statistics, The University of Chicago, Chicago, to be submitted to *Journal of the Royal Statistical Society, Series B*.
- [Le96a] Lehto, Harry, (1996), "Analysis of the optical light curve of OJ 287," to appear in *Proceedings of the Eleventh Florida Workshop in Non-linear Astronomy, Annals of the New York Academy of Sciences*.
- [Le96b] Lehto, Harry, (1996), "Wavelets in unevenly spaced data: OJ 287 light curve," this volume.
- [Ma93] Mallat, S., and Zhang, Z., (1993), "Matching Pursuits with Time-Frequency Dictionaries," *IEEE Transactions on Signal Processing*, **41**, pp. 3397-3415.
- [Mc87] McHardy, I. and B. Czerny, B., (1987), "Fractal X-ray Time Variability and Spectral Invariance of the Seyfert Galaxy NGC 5506", *Nature* **325**, pp. 696-698.
- [Me92a] Meyer, Y. (1992) *Wavelets and Applications*, Proceedings of the International Conference, Marseille, France, May 1989, Springer-Verlag: New York.
- [Me92b] Meyer, Y. (1992) *Wavelets and Operators*, Cambridge University Press: Cambridge (English translation).

- [Me93] Meyer, Y. (1993) *Wavelets. Algorithms and Applications*, SIAM: Philadelphia. (translation from French by R. Ryan)
- [Mo92] Moulin, Pierre, (1992), "Wavelets as a Regularization Technique for Spectral Density Estimation." *Proc. IEEE-Signal Processing Symposium on Time-Frequency and Time-Scale Analysis*, 73-76.
- [Mo93] Moulin, Pierre, (1993), "A Wavelet Regularization Method for Diffuse Radar-Target Image and Speckle-Noise Reduction," *Journal of Math. Imaging and Vision*, Special Issue on Wavelets, **3**, 123-134.
- [Mo94] Moulin, P., (1994), "Wavelet Thresholding Techniques for Power Spectrum Estimation," *IEEE-Trans-SP*, **42**, 3126-3136.
- [No94] Norris, J.P., Nemiroff, R.J., Scargle, J.D., Kouveliotou, C., Fishman, G.J., Meegan, C.A., Paciesas, W.S., and Bonnell, J.T. (1994), "Detection of Signature Consistent with Cosmological Time Dilation in Gamma-Ray Bursts," *Ap. J.*, **424**, pp. 540-545.
- [Ri92] Rioul, O., and Flandrin, P.. (1992), "Time-Scale Energy Distributions: A General Class Extending Wavelet Transforms," *I.E.E.E. Transactions on Signal Processing*, **40**, 1746-1757.
- [Sc82] Scargle, J. "Studies in Astronomical Time Series Analysis. II. Statistical Aspects of Spectral Analysis of Unevenly Spaced Data," (1982), *Ap. J.*, **263**, pp. 835-853 (Paper II).
- [Sc89] Scargle, J., "Studies in Astronomical Time Series Analysis. III. Fourier Transforms, Autocorrelation and Cross- correlation Functions of Unevenly Spaced Data," (1989), *Ap. J.*, **343**, pp. 874-887 (Paper III).
- [Sc92] Scargle, J., (1992), "Chaotic Processes in Astronomical Data," in *Statistical Challenges in Modern Astronomy*, ed. Feigelson and Babu, Springer-Verlag: New York, pp. 411-428.
- [Sc93] Scargle, J., Steiman-Cameron, Young, K., Donoho, D., Crutchfield, J., and Imamura, I. (1993) "The Quasi-Periodic Oscillations and Very Low-Frequency Noise of Scorpius X-1 as Transient Chaos: A Dripping Handrail?" *Ap. J. Lett.*, **411**, pp. L91-L94.
- [Sc96] Scargle, J., (1996), "Wavelet Methods in Astronomical Time Series Analysis," To appear in Proceedings of the conference: *Applications of Time Series Analysis in Astronomy and Metrology*, Padua, Italy, 3-10 Sept, 1995, Chapman and Hall.
- [Sz92a] Szatmary, K., and Gal, J. (1992) "Wavelet-Analysis of Some Pulsating Stars," poster at the IAU Coloquium No. 137, Inside the Stars, 13-17 April 1992, Vienna, Austria.
- [Sz92b] Szatmary, K., and Vinko, J. (1992), "Periodicities of the light curve of the semiregular variable star Y Lyncis." *M.N.R.A.S.*, **256**, pp. 321-328.
- [Sz93] Szatmary, K., Vinko, J., and Gal, J. (1993) 'Tests of Wavelet Analysis for Periodic Signals in Astronomy," poster presented at the conference "Applications of Time Series Analysis in Astronomy and Meteorology," Sept. 6-10, (1993), Padova, Italy.
- [Sz94] Szatmary, K., Vinko, J., and Gal, J. (1994) "Application of wavelet analysis in variable star research. I. Properties of the wavelet map of simulated variable star light curves," *Astronomy and Astrophysics*, in press.

- [VS96] <http://playfair.stanford.edu/~rvs>
- [Wi94] Wickerhauser, M.V., (1994), *Adapted Wavelet Analysis, from Theory to Software*, A. K. Peters, Wellesley, Massachusetts
- [Wo96] Wornell, G. W., (1996), *Signal Processing with Fractals: a Wavelet-Based Approach*, Prentice-Hall, Inc.
- [Yo96] Young, Karl, and Scargle, Jeffrey D., (1996), "The Dripping Handrail Model: Transient Chaos in Accretion Systems," *Ap. J.*, **468**, pp. 617-632.

Summaries

One esteemed member of each community – statistician PETER BICKEL of the University of California at Berkeley and astronomer VIRGINIA TRIMBLE of the University of California at Irvine – faced the difficult task of synthesizing the wealth of ideas expressed at the conference. Trimble reviews the many problems and methods raised by astronomers and statisticians at the conference (including the contributed papers). Bickel provides statistical advice for a number of astronomical projects faced with large and complex databases. Both give insightful views of the interrelationship of statistics and astronomy.

20

An Overview of “SCMA II”

P. J. Bickel¹

20.1 Introduction

The goals of this conference (and those of “SCMA I”) [FB1992] were three fold.

- i) To expose statisticians to the statistical challenges posed by the explosive growth in the amount and complexity of astronomical data and in astrophysical theory and experimentation during the second half of this century.
- ii) To expose astronomers to developments in statistical methodology which might be helpful in the analysis of astronomical data
- iii) To introduce the two communities to each other with a view to the establishment of fruitful collaborations.

I came, because I am interested in:

- (i) Important substantive questions which can only be attacked through the gathering and analysis of large very complex bodies of data.
- (ii) General methodological questions and approaches common to many different substantive fields,

and because, when young, I read with fascination George Gamow’s “Birth and Death of the Sun”. [G 52]

This conference, with its presentations of large astronomical projects rather than astrophysical theory, as in SCMA I, met my interests perfectly, so goal (i) was fulfilled in my case. In this paper I’ll do my initial bit towards goal (ii) and discuss goal (iii).

Specifically, I’ll mainly give this theoretician’s reaction to the wonderful array of projects and problems presented by the astronomers at this conference with occasional attempts to indicate where new statistical formulations and methodologies I’m somewhat familiar with might be of use. I’ll

¹Prepared with partial support of NSF Grant DMS 9504955 and NSA Grant MDA 904-94-11-2020

Department of Statistics, University of California, Berkeley

necessarily limit myself to discussing those papers by astronomers about which I believe I may be able to say something useful. At the end, I'll comment on some of the methodology papers by statisticians, make some general remarks on statistical inference, and on future relations between statistics and astronomy and more significantly, astronomers and statisticians. I'll begin in section 20.2 with discussing in general terms the striking features of astronomical data and statistical approaches actual and potential that were discussed and others that seem suitable to me. In section 20.3, I'll go over particular papers and close in section 20.4 with the discussion of the future.

20.2 Astronomical data/statistical approaches

The presentations brought home the evident:

Astronomical data sets tend to be both huge and very complex. MACHO (Massive Compact Halo Objects) is an ongoing project whose database already contains about 3 terabytes, HIPPARCOS, in its missions collected 130 million data points. The data are or will be of many types; "pictures" as in MACHO and HIPPARCOS, X ray spectra as in AXAF and ROSAT, laser pulses as in the Jefferys' data set and gravity waves as in LIGO.

Some particular features of astronomical (but not only astronomical!) data are:

- (i) They are often being gathered dynamically over time. A resulting theme appearing in several talks is that the resulting time series of objects have irregular spacing. This creates difficulties with standard frequency domain time series analysis.
- (ii) There are serious built in sampling biases generally stemming from the fact that there is a detection limit for any type of electromagnetic radiation i.e. the data are truncated. But other subtler selection biases also operate – for instance as in the MACHO project where there is a detection probability for lensing events as well as a probability that a so called lensing event is fraudulent.
- (iii) Photon noise for single sources is viewed as Poisson and hence when the number of photons is large approximately Gaussian. This approximation is evidently questionable in a number of the examples presented, in particular the HIPPARCOS and MACHO data for a variety of reasons including but not limited to superposition of noise from several sources and inadvertently or unavoidably included systematic effects.

More significantly and this can be singled out as a separate point:

- (iv) The data being analyzed has typically been subjected to a great deal of preprocessing, fitting of point spread functions, subtraction of background, corrections for movement of the platform.

Finally,

- (v) The signal form assumptions are guided by physics but often only vaguely. On the one hand this gives relatively crude statistical models a chance and on the other gives one firmer ground to stand on than in the social sciences where, as in astronomy, we generally cannot do experiments.

The papers in this conference exhibit statistical questions and approaches of almost every type.

- (i) Exploratory (“Model free”) data analysis, particularly curve fitting, from histograms to wavelets.
- (ii) Confirmatory data analysis
 - a) Error bars
 - b) Tests of significance
 - c) Confidence regions
- (iii) Classification and prediction

Methodologies of every kind appear, most notably frequency based time series analysis, but also state space methods, parametric models and maximum likelihood, nonparametric methods, Bayesian methods and Monte Carlo simulations of great complexity including and going beyond the usual bootstraps. Novel methodologies discussed at SCMA II include wavelet and multiresolution analysis and to some extent testing and estimation of chaotic behaviour.

Other methodologies or points of view which perhaps could have stood more presentation given the types of questions appearing in the various studies include semiparametric models and “hidden good data” methods such as the EM algorithm and related theory. As you see in my discussion of HIPPARCOS, empirical Bayes methods may be of use. Modern computational Bayes methods are worth learning. In all fairness to how up to date astronomers are, a poster paper [KT96] on EM in states space models was presented (Ch. 24) and a workshop on modern Bayesian computation organized! Of undoubtedly importance is simultaneous inference, setting thresholds when many related determinations are made and assessing the significance of features found by search. Although there has been a lot of theory developed in conjunction with social science applications, see [M 81], for example, these are issues that in general are still wide open.

20.3 Individual papers

My discussions of individual papers are inevitably impressionistic and flawed. In particular, given space limitations, the details of methodologies are necessarily omitted by authors and any approaches that I suggest may well already have been considered and rejected for good reason. To really contribute to these large, complex, and very interesting projects, statisticians have to be drawn into collaboration rather than casual consultation.

T. Axelrod et al.: The MACHO project (Ch. 12)

Axelrod's clear presentation of the search for massive dark objects through the gravitational lensing effect brought out in detail many of the themes I mentioned. My comments refer to the indicated sections of their paper.

Section 2. From the original enormous series of images photometric time series for individual objects are produced by elaborate fitting procedures which are sketched. The residuals from the fit of a composite of 68 light curves appear non Gaussian in being long tailed. The authors believe that temporal correlation is also present.

Given the departures described, lack of Gaussianity is no surprise. However, this may in part be an artifact of considering 68 light curves each of which may have satisfactorily Gaussian residuals but with differing scales. Mixing over scales produces long tailed distributions. An approach that a statistician might suggest here is to robustify the fitting by replacing least squares by least absolute deviations or by something more drastic in the fitting. Algorithms for least absolute deviations fits i.e. minimizing $\sum |y_i - f(t_i, \theta)|$ rather than $\sum (y_i - f(t_i, \theta))^2$ where y_i is observed flux at time t_i that are fast (using linear programming methods) are available from the STAT LIB server [<http://lib.stat.cmu.edu>]. More general algorithms for robustifying are discussed for instance in "Statistical Models in S" [CH 92].

Section 3.1. Axelrod et al. describe the evolution of their detection method from what can be viewed as acceptance in a simple test for the hypothesis that the star's light curve during the event is of the form prescribed by (1.2) - (1.4) to a two stage procedure in which final acceptance requires satisfaction or failure to be excluded by at least 10 criteria. The fundamental difficulty they face is that they are classifying light curves into ones exhibiting microlensing or not without benefit of a training sample. Or rather, the training sample is being built up as they go along with curves finally being decided as exhibiting microlensing by scientific consensus.

Given that a training sample of 80+ light curves which exhibit microlensing by scientific consensus now exist and arbitrarily large samples of variable star light curves are available with large amounts of covariate information, e.g. proximity to SN1987a, it seems reasonable to try to extract features determining classification systematically using algorithms such as CART [Classification And Regression Trees] and neural nets. Given the

small size of the training sample of curves, representing the information for each curve by a small number of parameters such as those entering into the ad hoc lensing rules first probably makes sense.

Another possibility is, as do Axelrod et al., to consider GSGL models whose rapidly growing number of parameters enable them to fit anything quickly but couple then with penalties which try to reduce this tendency. In the statistical and signal processing literature these are known as AIC (The Akaike Information Criterion). An elementary reference is Linhart and Zucchini [LZ 86].

Section 3.3. The testing questions posed here clearly require acquaintance with the astrophysics that I do not have. It’s clear that, without good estimates of the sampling biases caused by both failures to detect true events and false alarms, tests of models are problematic. However, Axelrod et al. do raise one generic question I can comment on. Are there more powerful tests of goodness of fit than K-S [Kolmogorov-Smirnov]? The answer is, of course. K-S is simply one of a number of omnibus tests which have some but very little power in all directions and considerable power in essentially one direction. A direction here is a one parameter family of alternatives. Given competing specific alternatives, e.g. u_{min} suitably scaled having a beta distribution, more powerful tests can be constructed.

Section 3.2.4. There seems little to suggest on the Axelrod et al. “bootstrap” procedure for estimating the probability of detection and their discussion of the difficulties of quantifying the probability of false alarm.

There is a possible dynamic alternative that I would like to suggest which I have to admit has a “Bayesian” flavor. Attach to each object in level 1.5 candidacy or above, not a 0 or 1 probability of being a microlensing event, but rather a probability given the data with prior probabilities corresponding to frequencies of microlensing events for level 1.5 candidates. Then rescan all objects at regular intervals to estimate the probability of false alarms given thresholds on these posterior probabilities.

Summary. Axelrod et al. present a paradigm of astronomical investigations; huge amounts of raw data, sophisticated data reduction involving statistical and other processing techniques, detection (“sampling”) biases, and models specified only partly by the underlying physics. The ultimate goals of classification, estimation of error probabilities and testing of hypotheses are familiar to statisticians but the environment is a very challenging one.

B. F. Schutz and D. Nicholson: LIGO (Ch. 13)

This is an exciting project, attempting for the first time ever to detect gravitational waves. Unfortunately, its uniqueness and the lack of analogous data make it difficult for statisticians to contribute at this crucial planning of experiment stage. The questions raised are largely theoretical, for instance, determination of lower bounds on the Bayes risk as carried out by van Trees and Ziv-Zakai. There are similar bounds in the statistical

literature due to Brown and Gajek [BG90] and Gill and Levit [GL 95], and others. However, there seems to be a long distance between anything available in the statistical (or I think signal processing) literature and realistic estimates of the detection efficiency for gravitational waves.

Other interesting theoretical questions were raised in connection with the all sky survey and setting appropriate thresholds when many tests are being carried out simultaneously. This is a fundamental problem in much of statistics. In this case a careful probabilistic analysis coupled with the methods of [A 89] may be of use.

R. L. White: Object classification in astronomical images (Ch. 8)

This excellent paper gives a very clear presentation of general classification procedures in the context of star/galaxy classification. It may be useful for statisticians as well as astronomers to put his discussion in the context of a standard unifying paradigm:

There is a training set index/training set of two samples X_1, \dots, X_m ; Y_1, \dots, Y_n one from each of the relevant populations of objects to be classified. The X 's (which can be thought of as vectors of features) are independent and identically distributed with common density f and the Y 's with common density g . An object with feature vector Z is to be classified. Z is known to be an X with probability π and Y with probability $1 - \pi$. It is well known that if f, g, π are all known the best classification rule is then:

$$\begin{aligned}\delta(Z) &= X \text{ if } \frac{f}{g}(Z) > \frac{1-\pi}{\pi} \\ &= Y \text{ if } \frac{f}{g}(Z) < \frac{1-\pi}{\pi}.\end{aligned}$$

In the great majority of applications, f and g are unknown and estimated by \hat{f} , \hat{g} using the training samples X_1, \dots, X_m , Y_1, \dots, Y_n . For example, nearest neighbour rules correspond to the following estimation procedure. Let $v_{\mathbf{X}}(z)$ be the volume of the sphere centered at z with radius the distance to the nearest member of X_1, \dots, X_m to z . Suppose we know $|Z| \leq M$, (for any possible Z). Then, let

$$\hat{f}(z) = 1/v_{\mathbf{X}}(z)c(\mathbf{X})$$

where $c(x) = \int_{\{|z| \leq M\}} v_x^{-1}(z)dz$. Define \hat{g} , $c(Y)$ similarly and let

$$\frac{\pi}{1-\pi} = \frac{c(X)}{c(Y)}.$$

Oblique trees which are equivalent at least in principle to flavors of CART [BFOS 84] can be thought of as making \hat{f} a histogram based on adaptively chosen parallelepiped bins. Even neural nets correspond to rather complex \hat{f} .

Two important issues touched on by White that are discussed more fully in the statistical literature in connection with adaptive procedures such as CART and neural nets are:

- i) The size of the tree/How long the neural net is run?
- ii) Estimation of probabilities of correct classification

There is an extensive discussion of one type of tree “pruning” in [BFOS 84]. A recent paper [B 95] shows how tree structure classification methods can be greatly improved by generating pseudosamples from the training set, growing trees and averaging them.

White discusses estimation of probabilities of misclassification using cross validation – the method of choice in statistics as well. He also in section 2.5 discusses estimation of the probability of each class given the data and seems to distinguish between the case of noisy and pure data. This is somewhat confusing. A given object is either a galaxy or a star. The probabilities White refers to are Bayesian ones given prior probabilities say π that the object is a star and $1 - \pi$ that it is a galaxy. Then, what White proposes to estimate is using our previous notations, $\frac{\pi f(Z)}{\pi f(Z) + (1-\pi)g(Z)}$ the posterior probability that the object is a star given Z . This is evidently doable by plugging in the estimates \hat{f} and \hat{g} defined by the algorithms and is more or less what I believe White advocates. However, π should also be specified and that is more reasonably chosen as the more or less known frequency of stars versus galaxies. This also seems a more reasonable choice than the one given by the nearest neighbours algorithm.

White concludes his discussion with an astronomically motivated extension of classification which I would expect to be relevant elsewhere, for example in creating standard maps of locations of brain features on the basis of dissection of individual brains. Given the characters of the observation process and plates, White points out that the image features of identical objects appear to change from one plate to another. He reports great success with a method in which raw intensities of objects on plates are replaced by their relative normalized ranks (for the same plate). Essentially, White has proposed a solution for an extension of the classification paradigm. For simplicity, suppose in the paradigm the X ’s and Y ’s are univariate, say, intensities. We do not observe the ideal intensities that we would get if the plate were perfect and exposed under ideal conditions, call these, $X_1, \dots, X_m; Y_1, \dots, Y_n$. Instead we have $\{\tilde{X}_{ij} : 1 \leq i \leq I, 1 \leq j \leq J_i\}$ $\{\tilde{Y}_{ik} : 1 \leq i \leq I, 1 \leq k \leq K_i\}$ where i refers to the plate and $\sum_{i=1}^I J_i = m$, $\sum_{i=1}^I K_i = n$. The relation between the \tilde{X} , \tilde{Y} , and X , Y ’s is that $\tilde{X}_{11} = a_1(X_1), \dots, \tilde{X}_{21} = a_2(X_{J_1+1}), \dots, \tilde{Y}_{11} = a_1(Y_1), \dots, \tilde{Y}_{21} = a_2(Y_{K_1+1})$ etc where the a ’s are unknown functions. If we assume they are monotone increasing (order of intensity within plates is preserved) then it is natural and can be justified on theoretical grounds – see [L 86] Ch. 6 for instance,

to replace $(\tilde{X}_{i1}, \dots, \tilde{X}_{iJ_i}, \tilde{Y}_{i1}, \dots, \tilde{Y}_{iK_i})$ by their ranks and then to standardize to $J_i + K_i$, the number of objects on plate. This kind of situation where one does classification based on structured training samples is novel to me. It bears further investigation by statisticians for theoretical properties and use in other contexts. There is also the possibility that other groups of transformations than monotone ones, leading to invariants other than ranks, may be reasonable in other contexts.

A. Siemiginowska et al.: Astrostatistics issues for AXAF (Ch. 14)

In this account of existing statistical problems in X ray astronomy and future expectations of data of high resolution both spatially and in the frequency domain, Siemiginowska et al. pose a number of crisp questions some of which I believe I have translated successfully. I can't resist commenting although it is clear that in all instances much more detail on what is actually being done is really needed.

4.1 Searching many parameter spaces Dealing with functions with many local maxima some but not all of which may be reasonable is a pervasive problem in statistics as well as computer science and astronomy. Going further in searching among regression models, it is customary and sensible not just to look at the best model as measured by fit penalized for the number of parameters being fitted, but at all models whose penalized fits are close to the "optimum" and then look more closely at physically reasonable ones – see for example the discussion in [MT 77] p. 304. Assessing probabilistically the significance of such fits and in general doing inference is difficult. An approach which may be reasonable is to do what astronomers often do and that is Monte Carlo ("bootstrap" (ET 1994)) the whole selection process. This can be done from either a Bayesian or frequentist point of view.

4.2 - 4.3. Uncertainties in model – Weighting Data. This is a set of questions where closer acquaintance with what is being done is really necessary. It does sound as if using mixture models for the received emissions, postulating a mixture of emission lines corresponding to point masses and some suitable parametric model for the continuum spectrum and then building in the binning process to give the likelihood of the data may be reasonable. Such models can often be fit reasonably using the EM algorithm [DLR 77], [BPSW70], see [KT 96] for an example at this conference.

4.4. Correlated Residuals The approach above may be reasonable here also. Alternatively, it may be plausible to bring in the presence of emission lines via Markov models and this might lead to runs tests.

5.1. N space The area of N dimensional random fields is of great interest in both probability and statistics. Joe Horowitz in this volume has prepared a bibliography for astronomers.

5.2. - 6. Wavelets – Adaptive Smoothing Here the questions seem to be mainly about inference and again Monte Carlo/Resampling methods may well be the only way to go.

7. Instrument Related Issues Again this may be a set of situations to which “Hidden data” models including censoring, grouping etc. may be applicable but “the Devil is in the details”.

M. van der Klis: Quantifying rapid variability (Ch. 18)

This was an extremely clear and interesting presentation of what amazing information can apparently be gotten with low signal to noise ratio data. Issues of the evolution of spectra and at most local stationarity of series in time were then introduced and followed up in the subsequent discussion by Swank, van der Klis and Sibul.

I’d like to add to the conceptual framework for nonstationary time series of the type discussed in his presentation and elsewhere in the conference the following. Suppose the series one observes can be thought of as M_t where

$$\langle M_t \rangle = \mu_t$$

and μ_t is the signal which one thinks of as deterministic and $\epsilon_t \equiv M_t - \mu_t$ as “noise”. In the astronomical context what appears to often be assumed is that M_t has a Poisson distribution with mean μ_t and it is at least hoped that the M_t are independent. From the van der Klis description of the spectrum of M_t , it seemed reasonable to model μ_t ($t = 1, 2, \dots$) as

$$\mu_t = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{itw} f(w) dw$$

where f is largely 0 but has 1 or more bumps. This suggests fitting a (semi)parametric model by maximum likelihood in the time domain where f is fit by a part of its periodic wavelet expansion. Thus replacing the Poisson by the Gaussian approximation and taking $t = 1, 2, \dots, T$, we would minimize

$$\sum_{t=1}^T \frac{(M_t - \mu_{t\ell}(\theta))^2}{M_t}$$

where

$$\mu_{t\ell}(\theta) = \sum \{\theta_\rho \hat{g}_\ell(t) : \ell \leq 2^p\}$$

and the \hat{g}_ℓ are the Fourier transforms of the g_ℓ defined in [D 92] p. 304. Here p would be small, of order $\log_2 T$ or smaller and chosen adaptively using the data and experience. As Jeff Scargle has pointed out, if M_t is assumed to be Poisson then one could more appropriately maximize $\sum_{t=1}^T \{-\mu_{t\ell}(\theta) + M_t \log \mu_{t\ell}(\theta)\}$.

This is a special case of a paradigm which has been found useful in modeling in all sorts of contexts. Somewhat specialized it reads as follows:

Suppose,

$$M_t = \mu_t + \epsilon_t, 1 \leq t \leq T$$

Assume a simple parametric form for the joint likelihood of $\epsilon_1, \dots, \epsilon_T$ and call it $f(M_t - \mu_t ; 1 \leq t \leq T, \theta)$. Assume also that μ_t is of the form,

$$\mu_t = h(g, t), g \in \mathcal{G}$$

where \mathcal{G} is a big set of functions.

Finally, assume every $g \in \mathcal{G}$ is arbitrarily approximable by a parametric form $g_k(\cdot, \eta_1, \dots, \eta_k)$, for k suitably large and appropriate η_1, \dots, η_k . Let $\mathcal{G}_k = \{g(\cdot, \eta_1, \dots, \eta_k : \eta_j \text{ real}\}$. Then pick \hat{K} adaptively (not trivial!) using the data and act as if $\mathcal{G}_{\hat{K}}$ were true. That is estimate g (and θ) and hence h by maximizing $\log f(M_t - h(g(\cdot, \eta_1, \dots, \eta_{\hat{K}}), t), 1 \leq t \leq T, \theta)$. This is Grenander's method of sieves [G83].

J. O. Berger: Recent developments in Bayesian analysis (Ch. 2)

W. H. Jefferys: Bayesian analysis of lunar ranging data (Ch. 3)

"Bayesian" methods have, I think, rightly gained favor in astronomy as they have in other fields of statistical application. I put "Bayesian" in quotation marks because I do not believe that this marks a revival in the sciences in the belief in personal probability. To me it rather means that all information on hand should be used in model construction, coupled with the view of Box [B 79], who considers himself a Bayesian: "Models, of course, are never true but fortunately it is only necessary that they be useful". The Bayesian paradigm permits one to construct models and hence statistical methods which reflect such information in an, at least in principle, marvellously simple way. A frequentist such as myself feels as at home with these uses of Bayes principle as any Bayesian.

Jefferys' example shows the utility of building a model which takes the physical information into account. The asymptotic equivalence of vague prior procedures and maximum likelihood suggests that the same conclusions would have been drawn with his model via maximum likelihood. Of course, if the physics suggests weighting of the likelihood a "Bayes prior", that's very reasonable for procedure generation also.

Where I find myself uncomfortable is with discussions such as that of the first part of Berger's paper where it is implicitly postulated that users of frequentist criteria do so mindlessly and a Bayesian solution which is supposed to guard against all mindlessness is advanced.

My answer to Berger's example is that a p value between .04 and .05 is implausible under both $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 2)$ but a p value $\leq .05$ is much less implausible under $\mathcal{N}(0, 2)$. The bottom line is, I think, that getting $|Z| \equiv 1.96$ is surprising if $\theta = 0$ and nothing more. Having said this, I'd like to compliment Berger on the intrinsic Bayes factor approach which looks like a useful tool for the generation of procedures.

Finally, I’d like to state a credo with which I hope most experimental scientists would agree.

1. Just like frequentist model construction, Bayes priors have to be taken with a large grain of salt when external more objective validation is not possible.
2. Even if Bayesian and frequentist model based inferences coincide, they may both be wrong. A famous example is the extremely small χ^2 obtained in testing the hypothesis of 3:1 red:white segregation in one of Mendel’s famous sweet pea experiments. If one postulates a binomial model with $p = \frac{3}{4}$ the chance of getting a χ^2 as large or larger than the one observed is essentially 1. However, by the same token the chance of getting a χ^2 as small or smaller is essentially 0. The plausible conclusion drawn by R.A. Fisher, see [JGI] p. 308 is that someone cooked the data – the binomial model was wrong whatever be p . Of course, if one’s prior had positive probability on the data being cooked this would be a triumph for personal probability. But Fisher was not a Bayesian just an excellent scientist.

D. Guégan: Nonparametric methods for time series (Ch. 17)

Guégan presents a novel approach to testing for the presence of chaos and estimation of features of chaos. Her approach looks promising but I believe there may be difficulties with the curse of dimensionality in estimating higher order densities.

A possible alternative to her approach for estimating chaotic dimension, but subject to the same difficulty, is the following: Estimate $E(X_t|X_{t-1}, \dots, X_{t-d})$ and $E(X_t^2|X_{t-1}, \dots, X_{t-d})$ and take the chaotic dimension to be the first d for which the estimated conditional variance of X_t is uniformly small as a function of X_{t-1}, \dots, X_{t-d} .

I do have a serious reservation about attempts to determine underlying chaotic behaviour in situations where noise is present and there is little physical guidance as to the nature of the underlying deterministic chaotic process that might be in force. In a recent paper [BB 96] we argue that without restrictions any stationary process can be approximated by finite MA processes with independent innovations, with probability $= \frac{1}{e}$.

There are a number of other excellent papers by astronomers and statisticians which I shall not comment on as part of this paper. On one, **Statistical aspects of the Hipparcos photometric data** by van Leeuwen et al. (Ch. 15), I have written a separate discussion.

I participated in the discussion of the wavelet and other techniques for irregularly spaced observations group of papers by Thomson, Priestley (Ch. 16), Scargle (Ch. 19), Bijaoui (Ch. 10) but find that I had little to add other than a philosophical point I believe I learned from David Donoho: Wavelets, Fourier series, wavelet packets need to be thought of first in terms of their utility as *sparse* representations of the signals you expect to have

in astronomy. Fitting in the presence of noise and statistical inference are of interest only after you have a model for the unobservable ideal which captures the features you expect (and has the flexibility to reveal features you don't expect!).

A second pair of papers by Murtagh and Aussem (Ch. 7), and Wegman et al. (Ch. 11) on large databases and algorithm resources are far from my sphere of competence. I enjoyed the statistical papers by Rao (Ch. 1), Efron, Akritas (Ch. 6), and Yang (Ch. 5) but have little to add.

My lack of knowledge and/or some exhaustion prevented me from commenting intelligently on the papers of Martínez (Ch. 9), Segal (Ch. 4), and Swank or on the many excellent posters displayed at the conference.

20.4 Bridging the interdisciplinary gap

As this conference and its predecessor demonstrate astronomers are excellent scientists eager to adopt statistical methods which are useful and able to implement and interpret these methods on more than a naive level.

So where can statisticians contribute? Certainly not in instrument design, basic physics and engineering. Primarily, it seems to me on two fronts

- a) By probabilistic stories – helping in the conceptual structuring of aspects of massive data sets
- b) By introducing novel methodology sometimes transported from other fields.

To a lesser extent statisticians may help in algorithm construction and software implementation, but the level of skills of astronomers here is, I think, as high or higher than most statisticians.

What does working with astronomers and astronomical data do for statisticians? It exposes them to fascinating substantive questions, very complex and large bodies of data calling for new methods which may well be transportable to other fields and enrich the base core of statistics as has happened in the past.

So I think the development of astrostatistics is very desirable and I encourage my statistician colleagues to find questions and collaborators, not usually in that order. I do expect further cooperation as at Penn State and to some extent Stanford and Berkeley. Nevertheless, I share Paul Hertz's doubts that collaborations will sprout all over and all will be happy. The reasons are

- Too many astronomers with interesting problems
- Too few statisticians
- Too little time.

Statistical questions are coming to the fore in many fields and most statisticians are fully committed.

It does seem that there is another approach worth thinking about: Astronomy graduates with research expected to have a substantial statistical component should consider taking applied (and even theoretical!) statistics and probability courses obtaining M.A.'s in statistics or even joint Ph.D.'s. Vice versa, statisticians looking towards astrostatistics should consider MA's in astronomy and joint Ph.D.'s. But that's harder without an undergraduate background in physics or astronomy.

There is, I believe, at least now a crass commercial incentive for astronomy graduate students to take Master's or higher degrees in statistics. Statistics is a transportable set of ideas giving a perspective and skills valuable from finance to molecular biology, from public policy to astronomy ... and, of course, to core statistics itself to which I would hope we may be able to attract some renegade astronomers.

REFERENCES

- [A 89] Aldous, D. **Probability Approximations via the Poisson Clumping Heuristic**. Springer Verlag (1989).
- [BB 96] Bickel, P. J. and Bühlmann, P. What is a linear process? To appear in *Proceedings National Academy of Sciences USA* (1996).
- [BFOS 84] Breiman, L., Friedman, J., Olshen, R. and Stone, C. **Classification and Regression Trees**. Wadsworth (1984).
- [J 61] Jeffreys H. **Theory of Probability**. 3rd Edition. Oxford U. Press (1961).
- [FB 92] Feigelson E. D. and Babu, G. J. **Statistical Challenges in Modern Astronomy**. Springer Verlag (1992).
- [G 52] Gamow, G. **The Birth and Death of the Sun**. Viking Press (1952).
- [DLR 77] Dempster, A., Laird, N., and Rubin, D. Maximum likelihood estimation for complete data via the EM algorithm. *JRSS B* **39**, 1-38.
- [BEP SW 70] Brown, L., Eagon, J., Petrie, T., Soules, G., and Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov Chain. *Ann. Math. Statist.* **41**, 164-171.
- [D 92] Daubechies, J. **Ten Lectures on Wavelets**. Society for Industrial and Applied Mathematics (1992).
- [G 81] Grenander, U. **Abstract Inference**. J. Wiley (1981).
- [B 79] Box, G. E. P. Some problems of statistics and every day life. *J. Amer. Statist. Assoc.* **74**, 1-4 (1979).
- [CH 92] Chambers, J. M. and Hastie, T. J. **Statistical Models in S**. Wadsworth/Brooks Cole (1992).
- [LZ 86] Linhart, H. and Zucchini, W. **Model Selection**. J. Wiley (1986).
- [B 95] Breiman, L. Bagging predictors. *Machine Learning* (in press) (1996).

- [L 86] Lehmann, E. L. **Testing Statistical Hypotheses**. Second Edition. J. Wiley/Chapman Hall/Springer Verlag.
- [MT 77] Mosteller, F. and Tukey, J. W. **Data Analysis and Regression**. Addison Wesley (1977).
- [BRG 90] Brown, L. D. and Gajek, L. Information inequalities for the Bayes risk. *Ann. Stat.* **18**, 1578-1594 (1990).
- [GL 95] Gill, R. D. and Levit, B. Y. Applications of the van Trees inequality: a Bayesian Cramer-Rao bound. *Bernoulli* **1**, 059-079 (1995).
- [KT 96] König, J. and Timmer, J. Analyzing x-ray variability by linear state space models. *Astronomy and Astrophysics*. To appear (1996).
- [M 81] Miller, R. G. **Simultaneous statistical inference**. 2nd Edition. Springer Verlag (1981).

21

Late-Night Thoughts of a Classical Astronomer

Virginia Trimble¹

ABSTRACT Conference participants from the astronomical side of the fence were astounded by the wide range of statistical concepts and techniques available, at least in principle, for use. Complimentarily, those from the statistical side expressed surprise at the enormous range of kinds of astronomical data (in spatial, temporal, wavelength, and other domains) crying out for more sophisticated analysis. Nevertheless, bringing the two together is not to be done over afternoon tea, and real collaborations, not just advisors, are needed.

21.1 Introduction: The astronomical data base

I arrived at Penn State with a prediction at hand: “Nobody is going to learn anything at this meeting.” That is, I expected that the astronomers would not actually be able to carry out any analyses that they hadn’t been able to do before, and that the statisticians would not be able to take home any specific databases and do anything with them. I believe this turned out to be at least roughly the case. What then was the purpose in coming? To find out what is out there, in terms of methods and problems and, perhaps more important, who is out there who knows something about the techniques or the data that one might be interested in. One of the speakers reminded us that statisticians should not be regarded as shoe salesmen, who might or might not have something that fits. But at least some of the leather merchants have had a chance to meet some of the shoemakers.

The rawest possible astronomical information consists of individual photons (or the radio-wavelength equivalent in Stokes parameters) labeled by energy, time of arrival, two-dimensional direction of arrival, and (occasionally) polarization. Such photons may come from the sources (stars, galaxies, ...) you are interested in, but also from unwanted sky, telescope, and detec-

¹Physics Department, University of California, Irvine CA 92697-4575 and Astronomy Department, University of Maryland, College Park MD 20742

tor backgrounds, each also with its own temporal, spatial, and wavelength patterns, often not well described as standard Gaussian noise.

Sources, telescopes, detectors, and backgrounds are very different beasts in the different wavelength regimes: radio, microwave, infrared, visible, ultraviolet, X-ray, and gamma-ray. For instance, both spatial and energy resolution are generally poor for X- and gamma-rays, but you can record photon arrival times to a microsecond or better. Radio interferometry buys you enormous spatial resolution, but at the cost of losing extended emission. Ground-based work in the infrared is always background-limited, because air, ground, telescope, and all the rest are 300 K emitters. Excellent wavelength resolution is possible for visible light, but only if you are prepared to integrate the photons from faint sources for many hours. Unless “source” photons greatly outnumber “background” photons, statistical care is already needed at the raw data level to decide whether there is a source real enough to worry about, whether it is compact or extended, and where it is in the sky.

The next level of processing takes the individual photons coming from some part of the sky and assembles them into spectra (number of photons or brightness *vs.* wavelength during a given time interval), light curves (brightness *vs.* time in a fixed wavelength band, that is, time series), or maps (brightness *vs.* position in the sky during a given time interval and in a fixed wavelength band, also called images or pictures). At this stage, one asks statistical questions about the reality of particular emission or absorption features in spectra, about whether a source is truly variable and whether periodicity can be found in the light curve, and about the reality and morphology of apparent features in the maps. An important question at this level is whether adding another parameter to your fit (emission line, pulsation mode, subcluster, ...) improves the fit enough to be believable. Astronomers have historically used the chi-squared test for this purpose, and indeed are typically not aware of any alternatives. One virtue of Bayesian methods is that they automatically vote against extra parameters for you unless the addition makes an enormous improvement.

Stage three compares processes spectra, light curves, maps, etc. with previously-existing templates and assigned sources to known classes like peculiar A stars (ones with surface temperatures near 10,000 K and anomalously strong absorption lines of europium and related elements) or BL Lac objects (active galactic nuclei characterized by rapid variability and weak emission lines). Often data from more than one wavelength band must be combined at this stage to recognize classes like low-mass X-ray binaries (pairs of stars with one neutron star or black hole accreting hot gas from a solar-type companion), gamma-ray pulsars (which show the same rotation period at radio and gamma-ray wavelengths), or Fanaroff-Riley type 2 radio galaxies (ones whose radio emission is very bright and has a particular double-lobed morphology). The main statistical questions at this stage concern whether the chosen template is a “good enough” match to

the new object (given the noise level), would another template be better, or have you found a genuine new class of source? In this last case, the two most immediate goals are (a) publish in *Nature* and (b) find two more, at which point, with three examples, you have discovered a well-known class of astrophysical object.

At stage four, we are ready to begin carrying out serious astronomical research, looking for correlations of properties among objects within and between template-defined classes in spatial, temporal, or other domains. Considerable theoretical firepower must be brought to bear at this stage. For instance, (a) given the distribution of brightnesses and colors of stars in a nearby galaxy, a luminosity-mass relation, and calculated lifetimes of stars as a function of mass, figure out the history of star formation $N(M, t)$ in the galaxy; (b) given the range of galaxy types and the current chemical composition of the X-ray emitting gas in a cluster of galaxies, figure out the history of nucleosynthesis in the cluster; or (c) given the sky positions and redshifts of a large number of galaxies, decide whether the pattern could have arisen from a particular spectrum of perturbations in a universe dominated by cold dark matter. Noise from all the previous levels is, of course, still with us, and the statistical nature of many conclusions shows in their being expressed in the form “we can exclude constant star formation rate (or dominance of Type Ia supernovae, or mixed dark matter) at the 95% confidence level”.

Invited review talks addressed problems at all of these levels, from deciding whether even one photon in a box belongs to your source (Jefferys, Ch. 3) to the topology of very large scale cosmic structure (Martinez, Ch. 9).

21.2 Astronomers on their own

The astronomer on the street is likely to be of the opinion (a) that statistics began with Gauss and (b) that the one thing we all understand is the least squares method. Prof. Rao’s (Ch. 1) historical introduction quickly disabused participants of both of these illusions. Soon thereafter, we became aware that the first two associations that the word “statistics” triggers in many of our minds (statistical parallax and Bose-Einstein, Fermi-Dirac, etc. statistics) were not terribly relevant to the purposes of the meeting, though the Luri et al. poster (Ch. 39) dealt with statistical parallaxes.

As always nomenclature (or jargon if you want to be insulting), including mathematical notation, was a significant part of the communications barrier between the two communities participating. While reading the poster presentations, I started making a list of the terms that I was not prepared to define accurately. When the total reached 99 (Table 21.1), I stopped, keeping in mind the Islamic tale that The Almighty has 100 names, of which

99 are known to mankind. The camel is so smug because only he knows the 100th name. At this point I encountered the word heteroscedastic, and can say only that the camel deserves it. Undoubtedly a comparable set of mysterious phrases from the astronomical dictionary could have been compiled with equal ease. Always mysterious are the eponyms (ideas, classes, methods, etc. named for a discoverer). Some terms seem to have an obvious meaning that is clearly wrong, like adaptive regularization (marrying your mistress before she shoots you?), and oriented pyramids (toward the east?). Others sound remarkably oxymoronic from outside (decision tree, annealing simplex method).

Astronomers, we have told ourselves repeatedly and have been told by other sorts of scientists and mathematicians, are particularly poorly equipped with statistical tools and careless in the use of the ones we have. I think there is some truth in this – scanning of several weeks issues of the *New England Journal of Medicine* yielded about twice as many separate methods and concepts as a couple of issues of the thrice-monthly *Astrophysical Journal* (which is **much** thicker). And, apart from Student's t-test and a couple of rank, regression, and correlation analyses, there were not the same methods either.

This is not to say that we are totally hopeless when left to our own devices. Eponymous examples include Malmquist (1920, 1924, who had a bias), Scott (1956, who had an effect), and Lutz & Kelker (1973, who had a correction). Malmquist noted that flux-limited samples (i.e. ones censored/truncated by the number of photons you can gather from a source in the allowed time) will always be deficient in intrinsically faint objects far away. Scott noted that, only when you have really big samples which requires looking at distant objects, will you find the rarest (including brightest) examples of a class. Lutz and Kelker pointed out that there are many distant stars with small parallaxes and few nearby ones with large parallaxes. Thus even symmetric errors will make it look like there are more nearby stars than is really the case. Errors are asymmetric in fact (since no parallax can be negative), making it worse. All three of these items deceive you into thinking that a distant set of stars, galaxies, etc. is closer than it is, and the first remains a major unresolved issue in establishing the extragalactic distance scale. If there are three cosmologists in the room, at least one will believe that at least one other does not understand Malmquist bias.

All three can be thought of as examples of truncated or censored data, many different approaches to which were discussed by Yang (Ch. 5), Segal (Ch. 4), Jefferys (Ch. 3), Duari, and Caditz (Ch. 38), but none is usually thought of as a statistical issue by practicing astronomers. We see them as akin to physical effects, like interstellar absorption, which you ignore at your peril.

Three more obviously statistical items were mentioned during the meeting that seem also to have been first thought of within the astronomical

TABLE 21.1. A subset of unfamiliar statistical terminology

hyperparameter	entropy prior
segmentation	Kahonen net
neural network	non-parametric
correlograms	differential distribution
gapped time sequence	siftware
multifractal	flatness spectra
adaptive regularization	Holden experiment
Slepian sequence	Gibbs sampler
à trous algorithm	embedding dimension
scalogram	scalegram
noise grams	errors of realization
maximum entropy methods	periodogram
autoregressive	multiscale entropy
pyramid, Gaussian	intermittency
pyramid, Burt	random coefficient
pyramid, Laplacian	Lebesgue density
pyramid, gradient	Fourth Minkowski functional
pyramidal vision	vision models
pivotal	i.i.d.
Huber estimator	intrinsic Bayes factor
minimum distance estimator	downhill simplex method
genetic program	annealing simplex method
multiscale edge detector	Lipschitz exponent
Lyapounov exponent	Sterrs
Gabor transform	gaussian window function
Boniferroni inequality	nested Lampton method
Ruger inequality	linear least squares estimator
gamma kernel	normalized fluctuation spectrum
Mallat's algorithm	Sobolev constraint
Wiener interpolator	Sterne-Campbell contingency test
binomial assumption	second-order intensity function
pursuit methods	Markov chain
supervised classification	Poisson regression
conjugate gradient algorithm	outliers
wavelet	axis-parallel decision tree
mother wavelet	K-function
Morelet wavelet	C-function
Mexican hat wavelet	Occam function
Haar wavelet	posterior space
Daubechies wavelet	Greenwood's formula
self-organized mapping	probability map
cumulative hazard rate	Heaviside's step function
B-spline function	Poisson bias
censored	proper weighting function
truncation	regressogram method
Lynden-Bell estimator	fuzzy correspondence analysis
Epanechnikov kernel	basis pursuit
linear state scale model	matching pursuit
Metropolis-Hastings algorithm	Wilcoxon statistic
Fisher matrix	

community. Historically first is $\log N - \log S$ (Ryle 1955), a plot of number of sources with brightness equal to or greater than flux S . Historically, it was used to show that radio galaxies and quasars have changed their average properties over billions of years. It is an integral method, with the advantage of not introducing binning noise but the disadvantage of propagating errors forward through the whole distribution.

$P(D)$ (Scheuer 1957) means “probability of deviation” and comes from the language of radio astronomy. It has been independently (re)discovered in the optical community as the method of surface brightness fluctuations for measuring distances to galaxies. The idea is that, when you try to count faint things (stars, radio sources, etc.) there get to be so many that you cannot resolve them as individual objects. But you can still say something about the number as a function of apparent brightness by looking hard at the fluctuations across the sky of their summed brightness, assuming that positions are random and the number of objects per resolution element is therefore describable by Poisson statistics.

Finally, V/V_m (Schmidt 1965) is a way of learning about how some sort of astronomical object is distributed in space even though your data sample is truncated in two or more dimensions (radio flux and optical brightness for the original quasar case that Schmidt considered). A modification can be used to construct luminosity functions from flux-limited data. Both V/V_m and $\log N - \log S$ have been used more recently to rule out certain possible models of gamma ray bursters.

21.3 Some common problems

This section includes most of the specific astrophysical issues discussed at the symposium, classified (very crudely) in parallel with the phases of data analysis mentioned in §1. Many sections are introduced by quotations from the conference participants and other players in the statistical arena. The content of the conference occupied (at least) a two-dimensional surface, where the y axis = statistical techniques and the x axis = astronomical problems. This summary is necessarily one-dimensional, and the “across-the-rows-and-down-the-columns” format of §3–4 results in many presentations being mentioned twice.

21.3.1 Source detection and image restoration

“Most of the bins have zero [source] photons in them, which is difficult even for statistics.” W. Jefferys

Indeed, a very common problem is one of deciding about the reality of a source (spectral line, etc.) seen against a non-randomly noisy background. The Theiler & Bloch (Ch. 30) and Damiani et al. (Ch. 33) posters addressed

precisely this issue for X-ray sources, the former using a new interpolation technique and the latter wavelet transforms.

A closely related task is that of cleaning and restoring images that have been messed up by processes of known (non-Gaussian) properties. The Hubble Space Telescope, with its improperly figured main mirror, has driven many recent efforts in this area. The Núñez & Llacer (Ch. 28) poster considered Bayesian methods, the Starck & Pantin poster (Ch. 29) multiscale maximum entropy methods, and the Anderson & Langer poster an existing package of pyramid and wavelet methods for image reconstruction. Wavelets are particularly appropriate for noise suppression (one form of which is called flat-fielding), and both Murtagh's talk (Ch. 7) and Kashyap's poster (Ch. 34) described wavelet approaches to simultaneous noise removal and source identification.

21.3.2 *Image (etc.) classification*

"If you have a spectrum, you immediately know if it's a star or a galaxy."
R. White (This is funny only if you know about cases like BL Lac, whose name says that it was first classified as a variable star and only later as the prototype of a subtype of active galactic nuclei, because its spectrum and light curve, considered in isolation, were every bit as ambiguous as its image.)

OK, you have decided you have a real source, and you are reasonably certain (perhaps even rightly so) that it belongs to one of a small number of discrete classes. How can you carry out the classification efficiently and automatically? Three problems of this sort were addressed. In each case, one starts with some sort of training set of images or spectra known to be correctly classified and an algorithm for deciding which class a new example belongs to (neural network, nearest neighbor, decision tree, matrix multiplication, rank order, or many others). The choice of critical parameters for distinguishing classes can be made by the programmer or by some of the types of programs.

White (Ch. 8) considered the most basic distinction – is an image that of a star or a galaxy (or plate flaw or cosmic ray hit)? – and a number of ways of selecting training sets and classification parameters (e.g. the rank-order of ratio of central to total brightness of an image in a field as a star/galaxy criterion). The Nail poster dealt with the next stage of classification, separating galaxies into normal and peculiar. Closely related is a recent exercise coordinated by Opher Lahav at Cambridge [*Science* 267, p.859, 1995], in which six senior extragalactic astronomers classified a number of galaxy images on a fine grid of subtypes, and then an artificial neural network did the same. It was roughly a tie. The Bailer-Jones poster (Ch. 42) presented a fairly preliminary attempt to automated classification of stellar spectra with a neural network. The "official" MK types are defined primarily by a small set of examples and secondarily by a small number of

specific line strength ratios. It was not clear whether the network had access to these official rules or was expected to find its own from the training set. The difficulty it found in separating type IV (subgiant) stars from main sequence and giant ones suggests the latter.

21.3.3 Pattern recognition and description

“What would you think if you looked at the 11 brightest constellations and saw 7 Big Dippers?” H. C. Arp, 1965 (in connection with arguments for non-cosmological redshifts of quasars). *“Well, you do see three.”* J. E. Gunn, 1965 (or even four – the Big and Little Dippers, Pleiades, and great square of Pegasus).

The topics mentioned here are distinct from those of §3.2 in that one does not start with a small set of *a priori* templates, but rather is trying to decide whether there is any interesting structure (in an image, spectrogram, time series, or whatever) and, if so, how should it be described.

The Lawrence et al. (Ch. 35) and Turmon (Ch. 31) posters both considered the patterns of magnetic activity on the solar surface from this point of view, using different statistical methods (wavelets, multifractals, and Markov random fields), but coming to rather similar conclusions, that the patterns are too complex for either verbal description or modelling to be very successful. The length scales on the Sun range from a little less than 1000 to about 1,000,000 km. A very similar problem, but on a scale of $10^{19} - 10^{22}$ km, arises when you attempt to characterize the large scale distribution of galaxies in the universe in terms of clusters, voids, filaments, sheets, or whatever. Betancort-Rijo’s poster (Ch. 26) considered this (though the inclusion of the phrase “random fields” in the title leaves me puzzled), as did the talk by Martínez (Ch. 9). One well-defined question is the basic topology of the large scale structure. Are clusters studded through spade like meatballs, voids scattered through denser regions like holes in swiss cheese, or something else? The answer, according to Martínez, seems to be a sponge-like topology, so that both over-dense and under-dense regions are connected up.

Pattern recognition is not, of course, a problem unique to astronomy, but has numerous applications in military reconnaissance, machine reading of handwriting, and so forth. Many of the methods developed in these areas, however, make heavy use of edges in the field, not very appropriate for astronomy, most of whose structure dribble off from dense cores to tenuous envelopes.

Whatever data set and methodology is under consideration, one has, as van der Klis (Ch. 18) reminded us, to worry about two different kinds of errors, equally serious. The first is seeing patterns that aren’t really there. Martian canals and other examples have made astronomers sensitive to these. Not seeing something that really is there is less embarrassing, but just as likely to impede physical understanding. These are called type 1

and type 2 errors, and, like Faranoff-Riley type 1 and type 2 radio galaxies, I forget which is which. But, in case you should ever need to know, Type II supernovae are found among population I stars, and type 1 Seyfert galaxies have broader emission lines than type 2's.

21.3.4 Fitting known functions

“The need for statistics arises because nothing in life is certain except death and taxes, unfortunately not in that order.” Anon.

Function fitting is the quintessential problem in statistical astronomy, since we were all taught that Gauss invented least squares in order to reduce a large number of observations of the first asteroid, Ceres, to a single best orbit. Newton's laws guarantee that (apart from perturbations due to Jupiter and such) the orbit can be described with a handful of numbers for semi-major axis, eccentricity, inclination of the orbit to the plane of the ecliptic, longitude of perihelion, and time of perihelion passage. To generalize from the solar system to any other pair of point masses you add the two masses as numbers 6 and 7. Dikova's poster (Ch. 48) concerned an outgrowth of Gauss's problem, identifying groups of asteroids with similar orbit parameters, while Ruymakers & Cuypers (Ch. 40) considered the case of the reliability of the orbits of binary stars, but using a bootstrap method, not least squares. Acuna & Horowitz showed that you can fool some of the photons some of the time. If the image of a single point source seen with a given telescope/detector combination is very accurately known, then you can reliably recognize and assign brightnesses to two point source images even when their angular separation is somewhat less than the traditional Rayleigh criterion.

A great many other astronomical tasks are also of this general form, because one knows in advance, for instance, that the widths of magnetically broadened hydrogen lines will be linear in field strength, while their centroids shift in proportion to the square of the field; rotational broadening is linear in rotation speed, while turbulent broadening scales with the square root of the masses of the atoms responsible for the line; and so forth.

21.3.5 Fitting unknown function, additional parameters, and goodness of fit

“With five parameters, you can fit an elephant.” George Gamow (attributed)

Siemiginowska et al. (Ch. 14) asked a number of critical questions in this area, drawing from examples in X-ray astronomy, though they apply at all wavelengths. How do you find your way through many-dimensional parameter space to a “best fit”? Support the template you are trying to fit itself has uncertainties (e.g. in atomic data for spectral lines); how can

this be included in error estimates in Bayesian and frequentist methods? What should replace chi-squared as a test of goodness of fit when most of the information is in a few data points? (The Wilcoxon test was mentioned by several speakers as being suitable for picking out regions of largest deviation.) And so forth. Wheaton's poster (Ch. 41) suggested an alternative weighting scheme to be used in place of the standard chi-squared one for the specific case of bins with very few photons in them.

A common astronomical approach to the problem of finding relationships when you don't know what to expect in advance goes back to the early days of radio astronomy and says that, when you don't understand a phenomenon, the first thing to do is to plot it on log-log paper. Indeed, plots of \log this *vs.* \log that quite often yield clusters, correlations, and principal planes (for instance the Tully-Fisher, $D_n - \sigma$, Faber-Jackson, and Fundamental Plane methods of measuring distances to galaxies). The disadvantage is that one is often left with relationships that are not in any way physically understood. Mukherjee & Kembhavi's poster on the decay of pulsar magnetic fields considers a problem of this type. I believe that Qian's poster presents a relevant methodology, but found it difficult to interpret.

Other posters that addressed looking for correlations, structures, etc. that are not understood in advance include those of Starck & Pankin, Luri, Arenou et al., Chernenko, Csabai & Szalay, Betancort-Rijo (*cf.* Ch. 29, 39, 46, 47, 43, 26), and Schaefer et al. (not actually posted, though represented in the abstract booklet, and perhaps containing the last Fourier transform in the world). Most of these are also mentioned later under types of tests.

By the way, you really can fit an elephant with five parameters. Or, at least, the contour of the upper outline from trunk to tail tip can be traced fairly well with portions of five sine waves of different wavelength and amplitude.

21.3.6 Time series

“Yes, I can see the canals.” J. P. Ostriker contemplating one of the results of J. D. Scargle's 1969 thesis analysis of periodic motions of wisps in the Crab Nebula.

Time series (or light curves of various kinds) was the astrophysical problem addressed in the largest number of talks and posters. It includes finding periods and power spectra, especially for data with large gaps, irregular intervals, or intervals near an actual period (leading to aliases). Because the astronomical community is small (and there are lots of stars), the loss of one productive observer can create such a data set from what would otherwise have been a much more satisfactory one. We can call this the ‘Grant Foster effect’, since he mentioned it during a discussion section. I had previously thought of it in connection with the Crab Nebula as “and then Baade died and stopped taking plates so often”.

The intuitive method of period fitting has had some spectacular failures. In the 1930's for instance, Harlow Shapley found a number of variable stars in the Small Magellanic Cloud (SMC) that he thought should be RR Lyrae variables. And indeed with a few dozen irregularly spaced observations per star, he was able to fit every one of them with a reasonable RR Lyrae period (0.2 to 0.8 days or thereabouts). This implied a small distance to the SMC consistent with his previous ideas based on Cepheid (brighter) variables, and so helped to preserve a wrong extra-galactic distance scale for about 20 more years, until David Thackeray finally found the real SMC RR Lyraes in 1952, a factor of four fainter than Shapley's stars. Subsequent studies showed that each of Shapley's variables had a real period longer than 1 day, implying that the stars are larger and brighter than RR Lyraes and the SMC much further away than he thought.

A number of posters focussed on time series analysis for particular objects or classes, including Vityazav (Ch. 25) (correlograms as a problem in aperture synthesis), Lawrence et al. (temporal patters of solar activity with wavelets and multifractals, Ch. 35), Lehto (light curve of the gravitationally lensed quasar OJ 287 with Haar wavelets, Ch. 36), Koenig (X-ray variability of the active galaxy NGC 5506 with linear state space models, Ch. 24), Young et al. (gamma ray bursts with wavelets, Ch. 37), Schaefer & Bond (quasi-periodic oscillations in AM Her binaries with Gabor transforms – like any eponym, this one invites the question, Eva, Magda, or Zaza?), Nair & Principe (long and short term predictions with neural networks for 3C 345 and PG 1159 stars), Hertz (reality of orbit period changes in X-ray binaries using several tests, Ch. 23), and Ringwald et al. (possible long-term variability of unknown origin in the cataclysmic variable GK Per; these data are nearly continuous and uncensored, but the authors still feel some residual uncertainty about the reality of the hecto-day oscillations, which were not at all expected).

Five of the oral presentations also concerned time series analysis for a variety of objects using a variety of methods. Van Leeuwen (Ch. 15) was concerned with the problem of enormous gaps in a time series in connection with trying to characterize variable stars from the HIPPARCOS data base (not, of course, the primary purpose of this astrometric mission). This has indeed proven possible in a somewhat limited way (though the inventories are small compared, for instance, to the variable star byproducts of MACHO and OGLE searches for gravitational lensing). But his primary concern was with later astrometry missions and how them might be designed to provide more useful light curves without compromising the primary goals.

Guégan (Ch. 17) addressed embedding dimensions, Lyapunov exponents, and other ways of recognizing chaos (meant as a technical term, not a description of the normal state of astronomical research). A strictly positive Lyapunov exponent, for instance, means that two trajectories starting with arbitrarily similar initial conditions can eventually wander arbitrarily far

from each other. She noted that astronomy has one advantage over other disciplines (e.g. economics) in which chaotic behavior has been sought in having at least one authentic example: the tumbling rotation of Saturn's moon Hyperion (the analysis of which makes use of data collected as far back as the mid 1920's).

As Thomson pointed out, there is more than one way to handle gapped time series data. One set of methods looks only at the actual data points (e.g. the Lomb-Scargle periodogram), while the other begins by interpolating to fill in the gaps. Thomson favors interpolation methods and showed examples of their application to the light curves of the two main components of the gravitationally lensed quasar OJ 287 and to variability in the solar wind as charted by the Ulysses "over the pole" mission. The former bears on the value of the Hubble constant, the latter on the existence and properties of certain solar oscillations. Swank and van der Klis (Ch. 18) both addressed X-ray light curves, especially the ones with extremely fine time resolution now coming from the Rossi X-ray Timing Explorer (XTE or RXTE) and eventually to come from the Advanced X-ray Astrophysics Facility (AXAF). A wide range of high frequency (up to kHz) phenomena are turning up, and one needs to be able to answer questions like how to identify the shortest time scale present, how to set upper limits in power spectra, and how to characterize red noise in the presence of low frequency leakage, before the "botany" stage of studying X-ray binaries and other sources can yield astrophysical understanding.

21.3.7 Recognizing rare events and new classes of sources

"If you believe something one sigma is enough; if you don't, then 15 sigma won't help." Richard P. Feynman c. 1975

A surprising number of astronomers have worked over the years on what many of their colleagues believed to be empty data sets. Examples include the first ten years of extra-solar-system gamma ray astronomy (until the first source was found), SETI (the search for extraterrestrial intelligence), brown dwarfs (until about a year ago), the extragalactic infrared background, and gravitational radiation (known to exist and do what Einstein's equations predict, but through indirect evidence on binary neutron stars, not through direct detection). In terms of the phases of astronomical data processing described in §1, these are problems in which the first task is to create a template with which candidate events or sources can be compared.

Axelrod (Ch. 12) discussed the searches for gravitational lensing of stars in the Large Magellanic Cloud and the galactic bulge by "MAssive Compact Halo ObjectS" or MACHOS (deliberately coined to provide a contrast with Weakly Interacting Massive Particles or WIMPs). Potential lenses include known stars, substellar objects, planets, or anything else in the disk or halo of our own galaxy with masses in the range 10^{-3} to 10 solar masses. The team began with a template for a MACHO event that required the

brightening and fading to be time symmetric, colorless, and other various definite things. Actual observations and further thought have required the template to evolve and include lensing of and by binary and moving stars (not time symmetric) and lensing of stars with blended images and finite disk size (not colorless), and so forth. Luckily all the raw data are being archived, and so it has proven possible to go back and re-evaluate all the observations with the new templates. It nevertheless remains quite difficult for the observers to evaluate precisely what their level of completeness is, and the methodology can be described as that of successive approximations. At least three other similar searches are underway; they are acronymed OGLE (Optical Gravitational Lens Experiment), EROS, and DUO.

Astone's poster (Ch. 22) dealt with an on-going search for gravitational radiation bursts using bar-type antennas located in Europe. Schutz's talk (Ch. 13) discussed the somewhat similar problems of the planned search using interferometric detectors to be located both in the US (LIGO) and in Europe (VIRGO and perhaps others). Here the problem is that the template must come from theoretical calculations of what you expect from the merger of a pair of neutron stars or the collapsing core of a supernova in terms of time-varying quadrupole potentials. Numerical general relativity is not yet fully equal to the task, and the theoretical part of the project is currently in the "method of recurrent simulations" stage.

Curiously, many participants were left with the impression that LIGO and VIRGO would constitute the first deliberate searches for gravitational radiation. In fact, J. Weber has operated two or more bar-type antennas continuously for more than 30 years, and he first described the design at a meeting on general relativity and gravitation in Warsaw in 1962. Peter Bergmann, who worked with Einstein, predicted at the time that it would be a century before anything came of the effort. He is, so far, 1/3 right.

A related problem arises when you are carrying out one of the classification projects mentioned in §3.2 or a corresponding program in the time domain and discover that none of your templates is a good fit to a particular object or image or light curve. The poster of Arenou et al. (Ch. 46) concerned one of these cases. They have been involved in the primary (astrometric) part of the HIPPARCOS mission and discovered that a few of the objects in the catalog acted neither like single stars nor like double ones. They had in fact (re)discovered what are called astrometric binaries, ones where you see only a single image, but the orbital image makes it wiggle around in the sky over a few years.

21.3.8 One population or two?

"Alan of Lille (de L'Isle): Alan of Tewksbury? Alan of Tewksbury: Alan of Lille?" Robert K. Merton 1965

An astronomer quite often finds himself holding two handfuls of beans and wanting to know whether they came out of the same bean bag. Exam-

ples include active galaxies selected at different wavelengths (radio, optical, X-ray); the Lyman α forest of absorption at low *vs.* high redshift; the planetary nebulae in different types of galaxies; the chemical compositions of meteorites *vs.* those of asteroids; quasars with and without broad absorption lines; and many others that would take more words to explain. Problems of this sort were not discussed, apart from Rood's poster, which attempted to decide whether gamma ray bursts and Abell clusters are drawn from the same parent population (his answer is, partly, probably).

Sometimes one population is a real set of observed sources (etc.) and the other a theoretical model that predicts a range of properties. In this case, the customary astronomical approach is the Kolmogorov-Smirnov test, which is, however, not necessarily the best available.

Deciding whether an apparent cluster of data points can or cannot have arisen by chance out of a smoother underlying population is a somewhat similar problem. Efron's example (childhood cancers in California towns) was not astronomical, and Berger's (Ch. 2) "bivariate observations" came from an astronomical, but otherwise undefined, context. We were assured that the former cluster was not significant and that the latter ones were.

21.3.9 Censored and truncated data: The problem of sample selection

"Most mistakes are made before ever putting pencil to paper." VT (frequently)

Which is which, for starters? Censored data occur when patients are lost to follow-up (but you have no reason to support *a priori* that they are different from the ones you can discuss). Truncated data occur because patients enter a trial after the first year, and so you have only 3 or 4 or 5 years of follow-up, not the full six years of the study (and, having entered later than the others, they might be somehow systematically different). To take an astronomical example, if you search for X-ray quasars by staring with an optical sample, then your data on non-detected ones will be censored — you will know that the quasars exist and that they have fluxes less than your detection limit, but you will not know the actual flux values (and have no reason to suppose that they are systematically different from the nearby sources whose fluxes you can measure). An X-ray search will lead to a truncated sample — below your flux limit, you will not only have no numbers, but you will also not be sure that any such objects exist.

One must be clear at the outset that there are no foolproof ways of dealing with these situations, except to work harder and find the missing or undetected objects (and then, of course, you will try to analyze the new, larger set, and be right back where you start from, except with a lower flux limit). Thus discussions of statistical methods can address only which approaches are likely to be best for a given, physically-defined, situ-

ation, and, very importantly, which estimators are equivalent to or reduce to others in specific situations. Yang (Ch. 5) provided a quite-technical review of some ways of dealing with such incomplete samples, including the Kaplan-Meier (two-people) survival curve for right censored data and the Lynden-Bell (one person) estimator for non-parametric analysis of incomplete data. Apparently all methods that bin data points reduce to the Lynden-Bell method where there is only one point per bin. Caditz's poster (Ch. 37) considered a smoothing (*vs.* binning) non-parametric approach to estimating luminosity functions from truncated (flux-limited) samples. The topic was also reviewed by V. Petrosian in SCMA I.

Many real astronomical samples are bounded in such complex ways that knowing how to cope with "simple" truncation or censoring is not very much help. In these cases, what we really need is guidance on how to choose the samples of stars, galaxies, or whatever to look at. The brute-force approach of gathering enormous quantities of data and throwing out all but the many-dimensional corner of completeness is essentially never possible.

A typical project (one I have been involved in for 20 years or so) is attempting to determine the distribution of mass ratios of binary stars (as a guide to understanding the processes of star formation). Recognizing that a star is a binary and getting enough information about it to know the mass ratio is dependent on apparent brightness (that is, real brightness/mass and distance), the period of the binary (with two or three peaks of high detectability), the mass ratio itself, the eccentricity of the orbit, the evolutionary stages of the two stars, and probably other things; and there is every reason to expect the mass ratio distribution to be a function of at least some of these. The result is that people who have examined different subsets of archival data and/or added to it get wildly different answers and feel so strongly about them that we can only just barely be brought to acknowledge that it is a problem in sample selection (and no samples are perfect), and not one in abnormal psychology.

21.3.10 Data compression and storage

"Archiving is easy; retrieval is difficult." Speaker at a 1993 IAU Symposium on Wide Field Imaging

Actually, with the advent of terabyte projects like the Sloan Digital Sky Survey, MACHO searches, 2MASS, and so forth, even archiving isn't all that easy, resulting in the need for very efficient ways of compressing data with little or no loss of information. Murtagh and Scargle (Ch. 7 & 19) in their talks both mentioned that wavelets show great promise for efficient compression and storage of many different kinds of astronomical data.

21.4 Powerful methods

*"In my head are many facts that, as a student. I have studied to procure.
In my head are many facts of which I wish I was more certain I was sure."*
Oscar Hammerstein II (The King and I)

Historically, astronomers have found their correlations, templates, clusters, and all the rest by methods that could not be easily quantified or even taught, and turned to quantifiable, statistical ideas only afterward to answer the question "how sure am I that this is right?" For instance, only Fritz Zwicky could find Zwicky clusters on Palomar sky survey plates, but the catalog remains useful down to the present, and we can now say a good deal about how complete it is for different kinds of clusters at different distances from us. Kolmogorov-Smirnov, chi-squared, and other tests are meant to answer that question. In contrast, least-squares, by the time you push the last button on your calculator, has provided not only the coefficients you asked for but also some kind of error estimate. Some, but not all, of the more elegant techniques presented at the symposium simultaneously fit something to something and evaluate the goodness of fit.

21.4.1 Bayesian (*vs. frequentist*) statistical methods

"At some level we are all Bayesians" Michael S. Turner at a 1996 symposium on cosmic distance scales. *"Yes. Those who are not were run over by trucks at an early age."* VT. same symposium. By this was meant that we all carry around with us a great many prejudices about the relative likelihood of various outcomes to everyday experiments, like crossing the street during the rush hour, and that we would be worse off without them. The other side of this coin is that it is possible to come to believe something so firmly that no amount of later data can influence one's opinion; on both scientific and non-scientific issues.

My own impression (prejudice?) is that Bayesian methods are most useful when one expects to change one's mind only slightly. Jeffery's talk (Ch. 3) on use of lunar laser ranging data to improve our knowledge of lunar motions and earth rotation is a good example. In contrast, I do not think there has ever been a time (including now) when applying such methods would have improved our knowledge of the Hubble constant, for the problems have always been wildly wrong choice of *a priori* probabilities, neglect of important physics, and inappropriate data selection, and I suspect they still are. Religious conversion is probably also not amenable to Bayesian treatment (perhaps it is a critical point phenomenon?).

Berger's review (Ch. 2) of Bayesian analysis was, on the whole, comforting to an outsider, in the sense that it left impressions (a) that exact choice of prior probabilities makes relatively little difference if you have a reasonably large data set, (b) that modern computational methods make it relatively easy to get hold of the terms that enter into the posterior proba-

bilities, and (c) that a Bayesian answer will differ wildly from a frequentist one about how probable a particularly hypothesis/event/etc. is only under rather pathological circumstances. Pursuant to (a) and (b), he present a default option for choosing the prior distributions for the unknown parameters of each model, which can be used when you don't quite know what else to do. One of the surprises was that there are conceivable data sets that are not very probable under any hypothesis, and I am not sure whether the point being made was much more profound than that it is "very unlikely" that I shall see a license plate number AAA 222 on the street today, but no more unlikely than PZD 181.

Bayesian analyses of specific problems were presented in the posters of Núñez & Llacer (Ch. 28) (image reconstruction) and Kester (deconvolving camera and source contributions to spectra from the Infrared Space Observatory).

21.4.2 Wavelets and related transforms

"Mathematics may be compared to a mill of exquisite workmanship, which grinds you stuff of any degree of fineness; but, nevertheless, what you get out depends on what you put in; and as the grandest mill in the world will not extract wheat-flour from peascod [peapods], so pages of formulae will not get a definite result out of loose data." Thomas Huxley, 1869

And wavelets, the topic most discussed at the symposium, can grind exceedingly small (but, of course, still subject to the Huxley limit that, in modern language, is generally described as GIGO: garbage in, garbage out). The prominence is apparently a recent development. Wavelets had only a single index entry in SCMA I and no definition in the glossary. Thus I provide my own crude one: wavelet transforms are a lot like Fourier transforms, but you get more choices. In particular, the shapes can be chosen to be especially good at localizing edges, bumps, and discontinuities (at which sine and cosine waves are rather poor) and to pick out several widely spaced length or time scales where there is lots of information, without having to worry too much about the ones in between.

One or both of these virtues was conspicuous in most of the poster presentations where wavelets were applied to specific kinds of astronomical data. These included Lawrence et al. (temporal structure of solar activity, Ch. 35), Lehto (light curve of OJ 287, Ch. 36), Anderson & Langer (image processing), Young et al. (gamma ray burst time series, Ch. 37), Chereul et al. (the local gravitational potential determined from HIPPARCOS data, Ch. 32), Damiani et al. (source detection with photon counting detectors, Ch. 33), Kashyap (flat fielding and source identification, Ch. 34), and Bijaoui (multiscale analysis, Ch. 10).

The speakers addressed both general properties and specific applications of wavelets. Several of them mentioned that the "Mexican hat" is quite commonly used, but not entirely proper for most applications; Haar wavelets,

in contrast, are like falling off a log. Murtagh (Ch. 7) emphasized the problems of data selection and coding in large data bases, using a number of IAU spectra of the Seyfert galaxy NGC 4151 as an example of how principal component analysis can be achieved economically. Orthogonal and discrete wavelet transforms are particularly suited to some applications.

Scargle reviewed (Ch. 19) both methods and existing astronomical applications. These range from initial data processing through image and time series analysis to data compression, transmission, and storage. For instance (later than the time frame Scargle reviewed) the Mars96 camera data will be sent back in wavelet transform. Hazards abound: scalograms (analogous to power spectra in Fourier transforms) are not the same as scalograms (the absolute values of the wavelet coefficients), and “matching pursuit” is not at all what you might think. But his main point was that wavelets provide a whole new way of thinking about astronomical (etc.) data, and not just a library of specialized techniques.

Priestley (Ch. 16) addressed wavelet analysis for the study of time-dependent spectra and presented a number of caveats about things that wavelets cannot do, starting with the analog of the Heisenberg uncertainty principle, that you buy high resolution in the time domain at the expense of resolution in the frequency domain, and conversely. Bijaoui (Ch. 10) focussed on analysis of images that have interesting structures on many different scales, using, for instance, wavelets to pick out a spiral arm in a poorly-resolved galaxy image.

Self-respecting wavelets come in complete sets. There are also interesting transforms with respect to incomplete sets of basis functions, for instance the Gabor transforms mentioned by Scargle and used by Schaefer & Bond in their search for quasi-periodic oscillations in AM Her stars.

21.4.3 Neural networks

“If I see a sheep on a hill, I think there is likely to be a wolf nearby.” Leon Sibul (during panel discussion)

Artificial neural networks are programs/algorithms that are supposed to work somewhat the same what the human brain does. That is, one has a bunch of digital nodes (neurons) connected by input/output channels (axons and dendrites) through a bunch of synapses. and what goes out the axons depends in some complex way on what came in the dendrites. Artificial neural networks have in common with humans that they can, in some sense, learn from experience and get better at a task after doing it for a while. Eventually, perhaps, we should expect boredom to set in and then, as the networks become more human, they ought to start generating totally unexpected outputs, as, for instance, the wolf in a discussion of what you can infer about the colors of all sheep on the basis of seeing one that is black on at least one side.

Among the poster presentations, Nair & Principe applied ANNs to long and short time scale prediction, Naim (Ch. 44) to identifying peculiar galaxies, and Bailer-Jones (Ch. 42) to classification of stellar spectra. I believe that all poster contributors and invited speakers has applied biological neural networks (of the highest quality) to their presentations.

21.4.4 Other methods

“You can trust us ...” James Berger (at coffee break). If only because there are so many ways to do things, one of them must give the right answer.

Statistical techniques that were neither obviously Bayeslet or wavesian appeared in a number of posters. These included multifractals (Lawrence et al. on solar activity, Ch. 35), multiscale maximum entropy (Starck & Pantin on image reconstruction, Ch. 29), minimum distance indicators (Qian on autoregressive models), genetic programs (Kester on the ISO camera), maximum likelihood (Luri et al. on statistical parallaxes from HIPPARCOS data, Ch. 39), bootstraps and other Monte Carlo methods (Hertz on X-ray binary periods, Ch. 23, and Hesterberg on error estimation), two-point correlation functions and nearest neighbors (Brainerd on the existence of gamma-ray repeaters), and probably some others. I am not claiming with high confidence that all of these are truly distinct, only that they have different names.

21.4.5 Comparisons and relationships of methods

“How to lie with statistics” Title of a 1960’s book. *Statistics and Truth*
Title of a 1995 book (by C. R. Rao)

A particularly interesting aspect of the talks by Efron and Martinez (Ch. 9) was their comparisons between methods and applications from different universes of discourse. Efron alternated between biomedical and astronomical problems, showing, for instance, that survival curves are a good deal like $\log N - \log S$, and that hazard rates, Poisson regression, nearest neighbors, and Mantel-Haenszel or log-rank tests can be useful concepts in both.

Martinez attempted to identify the standard statistical methods that correspond most closely to a variety of techniques that astronomers seem to have invented for themselves. For instance, “our” two-point correlation function, $\xi(r)$, is quite similar to “your” second order intensity function. The speaker did not mention the histories of the various methods, so that no one could be tempted to say about anyone else, “Ah yes; he has re-invented the wheel. Only his is square.”

21.4.6 Plausibility tests

“Ask a silly question, you get a silly answer.” Virginia Farmer Trimble (author’s mother) c. 1953

That one’s results should make sense is obvious, though only about four presentations specifically mentioned the point (posters by Starck & Pantin, Valdes, and Hertz, which last revived two significance tests from the pre-war astronomical literature, and Sibul during one of the panel discussions, Ch. 29, 23). It is a very old problem that pops up again and again. For instance, the first spectrogram of quasar 3C 273, taken in 1963, had a handful of emission lines whose wavelengths were fit quite reasonably well by high ionization states of neon and such. The fit to part of the Balmer series of hydrogen was statistically no better, but, since hydrogen is the commonest element in the universe, there was really no competition.

Human beings make such judgments easily, quickly, and naturally (and quite often wrongly). Building physical sense into processing algorithms seems to be rather difficult, and many published astronomical papers mention a “biomechanical servo” stage, in which things that sound silly are thrown out of the sample. Ancient examples include pixels with negative fluxes, supernova candidates nowhere near any galaxy, spectra with no features, and binary systems in which the star you don’t see is more massive, and so should be brighter than the one you do see. This is, of course, a good way to fail to discover new classes of objects. And no, I don’t know what the solution is, except that a group of systems of the fourth type were once a set of black hole candidates called Trimble-Thorne systems. None actually has a black hole, but they all turned out to be interesting in terms of stellar physics.

21.4.7 Implementation and looking ahead

“The future is not yet in existence.” Edward Wegman

Most day-to-day analysis, from image processing on up to simulations of interstellar clouds and large scale structure of the universe, is done with standard software packages (and a very good thing, too, cf. the square wheels of §4.5). Wegman (Ch. 11) described a number of these packages, what they do, and how they can be grabbed out of the ether. Not everything he mentioned is younger than yesterday’s newspaper. Though the websites to which he provided points all date from the last couple of years, his paper references go back to 1953.

The scientific organizing committee left the participants with several “thought problems” in the area of what should be done next. Would an actual, paper, book of modern statistical recipes be useful? What about a web site, and who would maintain it? [*A web site with on-line statistical software for astronomy has now been initiated at <http://www.astro.psu.edu/statcodes>. Eds.*] How can we foster collaborations between statisticians and astronomers

that will be attractive to both in the sense of advancing basic knowledge in both fields so that each collaborator has something significant to add to his CV at the end? Participants nodded solemnly and promised to Think About it All.

Acknowledgments: It is the traditional prerogative of the last speaker at a conference to extend the collective thanks of the participants to all those who have made the gathering possible. Our sincere gratitude goes to Dean Geoffroy and Pennsylvania State University for their hospitality; to the seven sources of silver who provided financial support, NASA, NSF, ISF, IAU (but not much), Eberly College, ISI, and IMS; to Drs. Babu and Feigelson for putting all the science together; and to Debbie Noyes (and undoubtedly others) for help with logistic arrangements. We all look forward to gathering again at SCMA III in about 2001.

REFERENCES

- [1] Lutz, T. E. and Kelker, D. H., 1973, *Publ. Astron. Soc. Pacific* 85, 573.
- [2] Malmquist, G., 1920, *Lund (Observatory) Medd. Ser.* 2, No. 22, 32.
- [3] Ryle, M., 1955, *Observatory*, 75, 137.
- [4] Scheuer, P. A. H., 1953, *Proc. Cambridge Phil. Soc.* 53, 764.
- [5] Schmidt, M., 1968, *Astrophys. J.* 151, 393.
- [6] Scott, E. L., 1956, *Astron. J.* 61, 190.

Contributed Papers

Several dozen papers, primarily by astronomers, were presented at the conference on a wide range of specific projects and methodologies. ASTONE, HERTZ, KÖNIG and VITYAZEV discuss time series problems, and BETANCORT-RIJO, NAIR, NÚÑEZ, STARCK, THEILER and THURMON consider image analysis issues. Wavelet analyses of both temporal and imaging problems are addressed by CHEREUL, DAMIANI, KASHYAP, LAWRENCE, LEHTO and YOUNG. CADITZ, LURI, RUYMAEKERS and WHEATON present estimation problems using nonparametric, maximum likelihood, bootstrap and least squares methods. Multivariate classification issues are discussed by BAILER-JONES, CSABAI, NAIM and WALSHAW. The final papers cover outlier rejection (ARENOU), non-linear analysis (CHERNENKO), dynamical analysis (DIKOVA), point processes (HOROWITZ), and luminosity functions (ULLMANN).

Algorithms for the Detection of Monochromatic and Stochastic Gravitational Waves

Pia Astone¹

ABSTRACT Some aspects of data analysis are discussed in gravitational wave (g. w.) experiments obtained with the presently working cryogenic detectors, resonant bars of the type invented by Joe Weber. Until now the main scientific goal has been the burst detection. The improved performances, in terms of duty cycles, of the existing detectors and the expected sensitivities of bars and interferometers has lead us to consider also the search of other sources, such as pulsars and relic g. w.

The sensitivity of a g. w. antenna may be given in terms of its noise spectral amplitude, $\tilde{h}(f)$, in unit $1/\sqrt{\text{Hz}}$. Using $\tilde{h}(f)$ it is possible to compare different families of detectors and, for example, to derive the sensitivity of a bar-interferometer experiment. From $\tilde{h}(f)$ it is simple to obtain the detector sensitivity to:

Burst: $h = (1/\tau_g)\sqrt{S_h(f)/2\pi\Delta f}$ for a burst of duration τ_g . $S_h(f) = \tilde{h}(f)^2$, Δf is the detector bandwidth. The analysis is based on the *coincidence* between two or more experiments.

Pulsars: a) known source location, $h_0 = \sqrt{2S_h(f)/t_m}$; b) unknown source location, $h_0 = \sqrt{2S_h(f)/\sqrt{t_m\Delta t}}$, where t_m is the observation time. In b) we divide t_m in n sub periods of length $\Delta t = t_m/n$. The analysis is based on *pattern recognition* from given sources. At least in principle, it can be done using only one detector, obtaining important information on the source.

Relic g. w. (stochastic background): $\Omega(f) = S_h(f)f^34\pi^2/3H_0^2$, where H_0 is the Hubble constant and $\Omega(f)$ is the ratio of the g.w. energy density to the critical density needed for a closed Universe. The measurement is based on the *cross-correlation* of two (or more) detectors.

In a bar detector the spectral density has its minimum at the two resonance frequencies, $\simeq 904$ and 920 Hz. The bandwidth is actually of the order of 2 Hz, but it is planned to obtain $\simeq 60$ Hz with improved electronics. The search for pulsars [1], has begun using the data of the Explorer detector [2],

¹Dip. di Fisica, Univ. “La Sapienza”, P.A. Moro 2, 00185, Rome, Italy

which cover a period of a few years since 1990. Using one year of data, and the parameters of the 1991 run we obtain $h_0 \simeq 2.3 \cdot 10^{-25}$ in $\simeq 2$ Hz around the resonances and $h_0 \simeq 10^{-24}$ between them. The planned sensitivity with the Nautilus detector is $h_0 = 1.1 \cdot 10^{-26}$, using $t_m = 1$ year and $\Delta f = 5$ Hz. We are analysing the data for unknown sources, using a spectral data base, where each spectrum is estimated over a time $\Delta t \simeq 0.66$ hours (Doppler effect negligible) and has a header with information for vetoing the data. We compare the amplitudes (averaged as a function of the sidereal time) and frequencies of spectral lines with those expected from sources uniformly distributed over the sky. For the stochastic search using one detector it is only possible to put upper limits on $\Omega(f)$ [1, 4]. The measurement needs two (or more) detectors and a possible correlation might be derived using the coherence function, whose standard deviation reduces as $(t_m \Delta f)^{-1/4}$. Cross-correlating two “near” antennas like Nautilus, in operation for one year, it can be reached a limit on $\Omega(f)$ of $6.5 \cdot 10^{-4}$, an interesting result in the string scenario [5]. The sensitivity depends weakly on the bandwidth, while it decreases with the detectors distance, hence it is interesting to consider experiments with aligned interferometers and near bars [6].

It has to be remarked the difficulty of the problem since we do not expect “many” or “big” signals and we have experienced how critical the procedures and the comparison criteria are. For example, in a coincidence analysis, one has to define a priori what is an event, the coincidence window and all the data selection criteria (vetoes, pattern in time). In my opinion, the use of Bayesian analysis, that force the experimenter to declare, defining the priors, the choises done in the analysis, leads to a meaningful estimate of probability and hence is useful when assigning a probability, i. e., to a coincidence result or to different source models. An analysis based on this framework is in progress, using the Explorer data in the pulsar search described above.

REFERENCES

- [1] P. Astone, S. Frasca, G.V. Pallottino, G. Pizzella. Proc. of the International Conference on Gravitational Waves, Cascina, 19-23 March 1996.
- [2] G. V. Pallottino. Proc. of the International Conference on Gravitational Waves: Sources and Detectors, Cascina, 19-23 March 1996.
- [3] P. Astone et al. Upper limit for a g. w. stochastic background measured with the Explorer and Nautilus gravitational wave resonant detectors. To appear in *Physics letters B* (1996).
- [4] P. Astone, G.V. Pallottino, G. Pizzella. Detection of impulsive, monochromatic and stochastic gravitational waves by means of resonant antennas, LNF-96/001 IR 1996, submitted for publication (1996).
- [5] R. Brunstein, M. Gasperini, M. Giovannini, G. Veneziano. *Phys. Lett. B* **361**, 44 (1995).
- [6] B. F. Schutz. Proc. of the International Conference on Gravitational Waves, Cascina, 19-23 March 1996.

23

Statistical Tests for Changing Periods in Sparsely Sampled Data

Paul Hertz

23.1 An astronomical time series problem

The rate of change of physical periods in stars and stellar systems is arguably the most important diagnostic of the physical processes driving the evolution of those systems. For periods measured from a fiducial phase marker (eclipse, maximum brightness, etc.), the period derivative is measured by fitting a non-linear ephemeris to the fiducial phase timings. This method assumes that the uncertainties in the residuals are uncorrelated measurement errors. Variations in the residuals which mimic period change can be caused by intrinsic or correlated variability in the fiducial timing mark. We have developed statistical tests which can distinguish between these two cases and applied these tests to orbital period measurements of low mass X-ray binaries which have been reported to have orbital period evolution.

23.2 The basic idea

Orbital ephemerides are traditionally determined by fitting timing residuals to candidate ephemerides. If T_n is the time of the fiducial marker in cycle n ($0 \leq n \leq N$), e_n is the measurement error for T_n , and $F(n)$ is the predicted time for T_n from the ephemeris, then $R_n = T_n - F(n)$ are the timing residuals. In the astronomers' traditional "O-C" method, $F(n)$ is chosen to minimize $\sum_{n=0}^N R_n^2$. If the $\{e_n\}$ are the only noise and are iid, then $E(R_n) = e_n$ and the method is justified.

What if there is an additional noise term intrinsic to the system? Postulate an additional intrinsic noise term ϵ_n in cycle n . Then $T_n - T_{n-1} = F(n) - F(n-1) + \epsilon_n + e_n - e_{n-1}$, and $E(R_n) = \sum_{i=0}^n \epsilon_i + e_n - e_0$. The "O-C" residuals R_n are no longer independent and the R_n are no longer expected to be white and independent with mean 0 and identical variances. Instead the $\{R_n\}$ describe a random walk process controlled by the intrinsic noise process $\{\epsilon_n\}$. Lombard & Koen (1993) describe the whole problem.

23.3 Statistical methods

Bootstrap calculations: We ran simulations to bootstrap the probability that traditional “O-C” residual fits would imply that orbital period evolution is present. The models used the actual distribution of cycle timings from the data, the actual measurement uncertainties from the data, the actual (constant) orbital period from the data, and intrinsic noise with an assumed (constant) standard deviation. We calculated the following as a function of intrinsic noise standard deviation: (i) the probability that the χ^2 for a linear ephemeris is acceptable. (ii) the probability that the χ^2 is acceptable for a quadratic ephemeris, (iii) the probability that the χ^2 is unacceptable for a linear ephemeris and acceptable for a quadratic ephemeris.

Contingency test: Sterne & Campbell (1937) devised a simple test based on a two dimensional contingency table. It can only be applied to measurements of consecutive cycles of the binary. The table entries and marginal totals are used to construct a statistic with a known distribution. Using a Kolmogorov-Smirnov test, this statistic is tested against the null hypothesis of a constant period.

Correlation test: Eddington & Plakidis (1929) noted that the difference in “O-C” residuals is a function of the number of cycles between the observations. If the intrinsic noise process has constant standard deviation σ , and the measurement noise has constant standard deviation s , then the difference in residuals is a function only of the cycle span m and not of the start cycle n . The expected standard deviation of the difference in residuals over m cycles is $U_m = \sqrt{m\sigma^2(1 - m/N) + 2s^2}$.

23.4 Application to low mass X-ray binaries

We have examined the sources EXO0748-676, 4U1820-30, 4U1822-37, and Cyg X-3. The bootstrap calculation shows that all four sources have a $> 90\%$ probability of requiring orbital period evolution using the “O-C” method. For each source there are values of σ and s (correlation test) for which a Kolmogorov-Smirnov test yields an acceptable fit. The contingency test is inconclusive for all sources. We conclude that only Cyg X-3 shows evidence for orbital period evolution. For details, see Hertz (1994) and Hertz et al. (1995).

REFERENCES

- [1] Eddington, A.S., & Plakidis, S. 1929, MNRAS, 90, 65.
- [2] Hertz, P. 1994, HEAD-AAS Meeting, Napa, CA.
- [3] Hertz, P., et al. 1995, ApJ, 438, 385.
- [4] Lombard, F., & Koen, C. 1993, MNRAS, 263, 309.
- [5] Sterne, T. E., & Campbell, L. 1937, Ann.Harvard Coll.Obs. 105, 459.

Analyzing X-ray Variability by Linear State Space Models

Michael König and Jens Timmer

ABSTRACT In recent years, autoregressive models have had a profound impact on the description of astronomical time series as the observation of a stochastic process. However, it has to be taken into account that real data always contain observational noise often obscuring the intrinsic time series of the object. We apply the technique of a Linear State Space Model which explicitly models the noise of astronomical data and allows to estimate the hidden autoregressive process. As an example, we have analysed the X-ray flux variability of the Active Galaxy NGC 5506 observed with EXOSAT.

24.1 Introduction and mathematical background

A common phenomenon of Active Galactic Nuclei is the strong flux variability which can be observed in X-ray lightcurves. We present the Linear State Space Model (LSSM) based on the theory of autoregressive (AR) processes (Honerkamp [2]) to analyse this variability.

A given discrete time series $x(t)$ is considered as a sequence of correlated random variables. The AR model expresses the temporal correlations of the time series in terms of a linear function of its past values plus a noise term and is closely related to the differential equation describing the dynamics of the system (for a detailed discussion see Scargle [4]). LSSMs generalize the AR process by explicitly modelling observational noise. The variable $x(t)$ that has to be estimated cannot be observed directly since it is covered by observational noise $\eta(t)$.

$$\vec{x}(t) = \vec{A} \vec{x}(t-1) + \vec{\epsilon}(t) \quad \vec{\epsilon}(t) \sim N(0, \vec{Q}) \quad (24.1)$$

$$y(t) = \vec{C} \vec{x}(t) + \eta(t) \quad \eta(t) \sim N(0, R) \quad (24.2)$$

The terms $\vec{\epsilon}(t)$ and $\eta(t)$ represent the dynamical noise with covariance matrix \vec{Q} and the observational noise with variance R , respectively. We have used the Expectation-Maximization algorithm to estimate the dynamics and a Kolmogorov-Smirnov test to quantify the order of the hidden AR process (Honerkamp [2], König and Timmer [3]).

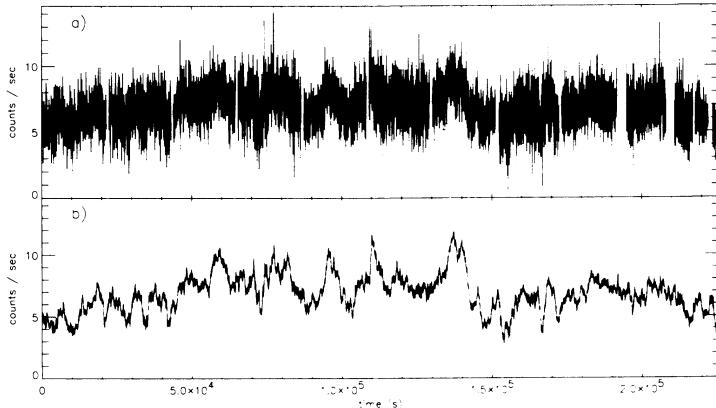


FIGURE 1. a) EXOSAT ME X-ray lightcurve of NGC 5506 (Jan. 1986), b) Hidden AR[1]-process, estimated with the LSSM fit.

24.2 Results and conclusion

We have used the X-ray lightcurve from EXOSAT of the Seyfert galaxy NGC 5506 for applying the LSSM. The lightcurve of NGC 5506 can be well modelled with a LSSM AR[1] model parameter of $a_1 = 0.9938 \pm 0.0007$ which corresponds to a relaxation time of $\tau = 4799^{+632}_{-472}$ s (König and Timmer [3]).

The covered AR[1] process indicates that the stochastic process is dominated by a single relaxation timescale, i.e. resulting from the exponentially decaying intensity of stochastically superimposed shots which are generated in the accretion process at the innermost region in the accretion disk (Green et al. [1]).

REFERENCES

- [1] Green A.R., McHardy I.M., Lehto H.J., 1993, MNRAS 265, 664
- [2] Honerkamp J., 1993, Stochastic Dynamical Systems, VCH Publ. New York, Weinheim
- [3] König M., Timmer J., 1996, A&A, submitted
- [4] Scargle J.D., 1981, ApJ Suppl. 45, 1.

The Time Interferometer: Spectral Analysis of the Gapped Time Series from the Stand Point of Interferometry

V. V. Vityazev

ABSTRACT The paper presents a comparative study of the fundamentals, problems and techniques common to the spectral analysis of time series and interferometry. The aperture synthesis techniques well known in radio astronomy is adapted to the spectral analysis of the time series. The method of iterated correlograms is proposed as a tool to obtain clean spectra of gapped and noisy observations.

Upon close examination one can see that in both sciences under consideration the rigorous (theoretical) quantities are introduced at the first level. In the spectral-analysis case they are the *power spectrum* and the *correlation function*; in the interferometry their counterparts are the *distribution of brightness* and the *spatial spectrum*. At the second level we have estimators of the strict quantities. In spectral analysis, they are the *periodogram* and the *correlogram*, whereas in the interferometry these are the *map* and the *visibility data*, respectively. Finally, equations which connect the quantities of the two levels are identical (convolution and multiplication) and include the characteristics of observations: the *beam* and the *transfer function* and their analogs, i.e. the *spectral window* and the *correlation window*.

In reality, due to finite dimensions of mirrors and finite time spans of observations we cannot get the distributions of brightness on the sky and the power spectra of the time series, and what we can do is to find their as good as possible approximations. In optics or in radio astronomy, when the filled apertures are used, the maps are produced directly in the focal plane of a telescope. Analogously, when the time series is given at all points of some interval or at time points regularly spaced within the interval, the evaluation of the periodogram can be made quite easily. When an interferometer is used, the aperture is not solid, and what we can measure is the visibility data, i.e. the estimator of the spectrum of spatial frequencies. The longer the baseline, the less area in the $(u - v)$ plane can be filled and the more dirty the resulting map becomes.

To overcome this various techniques of the *aperture synthesis* are used, and this leads to complete solution of the problem since it allows to fill the $(u-v)$ plane completely. If the *aperture synthesis* provides the partial filling of the $(u-v)$ -plane, the *cleaning* procedures can be used with the aim to eliminate from the map the artifacts of the “holes” in the $(u-v)$ -plane. The same problems we meet in the spectral-analysis case, when the time points are distributed irregularly or have long gaps. In this case the correlograms cannot be determined for all values of time lag τ , and this would give false features in the resulting periodograms. This is the point where the main idea of the present paper is hidden. It is: whether it is possible to apply the *aperture synthesis* method to spectral analysis of time series ?

On the basis of the conceptual identity of the correlogram and of the visibility data an attempt is made to adapt the aperture synthesis techniques to the spectral analysis of time series. Two methods of synthesizing a correlogram are proposed. The first one reproduces the idea of Ryle's interferometer and can be realized when averaging over statistical ensemble is possible. The second method is based on the iterated correlograms and can be applied to a single curve with gaps. It is shown that our method yields, at least by iterations, the correlogram at all the points of the time span without gaps, and consequently clean spectra can be obtained. The method is described in detail and the situations frequently met in astronomy are considered: a long gap and periodic gaps in observations as well as irregularly missed points.

In fact, the method of iterated correlograms solves the so-called restoration problem. The simplest inverse method to solve this problem is based on the deconvolution technique, but in the presence of inevitable noise associated with any measurements, the simple deconvolution method proves to be unsuccessful. On the contrary, our method is based essentially on a direct correlation transforms which can only suppress noise, not to enhance it. From this follows that our method is stable in the presence of noise.

Numerical examples are presented to illustrate the application of the algorithm to gapped astrometrical time series and to time series with uneven precision of the measurements.

The intensive checking of our method gives evidence that it can be applied to time series for which the total length of gaps does not exceed 50-70 per cent of the total time span. For further details the reader is referred to Vityazev (1993) and Vityazev (1996).

REFERENCES

- [1] Vityazev V.V., 1993, Astronomical and Astrophysical Transactions, Vol.5, pp.177-210.
- [2] Vityazev V.V., 1996, Astronomical and Astrophysical Transactions, (in press).

Structures in Random Fields

Juan E. Betancort-Rijo¹

ABSTRACT We have developed a framework within which the probability density for the occurrence of structures within a random field may be assessed. This framework is rather general, but we have used it mainly in the context of the large scale structure of the Universe, where the most relevant structures are: voids, galaxy clusters, filaments and laminae.

We have also considered the probability densities of structures in point distributions, starting with the Poissonian case and generalizing the results found in this case to a general distribution. The comparison, in the high density limit, with the appropriate continuous random field renders several useful results.

We briefly review here the basic content and the main applications of a line of research developed in principle to solve specific problems in the large scale structure of the Universe (LSSU), but that have found applications in a wide variety of problems.

When we look at a map of the distribution of galaxies on large scale and find some conspicuous features (structures) like large voids, filaments, laminae, rich clusters, we may wonder about the likelihood for their occurrence within a particular model for the LSSU. To answer this question we must deal with two different problems. Firstly, we have to obtain the present clustering properties of the galaxies (as given by all correlation functions or otherwise) from the initial conditions and the processes (let's say gravitational clustering) prescribed by the model under consideration. Then, we need to derive from the clustering properties the probability density for the structures in question. This is by far the most difficult part of the problem, representing a rather general question not specific to the LSSU and is the one that will concern us here.

To solve this problem we first considered the Poissonian case, where the positions of the galaxies are uncorrelated. The solution to this question have been given by Presskill & Politzer (P&P, 1986) for a d-dimensional distribution, but only for the cases where the mean density of points within the structures was much larger or much smaller than the average density. However, for most interesting structures (except for voids) this approximation does not work. To see this, take the case of an Abell cluster of richness

¹Instituto de Astrofísica de Canarias. E-38200. La Laguna (Tenerife).

type 31. Every distinct spot where a sphere of 1.5 Mpc of radius may be placed so as to contain 50 galaxies is called a cluster of this kind. But it is clear that by moving slightly a sphere containing 50 galaxies we may find a large number, N , of distinct collections of 50 galaxies enclosed within the sphere in question and sharing all but a few of their galaxies.

The expression presented by P&P give the probability density for the distinct collections of 50 galaxies. But, since all collections sharing most of their galaxies are formed out of the same physical structures (cluster) it is clear that to obtain the density of clusters we must divide this expression into the mean value, over all clusters, of N . The computation of this number is rather difficult representing the main complication of the present problem. To obtain Q , we first considered the one-dimensional problem, and then showed how to use it to obtain the value of Q corresponding to the d -dimensional case (Betancort-Rijo 1991a). The results of this work allow us to compute the probability of finding within a given population n individuals having values for a set of m properties Π within a distance $\Delta\Pi$ from each other.

The computation of Q for a wider class of distribution was presented in Betancort-Rijo (1994). Q was given for any distribution conforming to a Poissonian model, that include all those relevant to the LSSU in Betancort-Rijo (1995a). An example of applications may be found in Betancort-Rijo (1996a). Another example may be found in Betancort-Rijo (1991b) where we computed the densities of large voids in the LSSU within the cold dark matter models.

So far we have considered structures in point distributions. A closely related problem is that of obtaining the densities and correlations of structures in continuous random fields. In Betancort-Rijo (1990), we considered the special case of cylindrical filaments. The general procedure for Gaussian fields was presented in Betancort-Rijo (1992), where both differentiable and non-differentiable random fields were considered. The implications of this work for the galaxy clustering models are discussed in Betancort-Rijo (1995b), where we showed how the difficulties of the peak models disappear when galaxies are identified with structures rather than with peaks.

The detailed study of structures in three dimensional Gaussian fields is given in a recent work (Betancort-Rijo 1996c) and its main application is treated in Betancort-Rijo (1996b).

REFERENCES

- [1] Preskill J. & Politzer H.: 1986, Phys.Rev.Lett. 56, 99
- [2] Betancort-Rijo J.E.: 1990. Filament in Gaussian Fields. M.N.R.A.S. 243, 431.
- [3] Betancort-Rijo J.E.: 1991a. Probabilities of Structures in Poissonian Distributions, Phys.Rev.A., 43, 2694.
- [4] Betancort-Rijo J.E.: 1991b. Probabilities of Voids, M.N.R.A.S., 246, 608.

- [5] Betancort-Rijo J.E.: 1992. Structures in Random Fields: Gaussian Fields, Phys.Rev.A., 45, 3447.
- [6] Betancort-Rijo J.E.: 1994. Cluster in Point Distributions, Phys.Rev.E., 50, 4410.
- [7] Betancort-Rijo J.E.: 1995a. Cluster in General Distribution of Points, (Unpublished).
- [8] Betancort-Rijo J.E.: 1995b. The Structure Model for Galaxy Clustering, Astron.Astrophys., 299, 635.
- [9] Betancort-Rijo J.E.: 1996a. Relationship between Cluster Densities and other Statistical Quantities, Astron.Astrophys. (In press)
- [10] Betancort-Rijo J.E.: 1996b. Excursion Sets Size Distributions and the Cosmic Mass Function, Ap.J. (In press)
- [11] Betancort-Rijo J.E.: 1996c. Structures in Three-dimensional Random Fields, (In preparation)

The New γ -CFAR Detector For Astronomical Image Processing

**A. D. Nair¹, Jose C. Principe and Munchurl
Kim^{2 3}**

ABSTRACT In this paper, we propose a new constant false alarm rate (CFAR) detector based on a family of gamma kernels which provide the ability to change the scale and shape of a stencil that models the local area in an image. This new detector is called the CFAR detector. The scale and shape can be selected adaptively after training in such a way so as to give minimum false alarms. The detector makes the maximal use of intensity information to discriminate targets from background clutter. A significant increase in the signal-to-noise ratio can be attained.

27.1 Introduction

A fundamental problem in image processing is to extract relevant features from an image contaminated with noise. To extract the maximum information from an image one needs to develop a method for noise suppression and feature extraction. The γ -CFAR detector constitutes the first stage of a three-stage imaging system currently being developed at Computational NeuroEngineering Laboratory, Univ. of Florida (Kim et al., 1996). The three stages include a prescreening stage, a discrimination stage and a classification stage.

The CFAR is a novel target detector that makes the maximal use of intensity information by employing a class of analyzing functions which can be adaptively scaled (or shaped) to capture the regions around a test cell with most discriminatory information in images. The analyzing functions serve as a stencil for extracting the 1st and 2nd order statistics and are 2-D circularly symmetric gamma kernels. Features are extracted from a local neighborhood of a cell under test by centering gamma kernels on that point

¹Department of Astronomy, University of Florida

²Computational NeuroEngineering Laboratory, University of Florida

³This work was supported by the Florida Space Grant Consortium.

and taking the inner product of the local image and the power of the image with kernels. These values can be viewed as estimates of the local first and second momenta which contain all the information necessary to compute local variance.

A 2-D gamma function is described by the equation

$$g_{n,\mu}(k,l) = \frac{\mu^{(n+1)}}{2\pi n!} (\sqrt{k^2 + l^2})^{(n-1)} e^{(\mu\sqrt{k^2 + l^2})} \quad (27.1)$$

where $\Omega = \{(k, l); -N \leq k, l \leq N\}$ is the region of support of the kernel, n is the kernel order and the parameter that controls the shape and scale of the kernel.

27.2 The γ -CFAR detector

The CFAR detector defines a local region, selected heuristically, in the image where it locates potential targets based on estimates of the statistics of the pixel under test as well as a local clutter region. The stencil of the CFAR detector is adapted based on example images so as to provide the minimum false alarm rate. The CFAR detector implements a decision rule defined as

$$y = ((g_{1,\mu_1} \cdot X - g_{n,\mu_n} \cdot X) / \hat{\sigma}) \geq \text{target: } \leq \text{clutter}$$

where \cdot is the inner product operator and $\hat{\sigma}$ is the variance.

Once the value of n in equation 27.1 is fixed, the detector is trained on two sets of data. The first set of data contain the targets or the object of interest; the second set contains the non-targets. Now the task is to scan the parameter space incrementally and pick the optimal value of μ that can recognize a maximum number of targets in the first data set and at the same time recognize a maximum number of non-targets in the second data set.

27.3 Conclusion

The CFAR can be thought of as an estimator of the local intensity statistics. Here, the optimal value was found through an exhaustive search to quantify what the best possible performance is. Further refinements and suitability for photometry are being examined. Application to data and results will be discussed in an upcoming paper. The next stage, the discrimination stage, QGD, is currently being implemented.

REFERENCES

- [1] Kim, M., Fisher, John III, and Principe, Jose C., 1996, in Proc. SPIE.

Bayesian Image Reconstruction with Noise Suppression

Jorge Núñez and Jorge Llacer

In this paper we present a Bayesian image reconstruction algorithm with entropy prior (FMAPE) that uses a space-variant hyperparameter. The spatial variation of the hyperparameter allows different degrees of resolution in areas of different signal-to-noise ratio, thus avoiding the large residuals resulting from algorithms that use a constant hyperparameter. The space variant hyperparameter determines the relative weight between the prior information and the likelihood and it determines the degree of smoothness of the solution in a particular region of the image.

To compute the variable hyperparameter we first carry out a Maximum Likelihood Estimator (MLE) reconstruction. We then segment the resultimg image by a technique based on wavelet decomposition and self-organizing neural networks (Kohonen), resulting in a number of extended regions (9 to 25) plus the stars. A different value of the hyperparameter is assigned to each region, leading to a reconstruction with decreased bias and excellent visual characteristics.

As example, we demonstrate the reconstruction of a real image of Saturn obtained with the WF/PC camera of the aberrated HST. Figure 1 (left) shows the raw data to be reconstructed. The PSF is shown in a logarithmic scale in Figure 1 (right). Figure 2 (left) shows 9 regions resulting from the segmentation. Figure 2 (right) shows the result of the 9-region reconstruction. The results show that we obtained a smooth background, a well reconstructed image of the planet and sharp images of the divisions in the rings. By the use of the variable resolution FMAPE algorithm, we have avoided noise amplification in all the regions of the image.

To compare the results with the MLE and the constant hyperparameter approach, we used the FMAPE with constant values of the hyperparameter of 200 and 500, and reconstructed by the MLE at 100 iterations. Figure 3 shows the plot of the mean normalized residuals against the region number. The reconstruction obtained by MLE has mean residuals that are dispersed arround the ideal value of 1.0. The reconstruction with constant hyperparameter of 200 has too high mean residuals. With a hyperparameter of 500 some residuals are too low but others are still too high. However, the reconstruction with the variable hyperparameter, although not perfect, is much improved, with most of the mean residuals close to 1.0.

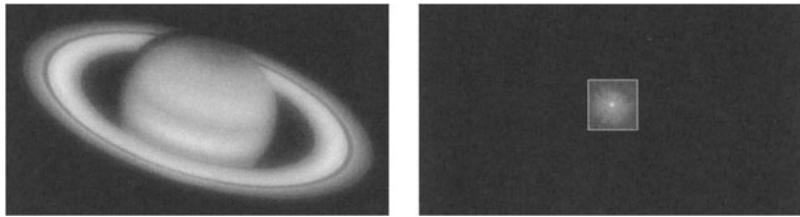


FIGURE 1. Left: Raw image for the planet Saturn. Right: observed PSF used for the reconstruction.

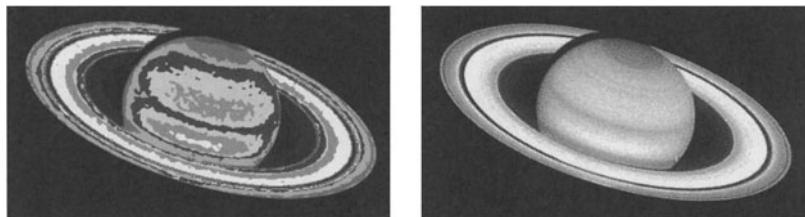


FIGURE 2. Left: Segmentation of the image in 9 regions. Right: Reconstruction of the image of Saturn using the FMAPE algorithm with the 9 channels.

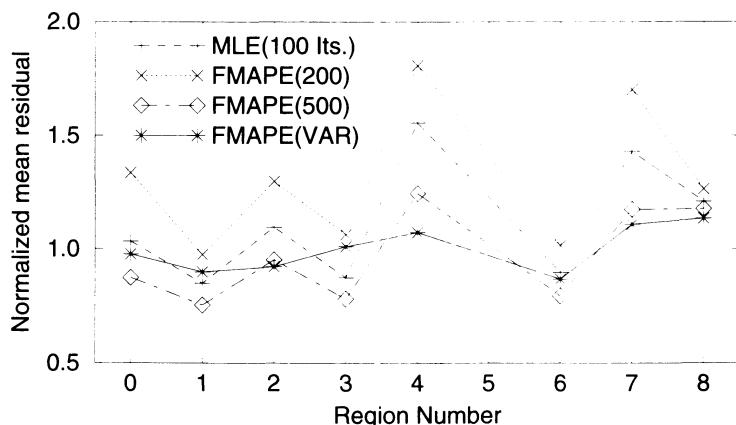


FIGURE 3. Mean residuals for the 9 regions obtained with the FMAPE algorithm with variable hyperparameter in comparison with other methods.

Astronomical Images Restoration by the Multiscale Maximum Entropy Method

Jean-Luc Starck and Eric Pantin ¹

ABSTRACT We describe in this paper the Multiscale Maximum Entropy Method which is based on the concept of multiscale entropy derived from the wavelet decomposition of a signal into different frequencies bands. It leads to a method which is flux conservative, and the use of a multiresolution support solves the problem of MEM to chose the α parameter, i.e. relative weight between the goodness-of-fit and the entropy.

The concept of entropy following Shannon's or Skilling and Gull's definition [2] is a global quantity calculated on the whole image I . It is not matched to quantify the distribution of the information at different scales of resolution. Therefore, we have proposed the concept of multiscale entropy [3] of a set of wavelet coefficients $\{w_j\}$ by

$$S(I) = \frac{1}{\sigma_I} \sum_{\text{scales } j} \sum_{\text{pixels}} A(j, x, y) \sigma_j (w_j(x, y) - m_j - |w_j(x, y)| \ln \frac{|w_j(x, y)|}{m_j})$$

where σ_I is the standard deviation of the noise in the image I , w_j are the wavelet coefficients, and σ_j is the standard deviation of the noise at scale j . The A is the reciprocal of the multiresolution support M [4], $A(j, x, y) = 1 - M(j, x, y)$.

The multiscale entropy is the addition of the entropy at each scale. We take the absolute value of w_j in that definition because the values of w_j can be positive or negative and a negative signal contains also some information in the wavelet transform. The advantage of such a definition entropy is the fact we can use previous works concerning the wavelet transform and image restoration [4]. The noise behaviour has been studied in the wavelet transform and we can estimate the standard deviation of the noise σ_j at the scale j . These estimations can be naturally introduced in our models $m_j : m_j = k_m \sigma_j$. The model m_j at the scale j represents the value taken by a wavelet coefficient in the absence of any relevant signal and in practice, it must be a value small compared to any significant signal value. Following

¹CEA/DSM/DAPNIA F-91191 Gif-sur-Yvette cedex

Gull and Skilling [2] procedure, we take m_j as a fraction of the noise ($k_m = 1/100$).

In this application, we use the discrete *à trous* algorithm (described in [4]) for its simplicity. An image $I(x, y)$ is decomposed into $w_j(x, y), j = 1, \dots, n_p$ scales (where n_p is the total number of wavelet scales) and a smooth image $c_{n_p}(x, y)$ and we can write $I(x, y) = c_{n_p}(x, y) + \sum_{j=1}^{n_p} w_j(x, y)$.

The entropy S measures the amount of information only at scales and in areas where we have a low signal-to-noise ratio. We will show in the next section how these notions can be taken together to yield efficient methods for filtering and image deconvolution.

As in the ME method, we will minimize a functional of O , but considering an image as a pyramid of different scales of resolution in which we try to maximize its contribution to the multiscale entropy. The functional to minimize is

$$J(O) = \sum_{\text{pixels}} \frac{(I - P * O)^2}{2\sigma_I^2} - \alpha S(O).$$

The solution can be found by performing the following iterative schema $O^{n+1} = O^n - \gamma \nabla(J(O^n))$, where $\nabla(J(O^n))$ is the gradient of $J(O^n)$.

Compared to the classical MEM, our method has a fixed α parameter and there is no need to determine it: it is the same for every image. Furthermore, this new method is flux-conservative and thus reliable photometry can be done on the deconvolved image. In [1], it was noticed that the “models” in the multi-channel MEM deconvolution should be linked to a physical quantity. We have shown here that this is the case since it is a fraction of the standard deviation of the noise at a given scale of resolution. Using wavelets decomposition, we have proven that many problems they encountered are naturally solved, especially the model and the α estimation.

REFERENCES

- [1] Bontekoe, T.J.R. , Koper, E. , and Kester,D.J.M., 1994, "Pyramid Maximum Entropy Images of IRAS Survey Data", *Astronomy and Astrophysics*, 294, pp 1037-1053.
- [2] Gull, S.F., Skilling, J., 1991, MEMSYS5 User's Manual.
- [3] Pantin, E. and Starck, J.L., "Deconvolution of Astronomical Images using the Multiresolution Maximum Entropy Method". to appear in *Astronomy and Astrophysics*.
- [4] Starck, J.L., Murtagh, F., and Bijaoui, A. "Multiresolution Support Applied to Image Filtering and Deconvolution", in *CVIP: Graphical Models and Image Processing*, Vol. 57, 5, pp 420-431, Sept. 1995.

Nested Test for Point Sources

James Theiler and Jeff Bloch

ABSTRACT We describe an extension of a test proposed by Lampton for detecting weak and/or short-lived transient point sources in finite-resolution photon-limited maps of the sky. We test the null hypothesis that the count density is spatially uniform (*i.e.*, no point sources) over a limited region of the sky by decomposing that region into a nested set of source regions and background annuli, and testing whether the count density in each source region differs significantly from the count density in each associated background. The tests are by construction *independent*, so a combined *p*-value can be rigorously defined.

Our aim is to discriminate real point sources from Poisson fluctuations in the background. The maximum-likelihood formalism prescribes a “matched filter” in which the telescope point-spread function is convolved with the sky map. Points on the convolved map that exceed a background-dependent threshold are identified as sources. If the threshold is set properly (and that is straightforward in the high count limit or whenever the uncertainties are Gaussian), then this technique is demonstrably optimal. The low count regime (in which the statistics are Poisson) is more problematic, though empirical corrections have been proposed [2].

Lampton [1] described a statistical test for point source detection that is in general less powerful than the matched filter, but it does not require a separate estimate of the background, and it is strictly valid even for arbitrarily low counts. In this test, a source region is surrounded by a (typically annular) background region, and the number of photons in each region is counted. If c_S (resp. c_B) is the number in the source (resp. background) region, then $p = I_f(c_S, c_B + 1)$ is the *p*-value associated with the null hypothesis that the source and background count densities are the same. Here, I_f is the incomplete beta function, and $f = A_S/(A_S + A_B)$ is the ratio of the area of the source region to the sum of the areas of the source and background regions.

In the nested Lampton test, the source region itself is re-divided into source and background regions, and the Lampton test is then applied to these two subregions. This second test assesses whether there are more counts in the center of the source region than there are around the edges. It is important to realize that this second test is strictly independent of the first test. In fact, it is this independence that permits us to reliably and efficiently combine the two *p*-values into a single measure of confidence.

If we have two p -values, p_1 and p_2 , we can always multiply them (let $p_* = p_1 p_2$), but we cannot interpret their product p_* as a p -value, because the product is not uniformly distributed on the unit interval. But it is straightforward to correct for this, and if we let $p = p_*(1 - \log p_*)$, then this “corrected product” can be used as a valid p -value. To combine k independent p -values, use $p = R_k(p_*)$, where $p_* = p_1 p_2 \dots p_k$ is the product, and $R_k(p_*) = p_* + \int_{p_*}^1 R_{k-1}(p_*/p) dp$. It is also possible to use weighted combinations. If we write $p_* = p_1^a p_2^b$, then $a > b$ more heavily weights the first test. Here, the corrected value is given by $p = (ap_*^{1/a} - bp_*^{1/b})/(a - b)$.

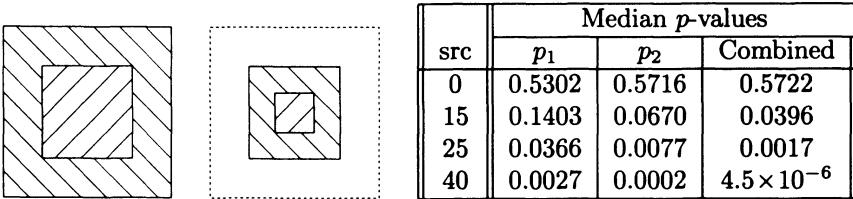


FIGURE 1. The left two figures illustrate the decomposition of the 13×13 region of interest into a 9×9 source region and a background annulus, and the further division of the source region again into a 3×3 source and background. The table on the right shows the results of a Monte-Carlo experiment in which a Gaussian point-spread source (with $\sigma = 1.5$ pixels) is placed at the center of this region with low count density background (1 count per pixel), and p -values are computed for the individual and for the combined tests.

Compared to traditional methods (e.g., see Simes [3]), the corrected product rule is not as robust to non-independence of the separate tests, but it does produce a smaller combined p -value in the regime where p_1 and p_2 are both small. This is crucial in large sky maps, because individual point source detections require very high levels of significance.

Acknowledgments: We are grateful to Xiaoyi Wu for useful discussions. This work was supported by the United States Department of Energy.

REFERENCES

- [1] M. Lampton. Two-sample discrimination of Poisson means. *Astrophys. J.*, 436:784–786, 1994.
- [2] H. L. Marshall. Tools for use with low signal/noise data. In D. R. Crabtree, R. J. Hanisch, and J. Barnes, editors, *Astronomical Data Analysis Software and Systems III*, volume 61 of *ASP Conference Series*, 1994.
- [3] R. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754, 1986.

Segmenting Chromospheric Images with Markov Random Fields

Michael J. Turmon¹ and Judit M. Pap²

ABSTRACT The solar chromosphere roughly consists of three types of region: plage, network, and background. Thresholding individual pixel intensities is typically used to identify these regions in solar images. We have incorporated spatial information by using a Bayesian setup with an image prior that prefers spatially coherent labelings; resulting segmentations are more physically reasonable. These priors are a first step in developing an appropriate model for chromospheric images.

The solar chromosphere, observable in ultraviolet light, roughly consists of three classes: plage (bright magnetic disturbances), network (hot boundaries of convection cells), and background (cooler interiors of cells). Plages appear as irregular groups of clumps, seldom near the solar poles. The cell-structured network has little contrast with the background, and is spatially homogeneous. The classes contribute differently to the UV radiation reaching Earth's upper atmosphere. It is of scientific interest (e.g., in studying global warming) to relate plage and network area and intensity to total UV irradiance. To do this, spatially resolved images are needed.

We treat this problem in a Bayesian framework as inference of the underlying pixel classes based on the observed intensity. Denoting pixel sites $s \in N$, and defining matrices of class labels $\mathbf{x} = \{x_s\}_{s \in N}$ and observed intensities \mathbf{y} , the posterior probability of labels given data is

$$P(\mathbf{x} | \mathbf{y}) = P(\mathbf{y} | \mathbf{x})P(\mathbf{x}) / P(\mathbf{y}) \propto P(\mathbf{y} | \mathbf{x})P(\mathbf{x}) .$$

The maximum a posteriori (MAP) rule maximizes this probability:

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x}} \log P(\mathbf{y} | \mathbf{x}) + \log P(\mathbf{x}) .$$

The first term is the familiar likelihood function, telling how the data is obtained from the labels; the second is the prior probability of a given

¹Jet Propulsion Laboratory, Pasadena, CA 91109

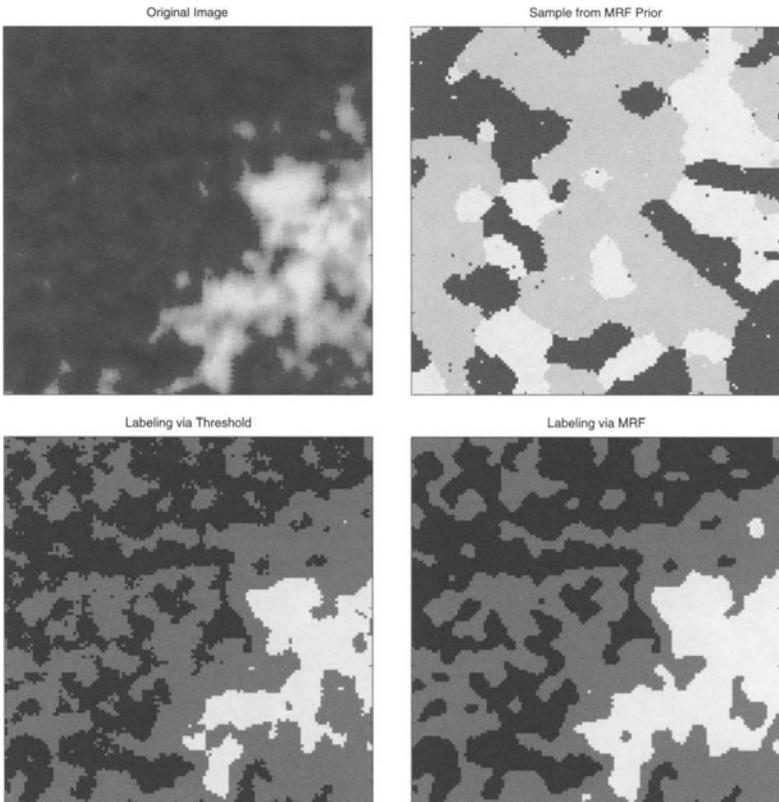
²UCLA Department of Astronomy, Los Angeles, CA 90095

labeling. In practice, the first term forces fidelity to the data while the second penalizes unlikely rough labelings.

Prior models may be specified in many ways: we have used the Markov field models introduced by Besag and others for image analysis. These models are defined by the conditional distributions

$$P(x_s = k \mid x_{N \setminus \{s\}}) = P(x_s = k \mid x_{N(s)}) = Z_s^{-1} \exp[-\beta \sum_{s' \in N(s)} 1(x_{s'} \neq k)]$$

where $N(s)$ is the 8-pixel neighborhood centered around a site s , and $Z(s)$ is a constant chosen to make the distribution sum to unity. The first equality expresses the Markov property that far-off sites do not influence the distribution of labels when the local neighbors are known, while the second favors ‘smooth’ labelings. Priors more tailored to this application can be built; of some interest is capturing the network structure.



Sample results are shown above. The first panel shows a piece of a chromospheric image from January 1980 with a plage in the lower-right corner. Below this is the corresponding threshold segmentation. The top-right panel shows a typical (random) image from the MRF prior $P(\mathbf{x})$ (no data is used in generating it). While this image does not precisely match any

expected plage/network pattern, the match is much better than a field of independent labels at each site. The MAP/MRF segmentation is in the final panel; we note that it has eliminated many of the tiny gaps in the large plage and made the network structure more apparent.

Analysis of Hipparcos Data in the Orthonormal Wavelet Representation

E. Chereul, M. Crézé, and O. Bienaymé

ABSTRACT Hipparcos data will provide the first unbiased probe of the phase space distribution of stars within 150 pc almost devoid of systematic errors. This probe will be used to trace small scale inhomogeneities of the gravitational potential which may play a role in the unexplained mechanism of the stellar disc heating. The phase space distribution of an homogeneous sample of tracer stars will be analysed in the Orthonormal Wavelet Representation. We present the method used and its calibration in terms on limits of detection in size, signal to noise ratio of the inhomogeneities on 2D simulations.

32.1 Introduction

Due to the lack of reliable data, the description of the phase space distribution of stars in our galaxy has long been bound to a naive approach: the density distribution and the first two moments of the velocity distribution. Attempts to go any further would come up against insuperable limitations imposed by systematic errors in velocity data and extremely poor distance data. Thus, small scale inhomogeneities of the distributions possibly reflecting inhomogeneities of the potential would remain definitely out of reach. As a consequence, while star velocity dispersions have long been known to be larger for elder star samples, possible causes of this secular “heating” cannot be discriminated from each other on the basis of any testable evidence. Mechanisms like Molecular Clouds, transient dynamical instabilities or Massive Halo Black Holes [CGL85] have been advocated while other theories refer to a slow collapse of the disc. All such mechanisms would leave different signatures in small scale. The **Hipparcos** space mission has produced an unprecedented set of distance and tangential motion measurements for nearly all moderately bright disc stars within 150 pc of the sun. This sample is an ideal material to trace small scale imhomogeneities.

The orthonormal wavelet representation [Mal89] is a promising framework to search for a-periodic structures and unknown scales and probably

low amplitudes. The search will be performed in various 2D and 3D cuts of the 6D phase space. Here, we present the method and report first numerical experiments starting with fake fully homogeneous distributions and introducing gradually subliminal inhomogeneities. Thus, we calibrate the sensitivity, selectivity and response of the method. The results may permit detection of low amplitude signatures of inhomogeneities in the phase space distribution of our Hipparcos sample. Then, their typical sizes and amplitudes will allow us to derive the main mechanism responsible of the disc heating.

32.2 Implementation of the wavelet analysis

The wavelet analysis of a given 2D signal provides a spectrum of the energy (the sum of the square wavelet coefficients) found at each scale. The significance of these energies is first calibrated empirically: sixty uniform simulated samples are generated and wavelet analysed, each giving a spectrum of a 0-fluctuation signal. The average 0-fluctuation spectrum will then be subtracted from each spectrum obtained with Hipparcos data and divided by the standard deviation of the 0-fluctuation energy distribution, giving a normalized energy spectrum:

$$E_l^{norm} = \frac{(E_l - \overline{E_l^{0-fluc.}})}{\sigma_{E_l^{0-fluc.}}}$$

Since the distribution of normalized energies obtained from a set of uniform random simulations is almost Gaussian at each scale l , a E_l^{norm} exceeding ± 2 is considered significant at the 2 sigma level.

32.3 Preliminary results

Inhomogeneities are simulated by concentrating part of the data points in Gaussian clusters. Characteristics of the structures are controlled via three parameters: the number of stars per cluster, the number of clusters and the size (standard deviation) of individual clusters. Playing with these three parameters, we explore the performances of the method along two basic characteristics of the structures: the signal to noise ratio of one structure (defined as number of cluster stars within 2 sigma divided by the square root of the number of non-cluster stars in the same area) and the degree of clustering. For each selected characteristics we produce a set of 30 fake samples of 5000 star points, then we count the probability of detecting a significant signal in the normalized spectrum.

Cluster sizes ranging from 0.8 through 12.5 pc, signal to noise ratios from 0.5 to 20 have been explored with different cluster numbers. The

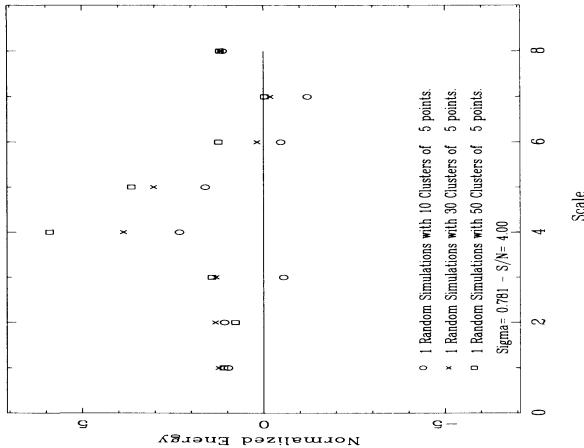


FIGURE 1. Energy spectrum evolution when the number of clusters increases.

window size is 200×200 pcs so that a cluster size $\sigma = 12.5$ is expected to reflect at scale $4\sigma/200$ which is scale 7.

The general trend of conclusions is the following: individual cluster detection is efficient as long as signal to noise ratio is as high as 3-4 or more, nearly all such structures produce signal above the 1.5 threshold at the appropriate scale in the normalized energy spectrum, with very few false detections.

If the degree of clustering is high (many low S/N clusters), then energy spectra show significant signal at the appropriate scale at S/N as low as 0.5 even though no individual detection can be made. In Figure 1 we give three energy spectra from simulations of 5000 star points featuring 10, 30 and 50 clusters with $\sigma_{clust}=0.8$ which corresponds to scale 4. The theoretical S/N per cluster is 4.

REFERENCES

- [CGL85] Jeremiah P. Ostriker, Cedric G. Lacey. Massive black holes in galactic halos ? *Astrophysical Journal*, 299:633–652, 1985.
- [Mal89] Stéphane G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Patt. Anal. Mach. Int.*, 11(7):674, 1989.

Statistical Properties of Wavelet Transforms Applied to X-Ray Source Detection

F. Damiani, A. Maggio, G. Micela and S. Sciortino

ABSTRACT We have developed a method for detection of sources in X-ray images, based on wavelet transforms (WT). After having computed the WT of an image, for various values of the wavelet scale parameter, candidate sources are selected as local maxima in the wavelet-transformed image. The reliable discrimination between true sources and random background fluctuations requires a detailed knowledge of the distribution of WT values arising only from background noise. The shape of this distribution may be very different from a Gaussian, especially in the limit of few counts per image resolution element, which is the case in most X-ray images.

By means of both analytical means and numerical simulations, we have therefore studied the WT probability distribution for a wide range of background density values. This enables us to derive thresholds for source detection in the WT images, for a range of confidence levels. These detection thresholds are now being used in our detection algorithm with good results.

33.1 Description of the method

With the advent of new X-ray detectors having higher spatial resolution, and therefore fewer background photons per resolution element, the case of sparse-photon X-ray images is becoming ever more common. Under such circumstances, it may be possible to detect sources showing just a few imaged photons, provided that the analysis method is powerful enough. To this aim, we have recently developed a source detection method, based on wavelet transforms (WT), for images produced by photon-counting detectors, that has been tuned accurately in order to fully exploit its capabilities.

Our algorithm selects sources as spatial maxima in the WT of the analyzed image (at various spatial scales a) rather than in the image itself. The statistical distribution of photons in a pixel of the original image (expected to be locally a Poissonian) is thus converted into another distribution $P(w)$ of values w in the wavelet-transformed image. Therefore, in order to discriminate efficiently WT maxima due to real sources from those arising from background fluctuations we must know in detail the distribution $P(w)$ of WT values due only to background fluctuations.

33.2 The distribution $P(w)$

Adopting as a generating wavelet the so-called ‘mexican hat’ function, best suited for our problem, we find that the distribution $P(w)$ has a complicated shape, that depends on the background intensity. More specifically, it can be seen that the shape of $P(w)$ depends only on the product $q = b \times a^2$, namely the number of background photons per squared wavelet scale, so we denote this distribution as $P_q(w)$. As the number q becomes lower, the image photons are more sparse, their statistics is more markedly Poissonian, and the WT distribution is increasingly different from a Gaussian (we call this the ‘discrete photon limit’). Instead, as q rises (say $q \geq 300$) the distribution $P_q(w)$ approaches slowly a Gaussian, as it can be shown analytically, since the distribution of photons in the analyzed image now approaches a Gaussian (the ‘continuous limit’). Such an analytic description has not however been found for the discrete photon limit.

Since current X-ray detectors have many thousands independent resolution elements, source detection involves a comparably large number of ‘trials’, whichever method one is using. Therefore, in order to keep the number of expected spurious detections reasonably low (1-2 per field), the appropriate detection threshold must be derived from the extreme positive tail of the distribution $P_q(w)$, that must be therefore known in detail. Since in the discrete photon limit we have no analytic expression for $P_q(w)$, we have computed it numerically, through large sets of Monte Carlo simulations (10^7 realizations for each background value), needed to describe accurately even the tail of $P_q(w)$. As a result, we have derived detection thresholds (in the WT space) for a large range of background values, and for any desired confidence level (namely, the probability that a given detection is due to a background fluctuation). Eventually, since our algorithm involves a cross-identification between sources detected at various scales a , we have applied the entire algorithm to simulated background images, to derive the overall number of spurious detections corresponding to a chosen confidence level.

33.3 Conclusions

Currently, our algorithm works successfully on ROSAT PSPC and HRI images. Especially in the center of ROSAT HRI images, where spatial resolution is highest, the number of background photons per resolution element (of size given by the PSF FWHM) is very low, and the shape of the WT at small scales a is most different from a Gaussian. The present study, moreover, will be even more relevant for the analysis of X-ray images with even higher resolution, (e.g. using the AXAF, ACIS and HRC), where we expect to detect reliably sources with only a handful of photons. For more details, we refer the reader to Damiani et al. (1996; recently submitted to ApJ).

Wavelet Based X-Ray Spatial Analysis: Statistical Issues

V. Kashyap and P. Freeman¹

34.1 Algorithm

We use a wavelet-based algorithm to detect and characterize sources in astronomical X-ray images, which consists of correlating binned data with the Mexican-Hat function (MexHat):

$$MH(x, y) = \left(2 - \frac{x^2}{\sigma_x^2} - \frac{y^2}{\sigma_y^2} \right) e^{-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}}.$$

Use of the algorithm over a grid of scales (σ_x, σ_y) in principle allows us to detect and characterize sources.

We identify putative source pixels by applying a detection threshold based on the value of the local background; the background is estimated by correlating image data with a suitably normalized negative annulus of the MexHat. All pixels in the data so identified are reset to the value of the local background, and the procedure is repeated to progressively find pixels containing weaker sources. This iterative process continues until the background estimate converges or no new source pixels are identified.

After source pixel identification over the grid of scales, we group contiguous pixels, and use their wavelet-coefficients to reconstruct a scale-independent final image while determining source size, shape, intensity, and clustering properties.

34.2 Statistical issues

34.2.1 Detection threshold

We follow Damiani et al. (*this volume*) in establishing a source detection threshold by comparing wavelet correlation coefficients of the data ($W(x, y; \sigma_x, \sigma_y) \equiv < MH * data >$) with the correlation coefficients computed from source-free, locally-flat backgrounds. We use simulations to determine, as a function of background count density C , the sampling proba-

¹AXAF Science Center at the University of Chicago

bility distribution for the wavelet coefficients: we use this distribution to determine the significance of source detection, and to set a threshold $W_T(C)$ based upon limiting the number of false source pixels to (say) one per image. Note that the finite number of the simulations imply that the predicted correlation value is subject to a non-trivial uncertainty. We include this effect by computing the variance $\sigma_w^2(C)$ of W_T during the simulations, and flagging source pixels found with $W_T(C) - 3\sigma_w < W(x, y) < W_T(C) + 3\sigma_w$ as marginal sources.

34.2.2 Exposure corrections

Variations in exposure $E(x, y)$ across the field-of-view lead to deterministic changes in the wavelet correlation coefficients of the image. Correcting for these changes allows us to analyze data in the entire field-of-view without regard to the presence of instrumental structure such as the ribs seen in ROSAT/PSPC.

A *model* of the background count rate $b(x, y)$ is first constructed, e.g., by flat-fielding the background map obtained iteratively. The contribution of the exposure variations to the wavelet correlation coefficients at some reference pixel (x_0, y_0) is

$$\langle MH * (b \cdot \Delta E) \rangle = \langle MH * (b \cdot E) \rangle - \langle MH * b \rangle \cdot E.$$

This value is *subtracted* from the observed correlation.

34.2.3 Image reconstruction

Image reconstruction from a wavelet decomposition presents several unresolved statistical issues. Because the correlation values of neighboring pixels are not statistically independent quantities, it is a non-trivial task to move from stating the significance of source detection in each pixel to stating the overall significance of a “detected” source. While the significance of detection of point sources modeled with Gaussian spread profiles has been analyzed by Damiani et al. (1996, submitted to ApJ), the case of extended sources and asymmetrical point sources remains an open question. This also presents the related problem of identifying the size, shape, and strength of extended sources. Beyond that, we have to deal with image reconstruction across scales, for which we must not only estimate final source significance but also how the choice of method (selection of coefficients, scales, grouping of pixels into sources, etc.) effects this final result. Work on resolving these problems is in progress.

Acknowledgments: We acknowledge useful discussions with Robert Rosner, Don Lamb, Salvatore Sciortino, Albert Bijaoui, and Francesco Damiani. This study was supported by the AXAF Science Center.

Wavelet and Multifractal Analyses of Spatial and Temporal Solar Activity Variations

J. K. Lawrence, A. C. Cadavid and
A. A. Ruzmaikin¹

The spatial distribution of magnetic fields on the solar surface and the temporal distribution of solar magnetic activity each show the dual properties of intermittence and scaling symmetry. Both distributions therefore define measures amenable to multifractal and wavelet analyses (Cadavid, et al 1994; Lawrence, et al 1993, 1995a, 1995b, 1996). We present examples of applications of these techniques to the monthly Wolf Sunspot Number from 1749 to 1993, to global temperature fluctuations of the Earth, and to a high-resolution ($\sim 2''$), digital, polarimetric image of line-of-sight magnetic field in quiet solar photosphere made with the San Fernando Observatory vacuum solar telescope and video spectra-spectroheliograph system.

A display of the modulus of the complex Morlet wavelet transform (for a review of wavelet methods see Farge 1992) of the Wolf number shows, in addition to the well-known 11-year sunspot cycle, numerous episodes of magnetic flux eruption on the Sun with characteristic time scales between a few months and a few years. We also display a spectrum of the “flatness” $\Phi \equiv <\mu^4> / <\mu^2>^2$ of the variation of the Wolf number calculated from its Haar wavelet transform. This indicates a transition from an intermittent regime ($\Phi > 3$) on time scales shorter than 2 years to a Gaussian regime ($\Phi = 3$) at longer times.

The same analyses have been applied to global average temperature anomalies for the Northern and Southern hemispheres of the Earth between 1858 and 1993 (Jones 1994). The modulus of the wavelet transform shows considerable structure on time scales of 2 - 7 years. Early and late in both time series there also is structure on time scales near the 11-year sunspot cycle period. Examination of the real part of the wavelet transform indicates that in the period 1860 - 1920 the terrestrial temperature fluctua-

¹San Fernando Observatory, Department of Physics and Astronomy, California State University, Northridge, Northridge, CA 91330-8268

tions are out of phase with the sunspot number, but from 1950 - 1993 they are in phase. Interestingly, on time scales < 2 years, where the sunspot number series is intermittent, the terrestrial temperature anomalies show a flatness $\Phi \approx 3$ indicating a random, Gaussian distribution.

One can also calculate a flatness spectrum of the two-dimensional magnetic image using a simplified “Mexican Hat” wavelet. This indicates characteristic scales in the image of sizes ~ 2 Mm and ~ 6 Mm. A transformed image at the smaller scale enables the detection of small magnetic flux features. A more powerful, multifractal approach involves calculation of the pointwise Hölder exponents α of the magnetic flux and looking for localized singularities (points where $\alpha < 2$). If this is carried out for the distribution of flux irrespective of polarity, then the singular points reveal the locations of otherwise inconspicuous flux elements. If, on the other hand, fluxes of opposite polarity are cancelled, then the singular points reveal locations of bipolarity or neutral lines between regions dominated by opposite polarities. Such locations may be the sites of physical flux cancellation and energy release in the form of microflares or X-ray bright points.

REFERENCES

- [1] Cadavid, A.C., Lawrence, J.K., Ruzmaikin, A.A. & Kayleng-Knight, A. 1994, ApJ, 429, 391.
- [2] Farge, M. 1992, Ann. Rev. Fluid Mech., 24, 395.
- [3] Jones, P.D. 1994, J. Climate, 7, 1794.
- [4] Lawrence, J.K., Cadavid, A.C. & Ruzmaikin, A.A. 1995a, Phys. Rev. E, 51, 316.
- [5] Lawrence, J.K., Cadavid, A.C. & Ruzmaikin, A.A. 1995b, ApJ, 455, 366.
- [6] Lawrence, J.K., Cadavid, A.C. & Ruzmaikin, A.A. 1996, ApJ, 465, 425.
- [7] Lawrence, J.K., Ruzmaikin, A.A. & Cadavid, A.C. 1993, ApJ, 417, 805.

Wavelets in Unevenly Spaced Data: OJ 287 light curve

Harry J. Lehto¹

ABSTRACT The three factors that often characterize astronomical time series are uneven spacing, noise and the small number of data points. Because of these limiting factors data analysis methods should themselves be studied before applying them to the data. We have investigated scalograms using Haar wavelets in simulated data sets with various samplings. We point out sampling artifacts and introduce a revised scalogram which we call noisegram. We also apply this method to the century long optical V-band flux density light curve of the blazar OJ 287 to analyze its variations on time scales from a day to a century and indicate the presence of variability over the entire range.

36.1 Introduction

The century long light curve consists of three differently sampled sections. The ~ 180 photographic observations from the 1890's to 1971 were all serendipitous. The detection of optical variability (Folsom et al, 1971) and subsequent identification as a BL Lac-type object ($z = 0.306$, Miller et al. 1978) resulted in a higher sampling rate with about 1000 data points in 1971-1993 with a yearly sampling modulation.

During 1993-1996 the intense observing of OJ 287 by the OJ-94 campaign has resulted in over 3000 V band points showing two predicted outbursts (Sillanpää, 1988, Lehto and Valtonen, 1996). The light curve has a strong diurnal modulation in sampling due to lack of data from Far East. Attempts to analyze the full century long data set as a single entity have so far been defied by the pathological sampling.

36.2 Scalograms

Wavelet analysis is in principle capable of disentangling variability on various timescales. We have used Haar wavelets and calculated scalograms as defined by Scargle (1996). First we applied the method to artificial data

¹Tuorla Observatory, FIN-21500 Piikkiö, Finland

sets with different kind of samplings each consisting of 1000-2000 points initially, uniformly spaced, randomly spaced, with gaps or with higher density of points at one end. Then we removed points from the light curve each time recalculating the scalegram. Only white noise was used for our simulated data points.

For evenly spaced data white noise is expected to create a flat scalegram, which we detect. Using a large number of realizations (~ 500) we identify the following artefacts that are caused only by non-uniformity of sampling:

- 1) As points are removed from initially equally spaced data sets the scalegram remains flat on timescales longer than $\sim 4\langle t \rangle$, where $\langle t \rangle$ is the average sampling time. On shorter timescales the scalegram has a slope equal to $\Delta \log(V)/\Delta \log(t) = 1$. This effect becomes noticeable in individual scalegrams when about 50% of the original data set has been removed. Initially unevenly spaced data sets showed weakly positive slope from the very beginning of the simulation.
- 2) If the light curve consists of two parts with each of them having very different sampling rates, then the scalegram shows locally a rapid jump in the scalegram at the longer of the sampling rates.
- 3) If the light curve has no single characteristic sampling interval then the scalegram of white noise has a slope close to unity. White noise sampled at the same rate as data in the light curve of OJ 287 shows also a slope of unity up to $t = 2500d$, the size of the largest gaps in the data.

All these artefacts can be understood as being caused when the light curve has locally a time scale below which the light curve is not sampled. This will cause an overestimation in the amplitude of the wavelet components and thus an overestimation in the value of the scalegram at long time scales.

We also estimated the error of one realization from the simulations. Changing the variance of the white noise in the simulation, shifts the scalegram along the ordinate only. The shape and the errors remain identical, thus we adopt $\sigma^2 = 1$ without loosing generality.

36.3 Noisegrams

If we consider two scalegrams of white noise and a fixed time series and further divide them by each other we obtain a flat curve, i.e. a constant, independent of timescale. The constant has a numerical value equal to the ratio of the variances of the two light curves. If we choose the divisor to have unit variance, then this ratio gives directly the variance of the other light curve. We can define the noisogram of a given light curve as the ratio of the scalegram of the light curve to the scalegram of a unit variance white noise light curve with the same sampling (Lehto, 1996).

The noisogram tells us how much variability is present on average at a given local timescale.

The noisogram of OJ 287 gives a gentle slope of ~ 0.6 . The maximum of the scalegram is at about 10^4 days or 30 years. This corresponds to a ~ 60 year variation seen in the base brightness of OJ 287 (Lehto and Valtonen, 1996). No variability is seen on longer timescales. The gentle slope in the noisogram continues down to a timescale of about 1 day. This means that the variability amplitude on a given timescale corresponds to the variability one would get from white noise with an rms of $\sigma = (t_d)^{0.3} \text{mJy}$.

The noisogram level we expect from measurement errors is about $\log(\sigma^2) = -2$. This well below the levels at which we detect variability.

The noisograms of OJ 287 indicate that variability occurs over 4dex in timescales.

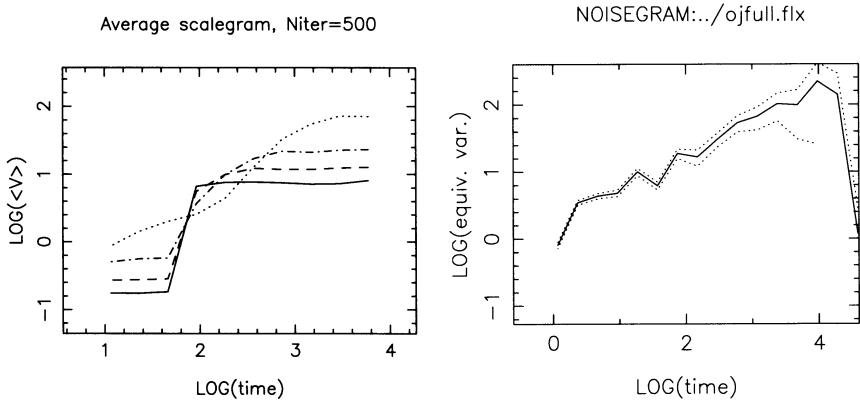


FIGURE 1. Left: Scalegram of 1100 points drawn from a Normal distribution. The first 100 are equally spaced and separated by 50 timesteps followed by the remaining points separated by 1 timesteps. Continuous line is the average scalegram of 500 realizations of the data set. The other lines represent scalegrams, when 30% (dash), 60% (dot-dash), and 90% (dot) of the data points of each simulated realization are removed. Right: Noisogram of OJ 287 covering timescales from 1d to 100 years.

Acknowledgments: The OJ-94 project observers have provided the recent data. Drs. Leo Takalo and Aimo Sillanpää have compiled most of the historical light curve. I wish to thank Jeff Scargle for constructive discussions on wavelets.

REFERENCES

- [1] Folsom, G.H. et al. 1971, *Ap.J.*, **169**, L131.
- [2] Lehto, H.J. and Valtonen, M.J. 1996 *Ap.J.*, **460**, 207.
- [3] Lehto, H.J. 1996 “Analysis of the OJ 287 V band flux density curve” in *Workshop on Blazar Variability* ed. H.R. Miller and J. Webb, in press.

- [4] Miller, J.S. et al. 1978 "Optical spectra of BL Lacertae objects" in *Pittsburg conf. on BL Lac objects*, ed. M.A. Wolfe, Univ. of Pittsburgh, p. 176.
- [5] Sillanpää, A. et al. 1988 *Ap.J.*, **325**, 628.
- [6] Scargle, J. 1996 in these proceedings.

Wavelet Based Analysis of Cosmic Gamma-Ray Burst Time Series

C. Alex Young, Dawn C. Meredith, and James M. Ryan

ABSTRACT Multiscale edge detection using wavelet transform maxima provides a robust method to compress information in a transient signal. We apply this method to Gamma-Ray Burst (GRB) time series data from the Compton Gamma Ray Observatory (CGRO). This provides a method to quantify the variability, identify structures, significantly suppress noise and compress the volume of data by as much as a factor of 10.

37.1 Multiscale edge detection

We explore the use of new tools for the study of gamma-ray burst (GRB) lightcurves. GRBs are transient bursts of gamma rays observed from uniformly distributed directions in the sky. Little is known about their origin. In addition to GRB's isotropy, the energy spectrum of GRBs lacks interesting features such as emission or absorption lines. The temporal observations of GRBs (lightcurves) promise to be enlightening because they contain the dynamics of the bursts. GRB lightcurves are nonstationary time series. Consequently, traditional time series analysis tools (e.g. Fourier transform) provide little useful information. The need for an adaptive transform that provides for the time dependence of spectral information leads to the use of the wavelet transform [1].

Our approach is to quantify the information in GRB lightcurves using the same method as the low level process of human vision [2, 3]. This means quantifying the sharp changes (edges) in a signal on multiple resolutions (MSED). We implemented this process using the extrema of a wavelet as an edge detector. We are able to make differentiations between the regularity of noise and real signals using properties of wavelets [2, 3]. By reconstructing this reduced extrema set we obtained an approximation of our signal which has reduced noise [2, 3].

Figure 1 shows the time series for GRB910602 and its time series after applying the wavelet noise reduction. This new signal is represented by a factor of 10 less data. We then choose the amount of information, the num-

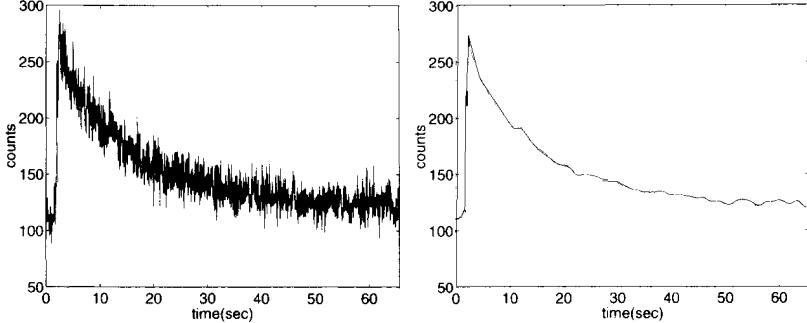


FIGURE 1. The left time series corresponds to the original burst GRB910602 and the right corresponds to the reconstructed burst GRB910602.

ber of parameters needed to describe the burst, as a quantifier of variability. By this definition, noise provides no information. The parameter representing this is the compression ratio CR which is the ratio of the amount of data before and after the MSED noise reduction [3]. We find no correlation between duration and CR . There is a strong correlation between peak flux at the 64ms timescale and CR . We believe this correlation is due both to a physical correlation between peak flux and variability and the fact that signals with less noise by definition have more information. Thus, we suggest using CR as a parameter to classify GRBs, helping to achieve the goal of understanding their generating mechanism.

We have shown MSED to be a promising tool for the study of GRB lightcurves. It enables us to reduce the noise content of a burst. It may also provide a means to parameterize the variability of GRBs. The compression ratio with some refinement might provide this measure, which will be analyzed in more detail. The compression ratio shows an interesting correlation with peak flux but we must first decide how much of this is due to a correlation between CR and signal-to-noise ratio. One solution to this may be the use of the Lipschitz regularity of the features in the burst in conjunction with CR . This connection may then allow us to quantify the bursts' variability as well as to describe the individual features in the burst.

Acknowledgments: This work was supported through NASA contract NAS5-26645 and through NASA Compton GRO guest investigation under grant NAG5-2731. Thanks also to Peter Bickel and Lianfen Qian.

REFERENCES

- [1] G. Kaiser, *A Friendly Guide to Wavelets* Birkhauser: Boston (1994).
- [2] S.G. Mallat and S. Zhong, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **7** (1992).
- [3] C.A. Young *Masters Thesis* University of New Hampshire (1996).

Smoothed Nonparametric Density Estimation for Censored or Truncated Samples

David M. Caditz

38.1 Introduction

Nonparametric distribution estimators such as the Kaplan-Meier [5] estimator for censored data and the Lynden-Bell [6] estimator for truncated data have been applied to astronomical data with good results. Such estimators, however, have limitations including failure to gracefully account for uncertainties in the detected or censored data [3] and the inability to directly estimate the differential distribution or frequency function. Both of these limitations are related to the treatment of the empirical source distribution, $\rho(\mathbf{x})$, as discrete, i.e., as a sum of delta functions:

$$\rho(\mathbf{x}) = \sum_{i=1}^N \delta(\mathbf{x} - \mathbf{x}_i), \quad (38.1)$$

where $\{\mathbf{x}_i\}$ are the reported observations. While such treatment is appropriate for, e.g., survival analysis, where loss and failure times are well known, it may be problematic in astronomical applications where luminosities, source distances, and flux limits have large uncertainties. Data smoothing has been applied in similar situations to uncensored and untruncated data [4, 7]. This poster and the accompanying reprints describe the application of data smoothing to truncated samples using the QSO luminosity function as a specific example.

38.2 The QSO luminosity function

Truncated QSO luminosity data is depicted in Figure 1 where the parameter space $\mathbf{x} = (L, L_{min})$, the source luminosity and minimum detectable luminosity, respectively. This sample is truncated by the constraint

$L \geq L_{min}$. The LB cumulative distribution function for L may be estimated from this data by the product

$$\Phi(> L_i) = \prod_{j=1}^i \left(1 + \frac{1}{N_j}\right),$$

where N_j is the number of detections in the region $L > L_j$ and $L_{min} < L_j$ and the sources are ordered by decreasing L . This cumulative distribution function is shown in Figure 1.

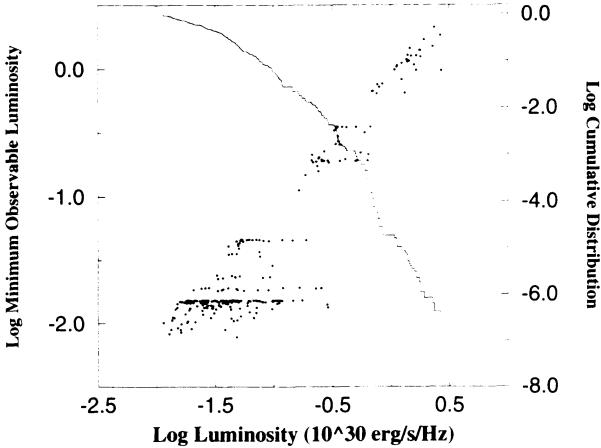


FIGURE 1. QSO survey data plotted on the L , L_{min} plane (left axis) and cumulative luminosity function obtained using the LB estimator.

38.3 Data smoothing

This poster describes a generalization to the LB estimator, based on data smoothing. The fundamental approach is to treat each observation as a smooth 'kernel', $K(\mathbf{x})$, of width h as opposed to a delta function distribution of equation (38.1). The observed distribution is then given by:

$$\rho(\mathbf{x}) = \sum_{i=1}^N \frac{1}{h_i^d} K\left(\frac{1}{h_i}(\mathbf{x} - \mathbf{x}_i)\right).$$

which is depicted in Figure 2 for the QSO data mentioned above. For this calculation, we use a Gaussian kernel:

$$K(\mathbf{x}) = \frac{1}{(2\pi)^2} \exp\left(-\frac{1}{2}\mathbf{x}^2\right).$$

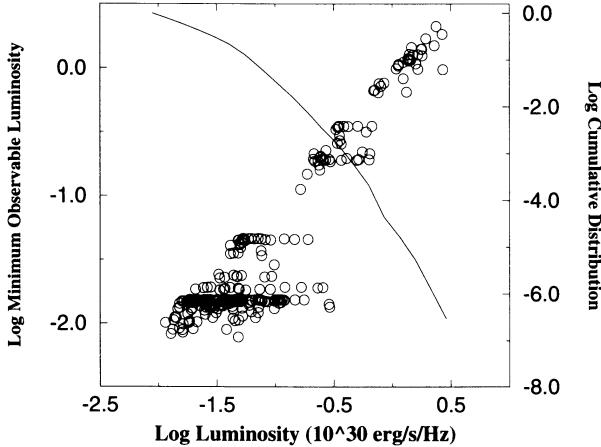


FIGURE 2. Smoothed survey data plotted on the L , L_{\min} plane (circles) and smoothed cumulative luminosity function.

The kernel width can be adjusted to reflect observational uncertainties (which may vary from source to source or across the observable parameter space) providing a natural means of incorporating these uncertainties into the distribution function. The radii of the circles in Figure 2 represent the smoothing width, h , used in this calculation.

The smoothed cumulative distribution function for truncated samples may be written [1, 2]:

$$\Phi(>x) = \exp \int_x^{\infty} \frac{\int_{x'}^{x'} \rho(x', y) dy}{\int_{x'}^{\infty} \int_{x'}^{x''} \rho(x'', y) dy dx''} dx' \quad (38.2)$$

where, for simplicity, we have assumed a bivariate distribution with observable range $x \geq y$. The cumulative distribution function for the QSO data (with $x = L$ and $y = L_{\min}$) is shown in Figure 2.

38.4 The differential distribution

Data smoothing allows the advantage of direct calculation of the differential distribution function (or density function) from the observed data. Differentiating equation (38.2), we find:

$$\Psi(x) = -d\Phi/dx = \Phi(x) \frac{\int_x^x \rho(x, y) dy}{\int_x^{\infty} \int_x^y \rho(x', y) dy dx'}. \quad (38.3)$$

The differential QSO luminosity distribution is shown in Figure 3.

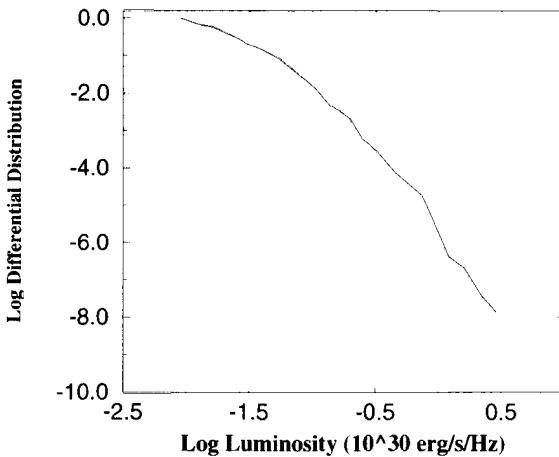


FIGURE 3. Smoothed differential QSO luminosity function obtained by application of equation (38.3) to the smoothed QSO survey data.

38.5 Conclusions

This poster presents data smoothing as a useful generalization to the well-known Lynden-Bell estimator for truncated data commonly encountered astronomical data analysis. The advantages of data smoothing include the ability to incorporate observational uncertainties into the analysis and the ability to estimate the differential distribution function from the observed data. As an example, we have applied smoothing to QSO luminosity data to construct the smoothed cumulative and differential luminosity functions. Other examples are described in the reprints provided as handouts.

REFERENCES

- [1] Caditz, D. M. & Petrosian, V. 1993, *Ap. J.*, 416, 450
- [2] Caditz, D. M. 1995, *Ap. J.*, 452, 140
- [3] Feigelson, E. D. & Nelson, P. I., 1985, *Ap. J* 293, 192
- [4] Hand, D. J. 1982, *Kernel Discriminant Analysis*, (New York: Wiley)
- [5] Kaplan, E. L. & Meier, P. 1958, *J. Am. Stat. Assoc.*, 53, 457
- [6] Lynden-Bell, D. 1971, *MNRAS*, 155, 95
- [7] Silverman, B. W. 1986, *Density Estimation for Statistics and Data Analysis*, (London: Chapman & Hall)

Luminosity and Kinematics: A Maximum Likelihood Algorithm for Exploitation of the Hipparcos Data

X. Luri, M. O. Mennessier, F. Figueras,
J. Torra, and A. E. Gómez

39.1 Introduction

The Maximum Likelihood (ML) algorithm presented here has been developed as a Ph. D. thesis work – Luri (1995) – in order to prepare a tool to obtain luminosity calibrations using the Hipparcos Mission results.

This method is conceived to complement the ones based purely on trigonometric parallaxes: it can use kinematical data (proper motions and radial velocities) in addition to trigonometric parallaxes to determine luminosity calibrations.

39.2 The method

The method is based in the *Maximum Likelihood* principle. Given a sample of stars, its likelihood \mathcal{L} is defined as $\mathcal{L}(\vec{\theta}) = \prod_{i=1}^{n_s} P(\vec{x}_i | \vec{\theta})$ where P is the *Probability Density Function* (PDF) of the random variable $\vec{x} = (m, \alpha, \delta, \pi_t, \mu_\alpha, \mu_\delta, v_r)$ and $\vec{\theta}$ is a vector composed by the parameters on which P depends. **The ML estimator of these parameters $\vec{\theta}_{ML}$ is defined as the one which maximizes the function $\mathcal{L}(\vec{\theta})$.**

To define the PDF of a sample several factors have to be taken into account. The first factor defining the PDF is the physics of the stars in the sample. In the examples of application of the method published in Luri (1995), Luri et al. (1996b) and Gómez et al. (1996) the following hypothesis have been made, but others can be taken when necessary: a Gaussian absolute magnitude distribution, a Schwarzschild velocity distribution, modified to include the effects of galactic differential rotation (Oort-Lindblad at first order) and a spatial distribution in exponential disk.

The way a sample has been selected has also a strong influence in the distribution of the observables in \vec{x} . If these effects are not taken into account, the ML estimation (and so the luminosity calibrations obtained) will be biased – see Luri (1993) for an example –. The selection of a sample is modeled using a function, the *Selection Function* S , giving for a star with a given value of \vec{x} the probability of being included in the sample. This function $S(\vec{x})$ is then included in the PDF of \vec{x} . The observational errors also affect the distribution of the quantities in \vec{x} , and ignoring them will lead to a bias in the ML estimation. Their distributions have been taken as being Gaussian and then included in the PDF of \vec{x} .

Finally, the interstellar absorption affects the values of the apparent magnitudes and so has to be taken into account. The Arenou et al. (1992) model has been included in the PDF of \vec{x} . For more information, a complete description of the method and its implementation can be found in Luri (1995) and Luri et al. (1996a).

39.3 Group separation

The samples of stars are seldom homogeneous, even if they are selected with a definite criteria as the spectral type. They are usually the mixture of several groups of stars with different physical characteristics. The PDF defined above does not properly describe an inhomogeneous sample and then can not provide good ML estimations. However, the PDF of such a sample can be written as the sum of the PDFs of the groups composing it. $\mathcal{P}(\vec{x}|\vec{\Theta}) = \sum_{j=1}^{n_g} w_j P_j(\vec{x}|\vec{\theta}_j)$, where w_j is the fraction of stars of the sample belonging to group j , P_j the PDF describing this group – constructed as detailed above – and $\vec{\Theta}$ a vector containing the parameters of all the groups and the w_j s. If the number of groups n_g is not known beforehand, a likelihood test like Wilks test – see Soubiran et al. (1990) – can be used to determine it.

Acknowledgments: This work was supported by the CICYT under contract ESP95-0180, and by the *Ayudas para la utilización de recursos científicos de carácter específico* by the DGICYT.

REFERENCES

- [1] Arenou, F. , Grenon, M. , Gómez, A.E. 1992. A&A, 258, 104
- [2] Gómez, A.E., Luri, X., Mennessier, M.O., Torra, J., Figueras, F. 1996, A&A. (submitted)
- [3] Luri X., 1995, Ph. D. Thesis, Univ. of Barcelona
- [4] Luri, X. , Mennessier, M.O., Torra, J., Figueras, F. 1993. A&A, 267, 305

- [5] Luri X., Mennessier M.O., Torra, J., Figueras, F., 1996a, A&A Sup. Ser., 117, 405
- [6] Luri X., Mennessier M.O., Torra, J., Figueras, F., 1996b, A&A , (in press)
- [7] Soubiran C., Gómez A.E., Arenou F. & Bougeard M.L. 1990, "Errors, bias and uncertainties in Astronomy", Eds. C. Jaschek y F. Murtagh, Cambridge University Press

Assessing Statistical Accuracy to the Orbital Elements of Visual Double Stars by Means of Bootstrap

G. Ruymakers¹ and J. Cuypers

The Keplerian orbit of the companion B of a visual binary with respect to the primary star A, can be described by the orbital elements: the revolution period P (in years), an epoch of passage through periastron T (in years), the semi-major axis a of the true orbit (in arcseconds), the eccentricity e (dimensionless), the (ascending) node Ω , the inclination i and the longitude of periastron ω (all in degrees) (see e.g. Dommangé 1992).

The estimation of accurate orbital elements is difficult (Eichhorn and Xu 1990): the equations to be solved are nonlinear and transcendental, and the arcs covered by the companion are usually short and sometimes not sufficiently curved to estimate reliable parameters.

According to the third law of Kepler, the total mass $M_A + M_B$ of the components A and B can be calculated from the orbital elements. Not much attention has been paid yet to calculate realistic estimates of the errors on the orbital elements and therefore not on the sum of the masses. Nowadays, algorithms are developed making use of modern computational facilities to calculate estimates of the orbital elements of visual binaries. For the investigation presented here, we have used the algorithm implemented by Pourbaix (1994).

We have studied the orbit of κ Peg (graded as ‘definitive orbit’ in Worley and Heintz (1984)). Modern observing techniques like speckle interferometry and astrometric satellites provide a significant increase in resolving power and in measurement accuracy for binary stars. We have computed the orbit using the observations that are not obtained by a speckle technique, the orbit based on speckle data alone, and the orbit based on the combined data.

The orbital elements and standard errors obtained by Pourbaix’s algorithm in combination of theory of propagation of error are:

¹Supported by The Belgian Federal Office for Scientific, Technical and Cultural Affairs

	non-speckle data 257 data		speckle data 55 data		combined data 312 data	
	estimated value	standard error	estimated value	standard error	estimated value	standard error
a	0.2386	0.0020	0.23484	0.00097	0.2380	0.0013
i	111.52	0.54	108.63	0.28	111.02	0.34
Ω	110.70	0.52	109.86	0.21	110.59	0.31
ω	124.55	0.88	122.84	0.47	124.16	0.59
e	0.2816	0.0051	0.3189	0.0022	0.2898	0.0031
P	11.5977	0.0046	11.549	0.11	11.5968	0.0033
T	1979.208	0.039	1979.189	0.11	1979.198	0.030

The orbital elements seem to be more accurately determined for the speckle data, except for the orbital elements T and P ! Because the non-speckle data range over a longer time interval (1880 - 1991) than the speckle data (1975 - 1993) the time aspect of the revolution is better determined in case of the non-speckle data. While other results agree within the errors, the estimates of inclination and eccentricity seem to be discrepant.

To study the reality of the error estimates, we have used bootstrap (Efron and Tibshirani 1993) to assess statistical accuracy to the orbital elements of κ Peg. We have made a bootstrap estimate of standard error of each of the orbital elements based on 5000 bootstrap samples from the two sets of observations and the combined set (see table). The orbits calculated for the speckle data and non-speckle data are in agreement, when we consider these error estimates.

Bootstrap estimates of standard error			
	non-speckle data	speckle data	combined data
a	0.0051	0.0013	0.0041
i	1.1	0.41	0.86
Ω	1.1	1.0	0.89
ω	3.4	1.4	2.6
e	0.014	0.0089	0.011
P	0.013	0.034	0.0083
T	0.13	0.04	0.12

We have to conclude that the measurements are still of great value, but which scheme of weights (if any) should be used to combine all observations? Bootstrap yields apparently more realistic estimates for the standard error than the ‘classic’ error estimation. More research will be done for other visual binaries (with less data and/or less covered arcs).

REFERENCES

- [1] Dommangé, J.:1992, in Benest D., Froeschlé (eds.), *HIPPARCOS*, 244–285
- [2] Efron, B. and Tibshirani, R.J.: 1993, *An Introduction to the Bootstrap*
- [3] Eichhorn, H.K. and Xu Yu-Lin: 1990, *Ap.J.* **358**, 575–587
- [4] Pourbaix, D.: 1994, *A&A* **290**, 682–691

- [5] Worley, C. and Heintz, W.D.: 1984, *4th Catalog of Orbits of Visual Binary Stars*

A Poisson Parable: Bias in Linear Least Squares Estimation

Wm. A. Wheaton

ABSTRACT A standard problem in high-energy astronomy data analysis is the decomposition of a set of I observed counts, n_i , described by Poisson statistics, for $i = 1, \dots, I$, according to some known J -component linear model,

$$\bar{n}_i = E[n_i] = \sum_{j=1}^J A_{ij} r_j, \quad (41.1)$$

with underlying physical count rates r_j or fluxes which are to be estimated from the data, the A_{ij} being known experiment constants. This problem is often solved by Linear Least Squares (LLSQ), but limited to situations where the number of counts per bin i is not too small.

For the simplest possible case, $J = 1$, which is just a counting experiment with no background, it is interesting to attempt a direct application of the weighted average formula using $\sqrt{\bar{n}_i} \approx \sigma_i$. However, the resulting formula is completely wrong! Using, instead of the observed n_i , the *expected count*, $E[n_i] = \sigma_i^2 = rt_i$ in the weighting, where t_i is the observing time in bin i , it turns out that the unknown rate r cancels from the weighted average sums, and we recover the obviously correct estimate $\hat{r} = N/T = \sum n_i / \sum t_i$.

41.1 Introduction

The problem of estimation from Poisson data has been solved by variations of linear least squares (LLSQ, often termed the “minimum chi-square method” or something similar), for many years in nuclear physics, high energy physics, and high-energy astronomy. In this paper I wish to draw attention once again to the problems caused by use of the observed counts to directly weight the LLSQ equations. I do so by examining the simplest possible Poisson estimation problem, and show that directly weighting each bin with $1/\sigma_i^2$ from the observed data n_i gives a manifestly wrong result.

41.2 A paradox

Thus, consider the problem of estimating a single count rate without any background at all, when the data have been binned into I bins, $i = 1, \dots, I$, each with n_i counts observed in livetime t_i (= total bin time - deadtime). It is known that $N = \sum n_i$ and $T = \sum t_i$ are “sufficient statistics” ([Leh59], pp 17–20) for this problem. That is, the maximally efficient estimator for the true rate r is a function of N and T only, so that the extra information due to the binning is superfluous. Nevertheless it is interesting to compare algorithms for handling binned data in this simple situation, for which equation (41.1) reduces to

$$\bar{n}_i = t_i r. \quad (41.2)$$

41.2.1 Weighted averaging

Consider first the weighted average of the count rate estimates for each bin, $\hat{r}_i = n_i/t_i$, with weights $w_i = 1/\sigma_i^2$. This is a plausible approach, since it is known that the weighted average, with weights $1/\sigma_i^2$, is the optimal (minimum variance) average. For a single sample, $\sigma = \hat{r}/\sqrt{N}$. Thus we take the weights to be $w_i = t_i^2/n_i$. The weighted average formula gives

$$\hat{r} = \frac{\sum w_i \hat{r}_i}{\sum w_i} = \frac{\sum t_i}{\sum (t_i^2/n_i)}, \quad (41.3)$$

which looks somewhat strange.

41.2.2 “Modified chi-square” method

What [EDJ⁺71] term the “modified χ^2 method” of LLSQ is the minimization of

$$\chi^2 \approx \sum_{i=1}^I (n_i - m_i)^2 / \sigma_i^2, \quad (41.4)$$

where m_i are the fitted or model counts in bin i , using the approximation that $\sigma_i = \sqrt{n_i}$, the observed data. Considering equations (41.2) as an $I \times J$ LLSQ system in the trivial case where $J = 1$, with one equation for each bin, we begin by weighting each equation by $1/\sigma_i \approx 1/\sqrt{n_i}$. Thus, after multiplication of both sides by the transpose of the weighted matrix (bringing in another factor of $1/\sigma_i$, note), we obtain:

$$\sum t_i = \sum \left\{ \frac{t_i^2}{n_i} \right\} r. \quad (41.5)$$

and obtain for our estimate of r :

$$\hat{r} = \frac{\sum t_i}{\sum (t_i^2/n_i)}. \quad (41.6)$$

the same as that derived from Eqn. 41.6.

Besides their unfamiliar look, these results cannot be expressed in terms of $\sum n_i$ and $\sum t_i$ alone, and are not even defined if $n_i = 0$ for any bin i . Yet the weighted average seems very reasonable, and is supported by a method which is just the trivial 1-parameter case of an algorithm which has been standard for many years. What has gone wrong?

41.2.3 Yet a third method

To answer this question we introduce a third method, again weighted averaging, but which recognizes that the true count rate r is *the same, by hypothesis*, in every bin. Thus the uncertainty should be derived from the expected counts \bar{n}_i in each bin, with

$$\bar{n}_i = t_i r. \quad (41.7)$$

Then $\sigma_i = r/\sqrt{\bar{n}_i}$, $w_i = t_i/r$, and we obtain

$$\hat{r} = \frac{\sum(n_i/r)}{\sum(t_i/r)}, = \frac{\sum n_i}{\sum t_i}, \quad (41.8)$$

as the unknown r cancels.

This is clearly the right answer, defined for all n_i , and the only one consistent with the known sufficiency of N and T , so we conclude that the answer (41.6) and (41.3) found by the other two methods is simply wrong.

41.3 Reconciliation of paradox

The failure of the weighted average method of Section 41.2.1 is easily understood. The theorem on the optimality and unbiasedness of the weighted average requires that we use the *true* variances σ_i in computing the weights, whereas in Sec. 41.2.1 we used an estimate. Most importantly, the estimate we used there was *correlated with the data*—that is, if \hat{r}_i happened to be low, it would receive a higher weight, and vice versa. Thus the estimate \hat{r} of (41.3) is systematically biased low.

The problem with the argument of Sec. 41.2.2 is basically the same, since we again have used the observed counts to form the weights. Since the $J = 1$ case of LLSQ turns out to be just the weighted average, both methods give the same wrong answer. Since the “modified chisquare method” is plainly incorrect even for $J = 1$, its use for larger values of J cannot be expected to be right. This has nothing to do with any assumption about the approximate normality of the Poisson distribution.

41.4 Summary and conclusions

In summary, the moral of our story is

1. The “modified χ^2 method” should be used with caution, or better, revised to used weights which are uncorrelated with the data in each bin.
2. Weighted averaging must be done in general with weights w_i which are uncorrelated with the data x_i , or else the weighted average will be biased.

It is seldom difficult in practice to find weights for the LLSQ equations which avoid the above problems. Simply making the weights uncorrelated with the data guarantees that the resulting estimate will be unbiased, and the efficiency of the estimate—that is its variance, compared to the minimum value in the ideal case when when $w_i = 1/\sigma_i^2$ —turns out to be quite insensitive to the exact choice of w_i . These matters have been discussed in more detail by [WDJ⁺95], who show that using roughly optimal but uncorrelated weights, the LLSQ method is effective even in the limit as the expected number of counts becomes arbitrarily small.

Acknowledgments: The work described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

REFERENCES

- [EDJ⁺71] W. T. Eadie, D. Drijard, F. E. James, M. Roos, and B. Sadoulet. *Statistical Methods in Experimental Physics*. North-Holland, Amsterdam, 1971. See especially chapters 7 and 8.
- [Leh59] E. L. Lehmann. *Testing Statistical Hypotheses*. John Wiley and Sons, New York, 1959.
- [WDJ⁺95] Wm. A. Wheaton, Alfred D. Dunklee, Allen S. Jacobson, James C. Ling, William A. Mahoney, and Robert G. Radocinski. Multiparameter linear least squares fitting to poisson data one count at a time. *The Astrophysical Journal*, 438(322), jan 1995.

Neural Network Classification of Stellar Spectra

Coryn Bailer-Jones¹, Mike Irwin², and Ted von Hippel³

ABSTRACT We have developed an automated stellar spectral classifier using a feed-forward artificial neural network and principal components analysis for front-end data compression. This classifier has been developed to classify spectra in the two-parameter domain (spectral type and luminosity class) of the MK system. We report a spectral type classification precision of 0.86 subtypes, and correct dwarf–giant classification in about 95% of cases, over a wide-range of spectral types (B–M).

42.1 Data acquisition and compression

We have digitized some 100 IIaO plates taken as part of the Michigan Spectral Survey⁴. This has provided a set of $\simeq 15000$ spectra down to $V \sim 11$ with a spectral coverage of $3800\text{\AA} - 5200\text{\AA}$ (two pixel resolution $\simeq 3\text{\AA}$). From these we selected 5000 high quality spectra across a wide range of spectral types (B3 to M5) which have 2D classifications from Houk⁵.

For spectral classification purposes a spectrum contains redundant information. Thus we can increase the speed and generalization ability of a parameterized classifier by compressing the spectra. Principal Components Analysis is a method of expressing the spectra in terms of a set of linearly independent basis vectors: The basis vectors have the convenient property that they are the eigenvectors of the covariance matrix of the data. Furthermore, the eigenvalues are proportional to the data variance explained by each eigenvector, so we can rank the eigenvectors according to their significance in representing the spectra. The spectra are then represented by their projections onto these eigenvectors: the admixture coefficients. We

¹Institute of Astronomy, Cambridge, UK, cobj@ast.cam.ac.uk

²Royal Greenwich Observatory, Cambridge, UK

³WIYN Telescope, National Optical Astronomical Observatory, Tucson

⁴Houk, N. 1984, in Garrison R.F., ed., *The MK Process and Stellar Classification*. David Dunlop Observatory, Toronto, p. 136.

⁵Houk, N. & Smith-Moore, M. 1988, *University of Michigan Catalogue of 2D Spectral Types for the HD Stars*, Vol. 4, and earlier volumes.

found that just the first 25 eigenvectors (out of 820: a compression factor of over 30) explained 95.8% of the variance in the data set, with the remaining eigenvectors being dominated by noise.

42.2 Artificial neural networks

An artificial neural network is an algorithm which can be trained to give an appropriate non-linear mapping between a set of data and their corresponding classifications. A neural network is a convenient approach to classification as it can give an arbitrary mapping to arbitrary precision⁶. The appropriate mapping is found by ‘training’ the network on a set of data, after which the network is used to classify ‘unknown’ data. We trained a neural network with three adaptive layers on half of our data set, using the 25 admixture coefficients as input to the network, and evaluated its performance on the other half of the data set (the *validation* set). For the spectral type problem, the neural network was used in a regression mode, because spectral types are better defined as subdivisions on a continuous sequence rather than as discrete classification bins. When the network classifications of the validation data set were compared with their classifications in the catalogue, the 1σ classification error was 0.86 spectral subtypes⁷. This compares with an intrinsic error of 0.63 spectral subtypes from the catalogue⁸. Luminosity classifications were obtained by using the neural network in probabilistic mode: each output gives an approximation of the posterior probability of class membership. The percentage of spectra in the validation data set correctly classified for classes III, IV and V were 91.0%, 18.9% and 96.5% respectively. Classes III and V were typically assigned a high posterior probability (84.2% of IIIs and 88.7% of Vs were assigned a probability of > 0.75). The success rate for class IV is very poor. Further tests imply that class IVs are not spectroscopically distinct from IIIs or Vs at this resolution, and so are not seen as a separate class.

These results were achieved using 5 nodes in each of the hidden layers: larger numbers of nodes (corresponding to a potentially more complex mapping) did not improve performance. We also trained networks with the 1,2,3...50 most significant eigenvectors, but found that classification performance using 26–50 components was no better than using just 25⁹.

Further work includes training neural networks on synthetic spectra and then using the networks to classify real spectra directly onto the physical parameters of T_{eff} , $\log g$ and [Fe/H].

⁶Bishop, C.M. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.

⁷One subtype is the difference between, for example, A6 and A7.

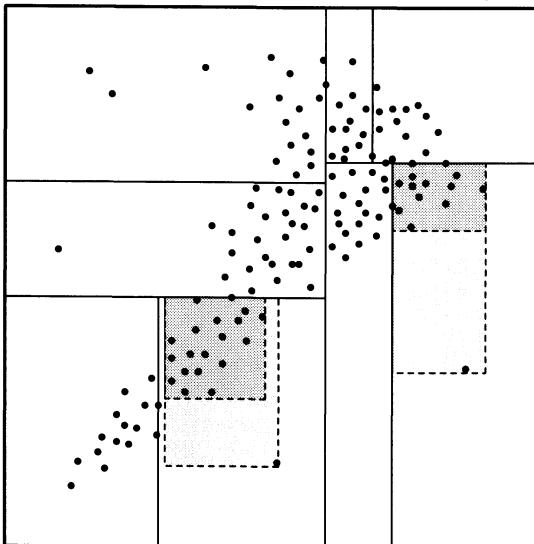
⁸Houk, N. 1995. Private communication.

⁹Bailer-Jones, C.A.L. 1996. *Ph.D. thesis*, University of Cambridge. In preparation.

Multidimensional Index for Highly Clustered Data with Large Density Contrasts

I. Csabai^{1,2}, A. Szalay¹, R. Brunner¹, and K. Ramaiyer¹

The SDSS [Sloan Digital Sky Survey] archive will contain multicolor data for over 100 million galaxies, with a volume of close to a Terabyte. Efficient searches in multicolor space will require novel indexing. Techniques such as the k-d tree are applicable, but less than optimal, since the most of the data is highly clustered, but a small fraction is randomly distributed, causing cells to be very imbalanced. We propose to resolve this by implementing an algorithm which splits the population into two high/low density parts.



The figure shows the distribution of galaxies in the two-color space. Points are highly concentrated with a few outliers. The solid lines show the k-d tree partitioning. In the selected two cells the optimal bounding box (dark grey) is smaller than the bounding box with the outliers (light

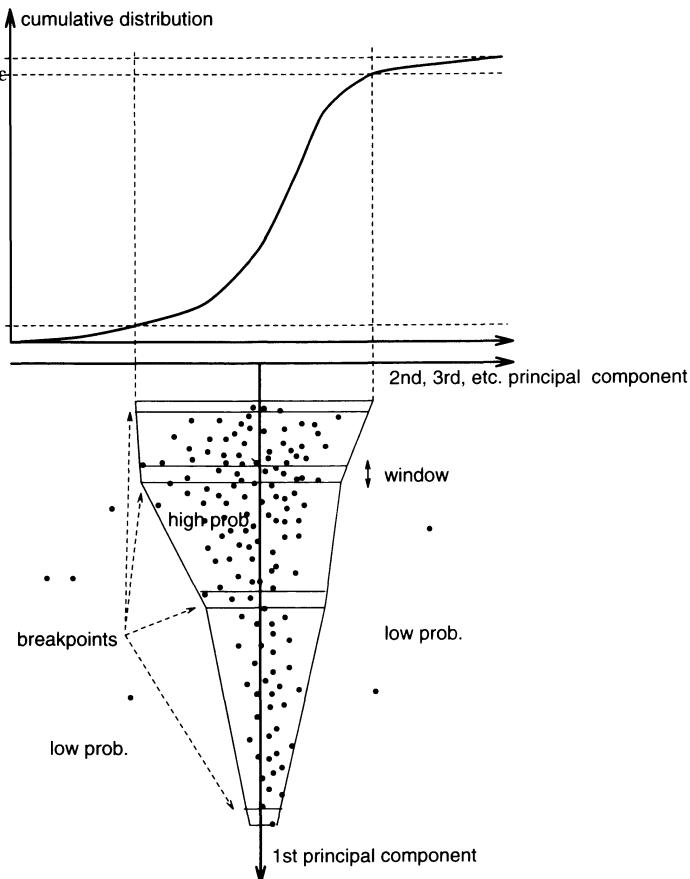
¹The Johns Hopkins University

²Roland Eötvös University, Budapest

grey). The difference of the volume of these bounding boxes are even larger in the multidimensional space.

The high/low density split algorithm:

- transform data into principal component system
- set n breakpoints along the principal axis. Each segment should contain equal number of objects
- select data in a *small* window around the breakpoint
- create cumulative distribution of selected data in all but the principal component
- find values where cumulative distribution reaches ϵ and $(1 - \epsilon)$ level. These points will serve as the vertices of the frame that separates high and low density data
- make the partitioning (e.g. k-d tree) separately for the two sets



Quantitative Morphology of Moderate Redshift Peculiar Galaxies

Avi Naim, Kavan U. Ratnatunga and Richard E. Griffiths¹

We report the results of training an Artificial Neural Network (ANN) classifier to distinguish between peculiar and normal galaxies on Hubble Space Telescope Wide Field Planetary Camera 2 images. 978 I band (filter F814W) images from 9 Groth-Westphal strip fields were classified by eye into five broad types (E/S0; early S; late S and two peculiar types). We examine the light concentration and asymmetry parameters (the C-A set) suggested by Abraham et al. (1995), as well as design a set of four morphological parameters (the 4P set) to describe each galaxy. Our parameters are:

1. Overall Texture (or “blobbiness”) : The degree of departure from a smooth, radially decreasing light distribution. Picks up bright localised structures.
2. The distortion of isophotes : The distance between the geometrical centers of different isophotes. Indicates overall distortions.
3. The filling-factor of isophotes : The degree to which regions enclosing isophotes are filled by pixels of that isophotal level. Indicates the existence of structures.
4. Skeleton ratios of detected structures : the ratio between the size of the skeleton and that of the corresponding region. Measures elongation of structures.

The ANN is trained on half of the galaxies and tested on the other half. There is a considerable mixture between normal and peculiar galaxies, both in the space spanned by the C-A set and in the 4P-set space. However, using

¹The Johns Hopkins University, Department of Physics & Astronomy, Baltimore, MD 21218, U.S.A.

the 4P set we manage to reduce the overlap and nearly double the success rate for classifying peculiar galaxies. This allows us to classify much larger sets, for which no eyeball classification is available.

We analyse the dominance of the bulge component in both peculiar and normal galaxies by examining the bulge-to-disk ratio derived from a maximum-likelihood fit of a photometric model to the images. While the majority of peculiar galaxies are disk-dominated, we also find evidence for a significant population of bulge-dominated peculiars. We conclude that peculiar galaxies do not all form a “natural” continuation of the Hubble sequence beyond the late spirals and the irregulars.

The trained neural network is applied to a second, larger sample of 1999 WFPC2 images and its probabilistic capabilities are used to estimate the frequency of peculiar galaxies at moderate redshifts as $35 \pm 15\%$.

Bayesian Inference on Mixed Luminosity Functions

David Walshaw¹

ABSTRACT A major area of statistical astronomy concerns the study of luminosity functions. A number of problems have been addressed in the literature, including estimation of intrinsic luminosity functions from surveys based on apparent brightness, and inferences based on censored samples. In this paper we consider an approach to inference based on mixture models. These are particularly appropriate when it is known or suspected that a class of observations does in fact consist of more than one category of object. A prime example concerns the brightest visible objects in the universe, namely the first ranked members of rich clusters of galaxies. Bhavsar proposes a model in which these are comprised of two distinct populations, and demonstrates that this performs considerably better than hypothesized models based on a single population. Here we take a Bayesian approach to making inferences on the individual components in such a mixture.

45.1 The model

The first-ranked galaxies of rich clusters are modeled using a two-component mixture. The bright extremes of ‘ordinary’ galaxies are assumed to follow a Generalized Extreme Value distribution (GEV). In a fraction α of the clusters, a second type of ‘special’ galaxy competes for first ranking. Following Bhavsar (1994) we model the magnitudes of these as being normally distributed. The cumulative distribution function for a first-ranked galaxy’s intrinsic luminosity X is then given by

$$F(x) = (1 - \alpha)G(x) + \alpha H(x), \quad (45.1)$$

where G is a GEV distribution function given by

$$G(x; \nu, \rho, \xi) = \exp(-[1 + \xi\{(x - \nu)/\rho\}]^{-1/\xi}),$$

and H is a normal distribution function with mean μ and standard deviation σ .

¹University of Newcastle upon Tyne, UK

45.2 Bayesian inference

As with most non-trivial Bayesian problems, an analytical approach to inference on the model (45.1) is not feasible. However, using iterative Markov Chain Monte Carlo methods it is possible to explore the posterior distributions arising in such complex situations. Here we will base inferences on a Gibbs sample drawn from the posterior distribution of the full parameter vector

$$\theta = (\alpha, \nu, \rho, \xi, \mu, \sigma) \quad (45.2)$$

using the method of *data augmentation*. This involves an extra step in each iteration of the Gibbs algorithm to simulate the classification of each of the first-ranked galaxies as either ‘ordinary’ or ‘special’. Each galaxy is randomly assigned to one particular class according to a binomial distribution whose weighting is determined by the posterior classification probability conditional on the luminosity, and the current estimate of the parameter vector. Non-informative prior information is provided by using a modified Jeffreys prior for each mixture component, and a conjugate Beta distribution for the mixing parameter α . For a full exposition of this approach, see Diebolt and Robert (1994).

45.3 Results

We use the same sample of 93 first-ranked cluster galaxies studied by Bhavsar (1994), with luminosities shown in Figure 1 (left). Results are based on a Gibbs sample of size 1000 drawn from the joint posterior for θ , after applying appropriate convergence criteria. Figure 1 (right) shows the sample from the marginal distribution of α , giving a good idea of the proportion of first-ranked galaxies which are ‘special’. In Figure 2 (left), we present the posterior sample for the function $\nu + \rho/\xi$, which corresponds

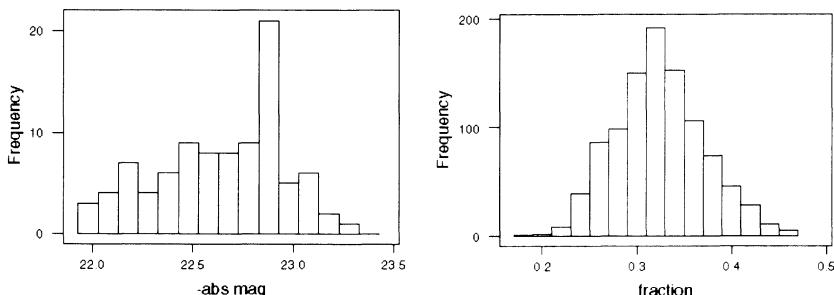


FIGURE 1. Left: Intrinsic luminosities of 1st-ranked galaxies. Right: First-ranked galaxies which are ‘special’.

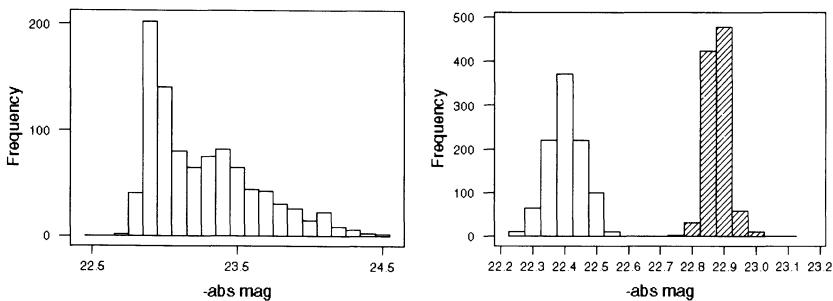


FIGURE 2. Left: Upper endpoint of ‘ordinary’ galaxy luminosity function. Right: Mean luminosities of ‘ordinary’ and ‘special’ galaxies.

to the upper endpoint of the distribution of intrinsic luminosities for ‘ordinary’ galaxies when $\xi > 0$. Finally, Figure 2 (right) shows the posterior distribution of the mean luminosity for each of the two classes of galaxy. The lack of overlap in these marginal distributions supports the argument that there really are two distinct types of object present.

REFERENCES

- [1] Bhavsar, S. P. (1994). Probing the nature of the brightest galaxies using extreme value theory, in *Extreme Value Theory and Applications*, (eds. J. Galambos et al.), 463-470.
- [2] Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling, *J. R. Statist. Soc. B*, **56**, 363-376.

Stochastic Solutions for the Hipparcos Astrometric Data Merging

F. Arenou, L. Lindegren, and R. Wielen

ABSTRACT During the Hipparcos astrometric reduction, an acceptable single or double star solution in agreement with the standard errors of the individual measurements could not be found for some stars. A stochastic model has been applied, which provides an estimate of the intrinsic scatter.

46.1 Introduction

The Hipparcos mission has determined accurate astrometric data – 5 parameters: positions, annual proper motions and absolute trigonometric parallaxes – for about 100 000 stars. The global data analysis task has been undertaken by two scientific Consortia, FAST and NDAC. They obtained independently two astrometric solutions which were merged in order to produce the Hipparcos Catalogue (118 218 entries). A comprehensive description of the mission may be found in [ES97].

Although satisfactory results were obtained for almost all stars, there existed some stars where none of the various models (single star, double star, orbital binary, etc) was adequate, thus leading to a high outlier rejection rate or a high goodness of fit. This may be due for instance to unrecognized duplicity or orbital motion.

Lacking an acceptable deterministic model, one solution could have been to apply robust procedures, which are designed to deal for instance with strongly non-Gaussian tails of the error distribution. In contrast, a stochastic model was assumed for the photocentric displacements superposed on a uniform motion of the concerned stars.

This was implemented by adding quadratically an intrinsic, excess scatter (cosmic error) to the standard errors of all observations. The correlation coefficients between FAST and NDAC observations, calibrated as a function of formal errors, increase with the cosmic error. These coefficients, together with the formal errors, provided the covariance matrix between observations.

For each star, the following procedure was adopted:

1. A normal astrometric solution was performed, still allowing the rejection of exceptional outliers:
 - (a) A weighted 5-parameter least-square solution was computed, using the individual observations (abscissae).
 - (b) The abscissae residuals were computed: the highest residual was rejected if it was greater than 3 times its standard error. If this was not the case, pairs of observations were tested: if the normalized difference was greater than $3\sqrt{2}$, the pair was rejected.
 - (c) The procedure was iterated until no outlier or more than 2 outliers was found. If no stochastic solution was found, the maximum number of outliers, 2, was decreased to 1, then 0.
2. Using the non-rejected observations, the cosmic error was computed, together with its standard error. This was done by iteration, finding when possible the value for which the unit-weight variance of the solution was 1.
3. Until the convergence of the cosmic error, the covariance matrix was computed for a new iteration.

46.2 Application

This method has been applied to all ‘single’ stars from the Catalogue with a good astrometric solution in order to define a very significant normalized cosmic error (corresponding approximately to a 3σ level of a Gaussian two-sided test). This criteria has been used to decide if the stochastic solution was retained or not.

Using the orbital binaries in the Catalogue, the cosmic error was found to be clearly correlated with their semi-major axis. This method also allowed the detection of stars with remaining bad input positions, which were found because of their high cosmic error.

The retained stochastic solutions have been applied as a last resort to the stars where all other models failed, 1561 stars of the final Hipparcos Catalogue, of which 643 are suspected doubles. These stars are possibly astrometric binaries with a period of less than a few years, and an estimate of the semi-major axis may then be obtained using the cosmic error found. This would not have been achieved with the blind use of other standard or robust methods.

REFERENCES

- [ES97] ESA. *The Hipparcos Catalogue*. ESA SP-1200, Vol I, 1997.

Identification of Nonlinear Factors in Cosmic Gamma-ray Bursts

Anton M. Chernenko

47.1 Introduction

There exist two approaches to time resolved spectroscopy of astrophysical phenomena. If the physics is understood well enough then it is possible to build a model spectral shape that contains real physical parameters of the source. Analysis of the observed spectra would then allow to directly determine variation of the physical parameters with time. If only a part M of total number of the parameters N are variable then it will be seen automatically.

If the physics is unknown, one is left with just empirical spectral model that would be ample in free parameters in order to fit almost any spectrum. If the number of the variable physical parameters M is less than number of the free empirical parameters then one will see some correlations between the latter ones. Only in case $M = 1$ (single variable parameter) the correlation is simple (linear or nonlinear). If $M > 1$ dependency of phenomenological parameters on each other becomes complicated for being described and understood. Investigation of such correlations is one of the main goals of the spectroscopy of gamma-ray bursts (GRBs) (e.g. [1], [3], [2]).

In this paper I describe an alternative approach: a nonlinear factor model of spectral variability that has been successfully used in [4].

47.2 Nonlinear factor model

Let us consider a matrix of accumulated spectra S_{ij} where i numbers energy channels and j numbers time intervals. This matrix could be viewed as a stack of time histories or a stack of spectra. Traditional approaches deal with spectra, while a factor model addresses time histories. A one-factor model here defines every time history as a power law function of a single hidden factor $X(t)$:

$$S_{ij} = a_i X(t_j)^{b_i}.$$

In such a model the variable parameter X is isolated from constant parameters a_i and b_i that could be used for comparison of different GRBs. However only a few of GRBs could be approximated by the single factor model. In other cases we tested so far, the following two-factor model works well:

$$S_{ij} = a_{1i}X_1(t_j)^{b_{1i}} + a_{2i}X_2(t_j)^{b_{2i}}.$$

Diversity of interrelations between time histories would produce a variety correlation patterns between parameters of empirical spectral models in the traditional approach. Factor analysis shows that they could be described by as few as 2 hidden apparently independent parameters, $X_1(t)$ and $X_2(t)$.

It should be noted that the first attempt to use factor models for GRB spectroscopy dealt with linear factors: $S_{ij} = \sum_k a_{ki}X_j$ [5]. The main drawback of the linear model, however, is that it does not allow the spectral shape of a given factor to change. This is known to not apply to GRBs.

47.3 Discussion

An obvious physical interpretation relates factors with separate physical emission components that might be further associated with separate emission processes. The fact that the spectrum of a component is completely determined by just a single parameter (flux) implies that during all of the burst the associated emission is controlled by a single physical parameter. Moreover, since the spectrum changes its shape with the flux, this physical parameter can not be additive in origin (like the emitting area, amount of emitting matter, etc.). Rather, it must be related to a physical property like temperature, strength of the magnetic field, or the like.

The analysis of the global relationship between the time histories $X_1(t)$, $X_2(t)$ of the two spectral components may allow us to determine whether they are related to a single energy reservoir or not, and may reveal spatial separation between distinct emission sites within a single source. An accurate determination of the relative contributions of the components to the bursts could allow us to estimate the distribution of energy between the emitters of the two components.

REFERENCES

- [1] Ford L. et al., 1994, in AIP Conference Proc **307**, p.268.
- [2] Briggs M., 1995, in Ann. of NY Acad. of Sci. **759**, p.416.
- [3] Pendleton G., 1996, in the Proc. of 3rd Huntsville Symposium on Gamma Ray Bursts
- [4] Cherenko A. and Mitrofanov I., MNRAS, **274**, 361, 1994.
- [5] Kozlenkov A., et al., 1992, in Proc. of Los Alamos Workshop on Gamma Ray Bursts, p.225.

Statistical Challenges in Asteroid Dynamics

S. Dikova¹

Observational data on asteroids, orbits especially, have to be interpreted in terms of origin and evolutionary processes. Cluster of asteroids are part of the puzzle. They are groupings of asteroids in the phase space of proper orbital elements. The lasts are quasi-integrals of motion, stable over very long intervals of time. Therefore they present a sort of "average" characteristics of motion in sense that they are result from a procedure of elimination of short and long periodic perturbations. The osculating elements represent the orbits in a fixed epoch.

We present a survey on a dynamical structure of the main asteroid belt comprising both distributions by proper and by osculating orbital elements. The results are displayed as two-dimensional plots. An argument is found versus Hirayama's hypothesis of common origin of the asteroid clusters through some primordial explosive event. This argument consists in the existence of an uniform background of asteroids in their distribution by proper elements. The plots are in good agreement with our thesis.

¹Institute for Astronomy, Bulgarian Academy of Sciences, Tzarigradsko Shousee 72, 1784 Sofia, Bulgaria; E-mail: SKYDYN@BGEARN.ACAD.BG

Brief Annotated Bibliography on Point Processes

Joseph Horowitz¹

49.1 Introduction

Three topics from mathematics and statistics arose repeatedly in the astronomical papers at the Statistical Challenges in Modern Astronomy II conference, namely, wavelets (not intrinsically statistical), time series, and point processes, in particular, the Poisson process and its many variants.

Although *Numerical Recipes* is touted as the statistical “bible of all astronomers” (Eric Feigelson’s phrase), it does not cover either time series or the Poisson process, much less its more complicated relatives, adequately. It does have useful material on spectral analysis of time series, and a smattering of other statistical topics; the famous ch. 15 is primarily concerned with nonlinear regression. There is virtually nothing on the Poisson process. Since it is a book on numerical methods, one does not expect *Numerical Recipes* to treat all these topics in any depth; on the other hand, it should therefore not be used as one’s sole source of statistical knowledge.

During the conference I told several people that I would send them a list of references on point processes, and here is the result. Professor Feigelson suggested that it be published in the conference proceedings as a convenience to the conferees.

I was somewhat disappointed, when I went through the point process literature, to find that much of it is at a rather sophisticated mathematical or statistical level that is often not very reader-friendly. Thus there are many lovely theorems, for example, on the asymptotic behavior of various estimators and hypothesis tests, but not too much of a directly utilitarian nature. Of course, such results provide the necessary theoretical guidance for the application of statistical methods, but many of them need quite a bit of decoding before they can be applied. This is a real failing of the statistical literature.

The following list of references consists of eleven books and three survey papers. Needless to say, it is my personal selection from the literature, and

¹Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003

I may have omitted some good items. There are many more books on the subject, some of which were excluded on the grounds that they were of more theoretical than applied interest. Papers on point processes appear regularly in many journals, so there is no doubt that I have missed many things that would be of direct use to astronomers.

REFERENCES

- [1] I. Basawa and B.L.S. Prakasa Rao, *Statistical Inference for Stochastic Processes*. Academic Press 1980. (Ch. 6 and section 8.4.)
- [2] D. R. Cox and V. Isham, *Point Processes*. Chapman and Hall 1980. (Theoretical but digestible.)
- [3] D. R. Cox and P.A.W. Lewis, *The Statistical Analysis of Series of Events*. Chapman and Hall 1966. (Old-fashioned, but lots of good, very direct, stuff.)
- [4] J. Goodman, *Statistical Optics*. Wiley 1984. (Not a statistics book, but has a nice discussion of the semiclassical model of radiation, in which several variants of the Poisson process play a central role.)
- [5] S. Karlin and H. Taylor, *A Course in Stochastic Processes*. (2 vols.) Academic Press 1981. (A general, intermediate-level work, vol. II contains some fairly elementary material on spatial Poisson processes.)
- [6] A. Karr, *Point Processes and their Statistical Analysis*. Marcel Dekker 1986. (Lots of interesting material, but quite theoretical; well-written.)
- [7] J. F. C. Kingman, *Poisson Processes* Oxford Univ. Press 1993. (Theoretical but clear.)
- [8] Yu. A. Kutoyants, *Parameter Estimation for Stochastic Processes*. Helder-mann 1984. (Ch.4 has some useful results on inference for Poisson processes; difficult going.)
- [9] P. A. W. Lewis (ed.), *Stochastic Point Processes: Statistical Analysis, Theory, and Applications* (Although dating from 1972, still very useful for practice.)
- [10] E. Parzen, *Stochastic Processes*. Holden-Day 1962. (An ancient, but very nice, intermediate book on stochastic processes, includes useful scattered material on Poisson and related processes and statistical analysis thereof.)
- [11] D. Snyder and M. Miller, *Random Point Processes in Time and Space* Springer 1991. (The 2nd ed. of an old, very applied book by Snyder; should be very useful for astronomers.)
- [12] D. Brillinger, Comparative aspects of the study of ordinary time series and of point processes. In *Developments in Statistics*, vol. I. P. Krishnaiah (ed.). Academic Press 1978. (Theoretical but interesting, full of useful references.)
- [13] M. Brown, Statistical analysis of nonhomogeneous Poisson processes. In [9]. (The basic theory for these processes, which arise again and again in astronomy.)
- [14] P. A. W. Lewis, Recent results in the statistical analysis of univariate point processes. In [9]. (This, along with [10] and [11], should be directly useful; see also the references in this paper, especially to Lewis's paper in *Theory of Sound and Vibr.*, 12, 1970.)

Testing the Hubble Law from Magnitude-Redshift Data of Field Galaxies: The Effect of the Shape of the Luminosity Function

O. Ullmann¹

The calibration of the redshift-distance relation $cz = Hr^p$, with z the redshift, r the distance and H and p two parameters, from (m,z) -data requires an initial guess on the luminosity function (LF) of the sample. The simplest assumption is that the LF is a Dirac function, that means that all considered objects have the same absolute magnitude M_0 (“standard candles”). Once the relation is calibrated it may then be used to determine vice versa the distance of a particular object or the LF of a total sample of objects. This is obviously a circular approach since a wrong initial assumption on the LF may have biased the calibration. Almost 20 years ago Segal and his co-workers have started a debate with their statement that, if the LF is only a function of M (no evolutionary effects), it is possible to determine the LF and the exponent p galaxy data.

We compared by simulation the classical standard candle approach and Segal's method and calculated the resulting bias in p as a function of the initial error in the LF. We started from the general statistical relation for the expectation of the apparent magnitude m under the absolute magnitude M and the Hubble velocity $v = cz$

$$E(m|M, v) = \frac{\int m f(M - (m - u(v))) P(m, v) A(m, v) dm}{\int f(M - (m - u(v))) P(m, v) A(m, v) dm} \quad (50.1)$$

where f is the distribution of the “true” M around the “calculated” $M = m - u(v)$ or the “intrinsic scatter of the distance-redshift relation”, $P(m, v)$ the joint probability that an object has the observables m and v , u the distance modulus and $0 \leq A(m, v) \leq 1$ a filter function which corrects for the sampling bias.

¹Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Straße 1, D-85740 Garching bei München, Federal Republic of Germany

Under certain assumptions on the functional forms of f , u , P (which contains implicitly the LF) and A we developed explicit expressions for $E(m|M, v)$. If one has for instance no intrinsic scatter, a of the data at m_g one gets with the above mentioned general form of the Gaussian LF with a dispersion σ and a mean M_0 , and a truncation redshift-distance relation from Equation 50.1

$$E(m|\log v) = \frac{5}{p} \log v + b - \sigma \sqrt{\frac{2}{\pi}} \frac{\exp \left[-\left(m_g - \left(\frac{5}{p} \log v + b \right) \right)^2 / 2\sigma^2 \right]}{1 + \operatorname{erf} \left[\left(m_g - \left(\frac{5}{p} \log v + b \right) \right) / \sqrt{2}\sigma \right]} \quad (50.2)$$

with $b = 25 - \frac{5}{p} \log H + M_0$. There are three parameters in Equation 50.2: p , b and σ which may be fitted to the data (it is determined simultaneously!). From these parameters the dispersion only shows up in the correction term, therefore it may be estimated only if the sample is biased, whilst the other two parameters are variables of both the naive formula and the correction term.

In order to demonstrate the effects on the estimation of p a series of simulations for various LFs, particularly asymmetrical distributions like Schechter functions, and spatial distributions of the objects have been carried out. It turned out that besides the dispersion the third moment of the distribution, the skewness, is an important parameter. Changes in skewness by 50% may well double the estimate of p .

Index

- Active galactic nuclei (AGN), 77ff, 126, 366, 371ff, 393ff
BL Lac objects, 340, 423ff
quasars (QSOs), 77ff, 87, 100ff, 170, 429ff
Advanced Satellite for Cosmology and Astrophysics (ASCA), 242
Advanced X-ray Astrophysics Facility (AXAF), 241ff, 358, 418ff
Akaike Information Criterion, 26, 355
aliasing, 270
analysis of variance, 270ff
Anderson-Darling test, 246
asteroids, 459ff
atomic representation, 341
autocorrelation function, 283ff
autoregressive (ARMA) models, 25ff, 130, 189, 317ff, 361, 383, 393ff
autoregressive process, 24ff
- Bayes' theorem, 19ff, 39ff, 54, 120ff
Bayesian image reconstruction, 403ff
Bayesian methods, 15ff, 39ff, 49ff, 63ff, 118ff, 247ff, 279, 355ff, 403ff, 409ff, 451ff
intrinsic Bayes factor, 22ff, 40ff
prior distribution, 19ff, 40ff, 54ff, 119ff, 228ff, 360ff, 380, 403ff, 451ff
Bayesian vs. frequentist, 35, 47, 63ff, 360ff, 373ff
- binary stars, 270ff, 379, 437ff
black holes, 226ff, 237ff, 321ff
bootstrap methods, 94, 169, 205, 277, 355ff, 373ff, 383, 392, 437ff
- CCD detectors, 124, 136, 210
celestial mechanics, 3ff, 15
censoring, 83ff, 100ff, 204, 368ff, 429ff
Central Limit Theorem, 6, 42
chaos, 303ff, 317ff, 339, 361, 376
chi-squared distribution, 109, 275, 324, 374
chi-squared test, 19, 40ff, 116, 205, 211ff, 246ff, 254ff, 269, 326ff, 366, 392, 441ff
CLEAN, 395ff
Compton Gamma Ray Observatory satellite, 427ff
confidence intervals, 46, 93, 120, 437ff
contingency table, 392
correlation integral, 162
correlogram, 395
correspondence analysis, 124
cosmic microwave background radiation, 79
cosmology, 67ff, 153ff, 226, 463ff
chronometric cosmology, 67ff, 104
covariance matrix, 8, 108, 227
Cox process, 162, 167
Cramer-Rao bound, 227
Cramer-von Mises test, 246
cross-validation, 143, 357

- Data coding and selection**, 123ff, 382
data compression, 338, 379
decision trees, 138ff, 354ff, 371
deconvolution, 111ff, 124ff, 253, 330, 395ff
density estimation, 111ff, 166ff, 429ff
 nearest neighbor methods, 138, 309ff, 356, 383
digital sky surveys, 135ff
dimensionality reduction, 142, 192ff
discrepancy function, 11
discriminant analysis, 124, 189
- Earth orientation**, 50
ecology, 170
edge effects, 161, 168
EM algorithm, 353, 358, 393
empirical Bayes, 111, 279, 353
empty space function, 154
errors-in-variables model (see
 measurement errors,
 heteroscedasticity)
estimation, 3ff, 39ff, 51ff, 225ff, 259ff, 275ff, 358, 373, 433ff, 437ff
Euler-Poincaré characteristic, 159
extreme value distribution, 451ff
- Fisher information matrix**, 238
flux-limited surveys, (see censoring
 and truncation)
Fourier analysis, 111ff, 129, 205, 225ff, 237ff, 271ff, 283ff, 321ff, 333ff, 352, 359, 361, 375, 381, 395ff
fractal analysis, 162ff, 336
- GAIA satellite**, 272
galaxy clustering (see also spatial
 point processes), 153ff, 166ff, 372, 380, 397ff
 topological genus, 154ff, 168, 372
void probability function, 154
galaxy clusters, 69ff, 153ff, 166ff, 397ff, 451
galaxy morphology, 181, 449ff
gamma function, 402
gamma-ray astronomy, 41ff, 337, 366, 376, 427ff, 457ff
gamma-ray bursts, 338, 381, 427ff, 457ff
Gauss-Bonnet theorem, 158
Gauss-Markoff Theory, 7ff
Gaussian filter, 158
Gaussian process, 226
Gibbs sampling, 32, 58ff, 64ff, 452ff
goodness-of-fit, 116, 380
gravitational lensing, 209ff, 352, 354, 375ff
gravitational wave astronomy, 225ff, 237ff, 352, 355, 376, 389ff
- Hazard function**, 89, 100ff
heteroscedasticity, 105ff, 118ff, 204, 262
hierarchical clustering, 249
Hipparchos satellite, 259ff, 275ff, 352, 375, 381, 413ff, 433ff, 455ff
history, astronomy, 3ff
history, statistics, 3ff, 15
Hubble constant, 153ff, 380
Hubble expansion, 67ff, 105ff, 153ff, 463ff
Hubble Space Telescope satellite, 339, 403ff, 449ff
hypothesis testing, 17ff, 39ff, 355
- Identifiability**, 91
image processing, 31, 45ff, 112, 124ff, 135ff, 173ff, 191, 204ff, 245ff, 319, 333ff, 366, 371ff, 381ff, 401ff, 403ff, 405ff, 407ff, 409ff, 417ff, 419ff, 421ff
incomplete β function, 407ff
infrared astronomy, 74ff, 272, 366

- intensity function (see also two-point correlation), 89ff, 154ff, 167
- interferometry, 272, 375
- intrinsic scatter, 116, 118ff, 455ff
- Kaplan-Meier estimator**, 83ff, 100ff, 379, 429ff
- kernel methods, 111ff, 158ff, 166ff, 299ff, 306ff, 401ff, 430ff
- Kolmogorov-Smirnov test, 205, 246, 355, 378, 393ff
- Large-scale structure** (see galaxy clustering)
- Laser Interferometry Gravitational-wave Observatory, 225ff, 237ff, 355ff, 377
- least absolute deviation, 6
- least squares, 3ff, 114ff, 118ff, 263, 271, 367, 373, 441ff
- likelihood (see also maximum likelihood), 55ff, 409ff
- likelihood ratio, 40, 229, 249, 255
- linear regression, 7ff, 118ff, 121ff, 188ff, 205
- log N -log S plot, 370, 383
- luminosity function, 72ff, 83ff, 100ff, 119ff, 379, 429ff, 433ff, 451ff, 463ff
- Lyapunov exponent, 304ff, 310ff, 375
- Lynden-Bell estimator, 72ff, 84ff, 100ff, 379, 429ff
- M -estimation**, 11
- MACHO project, 209ff, 354ff, 375, 377
- Magellanic Clouds, 211ff, 375
- magnitude, astronomy, 70ff, 88, 103ff, 155
- Malmquist bias, 69ff, 368
- Markov chain, 58
- Markov Chain Monte Carlo, 32ff, 47, 64ff, 120, 167, 256, 452ff
- Markov random field, 409ff
- maximum a posteriori (MAP), 122, 409ff
- maximum entropy, 21, 405
- maximum likelihood, 40, 72, 92, 111, 255, 278, 359ff, 383, 403ff, 407ff, 433ff
- measurement errors, 3ff, 105ff, 118ff, 121ff, 204, 243ff, 262ff, 276ff, 314, 323, 368, 434ff
- method of sieves, 360
- Metropolis-Hastings algorithm, 32ff, 58ff, 64
- Milky Way galaxy, 147, 413ff
- Minkowski functional, 154
- mixture models, 27ff, 111, 358, 451ff
- model uncertainties, 358
- moments, 9ff, 95, 108ff, 115, 119, 162, 168
- Moon, 49ff, 360, 380
- multi-resolution methods, 286, 333ff, 371, 381, 427ff
- multifractals, 372, 375, 383, 421ff
- multiscale analysis, 123ff, 173ff, 249, 405ff
- à trous algorithm, 129, 175ff, 406ff
- multivariate analysis, 109ff, 123ff, 187ff, 246ff, 447ff
- principal components analysis, 128, 149ff, 445ff, 448ff
- multivariate classification, 135ff, 149ff, 249, 356ff, 367, 371, 378, 383, 445ff, 447ff
- k -means partitioning, 128ff
- Wilks Λ test, 434ff
- N -body simulation**, 153, 159
- neural networks, 123ff, 130, 138, 149, 313ff, 354, 356ff, 371, 382, 403ff, 445ff, 449ff
- training sets, 137ff, 372
- neutrino astronomy, 42
- neutron stars, 225ff, 237ff, 321ff
- noise reduction, 43, 49ff, 124ff, 140, 178, 225ff, 237ff, 315, 333ff, 352ff, 371, 389ff, 401ff, 424ff, 427ff

- nonparametric methods, 72ff, 83ff, 111ff, 279, 303ff, 317ff, 358, 361, 374, 429ff
- normal distribution, 4ff, 17, 27, 93ff, 111, 115, 120, 167, 204, 243, 275, 280, 414ff, 418ff, 422ff, 434ff, 451ff, 463ff
- Object classification, 135ff, 149ff
- Ockham's razor, 25
- optical astronomy, 74ff, 135ff, 149ff, 154ff, 173ff, 366, 403ff, 445ff
- outliers, 29
- Photon counting, 118, 241ff, 259ff
- point spread function, 149, 211, 242, 246, 373, 403ff
- Poisson distribution, 94, 119, 163, 204, 249, 256, 331, 336, 359, 441ff
- Poisson process, 43ff, 49ff, 118ff, 124ff, 166ff, 178, 243ff, 256, 324ff, 352, 359, 397ff, 407ff, 417ff, 461ff
- projection pursuit, 189
- psychotherapist, 257
- pulsar, 41, 366
- pyramidal algorithm, 177
- Röntgen Satellite (ROSAT), 242, 420ff
- radio astronomy, 77ff, 366, 370
- rank methods, 145, 357, 371
- Rayleigh statistic, 44
- redshift, 69ff, 101, 154ff
- redshift surveys, 155
 - redshift-distance relation, 67ff, 463ff
- regression, nonlinear, 115ff, 188ff, 205, 246ff, 254ff, 358, 373, 437ff
- robust methods, 3ff, 21, 143ff, 279, 354, 455ff
- Saturn, 403ff
- scalogram, 335ff
- seeing, astronomy, 124, 143
- self-organizing feature map, 126ff
- semiparametric models, 353
- shoe clerk, 257
- smoothing, 125, 333ff, 410ff, 429ff
- software, astronomy, 150
- software, statistics, 34ff, 72ff, 140, 185ff, 203ff, 334, 337ff, 384
- source detection, 179, 213ff, 225ff, 249ff, 371, 381, 407ff, 417ff, 419ff
- spatial point processes (see also galaxy clustering), 153ff, 166ff, 367, 372, 397ff, 461ff
- speckle interferometry, 112
- spectra, astronomy, 124ff, 154ff, 243ff, 366, 372, 445ff, 457ff
- spline methods, 177, 271, 278, 313
- star-galaxy discrimination, 135ff, 149ff, 356
- state space models, 303, 315, 320, 375, 393ff
- stellar kinematics, 259ff, 352, 413ff, 433ff, 455ff
- stellar photometry, 259ff
- stochastic processes, 283ff
- Strong Law of Large Numbers, 97
- Sun, 339, 351, 372, 409ff, 421ff
- supernovae, 70
- Supernova 1987A, 42, 55
- survival analysis, 90ff, 429
- Time series analysis, 24ff, 43ff, 129, 188ff, 209ff, 225ff, 237ff, 259ff, 275ff, 283ff, 303ff, 317ff, 321ff, 333ff, 352, 354, 359, 361, 374ff, 377, 381ff, 389ff, 391ff, 393ff, 395ff, 421ff, 423ff, 427ff, 457ff, 461ff
- cross-correlation, 389
 - Lomb-Scargle periodogram, 275, 376
 - multivariate time series, 129
 - non-stationary time series, 283ff, 303ff, 317, 359

- periodicities, 42, 230ff, 259ff, 275ff, 359, 391ff
 prediction, 312ff
 quasi-periodic oscillations, 321ff, 376
 spectral analysis (see Fourier analysis)
 time-frequency analysis, 283ff, 338
 unevenly spaced data, 210ff, 259ff, 340, 352, 361, 375, 376, 395ff, 423ff
 Wolf sunspot series, 24, 130, 421ff
 truncation, 67ff, 83ff, 100ff, 155, 169, 204, 352, 368, 370, 378, 429ff
 two-point correlation (see also intensity function), 160, 169
 two-sample tests, 98ff, 108ff, 407ff
- Ultraviolet astronomy**, 126, 409ff
Ulysses satellite, 376
- V/V_{max} method**, 71ff, 370
variable stars, 335
variance-covariance matrix, 128
Very Long Baseline Interferometry, 50
- visualization methods**, 185ff, 334
- Wavelet analysis**, 123ff, 173ff, 189ff, 235, 249, 283ff, 333ff, 359, 361, 371ff, 375, 381, 403ff, 405ff, 413ff, 417ff, 419ff, 421ff, 423ff, 427ff, 461ff
weighted methods, 110ff, 118ff, 161ff, 248, 264, 358, 408ff, 441ff
Wilcoxon-Mann-Whitney test, 111, 189
World Wide Web sites, 185ff, 319, 342, 384
- X-ray astronomy**, 74ff, 241ff, 321ff, 337, 352, 358, 366, 373ff, 391ff, 393ff, 417ff, 419ff, 422ff
X-ray binary star systems, 321ff, 337, 366ff, 375, 382, 391ff
X-ray Timing Explorer satellite, 376
- Ziv-Zakai bound**, 228ff