

FESTSCHRIFT FOR LUCIEN LE CAM

Research Papers in Probability
and Statistics

DAVID POLLARD

ERIK TORGERSEN

GRACE L. YANG

Editors

FESTSCHRIFT FOR LUCIEN LE CAM

**Research Papers in
Probability and Statistics**

Springer Science+Business Media, LLC

FESTSCHRIFT FOR LUCIEN LE CAM

**Research Papers in
Probability and Statistics**

Editors

**David Pollard
Erik Torgersen
Grace L. Yang**



David Pollard
Department of Statistics
Yale University
New Haven, CT 06520
USA

Erik Torgersen
Department of Mathematics
University of Oslo
Blindern, Oslo 3
Norway

Grace L. Yang
Department of Mathematics
University of Maryland
College Park, MD 20742
USA

Library of Congress Cataloging-in-Publication Data
Festschrift for Lucien Le Cam : research papers in probability and statistics / editors, David Pollard, Erik Torgersen, Grace L. Yang.
— Research papers in probability and statistics.

p. cm.

Includes bibliographical references.

ISBN 978-1-4612-7323-3 ISBN 978-1-4612-1880-7 (eBook)

DOI 10.1007/978-1-4612-1880-7

1. Probabilities. 2. Mathematical statistics. I. Le Cam, Lucien M. (Lucien Marie), 1924– . II. Pollard, David, 1950– III. Torgersen, Erik N. IV. Yang, Grace L.

QA273.18.F47 1997

519.2—dc21

96-52745

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

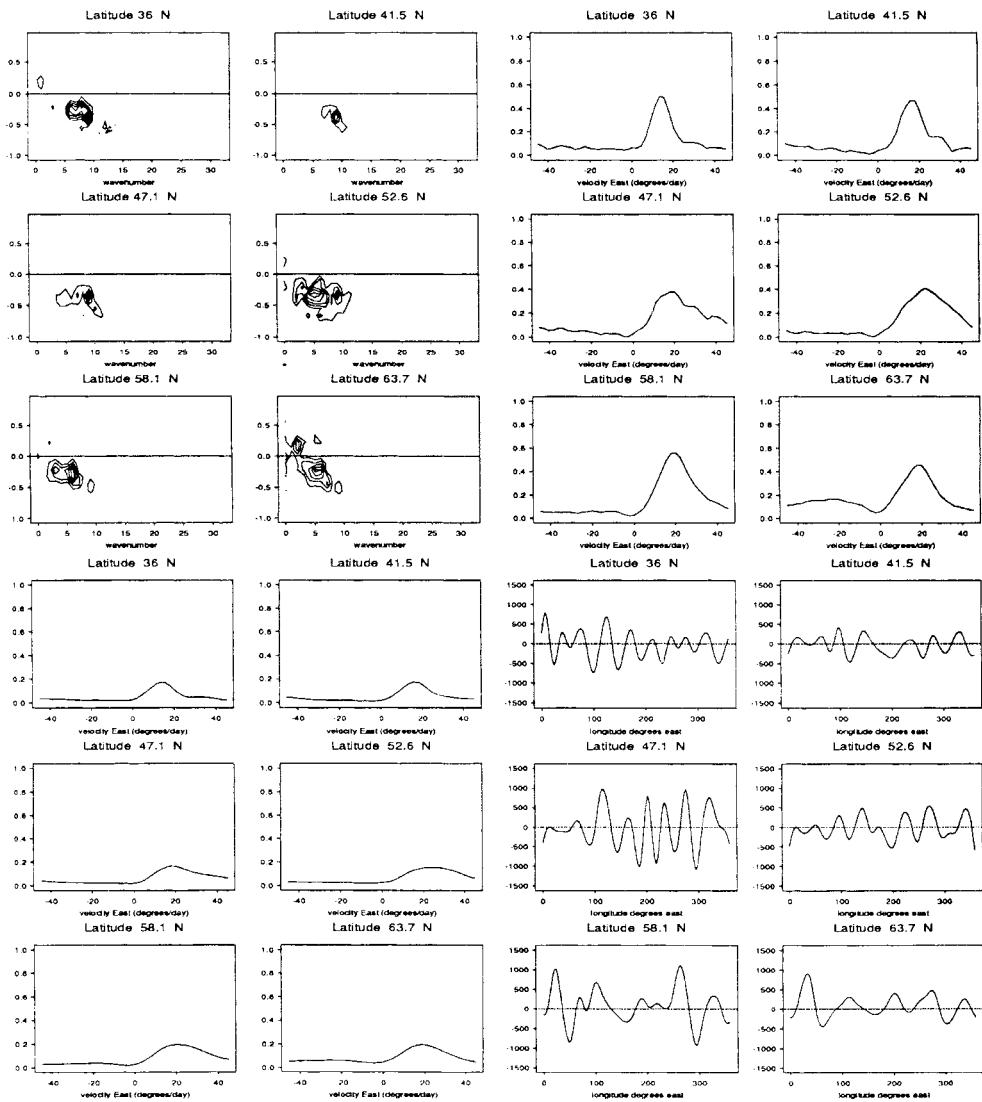
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage 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 Victoria Evarretta; manufacturing supervised by Joe Quatela.
Photocomposed copy prepared by the editors using LaTeX.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4612-7323-3



Figures 2-5 for Brillinger paper in *Festschrift for Lucien LeCam*

Preface

The articles in this volume were contributed by the friends of Lucien Le Cam on the occasion of his 70th birthday in November 1994. We wish him a belated happy birthday.

In addition to all the usual excuses for our tardiness in the preparation of the volume, we must point to the miracles of modern computing. As the old proverb almost put it: there's many a slip 'twixt \cup and \baselineskip. We beg forgiveness of any of our infinitely patient contributors who find that the final product does not quite match with the galley proofs.

Our task was also made harder by the sad death of our friend and fellow editor, Erik Torgersen.

We greatly appreciate the editorial help of David Donoho with one of the more troublesome contributions.

In addition to the 29 contributed articles, we have included a short vita, a list of publications, and a list of Lucien's Ph.D. students. We are also pleased that Lucien allowed us to include a private letter, written to Grace Yang, in response to a query about the extent of his formal mathematical training. The letter gives some insights into what made Lucien one of the leading mathematical statisticians of the century.

David Pollard and Grace Yang

Contents

Preface	v
Contributors	xi
Letter from Lucien Le Cam	xv
Biography of Lucien Le Cam	xix
Publications of Lucien Le Cam	xxiii
Students of Lucien Le Cam	xxxi
1 Counting Processes and Dynamic Modelling <i>Odd O. Aalen</i>	1
2 Multivariate Symmetry Models <i>R.J. Beran and P.W. Millar</i>	13
3 Local Asymptotic Normality of Ranks and Covariates in Transformation Models <i>P.J. Bickel and Y. Ritov</i>	43
4 From Model Selection to Adaptive Estimation <i>Lucien Birgé and Pascal Massart</i>	55
5 Large Deviations for Martingales <i>D. Blackwell</i>	89
6 An Application of Statistics to Meteorology: Estimation of Motion <i>David R. Brillinger</i>	93

7	At the Interface of Statistics and Medicine: Conflicting Paradigms <i>Vera S. Byers and Kenneth Gorelick</i>	107
8	Points Singuliers des Modèles Statistiques <i>D. Dacunha-Castelle</i>	135
9	Exponential Tightness and Projective Systems in Large Deviation Theory <i>A. de Acosta</i>	143
10	Consistency of Bayes Estimates for Nonparametric Regression: A Review <i>P. Diaconis and D.A. Freedman</i>	157
11	Renormalizing Experiments for Nonlinear Functionals <i>David L. Donoho</i>	167
12	Universal Near Minimaxity of Wavelet Shrinkage <i>D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard</i>	183
13	Empirical Processes and p-Variation <i>R.M. Dudley</i>	219
14	A Poisson Fishing Model <i>Thomas S. Ferguson</i>	235
15	Lower Bounds for Function Estimation <i>Catherine Huber</i>	245
16	Some Estimation Problems in Infinite Dimensional Gaussian White Noise <i>I. Ibragimov and R. Khasminskii</i>	259
17	On Asymptotic Inference in AR and Cointegrated Models With Unit Roots and Heavy Tailed Errors <i>P. Jeganathan</i>	275
18	Le Cam at Berkeley <i>E.L. Lehmann</i>	297
19	Another Look at Differentiability in Quadratic Mean <i>David Pollard</i>	305
20	On a Set of the First Category <i>Hein Putter and Willem R. van Zwet</i>	315

	Contents	ix
21 A Limiting Distribution Theorem <i>C.R. Rao and L.C. Zhao</i>	325	
22 Minimum Distance Estimates with Rates under ϕ -Mixing <i>George G. Roussas and Yannis G. Yatracos</i>	337	
23 Daniel Bernoulli, Leonhard Euler, and Maximum Likelihood (including a new translation of a paper by D. Bernoulli) <i>Stephen M. Stigler</i>	345	
24 Asymptotic Admissibility and Uniqueness of Efficient Estimates in Semiparametric Models <i>Helmut Strasser</i>	369	
25 Contiguity in Nonstationary Time Series <i>A.R. Swensen</i>	377	
26 More Optimality Properties of the Sequential Probability Ratio Test <i>E. Torgersen</i>	385	
27 Superefficiency <i>A.W. van der Vaart</i>	397	
28 Le Cam's Procedure and Sodium Channel Experiments <i>Grace L. Yang</i>	411	
29 Assouad, Fano, and Le Cam <i>Bin Yu</i>	423	

Contributors

ODD O. AALEN, University of Oslo, Institute for Basic Medical Sciences,
Blindern, N-06317 Oslo, PO Box 1122, Norway (*odd.aalen@basalmed.
uio.no*)

R. BERAN, Statistics Department, University of California, Berkeley, California 94720, USA (*beran@stat.berkeley.edu*)

PETER BICKEL, Statistics Department, University of California, Berkeley, California 94720, USA (*bickel@stat.berkeley.edu*)

LUCIEN BIRGÉ, URA 1321 "Statistique et modèles aléatoires," L.S.T.A., boîte 158, Université Paris VI, 4 Place Jussieu, F-75252 Paris Cedex 05, France (*lb@moka.ccr.jussieu.fr*)

DAVID BLACKWELL, Statistics Department, University of California, Berkeley, California 94720, USA (*blackwell@stat.berkeley.edu*)

DAVID BRILLINGER, Statistics Department, University of California, Berkeley, California 94720, USA (*brill@stat.berkeley.edu*)

VERA S. BYERS, M.D., Ph.D. President, Allergene Inc. and Adjunct Professor of Medicine University of California at San Francisco. Allergene Inc., 1650 Borel Place, Suite 234, San Mateo, California 94402, USA (*itsa.ucsf.edu*)

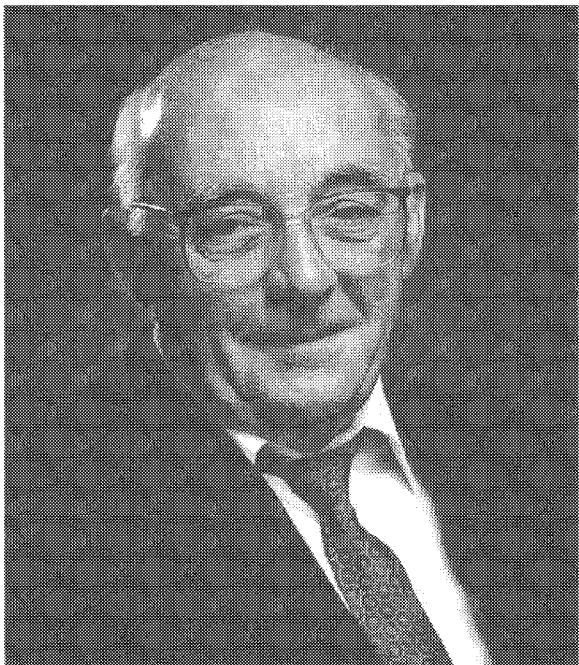
D. DACUNHA-CASTELLE, Département de Mathématiques, Bât. 425 91405 Orsay Cedex, France (*dacunha@cristal.matups.fr*)

ALEJANDRO DE ACOSTA, Department of Mathematics, Case Western Reserve University, Cleveland, Ohio 44106, USA (*add3@po.cwru.edu*)

- P. DIACONIS, Mathematics Department, Harvard University, Cambridge, Massachusetts 02138, USA
- DAVID L. DONOHO, Department of Statistics, Stanford University, Stanford 94305 and University of California at Berkeley, Berkeley, California 94720, USA (*donoho@playfair.stanford.edu*)
- R.M. DUDLEY, Mathematics Department, MIT, Room 2-245, MIT, Cambridge, Massachusetts 02139, USA (*rmd@math.mit.edu*)
- THOMAS S. FERGUSON, Mathematics Department, University of California at Los Angeles, Los Angeles, California 90024, USA (*tom@math.ucla.edu*)
- D.A. FREEDMAN, Statistics Department, University of California, Berkeley, California 94720, USA (*census@stat.berkeley.edu*)
- KENNETH GORELICK, M.D., Vice President of Drug Development, Chief Medical Officer, Genelabs Technologies, Inc. and Clinical Associate Professor Medicine, Stanford University. Genelabs Technologies, Inc. 505 Penobscot Drive, Redwood City, California 94063, USA (*pulmon@aol.com*)
- CATHERINE HUBER, Université Paris V, 45 rue des Saints-Pères, 75 270 Paris Cedex 06, France (*huber@citi2.fr*)
- I. IBRAGIMOV, St. Petersburg branch of the Steklov Mathematical Institute, 27 Fontanka, St. Petersburg, 191011, Russia (*ibr32@pdmi.ras.ru*)
- P. JEGANATHAN, Statistics Department, University of Michigan, Ann Arbor, Michigan 48109, USA (*jegan@stat.lsa.umich.edu*)
- I.M. JOHNSTONE, Statistics Department, Sequoia Hall, Stanford University, Stanford, California 94305, USA (*imj@playfair.stanford.edu*)
- G. KERKYACHARIAN, URA CNRS 1321, Mathématiques et Informatiques, Université de Picardie, 80039 Amiens, France
- R. KHASMINSKII, Department of Mathematics, Wayne State University, Detroit, Michigan 48202, USA (*rafail@math.wayne.edu*)
- E.L. LEHMANN, Statistics Department, University of California, Berkeley, California 94720, USA
- PASCAL MASSART, URA 743 “Modélisation stochastique et Statistique,” Bât. 425, Université Paris Sud, Campus d’Orsay, F-91405 Orsay Cedex, France. (*massart@stats.matups.fr*)
- P.W. MILLAR, Statistics Department, University of California, Berkeley, California 94720, USA (*millar@stat.berkeley.edu*)
- D. PICARD, URA CNRS 1321, Mathématiques, Université de Paris VII, 2 Place Jussieu, 75221 Paris Cedex 05, France
- DAVID POLLARD, Statistics Department, Yale University, Box 208290 Yale Station, New Haven, Connecticut 06520, USA (*pollard@stat.yale.edu*)

- HEIN PUTTER, Department of Mathematics and Computer Science, University of Leiden, PO Box 9512, Leiden 2300 RA, The Netherlands (*putter@cs.vu.nl*)
- C.R. RAO, Professor of Statistics and Holder of Eberly Chair, Department of Statistics, 326 Classroom Building, Pennsylvania State University, University Park, Pennsylvania 16802, USA (*crr1@psuvm.psu.edu*)
- Y. RITOV, Department of Statistics, Hebrew University, Jerusalem 91905, Israel (*yaacov@olive.mscc.huji.ac.il*)
- GEORGE G. ROUSSAS, Division of Statistics, 380 Kerr Hall, University of California, Davis, California 95616, USA (*roussas@wald.ucdavis.edu*)
- STEPHEN M. STIGLER, Ernest DeWitt Burton Professor, Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, Illinois 60637, USA (*stigler@galton.uchicago.edu*)
- HELMUT STRASSER, Department of Statistics, University of Economics and Business Administration, A-1090 Vienna, Austria (*strasser@stat2.wu-wien.ac.at*)
- ANDERS R. SWENSEN, Central Bureau of Statistics of Norway and University of Oslo, Department of Mathematics, P.O. Box 1053, University of Oslo, Blindern, 1 N-0316, Oslo, Norway (*swensen@math.uio.no*)
- E. TORGERSEN[†]
- A.W. VAN DER VAART, Faculteit der Wiskunde en Informatica, Vrije Universiteit, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands (*aad@fluit.cs.vu.nl*)
- WILLEM R. VAN ZWET, Department of Statistics, University of North Carolina, Chapel Hill, North Carolina 27599; Department of Mathematics, University of Leiden, PO Box 9512, Leiden 2300 RA, The Netherlands (*vanzwet@wi.leidenuniv.nl*)
- YANNIS G. YATRACOS, Department of Mathematics and Statistics, Université de Montréal, Montréal H3C 3J7, Canada (*yatracos@dms.umontreal.ca*)
- GRACE L. YANG, Department of Mathematics, University of Maryland, College Park, Maryland 20742, USA (*gly@math.umd.edu*)
- BIN YU, Statistics Department, University of California, Berkeley, California 94720, USA (*binyu@stat.berkeley.edu*)
- LIN CHENG ZHAO, Department of Mathematics, University of Science and Technology of China, Hefei, Anhui 230026, China

[†](Born 2 January 1935; died 25 January 1994). He received a Ph.D. in Statistics from University of California, Berkeley in 1968. He was Professor of Mathematics, University of Oslo, Norway; an elected member of ISI; a fellow of IMS; a member of the Norwegian Academy of Science; the author of the book "Comparison of Statistical Experiments," (Cambridge University Press, 1991). He was a co-editor of this Festschrift.



Letter from Lucien Le Cam

August 31, 1992

Dear Grace,

You made references to the fact that I don't have a formal mathematical education. This is true, and yet it is not. Writing after dinner and a glass of wine, I will try to fill in some of the details.

My elementary Math education was, I believe, standard for the country French of that time. We learned multiplication tables, up to 12×12 , and learned to perform long division. We also learned a few things about geometrical shapes and their names, trapezoid, parallelogram, square, circle, etc. In the meantime, the teacher had gotten a heart (a cow's heart) from a local butcher, and was teaching us about the functions of the four chambers. He also created some hydrogen from zinc and sulfuric acid and had an explosion.

By high school time, it was different. By the time I was 11 or 12 years old, I learned lots about cells; but cells in those days had no organelles, just a nucleus whose function was unknown. I learned also about Linnaeus classification of animals and that a lion and a cat are in the same family. Two years later, or so, we started on geometry, Euclidean type, similarity of triangles and such. The next year brought about Pythagoras' theorem. All of that was in the plane. We hit three dimensions only in my last year, 1941–42.

There we were subjected to an abundance of "conic sections." In the meantime, in the physics class, I had learned to solve linear differential

equations. You need them to understand the simplest electric circuits and the form of a “catenary.”

Some of my friends were stunning teachers by proving Bezout’s theorem, but I was more interested in other things. We had just learned how to solve quadratic equations; I presented my teacher with what amounts to $\int 1/\sqrt{a + bx + x^2} dx$, from geometric considerations.

Also I found, by chance, that the roots of algebraic quadratic equations were given by periodic continued fractions. This and the quadratic integral earned me a trip to the house of l’Abbé Mirguet, a mathematician priest, kept out of the regular priesthood, who ruled that I had better read some books, which, of course, were not to be found around.

I did not have any bent for math, but was very interested, fascinated, by Chemistry and Physics. I remember reading avidly a chemistry textbook written around 1830. It was most fascinating. The author had HO as [the] formula for water, [and] did not believe much in Dalton’s atomic theory. Besides, he had a few mistakes, like about ammonia gas dissolving in water and a wrong relation between the resulting volumes. I was able to disprove that in the chemistry lab, for which I had made a pass key. The Chemistry-Physics teacher found me there at odd times, but never said a word. Instead, he enlisted me to repair radio sets and burned out electric motors. He even asked me to construct a torsion balance, where to judge the tension of the wire, I had to use the pitch of a diapason. I broke some by tightening too much. This was interspersed with requests that I should saw up to reasonable lengths his supply of firewood. Old apple trees can be hard to saw by hand.

On the final high school exam, I was asked whether electrons exist. My teacher did not believe in them. I said “yes” and was asked “but, pray, tell me in what direction they orbit the atom?”

By that time, I had bought books in a store that took refuge from the Germans. One was Thompson (??) about “valence.” He computed forces between atoms by some exotic quantum mechanical technique. Another was a french translation of Max Planck lectures on electromagnetism. I still have that one. It was easy to read in 1940–41. I cannot read it now. The third book was lectures of Gauss on “Theorie der quadratischen Körper,” which was impossible and boring to read. There was a fourth book, about what is now called “radar.” It gave very detailed instructions about how to go about it. Since it had been written in 1937, I wonder what that means for the “inventors” of radar. Anyway, in 1942, the Thompson book and the “radar” book left me for undisclosed reasons.

The Max Planck lectures were full of “Curls” and “Divergences” and the like. I must have known or inferred what they meant. It certainly was not taught in my high school.

Going to Clermont-Ferrand in October 1942, I found I could not be accepted in the Chemistry program of the University. I was too late. They suggested I study math at the Lycée. Those kind people would even offer

me board and room. It was too late, but they offered me my noon meal for two years. There we had 16 hours of math lectures and 7 of physics, 6 of chemistry every week. The math was mostly 1800 style for engineers. Lots of drawing curves. The physics had a lot to do with the hydrogen thermometer and the chemistry was to puzzle out what kind of chemical the teacher may have put in a solution that looked like copper sulfate, but was not. It turned out to be some sort of methylene blue.

The very idea of "vectors" and linear algebra had not hit those teachers, even though they used "vectors" to represent forces in mechanics.

Then, after some difficulties, I went to Paris where I enrolled in the "Calculus" class. This was taught by two people, Valiron and Garnier. Garnier had surfaces that rolled onto each other. Valiron had a combination of real variables, complex variables, differential equations, partial differential equations, integral equation, and calculus of variations. Unfortunately he mostly repeated his book word for word and I hardly ever went to the lectures.

The Lebesgue integral was barely mentioned. Valiron said: "It is very simple. Let us work instead on the improper Riemann integrals."

I passed, just by chance, the "Calculus" exam and the "Rational mechanics" exam. However, I needed another "certificate." I had attempted to follow some lectures of Jean-Louis Destouches on quantum mechanics, but it was no hope. So, after some problems, I took the Statistics exam. Darmois was my examiner. He asked me to prove the multidimensional (matrix) version of Cramér-Rao. That was 1945. Cramér and Rao's papers were available in 1946. I did all right. I had never heard about matrices. It is true that Marcel Paul Schutzenberger had been holding seminars on van der Waerden's book on algebra, but he was a student, the crowd attending was uppity and I did not go.

I had more success with analysis. I had proved a few odd theorems, such as "If a set can be well ordered in two opposite directions, it is finite." I mentioned it to Colette Rothschild who said I better talk to "le Choc." That was Choquet. He was a student at the time and lived in a basement, working on industrial drawings. Choquet told me I better read de la Vallée Poussin's book on Baire classes.

Then I got employed by what was to be Electricité de France. That was a very nice employer, under the guidance of Pierre Massé for scientific ideas. He had anticipated much of Bellman's dynamic programming.

In the Spring of 1947, Electricité de France told me I could take courses at the Université if I wanted. I took a course from Julia on Hilbert Space. He was a phenomenal lecturer (and a Nazi) but covered only "elementary Hilbert Space theory." I learned mostly that Hilbert norms have special properties and that Hilbert spaces have orthonormal bases.

That in a way, is the extent of my formal mathematical formation. But there is more. I had taken a "subscription" to the Bourbaki books and liked to read them. For my needs at Electricité de France, I borrowed, and then

bought, Watson's on Bessel functions. I did read with interest parts of Paul Lévy's book of 1937, "Addition of random variables." In a fit of despair about the Navier-Stokes equations, I attended some of Leray's lectures on "fixed point theorems" at the Collège de France. We were worried about turbulence.

By 1947 Halphen persuaded me to publish a note about "characteristic functionals" in the Comptes Rendus. That was a bit miserable. After I came to Berkeley and met Bochner, he described it as "the work of an old man." I am sorry, I did not know any better. Once on an occasion, the appointed speaker in Darmois' weekly seminar did not show. So, being in charge of speakers, I spoke about Bochner's work on "Stochastic processes", because it was close to mine. Unfortunately, I could not answer several questions. Bochner had mentioned Kolmogorov's consistency theorem. I did not know what that was. Later on, I was to present a paper of Barankin, under similar circumstances. It referred to the Hahn-Banach theorem. I did not know what that was. Somebody suggested I should look at Banach's book, but it was missing from the library. This may have been in 1949.

At the same time Edith Mourier who was writing her thesis on probability on Banach spaces was pressing me for instances of practical applications. Not even knowing what a Banach space was induced some difficulties for applications.

All of that changed when I came to Berkeley. I think Neyman hired me as an applied statistician. However, Loève was lecturing on "measure theory," "Stochastic processes" and the like. Neyman was lecturing on uniformly most powerful unbiased tests. I had no idea what measure theory was about. My "practical" bent, from Electricité de France, made Neyman's lectures seem a bit spurious. Then, I was assigned a topic for qualifying examinations: "Fixed point theorems." I went at it with a vengeance, reading most of *Fundamenta Mathematicae*, the *Annals of Mathematics*, and such like, plus Saks "Integral," Kuratowski's topology, and a few other things.

Then, in April 1951, I flunked my qualifying examination.

I will tell you more some other time like my brush with Kantorovitch, Vulich, and Pinsker, but those were in Russian. I don't read Russian very well, if at all.

In those days, I could read, fast, very fast and I could remember. It is not sure how well I remembered. For instance I could remember in *English* stuff I read in Kuratowski in *French*, complete with the number of the page. The reverse occurred too.

In France, I had read the (Neyman-Pearson) Statistical Research Memoirs. At Berkeley, I became one of Neyman's students, though not a particularly cooperative one. Neyman complained at times that we never wrote a joint paper.

So that is the way it turned out.

Biography of Lucien Le Cam

Personal Data

Date of Birth November 18, 1924
Place of Birth Croze Creuse, France
Citizenship French

Education

1945 Licence es Sciences, University of Paris, France
1947-48 Graduate Studies, University of Paris, France
1952 Ph.D., University of California, Berkeley, California

Positions Held

1945-50	Statistician, Electricité de France, Paris
1950-52	Lecturer in Mathematics and Research Assistant, University of California, Berkeley
1952-53	Instructor in Mathematics and Junior Research Statistician, University of California, Berkeley
1953-55	Assistant Professor of Mathematics, Statistical Laboratory, University of California, Berkeley
1955-58	Assistant Professor, Department of Statistics, University of California, Berkeley
1958-60	Associate Professor, Department of Statistics, University of California, Berkeley

1960–	Professor, Department of Statistics, University of California, Berkeley
1961–65	Chairman, Department of Statistics, University of California, Berkeley
1957–58	Fellow of the Alfred P. Sloan Foundation
1971–72	Miller Professor, Department of Statistics, University of California, Berkeley
1972–73	Director, Centre de Recherches Mathématiques, Université de Montréal
1973–	Professor of Mathematics and Statistics, University of California, Berkeley

Professional Activities

1965	Editor (with J. Neyman) of Bernoulli (1723)-Bayes (1763)-Laplace (1813) Anniversary Volume, Springer, 1965
1967	Editor (with J. Neyman) of <i>Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, Vols. 1-V, (1967)</i>
1968–86	Associate Editor, <i>Zeitschrift für Wahrscheinlichkeitstheorie u.v. Gebiete</i>
1968–70	Member of Council, Institute of Mathematical Statistics
1970–72	Editor (with J. Neyman and E.L. Scott) of <i>Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley</i>
1973	President, Institute of Mathematical Statistics
1973	Member of Council, I.A.S.P.S.
1974–	Member comité aviseur du Centre de Recherches Mathématiques, Université de Montréal, Montréal, Canada
1974	Founder, Publications de la Chaire Aisenstadt, Les Presses de l'Université de Montréal, 1974
1976	Elected to American Academy of Arts and Sciences
1977	Elected Fellow of the American Association for the Advancement of Science
1979–	Associate Editor, Polish Journal of Probability and Mathematical Statistics
1981	Associate Director, Statistical Laboratory, University of California, Berkeley
1982	Member, New York Academy of Science

Member of Professional Societies

International Statistical Institute

Institute of Mathematical Statistics

American Mathematical Society

American Statistical Association

International Chinese Statistical Association

Publications of Lucien Le Cam

1. "Un instrument d'étude des fonctions aléatoires, la fonctionnelle caractéristique," *Comptes rendus des séances de l'Académie des Sciences*, Paris, **224** (1947), pp. 710–711.
2. "Sur certaines classes de fonctions aléatoires," *Comptes rendus des séances de l'Académie des Sciences*, Paris, **227** (1948), pp. 1206–1208, (with J. Bass).
3. "Les lois des débits des rivières francaises," *Houille Blanche*, Numéro Special B (1949), pp. 733–740, (with G. Morlat).
4. "On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates," *Univ. Calif. Publ. in Stat.*, **1**, No.11 (1953), pp. 277–329.
5. "Note on a theorem of Lionel Weiss," *Annals of Mathematical Statistics*, **25** (1954), pp. 791–794.
6. "An extension of Wald's theory of statistical decision functions," *Annals of Mathematical Statistics*, **26**, No. 1 (1955), pp. 69–81.
7. "On the asymptotic theory of estimation and testing hypotheses," *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, **I**, (1956), pp. 129–156.
8. "A remark on the roots of the maximum likelihood equation," *Annals of Mathematical Statistics*, **27**, No. 4 (1956), pp. 1174–1177, (with C. Kraft).
9. "Convergence in distribution of stochastic processes," *Univ. Calif. Publ. in Stat.*, **2**, No. 11 (1957), pp. 207–236.
10. "Remarques sur les variables aléatoires dans les espaces vectoriels non

- séparables," *Publications de l'Institut de Statistique de l' Université de Paris*, **VII**, Pts. 1–2 (1958), pp. 39–53.
11. "Une théorème sur la division d'un intervalle par des points pris au hasard," *Publications de l'Institut de Statistique de l'Université de Paris*, **VII**, Pts. 3–4 (1958), pp. 7–16.
 12. "Les propriétés asymptotiques des solutions de Bayes," *Publications de l'Institut de Statistique de l'Université de Paris*, **VII**, Pts. 3–4 (1958), pp. 17–35.
 13. "Locally asymptotically normal families of distributions," *Univ. Calif. Publ. in Stat.*, **3**, No. 2 (1960), pp. 37–98.
 14. "A necessary and sufficient condition for the existence of consistent estimates," *Annals of Mathematical Statistics*, **31** (1960), pp. 140–150, (with L. Schwartz).
 15. "The Poisson approximation to the Poisson binomial distribution," *Annals of Mathematical Statistics*, **31** (1960), pp. 737–740, (with J. L. Hodges, Jr.).
 16. "An approximation theorem for the Poisson binomial distribution," *Pacific Journal of Math.*, **10**, No. 4 (1960), pp. 1181–1197.
 17. "A stochastic description of precipitation," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. III (1961), pp. 165–186, Univ. of Calif. Press, Berkeley.
 18. "A note on the distribution of sums of independent random variables," *Proc. of Nat'l Acad. Sci.*, Vol. 50, No. 4 (1963), pp. 601–603.
 19. "Sufficiency and approximate sufficiency," (Special invited address Inst. Math. Stat., Dec. 1959.) *Annals of Mathematical Statistics*, Vol. 35, No. 4 (1964), pp. 1419–1455.
 20. "Consistent estimates and zero-one sets," *Annals of Mathematical Statistics*, Vol. 35, No. 1 (1964), pp. 157–161 (with L. Breiman and L. Schwartz).
 21. "On the distribution of sums of independent random variables," in *Bernoulli (1723) Bayes (1763) Laplace (1813)*, J. Neyman and L. Le Cam eds., Springer-Verlag (1965), pp. 179–202.
 22. "A remark on the central limit theorem," *Proc. Nat. Acad. Science*, August 1965, pp. 354–359.
 23. *Bernoulli (1723) Bayes (1763) Laplace (1813)*, J. Neyman and L. Le Cam eds., Springer-Verlag (1965).
 24. "Generalizations of Chernoff-Savage theorems on asymptotic normality on nonparametric test statistics," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I (1967), L. Le Cam and J. Neyman eds., Univ. of Calif. Press, Berkeley, (with M. Raghavachari and Z. Govindarajulu).
 25. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. Le Cam and J. Neyman eds., (1967), Univ. of Calif. Press, Berkeley and Los Angeles.
Vol. I: Theory of Statistics, pp. 1–666,

- Vol. II: Probability Theory - Part I, pp. 1-447, Part II, pp. 1-483,
 Vol. III: Physical Sciences and Engineering, pp. 1-324,
 Vol. IV: Biology and Problems of Health, pp. 1-934,
 Vol. V: Weather Modification Experiments, pp. 1-451.
26. "Likelihood functions for large numbers of independent observations," in *Research Papers in Statistics*, F.N. David, ed. John Wiley, New York (1966), pp. 167-187.
27. "Panel discussion on statistical inference," in *The Future of Statistics*, D. Watts, ed. Academic Press, New York (1968), pp. 139-160.
28. "Théorie asymptotique de la décision statistique," *Séminaire de Mathématiques Supérieures*-Été 1968, University of Montréal Press, Montréal, Canada (1969), pp. 7-143.
29. "Remarques sur le théorème limite central dans les espaces localement convexes," Colloques internationaux de Centre National de la Recherche Scientifique No. 186, June 1969, *Les probabilités sur les structures algébriques*. Editions du C.N.R.S., Paris, (1970), pp. 233-249.
30. "On the assumptions used to prove asymptotic normality of maximum likelihood estimates," *Annals of Mathematical Statistics*, Vol. 41, No. 3 (1970), pp. 802-828.
31. "On the weak convergence of probability measures," *Annals of Mathematical Statistics*, Vol. 41, No. 2 (1970), pp. 621-625.
32. "On semi-norms and probabilities, and abstract Wiener spaces," *Ann. Math.*, Vol. 93, No. 2 (1971), pp. 390-408, (with R.M. Dudley and Jacob Feldman).
33. "Limits of experiments," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I (1971), L. Le Cam, J. Neyman, and E. L. Scott, eds., Univ. of Calif. Press, Berkeley, pp. 245-261.
34. "Paul Lévy 1886-1971," in *Proceedings of the Sixth Symposium on Mathematical Statistics and Probability*, Vol. III (1972), L. Le Cam, J. Neyman, and E. L. Scott, eds., University of California Press, Berkeley, pp. xv-xx.
35. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, L. Le Cam, J. Neyman & E. L. Scott, eds, (1972), Univ. of Calif. Press, Berkeley and Los Angeles.
- Vol. 1: Theory of Statistics, 760 pp.
- Vol. 2: Probability Theory - Part one, 605 pp.
- Vol. 3: Probability Theory - Part two, 711 pp.
- Vol. 4: Biology And Problems of Health, 353 pp.
- Vol. 5: Darwinian, Neo-Darwinian, and Non-Darwinian Evolution, 369 pp.
- Vol. 6: Pollution And Health, 599 pp.
36. "Convergence of estimates under dimensionality restrictions," *Annals of Statistics*, Vol. 1, No. 1 (1973), pp. 38-53.

37. "Sur la loi des grands nombres pour des variables aléatoires de Bernoulli attachées à un arbre dyadique," (with A. Joffe and J. Neveu), *C.R. Acad. Sciences, Paris*, Vol. 277 (1973), pp. 963–964.
38. "Sur les contraintes imposées par les passages à la limite usuels en statistique," *Bulletin of the International Statistical Institute* (1973), Vienna, pp. 169–180.
39. "On the information contained in additional observations," *Annals of Statistics*, Vol. 2 (1974), pp. 630–649.
40. "Asymptotic methods in statistical decision theory, Vol. 1: Basic Structures," xvi + 270 pages, *Publications du C.R.M.*, Université de Montréal, Canada (1974).
41. "J. Neyman: on the occasion of his 80th birthday," *Annals of Statistics*, Vol. 2, No. 3 (1974), pp. vii–xiii, (with E. L. Lehmann).
42. "Distances between experiments," in *A Survey of Statistical Design and Linear Models*, J.N. Srivastava ed., North Holland, (1975), pp. 383–396.
43. "Construction of asymptotically sufficient estimates in some non-Gaussian situations," in *Proceedings of the Prague Symposium on Asymptotic Statistics*, J. Hájek ed., Academia Prague (1975).
44. "On local and global properties in the theory of asymptotic normality of experiments," in *Stochastic Processes and Related Topics*, Vol. 1 (1975), M. L. Puri ed., Academic Press, pp. 13–53.
45. Comment on Brad Efron's paper on "Defining the curvature of a statistical problem," *Annals of Statistics*, Vol. 3, No. 6 (1975), pp. 1223–1224.
46. "Circadian rhythm of stimulated lymphocyte blastogenesis," *Journal of Allergy and Clinical Immunology*, Vol. 58, No. 1, Part 2 (1976), pp. 181–189, (with M. Kaplan, *et al.*).
47. "An unusual metastatic lesion in a patient with osteosarcoma receiving tumor specific transfer factor," in *Transfer Factor*, M. S. Ascher, A. A. Gottlieb, and C. H. Kirkpatrick, eds., Academic Press, New York (1976), pp. 537–542, (with A.S. Levin, V.S. Byers, J.O. Johnson).
48. "Tumor specific transfer factor therapy in osteogenic sarcoma," *Annals of the New York Academy of Sciences*, Vol. 277 (1976), pp. 621–627, (with V.S. Byers, A.S. Levin, J.O. Johnston and A.J. Hackett).
49. "A reduction theorem for certain sequential experiments," in *Statistical Decision Theory and Related Topics*, Vol. II, S. Gupta and D. Moore, eds., Academic Press, N.Y. (1977).
50. "On the asymptotic normality of estimates," *Proceedings of the symposium to honour Jerzy Neyman*, R. Bartoszyński, E. Fidelis, and W. Klonecki, eds., Polish Scientific Publishers, Warsaw (1977).
51. "Identification of human populations with a high incidence of immunity against breast carcinoma," *Cancer Immunology and Immunotherapy*, Vol. 2 (1977), pp. 163–172 (with V.S. Byers, A.S. Levin, W.H. Stone, A.J. Hackett).
52. "A note on metastatistics—or, An essay towards stating a problem in the doctrine of chances," *SYNTHESE*, Vol. 36 (1977), pp. 133–160.

53. "On the asymptotic behavior of mixtures of Poisson distributions," *Zeitschrift für Wahrscheinlichkeitstheorie*, Vol. 44, (1978), pp. 1–45, (with R. Traxler).
54. "A reduction theorem for certain sequential experiments, II" *Annals of Statistics*, Vol. 7 (1979), pp. 847–859.
55. "On a theorem of J. Hájek," in *Contributions to Statistics* (Jaroslav Hájek Memorial Volume), J. Jurecková ed., Academia, Prague (1979), pp. 119–135.
56. "Immunotherapy of Osteogenic Sarcoma with Transfer Factor, Long-Term Follow-up," *Cancer Immunology and Immunotherapy*, Vol. 6 (1979), pp. 243–253, (with V.S. Byers, A.S. Levin, J.O. Johnson, A.J. Hackett).
57. *Jerzy Neyman - Biographical supplement to the International Encyclopedia of the Social Sciences*, Vol. 18, D. L. Sills, ed., Free Press (1979).
58. "Minimum chi-square - not maximum likelihood. Discussion of J. Berkson's paper," *Annals of Statistics*, Vol. 8 (1980), pp. 473–478.
59. "Limit theorems for empirical measures and Poissonization," in *Statistics and Probability, Essays in the honor of C.R. Rao*, G. Kallianpur, P. R. Krishnaiah, and J. K. Ghosh, eds., North-Holland (1982), pp. 455–463.
60. "On the risk of Bayes estimates," in *Statistical Decision Theory and Related Topics III*, Vol. II, S. S. Gupta and J. O. Berger, eds., Academic Press (1982), pp. 121–137.
61. "On some stochastic models of tumor growth and metastasis," in *Probability Models and Cancer*, L. Le Cam & J. Neyman, eds., North-Holland (1982), pp. 265–286.
62. *Probability Models and Cancer*, L. Le Cam and J. Neyman editors, North-Holland (1982), 301 pages.
63. "A remark on empirical measures," in *Festschrift in the honor of E.L. Lehmann*, P. Bickel, K. Doksum & J. L. Hodges, Jr. eds., Wadsworth (1982), pp. 305–327.
64. "Extension of a theorem of Chernoff and Lehmann," *Recent Advances in Statistics*, M. H. Rizvi, J. Rustagi & D. Siegmund, eds., Academic Press (1983), pp. 303–337, (with C. Mahan and A. Singh).
65. Review of *Statistical estimation: asymptotic theory*, by I. A. Ibragimov and R. Z. Has'minskii and *Contributions to a general asymptotic statistical theory* by J. Pfanzagl and W. Wefelmeyer, *Bull. Amer. Math. Soc.*, Vol. 11, No. 2, October 1984, pp. 392–400.
66. *Proceedings of the Berkeley Conference in honor of Jerzy Neyman and Jack Kiefer*, L. Le Cam and R.A. Olshen, eds., Vol. I and II, Wadsworth (1985).
67. "Sur l'approximation de familles de mesures par des familles gaussiennes," *Ann. Inst. Henri Poincaré*, Vol. 21, No. 3 (1985) pp. 255–287.
68. "On Lévy's martingale central limit theorem," *Sankhya*, Vol. 47, Series A, Part 2 (1985), pp. 141–155, (with P. Jeganathan).

69. "The central limit theorem around 1935," *Statistical Science*, Vol. 1, No. 1, (1986), pp. 78–96.
70. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag (1986).
71. Discussion of D. A. Freedman and P. Diaconis, "Consistency of Bayes Estimates," *Annals of Statistics* Vol. 14, No. 1 (1986), pp. 59–60.
72. "Convergence of stochastic empirical measures," *J. Multivariate Analysis*. Vol. 23 (1987), pp. 159–168, (with R.J. Beran and P.W. Millar).
73. Discussion of "The likelihood principle," by J. O. Berger and R. L. Wolpert, Institute of Mathematical Statistics Lecture Notes Monograph Series, Vol. 6, S. S. Gupta, ed., (1988), pp 182–185.2.
74. "Distinguished statistics, loss of information and a theorem of Robert B. Davies," in *Statistical decision theory and related topics IV*, S. S. Gupta and J. O. Berger, editors, Vol. II (1988), pp. 163–75, Springer-Verlag, (with Grace L. Yang).
75. "On the preservation of local asymptotic normality under information loss," *Annals of Statistics*, Vol. 16, No. 2 (1988), pp. 483–520, (with Grace L. Yang).
76. "*Asymptotics in statistics: Some basic concepts*," Springer-Verlag Series in Statistics (1990). Chinese version, Science Press, Beijing (December 1994), (with Grace L. Yang).
77. "Maximum likelihood—an introduction," Lecture Notes #18, Dept. of Math., Univ. of Maryland, and ISI Review Vol 58 #2 (1990), pp. 153–171.
78. "On the standard asymptotic confidence ellipsoids of Wald." ISI Review Vol 58 #2 (1990), pp. 129–152.
79. "Some recent results in the asymptotic theory of statistical estimation" Invited paper, International Congress of Mathematicians, Kyoto, August 21–29, 1990. Published in *Proceedings of the International Congress of Mathematicians*, Springer Verlag (1991), pp. 1083–1090.
80. Review of "Comparison of statistical experiments" by Erik Torgersen, *SIAM Review* Vol. 34 #4 (1992), pp. 669–671.
81. "Stochastic models of lesion induction and repair in yeast." *Mathematical Biosciences*, Vol. 112, 261–270 (1992).
82. "An infinite dimensional convolution theorem," in *Statistical decision theory and related topics V*, S. S. Gupta, and J. O. Berger, eds, 401–411. (1994).
83. "Neyman and stochastic models," *Probability and Mathematical Statistics*. Vol. 15, 37–45 (1995).

Publications to Appear

1. "Metric dimension and estimation," *The Proceeding 25th Anniversary Centre de Recherches Math. Univ. Montréal*, 1994.
2. "Comparison of experiments—A short review," in *Festschrift for David Blackwell*.
3. "La Statistique Mathématique depuis 1950," *Mathematics 1950–2000*.
4. "Asymptotic normality of experiments," *Encyclopedia of Statistical Sciences Up-date*, S. Kotz, C. Read, and D. Banks, eds., Wiley.

Technical Reports

1. "Harald Cramér and sums of independent random variables" Technical Report #103, Statistics, U.C. Berkeley, August 1987.
2. "On some stochastic models of the effect of radiation on cell survival," Technical Report #136, Statistics, U.C. Berkeley.
3. "On the Prokhorov distance between the empirical process and the associated Gaussian bridge." Technical Report #170, Statistics, U.C. Berkeley, September 1988.
4. "On measurability and convergence in distribution." Technical Report #211, Statistics, U.C. Berkeley.
5. "Some special results of measure theory" Technical Report #265, Statistics, U.C. Berkeley, August 1990.
6. "On the variance of estimates with prescribed expectations," Technical Report # 393, Statistics, U.C. Berkeley, July 1993.

In Preparation

"Stochastic models for sodium channels in nerve cells." To be submitted
(with Grace L. Yang).

Students of Lucien Le Cam

JULIUS RUBIN BLUM

Strong consistency of stochastic approximation methods, 1953

CHARLES HALL KRAFT

On the problem of consistent and uniformly consistent statistical procedures, 1954

BAYARD RANKIN

The concept of sets enchainied by a stochastic process and its use in cascade shower theory, 1955

GEORGES POWELL STECK

Limit theorems for conditional distributions, 1955

THOMAS SHELBYNE FERGUSON

I. On the existence of linear regression in linear structural relations. II. A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities, 1956

ISRAEL JACOB ABRAMS

Contributions to the stochastic theory of inventory, 1957

JAMES D. ESARY

A stochastic theory of accident survival and fatality, 1957

CHARLOTTE T. STRIEBEL

Efficient estimation of regression parameters for certain second order stationary processes, 1960

LORRAINE SCHWARTZ

Consistency of Bayes' procedures, 1960

HELEN WITTENBERG

Limiting distributions of random sums of independent random variables, 1963

GIAN DOMENICO MAJONE

Asymptotic behavior of Bayes' estimates in Borel sets, 1966

NORA SNOECK SMIRIGA

Stochastic processes with independent pieces, 1966

GRACE LO YANG

Contagion in stochastic models for epidemics, 1966

DAVID GOMBERG

Estimation of asymptotes, 1967

STEPHEN MACK STIGLER

Linear functions of order statistics, 1967

MIN-TE CHAO

Nonsequential optimal solutions of sequential decision problems, 1967

ERIK NIKOLAI TORGERSEN

Comparison of experiments when the parameter space is finite, 1968

ALEJANDRO DANIEL DE ACOSTA

Existence and convergence of probability measures in Banach spaces, 1969

PEDRO JESUS FERNANDEZ

On the weak convergence of random sums of independent random elements, 1970

ARNOLDO GUEVARA DE HOYOS

Continuity of some Gaussian processes parametrized by the compact convex sets in R^s , 1970

DETLEV LINDAE

Distributions of likelihood ratios and convergence of experiments, 1972

MOHD NAWAZ GORIA

Estimation of the location of discontinuities, 1972

ALOISIO PESSOA DE ARAUJO

On the central limit theorem in $C(0, 1)$, 1974

ROBERT HENRY TRAXLER

On tests for trend in renewal processes, 1974 (Biostatistics)

ODD OLAI AALEN

Statistical inference for a family of counting processes, 1975

ERIK JAN BALDER

An extension of duality-stability relations to nonconvex optimization problems, 1976

ERROL CHURCHILL CABY

Convergence of measures on uniform spaces, 1976

MOHAMED WALID MOUSSATAT

On the asymptotic theory of statistical experiments and some of its applications, 1976

NENG HSIN CHEN

On the construction of a non-parametric efficient location estimator, 1980

ANDERS SWENSEN

Asymptotic inference for a class of stochastic processes, 1980

SHAW HWA LO

Locally asymptotically minimax estimation for symmetric distribution functions and shift parameters, 1981

JANE LING WANG

Asymptotically minimax estimators for distributions with increasing failure rate, 1982

YANNIS YATRACOS

Uniformly consistent estimates and rates of convergence via minimum distance methods, 1983

IMKE JANSSEN

Stochastic models for the effects of radiation on cells in culture, 1984

BIN YU

Some results on empirical processes and stochastic complexity, 1990 (Jointly with Professor Terence Speed)

YU-LIN CHANG

Local behavior of mixtures of normal distributions, 1991

YU-LI GU

Minimax estimation for Poisson experiments, 1992

JAMES SCHMIDT

A computer simulation of the effect of radiation on a coil of DNA, 1993
(jointly with Professor Cornelius Tobias, Biophysics)

FESTSCHRIFT FOR LUCIEN LE CAM

**Research Papers in
Probability and Statistics**

1

Counting Processes and Dynamic Modelling

Odd O. Aalen¹

ABSTRACT I give some historical comments concerning the introduction of counting process theory into survival analysis. The concept of dynamic modelling of counting processes is discussed, focussing on the advantage of models that are *not* of proportional hazards type. The connection with a statistical definition of causality is pointed out. Finally, the concept of martingale residual processes is discussed briefly.

1.1 Introduction

In this article I shall discuss some issues related to the Ph.D. thesis I wrote under the supervision of Lucien Le Cam in Berkeley in 1973–75. The subject of the thesis was the use of martingale based counting process theory to develop statistical methodology for event history analysis. Since this approach has received a large amount of interest and is now considered the natural mathematical basis of survival and event history analysis, I will first give a brief description of the background and circumstances of the work done at Berkeley. Thereafter, I will discuss some areas where the established theory is still not sufficiently developed and where more work is needed. I will focus on three subjects: firstly, non-proportional regression models and time-dependent covariates, secondly, dynamic stochastic processes and causality, and thirdly, martingale residual processes. I will point out severe limitations in the Cox model when it comes to modelling the effect of stochastic processes on the hazard rate.

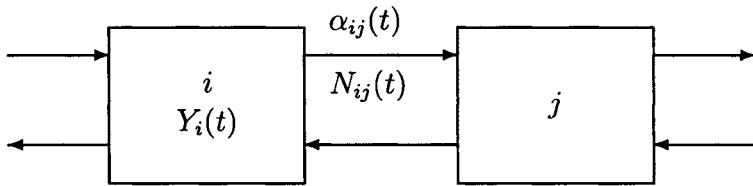
1.2 Some historical comments

The start of my own work can be traced back to the master thesis I finished in 1972 at the University of Oslo under the supervision of Jan M. Hoem. The subject of the thesis was to develop methodology to estimate the efficacy

¹University of Oslo

and risks of the intrauterine contraceptive device. Hoem had suggested to model the occurrence of various events that could happen to a woman who had had this device inserted, by a time-continuous Markov chain. This was a model he had used in a number of other contexts.

In my master thesis I suggested an estimator for the cumulative intensity, or hazard rate. It turned out that this estimator had previously been proposed by W. Nelson, and it was later termed the “Nelson-Aalen” estimator by other authors (see, for example, Andersen, Borgan, Gill & Keiding 1993). After finishing my master thesis, I continued studying this estimator, and also pondering under which circumstances it would be valid. It appeared obvious that it must be valid beyond the competing risks model for which I had originally developed its theory. Quite clearly, one should expect it to hold for any Markov chain when estimating the cumulative intensity of going from one state to another, under the assumption of complete observation.



Segment of a Markov process

Hoem had always insisted that when applying Markov chain models, one should make a figure of the state space with arrows indicating the possible transitions. This was of course not an original idea, but his insistence on always making the actual drawing is I think decisive for the next step that I made. A segment of a state space of an arbitrary time-continuous Markov chain showing two of the (possibly many) states and the accompanying transitions is given in the Figure. Assume that a number of individuals are moving independently on the state space. Let $Y_i(t)$ denote the number of individuals (the risk set) in state i , let $N_{ij}(t)$ denote the number of transitions that has taken place from state i to j up to time t , and let $\alpha_{ij}(t)$ denote the individual intensity of transition from i to j . It is then the integral of $\alpha_{ij}(t)$ which is estimated by the Nelson-Aalen estimator in the form $\int_0^t Y_i(t)^{-1} \mathbf{1}_{\{Y_i(t)>0\}} dN_{ij}(t)$. Looking at the figure one realizes that the Markov property of the movements of each individual on the total state space is really irrelevant for this estimation. The Nelson-Aalen estimator is sensible whenever the “force of transition” $\alpha_{ij}(t)$ operates on $Y_i(t)$ independent individuals at time t . How the risk set $Y_i(t)$ comes about can not be of any importance. In other words, the nature of the changes in the

risk set $Y_i(t)$ over time, whether it be part of a larger Markov process or not, is irrelevant. To put it in a slightly different way, what matters is that the process counting the transitions, $N_{ij}(t)$, has an intensity (or intensity process) $\alpha_{ij}(t) Y_i(t)$, where the intensity is now considered as a function of the past history. One advantage of this formulation is that it allows for almost arbitrarily complicated censoring mechanisms to operate.

When I came to Berkeley in 1973, to start my Ph.D. work, my hope was to develop a theory of the Nelson-Aalen estimator for this general setting. The trouble was that there did not exist in the published literature any proper theory for point processes (later counting processes) of the $N_{ij}(t)$ kind based on intensities defined as functions of the past history. The stationary type point process theory was clearly not useful. I therefore soon came to a dead end and decided to use a Markovian framework after all. However, I felt this very unsatisfactory since the Markov assumption was really unnecessary.

Lucien Le Cam became my supervisor at Berkeley. I had a number of discussions with him on various aspects of my thesis work. He was always very kind and forthcoming. Since he always had an open door one felt encouraged to disturb him at any time. I certainly took this liberty, and experienced that my often ill-formulated questions and ideas were answered in a serious and deeply thoughtful manner. Le Cam always took time to discuss, never seeking an easy way out with a superficial answer.

One of his suggestions stand out in view of the later developments. At one stage, when I was still struggling in the Markovian framework trying to develop some asymptotic theory for the Nelson-Aalen estimator, he mentioned that I ought to look at a paper by McLeish (1974) on central limit theorems for martingales that had just come out. Because surely, he said, there must be some martingales in what I was working on. At that stage I still could not quite see the relevance of McLeish's paper. However, somewhat later, I mentioned my search for a suitable theory of point processes to David Brillinger. He told me that he had recently received some papers concerning this from the Electronics Research Laboratory at Berkeley. These were the Ph.D. thesis by P. M. Brémaud (1972) and technical reports by R. Boel, P. Varaiya and E. Wong (later published, see references Boel, Varaiya & Wong 1975a and Boel, Varaiya & Wong 1975b) and it turned out to be exactly what was needed for my purpose. I was also given a copy of a technical report by Dolivo (1974) written at the College of Engineering, University of Michigan, which gave a particularly clear presentation. These papers gave the first proper mathematical theory for point processes, or counting processes, with intensities defined as functions of the past. The development was based on the theory of stochastic integrals for martingales. It was quite easy to see that the Nelson-Aalen estimator minus the cumulative intensity was essentially a stochastic integral and hence a martingale. The finite sample properties, like expectation, variance and the covariance structure of the process, which I had previously struggled

so much in deriving (Aalen 1976), fell out immediately. And now the suggestion of Le Cam to use McLeish's work came in very handy. On the basis of McLeish's paper, which concerned itself with time-discrete martingales, I developed a central limit theorem for continuous martingales which gave asymptotic theory for the Nelson-Aalen estimator.

It soon became apparent to me that this framework encompassed a lot more than the Nelson-Aalen estimator. Especially, two-sample tests for censored survival data could be shown to have natural representations as stochastic integrals, creating a unifying approach for a rather confusing field.

After finishing my Ph.D. thesis in Berkeley, I stayed in Copenhagen for some months. Here the newly developed theory caught the attention of Niels Keiding who later inspired the further developments in cooperation with Per Kragh Andersen, Ørnulf Borgan, Richard Gill and many others. One of the problems I brought with me to Berkeley found its solution in Copenhagen. I had developed an empirical transition matrix for Markov chains, generalizing the Kaplan-Meier estimator. After writing my thesis, I realized that the estimator had a simple martingale representation and should hence have a stochastic integral representation. However, I was unable to find this. In Copenhagen this problem was solved in a few minutes by Søren Johansen, who derived a matrix stochastic integral representation (see Aalen & Johansen 1978).

The recent book by Andersen et al. (1993) gives a very comprehensive review of the application of counting processes to statistics. Below I will take up some important subjects which are not yet well developed and require further research.

1.3 Dynamic modelling

The counting process formulation of survival and event history analysis is not merely a technical device opening up for the use of some powerful mathematical tools. It is also of considerable conceptual importance since it focusses the attention on the dynamic aspects. By this I mean that one is naturally led to a formulation based on how the past of the process influences the future. The intensity process of the counting process theory is just such a dynamic quantity, strongly distinguishing this approach from other ways of formulating point processes, e.g. stationary formulations and related ones. In defining the intensity process $\lambda(t)$ of a (multivariate) counting process one conditions with respect to past events and then asks what is the probability per time unit of observing another event. Statistical models based on this kind of thinking shall here be termed dynamic models.

The conditioning with respect to the past allows one to naturally incorporate events that have happened up to a certain time in the further analysis. One such event of great importance in event history analysis is

censoring, meaning that an individual is not observed beyond a certain time. Note that it is necessary to assume that the censoring event is really determined by the observed past, or by events that are independent of the process of interest. This means that the factors determining the censoring can really be considered part of (are measurable with respect to) a suitable sigma-algebra of past history. In mathematical terms the censoring process may be any predictable indicator process. In practice, however, censoring may sometimes be dependent on unobserved underlying processes that are related to the chance of survival. In this case, the standard methods of survival and event history analysis may yield biased results. This is essentially well known, but not presented very clearly in popular presentations of survival and event history analysis, and no doubt it is often ignored in practice.

Dynamic modelling has a far broader use, however. One may make dynamical statistical models based on the intensity process, that is, one may write it as a function of statistical parameters and observed stochastic processes. By estimating the parameters one will get a measure of the influence of the observed processes on the likelihood of the occurrence of the events of interest.

To be more concrete, consider a number of individuals, and let individual i have an intensity process $\lambda_i(t)$ of an event happening. This is related to the hazard rate $\alpha_i(t)$ through the multiplicative intensity model $\lambda_i(t) = \alpha_i(t)Y_i(t)$ where $Y_i(t)$ is an indicator, equal to one if individual i is still under observation and the event in question has not happened. If a vector of covariates $Z_i(t)$ is given at time t , then a general intensity based regression model can be formulated by

$$\alpha_i(t) = f(\gamma(t), \beta(t)'Z_i(t))$$

for a vector of regression functions $\beta(t)$.

Several parametrizations of the intensity process have been proposed. The most famous one is the Cox model where, assuming (usually) that the regression functions $\beta(t)$ are independent of t , one has $f(\gamma(t), \beta(t)'Z_i(t)) = \gamma(t)\exp(\beta'Z_i(t))$. Andersen & Gill (1982) were the first to extend counting process theory to this model.

Often the observed covariate processes are merely single measurements taken at time zero. More elaborate covariate processes with repeated measurements have also been used, but this is not so common and it appears somewhat unresolved how so-called time-dependent covariates are really handled in a good way (see pages 172 and 531 of Andersen et al. 1993). One difficulty may lie in the proportional hazards structure of the Cox model which, for instance, makes it difficult to connect the values of a regression parameter for a covariate when this is measured at two different times. As an example, consider high blood pressure as a risk factor of myocardial infarction, and assume the blood pressure is measured at time 0 and at a later time t_1 . After time t_1 it would be natural to use both blood

pressure measurements as covariates, but one might be interested in how the regression parameters in that case corresponds with what one might get had one continued using only the first measurement (i.e. if only this was available). If the model with both measurements are assumed to follow a true Cox model, then this problem corresponds to omitting a covariate. It is well known that the proportional hazards structure is not valid any longer when covariates are omitted, except for one special case related to the stable distribution (Hougaard 1986); this is the same phenomenon as in frailty models where proportionality is in general not preserved under mixing (see, for example, Aalen 1988). Also, omission of covariates may lead to bias (Bretagnolle & Huber-Carol 1988, Struthers & Kalbfleisch 1986). So, one cannot make consistent Cox models dependent on various amounts of available information. Of course, in practice one will often assume a pragmatic and relaxed view of the model and of these limitations, and nobody denies the great usefulness of the Cox model. Nevertheless, one should not ignore these conceptual issues.

Also, when considering time-dependent covariates one is interested in the marginal survival function after integrating out the distribution of the covariates. Deriving this marginal function is in general not tractable within the proportional hazards specification.

It is my impression that the Cox model has for a number of years had a great and positive influence on the statistical analysis of survival data, but that it may now, at least in medical statistics, have reached a point where it has become a conservative factor impeding further development. The reason for this may be precisely its great success. Proportional hazards has defined the terms in which one thinks about survival data, and issues not fitting easily into this framework may get suppressed.

The basic issue here is how to model in a useful way the relationship between the intensity process and the covariate processes $Z_i(t)$. In medical statistics the problem has recently been studied in the context of marker processes, that is processes indicating the development of a disease process, for example CD4 cell counts as a marker of HIV disease. Jewell & Kalbfleisch (1992) point out that the study of such relationships is made easier when the marker process is assumed to influence the intensity process in a linear way, that is, if one has a structure of the kind

$$f(\gamma(t), \beta(t)' Z_i(t)) = \gamma(t) + \beta(t)' Z_i(t). \quad (1)$$

Slightly different linear models for the marker process have been studied by Self & Pawitan (1992) and Tsiatis, Dafni, DeGruttola, Propert, Strawderman & Wulfson (1992). Regression models based on the linear parametrization in (1) have been studied by me (Aalen 1989, 1993), resulting in a simple graphical method.

A very detailed specification of a linear model was suggested several years ago by Woodbury and Manton (1977, 1983) and Myers (1981). The covari-

ate processes are modelled by linear stochastic differential equations, and go into equation (1) either directly or in a squared form (the model is still linear in the parameters). A nice mathematical structure is created, which also appears to be of considerable practical use, see Manton & Stallard (1988). This interesting approach seems to have been largely ignored in the medical statistical literature. Attempts like these at really modelling the covariate processes seem to be feasible only outside the proportional hazards framework. For pertinent comments on this, see Myers (1981, p.528).

The problem mentioned above about connecting regression coefficients for a covariate measured at different times, can be solved for suitable linear specifications. Consider a covariate measured at times 0 and t_1 yielding values C_0 and C_1 , and assume that the covariate C_0 is used up to time t_1 and that both covariates are used after this time. It can be shown that if one covariate, say C_0 is dropped from the specification after time t_1 , then the linear structure is preserved if the covariates have a multinormal distribution, and explicit formulas can be given for the new regression coefficients, see Aalen (1989). Note that the assumption of multinormality must here be made conditional on survival up to t_1 . However, with a linear structure with normal covariates the survivors at any time will also have normally distributed covariates, but with altered parameters.

In fact, a nice mathematical structure arises when the model is linear and the covariates are normal processes (or squares of such processes). In this case, analytical expressions may also be derived for marginal survival functions. The Woodbury and Manton model mentioned above is one general example of such a normal model.

The main point here is that dynamic modelling, that is, statistical modelling of the intensity process, gives one a number of possibilities that should be followed up without one being tied to a particular dominant type of model. However, much more work is certainly needed in order to develop good models for the relationship between stochastic covariate processes and the intensity process.

Next I will point out that dynamic modelling may be seen in a broader context than counting processes, and is really very closely connected with the problem of causality.

1.4 Causality

Although statistics, based as it is on counting and measuring, should be expected to give only a rather superficial understanding of biological and medical phenomena, it is still (surprisingly?) true that statistical analysis contributes to the understanding of causality in these fields (a classical example being smoking and lung cancer which was first derived as a statistical association and is now believed to be a causal relationship). A limited amount of theory concerning the relationship between statistics and causal-

ity exists. I will focus here on a particular view, namely the role of stochastic processes in this respect.

A causal understanding is naturally tied up with dynamic modelling. In both cases the focus is on how past events influence the occurrence of future events.

If there is a statistical association between two stochastic processes, say X and Y , then this association may be of two major types. The first possibility is that the two processes are (partly) reflecting the same phenomenon. This is not a causal relationship. A medical example might be the presence of congested sinuses and of eye irritation in allergy, which might both express an underlying allergic process.

The second type of association is where one process influences the likelihood of changes in the other process in a causal way. This may be a one-sided relationship with the influence going only in one direction, or it may be a two-sided relationship where both processes influence each other. In medicine it is usually assumed that cholesterol level, which can be considered a stochastic process, is causally influencing the development of heart disease in a mainly one-sided fashion. This is expressed in the belief that lowering the cholesterol level reduces the risk of heart disease. (The reality may be more complex, however, with high cholesterol level, as well as high blood pressure, being not only a cause but also a symptom of heart disease.)

Under some regularity conditions on the two processes they can be represented by Doob-Meyer decompositions (defined with respect to an increasing family of σ -algebras to which *both* processes are adapted).

$$X(t) = \int_0^t \lambda_x(s)ds + M_x(t), \quad Y(t) = \int_0^t \lambda_y(s)ds + M_y(t) \quad (2)$$

where the λ 's denote the local characteristics of the two processes and the M 's are martingales (their “differentials” are often called innovations since they represent the new and “surprising” changes in the process). In the case of a counting process the local characteristics are just the intensity processes. The concept of dynamic modelling really belongs to the general setting given here, counting processes being just one (particularly important) example.

One way of formalizing the first type of association discussed above, is in terms of the martingales, or innovations. This kind of association is assumed *not* to be present when the martingales are orthogonal, meaning that the product of the two processes is another martingale. If X and Y are counting processes this reduces to the assumption that the two processes never jump at the same time. If the two processes are diffusion processes driven by Wiener processes orthogonality corresponds to stochastic independence of the Wiener processes. In practical terms it means that the basic changes in the process, created by the innovations, are unrelated between the two

processes. Orthogonality is thus assumed to assure that the two processes represent different phenomena.

As suggested by Aalen (1987), it seems reasonable to assume orthogonality, before starting to talk about causal connections. An orthogonalization process may sometimes be applied to achieve this assumption.

For orthogonal processes, define $X(t)$ to be locally independent of Y at time t if $\lambda_x(s)$ is only a function of X up to time t (and possibly of some extraneous events), but *not* a function of Y (for a more formal definition, see Aalen 1987). If $X(t)$ is not locally independent of Y it is locally dependent. This concept was first introduced by Schweder (1970) for Markov chains. Similar ideas may be found in Mykland (1986) and also appear in the concept of Granger causality in time-series analysis.

Clearly, local dependence may be one-sided (only one process being dependent on the other) or two-sided. If there is only a one-sided local dependence over a time interval it is tempting to talk about a causal connection of one process on the other. A practical medical application is presented in Aalen, Borgan, Keiding & Thormann (1980).

A detailed study of this kind of statistical approach to causality, has been made by Arjas & Eerola (1993). It seems clear that statisticians ought to pay much more attention to causality, and that such a dynamic framework constitutes a very natural setting. Appropriate statistical models can be made and analyzed along the lines discussed in the previous section. In practice data may not be easily available, but one sees today increasingly that measurements are taken repeatedly over time so that one has in effect quite detailed observations of stochastic processes. This is especially the case in medical contexts.

1.5 Martingale residual processes

The attention so far has been on the predictable part of the Doob-Meyer decomposition (the local characteristic, intensity process) and statistical models for this. However, the innovation or martingale part is also a very useful quantity for statistical purposes.

When fitting dynamic statistical models to data, one tries to get as close a fit as possible between the predictable part and the observed data by some statistical principle. Consider a process X , which may well be a vector of several components, and its Doob-Meyer decomposition as given in equation (2). Let $\hat{\lambda}_x(s)$ be the estimated local characteristic. The goodness of fit may be judged by studying the remainder

$$\hat{M}_x(t) = X(t) - \int_0^t \hat{\lambda}_x(s)ds$$

This estimated innovation martingale is indeed a very useful tool. It has been used in the context of Cox models in different ways, see Fleming &

Harrington (1991) and Arjas (1988), and in linear models, see Aalen (1993). Goodness-of-fit tests based on this martingale have been studied by Hjort (see Section VI.3 of Andersen et al. 1993). The word "martingale residuals" has been coined, but certainly these residuals do not share the property of linear theory residuals of being (approximately) normally distributed, indeed their distribution can be very odd. They therefore have to be used in appropriate ways to be useful. It is argued by Aalen (1993) that plots should preferably be made based on aggregating individuals over several groups, instead of plotting individual residuals as done by Fleming & Harrington (1991).

Usually, the estimated innovation martingale will not itself be a martingale which is a difficulty in its use. However, the estimated innovation martingale *is* an exact martingale in one special case, namely when the statistical parameters are functions varying freely in time and the local characteristic is a linear function of these parameters, that is, for a model like the one in (1). A special case is the nonparametric linear regression model for counting processes (Aalen 1989, 1993). This yields another advantage for the linear model. The Hjort type goodness-of-fit test mentioned above will in this case have a simple limiting distribution as opposed to the complications arising in nonlinear models.

Considering the great research activity in survival analysis, it is surprising that definitive ways of checking goodness of fit by graphical procedures have not yet materialized, although some of the methods based on martingale residuals seem to point in the right direction. Certainly more work in this field is needed.

1.6 REFERENCES

- Aalen, O. O. (1976), 'Nonparametric inference in connection with multiple decrement models', *Scandinavian Journal of Statistics* **3**, 15–27.
- Aalen, O. O. (1987), 'Dynamic modelling and causality', *Scandinavian Actuarial Journal* pp. 177–190.
- Aalen, O. O. (1988), 'Heterogeneity in survival analysis', *Statistics in Medicine* **7**, 1121–1137.
- Aalen, O. O. (1989), 'A linear regression model for the analysis of life times', *Statistics in Medicine* **8**, 907–925.
- Aalen, O. O. (1993), 'Further results on the non-parametric linear regression model in survival analysis', *Statistics in Medicine* **12**, 1569–1588.
- Aalen, O. O. & Johansen, S. (1978), 'An empirical transition matrix for nonhomogeneous Markov chains based on censored observations', *Scandinavian Journal of Statistics* **5**, 141–150.

- Aalen, O. O., Borgan, Ø., Keiding, N. & Thormann, J. (1980), 'Interaction between life-history events: nonparametric analysis of prospective and retrospective data in the presence of censoring', *Scandinavian Journal of Statistics* **7**, 161–171.
- Andersen, P. K. & Gill, R. D. (1982), 'Cox's regression model for counting processes: A large sample study', *Annals of Statistics* **10**, 1100–1120.
- Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993), *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.
- Arjas, E. (1988), 'A graphical method for assessing goodness of fit in Cox's proportional hazards model', *Journal of the American Statistical Association* **83**, 204–212.
- Arjas, E. & Eerola, M. (1993), 'On predictive causality in longitudinal studies', *Journal of Statistical Planning and Inference* **34**, 361–386.
- Boel, R., Varaiya, P. & Wong, E. (1975a), 'Martingales on jump processes I: Representation results', *SIAM Journal of Control* **13**, 999–1021.
- Boel, R., Varaiya, P. & Wong, E. (1975b), 'Martingales on jump processes II: Applications', *SIAM Journal of Control* **13**, 1022–1061.
- Brémaud, P. (1972), A martingale approach to point processes, PhD thesis, Electrical Research Laboratory, University of California, Berkeley.
- Bretagnolle, J. & Huber-Carol, C. (1988), 'Effects of omitting covariates in Cox's model for survival data', *Scandinavian Journal of Statistics* **15**, 125–138.
- Dolivo, F. (1974), Counting processes and integrated conditional rates: a martingale approach with application to detection, PhD thesis, University of Michigan.
- Fleming, T. R. & Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, Wiley, New York.
- Hougaard, P. (1986), 'Survival models for heterogeneous populations derived from stable distributions', *Biometrika* **73**, 387–396.
- Jewell, N. P. & Kalbfleisch, J. D. (1992), Marker models in survival analysis and applications to issues associated with AIDS, in N. Jewell, K. Dietz & V. Farewell, eds, 'AIDS Epidemiology: Methodological Issues', Birkhäuser, Boston.
- Manton, K. G. & Stallard, E. (1988), *Chronic Disease Modelling*, Griffin & Company, New York.

- McLeish, D. L. (1974), 'Dependent central limit theorems and invariance principles', *Annals of Probability* **2**, 620–628.
- Myers, L. E. (1981), 'Survival functions induced by stochastic covariate processes', *Journal of Applied Probability* **18**, 523–529.
- Mykland, P. (1986), Statistical causality, Technical report, Department of Mathematics, University of Bergen, Norway.
- Schweder, T. (1970), 'Composable Markov processes', *Journal of Applied Probability* **7**, 400–410.
- Self, S. & Pawitan, Y. (1992), Modeling a marker of disease progression and onset of disease, in N. Jewell, K. Dietz & V. Farewell, eds, 'AIDS Epidemiology: Methodological Issues', Birkhäuser, Boston.
- Struthers, C. A. & Kalbfleisch, J. D. (1986), 'Misspecified proportional hazard models', *Biometrika* **73**, 363–369.
- Tsiatis, A. A., Dafni, U., DeGruttola, V., Propert, K. J., Strawderman, R. L. & Wulfson, M. (1992), The relationship of CD4 counts over time to survival in patients with AIDS: Is CD4 a good surrogate marker?, in N. Jewell, K. Dietz & V. Farewell, eds, 'AIDS Epidemiology: Methodological Issues', Birkhäuser, Boston.
- Woodbury, M. A. & Manton, K. G. (1977), 'A random walk model of human mortality and aging', *Theoretical Population Biology* **11**, 37–48.
- Woodbury, M. A. & Manton, K. G. (1983), 'A theoretical model of the physiological dynamics of circulatory disease in human populations', *Human Biology* **55**, 417–441.

2

Multivariate Symmetry Models

R. J. Beran¹
P. W. Millar²

2.1 Introduction

Let Γ_0 be a fixed, compact subgroup of the group Γ of orthogonal transformations on R^d . A random variable x , with values in R^d and distribution P , is Γ_0 -symmetric if x and γx have the same distribution for all $\gamma \in \Gamma_0$. In terms of P , this means $P(A) = P(\gamma A)$ for all Borel sets A and all $\gamma \in \Gamma_0$. Let x_1, \dots, x_n be iid random variables with values in R^d . The Γ_0 -symmetry model asserts that the x_i have an unknown common distribution that is Γ_0 -symmetric. The Γ_0 -location model specifies that for some unknown $\eta \in R^d$, the random variables $x_1 - \eta, \dots, x_n - \eta$ have an unknown common distribution P which is Γ_0 -symmetric. This paper develops some methods of inference for these multivariate symmetry models. Unlike the one dimensional case, there are a large number of “symmetry” notions in R^d , $d > 1$; Section 2.3 provides a few simple, useful examples, which figure in subsequent development.

Our methods of inference are based on the notion of “probabilities over half-spaces”, a concept explained in Section 2.2. Half-spaces in R^d are relevant for our inference problem because they form the simplest class of sets in R^d that both separates measures and remains unchanged by orthogonal transformations. Equally important is the fact that a particular half-space formulation capitalizes in a simple way on the underlying assumed “symmetry”, yielding easier analysis and an economic formulation of results. This inter-relation of the group structure and “half-space” concepts is described in Section 2.3.

Utilizing this basic framework, we develop two results on inference for these nonparametric multivariate symmetry models. The first generalizes a classical result. Given a Γ_0 -symmetry model, consider estimation of the underlying data distribution. For this estimation problem, the empirical measure will not be an efficient estimate under any non-trivial symmetry

¹University of California at Berkeley

²University of California at Berkeley

hypothesis. Section 2.4 explains how the empirical measure may be modified to yield an asymptotically efficient Γ_0 -symmetric estimator. The second result on inference concerns goodness of fit tests for the Γ_0 -location model. Let x_1, \dots, x_n be i.i.d. with common distribution G . The testing problem is to decide whether $G(\cdot) = P(\cdot + \eta)$ for some $\eta \in R^d$ and some Γ_0 -symmetric P .

In Sections 2.6 and 2.7 we devise a computationally feasible test of this null hypothesis, based on the half-space metric between distributions. Its asymptotic behavior is described in Sections 2.7 and 2.8. Critical values of this test can be obtained by special “conditional bootstrap” methods. These are explained in subsections 2.6.1–2.6.3. The test statistic itself, and the calculation of critical values, are computationally feasible, but require the availability of high speed computing facilities. Certain asymptotic optimality properties of these tests are implied by the optimality of associated confidence sets (Beran & Millar 1985).

Symmetries indexed by subgroups of the orthogonal group arise in several ways in multivariate analysis. In the context of classical multivariate analysis, Anderssen (1975), Perlman (1988) and others have studied “group symmetry covariance models” where the groups act on the covariance matrices of the multivariate normal distributions; see also Eaton (1983, Chapter 9). In spherical multivariate analysis, use of orthogonal transformations has a long history (Watson 1983); here the relevant distribution is not the normal, but the Langevin-von Mises-Fisher distribution. By capitalizing on assumed symmetries within these models, it has been possible to find, in an elegant way, precise expressions for the MLE, likelihood tests with high power, and so forth.

The development of this paper differs from that in the foregoing references in several ways. First, our approach is completely nonparametric. Likelihood concepts are replaced by methods based on empirical measures and notions of distance. Systematic use of half-spaces aids the group theoretic analysis, suggests a possible connection with current ideas in projection pursuit (Huber 1985, Donoho & Gasko 1987), and points the way to numerical implementation of some of the procedures (cf Section 2.6). Certain asymptotic optimality results are developed below; however, an important use of the asymptotics here is to give some theoretical grounding and justification for the computationally intensive procedures proposed. In particular, since it appears impossible in this non-classical framework to obtain useful closed form or asymptotic formulae for the procedures, we systematically employ iterated bootstrap methods to construct our procedures. (cf, Efron (1979), who pioneered the bootstrap concept; see also Beran (1984), Bickel & Freedman (1981), and Singh (1981) for early contributions to the idea).

Part of this paper develops a goodness of fit test for a very complicated null hypothesis: namely that the underlying distribution be given by a particular multivariate “location model”. The test statistic is based on a minimum distance method, whose computational feasibility is ensured in

part by a “local stochastic search” technique (Beran & Millar 1987, Millar 1993). Other studies of this particular idea applied to complicated null hypotheses include: (i) multivariate parametric models (Beran & Millar 1989) (ii) logistic models (Beran & Millar 1992) (iii) ellipsoidal symmetry models (Loranger 1989, but here the stochastic search was set at a finite number) (iv) censored data (Chow 1991).

2.2 Half-space and probabilities

The methods of inference in this paper are based on empirical measures indexed by half-spaces. Half-spaces, rather than some more conventional class of sets (such as lower left octants), were selected because they are the simplest class of sets in R^d that remains unchanged by orthogonal transformations and also determines measures on R^d : if $P(A) = Q(A)$ for every half-space A , then $P(A) = Q(A)$ for every Borel A . This section parametrizes half-spaces in a manner convenient for Γ_0 symmetry models.

Define the unit spherical shell of R^d by

$$S^d = \{s \in R^d : |s| = 1\}. \quad (1)$$

Here and throughout $|\cdot|$, $\langle \cdot, \cdot \rangle$ denote norm and inner product on the Euclidean space R^d . Define half-spaces $A(s, t)$ on R^d by

$$A(s, t) = \{x \in R^d : \langle s, x \rangle \leq t\} \quad (2)$$

where $s \in S^d$, $t \in R^1$. For notational convenience $\langle s, x \rangle$ will often be written $s'x$.

Any probability P on R^d can be identified with an element of $L_\infty(S^d \times R^1)$, the Banach space of real bounded functions on $S^d \times R^1$ with supremum norm: just define the map $(s, t) \rightarrow R^1$ by

$$P(s, t) \equiv P\{A(s, t)\}. \quad (3)$$

Let $X = (x_1, \dots, x_n)$ be a vector of iid random variables. The empirical measure \hat{P}_n puts mass n^{-1} at each point x_1, \dots, x_n . By (3), \hat{P}_n may be regarded as a random element of $L_\infty(S^d \times R^1)$. The *empirical process indexed by half-spaces* is then defined by

$$W_n(s, t) = n^{1/2} \left(\hat{P}_n(s, t) - P_n(s, t) \right), \quad (4)$$

assuming that P_n is the common distribution of x_1, \dots, x_n . The following result is known (see Beran & Millar 1986 and the references there).

Proposition 1 *Let P_n be a sequence of probabilities on R^d such that, for some P_0 ,*

$$\lim_{n \rightarrow \infty} \sup_{A, B} |P_n(A \cap B) - P_0(A \cap B)| = 0, \quad (5)$$

where the supremum is over all half-spaces A, B on R^d . Define W_n by (4). Then there is a mean zero Gaussian process

$$W = \{W(s, t) : (s, t) \in S^d \times R^1\}$$

such that

$$W_n \Rightarrow W, \quad (6)$$

convergence in $L_\infty(S^d \times R^1)$. If $P_0(\partial A) = 0$ for every half-space A , then W has continuous paths for the Euclidean topology of $S^d \times R^1$.

2.3 Symmetry models on R^d , $d \geq 2$

As noted in the introduction, there are many notions of “symmetric distribution” on R^d , $d \geq 2$. This section provides a development of the possibilities in terms of orthogonal subgroups and half-spaces.

Let

$$\Gamma = \{\text{all orthogonal transformations on } R^d\}. \quad (7)$$

Let Γ_0 be a compact subgroup of Γ . Define a random variable x to be Γ_0 -symmetric if

$$x \text{ and } \gamma x \text{ have the same distribution for all } \gamma \in \Gamma_0. \quad (8)$$

A probability P on R^d would then be Γ_0 -symmetric if (8) holds for any random variable having distribution P . Representation of P as an element of $L_\infty(S^d \times R^1)$, via half-spaces, (cf (3)) provides a neater formulation:

Proposition 2 *A probability P on R^d is Γ_0 symmetric if*

$$P(s, t) = P(\gamma s, t) \text{ for all } \gamma \in \Gamma_0. \quad (9)$$

Definition 1 Denote by $F(\Gamma_0)$ the collection of all probabilities on R^d that are Γ_0 -symmetric.

Since Γ_0 is a compact group, there is a “uniform probability distribution” on Γ_0 (Haar measure). Define

$$m_0 = \text{Haar measure on } \Gamma_0. \quad (10)$$

Recall that in compact groups, both left and right Haar measures are the same. Several examples of Γ_0 appear to be important for applications. In these cases, Γ_0 is actually commutative.

Example 1 (Simple symmetry) A R^d -valued rv x is *simply symmetric* if x and $-x$ have the same distribution. In this case the relevant Γ_0 has only two elements: the identity and multiplication by (-1) . Haar measure m_0 puts mass $1/2$ at each of these transformations. This is perhaps the

crudest useful notion of symmetry on R^d ; all other notions of symmetry in this paper reduce to this case on R^1 . Define the notation

$$\Gamma_S, m_S = \text{simple symmetry group and its Haar measure.} \quad (11)$$

It is clear that a probability P on R^d is Γ_S -symmetric iff

$$P(s, t) = P(-s, t) \text{ for all } (s, t) \in S^d \times R^1 \quad (12)$$

Example 2 (Isotropy, or spherical symmetry) Let

$$\Gamma_I = \{\text{all rotations in } R^d\}. \quad (13)$$

The random variables that are Γ_I -invariant are called isotropic. It is clear from (9) that

$$P \text{ is isotropic iff } P(s, t) \text{ does not depend on } s \quad (14)$$

Example 3 (Sign-change symmetry). A random variable $X = (x_1, \dots, x_d)$ in R^d is sign symmetric if X has the same distribution as $(\pm x_1, \dots, \pm x_n)$ for all choices of $+$, $-$. If P is the uniform distribution on the unit cube in R^d , centered at 0, with sides parallel to the axes, then P is sign symmetric but not isotropic.

For measures that are sign change invariant, the relevant group Γ_0 is discrete, containing 2^d points; each point may be represented as a sequence of length d , with $+$ or $-$ at each coordinate. Haar measure then puts mass 2^{-d} at each of these points. Define

$$\Gamma_P, m_P = \text{sign permutation group and its Haar measure.} \quad (15)$$

The following relationship holds for these examples:

$$\Gamma_S \subset \Gamma_P, \quad \Gamma_S \subset \Gamma_I \quad (16)$$

but Γ_P, Γ_I cannot be compared by set theoretic inclusion. On the other hand, (cf. Definition 1)

$$F(\Gamma_S) \supset F(\Gamma_P) \supset F(\Gamma_I) \quad (17)$$

Example 4 (Zonal symmetry) This notion of symmetry arises in certain geophysical problems. Let S^3 be the unit spherical shell in R^3 . The zonal symmetry group Γ_Z consists of all rotations γ in R^3 which leave the north pole-south pole axis of S^3 fixed. Thus each $\gamma \in \Gamma_Z$ corresponds to a rotation in R^2 ; with this identification, Haar measure m_Z for Γ_Z becomes the uniform distribution on the unit circle in R^2 .

2.4 Symmetrization of measures; efficiency

If Γ_0 is a compact subgroup of Γ , the group of orthogonal transformations on R^d , recall that $\mathbf{F}(\Gamma_0)$ is the collection of all probabilities on R^d that are Γ_0 symmetric. Let x_1, \dots, x_n be i.i.d. P , where $P \in \mathbf{F}(\Gamma_0)$. This section shows how to construct *efficient* estimators of P , where the estimator itself is Γ_0 -symmetric. This construction is also crucial to construction of the goodness of fit statistic in Section 2.6.

To begin, note first that, in general, the empirical measure \hat{P}_n obtained from x_1, \dots, x_n will *not* (a) belong to $\mathbf{F}(\Gamma_0)$ nor (b) be an efficient estimator of P . Therefore we develop here a simple improvement on \hat{P}_n ; this involves introducing a natural transformation that converts an ordinary probability on R^d into a “closest” Γ_0 -symmetric probability.

For fixed Γ_0 , define π_0 a mapping of $L_\infty(S^d \times R^1)$ to $L_\infty(S^d \times R^1)$ by

$$(\pi_0 f)(s, t) = \int f(\gamma s, t) m_0(d\gamma) \quad (18)$$

where m_0 is Haar measure for Γ_0 and $f \in L_\infty(S^d \times R^1)$. Of course π_0 depends on Γ_0 . If P is a probability on R^d , identified as an element of $L_\infty(S^d \times R^1)$ as described in (3), then its Γ_0 -symmetrization P_0 is defined by

$$P_0 = \pi_0 P \quad (19)$$

so that $P_0(s, t) = \int P(\gamma s, t) m_0(d\gamma)$. The following proposition asserts, among other things, that $\pi_0 P$ is actually a probability; explicit examples of $\pi_0 \hat{P}_n$ are collected below.

Proposition 3 *Let Γ_0 be a compact subgroup of Γ . then*

- (a) *π_0 is continuous, linear, $|\pi_0 f| \leq |f|$ for the $L_\infty(S^d \times R^1)$ norm, and $\pi_0(\pi_0 f) = \pi_0 f$*
- (b) *for each probability P on R^d , $\pi_0 P$ is a probability on R^d that is Γ_0 symmetric.*

Proof: Part (a) follows from

$$\begin{aligned} \sup_{s,t} |(\pi_0 f)(s, t)| &\leq \int \sup_{s,t} |f(\gamma s, t))| m_0(d\gamma) \\ &\leq \int \sup_{s,t} |f(s, t)| m_0(d\gamma) = |f|, \end{aligned}$$

since $\gamma \in \Gamma_0$, $s \in S^d$ implies $\gamma s \in S^d$, and since m_0 is a probability. In addition $\pi_0 f$ is “ Γ_0 symmetric” whenever $f \in L_\infty(S^d \times R^1)$, since $\gamma(\pi_0 f)(s, t) \equiv \pi_0 f(\gamma s, t) = \int f(\eta \gamma s, t) m_0(d\eta) = \int f(\eta s, t) m_0(d\eta) \equiv \pi_0 f(s, t)$, since m_0 is invariant.

To prove (b), define P^0 , a function on the Borel sets of \mathbb{R}^d by $P^0(C) = \int P\{\boldsymbol{x} : \gamma\boldsymbol{x} \in C\}m_0(d\gamma)$. Then P^0 is a probability, being a convex combination of such. On the other hand, as an element of $L_\infty(S^d \times \mathbb{R}^1)$, $P^0(s, t) = \pi_0 P(s, t)$ for all s, t , by (19), and this completes the proof, since half-spaces separate measures.

Example 5 Let \hat{P}_n be the empirical measure of x_1, \dots, x_n . Define π_S to be the projection obtained in (18) when $\Gamma_0 = \Gamma_S$, the “simple symmetry” group (Example 1). Then $\pi_S \hat{P}_n(\cdot)$ is the measure that puts mass $1/(2n)$ at each of the points $\{x_1, -x_1, x_2, -x_2, \dots, x_n, -x_n\}$.

Example 6 Let $\Gamma_0 = \Gamma_P$, the “sign permutation” group (Example 3) and π_P the corresponding projection in (18). If \hat{P}_n is the empirical of x_1, \dots, x_n , then $\pi_P \hat{P}_n$ assigns mass $n^{-1}2^{-d}$ to the points obtained by permuting the choice of sign on each of the d coordinates of the n data vectors.

Example 7 Let $\Gamma_0 = \Gamma_I$, the isotropy group (Example 2) and let π_I be the corresponding projection given in (18). Let \hat{P}_n be the empirical measure of x_1, \dots, x_n , with $x_i \in \mathbb{R}^d$, and let δ_a be unit mass distribution at the point $a \in \mathbb{R}^d$. Then $\pi_I \delta_a$ is uniform distribution on the shell $\{\boldsymbol{x} \in \mathbb{R}^d : |\boldsymbol{x}| = |a|\}$. By addition it is clear that then $\pi_I \hat{P}_n$ puts uniform mass of weight $1/n$ on each of the shells $\{\boldsymbol{x} \in \mathbb{R}^d : |\boldsymbol{x}| = |x_i|\}$, $1 \leq i \leq n$.

For the isotropic example, there is a more analytic expression for $\pi_I \hat{P}_n$. Let \mathbf{U}_r be the uniform distribution on the shell in \mathbb{R}^d of radius r , S_r^d . Let $q(\cdot; r)$ denote the probability measure on \mathbb{R}^d given by $q(C; r) = P\{\mathbf{U}_r \in C\}$. Then

$$\pi_I \hat{P}_n(C) = \frac{1}{n} \sum_{i=1}^n q(C; |x_i|).$$

If we choose C to be the halfspace $A(s, t)$, then since $\pi_I \hat{P}_n$ is isotropic, $\pi_I \hat{P}_n(s, t)$ does not depend on s . Indeed, by geometry, $q(A(s, t); r)$ already does not depend on s , and in fact is equal to the fraction of surface area on S_r^d cut out by the half-space $\{(y_1, \dots, y_d) \in \mathbb{R}^d : y_1 \leq t\}$. Thus $q(A(s, t); r) = q(A(\sigma, at); ar)$ for all $s, \sigma \in S^d$ and $a > 0$ (a helpful but simple scaling property); of course, precise formulas for $q(A(s, t); r)$ are available. These properties of q imply some remarkable smoothness properties for $\pi_I \hat{P}_n$ —for example, that its marginals have densities.

Example 8 Let Γ_Z be the zonal symmetry group on S^3 (Example 10), and let π_Z be the corresponding projection given in (18). Let \hat{P}_n be the empirical measure of x_1, \dots, x_n , where $x_i \in S^3$. Then $\pi_Z \hat{P}_n$ puts uniform mass $1/n$ on each of the zones Z_i on S^3 ; here the zone Z_i consists of the latitudinal circle on the surface of S^3 which runs through x_i , and whose plane is perpendicular to the north-south axis.

Having established a symmetrization method for \hat{P}_n , we now explain the asymptotic behavior of $\pi_0 \hat{P}_n$.

Proposition 4 Let P_n, P_0 satisfy the hypothesis of Proposition 1; let W be the process given there. Let Γ_0 be a compact subgroup of Γ , and suppose P_n, P_0 are Γ_0 symmetric. Then if \hat{P}_n is the empirical of iid observations from P_n :

$$(\pi_0 \hat{P}_n - P_n) \Rightarrow \pi_0 W,$$

convergence in $L_\infty(S^d \times R^1)$.

Proof: $\pi_0 \hat{P}_n - P_n = \pi_0 (\hat{P}_n - P_n)$, so the result follows from Proposition 1 and the continuous mapping theorem. In the isotropic case, it is easy to show (using the scaling property of q mentioned at the end of Example 7) that the class of functions $\{q(A; \cdot) : A = \text{half space}\}$ has V-C graphs; thus the result could in this case be deduced directly from standard CLT's for empirical processes (Pollard 1982).

We conclude this section with two results establishing $\pi_0 \hat{P}_n$ as an efficient estimator of a Γ_0 -symmetric distribution. To state a LAM result, let q be an increasing function on $[0, \infty)$, which for convenience is assumed bounded and continuous. Let P_0 be a fixed Γ_0 symmetric probability and let $N_n(c) = \{P : P \text{ is } \Gamma_0 \text{ symmetric and } \sup_{A,B} |P(A \cap B) - P_0(A \cap B)| \leq Cn^{-1/2}\}$ where the supremum is over all half spaces (cf., Proposition 1). Let x_1, \dots, x_n be iid P , where P is Γ_0 symmetric. Let D_n denote all estimates of P , based on x_1, \dots, x_n .

Proposition 5 (LAM efficiency) Under the conditions of the preceding paragraph

$$\begin{aligned} & \liminf_{\substack{c \uparrow \infty \\ \hat{Z}_n \in D_n}} \sup_{P \in N_n(c)} \int g(n^{1/2} |\hat{Z}_n - P|) dP^n \\ &= \lim_{n \rightarrow \infty} \sup_{P \in N_n(c)} \int g(n^{1/2} |\pi_0 \hat{P}_n - P|) dP^n \\ &= E_0 g(|\pi_0 W|). \end{aligned}$$

Here $|\cdot|$ denotes $L_\infty(S^d \times R^1)$ norm, and P^n the n -fold product measure of P .

The second efficiency result is a convolution theorem. For this, let \hat{Z}_n be a sequence of “regular” estimates of P_0 where \hat{Z}_n, P_0 are Γ_0 symmetric. See Millar (1985) for an appropriate definition of “regularity”.

Proposition 6 If \hat{Z}_n is a sequence of regular estimates of P_0 , and if $P_n \in N_n(c)$, then there is a random element L of $L_\infty(S^d \times R^1)$ such that

$$n^{1/2}(\hat{Z}_n - P_n) \Rightarrow (\pi_0 W) * L,$$

the $*$ denoting convolution.

By Proposition 4, $\pi_0 \hat{P}_n$ is efficient in the sense of this convolution theorem.

Proofs of Propositions 5, 6 are given in Section 2.10.

2.5 Symmetric shift models and identifiability

The Γ_0 -symmetric location model consists of probabilities Q on R^d such that

$$Q(A) = P(A - \eta), \quad (A \text{ a Borel set}) \quad (20)$$

where $P \in \mathbf{F}(\Gamma_0)$ (cf. Definition 1) and $\eta \in R^d$. This is an infinite dimensional parametric model with parameter θ given by

$$\theta = (\eta, P). \quad (21)$$

Here P is identified with an element of $L_\infty(S^d \times R^1)$, as described in (3). Therefore, the parameter points θ belong to certain subsets of $R^d \times L_\infty(S^d \times R^1)$. This latter is a normed linear space if, for $(\eta, \Delta) \in R^d \times L_\infty$, we define

$$|(\eta, \Delta)| = |\eta| \vee |\Delta|. \quad (22)$$

We shall provide a goodness of fit test for the Γ_0 location models in Section 2.6. The present section provides simple criteria for a model parametrized in this manner to have “identifiable” parameters. There are two useful notions of identifiability. To describe them, let Θ be a parameter set which is also a topological space, and let $\{Q_\theta, \theta \in \Theta\}$ be a family of probabilities on some metric space.

Definition 2

- (a) *The family $\{Q_\theta, \theta \in \Theta\}$ is strongly identifiable if, whenever $Q_{\theta_n} \Rightarrow Q_{\theta_0}$, with $\theta_n, \theta_0 \in \Theta$, then $\theta_n \rightarrow \theta_0$ in the topology of Θ .*
- (b) *The family $\{Q_\theta, \theta \in \Theta\}$ is (simply) identifiable if, whenever $Q_\theta = Q_{\theta'}$, then $\theta = \theta'$.*

Remark Definition (b) is the usual definition of identifiability. Definition (a) is important for certain considerations involving minimum distance methods (cf., Pollard (1980), for example). In general, a Γ_0 location model will *not* be identifiable unless conditions are placed on Γ_0 . For example, the transformations on R^2 that merely change the sign of the first coordinate do not lead to an identifiable symmetric location model.

Proposition 7 *Assume $\Gamma_0 \supset \Gamma_S$.*

- (i) *Let Θ consist of points (η, P) , $\eta \in R^d$, $P \in \mathbf{F}(\Gamma_0)$, so that $\Theta \subset R^d \times L_\infty(S^d \times R^1)$. If $\theta = (\eta, P)$, let $Q_\theta(A) = P(A - \eta)$. Then the model $\{Q_\theta, \theta \in \Theta\}$ is identifiable.*
- (ii) *Let Θ_c consists of points $\theta = (\eta, P)$, $\eta \in R^d$, $P \in \mathbf{F}(\Gamma_0)$, where $P(\delta A) = 0$ for every half space A . If Q_θ is defined as in (i), then $\{Q_\theta, \theta \in \Theta_c\}$ is strongly identifiable.*

Proof: Part (i) is easy so we consider (ii) only.

Suppose there exist $\theta_n = (\eta_n, P_n)$, $P_n \in \mathbf{F}(\Gamma_0)$, $\eta_n \in R^d$ and $\theta_0 = (\eta_0, P_0)$ such that

$$|\theta_n - \theta_0| \geq c \text{ for some } c > 0 \quad (23)$$

but yet

$$|P_n(\cdot - \eta_n) - P_0(\cdot - \eta_0)| \rightarrow 0. \quad (24)$$

Condition (23) means

$$|\eta_n - \eta_0| \vee |P_n(\cdot) - P_0(\cdot)| \geq c. \quad (25)$$

By elementary properties of the supremum norm, we may assume $\eta_0 = 0$. Let Z_n have distribution P_n , and Z_0 distribution P_0 so that by (24)

$$Z_n + \eta_n \Rightarrow Z_0. \quad (26)$$

Let us show first that $\{\eta_n\}$ is bounded. Let Z_{in}, η_{in} be the i th coordinate of Z_n, η_n , $1 \leq i \leq d$, and $F_{i,n}, F_{i0}$ the cdf's of Z_{in}, Z_{i0} . Then $Z_{in} + \eta_{in} \Rightarrow Z_{i0}$ so by simple symmetry $1/2 = F_{in}(\eta_{in} - \eta_{in}) = F_{i0}(\eta_{in})$, which implies $\{\eta_{in}, n \geq 1\}$ is bounded for each i . Thus $\{\eta_n\}$ is bounded.

It then follows that $\{Z_n\}$ is tight, since $\{Z_n + \eta_n\}$ is tight and $\{\eta_n\}$ is bounded. Let $\{Z_{n'}, \eta_{n'}\}$ be convergent subsequences such that

$$Z_{n'} \Rightarrow Y \quad (27)$$

$$\eta_{n'} \rightarrow \eta_\infty. \quad (28)$$

Since Z_n is simply symmetric, so is Y . By (26), (27)

$$Z_0 \xrightarrow{D} \eta_\infty + Y. \quad (29)$$

The symmetry of both Z_0, Y forces $\eta_\infty = 0$. Hence, $\eta_n \rightarrow 0 = \eta_0$ and $Z_n \Rightarrow Z_0$. Thus, for every half space A which is a continuity set for P_0 , $\lim_n P_n(A) = P_0(A)$. If P_0 satisfies $P_0(\partial A) = 0$ for every A , then in fact it can be shown that $\sup_A |P_n(A) - P_0(A)| \rightarrow 0$, which contradicts (25) and establishes part (b) of the proposition. The uniform convergence just mentioned is a generalization to the half-space framework of a classic result on R^1 (Chung 1968, p. 124; Bickel & Millar 1992, and references there).

Definition 3 Let $\theta = (\eta, P)$ $\eta \in R^d$, $P \in \mathbf{F}(\Gamma_0)$. For the rest of this paper, let Q_θ denote the measure on R^d given by

$$Q_\theta(A) = P(A - \eta). \quad (30)$$

Of course Q_θ may be regarded as an element of $L_\infty(S^d \times R^1)$ as in (3). The following relationship will be used repeatedly:

$$Q_\theta(s, t) = P(s, t - \langle s, \eta \rangle) \text{ if } \theta = (\eta, P). \quad (31)$$

2.6 A stochastic test for symmetry models

With the preparations of Sections 2.2–2.5 behind us, we can now describe precisely the goodness of fit test suggested in the Introduction.

Fix Γ_0 , a compact subgroup of Γ . Assume for the remainder of the paper that (cf., Section 2.3)

$$\Gamma_0 \supset \Gamma_S. \quad (32)$$

Let x_1, \dots, x_n be iid. we wish to test, on the basis of this data, whether the common distribution of the x_i 's is of the form Q_θ , where $\theta = (\eta, P)$, $\eta \in R^d$, $P \in \mathbf{F}(\Gamma_0)$ and (see (30)) $Q_\theta(A) \equiv P(A - \eta)$. A plausible test statistic is

$$\inf_{\theta \in \Theta} n^{1/2} |\hat{P}_n - Q_\theta| \quad (33)$$

where $\Theta = \{(\eta, P) : \eta \in R^d, P \in \mathbf{F}(\Gamma_0)\}$. The norm in (33) is that of $L_\infty(S^d \times R^1)$. This statistic, while intuitively attractive, is computationally intractable as it stands: the infimum over the infinite dimensional set Θ is particularly troublesome. In this section we develop a computationally feasible variant of (33) wherein

- (a) the $L_\infty(S^d \times R^1)$ norm is replaced by a maximum over a finite number of randomly chosen points (s_i, t) .
- (b) the infimum is replaced by the minimum over a finite number of randomly chosen points $\theta_i = (\eta_i, P_i)$.

For convenience, we call the maximization in (a) a “stochastic norm”, and the minimization in (b) a “stochastic search of Θ ”. Because Θ is infinite dimensional, some care must be taken in the development of (a), (b); see Millar (1988), for general discussion of some of the difficulties. See Beran & Millar (1987), for development in a finite dimensional situation of the notion of “stochastic procedure”, of which the procedure below is an example. The stochastic test statistics in this paper appear to be the first to be carried out in a genuinely nonparametric situation, with both stochastic search and stochastic norm.

2.6.1 STOCHASTIC NORM

For $k_n > 0$, let

$$s_1, s_2, \dots, s_{k_n} \text{ be iid uniform random variables on } S^d. \quad (34)$$

Assume that $\{s_i\}$ is independent of the data $\{x_i\}$. The stochastic norm $|\cdot|_n$ on $L_\infty(S^d \times R^1)$ is defined by

$$|f|_n = \max_{i \leq k_n} \sup_t |f(s_i; t)|, \quad f \in L_\infty(S^d \times R^1) \quad (35)$$

2.6.2 STOCHASTIC SEARCH

Here, for reasons explained at length in the references just cited, we choose the random θ 's in (33b) to be “bootstrap replicas” of a preliminary $n^{1/2}$ -consistent estimator of the parameter $\theta = (\eta, P)$; by such a means we can side step the difficulties of a search over an infinite dimensional parameter set. To give an explicit procedure at this time, let us adopt as our estimate of η a “trimmed mean”, and as our estimate for P a “ Γ_0 symmetrized” version of the empirical measure centered at the estimate of η . These estimates will now be described.

Let $X = (x_1, \dots, x_n)$ be iid Q , and let

$$\hat{Q}_n(\cdot) = \hat{Q}_n(X; \cdot) \quad (36)$$

be the empirical measure. For $0 < \alpha < 1/2$, let

$$\hat{\eta}_n(X) \equiv \alpha\text{-trimmed mean of } \hat{Q}_n. \quad (37)$$

Since there are many concepts of trimmed mean in R^d , let us for convenience take the one that is the random vector consisting of the α trimmed means of \hat{Q}_{in} , $1 \leq i \leq d$, where \hat{Q}_{in} is the empirical of the i th coordinates of the data vectors $x_j \in R^d$, $1 \leq j \leq n$. This gives a \sqrt{n} consistent estimator of η in the parameter $\theta = (\eta, P)$, $P \in \mathcal{F}(\Gamma_0)$.

To deal with the nonparametric part of $\theta = (\eta, P)$, it is convenient to introduce centering operators. Define, for $\eta \in R^d$, the mapping $\tau_\eta : L_\infty(S^d \times R') \rightarrow L_\infty(S^d \times R')$ by

$$(\tau_\eta f)(s, t) = f(s, t - \langle s, \eta \rangle). \quad (38)$$

Then (31) shows that, if $\theta = (\eta, P)$

$$\begin{aligned} Q_\theta &= \tau_{-\eta} \circ P \text{ and} \\ P &= \tau_\eta \circ Q_\theta. \end{aligned} \quad (39)$$

Thus τ_η applied to any measure Q “centres” Q at η . As a linear operator, it is clear that $|\tau_\eta f| = |f|$ and that $\tau_{\eta+\eta_1} = \tau_\eta \circ \tau_{\eta_1}$, so $\{\tau_\eta : \eta \in R^d\}$ is a group. The estimate of P in (η, P) is then given by

$$P_0^* \equiv \pi_0 \circ \tau_{\hat{\eta}_n} \circ \hat{Q}_n \quad (40)$$

where $\hat{\eta}_n$ is given by (36) and where π_0 is the symmetrizing operation of Γ_0 . For $\theta = (\eta, P)$, the proposed estimate for Q_θ is then $\tau_{-\hat{\eta}_n} \circ P_0^*$ which is easily evaluated as

$$\int_{\Gamma_0} \hat{Q}_n(\gamma s, t + \langle \gamma s - s, \hat{\eta}_n \rangle) m_0(d\gamma). \quad (41)$$

To construct the stochastic search, draw bootstrap samples $X_1^*, \dots, X_{j_n}^*$, each of size n , from the fitted distribution. Set $Q_i^* = \text{empirical of } X_i^*$, and

$$\eta_i^* = \hat{\eta}_n(X_i^*), \quad 1 \leq i \leq j_n; \quad \eta_0^* \equiv \hat{\eta}_n \quad (42)$$

$$P_i^* = \pi_0 \circ \tau_{-\hat{\eta}_n} \circ P_0^*, \quad 1 \leq i \leq j_n \quad (42)$$

$$\theta_i^* = (\eta_i^*, P_i^*), \quad 0 \leq i \leq j_n. \quad (43)$$

Note that the dependence on n has been suppressed in this notation. The search set Θ_n is then defined by

$$\Theta_n = \{\theta_0^*, \theta_1^*, \dots, \theta_{j_n}^*\}, \quad (44)$$

and the stochastic test statistic is

$$\min_{0 \leq i \leq j_n} n^{1/2} |\hat{Q}_n - Q_{\theta_i^*}|_n. \quad (45)$$

Example 9 If $\Gamma_0 = \Gamma_S$ (cf. Section 2.3) then it is easy to see from (41) that $Q_{\theta_i^*}$ is the empirical measure of $x_{1i}^*, x_{2i}^*, \dots, x_{ni}^*$, $2\eta_i^* - x_{1i}^*, \dots, 2\eta_i^* - x_{ni}^*$, the symmetrization familiar from R^1 . Here $X_i^* = (x_{1i}^*, x_{2i}^*, \dots, x_{ni}^*)$.

Remark The evaluation of $Q_{\theta_i^*}$, as exhibited in (41), is easy for $\Gamma_0 = \Gamma_S$, but may be difficult computationally for other choices of Γ_0 . For such cases, good approximations are desired; one possibility would be a Monte Carlo evaluation based on iid uniform Γ_0 random variables.

2.6.3 ESTIMATES OF CRITICAL VALUES

Computable, and asymptotically valid, critical values for the test statistic (45) can be obtained by bootstrap methods. Here are some possibilities.

- (a) **A conditional bootstrap** Given the statistic (45) fix the θ_i^* , $0 \leq i \leq j_n$ forming the search set Θ_n and fix the s_i determining the stochastic norm $|\cdot|_n$. Conditionally independently (given X_i) of $\{\theta_i^*\}$ and $\{s_i\}$, draw i_n bootstrap samples $X_u^{**} = (x_{1u}^{**}, \dots, x_{nu}^{**})$, $1 \leq u \leq i_n$, from the fitted distribution. Let \hat{Q}_u^{**} be the empirical of the u^{th} such bootstrap sample. Let D_n^{**} denote the empirical cdf of $\{\min_{\theta \in \Theta_n} n^{1/2} |\hat{Q}_u^{**} - Q_\theta|_n, 1 \leq u \leq i_n\}$. Then asymptotically valid critical values are given by the quantiles of D_n^{**} , if $i_n \rightarrow \infty$, by Theorem 2.
- (b) **Partially conditional bootstrap** For this method, fix $\{s_i\}$ which determine the stochastic norm. Independently of $\{s_i\}$, draw i_n samples X_n^{**} , $1 \leq u \leq i_n$ as in (a). From each of these, compute afresh a new search set Θ_n^{**} , of cardinality $j_n + 1$, as described earlier in this section. Denote the empirical cdf of $\{\min_{\theta \in \Theta_n^{**}} n^{1/2} |\hat{Q}_u^{**} - Q_\theta|_n, 1 \leq u \leq i_n\}$ by D_n^{**} . As before, the quantiles of D_n^{**} provide asymptotically valid critical values. This method provides extra probing of

the parameter set Θ , and, since Θ is infinite dimensional, this added effort might give comfort.

- (c) **Unconditional bootstrap** For this method, one draws additional bootstrap samples X_u^{**} , $1 \leq u \leq i_n$, as in the previous methods, and replicates afresh the entire statistic (41), including not only Θ_n but also construction of new s_i for the stochastic norm. Again the empirical of these replicated test statistics provides asymptotically valid critical values.

It is clear that computational complexity increases dramatically from method (a) to methods (b), (c). Whether or not there is any compensating additional asymptotic accuracy is an interesting open problem: the methods are equally accurate to first order approximation.

2.7 Location functionals and $n^{1/2}$ consistency

The trimmed means used in Section 2.6 to construct the stochastic search of Θ can be replaced by other, perhaps better, statistics. This section describes a more general approach, leading to two developments. First we develop the general notion of asymptotically smooth location functionals. These can be used, as were trimmed means in Section 2.6, to construct location estimates with a tractable asymptotic theory. Second, we give results concerning the joint asymptotic behavior of these location estimates $\hat{\eta}_n$ and the estimate of the underlying Γ_0 -symmetric distribution given in Section 2.6 as $\pi_0 \tau_{\hat{\eta}} \hat{Q}_n$. These asymptotic results culminate in Corollary 1 for the centred estimates, and in Corollary 2 for their bootstrap replicas. For the rest of this section, fix Γ_0 .

Definition 4 A general location functional η is a map from probabilities to R^d with the following properties

- (a) *Broad domain:* $\eta(\cdot)$ is well defined on all probabilities P on R^d
- (b) *Location property:* if $P \in \mathbf{F}(\Gamma_S)$ and $Q(\cdot) = P(\cdot - \eta)$, $\eta \in R^d$, then $\eta(Q) = \eta$
- (c) *Continuity:* if $Q_n \Rightarrow Q_0$ $Q_0(\cdot) = P_0(\cdot - \eta_0)$, $P_0 \in \mathbf{F}(\Gamma_0)$, $P(\delta A) = 0$ for all half spaces A , then $\eta(Q_n) \rightarrow \eta_0 \equiv \eta(Q_0)$.

Let x_1, \dots, x_n be iid, with common distribution Q_n and let \hat{Q}_n be their empirical measure. Define the statistic $\hat{\eta}_n$ by

$$\hat{\eta}_n = \eta(\hat{Q}_n). \quad (46)$$

We require for the estimator $\hat{\eta}_n$ a certain “asymptotic linearity”. To state this, define a norm $|\cdot|_2$ on probabilities by

$$|P - Q|_2 = \sup_{A, B} |P(A \cap B) - Q(A \cap B)| \quad (47)$$

where the supremum is over all half-spaces A, B . Let ψ_n, ψ be functions from R^d to R^d . Assume

(A1) (Asymptotic linearity) if $|Q_n - Q_0|_2 \leq cn^{-1/2}$, then $n^{1/2}(\hat{\eta}_n - \eta_n) = n^{-1/2} \sum \Psi_n(x_i - \eta_n) + o_{Q_n}(1)$ where $\eta_n = \eta(Q_n)$, $\psi_n \rightarrow \psi$ uniformly, ψ_n is bounded and continuous, $Q_0(\cdot) = P_0(\cdot - \eta_0)$, $P_0 \in \mathbf{F}(\Gamma_0)$, ψ depends on P_0 only, and $E_{Q_n} \psi_n(x_i - \eta_n) = 0$.

If $\Gamma_0 \supset \Gamma_S$, it seems clear that an additional property should hold:

$$\psi(x) = -\psi(-x) \text{ for } x \in R^d \quad (48)$$

if ψ corresponds to $P_0 \in \mathbf{F}(\Gamma_0)$. If x_1, \dots, x_n are iid Q_n , with empirical \hat{Q}_n , define $W_n = n^{1/2}(\hat{Q}_n - Q_n)$

Proposition 8 Assume Definition 4, 47, (A1), and that $|Q_n - Q_0|_2 \leq cn^{-1/2}$. Then $(n^{1/2}(\hat{\eta}_n - \eta_n), W_n)$ converges weakly in $R^d \times L_\infty(S^d \times R^1)$ to (Y, W) , where Y is normal with mean 0, covariance Γ_ψ and W is Gaussian mean 0 with covariance $Q_0(A \cap B) - Q_0(A)Q(B)$, A, B half spaces.

Here Γ_ψ is the matrix whose entries are $E_{Q_0} \psi_i(x_1) \psi_j(x_1)$, for $\psi = (\psi_1, \dots, \psi_d)$.

Proof: The sequence $n^{1/2}(\hat{\eta}_n - \eta_n)$ converges to Y in R^d by asymptotic linearity; W_n converges in L_∞ by the Proposition 1 in Section 2.2. Hence the vector $(n^{1/2}(\hat{\eta}_n - \eta_n), W_n)$ is tight. The finite dimensional distributions look like:

$$\left(\frac{\sum \psi(x_i - \eta_n)}{\sqrt{n}}, \frac{\sum \xi_1(x_i)}{\sqrt{n}}, \dots, \frac{\sum \xi_k(x_i)}{\sqrt{n}} \right) + o_{Q_n}(1)$$

where $\xi_i(x) = I\{s'_i x \leq t_i\} - Q_n\{x : s'_i x \leq t_i\}$, for $(s_1, t_1), \dots, (s_k, t_k)$ in $S^d \times R^1$. These fdd obviously converge by the ordinary CLT on R^{k+1} , completing the proof.

With more hypotheses, we can establish convergence of natural estimates of the parameters $\theta = (\eta, P)$. Let x_1, \dots, x_n be again iid Q_n . Assume

$$(Q_n - Q_0)n^{1/2} \rightarrow \zeta_0 \in L_\infty(S^d \times R^1) \text{ as } n \rightarrow \infty \quad (49)$$

and ζ_0 is uniformly continuous, in that

$$\sup_{s, t} |\zeta_0(s, t + \langle \eta_n, s \rangle) - \zeta_0(s, t + \langle \eta_0, s \rangle)| \rightarrow 0 \quad \text{as } \eta_n \rightarrow \eta_0,$$

where $Q_0(\cdot) = P_0(\cdot - \eta_0)$, $P_0 \in \mathbf{F}(\Gamma_0)$, $\eta_0 \in R^d$.

Assume also that

- (A2) the probability measure P_0 , specified in (49), satisfies: for each s , $P_0(s, t) = \int^t g_0(s, u)du$ where g_0 is bounded, and uniformly continuous in (s, t) , and $\sup_s |g(s, t_1) - g(s, t_2)| \leq C$

Let \hat{Q}_n again be the empirical of x_1, \dots, x_n ; and set $\hat{\eta}_n = \eta(\hat{Q}_n)$, where η satisfies the conditions of a location functional. If $Q_0 = Q_{\theta_0}$, $\theta_0 = (\eta_0, P_0)$, $P_0 \in F(\Gamma_0)$, then natural estimates of (η, P) are $\hat{\eta}_n = \eta(\hat{Q}_n)$ and $P_0^* \equiv \pi_0 \tau_{\hat{\eta}_n} \hat{Q}_n$ (see (38) ff).

Theorem 1 ($n^{1/2}$ consistency). Assume Definition 4, (A1), (47), (49), (A2). Set $\eta_n = \eta(Q_n)$, and let $P_n^0 = Q_n(\cdot + \eta_n)$. Then in $L_\infty(S^d \times R^1)$

$$n^{1/2}[P_0^* - \pi_0 P_n^0](s, t) \Rightarrow (\pi_0 \tau_{\eta_0} W)(s, t) + \int_{\Gamma} \langle \gamma s, Y \rangle g(\gamma s, t) m(d\gamma).$$

Here Y, W are given in Proposition 8.

Proof: By the development of (38) ff, the theorem asserts the convergence of

$$n^{1/2}[\pi_0 \tau_{\hat{\eta}_n} \hat{Q}_n - \pi_0 \tau_{\eta_0} Q_n]. \quad (50)$$

To analyze this, write it as the sum of

$$n^{1/2}[\pi_0 \tau_{\hat{\eta}_n} \hat{Q}_n - \pi_0 \tau_{\hat{\eta}_n} Q_n] \quad (51)$$

and

$$n^{1/2}[\pi_0 \tau_{\hat{\eta}_n} Q_n - \pi_0 \tau_{\eta_0} Q_n]. \quad (52)$$

To analyze the term in (51), let η'_n be any sequence, $\eta'_n \rightarrow \eta_0$. Then

$$n^{1/2} \left(\tau_{\eta'_n} \hat{Q}_n - \tau_{\eta'_n} Q_n \right) \Rightarrow \tau_{\eta_0} W \quad (53)$$

because the observations $x_1 - \eta'_n, \dots, x_n - \eta'_n$ satisfy the requirements of Proposition 1; that is, (53) is immediate from the triangular array version of the usual CLT for empirical processes, because of hypothesis (A1). By the substitution theorem of Beran & Millar (1987), it then follows that

$$n^{1/2} \left(\tau_{\hat{\eta}_n} \hat{Q}_n - \tau_{\hat{\eta}_n} Q_n \right) \Rightarrow \tau_{\eta_0} W \quad (54)$$

That is, (53) holds with a random substitution for η'_n , because $\hat{\eta}_n \rightarrow \eta_0$. Since π_0 is continuous, the continuous mapping theorem implies that the expression in (51) converges to

$$\pi_0 \tau_{\eta_0} W. \quad (55)$$

To analyze, (52), let $\{\eta'_n\}$ be any sequence such that $(\eta'_n - \eta_n)n^{1/2} \rightarrow \lambda$. Then

$$\begin{aligned} & n^{1/2}\{\tau_{\eta'_n}Q_n(s, t) - \tau_{\eta_n}Q_n(s, t)\} \\ &= n^{1/2}[Q_n(s, t + \langle s, \eta'_n \rangle) - Q_0(s, t + \langle s, \eta'_n \rangle)] \\ &\quad - n^{1/2}(Q_n(s, t) + \langle s, \eta_n \rangle) - Q_0(s, t + \langle s, \eta_n \rangle)) \\ &\quad + n^{1/2}(Q_0(s, t + \langle s, \eta'_n \rangle) - Q_0(s, t + \langle s, \eta_n \rangle)). \end{aligned} \quad (56)$$

By (49) and the uniform continuity of $Q_0(s, t)$, the first two bracketed terms in (56) add, as $n \rightarrow \infty$, to $\zeta(s, t + \langle s, \eta \rangle) - \zeta(s, t + \langle s, \eta_0 \rangle) = 0$. By (A2) the remaining term is, in the limit,

$$-\lim_n n^{1/2}\langle s, (\eta'_n - \eta_n) \rangle g(s, t) = -\lambda g(s, t) \quad (57)$$

By the substitution theorem in Beran & Millar (1987), we may replace η'_n with the random $\hat{\eta}_n$, to find from (57) that the term in (52) converges, as $n \rightarrow \infty$ to

$$-\langle s, Y \rangle g(s, t + \langle s, \eta_0 \rangle). \quad (58)$$

Adding (58) and (55) gives the final result.

Corollary 1 (a) Assume the hypotheses of Theorem 1. Then $(n^{1/2}(\hat{\eta}_n - \eta_n), n^{1/2}(P_0^* - \pi_0 P_n^0))$ converges weakly in $R^d \times L_\infty$ to a mean 0 gaussian element $(Y, Z) \in R^d \times L_\infty$, where Z is given by the limit in Theorem 1, and Y by Proposition 8.

(b) If (48) holds, then Y, Z are independent.

Proof: Part (a) follows by the argument given in Proposition 8. Given the convergence of Theorem 1, part (b) follows from the fact that the limit distribution in (a) is Gaussian and the fact that $n^{1/2}(\hat{\eta}_n - \eta_n)$ and $n^{1/2}\pi_0(P_0^* - P_n^0)(s, t)$ are (for each s, t) asymptotically uncorrelated because of (48) and the hypothesis that $\Gamma_0 \supset \Gamma_S$.

The corollary below extends Theorem 1 to the bootstrap replicas $\pi_0 \hat{P}_i^*$ of Section 2.6; the result is used to establish the results of Section 2.8.

Corollary 2 Let x_1, \dots, x_n be iid Q_0 , $Q_0(\cdot) = P_0(\cdot - \eta_0)$, $\eta_0 \in R^d$, $P_0 \in \mathcal{F}(\Gamma_0)$. Assume the hypotheses of Theorem 1 with $Q_n \equiv Q_0$ there. Let $\pi_0 \hat{P}_1^*$ be defined by (42). Then $n^{1/2}(\pi_0 \hat{P}_1^* - P_0)(s, t) \Rightarrow$

$$\pi_0 \tau_{\eta_0}(W' + W'')(s, t) - \langle s, Y' + Y'' \rangle g(s, t + \langle s, \eta_0 \rangle)$$

where (Y', W') , (Y'', W'') are iid copies of (Y, W) given in Proposition 8.

Remark A triangular array variant of this result is possible, but we omit this familiar development; the corollary as stated will justify the bootstrap calculations.

Proof: Write $n^{1/2}(\hat{P}_1^* - P_0)$ as $n^{1/2}(\hat{P}_1^* - \hat{P}_n^0) + n^{1/2}(\hat{P}_n^0 - P_0)$. Then Theorem 1 applies so that $n^{1/2}\pi_0(\hat{P}_1^* - \hat{P}_n^0) \Rightarrow \pi_0\tau_0 W'' - \langle \cdot, Y'' \rangle g(\cdot; \cdot + \langle \cdot, \eta_0 \rangle)$ and a similar limit holds for $n^{1/2}(\hat{P}_n^0 - P_0)$, with (W', Y') replacing (Y'', W'') . Independence of (W', Y') , (W'', Y'') is immediate from the bootstrap construction.

2.8 Asymptotics for the stochastic test

In this section we give the asymptotic form of the stochastic test statistic described in Section 2.6, except that we assume that it is derived from a general location functional, (instead of a trimmed mean): this merely means that the $\hat{\eta}$ of Section 2.6 is to be replaced in the recipe of Section 2.6 by the $\hat{\eta}$ of Section 2.7. This asymptotic result is used to justify bootstrap calculations of critical values for the test.

As in previous sections let $\theta = (\eta, P)$ where $P \in \mathcal{F}(\Gamma_0)$ and $\Gamma_0 \supset \Gamma_S$; $Q_\theta(\cdot) = P(\cdot + \eta)$. Fix $\theta_0 = (\eta_0, P_0)$. Let $\{Q_n\}$ be a sequence of probabilities on R^d such that (cf (A1))

$$|Q_n - Q_{\theta_0}|_2 \leq cn^{-1/2} \quad (59)$$

$$|\eta(Q_n) - \eta_0| \leq cn^{-1/2}. \quad (60)$$

As in (A2), suppose P_0 has a density g

$$P_0(s, t) = \int_{-\infty}^t g_0(s, u) du \quad (61)$$

so that

$$Q_{\theta_0}(s, t) = \int_{-\infty}^{t-s'\eta_0} g_0(s, u) du. \quad (62)$$

If x_1, \dots, x_n are iid Q_n , let \hat{Q}_n denote the empirical measure, as an element of $L_\infty(S^d \times R^1)$, and set

$$W_n = n^{1/2}[\hat{Q}_n - Q_n] \quad (63)$$

so that $W_n \Rightarrow W$ as described in Section 2.2.

For θ_0 fixed as above, define a linear map ξ from $R^d \times L_\infty(S^d \times R^1) \rightarrow L_\infty(S^d \times R^1)$, by

$$\xi(\eta, H)(s, t) = g_0(s, t - s'\eta_0)(s'\eta) - H(s, t - s'\eta_0) \quad (64)$$

where $\eta \in R^d$, $H \in L_\infty$. For $f \in L_\infty(S^d \times R^1)$, let

$$|f|_n = \max_{i \leq k_n} \sup_t |f(s_i, t)|$$

denote the stochastic norm of Section 2.6, and let $\Theta_n = (\theta_0^*, \theta_1^*, \dots, \theta_{j_n^*})$ be the stochastic search of the parameter set, given in the same section.

Finally, let

$$V = (Y, Z) \quad (65)$$

where (Y, Z) is the mean 0, Gaussian, $R^d \times L_\infty(S^d \times R^1)$ -valued random variable given by Corollary 1.

Recall that j_n, k_n are the sizes of the search sets for $\Theta_n, |\cdot|_n$, defined in Section 2.6.

Theorem 2 Assume $\Gamma_0 \supset \Gamma_S$, and (59), (61). Under hypotheses (i)–(viii) below, the following convergence in distribution holds under Q_n , if $k_n \rightarrow \infty$ and $j_n \rightarrow \infty$:

$$\lim_{n \rightarrow \infty} \min_{\theta \in \Theta_n} n^{1/2} |\hat{Q}_n - Q_\theta| = \inf_{v \in \text{support } V} |W - \xi(v; \cdot)|.$$

Before stating these hypotheses, define, for real functions f on a normed space B , the *modulus of continuity* $w(f; \delta)$ by

$$w(f; \delta) = \sup_{|x-y| < \delta} |f(x) - f(y)|, x, y \in B. \quad (66)$$

Then the hypotheses for Theorem 2 are given by (i)–(viii) below.

- (i) The location estimate, $\hat{\eta}_n \equiv \eta(\hat{Q}_n)$ satisfies Definition 4 and (A1), under Q_n .
- (ii) For every $\epsilon > 0$, there exists $\delta > 0$ such that $w(n^{1/2}[Q_n - Q_0]; \delta) < \epsilon$, all large n where Q_n, Q_0 are regarded as elements of $L_\infty(S^d \times R^1)$.
- (iii) Let $\hat{\eta}_n = \eta(\hat{Q}_n)$ and let η_i^* be the bootstrap replica of $\hat{\eta}_n$ described in Section 2.6, $i \geq 1$. Then,

$$\lim_{n \rightarrow \infty} j_n Q_n \{n^{1/2} |\eta_i^* - \eta_n| > a_n\} = 0 \quad \text{for } i = 0, 1$$

for some sequence $\{a_n\}$, $a_n \uparrow \infty$, $a_n^2 n^{-1/2} \rightarrow 0$.

Define δ_n by

$$\delta_n = a_n n^{-1/2}$$

so that the condition in (iii) is $a_n \delta_n \rightarrow 0$. The hypotheses can then be continued:

- (iv) The density g on $S^d \times R^1$ satisfies $a_n w(g(\cdot, \cdot); \delta_n) \rightarrow 0$
- (v) If $u = (s_1, t_1) \in S^d \times R^1$ and $v = (s_2, t_2) \in S^d \times R^1$ define $d(u, v) = |s_1 - s_2| \vee |t_1 - t_2|$, and $d_n(u, v) = Q_n(A(s_1, t_1) \Delta A(s_2, t_2))$ where $|\cdot|$ is Euclidean norm, $A(s, t)$ is the half-space given in Section 2.2, and Δ denotes symmetric difference. Let h be an increasing function on $[0, \infty)$, $h(0) = 0$, continuous at 0. Then, for some such h :

$$h(d(u, v)) \geq d_n(u, v), \quad \text{all large } n$$

(vi) The sequence δ_n defined before (iv), and the h of (v) satisfy

$$h(\delta_n) \log n \rightarrow 0$$

(vii) The size j_n of the search set Θ_n satisfies

$$j_n n^{-q} \rightarrow 0$$

for some $q > 0$.

(viii) The size k_n , which determines the stochastic norm, satisfies

$$k_n \geq \delta_n^{-d} \quad \text{for all large } n$$

Comments on the hypotheses. While Theorem 2 is of independent interest, its main application is to justify the calculation of critical values by bootstrap methods. For this, one must apply the theorem conditionally with Q_n taken to be the fitted distribution described above (42). Under these circumstances, some of the hypotheses are more or less automatically fulfilled, and essentially impose (a) regularity on Q_0 and (b) rate conditions on j_n, k_n the sizes of the search sets for $\Theta_n, |\cdot|_n$ respectively. Typical rate conditions are: $j_n = o(n^{+p/4})$ and $k_n \geq n^{d/4}$, where $0 < p < \infty$ depends on $\eta(\cdot)$; see Example 10 below.

To begin comment on the individual hypotheses, note first that the trimmed means of Section 2.6 satisfy (i), because its simplistic nature allows application of known one-dimensional results. Such single trimmed means may not be the best from other points of view (cf. Donoho & Gasko 1987), but these are easy to compute and do the required search. It was to allow for other possibilities (eg the midmean on R^d) that we took the more general approach of Section 2.7.

If $n^{1/2}(Q_n - Q_0)$ has a density f_n , bounded independently of n , then (ii) is immediate. It also holds in probability (under mild hypotheses) if the empirical measure \hat{Q}_n replaces Q_n , by known results on the modulus of the empirical process.

Regarding hypothesis (iii), note that j_n, a_n satisfying this condition certainly exist, because of the asymptotics of Section 2.7. Often if (iii) holds for $i = 0$, it will hold automatically for $i = 1$ by a familiar conditioning argument.

Condition (iv) will hold if $g(\cdot, \cdot)$, the density of Q_0 , has a bounded continuous derivative; then $w(g(\cdot, \cdot), \delta) \leq c\delta$ in which case the rate condition in (iv) follows from that in (iii). Thus (iv), a hypothesis on Q_0 only, typically adds no new condition. Hypotheses (i)–(iv) serve to ensure an asymptotic form of differentiability of the functional $\theta \rightarrow Q_\theta$, which maps $R^d \times L_\infty(S^d \times R^1)$ to $L_\infty(S^d \times R^1)$.

Hypothesis (v) ensures that the Euclidean distance between u, v is “comparable” to the Q_n -symmetric difference between the relevant half spaces.

Typical h are of the form $h(x) = |x|^\alpha$ for some $\alpha > 0$. The reason for this hypothesis is that we wish to avail ourselves of results involving moduli of continuity for the empirical process indexed by half-spaces; the known results here use the d_n metric. On the other hand, our need for differentiability of the functional $\theta \rightarrow Q_\theta$ requires a different metric on half spaces. Development of exact moduli of continuity in the “Euclidean” metric d for the empirical process on half spaces is an interesting problem with useful applications. Heretofore it has not been addressed by the empirical process community, and, it is beyond the scope of this paper. When Q_n is taken to be \hat{Q}_n in (v), further weakening of this hypothesis is possible, as will be evident from the proofs.

Hypothesis (vi) further slows the growth rate of j_n , the size of Θ_n . In particular, j_n cannot (according to the development here) increase faster than a power of n , in contrast to the parametric case. This restriction arises because of the estimation of the non-parametric part of the parameter $\theta = (\eta, P)$.

Finally, hypothesis (vii) gives a rate of growth for the search size of $|\cdot|_n$; this growth rate ensures a certain uniformity in which the stochastic norm approaches the true norm. In the case where the parameter set is finite dimensional, no such rate is needed, since the finite dimensionality of Θ automatically ensures the needed uniformity. Note also that the rate for k_n depends on the dimension d of R^d in a strong way. It remains to be seen whether this dimensionality dependence can be lessened to some extent by, e.g., devising a more efficient search method to generate the stochastic norm.

Example 10 We indicate simple conditions which give the rates for k_n , j_n mentioned in (vii) and (viii). Suppose that $Q_n(n^{1/2}|\eta_1^* - \eta_n|/\lambda) \leq c\lambda^{-p}$ for some $p > 0$, that g has a bounded continuous derivative, and that $h(x) = |x|^\alpha$ for some $\alpha > 0$. Then we may take $a_n = o(n^{+1/4})$, leading to $j_n = o(n^{p/4})$ and $k_n \geq n^{+d/4}$. Note that these rates are independent of h .

An important variant of Theorem 2 occurs when j_n remains fixed at, say, j_0 for all n . Then one may dispense with the rate condition on k_n :

Theorem 3 Assume $j_n = j_0 < \infty$ for all n , and that (59), (61), (i), (ii) hold. Let V_0, V_1, \dots, V_{j_0} be iid copies of V (cf (65)). Then under Q_n , as $n \rightarrow \infty$, $k_n \rightarrow \infty$: $\min_{\theta \in \Theta_{j_0}} n^{1/2} |\hat{Q}_n - Q_\theta|_n \Rightarrow \min\{|W - \xi(V_0)|, \min_{1 \leq i \leq j_0} |W - \xi(V_0 + V_i)|\}$.

This is proved by an easier version of the proof below.

2.9 Proof of Theorem 2

Lemma 4 Let Q be any probability on R^d , and $\eta \in R^d$. Define $P(\cdot) =$

$Q(\cdot + \eta)$. Then

$$\begin{aligned} & \sup_{s,t} |P(s, t - s'\eta) - P_0(s, t - s'\eta_0) - \eta(\eta - \eta_0; P - P_0)(s, t)| \\ & \leq |\eta - \eta_0|w(g; |\eta - \eta_0|) + w(P - P_0; |\eta - \eta_0|). \end{aligned} \quad (67)$$

Proof: Write $P(s, t - s'\eta) - P_0(s, t - s'\eta_0) = A_1 + A_2 + A_3$ where

$$\begin{aligned} A_1 &= P(s, t - s'\eta_s) - P_0(s, t - s'\eta_0), \\ A_2 &= P_0(s, t - s'\eta) - P_0(s, t - s'\eta_0), \\ A_3 &= [P(s, t - s'\eta) - P_0(s, t - s'\eta)] - [P(s, t - s'\eta_0) - P_0(s, t - s'\eta_0)]. \end{aligned}$$

Then by the mean value theorem (with s fixed)

$$A_2 = g(s, t - s'\eta_0)\langle s, \eta - \eta_0 \rangle \quad (68)$$

with an error which is less than

$$|\eta - \eta_0|w(g(s; \cdot); |\eta - \eta_0|). \quad (69)$$

It is also clear that

$$|A_3| \leq w(P - P_0; |\eta - \eta_0|). \quad (70)$$

The lemma then follows from (62), (63)

Lemma 5

$$\min_{\theta \in \Theta_n} n^{1/2} |\hat{Q}_n - Q_\theta|_n = \min_{\theta \in \Theta_n} |W_n - n^{1/2} \xi(\theta - \theta_n)|_n + o_{Q_n}(1)$$

where W_n is given in (63), and, if $\theta = (\eta, P)$, then $\xi(\theta) \equiv \xi(\eta, P)$; $\theta_n = (\eta_n, P_n)$, $\eta_n = \eta(Q_n)$, $P_n = Q(\cdot + \eta_n)$

Proof: Let $R_n(\eta - \eta_0, P - P_0)$ be the upper bound in (67). Then $n^{1/2}(\hat{Q}_n - Q_\theta) = W_n - n^{1/2}(Q_\theta - Q_{\theta_n})$, and using Lemma 4 we see that

$$|Q_\theta - Q_{\theta_n} - \xi(\eta - \eta_n, P - P_n)| \leq R(\eta - \eta_n, P - P_n) + R(\eta_n - \eta_0, P_n - P_0).$$

Thus

$$\min_{\theta \in \Theta_n} n^{1/2} |\hat{Q}_n - Q_\theta|_n = \min_{\theta \in \Theta_n} |W_n - \xi(n^{1/2}(\theta - \theta_n))|_n + \text{error} \quad (71)$$

where

$$\text{error} \leq \max_{0 \leq i \leq j_n} n^{1/2} R(\eta_i^* - \eta_n, P_i^* - P_n) + n^{1/2} R(\eta_n - \eta_0, P_n - P_0). \quad (72)$$

The second remainder term in (72) is bounded above by

$$n^{1/2} |\eta_n - \eta_0|w(g; |\eta_n - \eta_0|) + w(n^{1/2}(P_n - P_0); |\eta_n - \eta_0|).$$

Since $n^{1/2}|\eta_n - \eta_0|$ is bounded (cf 61), this goes to zero by (ii), (iv). To deal with the first error term in (65), note that because of (iii):

$$\max_{0 \leq i \leq j_n} n^{1/2} |\eta_i^* - \eta_n| \leq a_n \quad (73)$$

with probability approaching 1 as $n \rightarrow \infty$. This argument uses the fact that if Z_i are iid, non-negative, then $P\{\max_{i \leq j_n} Z_i > \lambda\} \leq j_n P\{Z_1 > \lambda\}$. On the other hand, methods of Alexander (1984), yield, under hypotheses (v), (vi), that for any $\epsilon > 0$: $Q_n\{w(n^{1/2}(P_1^* - P_n), \delta_n) > \epsilon\} \leq n^{-q}$ for all large n , where q is any fixed positive number (the “large n ” condition depends on q). Because of hypothesis (vii), this implies, by the same argument that gave (73), that

$$\max_{0 \leq i \leq j_n} w(n^{1/2}(P_i^* - P_n), \delta_n) \leq \epsilon \quad (74)$$

with probability approaching 1 as $n \rightarrow \infty$. Applying (73), (74) we find that, with probability approaching 1 as $n \rightarrow \infty$: $\max_{0 \leq i \leq j_n} n^{1/2} R(\eta_i^* - \eta_n, P_i^* - P_n) \leq a_n w(g, \delta_n) + \epsilon$. By hypothesis (iv), we thus conclude that the error term goes to 0 in probability, as $n \rightarrow \infty$, completing the proof of Lemma 5.

Continuing the proof, we now turn to the search set $\{s_i, 1 \leq i \leq k_n\}$ that determines $|\cdot|_n$. Define “mesh s_i ” to be $\sup_{s \in S^d} \min_{i \leq k_n} |s_i - s|$.

Lemma 6 *Let $\{b_n\}$ be a sequence of numbers, $b_n \uparrow \infty$. Let s_1, \dots, s_{k_n} be iid, uniform on S^d . If $k_n \geq (\log b_n)^2 b_n^{d-1}$ then mesh $\{s_1, \dots, s_{k_n}\} \leq (b_n)^{-1}$ with probability approaching 1 as $n \rightarrow \infty$.*

Proof: A proof can be carried out along these lines. First pave S^d with patches of equal diameter (possibly allowing some overlap) where the diameter is $\epsilon_i = 1/b_n$. A reasonably efficient paving will require roughly $C_d(b_n)^{d-1}$ such patches, where C_d is a constant depending on dimension (it is of order 2^d , but we do not need this) and each patch will have surface area ϵ_n^{d-1} , approximately. Let C_1, C_2, \dots be a list of these patches. Then $P\{\text{mesh } s_i \leq 2\epsilon_n\} \leq P\{\text{each } C_i \text{ contains at least one point of } \{s_i\}\} = 1 - P\{\text{at least one } C_i \text{ is empty}\}$. But $P\{\text{at least one } C_i \text{ is empty}\} \leq C_d(b_n)^{d-1} P(C_1 \text{ is empty}) = C_d(b_n)^{d-1} (\text{Area } C_1)^{k_n} \leq C_d(b_n)^{d-1} (1 - \epsilon_n^{d-1})^{k_n}$, $\epsilon_n = b_n^{-1}$. Thus $P\{\text{mesh } s_i \leq \epsilon_n\} \geq 1 - C_d(1/\epsilon_n)^{d-1} (1 - \epsilon_n^{d-1})^{k_n}$. Standard estimates then show that if k_n is chosen as in the statement of the Lemma 6, then $(1/\epsilon_n)^{d-1} (1 - \epsilon_n^{d-1})^{k_n} \rightarrow 0$, completing the proof.

To finish the proof of the theorem, define for $M = (\eta, H) \in R^d \times L_\infty(S^d \times R')$

$$Q_n(M; t, s) = |W_n(s, t) - \xi(\eta, H)(s, t)|. \quad (75)$$

Then

$$|\min_{\theta \in \Theta_n} \max_{1 \leq i \leq k_n} \sup_t Q_n(n^{1/2}(\theta - \theta_n); s_i, t)| \quad (76)$$

$$\begin{aligned} & - \min_{\theta \in \Theta_n} \sup_s \sup_t |\mathbf{Q}_n(n^{1/2}(\theta - \theta_n); s, t)| \\ & \leq \max_{\theta \in \Theta_n} w(\mathbf{Q}_n(n^{1/2}(\theta - \theta_n); \cdot, \cdot); \delta_n) \end{aligned} \quad (77)$$

since the choice of k_n and Lemma 6 force mesh $\{s_i\} \leq \delta_n$. Using the definition of ξ and an elementary argument:

$$\begin{aligned} & w(\mathbf{Q}_n(n^{1/2}\eta_i^* - \eta_n), n^{1/2}(P_i^* - P_n); (\cdot, \cdot)), \delta_n) \\ & \leq w(W_n; \delta_n) + w(\xi((\eta_i^* - \eta_n)\sqrt{n}, \sqrt{n}(P_i^* - P_n)); \delta_n) \end{aligned} \quad (78)$$

$$\begin{aligned} & \leq w(W_n; \delta_n) + K_1|\eta_i^* - \eta_n|\sqrt{n}\delta_n \\ & + w(g_0; \delta_n)n^{1/2}|\eta_i^* - \eta_n| + w(n^{1/2}(P_i^* - \hat{P}_n); \delta_n) \end{aligned} \quad (79)$$

where $K_1 = \sup_{s,t} g_0(s, t)$. Next, maximize the upper bound in (78) over $i \leq j_n$. Fix $\epsilon > 0$; argument similar to that in the proof of Lemma 5 gives as an upper bound to the second expression in (76):

$$K_1 a_n \delta_n + a_n w(g_0; \delta_n) + \epsilon \quad (80)$$

for all sufficiently large n . The hypotheses of Theorem 2 force the n -dependent terms of 80 to go to zero. By (76), (80) and Lemma 5, therefore, it suffices to show that

$$\min_{\theta \in \Theta_n} |W_n - \xi(n^{1/2}(\theta - \theta_n); \cdot)| \Rightarrow \inf_{V \in \text{support } V} |W_n - \xi(\cdot; \cdot)|. \quad (81)$$

In particular, the stochastic norm has been replaced by the true L_∞ norm by the argument (76)–(80). Let \hat{G}_n be the empirical measure of $(\theta_i - \theta_n)n^{1/2}$, $\theta_i \in \Theta_n$, $\theta_i \equiv (\eta_i^*, P_i^*)$. Then the left side of (81) may be rewritten as

$$\text{essinf}_{\hat{G}_n} |W_n - \xi(\cdot)|. \quad (82)$$

But $W_n \Rightarrow W$, $\hat{G}_n \Rightarrow G$, the distribution of V ; by arguments of Millar (1993), the expression in (82) converges to $\text{essinf}_G |W - \xi(\cdot)|$. Since ξ is continuous on $R^d \times L_\infty(S^d \times R^1)$ the essinf_G may be replaced by $\inf_{V \in \text{support } V}$.

This completes the proof of Theorem 2.

2.10 Proofs of Propositions 5 and 6

Let (U, \mathbf{U}, μ_0) be a probability space. Let \mathbf{F} denote a collection of uniformly bounded, real functions on U . Assume that

$$\mathbf{F} \text{ is precompact in the } L^2(\mu_0) \text{ metric.} \quad (83)$$

Let $W = \{W(f) : f \in \mathbf{F}\}$ be a Gaussian process with parameter set \mathbf{F} having mean 0 and covariance

$$EW_f W_g = \mu_0(fg) - (\mu_0 f)(\mu_0 g), f \in \mathbf{F}, g \in \mathbf{F}. \quad (84)$$

Here $\mu_0 f$ denotes the integral of f with respect to μ_0 . Let Q_0 be the distribution of W , and suppose that

$$Q_0 \text{ is supported by } C_0(\mathbf{F}) \quad (85)$$

where $C_0(\mathbf{F})$ is the collection of uniformly continuous, real function on \mathbf{F} , where \mathbf{F} is endowed with the $L^2(\mu_0)$ metric. Thus the process W has continuous paths on \mathbf{F} . Let H denote the Hilbert space of real functions h on U , such that $\int h^2 d\mu_0 < \infty$ and $\int h d\mu_0 = 0$. If f is a bounded real function on U , define

$$\bar{f} = \int f d\mu_0. \quad (86)$$

Define the mapping $\tau : H \rightarrow C_0(\mathbf{F})$ by

$$(\tau h)(f) = \langle f - \bar{f}, h \rangle_H \quad (87)$$

where $\langle \cdot, \cdot \rangle_H$ is the inner product of $H : \langle h, k \rangle_H = \int h k d\mu_0$. It is immediate that τ indeed maps into $C_0(\mathbf{F})$ because the metric on \mathbf{F} is that of $L^2(\mu_0)$.

Lemma 7 *Assume (83) - (87). Then the triplet (τ, H, B) , where B is the closure of τH in $C_0(\mathbf{F})$, is an abstract Wiener space, and its canonical normal distribution on B is Q_0 , the distribution of W .*

Proof: A description of abstract Wiener spaces and their canonical normal distributions suitable for the present development may be found in Millar (1983). Let E denote expectation with respect to the standard normal cylinder measure on H , and x a (cylinder) rv on H having the canonical normal distribution. Let m_1, m_2 denote elements of the dual of $C_0(\mathbf{F})$, so m_1, m_2 are (signed) measures on the space \mathbf{F} . Then

$$\begin{aligned} E \int (\tau x)(f) m_1(df) \int (\tau x)(g) m_2(dg) \\ = \iint E \langle f - \bar{f}, x \rangle_H \langle g - \bar{g}, x \rangle_H m_1(df) m_2(dg) \\ = \iint \langle f - \bar{f}, g - \bar{g} \rangle_H m_1(df) m_2(dg) \\ = \iint (\mu(fg) - (\mu f)(\mu g)) m_1(df) m_2(dg). \end{aligned}$$

This identifies Q_0 as the image, under τ , of the standard normal cylinder measure on H ; see Millar (1983, Chapter V, Section 2). Since Q_0 is countably additive by hypothesis, this completes the proof.

Define

$$\{Q_h, h \in H\} = \text{canonical normal shift family for } (\tau, H, B). \quad (88)$$

For $h \in H$, define $\mu_{n^{-1/2}h}$ to be the probability with density

$$(1 + n^{-1/2}h(u))\mu_0(du) \quad (89)$$

and let

$$\mu_h^{(n)} = n\text{-fold product of } \mu_{n^{-1/2}h}. \quad (90)$$

Then, the statistical experiments $\{\mu_h^{(n)}, h \in H\}$ converge, in the sense of Le Cam (1972), to $\{Q_h, h \in H\}$; see Millar (1983, Chapter VI, Section 1), for further explanation. From now on, any probability μ on U will be identified with an element of $L_\infty(\mathbf{F})$ by means of the mapping $\mathbf{F} \rightarrow R^1$ given by

$$f \rightarrow \mu(f). \quad (91)$$

Note that, with this convention,

$$n^{1/2} (\mu_{n^{-1/2}h}(f) - \mu_0(f)) = (\tau h)(f). \quad (92)$$

Let ψ be a functional taking values in a Banach space B_2 , and defined on probabilities μ which are regarded as elements of $L_\infty(\mathbf{F})$, as in (91). Thus, ψ is a B_2 -valued map from a subset of $L_\infty(\mathbf{F})$. Let x_1, \dots, x_n be i.i.d. U -valued random variables with common unknown distribution μ . The task is to estimate $\psi(\mu)$ efficiently.

To develop a general approach, fix μ_0 and suppose that ψ is *differentiable over H at μ_0* : there exists a continuous linear map $\mathbf{l} : L_\infty(\mathbf{F}) \rightarrow B_2$ such that for each $h \in H$

$$\psi(\mu_{n^{-1/2}h}) = \psi(\mu_0) + n^{-1/2}\mathbf{l}(\tau h) + o(n^{-1/2}). \quad (93)$$

Assume that

(A3) the range of \mathbf{l} is dense in B_2 .

This assumption can be weakened, but it suffices for proving Propositions 5, 6. Let \mathbf{D}_n denote the collection of estimates of $\psi(\mu)$ that are available after observing x_1, \dots, x_n . Let $M_{n,c}$ denote the collection of probabilities μ such that

$$\sup | \mu_n(fg) - \mu_0(fg) | \leq cn^{-1/2} \quad (94)$$

where $f, g \in \mathbf{F} \cup \{1\}$. The $n^{-1/2}$ rate here is kept only for historical reasons; in the present situation it can be replaced any sequence $\{\epsilon_n\}$, where $\epsilon_n \downarrow 0$. Finally, let q be an increasing function on $[0, \infty)$ with $q(0) = 0$. For convenience, assume that q is bounded and continuous—an assumption easily removed by standard arguments. For any μ , let μ^n denote the n -fold product measure.

Lemma 8 Assume (83)–(87), (A3) and (93)–(94). Then

$$\lim_{n \rightarrow \infty} \lim_{c \uparrow \infty} \inf_{\hat{T}_n \in \mathbf{D}_n} \sup_{\mu \in M_{nc}} \int q(n^{1/2}|\hat{T}_n - \psi(\mu)|) d\mu^n \geq \int q(|\log x|) Q_0(dx). \quad (95)$$

Proof: Let \mathbf{N} denote the null space of $\mathbf{l} \circ \tau$, and let \mathbf{N}^\perp be its orthocomplement in H . Then $(\mathbf{l} \circ \tau, \mathbf{N}^\perp, B_2)$ is an abstract Wiener space, because of (A3) and Lemma 7; the canonical normal on B_2 here is the image of Q_0 under \mathbf{l} . To obtain a lower bound for the left side of (95), replace the $\sup_{\mu \in M_{nc}}$ by $\sup_{\mu_{n-1/2,h}}$ where in the latter $\sup h$ ranges over those $h \in H$ satisfying $|\tau h| \leq c$. With this change, the integrand of $d\mu_h^n$ can then be replaced by $q(|\hat{V}_n - \mathbf{l}\tau h|)$, where $\hat{V}_n = n^{1/2}[\hat{T}_n - \psi(\mu_0)]$. Thus a lower bound of (95) is

$$\lim_{n \rightarrow \infty} \lim_{c \uparrow \infty} \inf_{\hat{V}_n \in \mathbf{D}'_n} \sup_{h: |\tau h| \leq c} \int q(|\hat{V}_n - \mathbf{l}\tau(h)|) d\mu_h^n \quad (96)$$

where \mathbf{D}'_n is the collection of all B_2 -valued estimators. Since $(\mathbf{l} \circ \tau, \mathbf{N}^\perp, B_2)$ is an abstract Wiener space, we may apply the Hajek-LeCam asymptotic minimax theorem (Le Cam 1972, or Millar 1983, Chapter VI, Section 2) to see that (96) is at least as large as the right side of (95). This completes the proof.

Remark The argument as given appears to require a stronger notion of differentiability than that of (93). The familiar argument, wherein one considers only finite dimensional subspaces of H , and later lets the dimension ∞ , has been deliberately omitted.

Next, suppose \mathbf{F} admits a central limit theorem: if $\{\mu_n\}$ satisfies (94), if $\hat{\mu}_n$ is the empirical measure of x_1, \dots, x_n , and if $W_n = n^{1/2}[\hat{\mu}_n - \mu_n]$, then

$$W_n \Rightarrow W, \quad (97)$$

weak convergence in $L_\infty(\mathbf{F})$, where W satisfies (84). See Le Cam (1983), for results on such triangular array theorems; see Dudley (1978), for the pioneering paper on the subject. Suppose that, under μ_n ,

$$n^{1/2}[\psi(\hat{\mu}_n) - \psi(\mu_n)] \rightarrow \mathbf{l}[W]. \quad (98)$$

Such convergence is not guaranteed under the weak differentiability condition (93).

Lemma 9 Under (97) and (98), the estimator $\psi(\hat{\mu}_n)$ of $\psi(\mu)$ is locally asymptotic minimax in the sense that

$$\lim_{n \rightarrow \infty} \lim_{c \uparrow \infty} \sup_{\mu \in M_{nc}} \int q(n^{1/2}|\psi(\hat{\mu}_n) - \psi(\mu)|) d\mu_n = \int q(|\log x|) Q_0(dx).$$

Remark Of course the second expression in Lemma 9 is $Eq(|\mathbf{1} \circ W|)$.

Since the proof of Lemma 9 is immediate we complete now the proof of Proposition 5. For this, take in the above lemmas, $\mathbf{F} = \{I_A : A = \text{half space in } \mathbb{R}^d\}$. Take $\psi \equiv \pi_0$, where π_0 is the symmetrization operation described in Section 2.4. Since π_0 is continuous and linear, it is immediate that the differentiability property is satisfied. That (98) holds in the present situation follows from Beran & Millar (1986, Section 4), for example, and the continuity of π_0 . For the present case one may take B_2 to be the subset of $L_\infty(\mathbf{F})$ left invariant by π_0 . Proposition 5 is now immediate from Lemma 8 and 9. Proposition 6 is also immediate from the above development and Millar (1985).

Acknowledgments: Research of R. J. Beran and P. W. Millar supported by National Science Foundation Grants DMS8701426 and DMS9224868. The authors thank Dr. M. Loranger for useful comments on an earlier draft.

2.11 REFERENCES

- Alexander, K. S. (1984), 'Probability inequalities for empirical processes and a law of the iterated logarithm', *Annals of Probability* **12**, 1041–1067.
- Anderssen, S. A. (1975), 'Invariant normal models', *Annals of Statistics* pp. 132–54.
- Beran, R. (1984), 'Bootstrap methods in statistics', *Jahresbericht der Deutschen Mathematischen Vereinigung* pp. 14–30.
- Beran, R. & Millar, P. (1989), 'A stochastic minimum distance test for multivariate parametric models', *Annals of Statistics* pp. 125–140.
- Beran, R. J. & Millar, P. W. (1985), Rates of growth for weighted empirical processes, in L. Le Cam & R. A. Olshen, eds, 'Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Volume II', Wadsworth, Belmont, CA, pp. 865–887.
- Beran, R. J. & Millar, P. W. (1986), 'Confidence sets for a multivariate distribution', *Annals of Statistics* pp. 431–443.
- Beran, R. J. & Millar, P. W. (1987), 'Stochastic estimation and testing', *Annals of Statistics* pp. 1131–1154.
- Beran, R. J. & Millar, P. W. (1992), Tests of fit for logistic models, in K. V. Mardia, ed., 'The Art of Statistical Science', Wiley, pp. 153–172.

- Bickel, P. & Freedman, D. (1981), 'Some asymptotic theory for the bootstrap', *Annals of Statistics* pp. 1196–1217.
- Bickel, P. J. & Millar, P. W. (1992), 'Uniform convergence of probabilities on classes of functions', *Statistica Sinica* pp. 1–15.
- Chow, E. D. (1991), Stochastic minimum distance tests for censored data, PhD thesis, University of California at Berkeley.
- Chung, K. (1968), *A Course in Probability Theory*, Harcourt, Brace and World, New York.
- Donoho, D. & Gasko, M. (1987), Multivariate generalizations of the median and trimmed mean, Technical Report 133, Statistics Department, University of California at Berkeley.
- Dudley, R. M. (1978), 'Central limit theorems for empirical measures', *Annals of Probability* **6**, 899–929.
- Eaton, M. (1983), *Multivariate Statistics: A vector space approach*, Wiley, New York.
- Efron, B. (1979), 'Bootstrap methods: another look at the jackknife', *Annals of Statistics* pp. 1–26.
- Huber, P. (1985), 'Projection pursuit', *Annals of Statistics* pp. 435–474.
- Le Cam, L. (1972), Limits of experiments, in L. Le Cam, J. Neyman & E. L. Scott, eds, 'Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 245–261.
- Le Cam, L. (1983), A remark on empirical measures, in P. J. Bickel, K. Doksum & J. L. Hodges, eds, 'Festschrift for Erich Lehmann', Wadsworth, Belmont, California, pp. 305–327.
- Loranger, M. (1989), A stochastic test of ellipsoidal symmetry, PhD thesis, University of California at Berkeley.
- Millar, P. (1985), 'Nonparametric applications of an infinite dimensional convolution theorem', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* pp. 545–556.
- Millar, P. W. (1983), 'The minimax principle in asymptotic statistical theory', *Springer Lecture Notes in Mathematics* pp. 75–265.
- Millar, P. W. (1988), Stochastic test statistics, in 'Proceedings of the 20th Symposium on the Interface: Computer Science and Statistics', American Statistical Association, pp. 62–68.

- Millar, P. W. (1993), Stochastic search and the empirical process,
Technical report, University of California at Berkeley.
- Perlman, M. D. (1988), 'Comment: group symmetry covariance models',
Statistical Science pp. 421–425.
- Pollard, D. (1980), 'The minimum distance method of testing', *Metrika*
pp. 43–70.
- Pollard, D. (1982), 'A central limit theorem for empirical processes',
Journal of the Australian Mathematical Society (Series A)
pp. 235–248.
- Singh, K. (1981), 'On the asymptotic accuracy of Efron's bootstrap',
Annals of Statistics pp. 1187–1195.
- Watson, G. S. (1983), *Statistics on Spheres*, John Wiley and Sons, New
York.

3

Local Asymptotic Normality of Ranks and Covariates in Transformation Models

P. J. Bickel¹
Y. Ritov²

3.1 Introduction

Le Cam & Yang (1988) addressed broadly the following question: Given observations $X^{(n)} = (X_{1n}, \dots, X_{nn})$ distributed according to $P_\theta^{(n)}$; $\theta \in R^k$ such that the family of probability measures $\{P_\theta^{(n)}\}$ has a locally asymptotically normal (LAN) structure at θ_0 and a statistic $Y^{(n)} = g_n(X^{(n)})$:

- (i) When do the distributions of $Y^{(n)}$ also have an LAN structure at θ_0 ?
- (ii) When is there no loss in information about θ in going from $X^{(n)}$ to $Y^{(n)}$?

In this paper we exhibit an important but rather complicated example to which the Le Cam-Yang methods may, after some work, be applied.

Semiparametric transformation models arise quite naturally in survival analysis. Specifically, let $(Z_1, T_1), (Z_2, T_2), \dots, (Z_n, T_n)$ be independent and identically distributed with Z , a vector of covariates, having distribution H , and T real. We suppose there exists an *unknown strictly monotone* transformation $a_0 : R \rightarrow R$ such that, given $Z = z$, $a_0(T)$ is distributed with distribution function $F(\cdot, z, \theta)$ where $\{F(\cdot, z, \theta) : \theta \in \Theta \subset R^d\}$ is a regular parametric model. That is, the $F(\cdot, z, \theta)$ are dominated by μ with densities $f(\cdot, z, \theta)$ and the map $\theta \rightarrow \sqrt{f(\cdot, z, \theta)}$ is Hellinger differentiable. As is usual in these models we take μ to be Lebesgue measure and $a'_0 > 0$. The most important special cases of these models are the regression models where $\theta = (\eta, \nu)$, and defining the distribution of T given Z structurally,

$$a_0(T) = \eta + \nu' Z + \epsilon,$$

¹University of California, Berkeley

²The Hebrew University of Jerusalem

where ϵ is independent of Z with fixed known distribution G_0 . If G_0 is an extreme value distribution this is the Cox proportional hazards model. If G_0 is log Pareto this is the Clayton & Cuzick (1986) model, which introduces a gamma distributed frailty into the Cox model. Finally, if G_0 is Gaussian this is the natural semiparametric extension of the Box-Cox model (see Doksum 1987 for instance).

If θ is fixed in any of these semiparametric models it is evident that a maximal invariant under the group of monotone transformations of T is the vector

$$\tilde{Z} \equiv (Z_{(1)}, \dots, Z_{(n)})$$

where $T_{(1)} < \dots < T_{(n)}$ are the ordered T_i and $Z_{(i)}$ is the covariate of $T_{(i)}$, that is, $(Z_{(i)}, T_{(i)})$ $i = 1, \dots, n$ is the appropriate permutation of (Z_i, T_i) , $i = 1, \dots, n$. Knowing \tilde{Z} is equivalent to knowing the ranks of the T_i and the corresponding Z_1, \dots, Z_n . It is intuitively plausible that, asymptotically, the (marginal) likelihood of \tilde{Z} which doesn't depend on a_0 can be used for inference about θ in the usual way, without any loss of information. That is, question (i) and (ii) can be answered affirmatively if $P_\theta^{(n)}$ is taken to be, in some sense, the least favorable family for estimation of θ_0 . That is, we take $a_\theta(T)$ given $Z = z$ to have distribution $F_\theta(\cdot|z)$ with a_θ so chosen as to make this the hardest parametric submodel of our semiparametric model at θ_0 , in the sense of Stein (1956)—see also Bickel, Klaassen, Ritov & Wellner (1993, pages 153–175). The detailed construction is given in what follows. Our result can be viewed as a generalization of a classical result of Hájek & Šidák (1967, section VII.1.2), in which LAN is established for the regression models when $\nu = 0$. The independence of ranks and covariates for this special case makes the problem much easier.

Implementation of this approach for inference is well known and simple only for the Cox model where the likelihood of \tilde{Z} is the Cox partial likelihood. In general, the likelihood is expressible only as an n -fold integral. Thus if Λ_n is the log likelihood ratio of $\theta_0 + sn^{-1/2}$ versus θ_0 , we have,

$$\begin{aligned} \Lambda_n(s) &= \log \int \cdots \int \prod_{j=1}^n \frac{f(t_j, Z_{(j)}, \theta_0 + sn^{-1/2})}{f(t_j, Z_{(j)}, \theta_0)} dt_j \\ &\quad - \log \int \cdots \int \prod_{j=1}^n f(t_j, Z_{(j)}, \theta_0) dt_j \end{aligned} \tag{1}$$

However, it is possible to use Monte Carlo methods in a subtle way, drawing on some of the information we develop for our result, to compute Λ_n accurately enough to use it for inference. Alternatively, the analytic approximation to Λ_n that we develop can be used more conveniently for this purpose also—see Bickel (1986). We shall pursue these approaches elsewhere.

The paper is organized as follows. In section 2 we introduce notation and the least favorable family a_θ and establish LAN directly under restrictive

conditions. In section 3 we state our general result and show how it follows from Le Cam and Yang's Theorem 4.

3.2 LAN—the bounded case

For simplicity we state our results for the case θ one dimensional. There is no real loss of generality since proofs carry over easily. Here is some notation. We denote the distribution of Z by H , which we take as known and independent of θ (this is irrelevant). Then if $P_{(\theta,a)}$ is the distribution of (T, Z) and (θ_0, a_0) is the true state of Nature,

$$\frac{dP_{(\theta,a)}}{dP_{(\theta_0,a_0)}}(t,z) = \frac{a'}{a'_0}(t) \frac{f(a(t), z, \theta)}{f(a_0(t), z, \theta_0)}.$$

It is convenient to reparametrize the model. Let $U = F_Y \circ a_0(T)$, where

$$F_Y(t) = \iint_{-\infty}^t f(s, z, \theta_0) ds dH(z)$$

is the marginal distribution function of $a_0(T)$. Suppose F_Y is strictly increasing (this is inessential). Then the conditional density of $U \mid Z = z$ under $P_{(\theta,a_0)}$ is

$$g(u, z, \theta) \equiv f(F_Y^{-1}(u), z, \theta) / f_Y(F_Y^{-1}(u)) \quad (2)$$

where $f_Y = F'_Y$. So if $b \equiv F_Y \circ a \circ a_0^{-1} \circ F_Y^{-1}$ and $Q_{(\theta,b)}$ is the distribution of U under $P_{(\theta,a)}$

$$\frac{dQ_{(\theta,b)}}{dQ_{(\theta_0,b_0)}} = b'(u) \frac{g(b(u), z, \theta)}{g(u, z, \theta_0)} \quad \text{for } 0 < u < 1,$$

the likelihood ratio for a transformation model where transformations are from $(0, 1)$ to $(0, 1)$ and $b_0 = F_Y \circ a_0 \circ a_0^{-1} \circ F_Y^{-1}$ is the identity. The distribution of the ranks of U under $Q_{(\theta,b)}$ is the same as the distribution of the ranks of T under $P_{(\theta,a)}$. We formulate our conditions in terms of derivatives of $\lambda \equiv \log g$ which hold in all cases we have mentioned when f is related to g by (2). For convenience in what follows, we let $\lambda_\theta = \partial \lambda / \partial \theta$, $\lambda_{u\theta} = \partial^2 \lambda / \partial u \partial \theta$ and, in general, let subscripts denote differentiation with primes used for functions of u only. We will use the following assumptions;

A1: The function $\lambda(u, z, \theta)$ is twice differentiable in (θ, u) on $V(\theta_0) \times (0, 1)$ where $V(\theta_0)$ is a neighbourhood of θ_0 .

B: $\lambda, \lambda_\theta, \lambda_u, \lambda_{\theta u}, \lambda_{uu}, \lambda_{\theta uu}$ are uniformly bounded in (u, z, θ) on $(0, 1) \times \text{supp} H \times V(\theta_0)$.

Assumption B is extremely restrictive. It permits essentially only families such that $g(u, z, \theta)$ is bounded away from 0 on $[0, 1]$ and in particular rules out all our examples. However, the argument here makes clear the essential computation which, in the next section, enables us to apply the Le Cam-Yang results to all our examples.

3.2.1 FORMAL DERIVATION OF LEAST FAVORABLE b_θ

Without loss of generality let $\theta_0 = 0$, but without any presumption that 0 corresponds to independence of Z and T . Let

$$b_\theta(u) = u + \theta\Delta(u) \quad (3)$$

where $\Delta \in \mathcal{D}$, and

$$\mathcal{D} \equiv \{\Delta : \Delta, \Delta' \text{ and } \Delta'' \text{ are all bounded on } [0, 1] \text{ and } \Delta(0) = \Delta(1) = 0\}.$$

Then for $|\theta| < \epsilon$, $\epsilon > 0$ the b_θ are transformations of u that depend on θ and b_0 is the identity. Under A1 and B the model $\{Q_{(\theta, b_\theta)} : |\theta| < \epsilon\}$ is regular and the score function at θ is,

$$v_\Delta(u, z, \theta) \equiv \lambda_\theta(u, z, \theta) + \Delta'(u) + \Delta(u)\lambda_u(u, z, \theta).$$

This follows since the map $\theta \rightarrow \sqrt{q_{(\theta, b_\theta)}}$, where $q_{(\theta, b_\theta)} \equiv dQ_{(\theta, b_\theta)}/dQ_{(0, b_0)}$, is pointwise differentiable by A1 and therefore Hellinger differentiable by B.

By standard theory (Bickel et al. 1993, page 70), if the Fisher information $\int v_\Delta^2(u, z, 0)g(u, z, 0) du dH(z)$ is minimized over the closure in $L_2(Q_{(0, b_0)})$ of $\{v_\Delta : \Delta \in \mathcal{D}\}$ by v_{Δ_0} with $\Delta_0 \in \mathcal{D}$ then

$$Ev_{\Delta_0}(U, Z, 0) \left(\Delta'(U) + \Delta(U)\lambda_u(U, Z, 0) \right) = 0 \quad \text{for all } \Delta \in \mathcal{D}. \quad (4)$$

Furthermore, Klaassen (1992) shows that, under regularity conditions, the equality (4) holds if and only if Δ_0 satisfies the Sturm-Liouville equation

$$\Delta_0''(u) - \alpha(u)\Delta_0(u) + \gamma(u) = 0 \quad \text{for } 0 < u < 1, \quad (5)$$

subject to the boundary conditions $\Delta_0(0) = \Delta_0(1) = 0$, where

$$\alpha(u) = -E(\lambda_{uu}(u, Z, 0) | U = u) \quad (6)$$

$$\gamma(u) = E(\lambda_{\theta u}(u, Z, 0) | U = u). \quad (7)$$

Expectations here and in what follows are under $\theta = 0$. Equation (5) is equivalent to,

$$E \left(\frac{\partial}{\partial u} v_{\Delta_0}(u, Z, 0) | U = u \right) = 0 \quad (8)$$

given that, as one would expect from $\int g(u, z, 0) dH(z) \equiv 1$,

$$E(\lambda_u(u, Z, 0) | U = u) = 0. \quad (9)$$

It is easy to see (Bickel 1986, Klaassen 1992) that A1 and B guarantee the validity of (4), (5) and (8), as well as the boundedness of Δ_0 , Δ''_0 . Then $b_\theta^0(u) \equiv u + \theta\Delta_0(u)$ are least favorable. Let $\Lambda_n(s)$ be the log likelihood ratio of the ranks as defined by (1). Note that

$$\Lambda_n(s) = \log E \left(\exp L_n(sn^{-1/2}) \mid \tilde{Z} \right) \quad (10)$$

where

$$L_n(\theta) = \sum_{i=1}^n \log \frac{q(\theta, b_\theta^0)}{q(0, b_0)} (U_i, Z_i),$$

the log likelihood ratio of (U_i, Z_i) , $i = 1, \dots, n$, for $Q_{(\theta, b_\theta^0)}$ and again, expectation is under θ_0 . We have noted earlier that the $\{Q_{(\theta, b_\theta^0)}; |\theta| < \epsilon\}$ family is LAN at $\theta = 0$, and, in fact,

$$L_n(sn^{-1/2}) = sn^{-1/2} \sum_{i=1}^n v_{\Delta_0}(U_i, Z_i, 0) - \frac{s^2}{2} E v_{\Delta_0}^2(U_1, Z_1, 0) + o_p(1). \quad (11)$$

We now establish,

Theorem 1 *Under A1 and B,*

$$\Lambda_n(s) = L_n(sn^{-1/2}) + o_p(1),$$

which establishes our claim in the bounded case.

The result will follow from the *Key Lemma*, which we prove first.

Key Lemma *If $w(v, z)$ is twice differentiable in u and,*

(i) $w(U, Z) \in L_2(Q_{(0, b_0)})$ and $Ew(U, Z) = 0$,

(ii) $w_u(U, Z) \in L_2(Q_{(0, b_0)})$ and $E(w_u(u, Z) \mid U = u) = 0$ for all u ,

(iii) $\sup_{u, z} |w_{uu}(u, z)| < \infty$,

then, if the $Z_{(i)}$ are the concomitants of the order statistics $U_{(i)}$, as in the introduction,

$$\sum_{i=1}^n w(U_i, Z_i) = \sum_{i=1}^n w\left(\frac{i}{n+1}, Z_{(i)}\right) + O_P(1).$$

Proof: Write, expanding around $(U_{(1)}, \dots, U_{(n)})$,

$$\begin{aligned} \sum_{i=1}^n w(U_i, Z_i) &= \sum_{i=1}^n w(U_{(i)}, Z_{(i)}) \\ &= \sum_{i=1}^n w\left(\frac{i}{n+1}, Z_{(i)}\right) + \sum_{i=1}^n \left[w_u(U_{(i)}, Z_{(i)}) \left(U_{(i)} - \frac{i}{n+1} \right) \right. \\ &\quad \left. - \frac{1}{2} w_{uu}(U_{(i)}^*, Z_{(i)}) \left(U_{(i)} - \frac{i}{n+1} \right)^2 \right] \end{aligned} \quad (12)$$

where $|U_{(i)}^* - i/(n+1)| \leq |U_{(i)} - i/(n+1)|$ for all i . Now,

$$\begin{aligned} & E \left(\left[\sum_{i=1}^n w_u(U_{(i)}, Z_{(i)}) \left(U_{(i)} - \frac{i}{n+1} \right) \right]^2 \mid U_{(1)}, \dots, U_{(n)} \right) \\ &= \sum_{i=1}^n E(w_u^2(U, Z) \mid U = U_{(i)}) \left(U_{(i)} - \frac{i}{n+1} \right)^2 \\ &= \sum_{i=1}^n E(w_u^2(U, Z) \mid U = U_i) \left(U_i - \frac{R_i}{n+1} \right)^2 = O_P(1), \end{aligned}$$

where $U_{(R_i)} = U_i$. Here we use (ii) for the first identity and also

$$\sum_{i=1}^n E(w_u^2(U_i, Z_i) \mid U_i) = O_P(n).$$

The third term in (12) is $O_P(1)$ by (iii). \square

The theorem follows readily from the Key Lemma by means of an appeal to Theorem 4 of Le Cam and Yang, applied to the distinguished statistics $n^{-1/2} \sum_{i=1}^n v_{\Delta_0}(\frac{i}{n+1}, Z_{(i)}, 0)$. Alternatively, we can argue directly. Write

$$\begin{aligned} \Lambda_n(s) &= sn^{-1/2} \sum_{i=1}^n v_{\Delta_0} \left(\frac{i}{n+1}, Z_{(i)}, 0 \right) \\ &\quad - \frac{s^2}{2} Ev_{\Delta_0}^2(U_1, Z_1, 0) + \log E \left(\frac{A_n}{B_n} \mid \tilde{Z} \right) \end{aligned}$$

where

$$\begin{aligned} A_n &= \exp L_n(sn^{-1/2}), \\ B_n &= \exp \left(sn^{-1/2} \sum_{i=1}^n v_{\Delta_0} \left(\frac{i}{n+1}, Z_{(i)}, 0 \right) - \frac{s^2}{2} Ev_{\Delta_0}^2(U_1, Z_1, 0) \right). \end{aligned}$$

Then

$$\begin{aligned} \left| E \left(\frac{A_n}{B_n} - 1 \mid \tilde{Z} \right) \right| &\leq E \left(\left| \frac{A_n}{B_n} - 1 \right| 1 \left\{ \frac{A_n}{B_n} \leq M \right\} \mid \tilde{Z} \right) \\ &\quad + \frac{1}{B_n} E \left(A_n 1 \left\{ \frac{A_n}{B_n} > M \right\} \mid \tilde{Z} \right) \\ &\quad + P \left(\frac{A_n}{B_n} > M \mid \tilde{Z} \right) \end{aligned}$$

But A_n is uniformly integrable by LAN and by the Key Lemma $B_n^{-1} = O_P(1)$ and $A_n/B_n = 1 + o_p(1)$. The theorem follows.

3.3 LAN—the general case

To encompass the examples of Section 1 we need to replace condition B. We do so with

A2: $\lambda, \lambda_\theta, \lambda_u, \lambda_{\theta u}, \lambda_{uu}$ are uniformly bounded in (u, z, θ) for $\epsilon \leq u \leq 1 - \epsilon$, $\theta \in V(\theta_0)$, all $\epsilon > 0$.

and

A3: $\lambda_\theta(U, Z, 0) \in L_2(Q_{0,b_0})$. The functions $\lambda_{uu}(U, Z, 0)$, $\lambda_{\theta u}(U, Z, 0)$, and $\lambda_\theta(U, Z, 0)\lambda_u(U, Z, 0)$ are all in $L_1(Q_{(0,b_0)})$. Further, γ and α given by (6) and (7) are continuous on $(0, 1)$ and satisfy:

$$\int_0^1 \alpha(t)t(1-t) dt < \infty \quad \text{and} \quad \sup_{t \in (0,1)} t^{3/2}(1-t)^{3/2}|\gamma(t)| < \infty.$$

It follows (see below) that (5) has a unique solution Δ_0 which is bounded and differentiable. Therefore there exists $\epsilon > 0$ such that b_θ given by (3) is a transformation for $|\theta| < \epsilon$. We require

A4: $v_{\Delta_0}(U, Z, 0) \in L_2(Q_{(0,b_0)})$ and the family $\{Q_{(\theta,b_\theta)} : |\theta| < \epsilon\}$ is regular (LAN) at $\theta = 0$. That is, $L_n(sn^{-1/2})$ obeys (11).

Finally, we require

A5: $\lambda_u(U, Z, 0) \in L_1(Q_{0,b_0})$ and (9) holds.

Note Klaassen (1992) shows that A1–A5 hold for the Clayton-Cuzick and normal (generalized Box-Cox) transformation models. We prove

Theorem 2 Under A1–A5 the conclusion of theorem 1 continues to hold.

It is possible under the conditions of Klaassen (1992) to extend our direct argument. However, under the general conditions A1–A5 it is much easier to appeal to Theorem 4 of Le Cam and Yang. By A3, A4 and (10) we can consider LAN for the ranks and covariates in the context of the parametric model $\{Q_{(\theta,b_\theta)}\}$. By the Le Cam-Yang theorem we need only exhibit statistics $T_n(\tilde{Z})$ such that

$$T_n(\tilde{Z}) = n^{-1/2} \sum_{i=1}^n v_{\Delta_0}(U_i, Z_i, 0) + o_p(1).$$

Let $s_m : [0, 1] \rightarrow [0, 1]$ such that $s_m \in C^\infty$ and

$$s_m(u) = \begin{cases} 1 & \text{if } m^{-1} \leq u \leq 1 - m^{-1} \\ 0 & \text{if } 0 \leq u \leq (2m)^{-1} \text{ or } 1 - (2m)^{-1} \leq u \leq 1 \end{cases}$$

Consider the Sturm-Liouville equation on $[0, 1]$,

$$\Delta''(u) - \alpha_m(u)\Delta(u) + \gamma_m(u) = 0, \quad (13)$$

subject to $\Delta(0) = \Delta(1) = 0$, where

$$\begin{aligned} \alpha_m(u) &= \alpha(u)s_m(u) \\ \gamma_m(u) &= \gamma(u)s_m(u) + E(\lambda_\theta(u, Z, 0) \mid U = u) s'_m(u). \end{aligned}$$

As discussed in Bickel (1986) and Klaassen (1992), the solution Δ_{0m} to (13) is unique and solves uniquely the integral equation,

$$\Delta(u) + \int_0^1 K(u, s)\alpha_m(s)\Delta(s) ds - \int_0^1 K(u, s)\gamma_m(s) ds = 0 \quad (14)$$

where $K(u, s) = s \wedge u - su$. Note that $\alpha > 0$ and let

$$\psi_{0m}(u) \equiv \alpha_m^{1/2}(u)\Delta_{0m}(u) \quad (15)$$

$$r_m(u) \equiv \alpha_m^{1/2}(u) \int_0^1 K(u, v)\gamma_m(v) dv \quad (16)$$

Then, equivalently,

$$L_m\psi_{0m} = r_m \quad (17)$$

where $L_m : L_2(0, 1) \rightarrow L_2(0, 1)$ is the operator $I + K_m$, for I the identity and K_m the bounded self-adjoint operator,

$$K_m(\chi)(u) = \int_0^1 \alpha_m^{1/2}(s)K(s, u)\alpha_m^{1/2}(u)\chi(s) ds.$$

The operators L_m have minimal eigenvalue ≥ 1 so that $\|L_m^{-1}\| \leq 1$.

Lemma 3 *If Δ_{0m}, ψ_{0m} are defined by (13), (17) and $\psi_0 \equiv \alpha^{1/2}\Delta_0$, then,*

$$\int (\psi_{0m} - \psi_0)^2(u) du \rightarrow 0 \quad (18)$$

$$\sup_m \|\Delta_{0m}\|_\infty < \infty \quad (19)$$

and, for every $\epsilon > 0$,

$$\sup\{|\Delta_{0m}(u) - \Delta_0(u)| : \epsilon \leq u \leq 1 - \epsilon\} \rightarrow 0 \quad (20)$$

$$\sup\{|\Delta'_{0m}(u) - \Delta'_0(u)| : \epsilon \leq u \leq 1 - \epsilon\} \rightarrow 0 \quad (21)$$

Proof: If L corresponds to α in the same way that L_m corresponds to α_m then

$$\|L_m - L\|^2 \leq \int_0^1 \int_0^1 K^2(s, u) \left((\alpha_m(s)\alpha_m(u))^{1/2} - (\alpha(s)\alpha(u))^{1/2} \right)^2 ds du.$$

The integrand converges to 0 and

$$\begin{aligned} & \int_0^1 \int_0^1 K^2(s, u) |\alpha_m(s)| |\alpha_m(u)| ds du \\ & \leq \int_0^1 \int_0^1 K^2(s, u) |\alpha(s)| |\alpha(u)| ds du \\ & \leq \left(\int_0^1 s(1-s) |\alpha(s)| ds \right)^2 < \infty \end{aligned}$$

for all m . So $\|L_m - L\| \rightarrow 0$ by Vitali's theorem. Since $\|L_m^{-1}\| \leq 1$,

$$\|L_m^{-1} - L^{-1}\| \rightarrow 0.$$

Next, let $r_m(\cdot) = r_{1m}(\cdot) + r_{2m}(\cdot) + r_{3m}(\cdot)$ where

$$\begin{aligned} r_{1m}(u) &= \alpha_m^{1/2}(u) \int_0^1 K(u, v) s_m(v) \gamma(v) dv \\ r_{2m}(u) &= \alpha_m^{1/2}(u) \iint \left((\lambda_{\theta u}(v, z, 0) + \lambda_\theta \lambda_u(v, z, 0)) \right. \\ &\quad \left. K(u, v)(1 - s_m(v)) g(v, z, 0) \right) dv dH(z) \\ r_{3m}(u) &= \alpha_m^{1/2}(u) \iint \lambda_\theta(v, z, 0) (1(v \leq u) - u) (1 - s_m(v)) \\ &\quad g(v, z, 0) dv dH(z) \end{aligned}$$

Now, by A3, $r_{1m} \rightarrow r$ a.e. where $r = \alpha^{1/2}(u) \int_0^1 K(u, s) \gamma(s) ds$, and, $r_{2m}, r_{3m} \rightarrow 0$ a.e. Further, by A3

$$\begin{aligned} |r_{1m}(u)| &\leq \alpha^{1/2}(u) \int_0^1 K(u, v) |s_m(v)| |\gamma(v)| dv \\ &\leq C \alpha^{1/2}(u) \left(\int_0^u \frac{1-u}{v^{1/2}(1-v)^{3/2}} dv + \int_u^1 \frac{u}{(1-v)^{1/2}v^{3/2}} dv \right) \\ &= 4C(\alpha(u)u(1-u))^{1/2} \end{aligned} \tag{22}$$

We conclude that $\int (r_{1m} - r)^2 \rightarrow 0$. Since $K(u, v) \leq u(1-u)$, it follows from A3 that $\int r_{2m}^2 \rightarrow 0$. Finally, $|r_{3m}| < (\alpha(u)u(1-u))^{1/2} \|\lambda_\theta\|_2$ and we obtain that

$$\int (r_m - r)^2 \rightarrow 0. \tag{23}$$

Finally, conclude by (22), (23),

$$L_m^{-1} r_m \rightarrow L^{-1} r$$

in L_2 , and (18) follows. Next, (19) follows from (14), since then

$$\begin{aligned} \|\Delta_{0m}\|_\infty &\leq \sup_u \left(\int_0^1 K^2(u, s) |\alpha_m(s)| ds \right)^{1/2} \left(\int_0^1 \psi_{0m}^2(s) ds \right)^{1/2} \\ &\quad + \int_0^1 s(1-s) |\gamma_m(s)| ds \end{aligned}$$

Finally, (20) and (21) follow from (18), (19) and (13) since, for $m > \epsilon^{-1}$,

$$\begin{aligned} \sup\{|\Delta''_{0m}(t)| : \epsilon \leq t \leq 1 - \epsilon\} &\leq \sup\{|\alpha(t)| : \epsilon \leq t \leq 1 - \epsilon\} \|\Delta_{0m}\|_\infty \\ &\quad + \sup\{|\gamma(t)| : \epsilon \leq t \leq 1 - \epsilon\}, \end{aligned}$$

so the families $\{\Delta'_{0m}(\cdot)\}$, $\{\Delta_{0m}(\cdot)\}$ are uniformly bounded and equicontinuous on $[\epsilon, 1 - \epsilon]$.

Proof: [of Theorem 2] Let

$$\begin{aligned} h_m(u, z, 0) &= \lambda_\theta(u, z, 0)s_m(u) + \Delta'_{0m}(u) + \lambda_u(u, z, 0)s_m(u)\Delta_{0m}(u) \\ &\quad - E\lambda_\theta(U, Z, 0)s_m(U) - E\Delta'_{0m}(U) \\ &\quad - E\lambda_u(U, Z, 0)s_m(U)\Delta_{0m}(U). \end{aligned}$$

By construction, for each m , the function h_m and all its derivatives with respect to u are bounded. Furthermore,

$$\begin{aligned} \frac{\partial h_m}{\partial u}(u, z, 0) &= \lambda_{\theta u}(u, z, 0)s_m(u) + \lambda_\theta(u, z, 0)s'_m(u) + \Delta''_{0m}(u) \\ &\quad + \lambda_{uu}(u, z, 0)s_m(u)\Delta_{0m}(u) \\ &\quad + \lambda_u(u, z, 0)(s'_m(u)\Delta_{0m}(u) + s_m(u)\Delta'_{0m}(u)). \end{aligned}$$

By (9) and (13),

$$E \left(\frac{\partial h_m}{\partial u}(U, Z, 0) \mid U = u \right) = 0.$$

Hence the Key Lemma applies and

$$n^{-1/2} \sum_{i=1}^n h_m(U_i, Z_i, 0) = n^{-1/2} \sum_{i=1}^n h_m \left(\frac{i}{n+1}, Z_{(i)}, 0 \right) + o_p(1).$$

To complete the proof of the theorem, we need only show that, for some sequence $m_n \rightarrow \infty$,

$$n^{-1/2} \sum_{i=1}^n (h_m(U_i, Z_i, 0) - v_{\Delta_0}(U_i, Z_i, 0)) = o_p(1)$$

or equivalently that, as $m \rightarrow \infty$,

$$\begin{aligned} E &\left(\lambda_\theta(U, Z, 0)(1 - s_m(U)) - (\Delta'_{0m} - \Delta'_0)(U) \right. \\ &\quad \left. - \lambda_u(U, Z, 0)(s_m\Delta_{0m}(U) - \Delta_0(U)) \right)^2 \rightarrow 0 \end{aligned} \tag{24}$$

Now by (4),

$$\begin{aligned} & E(\lambda_\theta(U, Z, 0) + \Delta'_0(U) + \lambda_u(U, Z, 0)\Delta_0(U))^2 \\ &= E\lambda_\theta^2(U, Z, 0) - E(\Delta'_0(U) + \lambda_u(U, Z, 0)\Delta_0(U))^2. \end{aligned} \quad (25)$$

On the other hand, a tedious calculation using integration by parts shows that,

$$\begin{aligned} & E(\lambda_\theta(U, Z, 0)s_m(U) + \Delta'_{0m}(U) + \lambda_u(U, Z, 0)\Delta_{0m}(U)s_m(U))^2 \\ &= E\lambda_\theta^2(U, Z, 0)s_m^2(U) - E(\Delta'_{0m}(U) + \lambda_u(U, Z, 0)\Delta_{0m}(U)s_m(U))^2 \\ &\quad - 2(E\lambda_\theta(U, Z, 0)\lambda_u(U, Z, 0)s_m(U)(1 - s_m(U)) \\ &\quad - E\lambda_u^2(U, Z, 0)\Delta_m^2(U)s_m(U)(1 - s_m(U))). \end{aligned} \quad (26)$$

The last two terms in (26) tend to 0 by dominated convergence in view of (19) and (25). Finally

$$\Delta'_{0m}(U) - \lambda_u(U, Z, 0)s_m(U) \xrightarrow{P} \Delta'_0(U) - \lambda_u(U, Z, 0)\Delta_0(U)$$

by (20) and (21) so that by Fatou's theorem and (25), (26),

$$\begin{aligned} & \liminf_m E(\lambda_\theta(U, Z, 0)s_m(U) + \Delta'_{0m}(U) + \lambda_u(U, Z, 0)\Delta_{0m}(U))^2 \\ & \leq E(\lambda_\theta(U, Z, 0) + \Delta'_0(U) + \lambda_u(U, Z, 0)\Delta_0(U))^2. \end{aligned} \quad (27)$$

Then, (24) and the theorem follow from (27).

Acknowledgments: We are grateful to Grace Yang and Chris Klaassen for some critical corrections. The research of both authors was partially supported by a US/Israel Bi-National Science Foundation Grant.

3.4 REFERENCES

- Bickel, P. J. (1986), Efficient testing in a class of transformation models, Report MS-R8614, Center for Mathematics and Computer Science, Amsterdam. Papers on semiparametric models at the ISI centennial session.
- Bickel, P. J., Klaassen, C. A. J., Ritov, Y. & Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press, Baltimore.
- Clayton, D. G. & Cuzick, J. (1986), The semi-parametric Pareto model for regression analysis of survival times, Report MS-R8614, Center for Mathematics and Computer Science, Amsterdam. Papers on semiparametric models at the ISI centennial session.

- Doksum, K. (1987), 'An extension of partial likelihood methods for proportional hazard models to general transformation models', *Annals of Statistics* **15**, 325–345.
- Hájek, J. & Šidák, Z. (1967), *Theory of Rank Tests*, Academic Press.
(Also published by Academia, the Publishing House of the Czechoslovak Academy of Sciences, Prague.).
- Klaassen, C. A. J. (1992), Efficient estimation in the Clayton-Cuzick model for survival data, Preprint, University of Amsterdam.
- Le Cam, L. & Yang, G. L. (1988), 'On the preservation of local asymptotic normality under information loss', *Annals of Statistics* **16**, 483–520.
- Stein, C. (1956), Efficient nonparametric testing and estimation, in J. Neyman, ed., 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 187–195.

4

From Model Selection to Adaptive Estimation

Lucien Birgé¹
Pascal Massart²

4.1 Introduction

Many different model selection information criteria can be found in the literature in various contexts including regression and density estimation. There is a huge amount of literature concerning this subject and we shall, in this paper, content ourselves to cite only a few typical references in order to illustrate our presentation. Let us just mention AIC, C_p or C_L , BIC and MDL criteria proposed by Akaike (1973), Mallows (1973), Schwarz (1978), and Rissanen (1978) respectively. These methods propose to select among a given collection of parametric models that model which minimizes an empirical loss (typically squared error or minus log-likelihood) plus some penalty term which is proportional to the dimension of the model. From one criterion to another the penalty functions differ by factors of $\log n$, where n represents the number of observations.

The reasons for choosing one penalty rather than another come either from information theory or Bayesian asymptotic computations or approximate evaluations of the risk on specific families of models. Many efforts were made to understand in what circumstances these criteria allow to identify the *right* model asymptotically (see Li (1987) for instance). Much less is known about the performances of the estimators provided by these methods from a *nonparametric point of view*. Let us consider the particular context of density estimation in \mathbb{L}^2 for instance. By a nonparametric point of view, we mean that the unknown density does not necessarily belong to any of the given models and that the best model should approximately realize the best trade-off between the risk of estimation within the model and the distance of the unknown density to the model. When the models have good approximation properties (following Grenander (1981) these models will be called sieves), an adequate choice of the penalty can produce *adaptive* estimators in the sense that they estimate a density of unknown

¹Université Paris VI and URA CNRS 1321

²Université Paris Sud and URA CNRS 743

smoothness at the rate which one would get if the degree of smoothness were known. Notable results in that direction have been obtained by Barron & Cover (1991) who use the MDL criterion when the models are chosen as ε -nets and by Polyak & Tsybakov (1990) who select the order of a Fourier expansion via Mallow's C_p for regression. One should also mention the results on penalized spline smoothing by Wahba and various coauthors (see Wahba (1990) for an extensive list of references).

This paper is meant to illustrate by a few theorems and applications, mainly directed towards adaptive estimation in Besov spaces, the power and versatility of the method of penalized minimum contrast estimation on sieves. A more general approach to the theory will be given in the companion paper Barron, Birgé & Massart (1995). We shall here content ourselves to consider linear sieves and the particular contrast which defines projection estimators for density estimation. These restrictions will allow us to make an extensive use of a recent and very powerful exponential inequality of Talagrand (1994) on the fluctuations of empirical processes which greatly simplifies the presentation and proofs. The choice of the penalty derives from the control of the risk on a fixed sieve. From that respect our approach presents some similarity with the method of structural minimization of the risk of Vapnik (1982). Minimum contrast estimators on a fixed sieve have been studied in great detail in Birgé & Massart (1994). For projection estimators their results can roughly be summarized as follows: s is an unknown density in $L^2(\mu)$ to be estimated using a projection estimator acting on a linear sieve S of dimension D and the loss function is proportional to the square of the distance induced by the norm. Under reasonable conditions on the structure of the space S one gets a quadratic risk of the order of $\|s - \pi(s)\|^2 + D/n$ if one denotes by $\pi(s)$ the projection of s on S . This is essentially the classical decomposition between the square of the bias and the variance. The presence of a D/n term corresponding to a D -dimensional approximating space is not surprising for those who are familiar with Le Cam's developments about the connections between the dimension (in the metric sense) of a space and the minimax risk on this space. One should see Le Cam (1973) and (1986, Chapter 16) for further details.

Our main purpose, in this paper, is to show that if we replace the single sieve S by a collection of linear sieves S_m , $m \in \mathcal{M}_n$, with respective dimensions D_m and suitable properties, and introduce a penalty function $\text{pen}(m)$ of the form $\mathcal{L}(m)D_m/n$, one gets a risk which, up to some multiplicative constant, realizes the best trade-off between $\|s - s_m\|^2$ and $\mathcal{L}(m)D_m/n$. Here s_m is the best approximant of s in S_m and $\mathcal{L}(m)$ is either uniformly bounded or possibly of order $\log n$ when too many of the sieves have the same dimension D_m . Note also that $\text{pen}(m)$ will be allowed to be random. We shall show that some more or less recently introduced methods of adaptive density estimation like the unbiased cross validation (Rudemo 1982), or the hard thresholding of wavelet empirical coefficients (Donoho, John-

stone, Kerkyacharian & Picard 1993) can be viewed as special instances of penalized projection estimators. In order to emphasize the flexibility and potential of the methods of penalization we shall play with different families of sieves and penalties and propose some new adaptive estimators especially in the context of wavelet expansions in nonhomogeneous Besov spaces and piecewise polynomials with non equally spaced knots.

4.2 The statistical framework

4.2.1 THE MODEL AND THE ESTIMATORS

We observe n i.i.d. random variables X_1, \dots, X_n with values on some measurable space \mathcal{X} and common density s with respect to some measure μ . We assume that s belongs to the Hilbert space $\mathbb{L}^2(\mu)$ with norm $\|\cdot\|$ and denote by $\|\cdot\|_p$ the norm in $\mathbb{L}^p(\mu)$ for $1 \leq p \leq \infty$ and $p \neq 2$. We first consider an N_n -dimensional linear subspace $\bar{\mathcal{S}}_n$ of $\mathbb{L}^2(\mu)$, then choose a finite family $\{\bar{\mathcal{S}}_m \mid m \in \mathcal{M}_n\}$ of linear subspaces of $\bar{\mathcal{S}}_n$, each $\bar{\mathcal{S}}_m$ being a D_m -dimensional subspace of $\mathbb{L}^2(\mu)$ and finally for each $m \in \mathcal{M}_n$ we take a convex subset $S_m \subset \bar{\mathcal{S}}_m$. In most cases, $S_m = \bar{\mathcal{S}}_m$. The set \mathcal{M}_n usually depends on n and more generally all the elements bearing a subscript (like m or m') which belong to \mathcal{M}_n . In order to keep the notations as simple as possible we shall systematically omit the subscript n when m is already present and also when the dependence on n is clear from the context. All real numbers that we shall introduce and which are not indexed by m or n are “fixed constants”. We shall also denote by \mathcal{S}_n the union of the S_m ’s, by s_m and \bar{s}_n the projections of s onto S_m and $\bar{\mathcal{S}}_n$ respectively, by \mathbb{P} the joint distribution of the observations X_i ’s when s obtains and by \mathbb{E} the corresponding expectation. The centered empirical operator ν_n on $\mathbb{L}^2(\mu)$ is defined by

$$\nu_n(t) = \frac{1}{n} \sum_{i=1}^n t(X_i) - \int_{\mathcal{X}} t(x)s(x)d\mu(x) \quad \text{for all } t \in \mathbb{L}^2(\mu).$$

Let us consider on $\mathcal{X} \times \mathcal{S}_n$ the contrast function $\gamma(x, t) = -2t(x) + \|t\|^2$ where $\|\cdot\|$ denotes the norm in $\mathbb{L}^2(\mu)$. The empirical version of this contrast is $\gamma_n(t) = (1/n) \sum_{i=1}^n \gamma(X_i, t)$. Minimizing $\gamma_n(t)$ over $\bar{\mathcal{S}}_m$ leads to the classical projection estimator \hat{s}_m on $\bar{\mathcal{S}}_m$ and we shall denote by \hat{s}_n the projection estimator on $\bar{\mathcal{S}}_n$. If $\{\varphi_\lambda\}_{\lambda \in \Lambda_m}$ is an orthonormal basis of $\bar{\mathcal{S}}_m$ one gets:

$$\hat{s}_m = \sum_{\lambda \in \Lambda_m} \hat{\beta}_\lambda \varphi_\lambda \quad \text{with} \quad \hat{\beta}_\lambda = \frac{1}{n} \sum_{i=1}^n \varphi_\lambda(X_i) \quad \text{and} \quad \gamma_n(\hat{s}_m) = - \sum_{\lambda \in \Lambda_m} \hat{\beta}_\lambda^2.$$

In order to define the penalty function, we associate to each S_m a weight $L_m \geq 1$. The use of those weights will become clear later but let us just

mention here that in the examples, either $L_m \asymp 1$ or $L_m \asymp \log n$ depending on the number of sieves with the same dimension D_m . For each $m \in \mathcal{M}_n$ the value of the penalty function $\text{pen}(m)$ is defined by

$$\text{pen}(m) = \tilde{K}_m(X_1, \dots, X_n) \frac{L_m D_m}{n} \quad (1)$$

where \tilde{K}_m is a positive random variable independent of the unknown s . Typically one must think of \tilde{K}_m as a fixed constant (independent of m and n) or as a random variable which is, with a large probability and uniformly with respect to m and n , bounded away from zero and infinity. Then, in both cases, $\text{pen}(m)$ is essentially proportional to $L_m D_m/n$.

A penalized projection estimator (PPE for short) is defined as any $\tilde{s} \in S_{\tilde{m}} \subset S_n$ such that

$$\gamma_n(\tilde{s}) + \text{pen}(\tilde{m}) = \inf_{m \in \mathcal{M}_n} \left(\inf_{t \in S_m} \gamma_n(t) + \text{pen}(m) \right) \quad \text{if } \tilde{s} \in S_{\tilde{m}}. \quad (2)$$

If such a minimizer does not exist one rather takes an approximate minimizer and chooses \tilde{s} satisfying

$$\gamma_n(\tilde{s}) + \text{pen}(\tilde{m}) \leq \inf_{m \in \mathcal{M}_n} \left(\inf_{t \in S_m} \gamma_n(t) + \text{pen}(m) \right) + \frac{1}{n}.$$

We shall assume in the sequel that (2) holds, the modifications needed to handle the extra $1/n$ term being straightforward.

In the sequel we shall distinguish between two different situations corresponding to different structures of the family of sieves: nested and non-nested. The nested situation can be described by the following assumption

N: Nested family of sieves *We assume that the integer N_n is given satisfying $N_n \leq n\Gamma^{-2}$ for some fixed constant Γ that $m \mapsto D_m$ is a one-to-one mapping, and that one of the two equivalent sets of assumptions holds:*

- (i) *$\|u\|_\infty \leq \Phi\sqrt{D_m}\|u\|$ for all $m \in \mathcal{M}_n$ and $u \in \bar{S}_m$ where Φ is a fixed constant and $D_m \leq N_n$ for all m . Moreover, $D_m < D_{m'}$ implies that $\bar{S}_m \subset \bar{S}_{m'}$ and $S_m \subset S_{m'}$;*
- (ii) *\bar{S}_n is a finite-dimensional subspace of $\mathbb{L}^2(\mu)$ with an orthonormal basis $\{\varphi_\lambda\}_{\lambda \in \bar{\Lambda}_n}$ and the cardinality of $\bar{\Lambda}_n$ is $|\bar{\Lambda}_n| = N_n$. A family of subsets $\{\Lambda_m\}_{m \in \mathcal{M}_n}$ of $\bar{\Lambda}_n$ with $|\Lambda_m| = D_m$ is given, \bar{S}_m is the linear span of $\{\varphi_\lambda\}_{\lambda \in \Lambda_m}$ and $\|\sum_{\lambda \in \Lambda_m} \varphi_\lambda^2\|_\infty \leq D_m \Phi^2$. Moreover for all m and m' , the inequality $D_m < D_{m'}$ implies that $S_m \subset S_{m'}$ and $\Lambda_m \subset \Lambda_{m'}$.*

The equivalence between (i) and (ii) follows from Lemma 6 of Birgé & Massart (1994). Assumption N will typically be satisfied when $\{\varphi_\lambda\}_{\lambda \in \bar{\Lambda}_n}$ is either a bounded basis or a localized basis in natural order. In this case, the

usual choices for \mathcal{M}_n will be either a finite subset of \mathbf{N} (and then $m \mapsto D_m$ is increasing) or a totally ordered family of sets. In some situations, the basis $\{\varphi_\lambda\}_{\lambda \in \bar{\Lambda}_n}$ is given (Fourier expansions for instance) from which one defines \bar{S}_m . In other cases (piecewise polynomials for instance) one starts with the family $\{\bar{S}_m\}_{m \in \mathcal{M}_n}$ which is the natural object to consider.

In the non-nested situation we shall distinguish a particular situation which is of special interest:

Case S: Non-nested subsets of a basis Let $\{\varphi_\lambda\}_{\lambda \in \bar{\Lambda}_n}$ be an orthonormal system in $\mathbb{L}^2(\mu)$ with $|\bar{\Lambda}_n| = N_n$ and \bar{S}_n be the linear span of $\{\varphi_\lambda\}_{\lambda \in \bar{\Lambda}_n}$. Each $m \in \mathcal{M}_n$ is a subset of $\bar{\Lambda}_n$ with cardinality D_m and $S_m = \bar{S}_m$ is the linear span of $\{\varphi_\lambda\}_{\lambda \in m}$.

Particular choices of \mathcal{M}_n and of the penalty function lead to various classical estimators. Here are three illustrations.

An analogue of Mallows' C_L

Assuming that **N** holds we define the penalty by $\text{pen}(m) = K\Phi^2 D_m/n$. This gives a sequence of parametric problems with an increasing number of parameters and a penalty proportional to the number of parameters. This is an analogue in density estimation of Mallows' C_L method for the regression framework—see for instance Mallows (1973) or Li (1987).

Cross-validation

Assume again that **N** holds. A particular choice of the penalty function leads to a well-known method of selecting the order of an expansion:

Proposition 1 Assume that we are in the nested situation described by Assumption **N** and that

$$\text{pen}(m) = \frac{2}{n(n+1)} \sum_{i=1}^n \sum_{\lambda \in \Lambda_m} \varphi_\lambda^2(X_i).$$

The resulting PPE \hat{s} is the projection estimator on $S_{\tilde{m}}$ where \tilde{m} is chosen by the unbiased cross-validation method.

Proof: Let us recall that the ideal m (in view of minimizing the quadratic loss) should minimize $\|s - \hat{s}_{m'}\|^2$ or equivalently $\int \hat{s}_{m'}^2 - 2 \int \hat{s}_{m'} s$ with respect to $m' \in \mathcal{M}_n$. Since this quantity involves the unknown s , it has to be estimated and the unbiased cross-validation method defines \hat{m} as the minimizer with respect to $m \in \mathcal{M}_n$ of

$$\int \hat{s}_m^2 d\mu - \frac{2}{n(n-1)} \sum_{i \neq i'} \sum_{\lambda \in \Lambda_m} \varphi_\lambda(X_i) \varphi_\lambda(X_{i'}).$$

Since

$$\int \hat{s}_m^2 d\mu = \frac{1}{n^2} \sum_{i, i'} \sum_{\lambda \in \Lambda_m} \varphi_\lambda(X_i) \varphi_\lambda(X_{i'})$$

one finds \hat{m} as the minimizer of

$$-\frac{n+1}{n-1} \int \hat{s}_m^2 d\mu + \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{\lambda \in \Lambda_m} \varphi_\lambda^2(X_i).$$

On the other hand the PPE selects \tilde{m} as the minimizer of

$$\begin{aligned} \gamma_n(\hat{s}_m) + \text{pen}(m) &= \int \hat{s}_m^2 d\mu - \frac{2}{n} \sum_{i=1}^n \hat{s}_m(X_i) + \text{pen}(m) \\ &= - \int \hat{s}_m^2 d\mu + \frac{2}{n(n+1)} \sum_{i=1}^n \sum_{\lambda \in \Lambda_m} \varphi_\lambda^2(X_i) \end{aligned}$$

which implies that $\hat{m} = \tilde{m}$ and the conclusion follows. \square

In this case, assuming that $L_m = 1$, the estimator $\tilde{K}_m(X_1, \dots, X_n)$ is given by

$$\frac{2}{n+1} \sum_{i=1}^n \frac{1}{m} \sum_{\lambda \in \Lambda_m} \varphi_\lambda^2(X_i).$$

Threshold estimators

We now consider the situation described by Case S, \mathcal{M}_n being the family of all (nonempty) subsets of $\bar{\Lambda}_n$ and $\text{pen}(m) = \tilde{L}_n D_m/n$ where \tilde{L}_n is a (possibly random) variable independent of m , we have to minimize over all possible subsets m of $\bar{\Lambda}_n$ the quantity

$$\gamma_n(\hat{s}_m) + \text{pen}(m) = - \sum_{\lambda \in m} \hat{\beta}_\lambda^2 + \frac{\tilde{L}_n D_m}{n} = - \sum_{\lambda \in m} \left(\hat{\beta}_\lambda^2 - \frac{\tilde{L}_n}{n} \right).$$

The solution \tilde{m} is the set of the λ 's such that $\hat{\beta}_\lambda^2 > \tilde{L}_n/n$ which leads to a threshold estimator introduced, in the context of white noise models, by Donoho & Johnstone (1994)

$$\tilde{s} = \sum_{\lambda \in \bar{\Lambda}_n} \hat{\beta}_\lambda \varphi_\lambda \mathbb{I}_{\{\hat{\beta}_\lambda^2 > \tilde{L}_n/n\}}.$$

These three examples are indeed typical of the two major types of model selection for projection estimators: selecting the order of an expansion or selecting a subset of a basis. We shall later give a formal treatment of these two problems.

4.2.2 BESOV SPACES AND EXAMPLES OF SIEVES

The target function s , in most of our illustrations, will be assumed to belong to some classical function spaces that we introduce below. We assume in this section that μ is Lebesgue measure.

Besov spaces

We shall consider here various forms of Besov spaces $B_{\alpha p \infty}(\mathcal{A})$ with $\alpha > 0$, $1 \leq p \leq \infty$, and three different types of supporting sets \mathcal{A} :

- Some compact interval which, without loss of generality, can be taken as $[0, 1]$ and then $\mathcal{A} = [0, 1]$;
- The torus \mathbb{T} which we shall identify to the interval $[0, 2\pi]$, then $\mathcal{A} = [0, 2\pi]$ and we deal with periodic functions;
- Some compact interval $[-A, A]$ ($\mathcal{A} = [-A, A]$) but in this case we shall consider it as the restriction of the Besov space on the whole real line to the set of functions which have a compact support in $(-A, A)$.

Let us first recall some known facts on Besov spaces which can be found in the books by DeVore & Lorentz (1993) or Meyer (1990). Following DeVore & Lorentz (1993, page 44) we define the r -th order differences of a function t defined on \mathcal{A} by

$$\Delta_h^r(t, x) = \sum_{k=0}^r \binom{r}{k} (-1)^{r-k} t(x + kh).$$

The Besov space $B_{\alpha p \infty}(\mathcal{A})$ will be the space of functions t on \mathcal{X} such that

$$\sup_{y>0} y^{-\alpha} \omega_r(t, y)_p < +\infty \quad \text{where} \quad \omega_r(t, y)_p = \sup_{0 < h \leq y} \|\Delta_h^r(t, \cdot)\|_p$$

and $r = [\alpha] + 1$ (DeVore & Lorentz 1993, page 55). As a particular case, we get the classical Hölder spaces when $p = \infty$. One should notice that since we always work on a compact interval \mathcal{A} , L^p -norms on \mathcal{A} are easy to compare and $\omega_r(t, y)_p \geq C(p) \omega_r(t, y)_2$ for $p \geq 2$. This implies that $B_{\alpha p \infty}(\mathcal{A}) \subset B_{\alpha 2 \infty}(\mathcal{A})$. Therefore, if $p \geq 2$ we can restrict ourselves to considering only the larger space $B_{\alpha 2 \infty}(\mathcal{A})$ since we are looking for upper bounds for the risk.

Wavelet expansions

Let us consider an orthonormal wavelet basis $\{\varphi_{j,k} | j \geq 0, k \in \mathbb{Z}\}$ of $L^2(\mathbb{R}, dx)$ (see Meyer (1990) for details) with the following conventions: $\varphi_{0,k}$ are translates of the father wavelet and for $j \geq 1$, the $\varphi_{j,k}$'s are affine transforms of the mother wavelet. One will also assume that these wavelets are compactly supported and have *regularity* r in the following sense: all their moments up to order r are 0. Let $t \in L^2(\mathbb{R}, dx)$ be some function with compact support in $(-A, A)$. Changing the indexing of the basis if necessary we can write the expansion of t on the wavelet basis as:

$$t = \sum_{j \geq 0} \sum_{k=1}^{2^j M} \beta_{j,k} \varphi_{j,k}, \tag{3}$$

where $M \geq 1$ is a finite integer depending on A and the lengths of the wavelet's supports. For any $j \in \mathbb{N}$, we denote by $\Lambda(j)$ the set of indices $\{(j, k) \mid 1 \leq k \leq 2^j M\}$ and if $m \subset \Lambda = \sum_{j \geq 0} \Lambda(j)$ we put $m(j) = m \cap \Lambda(j)$. Let B_0 denote the space of functions t such that $\Sigma_\infty(t) = \sum_{j \geq 0} 2^{j/2} \sup_{\lambda \in \Lambda(j)} |\beta_\lambda| < +\infty$. From Bernstein's inequality (Meyer 1990, Chapter 2, Lemma 8)

$$\|t\|_\infty \leq \Phi_\infty \Sigma_\infty(t) \quad \text{for all } t \in B_0, \quad (4)$$

where Φ_∞ only depends on the choice of the basis. We also define \bar{V}_J , for $J \in \mathbb{N}$, to be the linear span of $\{\varphi_\lambda \mid \lambda \in \Lambda(j), 0 \leq j \leq J\}$; then $2^J M \leq \text{Dim}(\bar{V}_J) = N < 2^{J+1} M$ and it follows from (4) that there exists a constant Φ , namely $\Phi^2 = 2\Phi_\infty^2/M$, such that

$$\|t\|_\infty \leq \Phi \sqrt{N} \|t\| \quad \text{for all } t \in \bar{V}_J. \quad (5)$$

Let t be given by (3) with $\alpha < r + 1$; if t belongs to the Besov space $B_{\alpha p \infty}([-A, A])$ then (Kerkyacharian & Picard 1992)

$$\sup_{j \geq 0} 2^{j(\alpha + \frac{1}{2} - \frac{1}{p})} \left(\sum_{\lambda \in \Lambda(j)} |\beta_\lambda|^p \right)^{1/p} = \|t\| < +\infty. \quad (6)$$

One derives from equation (6) that $\sup_{\lambda \in \Lambda(j)} |\beta_\lambda| \leq 2^{-j(\alpha + \frac{1}{2} - \frac{1}{p})} \|t\|$ which proves the inclusion $B_{\alpha p \infty}([-A, A]) \subset B_0$ provided that $\alpha > 1/p$.

Piecewise polynomials

Without loss of generality we shall restrict our attention to piecewise polynomial spaces on $[0, 1]$. A linear space S_m of piecewise polynomials is characterized by $m = (r, \{b_0 = 0 < b_1 < \dots < b_D = 1\})$ where r is the maximal degree of the polynomials (that we shall essentially keep fixed in the sequel) and $\{b_0 = 0 < b_1 < \dots < b_D = 1\}$ is a nondecreasing sequence which generates a partition of $[0, 1]$ into D intervals. Such a space has the dimension $D_m = D(r+1)$. We shall distinguish between regular piecewise polynomials for which all intervals have the same length and general piecewise polynomials with arbitrary intervals subject to the restriction that their lengths are multiples of a fixed value $1/N$ with $N \in \mathbb{N}$. In this case the b_j 's are of the form N_j/N where the N_j 's are integers and the corresponding set of m 's will be denoted by \mathcal{P}_N^r . The reasons for restricting the values of the b_j 's to a grid are given in Birgé & Massart (1994, Section 3) and we shall not insist on that. When dealing with regular partitions, we shall restrict, in order to get a nested family of sieves, to dyadic partitions generated by the grid $\{j2^{-J_m}, 0 \leq j \leq 2^{J_m}\}$ where J_m is an integer. The corresponding set of m 's for $0 \leq J_m \leq J$ will be denoted by $\tilde{\mathcal{P}}_J^r$.

We shall need hereafter some properties of these spaces of polynomials. Let us first recall (see Whittaker & Watson (1927, pp. 302-305) for details) that the Legendre polynomials $Q_j, j \in \mathbb{N}$ are a family of orthogonal polynomials in $\mathbb{L}^2([-1, 1], dx)$ such that Q_j has degree j and

$$|Q_j(x)| \leq 1 \quad \text{for all } x \in [-1, 1], \quad Q_j(1) = 1, \quad \int_{-1}^1 Q_j^2(t) dt = \frac{2}{2j+1}.$$

As a consequence, the family of polynomials $R_j(x) = \sqrt{2j+1}Q_j(2x-1)$ is an orthonormal basis for the space of polynomials on $[0, 1]$ and if H is a polynomial of degree r such that $H(x) = \sum_{j=0}^r a_j R_j(x)$,

$$|H(x)|^2 \leq \left(\sum_{j=0}^r a_j^2 \right) \left(\sum_{j=0}^r 2j+1 \right) = (r+1)^2 \sum_{j=0}^r a_j^2.$$

Hence $\|H\|_\infty \leq (r+1)\|H\|$. Therefore any polynomial H of degree r on an interval $[a, b]$ satisfies $\|H\|_\infty \leq (r+1)(b-a)^{-1/2}\|H\|$ from which one deduces that for $H \in S_m$

$$\|H\|_\infty \leq \frac{r+1}{\sqrt{h}} \|H\| \quad \text{where } h = \inf_{1 \leq j \leq D} \{b_j - b_{j-1} \mid b_j > b_{j-1}\}. \quad (7)$$

Therefore, if s is a function on $[a, b]$ and H_s its projection on the space of polynomials of degree $\leq r$ on $[a, b]$, one gets

$$\|H_s\|_\infty \leq \frac{r+1}{(b-a)^{1/2}} \|H_s\| \leq \frac{r+1}{(b-a)^{1/2}} \|s\| \leq (r+1)\|s\|_\infty \quad (8)$$

and this inequality remains true for the projections on spaces of piecewise polynomials since it only depends on the degree and not on the support.

4.3 Presentation of some of the results

From now on, we shall have to introduce various constants to set up the assumptions, describe the penalty function, state the results and produce the proofs. In order to clarify the situation we shall stick to some fixed conventions and give to the letters κ, C (or c) and K , with various sub- or superscripts, a special meaning. The constants used to set up the assumptions will be denoted by various letters but the three letters above will be reserved. κ_1, \dots denote universal (numerical) constants which are kept fixed throughout the paper. K, K', \dots are constants to be chosen by the statistician or to be used as generic constants in some assumptions. Finally C, c, C', \dots denote constants which come out from the computations and proofs and depend on the other constants given in the assumptions. One shall also use $C(\cdot, \cdot, \dots)$ to indicate more precisely the dependence on

various quantities and especially those which are related to the unknown s . The value of K or C is fixed throughout a proof but, in order to keep the notations simple, we shall use the same notation for different constants when one goes from one proof or example to another.

Before giving the formal results let us describe a few typical and illustrative examples (more will be given later) of applications of these results together with a sketch of proof in order to make them more appealing. We shall distinguish between the two situations described above: nested and non-nested.

4.3.1 NESTED MODELS

We assume that \mathbf{N} holds and $S_m = \bar{S}_m$ and we choose either a deterministic or a random penalty function of the form

$$\text{pen}(m) = K\Phi^2 \frac{D_m}{n} \quad \text{or} \quad \text{pen}(m) = \frac{K}{n(n+1)} \sum_{i=1}^n \sum_{\lambda \in \Lambda_m} \varphi_\lambda^2(X_i)$$

where K is a suitably chosen constant. We recall from Section 4.2.1 that the choice $K = 2$ corresponds to Mallows' C_L or cross-validated estimators. We shall prove below that under such assumptions, one gets, as expected

$$\mathbb{E}[\|\tilde{s} - s\|^2] \leq C \inf_{m \in \mathcal{M}_n} [\|s_m - s\|^2 + D_m/n].$$

Assuming that the true s belongs to some unknown Besov space $B_{\alpha, 2, \infty}$ with $\alpha > 0$ and choosing a convenient basis with good approximation properties with respect to such spaces (wavelet basis, dyadic splines or Fourier basis), we shall get the usual and optimal $n^{-\alpha/(2\alpha+1)}$ rate of convergence for our penalized estimator (see Example 1 below).

Remarks: The constant \sqrt{K} should be larger than some universal constant involved in some suitable exponential inequality. A reasonable conjecture (by analogy with the gaussian case) is that a lower bound for K is one which means that our results should hold for the classical cross-validated estimators.

4.3.2 SELECTING A SUBSET OF A WAVELET BASIS

We consider the wavelet basis of regularity r and the notations introduced in Section 4.2.2 and assume that Case \mathbf{S} obtains with $\bar{\Lambda}_n = \sum_{0 \leq j \leq J_n} \Lambda(j)$ where J_n is given by $2^{J_n} \asymp n/(\log^2 n)$. Then the dimension N_n of \bar{S}_n satisfies $2^{J_n} M < N_n < 2^{J_n+1} M$ and $D_m = |m|$.

Thresholding

\mathcal{M}_n is taken to be *all* the subsets of $\bar{\Lambda}_n$ and the penalty function is given by $K(\Phi_\infty \Sigma_\infty(\hat{s}_n) + K')(\log n)|m|/n$. As mentioned in Section 4.2.1, the PPE

is then a threshold estimator. We recall that \hat{s}_n is the projection estimator on the largest sieve $\bar{\mathcal{S}}_n$. It comes from (4) that $\Phi_\infty \Sigma_\infty(\hat{s}_n)$ is an estimator of an upper bound of $\|s\|_\infty$ provided that s belongs to B_0 . In this case we shall prove that

Proposition 2 *Let \tilde{s} be the threshold estimator given by*

$$\tilde{s} = \sum_{\lambda \in \Lambda_n} \hat{\beta}_\lambda \varphi_\lambda \mathbb{I}_{\{\hat{\beta}_\lambda^2 > \tilde{T}\}}, \quad \text{with } \tilde{T} = K(\Phi_\infty \Sigma_\infty(\hat{s}_n) + K') \log n / n$$

where K has to be larger than a universal constant and K' is an arbitrary positive number. Provided that s belongs to B_0 , the following upper bound holds for any $q \geq 1$,

$$\mathbb{E}[\|\tilde{s} - s\|^q] \leq C \inf_{m \in \mathcal{M}_n} \left[\|s_m - s\|^2 + \log n \frac{D_m}{n} \right]^{q/2} \quad (9)$$

as soon as

$$\Phi_\infty [\Sigma_\infty(s) - \Sigma_\infty(\hat{s}_n)] \leq K'. \quad (10)$$

Either one knows an upper bound for $\Phi_\infty \Sigma_\infty(s)$ and one should choose K' to be this upper bound or (10) will hold only for n large enough. Assuming that s belongs to some unknown Besov space $B_{\alpha,p,\infty}$, with $r+1 > \alpha > 1/p$ (and therefore $s \in B_0$) the resulting rate of convergence is $(\log n/n)^{\alpha/(2\alpha+1)}$. There is an extra power of $\log n$ in the rate but it should be noticed that (9) holds for a set of densities which is larger than the Besov spaces. With a different thresholding strategy, the same rates have been obtained by Donoho et al. (1993).

Special strategy for Besov spaces

We introduce a smaller family of sieves which has the same approximation properties in the Besov spaces than the previous one. It can be described as follows. Let us first introduce an auxiliary positive and decreasing function l defined on $[1, +\infty)$ with $l(1) < 1$. For each pair of integers J, j' with $0 \leq j' \leq J$, let $\mathcal{M}_J^{j'}$ be the collection of subsets m of Λ such that $m(j) = \Lambda(j)$ for $0 \leq j \leq j'$ and $|m(j)| = [\Lambda(j)|l(j-j')|]$ for $j' < j \leq J$, where $[x]$ denotes the integer part of x . We define $\mathcal{M}_n = \sum_{0 \leq j' \leq J_n} \mathcal{M}_{J_n}^{j'}$ and $\text{pen}(m) = K(\Phi_\infty \Sigma_\infty(\hat{s}_n) + K')L|m|/n$ where K will be larger than a universal constant, K' is an arbitrary positive number and L is a fixed weight depending on l . The resulting penalized estimator \tilde{s} will then satisfy

Proposition 3 *Let s belong to B_0 and (10) be satisfied. If the function l is such that $2^{-j'} \log |\mathcal{M}_{J_n}^{j'}|$ is bounded by a fixed constant then, for any $q \geq 1$,*

$$\mathbb{E}[\|\tilde{s} - s\|^q] \leq C \inf_{m \in \mathcal{M}_n} \left[\|s_m - s\|^2 + \frac{D_m}{n} \right]^{q/2}.$$

We shall see in Example 4 that one can choose l satisfying the required conditions and such that the resulting bias leads to the right rate of convergence $n^{-\alpha/(2\alpha+1)}$ for all Besov spaces $B_{\alpha p \infty}$, with $r+1 > \alpha > 1/p$ simultaneously.

4.3.3 VARIABLE WEIGHTS AND PIECEWISE POLYNOMIALS

Up to now we only considered situations where the weights L_m did not depend on m . The following example is meant to illustrate the advantage of letting the weights vary with the models in some cases. It deals with piecewise polynomials. Let us fix some maximal degree r for our polynomials and take $2^{J_n} \asymp n/\log^2 n$. We consider the family of sieves $\{S_m\}_{m \in \mathcal{M}_n}$ where $\mathcal{M}_n = \mathcal{P}_{2^{J_n}}^r$. The S_m 's are the corresponding piecewise polynomials of degree $\leq r$ described in Section 4.2.2. One should notice that since $\bar{\mathcal{P}}_{J_n}^r \subset \mathcal{M}_n$, this family of sieves includes in particular piecewise polynomials based on regular dyadic partitions. Let us define $L_m = 1$ when $m \in \bar{\mathcal{P}}_{J_n}^r$ and $L_m = \log n$ otherwise. In this situation, it is wiser to choose $\bar{\mathcal{S}}_n$ as the space of piecewise polynomials based on the finest possible partition generated by the sequence $\{j2^{-J_n}\}_{0 \leq j \leq 2^{J_n}}$ and with degree $2r$ instead of r ; then $N_n = (2r+1)2^{J_n}$. With such a choice the squares of the elements of all the sieves will belong to $\bar{\mathcal{S}}_n$.

Proposition 4 *Let us choose $\text{pen}(m) = K(\|\hat{s}_n\|_\infty + K')L_m D_m/n$ and assume that s is bounded, then the PPE \tilde{s} satisfies, for any $q \geq 1$,*

$$\mathbb{E}[\|\tilde{s} - s\|^q] \leq C \inf_{m \in \mathcal{M}_n} \left[\|s - s_m\|^2 + \frac{L_m D_m}{n} \right]^{q/2}.$$

For an arbitrary s , the method hunts for a partition which provides, up to a $\log n$ factor, the best trade-off between the dimension of the partition and the bias. But if s belongs to some Besov space $B_{\alpha 2 \infty}$ with $\alpha < r+1$, then the estimator achieves the optimal rate of convergence $n^{-\alpha/(2\alpha+1)}$.

4.3.4 SKETCH OF THE PROOFS

In order to prove results of the form

$$\mathbb{E}[\|s - \tilde{s}\|^q] \leq C_0 \inf_{m \in \mathcal{M}_n} \left[\|s - s_m\|^2 + L_m D_m/n \right]^{q/2}$$

we always follow the same basic scheme (with various additional technicalities). m is defined as the minimizer with respect to $m' \in \mathcal{M}_n$ of $\|s_{m'} - s\|^2 + L_{m'} D_{m'}/n$. Using a powerful result of Talagrand (1994), we begin to prove that with probability larger than $1 - p_{m'} \exp(-c\sqrt{\xi})$, for any $m' \in \mathcal{M}_n$ and uniformly for $t \in S_{m'}$

$$\nu_n(t - s_m) \leq \frac{1}{4} (\|t - s\|^2 + \|s_m - s\|^2) + \frac{\xi}{n} + C \left[\frac{L_{m'} D_{m'}}{n} + \frac{L_m D_m}{n} \right]. \quad (11)$$

By assumption, the $L_{m'}$'s are chosen in such a way that $\sum_{m' \in \mathcal{M}_n} p_{m'} \leq C_1$ which implies that the control (11) holds for all m' simultaneously with probability larger than $1 - C_1 \exp(-c\sqrt{\xi})$. In particular (11) holds with $t = \tilde{s}$ and $m' = \tilde{m}$. We then use the following simple lemma:

Lemma 1 *Let $\tilde{s} = \hat{s}_{\tilde{m}}$ be the PPE associated with the penalty function $\text{pen}(\cdot)$, m a given element of \mathcal{M}_n and s_m the projection of the true underlying density s onto S_m . The following inequality holds:*

$$\|s - \tilde{s}\|^2 \leq \|s - s_m\|^2 + \text{pen}(m) - \text{pen}(\tilde{m}) + 2\nu_n(\tilde{s} - s_m). \quad (12)$$

Proof: The conclusion follows from the fact that $\gamma_n(\tilde{s}) + \text{pen}(\tilde{m}) \leq \gamma_n(s_m) + \text{pen}(m)$ and the following inequalities:

$$\gamma_n(t) = \|t\|^2 - 2\nu_n(t) - 2 \int tsd\mu \quad \text{for all } t \in \mathcal{S}_n;$$

$$\begin{aligned} \|s - t\|^2 - \|s - s_m\|^2 &= \|t\|^2 - \|s_m\|^2 + 2 \int (s_m - t)s d\mu \\ &= \mathbb{E}(\gamma_n(t) - \gamma_n(s_m)). \end{aligned} \quad \square$$

Using (11) and (12) simultaneously we get

$$\|s - \tilde{s}\|^2 \leq 3\|s - s_m\|^2 + 2[\text{pen}(m) - \text{pen}(\tilde{m})] + \frac{4\xi}{n} + 4C \left[\frac{L_{\tilde{m}} D_{\tilde{m}}}{n} + \frac{L_m D_m}{n} \right].$$

If $\text{pen}(m')$ is defined in such a way that for all $m' \in \mathcal{M}_n$,

$$2C \frac{L_{m'} D_{m'}}{n} \leq \text{pen}(m') \leq K \frac{L_{m'} D_{m'}}{n},$$

one gets with probability larger than $1 - C_1 \exp(-c\sqrt{\xi})$

$$\|s - \tilde{s}\|^2 \leq 3\|s - s_m\|^2 + (4C + 2K) \frac{L_m D_m}{n} + \frac{4\xi}{n}.$$

One concludes using the following elementary lemma since $L_m D_m \geq 1$.

Lemma 2 *Let X be a nonnegative random variable satisfying $X^2 \leq a + K_1 t/n$ with probability larger than $1 - K_2 \exp(-K_3 \sqrt{t})$ for all $t > 0$. Then for any number $q \geq 1$*

$$\mathbb{E}[X^q] \leq 2^{(q/2-1)^+} \left[a^{q/2} + K_2 \Gamma(q+1) \left(\frac{K_1}{n K_3} \right)^{q/2} \right].$$

The case of a random penalty requires extra arguments but the basic ideas are the same.

4.4 The theorems

4.4.1 TALAGRAND'S THEOREM

All our results rely upon an important theorem of Talagrand (1994) which can be considered, if stated in a proper form, as an analogue of an inequality for Gaussian processes by Cirel'son, Ibragimov & Sudakov (1976). Let us first recall this inequality in the case of a real-valued non-centered process in order to emphasize the similarity between the two results.

Theorem 1 *Let $X_t, t \in T$ be a real valued gaussian process with bounded sample paths and $v = \sup_t \text{Var}(X_t)$. Then for $\xi > 0$*

$$\begin{aligned} \mathbb{P} \left[\sup_t (X_t - \mathbb{E}[X_t]) \geq \mathbb{E} \left[\sup_t (X_t - \mathbb{E}[X_t]) \right] + \xi \right] \\ \leq \frac{2}{\sqrt{2\pi v}} \int_{\xi}^{+\infty} e^{-x^2/(2v)} dx \leq \exp \left[-\frac{1}{2} \frac{\xi^2}{v} \right]. \end{aligned}$$

Although Talagrand did not state his theorem³ in such a form one can actually write it as follows:

Theorem 2 *Let X_1, \dots, X_n be n i.i.d. random variables, $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. random signs (+1 or -1 with probability 1/2) independent of the X_i 's and $\{f_t, t \in T\}$ a family of functions that are uniformly bounded by some constant b . Let $v = \sup_{t \in T} \text{Var}(f_t(X_1))$. There exists universal constants $\kappa_2 \geq 1$ and κ_1 such that for any positive ξ*

$$\begin{aligned} \mathbb{P} \left[\sup_t \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n f_t(X_i) - \mathbb{E}[f_t(X_i)] \right) \geq \kappa_2 \mathbb{E} \left[\sup_t \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f_t(X_i) \right| \right] + \xi \right] \\ \leq \exp \left[-\kappa_1 \left(\frac{\xi^2}{v} \wedge \frac{\xi \sqrt{n}}{b} \right) \right]. \end{aligned} \quad (13)$$

In the sequel κ_1 and κ_2 will always denote the two constants appearing in (13).

Talagrand's Theorem has many useful statistical consequences, such as the following extension of a result from Mason & van Zwet (1987): the control of χ^2 -type statistics $\mathcal{K}_{n,D}^2 = \sum_{j=1}^D (X_j - np_j)^2 / np_j$ with (X_1, \dots, X_k) a multinomial random vector with distribution $\mathcal{M}(n, p_1, \dots, p_k)$ and $D \leq k$. If $\delta = \inf_{1 \leq j \leq D} p_j$, $x > 0$ and $\varepsilon > 0$, the following inequality holds:

$$\mathbb{P} [\mathcal{K}_{n,D}^2 \geq (1 + \varepsilon) \kappa_2^2 D + x] \leq 2 \exp \left[\frac{-\kappa_1 \varepsilon x}{1 + \varepsilon} \left(1 \wedge \sqrt{\frac{n\delta}{x}} \right) \right]. \quad (14)$$

³During the final revision of our article we became aware of improvements by Talagrand (1995) and Ledoux (1995) of Theorem 2 that might lead to more explicit lower bounds for our penalty functions.

This inequality implies Mason and van Zwet's inequality (Mason & van Zwet 1987, Lemma 3) when $x \leq n\delta$ and provides more information on the tail distribution of $\mathcal{K}_{n,D}^2$ since it holds without any restriction on x . The proof is given in Section 4.6.

4.4.2 SELECTING THE ORDER OF AN EXPANSION

In this section, we shall restrict ourselves to the nested case. The first result deals with the analogue of Mallows' C_L .

Theorem 3 *Assume that **N** holds, choose some positive θ and define the penalty function by $\text{pen}(m) = (\kappa_2^2 \Phi^2 + \theta) D_m/n$. Then for any $q \geq 1$*

$$\mathbb{E}[\|\tilde{s} - s\|^q] \leq C(q, \|s\|, \Phi, \Gamma, \theta) \inf_{m \in \mathcal{M}_n} \left[\frac{D_m}{n} + \|s_m - s\|^2 \right]^{q/2}. \quad (15)$$

In view of Proposition 1, the following theorem applies to cross-validated projection estimators provided that the conjecture $\kappa_2 = 1$ is true, but cross-validation would also make sense with different values of the constant K_n in (16) below.

Theorem 4 *Assume that **N** is satisfied, that S_m is the linear span of $\{\varphi_\lambda\}_{\lambda \in \Lambda_m}$ for all m 's, and that*

$$\inf_{m \in \mathcal{M}_n} \frac{1}{D_m} \int \left(\sum_{\lambda \in \Lambda_m} \varphi_\lambda^2 \right) s d\mu = a > 0;$$

$$\text{pen}(m) = \frac{K_n}{n(n+1)} \sum_{i=1}^n \sum_{\lambda \in \Lambda_m} \varphi_\lambda^2(X_i) \quad \text{with} \quad K_n = \frac{n+1}{n}(\kappa_2^2 + \theta) \quad (16)$$

for some positive θ . Then for any $q \geq 1$

$$\mathbb{E}[\|\tilde{s} - s\|^q] \leq C(q, \|s\|, \Phi, \Gamma, a, \theta) \inf_{m \in \mathcal{M}_n} \left[\frac{D_m}{n} + \|s_m - s\|^2 \right]^{q/2}. \quad (17)$$

4.4.3 EXTENSION

In order to analyze some particular types of estimators which were first proposed by Efroimovich (1985) (see Example 2 below), it is useful to have some more general version of Theorem 3. \bar{S}_n is a finite-dimensional space with an orthonormal basis $\{\varphi_\lambda\}_{\lambda \in \bar{\Lambda}_n}$ and $|\bar{\Lambda}_n| = N_n$. Let $\bar{\mathcal{M}}_n$ be a finite collection of sets (not necessarily totally ordered by inclusion but having a maximal element \bar{m}_n) and $m' \mapsto \Lambda_{m'}$ be an increasing mapping from $\bar{\mathcal{M}}_n$ into the family of nonvoid subsets of $\bar{\Lambda}_n$. We define $S_{m'}$ as the linear span of $\{\varphi_\lambda\}_{\lambda \in \Lambda_{m'}}$; then $D_{m'} = |\Lambda_{m'}|$. Let \mathcal{M}_n be a totally ordered subfamily of $\bar{\mathcal{M}}_n$ containing \bar{m}_n and for which Assumption **N** holds. One defines for

each $m' \in \bar{\mathcal{M}}_n$ an associate $\tau(m') \in \mathcal{M}_n$ which is the smallest m such that $m \supset m'$. Assuming that the penalty function satisfies the inequality

$$(\kappa_2^2 \Phi^2 + \theta) D_{\tau(m')} / n \leq \text{pen}(m') \leq K D_{\tau(m')} / n,$$

it is easily seen that the bound (15) still holds for the PPE \tilde{s} based on the larger family of sieves $\{S_m\}_{m \in \bar{\mathcal{M}}_n}$, where s_m is defined as before, since the proof only involves the larger spaces $S_{\tau(m')}$ and the values of the penalty function. If we assume that for each $m' \in \bar{\mathcal{M}}_n$

$$D_{m'} \geq \delta D_{\tau(m')} \quad (18)$$

for some fixed positive constant δ , the penalty $\text{pen}(m') = \delta^{-1}(\kappa_2^2 \Phi^2 + \theta) D_{m'} / n$ will satisfy the above inequality and since $\|s_{m'} - s\| \geq \|s_{\tau(m')} - s\|$ the following bound remains valid:

$$\mathbb{E}[\|\tilde{s} - s\|^q] \leq \delta^{-q/2} C(q, \|s\|, \Phi, \Gamma, \theta) \inf_{m' \in \bar{\mathcal{M}}_n} \left[\frac{D_{m'}}{n} + \|s_{m'} - s\|^2 \right]^{q/2}.$$

4.4.4 SELECTING A SUBSET OF A BASIS

Let us now study the more general situation of a rich and possibly non-nested family of sieves. We shall use the assumption

B: $R_n(s) = \sup_{t \in \mathcal{S}_n} \|t\|^{-2} \int t^2 s d\mu$ is finite and there exists a family of weights $L_m \geq 1$, $m \in \mathcal{M}_n$ and a fixed constant Δ such that

$$\sum_{m \in \mathcal{M}_n} \exp(-L_m D_m) \leq \Delta. \quad (19)$$

Our first theorem deals with a bounded situation where $S_m \neq \bar{S}_m$.

Theorem 5 Assume that $\|t\|_\infty \leq B_n$ for all $t \in \mathcal{S}_n$ and that **B** holds with $R_n(s) \leq B_n$, $\text{pen}(m)$ being possibly random. Then, for any $q \geq 1$,

$$\begin{aligned} & \mathbb{E}[\|\tilde{s} - s\|^q \mathbb{I}_{\tilde{\Omega}}] \\ & \leq C(q) \left[\inf_{m \in \mathcal{M}_n} \left[\|s - s_m\|^q + \mathbb{E}[(\text{pen}(m))^{q/2} \mathbb{I}_{\tilde{\Omega}}] \right] + \Delta (B_n/n)^{q/2} \right] \end{aligned}$$

if $\tilde{\Omega}$ is defined by

$$\tilde{\Omega} = \{\text{pen}(m) \geq \kappa_1^{-1} (3 + 5\sqrt{2}\kappa_2 + 4\kappa_2^2) B_n L_m D_m / n \text{ for all } m \in \mathcal{M}_n\}.$$

The boundedness restrictions on \mathcal{S}_n and $R_n(s)$ being rather unpleasant we would like to be dispensed with them. A more general situation can be handled in the following way. We recall that \tilde{s}_n is the projection of s on $\tilde{\mathcal{S}}_n$, \hat{s}_n the projection estimator defined on $\tilde{\mathcal{S}}_n$ and t_m the projection of t on S_m .

Theorem 6 Assume that **B** holds, μ is a finite measure and there exists some orthonormal basis $\{\varphi_\lambda\}_{\lambda \in \bar{\Lambda}_n}$ of $\bar{\mathcal{S}}_n$ which satisfies $\|\varphi_\lambda\|_\infty \leq \Phi \sqrt{N_n}$ for all $\lambda \in \bar{\Lambda}_n$ with $N_n = |\bar{\Lambda}_n| = n/(\theta_n \log n)$ where Φ is a fixed constant and $\{\theta_k\}_{k \geq 1}$ a sequence converging to infinity. Suppose that a real function ψ is given on $\bar{\mathcal{S}}_n$ such that for all $t \in \bar{\mathcal{S}}_n$ and $m \in \mathcal{M}_n$, $\|t_m\|_\infty \leq \psi(t)$ and

$$|\psi(\bar{s}_n) - \psi(\hat{s}_n)| \leq \Phi' \sqrt{N_n} \sup_{\lambda \in \bar{\Lambda}_n} |\nu_n(\varphi_\lambda)|. \quad (20)$$

Let us define the penalty function by

$$\text{pen}(m) = K \frac{L_m D_m}{n} (\psi(\hat{s}_n) + K'), \quad \text{with } K \geq \frac{2}{\kappa_1} (3 + 5\sqrt{2}\kappa_2 + 4\kappa_2^2) \quad (21)$$

where K' is a positive constant to be chosen by the statistician. Then, for any $q \geq 1$,

$$\begin{aligned} \mathbb{E}[\|\tilde{s} - s\|^q] &\leq C(q, K, K', \Delta, \Psi(s)) \inf_{m \in \mathcal{M}_n} \left[\|s - s_m\| + \frac{L_m D_m}{n} \right]^{q/2} \\ &\quad + n^{-q/2} C'(q, K, K', \Phi, \Phi', \{\theta_k\}, \Psi(s), \|s\|) \end{aligned}$$

provided that the following conditions are satisfied:

$$R_n(s) \leq \psi(\bar{s}_n) + K' \quad \text{and} \quad \Psi(s) = \sup_n \psi(\bar{s}_n) < +\infty. \quad (22)$$

4.5 Examples

4.5.1 NESTED MODELS

Example 1 We assume here that the true s belongs to some unknown Besov space $B_{\alpha, 2, \infty}(\mathcal{A})$ and that $\mathcal{M}_n = \{0, \dots, J_n\}$. If $\mathcal{A} = [-A, A]$, let $J_n = [\log n]$, $\{\varphi_\lambda\}_{\lambda \in \Lambda}$ be a wavelet basis of regularity r and $\Lambda_m = \sum_{0 \leq j \leq m} \Lambda(j)$, then S_m is the linear span of $\{\varphi_\lambda\}_{\lambda \in \Lambda_m}$. If $\mathcal{A} = [0, 1]$, S_m is the space of piecewise polynomials of degree $\leq r$ based on the dyadic partition generated by the grid $\{j2^{-m}, 0 \leq j \leq 2^m\}$ and $J_n = [\log n]$. If $\mathcal{A} = \mathbb{T}$, S_m is the set of trigonometric polynomials of degree $\leq m$ and $J_n = n$. Provided that $\alpha < r + 1$, the approximation properties of s by s_m which are collected in Section 4.7.1 lead, for each of our three cases, to a bias control of the form $\|s - s_m\| \leq C(s) D_m^{-\alpha}$. Assumption N (ii) is satisfied by the Fourier basis and N (i) by the piecewise polynomials because of (7) and by the wavelets by (5). Therefore Theorem 3 applies and choosing m in such a way that $D_m \asymp n^{1/(1+2\alpha)}$ we get a rate of convergence of order $n^{-\alpha/(2\alpha+1)}$ provided that $\alpha < r + 1$ except for the trigonometric polynomials which allow to deal with all values of α simultaneously.

One can also use the cross-validated estimator defined in Theorem 5 if we assume that (16) is satisfied.

Example 2 $\bar{\mathcal{S}}_n$ is a finite-dimensional space with an orthonormal basis $\{\varphi_\lambda\}_{\lambda \in \bar{\Lambda}_n}$ and $|\bar{\Lambda}_n| = N_n$. Let $\bar{\mathcal{M}}_n$ be the collection of all subsets of $\{1, \dots, J_n\}$ and $\{\Lambda(j)\}_{1 \leq j \leq J_n}$ be a partition of $\bar{\Lambda}_n$. For any $m' \in \bar{\mathcal{M}}_n$ we define $\Lambda_{m'} = \sum_{j \in m'} \Lambda(j)$ and $S_{m'}$ to be the linear span of $\{\varphi_\lambda\}_{\lambda \in \Lambda_{m'}}$. Let \mathcal{M}_n be the collection of the sets $\{1, \dots, j\}$ with $1 \leq j \leq J_n$ and assume that N holds for \mathcal{M}_n . Let us choose the penalty function $\text{pen}(m') = K|\Lambda_{m'}|/n$. Then the corresponding PPE will minimize

$$\sum_{j \in m'} \left[- \sum_{\lambda \in \Lambda(j)} \hat{\beta}_\lambda^2 + |\Lambda(j)| \frac{K}{n} \right]$$

with respect to m' and the solution is clearly to keep only those indices j for which $\sum_{\lambda \in \Lambda(j)} \hat{\beta}_\lambda^2 \geq |\Lambda(j)|K/n$. This is the level thresholding estimator introduced by Efroimovich (1985) for trigonometric expansions and more recently in Kerkyacharian, Picard & Tribouley (1994) with wavelets expansions. We can deal with this example using the extension of Theorem 3 given in Section 4.4.3 provided that the dimension of $S_{m'}$ has the required property (18). This property will be clearly satisfied if $|\Lambda(j)| \geq (1 + \rho)|\Lambda(j - 1)|$ for all j 's and some $\rho > 0$. Comparing these results with those of Example 1 we notice that this method performs exactly as the methods described in Example 1. This means that in such a case one cannot do better with the larger family $\{S_{m'}\}_{m' \in \bar{\mathcal{M}}_n}$ than with the simpler one $\{S_m\}_{m \in \mathcal{M}_n}$.

4.5.2 SELECTING A SUBSET OF A WAVELET BASIS

We shall now provide some details and proofs about the results announced in Section 4.3.2. We follow hereafter the notations and assumptions of that section. We recall that it follows from (5) that $\|\varphi_\lambda\|_\infty \leq \Phi\sqrt{N_n}$ for all $\lambda \in \bar{\Lambda}_n$ and that N_n has been chosen of order $n/(\log^2 n)$ so that all the structural assumptions concerning the basis which are required to apply Theorem 6 are satisfied.

Example 3 (Thresholding) Following the set-up given in Section 4.3.2 we want to apply Theorem 6, \mathcal{M}_n being the family of *all* the subsets of $\bar{\Lambda}_n$ and the penalty function being given by $\text{pen}(m) = K(\Phi_\infty \Sigma_\infty(\hat{s}_n) + K') \log n D_m/n$.

Proof of Proposition 2: Note first that $R_n(s) \leq \|s\|_\infty < +\infty$ since $s \in B_0$. The number of models m with a given cardinality $|m| = D$ is bounded by $\binom{N_n}{D} < (eN_n/D)^D$. Hence Assumption B holds with our choice $L_m = \log n$. Following (4), let us choose $\psi(t) = \Phi_\infty \Sigma_\infty(t)$ for all $t \in \bar{\mathcal{S}}_n$. Then

$$|\psi(\bar{s}_n) - \psi(\hat{s}_n)| \leq \Phi_\infty \Sigma_\infty(\bar{s}_n - \hat{s}_n) \leq \Phi_\infty \sup_\lambda |\nu_n(\varphi_\lambda)| \sum_{j=0}^{J_n} 2^{j/2}$$

which implies (20). It remains to check (22) which is immediate from (10) and $R_n(s) \leq \|s\|_\infty \leq \Phi_\infty \Sigma_\infty(s)$. \square

When applied to Besov spaces, Proposition 2 gives

Corollary 3 *Assume that s belongs to some Besov space $B_{\alpha,p,\infty}$ with $r+1 > \alpha > 1/p$. Then the threshold estimator described above satisfies, for any $q \geq 1$,*

$$\mathbb{E}[\|\tilde{s} - s\|^q] = \mathcal{O}\left((\log n/n)^{q\alpha/(1+2\alpha)}\right)$$

provided that (10) holds.

Proof: Since s belongs to B_0 , Proposition 2 applies and provides an upper bound for the risk of the form $C[\|s_m - s\|^2 + \log n D_m/n]^{q/2}$ for any subset m of $\bar{\Lambda}_n$. Although Proposition 6 of Section 4.7.2 has been designed for the smaller family of sieves to be considered in the next example, we can a fortiori apply it with the larger collection of sieves that we are handling here. The choices $J = J_n$ and $2^{j'} \asymp (n/\log n)^{1/(1+2\alpha)}$ ensure the existence of some m such that

$$D_m = \mathcal{O}\left(\left(\frac{n}{\log n}\right)^{1/(1+2\alpha)}\right) \quad \text{and} \quad \|s - s_m\|^2 = \mathcal{O}\left(\left(\frac{n}{\log n}\right)^{-2\alpha/(1+2\alpha)}\right)$$

which leads to the expected rate. \square

Example 4 (Special strategy for Besov spaces) We follow the set-up given in Section 4.3.2. Let us first give more information about the computation of the estimator. Since the penalty has the form $\tilde{K}L|m|/n$, the estimator will take a rather simple form, despite the apparent complexity of the family of sieves. We have to minimize over the values of m in $\mathcal{M}_n = \sum_{0 \leq j' \leq J_n} \mathcal{M}_{J_n}^{j'}$ the quantity $\tilde{K}L|m|/n - \sum_{\lambda \in m} \hat{\beta}_\lambda^2$. This optimization can be carried out in two steps, first with respect to $m \in \mathcal{M}_{J_n}^{j'}$ for fixed j' and then with respect to j' . The first step amounts to minimize

$$\sum_{j' < j \leq J_n} \left[\frac{\tilde{K}L}{n} |m(j)| - \sum_{\lambda \in m(j)} \hat{\beta}_\lambda^2 \right].$$

Since for a given j , $|m(j)|$ is fixed, the operation amounts to selecting the set $\hat{m}^{j'}(j)$ corresponding to the largest $[|\Lambda(j)|l(j-j')]$ coefficients $|\hat{\beta}_\lambda|$ for each $j > j'$. This is analogous but different from a thresholding procedure. Instead of selecting the coefficients which are larger than some threshold, one merely fixes the number of coefficients one wants to keep equal to $|m(j)|$ and takes the largest ones. For each j' the minimization of the criterion leads to the element $\hat{m}^{j'}$ of $\mathcal{M}_{J_n}^{j'}$. One should notice that all the elements of

$\mathcal{M}_{J_n}^{j'}$ have the same cardinality $2^{j'} Q(j')$. Therefore one selects j' in order to minimize

$$-\sum_{\lambda \in \hat{\mathcal{M}}^{j'}} \hat{\beta}_\lambda^2 + \tilde{K}L2^{j'}Q(j')/n$$

which only requires a few comparisons because the number of j' 's is of the order of $\log n$. We now want to apply Theorem 6 to the family \mathcal{M}_n with $\text{pen}(m) = K(\Phi_\infty \Sigma_\infty(\hat{s}_n) + K')LD_m/n$.

Proof of Proposition 3: The proof follows exactly the lines of the preceding proof with the same function ψ , the only difference being the new choice of the weight L which is now independent of n . Since $|m| \geq M2^{j'}$ for all $m \in \mathcal{M}_{J_n}^{j'}$, we get

$$\sum_{m \in \mathcal{M}_n} \exp(-\kappa_1 L|m|) \leq \sum_{j'} \exp(-\kappa_1 LM2^{j'} + \log |\mathcal{M}_{J_n}^{j'}|).$$

(19) follows if we choose $L \geq 2M^{-1} \sup_{j'}(2^{-j'} \log |\mathcal{M}_{J_n}^{j'}|)$ which achieves the proof. \square

Let us now conclude with an evaluation of the risk of \tilde{s} when the target function s belongs to some Besov space. From now on, let l be the function $l(x) = x^{-3}2^{-x}$.

Corollary 4 Assume that s belongs to some Besov space $B_{\alpha,p,\infty}$ with $r+1 > \alpha > 1/p$. Then the estimator \tilde{s} satisfies, for any $q \geq 1$,

$$\mathbb{E}[\|\tilde{s} - s\|^q] = \mathcal{O}\left(n^{-q\alpha/(1+2\alpha)}\right)$$

provided that (10) holds.

Proof: It follows the lines of the proof of Corollary 1 by applying again Proposition 6 which is exactly tuned for our needs. One chooses $2^{j'} \asymp n^{1/(1+2\alpha)}$ and one concludes by Proposition 3. \square

4.5.3 VARIABLE WEIGHTS AND PIECEWISE POLYNOMIALS

Example 5 We exactly follow the definition of the family of sieves given in Section 4.3.3. and try to apply Theorem 6.

Proof of Proposition 4: $R_n(s)$ is clearly bounded by $\|s\|_\infty$. Since there is at most one sieve per dimension when $m \in \bar{\mathcal{P}}_{J_n}^r$ and since the number of different partitions including D nonvoid intervals (and therefore corresponding to a sieve of dimension $D(r+1)$) is bounded by $(e2^{J_n}/D)^D$, (19) is satisfied and Assumption B holds. Recalling that whatever m and $t \in S_m$, $t^2 \in \bar{\mathcal{S}}_n$, we conclude that $R_n(s) \leq \|\bar{s}_n\|_\infty$. Let $\psi(t) = \|t\|_\infty$. Since by (8) $\psi(\bar{s}_n) \leq (2r+1)\|s\|_\infty$, (22) is satisfied. It remains to find a basis of

$\bar{\mathcal{S}}_n$ with the required properties. We take the basis which is the union of the Legendre polynomials on each elementary intervals. Due to the properties of these polynomials mentioned in Section 4.2.2, the required bound on $\|\varphi_\lambda\|_\infty$ holds. Finally, denoting by I_j the interval $[(j-1)2^{-J_n}, j2^{-J_n})$ we get by (7)

$$\begin{aligned}\|\psi(\hat{s}_n) - \psi(\bar{s}_n)\|_\infty &= \sup_{1 \leq j \leq 2^{J_n}} \|(\hat{s}_n - \bar{s}_n)\mathbb{I}_{I_j}\|_\infty \\ &\leq \sup_{1 \leq j \leq 2^{J_n}} (2r+1)2^{J_n/2} \|(\hat{s}_n - \bar{s}_n)\mathbb{I}_{I_j}\| \\ &\leq (2r+1)2^{J_n/2} \sqrt{2r+1} \sup_{\lambda \in \bar{\Lambda}_n} \nu_n(\varphi_\lambda)\end{aligned}$$

which gives (20) and Theorem 6 applies. \square

Following the arguments of Example 1, one can conclude that the estimator will reach the optimal rate of convergence $n^{-\alpha/(2\alpha+1)}$ for all Besov spaces $B_{\alpha,2,\infty}([0,1])$ with $\alpha < r+1$ since in this case the best choice of m corresponds to a regular partition and therefore $L_m = 1$. For other densities, the risk comes within a $\log n$ factor to the risk obtained by the estimator build on the best partition if s were known.

Remarks: A similar strategy of introducing variable weights could be applied in the same way to deal with the situation described in Example 3. It would lead to similar results and give the right rate of convergence in Besov spaces $B_{\alpha,2,\infty}([0,1])$ when $1/2 < \alpha < r+1$. But the resulting estimator would not be a thresholding estimator anymore since the penalty would not be proportional to the dimension of the sieve.

4.6 Proofs

4.6.1 INEQUALITIES FOR χ^2 STATISTICS

Let $\|a\|$ denote the euclidean norm in \mathbb{R}^D . Inequality (13) implicitly contains the following bound on χ^2 -type statistics which is of independent interest.

Proposition 5 *Let X_1, \dots, X_n be i.i.d. random variables and $Z_n = \sqrt{n}\nu_n$ the corresponding normalized empirical operator. Let $\varphi_1, \dots, \varphi_D$ be a finite set of real functions. Let $v = \sup_{\|a\| \leq 1} \mathbb{E}[(\sum_{j=1}^D a_j \varphi_j(X_1))^2]$ and $b^2 = \|\sum_{j=1}^D \varphi_j^2\|_\infty$. The following inequality holds for all positive t and ε :*

$$\mathbb{P}\left[\sum_{j=1}^D Z_n^2(\varphi_j) \geq (1+\varepsilon)\kappa_2^2((Dv) \wedge b^2) + x\right] \leq 2 \exp\left[\frac{-\kappa_1\varepsilon}{1+\varepsilon} \left(\frac{x}{v} \wedge \frac{\sqrt{nx}}{b}\right)\right].$$

Proof: We denote by Z'_n the symmetrized empirical process defined by $Z'_n(f) = n^{-1/2} \sum_{i=1}^n \varepsilon_i f(X_i)$, where the ε_i 's are independent Rademacher random variables independent from the X_i 's. Let

$$Y = \sup_{\|a\| \leq 1} \left| Z_n \left(\sum_{j=1}^D a_j \varphi_j \right) \right| \quad \text{and} \quad Y' = \sup_{\|a\| \leq 1} \left| Z'_n \left(\sum_{j=1}^D a_j \varphi_j \right) \right|.$$

From the well known duality formula $\sup_{\|a\| \leq 1} |\sum_{j=1}^D a_j b_j| = \|b\|$ and the linearity of Z_n and Z'_n we derive that

$$(i) \quad Y = \left[\sum_{j=1}^D Z_n^2(\varphi_j) \right]^{1/2} \quad \text{and} \quad (i') \quad Y' = \left[\sum_{j=1}^D Z_n'^2(\varphi_j) \right]^{1/2}.$$

In order to apply Theorem 2 we first control Y' . It comes from (i') and Jensen's inequality that

$$\mathbb{E}(Y') \leq \left[\sum_{j=1}^D \mathbb{E}(Z_n'^2(\varphi_j)) \right]^{1/2} \leq \left[\sum_{j=1}^D \mathbb{E}(\varphi_j^2(X_1)) \right]^{1/2} \leq \sqrt{Dv} \wedge b$$

and therefore Theorem 2 yields

$$\mathbb{P}[Y \geq \kappa_2(\sqrt{Dv} \wedge b) + \xi] \leq 2 \exp \left[-\kappa_1 \left(\frac{\xi^2}{v} \wedge \frac{\xi \sqrt{n}}{b} \right) \right].$$

Since for $\varepsilon > 0$, $(\alpha + \beta)^2 \leq \alpha^2(1 + \varepsilon) + \beta^2(1 + \varepsilon^{-1})$, we get for $x = (1 + \varepsilon^{-1})\xi^2$

$$\mathbb{P}[Y^2 \geq (1 + \varepsilon)\kappa_2^2((Dv) \wedge b^2) + x] \leq 2 \exp \left[-\frac{\kappa_1}{1 + \varepsilon^{-1}} \left(\frac{x}{v} \wedge \frac{\sqrt{nx}}{b} \right) \right]$$

and the result follows from (i). \square

In order to enlight the power of this bound, let us see what it gives for the standard χ^2 statistics. We want to prove (14). Considering a partition of $[0, 1]$ by intervals $(I_j)_{1 \leq j \leq D}$ such that the length of each I_j is equal to p_j and applying Proposition 5 with X_i uniformly distributed on $[0, 1]$, $\varphi_j = (1/\sqrt{p_j})\mathbb{I}_{I_j}$, $v = 1$ and $b^2 = 1/\delta \geq D$ we get the required bound (14) for the χ^2 statistics $\mathcal{K}_{n,D}^2$ which has the same distribution as $\sum_{j=1}^D Z_n^2(\varphi_j)$.

4.6.2 PROOF OF THEOREMS 3 AND 4

Without loss of generality we can assume that the index set \mathcal{M}_n is chosen in such a way that $m = D_m$ and that \bar{S}_n is the largest of the S_m 's, which we shall assume throughout the proof. Let m be some element of \mathcal{M}_n which minimizes the sum $m/n + \|s - s_m\|^2$, m' an arbitrary element in \mathcal{M}_n and

$t \in S_{m'}$. We define $w(t) = \|s - t\| + \|s - s_m\|$ and apply Talagrand's Theorem to the family of functions $f_t = (t - s_m)/w(t)$ for $t \in S_{m'}$. It will be convenient to use the following form of Theorem 2 which is more appropriate for our needs:

$$\mathbb{P} \left[\sup_t \nu_n(f_t) \geq \kappa_2 \mathbb{E} + \xi \right] \leq \exp \left[-n \kappa_1 \left(\frac{\xi^2}{v} \wedge \frac{\xi}{b} \right) \right] \quad (23)$$

with

$$b = \sup_t \|f_t\|_\infty; \quad v = \sup_t \text{Var}(f_t(X)); \quad \mathbb{E} = \mathbb{E} \left[\sup_t \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f_t(X_i) \right| \right].$$

To control \mathbb{E} , we shall distinguish between two cases:

(a) If $m \leq m'$ then $s_m \in S_{m'}$. Let $\{\varphi_\lambda\}_{\lambda \in \Lambda_{m'}}$ be an orthonormal basis of $\bar{S}_{m'}$. Then $t - s_m = \sum_{\lambda \in \Lambda_{m'}} \beta_\lambda \varphi_\lambda$ and

$$\begin{aligned} \mathbb{E} \left[\left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i (t - s_m)(X_i) \right)^2 \right] &\leq \|t - s_m\|^2 \sum_{\lambda \in \Lambda_{m'}} \mathbb{E} \left[\left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi_\lambda(X_i) \right)^2 \right] \\ &= \frac{1}{n} \sum_{\lambda \in \Lambda_{m'}} \mathbb{E}[\varphi_\lambda^2(X_1)] \|t - s_m\|^2. \end{aligned}$$

Since $w(t) \geq \|t - s_m\|$ and Assumption N holds we get

$$\mathbb{E}^2 \leq \frac{1}{n} \int \Psi_{m'}^2 s d\mu \quad \text{where} \quad \Psi_{m'}^2 = \sum_{\lambda \in \Lambda_{m'}} \varphi_\lambda^2 \leq \Phi^2 m'. \quad (24)$$

(b) If $m > m'$, one uses the decomposition $t - s_m = (t - s_{m'}) + (s_{m'} - s_m)$ and the inequalities $w(t) \geq \|s - t\| \geq \|s - s_{m'}\| \geq \|s_m - s_{m'}\|$ and $\|s - t\| \geq \|s_{m'} - t\|$ to get by similar arguments

$$\begin{aligned} \mathbb{E} &\leq \mathbb{E} \left[\sup_t \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \frac{(t - s_{m'})(X_i)}{w(t)} \right| \right] + \mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \frac{(s_m - s_{m'})(X_i)}{\inf_t w(t)} \right| \right] \\ &\leq \left(\frac{1}{n} \int \Psi_{m'}^2 s d\mu \right)^{1/2} + \left(\frac{1}{n} \frac{\int (s_m - s_{m'})^2 s d\mu}{\|s_m - s_{m'}\|^2} \right)^{1/2}. \end{aligned} \quad (25)$$

One concludes in both cases that $\mathbb{E} \leq E_{m'}$ with

$$E_{m'} = \left(\frac{1}{n} \int \Psi_{m'}^2 s d\mu \right)^{1/2} + \mathbb{I}_{\{m' < m\}} \Phi \sqrt{\frac{m}{n}}. \quad (26)$$

Let us now fix $\eta > 0$, $\bar{m} = m \vee m'$ and for $\xi > 0$ define $x_{m'} = x_{m'}(\xi)$ by $nx_{m'}^2 = \xi^2 + \eta\bar{m}$. Notice that for any $u \in S_m$ and $t \in S_{m'}$, $\|t - u\|_\infty \leq \|t - u\|\Phi\sqrt{\bar{m}}$ from which we get

$$\|f_t\|_\infty \leq \frac{\|t - s_m\|_\infty}{\|t - s_m\|} \leq \Phi\sqrt{\bar{m}}; \quad \text{Var}(f_t(X)) \leq \frac{\int (s_m - t)^2 s d\mu}{\|t - s_m\|^2} \leq \|s\|\Phi\sqrt{\bar{m}}.$$

Then (23) implies that

$$\mathbb{P} = \mathbb{P} \left[\sup_{t \in S_{m'}} \frac{\nu_n(t - s_m)}{w(t)} \geq \kappa_2 E_{m'} + x_{m'} \right] \leq \exp \left[\frac{-n\kappa_1}{\Phi\sqrt{m}} \left(\frac{x_{m'}^2}{\|s\|} \wedge x_{m'} \right) \right].$$

Since $nx_{m'}^2 \geq \sqrt{\eta m/2}(\xi + \sqrt{\eta m})$; $x_{m'}\sqrt{2n} \geq \xi + \sqrt{\eta m}$ and $m' \leq \bar{m} \leq n\Gamma^{-2}$, one gets

$$\begin{aligned} \mathbb{P} &\leq \exp \left[-\frac{\kappa_1}{\Phi} (\xi + \sqrt{\eta m}) \left(\frac{\sqrt{\eta/2}}{\|s\|} \wedge \frac{\sqrt{n}}{\sqrt{2m}} \right) \right] \\ &\leq \exp \left[-\frac{\kappa_1}{\Phi\sqrt{2}} (\xi + \sqrt{\eta m'}) \left(\frac{\sqrt{\eta}}{\|s\|} \wedge \Gamma \right) \right]. \end{aligned}$$

Denoting by Ω_ξ the following event

$$\Omega_\xi = \left\{ \sup_{t \in S_{m'}} \frac{\nu_n(t - s_m)}{w(t)} \leq \kappa_2 E_{m'} + x_{m'}(\xi) \quad \text{for all } m' \in \mathcal{M}_n \right\} \quad (27)$$

we see that since the m' 's are all different positive integers

$$1 - \mathbb{P}[\Omega_\xi] \leq \exp \left[-\xi \frac{\kappa_1}{\Phi\sqrt{2}} \left(\frac{\sqrt{\eta}}{\|s\|} \wedge \Gamma \right) \right] \sum_{j=1}^{\infty} \exp \left[-\sqrt{j} \frac{\kappa_1\sqrt{\eta}}{\Phi\sqrt{2}} \left(\frac{\sqrt{\eta}}{\|s\|} \wedge \Gamma \right) \right].$$

If Ω_ξ is true, Lemma 1 implies that

$$\begin{aligned} \|s - \tilde{s}\|^2 &\leq \|s - s_m\|^2 + \text{pen}(m) - \text{pen}(\tilde{m}) \\ &\quad + 2(\kappa_2 E_{\tilde{m}} + x_{\tilde{m}}(\xi))(\|s - \tilde{s}\| + \|s - s_m\|). \end{aligned} \quad (28)$$

We shall again distinguish between two cases and repeatedly use the inequality $2ab \leq \alpha^2 a^2 + \alpha^{-2} b^2$.

(a) If $\tilde{m} < m$ one applies (26) and (24) to get

$$[\kappa_2 E_{\tilde{m}} + x_{\tilde{m}}(\xi)]^2 \leq 8\kappa_2^2 \Phi^2 \frac{m}{n} + \frac{2}{n} [\xi^2 + \eta m] = \frac{2}{n} [m(4\Phi^2 \kappa_2^2 + \eta) + \xi^2],$$

from which (28) becomes

$$\begin{aligned} \|s - \tilde{s}\|^2 &\leq \|s - s_m\|^2 + \text{pen}(m) + \frac{1}{2}(\|s - \tilde{s}\|^2 + \|s - s_m\|^2) \\ &\quad + \frac{8}{n} [m(4\Phi^2 \kappa_2^2 + \eta) + \xi^2] \end{aligned}$$

and finally

$$\|s - \tilde{s}\|^2 \leq \frac{16}{n} [m(4\Phi^2 \kappa_2^2 + \eta) + \xi^2] + 3\|s - s_m\|^2 + 2\text{pen}(m). \quad (29)$$

(b) If $\tilde{m} \geq m$, one chooses two real numbers α and $\beta \in (0, 1)$ and applies the following inequalities

$$\begin{aligned} 2\|s - s_m\| [\kappa_2 E_{\tilde{m}} + x_{\tilde{m}}(\xi)] &\leq \alpha^2 [\kappa_2 E_{\tilde{m}} + x_{\tilde{m}}(\xi)]^2 + \alpha^{-2} \|s - s_m\|^2; \\ 2\|s - \tilde{s}\| [\kappa_2 E_{\tilde{m}} + x_{\tilde{m}}(\xi)] &\leq \beta^2 [\kappa_2 E_{\tilde{m}} + x_{\tilde{m}}(\xi)]^2 + \beta^{-2} \|s - \tilde{s}\|^2; \\ [\kappa_2 E_{\tilde{m}} + x_{\tilde{m}}(\xi)]^2 &\leq (1 + \alpha^2) ((\kappa_2 E_{\tilde{m}})^2 + \alpha^{-2} x_{\tilde{m}}^2(\xi)) \end{aligned}$$

together with (28) to derive

$$\begin{aligned} (1 - \beta^{-2})\|s - \tilde{s}\|^2 &\leq (\alpha^2 + \beta^2)(1 + \alpha^2) [(\kappa_2 E_{\tilde{m}})^2 + \alpha^{-2} x_{\tilde{m}}^2(\xi)] \\ &\quad + (1 + \alpha^{-2})\|s - s_m\|^2 + \text{pen}(m) - \text{pen}(\tilde{m}). \end{aligned}$$

Since by (26), $nE_{\tilde{m}}^2 = \int \Psi_{\tilde{m}}^2 s d\mu$ and $nx_{\tilde{m}}^2 = \xi^2 + \eta\tilde{m}$ when $\tilde{m} \geq m$, we get

$$(1 - \beta^{-2})\|s - \tilde{s}\|^2 \leq (1 + \alpha^{-2}) \left(\|s - s_m\|^2 + (\alpha^2 + \beta^2) \frac{\xi^2}{n} \right) + \text{pen}(m) \quad (30)$$

provided that the penalty satisfies for all $m' \in \mathcal{M}_n$

$$\text{pen}(m') \geq \frac{(\alpha^2 + \beta^2)(1 + \alpha^2)}{n} \left[\kappa_2^2 \int \Psi_{m'}^2 s d\mu + \frac{\eta m'}{\alpha^2} \right]. \quad (31)$$

Under the assumptions of Theorem 3 we can apply (24) and (31) will hold provided that α, η and $1 - \beta$ are small enough depending on θ . One then derives from (29) and (30) that in both cases with probability greater than $1 - \mathbb{P}[\Omega_\xi]$

$$\|s - \tilde{s}\|^2 \leq C_1 \|s - s_m\|^2 + C_2 \frac{m}{n} + C_3 \frac{\xi^2}{n}.$$

Theorem 3 then follows from Lemma 2.

To prove Theorem 4 we first apply (24) and Hoeffding's inequality to get

$$\mathbb{P} [|\nu_n(\Psi_{m'}^2)| > \varepsilon m'] \leq 2 \exp \left[\frac{-2n\varepsilon^2 m'^2}{4\Phi^2 m'^2} \right]$$

for any positive ε , which implies that $\mathbb{P}(\Omega_n^\varepsilon) \leq 2n\Gamma^{-2} \exp[-(n\varepsilon^2)/(2\Phi^2)]$ if we denote by Ω_n the event

$$\left\{ \left| \frac{1}{n} \sum_{i=1}^n \sum_{\lambda \in \Lambda_{m'}} \varphi_\lambda^2(X_i) - \int \Psi_{m'}^2 s d\mu \right| \leq \varepsilon m' \quad \text{for all } m' \in \mathcal{M}_n \right\}.$$

If Ω_n is true

$$\begin{aligned} \text{pen}(m') &\geq \frac{K_n}{n+1} \left(\int \Psi_{m'}^2 s d\mu - \varepsilon m' \right) \\ &\geq \frac{1}{n} \left(\kappa_2^2 + \frac{\theta}{3} \right) \int \Psi_{m'}^2 s d\mu + \frac{\theta am'}{3n} + \frac{\theta m'}{3n} \left(a - \varepsilon \frac{3nK_n}{\theta(n+1)} \right) \end{aligned}$$

for all $m' \in \mathcal{M}_n$ and (31) will then be satisfied provided that we choose $\alpha, \eta, \varepsilon$ and $1 - \beta$ small enough, depending only on κ_2, θ and a . In order to conclude we notice that on Ω_n^c

$$\|s - \tilde{s}\|^2 \leq 2 \left(\|s\|^2 + \sum_{\lambda \in \bar{\Lambda}_n} \hat{\beta}_j^2 \right) \leq 2 \left(\|s\|^2 + \left(\frac{n\Phi}{\Gamma^2} \right)^2 \right)$$

since $\bar{\mathcal{S}}_n$ is one of the S_m 's and therefore $|\hat{\beta}_\lambda| \leq \|\varphi_\lambda\|_\infty \leq \Phi\sqrt{N_n}$. Hence

$$\mathbb{E} [\|s - \tilde{s}\|^q \mathbb{I}_{\Omega_n^c}] \leq [2(\|s\|^2 + n^2\Phi^2\Gamma^{-4})]^{q/2} \frac{2n}{\Gamma^2} \exp \left[-\frac{n\varepsilon^2}{2\Phi^2} \right].$$

On the other hand on Ω_n (31) is satisfied and $\text{pen}(m)$ is bounded by

$$\text{pen}(m) \leq \frac{K_n}{n+1} \left(\int \Psi_m^2 s d\mu + \varepsilon m \right) \leq \frac{1}{n} (\kappa_2^2 + \theta)(\Phi^2 m + \varepsilon m)$$

and finally by (30)

$$\|s - \tilde{s}\|^2 \mathbb{I}_{\Omega_n} \leq C_1 \|s - s_m\|^2 + C_2 \frac{m}{n} + C_3 \frac{\xi^2}{n}$$

which allows us to conclude by Lemma 2.

4.6.3 PROOF OF THEOREMS 5 AND 6

Let m be some fixed element of \mathcal{M}_n , m' an arbitrary element in \mathcal{M}_n and $t \in S_{m'}$. Once again, we want to apply Theorem 2 to the family of functions $f_t = (t - s_m)/w(t)$ where $w(t) = (\|s - t\| + \|s - s_m\|) \vee 2x_{m'}$ with $t \in S_{m'}$, $x_{m'}^2 = x_{m'}^2(\xi) = B_n(\xi^2 + \kappa_1^{-1}D_{m'}L_{m'})/n$ and $\xi \geq 1$. We get

$$\|f_t\|_\infty \leq \frac{\|t - s_m\|_\infty}{2x_{m'}} \leq \frac{2B_n}{2x_{m'}}; \quad \text{Var}(f_t(X)) \leq \frac{\int (s_m - t)^2 s d\mu}{\|t - s_m\|^2} \leq B_n;$$

$$\mathbb{E} \leq \sqrt{B_n D_{m'}/n} + \sqrt{B_n/n} \leq E_{m'} = \sqrt{2}x_{m'}$$

by the analogues of (24) and (25) since $R_n(s) \leq B_n$ and now $D_{m'} \neq m'$. Theorem 2 then implies that

$$\mathbb{P} \left[\sup_{t \in S_{m'}} \frac{\nu_n(t - s_m)}{w(t)} > \kappa_2 E_{m'} + x_{m'}(\xi) \right] \leq \exp \left[-\frac{n\kappa_1 x_{m'}^2}{B_n} \right]$$

and we use Assumption B with Ω_ξ defined by (27) to get

$$1 - \mathbb{P}[\Omega_\xi] \leq \sum_{m' \in \mathcal{M}_n} \exp[-\kappa_1 \xi^2 + D_{m'} L_{m'}] \leq \Delta \exp(-\kappa_1 \xi^2). \quad (32)$$

Let $2(\kappa_2 E_{m'} + x_{m'}) = \kappa x_{m'}$. Lemma 1 implies that, if Ω_ξ is true

$$\|s - \tilde{s}\|^2 \leq \|s - s_m\|^2 + \text{pen}(m) - \text{pen}(\tilde{m}) + \kappa x_{\tilde{m}}(\xi) w(\tilde{s}). \quad (33)$$

Then either $w(\tilde{s}) = 2x_{\tilde{m}}(\xi)$ and $x_{\tilde{m}}(\xi)w(\tilde{s}) = 2x_{\tilde{m}}^2(\xi)$ or

$$\begin{aligned} 2x_{\tilde{m}}(\xi)w(\tilde{s}) &= 2x_{\tilde{m}}(\xi)\|s - \tilde{s}\| + 2x_{\tilde{m}}(\xi)\|s - s_m\| \\ &\leq x_{\tilde{m}}^2(\xi) + \|s - s_m\|^2 + \kappa x_{\tilde{m}}^2(\xi) + \|s - \tilde{s}\|^2/\kappa. \end{aligned}$$

In both cases since $\kappa = 2(1 + \kappa_2\sqrt{2}) > 3$

$$2\kappa x_{\tilde{m}}(\xi)w(\tilde{s}) \leq \kappa\|s - s_m\|^2 + \kappa(1 + \kappa)x_{\tilde{m}}^2(\xi) + \|s - \tilde{s}\|^2$$

and (33) becomes with $\kappa(1 + \kappa) = 2(3 + 5\sqrt{2}\kappa_2 + 4\kappa_2^2)$

$$\|s - \tilde{s}\|^2 \leq (2 + \kappa)\|s - s_m\|^2 + 2[\text{pen}(m) - \text{pen}(\tilde{m})] + \kappa(1 + \kappa)x_{\tilde{m}}^2(\xi).$$

Therefore on the set $\Omega_\xi \cap \tilde{\Omega}$,

$$\|s - \tilde{s}\|^2 \leq 2 \left[(2 + \kappa_2\sqrt{2})\|s - s_m\|^2 + \text{pen}(m) \right] + \kappa(1 + \kappa)B_n\xi^2/n$$

and Theorem 5 follows from (32) and an analogue of Lemma 2.

Let us now turn to Theorem 6. Let $R = \psi(\bar{s}_n) + K' \geq R_n(s)$, $\varepsilon = R/(3\Phi'\sqrt{N_n})$ and $\tilde{\Omega}$ be the event $\{\sup_\lambda |\nu_n(\varphi_\lambda)| \leq \varepsilon\}$. From our assumptions and Bernstein's inequality we get

$$\mathbb{P}[\tilde{\Omega}^c] \leq 2N_n \exp \left[\frac{-n\varepsilon^2}{2R_n(s) + \frac{2}{3}\Phi\sqrt{N_n}\varepsilon} \right] \leq 2N_n \exp \left[\frac{-K'\theta_n \log n}{2\Phi'(9\Phi' + \Phi)} \right] \quad (34)$$

since $\|\varphi_\lambda\|_\infty \leq \Phi\sqrt{N_n}$ and $\text{Var}(\varphi_\lambda(X_1)) \leq \int \varphi_\lambda^2 s d\mu \leq R_n(s)$. Let $B_n = 2(\psi(\bar{s}_n) - R/3 + K')$ and assume that $\tilde{\Omega}$ is true; then by (20) $\psi(\bar{s}_n) - R/3 \leq \psi(\hat{s}_n) \leq \psi(\bar{s}_n) + R/3$. Since s_m and \hat{s}_m are the projections of \bar{s}_n and \hat{s}_n respectively on S_m , one derives that

$$R \leq B_n; \quad \sup_{m \in \mathcal{M}_n} \|\hat{s}_m\|_\infty \leq \psi(\hat{s}_n) \leq B_n; \quad \sup_{m \in \mathcal{M}_n} \|s_m\|_\infty \leq \psi(\bar{s}_n) \leq B_n$$

and for all $m \in \mathcal{M}_n$ simultaneously since $K \geq \kappa(1 + \kappa)/\kappa_1$

$$\kappa(1 + \kappa)B_n \frac{L_m D_m}{2n\kappa_1} \leq \text{pen}(m) \leq \frac{4KL_m D_m}{3n} (\psi(\bar{s}_n) + K').$$

If \tilde{s}' is the penalized projection estimator defined on the family of sieves S'_m , $m \in \mathcal{M}_n$ with a penalty given by (21) and $S'_m = \{t \in S_m \mid \|t\|_\infty \leq B_n\}$, it follows from the above inequalities that $\tilde{s} = \tilde{s}'$ and that Theorem 5 applies to \tilde{s}' . One can then conclude since $L_m D_m \geq 1$ that

$$\begin{aligned} \mathbb{E}[\|\tilde{s} - s\|^q \mathbb{I}_{\tilde{\Omega}}] &\leq C(q) \left[\|s - s_m\|^q + \mathbb{E}[(\text{pen}(m))^{q/2} \mathbb{I}_{\tilde{\Omega}}] + \Delta(B_n/n)^{q/2} \right] \\ &\leq C(q) \left[\|s - s_m\|^q + C'(q, \Delta, \Psi(s), K, K') \left(\frac{L_m D_m}{n} \right)^{q/2} \right]. \end{aligned}$$

On the other hand, if $\tilde{\Omega}$ does not hold one uses the crude estimate

$$\|\tilde{s}\|_\infty \leq \|\psi(\hat{s}_n)\|_\infty \leq \Psi(s) + \Phi' \sqrt{N_n} \sup_j \|\nu_n(\varphi_\lambda)\|_\infty \leq \Psi(s) + \Phi\Phi' N_n$$

from which one deduces by (34) since $C'^2 = \int d\mu < +\infty$ that

$$\mathbb{E}[\|s - \tilde{s}\|^q \mathbb{I}_{\tilde{\Omega}^c}] \leq 2N_n[C'(\Psi(s) + \Phi\Phi' N_n) + \|s\|]^q \exp\left[\frac{-K'\theta_n \log n}{2\Phi'(9\Phi' + \Phi)}\right],$$

which is bounded by $Cn^{-q/2}$ as required if C is large enough.

4.7 Some results in approximation theory for Besov spaces

4.7.1 LINEAR APPROXIMATIONS

We shall collect here some known results of approximation of Besov spaces $B_{\alpha,p,\infty}(\mathcal{A})$ defined in Section 4.2.2 by classical finite-dimensional linear spaces. We first assume that $p = 2$ and consider the following approximation spaces:

- If $\mathcal{A} = [0, 1]$ let S be the space of piecewise polynomials of degree bounded by r with $r > \alpha - 1$ based on the partition generated by the grid $\{j/D, 0 \leq j \leq D\}$;
- If $\mathcal{A} = \mathbb{T}$ let S be the space of trigonometric polynomials of degree $\leq D$;
- If $\mathcal{A} = [-A, A]$ let S be the space \bar{V}_J generated by a wavelet basis of regularity $r > \alpha - 1$ defined in Section 4.2.2 with $D = 2^J$.

Let $\pi(s)$ be the projection of s onto the approximating space S . Then in each of the three situations, with different constants $C(s)$ in each case, we get

$$\|s - \pi(s)\| \leq C(s)D^{-\alpha}. \quad (35)$$

The proof of (35) comes from DeVore & Lorentz (1993) page 359 for piecewise polynomials and page 205 for trigonometric polynomials. For the wavelet expansion we shall prove a more general result than (35) which holds when $p \leq 2$ and $\alpha > 1/p - 1/2$. From the classical inequality

$$\sum_{\lambda \in \Lambda(j)} |\beta_\lambda|^2 \leq (2^j M)^{(1-2/p)^+} \left[\sum_{\lambda \in \Lambda(j)} |\beta_\lambda|^p \right]^{2/p}$$

and (6) we derive that

$$\|s - \pi(s)\|^2 = \sum_{j>J} \sum_{\lambda \in \Lambda(j)} |\beta_\lambda|^2 \leq \|s\|^2 \sum_{j>J} 2^{-2j(\alpha + \frac{1}{2} - \frac{1}{p})} (2^j M)^{(1-2/p)^+}.$$

This implies for $p = 2$

$$\|s - \pi(s)\|^2 \leq \|s\|^2 \sum_{j>J} 2^{-2j\alpha} = \frac{\|s\|^2}{4^\alpha - 1} 2^{-2J\alpha},$$

which gives (35) since $D = 2^J$. Moreover for $p < 2$ and $\alpha > 1/p - 1/2$,

$$\|s - \pi(s)\|^2 \leq \|s\|^2 \sum_{j>J} 2^{-2j(\alpha + \frac{1}{2} - \frac{1}{p})} = \|s\|^2 \frac{2^{-2J(\alpha + \frac{1}{2} - \frac{1}{p})}}{4^{\alpha + \frac{1}{2} - \frac{1}{p}} - 1}. \quad (36)$$

4.7.2 NONLINEAR APPROXIMATIONS

Starting with a wavelet basis $\{\varphi_\lambda\}_{\lambda \in \Lambda}$ as described in Section 4.2.2, we follow the framework stated for Proposition 3 of Section 4.3.2 with $l(x) = x^{-3}2^{-x}$. We want to study the approximation properties of the union of linear spaces $S_m = \text{Span}(\{\varphi_\lambda \mid \lambda \in m\})$ when m belongs to $\mathcal{M}_J^{j'}$.

Proposition 6 *All elements of $\mathcal{M}_J^{j'}$ have the same cardinality bounded by $\kappa' M 2^{j'}$ and $\log |\mathcal{M}_J^{j'}|$ is bounded by $\kappa'' M 2^{j'}$. Moreover, if $M 2^{j'} \geq J^3$ and s belongs to $B_{\alpha p \infty}([-A, A])$ with $p \leq 2$ and $\alpha > 1/p - 1/2$ there exists $m \in \mathcal{M}_J^{j'}$ such that*

$$\|s - s_m\|^2 \leq C \|s\|^2 \left(2^{-2\alpha j'} + 2^{-2J(\alpha + \frac{1}{2} - \frac{1}{p})} \right). \quad (37)$$

Remarks

- From the bounds on the cardinalities of both $\mathcal{M}_J^{j'}$ and the elements of $\mathcal{M}_J^{j'}$ one can derive that $\cup_{m \in \mathcal{M}_J^{j'}} S_m$ is a metric space with finite metric dimension (in the sense of Le Cam) bounded by $D = C' 2^{j'}$ whatever J .
- Choosing $J \asymp \alpha j' / (\alpha + 1/2 - 1/p)$, this finite dimensional nonlinear metric space approximates $B_{\alpha p \infty}([-A, A])$ within $\mathcal{O}(D^{-\alpha})$. Hence (37) provides an analogue of (35), the difference being that a nonlinear finite-dimensional space instead of a linear vector space (both with dimensions of order D) is needed to get the $D^{-\alpha}$ -rate of approximation for $p < 2$.
- As a consequence, the ε -entropy of $\{s \in B_{\alpha p \infty}([-A, A]) \mid \|s\| \leq 1\}$ is of order $\varepsilon^{-1/\alpha}$ when $\alpha > 1/p - 1/2$.

Proof of Proposition 6: The bound on $|m|$ derives from

$$|m| = M \left(\sum_{j=0}^{j'} 2^j + 2^{j'} \sum_{k=1}^{J-j'} 2^k l(k) \right) < \kappa' M 2^{j'}.$$

The control of $|\mathcal{M}_J^{j'}|$ is clear for $j' = J$. Otherwise

$$|\mathcal{M}_J^{j'}| = \prod_{j=j'+1}^J \left(\frac{M 2^j}{[M 2^j l(j - j')]} \right) = \prod_{j=1}^{J-j'} \left(\frac{M 2^{j+j'}}{[M 2^{j+j'} l(j)]} \right)$$

and from the inequality $\log \left(\frac{n}{[nx]} \right) < nx(\log(1/x) + 1)$ which holds for $0 < x \leq 1$ one gets

$$\log |\mathcal{M}_J^{j'}| \leq M 2^{j'} \sum_{j=1}^{\infty} 2^j l(j) \left(\log \left(\frac{1}{l(j)} \right) + 1 \right)$$

and the series converges from our choice of l . If $p = 2$, the conclusion of Proposition 6 follows from (35). If $p < 2$ the bias can always be written as

$$\|s - s_m\|^2 = \sum_{j>J} \sum_{\lambda \in \Lambda(j)} \beta_{\lambda}^2 + \sum_{j=j'+1}^J \sum_{\lambda \in \Lambda(j) \setminus m(j)} \beta_{\lambda}^2$$

where the second term is 0 when $j' = J$. We can bound the first term by (36). In order to control the second term we shall need the following

Lemma 3 *Assume that we are given n nonnegative numbers $0 \leq b_1 \leq \dots \leq b_n$ with $\sum_{i=1}^n b_i = B$. For any number $r > 1$ and any integer k with $1 \leq k \leq n - 1$ one has*

$$\sum_{i=1}^{n-k} b_i^r \leq B^r \frac{k^{1-r}}{r-1} (1 - r^{-1})^r.$$

Proof of the lemma: One can assume without loss of generality that $B = 1$ and that for $i > n - k$ the b_i 's are equal to some $b \leq 1/k$. Then $\sum_{i=1}^{n-k} b_i + kb = 1$ which implies that

$$\sum_{i=1}^{n-k} b_i^r \leq b^{r-1} \sum_{i=1}^{n-k} b_i = b^{r-1} (1 - kb),$$

and the conclusion follows from a maximization with respect to b . \square

We choose m such that for $j > j'$, $m(j)$ corresponds to the $[M 2^j l(j - j')]$ largest values of the $|\beta_{\lambda}|$ for $\lambda \in \Lambda(j)$. Since by assumption $M 2^{j'} \geq (J - j')^3$

and $l < 1$, $1 \leq |m(j)| < |\Lambda(j)|$ and we may apply Lemma 3 with $r = 2/p$ and $k = |m(j)|$ to get

$$\begin{aligned} \sum_{\lambda \in \Lambda(j) \setminus m(j)} \beta_\lambda^2 &\leq C(p) \left(\sum_{\lambda \in \Lambda(j)} \beta_\lambda^p \right)^{2/p} (|m(j)|)^{1-2/p} \\ &\leq C(p, M) \|s\|^2 2^{-2j(\alpha + \frac{1}{2} - \frac{1}{p})} (2^j l(j-j'))^{1-2/p} \end{aligned}$$

from which we deduce that

$$\sum_{j=j'+1}^J \sum_{\lambda \in \Lambda(j) \setminus m(j)} \beta_\lambda^2 \leq C(p, M) \|s\|^2 2^{-2\alpha j'} \sum_{j=1}^{\infty} 2^{-2j(\alpha + \frac{1}{2} - \frac{1}{p})} (2^j l(j))^{1-2/p}$$

and the series converges since $2^j l(j) = j^{-3}$ and $\alpha + 1/2 - 1/p > 0$. \square

4.8 REFERENCES

- Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in P. N. Petrov & F. Csaki, eds, 'Proceedings 2nd International Symposium on Information Theory', Akademia Kiado, Budapest, pp. 267–281.
- Barron, A. R. & Cover, T. M. (1991), 'Minimum complexity density estimation', *IEEE Transactions on Information Theory* **37**, 1034–1054.
- Barron, A. R., Birgé, L. & Massart, P. (1995), Model selection via penalization, Technical Report 95.54, Université Paris-Sud.
- Birgé, L. & Massart, P. (1994), Minimum contrast estimation on sieves, Technical Report 94.34, Université Paris-Sud.
- Cirel'son, B. S., Ibragimov, I. A. & Sudakov, V. N. (1976), Norm of gaussian sample function, in 'Proceedings of the 3rd Japan-USSR Symposium on Probability Theory', Springer-Verlag, New York, pp. 20–41. Springer Lecture Notes in Mathematics 550.
- DeVore, R. A. & Lorentz, G. G. (1993), *Constructive Approximation*, Springer-Verlag, Berlin.
- Donoho, D. L. & Johnstone, I. M. (1994), 'Ideal spatial adaptation by wavelet shrinkage', *Biometrika* **81**, 425–455.
- Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1993), Density estimation by wavelet thresholding, Technical Report 426, Department of Statistics, Stanford University.

- Efroimovich, S. Y. (1985), 'Nonparametric estimation of a density of unknown smoothness', *Theory of Probability and Its Applications* **30**, 557–568.
- Grenander, U. (1981), *Abstract Inference*, Wiley, New-York.
- Kerkyacharian, G. & Picard, D. (1992), 'Estimation de densité par méthode de noyau et d'ondelettes: les lieus entre la géométrie du noyau et les contraintes de régularité', *Comptes Rendus de l'Academie des Sciences, Paris, Ser. I Math* **315**, 79–84.
- Kerkyacharian, G., Picard, D. & Tribouley, K. (1994), L^p adaptive density estimation, Technical report, Université Paris VII.
- Le Cam, L. (1973), 'Convergence of estimates under dimensionality restrictions', *Annals of Statistics* **19**, 633–667.
- Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.
- Ledoux, M. (1995). Private communication.
- Li, K. C. (1987), 'Asymptotic optimality for C_p , C_L , cross-validation, and generalized cross-validation: Discrete index set', *Annals of Statistics* **15**, 958–975.
- Mallows, C. L. (1973), 'Some comments on C_p ', *Technometrics* **15**, 661–675.
- Mason, D. M. & van Zwet, W. R. (1987), 'A refinement of the KMT inequality for the uniform empirical process', *Annals of Probability* **15**, 871–884.
- Meyer, Y. (1990), *Ondelettes et Opérateurs I*, Hermann, Paris.
- Polyak, B. T. & Tsybakov, A. B. (1990), 'Asymptotic optimality of the C_p -criteria in regression projective estimation', *Theory of Probability and Its Applications* **35**, 293–306.
- Rissanen, J. (1978), 'Modeling by shortest data description', *Automatica* **14**, 465–471.
- Rudemo, M. (1982), 'Empirical choice of histograms and kernel density estimators', *Scandinavian Journal of Statistics* **9**, 65–78.
- Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**, 461–464.
- Talagrand, M. (1994), 'Sharper bounds for Gaussian and empirical processes', *Annals of Probability* **22**, 28–76.

- Talagrand, M. (1995), New concentration inequalities in product spaces,
Technical report, Ohio State University.
- Vapnik, V. (1982), *Estimation of Dependences Based on Empirical Data*,
Springer-Verlag, New York.
- Wahba, G. (1990), *Spline Models for Observational Data*, Society for
Industrial and Applied Mathematics, Philadelphia.
- Whittaker, E. T. & Watson, G. N. (1927), *A Course of Modern Analysis*,
Cambridge University Press, London.

5

Large Deviations for Martingales

D. Blackwell¹

5.1 Introduction and Summary

Let X_1, X_2, \dots be variables satisfying

- (i) $|X_n| \leq 1$ and
- (ii) $E(X_n | X_1, \dots, X_{n-1}) = 0$,

and put $S_n = X_1 + \dots + X_n$.

Theorem 1 *For any positive constants a and b ,*

$$P\{S_n \geq a + bn \text{ for some } n\} \leq \exp(-2ab).$$

Theorem 2 *For any positive constant c ,*

$$P\{S_n \geq cn \text{ for some } n \geq N\} \leq r_1^N \leq r_2^N,$$

where $r_1 = 1/((1+c)^{1+c}(1-c)^{1-c})^{1/2}$ and $r_2 = \exp(-c^2/2)$.

Theorem 2, which gives uniform exponential rates for the strong law of large numbers for martingales with bounded increments, improves an old result (Blackwell 1954, page 347), replacing $(1+c)^{-c/(2+c)}$ by the smaller and neater r_2 and by the still smaller r_1 . The bounds r_1^N and r_2^N were obtained by Hoeffding (1963, page 15) for $P\{S_N \geq cN\}$, so the only novelty in Theorem 2 is in replacing the event $\{S_N \geq cN\}$ by the larger event $\{S_n \geq cn \text{ for some } n \geq N\}$, i.e. replacing the weak law by the strong law. I thank David Pollard for several helpful comments, and especially for calling my attention to Hoeffding's paper.

The idea of looking for bounds for $P\{S_n \geq a + bn \text{ for some } n\}$, as in Theorem 1, is due to Dubins & Savage (1965). They found sharp bounds for related probabilities under hypotheses weaker than ours. Their bound, specialized to our case, replaces our $\exp(-2ab)$ by the larger $1/(1+ab)$. Our proof of Theorem 1 uses the gambling ideas of Dubins & Savage (1964). To make this paper self-contained, we use their methods instead of invoking their results.

¹University of California, Berkeley

5.2 Proofs

Let $X_1, X_2, \dots, S_n, a, b$ be as in Theorem 1. Here are two games you could play. In each game you watch X_1, X_2, \dots as long as you like, but stopping eventually. If you stop after observing X_1, \dots, X_n , with $S_n = s$, in Game 1 you get 1 if $s \geq a + bn$ and 0 otherwise, while in Game 2 you get

$$f(s, n) = \exp(-d(a + bn - s)),$$

where $d = 2b$. Clearly Game 2 is better for you than Game 1, since your income is at least as large for every s, n . We shall show that for any bounded stop rule, your expected income in Game 2 does not exceed $\exp(-2ab)$. Since in Game 1 you can get nearly $P\{S_n \geq a + bn \text{ for some } n\}$ with a bounded stop rule, this probability cannot exceed $\exp(-2ab)$, which is what Theorem 1 asserts.

To show that your expected income in Game 2 cannot exceed $\exp(-2ab)$, suppose that you have observed X_1, \dots, X_n with $S_n = s$. We show that it is better to stop now than to observe X_{n+1} and then stop, i.e. that

$$f(s, n) \geq G = Ef(s + X_{n+1}, n + 1).$$

Since f is convex in s , G is maximized when X_{n+1} has values ± 1 , each with probability .5, making $G = (f(s + 1, n + 1) + f(s - 1, n + 1))/2$. The inequality $f(s, n) \geq (f(s + 1, n + 1) + f(s - 1, n + 1))/2$ reduces to

$$\exp(2b^2) \geq (\exp(2b) + \exp(-2b))/2,$$

which is true for all b (expand in power series). Thus in Game 2, in every position (s, n) , stopping now is better than observing one more X , then stopping. This implies that, in every position, stopping now is the best of all bounded stopping rules. In particular, at the starting position $(0, 0)$, any bounded stopping rule has expected income $\leq f(0, 0) = \exp(-2ab)$.

To prove Theorem 2 from Theorem 1, note that event $A = \{S_n \geq cn \text{ for some } n \geq N\}$ is a subset of $B = \{S_n \geq (cN/2) + (c/2)n \text{ for some } n\}$. From Theorem 1 with $a = cN/2$, $b = c/2$ we get $P(B) \leq \exp(-c^2N/2)$, giving the bound r_2^N . To obtain r_1 , do not put $d = 2b$ in $f(s, n)$, but instead let d be any positive number and define the slope b by

$$f(s, n) = (f(s + 1, n + 1) + f(s - 1, n + 1))/2,$$

that is, by

$$\exp(bd) = (\exp(d) + \exp(-d))/2.$$

With $a + bN = cN$, the bound $f(0, 0) = \exp(-ad)$ becomes h^N , where

$$h = \exp((b - c)d) = (\exp(d) + \exp(-d)) \exp(-cd)/2.$$

The minimum value of h occurs for $\exp(d) = \left(\frac{1+c}{1-c}\right)^{1/2}$, and is r_1 .

5.3 REFERENCES

- Blackwell, D. (1954), 'On optimal systems', *Annals of Mathematical Statistics* **25**, 394–397.
- Dubins, L. & Savage, L. (1964), *How to Gamble if You Must*, McGraw-Hill.
- Dubins, L. & Savage, L. (1965), 'A Tchebycheff-like inequality for stochastic processes', *Proceedings of the National Academy of Sciences* **53**, 274–275.
- Hoeffding, W. (1963), 'Probability inequalities for sums of bounded random variables', *Journal of the American Statistical Association* **58**, 13–30.

6

An Application of Statistics to Meteorology: Estimation of Motion

David R. Brillinger¹

ABSTRACT Concern is with moving meteorological phenomena. Some existing techniques for the estimation of motion parameters are reviewed. Fourier-based and generalized-additive-model-based analyses are then carried out for the global geopotential 500 millibar (mb) height field during the period 1–6 January 1986.

6.1 Introduction

Early in his professional career Lucien Le Cam was a statistician at Électricité de France. His Fourth Berkeley Symposium paper, “A stochastic description of precipitation”, (Le Cam 1961), describes a conceptual stochastic model for (perhaps moving) rainfields developed at that time. In particular, Professor Le Cam was concerned with the development of river stream flow following rainfall. The model involved a smoothing transformation of a point process with direction and velocity of movement included. There have since been many references building on his work including: Smith & Karr (1985), Gupta & Waymire (1987), Cox & Isham (1988), Phelan (1992).

In this paper the focus is on velocity estimation of moving disturbances. To begin consider a number, N , of plane waves, $g_n(\alpha_n x + \beta_n y - v_n t)$, moving across a surface. The model of interest is

$$Y(x, y, t) = \sum_{n=1}^N g_n(\alpha_n x + \beta_n y - v_n t) + \text{noise} \quad (1)$$

with (x, y) location, t time and the n th wave having velocity v_n and direction cosines (α_n, β_n) . The velocity is the parameter of particular interest in this work. The functions $g_n(\cdot)$ may be known up to a finite dimensional parameter, e. g. $g(u) = \rho \cos(mu + \phi)$ or simply may be assumed smooth. The first case suggests employing Fourier techniques, while the second suggests projection pursuit regression techniques to study the velocities.

¹University of California at Berkeley

The analyses presented here are for a worldwide spatial-temporal field. An interesting aspect is that, because the data are for the whole sphere, there is a basic periodicity in the x and y coordinates.

The next section reviews some of the previous related work. In the following sections Fourier-based and smoothing-based techniques are applied to a particular meteorological data set. One finds that each of these techniques has its advantages, but that the standard errors of the velocity estimates appear notably smaller for the smoothing-based technique.

6.2 Some History

Briggs, Phillips & Shinn (1950) studied the behavior of radio waves reflected from the ionosphere. They were concerned with a randomly changing pattern moving across the ground. For example they were after drift velocities. Time series were envisaged recorded via an array of 3 receivers, i.e. at 3 locations (x, y) . With $Y(x, y, t)$ denoting the signal recorded at time t and location (x, y) , the basic parameter suggested to be employed in the estimation was the correlation function

$$\text{corr}(Y(x + v, y + w, t + u), Y(x, y, t)) \quad (2)$$

assumed not to depend on x, y, t . Briggs (1968) extends the work and in particular assumes the function (2) has the form

$$\rho[A(v - V_x u)^2 + B(w - V_y u)^2 + Ku^2 + 2H(v - V_x u)(w - V_y u)]$$

with (V_x, V_y) denoting the drift velocities. Estimates of the V_x V_y and the other parameters are obtained via nonlinear regression. Briggs (1968) also discusses the case where dispersion occurs, that is the velocity depends on the wavenumber.

Leese, Novak & Clark (1971) study cloud motion via images taken from a geosynchronous satellite. They are interested in wind fields. Pictures, $Y(x, y, t)$, are taken at times $t = t_1$ and t_2 . The basic criterion proposed is (2) with $t = t_1$ and $t + u = t_2$. The correlation is assumed to be independent of (x, y) and the point, (\hat{v}, \hat{w}) , of maximum cross-correlation is determined. The speed estimate is then $\sqrt{\hat{v}^2 + \hat{w}^2}/(t_2 - t_1)$ and the direction estimate $\tan^{-1}(\hat{v}/\hat{w})$. These researchers found their procedure "Better than manual for speed". They noted that complications that could lead to difficulties of estimation included: growth, decay, rotations and layers.

Arking, Lo & Rosenfeld (1978) take a Fourier approach. They suppose that

$$Y(x, y, t_2) \approx Y(x - v, y - w, t_1)$$

The cross-spectrum of the two fields $Y(\cdot, t_2)$ and $Y(\cdot, t_1)$ is then given by

$$f_{21}(\mu, \lambda) \approx e^{-i(v\mu + w\nu)} f_{11}(\mu, \nu)$$

where $f_{11}(\cdot)$ denotes the spatial power spectrum of $Y(x, y, t)$. The pair (v, w) are estimated from neighboring wavenumber, (μ, ν) , data.

Marshall (1980) was concerned with speed and direction of storm rainfall patterns. Data were available from a rain gauge network, with sensors located at positions (x_j, y_j) , $j = 1, 2, \dots, J$ and measurements made at times $t = 0, 1, 2, \dots$. To carry through the analysis, the data were interpolated to a grid. This researcher also took a maximum cross-correlation approach, estimating for given u , the (v, w) maximizing (2) above. If that extreme point is (\hat{v}_u, \hat{w}_u) , the mean velocity of the storm is estimated by the average of \hat{v}_u and \hat{w}_u . Marshall (1980) used least squares to fit the model

$$h \exp(-a^2 r^2 - b^2 s^2)$$

to (2), with $r = v \sin \theta + w \cos \theta$ and $s = w \sin \theta - v \cos \theta$. Brillinger (1985) indicated extensions of the results of Hannan & Thomson (1973) to provide large sample distributions for maximum cross-correlation estimates in the case of two time slices. Brillinger (1993) is concerned with estimating the joint distributions of several successive motions given consecutive locations of moving particles. Brillinger (1995) is concerned with the estimation of the travel times of the effects of cloud seeding. In that paper a conceptual model is built, analogous to that of Le Cam (1961), for the transference of the effects, then both parametric and nonparametric estimation is carried out.

In the study of problems such as those just described, critical distinctions that arise include: Is the number of sensors, J , small or large? Is the velocity constant or dispersive? Is the number of time slices, T , small or large? The choices made affect the approximations to distributions of the estimates in important ways. When J or T are large, traditional asymptotic approximations are available.

Some references on the maximum cross-correlation approach are: Burke (1987), Kamachi (1989) and Tokmakian, Strub & McClean-Padman (1990). Estimation techniques, based on differential expressions of the motion, are reviewed in Aggarwal & Nandhakumar (1988). Carroll, Hall & Ruppert (1994) investigate penalized least squares and maximum cross-covariance methods for estimating the missalignment of a pair of images. Research continues on this type of problem, into the circumstances under which each method is to be preferred.

6.3 The Data

The particular data studied here are a five day sequence of 0000 and 1200 Greenwich Mean Time (GMT) geopotential analyses. These are spatially interpolated estimates of the height of the 500 millibar (mb) pressure field across the surface of the Earth. This quantity provides the thickness of

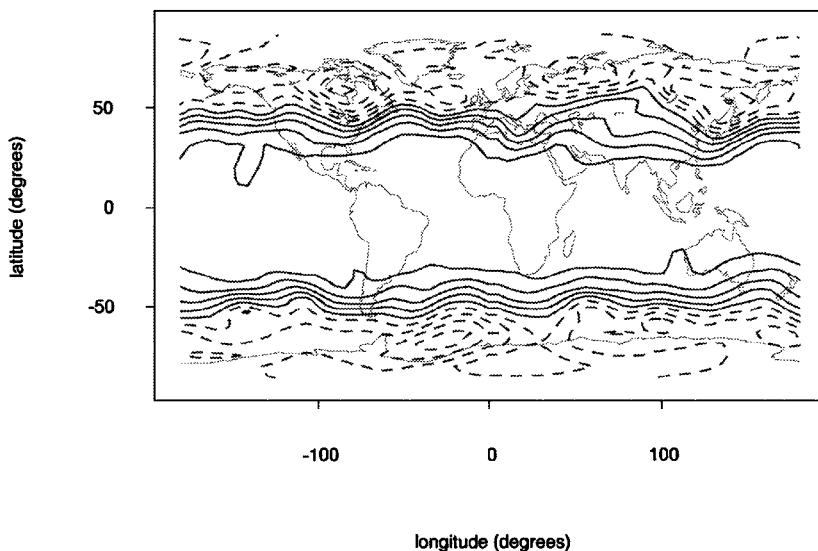


FIGURE 6.1. Contour plot of the 500 mb height at 1200 GMT 1 January 1986. Contours at levels 5300 meters and below are indicated by dashed lines. Contours are spaced 100 meters apart.

the atmosphere between the sea level and the 500 mb level. It relates to temperature, being low for cold values and high for warm values. The data were prepared by the National Meteorological Center in Washington and one reference to the method is Dey & Morone (1985). The period covered is 1200 GMT 1 January 1986 to 0000 GMT 6 January 1986. The time interval between data is 12 hours and there are 10 time slices. The geopotential is computed on a 64 by 32 global grid, (64 equispaced longitudes and the 32 latitudes 85.8, 80.3, 74.7, 69.2, 63.7, 58.1, 52.6, 47.1, 41.5, 36.0, 30.5, 24.5, 19.4, 13.8, 8.3, 2.8 North and South). The data are based on many observations and are interpolated to this regular array. They are meant to provide input values for numerical forecasts in particular.

The measurements of 1200 GMT 1 January are graphed as contours in an image in Figure 6.1. Values 5300 meters and below are indicated by dashed lines. One sees, for example, a depression over Hudson Bay. Further examination of the 10 such images shows the depression to move eastward and fill in over the eastern Atlantic on 5 January.

6.4 The Problem

The problem of concern is how to estimate the velocity of a moving phenomenon, such as the 500 mb field whose initial time slice is graphed in Figure 6.1. This field could be denoted $Y(x, y, t)$, but consideration will be restricted to motion along single latitudes. Denote the values along a given latitude, y , by

$$Y(x, t)$$

with t referring to time and x to longitude East. The model considered is

$$Y(x, t) = g(x - vt) + \epsilon(x, t) \quad (3)$$

with $\epsilon(\cdot)$ stationary noise. Here $g(\cdot)$ represents the moving shape and v its velocity. Because the Earth is a sphere, the function $g(\cdot)$ has period 360° . Fields that are periodic are considered in Yaglom (1962), Monin (1963), Hannan (1964), DuFour & Roy (1976).

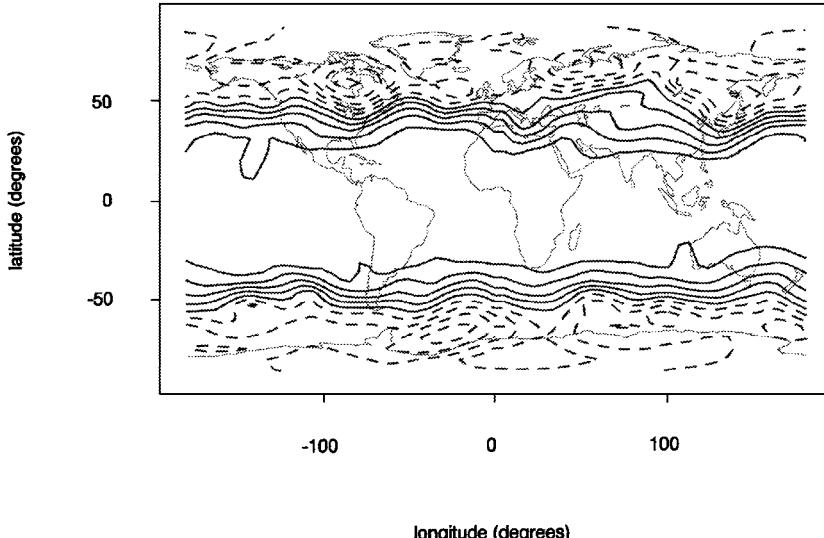


FIGURE 6.2. Lineal-temporal periodogram, $|d_Y^T(m, \lambda)|^2$, of the data values $Y(x, t)$ at 6 northern latitudes. Units of the y -axis are cycles/day.

6.5 A Fourier Approach

For the moment, let the longitude be expressed in radians, $\theta = 2\pi x/360$, rather than degrees. To reduce the effects of the presence of a fixed or slowly moving disturbance, the process $Y(\cdot)$ analyzed is that of the differences

$$Y(x, t) = Z(x, t + 1) - Z(x, t)$$

with $Z(\cdot)$ the original 500 mb values. The differencing operation enhances small features and makes the process values more nearly independent. Data are available for $\theta = 2\pi l/L$, $l = 0, \dots, L - 1$ and $t = 0, \dots, T - 1$. A first

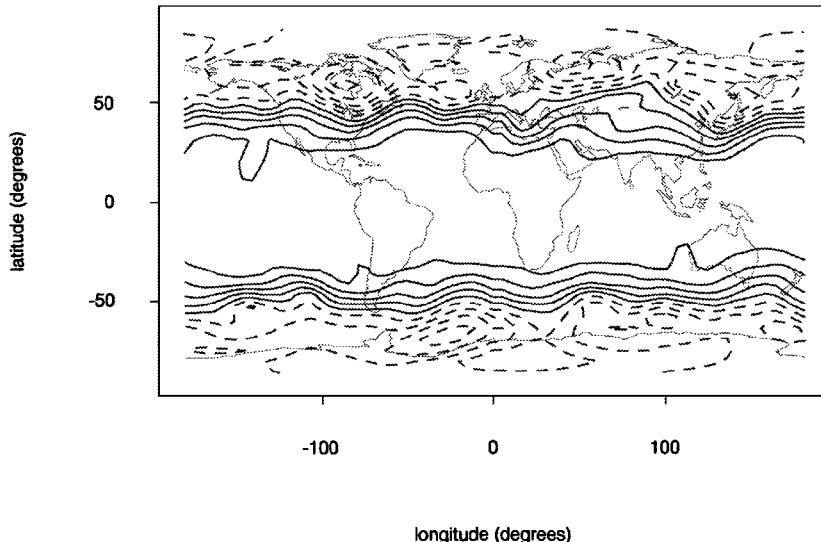


FIGURE 6.3. Plot of the multiple R-squared criterion (5) for the Fourier fitting procedure.

analysis will be based on the empirical Fourier transform

$$d_Y^T(m, \lambda) = \sum_l \sum_t Y\left(\frac{2\pi l}{L}, t\right) \exp\left(-i\frac{2\pi lm}{L}\right) \exp(-i\lambda t) \quad (4)$$

for $L = 64$ and $T = 9$. Suppose the Fourier series expansion of $g(\theta)$ is

$$g(\theta) = \sum_k \beta_k \exp(i\theta k)$$

$0 \leq \theta < 2\pi$ with k wavenumber. Evaluating the Fourier transform of $g(\theta - vt)$ as in (4) one obtains

$$\beta_m \Delta^T(\lambda + vm)$$

with

$$\Delta^T(\lambda) = \sum_{t=0}^{T-1} \exp(-i\lambda t)$$

The function $|\Delta^T(\cdot)|$ has principal mass near the origin, side lobes at $3\pi/T, 5\pi/T, \dots$ and period 2π . By inspection $|d_Y^T(m, \lambda)|^2$, as a function of m, λ , can be anticipated to have a ridge along $\lambda + vm \approx 0$.

Figure 6.2 presents contour plots of $|d_Y^T(m, \lambda)|^2$, the lineal-temporal periodogram, for the 6 northern latitudes 36.0, 41.5, 47.1, 52.6, 58.1, 63.7. Frequencies along the y -axis are in cycles/day. One sees principal peaks and suggestions of ridges, primarily at wavenumbers 5–10 and periods of 3–10 days.

The quantity $d_Y^T(m, \lambda)$ is a double Fourier transform. In the estimation of velocity v , it is simpler to work with the following single Fourier transform

$$Y_m(t) = \sum_l Y\left(\frac{2\pi l}{L}, t\right) \exp\left(-i\frac{2\pi lm}{L}\right)$$

$t = 0, \dots, T - 1$. This is $\beta_m \exp(-ivmt) + noise$ under the model (3). Elementary regression suggests consideration of the multiple R-squared type statistic

$$R(v)^2 = 1 - \sum_m \left| \sum_t Y_m(t) e^{ivmt} \right|^2 / \left(T \sum_m \sum_t |Y_m(t)|^2 \right) \quad (5)$$

as a measure of fit for a prespecified velocity value, v . To assess the optimal value of v , (5) is graphed in Figure 6.3 taking $m = 0, \dots, 12$. (From Figure 6.2 the principal mass is at $m \leq 12$.) The maxima of (5) are seen to occur at the velocities near 20° longitude/day. The largest value of the criterion is about .2 in each case and the peaks are fairly broad. This circumstance is reflected in the numerical estimates and associated uncertainties given in Section 7.

Hayashi (1982) reviews space-time spectral analysis methods and their applications to large-scale atmospheric waves.

6.6 A Nonparametric Approach

Consider again the model (3). In the case that the velocity v is known, but not $g(\cdot)$, (3) is the simplest case of the generalized additive model, (Hastie & Tibshirani 1990, Hastie 1992). In the case of unknown v , it is the simplest case of projection pursuit regression, see Friedman & Stuetzle (1981). The

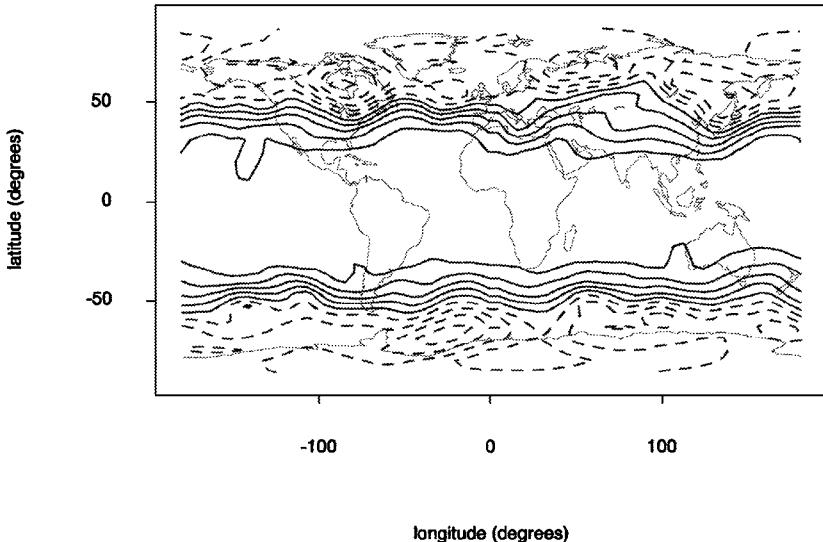


FIGURE 6.4. Plot of the multiple R-squared criterion (6) for the generalized additive model fitting procedure.

function $g(\cdot)$ may therefore be estimated in a variety of manners, see the preceding references. In the present case, natural cubic splines with equi-spaced knots are employed, (Hastie & Tibshirani 1990).

To start, v is viewed as known. Then $g(\cdot)$ is estimated, based on the data values $(x - vt, Y(x, t))$. As a measure of fit, multiple R-squared

$$R(v)^2 = 1 - \sum(Y - \hat{g})^2 / \sum(Y - \bar{Y})^2 \quad (6)$$

is again employed. Its values are graphed in Figure 6.4, as a function of v , for the 6 different latitudes. The peaks are much more prominent, but the maxima at the different latitudes are seen to occur at similar locations to those of Figure 6.3.

Figure 6.5 graphs the estimates of the function $g(\cdot)$ for the 6 latitudes. In a search for periodicities, Huber (1985) refers to such a technique as “A time series version of PPR (projection pursuit regression)”.

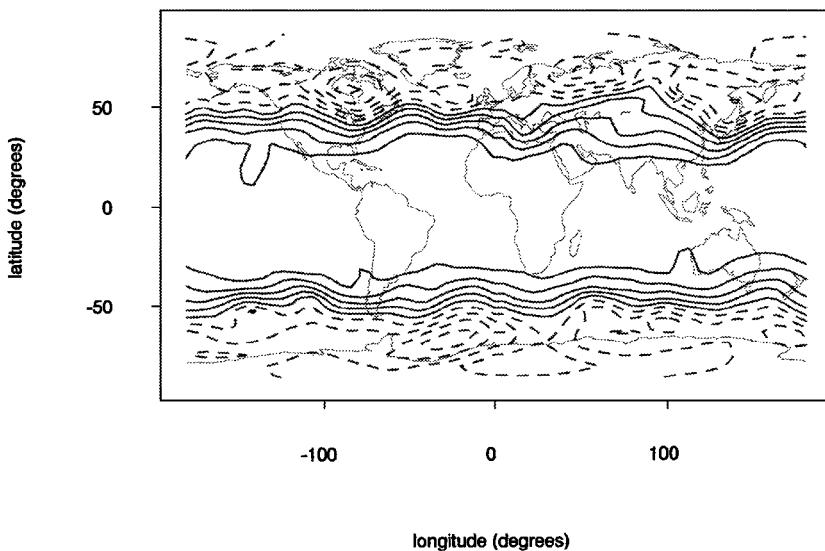


FIGURE 6.5. The estimated functions $\hat{g}(x)$ obtained via smoothing.

6.7 Uncertainty Estimation

For the moment it will be assumed that the noise process, $\epsilon(\cdot)$, consists of independent 0 mean common variance normals. To begin consider the Fourier procedure. The estimate maximizing (5) is maximum likelihood and there are classic formulas to approximate the standard error. In the present case, since the only parameter of concern is v , the simple procedure of Richards (1961) may be employed. It has the advantage that one can disregard the nuisance parameters. It involves expressing the estimates of the β 's, which are ordinary least squares, in terms of the parameter v and after this substitution determining the maximizing value of v . Surprisingly the second derivative of this likelihood of the single variable, v , may be used to approximate the standard error of \hat{v} . The values obtained are:

Latitude $^{\circ}N$	velocity $^{\circ}/\text{day}$	s.e. $^{\circ}/\text{day}$
36.0	14.6	28.4
41.5	16.5	23.4
47.1	17.4	31.2
52.6	21.5	45.7
58.1	19.2	33.8
63.7	19.7	37.5

These standard errors are large, compared to \hat{v} , as might have been expected from the broad peaks of Figure 6.3.

In the case of fitting the model (3) with $g(\cdot)$ smooth, the results are:

Latitude $^{\circ}N$	velocity $^{\circ}/\text{day}$	s.e. $^{\circ}/\text{day}$
36.0	13.9	5.9
41.8	17.3	7.4
47.1	18.7	14.0
52.6	22.8	16.6
58.1	19.4	10.4
63.0	18.8	10.6

These standard errors are notably smaller. This second procedure is apparently more sensitive and will perhaps pick up the fine features better. The spline employed had 24 nodes and hence 25 linear parameters. The Fourier technique also had 25 linear parameters.

6.8 Discussion and Summary

The results of the two analyses, the first a parametric Fourier and the second a nonparametric, are broadly similar. Advantages of the Fourier approach include: the model may also be examined by plots such as Figure 6.2, periodicity is handled directly, the whiteness of the noise may be studied and autocorrelation may be introduced into the model as necessary. The advantages of the generalized additive model approach include: the estimates are (apparently) more precise and an estimate of $g(\cdot)$ is provided directly.

In interpreting the results one needs to keep in mind the possibility of aliasing, i.e. that there is a disturbance going around the planet so quickly that it appears to be moving slowly when sampled but every 12 hours.

The model can be extended to

$$Z(x, t) = f(x) + g(x - vt) + \epsilon(x, t)$$

and the “fixed” component $f(\cdot)$ estimated. As indicated in Section 1 the model may also be extended to the case of several components moving with different velocities and directions, see (1). One could consider the automatic estimation of the smoothness and dimension parameters. Then the more general forms of generalized additive modelling and projection pursuit regression would be called for.

Acknowledgments: The 500 mb geopotential data and helpful comments on it were provided by Francis Zwiers of Environment Canada. Alastair Scott pointed out the broad usefulness of Richards (1961). The Editors gave

important assistance. The initial nonparametric analyses were carried out via the ace procedure of Breiman & Friedman (1985). Preliminary versions of some of the material were presented in the Wilks Lectures, Princeton University May 1989, and at the 1990 Annual Meeting of the Statistical Society of Canada in St. John's, Newfoundland.

Enfin, je veux dire, "Merci beaucoup Professor Le Cam pour tout le conseil, statistique et personnel, pendant tant des années."

The research was supported in part by the National Science Foundation Grant MCS-9300002 and the Office of Naval Research Grant N00014-94-1-0042.

6.9 REFERENCES

- Aggarwal, J. K. & Nandhakumar, N. (1988), 'On the computation of motion from sequences of images—a review', *Proceedings of the IEEE* pp. 917–935.
- Arking, A., Lo, R. C. & Rosenfeld, A. (1978), 'A Fourier approach to cloud motion estimation', *Journal of Applied Meteorology* pp. 735–744.
- Breiman, L. & Friedman, J. H. (1985), 'Estimating optimal transformations for multiple regression and correlation, (with discussion)', *Journal of the American Statistical Association* pp. 580–619.
- Briggs, B. H. (1968), 'On the analysis of moving patterns in geophysics—I. correlation analysis', *Journal of Atmospheric and Terrestrial Physics* pp. 1777–1794.
- Briggs, B. H., Phillips, G. J. & Shinn, D. H. (1950), 'The analysis of observations on spaced receivers of the fading of radio signals', *Proceedings of the Physical Society B* pp. 106–121.
- Brillinger, D. R. (1985), 'Fourier inference: some methods for the analysis of array and nonGaussian series data', *Water Resources Bulletin* pp. 743–756.
- Brillinger, D. R. (1993), 'Distributions of particle displacements via higher-order moment functions', *IEEE Proceedings-F* pp. 390–394.
- Brillinger, D. R. (1995), 'On a weather modification problem of Professor Neyman', *Probability and Mathematical Statistics*.
- Burke, M. J. (1987), 'Moving random surfaces and correlation analysis', *Radio Science* pp. 607–624.

- Carroll, R. J., Hall, P. & Ruppert, D. (1994), 'Estimation of lag in misregistered, averaged images', *Journal of the American Statistical Association* pp. 219–229.
- Cox, D. R. & Isham, V. S. (1988), 'A simple spatial-temporal model of rainfall', *Proceedings of the Royal Society A* pp. 317–328.
- Dey, C. H. & Morone, L. L. (1985), 'Evolution of the national meteorological center global data assimilation system: January 1982–December 1983', *Monthly Weather Review* pp. 304–318.
- DuFour, J.-M. & Roy, R. (1976), 'On spectral estimation for a homogeneous random process on the circle', *Stochastic Processes and their Applications* pp. 107–120.
- Friedman, J. H. & Stuetzle, W. (1981), 'Projection pursuit regression', *Journal of the American Statistical Association* pp. 817–823.
- Gupta, V. K. & Waymire, E. (1987), 'On Taylor's hypothesis and dissipation in rainfall', *Journal of Geophysics Research* pp. 9657–9660.
- Hannan, E. J. (1964), 'The statistical analysis of hydrological time series', *Proceedings of the National Symposium on Water Resources* pp. 233–243.
- Hannan, E. J. & Thomson, P. J. (1973), 'Estimating group delay', *Biometrika* pp. 241–252.
- Hastie, T. J. (1992), Generalized additive models, in J. M. Chambers & T. J. Hastie, eds, 'Statistical Models in S', Wadsworth, Belmont, California, chapter 7, pp. 249–308.
- Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman and Hall, London.
- Hayashi, Y. (1982), 'Space-time analysis and its applications to atmospheric waves', *Journal of the Meteorological Society of Japan* pp. 156–171.
- Huber, P. (1985), 'Projection pursuit', *Annals of Statistics* pp. 435–474.
- Kamachi, M. (1989), 'Advection surface velocities derived from sequential images for rotational flow field: limitations and applications of maximum cross correlation method with rotational registration', *Journal of Geophysical Research* pp. 18227–18233.
- Le Cam, L. (1961), A stochastic description of precipitation, in J. Neyman, ed., 'Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability', University of California Press, Berkeley, pp. 165–186.

- Leese, J. A., Novak, C. S. & Clark, B. B. (1971), 'An automated technique for obtaining cloud motion from geosynchronous satellite data using cross correlation', *Journal of Applied Meteorology* pp. 118-132.
- Marshall, R. J. (1980), 'The estimation and distribution of storm movement and storm structure, using a correlation analysis technique and rain-gauge data', *Journal of Hydrology* pp. 19-39.
- Monin, A. S. (1963), Stationary and periodic time series in the general circulation of the atmosphere, in M. Rosenblatt, ed., 'Time Series Analysis', Wiley, New York, pp. 144-153.
- Phelan, M. J. (1992), Aging functions and their nonparametric estimation in point process models of rainfall, in A. T. Walden & P. Guttorp, eds, 'Statistics in the Environmental and Earth Sciences', Arnold, London.
- Richards, F. S. G. (1961), 'A method of maximum-likelihood estimation', *Journal of the Royal Statistical Society B* pp. 469-475.
- Smith, J. A. & Karr, A. F. (1985), 'Parameter estimation for a model of space-time rainfall', *Water Resources Research* pp. 1251-1257.
- Tokmakian, R. T., Strub, P. J. & McClean-Padman, J. (1990), 'Evaluation of the maximum cross-correlation method of estimating sea surface velocities from sequential satellite images', *Journal of Atmospheric and Oceanic Technology* pp. 852-865.
- Yaglom, A. M. (1962), *An Introduction to the Theory of Stationary Random Functions*, Prentice Hall, New York.

7

At the Interface of Statistics and Medicine: Conflicting Paradigms

Vera S. Byers¹

Kenneth Gorelick²

7.1 Introduction

The association of Professor Le Cam with clinical investigators started in 1973, when he collaborated in the design of a clinical trial using an immunotherapeutic agent in patients with osteogenic sarcoma. Collaborations between statisticians and clinicians started quite early—British clinicians and statisticians were publishing studies together in the 1950's. However before the 1960's such collaborations were sporadic, and certainly not mandatory. By 1973 they were quite pervasive, so the history of Professor Le Cam's involvement in clinical research probably mirrored closely the development of collaborations between clinical researchers and statisticians. Those of us in clinical research have seen the growing influence of statisticians. The statistical-clinical relationship is an evolving one and remains, at times, rather rocky. This friendly antagonism is likely due to the institutional differences in perception and priorities between the two professions.

Statisticians by nature view the phenomenological world as a statistical universe. While individual nonconformists can muddy the waters, sufficient sample size can lead to a generalizable truth that can then be applied to future situations. The ultimate statistical design is after all, a prospective study in which the target Δ is anticipated.

Clinicians, on the other hand, view the world on a case-by-case basis, recognizing that each patient/pathology/therapy gestalt is unique, and must be addressed in a correspondingly unique manner. Fundamental discoveries concerning the workings of the body are constantly altering the paradigms of clinical practice. For example, prior to the discovery of slow vs rapid acetylation (a form of drug metabolism and inactivation) in different indi-

¹Allergene Inc.

²Stanford University

viduals only the clinician's experience could guide a patient's treatment to higher than usual doses of antituberculous chemotherapy.

Since statistical methodology evolves independently of medicine, clinicians often are bewildered by the ever changing statistical requirements for the design and interpretation of clinical trials. In theory, all clinical trials are subjected to intense ethical scrutiny. It is recognized that the study patient is subjected to the risks of a clinical trial, and these patients may gain little in direct benefits, particularly if they are placed on an inferior treatment arm. Their interests are balanced against those of the future patients who will benefit from a scientifically sound trial, regardless of outcome. Although a study's rationale often appears cogent to the institutional review boards (IRB's) that must approve all clinical studies, in actual practice they often fall short of the mark, leaving physicians caught between the Scylla of statistically "bulletproof" trials (often based on guesstimated treatment effects) and the Charybdis of medically optimal human research (predicated on the uniqueness of the individual). In the following pages, we present our view of some of the statistical principles that have gained ascendancy in this new age in which clinical trials are only found valid if their statistical underpinnings are sound. We also will describe some of the ethical dilemmas of matching and marrying ethics and statistics in these settings.

The sophisticated clinical studies of today had their roots in case reports, authored by general practitioners in the last century, and sent in to journals such as the British "Lancet", or the newly formed (1812) American "New England Journal of Medicine and Surgery". These reports, now termed "anecdotal" were simply a method by which busy general physicians could transfer ideas to each other in the absence of the medical meetings held today. The initial published reports of general anesthesia, or the value of antisepsis would never meet today's more rigorous standards.

With time it became apparent that the larger the numbers of patients in the group the more valid the data, but statistics were not really drawn into the picture until drugs began to be evaluated for toxicity and efficacy. Since the 1950s, a number of important statistical advances have taken place, from the development of the odds ratio and the Kaplan-Meier survival curve, to log rank test, Cox regression and many others (Altman & Goodman, 1994). There has been considerable lag between the development of a new statistical methodology, its introduction to clinical research, and finally its acceptance by clinicians as a useful tool (Altman & Goodman, 1994). This is consistent with the Hippocratic oath, which urges physicians to "...be neither the first nor the last..." to introduce new techniques into their armamentariums.

Because of the vast flexibility of the human mind, and the ways that it can muddy data interpretation, sophisticated statistical tools must be used to yield the greatest degree of objectivity in the analysis of clinical

data. A great leap forward in statistical analysis resulted from the work of Professor Jerzy Neyman, Professor Le Cam's mentor, and our whimsical and brilliant colleague. Professor Neyman established the basic statistical theory for hypothesis testing, laying the foundation for research techniques of randomized clinical trials. He also defined the two kinds of errors and the concept of power. These tools, together with others given to us by his statistical colleagues allowed us to integrate statistics into our clinical life. Once this was done, statistically correct clinical trials could be designed, and we could then for the first time assess how the human factor interfered with the dispassionate analysis of data.

7.2 The Jargon of Clinical Trials

Over the years the design, performance, and analysis of clinical trials has become a medical discipline of its own. While clinical trials can study a variety of objectives, we will focus on those intended to evaluate the safety and efficacy of new drugs. In therapeutic clinical trials, three phases of clinical research have been generally recognized, and even codified by the Food and Drug Administration (FDA). First, the toxicity of a drug is evaluated in a relatively small—6 to 30 patients—Phase I study. In many cases, e.g., for cancer chemotherapeutic agents, the purpose of Phase I is to discover the nature of drug toxicity, the maximum tolerated dose (MTD), and the way in which the human body handles the drug (pharmacokinetics). Patients usually have advanced disease rather than minimal disease, since the purpose of this phase is to conduct an initial test of toxicity in a population for which there is no alternative therapy. The safest study design is often a cohort dose escalation, in which sequential groups of 3–5 patients are treated at one dose, evaluated, then the next group is treated at a prospectively defined higher dose, until the MTD is determined. Phase II studies typically include 30 to 50 patients, and examine different variables such as stage of disease, concomitant medications etc., that must be optimized before initiation of definitive studies. Although demonstration of efficacy is not usually required at this stage, it is hoped that some efficacy will be seen to justify the resources required for a phase III study. If the drug has important toxicity, efficacy may be required even at this early phase to justify exposing the larger number of patients in the phase III study to this risk. Phase III, or pivotal studies consist of at least two arms, control and treatment, and optimally are randomized and blinded although the latter is often impossible. In the past historical controls were often used for the control arm, but these are so fraught with potential errors that it is considered unwise to invest in an expensive study using anything but concurrent controls.

Clinicians today agree that blinding of treatment arms is key to the con-

duct of an interpretable study. Ideally studies should be double-blinded; i.e., neither the investigator nor the patient is aware of the treatment assignment. Proper randomization prevents selection bias, so the sicker patients are not assigned to the treatment arm. Blinding patients to their treatment will control for the "placebo effect", which can be very powerful. Equally important is blinding the physician to the treatment, to prevent the introduction of bias into either concomitant therapy or treatment evaluation.

The characteristics of study patients are defined by the entry criteria, which include inclusion and exclusion criteria. Thus for example, a Phase I breast cancer study may be designed with inclusion criteria that only allow admission of patients who have failed standard therapy since these patients have no other hope of survival.

The exclusion criteria may be used to focus on a population that would be expected to benefit from the therapy. For example, breast cancer in pre and post menopausal women is very different in its response to treatment. Premenopausal women usually have cancer cells that have estrogen receptors. Since estrogen stimulates its growth, if the drug is an estrogen antagonist, the exclusion criteria in this study might list those who are post menopausal. In the phase III studies of drugs for pharmaceutical companies, patient selection criteria become critically important since they, together with the endpoints, define the label indications. This has important financial implications since advertisements for the drug must be limited to the label indications. Even though physicians may use the drug at other times in the same disease, or even for other diseases, these other indications cannot be referred to in advertisements, and without this, wider use is quite limited. A pharmaceutical company must make a decision as to whether it is better to obtain relatively quick drug approval for a small but life-threatening indication, or a slower approval for an indication which is not life threatening, but has larger commercial potential.

The interaction of study design, regulatory approval and insurance reimbursement has resulted in a very murky, amorphous area. While the FDA does not regulate the way physicians practice medicine, insurance companies find the restrictive nature of the approvals much to their liking. By limiting reimbursement for drugs to their use in "approved" indications, they limit their requirement to pay for new medical discoveries. While in the past, drugs could be approved for broad indications on the basis of a few suggestive studies, today there is a strong tendency to approve drugs only for the exact population selected in phase III trials, and for only the treatment of the exact disease studied. Thus, the evolution of statistical methodology, regulatory oversight and control of medical practice by third parties has turned scientific methodology into a tool of social engineering that limits the availability of novel medical treatment to those who can afford them out of pocket.

Data from clinical trials are collected in Case Report Forms (CRFs).

These documents, which can run as long as 100 pages/patient, are filled with demographic data, results of tests on the patient required at very specific times (study activities), adverse events, other medications taken, and other data defined by the study protocol. The CRFs are filled in by the physician running the study (Principal Investigator) and verified by a second person, a study monitor. Data quality assurance has become an independent specialty. This is extremely labor intensive, since the results of every test must be verified, and an explanation must be provided for every test that is not run, or is run at the wrong time, together with an explanation of why it does not interfere with the analysis. On a typical evaluation day, there may be 100 tests run, and it is not unusual to evaluate patients weekly or monthly for two years.

For a study intended to win FDA approval, the data are submitted to the FDA. Typically two Phase III studies are required for approval. These studies may be very large, on the order of 2000 patients, or as small as 100 patients, depending on the disease, the drug and the study design. It can be easily seen that the work of a clinical trial can be evenly divided between actually treating patients and documenting the results, and performing the rigorous data quality assurance and construction of analysis data sets required before analysis.

A joint frustration of clinician and statistician is the trial that fails to meet its prospectively defined objectives, but nonetheless contains tantalizing data that cry out for additional post hoc analyses. While the original analyses may be completed promptly once the trial is over, information gathering through incremental analysis can take a long time and may be necessary to turn a marginal study into one that reflects the Sponsor's beliefs in the true value of the test article.

7.3 Sample Size Calculation

The application of appropriate statistical methodology is not only needed to meet regulatory requirements for approval of new drugs, but also to define the best medical care for patients. Much medical practice falls under the heading of "medical art"—customs that arose from anecdotes of patient care, often supported by small, inadequate trials that lend an aura of respectability to essentially unproven ideas.

Thus, the impact of a well designed clinical study on the way in which medicine is practiced can be staggering. A very real example of the value of sample size calculations and applied statistical methodology is the EC/IC bypass study conducted between 1977 and 1985.

Stroke is one of the most feared complications of atherosclerosis. In 1967 a technique was developed to supplement blood flow to the brain by connecting blood vessels from outside the head (extracranial) to those inside the

head (intracranial), that were distal to a blockage. This procedure, termed extracranial/intracranial bypass (EC/IC bypass) became one of the most common neurosurgical procedures in the world. From 1975 to 1985 over 100 articles were published in the medical literature discussing the use of this technique in occlusive cerebrovascular disease. These consisted essentially of uncontrolled case reports of 20–100 patients successfully treated with this operation. In 1977 a worldwide, multicenter trial was initiated to compare this operation to standard medical therapy. The study was statistically based; there were rigorous entry criteria and prospective determinations of sample size and interim analyses. With a rate of stroke or death in a medically treated group estimated at 24%, and an expected benefit in the surgically treated group of 33%, setting a one-tailed alpha at 0.05 with power of 90%, the target sample size was estimated at 482 per arm. (EC/IC study group 1985)

Overall, 71 centers enrolled 1377 patients over 8 years. The study, reported in a variety of publications, indicated that this procedure was of no benefit to patients, (Haynes et al. 1987). In contrast to the small, nonrandomized and uncontrolled positive studies that supported the widespread use of this technique, a single well designed study resulted in its effective demise. From 1986 to 1994, fewer than 10 publications on EC/IC bypass for cerebrovascular ischemia have been listed in Index Medicus.

Clearly, the discipline of proper study design including planned sample sizes based on power and clinically significant differences, should be applied to define the value of other generally accepted surgical procedures, such as organ transplantation.

Statistical inference is critical to the design and evaluation of clinical trial data. Statisticians like to evaluate ever larger universes, but clinicians look to "n of one" clinical trials. (Hodgson 1993) The FDA guidelines for drug evaluation call for "adequate and well controlled clinical trials" and specifically address the issue of sample size. However to do this, one must predetermine the acceptable probabilities of type I and II errors, estimate the placebo group response, and the magnitude of treatment effect. Since Phase I and II studies are usually small, and the precise end points are often not well characterized in the literature, either extensive experience with the disease or a true understanding of the biochemical/cellular abnormalities causing the disease is needed to estimate the effect of treatment. Sample size is therefore not a statistical question, but rather a clinical one. The physician must decide the magnitude of difference expected between the treated arm, and control arm. Once this has been accomplished, the alpha is set at 0.05, the beta at 0.1 to 0.2, giving a power of 80 to 90%, and the sample size is calculated. Obviously a mistake in estimating the extent of difference between the groups may produce a false negative result. This will invalidate the entire trial, and may damn a drug forever simply because it is not quite as effective as had been predicted.

Acquiring the information necessary for these predictions is complex and time consuming. In our experience, even when clinical trials are based in medical truisms, it is often (if not always) impossible to determine reliably the true natural history of a disease or intervention.

A good example of the problems which can result from unknown types of heterogeneity in the patient population under study is shown from the clinical trials with monoclonal antibodies, thought to be effective against gram negative sepsis and septic shock, life-threatening complications of infection caused by gram negative bacteremia (bacteria in the blood), and an antagonist to interleukin 1, a critical mediator of sepsis and systemic inflammation. A crucial challenge lay in the selection of patients who actually had gram negative sepsis.

The syndrome of sepsis is clinically apparent long before an exact diagnosis of its cause can be made. It is clinically recognizable and characterized by several findings, including fever, hypotension, decreased urine flow and rapid breathing and heart rate. The syndrome can result in rapid death, or the development of complications that lead to death over several weeks. In either case, clinicians believe that intervention at the earliest possible point is essential for any new treatment to show effectiveness. Since proving an etiologic diagnosis can only be made by culturing infected tissue or fluids, and this takes 48 hours, it is necessary to treat the patients without a definitive diagnosis. A set of criteria was therefore developed (Bone et al. 1989) to provide consistency in the diagnosis of sepsis. This was considered a major advance in the ability to compare various therapeutic agents. Protocols were developed by which all patients meeting these criteria were treated.

On this basis, four huge trials were designed, again using similar entry criteria. (Greenman et al. 1991; Wenzel et al. 1990; Ziegler et al. 1991; Fisher et al. 1994) The results in all the studies were equivocal, and the subgroups of patients who benefited from the therapy were different among the different trials.

A study running simultaneously with the therapy studies investigated the predictive value of the Bone criteria for bacteremia (Bates & Lee, 1994). This study showed that the frequency of documented infections of the types sought was much lower than estimated from the Bone criteria. Had the clinical investigators waited for the results of this study, their trial designs would surely have been different. However, the total number of patients required would have been much larger, maybe prohibitively so, and their studies might have been delayed by up to seven years and the sample size estimates might have increased prohibitively. This illustrates the difficulties involved in using medical truisms as patient selection criteria, but also highlights some of the practical difficulties in trying to bring new medical technology to fruition in a rapidly changing environment.

The findings of the sepsis trials have led to some basic changes in the

way clinicians look at sepsis, septic shock and bacteremia. (Bone RC et al. 1992) While this is a worthy benefit, the companies that were developing the drugs experienced a decrease in market value of over \$4 billion, and as a result the development of new drugs for this still untreatable, lethal disease has slowed remarkably.

It is worth noting that clinical trials showing positive results in small sample sizes are often published. Several of the large sepsis trials were preceded by positive smaller trials (Fisher et al. 1994b; Greenberg et al. 1992). We do not know how many small clinical trials with negative results are not published, but would assume that the positive ones constitute a minority (5%) of the total. Determination of appropriate sample size is unquestionably necessary for the conduct of statistically meaningful trials. On the other hand, inflexible adherence to this principle may impede the development of needed treatments for uncommon or ill characterized diseases.

Another example is in the development of aerosolized pentamidine to prevent pneumocystis carinii pneumonia (PCP), a common cause of mortality in AIDS. While AIDS is arguably the most intensively studied disease of our time, in 1986, four years after the disease and its association with PCP had been identified, clinicians still did not know what the attack rate of PCP was in patients diagnosed with AIDS. This made design of a study impossible until the data were obtained.

7.4 The Placebo Effect

Clinical researchers have developed a healthy respect for the placebo effect. In strictest terms, this is a biologically based phenomena in which the conscious or subconscious mind is able to induce multiple endogenous chemicals which will improve health and survival simply because the patient anticipates a beneficial effect from being in the clinical trial. Examples of such "mind over body" situations abound in clinical medicine, and anyone who works with severely ill patients is familiar with the cancer patient who keeps himself alive long enough to see his son graduate; the AIDS patient who completes his long awaited cruise then dies within a day of its completion; the heart attack victim who dies the day after his birthday. The mechanisms by which this is accomplished are just beginning to be defined and range from induction of complex chemicals such as adrenaline, to neuro-immuno-endocrine protein cytokines.

The placebo effect has been well established as occurring in clinical situations which are relatively minor such as skin test reactivity used to assess allergic reactions. Its magnitude was almost inevitably on the order of 30%. Statisticians such as Professor Le Cam had begun to develop concern that this effect could be much more powerful than that—and more troublesome

for interpreting clinical trials. In point of fact, improvement in skin tests has much the same biologic basis as regression of tumors and increased survival, since both involve induction of the same cytokines, but it certainly confuses interpretation of the clinical trials.

In one study, levamisole, an antihelminthic agent, was evaluated in a randomized blinded study to assess its efficacy in treating malignant melanoma, one of the most virulent of the cancers (Spitler & Sagebiel, 1980). When the study was analyzed it was concluded that the drug was ineffective, with no difference observed between treated and control groups. A similar negative result was found in a subsequent study (Spitler 1991). However when the survival time of both treated and control groups was compared to those of the community (historical controls), both experimental and control groups had a 30% survival advantage (Benson & McCallie, 1979, Meyers, 1979).

This example clearly illustrates how the use of historical controls to design a study can be deceptive; in this example the drug outperformed historical, but not concurrent controls. Still—is there a placebo effect in cancer? A Canadian study of the same drug in the same population, run at about the same time, was positive, reporting a survival benefit among the levamisole treated patients of about 30% (Quirt et al. 1991). This study was not blinded, so the experimental patients and the investigators knew a potentially therapeutic drug was being given. Perhaps the Heisenberg principle has invaded clinical trials, and the simple fact of observing a phenomenon has altered it. This hypothesis is not facile; patients in control arms of clinical trials often outperform concurrent patients not in a clinical trial. For example, one site in an interventional trial of a sepsis drug also participated in a multicenter observational study to assess sepsis frequency and outcome. During the interventional trial, sepsis frequency and mortality plummeted in this center without changing in any of the other centers involved in the observational study (Iannini, P. personal communication). This simple observation modifies the way clinical trial data can be integrated into clinical practice.

How should clinicians interpret these apparently conflicting data? More detail than is published would be needed to reach a sound conclusion. Should this drug be used in patients who have no alternative hope for survival? Taking a population view, use of this drug would not be expected to yield a net benefit, and therefore it should not be indicated. As clinicians, we cannot dismiss the 30% benefit seen in patients treated with this drug; we cannot, however, dissect out the relative importance of the therapeutic intervention in the beneficial effect observed in the Canadian study. Perhaps the mere belief that something beneficial is being offered is sufficient to improve outcome. In this case one must balance risk and benefit; a relatively safe drug given to a terminally ill patient either for pharmacologic or psychologic benefit does not violate the principle of *primum non nocere*.

7.5 Placebo Controls

Ethical trial design mandates that subjects in the control arm be given appropriate therapy by current medical standards. Appropriate therapy may consist of a drug regimen which is recognized to be mediocre, since no better therapy is currently available, or a placebo which is truly inactive in cases where observation alone is considered appropriate therapy. Placebo treatment is considered superior to untreated controls, because blinding can be maintained. Sometimes placebos may carry their own risks, as in the case of intravenous administration of study drug to severely immunocompromised AIDS patients. The requirement for placebo controls is one of the most difficult ethical issues in clinical trial design. Apart from health issues, a clinical trial is labor intensive for the patient, who must submit to tests that are often painful, and disrupt his/her life to keep frequent clinic appointments. Thus, sometimes it is necessary to accept less than ideal statistical design to accommodate ethics and practicality.

An example of this is one of the vaccine trials currently being run by the Southwest Oncology Group in the U.S. Stage II malignant melanoma is not universally fatal; in fact the prognosis is sufficiently optimistic that there is no standard treatment that is not more harmful than observation. Patients with stage II melanoma are usually managed expectantly; they are requested to see a physician frequently, to increase the chances of early detection of a recurrence.

Vaccine therapy for melanoma is being evaluated in a randomized controlled trial of vaccine, given after the primary tumor is completely surgically removed. Patients in the experimental arm are given repeated injections of an allogeneic tumor vaccine prepared from lysed melanoma cell lines every week for two years. Since it is possible that vaccination may be of no benefit, or even increase the rate of recurrence of the tumors, a control arm consisting of patients receiving no treatment was considered both ethical and scientifically essential. However, if a placebo injection would have been used, patients in the control arm would have to endure the painful injections without prospect of benefit. This was considered unethical by the physicians responsible for the study's implementation, so the study is randomized but not blinded. If instead the drug could have been administered as a pill, then a control arm consisting of patients treated with a placebo would have been considered, and the study could have been blinded (V. Sondak, Personal Communication).

In a serious disease such as AIDS or cancer, particularly when the patient population being treated has no other alternative therapy, it is difficult for the clinician to justify treating patients with an inactive drug, especially when the trial is blinded so the patient is not told that the drug he/she is receiving is inactive. Of course the experimental arm may turn out to be inactive or even toxic, and this is the risk that is offered the study patient,

hopefully balanced by a potential benefit from the experimental therapy. In point of fact however there are often other agendas. For example, in the 019 study of AZT in patients with HIV disease, a previous blinded study had shown that high doses of AZT were successful in preventing a drop in CD4+ cells as compared to placebo. However high dose AZT had toxic side effects. Therefore it was necessary to design another trial comparing high dose and moderate dose AZT, to see if the moderate dose was as effective but less toxic. It was felt that a placebo arm, consisting of completely inactive drug, was necessary, so the trial was designed as a three armed study. At the time, enthusiasm was very high for AZT, and both physicians and AIDS activists felt AZT was better than no AZT, so it was difficult to persuade patients to enroll in the trial. They ultimately completed enrollment, because the drug was offered free to trial patients, and was very expensive for many patients to otherwise afford. The ethics of this were questionable, given the potentially coercive nature of the offer of free drug to indigent patients. AIDS activists developed their own solution; they had pills given to the study participants analyzed by an outside laboratory, and those patients who found they were on placebo simply dropped out of the study and paid for the active compound!

A placebo arm may be just as valuable in defining true toxicity as in defining true efficacy. For example, many breast cancers have their growth stimulated by estrogen. Therefore eliminating estrogen from the body will slow growth, and as a result will prevent growth—and produce death—of micro metastases. To avoid removing the ovaries, tamoxifen, which can block the estrogen receptors on the breast cancer cells, is very useful. However it is expected to have side effects resembling menopause.

Early toxicity and dosing (Phase I and II) studies are usually not blinded or randomized. These studies showed that tamoxifen produced a whole range of expected side effects including menstrual irregularity, vaginal discharges, , headaches, nausea, hot flashes, and abdominal cramping. In the blinded phase III study 2644 patients were treated in a randomized, double blind, placebo controlled study (Fisher et al. 1989). These same side effects were seen in the treated group. They were also seen in the control group. In fact, it required statistical analysis to determine the true side effects. When the reactions in the two groups were compared, it became obvious that the only true differences between them, and the only true side effects, were hot flashes (40% in the placebo arm, and 57% in the treated arm), vaginal discharge (12% in the placebo arm and 23% in the control arm), and menstrual irregularities (15% in the placebo arm, 19% in the treated arm).

Clearly, the application of blinded, placebo-controlled methodology has unmasked the nature of the beneficial effects of some drugs, and demonstrated that the true efficacy lay in a far more complex psycho-physiologic phenomenon, the placebo effect. While the nature of the placebo effect is

not fully understood, medical practitioners agree that an important component of this effect derives from the therapeutic relationship between physician and patient. By eliminating the beneficial effects of this interaction from the evaluation of new therapies, blinded, placebo controlled trials clearly establish the minimal level of effect that could be seen when a drug is administered by a physician who is unfortunate enough to be unable to establish this psychic bridge.

This leads to the first horn of the medical-statistical dilemma: how does one balance the requirements of proper statistical design with the medical mandate to act always in the best interest of each patient? The second horn is virtually a corollary of the first: once "data" in the form of controlled clinical trials exist, how can patients receive proper treatment that may fall outside the bounds of the trial's results?

7.6 Appropriate Treatment for Controls

One of the most important features in the design of the clinical trial is the treatment of the control group, and this is often ethically the most difficult feature. An experimental design in which an active drug is compared with a placebo is the simplest, and usually requires the fewest patients, leading to more rapid evaluation and ultimately drug approval. However if a standard therapy already exists, it is usually considered necessary to compare the new experimental therapy with the standard therapy. If the new therapy is expected to be very superior to the standard one, the number of patients required can be manageable. If the two drugs are expected to be equivalent, then the number of patients required to show equivalence can be very large indeed. One example of this ethical dilemma is described below:

Estrogen is truly a wonder drug for post menopausal women. Given early, it prevents hot flashes and other immediate results of estrogen deficiency such as vaginal dryness, and it also blocks the insidious and gradual onset of osteoporosis caused by estrogen deficiency. This leaching of calcium from the bone results in hip fractures, a major cause of mortality in older women, which condemns many more to spending their last years plagued by the crippling complications of frequent fractures of the spine, hips and other important structures. Nothing can restore calcium to the bone once it is removed, but use of estrogen early after menopause stabilizes it in the bone. However estrogen is contraindicated in women with a strong family history of breast cancer, and alternative non-estrogen drugs are needed.

Etidronate is a non-estrogen drug which was expected to be effective in osteoporosis. A randomized blinded clinical trial was designed in which 66 postmenopausal women (in whom estrogen therapy was not contraindicated) were given either etidronate, or placebo while estrogen was withheld. (Storm et al. 1990). Patients were studied for 150 weeks. During the first

60 weeks of the study, no differences were observed between the treatment groups in terms of spinal deformity (an indication of vertebral fracture) or fracture rate. Only after estrogen therapy had been withheld for nearly three years, was there enough deterioration in the control group to indicate a statistically significant benefit of etidronate compared with placebo. This trial, on a relatively small number of patients served as one of the pivotal studies on which the drug was approved. However it left the control group with significant osteoporosis and at risk of fractures, since the calcium they lost is irreplaceable.

It is clear that a statistical design to compare etidronate to standard, effective therapy would have required many more patients and much more time. From the perspective of drug development (fast time-to-market) and academic advancement (publish or perish), this was a well designed and conducted trial. However, from a more Helsinkian³ and Hippocratic perspective these physicians failed to provide their patients with the best available treatment. While patients did sign "informed consent" documents that apprised them of alternatives it strains the credulity of the observer to believe that these women made a thoughtful decision to accept the possibility of crippling fractures in order to permit the more rapid evaluation of a novel treatment. It is more likely that patients were influenced by the investigating physicians to accept a therapeutic option (no treatment) that would have been considered inappropriate outside the context of a "study". While a useful drug was thus more rapidly made available, the intentional withholding of beneficial treatment, even with the "informed consent" of the subjects makes this study ethically suspect.

Design, conduct and analysis of the larger and longer estrogen controlled trial which would meet standards of ethical research would have taken much longer, because of the complicated double-checking requirements of the FDA for record keeping and analysis. Etidronate is now on the market and available for those women who cannot take estrogen; It would probably not yet be there if the above study design had been used. Thus the ethical dichotomy—Is it more valuable to society to have 33 women with irreplaceable calcium loss and a useful drug brought swiftly and relatively inexpensively to market, or to protect the health of those volunteers at the cost of delayed access for the thousands of patients who stand to benefit from the drug?

³The Declaration of Helsinki is an international treaty that guarantees the rights of experimental subjects, including the right to receive standard therapy during a trial.

7.7 Randomization and Blinding

Physicians often object to double blinding a study in which the drug is toxic, or the untreated disease may be lethal. They may therefore design single-blinded trials, where the physician is aware of treatment group assignment. Unblinded investigators risk introducing bias into the study, and invalidating its results. The ways this can be done are legion, including assigning sicker patients or the treatment arm, or treating the patients in the experimental arm differently from those in the control arm. This can be minimized by very precise study design, but can never be eliminated in a trial that is not double-blinded. In a recent review (Schultz et al. 1995), 250 controlled trials were analyzed and it was found that when the trials were not double blinded, or when the assignment of patients was inadequately concealed from physicians, authors exaggerated the effects of the treatment by as much as 41%. They conclude that concealment of the allocation is essential if the goals of the study are not to be undermined, whether by subconscious bias or conscious subversion.

Randomization and blinding are necessary to eliminate confounding factors such as the placebo effect, on the side of the patient, and differential treatment on the part of the physician. The requirement by regulatory authorities that approval of drugs be made on the basis of "adequate and well controlled clinical trials" with randomization to a concurrent control and double blinding of treatment group assignment has been an important step in improving the scientific rigor behind clinical trials that support new drug approvals. The effect of this progress is strikingly shown by one example.

Khellin was a precursor of cromolyn sodium (Intal) an effective anti-allergy drug. Initially, however khellin was thought to be effective in angina pectoris (Bleich & Moore, 1979), a symptom of coronary artery disease that often leads to heart attacks if left untreated. An early report, which evaluated the efficacy of Khellin in 250 patients with angina pectoris, indicated that the drug was effective in 90% of cases (Anrep et al. 1949). In another 23 patient study, 83% of patients enjoyed complete remission or marked improvement of symptoms (Ayad, H. 1948). Although these studies were open label and non-randomized, investigators expressed confidence that their findings could not be imputed either to chance or to a placebo effect (Dewar & Grimson, 1951). This belief was then put to a rigorous test, in a series of randomized studies. In a harbinger of problems to come, the studies were open label, because physicians were reluctant to treat patients with potentially lethal symptoms such as angina, without knowing the nature of the treatment administered. These studies indicated clinical response rates of 70–80%. Outcomes were reported not only in terms of subjective criteria, like chest pain, but also in terms of nominally objective criteria such as exercise tolerance tests (Leiner & Dack, 1953, Osher et al. 1951).

Despite positive results of clinical trials and widespread acceptance of the

drug, skeptical investigators began to apply more sophisticated methodologies to the study of khellin in angina. These included proper controls, randomization and double blinding. The first of these (Greiner et al. 1950) concluded that khellin was no better than placebo pills for control of angina pain. A second study confirmed Greiner's findings Leiner & Dack, 1953). By 1954 reports on the efficacy of khellin in the treatment of angina pectoris had faded from the medical literature.

While the traditional means of introducing new techniques (experience and anecdote) still prevails for surgical procedures (e.g., coronary bypass surgery) and other nonpharmacological interventions, the development of a powerful regulatory infrastructure for pharmaceutical products has, over the past few decades, resulted in statistically rigorous (and perhaps overly rigorous) review of new drugs prior to their approval and marketing.

All this discussion highlights one point about clinical trials: they are often dangerous. A poignant and unusual example of this is shown by a recent study testing the ability to desensitize patients allergic to peanuts, by injecting them with gradually increasing doses of peanut extract. The objective was worthy since a surprising number of deaths caused by anaphylactic reactions to foods (a reaction in which patient develops acute airway closure and/or severe drop in blood pressure immediately after eating the food) was due to unwitting ingestion of prepared foods containing peanuts. It was well designed, being double-blind and placebo-controlled. Unfortunately a formulation error in the pharmacy resulted in the accidental administration of a high dose of peanut extract to a placebo treated subject. The subject, a young man, died of anaphylaxis (Oppenheimer, 1993).

Another facet of conducting controlled studies is illustrated in a study of an antifungal agent, itraconazole, that was studied in patients with pulmonary cryptococcosis. This often fatal fungal illness frequently affects the central nervous system. While the fungus enters the body through the lung, it can disseminate, and be especially devastating in HIV-infected individuals. An historical review of 37 patients with cryptococcosis treated in Rwanda showed that 29 had isolated pulmonary disease of whom four died shortly after hospitalization from unspecified causes. Ten of the remaining patients received no treatment for a variety of reasons including war and lack of medical care. Seven of these developed disseminated disease in less than one year and all but one were dead within three years. Fifteen patients were treated with an antifungal drug, itraconazole (12) or ketoconazole (3). One ketoconazole-treated patient developed disseminated disease, and seven of 15 patients were alive at follow up periods up to three years. (Batumwanayo et al. 1994) Given these data, it is hard to see how an ethical controlled study could be designed and it is equally hard to see how a regulatory agency could approve the drug for this indication. Thus, while the drug is available, it is not approved for this indication and it is

illegal for the manufacturer to inform physicians of the results of this trial or to suggest that it may be useful in patients with this illness.

Clinical trials such as these certainly emphasize difficulties in the ethics of subjecting experimental subjects to risky procedures in order to benefit the common good. The ever increasing number of clinical trials is adequate evidence that society still thinks the common good is dominant. The challenge is to make these trials less dangerous but still statistically sound by innovative design.

7.8 Risk-Benefit Analysis

Perhaps one of the most difficult parts of the dance between medical ethics and statistical requirements is the decision as to whether the risks and benefits of a therapeutic agent can be balanced to allow an ethical trial. For example, even an agent with very toxic side effects can be ethically tested if the patients have a potentially fatal disease, such as some types of cancer, and if the agent can be effectively tested in patients who have exhausted all other therapies and who are at a stage in their disease where they will almost certainly die anyway. The rationale is that there is a chance that the drug may benefit them, and if it does not, or even if they die from the toxic effects of the drug, the effect on outcome is insignificant.

This is illustrated by results of one of the first studies to introduce combination chemotherapeutic regimens to therapy of breast cancer (Fisher et al. 1981). Patients in the treatment arm were treated with phenylalanine mustard or with 5-fluorouracil, both highly toxic chemotherapeutic agents. At the end of five years, the disease free survival in the placebo treated group was just over 46%, that of the chemotherapy treated group was 55%. This was statistically significant ($p < 0.001$). However in the control group, of the 505 patients begun on the trial, 436 (86%) of the patients completed the study, with the majority dying of breast cancer. In the experimental group, of the 525 patients starting the study, only 299 (57%) completed the study, with the majority dying of the chemotherapy. Clearly the only advantage of this regimen was the ability to pick your cause of death—cancer or chemotherapy. Nevertheless it was widely used, with physicians telling their patients, “If you are tough enough to survive the chemotherapy you have a good chance of cure”.

The game changes greatly however if the disease is not lethal. For example, type I diabetes has a great deal of morbidity. Young patients have a relatively high incidence of diabetic coma because they have not yet learned to balance their exercise, food intake, and insulin dose. Every episode is damaging to the central nervous system. Older patients suffer complications including kidney failure, blindness, and loss of limbs. However, the life span is only slightly shortened; diabetes is not now a lethal disease.

Diabetes is an autoimmune disease, caused by an assault on the pancreatic cells which produce insulin. This process, called prediabetes, smolders for years before enough of the pancreas is destroyed to produce the disease. It is now possible to diagnose prediabetes about 3 years before onset of disease with a 95% certainty (Eisenbarth et al. 1994). Naturally physicians caring for these patients are under great pressure to intervene with some agent that will stop the inevitable progression of the disease. Several candidate drugs are available, for example cyclosporin which is routinely used in treating renal transplant rejection. However, the situation is very different with diabetes than with renal transplant rejection, or cancer. Cyclosporin has a 25–38% incidence of causing altered kidney function in transplant patients (Physician's Desk Reference 1994). Even if it is effective in delaying onset of clinical diabetes in 80% of the patients, but produces renal failure in 5%, it is difficult to justify its use. The drug was very effective in a 1 year study of patients with new onset diabetes. It was capable of increasing the frequency and length of remission from dependence on insulin so effectively that 1/2 the patients treated remained free of requirement for insulin for the length of treatment, 1 year (Bougneres et al. 1988). Moreover they had a significant decrease in episodes of diabetic ketoacidosis which can leave permanent damage to the brain (11 vs. 61%, p=0.001). No impaired kidney function was seen in the trial described here, where the patients were intensively (and expensively) monitored for cyclosporin levels. Cyclosporin was proposed for use in prediabetes, but because in actual practice it would have to be used lifelong, the incidence of renal failure would certainly be significant. For this reason, cyclosporin is not considered a candidate for use in prediabetes.

7.9 Interim Analysis

It is a medical imperative to protect the rights of the experimental subjects, who participate in trials of unproven therapies either for money, or for altruistic motives. Thus there is great pressure to ensure that a trial is terminated as soon as possible, to protect the experimental subjects. There are two reasons for this. If the drug is active against a disease with no other effective treatment, it is important to provide it to the patients in the control arm as soon as possible. On the other hand, if the drug is harmful, patient exposure to the agent should be terminated promptly. Interim analysis ensures that exceptional risks or benefits are identified as soon as possible. Another reason for interim analysis is that in a clinical trial, the duration of experimental therapy is carefully defined, despite the fact that it is difficult to predict when maximum clinical effect will be seen. Therefore protocols may include several interim “looks” unblinding the study and comparing the experimental with the control group. Several

problems with interim analysis arise.

7.9.1 EARLY TERMINATION OF CLINICAL TRIALS

Zidovudine (AZT) was the first anti-retroviral agent to be tested in HIV disease. A Phase III study was designed which was double-blind and placebo controlled, treating patients at the highest tolerated dose, 1500 mg/day. Endpoints were opportunistic infections, physical and mental performance (Karnofsky score), and number of CD4+ cells. Interim analysis indicated the drug was effective, the trial was prematurely terminated and patients in the control group were given benefit of the drug (Fischl et al. 1987). A similar result was found in a later trial (Volberding et al. 1990).

Retrospectively however this may have been a suboptimal decision. After 10 years of AZT use, two conflicting studies have been published regarding the efficacy of AZT, one indicating it is still beneficial, the other finding it is ineffective in long term therapy. (Aboulker et al. 1993; Cooper et al. 1994). To attempt to resolve the issue, a recent study (Lenderking et al. 1994) concluded that in patients treated with AZT, the increase in quality of life due to a delay in the progression of HIV disease is exactly balanced by the reduction in quality of life due to side effects of therapy. Because of these studies, some physicians recommended it not be used to treat HIV positive patients.

In clinical medicine, nothing is as simple as it seems. The studies showing lack of efficacy were performed using very high doses of drug, in the order of 1500 to 1000 mg/day. Newer studies indicate that 300–500 mg/day are effective with a significant decrease in side effects (Cooper et al. 1994). Thus the AZT controversy continues. If the initial trials had tested both high and low doses of drug, used more solid clinical endpoints such as progression to disease, and had not been prematurely terminated, probably millions of dollars and many lives would have been saved. Again the dilemma; to what extent do we go to protect the rights of clinical trial subjects.

7.9.2 A STATISTICAL HIT

A second difficulty with interim analysis is that the more frequently the data are examined, the more likely it is to find statistical significance at the 0.05 level. As a result, Bonferroni's rule and its corollaries are applied to these analyses. This introduces a painful penalty to eager clinicians, in terms of adjusted p values (affectionately referred to as a "statistical hit") that are required when interim analyses are conducted. Moreover the details of interim analyses must be prospectively identified in the protocol; this permits more limited statistical penalties in exchange for more limited invasion of the data. In one of the studies using monoclonal antibodies in septic shock, described above, an analysis time of two weeks post therapy

was planned. However, an unplanned interim analysis early in the study indicated that a difference between the two groups was only seen at four weeks post therapy, and the final analysis was conducted accordingly. This post hoc change in a primary end point was sufficient to invalidate the study's results for the FDA (Wenzel 1992; Wanner et al. 1992) Since it had not been made public it did not preclude publication in a peer-reviewed journal. (Ziegler et al, 1991), an unsettling comment on the validity of the peer-review process.

It is the clinicians whose invasion of the data invokes the need "to bonferroni" the results. This approach to data analysis has resulted in an outcry from physicians over the statistical cost of Bonferroni's rule. Statisticians have tried to provide more appropriate tools that permit interim analysis at low statistical cost, providing that they are applied rigorously (O'Brien & Fleming 1979). The bonferroning of clinical trials has led to the institution of data safety and monitoring boards (DSMBs). Since statistical penalties should only be applied when the results of the analysis can have impact on the conduct of the study, the DSMBs must have membership limited to individuals free of apparent conflict of interest. Ideally they should be the only individuals to whom the data analysis center provides interim results on relative efficacy of treatment regimens. (Fleming & DeMets 1993). DSMB guidelines indicate they may only act when predetermined effect levels have been reached, thus avoiding the temptation of "conflicted" investigators (such as pharmaceutical company employees) to act when data are positive, but below the prospectively determined threshold for action.

Bonferroni has taken a lot of the fun out of data dredging and unplanned subgroup analysis and creates an environment that often requires larger and more expensive clinical trials, the cost of which is used to justify the ever-increasing cost of therapy.

Clinicians, pharmaceutical companies, regulators and health care policy experts continue to wrangle with the question of appropriate balance between "protection" of the public by limiting access only to those drugs unquestionably proven both safe and effective by the most conservative analysis, and patients' rights to access to drugs, proven or not, that may benefit them. This dichotomy has been highlighted by the controversies surrounding accelerated approval for AIDS drugs, where a vocal, articulate group of people appears to have gained a relative advantage in regard to early access to drugs, in advance of the results of adequate and well controlled trials. In any case, it has been a clear trend for several years to approve ever narrower indications for drugs on the basis of ever larger studies, the size of which is in part driven by Bonferroni's inferences on sample size and multiple analyses.

7.10 Subgroup analysis

The application of statistical principles (see above) has led those who interpret clinical trials to require that any subgroup analysis must be prospectively defined by the investigator if it is to carry meaningful weight. This insures that investigators don't endlessly analyze subgroups, divided on the basis of unlimited clinical parameters, until one group out of these analyses proves to be statistically significant. Nevertheless, clinicians are intensely tempted to use their medical knowledge of the biology of the disease, to analyze the results. The justification for such analyses can be extremely medically compelling, but the results, flying in the face of medical reason, usually prove disastrous. There is always a positive (and/or negative) subgroup in any study that is large enough to be broken down into multivariate analysis. These data may then be used to design subsequent trials.

In the sepsis studies referenced above (Greenman 1991; Fisher 1994; Wenzel 1992; Ziegler 1991), benefit was shown in subgroups. Few took note of the corollary that when a "positive" subgroup is identified in a study that has shown no overall treatment effect, there must be an offsetting "negative" subgroup. Additional trials are needed to clarify which of the inferences (effective drug, ineffective drug or harmful drug) is correct. A clear example of this was shown in the second study of anti-endotoxin antibody HA-1A. The first study, (Ziegler et al. 1991) showed a significant survival benefit in a small subset of patients, with a non-significant worsening in the balance of the patients. The second study (McCloskey et al. 1994) was terminated at an interim analysis which showed no benefit in the target group of patients, but worsening in the non-target group at the target p value of $p < 0.10$. The inference was thus drawn that the drug, previously believed beneficial (and even marketed in several European countries) was actually harming patients. As a result of this study the manufacturer ceased to study the drug and voluntarily withdrew it from the markets where it was available.

The Second International Study of Infarct Survival (ISIS-2) investigating whether aspirin is beneficial for prevention of infarcts, presented results by the astrological sign under which patients were born. Aspirin was clearly beneficial overall and for persons born under all signs except Libra and Gemini, for which harmful effects were observed (ISIS-2 Collaborative Group, 1988)

7.11 Endpoints

Probably the bitterest arguments in all of the clinical trial arena come from disagreements over appropriate endpoints. Drug regulatory bodies have traditionally argued that in life threatening diseases, the only relevant

endpoint is life or death. Thus for example in cancer, the only drugs that would be approved are those that improve survival. As rationale for this they point to studies, such as the AZT trial described above, where use of surrogate endpoints—laboratory values—and clinical endpoints—decrease in infections—resulted in controversial drug approvals. The National Cancer Institute is probably the most vocal opponent of this practice, and they point out the numerous examples, in which a drug may not prolong survival of a cancer patient, but can dramatically shrink tumor size, thereby reducing pain, and requirement for narcotics. One drug falling in this category is 5-fluorouracil, approved for another indication, but used extensively in patients with colon cancer. By the criteria now used for approval, this drug would not be available for colon cancer. More relevant, many insurance companies will refuse to reimburse for so called “off label” indications—those indications not approved by regulatory authorities.

Other groups argue that surrogate endpoints such as the levels of blood glucose in diabetes have been shown to be excellent predictors of reduction in morbidity and mortality. They feel that a requirement for improved survival in the treated groups is unethical in many cases such as diabetes studies, since it forces the trial to continue until the members of the control group have suffered unacceptable levels of disastrous disease related events. This argument is further clouded by debate about the generalizability of surrogate marker changes. If drugs of one class improve both surrogate and clinical endpoints, does this permit the inference that different drugs with similar effects on the surrogate will offer equivalent clinical benefit?

Statistical analysis becomes critically important in this debate—but the most important limiting factor is often the lack of detailed medical knowledge to estimate expected benefit from a drug with respect to a selected end point. For example, Phase II cancer trials often evaluate treatment effect in terms of “response”. Positive outcomes are “complete responders” (no evidence of residual tumor) and “partial responders” (reduction in total tumor mass by > 50%). This information does not necessarily help one determine the increased survival to be expected in Phase III trial; thus, in a most critical design issue the study’s designers are still shooting in the dark.

7.12 Intent To Treat

While each of the issues we have discussed are points of medical-statistical contention, we have found few concepts that raise clinical hackles higher than the notion of “intent to treat” analysis. This means that all patients in each arm of a study must be analyzed as if they actually were treated with the assigned therapy, even if through some mishap some patients in the treatment arm did not receive all the drug, any of the drug, or even

the wrong drug. Similarly, all patients in the control arm must be analyzed alike, even those who discovered they were in the control arm, and obtained active drug through the medical "underground". While analysis of patient data by treatment group assignment has an impeccable statistical basis, it is hard for a clinician to understand why a patient's results should be considered in a way that differs from the actual treatment he received. For example, if we were evaluating the use of penicillin in the treatment of pneumonia, and a placebo patient were to accidentally receive active drug, clinicians would consider it valid to evaluate that patient's results with other penicillin treated patients, not with the other placebo treated patients.

The strongest argument for intent-to-treat analysis lies in its elimination of bias. A recent paper discussed differing analyses of several clinical trials that compared surgical to medical treatment (Diamond & Denton, 1993). In these studies patients could be transferred from a medical arm to a surgical arm if medically indicated. Sicker patients tended to cross over into the surgical arm (since they medically needed more aggressive therapy) while healthier patients crossed into the medical arm (since more aggressive treatment was not warranted). As a result the studies were strongly biased to show better outcome in medical treatment under an "as treated" analysis. Using computer simulation (Ferguson et al. 1988) developed a model of these clinical trials. Patients were randomly assigned to "surgical" or "medical" groups, and crossover was modeled based on a logistic prediction model developed from an actual patient data base. Data were then analyzed both from intent-to-treat and as-treated populations. Results of the intent to treat analysis showed no difference in outcome based on assignment to medical or surgical therapy. On the other hand, "as-treated" analysis showed a clear survival benefit among patients who received medical, as opposed to surgical, treatment ($p < 0.00001$). In this example, computer modeling allows us to explore the impact of bias introduced by "medically warranted" crossovers. At the same time, ethical considerations require us to permit patients to receive medically indicated treatment. In the example given, which treatment is superior? At the interface of statistical and medical judgment, these clinicians prefer to believe that proper medical judgment resulted in the best treatment for the patients and that statistical modeling is not yet sophisticated enough to account for all of the factors in medical decision making.

The other important benefit of intent to treat is to factor in the unknown factors that cause patients to be noncompliant to studies (or treatment). In a pivotal trial of new treatment for colon cancer, the standard no treatment control was compared to 5-fluorouracil (5-FU) plus levamisole (Moertel et al. 1990). The study demonstrated that combination treatment was superior. However, 13 patients randomized to the no treatment control refused to take their assigned treatment, with many seeking out combination therapy by obtaining 5-FU and by purchasing veterinary levamisole which had

long been available for veterinary use. Surprisingly, these patients fared worse than control group patients that followed their assigned treatment (Fleming T, personal communication). The existence of this phenomenon is indisputable; its explanation, on the other hand, has yet to be found. Nonetheless, it is evident that including these patients in an as-treated analysis would have weakened the study's findings.

7.13 Conclusions

Statistics ... can't live with 'em, can't live without 'em. Statisticians like large samples to add power to their hypotheses/inferences. Clinicians study small numbers either because the patient supply is limited, or it is too expensive and time consuming to conduct larger studies. Since the variables in a single person typically exceed those in models, perhaps the "n of one" trial in which each patient is used as his/her own control, is the right way to go, and statistics be damned. This is certainly the paradigm for clinical practice where each patient's disease and internal milieu are unique, and therefore do not fit exactly into models derived from formal clinical trials.

On the other hand, drug regulators are firmly committed to statistically rigorous study design and analysis as a precondition to the approval of new drugs. For the foreseeable future, no drug will be approved without firm statistical foundation, and we predict that within the next 10 years no medical or scientific journal will publish an article without the same firm statistical underpinnings. This coming rigor will assure the validity and robustness of any conclusions reached, but may have a chilling effect on the development of new approaches to treating unusual or unusually heterogeneous conditions which do not lend themselves to large scale, controlled clinical trials.

In touching his toes to the waters of clinical research, Professor Le Cam found himself on the leading edge of the clinical/statistical interface. To the end of his life, Jerzy Neyman took gleeful pleasure in promoting these very practical research efforts of Le Cam, delighting in the dichotomy between the practical, irritating problems raised by clinical research, and the highly theoretical problems that have been Le Cam's true love.

7.14 REFERENCES

- Aboulker, J. P., Swart, A. M. & the Concorde Coordinating Committee (1993), 'Preliminary analysis of the concorde trial', *The Lancet* 41(3), 889-890.
- Altman, D. G. & Goodman, S. N. (1994), 'Transfer of technology from statistical journals to the biomedical literature', *Journal of the American Medical Association* 272, 129-132.

- Ayad, H. (1948), 'Khellin in angina pectoris', *The Lancet* **1**, 305.
- Bates, D. W. & Lee, T. H. (1994), 'Projected impact of monoclonal anti-endotoxin antibody', *Archive of Internal Medicine* **154**, 1241-1249.
- Batungwanayo, J., Taelman, H., Bogaerts, J. & et al. (1994), 'Pulmonary cryptococcosis associated with HIV-1 infection in Rwanda: a retrospective study of 37 cases', *AIDS* **8**, 1271-1276.
- Benson, H. & McCallie, D. P. (1979), 'Angina pectoris and the placebo effect', *New England Journal of Medicine* **300**, 1424-1429.
- Bone, R. C., Cerra, F. B. & et al. (1992), 'Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis: The ACCP/SCCM Consensus Conference Committee American College of Chest Physicians/Society of Critical Care Medicine', *Chest* **101**, 1644-1655.
- Bone, R. C., Fisher, C. J. J., Clemmer, T. P., Slotman, G. J., Metz, C. A. & Balk, R. A. (1989), 'Sepsis syndrome: a valid clinical entity', *Critical Care Medical Journal* **17**, 389-393. (Methylprednisolone Severe Sepsis Study Group).
- Bougnères, P. F., Carel, J. C., Castano, L., Boitard, C., Gardin, J. P., Landais, P., Hors, J., Mihatsch, M. J., Paillard, M., Chaussain, J. L. & Bach, J. F. (1988), 'Factors associated with early remission of type I diabetes in children treated with cyclosporine', *New England Journal of Medicine* **318**, 663-670.
- Cooper, D. A., Gatell, J. M., Droon, S., Clumeck, N., Millard, J., Goebel, F. D., Bruun, J. N., Stingl, G., Melville, R. S., Gonzalez-Lahoz, J., Stevens, J. W., Fidden, A. P. & the European-Australian Collaborative Group (1993), 'Zidovudine in persons with asymptomatic HIV infection and CD4+ cell counts greater than 400 per cubic millimeter', *New England Journal of Medicine* **329**, 295-303.
- Dewar, H. A. & Grisson, T. A. (1951), 'Further experiences with khellin in angina of effort', *British Heart Journal* **13**, 348-352.
- Diamond, G. A. & Denton, T. A. (1993), 'Alternative perspectives on the biased foundations of medical technology assessment', *Archive of Internal Medicine* **118**, 455-464.
- Eisenbarth, G. S., Gianani, R., Pugliese, A., Verge, C. F. & Pietropaolo, M. (1994), 'Prediction and prevention of type I diabetes', *Transplantation Proceedings* **2**(6), 361-362.

- Fischl, M. A., Richman, D. D., Grieco, M. H., Gottlieb, M. S., Volberding, P. A., Laskin, O. L., Leedom, J. M., Groopman, J. E., Mildvan, D. & Schooley, R. T. (1987), 'The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. a double-blind, placebo controlled trial', *New England Journal of Medicine* **31**(7), 185–191.
- Fisher, B., Costantino, J., Redmond, C., Poisson, R., Bowman, D., Couture, J., Dimitrov, N. V., Wolmark, N., Wickerham, D. L., Fisher, E. R., Margolese, R., Ribidoux, A., Shibata, H., Terz, J., G., P. A. H., Feldman, M., Farrer, W., Evans, J., Lickley, H. L., Ketner, M. et al. (1989), 'A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen receptor-positive tumors', *New England Journal of Medicine* **320**, 479–484.
- Fisher, B., Redmond, C., Wolmark, N. & Wieand, S. H. (1981), 'Disease-free survival at intervals during and following completion of adjuvant chemotherapy: The NSABP experience from three breast cancer protocols', *Cancer* **48**, 1273–1280.
- Fisher, C. J. J., Slotman, G. J., Opal, S. M., Pribble, J. P., Bone, R. C., Emmanuel, G., Ng, D., Bloedow, D. C., Catalano, M. A. & the IL-1RA Sepsis Syndrome Study Group (1994a), 'Initial evaluation of human recombinant interleukin-1 receptor antagonist in the treatment of sepsis syndrome: a randomized, open-label, placebo-controlled multicenter trial', *Critical Care Medical Journal* **22**, 12–21.
- Fisher, G. J., Dhainaut, J. F., Opal, S. M., Pribble, J. P., Balk, R. A., Slotman, G. J., Iberti, T. J., Rackow, E. C., Shapiro, M. J., Greenman, R. L. & rhIL-1ra Sepsis Syndrome Study Group, P. I. (1994b), 'Recombinant human interleukin 1 receptor antagonist in the treatment of patients with sepsis syndrome. results from a randomized, double-blind, placebo-controlled trial', *Journal of the American Medical Association* **271**, 1836–43.
- Fleming, T. R. & DeMets, D. L. (1993), 'Monitoring of clinical trials; issues and recommendations', *Controlled Clinical Trials* **14**, 183–197.
- Fleming, T. R., Harrington, D. P. & O'Brien, P. C. (1984), 'Designs for group sequential tests', *Controlled Clinical Trials* **5**, 348–361.
- Greenberg, R. N., Wilson, K. M., Kunz, A. Y., Wedel, N. I. & Gorelick, K. J. (1992), 'Observations using antiendotoxin antibody (E5) as adjuvant therapy in humans with suspected, serious, gram-negative sepsis', *Critical Care Medical Journal* **20**, 730–735.

- Greenman, R. L., Schein, R. M., Martin, M. A., Wenzel, R. P., MacIntyre, N. R., Emmanuel, G., Chmel, H., Kohler, R. B., McCarthy, M., Plouffe, J. & Group, T. X. S. S. (1991), 'A controlled clinical trial of E5 murine monoclonal IgM antibody to endotoxin in the treatment of gram-negative sepsis', *Journal of the American Medical Association* **266**, 1097-1102.
- Greiner, T., Gold, H., Cattell, M. & et al (1950), 'A method for the evaluation of the effects of drugs on cardiac pain in patients with angina of effort: a study of khellin (visammin)', *American Journal of Medicine* **9**, 143-155.
- Group, I.-C. (1988a), 'Randomized trial of IV streptokinase, oral aspirin, both or neither among 17,187 cases of suspected acute myocardial infarction', *The Lancet* **2**, 849-860.
- Group, T. B.-B. P. R. (1988b), 'The beta-blocker pooling project (BBPP) subgroup findings from randomized trials in post infarction patients', *European Heart Journal* **9**, 9-16.
- Group, T. E. B. S. (1985), 'The international cooperative study of extracranial/intracranial arterial anastomosis (EC/IC bypass study): Methodology and entry characteristics', *Stroke* **16**, 397-406.
- Haynes, R. B., Mukherjee, J., Sackett, D. L., Taylor, D. W., Barnett, H. J. M. & Peerless, S. J. (1987), 'Functional status changes following medical or surgical treatment of cerebral ischemia', *Journal of the American Medical Association* **257**, 2043-2046.
- Hodgson, M. & of-one clinical trials, N. (1993), 'The practice of environmental and occupational medicine', *Journal of Occupational Medicine* **35**, 375-380.
- Leiner, G. C. & Dack, S. (1953), 'The ineffectiveness of khellin in the treatment of angina pectoris', *Journal of Mt. Sinai Hospital* **20**, 41-45.
- McCloskey, R. V., Straube, R. C., Sanders, C., Smith, S. M., Smith, C. R. & Group, C. T. S. (1994), 'Treatment of septic shock with human monoclonal antibody HA-1A. a randomized, double-blind, placebo-controlled trial', *Archive of Internal Medicine* **121**, 1-5.
- Moertel, C. G., Fleming, T. R., Macdonald, J. S., Haller, D. G., Laurie, J. A., Goodman, P. J., Ungerleider, J. S., A., E. W., Tormey, D. C. & Glick, J. H. (1990), 'Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma', *New England Journal of Medicine* **22**, 252-258.

- O'Brien, P. C. & Fleming, T. R. (1979), 'A multiple testing procedure for clinical trials', *Biometrics* **35**, 549-556.
- Oppenheimer, J. J., Nelson, H. S., Bock, S. A., Christensen, F. & Leung, D. Y. M. (1992), 'Treatment of peanut allergy with rush immunotherapy', *Journal of Allergy Clinical Immunology* **90**, 256-262.
- Osher, H. L., Katz, K. H. & Wagner, D. L. (1951), 'Khellin in the treatment of angina pectoris', *New England Journal of Medicine* **244**, 315-321.
- Quirt, L. C., Shelly, W. E. & Pater, J. L. e. a. (1991), 'Improved survival in patients with poor prognosis malignant melanoma treated with adjuvant levamisole: a phase III study by the National Cancer Institute of Canada Clinical Trials Group', *Journal of Clinical Oncology* **9**, 729-735.
- Schultz, K. F., Chalmers, I., Hayes, R. J. & Altman, D. G. (1995), 'Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials', *Journal of the American Medical Association* **273**, 408-412.
- Spitler, L. E. (1991), 'A randomized trial of levamisole versus placebo as adjuvant therapy in malignant melanoma', *Journal of Clinical Oncology* **9**, 736-740.
- Spitler, L. E. & Sagebiel, R. A. (1980), 'A randomized trial of levamisole versus placebo as adjuvant therapy in malignant melanoma', *New England Journal of Medicine* **303**, 1143-1147.
- Storm, T., Thamsberg, G., Steiniche, T., Genant, H. K. & Sorensen, O. H. (1989), 'Effect of intermittent cyclical etidronate therapy on bone mass and fracture rate in women with postmenopausal osteoporosis', *New England Journal of Medicine* **322**, 1265-1271.
- Volberding, P. A., Lagakos, S. W. & Koch, M. A. e. a. (1990), 'Zidovuine in asymptomatic human immunodeficiency virus infection: a controlled trial in persons with fewer than 500 CD4-positive cells per cubic millimeter', *New England Journal of Medicine* **332**, 941-949.
- Warren, H. S., Danner, R. L. & Munford, R. S. (1992), 'Anti-endotoxin monoclonal antibodies', *New England Journal of Medicine* **326**, 1153-1740.
- Wenzel, R., Bone, R., Fein, A., Quenzer, R., Schentag, J., Gorelick, K. J. & Perl, T. (1991), 'Results of a second double-blinded randomized controlled trial of antiendotoxin antibody, E5 in gram-negative

- sepsis, in '31st Interscience Conference on Antimicrobial Agents and Chemotherapy', Chicago, IL.
- Wenzel, R. P. (1992), 'Anti-endotoxin monoclonal antibodies—a second look', *New England Journal of Medicine* **326**, 1151–1153.
- Yusuf, S., Peto, R., Lewis, J., Collins, R. & Sleight, P. (1985), 'Beta blockade during and after myocardial infarction: an overview of the randomized trials', *Prog Cardiovascular Disease* **27**, 885–871.
- Yusuf, S., Wittes, J., Probstfield, J. & Tyroler, H. A. (1991), 'Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials', *Journal of the American Medical Association* **266**, 938.
- Ziegler, E. J., Fisher, C. J. J., Sprung, C. L., Straube, R. C., Sadoff, J. C., Foulke, G. E., Wortel, C. H., Fink, M. P., Dellinger, R. P., H., T. N. & the HA-1A Sepsis Study Group (1991), 'Treatment of gram-negative bacteremia and septic shock with HA-1A human monoclonal antibody against endotoxin. a randomized, double-blind, placebo-controlled trial', *New England Journal of Medicine* **324**, 429–436.

8

Points Singuliers des Modèles Statistiques

D. Dacunha-Castelle¹

8.1 Introduction

Lucien Le Cam a donné un cadre fondamental au développement d'une théorie unifiée de la statistique asymptotique. Il a aussi donné (quelque fois malicieusement) des exemples où les méthodes générales s'appliquent mal, ces difficultés étant le plus souvent liées soit à la complexité topologique de l'espace des paramètres, soit à l'absence de domination du modèle.

Nous voulons dans cet article indiquer une piste sur l'utilisation en statistique de ce qui peut être considéré comme une singularité du paramètre. Nous prendrons comme premier exemple la version statistique du problème des moments, que l'on peut considérer comme un problème semi paramétrique. Donnons deux situations statistiquement importantes relevant de ce problème. La première est celle de la cristallographie. Connaissant des moments $\{\phi_k(\mu), k \in \Lambda\}$ d'une distribution μ formée de r masses de Dirac du cube T^3 , reconstruire cette mesure. Les moments sont mesurés avec une certaine imprécision que l'on peut qualifier de bruit ϵ (ou même ne sont connus, pour certains, en cristallographie que par leur module). Et le fondement des méthodes mathématiques est le caractère singulier (fini dimensionnel) de la mesure μ .

D'un point de vue statistique, évidemment, on ne pourra obtenir au mieux qu'un estimateur $\hat{\mu}_\epsilon$ de μ et comme le but essentiel est la localisation de μ (connaître son support) il sera convenable de choisir comme fonction de perte du problème, $\pi(\mu, \hat{\mu}_\epsilon)$ où π est la distance de Prokhorov ($\hat{\mu}_\epsilon$ sera absolument continue) lorsque $\epsilon \rightarrow 0$, $\pi(\mu, \hat{\mu}_\epsilon) \rightarrow 0$ et nous pourrons donner des bornes à cette perte.

Un deuxième exemple est celui des mélanges de populations, décrits par des modèles statistiques du type $\sum_{i=1}^r p_i G_{\theta_i}$, où $(G_\theta)_{\theta \in \Theta}$ est une famille connue, par exemple une famille de translation $G(\cdot - \theta)$ mais non dominée, r , p_i , θ_i sont inconnus. Même la méthode classique d'estimation par les moments s'applique très mal si r est inconnu. Pour déterminer r , on utilisera la singularité de la mesure de mélange $\sum_{i=1}^r p_i \delta_{\theta_i}$ (mesure de Bayes ou mieux de transition au sens de Le Cam) dans l'ensemble de toutes les

¹Université de Paris-Sud

mesures de transition. Nous présentons d'autres applications relevant de cette technique, notamment la déconvolution aveugle lorsque le signal prend un nombre fini de valeurs.

8.2 Statistiques et singularités dans le problème des moments

Les "singularités" où les situations exceptionnelles surviennent en statistique le plus souvent à la frontière de l'espace des paramètres, par dégénérescence de l'information de Fisher et (ou) par perte de domination. Les modèles statistiques peuvent être de ce point de vue extrêmement compliqués mais le point de vue statistique peut être éclairant pour le problème analytique. Nous traitons ici d'un exemple important: le modèle (semi-paramétrique) des moments et les modèles exponentiels.

Soit $\Phi = (\varphi_1 \dots \varphi_r)$ une famille de r fonctions mesurables définie sur le cube T^s .

Pour $u \in \mathbb{R}^r$, on note $\langle u, \Phi \rangle = \sum_{j=1}^r u_j \varphi_j$. Soit \mathcal{P} les probabilités sur T^s , et pour $\mu \in \mathcal{P}$, $c(\mu) = E_\mu \Phi \in \mathbb{R}^r$. Soit $K = \{c, \exists \mu \in \mathcal{P}, E_\mu \Phi = c\}$, $\overset{\circ}{K}$ sa fermeture dans \mathbb{R}^r , K^* sa frontière, et K^{**} l'ensemble des points d'unicité défini par $K^{**} = \{c \in \mathbb{R}^r, \exists! \mu \in \mathcal{P}, E_\mu \Phi = c\}$. Un problème majeur posé dans ce cadre est le suivant. Sachant a priori que $c \in K^{**}$, ayant mesuré (observé) c avec une imprécision ϵ , on obtient \hat{c}_ϵ , que peut-on dire de $\mu(c)$ et plus précisément comment peut-on estimer des fonctionnelles à valeur ensemble comme le support de $\mu(c)$. Un préalable est d'avoir un critère analytique permettant de définir K et K^* .

Théorème 1. Si Φ est un système de Haar (Lewis 1993) borné,

I. les conditions suivantes sont équivalentes

- (i) $c \in \overset{\circ}{K}$, intérieur de K
- (ii) $\exists P_\lambda, dP_\lambda = Z^{-1}(\lambda) \exp\langle\lambda, \Phi\rangle dx$ tel que $E_{P_\lambda} \Phi = c$.
- (iii) $h^*(c) > 0$, où h^* est une fonction semi continue inférieurement et continue sur $\overset{\circ}{K}$.

II. Les conditions suivantes sont équivalentes

- (i) $c \in K^*$
- (ii) $h^*(c) = 0$.

III. Les conditions suivantes sont équivalentes

- (i) $c \in \overline{K}^c$
- (ii) $h^*(c) = -\infty$.

Ce théorème a été démontré par Dacunha-Castelle & Gamboa (1990) par une technique de grandes déviations puis par Lewis (1993) par des arguments d'analyse convexe.

La fonction h peut être construite ainsi. Soit F une probabilité sur \mathbb{R}^+ telle que $F(\{0\}) > 0$, de transformée de Laplace $\exp(\psi(t))$, définie sur $]-\infty, \alpha[$, pour $\alpha < \infty$. On pose

$$h(c) = \alpha - \log(F\{0\}) - \sup_{v \in \mathbb{R}^{2r+1}} \left(\langle v, \tilde{c} \rangle - \int_{T^s} \psi(\langle v, \Phi(x) \rangle) dx \right)$$

où $\tilde{c} = (c, 1)$.

D'autres exemples de fonctions h peuvent être trouvés dans Dacunha-Castelle & Gamboa (1990) et Gamboa & Gassiat (1996b).

On peut en déduire une amorce de l'analyse de K^* et K^{**} qui reste très largement à faire dans le cas général. Supposons les éléments de Φ bornés. Si $\mu \in K^*$, $E_\mu \Phi = c$, il existe donc une suite λ_n telle que $\|\lambda_n\| \rightarrow \infty$ et $\lim_n P_{\lambda_n} = \mu$.

Analyser la frontière d'un domaine des moments équivaut donc à analyser les points frontières d'une famille exponentielle. Outre l'analyse convexe classique, cela permet d'utiliser des méthodes comme celles de Laplace pour les limites d'intégrales exponentielles et des méthodes géométriques plus profondes.

Donner à titre d'exemple la proposition suivante qui résulte de l'application directe et dans ce cadre sans problème de la méthode de Laplace et pour la deuxième partie du lemme de Morse.

Théorème 2. *Supposons Φ de classe C^3 sur T^s .*

- (i) *Toute probabilité μ telle que $E_\mu \Phi = c \in K^*$ est porté par une variété $V_u = \{x, \langle u, \phi(x) \rangle \text{ atteint son maximum}\}$ où $u \in S(r)$ sphère unité de \mathbb{R}^r .*
- (ii) *Si $\langle u, \Phi(x) \rangle$ n'a qu'un nombre fini de maxima $x_1 \dots x_\ell$, si $h_j = \det D^2(\langle u, \Phi(x_j) \rangle) \neq 0$ pour $j = 1 \dots \ell$, D^2 étant la hessienne, si $h = \sum_{j=1}^\ell h_j^{-1/2}$ alors $\mu = \sum_{j=1}^\ell (h_j^{-1/2}) \delta_{x_j}$.*
- (iii) *Dans le cas de $s = 2$ par exemple si rang $h(x) = s$ pour tout $x \in V_u$, si $\hat{h}(x)$ est la valeur de $\det D^2(\langle \hat{u}, \Phi(x) \rangle)$ dans la direction \hat{u} transverse à V_u en x , alors μ est la probabilité portée par la variété V_u de dimensions s qui vaut*

$$\frac{\tilde{h}^{-1/2}(x)}{\int_{V_u} \tilde{h}^{-1/2}(x) ds(x)}$$

où $ds(x)$ est la mesure curviligne canonique portée par V_u .

Remarquons pour terminer que si $P_{\lambda_n} \rightarrow \mu$, $E_\mu \mu = c$, $c \in K^*$ alors

$$E|\phi - c(\lambda_n)|^2 \rightarrow E_\mu(\phi - c)^2$$

la covariance paramétrique (inverse de l'information de Fisher) converge vers la covariance empirique, l'information de Fisher dégénère en K^* .

8.3 Les systèmes de Tchebychev et l'ordre d'un modèle statistique

Soit $\phi_1 \dots \phi_{2r}$ un système de Tchebychev sur \mathbb{R} (voir Krein & Nudelman 1977, Karlin & Studden 1966 pour la définition), et le problème des moments associés. Alors $K^* = K^{**}$ et l'on a de plus un renforcement du théorème 1 sous forme multilinéaire.

Théorème 3. (Krein & Nudelman 1977) $c \in K$ (resp. $c \in K^{**}$) équivaut à $\det H_r(c) > 0$ (resp. = 0) où $H_r(c)$ est la matrice de Hankel, $H_{i,j} = c_{i+j}$, $1 \leq i, j \leq r$.

Les problèmes d'ordre des processus se posent lorsque l'espace des paramètres est du type $\Theta = \bigcap_{r=1}^P \Theta_r$ où $r, P \in \mathbb{N}$, Θ_r est un ouvert de \mathbb{R}^r , et $\Theta_{r'} \subset \Theta_r^*$ pour $r' < r$.

Le modèle le plus simple est celui de l'ensemble des lois multinomiales $\sum_{j=1}^r p_j \delta_{\alpha_j}$, $\alpha_j \in A$, $0 < p_j < 1$ pour $r > 1$, $\sum_{j=1}^r p_j = 1$ c'est un modèle exponentiel tel $\Theta_r^* = \bigcup_{j=1}^{r-1} \Theta_j$. Donnons, pour des modèles statistiques très simples une utilisation du caractère "singulier" des points de Θ_r^* dans ce cadre.

Appliquons d'abord le résultat précédent au problème de la détermination de l'ordre dans un mélange de population. Soit $Q = \sum_{i=1}^r p_i G_{\theta_i}$ un modèle de mélange, où $\theta_i \in (G_\theta)_{\theta \in \Theta}$ famille connue de distributions r , $(p_i)_{i=1 \dots r}, (\theta_i)_{i=1}$ sont inconnus. On observe un n -échantillon de Q , soit $X_1 \dots X_n$ et le but est d'estimer le modèle. On ne le suppose pas dominé, ce qui exclut d'utiliser la vraisemblance (c'est le cas par exemple d'un modèle de translation avec G_θ ayant une composante atomique) et r étant inconnu les méthodes semi-paramétriques usuelles comme celles des moments ne peuvent être mises en œuvre en général.

Exposons la méthode dans le cas particulier d'un modèle de translation, $\Theta = \mathbb{R}$. Soit $(x, x^2, \dots, x^{2r}) = \Phi_r$ et supposons $E_{G_0}(x)^{2r} < \infty$. On a pour tout $p \in \mathbb{N}$,

$$\begin{aligned} E_Q(X^p) &= \sum_{i=1}^r p_i \int x^p dG(x - \theta_i) \\ &= \sum_{i=1}^r p_i \sum_{k=0}^p \binom{p}{k} E_\mu \theta^k E_{G_0}(X^{-p-k}) \end{aligned}$$

où $\mu = \sum_{i=1}^r p_i \delta_{\theta_i}$ est la mesure de mélange. Donc pour tout r , il existe $A_r(G_0)$ inversible (ou qui se ramène aisément à ce cas) tel que:

$$A_r(G_0) E_\mu \Phi_r = E_Q \Phi_r.$$

On peut donc estimer $E_\mu \Phi_r$ par

$$\widehat{\mu_n(\phi_r)} = [A_r(G_0)]^{-1} \hat{\Phi}_r$$

où

$$\hat{\Phi}_r = \frac{1}{n} \sum_{j=1}^n \Phi_r(X_j).$$

La singularité de μ intervient alors sous la forme suivante:

Lemme. r est caractérisé par $\det H_\mu(r) = 0$ et $\det H_\mu(r') > 0$ pour $r' < r$ (avec $H_\mu(r) = \mu(x^{i+j})$, $1 \leq i, j \leq r$).

On en déduit:

Théorème 4. (Dacunha-Castelle & Gassiat 1996) Soit $\ell(n)$ une suite tendant vers l'infini et $A(p)$ une suite positive strictement croissante.

Soit $J_n(p) = \det H_p(\hat{\mu}_n(\Phi_r)) + A(p)\ell(n)$ et $\hat{r}_n = \operatorname{argmin} J_n(p)$. Alors si $\ell(n)/\sqrt{n} \log_2 n \rightarrow 0$, $\hat{r}_n \rightarrow r$ p.s. et si G a des moments exponentiels l'erreur d'estimation sur r admet une majoration de type $\exp(-an\ell^2(n))$.

Ce théorème est une application directe des calculs associés à l'habituelle compensation par une expression du type $A(p)\ell(n)$ destinée à éviter une surestimation de l'ordre lorsque l'on passe du critère analogue donné par le lemme à la situation statistique. La méthode s'étend à des cas de modèles de mélange beaucoup plus généraux (Dacunha-Castelle & Gassiat 1996).

De manière très analogue, si l'on a un processus stationnaire associé à un spectre ponctuel formé de r raies on peut estimer r par la méthode précédente en estimant à partir des covariances empiriques les déterminants de Toeplitz de la mesure aléatoire associée au spectre, la méthode reste valable en l'absence d'ergodicité.

Un autre exemple de modèle utilisé en communications est le suivant.

Soit $Y_t = \sum_{k \in \mathbb{Z}} u_k X_{t-k}$, $u \in \ell^2$ où (X_t) est une suite de variables aléatoires indépendantes dont on sait seulement qu'elles prennent exactement pour r valeurs distinctes les coefficients u_k de filtre sont inconnus et observe Y_t , $t = 1 \dots n$. On veut estimer (u_k) et restituer X_t . Il existe une littérature très abondante sur le sujet, notamment les résultats de Gassiat (1993) sur le minimax dans le cas le plus général (non causal), résultats qui utilisent profondément les idées de Le Cam et les modèles L.A.N. mais comme tous ces résultats, la domination est essentielle, elle est perdue ici par le filtrage. En utilisant le caractère singulier pour le problème des moments de la loi des X_t dont le déterminant de Toeplitz (ou de Hankel) s'annule à l'ordre r , on peut obtenir de très nombreux résultats spécifiques à cette situation.

Le plus simple est le suivant (Gamboa & Gassiat 1996a). Supposons le filtre $u = (u_x)$ inversible c'est-à-dire tel qu'il existe $\theta \in \ell^2$ avec $\theta * u = \delta$. Fixons une échelle pour déterminer le problème, par exemple $u_0 = 1$. L'idée est alors très simple comme l'addition de variables à support fini augmente la taille du support si u_0 est la vraie valeur, en estimant empiriquement à partir des Y_t et de la relation $X_t = \theta * Y_t$ les moments de la loi μ de X_t , seul $\theta = \theta_0$ conduira à un support fermé de r points, les θ différents de θ_0 donnent un support strictement plus grand donc un déterminant non nul.

Techniquement, on pose

$$\begin{aligned} m(n) &\rightarrow \infty \text{ avec } \frac{m(n)}{n} \rightarrow 0, \quad t \in [1 + m(n), n - m(n)] \\ \hat{Z}_t(\theta) &= \sum_{k=-m(n)}^{m(n)} \theta_k Y_{t-k} \\ c_n(\theta) &= \frac{1}{n - 2m(n)} \sum_{t=1+m(n)}^{n-m(n)} \Phi_r(\text{Arctg } \hat{Z}_t(\theta)) \end{aligned}$$

Φ_r est le système de Toeplitz des $2r$ premières fonctions trigonométriques et $T_n(\theta)$ la matrice de Toeplitz associée à la valeur $c_n(\theta)$.

Théorème 5. Si Θ est un compact de ℓ^2 , et $\theta = (\theta_n)$,

$$\begin{aligned} \Theta_n &= \{\theta, \theta_k = 0 \text{ par } |k| > m(n)\} \quad \text{with } m(n) \rightarrow \infty \\ \hat{\theta}_n &= \underset{\theta \in \Theta_n}{\operatorname{argmin}} |\det T_n(\theta)| \end{aligned}$$

alors $\hat{\theta}_n \rightarrow \theta$ p.s.

8.4 Conclusion

La théorie asymptotique générale telle que Le Cam l'a développée, notamment en ce qui concerne les expériences gaussiennes s'applique moyennant le reparamétrage des variétés constituant la frontière des modèles dans les problèmes où intervient un paramètre comme l'ordre. Schématiquement l'espace $\Theta = \bigoplus_{r=1}^p \Theta_r$, Θ_r est ouvert et $\Theta_r^* = \bigcup_{r' < r} \Theta_{r'}$ est sa frontière. S'il existe une vraisemblance, sur Θ_r^* la matrice d'information dégénère, les tests d'ordre des modèles ont des lois limites qui ne sont pas des χ^2 mais sont associées à des expériences gaussiennes plus compliquées. C'est ce que Hannan a montré pour le cas des processus ARMA (Azencott & Dacunha-Castelle 1987), de même que Ghosh & Sen (1985) en ce qui concerne les mélanges mais sous des hypothèses trop restrictives. Ceci vaut lorsque la variété sur laquelle $I(\theta)$ dégénère est assez simple, et si la dimension de la variété transverse sur laquelle la restriction de $I(\theta)$ est strictement positive et constante. Nous avons indiqué dans le cas général où il n'y a pas domination que l'utilisation du caractère singulier des points frontières permet d'estimer l'ordre et de revenir, à partir de là, à des situations plus classiques.

Dans Dacunha-Castelle & Gassiat (1996) nous avons montré qu'il existe une théorie de la vraisemblance au sens de Le Cam pour des modèles non identifiables comme les modèles les plus généraux ou les modèles ARMA.

L'existence de cette théorie passe par un reparamétrage permettant d'*éclater* la singularité à l'origine, origine représentant la vraie probabilité P_0 . En utilisant des classes de Donsker, obtenues par des calculs

d'entropie à crochet, on montre que la statistique du maximum de la vraisemblance est alors le plus souvent du type $\sup_{d \in D} \xi_d^2 1_{\{\xi_d > 0\}}$, où ξ_d est un processus gaussien canonique indexé pour un compact D de la sphère unité de $L^2(P_0)$. La forme de la frontière de D est essentielle pour déterminer la loi du maximum de la vraisemblance. Le reparamétrage est donc ici le moyen de traiter des singularités assez compliquées de problèmes statistiques et ils sont directement inspiré des travaux de Lucien Le Cam.

8.5 BIBLIOGRAPHIE

- Azencott, R. & Dacunha-Castelle, D. (1987), *Sequences of Irregular Observations*, Springer-Verlag, New York.
- Dacunha-Castelle, D. & Gamboa, F. (1990), ‘Maximum d'entropie et problème des moments’, *Annales de l'Institut Henri Poincaré* **26**, 567–596.
- Dacunha-Castelle, D. & Gassiat, E. (1996), Estimation de l'ordre d'un mélange de populations, Technical report, Université de Paris-Sud, Orsay. (To appear in *Bernoulli*).
- Gamboa, F. & Gassiat, E. (1996a), Blind deconvolution of discrete linear systems, Technical report, Université de Paris-Sud, Orsay. (To appear in *Annals of Statistics*).
- Gamboa, F. & Gassiat, E. (1996b), Super resolution via MEM technics, Technical report, Université de Paris-Sud, Orsay. (To appear in *Siam Journal of Mathematical Analysis*).
- Gassiat, E. (1993), ‘Adaptative estimation in noncausal AR processes’, *Annals of Statistics* **21**, 2022–2042.
- Ghosh, J. K. & Sen, P. K. (1985), On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results, in L. Le Cam & R. A. Olshen, eds, ‘Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Volume II’, Wadsworth, Belmont, CA, pp. 789–806.
- Karlin, S. & Studden, N. J. (1966), *Tchebycheff Systems with Applications to Analysis and Statistics*, Wiley, New York.
- Krein, M. G. & Nudelman, A. A. (1977), *The Markov moment problem and extremal problems*, Vol. 50 of *Translations of Mathematical Monographs*, American Mathematical Society.
- Lewis, A. S. (1993), Consistency of moment systems, Technical report, Waterloo University, Canada.

9

Exponential Tightness and Projective Systems in Large Deviation Theory

A. de Acosta¹

9.1 Introduction

Let $\{E_\alpha, p_\alpha^\beta\}$ be a projective system of Hausdorff topological spaces; here $\alpha, \beta \in A$, a directed set, and $p_\alpha^\beta : E_\beta \rightarrow E_\alpha$ is a continuous surjective map for $\alpha < \beta$ (for details, see Section 3). Let E be a Hausdorff topological space endowed with a σ -algebra \mathcal{E} (possibly smaller than the Borel σ -algebra), and for each $\alpha \in A$, let $p_\alpha : E \rightarrow E_\alpha$ be a continuous measurable surjective map such that $p_\alpha = p_\alpha^\beta \circ p_\beta$ for $\alpha < \beta$. Let $\{\mu_n\}$ be a sequence of probability measures on \mathcal{E} , and assume that for each $\alpha \in A$, the sequence $\{\mu_n \circ p_\alpha^{-1}\}$ satisfies the large deviation principle (see Section 2). In this paper we show that under suitable additional assumptions, the large deviation principle for $\{\mu_n\}$ follows.

In Theorem 1 of Section 2 we generalize an interesting analogue of Prohorov's tightness theorem recently obtained by Pukhalskii (1991). Corollary 5 is useful for the proof of one of the projective system theorems, Theorem 2.

Section 3 contains the projective system theorems. Theorem 2 is proved under the assumption of exponential tightness; E is endowed with an arbitrary Hausdorff topology. Theorem 4 is specific for the initial topology induced by $\{p_\alpha\}_{\alpha \in A}$ on E ; it encompasses a result of Dawson & Gärtner (1987) (see also Dembo & Zeitouni 1993) on large deviations for projective limits as well as the general form of Sanov's theorem in de Acosta (1994b). In applications, the probability measures $\{\mu_n\}$ typically live in a space which has a special structure and is much smaller than the projective limit of the system $\{E_\alpha, p_\alpha^\beta\}$; also, a natural candidate for rate function may be available. Then Theorems 2 and 4 may provide a more direct and flexible approach to the proof of large deviation principles than indirect arguments involving projective limits (as in Dembo & Zeitouni 1993). The applications in Section 4 and de Acosta (1994b) are examples of such situations.

¹Case Western Reserve University

In Section 4 we give two applications which illustrate the use of Theorem 2. The first one is a new proof of the large deviation principle for the averages of an i.i.d. sequence taking values in a separable Banach space (Donsker & Varadhan 1976, Bahadur & Zabell 1979, Azencott 1980, de Acosta 1985) The second one is a large deviation principle for a sequence of Lévy processes in \mathbb{R}^d ; this generalizes one of the results in de Acosta (1994a), by a different method.

9.2 Exponential tightness and subnets

Let us recall that a net $\{\mu_d\}_{d \in D}$ (D is a directed set) of probability measures (p.m.'s) on a σ -algebra \mathcal{E} of subsets of a Hausdorff topological space *satisfies the large deviation principle* with normalizing constants $\{r_d\}_{d \in D}$ (where $\lim_d r_d = \infty$) and rate function $I : E \rightarrow [0, \infty]$ if I is lower semi-continuous and for every $B \in \mathcal{E}$,

$$\begin{aligned} -\inf_{x \in \text{int } B} I(x) &\leq \liminf_d r_d^{-1} \log \mu_d(B) \\ &\leq \limsup_d r_d^{-1} \log \mu_d(B) \leq -\inf_{x \in \text{cl } B} I(x). \end{aligned}$$

(Here int (resp., cl) stands for interior (resp., closure)). The rate function I is *good* if for every $a \geq 0$, $\{x \in E : I(x) \leq a\}$ is compact.

The following result generalizes Theorem (P) of Pukhalskii (1991), which was proved for a sequence of p.m.'s on the Borel σ -algebra of a metric space and $r_n = n$ (see the Remark following Theorem 1). Our method of proof uses some well known arguments in the large deviations literature (Azencott 1980, Dembo & Zeitouni 1993) and is simpler.

Theorem 1 *Let E be a Hausdorff topological space and let \mathcal{E} be a σ -algebra of subsets of E such that*

- (i) \mathcal{E} contains the class of compact sets,
- (ii) \mathcal{E} contains a base \mathcal{U} for the topology.

Let D be a directed set, let $\{\mu_d\}_{d \in D}$ be a net of p.m.'s on \mathcal{E} and let $r : D \rightarrow \mathbb{R}^+$ be such that

$\lim_{d \in D} r_d = \infty$. Assume that $\{\mu_d\}_{d \in D}$ is exponentially tight: for every $a > 0$, there exists a compact set K_a such that

$$\limsup_{d \in D} r_d^{-1} \log \mu_d(K_a^c) \leq -a.$$

Then there exists a subnet $\{\mu_{\varphi(h)}\}_{h \in H}$ (with H a directed set, $\varphi : H \rightarrow D$) which satisfies the large deviation principle with normalizing constants $\{r_{\varphi(h)}\}_{h \in H}$ and a certain good rate function $I : E \rightarrow [0, \infty]$.

Proof: Consider the net $F : D \rightarrow [-\infty, 0]^{\mathcal{U}}$ defined by

$$F(d) = \{r_d^{-1} \log \mu_d(U)\}_{U \in \mathcal{U}}.$$

By Tichonov's theorem applied to the compact space $[-\infty, 0]^\mathcal{U}$ there is a convergent subnet $f : H \rightarrow [-\infty, 0]^\mathcal{U}$,

$$f(h) = \{r_{\varphi(h)}^{-1} \log \mu_{\varphi(h)}(U)\}_{U \in \mathcal{U}},$$

where H is a suitable directed set and $\varphi : H \rightarrow D$ a suitable map (Kelley 1955, pages 70, 136). We define, for $U \in \mathcal{U}$ and $x \in E$,

$$\begin{aligned}\ell(U) &= \lim_{h \in H} \rho(h) \log \nu_h(U) \in [-\infty, 0], \\ I(x) &= -\inf\{\ell(U) : x \in U, U \in \mathcal{U}\},\end{aligned}$$

where $\rho(h) = r_{\varphi(h)}^{-1}$, $\nu_h = \mu_{\varphi(h)}$. It is easily shown that I is lower semicontinuous.

We first prove the lower bound. In fact, given $B \in \mathcal{E}$, $x \in \text{int } B$, let $U \in \mathcal{U}$ be such that $x \in U \subset B$. Then

$$\liminf_h \rho(h) \log \nu_h(B) \geq \lim_h \rho(h) \log \nu_h(U) = \ell(U) \geq -I(x),$$

and therefore

$$\liminf_h \rho(h) \log \nu_h(B) \geq -\inf_{x \in \text{int } B} I(x).$$

By a well-known argument, (Dembo & Zeitouni 1993, page 8), using the assumption of exponential tightness and the obvious inequality

$$\nu_h(B) \leq \nu_h(\text{cl } B \cap K_a) + \nu_h(K_a^c),$$

it is enough to prove the upper bound for compact sets. Let K be compact, and let $a > -\inf_{x \in K} I(x)$. Then for every $x \in K$ there exists $U_x \in \mathcal{U}$ such that $x \in U_x$ and $a > \ell(U_x)$. Let $\{U_{x_j} : j = 1, \dots, k\}$ be a finite subcover of $\{U_x\}_{x \in K}$. Then

$$\begin{aligned}\limsup_h \rho(h) \log \nu_h(K) &\leq \lim_h \rho(h) \log \nu_h(\bigcup_{j=1}^k U_{x_j}) \\ &= \max\{\ell(U_{x_1}), \dots, \ell(U_{x_k})\} \\ &< a.\end{aligned}$$

This completes the proof of the upper bound. The fact that the rate function I is good follows from exponential tightness and the lower bound by a well-known argument (Dembo & Zeitouni 1993, page 8).

Remark In the framework of Theorem 1, assume furthermore that \mathcal{U} is countable and $D = \mathbb{N}$, that is, an exponentially tight sequence $\{\mu_n\}_{n \in \mathbb{N}}$ is given. Then $[-\infty, 0]^\mathcal{U}$ is a compact metrizable space and an obvious rephrasing of the arguments in Theorem 1 shows that its conclusion holds for a subsequence of $\{\mu_n\}_{n \in \mathbb{N}}$. The result of Pukhalskii (1991) follows: in fact, the assumptions in Pukhalskii (1991, Theorem (P)) imply that $\{\mu_n\}_{n \in \mathbb{N}}$ is tight and therefore, as is easily seen, it is enough to prove the

result for a σ -compact metric space, endowed with its Borel σ -algebra. But the topology of such a space has a countable base.

We omit the straightforward proof of the following corollary, which is obtained from Theorem 1 by a contradiction argument.

Corollary 5 *Let E, \mathcal{E}, r be as in Theorem 1. Assume*

- (i) $\{\mu_d\}_{d \in D}$ is exponentially tight,
- (ii) There exists $I : E \rightarrow [0, \infty]$ such that if a subnet $\{\mu_{\varphi(h)}\}_{h \in H}$ satisfies the large deviation principle with normalizing constants $\{r_{\varphi(h)}\}_{h \in H}$ and good rate function J , then $J = I$.

Then $\{\mu_d\}_{d \in D}$ satisfies the large deviation principle with normalizing constants $\{r_d\}_{d \in D}$ and good rate function I .

9.3 Projective systems and large deviations

We shall call a family $\{E_\alpha, p_\alpha^\beta\}$ where $\alpha, \beta \in A$, a directed set, a *projective system* if

- (i) for each $\alpha \in A$, E_α is a Hausdorff topological space,
- (ii) for each $\alpha, \beta \in A$, $\alpha < \beta$, $p_\alpha^\beta : E_\beta \rightarrow E_\alpha$ is a continuous surjective map, p_α^α is the identity map on E_α and $p_\alpha^\gamma = p_\alpha^\beta \circ p_\beta^\gamma$ for $\alpha < \beta < \gamma$.

We shall also consider a set E and surjective maps $p_\alpha : E \rightarrow E_\alpha$ such that for $\alpha < \beta$, $p_\alpha = p_\alpha^\beta \circ p_\beta$ (this condition implies the second condition in (ii) above) and for $x, y \in E$, $x \neq y$, there exists $\alpha \in A$ such that $p_\alpha(x) \neq p_\alpha(y)$.

Theorem 2 *Let E, \mathcal{E} be as in Theorem 1. Let $\{E_\alpha, p_\alpha^\beta\}$ be a projective system and let $p_\alpha, \alpha \in A$, be as above; we assume that p_α is a continuous surjective map which is measurable when E is endowed with \mathcal{E} and E_α with the Borel σ -algebra.*

Let $\{\mu_d\}_{d \in D}$ (D is a directed set) be a net of p.m.'s on \mathcal{E} and $r : D \rightarrow \mathbb{R}^+$ be such that $\lim_d r_d = \infty$. Assume:

- (i) *For each $\alpha \in A$, the net $\{\mu_d \circ p_\alpha^{-1}\}_{d \in D}$ satisfies the large deviation principle with normalizing constants $\{r_d\}_{d \in D}$ and rate function $I_\alpha : E_\alpha \rightarrow [0, \infty]$.*
- (ii) *$\{\mu_d\}_{d \in D}$ is exponentially tight (see Theorem 1).*

Then $\{\mu_d\}_{d \in D}$ satisfies the large deviation principle with normalizing constants $\{r_d\}_{d \in D}$ and good rate function

$$I(x) = \sup_\alpha I_\alpha(p_\alpha(x)), \quad x \in E.$$

For the proof we need the following lemma.

Lemma 3 *In the context of the statement of Theorem 2, let $J : E \rightarrow [0, \infty]$ be such that for all $a \geq 0$, $\{x \in E : J(x) \leq a\}$ is compact. For each $\alpha \in A$, define $J_\alpha : E_\alpha \rightarrow [0, \infty]$ by $J_\alpha(z) = \inf\{J(x) : x \in p_\alpha^{-1}(z)\}$. Then*

- (i) *for all $a \geq 0$, $\{x \in E_\alpha : J_\alpha(x) \leq a\}$ is compact,*
- (ii) *for all $x \in E$, $J(x) = \sup_\alpha J_\alpha(p_\alpha(x))$.*

Proof: It is easily seen that $p_\alpha(\{x \in E : J(x) \leq a\}) = \{x \in E_\alpha : J_\alpha(x) \leq a\}$, proving (i).

By the definition of J_α , we have $J(x) \geq J_\alpha(p_\alpha(x))$, and therefore $J(x) \geq \sup_\alpha J_\alpha(p_\alpha(x))$ for all $x \in E$. To prove the opposite inequality, fix $x \in E$ and let $c = \sup_\alpha J_\alpha(p_\alpha(x))$; obviously we may assume $c < \infty$. For $b > c$, let $K_b = \{y \in E : J(y) \leq b\}$. By the definition of J_α , we have: for every $\alpha \in A$,

$$p_\alpha^{-1}(p_\alpha(x)) \cap K_b \neq \emptyset.$$

But $\{p_\alpha^{-1}(p_\alpha(x)) \cap K_b\}_{\alpha \in A}$ is then a directed decreasing family of non-empty compact sets, and consequently

$$\{x\} \cap K_b = \bigcap_\alpha p_\alpha^{-1}(p_\alpha(x)) \cap K_b \neq \emptyset,$$

so $x \in K_b$, that is, $J(x) \leq b$.

Proof of Theorem 2: We apply Corollary 5 with $I(x) = \sup_\alpha I_\alpha(p_\alpha(x))$. Suppose that a subnet $\{\mu_{\varphi(h)}\}_{h \in H}$ satisfies the large deviation principle with normalizing constants $\{r_{\varphi(h)}\}_{h \in H}$ and good rate function J . By the contraction principle (Dembo & Zeitouni 1993, page 110), $\{\mu_{\varphi(h)} \circ p_\alpha^{-1}\}_{h \in H}$ satisfies the large deviation principle with the same normalizing constants and good rate function $J_\alpha(z) = \inf\{J(x) : x \in p_\alpha^{-1}(z)\}$. By assumption (i) and the uniqueness of rate functions, we have $J_\alpha = I_\alpha$ (it is easily seen that the usual argument for uniqueness (Dembo & Zeitouni 1993, page 103) can be modified to apply to two good rate functions on a Hausdorff space; that the rate function I_α is good follows from the exponential tightness of $\{\mu_d \circ p_\alpha^{-1}\}_{d \in D}$ and the lower bound (Dembo & Zeitouni 1993, page 8)).

Now Lemma 3 yields: for all $x \in E$,

$$J(x) = \sup_\alpha J_\alpha(p_\alpha(x)) = \sup_\alpha I_\alpha(p_\alpha(x)) = I(x),$$

and the conclusion follows from Corollary 5.

Remark Theorem 4.5 of Pukhalskii (1991) is a result in the spirit of Theorem 2 for a particular projective system.

Theorem 4 Let E , $\{E_\alpha, p_\alpha^\beta\}$ be as in the paragraph preceding Theorem 2. Let E be endowed with the initial topology induced by the maps $\{p_\alpha\}_{\alpha \in A}$, and let \mathcal{E} be a σ -algebra of subsets of E such that for all $\alpha \in A$, p_α is measurable, where E_α is endowed with its Borel σ -algebra.

Let $\{\mu_d\}_{d \in D}$ (D is a directed set) be a net of p.m.'s on \mathcal{E} and $r : D \rightarrow \mathbb{R}^+$ be such that $\lim_d r_d = \infty$. Assume:

(i) For each $\alpha \in A$, the net $\{\mu_d \circ p_\alpha^{-1}\}_{d \in D}$ satisfies the large deviation principle with normalizing constants $\{r_d\}_{d \in D}$ and rate function $I_\alpha : E_\alpha \rightarrow [0, \infty]$.

(ii) There exists $I : E \rightarrow [0, \infty]$ such that for all $a \geq 0$, $\{x \in E : I(x) \leq a\}$ is compact and for all $\alpha \in A$, $z \in E_\alpha$

$$I_\alpha(z) = \inf\{I(x) : x \in p_\alpha^{-1}(z)\}.$$

Then $\{\mu_d\}_{d \in D}$ satisfies the large deviations principle with normalizing constants $\{r_d\}_{d \in D}$ and good rate function I . Moreover, for all $x \in E$

$$I(x) = \sup_\alpha I_\alpha(p_\alpha(x)). \quad (1)$$

Proof: For $B \subset E$, let $I(B) = \inf_{x \in B} I(x)$. We first prove the lower bound. Let $B \in \mathcal{E}$, and suppose $p_\alpha^{-1}(U) \subset B$ for some $\alpha \in A$, U open in E_α . Then by assumptions (i) and (ii),

$$\begin{aligned} \liminf_d r_d^{-1} \log \mu_d(B) &\geq \liminf_d r_d^{-1} \log(\mu_d \circ p_\alpha^{-1})(U) \\ &\geq -\inf_{z \in U} I_\alpha(z) \\ &= -\inf_{z \in U} \inf\{I(x) : x \in p_\alpha^{-1}(z)\} \\ &= -I(p_\alpha^{-1}(U)). \end{aligned}$$

The family $\{p_\alpha^{-1}(V) : \alpha \in A, V \text{ open in } E_\alpha\}$ is a base for the initial topology on E ; therefore

$$\liminf_d r_d^{-1} \log \mu_d(B) \geq -I(\text{int } B).$$

To prove the upper bound, let $B \in \mathcal{E}$, $C = \text{cl } B$. Then for all $\alpha \in A$, $C \subset p_\alpha^{-1}(\text{cl } p_\alpha(C))$ and therefore by assumptions (i) and (ii)

$$\begin{aligned} \limsup_d r_d^{-1} \log \mu_d(B) &\leq \limsup_d r_d^{-1} \log(\mu_d \circ p_\alpha^{-1})(\text{cl } p_\alpha(C)) \\ &\leq -\inf_{z \in \text{cl } p_\alpha(C)} I_\alpha(z) \\ &= -\inf_{z \in \text{cl } p_\alpha(C)} \inf\{I(x) : x \in p_\alpha^{-1}(z)\} \\ &= -I(p_\alpha^{-1}(\text{cl } p_\alpha(C))) \end{aligned}$$

But, as is easily shown, $\{p_\alpha^{-1}(\text{cl } p_\alpha(C))\}_{\alpha \in A}$ is a directed decreasing family of closed sets and

$$C = \bigcap_\alpha p_\alpha^{-1}(\text{cl } p_\alpha(C)),$$

and by a well-known property of good rate functions (Dembo & Zeitouni 1993, page 104), assumption (ii) implies

$$I(C) = \sup_{\alpha} I(p_{\alpha}^{-1}(clp_{\alpha}(C))).$$

It follows that

$$\limsup_d r_d^{-1} \log \mu_d(B) \leq -I(clB).$$

The expression for I follows from Lemma 3.

Remarks

- It is of interest to note that the proof of the lower bound depends only on assumption (i) and the inequality

$$I_{\alpha}(z) \leq \inf\{I(x) : x \in p_{\alpha}^{-1}(z)\}.$$

In applications, this inequality is often easy to verify; this is the case in de Acosta (1994b). The inequality is of course obvious if I is given a priori by (1).

- Lemmas 2.1 and 2.2 of de Acosta (1994b) show that, with the obvious definition of the projective system in that case, assumption (ii) of Theorem 4 holds when $I = I_{\mu}$ (in the notation of de Acosta 1994b). In the context of de Acosta (1994b), assumption (i) of Theorem 4 is just Cramér's theorem in a finite-dimensional vector space. Therefore Theorem 1.1 of de Acosta (1994b) may be obtained from Theorem 4.
- If in Theorem 4 we take $E = \lim_{\leftarrow} E_{\alpha}$, the projective limit of the system $\{E_{\alpha}, p_{\alpha}^{\beta}\}$, (endowed with the Borel σ -algebra), and $I(x) = \sup_{\alpha} I_{\alpha}(p_{\alpha}(x))$, then it follows from elementary properties of projective limits that assumption (ii) of Theorem 4 is automatically satisfied. Thus Theorem 4 encompasses the result of (Dawson & Gärtner 1987) on large deviations for projective limits.

9.4 Applications

We first apply Theorem 2 to

Theorem 5 (Donsker & Varadhan 1976, Bahadur & Zabell 1979, Azencott 1980, de Acosta 1985). *Let E be a separable Banach space endowed with its Borel σ -algebra. Let $\{X_j\}$ be i.i.d. E -valued r.v.'s, $S_n = \sum_{j=1}^n X_j$. Assume: for all $t > 0$,*

$$\int_E \exp(t\|x\|) \mu(dx) < \infty, \quad (2)$$

where $\mu = \mathcal{L}(X_1)$. Then $\{\mathcal{L}(S_n/n)\}_{n \in \mathbb{N}}$ satisfies the large deviation principle (with normalizing constants $\{n\}$) and good rate function

$$I(x) = \sup_{\xi \in E^*} [\langle x, \xi \rangle - \log \hat{\mu}(\xi)] \quad x \in E,$$

where E^* is the dual space of E and $\hat{\mu}(\xi) = \int \exp(\xi) d\mu$ for $\xi \in E^*$.

Proof: Let \mathcal{N} be the family of finite-dimensional subspaces of E^* , directed upward by inclusion. For each $N \in \mathcal{N}$, let $N^\perp = \{x \in E : \langle x, \xi \rangle = 0 \text{ for all } \xi \in N\}$, and let $\Pi_N : E \rightarrow E/N^\perp$ be the canonical projection; for each $N, M \in \mathcal{N}, N \subset M$, let $\Pi_N^M : E/M^\perp \rightarrow E/N^\perp$ be the canonical projection. Then $\{E/N^\perp, \Pi_N^M\}$ is a projective system of finite-dimensional normed spaces and clearly $\{E/N^\perp, \Pi_N^M\}$, E and $\{\Pi_N\}$ satisfy the assumptions of Theorem 2.

Let $\mu_k = \mathcal{L}(S_k/k)$. Then for each $N \in \mathcal{N}$, $\mu_k \circ \Pi_N^{-1} = \mathcal{L}(\sum_{j=1}^k \Pi_N(X_j)/k)$ and therefore by assumption (2) and Cramér's theorem in E/N^\perp (de Acosta 1985, Dembo & Zeitouni 1993), $\{\mu_k \circ \Pi_N^{-1}\}_{k \in \mathbb{N}}$ satisfies the large deviation principle with good rate function

$$I_N(z) = \sup_{\eta \in N} (\langle z, \eta \rangle - \log(\mu \circ \Pi_N^{-1})^\wedge(\eta)). \quad (3)$$

Note that N may be identified with the dual of E/N^\perp .

Assumption (2) implies that $\{\mu_k\}_{k \in \mathbb{N}}$ is exponentially tight (de Acosta 1985). Therefore by Theorem 2, $\{\mu_k\}_{k \in \mathbb{N}}$ satisfies the large deviation principle with good rate function

$$I(x) = \sup_N I_N(\Pi_N(x)).$$

By (3), taking into account that for $\eta \in N$ obviously $\langle \Pi_N(x), \eta \rangle = \langle x, \eta \rangle$ and $(\mu \circ \Pi_N^{-1})^\wedge(\eta) = \hat{\mu}(\eta)$, we have

$$\begin{aligned} I(x) &= \sup_N \sup_{\eta \in N} [\langle \Pi_N(x), \eta \rangle - \log(\mu \circ \Pi_N^{-1})^\wedge(\eta)] \\ &= \sup_N \sup_{\eta \in N} [\langle x, \eta \rangle - \log \hat{\mu}(\eta)] \\ &= \sup_{\xi \in E^*} [\langle x, \xi \rangle - \log \hat{\mu}(\xi)]. \end{aligned}$$

In our second application we use Theorem 2 to prove a generalization of Theorem 1.2 of de Acosta (1994a); for simplicity we will take the Lévy processes to be \mathbb{R}^d -valued. We will use the terminology and some arguments from de Acosta (1994a). The proof by the projective limit method of a closely related result in (Dembo & Zeitouni 1993, Theorem 5.1.2), uses arguments which are very similar to the second part of the proof below. Let $T = [0, 1]$, and let $D(T, \mathbb{R}^d)$ be the space of cadlag \mathbb{R}^d -valued functions on T , endowed with the uniform norm $\|f\|_\infty = \sup_{s \in T} \|f(s)\|$, where $\|\cdot\|$ is the Euclidean norm on \mathbb{R}^d (say), and with the σ -algebra \mathcal{D} generated by the evaluations $\pi_t(f) = f(t), t \in T, f \in D(T, \mathbb{R}^d)$.

Theorem 6 Let $\{X_n(t) : t \geq 0\}, n \in \mathbb{N}$, be a sequence of \mathbb{R}^d -valued Lévy processes. Assume:

(i) For every $n \in \mathbb{N}$ and every $\xi \in \mathbb{R}^d$,

$$E \exp\langle X_n(1), \xi \rangle < \infty.$$

(ii) For every $\xi \in \mathbb{R}^d$,

$$\phi(\xi) = \lim_n n^{-1} \log E \exp\langle X_n(1), n\xi \rangle$$

exists in \mathbb{R} and ϕ is differentiable.

Let $\mu_n = \mathcal{L}(\{X_n(t) : t \in T\})$. Then $\{\mu_n\}_{n \in \mathbb{N}}$ satisfies the large deviation principle on $D(T, \mathbb{R}^d)$ (endowed with the uniform norm $\|\cdot\|_\infty$ and the σ -algebra \mathcal{D}) with normalizing constants $\{n\}$ and the good rate function

$$I(f) = \begin{cases} \int_T \phi^*(f'(s)) ds & \text{if } f(0) = 0 \text{ and } f \text{ is absolutely continuous} \\ \infty & \text{otherwise,} \end{cases}$$

where $\phi^*(x) = \sup_{\xi \in \mathbb{R}^d} [\langle x, \xi \rangle - \phi(\xi)]$ for $x \in \mathbb{R}^d$.

Lemma 7 Let $Z_n(t) = X_n([nt]/n), t \in T$. Then $\{\mathcal{L}(Z_n)\}_{n \in \mathbb{N}}$ is exponentially tight for the uniform topology on $D(T, \mathbb{R}^d)$.

Proof: The argument is a slight modification of de Acosta (1994a, Lemma 4.1). Let $B_r = \{z \in \mathbb{R}^d : \|z\| \leq r\}$. Following de Acosta (1994a), it suffices to prove: for every $\varepsilon > 0, a > 0$, there exist $r > 0, c > 0, m \in \mathbb{N}$ such that for $n \geq m$,

$$P\{d(Z_n, H_m(B_r)) > \varepsilon\} \leq ce^{-na}. \quad (4)$$

Arguing as in de Acosta (1994a) and using the property of stationary independent increments, we have

$$\begin{aligned} P\{Z_n \notin H_m(B_r)\} &\leq 4 \sup_{1 \leq j \leq n} P\{\|X_n(j/n)\| > r/4\} \\ &\leq 4e^{-nr/4} \sup_{1 \leq j \leq n} E \exp(n\|X_n(j/n)\|) \\ &\leq 4e^{-nr/4} \{E \exp(n\|X_n(1/n)\|)\}^n \\ &\leq 4e^{-nr/4} \beta^n, \end{aligned} \quad (5)$$

where $\beta = \sup_n E \exp(n\|X_n(1/n)\|) < \infty$, as it easily follows from assumptions (i) and (ii) and the fact that

$$E \exp\langle X_n(1/n), \xi \rangle = \{E \exp\langle X_n(1), \xi \rangle\}^{1/n}.$$

Again as in de Acosta (1994a), for any $n \geq m$ and $\sigma > 0$,

$$\begin{aligned}
& P\{Z_n \in H_n(B_r), d(Z_n, H_m(B_r)) > \varepsilon\} \\
& \leq 4m \sup_{1 \leq j \leq \frac{n}{m} + 1} P\{\|X_n(j/n)\| > \varepsilon/4\} \\
& \leq 4me^{-n\sigma\varepsilon/4} \sup_{1 \leq j \leq \frac{n}{m} + 1} E \exp(\sigma(n\|X_n(j/n)\|)) \\
& \leq 4me^{-n\sigma\varepsilon/4} \{E \exp(\sigma n\|X_n(1/n)\|)\}^{\frac{n}{m}+1} \\
& \leq 4m\tau \exp\left\{-n\left(\frac{\sigma\varepsilon}{4} - \frac{\log\tau}{m}\right)\right\}, \tag{6}
\end{aligned}$$

where $\tau = \sup_n E \exp(\sigma n\|X_n(1/n)\|) < \infty$, as argued above. Choose now $r = 4(\log\beta + a)$, $\sigma \geq 8a\varepsilon^{-1}$, and then $m \geq (\log\tau)/a$. Then if $c = 4(1+m\tau)$ and $n \geq m$, we have from (5) and (6):

$$\begin{aligned}
& P\{d(Z_n, H_m(B_r)) > \varepsilon\} \\
& \leq P\{Z_n \notin H_n(B_r)\} + P\{Z_n \in H_n(B_r), d(Z_n, H_m(B_r)) > \varepsilon\} \\
& \leq 4e^{-na} + 4m\tau e^{-na} \\
& = ce^{-na},
\end{aligned}$$

proving (4).

Proof of Theorem 6: Since for every $f_0 \in D(T, \mathbb{R}^d)$, the map $f \rightarrow \|f - f_0\|_\infty$ is \mathcal{D} -measurable, it is easily seen that $\mathcal{E} = \mathcal{D}$ satisfies conditions (i) and (ii) of Theorem 2. Let \mathcal{F} be the family of finite subsets of T , directed upward by inclusion. For $f \in D(T, \mathbb{R}^d)$, $F \in \mathcal{F}$, we define (with an obvious abuse of notation)

$$p_F(f) = f|F = (f(t_1), \dots, f(t_k)) \in (\mathbb{R}^d)^k = (\mathbb{R}^d)^F$$

where $F = \{t_1, \dots, t_k\}$ with $0 \leq t_1 < t_2 < \dots < t_k \leq 1$. For $F, G \in \mathcal{F}$, $F \subset G$, we define $p_F^G : (\mathbb{R}^d)^G \rightarrow (\mathbb{R}^d)^F$ in the obvious way. Then $\{(\mathbb{R}^d)^F, p_F^G\}$ is a projective system which together with $E = D(T, \mathbb{R}^d)$, $\{p_F\}$, satisfies the assumptions of Theorem 2.

We will show now that $\{\tilde{\mu}_n\}_{n \in \mathbb{N}}$ satisfies assumption (i) of Theorem 2, where $\tilde{\mu}_n = \mathcal{L}(Z_n)$ on \mathcal{D} . Assumption (ii) has already been proved in Lemma 7. Let $F = \{t_1, \dots, t_k\}$ be as above, and let $L_k : (\mathbb{R}^d)^k \rightarrow (\mathbb{R}^d)^k$ be the injective linear map

$$L_k(x_1, \dots, x_k) = (x_1, x_1 + x_2, \dots, x_1 + x_2 + \dots + x_k).$$

Then

$$\begin{aligned}
p_F(Z_n) &= (X_n([nt_1]/n), \dots, X_n([nt_k]/n)) \\
&= L_k(Y_n^{(1)}, \dots, Y_n^{(k)}),
\end{aligned}$$

where $Y_n^{(j)} = X_n([nt_j]/n) - X_n([nt_{j-1}]/n)$, $j = 1, \dots, k$, and $t_0 = 0$.

For $\xi_j \in \mathbb{R}^d$, $j = 1, \dots, k$, by the independence and stationarity of the increments,

$$\begin{aligned} E \exp \sum_{j=1}^k \langle Y_n^{(j)}, n\xi_j \rangle &= \Pi_{j=1}^k E \exp \langle Y_n^{(j)}, n\xi_j \rangle, \\ &= \Pi_{j=1}^k (E \exp \langle X_n(1), n\xi_j \rangle)^{m_j} \end{aligned}$$

where $m_j = ([nt_j] - [nt_{j-1}])/n$, and therefore by assumption (ii)

$$\lim_n n^{-1} \log E \exp \sum_{j=1}^k \langle Y_n^{(j)}, n\xi_j \rangle = \sum_{j=1}^k (t_j - t_{j-1}) \phi(\xi_j).$$

Now by a well-known result (Dembo & Zeitouni 1993, page 45), it follows that $\{\mathcal{L}(Y_n^{(1)}, \dots, Y_n^{(k)})\}_{n \in \mathbb{N}}$ satisfies the large deviation principle with good rate function $J_F : (\mathbb{R}^d)^k \rightarrow [0, \infty]$ given by

$$\begin{aligned} J_F(y_1, \dots, y_k) &= \sup_{\xi_1, \dots, \xi_k} [\sum_{j=1}^k \langle y_j, \xi_j \rangle - \sum_{j=1}^k (t_j - t_{j-1}) \phi(\xi_j)] \\ &= \sup_{\xi_1, \dots, \xi_k} \sum_{j=1}^k (t_j - t_{j-1}) [\langle (t_j - t_{j-1})^{-1} y_j, \xi_j \rangle - \phi(\xi_j)] \\ &= \sum_{j=1}^k (t_j - t_{j-1}) \phi^*((t_j - t_{j-1})^{-1} y_j), \end{aligned}$$

where ϕ^* is the convex conjugate of ϕ , $\phi^*(x) = \sup_{\xi \in \mathbb{R}^d} [\langle x, \xi \rangle - \phi(\xi)]$. By the contraction principle (Dembo & Zeitouni 1993, page 110) it follows that $\{\tilde{\mu}_n \circ p_F^{-1}\}_{n \in \mathbb{N}} = \{\mathcal{L}(p_F(Z_n))\}_{n \in \mathbb{N}}$ satisfies the large deviation principle with good rate function $I_F : (\mathbb{R}^d)^k \rightarrow [0, \infty]$ given by

$$I_F(x_1, \dots, x_k) = \sum_{j=1}^k (t_j - t_{j-1}) \phi^*((t_j - t_{j-1})^{-1} (x_j - x_{j-1})).$$

Now Theorem 2 applies and $\{\tilde{\mu}_n\}_{n \in \mathbb{N}}$ satisfies the large deviation principle (on $D(T, \mathbb{R}^d)$, endowed with $\|\cdot\|_\infty$ and \mathcal{D}) with good rate function

$$\begin{aligned} I_0(f) &= \sup_{F \in \mathcal{F}} I_F(p_F(f)) \\ &= \sup_{F \in \mathcal{F}} \sum_{j=1}^k (t_j - t_{j-1}) \phi^*((t_j - t_{j-1})^{-1} (f(t_j) - f(t_{j-1}))). \end{aligned}$$

We must now show: $I_0 = I$. Assume first that $I_0(f) < \infty$. Then, arguing in a fashion similar to de Acosta (1994a, Theorem 3.1), we have: if $0 \leq a_1 <$

$b_1 \leq a_2 < b_2 \leq \dots \leq a_n < b_n \leq 1$, $\rho > 0$, $\eta_j \in \mathbb{R}^d$, $\|\eta_j\| \leq 1$, and $\xi_j = \rho\eta_j$,

$$\sum_{j=1}^n (b_j - a_j) (\langle (b_j - a_j)^{-1}(f(b_j) - f(a_j)), \xi_j \rangle - \phi(\xi_j)) \leq I_0(f),$$

and

$$\begin{aligned} \sum_{j=1}^n \langle f(b_j) - f(a_j), \eta_j \rangle &\leq \rho^{-1} \sup_{1 \leq j \leq n} |\phi(\rho\eta_j)| \sum_{j=1}^n (b_j - a_j) + \rho^{-1} I_0(f), \\ \sum_{j=1}^n \|f(b_j) - f(a_j)\| &\leq \rho^{-1} \sup_{\|\xi\| \leq \rho} |\phi(\xi)| \sum_{j=1}^n (b_j - a_j) + \rho^{-1} I_0(f) \end{aligned} \quad (7)$$

But $\sup_{\|\xi\| \leq \rho} |\phi(\xi)| < \infty$ for any $\rho > 0$ by the continuity of ϕ , and the absolute continuity of f follows from (7). Also, $I_0(f) < \infty$ implies $f(0) = 0$, since $\tilde{\mu}_n(\{f \in D(T, \mathbb{R}^d) : f(0) = 0\}) = 1$ by the definition of Lévy process.

Assume now that $f(0) = 0$ and f is absolutely continuous. Writing $f(t_j) - f(t_{j-1}) = \int_{t_{j-1}}^{t_j} f'(s)ds$, the convexity of ϕ^* and Jensen's inequality yield $I_0(f) \leq I(f)$. To prove the opposite inequality, introduce as in de Acosta (1994a) the martingale

$$g_n = \sum_{j=1}^{2^n} 2^n \left[f\left(\frac{j}{2^n}\right) - f\left(\frac{j-1}{2^n}\right) \right] I_{[\frac{j-1}{2^n}, \frac{j}{2^n})}$$

on the probability space T , endowed with the Borel σ -algebra and Lebesgue measure. Then by a classical result, $g_n \rightarrow f'$ a.e. (and in $L^1(T)$). Since ϕ^* is non-negative and lower semicontinuous, it follows from Fatou's lemma that

$$\begin{aligned} \int_T \phi^*(f'(s))ds &\leq \int_T \liminf_n \phi^*(g_n(s))ds \\ &\leq \liminf_n \int_T \phi^*(g_n(s))ds \\ &= \liminf_n \sum_{j=1}^{2^n} 2^{-n} \phi^* \left(2^n \left[f\left(\frac{j}{2^n}\right) - f\left(\frac{j-1}{2^n}\right) \right] \right) \\ &\leq I_0(f). \end{aligned}$$

This completes the proof that $I_0 = I$.

We must finally show that the large deviation principle, already proved for $\{\tilde{\mu}_n\}$, also holds for $\{\mu_n\}$. To this end it is enough to show (Dembo & Zeitouni 1993, Section 4.4.2): for every $\varepsilon > 0$,

$$\limsup_n n^{-1} \log P\{\|W_n - Z_n\|_\infty > \varepsilon\} = -\infty, \quad (8)$$

where $W_n = X_n|T$. By an easy variant of de Acosta (1994a, Lemma 4.3), for any $a > 0$

$$\begin{aligned} P\{\|W_n - Z_n\|_\infty > \varepsilon\} &\leq 4n \sup_{t \leq 1/n} P\{\|X_n(t)\| > \varepsilon/4\} \\ &\leq 4ne^{-na\varepsilon/4} \sup_{t \leq 1/n} E \exp(na\|X_n(t)\|). \end{aligned} \quad (9)$$

For any $b > 0, \xi \in \mathbb{R}^d$ with $\|\xi\| \leq 1, t \leq 1/n$

$$\begin{aligned} E \exp\langle X_n(t), nb\xi \rangle &= (E \exp\langle X_n(1/n), nb\xi \rangle)^{tn} \\ &\leq E \exp |\langle X_n(1/n), nb\xi \rangle| \\ &\leq E \exp(nb\|X_n(1/n)\|) \end{aligned}$$

and it follows that

$$M_a = \sup_n \sup_{t \leq 1/n} E \exp(na\|X_n(t)\|) < \infty. \quad (10)$$

Now (8) follows from (9) and (10).

9.5 REFERENCES

- Azencott, R. (1980), ‘Grandes déviations et applications’, *Springer Lecture Notes in Mathematics* **774**, 1–176.
- Bahadur, R. R. & Zabell, S. (1979), ‘Large deviations of the sample mean in general vector spaces’, *Annals of Probability* **7**, 587–621.
- Dawson, D. & Gärtner, J. (1987), ‘Large deviations from the McKean-Vlasov limit for interacting diffusions’, *Stochastics* **20**, 247–308.
- de Acosta, A. (1985), ‘On large deviations of sums of independent random vectors’, *Springer Lecture Notes in Mathematics* **1153**, 1–14.
- de Acosta, A. (1994a), ‘Large deviations for vector valued Lévy process’, *Stochastic Processes and their Applications* **51**, 75–115.
- de Acosta, A. (1994b), ‘On large deviations of empirical measures in the τ -topology’, *Journal of Applied Probability* **31A**, 41–47. (Lajos Takács Festschrift volume).
- Dembo, A. & Zeitouni, O. (1993), *Large deviations techniques and applications*, Jones and Bartlett, Boston and London.
- Donsker, M. & Varadhan, S. R. S. (1976), ‘Asymptotic evaluation of certain Markov process expectations for large time III’, *Communications in Pure and Applied Mathematics* **29**, 389–461.

Kelley, J. (1955), *General Topology*, Van Nostrand, Princeton, NJ.

Pukhalskii, A. (1991), On functional principle of large deviations, *in* V. Sazonov & T. Shervashidze, eds, ‘New Trends in Probability and Statistics’, VSP, Moks’las, Vilnius, pp. 198–218.

10

Consistency of Bayes Estimates for Nonparametric Regression: A Review

P. Diaconis¹

D. A. Freedman²

ABSTRACT This paper reviews some recent studies of frequentist properties of Bayes estimates. In nonparametric regression, natural priors can lead to inconsistent estimators; although in some problems, such priors do give consistent estimates.

10.1 Introduction

Consider a sequence of iid pairs $(Y_1, \xi_1), (Y_2, \xi_2), \dots$ with $E(Y_i | \xi_i) = f(\xi_i)$. Here, f is an unknown function to be estimated from the data. A Bayesian approach postulates that f lies in some class of functions Θ and puts a prior distribution π on Θ . This generates a posterior distribution $\tilde{\pi}_n$ on Θ : the conditional law of the regression function f given the data $(Y_1, \xi_1), (Y_2, \xi_2), \dots, (Y_n, \xi_n)$. The prior π is said to be consistent at f if $\tilde{\pi}_n$ converges to point mass at f almost surely as $n \rightarrow \infty$.

When Θ is finite-dimensional, π will be consistent at any f in the support of π ; some additional regularity conditions are needed. If Θ is infinite-dimensional, the situation is quite different, and inconsistency is the rule rather than the exception. Section 2 reviews examples where commonly-used “hierarchical priors” lead to inconsistency in infinite-dimensional binary regression problems (the response variable Y_i takes only the values 0 and 1).

Section 3 discusses the root problem behind these inconsistencies, which is deceptively simple: if you choose p at random and toss a p -coin once, that is just like tossing a \bar{p} -coin, where \bar{p} is the average of p . In other words, a mixture of Bernoulli variables is again Bernoulli. When the relevant class of measures is not closed under mixing, in a sense to be made precise below, we believe that hierarchical priors will be consistent under standard regularity

¹Harvard University

²University of California at Berkeley

conditions. Section 4 contains one such theorem, for normal regression. Our mixing condition is satisfied, because a mixture of $N(\mu, 1)$ variables cannot be $N(\mu, 1)$.

In the balance of this section, we discuss some history. Lucien Le Cam is a major contributor to the study of frequentist properties of Bayes procedures. Le Cam's first boss in France, Etienne Halphen, was a staunch Bayesian who sought to convert the young Lucien. This was not to be, but it did stimulate a lifelong interest in Bayes procedures.

Formal work began with the thesis: Le Cam (1953) proved a version of what has come to be known as the Bernstein-von Mises theorem. Le Cam's theorems were almost sure results, with respect to the true underlying measure that had generated the data, and he proved convergence in total variation norm. Previous authors had demonstrated only convergence of distribution functions, in probability. Furthermore, Le Cam seems to have been the first to condition on all the data, not just a summary statistic (like the sample mean).

Le Cam (1958) explained how localizing the prior affects convergence of the posterior. Breiman, Le Cam & Schwartz (1964) gave versions of Doob's theorem showing consistency, starting from the joint distribution of parameters and data. Le Cam (1982) gave bounds—rather than asymptotic theory—for Bayes risk. Also see Le Cam & Yang (1990).

A more complete exposition of these results can be found in Lucien's book (Le Cam 1986). Convergence properties of Bayes estimates are closely related to the behavior of maximum likelihood estimates. Le Cam (1990) gives a beautiful overview of counter-examples in the latter area.

Frequentist properties of Bayes rules have been studied since Laplace (1774), who showed that in smooth, finite-dimensional problems, the posterior concentrates in a neighborhood of the maximum likelihood estimates. Modern versions of the result can be found in Bernstein (1934), von Mises (1964), Johnson (1967, 1970), Le Cam (1982), or Ghosh, Sinha & Joshi (1982). These results hold for almost all data sequences. In very simple settings, we obtained bounds that hold for all sequences (Diaconis & Freedman 1990).

Freedman (1963) was an early paper on nonparametric Bayes procedures, with a counter-example: there is a prior supported on all of the parameter space, whose posterior converges almost surely to the wrong answer. This paper introduced the Dirichlet and tail free priors, and showed them to be consistent. For reviews, see Ferguson (1974) or Diaconis & Freedman (1986).

Bayesian regression, with hierarchical priors, was developed in finite-dimensional settings by Lindley & Smith (1972). In non-parametric regression, there is an early paper by Kimeldorf & Wahba (1970), who use Gaussian processes to define priors; for a review, see Wahba (1990). Cox (1993) studies frequentist coverage properties of posterior confidence sets. Also see Kohn & Ansley (1987). Diaconis (1988) traces the history back to

Poincaré and gives many further references.

The simplest possible regression problem has a constant regression function. That is the location problem: $Y_i = \mu + \epsilon_i$, where μ is an unknown constant and the errors ϵ_i are iid. Diaconis & Freedman (1986) studied nonparametric priors on μ and the law of the errors; also see Doss (1984, 1985a, 1985b). Some natural priors lead to inconsistent estimates, while other priors give consistent results.

10.2 Binary regression

This section summarizes results from Diaconis & Freedman (1993a, 1993b). There is a binary response variable Y , which is related to a covariate ξ :

$$P\{Y = 1|\xi\} = f(\xi) \quad (1)$$

The problem is to estimate f from the data.

Following de Finetti (1959, 1972), we think of ξ as a sequence of 0s and 1s. Sequence space is given the usual product topology, and the parameter space Θ is the set of measurable functions f from sequence space to $[0, 1]$. The L_2 topology is installed on Θ , relative to coin-tossing measure λ in sequence space. A basic neighborhood of $f \in \Theta$ is

$$N(f, \epsilon) = \{g : \int (g - f)^2 d\lambda < \epsilon\} \quad (2)$$

We will consider a prior π on Θ , with posterior $\tilde{\pi}_n$. Then π is consistent at f provided $\tilde{\pi}_n\{N(f, \epsilon)\} \rightarrow 1$ almost surely, for all positive ϵ .

The next step is to define the hierarchical priors on Θ . Begin with a prior π_k supported on the class of functions f that depend only on the first k coordinates, or bits, in ξ . Under π_0 , the function f does not depend on ξ at all. Under π_1 , f depends only on ξ_1 . And so forth. Then treat k as an unknown “hyper-parameter”, putting prior weight w_k on k . We refer to k as the “theory index”; theory k says that $f(x)$ depends only on the first k bits of x ; and w_k are the “theory weights.” Our prior is of the form

$$\pi = \sum_{k=0}^{\infty} w_k \pi_k / \sum_{k=0}^{\infty} w_k \quad (3)$$

where

$$w_k > 0 \text{ for all } k \text{ and } \sum_{k=0}^{\infty} w_k < \infty \quad (4)$$

To complete the description of the prior, π_k must be specified. According to π_k , only the first k bits in ξ matter, so f depends only on ξ_1, \dots, ξ_k . Thus, π_k is determined by specifying the joint distribution of the 2^k possible values for f . Here, we take these to be independent and uniformly

distributed over $[0, 1]$. More general priors are considered in Diaconis & Freedman (1993a, 1993b).

We turn now to the data. For technical reasons, it is simplest to consider “balanced” data, as in Diaconis & Freedman (1993a); more conventional sampling plans are discussed in Diaconis & Freedman (1993b). At stage n , there are 2^n subjects. Each has a covariate sequence; the first n bits of these covariate sequences cover all possible patterns of length n ; each pattern appears once and only once. The remaining bits from $n+1$ onward are generated by coin tossing. Given the covariates, response variables are generated from (1); the response of subject i depends only on the covariates for that subject. The preliminaries are now finished, and we can state a theorem.

Theorem 1 *With nonparametric binary regression, balanced data, and a hierarchical uniform prior:*

- (a) π is consistent at f unless $f \equiv 1/2$;
- (b) Suppose $f \equiv 1/2$. Then π is consistent at f provided that for some $\delta > 0$, for all sufficiently large n ,

$$\sum_{k=n}^{\infty} w_k < 2^{-n(\frac{1}{2}+\delta)}$$

On the other hand, π is inconsistent at f provided that for some $\delta > 0$, for infinitely many n ,

$$\sum_{k=n}^{\infty} w_k > 2^{-n(\frac{1}{2}-\delta)}$$

The surprising point is the inconsistency result in part (b). Suppose the data are generated by tossing a fair coin: $f \equiv 1/2$. Theory 0 is true: f does not depend on ξ at all. You don’t know that, and allow theories of finite but arbitrary complexity in your prior, according to (3) and (4). In the face of all these other theories, the posterior loses faith in theory 0. The curse of dimensionality strikes again.

Regression is a natural problem, hierarchical priors are often used, and the one defined by (3) and (4) charges every weak star neighborhood of the parameter space Θ . Still, inconsistency may result. In high-dimensional problems, little can be taken for granted. “Rational use of additional information” is not a slogan to be adopted without reflection.

10.3 Why inconsistency?

What is the root cause of the inconsistency? Suppose $f \equiv 1/2$, so the data result from coin tossing, and the covariates do not matter. Thus, theory 0

is the truth. The statistician does not know this, however, and high-order theories may be deceptively attractive because they have many parameters.

To make this a little more precise, consider a design of order n , so there are 2^n subjects. According to theory n , the response of each subject is determined by the toss of a coin, where the probability is uniform on $[0, 1]$. Now one toss of a uniform coin is like one toss of a fair coin—you get heads with probability $1/2$ and tails with probability $1/2$. Thus, theory n competes with theory 0. Indeed, the predictive probability of the data under theory n is

$$\pi_n\{\text{data}\} = 1/2^{2^n}.$$

Let S be the sum of the response variables—the total number of heads. Under theory 0, the predictive probability of the data is

$$\pi_0\{\text{data}\} = \left[(2^n + 1) \binom{2^n}{S} \right]^{-1} \approx \frac{\sqrt{\pi/2}}{2^{n/2}} \pi_n\{\text{data}\}$$

because $S \approx 2^n/2$. Thus,

$$\pi_n\{\text{data}\} = \text{const. } 2^{n/2} \pi_0\{\text{data}\} \quad (5)$$

The prior π is a mixture $\sum_{k=0}^{\infty} w_k \pi_k / \sum_{k=0}^{\infty} w_k$. The posterior is a similar mixture, the posterior weight on theory k being w_k times the predictive probability of the data under π_k . If $f \equiv 1/2$, then, it is the theory weights w_k that decide consistency. If w_k declines rapidly, for example, $w_k = 1/2^k$, the weight on theory n compensates for the factor $2^{n/2}$ in (5); and the prior is consistent at $f \equiv 1/2$. On the other hand, if w_k declines slowly, for example, $w_k = 1/(k+1)^2$, the factor $2^{n/2}$ dominates, and inconsistency is the result.

The heart of the problem seems to be that a mixture of Bernoulli variables is again Bernoulli. For example, suppose the response variable takes three values, 0, 1 and 2; and, given the covariates ξ , the response is distributed as the number of heads when an $f(\xi)$ -coin is tossed twice. A mixture of bin(2, p) variables cannot be bin(2, p); the heuristic suggests that Bayes estimates will be consistent.

To discuss this kind of theorem in any degree of generality, we would need to impose smoothness conditions like those which underly the usual asymptotics of maximum likelihood estimates, including the Bernstein-von Mises theorem; and integrability conditions of the kind which underly the usual theory of entropy bounds. The second set of conditions would enable us to localize the problem, and the first set would enable us to make local estimates. Rather than pursue such technical issues here, we discuss one simple form of the theorem, for normal response variables.

10.4 Normal regression

Suppose the response variable is normal with mean μ and variance 1, where $\mu = f(\xi)$. The covariates are a sequence of 0's and 1's. We require the subjects to satisfy the balance condition as before. We assume:

Assumption 1 *Given the covariates, the response variables are independent across subjects, and normally distributed, with common variance 1 and $E\{Y | \xi\} = f(\xi)$.*

The function f is assumed to be measurable; f may be unbounded, but we require f to be square integrable (relative to coin-tossing measure). We define hierarchical priors, consistency, etc., as before. However, the π_k are assumed for convenience to be “normal” in the following sense: theory k says that $f(x)$ depends only on the first k bits of x ; under π_k , the 2^k possible values of f are independent $N(0, 1)$ variables. Thus, π is a conventional hierarchical normal prior.

This completes the setup. The main theorems can now be stated.

Theorem 2 *Suppose the design is balanced and normal in the sense of Assumption 1. Suppose the prior π is hierarchical, and the π_k are normal. Then π is consistent at all f .*

Let C_k be the class of L_2 functions f which depend only on the first k bits of the argument x . Recall that $\tilde{\pi}_n$ is the posterior given the data at stage n .

Theorem 3 *Suppose the design is balanced, and normal in the sense of Assumption 1. Suppose the prior π is hierarchical, the π_k are normal, and $f \in C_k$ for some k . Then $\tilde{\pi}_n\{C_k\} \rightarrow 1$ a.e. as $n \rightarrow \infty$.*

Theorem 2 demonstrates consistency, while Theorem 3 says that the Bayesian gets the order of a finite model right. This is a bit surprising, because many model selection algorithms over-estimate the order of a finite model. For proofs, see Diaconis & Freedman (1994).

Cox (1993) has results for a similar problem, and it may be worth a moment to indicate the differences. That paper uses a different prior, based on Gaussian processes; the covariates are deterministic and equally spaced, rather than completely at random; and results depend on the behavior of $f(x)$ at rational x , rather than a.e. properties of f .

Acknowledgments: Research of Diaconis partially supported by NSF Grant DMS 86-00235. Research of Freedman partially supported by NSF Grant DMS 92-08677.

10.5 REFERENCES

- Bernstein, S. (1934), *Theory of Probability*, GTTI, Moscow. (Russian).
- Breiman, L., Le Cam, L. & Schwartz, L. (1964), 'Consistent estimates and zero-one sets', *Annals of Mathematical Statistics* **35**, 157–161.
- Cox, D. (1993), 'An analysis of Bayesian inference for nonparametric regression', *Annals of Statistics* **21**, 903–923.
- de Finetti, B. (1959), *La probabilità, la statistica, nei rapporti con l'induzione, secondo diversi punti di vista*, Centro Internazionale Matematica Estivo Cremonese, Rome. English translation in de Finetti (1972).
- de Finetti, B. (1972), *Probability, Induction, and Statistics*, Wiley, New York.
- Diaconis, P. (1988), Bayesian numerical analysis, in S. S. Gupta & J. O. Berger, eds, 'Statistical Decision Theory and Related Topics IV', Vol. 1, pp. 163–177.
- Diaconis, P. & Freedman, D. (1986), 'On the consistency of Bayes estimates (with discussion)', *Annals of Statistics* **14**, 1–67.
- Diaconis, P. & Freedman, D. (1990), 'On the uniform consistency of Bayes estimates for multinomial probabilities', *Annals of Statistics* **18**, 1317–1327.
- Diaconis, P. & Freedman, D. (1993a), 'Nonparametric binary regression: a Bayesian approach', *Annals of Statistics* **21**, 2108–2137.
- Diaconis, P. & Freedman, D. (1993b), Nonparametric binary regression with random covariates, Technical Report 291, Department of Statistics, University of California, Berkeley. (To appear in *Probability and Mathematical Statistics*.).
- Diaconis, P. & Freedman, D. (1994), Consistency of Bayes estimates for nonparametric regression: normal theory, Technical Report 414, Department of Statistics, University of California, Berkeley.
- Doss, H. (1984), 'Bayesian estimation in the symmetric location problem', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **68**, 127–147.
- Doss, H. (1985a), 'Bayesian nonparametric estimation of the median; part I: Computation of the estimates', *Annals of Statistics* **13**, 1432–1444.

- Doss, H. (1985b), 'Bayesian nonparametric estimation of the median; part II: Asymptotic properties of the estimates', *Annals of Statistics* **13**, 1445–1464.
- Ferguson, T. (1974), 'Prior distributions on spaces of probability measures', *Annals of Statistics* **2**, 615–629.
- Freedman, D. (1963), 'On the asymptotic behavior of Bayes estimates in the discrete case', *Annals of Mathematical Statistics* **34**, 1386–1403.
- Ghosh, J. K., Sinha, B. K. & Joshi, S. N. (1982), Expansions for posterior probability and integrated Bayes risk, in S. S. Gupta & J. O. Berger, eds, 'Statistical Decision Theory and Related Topics III', Vol. 1, Academic Press, New York, pp. 403–456.
- Johnson, R. (1967), 'An asymptotic expansion for posterior distributions', *Annals of Mathematical Statistics* **38**, 1899–1906.
- Johnson, R. (1970), 'Asymptotic expansions associated with posterior distributions', *Annals of Mathematical Statistics* **41**, 851–864.
- Kimeldorf, G. & Wahba, G. (1970), 'A correspondence between Bayesian estimation on stochastic processes and smoothing by splines', *Annals of Mathematical Statistics* **41**, 495–502.
- Kohn, R. & Ansley, C. (1987), 'A new algorithm for spline smoothing and interpolation based on smoothing a stochastic process', *SIAM Journal on Scientific and Statistical Computing* **8**, 33–48.
- Laplace, P. S. (1774), 'Memoire sur la probabilité des causes par les événements', *Mémoires de mathématique et de physique présentés à l'académie royale des sciences, par divers savants, et lus dans ses assemblées*. Reprinted in Laplace's *Oeuvres Complètes* 8 27–65. English translation by S. Stigler (1986) *Statistical Science* **1** 359–378.
- Le Cam, L. (1953), 'On some asymptotic properties of maximum likelihood estimates and related Bayes estimates', *University of California Publications in Statistics* **1**, 277–330.
- Le Cam, L. (1958), 'Les propriétés asymptotiques des solutions de Bayes', *Publications de l'Institut de Statistique de l'Université de Paris* **7**, 17–35.
- Le Cam, L. (1982), On the risk of Bayes estimates, in S. S. Gupta & J. O. Berger, eds, 'Statistical Decision Theory and Related Topics III', Vol. 2, Academic Press, New York, pp. 121–138.
- Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

- Le Cam, L. (1990), 'Maximum likelihood: an introduction', *International Statistical Review* **58**, 153–172.
- Le Cam, L. & Yang, G. L. (1990), *Asymptotics in Statistics: Some Basic Concepts*, Springer-Verlag.
- Lindley, D. & Smith, A. (1972), 'Bayes estimates for the linear model', *Journal of the Royal Statistical Society* **67**, 1–19.
- von Mises, R. (1964), *Mathematical Theory of Probability and Statistics*, Academic Press, New York. H. Geiringer, ed.
- Wahba, G. (1990), *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, Philadelphia.

11

Renormalizing Experiments for Nonlinear Functionals

David L. Donoho¹

11.1 Introduction

Let $f = f(t)$, $t \in \mathbf{R}^d$ be an unknown “object” (real-valued function), and suppose we are interested in recovering the nonlinear functional $T(f)$. We know *a priori* that $f \in \mathcal{F}$, a certain convex class of functions (e.g. a class of smooth functions). For various types of measurements $Y_n = (y_1, y_2, \dots, y_n)$, problems of this form arise in statistical settings, such as nonparametric density estimation and nonparametric regression estimation; but they also arise in signal recovery and image processing. In such problems, there generally exists an “optimal rate of convergence”: the minimax risk from n observations, $R(n) = \inf_{\hat{T}} \sup_{f \in \mathcal{F}} E(\hat{T}(Y_n) - T(f))^2$, tends to zero as $R(n) \asymp n^{-r}$. There is an extensive literature on the determination of such optimal rates for a variety of functionals T , function classes \mathcal{F} , and types of observation Y_n ; the literature is really too extensive to list here, although we mention Ibragimov & Has’minskii (1981), Sacks & Ylvisaker (1981), and Stone (1980). Lucien Le Cam (1973) has contributed directly to this literature, in his typical abstract and profound way; his ideas have stimulated the work of others in the field, e.g. Donoho & Liu (1991a).

In many cases such rate results can be derived by studying the so-called *white-noise model*, where we suppose we observe the process Y_ϵ characterized infinitesimally by

$$Y_\epsilon(dt) = f(t)dt + \epsilon W(dt), \quad t \in \mathbf{R}^d \quad (1)$$

where W is a standard Wiener process and ϵ is the “noise level”, which we think of as small. At the abstract level, one can often establish that the white noise experiment $(Y_\epsilon, \mathcal{F})$ is globally asymptotically equivalent to the concrete experiment (Y_n, \mathcal{F}) , under a calibration of the form $\epsilon = \sigma/\sqrt{n}$ for some constant $\sigma > 0$; this is done for certain instances of nonparametric regression in Brown & Low (1992) and for certain instances of nonparametric density estimation in Nussbaum (1993). (These results are, to us, the most beautiful applications of the Le Cam Equivalence of

¹Stanford and University of California at Berkeley

Experiments theory.) It then follows that the minimax risk in the white noise model, $R^*(\epsilon) = \inf_{\hat{T}} \sup_{f \in \mathcal{F}} E(\hat{T}(Y_\epsilon) - T(f))^2$, will generally obey

$$R(n) \sim R^*(\sigma/\sqrt{n}), \quad n \rightarrow \infty. \quad (2)$$

Even when global asymptotic equivalence of experiments doesn't hold (e.g. when boundaries of the observation domain occur in the practical problem), the relation (2) may still hold, for the specific functional T of interest.

Owing to its connection with continuous space, $R^*(\epsilon)$ is often easy to compute. Donoho & Liu (1991b) and Donoho & Low (1992) gave several examples where, for T a linear functional, we have an exact power law

$$R^*(\epsilon) = R^*(1)(\epsilon^2)^r, \quad \epsilon > 0, \quad (3)$$

and the exponent r can be easily computed from scaling properties of the functional T and the class \mathcal{F} . Ultimately, such power laws are again due to equivalence of experiments with their own dilations and rescalings Low (1992)—another beautiful application of Le Cam's ideas.

It follows from (2)–(3) that ultimately, many rates-of-convergence results in nonparametric estimation are simply a consequence of simple scaling laws and abstract equivalence of experiments.

In this paper, we discuss the extent to which this situation continues to hold for *nonlinear* functionals T . We exhibit several nonlinear functionals T which obey exact scaling relations, and show that in such cases, the optimal rate can be derived from scaling arguments. We give examples in signal processing (finding modes, zeros, change points), in image processing (estimating Horizons and convex sets), and in surface processing (estimating the curvature and contours of surfaces). However, we also give examples to show that the renormalization argument is not always enough; it fails to work with the same generality as in the linear case.

11.2 Renormalization

We now describe a class of situations where the scaling law (3) holds. We introduce the renormalization operator $\mathcal{U}_{a,b}$ defined by $(\mathcal{U}_{a,b}f)(t) = af(bt)$.

Definition 5 *The problem $(T, \mathcal{F}, \epsilon)$ renormalizes, with renormalization exponents e_1, e_2, e_3, d_3 , if the following hold:*

$$\mathcal{U}_{a,b}\mathcal{F} = \mathcal{F} \quad \text{iff} \quad ab^{e_1} = 1 \quad (4)$$

$$\epsilon \mathcal{U}_{a,b}W =_D W \quad \text{iff} \quad \epsilon ab^{e_2} = 1 \quad (5)$$

$$T(\mathcal{U}_{a,b}f) = a^{d_3}b^{e_3}T(f) \quad \text{if } a, b \text{ obey (4)–(5)}. \quad (6)$$

For an example, let \mathcal{F} be the class of *contractions*: $\{|f(t) - f(s)| \leq |t - s|\}$. Then (4) holds with $e_1 = 1$. Let W be the d -dimensional Wiener sheet process. Then $W(bt) =_D b^{d/2} \cdot W(t)$, and so $e_2 = d/2$. Finally, let $T(f) = f(0)$. Then $T(af(b \cdot 0)) = af(0)$, so $d_3 = 1$, $e_3 = 0$. Donoho & Low (1992) gave an array of examples of this type with T a linear functional (e.g. function or derivative at a point), and \mathcal{F} a standard class of smooth functions (Sobolev or Hölder ball).

About the definition. (4)–(6) are three coupled equations, each describing invariances of certain rescalings of certain objects. Conceptually the i th equation would contain exponents d_i and e_i asserting that rescalings were invariant when $a^{d_i} \cdot b^{e_i}$ met certain conditions. In the present setting, where \mathcal{F} will always be homogeneous under scaling and W will always mean Wiener process, it turns out $d_1 = d_2 = 1$ always, and that $e_2 = d/2$ always, and so we explore only a limited range of behavior; for this reason in our definition the d_i exponent is only “visible” in equation (6).

Our aim in isolating these properties is the following

Theorem 1 Suppose (4)–(6) hold. Then

$$R^*(\epsilon) = R^*(1) \cdot (\epsilon^2)^r, \quad \epsilon > 0,$$

where

$$r = \frac{d_3 \cdot e_1 - e_3}{e_1 + d - e_2}. \quad (7)$$

Before giving the proof, we indicate the interpretation. Suppose that $R^*(1) \notin \{0, \infty\}$. Then $R^*(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow 0$; in fact at rate $(\epsilon^2)^r$. On the other hand, if $R^*(\epsilon) \in \{0, \infty\}$ for one particular value of $\epsilon > 0$, then the same is true at any other value of $\epsilon > 0$. Qualitatively, then, this result reduces the work of calculating rates of convergence into verifying the rescaling relations (4)–(6), and checking $0 < R^*(1) < \infty$.

The proof of this result, conceptually, depends on a certain rescaling process which renormalizes an estimation problem at noise level ϵ to become an equivalent estimation problem at noise level 1. The basic principle is already implicit in Low (1992). The framework we give here, spelling out exactly which objects are required to be invariant and spelling out exactly how the invariances affect the rates, seems new and useful. Our point of view and notation are closest to Donoho & Low (1992), which studies the case where T is a linear functional.

Define $\alpha = \alpha(\epsilon)$ and $\beta = \beta(\epsilon)$ as solutions to the simultaneous equations

$$\alpha \beta^{e_1+d} = 1, \quad \epsilon \alpha \beta^{e_2} = 1. \quad (8)$$

Evidently, $\alpha = \epsilon^{(e_1+d)/(e_1+d-e_2)}$ and $\beta = \epsilon^{1/(e_1+d-e_2)}$. Then the operation

$$\tilde{Y} = \mathcal{U}_{\alpha, \beta} Y_\epsilon$$

yields a new stochastic process with

$$\tilde{Y}(dt) = \alpha f(\beta t)\beta^d dt + \epsilon\alpha W(\beta dt), \quad t \in \mathbf{R}^d. \quad (9)$$

By (5) and (8), $\tilde{W}(dt) \equiv \epsilon\alpha W(\beta dt)$ defines a new standard Wiener process; setting $a = \alpha\beta^d$, $b = \beta$, $\tilde{f} = \mathcal{U}_{a,b}f$ we see that this is a standard white noise model, at noise-level 1:

$$\tilde{Y}(dt) = \tilde{f}(t)dt + \tilde{W}(dt), \quad t \in \mathbf{R}^d.$$

Now, by (6),

$$T(f) = a^{-d_3}b^{-e_3}T(\tilde{f}) = \epsilon^r T(\tilde{f}).$$

Hence the problem $(T, \mathcal{F}, \epsilon)$ of estimating T from noise-level- ϵ observations Y_ϵ is *equivalent in Le Cam's sense* to the problem $(\epsilon^r \cdot T, \tilde{\mathcal{F}}, 1)$ of estimating $\epsilon^r T(\tilde{f})$ from noise-level-1 observations. Explicitly, if \tilde{T} is any estimator of T for the noise-level 1 problem, with risk $R_1(\tilde{T}, \tilde{f}) \equiv E(\tilde{T}(\tilde{Y}) - T(\tilde{f}))^2$, the subscript indicating the noise level, then

$$\hat{T}(Y_\epsilon) \equiv \epsilon^r \tilde{T}(\tilde{Y})$$

defines an estimator for the noise level ϵ problem, with risk

$$R_\epsilon(\hat{T}, f) = (\epsilon^2)^r R_1(\tilde{T}, \tilde{f}), \quad (10)$$

where $R_\epsilon(\hat{T}, f) \equiv E(\hat{T}(Y_\epsilon) - T(f))^2$. Similarly, if \hat{T} is any estimator for the noise-level ϵ problem, then

$$\tilde{T}(\tilde{Y}) \equiv \hat{T}(Y)/\epsilon^r$$

defines an estimator for the noise level 1 problem, with risk

$$R_1(\hat{T}, f) = R_\epsilon(\tilde{T}, \tilde{f})/(\epsilon^2)^r \quad (11)$$

Combining (10)–(11) gives equivalence of minimax risks:

$$R^*(\epsilon; T, \mathcal{F}) = (\epsilon^2)^r \cdot R^*(1; T, \tilde{\mathcal{F}}).$$

Now $ab^{e_1} = \alpha\beta^{e_1+d} = 1$, so, by (4), $\tilde{\mathcal{F}} = \mathcal{U}_{a,b}\mathcal{F} = \mathcal{F}$. Hence, obviously,

$$R^*(1; T, \tilde{\mathcal{F}}) = R^*(1; T, \mathcal{F}),$$

and the theorem is proven.

As a simple example, we use the calculations in an earlier paragraph. Let $T(f) = f(0)$, $\mathcal{F} = \{\text{CONTRACTIONS}\}$. Then $r = 2/(2+d)$.

11.3 Signals: Dimension 1

We now study some functionals which arise in nonparametric problems: the mode, the maximum, the root, and the change point. These have been studied before in density estimation, nonparametric regression, and even in the white noise model (Ibragimov & Has'minskii 1980); but here we emphasize renormalizability, which forces us to study different functional classes \mathcal{F} than previously.

11.3.1 MODE OF OBJECT; PARABOLIC BOUNDARIES

Let C^-, C^+ satisfy $0 < C^- < C^+ < \infty$. Let $\text{MODE}(C^-, C^+)$ be the class of *unimodal* functions on \mathbf{R} : increasing to the left of the mode, decreasing to the right of the mode, and with a unique point of maximum. In addition, functions in this class all satisfy the quantitative regularity condition

$$C^+(t - \text{MODE})^2 \leq f(\text{MODE}) - f(t) \leq C^-(t - \text{MODE})^2.$$

Let $T(f) = \text{MODE}(f)$.

This problem renormalizes to noise level $\epsilon = 1$ as follows. If $ab^2 = 1$, and $h(x) \leq C^-x^2$, then $ah(bx) \leq aC^-(bx)^2 = C^-x^2$ also. Similarly for the inequality $h(x) \geq C^-x^2$. Consequently, (4) holds, with $e_1 = 2$. We are in dimension 1, so (5) holds with $e_2 = 1/2$. Finally,

$$\text{MODE}(\mathcal{U}_{a,b}f) = b^{-1}\text{MODE}(f)$$

so (6) holds with $d_3 = 0$, $e_3 = -1$. Applying (7) we get

$$r_{\text{MODE}} = 2/5.$$

11.3.2 MAXIMUM OF OBJECT; PARABOLIC BOUNDARIES

Let again $\mathcal{F} = \text{MODE}(C^-, C^+)$. Let $T(f) \equiv \text{Max}(f) \equiv \sup_t f(t)$ be the maximum value of f . As in the previous example, we have $e_1 = 2$, and $e_2 = 1/2$; now, however,

$$\text{Max}(\mathcal{U}_{a,b}f) = a\text{Max}(f)$$

so $d_3 = 1$, $e_3 = 0$. The problem of estimating the maximum value of the signal renormalizes to noise level $\epsilon = 1$, and by (7),

$$r_{\text{Max}} = 4/5.$$

11.3.3 ZERO OF OBJECT; LINEAR BOUNDARIES

Again with C^-, C^+ satisfy $0 < C^- < C^+ < \infty$, let $\text{ZERO}(C^-, C^+)$ be the set of functions $f(t)$ such that $f(t)$ is monotone increasing for $t \in \mathbf{R}$ and

has a unique zero $\text{ZERO}(f)$. Additionally, functions in this class satisfy the quantitative constraint

$$C^-(t - \text{ZERO}) \leq f(t) - f(\text{ZERO}) \leq C^+(t - \text{ZERO}).$$

The class $\mathcal{F} = \text{ZERO}(C^-, C^+)$ renormalizes (4) with exponents $e_1 = 1$. In dimension $d = 1$, we have (5) with exponent $e_2 = 1/2$; and finally we have

$$\text{ZERO}(\mathcal{U}_{a,b}f) = b^{-1}\text{ZERO}(f)$$

so (6) holds for $T(f) = \text{ZERO}(f)$, with exponents $d_3 = 0, e_3 = 1$. Combining these in (7) gives

$$r_{\text{ZERO}} = 2/3.$$

11.3.4 CHANGE POINT OF A BINARY OBJECT

Let $\theta \in \mathbf{R}$ and let $f_\theta = 1_{\{t \geq \theta\}}$ be a binary object with upward jump at θ . Let CHANGE denote the class $\{f_\theta\}$ of such binary objects. Let $Ch(f_\theta) = \theta$ be the changepoint functional. Compare Korostelev (1987).

If $f \in \text{CHANGE}$ then $\mathcal{U}_{a,b}f \in \text{CHANGE}$ if and only if $a = 1$ and $b > 0$. Hence (4) holds, with $e_1 = 0$. Again, $d = 1$, so (5) holds with $e_2 = 1/2$. Finally, if f has changepoint θ , then $Ch(\mathcal{U}_{a,b}f) = \theta/b$, so (6) holds for $T(f) = Ch(f)$ with $d_3 = 0, e_3 = -1$. We get

$$r_{\text{CHANGE}} = 2.$$

Notice that $r > 1$; the rate for estimating parameters of rough objects is intrinsically faster than for estimating smooth objects.

11.3.5 COMPARISON WITH TRADITIONAL LITERATURE

As indicated in the introduction, because of (2), one naturally expects these results to match up with nonparametric regression and nonparametric density estimation results. Consider for example nonparametric regression $y_i = f(t_i) + \sigma z_i, i = 1, \dots, n$ with t_i equispaced on $[0, 1]$ and consider any of the (T, \mathcal{F}) pairs just mentioned. Assuming risk-equivalence (2), the above rate calculations, which derive from trivial renormalization arguments, imply results for nonparametric regression. So for estimation of the mode of f over the class $\text{MODE}(C^-, C^+)$, we get the rate $n^{-2/5}$ in mean squared error; for estimation of the maximum, we get $n^{-4/5}$ in mean squared error; for estimation of the change-point we get n^{-2} in mean squared error. Historically, rate calculations proceeded quite differently; but these simple renormalization arguments reproduce them.

11.4 Binary Images: $d = 2$

We now consider a 2-dimensional model which represents a caricature of problems arising in image processing. Here $t = (x, y) \in \mathbf{R}^2$, and one imagines that $f(t) = 1$ or 0 according as the image is “white” or “black” at t . The idea of studying a white noise model with such functions f , and some preliminary results, were enthusiastically discussed in an oral presentation, by R. Z. Khas'minskii, at the celebration for Le Cam's 65th birthday held in Berkeley in November, 1989. See the published article (Khasminskii & Lebedev 1990); the monograph (Korostelev & Tsybakov 1993) contains many recent developments on this theme.

11.4.1 RECOVERY OF A HORIZON

Let $\theta(x)$, $x \in \mathbf{R}$ be a real-valued function, and define

$$f(x, y) = \begin{cases} 1 & y > \theta(x) \\ 0 & y \leq \theta(x) \end{cases}$$

Thus, f is an image of a bright sky above a dark landscape, with horizon θ . Let $\text{HORIZONS} = \{f : \theta \text{ is a contraction}\}$; the class of binary images with Lipschitz horizons.

Note that $by > \theta(bx)$ iff $y > \theta(bx)/b$; and that $\theta(bx)/b$ is a contraction if θ is. Hence for $\mathcal{F} = \text{HORIZONS}$, (4) holds, with renormalization exponent $e_1 = 0$ (i.e. a must be 1, but b can be anything positive).

Let $T(f) = \theta(x_0)$ be the height of the horizon at x_0 . By the obvious translation invariance of the problem, the difficulty of estimation at any fixed x_0 is the same as at $x_0 = 0$, so we set $x_0 = 0$ without loss of generality.

Evidently, $T(\mathcal{U}_{a,b}f) = b^{-1}T(f)$ if $a = 1$ and $b \neq 0$, so (6) holds, with $d_3 = 0$, $e_3 = -1$. Finally, as we have $d = 2$ (5) holds, with $e_2 = 1$. We conclude from (7) that

$$r_{\text{HORIZON}} = 1.$$

11.4.2 RECOVERY OF A CONVEX SET

Let $\mathcal{C}(\theta)$ be the class of bounded convex sets C containing 0 in the interior and with the additional property that, for any boundary point $b \in \partial C$, C contains a triangle $\Delta_{b,\theta}$ with one vertex at b , with $0 \in \Delta_{b,\theta}$ and opening angle θ at b .

The class $\mathcal{C}(\theta)$ consists of convex sets which are fairly “round” because the triangle constraint forces a certain maximum to the aspect ratio.

Let \mathcal{F} be the class of indicators functions of sets in $\mathcal{C}(\theta)$. As bC is in the class $\mathcal{C}(\theta)$ whenever C is in $\mathcal{C}(\theta)$, the class \mathcal{F} is invariant under every $\mathcal{U}_{1,b}$, $b > 0$. Hence (4) holds, with $e_1 = 0$.

We now consider two possibilities for $T(f)$.

(1) $T(f) = \text{DIAMETER}(C)$. Here $\text{DIAMETER}(C) = \sup\{d(t_1, t_2) : t_i \in C\}$. Now if $f = 1_C$, $\mathcal{U}_{1,b}f = 1_{b^{-1}C}$, $\text{DIAMETER}(b^{-1}C) = b^{-1}\text{DIAMETER}(C)$, so $d_3 = 0$, $e_3 = -1$. Hence,

$$r_{\text{DIAMETER}} = 1.$$

(2) $T(f) = \text{BDRY}_\theta(C)$. Here $\text{BDRY}_\theta(C) = \sup\{r : (r \cos(\theta), r \sin(\theta)) \in C\}$. Now if $f = 1_C$, $\mathcal{U}_{1,b}f = 1_{b^{-1}C}$, $\text{BDRY}_\theta(b^{-1}C) = b^{-1}\text{BDRY}_\theta(C)$, so $d_3 = 0$, $e_3 = -1$. Hence, again,

$$r_{\text{BDRY}} = 1.$$

11.5 Surfaces: $d = 2$

Still in dimension $d = 2$, we now consider the case where f , rather than a black-or-white image, represents a smooth surface.

11.5.1 CURVATURE FORMS

Let $\mathcal{F}(m, p, C)$ denote the class of all locally C^∞ functions, tending to zero as $|t| \rightarrow \infty$ and satisfying $\|f^{(m)}\|_p \leq C$. This class has renormalization exponent $e_1 = m - d/p$. We suppose that $m > 3 + 2/p$. We discuss measures of the curvature of the surface f at point t_0 . Let $(Hf)(t_0)$ denote the Hessian matrix of f at t_0 . In dimension d , the Hessian obeys the scaling relation $[(H\mathcal{U}_{a,b}f)(0)]_{ij} = ab^d[(Hf)(0)]_{ij}$. Consequently, the determinant form

$$T_{\text{DET}}(f) = \text{DET}^{1/d}((Hf)(0))$$

satisfies

$$T_{\text{DET}}(f) = (ab^d)T_{\text{DET}}(f);$$

and the maximum eigenvalue form

$$\begin{aligned} T_{\text{MAXEV}}(f) &= \text{MAXEV}((Hf)(0)) \\ &= \sup u^T(Hf)(0)u : u^T u = 1 \end{aligned}$$

satisfies

$$T_{\text{MAXEV}}(f) = ab^d T_{\text{MAXEV}}(f).$$

Hence, for inference about T_{DET} in $d = 2$ we get the rate

$$r_{\text{DET}} = \frac{m - 2/p - 3}{m - 2/p + 1},$$

and for inference about T_{MAXEV} we get the same rate

$$r_{\text{MAXEV}} = \frac{m - 2/p - 3}{m - 2/p + 1}.$$

11.5.2 CONTOUR ESTIMATION

For constants C^-, C^+ , with $0 < C^- < C^+ < \infty$, let $\text{CONTOUR}(C^-, C^+)$ denote the class of functions f on \mathbf{R}^2 satisfying: (i) f is concave; (ii) f attains its maximum at 0; (iii) $f(0) > 0$; (iv) for each θ in $[0, 2\pi]$ there is a unique $r_\theta > 0$ so that $f(r_\theta \cos(\theta), r_\theta \sin(\theta)) = 0$; (v) we have the Lipschitz behavior

$$C^+(r_\theta - r) \geq f(r \cos(\theta), r \sin(\theta)) \geq C^-(r_\theta - r) \quad (12)$$

valid for $r > 0$.

Functions in this class have well-defined zero-level sets (which are in fact convex). We note that conditions (i)–(iv) are invariant under every renormalization $\mathcal{U}_{a,b}$; and that (12) is invariant under renormalizations satisfying $ab = 1$. Hence, with $\mathcal{F} = \text{CONTOUR}(C^-, C^+)$, $\mathcal{U}_{a,b}\mathcal{F} = \mathcal{F}$ when $ab = 1$; so $e_1 = 1$.

For the functional $T(f) = r_\theta$, we have $d_3 = 0$, $e_3 = -1$. Hence,

$$r_{\text{CONTOUR}} = 1/2$$

11.6 Subtleties

The examples so far, though simple and easy to understand, were constructed by a process of trial and error, which revealed that renormalization in the nonlinear case is more delicate and less general than in the linear case.

11.6.1 MODE OF A SMOOTH UNIMODAL OBJECT

Return to the 1-dimensional setting of estimating the mode. Let $C > 0$ and $\beta > 0$, and let $\text{SM}(C, \beta)$ be the class of unimodal objects satisfying

$$f(x) = f(\text{MODE}) + C(x - \text{MODE})^2 + r(x)$$

where

$$|r(x)| \leq \beta|x|^3.$$

If $ab^2 = 1$ then $\tilde{f} = \mathcal{U}_{a,b}f$ satisfies $\text{MODE}(\tilde{f}) = \widetilde{\text{MODE}} = \text{MODE}/b$ and

$$\tilde{f}(x) = \tilde{f}(\widetilde{\text{MODE}}) + C(x - \widetilde{\text{MODE}})^2 + \tilde{r}(x)$$

only now

$$\begin{aligned} |\tilde{r}(x)| &= a|r(bx)| \\ &\leq a\beta|bx|^3 = ab^3\beta|x|^3 \\ &\leq b\beta|x|^3. \end{aligned}$$

Consequently

$$\mathcal{U}_{a,b} \text{SM}(C, \beta) = \text{SM}(C, b\beta).$$

If $b < 1$, $\text{SM}(C, b\beta) \subset \text{SM}(C, \beta)$. Hence we renormalize to a problem at noise level 1, but over a smaller functional class. Consequently,

$$R^*(\epsilon) = R^*(1; T, \text{SM}(C, b\beta)) (\epsilon^2)^{2/5}.$$

Now if $b < 1$, then $R^*(1; T, \text{SM}(C, b\beta)) \leq R^*(1; T, \text{SM}(C, \beta))$. On the other hand, $\text{SM}(C, b\beta)$ shrinks down to the vicinity of a parabola as $b \rightarrow 0$, it seems plausible that $R^*(1; T, \text{SM}(C, b\beta)) \rightarrow 0$ as $b \rightarrow 0$. If so, then the true state of affairs is

$$R^*(\epsilon) = o(1) (\epsilon^2)^{2/5} \quad \epsilon \rightarrow 0,$$

while renormalization alone tells us only that

$$R^*(\epsilon) \leq R^*(1) (\epsilon^2)^{2/5} \quad \epsilon \in (0, 1).$$

11.6.2 RECOVERING A SMOOTH HORIZON

Similar conclusions apply in the image model. Let $\text{SH}(C)$ denote the class of images with smooth horizons θ , with $|\theta''(x)| \leq C$ for all x ; $T(f) = \theta(0)$. Then

$$\mathcal{U}_{1,b} \text{SH}(C) = \text{SH}(bC).$$

Consequently,

$$R^*(\epsilon; T, \text{SH}(C)) = \epsilon^2 \cdot R^*(1; T, \text{SH}(bC))$$

and so renormalization establishes that $R^*(\epsilon) \leq R^*(1) \epsilon^2$, $\epsilon \in (0, 1)$. It seems plausible that $R^*(1; T, \text{SH}(bC)) \rightarrow 0$ as $b \rightarrow 0$, and so $R^*(\epsilon) = o(1) \epsilon^2$.

11.6.3 HÖLDER EXPONENT

Let $0 < \alpha_0 < \alpha_1 < \infty$ and $0 < C^- < C^+ < \infty$. Consider the class $\text{HÖLDER}(\alpha_0, \alpha_1, C^-, C^+)$ of all functions $f(t)$ on $t \in [0, \infty)$ satisfying $f(0) = 0$ and

$$C^- t^\alpha \leq f(t) \leq C^+ t^\alpha, \quad t \geq 0; \tag{13}$$

where $\alpha \in [\alpha_0, \alpha_1]$.

Our aim is to estimate $T(f) = \alpha$. Evidently $d_3 = e_3 = 0$, and rescaling which preserves C^- and C^+ must obey

$$ab^\alpha = 1.$$

However, this scaling is *parameter-dependent*. Hence, the estimation of the Hölder exponent does not renormalize in the sense of our definition.

11.7 Rigor?

Renormalization is a delicious piece of “general nonsense”. Is there any more to it than this? Well, in order for the renormalization approach to have any interest, one must, in addition to the “soft” process of calculating renormalization exponents, engage in the “hard” process of *showing that* $R^*(1) < \infty$. We have ignored this question so far; it may be established in concrete cases “by hand”, as one might expect. Here we only reflect on “general” ideas true to the Le Cam spirit.

11.7.1 THE MODULUS OF CONTINUITY

Finiteness of $R^*(1)$ can be connected with another question: continuity of T as a functional of f .

Definition 6 *The L^2 -modulus of continuity of T over the class \mathcal{F} is*

$$\omega(\epsilon; T, \mathcal{F}) = \sup\{|T(f_1) - T(f_0)| : \|f_1 - f_0\|_2 \leq \epsilon, f_i \in \mathcal{F}\}$$

If $\omega(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow 0$ then T is a continuous functional over the class \mathcal{F} in L_2 distance. In the jargon of applied mathematics, the problem of recovering T from observations $y = f + z$ is *stable* under perturbations z with $\|z\|_2$ small.

The modulus of continuity obeys precisely the same scaling relations as $R^*(\epsilon)$:

Theorem 2 *Suppose (4) and (6) hold. Then*

$$\omega(\epsilon) = \omega(1)\epsilon^r, \quad \epsilon > 0$$

where r is the same as (7).

The proof uses a rescaling argument similar to Theorem 1 of Donoho & Low (1992)

The following lower bound is along the lines of Donoho & Liu (1991a, 1991b).

Theorem 3 *Let T and \mathcal{F} be arbitrary. Then*

$$R^*(\epsilon) \geq \rho \cdot \omega^2(\epsilon)$$

for an absolute constant ρ .

The proof is as follows. Without loss of generality, let (f_0, f_1) attain the modulus at ϵ . The minimax risk, for squared error, over the two-element family $\mathcal{F} = \{f_0, f_1\}$, for estimating the parameter θ taking the value 0 at f_0 and 1 at f_1 can be calculated using a formula of Le Cam (1986, p. 49); let this constant be ρ . (It is actually independent of the f_i). The minimax risk for estimating a parameter taking two values $\omega(\epsilon)$ apart is simply $\omega^2(\epsilon)$ times ρ .

11.7.2 CONSTRUCTION OF OUR EXAMPLES

We can now explain the trial-and-error process that led to the specific examples we have used in this paper. For each couple (T, \mathcal{F}) , we planned to use in this paper, we checked for $\omega(1; T, \mathcal{F}) < \infty$. When this didn't hold, we knew that necessarily $R^*(1) = \infty$, and so we considered modifications to our class \mathcal{F} to make ω finite.

11.7.3 HARDEST SUBPROBLEMS AND RENORMALIZATION

We digress to make a conceptual point. In the renormalizable case, if the functional T can be estimated at all, then it can be estimated at the rate $\omega^2(\epsilon)$. Indeed, if $R^*(1) < \infty$, then $\omega(1) < \infty$, so

$$R^*(\epsilon) = \text{CONST} \cdot \omega^2(\epsilon) \quad \epsilon > 0;$$

(in fact $\text{CONST} = R^*(1)/\omega^2(1)$). Now Donoho & Liu (1991a, 1991b) show that for linear functionals, we must have $R^*(\epsilon) \asymp \omega^2(\epsilon)$. In this respect, if renormalization works for nonlinear functionals, it entails a kind of quasi-linearity of the functional T over the class \mathcal{F} .

There is another way to put this. We borrow terminology from Donoho & Liu (1991a). The quantity $\omega^2(\epsilon)$ measures, up to a constant, the difficulty of the hardest 2-point subproblem in \mathcal{F} . Hence, if renormalization works, *then the difficulty for estimating T over \mathcal{F} is equivalent, to within constants, to the difficulty of estimating T over the hardest two-point subproblem*. This is another sense in which success of renormalization entails quasi-linearity.

We know that already for global quadratic functionals, the difficulty of the full problem can fail to be equivalent to the difficulty of the hardest two-point subproblem; see the discussion in Donoho & Nussbaum (1990). Hence we expect that the renormalization approach of this paper fails in such cases.

11.7.4 GLOBAL T

In principle, the renormalization apparatus extends to global functionals, whose value is determined by the cumulative effects of contributions from various regions. Unfortunately, it seems that renormalization is less successful in such settings. Our opinion is that the renormalization approach is well-suited only for problems of estimating functionals which are *local* in the sense that they deal with locations, slopes, curvatures, etc.

For example, in the convex contour setting considered earlier, we might have considered the functional $T(f) = \text{AREA}(C)$, which yields $d_3 = 0$, $e_3 = -2$. This gives rise to a rate prediction

$$r_{\text{AREA}} = 2$$

which is the “parametric” rate for parameters of discontinuous images in white noise. However, $R^*(1) = +\infty$ for this functional. The culprit is heteroscedasticity: reasonable estimators for the Area have a variance which depends on the length of the perimeter, and so larger sets have more variable estimates – the variance growing without bound as the size of the set increases. As the class $\mathcal{C}(\theta)$ is scale invariant, the variance of reasonable estimators is unbounded over this class, and so for this global functional, $R^*(1) = +\infty$.

It seems that a transformation of the functional considered, to $T(f) = \sqrt{\text{AREA}}(C)$, is variance stabilizing, and leads to finite minimax risk.

11.8 Speculation

The “general nonsense” we have discussed here extends “automatically” to certain *inverse problems*, where we observe noisy data, not on f directly, but on a transform Kf . Linear inverse problems which can be treated easily include the cases where K is the Radon Transform or Abel Transform. The key condition is that K be homogeneous under renormalization, either exactly or approximately; compare Donoho & Low (1992). But “general nonsense” suggests (to our peril?) that there is no reason to stop with homogeneous *linear* transforms. So, we discuss a homogeneous nonlinear inverse problem arising in computer vision.

Let BUMP be the set of unimodal functions $\phi(t)$ with (i) nested convex level sets; (ii) boundary conditions $\phi(t) \rightarrow 0$ as $|t| \rightarrow \infty$; (iii) $\phi \in \mathcal{F}(m, p, C)$, $m \geq 3$.

Suppose we are given data as $f(t) = |\nabla \phi(t)|^2$, and we wish to recover $\phi(t)$. In the jargon of applied mathematics, this is the problem of solving the *Eikonal equation*, to recover ϕ from data $|\nabla \phi(t)|^2$. In the jargon of computer vision, this is the problem of *shape-from-shading*.

Let $T(f) = \phi(0)$. Then notice that the data $\mathcal{U}_{a,b}f$ arise as $|\nabla \mathcal{U}_{\alpha,\beta}\phi|^2$ with $b = \beta$, $a = (\alpha\beta)^2$. Also

$$T(\mathcal{U}_{a,b}f) = (\mathcal{U}_{\alpha,\beta}\phi)(0) = \alpha \cdot \phi(0) = \sqrt{a}/b \cdot T(f)$$

Consequently, we may view T as homogeneous, with exponents $d_3 = 1/2$, $e_3 = -1$. Also, $e_1 = m - d/p$ and $e_2 = d/2 = 1$. Evidently, the whole problem renormalizes, with rate $r_{\text{shape}} = (m/2 + 1)/(m - d/p + d/2)$.

The applied mathematics community’s study of recovering ϕ from f is currently focused on the question of whether there is a unique solution to the problem with *noiseless* data. It would be interesting to know if $\omega(1; T, \text{BUMP}) < \infty$. This would imply that recovery of shape-from-shading is stable, in the sense of applied mathematics; we could next investigate the question whether $R^*(1, T, \text{BUMP}) < \infty$. This would imply that we can recover shape from shading successfully in the presence of white Gaussian noise, and in other asymptotically equivalent experiments.

Acknowledgments: Originally presented as a talk in the Neyman seminar. Thanks to Charles Kooperberg for several discussions on white-noise image processing and to NSF-DMS 92-09130 for support. Thanks also to the editor, David Pollard, who read the manuscript carefully and proposed many improvements.

11.9 REFERENCES

- Brown, L. D. & Low, M. G. (1992), Asymptotic equivalence of nonparametric regression and white noise, Technical report, Department of Mathematics, Cornell University.
- Donoho, D. L. & Liu, R. C. (1991a), ‘Geometrizing rates of convergence, II’, *Annals of Statistics* **19**, 633–667.
- Donoho, D. L. & Liu, R. C. (1991b), ‘Geometrizing rates of convergence, III’, *Annals of Statistics* **19**, 668–701.
- Donoho, D. L. & Low, M. (1992), ‘Renormalization exponents and optimal pointwise rates of convergence’, *Annals of Statistics* **20**, 944–970.
- Donoho, D. L. & Nussbaum, M. (1990), ‘Minimax quadratic estimation of a quadratic functional’, *Journal of Complexity* **6**, 290–323.
- Ibragimov, I. A. & Has’minskii, R. Z. (1980), ‘Estimates of the signal, its derivatives, and point of maximum for gaussian distributions’, *Theory of Probability and Its Applications* **25**, 703–720.
- Ibragimov, I. A. & Has’minskii, R. Z. (1981), *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.
- Khasminskii, R. Z. & Lebedev, V. S. (1990), ‘On the properties of parametric estimators for areas of a discontinuous image’, *Problems of Control and Information Theory* pp. 375–385.
- Korostelev, A. P. (1987), ‘Minimax estimation of a discontinuous signal’, *Theory of Probability and Its Applications* **32**, 796–799.
- Korostelev, A. P. & Tsybakov, A. B. (1993), *Minimax Theory of Image Reconstruction*, Vol. 82 of *Lecture Notes in Statistics*, Springer-Verlag, New York.
- Le Cam, L. (1973), ‘Convergence of estimates under dimensionality restrictions’, *Annals of Statistics* **19**, 633–667.
- Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

- Low, M. G. (1992), 'Renormalization and white noise approximation for nonparametric functional estimation problems', *Annals of Statistics* **20**, 545–554.
- Nussbaum, M. (1993), Asymptotic equivalence of density estimation and white noise, Technical report, Institute for Applied Stochastic Analysis, Berlin.
- Sacks, J. & Ylvisaker, N. D. (1981), 'Asymptotically optimum kernels for density estimation at a point', *Annals of Statistics* **9**, 334–346.
- Stone, C. J. (1980), 'Optimum rates of convergence for nonparametric estimators', *Annals of Statistics* **8**, 1348–1360.

12

Universal Near Minimaxity of Wavelet Shrinkage

D. L. Donoho¹, I. M. Johnstone²
G. Kerkyacharian³, D. Picard⁴

ABSTRACT We discuss a method for curve estimation based on n noisy data; one translates the empirical wavelet coefficients towards the origin by an amount $\sqrt{2 \log(n)} \cdot \sigma / \sqrt{n}$. The method is nearly minimax for a wide variety of loss functions—e.g. pointwise error, global error measured in L^p norms, pointwise and global error in estimation of derivatives—and for a wide range of smoothness classes, including standard Hölder classes, Sobolev classes, and Bounded Variation. This is a broader near-optimality than anything previously proposed in the minimax literature. The theory underlying the method exploits a correspondence between statistical questions and questions of optimal recovery and information-based complexity. This paper contains a detailed proof of the result announced in Donoho, Johnstone, Kerkyacharian & Picard (1995).

12.1 Introduction

In recent years, mathematical statisticians have been interested in estimating infinite-dimensional parameters—curves, densities, images, A paradigmatic example is the problem of *nonparametric regression*,

$$y_i = f(t_i) + \sigma \cdot z_i, \quad i = 1, \dots, n, \tag{1}$$

where f is the unknown function of interest, the t_i are equispaced points on the unit interval, and $z_i \stackrel{iid}{\sim} N(0, 1)$ is a Gaussian white noise. Other problems with similar character are *density estimation*, recovering the density f from $X_1, \dots, X_n \stackrel{iid}{\sim} f$, and *spectral density estimation*, recovering f from X_1, \dots, X_n a segment of a Gaussian zero-mean second-order stationary process with spectral density $f(\xi)$.

¹Stanford University

²Stanford University

³Université de Picardie

⁴Université de Paris VII

For simplicity, we focus on this nonparametric regression model (1) and a proposal of Donoho & Johnstone (1994); similar results are possible in the density estimation model (Johnstone, Kerkyacharian & Picard 1992, Donoho, Johnstone, Kerkyacharian & Picard 1996). We suppose that we have $n = 2^{J+1}$ data of the form (1) and that σ is known.

1. Take the n given numbers and apply an empirical wavelet transform W_n^n , obtaining n empirical wavelet coefficients $(w_{j,k})$. This transform is an order $O(n)$ transform, so that it is very fast to compute; in fact faster than the Fast Fourier Transform.
2. Set a threshold $t_n = \sqrt{2 \log(n)} \cdot \sigma / \sqrt{n}$, and apply the soft threshold nonlinearity $\eta_t(w) = \text{sgn}(w)(|w| - t)_+$ with threshold value $t = t_n$. That is, apply this nonlinearity to each one of the n empirical wavelet coefficients, obtaining $\hat{\theta}_{jk} = \eta_t(w_{jk})$. [In practice, shrinkage is only applied at the finer scales $j \geq j_0$.]
3. Invert the empirical wavelet transform, getting the estimated curve $\hat{f}_n^*(t)$.

The empirical wavelet transform is implemented by a pyramidal filtering scheme: for the reader's convenience, we recall some of its features in the Appendix. The output of the wavelet thresholding procedure may be written

$$\hat{f}_n^* = \sum_k w_{j_0,k} \phi_{j_0,k} + \sum_{j \geq j_0, k} \hat{\theta}_{j,k} \psi_{j,k} \quad (2)$$

The 'scaling functions' $\phi_{j_0,k} = (\phi_{j_0,k}(t_i), i = 1, \dots, n)$ and 'wavelets' $\psi_{j,k} = (\psi_{j,k}(t_i), i = 1, \dots, n)$ appearing in (2) are just the rows of W_n^n as constructed above. The 'wavelet' is typically of compact support, is (roughly) located at $k2^{-j}$ and contains frequencies in an octave about 2^j . [We use quotation marks since the true wavelet ψ and scaling function ϕ of the mathematical theory are not used explicitly in the algorithms applied to finite data: rather they appear as limits of the infinitely repeated cascade.]

A number of examples and properties of this procedure are set out in Donoho et al. (1995). In brief, the rationale is as follows. Many functions $f(t)$ of practical interest are either globally smooth, or have a small number of singularities (e.g. discontinuities in the function or its derivatives). Due to the smoothness and localisation properties of the wavelet transform, the wavelet coefficients $\theta_{j,k}(f)$ of such functions are typically sparse: most of the energy in the signal is concentrated in a small number of coefficients—corresponding to low frequencies, or to the locations of the singularities. On the other hand, the orthogonality of the transform W_n^n guarantees that the noise remains white and Gaussian in the transform domain; that is, more or less evenly spread among the coefficients. It is thus appealing to use a thresholding operation, which sets most small coefficients to zero, while allowing large coefficients (presumably containing signal) to pass unchanged or slightly shrunken.

The purpose of this paper is to set out a broad near-minimax optimality property possessed by this wavelet shrinkage method. We consider a large range of error measures and function classes: if a single estimator is near optimal whatever be the choice of error measure and function class, then it clearly enjoys a degree of robustness to these parameters which may not be known precisely in a particular application.

The empirical transform corresponds to a theoretical wavelet transform which furnishes an orthogonal basis of $L^2[0, 1]$. This basis has elements (wavelets) which are in C^R and have, at high resolutions, D vanishing moments. The fundamental discovery about wavelets that we will be using is that they provide a “universal” orthogonal basis: an unconditional basis for a very wide range of smoothness spaces: all the Besov classes $B_{p,q}^\sigma[0, 1]$ and Triebel classes $F_{p,q}^\sigma[0, 1]$ in a certain range $0 \leq \sigma < \min(R, D)$. Each of these function classes has a norm $\|\cdot\|_{B_{p,q}^\sigma}$ or $\|\cdot\|_{F_{p,q}^\sigma}$, which measures smoothness. Special cases include the traditional Hölder (-Zygmund) classes $\Lambda^\alpha = B_{\infty,\infty}^\alpha$ and Sobolev Classes $W_p^m = F_{p,2}^m$.

These function spaces are relevant to statistical theory since they model important forms of spatial inhomogeneity not captured by the Sobolev and Hölder spaces alone (cf. Donoho & Johnstone (1992), Johnstone (1994)). For more about these spaces and the universal basis property, see the Lemarié & Meyer (1986) or the books of Frazier, Jawerth & Weiss (1991) and Meyer (1990). Some relevant facts are summarized in the Appendix, including the unconditional basis property and sequence space characterisations of spaces that we use below.

Definition $\mathcal{C}(R, D)$ is the scale of all spaces $B_{p,q}^\sigma$ and all spaces $F_{p,q}^\sigma$ which embed continuously in $C[0, 1]$, so that $\sigma > 1/p$, and for which the wavelet basis is an unconditional basis, so that $\sigma < \min(R, D)$.

Now consider a global loss measure $\|\cdot\| = \|\cdot\|_{\sigma', p', q'}$ taken from the $B_{p,q}^\sigma$ or $F_{p,q}^\sigma$ scales, with $\sigma' \geq 0$. With $\sigma' = 0$ and p', q' chosen appropriately, this means we can consider L^2 loss, L^p loss $p > 1$, etc. We can also consider losses in estimating the derivatives of some order by picking $\sigma' > 0$. We consider a priori classes $\mathcal{F}(C)$ taken from norms in the Besov and Triebel scales with $\sigma > 1/p$ —for example, Sobolev balls.

In addition to the above constraints, we shall refer to three distinct zones of parameters $\mathbf{p} = (\sigma, p, q, \sigma', p', q')$:

$$\begin{aligned} \text{regular: } \mathcal{R} &= \{p' \leq p\} \cup \{p' > p, (\sigma + 1/2)p > (\sigma' + 1/2)p'\} \\ \text{logarithmic: } \mathcal{L} &= \{p' > p, (\sigma + 1/2)p < (\sigma' + 1/2)p'\} \\ \text{critical: } \mathcal{C} &= \{p' > p, (\sigma + 1/2)p = (\sigma' + 1/2)p'\}. \end{aligned}$$

The regular case \mathcal{R} corresponds to the familiar rates of convergence usually found in the literature: for example with quadratic loss ($\sigma' = 0, p' = 2$) the regularity condition $\sigma > 1/p$ forces us into the regular case. The existence of the logarithmic region \mathcal{L} was noted in an important paper by Nemirovskii (1985): this corresponds to lower degrees of smoothness σ of

the function space \mathcal{F} . The critical zone \mathcal{C} separates \mathcal{R} and \mathcal{L} and exhibits the most complex phenomena.

In general, an exactly minimax estimation procedure would depend on which error measure and function class is used (e.g. Donoho & Johnstone (1992)). The chief result of this paper says that for error measures and function classes from either Besov or Triebel scales, the specific wavelet shrinkage method described above is always within a logarithmic factor of being minimax.

Theorem 1 *Pick a loss $\|\cdot\|$ taken from the Besov and Triebel scales $\sigma' \geq 0$, and a ball $\mathcal{F}(C; \sigma, p, q)$ arising from an $\mathcal{F} \in \mathcal{C}(R, D)$, so that $\sigma > 1/p$; and suppose the collection of indices obey $\sigma > \sigma' + (1/p - 1/p')_+$, so that the object can be consistently estimated in this norm. There is a rate exponent $r = r(\mathbf{p})$ with the following properties:*

[1] *The estimator \hat{f}_n^* attains this rate within a logarithmic factor; with constants $C_1(\mathcal{F}(C), \psi)$,*

$$\sup_{f \in \mathcal{F}(C)} P(\|\hat{f}_n^* - f\| \geq C_1 \cdot \log(n)^{e_1 + e_{C+}} \cdot C^{1-r} \cdot (\sigma \sqrt{\log n/n})^r) \rightarrow 0.$$

[2] *This rate is essentially optimal: for some other constant $C_2(\|\cdot\|, \mathcal{F})$*

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}(C)} P(\|\hat{f} - f\| \geq C_2 \cdot \log(n)^{e_{LC} + e_{C-}} \cdot C^{1-r} \cdot (\sigma / \sqrt{n})^r) \rightarrow 1.$$

The rate exponent $r = r(\mathbf{p})$ satisfies

$$r = \frac{\sigma - \sigma'}{\sigma + 1/2} \quad \mathbf{p} \in \mathcal{R} \quad (3)$$

$$r = \frac{\sigma - \sigma' - (1/p - 1/p')_+}{\sigma + 1/2 - 1/p} \quad \mathbf{p} \in \mathcal{L} \cup \mathcal{C}. \quad (4)$$

The logarithmic exponent e_1 may be taken as:

$$e_1 = \begin{cases} 0 & \sigma' > 1/p' \\ 1/\min(1, p', q') - 1/q' & 0 \leq \sigma' \leq 1/p', \text{ Besov Loss} \\ 1/\min(1, p', q') - 1/\min(p', q') & 0 \leq \sigma' \leq 1/p', \text{ Triebel Loss} \end{cases}; \quad (5)$$

On \mathcal{R} , all exponents $e_{LC}, e_{C\pm}$ vanish. On \mathcal{L} , $e_{LC} = r/2$ and $e_{C\pm}$ vanish. On \mathcal{C} , $e_{LC} = r/2$ and the bounds $e_{C\pm}$ both have the form:

$$e_{C\pm} = \frac{1}{p'} \left(\frac{p'}{\tilde{q}'} - \frac{p}{\tilde{q}} \right)_+, \quad \begin{array}{ll} \text{for } e_{C+}: & \tilde{q}' = p' \wedge q', \quad \tilde{q} = p \vee q \\ \text{for } e_{C-}: & \tilde{q}' = p' \vee q', \quad \tilde{q} = p \wedge q. \end{array}$$

On \mathcal{C} , in certain cases, sharper results hold: (i) For Besov \mathcal{F} and norm, $e_{C+} = e_{C-}$, with $\tilde{q}' = q'$ and $\tilde{q} = q$. (ii) For Besov \mathcal{F} and Triebel norm, for e_{C-} , $\tilde{q}' = p'$ and $\tilde{q} = q$, and further if $q' \geq p$, $e_{C+} = e_{C-}$.

Remarks The index suffices are mnemonic: e_{LC} is non-zero only on $\mathcal{L} \cup \mathcal{C}$, while $e_{C\pm}$ are non-zero only in the critical case \mathcal{C} .

Thus in the regular case \mathcal{R} , when $\sigma' > 1/p'$, all indices e_1, e_{LC}, e_\pm vanish, and the upper and lower rates differ by $(\log n)^{\gamma/2}$. In fact, thresholding at the fixed level $\sqrt{2 \log n}$ is necessarily sub-optimal by this amount (cf. e.g. Hall & Patil (1994)): this is a price paid for such broad near-minimality.

In the logarithmic case \mathcal{L} , if $\sigma' > 1/p'$, the upper and lower rates agree, and so the rate of convergence result for wavelet shrinkage is in fact sharp.

In the critical case \mathcal{C} , the results to date are sharp (at the level of rates) in certain cases when $\sigma' > 1/p'$: (a) for Besov \mathcal{F} and norm, and (b) for Besov \mathcal{F} and Triebel norm if also $q' \geq p$.

By elementary arguments, these results imply similar results for other combinations of loss and a-priori class. For example, we can reach similar conclusions for L^1 loss, though it is not nominally in the Besov and Triebel scales; and we can also reach similar conclusions for the a-priori class of functions of total variation less than C , also not nominally in $\mathcal{C}(R, D)$. Such variations follow immediately from known inequalities between the desired norms and relevant Besov and Triebel classes.

At a first reading, all material relating to Triebel spaces and bodies can be skipped: we note here simply that they are of interest since the L_p -Sobolev norms (for $p \neq 2$), including the L_p norm itself, lie within the Triebel scale (and not the Besov).

Theorem 1 is an extended version of Theorem 4, announced without proof in Donoho et al. (1995): to make this paper more self-contained, some material from that paper is included in this one.

12.2 A Sequence Space Model

Consider the following *Sequence Model*. We start with an index set \mathcal{I}_n of cardinality n , and we observe

$$y_I = \theta_I + \epsilon \cdot z_I, \quad I \in \mathcal{I}_n, \tag{6}$$

where $z_I \stackrel{iid}{\sim} N(0, 1)$ is a Gaussian white noise and ϵ is the noise level. The index set \mathcal{I}_n is the first n elements of a countable index set \mathcal{I} . From the n data (6), we wish to estimate the object with countably many coordinates $\theta = (\theta_I)_I$ with small loss $\|\hat{\theta} - \theta\|$. The object of interest belongs *a priori* to a class Θ , and we wish to achieve a *Minimax Risk* of the form

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| > \omega\}$$

for a special choice $\omega = \omega(\epsilon)$. About the error norm, we assume that it is *solid* and *orthosymmetric*, namely that the coordinates

$$|\xi_I| \leq |\theta_I| \quad \forall I \quad \implies \quad \|\xi\| \leq \|\theta\|. \tag{7}$$

Moreover, we assume that the *a priori* class is also solid and orthosymmetric, so

$$\theta \in \Theta \quad \text{and} \quad |\xi_I| \leq |\theta_I| \quad \forall I \quad \Rightarrow \quad \xi \in \Theta. \quad (8)$$

Finally, at one specific point (21) below we will assume that the loss measure is either convex, or at least ρ -convex $0 < \rho \leq 1$, in the sense that $\|\theta + \xi\|^\rho \leq \|\theta\|^\rho + \|\xi\|^\rho$; 1-convex is just convex.

Results for this model will imply Theorem 1 by suitable identifications. Thus we will ultimately interpret

- [1] (θ_I) as wavelet coefficients of f ;
- [2] $(\hat{\theta}_I)$ as empirical wavelet coefficients of an estimate \hat{f} ; and
- [3] $\|\hat{\theta} - \theta\|$ as a norm equivalent to $\|\hat{f} - f\|$.

We will explain such identifications further in section 12.7 below.

12.3 Solution of an Optimal Recovery Model

Before tackling data from (6), we consider a simpler abstract model, in which noise is deterministic (Compare Micchelli (1975), Micchelli & Rivlin (1977), Traub, J., Wasilkowski, G. & Woźniakowski (1988)). The approach of analyzing statistical problems by deterministic noise has been applied previously in (Donoho 1994b, Donoho 1994a). Suppose we have an index set \mathcal{I} (not necessarily finite), an object (θ_I) of interest, and observations

$$x_I = \theta_I + \delta \cdot u_I, \quad I \in \mathcal{I}. \quad (9)$$

Here $\delta > 0$ is a known “noise level” and (u_I) is a nuisance term known only to satisfy $|u_I| \leq 1 \quad \forall I \in \mathcal{I}$. We suppose that the nuisance is chosen by a clever opponent to cause the most damage, and evaluate performance by the worst-case error:

$$E_\delta(\hat{\theta}, \theta) = \sup_{|u_I| \leq 1} \|\hat{\theta}(x) - \theta\|. \quad (10)$$

12.3.1 OPTIMAL RECOVERY—FIXED Θ

The existing theory of optimal recovery focuses on the case where one knows that $\theta \in \Theta$, and Θ is a fixed, known *a priori* class. One wants to attain the minimax error

$$E_\delta^*(\Theta) = \inf_{\hat{\theta}} \sup_{\Theta} E_\delta(\hat{\theta}, \theta).$$

Very simple upper and lower bounds are available.

Definition 7 *The modulus of continuity of the estimation problem is*

$$\Omega(\delta; \|\cdot\|, \Theta) = \sup \{ \|\theta^0 - \theta^1\| : \theta^0, \theta^1 \in \Theta, |\theta_I^0 - \theta_I^1| \leq \delta, \forall I \in \mathcal{I} \}. \quad (11)$$

When the context makes it clear, we sometimes write simply $\Omega(\delta)$.

Proposition 1

$$E_\delta^*(\Theta) \geq \Omega(\delta)/2. \quad (12)$$

Proof: Suppose θ^0 and θ^1 attain the modulus. Then under the observation model (9) we could have observations $x = \theta^0$ when the true underlying $\theta = \theta^1$, and vice versa. So whatever we do in reconstructing θ from x must suffer a worst case error of half the distance between θ^1 and θ^0 . \square

A variety of rules can nearly attain this lower bound.

Definition 8 *A rule $\hat{\theta}$ is feasible for Θ if, for each $\theta \in \Theta$ and for each observed (x_I) satisfying (9),*

$$\hat{\theta} \in \Theta, \quad (13)$$

$$|\hat{\theta}_I - x_I| \leq \delta. \quad (14)$$

Proposition 2 *A feasible reconstruction rule has error*

$$\|\hat{\theta} - \theta\| \leq \Omega(2\delta), \quad \theta \in \Theta. \quad (15)$$

Proof: Since the estimate is feasible, $|\hat{\theta}_I - \theta_I| \leq 2\delta \forall I$, and $\theta, \hat{\theta} \in \Theta$. The bound follows by the definition (11) of the modulus. \square

Comparing (15) and (12) we see that, quite generally, *any feasible procedure is nearly minimax*.

12.3.2 SOFT THRESHOLDING IS AN ADAPTIVE METHOD

In the case where Θ might be any of a wide variety of sets, one can imagine that it would be difficult to construct a procedure which is near-minimax over each one of them—i.e. for example that the requirements of feasibility with respect to many different sets would be incompatible with each other. Luckily, if the sets in question are all orthosymmetric and solid, a single idea—shrinkage towards the origin—leads to feasibility independently of the details of the set’s shape.

Consider a specific shrinker based on the soft threshold nonlinearity $\eta_t(y) = \text{sgn}(y)(|y| - t)_+$. Setting the threshold level equal to the noise level $t = \delta$, we define

$$\hat{\theta}_I^{(\delta)}(y) = \eta_t(x_I), \quad I \in \mathcal{I}. \quad (16)$$

This pulls each noisy coefficient x_I towards 0 by an amount $t = \delta$, and sets $\hat{\theta}_I^{(\delta)} = 0$ if $|x_I| \leq \delta$. Because it pulls each coefficient towards the origin by at least the noise level, it satisfies the *uniform shrinkage condition*:

$$|\hat{\theta}_I| \leq |\theta_I|, \quad I \in \mathcal{I}. \quad (17)$$

Theorem 2 *The Soft Thresholding estimator $\hat{\theta}^{(\delta)}$ defined by (16) is feasible for every Θ which is solid and orthosymmetric.*

Proof: $|\hat{\theta}_I^{(\delta)} - x_I| \leq \delta$ by definition; while (17) and the assumption (8) of solidness and orthosymmetry guarantee that $\theta \in \Theta$ implies $\hat{\theta}^{(\delta)} \in \Theta$. \square

This shows that soft-thresholding leads to nearly-minimax procedures over all combinations of symmetric *a priori* classes and symmetric loss measures.

12.3.3 RECOVERY FROM FINITE, NOISY DATA

The optimal recovery and information-based complexity literature generally posits a finite number n of noisy observations. And, of course, this is consistent with our model (6). So consider observations

$$x_I = \theta_I + \delta \cdot u_I, \quad I \in \mathcal{I}_n. \quad (18)$$

The minimax error in this setting is

$$E_{n,\delta}^*(\Theta) = \inf_{\hat{\theta}} \sup_{\Theta} \|\hat{\theta} - \theta\|.$$

To see how this setting differs from the “complete-data” model (9), we set $\delta = 0$. Then we have the problem of inferring the complete vector $(\theta_I : I \in \mathcal{I})$ from the first n components $(\theta_I : I \in \mathcal{I}_n)$. To study this, we need the definition

Definition 9 *The tail- n -width of Θ in norm $\|\cdot\|$ is*

$$\Delta(n; \|\cdot\|; \Theta) = \sup\{\|\theta\| : \theta \in \Theta, \theta_I = 0, \forall I \in \mathcal{I}_n\}.$$

We have the identity

$$E_{n,0}^*(\Theta) = \Delta(n; \|\cdot\|; \Theta),$$

which is valid whenever both $\|\cdot\|$ and Θ are solid and orthosymmetric.

A lower bound for the minimax error is obtainable by combining the $n = \infty$ and the $\delta = 0$ extremes:

$$E_{n,\delta}^*(\Theta) \geq \max(\Omega(\delta)/2, \Delta(n)). \quad (19)$$

Again, soft-thresholding comes surprisingly close, under surprisingly general conditions. Consider the rule

$$\hat{\theta}^{n,\delta} = \begin{cases} \eta_\delta(x_I), & I \in \mathcal{I}_n, \\ 0, & I \in \mathcal{I} \setminus \mathcal{I}_n \end{cases} . \quad (20)$$

Supposing for the moment that the loss measure $\|\cdot\|$ is convex we have

$$\|\hat{\theta}^{n,\delta} - \theta\| \leq \Omega(2\delta) + \Delta(n), \quad \theta \in \Theta. \quad (21)$$

[If the loss is not convex, but just ρ -convex, $0 < \rho < 1$, we can replace the right hand side by $(\Omega(2\delta)^\rho + \Delta(n)^\rho)^{1/\rho}$.]

Comparing (21) and (19), we again have that soft-thresholding is nearly minimax, simultaneously over a wide range of a-priori classes and choices of loss.

12.4 Evaluation of the modulus of continuity

To go farther, we specialize our choice of possible losses $\|\cdot\|$ and *a priori* classes Θ to members of the Besov and Triebel scales of sequence spaces. and calculate moduli of continuity and tail n -widths.

These spaces are defined as follows. First, we specify that the abstract index set \mathcal{I} is of the standard multiresolution format $I = (j, k)$ where $j \geq -1$ is a resolution index, and $0 \leq k < 2^j$, is a spatial index. We write equally (θ_I) or $(\theta_{j,k})$, and we write $\mathcal{I}^{(j)}$ for the collection of indices $I = (j, k)$ with $0 \leq k < 2^j$. We define the Besov sequence norm

$$\|\theta\|_{\mathbf{B}_{p,q}^\sigma}^q = \sum_{j \geq -1} (2^{js} (\sum_{I \in \mathcal{I}^{(j)}} |\theta_I|^p)^{1/p})^q \quad (22)$$

where $s \equiv \sigma + 1/2 - 1/p$, and the Besov body

$$\Theta_{p,q}^\sigma(C) \equiv \{\theta : \|\theta\|_{\mathbf{B}_{p,q}^\sigma} \leq C\}.$$

Similarly, the Triebel body $\Phi_{p,q}^\sigma = \Phi_{p,q}^\sigma(C)$ is defined by

$$\|\theta\|_{\mathbf{F}_{p,q}^\sigma} \leq C,$$

where $\mathbf{F}_{p,q}^\sigma$ refers to the norm

$$\|\theta\|_{\mathbf{F}_{p,q}^\sigma} = \|(\sum_{I \in \mathcal{I}} 2^{jsq} |\theta_I|^q \chi_I)^{1/q}\|_{L^p[0,1]}, \quad (23)$$

χ_I stands for the indicator function $1_{[k/2^j, (k+1)/2^j]}$, and $s \equiv \sigma + 1/2$. We remark, as an aside, that Besov and Triebel norms are ρ -convex, with $\rho = \min(1, p, q)$, so that in the usual range $p, q \geq 1$ they are convex.

These sequence norms are solid and orthosymmetric, the parameters (σ, p, q) allow various ways of measuring smoothness and spatial inhomogeneity, and they correspond to function space norms of scientific relevance (references in Appendix).

Theorem 3 (Besov Modulus) *Let $\|\cdot\|$ be a Besov norm with parameter (σ', p', q') (cf. (22)). Let Θ be a Besov body $\Theta_{p,q}^\sigma(C)$, and suppose that $\sigma > \sigma' + (1/p - 1/p')_+$. Then for $0 < \delta < \delta_1(C)$,*

$$\Omega(\delta, C) \asymp \begin{cases} C^{1-r}\delta^r & \mathbf{p} \in \mathcal{R} \cup \mathcal{L} \\ C^{1-r}\delta^r \log(C/\delta)^{e_C} & \mathbf{p} \in \mathcal{C} \end{cases} \quad (24)$$

where the rate exponent r is given in (3,4), $e_C = (1/q' - (1-r)/q)_+$, and the constants of equivalence $c_i = c_i(\mathbf{p})$.

[Here $A(\eta) \asymp B(\eta)$ means that there exist constants c_i such that $0 < c_1 \leq A(\eta)/B(\eta) \leq c_2 < \infty$ for all η .]

Here is the plan for the proof of Theorem 3—We first consider an optimal recovery problem corresponding to a single resolution level (Lemma 4). Using a modified modulus Ω^o defined below this is applied to the regular and logarithmic cases. The critical case is deferred to the Appendix.

Definition 10 *$W(\delta, C; p', p, n)$ is defined as the value of the n -dimensional constrained optimization problem*

$$\sup \|\xi\|_{p'} \quad \text{s.t.} \quad \xi \in R^n, \quad \|\xi\|_p \leq C, \quad \|\xi\|_\infty \leq \delta. \quad (25)$$

A vector ξ which satisfies the indicated constraints is called feasible for $W(\delta, C; p', p, n)$.

Remark This quantity describes the value of a certain optimal recovery problem. Let $\Theta_{n,p}(C)$ denote the n -dimensional ℓ^p ball of radius C ; then $W(\delta, C; p', p, n) = \Omega^o(\delta; \|\cdot\|_{p'}, \Theta_{n,p}(C))$. Our approach to Theorems 3 and 8 will be to reduce all calculations to calculations for $W(\delta, C; p', p, n)$ and hence to calculations for ℓ^p balls. In some sense the idea is that Besov bodies are built up out of ℓ^p balls.

Lemma 4 *We have $W \leq W^*$, where we define*

$$W^*(\delta, C; p', p, n) = \begin{cases} \min(\delta n^{1/p'}, C n^{1/p'-1/p}), & 0 < p' \leq p \leq \infty; \\ \min(\delta n^{1/p'}, \delta^{1-p/p'} C^{p/p'}, C), & 0 < p \leq p' \leq \infty. \end{cases} \quad (26)$$

In the first case $W = W^*$, and moreover even the second case is a near-equality. In fact in both cases of (26) there are an integer n_0 and a positive number δ_0 obeying

$$1 \leq n_0 \leq n, \quad 0 < \delta_0 \leq \delta$$

so that the vector ξ defined by

$$\xi_1 = \xi_2 = \dots = \xi_{n_0} = \delta_0; \quad \xi_{n_0+1} = \dots = \xi_n = 0$$

is feasible for $W(\delta, C; p', p, n)$ and satisfies $\|\xi\|_{p'} = \delta_0 n_0^{1/p'}$, and

$$\delta_0 n_0^{1/p'} \leq W(\delta, C; p', p, n) \leq \delta_0 (n_0 + 1)^{1/p'}. \quad (27)$$

Moreover, if $0 < p' \leq p \leq \infty$, we have

$$n_0 = n, \quad \delta_0 = \min(\delta, Cn^{-1/p}),$$

and there is exact equality $\delta_0 n_0^{1/p'} = W(\delta, C; p', p, n)$ and $W = W^*$. On the other hand, if $0 < p \leq p' \leq \infty$ then

$$n_0 = \min(n, \max(1, \lfloor (C/\delta)^p \rfloor)), \quad \text{and } \delta_0 = \min(\delta, C), \quad (28)$$

and W^* also satisfies (27). In both cases,

$$W(\delta, C) \leq \delta^{1-p/p'} C^{p/p'}. \quad (29)$$

Thus, if $p' \leq p$, a 'least favorable' feasible vector is *dense*, lying along the diagonal $c(1, \dots, 1)$, with the extremal value of c determined by the more restrictive of the ℓ_∞ and ℓ_p constraints. On the other hand, if $p' \geq p$, the (near) least favorable vectors may be *sparse*, with the degree of sparsity determined by C/δ . Beginning with the dense case with all non-zero coordinates when $\delta < Cn^{-1/p}$, the sparsity increases with δ through to the extreme case, having a single non-zero value when $\delta > C$. Geometrically, picture an expanding cube of side 2δ , at first wholly contained within the ℓ_p ball of radius C , then puncturing it and finally wholly containing it. We omit the formal proof, which amounts to applying standard inequalities (upper bounds) and verifying the stated results (lower bounds).

We now define a modified modulus of continuity which is more convenient for calculations involving Besov and Triebel norm balls.

$$\Omega^\circ(\delta; \|\cdot\|, \Theta) = \sup\{\|\theta\| : \theta \in \Theta, |\theta|_I \leq \delta \forall I \in \mathcal{I}\}.$$

Assuming that $0 \in \Theta$, that $\Theta = \Theta(C) = \{\theta : \|\theta\|_\Theta \leq C\}$, and that $\|\cdot\|_\Theta$ is ρ -convex, then it follows easily that

$$\Omega^\circ(\delta) \leq \Omega(\delta) \leq 2^{1/\rho} \Omega^\circ(2^{-1/\rho} \delta). \quad (30)$$

We sometimes write $\Omega^\circ(\delta, C)$ to show explicitly the dependence on C .

We now apply this result to Theorem 3. We assume that we are not in the critical case (which itself is treated in the Appendix.) We will use the

following notational device. If $\theta = (\theta_I)_{I \in \mathcal{I}}$ then $\theta^{(j)}$ is the same vector with coordinates set to zero which are not at resolution level j :

$$\theta_I^{(j)} = \begin{cases} \theta_I & I \in \mathcal{I}^{(j)} \\ 0 & I \notin \mathcal{I}^{(j)} \end{cases}.$$

We define $\Omega_j \equiv \Omega_j(\delta, C; \mathbf{p})$ by

$$\Omega_j \equiv \sup\{\|\theta^{(j)}\|_{\mathbf{B}_{p',q'}^\sigma} : \|\theta^{(j)}\|_{\mathbf{B}_{p,q}^\sigma} \leq C, \|\theta^{(j)}\|_\infty \leq \delta\}.$$

Then, using the definition of Ω and of the Besov norms along with (30)

$$\|(\Omega_j)_j\|_{\ell^\infty} \leq \Omega \leq 2^{1/p} \|(\Omega_j)_j\|_{\ell^{q'}}.$$

Now applying the definitions,

$$\Omega_j = 2^{js'} W(\delta, C2^{-js}; p', p, 2^j) = 2^{js'} W_j(\delta, C2^{-js}),$$

say, with $W_j^*(\delta, C)$ defined similarly. Here is the key observation. Define a function of a real variable j :

$$\Omega^*(j) = 2^{js'} W^*(\delta, C2^{-js}; p', p, 2^j),$$

Then, as soon as $\delta < C$,

$$\sup_{j \in R} \Omega^*(j) = \delta^r C^{1-r},$$

as may be verified by direct calculation in each of the cases concerned. Let j^* be the point of maximum in this expression. Using the formulas for $W_j^*(\delta, C2^{-js})$, we can verify that, because we are not in the critical case, $s'p' \neq sp$, and

$$2^{-\eta_0 |j - j^*|} \leq \Omega^*(j)/\Omega^*(j^*) \leq 2^{-\eta_1 |j - j^*|} \quad (31)$$

with exponents $\eta_i = \eta_i(\mathbf{p}) > 0$. We can also verify that for $\delta < C$, $j^* > 1$. In the regular case, $2^{j^*} = (C/\delta)^{1/(s+1/p)}$ and we choose $j_0 = \lfloor j^* \rfloor$ so that (in the notation of Lemma 4) $n_{j_0} = n$: we call this the 'dense' case in Theorem 8 below. In the logarithmic case, $2^{j^*} = (C/\delta)^{1/s}$ and we choose $j_0 = \lceil j^* \rceil$, so that $n_{j_0} = 1$: this is called the 'sparse' case below. In each case, $|j_0 - j^*| < 1$, and from (27)

$$(1 + 1/n_{j_0})^{1/p'} \cdot \Omega_{j_0} \geq \Omega^*(j_0) \geq 2^{-\eta_0} \delta^r C^{1-r};$$

on the other hand, using the formulas for $W_j^*(\delta, C2^{-js})$,

$$\Omega_j \leq \Omega^*(j) \leq 2^{-\eta_1 (|j - j_0| - 1)} \cdot \delta^r C^{1-r}.$$

Because (28) guarantees $n_{j_0} \geq 1$, it follows that

$$c_0 \delta^r C^{1-r} \leq \Omega_{j_0} \leq \Omega \leq 2^{1/\rho} \|(\Omega_j)_j\|_{q'} \leq c_1 \delta^r C^{1-r}. \square$$

What if $\|\cdot\|$ or Θ , or both, come from the Triebel Scales? A norm from the Triebel scale is bracketed by norms from the Besov scales with the same σ and p , but different q 's:

$$a_0 \|\theta\|_{B_{p,\max(p,q)}^\sigma} \leq \|\theta\|_{F_{p,q}^\sigma} \leq a_1 \|\theta\|_{B_{p,\min(p,q)}^\sigma} \quad (32)$$

(compare Peetre (1975, page 261) or Triebel (1992, page 96)). Hence, for example,

$$\Theta_{p,\min(p,q)}^\sigma(C/a_1) \subset \Phi_{p,q}^\sigma(C) \subset \Theta_{p,\max(p,q)}^\sigma(C/a_0),$$

and so we can bracket the modulus of continuity in terms of the modulus from the Besov case, but with differing values of q, q' . By (24), the qualitative behavior for the modulus in the Besov scale, outside the critical case $(\sigma + 1/2)p = (\sigma' + 1/2)p'$, $p' > p$, does not depend on q, q' . Hence, the modulus of continuity continues to obey the same general relations (24) even when the Triebel scale is used for one, or both, of the norm $\|\cdot\|$ and class Θ .

In the critical case, we can at least get bounds; for example in the Triebel norm, Triebel body case, combining (24) with (32) gives, for $0 < \delta < \delta_1(C)$

$$c_0 \cdot C^{(1-r)} \delta^r \log(C/\delta)^{e_2^-} \leq \Omega(\delta) \leq c_1 \cdot C^{(1-r)} \delta^r \log(C/\delta)^{e_2^+} \quad (33)$$

with $e_2^+ = (1/\min(q', p') - (1-r)/\max(p, q))_+$ and $e_2^- = (1/\max(p', q') - (1-r)/\min(p, q))_+$.

The next result shows that the Triebel norms can lead to genuinely different results than the Besov: it would apply, for example, to certain L_p and L_p- Sobolev loss functions when $p \neq 2$.

Theorem 5 (Triebel modulus, critical case) Let $\|\cdot\|$ be a member of the Triebel scale and Θ a Besov body $\Theta_{p,q}^\sigma(C)$. Suppose that $\sigma > \sigma' + (1/p - 1/p')$ and that we are in the critical case $p' \geq p$, $(\sigma + 1/2)p = (\sigma' + 1/2)p'$. Then

$$\Omega(\delta, C) \geq c_0 C^{1-r} \delta^r (\log C/\delta)^{(1/p' - (1-r)/q)_+} \quad \delta < \delta_0(C) \quad (34)$$

and if $q' \geq p$, the right side (with a different constant c_1) is also an upper bound for $\Omega(\delta)$.

Remark This result is an improvement on the lower bound of (33) when $p' < q'$ and an improvement on the upper bound when $q' < p'$ (and $q' \geq p$).

Proof: We establish the lower bound here, and defer the upper bound to the Appendix. Write

$$\|\theta\|_f^{p'} = \|\theta\|_{f_{p',q'}^{\sigma'}}^{p'} = \int_0^1 \left(\sum_{jk} |\theta_{jk}|^{q'} 2^{s'q'j} \chi_{j,k} \right)^{p'/q'} dt = \int_0^1 f_\theta^{p'/q'}(t),$$

where we identify $\chi_{j,k}$ with χ_I and (j,k) , $0 \leq k < 2^j$ with I . We take the choice of parameters used in the Besov modulus lower bound (critical case), and attempt to maximise $\|\theta\|_f$ by “stacking” up the coefficients θ_I . Thus, let j_a, j_b be defined as above, and set

$$\theta_{jk} = \begin{cases} \delta & 1 \leq k \leq n_j, \quad j_a \leq j < j_b \\ 0 & \text{otherwise.} \end{cases}$$

where $n_j = [m_j]$, $m_j = (\bar{c}\delta^{-1}2^{-js})^p$ and $\bar{c} = C(j_b - j_a)^{-1/q}$. By construction, $\theta \in \Theta_{p,q}^\sigma(C)$ and $\|\theta\|_\infty = \delta$. The sequence $\alpha_j = n_j 2^{-j}$ is decreasing, and so

$$\begin{aligned} \|\theta\|_f^{p'} &= \delta^{p'} \int_0^1 \left(\sum_j 2^{s'q'j} I\{t \leq \alpha_j\} \right)^{p'/q'} dt \\ &= \delta^{p'} \sum_j \int_{\alpha_{j+1}}^{\alpha_j} \left(\sum_{j \leq j} 2^{s'q'j} \right)^{p'/q'} dt \\ &\geq \delta^{p'} \sum_{j_a}^{j_b} 2^{s'p'j} (\alpha_j - \alpha_{j+1}). \end{aligned}$$

Since $C\delta^{-1}2^{-sj+} \asymp 1$, it is easily checked that $n_j \gg 1$ for $j \leq j_b$ and that $\alpha_{j+1} < \alpha_j/2$. Consequently

$$\|\theta\|_f^{p'} \geq \frac{1}{2} \delta^{p'} \sum_{j_a}^{j_b} 2^{(s'p'-1)j} n_j \geq \frac{1}{2} \delta^{p'-p} C^p (j_b - j_a)^{1-p/q}, \quad (35)$$

since, in the critical case $s'p' = (\sigma' + 1/2)p' = (\sigma + 1/2)p = sp + 1$, and so $2^{(s'p'-1)j} m_j = \bar{c}^p \delta^{-p}$. Hence Ω^0 and hence Ω satisfy the claimed (34). \square

In addition to concrete information about the modulus, we need concrete information about the tail- n -widths.

Theorem 6 Let $\|\cdot\|$ be a member of the Besov or Triebel scales, with parameter (σ', p', q') . Let Θ be a Besov body $\Theta_{p,q}^\sigma(C)$ or a Triebel Body $\Phi_{p,q}^\sigma(C)$. Suppose $\eta = \sigma - \sigma' - (1/p - 1/p')_+ > 0$. Then

$$\Delta(n; \|\cdot\|, \Theta) \asymp n^{-\eta} \quad n = 2^{J+1},$$

with constants $c_i = c_i(\mathbf{p})$.

Definition 11 $D(C; p', p, n)$ is the value of the n -dimensional constrained optimization problem

$$\sup \|\xi\|_{p'} \quad s.t. \quad \xi \in R^n, \quad \|\xi\|_p \leq C. \quad (36)$$

A vector ξ which satisfies the indicated constraints is called feasible for $D(C; p', p, n)$.

Since $D(C; p', p, n) = W(\infty, C; p', p, n)$, we have immediately upper bounds from Lemma 4. More careful treatment gives the exact formula

$$D(C; p', p, n) = Cn^{(1/p' - 1/p)_+}. \quad (37)$$

Proof of Theorem 6: We consider the case where both loss and a priori class come from the Besov scale. Other cases may be treated using (32) and the observation that the exponent in (38) does not depend on (q, q') . Define

$$\Delta_j = \Delta_j(C; \mathbf{p}) = \sup \{\|\theta^{(j)}\|_{\mathbf{b}_{p', q'}^{\sigma'}} : \|\theta^{(j)}\|_{\mathbf{b}_{p, q}^\sigma} \leq C\}$$

we note that

$$\Delta_{J+1} \leq \Delta(n) \leq \|(\Delta_j)_{j \geq J+1}\|_{q'}.$$

Now comparing definitions and then formula (37), we have

$$\Delta_j = 2^{js'} D(C 2^{-js}; p', p, 2^j) = C 2^{-j\eta}, \quad j \geq 0.$$

Consequently

$$\|(\Delta_j)_{j \geq J+1}\|_{q'} \leq \Delta_{J+1} \cdot (\sum_{h \geq 0} 2^{-h\eta})^{1/q}, \quad \eta = \eta(\mathbf{p}).$$

Combining these results, we have

$$\Delta(n) \asymp 2^{-(J+1)\eta}, \quad n = 2^{J+1} \rightarrow \infty. \square \quad (38)$$

12.5 Statistical Sequence Model: Upper Bounds

We now translate the results on optimal recovery into results on statistical estimation.

The basic idea is the following fact (Leadbetter, Lindgren & Rootzen 1983): Let (z_I) be i.i.d. $N(0, 1)$. Define

$$A_n = \left\{ \|(z_I)\|_{\ell_n^\infty} \leq \sqrt{2 \log n} \right\},$$

then

$$\pi_n \equiv \text{Prob}\{A_n\} \rightarrow 1, \quad n \rightarrow \infty. \quad (39)$$

In words, we have very high confidence that $\|(z_I)_I\|_{\ell_n^\infty} \leq \sqrt{2 \log(n)}$. This motivates us to act as if noisy data (6) were an instance of the deterministic model (18), with noise level $\delta_n = \sqrt{2 \log n} \cdot \epsilon$. Accordingly, we set $t_n = \delta_n$, and define

$$\hat{\theta}_I^{(n)} = \begin{cases} \eta_{t_n}(y_I), & I \in \mathcal{I}_n, \\ 0, & I \in \mathcal{I} \setminus \mathcal{I}_n \end{cases} \quad (40)$$

Recall the optimal recovery bound (21) (case where triangle inequality applies). We get immediately that whenever $\theta \in \Theta$ and the event A_n holds,

$$\|\hat{\theta}^{(n)} - \theta\| \leq \Omega(2\delta_n) + \Delta(n);$$

as this event has probability π_n we obtain the risk bound

Theorem 7 *If $\|\cdot\|$ is convex then for all $\theta \in \Theta$,*

$$P\{\|\hat{\theta}^{(n)} - \theta\| \leq \Omega(2\delta_n) + \Delta(n)\} \geq \pi_n; \quad (41)$$

with a suitable modification if $\|\cdot\|$ is ρ -convex, $0 < \rho < 1$.

This shows that statistical estimation is not really harder than optimal recovery, except by a factor involving $\sqrt{\log(n)}$.

12.6 Statistical Sequence Model: Lower Bounds

With noise levels equated, $\epsilon = \delta$, statistical estimation is not easier than optimal recovery:

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| \geq \max(\Delta(n), c\Omega(\epsilon))\} \rightarrow 1, \quad \epsilon = \sigma/\sqrt{n} \rightarrow 0. \quad (42)$$

Half of this result is nonstatistical; it says that

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| \geq \Delta(n)\} \rightarrow 1 \quad (43)$$

and this follows for the reason that (from section 12.3.3) this holds in the noiseless case. The other half is statistical, and requires a generalization of lower bounds developed by decision theorists systematically over the last 15 years—namely the embedding of an appropriate hypercube in the class Θ and using elementary decision-theoretic arguments on hypercubes. Compare (Samarov 1992, Bretagnolle & Huber 1979, Ibragimov & Khas'minskii 1982, Stone 1982).

Theorem 8 *Let $\|\cdot\|$ come from the Besov or Triebel scale, with parameter (σ', p', q') . Let Θ be a Besov body $\Theta_{p,q}^\sigma(C)$. Then with a $c = c(\mathbf{p})$*

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| \geq c\Omega(\epsilon, C)\} \rightarrow 1. \quad (44)$$

Moreover, when $p' > p$ and $(\sigma + 1/2)p \leq (\sigma' + 1/2)p'$, (and, in the critical Triebel case, under the extra hypothesis that $q' \geq p$) we get the even stronger bound

$$\inf_{\hat{\theta}} \sup_{\Theta} P\{\|\hat{\theta} - \theta\| \geq c\Omega(\epsilon\sqrt{\log(\epsilon^{-1})}, C)\} \rightarrow 1. \quad (45)$$

The proof of Theorem 3 constructed a special problem of optimal recovery: recovering a parameter θ known to lie in a certain 2^{j_0} -dimensional ℓ^p ball ($j_0 = j_0(\mathbf{p})$), measuring loss in $\ell^{p'}$ -norm. The construction shows that this finite-dimensional subproblem is essentially as hard (under model (9)) as the full infinite-dimensional problem of optimal recovery of an object in an σ, p, q -ball with an σ', p', q' -loss. The proof of Theorem 8 shows that, under the calibration $\epsilon = \delta$, the statistical estimation problem over this particular ℓ^p ball is at least as hard as the optimal recovery problem, and sometimes harder by an additional logarithmic factor.

12.6.1 PROOF OF THEOREM 8

The proof of Theorem 3 identifies a quantity Ω_{j_0} , which may be called the difficulty of that single-level subproblem for which the optimal recovery problem is hardest. In turn, that subproblem, via Lemma 4, involves the estimation of a 2^{j_0} -dimensional parameter, of which $n_0(j_0)$ elements are nonzero a priori. The present proof operates by studying this particular subproblem and showing that it would be even harder when viewed in the statistical estimation model.

We study several cases, depending upon whether this least favorable subproblem represents a “dense”, “sparse”, or “transitional” case. The phrases “dense”, “sparse”, etc. refer to whether $n_0 \asymp 2^{j_0}$, $n_0 = 1$, or $n_0 \asymp 2^{j_0(1-a)}$, $0 < a < 1$. In view of (32), we may confine attention to Besov norms $\|\cdot\|$, except in the critical Case III (which will again be deferred to the Appendix).

Case I: The least-favorable ball is “dense”: $\mathbf{p} \in \mathcal{R}$

We describe a relation between the minimax risk over ℓ^p balls and the quantity $W(\delta, C)$. We have observations

$$v_i = \xi + \delta \cdot z_i, \quad i = 1, \dots, n \quad (46)$$

where z_i are i.i.d. $N(0, 1)$ and we wish to estimate ξ . We know that $\xi \in \Theta_{n,p}(C)$. Because of the optimal recovery interpretation of $W(\delta, C)$, the following bound on the minimax risk says that this statistical estimation model is not essentially easier than the optimal recovery model.

Lemma 9 Let $\pi_0 = \Phi(-1)/2 \approx .08$. Let $n_0 = n_0(\delta, C; p', p, n)$ be as in Lemma 4. Then

$$\inf_{\xi} \sup_{\Theta_{n,p}(C)} P\{\|\hat{\xi} - \xi\|_{p'} \geq \frac{1}{2}\pi_0^{1/p'} W(\delta, C; p', p, n)\} \geq 1 - e^{-2n_0\pi_0^2}. \quad (47)$$

Proof: Let n_0 and δ_0 be as in Lemma 4. Let the s_i be random signs, equally likely to take the values ± 1 independently of each other and of the (z_i) . Define the random vector $\xi \in R^n$ via

$$\xi_i = \begin{cases} s_i \delta_0 & 1 \leq i \leq n_0, \\ 0 & i > n_0 \end{cases}.$$

Note that $\xi \in \Theta_{n,p}(C)$ with probability 1. Here and later, P_μ denotes the joint distribution of (ξ, v) under the prior.

Because a sign error in a certain coordinate implies an estimation error of size δ_0 in that coordinate, for any estimator $\hat{\xi}$ we have

$$\|\hat{\xi} - \xi\|_{p'}^{p'} \geq \delta_0^{p'} N(\hat{\xi}, \xi) \quad (48)$$

where we set

$$N(\hat{\xi}, \xi) = \sum_{i=1}^{n_0} I\{\hat{\xi}_i \xi_i < 0\} = \sum_{i=1}^{n_0} V_i(v, \xi_i),$$

say. Let $\hat{\xi}_i^*$ be the Bayes rule for loss function $I\{\hat{\xi}_i \xi_i < 0\}$, and let V_i^* be defined as for V_i . Then

$$P_\mu(V_i^* = 1|v) \leq P_\mu(V_i = 1|v).$$

Conditional on v , the variables $\{V_i\}$ are independent, as are $\{V_i^*\}$, and so it follows that $N(\hat{\xi}^*, \xi)$ is stochastically smaller than $N(\hat{\xi}, \xi)$. Hence

$$\inf_{\hat{\xi}} P_\mu\{\|\hat{\xi} - \xi\|_{p'}^{p'} \geq \delta_0^{p'} a\} \geq P_\mu\{N(\hat{\xi}^*, \xi) \geq a\}.$$

From the structure of the prior and the normal translation family, it is evident that the Bayes rule $\hat{\xi}_i^*(v) = \text{sgn}(v_i)\delta_0$, so that

$$N(\hat{\xi}^*, \xi) = \#\{i : v_i \xi_i < 0\} \sim \text{Bin}(n_0, \pi)$$

where

$$\pi = P_\mu\{v_i \xi_i < 0\} = P\{\delta \cdot z_i < -\delta_0\} \geq \Phi(-1) = 2 \cdot \pi_0.$$

Choose $c = n_0 \pi_0$ and note that by the Cramér-Chernoff large deviations principle, the number of sign errors is highly likely to exceed $\pi_0 n_0$:

$$P_\mu\{N(\hat{\xi}^*, \xi) < \pi_0 n_0\} \leq e^{-n_0 H(\pi_0, 2\pi_0)},$$

where $H(\pi, \pi') = \pi \log(\pi/\pi') + (1-\pi) \log((1-\pi)/(1-\pi'))$. As $H(\pi, \pi') \geq 2(\pi - \pi')^2$, we get

$$P_\mu\{N(\hat{\xi}^*, \xi) \geq \pi_0 n_0\} \geq 1 - e^{-2n_0 \pi_0^2}. \quad (49)$$

Hence (48) implies the bound

$$\inf_{\hat{\xi}} P_\mu \{ \|\hat{\xi} - \xi\|_{p'} \geq \delta_0 (\pi_0 n_0)^{1/p'} \} \geq 1 - e^{-2n_0 \pi_0^2}.$$

Recalling that n_0 and δ_0 satisfy

$$W(\delta, C; p', p, n) \leq \delta_0 (n_0 + 1)^{1/p'}$$

and noting that $\sup_{\Theta} P(A) \geq P_\mu(A)$ for any A gives the stated bound on the minimax risk (47). \square

This lemma allows us to prove the dense case of Theorem 8 by choosing the n -dimensional ℓ^p balls optimally. Using now the notation introduced in the proof of Theorem 3, there is $c_0 > 0$ so that for $\epsilon < \delta_1(C, c_0)$ we can find j_0 giving

$$\Omega_{j_0}(\epsilon, C) > c_0 \cdot \Omega(\epsilon, C). \quad (50)$$

Let $\Theta^{(j_0)}(C)$ be the collection of all sequences $\theta^{(j_0)}$ whose coordinates vanish away from level j_0 and which satisfy

$$\|\theta^{(j_0)}\|_{\mathbf{B}_{p,q}^\sigma} \leq C.$$

For θ in $\Theta^{(j_0)}(C)$, we have

$$\|\theta\|_{\mathbf{B}_{p,q}^\sigma} = 2^{j_0 s} \|\theta\|_p;$$

geometrically, $\Theta^{(j_0)}(C)$ is a 2^{j_0} -dimensional ℓ^p -ball inscribed in $\Theta_{p,q}^\sigma(C)$. Moreover, for θ, θ' in $\Theta^{(j_0)}(C)$,

$$\|\theta - \theta'\|_{\mathbf{B}_{p',q'}^{\sigma'}} = 2^{j_0 s'} \|\theta - \theta'\|_{p'}$$

hence, applying Lemma 9, and appropriate reductions by sufficiency, we have that, under the observations model (6), the problem of estimating $\theta \in \Theta^{(j_0)}(C)$ is no easier than the problem of estimating $\xi \in \Theta_{n,p}(2^{-j_0 s} C)$ from observations (46), with noise level $\delta = \epsilon$, and with an $\ell^{p'}$ loss scaled by $2^{j_0 s'}$. Hence, (47) gives

$$\inf_{\hat{\theta}} \sup_{\Theta^{(j_0)}} P \{ \|\hat{\theta} - \theta\|_{\mathbf{B}_{p',q'}^{\sigma'}} \geq \frac{1}{2} 2^{j_0 s'} \pi_0^{1/p'} W_{j_0}(\epsilon, C 2^{-j_0 s}) \} \geq 1 - e^{-2n_0 \pi_0^2}$$

Now

$$\Omega_{j_0} = 2^{j_0 s'} W_{j_0}(\epsilon, C 2^{-j_0 s})$$

so from (50)

$$\inf_{\hat{\theta}} \sup_{\Theta_{p,q}^\sigma(C)} P \{ \|\hat{\theta} - \theta\|_{\mathbf{B}_{p',q'}^{\sigma'}} \geq c_0 \cdot \pi_0^{1/p'} \cdot (1 + 1/n_0)^{-1/p'} \cdot \Omega(\epsilon, C) \} \geq 1 - e^{-2n_0 \pi_0^2}$$

In the regular case, $n_0 = 2^{j_0} \rightarrow \infty$ as $\epsilon \rightarrow 0$, so that setting $c = c_0 \cdot \pi_0 \cdot (1 - \gamma)$, $\gamma > 0$, we get (44).

Case II: The least-favorable ball is “sparse”: $\mathbf{p} \in \mathcal{L}$

Our lower bound for statistical estimation follows from a special needle-in-a-haystack problem. Suppose that we have observations (46), but all the ξ_i are zero, with the exception of at most one; and that one satisfies $|\xi_i| \leq \delta_0$, with δ_0 a parameter. Let $\Theta_{n,0}(1, \delta_0)$ denote the collection of all such sequences. The following result says that we cannot estimate ξ with an error essentially smaller than δ_0 , provided δ_0 is not too large. In the sparse case, we have $n_0 = 1$ and so this bound implies that statistical estimation is not easier than optimal recovery.

Lemma 10 *With $\eta \in (0, 2)$, let $\delta_0 < \sqrt{(2 - \eta) \log n} \cdot \delta$ for all n*

$$\inf_{\hat{\xi}(v)} \sup_{\Theta_{n,0}(1, \delta_0)} P\{\|\hat{\xi}(v) - \xi\|_{p'} \geq \delta_0/3\} \rightarrow 1. \quad (51)$$

Proof: We only sketch the argument. Let $P_{n,\delta}$ denote the measure which places a nonzero element at one of the n sites, I say, uniformly at random, with a random sign. Let $\gamma = \delta_0/\delta$. By a calculation,

$$\begin{aligned} \frac{dP_{n,\delta}}{dP_{n,0}}(v) &= e^{-\gamma^2/2} n^{-1} \sum_i \cosh(\delta_0 v_i / \delta^2), \\ &= e^{-\gamma^2/2} n^{-1} \sum_i \cosh(\gamma z_i) + \theta n^{-1} e^{\gamma^2/2} e^{\gamma|z_I|}, \quad |\theta| < 1, \end{aligned} \quad (52)$$

where the constant $\theta = \theta_n(\gamma) \leq 1$. The first term converges a.s. to 1. Since $n^{-1} e^{-\gamma^2/2} < n^{-\eta/2}$, the remainder term obeys the probabilistic bound

$$P\{e^{\gamma|z_I|} > \epsilon n^{\eta/2}\} \leq P\{\sqrt{2 \log(n)} |z_I| > \log(n) \eta/2 + \log(\epsilon)\} \rightarrow 0.$$

Consequently

$$P_{n,\delta}\{|1 - \frac{dP_{n,\delta}}{dP_{n,0}}(v)| > \epsilon\} \rightarrow 0.$$

Consequently, any rule has essentially the same operating characteristics under $P_{n,\delta}$ as under $P_{n,0}$ and must therefore make, with overwhelming probability an error of size $\geq \delta_0/3$ in estimating ξ . \square

To apply this, we argue as follows. Let $\eta \in (0, 2)$ and let j_0 be the largest integer satisfying

$$\sqrt{(2 - \eta) \log_2(2^j)} \cdot \epsilon \cdot 2^{js} \leq C$$

so that roughly $j_0 \sim s^{-1} \log_2(C/\epsilon) + O(\log(\log(C/\epsilon)))$, and set $\delta_{j_0}^2 = (2 - \eta) \log_2(2^{j_0}) \cdot \epsilon^2$. Then for some $a > 0$,

$$\delta_{j_0} \geq a \cdot C \cdot 2^{-j_0 s} \quad \delta < \delta_1(C, a). \quad (54)$$

Now, define the random variable $\theta^{(j_0)}$ vanishing away from level j_0 : $\theta_I^{(j_0)} = 0$, $I \notin \mathcal{I}^{(j_0)}$; and having one nonzero element at level j_0 , of size δ_{j_0} and random polarity. Then, from the previous lemma we have

$$\inf_{\hat{\theta}} P\{\|\hat{\theta}^{(j_0)} - \theta^{(j_0)}\|_{p'} \geq \delta_{j_0}/3\} \rightarrow 1$$

as $\delta \rightarrow 0$. Since $\theta^{(j_0)} \in \Theta^{(j_0)}(C)$ by the choice of j_0 , we also have

$$\inf_{\hat{\theta}} \sup_{\Theta_{p,q}^s} P\{\|\hat{\theta} - \theta\|_{\mathbf{b}_{p',q'}^{\sigma'}} \geq (\delta_{j_0}/3) \cdot 2^{js'}\} \rightarrow 1.$$

Using (54) gives

$$\delta_{j_0} 2^{js'} \geq aC 2^{-j_0(s-s')} = c_2 a C \left(\frac{\epsilon}{C} \sqrt{\log(C/\epsilon)}\right)^{(s-s')/s} (1 + o(1)) \quad (55)$$

$$\geq c_2 a C^{1-r} (\epsilon \sqrt{\log \epsilon^{-1}})^r (1 + o(1)), \quad (56)$$

as $\delta \rightarrow 0$, which proves the theorem in this case.

12.7 Translation into Function Space

Our conclusion from Theorems 3-8:

Corollary 6 *In the sequence model (6), the single estimator (40) is within a logarithmic factor of minimax over every loss and every a priori class chosen from the Besov and Triebel sequence scales. For a certain range of these choices the estimator is within a constant factor of minimax.*

Theorem 1 is the translation of this conclusion back from sequences to functions. Fundamental to our approach, in section 12.3.1 above, is the heuristic that observations (1) are essentially equivalent to observations (6). This contains within it three specific sub-heuristics:

1. That if we apply an empirical wavelet transform, based on pyramid filtering, to n noiseless samples, then we get the first n coefficients out of the countable sequence of all wavelet coefficients.
2. That if we apply an empirical wavelet transform, based on pyramid filtering, to n noisy samples, then we get the first n theoretical wavelet coefficients, with white noise added; this noise has standard deviation $\epsilon = \sigma/\sqrt{n}$.
3. That the Besov and Triebel norms in function space (e.g. L^p , W_p^m norms) are equivalent to the corresponding sequence space norms (e.g. $f_{p,2}^0$ and $f_{p,2}^m$).

Using these heuristics, the sequence-space model (6) may be viewed as just an equivalent representation of the model (1); hence errors in estimation of wavelet coefficients are equivalent to errors in estimation of functions, and rates of convergence in the two problems are identical, when the proper calibration $\epsilon = \sigma/\sqrt{n}$ is made.

These heuristics are just approximations, and a number of arguments are necessary to get a full result, covering all cases. We now give a detailed sketch of the connection between the nonparametric and sequence space problems.

12.7.1 EMPIRICAL WAVELET TRANSFORM

1°. In (Donoho 1992a, Donoho 1992b) it is shown how one may define a theoretical wavelet-like transform $\theta^{[n]} = W_n f$ taking a continuous function f on $[0, 1]$ into a countable sequence $\theta^{[n]}$, with two properties:

- (a) *Matching.* The theoretical transform of f gives a coefficient sequence $\theta^{[n]}$ that agrees exactly with the empirical transform $\theta^{(n)}$ of samples of f in the first n places. Here n is dyadic, and $\theta^{[n]}(f)$ depends on n .
- (b) *Norm Equivalence.* Provided $1/p < \sigma < \min(R, D)$, the Besov and Triebel sequence norms of the full sequence $\theta^{[n]}$ are equivalent to the corresponding Besov and Triebel function space norms of f , with constants of equivalence that do not depend on n , even though in general $\theta^{[n]}$ depends on n .

In detail, this last point means that if \hat{f} and f are two continuous functions with coefficient sequences $\hat{\theta}^{[n]}$ and $\theta^{[n]}$ respectively, and if $\|\theta\|$ and $|f|$ denote corresponding sequence-space and function-space norms, respectively, then there are constants B_i so that

$$B_0 \|\hat{\theta}^{[n]} - \theta^{[n]}\| \leq |\hat{f} - f| \leq B_1 \|\hat{\theta}^{[n]} - \theta^{[n]}\|; \quad (57)$$

the constants do not depend on f or n . In particular, the coefficient sequences, though different for each n , bear a stable relation to the underlying functions.

2°. The empirical wavelet transform of noisy data $(d_i)_{i=1}^n$ obeying (1) yields data

$$\tilde{y}_I = \theta_I + \epsilon \cdot \tilde{z}_I, \quad I \in \mathcal{I}_n, \quad (58)$$

with $\epsilon = \sigma/\sqrt{n}$. This form of data is of the same general form as supposed in the sequence model (6). Detailed study of the Pyramid Filtering Algorithm of Cohen, Daubechies, Jawerth & Vial (1993) reveals that all but $O(\log(n))$ of these coefficients are a standard Gaussian white noise with variance σ^2/n ; the other coefficients “feel the boundaries”, and have a slight covariance among themselves and a variance which is roughly, but not exactly, σ^2/n . Nevertheless, the analog of (39) continues to hold for this (very slightly) colored noise:

$$P\{\sup_{\mathcal{I}_n} |\tilde{z}_I| \geq \sqrt{2 \log(n)}\} \rightarrow 0. \quad (59)$$

In fact, our upper risk bound (41) depended on properties of the noise only through (39), so this is all we need in order to get risk upper bounds paralleling (41).

12.7.2 RISK UPPER BOUND

To see the implications, suppose we pick a function ball $\mathcal{F}(C)$ and a loss norm $|\cdot|$, both arising from the Besov scale, with indices σ, p, q and

σ', p', q' , respectively. Consider the corresponding objects $\Theta_{p,q}^\sigma$ and $\|\cdot\|$ in the sequence space. (57) assures that sequence space losses are equivalent to function space losses. Also, with $\Theta^{[n]}$ the set of coefficient sequences $\theta^{[n]} = \theta^{[n]}(f)$ arising from $f \in \mathcal{F}(C)$, for constants A_i , (57) yields the inclusions

$$\Theta_{p,q}^\sigma(A_0 \cdot C) \subset \Theta^{[n]} \subset \Theta_{p,q}^\sigma(A_1 \cdot C). \quad (60)$$

Now suppose we estimate f by applying the prescription (40) to the data $(\tilde{y}_I)_{I \in \mathcal{I}_n}$, producing $\hat{\theta}_n^*$. By (60), $\theta^{[n]}(f) \in \Theta_{p,q}^\sigma(A_1 \cdot C)$. By (59), the estimation error in sequence space obeys, with overwhelming probability,

$$\|\hat{\theta}_n^* - \theta^{[n]}\| \leq \Omega(2t_n) + \Delta(n),$$

where Ω is the modulus for $\|\cdot\|$ over $\Theta_{p,q}^\sigma(A_1 \cdot C)$, etc. Combining with (57) and Theorem 3 we get that with overwhelming probability, for large n ,

$$|\hat{f}_n^* - f| \leq 2B_1 \cdot \Omega \left(2 \cdot \sigma \cdot \sqrt{\frac{2 \log(n)}{n}} \right). \quad (61)$$

Completely parallel statements hold if either or both $|\cdot|$ and $\mathcal{F}(C)$ come from the Triebel scales with $\sigma' > 1/p'$.

To finish the upper risk bound, we consider the case where $|\cdot|$ comes from the Besov scale with $0 \leq \sigma' \leq 1/p' < \min(R, D)$. We remark that if $f^{(j)}$ is a function whose wavelet transform vanishes away from resolution level j and $\theta^{(j)}$ denotes the corresponding coefficient sequence, then

$$b_0 \|\theta^{(j)}\| \leq |f^{(j)}| \leq b_1 \|\theta^{(j)}\|, \quad (62)$$

with constants of equivalence independent of f and j . See Meyer (1990, page 46, Théorème 7). At the same time $|\cdot|$ is ρ -convex, $\rho = \min(1, p, q)$. Hence, if f is a function whose wavelet coefficients vanish at levels $j > J$, then

$$|f|^\rho \leq b_1^\rho \sum_{j \leq J} \|\theta^{(j)}\|^\rho.$$

This bears comparison with

$$\|\theta\| = \left(\sum_{j \leq J} \|\theta^{(j)}\|^{q'} \right)^{1/q'}. \quad (63)$$

Now from $n^{1/\rho-1/q'} \|\xi\|_{\ell_n^{q'}} \geq \|\xi\|_{\ell_n^\rho}$, valid for $q' \geq \rho$ and $\xi \in R^n$, we have

$$|f| \leq C \cdot (J+2)^{1/\rho-1/q'} \|\theta\|.$$

Applying this in place of (57) gives, instead of (61),

$$|\hat{f}_n^* - f| \leq b_1 \cdot \log(n)^{1/\rho-1/q'} \Omega \left(2 \cdot \sigma \cdot \sqrt{\frac{2 \log(n)}{n}} \right). \quad (64)$$

In the Triebel case, we use (32),

$$\|\theta\|_{\mathbf{f}_{p,q}^\sigma} \leq C\|\theta\|_{\mathbf{b}_{p,\min(p,q)}^\sigma}$$

so that we may continue from the point (63) with $\min(p', q')$ in place of q' to conclude that with overwhelming probability

$$|\hat{f}_n^* - f| \leq b_1 \cdot \log(n)^{1/\rho-1/\min(p',q')} \Omega \left(2 \cdot \sigma \cdot \sqrt{\frac{2 \log(n)}{n}} \right). \quad (65)$$

12.7.3 RISK LOWER BOUND

We remark again that the noise in the wavelet coefficients (58) is exactly a Gaussian white noise except for $O(\log(n))$ terms which “feel the boundary”. Modifying the lower bound argument (44) by avoiding those coordinates which “feel the boundary” does not change the general conclusion, only the constants in the expressions. Hence (44) is a valid lower bound for estimating the parameter vector θ from observations (1).

To translate the sequence statement into a function statement (and complete the proof of Theorem 1), we again distinguish cases.

1. In the case where the loss comes from the scale $\mathcal{C}(R, D)$, the translation follows from norm equivalence [(b) above].
2. For the case where the loss does not come from the scale $\mathcal{C}(R, D)$, and $(\sigma' + 1/2)p' \neq (\sigma + 1/2)p$, we use the single-level norm equivalence (62). Because the lower bound (44) operates by arguing only with objects $\theta^{(j_0)}$ that are nonzero at a single resolution level j_0 , this establishes the lower bound.
3. For the case where the loss does not come from the scale $\mathcal{C}(R, D)$, and $(\sigma' + 1/2)p' = (\sigma + 1/2)p$, we use a more involved argument. Owing to the regularity of the wavelets, we have, even when $\sigma' < 1/p'$, the norm inequality

$$\|\theta\|_{\mathbf{b}_{p',q'}^{\sigma'}} \leq C \left| \sum_I \theta_I \psi_I \right|_{B_{p',q'}^{\sigma'}} \quad (66)$$

even though no inequality in the opposite direction can be expected. Similar results hold in the Triebel scale. Consequently, lower bounds on the risk in sequence space offer lower bounds on the risk in function space. A careful proof of the inequality requires study of the functions ψ_I as constructed in together with arguments given there, which depend on techniques of Meyer (1990, page 50 et. seq.). Another argument would use Frazier et al. (1991) to show that $(\psi_I)_I$ is a collection of “smooth molecules”.

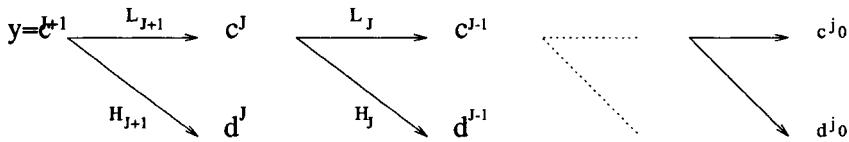


FIGURE 12.1. Cascade structure of discrete wavelet transform

12.8 Appendix

A discrete wavelet transform. For the reader's convenience, we give here a short account of a particular form of the empirical wavelet transform W_n^n used in Section 12.1. Our summary is derived from Daubechies (1992, Section 5.6), which contains a full account. The original papers by Stephane Mallat are also valuable sources (e.g. Mallat (1989)).

We consider a periodised, orthogonal, discrete transform. Our implementation of this transform (along with boundary corrected and biorthonormal versions) are available as part of a larger collection of MATLAB routines, *WaveLab*, which may be obtained over Internet from the authors via anonymous ftp from `stat.stanford.edu` in directory `/pub/software`.

The forward transform maps data y , of length $n = 2^{J+1}$ onto wavelet coefficients $w = (c^{(j_0)}, d^{(j_0)}, d^{(j_0+1)}, \dots, d^{(J)})$ as diagrammed in Figure 1. If we associate the data values y_i with positions $t_i = i/n \in [0, 1]$, the coefficients $\{w_{jk} : j = j_0, \dots, J; k = 0, \dots, 2^{j-1}\}$ may loosely be thought of as containing information in y at location $k2^{-j}$ and frequencies in an octave about 2^j .

Thus $d^{(j)}$ is a vector of 2^j ‘detail’ coefficients at resolution level j . [Note that our convention for indexing j is the reverse of that of Daubechies’!] Let $\mathbf{Z}_r = \{0, 1, \dots, r-1\}$. The operators L_j and H_j map \mathbf{Z}_{2^j} onto $\mathbf{Z}_{2^{j-1}}$ by convolution and downsampling:

$$c_k^{(j-1)} = \sum_s h_{s-2k} c_s^{(j)} \quad d_k^{(j-1)} = \sum_s g_{s-2k} c_s^{(j)}. \quad (67)$$

The summations run over \mathbf{Z}_{2^j} , and subscripts are extended periodically as necessary. The “low pass” filter (h_s) and “high pass” filter ($g_s = (-1)^s h_{1-s}$) are finite real-valued sequences subject to certain length, orthogonality and moment constraints associated with the construction of the scaling function ϕ and wavelet ψ : longer filters are required to achieve greater smoothness properties. Daubechies (1992) gives full details, along with tables of some of the celebrated filter families. In the simplest (Haar) case, $h_s \equiv 0$ except for $h_0 = h_1 = 1/\sqrt{2}$, and (67) becomes

$$c_k^{(j-1)} = (c_{2k}^{(j)} + c_{2k+1}^{(j)})/\sqrt{2}, \quad d_k^{(j-1)} = (c_{2k}^{(j)} - c_{2k+1}^{(j)})/\sqrt{2}.$$

However this choice entails no smoothness properties, and so is in practice generally replaced with longer filter sequences (h_s).

Regardless of the value of j_0 at which the cascade is stopped, the forward transform W_n^n is an orthogonal transformation. Thus the inverse wavelet transform may be implemented using the adjoint, yielding the equations

$$c_s^{(j+1)} = \sum_k h_{s-2k} c_k^{(j)} + g_{s-2k} d_k^{(j)}, \quad j_0 \leq j \leq J, \quad (68)$$

which in the Haar case reduce to

$$c_{2r}^{(j+1)} = (c_r^{(j)} + d_r^{(j)})/\sqrt{2}, \quad c_{2r+1}^{(j+1)} = (c_r^{(j)} - d_r^{(j)})/\sqrt{2}.$$

Since the filter sequences (h_s) and (g_s) appearing in (67) and (68) are of (short) finite length, the transform and its inverse involve only $O(n)$ operations.

Unconditional bases and Besov/Triebel spaces. Again, for convenience, we summarise a few definitions and consequences from the references cited in Section 12.1. A sequence $\{e_n\}$ of elements of a separable Banach space E is called a Schauder basis if for all $v \in E$, there exist unique $\beta_n \in \mathcal{C}$ such that $\sum_1^N \beta_n e_n$ converges to v in the norm of E as $N \rightarrow \infty$. A Schauder basis is called *unconditional* if there exists a constant C with the following property: for every n , sequence (β_j) and constants (α_j) with $|\alpha_j| \leq 1$,

$$\left\| \sum_1^n \alpha_j \beta_j e_j \right\| \leq C \left\| \sum_1^n \beta_j e_j \right\|. \quad (69)$$

Thus, shrinking the coefficients of any element of E relative to an unconditional basis can increase its norm by at most the factor C .

Here is one of the classical definitions of Besov spaces. We follow DeVore & Popov (1988). Let $\Delta_h^{(r)} f(t)$ denote the r -th difference $\sum_{k=0}^r \binom{r}{k} (-1)^k f(t + kh)$. The r -th modulus of smoothness of f in $L^p[0, 1]$ is

$$w_{r,p}(f; t) = \sup_{h \leq t} \left\| \Delta_h^{(r)} f \right\|_{L^p[0, 1-rh]}.$$

The *Besov* seminorm of index (σ, p, q) is defined for $r > \sigma$ by

$$|f|_{B_{p,q}^\sigma} = \left(\int_0^1 \left(\frac{w_{r,p}(f; h)}{h^\sigma} \right)^q \frac{dh}{h} \right)^{1/q}$$

if $q < \infty$, and by

$$|f|_{B_{p,\infty}^\sigma} = \sup_{0 < h < 1} \frac{w_{r,p}(f; h)}{h^\sigma}$$

if $q = \infty$. The Besov norm $\|f\|_{B_{p,q}^\sigma}$ is then defined as $\|f\|_{L^p[0,1]} + |f|_{B_{p,q}^\sigma}$.

An important consequence of the results of Lemarié and Meyer is that this norm is equivalent to the sequence norm (22). That is, given a wavelet

transform of sufficient regularity that associates to f coefficients $(\theta_I(f))$, there exist constants C_1, C_2 , not depending on f , so that

$$C_1 \|f\|_{B_{p,q}^\sigma} \leq \|\theta\|_{\mathbf{b}_{p,q}^\sigma} \leq C_2 \|f\|_{B_{p,q}^\sigma}.$$

(For the original equivalence result on \mathcal{R} see Lemarié & Meyer (1986); for a comprehensive development of the ideas see Frazier et al. (1991); for a version applying to $[0, 1]$, see Meyer (1991); and for the version adapted to the statistical application in Theorem 1, see Donoho (1992b).)

The norm equivalence means that we may work with the sequence norms (22) and (23). These are clearly solid and orthosymmetric in the sense (7) (and so the unconditional basis property (69) for the original norm $\|f\|_{B_{p,q}^\sigma}$, and all equivalent norms, follows.) Thus it is the wavelet transform that renders the pleasant properties of soft thresholding in solid, orthosymmetric norms applicable to the Besov and Triebel functions spaces of statistical and scientific interest.

Proof of Theorem 3: Now we turn to the critical case $p' > p$ and $s'p' = sp$. Let $j_-(\delta, C)$ denote the smallest integer, and $j_+(\delta, C)$ the largest integer, satisfying

$$(C/\delta)^{1/(s+1/p)} \leq 2^{j_-} \leq 2^{j_+} \leq (C/\delta)^{1/s} \quad (70)$$

evidently, $j_- \sim \log_2(C/\delta)/(s + 1/p)$ and $j_+ \sim \log_2(C/\delta)/s$. We note, from (26), that $\Omega^*(j) = \delta^r C^{(1-r)}$ for $j_- \leq j \leq j_+$, so that a unique maximizer j_* does not exist, and exponential decay (31) away from the maximizer cannot apply. On the other hand, we have that for some $\eta_1 > 0$,

$$\Omega^*(j)/\Omega^*(j_+) \leq 2^{-\eta_1(j-j_+)}, \quad j > j_+ \quad (71)$$

$$\Omega^*(j)/\Omega^*(j_-) \leq 2^{-\eta_1(j_--j)}, \quad j < j_- \quad (72)$$

which can be applied just as before, and so we focus on the zone $[j_-, j_+]$.

We now recall the fact that

$$\Omega^o \equiv \sup \left(\sum_j (2^{js'} W_j(\delta, c_j))^{q'} \right)^{1/q'} : \left(\sum_j (2^{js} c_j)^q \right)^{1/q} \leq C.$$

Let $(c_j)_j$ be any sequence satisfying $c_j = 0$, $j \notin [j_-, j_+]$ and satisfying $\sum_{j=j_-}^{j_+} (2^{js} c_j)^q \leq C^q$. Using (29), and because in the critical case $s' = s(1-r)$ and $r = 1 - p/p'$,

$$\begin{aligned} \left(\sum_{j=j_-}^{j_+} (2^{js'} W_j(\delta, c_j))^{q'} \right)^{1/q'} &\leq \left(\sum_{j=j_-}^{j_+} (2^{js'} \delta^r c_j^{1-r})^{q'} \right)^{1/q'} \\ &= \delta^r \left(\sum_{j=j_-}^{j_+} (2^{js} c_j)^{q'(1-r)} \right)^{1/q'} \\ &\leq \delta^r (j_+ - j_- + 1)^{ec} C^{1-r} \end{aligned}$$

where the last step follows from $\|x\|_{\ell_n^{q'(1-r)}} \leq \|x\|_{\ell_n^q} \cdot n^{(1/q'(1-r)-1/q)_+}$; see (37) below. Combining the three ranges $j < j_-$, $j > j_+$ and $[j_-, j_+]$

$$\Omega^0 \leq C_1 \cdot (\log_2(C/\delta))^{ec} \cdot \delta^r C^{(1-r)} + C_2 \cdot \delta^r C^{(1-r)}, \quad \delta < \delta_1(C).$$

When $q'(1-r) \geq q$, the upper bound is of order $\delta^r C^{1-r}$ as in the non-critical case, so that a lower bound for Ω^0 of the same order is obtained by considering a single level as before. On the other hand, when $q'(1-r) < q$, a lower bound combining levels j_- to j_+ is needed, so we let $(c_j^*)_j$ be the particular sequence

$$c_j^* = 2^{-js} C(j_+ - j_- + 1)^{-1/q}, \quad j_a \leq j \leq j_b;$$

where we have set $j_a = \frac{3}{4}j_- + \frac{1}{4}j_+$, $j_b = \frac{1}{4}j_- + \frac{3}{4}j_+$. For such j , it follows from (26) that $W_j^*(\delta, c_j^*) = \delta^r (c_j^*)^{1-r}$. Then, as $W \geq 2^{-1/p'} W^*$,

$$\begin{aligned} \left(\sum_{j_-}^{j_+} (2^{js'} W_j(\delta, c_j^*))^{q'} \right)^{1/q'} &\geq 2^{-1/p'} \left(\sum_{j_-}^{j_+} (2^{js'} W_j^*(\delta, c_j^*))^{q'} \right)^{1/q'} \\ &\geq c_0 \cdot (\log_2(C/\delta))^{ec} \delta^r C^{(1-r)} \quad \delta < \delta_1(C). \square \end{aligned}$$

Proof of Theorem 5: Upper Bound. 1°. On the assumption that $p' \geq q'$, the maximum of $\|\theta\|_{f'}^{p'}$ over Θ must lie among sequences $\{\theta_{jk}\}$ with $k \rightarrow |\theta_{jk}|$ decreasing in k for each j . This follows from the inequality

$$(a_0 + b_0)^\lambda + (a_1 + b_1)^\lambda \geq (a_0 + b_1)^\lambda + (a_1 + b_0)^\lambda \quad (73)$$

valid for $a_0 \geq a_1$, $b_0 \geq b_1$ and $\lambda \geq 1$, which in turn follows from convexity of $x \rightarrow x^\lambda$ ($\lambda \geq 1$). Inequality (73) shows that replacing $\{\theta_{jk}\}_{k=0}^{2^j-1}$ by its decreasing rearrangement can only increase $\|\theta\|_{f'}$.

2°. Suppose that we freeze all coefficients θ_{jk} except at level $j = j_0$, and maximise $\|\theta\|_{f'}$ subject to the constraints $|\theta_{j_0 k}| \leq \delta$ and $\sum_k |\theta_{j_0 k}|^p \leq \gamma^p$. Introduce variables $x_k = |\theta_{j_0 k}|^p / \gamma^p$ and the function $x(t) = \sum_k x_k \chi_{jk}(t)$ and write

$$f_\theta(t) = \tilde{\gamma} \cdot (g(t) + x^{q'/p}(t)), \quad \tilde{\gamma} = 2^{s' q' j_0} \gamma^{q'},$$

where $g(t)$ contains all the coefficients from levels other than j_0 , and by the previous step, both $g(t)$ and $x(t)$ may be taken to be decreasing in t . We may thus regard $\|\theta\|_{f'}^{p'}$ as a function $F(x)$ defined on the set \mathcal{D} of decreasing sequences $x_1 \geq x_2 \geq \dots \geq x_N \geq 0$ ($N = 2^{j_0}$) with $\sum_1^N x_k = 1$. We have

$$\frac{\partial F}{\partial x_k} = \tilde{\gamma} \frac{p'}{p} \int_{I_{j_0 k}} (g + x^{q'/p})^{(p'/q')-1} \cdot x_k^{(q'/p)-1}.$$

If $k' > k$, then from the hypothesis $q' \geq p$ and our monotonicity assumptions, $\partial F / \partial x_k \geq \partial F / \partial x_{k'}$, and so $F(x)$ is Schur-convex (e.g. Marshall

& Olkin (1979, Theorem A.3., p. 56)). It follows that the maximum of $F(x)$ over $\mathcal{D} \cap \{x : x_k \leq \bar{x} \triangleq (\delta/\gamma)^p \forall k\}$ is attained at the vector $x = (\bar{x}, \dots, \bar{x}, \eta\bar{x}, 0, \dots, 0)$ with $0 \leq \eta < 1$ and $\sum_1^N x_k = 1$. Returning to evaluation of $\Omega^0(\delta, C)$ and “unfreezing” the other coefficients, it now suffices to maximise $\|\theta\|_f^{p'}$ over $\theta \in \Theta(C) \cap \Theta^0(\delta)$, where $\Theta^0(\delta)$ is defined as the set of (θ_{jk}) for which there exists a sequence (n_j, δ_j) with $0 \leq n_j < 2^j$, $0 \leq \delta_j \leq \delta$ and

$$\theta_{jk} = \begin{cases} \delta & 1 \leq k \leq n_j \\ \delta_j & k = n_j + 1 \\ 0 & n_j + 1 < k \leq 2^j \end{cases}.$$

Let \mathcal{A} be the collection of decreasing sequences with $\alpha_j \in \{0\} \cup [2^{-j}, 1]$ and $\lim_{j \rightarrow \infty} \alpha_j = 0$. Given $\alpha \in \mathcal{A}$, let $n_j = \lceil 2^j \alpha_j \rceil$ and define

$$\theta_{jk} = \begin{cases} \delta & \text{if } k \leq n_j \\ 0 & \text{otherwise.} \end{cases}$$

Clearly

$$\begin{aligned} \|\theta\|_f^{p'} &\geq \delta^{p'} \int \left(\sum_j 2^{s'q'j} I\{t \leq \alpha_j\} \right)^{p'/q'} \\ &\triangleq N'(\alpha), \end{aligned}$$

and since either $n_j = 0$ or $n_j \leq 2.2^j \alpha_j$,

$$\|\theta\|_b^q = 2^{q/p} \delta^q \sum_j 2^{sqj} (2^j \alpha_j)^{q/p} \triangleq 2^{q/p} N(\alpha).$$

Define now

$$\Omega^\#(\delta, C) = \sup\{N'(\alpha) : \alpha \in \mathcal{A}, N(\alpha) \leq C^q\};$$

we have just established the left half of the estimates

$$\Omega^\#(\delta, c_0 C) \leq \Omega^0(\delta, C)^{p'} \leq \Omega^\#(\delta, c_1 C) \tag{74}$$

with $c_0 = 2^{1/p}$.

For the right-hand estimate, assume that $\theta \in \Theta(C) \cap \Theta^0(\delta)$ and define $j_{\max} = \sup\{j : n_j > 0\} < \infty$, and

$$\alpha_j = \begin{cases} \sup\{(n_{\bar{j}} + 1)2^{-\bar{j}} : j \leq \bar{j} \leq j_{\max}\} & j \leq j_{\max} \\ 0 & j > j_{\max} \end{cases}.$$

This construction guarantees that $\alpha \in \mathcal{A}$, and since $\delta_j \leq \delta$, and $n_j + 1 \leq 2^j \alpha_j$, we also have $\|\theta\|_f^{p'} \leq N'(\alpha)$. To bound $N(\alpha)$, let $\{j_L\}$ be the locations of jumps in $\{\alpha_j\}$: $\alpha_{j_L+1} < \alpha_{j_L}$. Since $\alpha_{j_L} = (n_{j_L} + 1)2^{-j_L}$ with $n_{j_L} \geq 1$,

$$\begin{aligned} \sum_{j_{L-1}+1}^{j_L} 2^{sqj} (2^j \alpha_j)^{q/p} &\leq c(\bar{s}) 2^{sqj_L} (2^{j_L} \alpha_{j_L})^{q/p} \\ &\leq 2^{q/p} c(\bar{s}) 2^{sqj_L} n_{j_L}^{q/p}. \end{aligned}$$

From this it follows that $N(\alpha) \leq 2^{q/p} c(\bar{s}) \|\theta\|_b^q$, which establishes (74) with $c_1 = 2^{1/p} c^{1/q}(\bar{s})$.

Evaluation of $\Omega^\#(\delta, C)$. For $\alpha \in \mathcal{A}$, we have

$$\begin{aligned} N'(\alpha) &= \delta^{p'} \sum_j \int_{\alpha_{j+1}}^{\alpha_j} \left(\sum_{j \leq j} 2^{s' q' j} \right)^{p'/q'} dt \\ &\leq c(s' q') \delta^{p'} \sum_j 2^{s' p' j} \alpha_j \stackrel{\Delta}{=} c(s' q') N''(\alpha). \end{aligned}$$

We now maximise $N''(\alpha)$ subject to $N(\alpha) \leq C^q$. This constraint together with the requirement that nonzero $\alpha_j \geq 2^{-j}$ implies that $\alpha_j = 0$ for $j > j_+$ (as defined at (70)). Making the change of variable $u_j = (\delta C^{-1} 2^{\bar{s} j} \alpha_j^{1/p})^q$, and noting that in the critical case $\bar{s} p = s' p'$, we have

$$N''(\alpha) = \delta^{p'-p} C^p \sum_j^{j_+} u_j^{p/q},$$

and an upper bound to $\Omega^\#(\delta, C)$ is obtained by maximising $\sum u_j^{p/q}$ over non-negative sequences $\{u_j, j \leq j_+\}$ with $\sum u_j = 1$ and $u_j \leq (\delta C^{-1} 2^{\bar{s} j})^q$. From (70), j_- is the smallest value of j for which $\delta C^{-1} 2^{\bar{s} j} \geq 1$. When $p < q$, $\sup\{\sum_{j=j_-}^{j_+} u_j^{p/q} : \sum_{j=j_-}^{j_+} u_j = 1\} = (\Delta_j)^{1-p/q}$, where $\Delta_j = j_+ - j_- + 1 \asymp \log(C/\delta)/s(ps + 1)$. On the other hand, the constraint $u_j \leq (\delta C^{-1} 2^{\bar{s} j})^q$ forces

$$\sum_{j < j_-} u_j^{p/q} \leq \sum_{j < j_-} [\delta C^{-1} 2^{\bar{s} j} - 2^{-\bar{s}(j-j_-)}]^p \leq c(\bar{s} p)$$

Combining the ranges below and above j_- , we conclude that

$$\Omega^\#(\delta, C) \leq \frac{c(s' q')}{[s(1 + sp)]^{1-p/q}} \delta^{p'-p} C^p (\log C/\delta)^{1-p/q} (1 + o(1)).$$

On the other hand, the sequence α_j corresponding to $u_j = (\Delta_j)^{-1} I\{j_a \leq j \leq j_+\}$ belongs to \mathcal{A} and shows that in fact $\Omega^\#(\delta, C) \asymp \delta^{p'-p} C^p (\log C/\delta)^{1-p/q}$.

Theorem 8: Case III: The least-favorable ball is “transitional”: $p \in \mathcal{C}$

Our lower bound for statistical estimation follows from a multi-needle-in-a-haystack problem. Here the variable n_0 tends to ∞ , but much more slowly than the size of the subproblems. Suppose that we have observations (46), but that most of the ξ_i are zero, with the exception of at most n_0 ; and that the nonzero ones satisfy $|\xi_i| \leq \delta_0$, with δ_0 a parameter. Let $\Theta_{n,0}(n_0, \delta_0)$ denote the collection of all such sequences. The following result says that, if $n_0 \ll n$ we cannot estimate ξ with an error essentially smaller than $\delta_0 n_0^{1/p'}$, provided δ_0 is not too large. This again has the interpretation

that a statistical estimation problem is not easier than the corresponding optimal recovery problem.

For the proof, and for later use, we introduce a prior distribution μ on $(\xi_i, i = 1, \dots, n)$. Set $\epsilon_n = n_0/(2n)$. Consider the law making ξ_i i.i.d. taking values 0 with probability $1 - \epsilon_n$, and with probability ϵ_n taking values $s_i \delta_0$, where the $s_i = \pm 1$ are random signs, independent and equally likely to take values +1 and -1. Then let v_i be as in (46), and let $\gamma = \delta_0/\delta$ be the signal-to-noise ratio. The argument is similar in structure to that of Lemma 9. Given an estimator $\hat{\xi}$, count the number of errors of magnitude at least $\delta_0/2$:

$$N(\hat{\xi}, \xi) = \sum_i I\{|\hat{\xi}_i - \xi_i| > \delta_0/2\}.$$

Lemma 11 *If $n_0 \leq A \cdot n^{1-a}$, and, for $\eta \in (0, a)$ we have $\delta_0 \leq \sqrt{2(a-\eta)\log(n)} \cdot \delta$ then there exist constants b_i such that*

$$\inf_{\hat{\xi}(v)} P_\mu\{N(\hat{\xi}, \xi) \geq n_0/10\} \geq 1 - e^{-b_1 n_0}, \quad (75)$$

$$\inf_{\hat{\xi}(v)} \sup_{\Theta_{n,0}(n_0, \delta_0)} P\{\|\hat{\xi}(v) - \xi\|_{p'} \geq (\delta_0/2)(n_0/10)^{1/p'}\} \geq 1 - 2e^{-b_2 n_0}. \quad (76)$$

Proof: The posterior distribution of ξ_i given v satisfies

$$P(\xi_i \neq 0 | v) = (\epsilon_n e^{-\gamma^2/2} \cosh(\gamma v/\delta)) / ((1 - \epsilon_n) + \epsilon_n e^{-\gamma^2/2} \cosh(\gamma v/\delta)).$$

Under our assumptions on ϵ_n and δ_0 , $\epsilon_n e^{-\gamma^2/2} \cosh(\gamma^2) \rightarrow 0$, so for all sufficiently large n ,

$$\epsilon_n e^{-\gamma^2/2} \cosh(\gamma v/\delta) < (1 - \epsilon_n), \quad \text{for } v \in [-\delta_0, \delta_0].$$

Therefore, the posterior distribution has its mode at 0 whenever $v \in [-\delta_0, \delta_0]$. Let $\hat{\xi}_i^*$ denote the Bayes estimator for ξ_i with respect to the 0-1 loss function $1_{|\hat{\xi}_i - \xi_i| > \delta_0/2}$. By the above comments, whenever $\xi_i \neq 0$ and $v_i \in [-\delta_0, \delta_0]$, then the loss is 1 for the Bayes rule. We can refine this observation, to say that whenever $\xi_i \neq 0$ and $\text{sgn}(\xi_i)v_i \leq \delta_0$, the loss is 1. On the other hand, given $\xi_i \neq 0$ there is a 50% chance that the corresponding $\text{sgn}(\xi_i)v_i \leq \delta_0$. Let $\pi_0 = \epsilon_n/5$. Then $\pi_0 \leq \epsilon_n/4 = P\{\xi_i \neq 0 \& \text{sgn}(\xi_i)v_i \leq \delta_0\}/2$. For the Bayes risk we have, because $0 < \pi_0 < 2\pi_0 < P\{\xi_i \neq 0 \& \text{sgn}(\xi_i)v_i \leq \delta_0\}$, and $H(\pi_0, \pi)$ is increasing in π for $\pi > \pi_0$,

$$P_\mu\{N(\hat{\xi}^*, \xi) < \pi_0 n\} \leq e^{-nH(\pi_0, 2\pi_0)} = e^{-nH(\epsilon_n/5, 2\epsilon_n/5)}. \quad (77)$$

The same inequality holds for an arbitrary estimator $\hat{\xi}$ since, just as in the proof of Lemma 9, $N(\hat{\xi}, \xi)$ is stochastically larger than $N(\hat{\xi}^*, \xi)$.

To obtain (76) we must take account of the fact that the prior μ does not concentrate on $\Theta = \Theta_{n,0}(n_0, \delta_0)$. However,

$$P(\Theta^c) = P\{\#\{i : \xi_i \neq 0\} > n_0\} \leq e^{-nH(2\epsilon_n, \epsilon_n)}.$$

Define $\bar{\mu} = \mu(\cdot|\Theta)$; since

$$P_{\bar{\mu}}(A^c) \leq P_{\mu}(A^c) + P_{\mu}(\Theta^c), \quad (78)$$

we have

$$\sup_{\hat{\xi}} P_{\bar{\mu}}\{N(\hat{\xi}, \xi) < \pi_0 \cdot n\} \leq e^{-nH(\epsilon_n/5, 2\epsilon_n/5)} + e^{-nH(2\epsilon_n, \epsilon_n)}.$$

By a calculation, for $k \neq 1$, there is $b(k) > 0$ so that

$$e^{-nH(\epsilon_n, k\epsilon_n)} \leq e^{-b(k)n\epsilon_n} = e^{-b(k)n_0},$$

as $n_0 \rightarrow \infty$. Because an error in a certain coordinate implies an estimation error of size $\delta_0/2$ in that coordinate,

$$N(\hat{\xi}, \xi) \geq m \implies \|\hat{\xi}(v) - \xi\|_{p'} \geq (\delta_0/2)m^{1/p'}.$$

Hence, with $m = \pi_0 n = n_0/10$,

$$\inf_{\hat{\xi}} P_{\bar{\mu}}\{\|\hat{\xi}(v) - \xi\|_{p'} \geq (\delta_0/2) \cdot (n_0/10)^{1/p'}\} \geq 1 - 2e^{-b'n_0}.$$

Finally (76) follows since $\sup\{P_{\xi}(A) : \xi \in \Theta_{n,0}(n_0, \delta_0)\} \geq P_{\bar{\mu}}(A)$. \square

To prove the required segment of the Theorem, we begin with the Besov case, and recall notation from the proof of the critical case of Theorem 3. For $j_-(\epsilon, C) \leq j \leq j_+(\epsilon, C)$, there are constants c_j such that $\sum_{j_-}^{j_+} 2^{sjq} c_j^q \leq C^q$. There corresponds an object supported at level $j_- \leq j \leq j_+$ and having $n_{0,j}$ nonzero elements per level, each of size ϵ , satisfying $\epsilon n_{0,j}^{1/p} \leq c_j$. This object, by earlier arguments attains the modulus to within constants, i.e.

$$\sum_{j_-}^{j_+} (2^{js'} \epsilon n_{0,j}^{1/p'})^{q'} \geq c \Omega^{q'}(\epsilon), \quad \epsilon < \delta_1(C, \delta) \quad (79)$$

A side calculation reveals that we can find $0 < a_0 < a_1 < 1$ and $A_i > 0$ so that for $j_a \leq j \leq j_b$, $\epsilon < \epsilon_1(C)$,

$$A_1 2^{j(1-a_1)} \leq n_{0,j} \leq A_0 2^{j(1-a_0)}.$$

Define now $\delta_j = \lambda_j \epsilon$, where $\lambda_j^2 = 2(1 - a_1 - \eta) \log 2^j$, and define $m_{0,j}$ such that $\delta_j m_{0,j}^{1/p} = c_j$. Then set up a prior for θ , with coordinates vanishing outside the range $[j_a, j_b]$ and with coordinates inside the range independent from level to level. At level j inside the range, the coordinates

are distributed, using Lemma 4, according to our choice of $\delta_0 \equiv \delta_j$ and $n_0 \equiv \lfloor m_{0,j} \rfloor$.

Let $N_j \sim \text{Bin}(2^j, \epsilon_j)$ be the number of non-zero components at level j , and let $B_j = \{N_j \leq n_{0j} = 2.2^j \epsilon_j\}$. On $B = \cap B_j$, it is easily checked that $\|\theta\|_b^q \leq C^q$, and since

$$P(B_j^c) \leq e^{-2^j H(2\epsilon_j, \epsilon_j)} \leq e^{-cn_{0j}},$$

it follows that $P_\mu(\Theta) \rightarrow 1$ as $\epsilon \rightarrow 0$.

From Lemma 4, at each level, the $\ell^{p'}$ error exceeds $(\delta_j/2)(m_{0,j}/10)^{1/p'}$ with a probability approaching 1. Combining the level-by-level results, we conclude that, uniformly among measurable estimates, with probability tending to one, the error is bounded below by

$$\|\hat{\theta} - \theta\|^{q'} \geq \sum_{j_a}^{j_b} (2^{js'} (\delta_j/2)(m_{0,j}/10)^{1/p'})^{q'}.$$

Now we note that

$$\delta_j m_{0,j}^{1/p'} = \lambda_j^{(1-p/p')} \epsilon n_{0,j}^{1/p'}$$

hence this last expression is bounded below by

$$\|\hat{\theta} - \theta\| \geq \lambda_j^{(1-p/p')} \cdot c_0 \cdot \Omega(\epsilon).$$

In this critical case, $r = (1 - p/p')$ and $j_- \sim \log_2(C/\epsilon)/(s + 1/p)$. Hence with overwhelming probability, $\|\hat{\theta} - \theta\| \geq c' \cdot \Omega(\epsilon \sqrt{\log(C/\epsilon)})$.

In the case of Triebel loss, we use a prior of the same structure, but set $\delta_j \equiv \delta_0 = \epsilon \lambda_j$. Also, in accordance with the proof of the modulus bound in Theorem 5 we set $\bar{c} = C(j_b - j_a)^{-1/q}$, $m_{0j} = (\bar{c} \delta_0^{-1} 2^{-js})^p$ and $n_{0j} = [m_{0j}]$.

For given $\hat{\theta}$, let $N_j(\hat{\theta}, \theta) = \#\{k : |\hat{\theta}_{jk} - \theta_{jk}| > \delta_0/2\}$. Note also that when $a \neq 0$ and (I_j) are arbitrary 0–1 valued random variables

$$(\sum_j 2^{aj} I_j)^\alpha \asymp \sum_j 2^{a\alpha j} I_j.$$

It follows that

$$\begin{aligned} \|\hat{\theta} - \theta\|_f^{p'} &\geq (\delta_0/2)^{p'} \int (\sum_{jk} 2^{s'q'j} I\{|\hat{\theta}_{jk} - \theta_{jk}| \geq \delta_0/2\} I_{jk})^{p'/q'} \\ &\asymp (\delta_0/2)^{p'} \int \sum_{jk} 2^{s'p'j} I\{|\hat{\theta}_{jk} - \theta_{jk}| \geq \delta_0/2\} I_{jk} \\ &= (\delta_0/2)^{p'} \sum_j 2^{(s'p'-1)j} N_j(\hat{\theta}, \theta). \end{aligned}$$

Let $A_j = \{N_j(\hat{\theta}, \theta) \geq m_{0j}\}$. On $A = A(\hat{\theta}) = \cap_{j_a}^{j_b} A_j$, and referring to the argument following (35),

$$\begin{aligned} \|\hat{\theta} - \theta\|_f^{p'} &\geq (\delta_0/2)^{p'} \sum_j 2^{(s'p'-1)j} m_{0j} \\ &\geq c_2 C^{1-r} \delta_0^r (\log C/\delta_0)^{1-p/q}, \\ &\geq c_3 \Omega(\epsilon \sqrt{\log \epsilon^{-1}}, C) \end{aligned}$$

with the final inequality holding if $q' \geq p$.

From (75), we conclude that $P_\mu(A^c) \rightarrow 0$ uniformly in $\hat{\theta}$ and the conclusion of the theorem now follows from (78). This completes the proof in the transitional case; the proof of Theorem 9 is complete.

Acknowledgments: This work was supported in part by NSF grants DMS 92-09130, 95-05151, and NIH grant CA 59039-18. The authors are grateful to Université de Paris VII (Jussieu) and Université de Paris-Sud (Orsay) for supporting visits of DLD and IMJ. The authors would also like to thank David Pollard and Grace Yang for their suggested improvements in presentation.

12.9 REFERENCES

- Bretagnolle, J. & Huber, C. (1979), ‘Estimation des densites: risque minimax’, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **47**, 119–137.
- Cohen, A., Daubechies, I., Jawerth, B. & Vial, P. (1993), ‘Multiresolution analysis, wavelets, and fast algorithms on an interval’, *Comptes Rendus Acad. Sci. Paris (A)* **316**, 417–421.
- Daubechies, I. (1992), *Ten Lectures on Wavelets*, number 61 in ‘CBMS-NSF Series in Applied Mathematics’, SIAM, Philadelphia.
- DeVore, R. A. & Popov, V. A. (1988), ‘Interpolation of Besov spaces’, *Transactions of the American Mathematical Society* **305**, 397–414.
- Donoho, D. & Johnstone, I. M. (1992), Minimax estimation via wavelet shrinkage, Technical report, Stanford University.
- Donoho, D. L. (1992a), ‘De-noising via soft-thresholding’, *IEEE transactions on Information Theory*. To appear.
- Donoho, D. L. (1992b), Interpolating wavelet transforms, Technical Report 408, Department of Statistics, Stanford University.

- Donoho, D. L. (1994a), 'Asymptotic minimax risk for sup-norm loss; solution via optimal recovery', *Probability Theory and Related Fields* **99**, 145–170.
- Donoho, D. L. (1994b), 'Statistical estimation and optimal recovery', *Annals of Statistics* **22**, 238–270.
- Donoho, D. L. & Johnstone, I. M. (1994), 'Ideal spatial adaptation via wavelet shrinkage', *Biometrika* **81**, 425–455.
- Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1995), 'Wavelet shrinkage: Asymptopia?', *Journal of the Royal Statistical Society, Series B* **57**, 301–369. With Discussion.
- Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1996), 'Density estimation by wavelet thresholding', *Annals of Statistics*. To appear.
- Frazier, M., Jawerth, B. & Weiss, G. (1991), *Littlewood-Paley Theory and the study of function spaces*, NSF-CBMS Regional Conf. Ser in Mathematics, **79**, American Mathematical Society, Providence, RI.
- Hall, P. & Patil, P. (1994), Effect of threshold rules on performance of wavelet-based curve estimators, Technical Report CMA-SRR13-94, Australian National University. Under revision, Statistica Sinica.
- Ibragimov, I. A. & Khas'minskii, R. Z. (1982), 'Bounds for the risks of non-parametric regression estimates', *Theory of Probability and Its Applications* **27**, 84–99.
- Johnstone, I. M. (1994), Minimax bayes, asymptotic minimax and sparse wavelet priors, in S. Gupta & J. Berger, eds, 'Statistical Decision Theory and Related Topics, V', Springer-Verlag, pp. 303–326.
- Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1992), 'Estimation d'une densité de probabilité par méthode d'ondelettes', *Comptes Rendus de l'Academie des Sciences, Paris (A)* **315**, 211–216.
- Leadbetter, M. R., Lindgren, G. & Rootzen, H. (1983), *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, New York.
- Lemarié, P. G. & Meyer, Y. (1986), 'Ondelettes et bases Hilbertiennes', *Revista Matematica Iberoamericana* **2**, 1–18.
- Mallat, S. G. (1989), 'A theory for multiresolution signal decomposition: The wavelet representation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 674–693.

- Marshall, A. W. & Olkin, I. (1979), *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York.
- Meyer, Y. (1990), *Ondelettes et Opérateurs, I: Ondelettes, II: Opérateurs de Calderón-Zygmund, III: (with R. Coifman)*, *Opérateurs multilinéaires*, Hermann, Paris. English translation of first volume is published by Cambridge University Press.
- Meyer, Y. (1991), 'Ondelettes sur l'intervalle', *Revista Matematica Iberoamericana* **7**, 115–133.
- Micchelli, C. A. (1975), Optimal estimation of linear functionals, Technical Report 5729, IBM.
- Micchelli, C. A. & Rivlin, T. J. (1977), A survey of optimal recovery, in C. A. Micchelli & T. J. Rivlin, eds, 'Optimal Estimation in Approximation Theory', Plenum Press, New York, pp. 1–54.
- Nemirovskii, A. S. (1985), 'Nonparametric estimation of smooth regression function', *Izv. Akad. Nauk. SSR Teckhn. Kibernet.* **3**, 50–60. (in Russian). *J. Comput. Syst. Sci.* **23**, 6, 1–11, (1986) (in English).
- Peetre, J. (1975), *New Thoughts on Besov Spaces, I*, Duke University Mathematics Series, Raleigh, Durham.
- Samarov, A. (1992), Lower bound for the integral risk of density function estimates, in R. Z. Khasminskii, ed., 'Advances in Soviet Mathematics' **12**, American Mathematical Society, Providence, R.I., pp. 1–6.
- Stone, C. (1982), 'Optimal global rates of convergence for nonparametric regression', *Annals of Statistics* **10**, 1040–1053.
- Traub, J., Wasilkowski, G. & Woźniakowski (1988), *Information-Based Complexity*, Addison-Wesley, Reading, MA.
- Triebel, H. (1992), *Theory of Function Spaces II*, Birkhäuser Verlag, Basel.

13

Empirical Processes and p-variation

R. M. Dudley¹

ABSTRACT Remainder bounds in Fréchet differentiability of functionals for p -variation norms are found for empirical distribution functions. For $1 \leq p \leq 2$ the p -variation of the empirical process $n^{1/2}(F_n - F)$ is of order $n^{1-p/2}$ in probability up to a factor $(\log \log n)^{p/2}$. For $(F, G) \mapsto \int F dG$ and for $(F, G) \mapsto F \circ G^{-1}$ this yields nearly optimal remainder bounds. Also, p -variation gives new proofs for the asymptotic distributions of the Cramér-von Mises-Rosenblatt and Watson two-sample statistics when the two sample sizes m, n go to infinity arbitrarily.

13.1 Introduction

This paper is a continuation of Dudley (1994), in which specific remainder bounds are found for the differentiation of operators on distribution functions with special regard for empirical distribution functions.

Let F be a probability distribution function and F_n an empirical distribution function for it. Then $n^{1/2}(F_n - F)$ is an *empirical process*. The p -variation of the process (defined in Section 2) is bounded in probability as $n \rightarrow \infty$ if and only if $p > 2$: Dudley (1992, Corollary 3.8; 1994). Theorem 2 below shows that for $1 \leq p \leq 2$, the p -variation of the empirical process is at least $n^{1-p/2}$ (for F continuous) and is at most of the order $n^{1-p/2}(\log \log n)^{p/2}$ in probability.

Then, Section 3 treats the operator $(F, G) \mapsto \int F dG$ and finds a bound in probability for the integral of one empirical process based on observations i.i.d. with distribution F with respect to another such process for a distribution G (where the two processes have an arbitrary joint distribution). Section 4 treats composition $G \mapsto F \circ G$, for fixed F , where $(F \circ G)(x) \equiv F(G(x))$. The operator $(f, g) \mapsto (F + f) \circ (G + g)$, with p -variation norms on f , was considered in Dudley (1994). Section 5 treats the operator $(F, G) \mapsto F \circ G^{-1}$ and Section 6, the two-sample Cramér-von Mises statistics of Rosenblatt (1952) and Watson (1962).

The inverse operator $F \mapsto F^{-1}$ was known to be compactly but not Fréchet differentiable with respect to the sup norm on distribution functions

¹Mathematics Department, Massachusetts Institute of Technology

F and L^p norm on the range, and likewise for the composition operator $(F, G) \mapsto F \circ G$ with respect to F , for compact differentiability with respect to the L^p norm also on G : Reeds (1976), Fernholz (1983, Props. 6.1.1, 6.1.6). The use of p -variation norms yields Fréchet differentiability and explicit bounds on remainders which for empirical distribution functions F_m and G_n give for these operators, as shown in Dudley (1994), the correct powers of m and n in remainder bounds, although not necessarily the right logarithmic factors or the strongest norms on the ranges of the operators. The present paper does the same for the operator $(F, G) \mapsto \int F dG$. Recall that as shown in Dudley (1994, Proposition 2.1), when Fréchet differentiability fails, as for the sup norm, compact differentiability yields no such remainder bounds.

When an operator is a composition of other operators, the best possible remainder bounds may not follow from the chain rule. For example, as will be seen at the end of Section 5, to represent $\int F dG$ as $\int_0^1 F \circ G^{-1} dt$ gives a non-optimal remainder bound.

13.2 Empirical Processes' p-variation

Let ψ be a convex, increasing function from $[0, \infty)$ onto itself and let f be a function from an interval J into \mathbf{R} . Recall that the ψ -variation of f is defined by

$$v_\psi(f) := \sup \left\{ \sum_{k=1}^m \psi(|f(t_k) - f(t_{k-1})|) : t_0 \in J, \right. \\ \left. t_0 < t_1 < \dots < t_m \in J, m = 1, 2, \dots \right\}.$$

Let $Ly := \max(1, \log y)$. Recall that as noted in Dudley (1994, Section 5), there is a convex, increasing function ψ_1 with $\psi_1(x) = x^2/LL(1/x)$ for $0 < x \leq e^{-e}$. Letting

$$\psi_1(x) := e^{-2e} + C(x^2 - e^{-2e}) \text{ for } x \geq e^{-e}$$

for a large enough constant C , we will have $\psi_1(x) \geq x^2/LL(1/x)$ for all $x > 0$. Taylor (1972) showed that for the Brownian motion process $x(\cdot) : t \mapsto x_t$ on $0 \leq t \leq 1$, almost surely $v_{\psi_1}(x(\cdot)) < \infty$, while if $\psi_1(x) = o(\psi(x))$ as $x \downarrow 0$, then $v_\psi(x(\cdot)) = +\infty$ almost surely. So for the Brownian bridge process $y_t = x_t - tx_1$, $0 \leq t \leq 1$, the same holds. It follows that if $\psi_1 = o(\psi)$, $v_\psi(n^{1/2}(F_n - F))$ is not bounded in probability as $n \rightarrow \infty$, while $v_{\psi_1}(n^{1/2}(F_n - F))$ is bounded in probability: Dudley (1994, Section 5).

For any ψ , the space of all functions f on J such that $v_\psi(cf) < \infty$ for some $c > 0$ will be called $W_\psi(J)$. For $\psi(x) \equiv x^p$, $1 \leq p < \infty$, the ψ -variation is called p -variation and denoted $v_p(f)$. Then $W_p(J) := W_\psi(J)$ is a Banach space with the p -variation norm defined by $\|f\|_{[p]} := v_p(f)^{1/p} + \|f\|_\infty$ where $\|f\|_\infty := \sup\{|f(x)| : x \in J\}$.

It follows from the results about ψ_1 stated above that

$$(1) \quad \|n^{1/2}(F_n - F)\|_{[p]} \text{ is bounded in probability as } n \rightarrow \infty \\ \text{if and only if } p > 2.$$

("If" was also proved in Dudley (1992, Corollary 3.8).) For $p \leq 2$, although the p -variation norm of the empirical process will not be bounded in probability, its order of growth in n can be bounded as follows.

(2) **Theorem** *For any distribution function F on \mathbf{R} and $1 \leq p \leq 2$, $v_p(n^{1/2}(F_n - F)) = O_p(n^{1-p/2}(LLn)^{p/2})$ as $n \rightarrow \infty$. Conversely if F is continuous, then almost surely for all n , $v_p(n^{1/2}(F_n - F)) \geq n^{1-p/2}$.*

Proof. To prove the first statement, let U be the uniform $U[0, 1]$ distribution function, $U(x) \equiv \max(0, \min(x, 1))$, and let U_n be empirical distribution functions for it. Then for any distribution function F , $U \circ F \equiv F$ and $U_n \circ F$ have all the properties of the F_n . So we can take $n^{1/2}(F_n - F) = n^{1/2}(U_n - U) \circ F$. Since F is non-decreasing, we get $v_p(F_n - F) \leq v_p(U_n - U)$ for any p . So we can assume F is continuous.

First let $p = 2$. For a given n let $f := n^{1/2}(F_n - F)$, and for any $t_0 < t_1 < \dots < t_m$ for any positive integer m let $\Delta_i := f(t_i) - f(t_{i-1})$, $i = 1, \dots, m$. In the supremum of 2-variation sums we can assume that the Δ_i are alternating in sign, since $(A + B)^2 > A^2 + B^2$ for $AB > 0$, so any adjoining differences of the same sign should be combined.

Claim In the supremum we can assume

- (a) $\Delta_i > 2/(3n^{1/2})$ whenever $\Delta_i > 0$, and
- (b) If $\Delta_i < 0$, then either $i = 1$ or m or $|\Delta_i| \geq 1/(3n^{1/2})$.

Proof of (a). Since $\Delta_i > 0$, there is at least one X_j (in the sample on which F_n is based) with $t_{i-1} < X_j \leq t_i$. For the largest such X_j , we can take $t_i = X_j$ since this can only enlarge both Δ_i and $|\Delta_{i+1}|$. Likewise, we can let t_{i-1} increase toward (but not quite equal) some $X_k \leq X_j$. If $X_k = X_j$ then (a) holds. Otherwise $X_k < X_j$ and if $\Delta_i \leq 2/(3n^{1/2})$ we can move t_{i-1} to a point just less than X_j . This will increase both $|\Delta_{i-1}|$ and Δ_i and make $\Delta_i > 2/(3n^{1/2})$, so (a) is proved.

Proof of (b). If $\Delta_i < 0$ and $i \geq 2$ the sum can only be increased by letting t_{i-1} decrease until $t_{i-1} = X_j$ for some j , and if $i < m$, we can also let t_i increase nearly up to another observation X_k . Suppose (b) fails for Δ_i and there is another r with $X_j < X_r < X_k$. Then by inserting new division points at X_r and just below it, we can enlarge the 2-variation sum. If there are then adjoining positive Δ_j , they can be combined as before. Then (a) is preserved. Iterating, we can assume that there is no such r . Then, I claim the sum is increased by deleting t_{i-1} and t_i : by (a), we have $A := \Delta_{i-1} > 2/(3n^{1/2})$ and $B := \Delta_{i+1} > 2/(3n^{1/2})$, while $0 < C := -\Delta_i < 1/(3n^{1/2})$. Then $(f(t_{i+1}) - f(t_{i-2}))^2 = (A + B - C)^2 > A^2 + B^2 + C^2$, in other words $AB > C(A + B)$ since $C < \frac{1}{2} \min(A, B)$ and $A + B \leq 2 \max(A, B)$. Here

$A + B - C > 2/(3n^{1/2}) > 0$, so again (a) is preserved, and, iterating, (b) is proved.

Now, continuing with the proof of Theorem 2, since

$$\psi_1(x) \geq x^2/(LL(1/x)) \text{ for all } x > 0,$$

we have for any set $I \subset \{1, \dots, m\}$,

$$\sum_{i \in I} \psi_1(|\Delta_i|) \geq \sum_{i \in I} \Delta_i^2/(LL(1/|\Delta_i|)).$$

Taking I to be the set of all i for which $|\Delta_i| \geq 1/(3n^{1/2})$, which is all $i = 1, \dots, m$ except possibly for $i = 1, m$, we get

$$v_2(f) = O\left(v_{\psi_1}(f)LLn + \frac{1}{n}\right) \text{ as } n \rightarrow \infty.$$

Since $v_{\psi_1}(f)$ is bounded in probability the case $p = 2$ follows.

Now for $1 \leq p < 2$, the p -variation sums can be bounded by the Hölder inequality as follows:

$$\sum_{k=1}^m |f(t_k) - f(t_{k-1})|^p \leq \left(\sum_{k=1}^m |f(t_k) - f(t_{k-1})|^2\right)^{p/2} m^{1-p/2}.$$

The n observations X_1, \dots, X_n divide the line into at most n left closed, right open intervals and one interval $(-\infty, \min_i X_i)$. In each such interval f is nonincreasing, and in the supremum of the sums we can assume there are at most two values of t_k in each of the $n+1$ intervals, from the inequality $(A+B)^p \geq A^p + B^p$, $p \geq 1$, $A, B \geq 0$. Thus we can take $m \leq 2n+2$. Then from the $p = 2$ case, the upper bound follows.

In the other direction, if F is continuous, then for a given n , almost surely $f := n^{1/2}(F_n - F)$ will have jumps of height $n^{-1/2}$ at n distinct points. For $m = 2n$, putting t_k at and just below each jump shows that $v_p(f) \geq n^{1-p/2}$. \square

Since $(F_n - F)(t) \rightarrow 0$ as $t \rightarrow -\infty$, $\|F_n - F\|_\infty \leq v_p(F_n - F)^{1/p}$ and $\|F_n - F\|_{[p]} \leq 2v_p(F_n - F)^{1/p}$. Thus we have

(3) **Corollary** For $1 \leq p \leq 2$,

$$\|n^{1/2}(F_n - F)\|_{[p]} = O_p(n^{(2-p)/(2p)}(LLn)^{1/2}) \quad \text{as } n \rightarrow \infty.$$

13.3 The operator $(F, G) \mapsto \int F dG$

This operator is bilinear in F and G . Specifically, we have, if all the integrals are defined,

$$(4) \quad \int (F + f)d(G + g) = \int FdG + \int fdG + \int Fdg + \int f dg.$$

Here if we think of F and G as fixed and f, g as small, approaching 0, then $\int FdG$ is the value of the operator at the point F, G , while $\int fdG$ is a

partial derivative term in the direction f , $\int F dg$ is a partial derivative term in the direction g , and $\int f dg$ is the remainder. To apply the operator to empirical distribution functions F_n, G_m we have $f = F_n - F$, $g = G_m - G$.

Young (1936, (10.9)) proved a basic inequality, given here in a slightly different form as in Dudley (1992, Theorem 3.5), for $\int F dG$ in terms of p -variation norms:

$$(5) \quad \begin{aligned} |\int F dG| &\leq C_{p,q} \|F\|_{[p]} \|G\|_{[q]} \text{ if } \frac{1}{p} + \frac{1}{q} > 1 \\ \text{where } C_{p,q} &:= 1 + \zeta\left(\frac{1}{p} + \frac{1}{q}\right) < \infty \end{aligned}$$

and $\zeta(s) \equiv \sum_{n \geq 1} n^{-s}$. Here $\int F dG$ is an extended Riemann-Stieltjes integral defined by Young (1936); for some small clarifications see Dudley (1992, Section 3). Inequality (5) provides a duality between p -variation spaces. But, by (1), the empirical process $n^{1/2}(F_n - F)$ has p -variation norms bounded in probability as $n \rightarrow \infty$ only for $p > 2$, and we can't take $p > 2$ and $q > 2$ simultaneously in Young's bound. To bound the remainder here we can apply Theorem 2 to make $p < 2$ or $q < 2$. We have from (4)

$$(6) \quad \begin{aligned} \int F_n dG_m &= \\ \int F dG + \int (F_n - F) dG + \int F d(G_m - G) + \int (F_n - F) d(G_m - G). \end{aligned}$$

Here if F_n is based on observations X_1, \dots, X_n and G_m on observations Y_1, \dots, Y_m , and if $X_i \neq Y_j$ for all i and j , then $mn \int F_n dG_m$ is the number of pairs (i, j) such that $X_i < Y_j$, which is the well-known Mann-Whitney statistic, e.g. Randles & Wolfe (1991, (2.3.8)). The remainder term can be bounded as follows:

(7) **Proposition** For any $\epsilon > 0$, for any distribution functions F, G and any possible joint distribution of F_n and G_m , as $m, n \rightarrow \infty$,

$$(mn)^{1/2} \int (F_n - F) d(G_m - G) = O_p(\min(m, n)^\epsilon).$$

Proof. Choose $1 < p < 2$ such that $1 - p/2 < \epsilon$, then $r > 2$ such that $\frac{1}{p} + \frac{1}{r} > 1$. Apply Corollary 3 to bound the p -variation norm of $F_n - F$ by $O_p(n^{\epsilon-1/2})$. The r -variation norm of $G_m - G$ is of order $O_p(m^{-1/2})$ since $r > 2$: Dudley (1992, Corollary 3.8). So the Young duality inequality (5) gives the conclusion for $n \leq m$. Otherwise, we can consider $\|G_m - G\|_{[p]}$ and $\|F_n - F\|_{[r]}$ and so multiply by m^ϵ rather than n^ϵ . \square

Proposition 7 implies the following asymptotic behavior of $\int F_n dG_m$: in (6), $\int (F_n - F) dG$ and $\int F d(G_m - G)$ are asymptotically normal and of orders $O_p(n^{-1/2})$ and $O_p(m^{-1/2})$ respectively. If $0 < \epsilon < 1/2$, $m \rightarrow \infty$ and $n \rightarrow \infty$, then Proposition 7 implies that the remainder is

$$O_p(\max(m, n)^{-1/2} \min(m, n)^{\epsilon-1/2}) = o_p(\min(m, n)^{-1/2}).$$

The o_p bound on the right, which is enough to imply asymptotic normality of $\int F_n dG_m$ when F_n and G_m are independent, also follows from results of

Gill (1989, pp. 110-111), proved by way of compact differentiability of (a modification of) $(F, G) \mapsto \int F dG$ in the supremum norm.

The stronger bound for the remainder gives directly the following:

(8) **Corollary** *For any possible joint distributions of F_m and G_n , if $m, n \rightarrow \infty$ in such a way that $m/n \rightarrow c$ where $0 < c < \infty$, then $n^{1/2}(\int F_n dG_m - \int F dG)$ converges in distribution to a sum of two normal variables. The remainder term $n^{1/2}(\int F_n - F d(G_m - G))$ is $O_p(n^{\epsilon - \frac{1}{2}})$ for any $\epsilon > 0$.*

If F_n and G_m are independent, then the normal variables are independent, so their sum is normal.

If F_m and G_n are independent and G is continuous, then (as R. Pyke kindly pointed out to me) the expression in Proposition 7 has mean 0 and finite variance: the variance is $\int F - F^2 dG - \int G^2 dF + (\int G dF)^2$, for all m and n , so the expression is $O_p(1)$. So Proposition 7, although not quite optimal, at least indicates the correct powers of m and n . When F_n and G_m have an arbitrary joint distribution, $\min(m, n)^\epsilon$ can at any rate be replaced, not surprisingly, by a logarithmic factor, using Orlicz-variation duality (Young 1938, Theorem 5.1). The details are not carried through here since it isn't clear that the result would be optimal.

13.4 The operator $\mathbf{G} \mapsto \mathbf{F} \circ \mathbf{G}$

In this section we consider differentiability of the operator $h \mapsto F \circ (H+h)$ at $h = 0$ from L^s to L^p , $1 \leq p < s$, when F is twice continuously differentiable on a bounded interval $[a, b]$ including the range of H . Such operators have been much studied and are special cases of what are often called Nemitsky operators, e.g. Appell & Zabrejko (1990). We get a bound for the remainder in case $H+h$ also has values in $[a, b]$, which will be applicable when F , H and $H+h$ are all distribution functions. In a more general case, when F (like the uniform $U[0, 1]$ distribution function) is non-differentiable at the endpoints a, b and this turns out to contribute the leading error term in the remainder, we get a larger bound.

A function F on a closed interval $[a, b]$ will be called C^k if it is continuous on $[a, b]$ and has derivatives through k th order continuous on the open interval (a, b) which extend to continuous functions on $[a, b]$.

(9) **Theorem** *Let F be a function from \mathbf{R} into \mathbf{R} whose restriction to a bounded interval $[a, b]$ is C^2 .*

(i) *Let (X, \mathcal{A}, μ) be a finite measure space and H a measurable function from X into $[a, b]$. For $1 \leq p < s$ and $h \in L^s(X, \mu)$, define the remainder*

$$R(h) := F \circ (H + h) - F \circ H - (F' \circ H)h.$$

If $\|h\|_s \rightarrow 0$ for h such that $H+h$ also takes values in $[a, b]$, then

$$\|R(h)\|_p = O(\|h\|_s^{\min(2, s/p)}).$$

Here $O(\cdot)$ cannot be replaced by $o(\cdot)$.

(ii) Suppose that $F(y) = F(a)$ for $-\infty < y \leq a$, $F(y) = F(b)$ for $b \leq y < \infty$, F is nondecreasing, $X = [c, d]$ for some $c < d$ with $\mu = \text{Lebesgue measure}$, and H is an increasing C^1 function from $[c, d]$ onto $[a, b]$ with $H'(x) \geq \delta > 0$ for $c < x < d$. Then the operator $h \mapsto F \circ (H + h)$ is Fréchet differentiable at $h = 0$ from L^s to L^p with usual derivative $h \mapsto (F' \circ H)h$, and remainder of order $\|R(h)\|_p = O(\|h\|_s^\zeta)$ where $\zeta = s(p+1)/(p(s+1))$. Again, $O(\cdot)$ cannot be replaced by $o(\cdot)$.

Proof. (i) Since $H + h$ and H take values in $[a, b]$ and F is C^2 there, $|R(h)| \leq C|h|^2$ as $|h| \rightarrow 0$ for some constant $C < \infty$, and $|h| \leq b - a$. If $s \geq 2p$, then $(\int |h|^{2p} d\mu)^{1/p} = O(\|h\|_s^2)$. If $s < 2p$, then the inequality $\int |h|^{2p} d\mu \leq \int |h|^s d\mu (b - a)^{2p-s}$ implies that $\|R(h)\|_p = O(\|h\|_s^{s/p})$ and the first conclusion follows.

To see that $O(\cdot)$ cannot be replaced by $o(\cdot)$, let $F(x) = x^2$, $X = [a, b] = [0, 1]$, $\mu = \text{Lebesgue measure}$, $H(x) \equiv x$. Then $R(h) \equiv h^2$. For $s \geq 2p$, take $h = c1_{[0,1-c]}$ as $c \downarrow 0$. Then $\|h^2\|_p \sim c^2 \sim \|h\|_s^2$, so $O(\cdot)$ cannot be replaced by $o(\cdot)$. Or if $p < s < 2p$, let $h = (1-c)1_{[0,c]}$. Then as $c \downarrow 0$, $\|h^2\|_p \sim c^{1/p}$ and $\|h\|_s \sim c^{1/s}$, so again $O(\cdot)$ cannot be replaced by $o(\cdot)$. In both cases $H + h$ takes values in $[0, 1]$. (If desired, F can be a probability distribution function with $\inf_{[0,1]} F' > 0$: let $F(x) = \frac{1}{2}(x + x^2)$ for $0 \leq x \leq 1$, and the same examples apply up to constants.) So (i) is proved.

(ii) Here $H + h$ can take values outside of $[a, b]$. Let's assume without loss of generality that $a = c = 0$, $F(0) = 0$ and $b = d = 1$. Then since H is onto, $H(0) = 0$ and $H(1) = 1$. Let A be the measurable set of $x \in [0, 1]$ where $(H + h)(x) < 0$. Now since $H' \geq \delta > 0$, $H(x) \geq \delta x$ for $0 \leq x \leq 1$. So on A , $h(x) < -\delta x$ and $F(H + h)(x) = 0$. Also on A , $R(h)(x) = -F(H(x)) - F'(H(x))h(x)$, $0 \leq F(H(x)) \leq Cx$ for some C since F and H are C^1 , and $0 \leq -F'(H(x))h(x) \leq D|h(x)|$ for some constant D . Then since $|h(x)| > \delta x$, there is a constant $M < \infty$ ($M = D + C/\delta$) such that $|R(h)| \leq M|h|$ on A . By Hölder's inequality

$$\int_A |R(h)(x)|^p dx \leq M^p \int_0^1 |h|^p 1_{\{|h| > \delta x\}} dx \leq M^p \|h\|_s^p (\lambda(|h| > \delta x))^{1-p/s}$$

where λ is Lebesgue measure. For fixed $\|h\|_s$, $\lambda(|h| \geq \delta x)$ is maximized when $|h| = \delta x$ on an interval $[0, \alpha]$ and $h(x) = 0$ otherwise, when it is $O(\|h\|_s^{s/(s+1)})$. Combining terms and taking p th roots gives, on A , the stated bound for the p -norm of the remainder. By symmetry, we get a bound of the same order for the remainder on the set where $H + h > 1$. Now $p < s$ implies $s(p+1)/(p(s+1)) < \min(s/p, 2)$, so that the remainder bound coming from $H + h$ outside $[a, b]$ dominates the bound from $H + h$ in $[a, b]$. To see that $O(\cdot)$ cannot be replaced by $o(\cdot)$, consider the functions $F(x) = H(x) = x$, $0 \leq x \leq 1$, and let $h(x) = -2x$ on an interval $[0, \alpha]$, $h(x) = 0$ for $x > \alpha$, and let $\alpha \downarrow 0$. \square

Andersen, Borgan, Gill & Keiding (1993, Proposition II.8.8) consider

(compact) differentiability of composition with sup norm on the range. Then in the situation of the second half of Theorem 9, differentiability fails, so H is required to have range in a proper subinterval $[s, t]$, $a < s < t < b$.

13.5 The operator $(F, G) \mapsto F \circ G^\leftarrow$

Given an interval $[a, b]$, and a real-valued function G on $[a, b]$, and any real y let

$$G^\leftarrow(y) := G_{[a,b]}^\leftarrow(y) := \inf\{x \in [a, b] : G(x) \geq y\},$$

or $G_{[a,b]}^\leftarrow(y) := b$ if $G(x) < y$ for all $x \in [a, b]$. The notation G^\leftarrow is used instead of G^{-1} to specify a particular definition of “inverse,” e.g. Beirlant & Deheuvels (1990).

Suppose that inverses are taken with respect to a bounded interval $[a, b]$. If G is C^2 from $[a, b]$ onto an interval $[c, d]$, $G' \geq \delta$ for some $\delta > 0$, and $1 \leq s < \infty$, then by Corollary 2.4 of Dudley (1994), the inverse operator taking g to $(G + g)^\leftarrow$ restricted to $[c, d]$ is Fréchet differentiable at $g = 0$ from $(W_s([a, b]), \|\cdot\|_{[s]})$ into $(L^s[c, d], \|\cdot\|_s)$ (for the L^s norm with respect to Lebesgue measure on $[c, d]$), with derivative $g \mapsto -(g \circ G^\leftarrow)/(G' \circ G^\leftarrow)$. The remainder

$$R_g := (G + g)^\leftarrow - G^\leftarrow + (g \circ G^\leftarrow)/(G' \circ G^\leftarrow)$$

satisfies

$$(10) \quad \|R_g\|_s = O(\|g\|_{[s]}^{1+1/s}) \text{ as } \|g\|_{[s]} \rightarrow 0.$$

Let U be the $U[0, 1]$ distribution function, $U(x) = \max(0, \min(x, 1))$ for all x . In the following $\|\cdot\|_\infty$ denotes the sup norm $\|h\|_\infty := \sup_{0 < t < 1} |h(t)|$. On $[0, 1]$ let

$$R_{f,g} := (U + f) \circ (U + g)^\leftarrow - U - f \circ U + g,$$

the remainder in differentiating $(f, g) \mapsto (U + f) \circ (U + g)^\leftarrow$ at $f = g = 0$. Let $\|\cdot\|_p$ be the L^p norm with respect to Lebesgue measure on $[0, 1]$.

(11) **Proposition** *For some constant $C < \infty$, $1 \leq p < \infty$, any measurable real functions f on \mathbf{R} and g, h on $[0, 1]$, let inverses be taken with respect to $[0, 1]$. Then*

- (a) $\|f \circ (U + h) - f \circ U\|_p \leq C\|f\|_{[p]}\|h\|_\infty^{1/p}$.
- (b) $\|(U + g)^\leftarrow - U\|_\infty \leq \|g\|_\infty$.
- (c) $\|f \circ (U + g)^\leftarrow - f \circ U\|_p \leq C\|f\|_{[p]}\|g\|_\infty^{1/p}$.
- (d) $\|R_{f,g}\|_p = O(\|f\|_{[p]}^{1+1/p} + \|g\|_{[p]}^{1+1/p})$ as $\|f\|_{[p]} + \|g\|_{[p]} \rightarrow 0$.

Proof. For (a), although Theorem 2.2 of Dudley (1994) is stated only for $s < \infty$, it actually holds (the proof only gets a little easier) for $s = \infty$, so (a) holds.

For (b), let $\delta := \|g\|_\infty$. If $x + g(x) \geq y$ then $x + \delta \geq y$, so $(U + g)^{\leftarrow}(y) \geq y - \delta$. Conversely if $y + \delta \leq 1$ then $y + \delta + g(y + \delta) \geq y$ so $(U + g)^{\leftarrow}(y) \leq y + \delta$, which also holds if $y + \delta > 1$, and (b) follows.

Then (c) follows from (a) and (b). For (d),

$$R_{f,g} = f \circ (U + g)^{\leftarrow} - f \circ U + R_g.$$

So from (10), (c), and the inequality $AB^\alpha \leq A^{1+\alpha} + B^{1+\alpha}$ for $A, B, \alpha > 0$, (d) follows. \square

Now consider the case when F , $F + f$, G and $G + g$ are all probability distribution functions, and specifically $F + f = F_m$ (an empirical distribution function) and $G + g = G_n$. Then $F \circ G^{\leftarrow}$ is a procentile-procentile function (P-P plot, e.g. Beirlant & Deheuvels (1990)) and $F_m \circ G_n^{\leftarrow}$ its empirical counterpart. The special case where both F and G equal U is in fact not so special: if F_m and G_n are both empirical distribution functions from the same continuous distribution function F , strictly increasing on an interval $[a, b]$ with $F(a) = 0$ and $F(b) = 1$, then we can write $F_m = U_m \circ F$ and $G_n = V_n \circ F$ where U_m and V_n are empirical distribution functions for $U = U[0, 1]$. So $F_m \circ G_n^{\leftarrow} \equiv U_m \circ V_n^{\leftarrow}$. Thus, as is known, any results for $U_m \circ V_n^{\leftarrow}$ yield conclusions for such $F_m \circ G_n^{\leftarrow}$.

(12) **Proposition** *If U_m and V_n are empirical distribution functions for U , which can have any joint distribution, and $m \wedge n := \min(m, n)$,*

$$\|U_m \circ V_n^{\leftarrow} - U_m + V_n - U\|_2 = O_p([(LL(m \wedge n))/(m \wedge n)]^{3/4}) \text{ as } m, n \rightarrow \infty.$$

Proof. Apply Proposition 11(d) and Corollary 3 for $p = 2$. \square

In the situation of Proposition 12, the exponent $-3/4$ for $m \wedge n$ is correct, for example when $m \gg n$ by Theorem 2.6 of Dudley (1994) and the comment after it. If we apply the latter Theorem instead of (10) in the proof of Proposition 11(d) in the case of Proposition 12, the bound in the latter is replaced by

$$(13) \quad O_p((LLm)^{1/2}m^{-1/2}n^{-1/4} + n^{-3/4}),$$

which is better by a factor $(LLm)^{1/4}$ if m and n are of the same order of magnitude and still better if they are not. I don't know whether the $(LLm)^{1/2}$ factor is needed.

As in the case of the inverse operator $g \mapsto (G + g)^{\leftarrow}$ [Dudley (1994, Theorem 2.6 and the discussion after it)] it seems that to treat L^p norms of remainders for $2 < p \leq \infty$, one must go beyond the kinds of proofs treated so far in this paper. From Beirlant & Deheuvels (1990, (1.9), (2.1), (2.2), (2.4), (2.6) and (2.7)) it follows that if $m/(n^{1/2} \log n) \rightarrow \infty$ and $n^{3/2}/(m \log n) \rightarrow \infty$ then

$$\|U_m \circ V_n^{\leftarrow} - U_m + V_n - U\|_\infty = O_p\left(\frac{(\log n)^{1/2}}{m^{1/2}n^{1/4}} + \frac{\log m}{m} + \frac{\log n}{n}\right).$$

(For this conclusion, independence of U_m and V_n is not needed.) If m and n are of the same order of magnitude, the bound for the sup norm becomes $O_p((\log n)^{1/2}/n^{3/4})$, larger than (13) for the L^2 norm.

Suppose $(X_1, \|\cdot\|)$, $(X_2, \|\cdot\|)$ and $(X_3, \|\cdot\|)$ are normed spaces and for $i = 1, 2$, T_i is a map from an open set $U_i \subset X_i$ into X_{i+1} , Fréchet differentiable at a point $x_i \in U_i$ with derivative $T'_i(x_i)$. Suppose $T(x_1) = x_2$ and $T(x) \in U_2$ for all $x \in U_1$. Then by a well-known chain rule, $T_2 \circ T_1$ is Fréchet differentiable at x_1 from U_1 into X_3 . Suppose that for some $\alpha > 1$ we have remainder bounds

$$\|T_i(x_i + u_i) - T_i(x_i) - T'_i(x_i)(u_i)\| = O(\|u_i\|^\alpha)$$

as $\|u_i\| \rightarrow 0$, $u_i \in X_i$, for $i = 1, 2$. Then one can get such a remainder bound also for $T_2 \circ T_1$. But such general chain rule remainder bounds may not give bounds of the best possible order for the composition, as in the following example.

It is well known, and not hard to check, that we have the probability integral transformation (for a Lebesgue-Stieltjes integral) $\int_{-\infty}^{\infty} FdG = \int_0^1 (F \circ G^{-})(t)dt$ for any two distribution functions F , G of probability measures. The results for $F \circ G^{-}$ from this section could then be applied to $\int FdG$, as follows: $\int_0^1 U_m dV_n = \int_0^1 U_m \circ V_n^{-} dt$. The integral from 0 to 1 is a linear operation and so differentiable with no remainder. Composing and applying Proposition 12 yields

$$\left| \int_0^1 U_m dV_n - \frac{1}{2} - \int_0^1 U_m - V_n dt \right| = O_p \left(\left[\frac{LL(m \wedge n)}{m \wedge n} \right]^{3/4} \right).$$

But Proposition 7 gives the stronger bound $O_p((m \wedge n)^\epsilon (mn)^{-1/2})$.

13.6 The 2-sample Cramér-von Mises-Rosenblatt and Watson statistics

Suppose again that F, G are distribution functions. Let F_m be an empirical distribution function for F and G_n an empirical distribution function for G , independent of F_m . Then the two-sample empirical process will be defined by

$$(F, G)_{m,n} := \left(\frac{mn}{m+n} \right)^{1/2} (F_m - G_n).$$

Let $N := m+n$ and $K_N := K_{m,n} := (mF_m + nG_n)/N$. Lehmann (1951, p. 174) and Rosenblatt (1952) proposed a test for $F = G$ based on the statistic $\int(F, G)_{m,n}^2 d(F_m + G_n)/2$. Later authors (mentioned in the Notes below) have preferred the closely related statistic

$$W^2 := W_{m,n}^2 := \int(F, G)_{m,n}^2 dK_{m,n},$$

which has the same asymptotic distribution by nearly the same proof. Earlier A. M. Mood, according to Dixon (1940), suggested for $m = n$ the statistic with F_m (or G_n) instead of K_N .

(14) **Theorem (Rosenblatt-Fisz-Kiefer)** *For any continuous distribution function $F = G$, the statistic $W_{m,n}^2$ converges in distribution as $m, n \rightarrow \infty$ to $\int_0^1 y_t^2 dt$, where y_t is the Brownian bridge.*

Notes So, the asymptotic distribution is the same as for the one-sample Cramér-von Mises statistic $n \int (F_n - F)(t)^2 dF(t)$. Rosenblatt (1952) proved the conclusion under the further assumption

$$(15) \quad \text{as } m, n \rightarrow \infty, m/n \rightarrow \lambda \text{ where } 0 < \lambda < \infty.$$

Fisz (1960) gave a correction to the proof. Aki (1981) gave another proof, also assuming (15), based on functional differentiability methods, extending the work of Filippova (1961). Darling (1957, p. 827) indicated that the weaker assumption

$$(16) \quad 0 < \liminf m/n \leq \limsup m/n < \infty$$

would suffice, even for weighted forms of the statistic. Theorem 14 holds as stated, with $m \rightarrow \infty$ and $n \rightarrow \infty$ with no restriction such as (15) or (16) on m/n , as was first shown apparently by Kiefer (1959) and will also follow from the proof below by way of p -variation.

Proof. As before let U be the uniform $U[0, 1]$ distribution function. Let U_m, V_n be independent empirical distribution functions for U . Then as is well known, we can set $F_m = U_m(F)$, $G_n = V_n(G) = V_n(F)$. If X_i are i.i.d. with distribution F , then since F is continuous, $F(X_i)$ are i.i.d. $U[0, 1]$. It follows by these transformations that we can assume $F = G = U$. It's easy to check that

$$(17) \quad (F, G)_{m,n} = \left(\frac{n}{m+n} \right)^{1/2} m^{1/2} (F_m - F) - \left(\frac{m}{m+n} \right)^{1/2} n^{1/2} (G_n - F).$$

A sequence Y_n of possibly non-measurable functions from a probability space into a metric space (S, d) is said to converge *in outer probability* to a random variable Y_0 if for every $\epsilon > 0$, $P^*(d(Y_n, Y_0) > \epsilon) \rightarrow 0$ as $n \rightarrow \infty$, where

$$P^*(A) := \inf\{P(B) : B \text{ measurable}, A \subset B\}.$$

By the Donsker property of $\{f \in W_p(\mathbf{R}) : \|f\|_{[\mathbf{p}]} \leq 1\}$ for $p < 2$, the Love-Young duality for q -variation spaces converse to (5) above (Dudley 1992, Theorem 2.2 and Theorem 3.6), and existence of almost surely convergent realizations, see Dudley (1985, Theorem 4.1), for any $r > 2$, taking $(1/p) + (1/r) = 1$, there exist on some probability space processes α_m having all the properties of $m^{1/2}(F_m - F)$ for each fixed m (but not the same joint distribution for different m) and a Brownian bridge α such that $\|\alpha_m - \alpha\|_{[r]} \rightarrow 0$ in outer probability. By right continuity, the r -variation

is determined by restriction to rational $t \in [0, 1]$, so we have convergence in probability (outer probability is unnecessary here). On another copy of the probability space, let α_n be called β_n and let α be called β . Taking a product space we have (α_m, α) independent of (β_n, β) . Then

$$C_{m,n} := \left(\frac{n}{m+n} \right)^{1/2} \alpha - \left(\frac{m}{m+n} \right)^{1/2} \beta$$

is a Brownian bridge since α and β are independent Brownian bridges and the squares of the coefficients add up to 1. Also, by (17),

$$\gamma_{m,n} := \left(\frac{n}{m+n} \right)^{1/2} \alpha_m - \left(\frac{m}{m+n} \right)^{1/2} \beta_n$$

has all the properties of $(F, G)_{m,n}$. We have

$$\|\gamma_{m,n} - C_{m,n}\|_{[r]} \rightarrow 0$$

in (outer) probability as $m, n \rightarrow \infty$.

Let $J_N := F + (m^{1/2}\alpha_m + n^{1/2}\beta_n)/N$. Then $W_{m,n}^2$ is equal in distribution to $\int \gamma_{m,n}^2 dJ_N$. Now, J_N is an empirical distribution function for F for sample size N (just as K_N is). For any $1 < q < 2$ there is an $r > 2$ such that $(1/q) + (1/r) > 1$. Then $\|\gamma_{m,n}\|_{[r]}$ is bounded in probability by (1), and

$$\|\gamma_{m,n}\|_{[r]} \leq \|\gamma_{m,n}\|_{[r]}^2$$

by Lemma 5.1 of Dudley (1994). We have $J_N = F + (J_N - F)$ where

$$\|J_N - F\|_{[q]} = O_p(N^{(1-q)/q}(LLN)^{1/2}) \text{ as } N \rightarrow \infty$$

by Corollary 3. Thus by the inequality of Young (1936) ((5) above) again,

$$\int \gamma_{m,n}^2 d(J_N - F) = O_p(N^{(1-q)/q}(LLN)^{1/2}) \rightarrow 0$$

in probability as $N \rightarrow \infty$. Now $\|\gamma_{m,n} + C_{m,n}\|_2$ is bounded in probability and

$$\|\gamma_{m,n} - C_{m,n}\|_2 \leq \|\gamma_{m,n} - C_{m,n}\|_\infty \rightarrow 0$$

in probability, so $\int_0^1 \gamma_{m,n}^2 - C_{m,n}^2 dt \rightarrow 0$ in probability. Since $\int_0^1 C_{m,n}^2(t) dt$ has for all m, n the distribution of $\int_0^1 y_t^2 dt$, W^2 has the same limit distribution. \square

For the convergence of the uniform empirical process to the Brownian bridge in r -variation norm for $r > 2$, used in the above proof, Huang (1994, Section 3.7; 1995) gave rates: he showed that from the construction of Komlós, Major & Tusnády (1975) one obtains Brownian bridges $B_{(m)}$ such that for some constant $C(r)$,

$$E\|\alpha_m - B_{(m)}\|_{[r]} \leq C(r)m^{\frac{1}{r}-\frac{1}{2}}$$

where the exponent of m is best possible since for any Brownian bridge (or sample-continuous process) B , almost surely

$$\|\alpha_m - B\|_{[r]} \geq m^{\frac{1}{r}-\frac{1}{2}}.$$

Watson (1962) introduced a statistic similar to Rosenblatt's, suitable for use on the circle since it doesn't depend on the choice of initial angle, namely

$$\begin{aligned} U^2 &:= U_{m,n}^2 := \frac{mn}{m+n} [f(F_m - G_n)^2 dK_N - \{f(F_m - G_n)dK_N\}^2] \\ &= f(F, G)_{m,n}^2 dK_N - (f(F, G)_{m,n} dK_N)^2. \end{aligned}$$

In the same way as above it can be seen that the asymptotic distribution of $U_{m,n}^2$ for $m, n \rightarrow \infty$ arbitrarily, as Persson (1979) and Janson (1984, p 504) showed, and as Watson (1962) had shown under (15), is that of

$$U_\infty^2 := U_{\infty,\infty}^2 := \int_0^1 y_t^2 dt - (\int_0^1 y_t dt)^2.$$

The Brownian bridge has a Fourier series representation

$$y_t = \frac{1}{\pi 2^{1/2}} \sum_{k=1}^{\infty} \frac{1}{k} \{X_k(1 - \cos(2\pi kt)) + Y_k \sin(2\pi kt)\}$$

where $X_1, X_2, \dots, Y_1, Y_2, \dots$, are i.i.d. $N(0, 1)$. Then

$$y_t - \int_0^1 y_s ds = \frac{1}{\pi 2^{1/2}} \sum_{k=1}^{\infty} \frac{1}{k} \{-X_k \cos(2\pi kt) + Y_k \sin(2\pi kt)\}, \text{ so}$$

$$(18) \quad U_\infty^2 = \int_0^1 y_t^2 dt - (\int_0^1 y_t dt)^2 = \frac{1}{4\pi^2} \sum_{k=1}^{\infty} \frac{1}{k^2} (X_k^2 + Y_k^2).$$

Each term $X_k^2 + Y_k^2$ has the exponential distribution with density $\frac{1}{2}e^{-x/2}$ for $x \geq 0$. A convolution of exponential densities with different scale parameters is a linear combination of those densities. It turns out that

$$(19) \quad P(U_\infty^2 > x) = 2 \sum_{i=1}^{\infty} (-1)^{i-1} \exp(-2\pi^2 i^2 x), \quad x \geq 0$$

(Watson 1962, (2)). A curiosity here is that $\pi^{-2} \sup_{0 \leq t \leq 1} y_t^2$ has the same distribution [e.g. Dudley (1993, (12.3.4))], while for the Bahadur-Kiefer remainder $\mathcal{K}_n := U_n^\leftarrow + U_n - 2U$, $n^3 \|\mathcal{K}_n\|_\infty^4 / (\pi(\log n))^2$ has the same asymptotic distribution (Kiefer 1970, Theorem 1), where $\|\mathcal{K}_n\|_\infty$ can also be replaced by $\sup \mathcal{K}_n$ or $-\inf \mathcal{K}_n$. The series in (19) converges rapidly, so only a few terms give a fine approximation unless x is small, while the series in (18) converges slowly.

Acknowledgments: This research was partially supported by National Science Foundation Grants.

I'd like to thank Peter Bickel, Richard Gill, Ron Pyke, Jinghua Qian and Jon Wellner for helpful discussions.

13.7 REFERENCES

- Aki, S. (1981), ‘Asymptotic behavior of functionals of empirical distribution functions for the two-sample problem’, *Annals of the Institute of Statistical Mathematics* **33**, 391–403.
- Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993), *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.
- Appell, J. & Zabrejko, P. P. (1990), *Nonlinear Superposition Operators*, Cambridge University Press.
- Beirlant, J. & Deheuvels, P. (1990), ‘On the approximation of P-P and Q-Q plot processes by Brownian bridges’, *Statistics and Probability Letters* **9**, 241–251.
- Darling, D. A. (1957), ‘The Kolmogorov-Smirnov, Cramér-von Mises tests’, *Annals of Mathematical Statistics* **28**, 823–838.
- Dixon, W. J. (1940), ‘A criterion for testing the hypothesis that two samples are from the same population’, *Annals of Mathematical Statistics* **11**, 199–204.
- Dudley, R. M. (1985), ‘An extended Wichura theorem, definitions of Donsker class, and weighted empirical distributions’, *Springer Lecture Notes in Mathematics* **1153**, 141–178.
- Dudley, R. M. (1992), ‘Fréchet differentiability, p -variation and uniform Donsker classes’, *Annals of Probability* **20**, 1968–1982.
- Dudley, R. M. (1993), *Real Analysis and Probability*, Chapman and Hall, New York. Second printing, corrected.
- Dudley, R. M. (1994), ‘The order of the remainder in derivatives of composition and inverse operators for p -variation norms’, *Annals of Statistics* **22**, 1–20.
- Fernholz, L. T. (1983), *von Mises Calculus for Statistical Functionals*, Vol. 19 of *Lecture Notes in Statistics*, Springer, New York.
- Filippova, A. A. (1961), ‘[the von] mises theorem on the asymptotic behavior of functionals of empirical distribution functions and its statistical applications’, *Theory of Probability and Its Applications* **7**, 24–57.
- Fisz, M. (1960), ‘On a result by M. Rosenblatt concerning the von Mises-Smirnov test’, *Annals of Mathematical Statistics* **31**, 427–429.
- Gill, R. D. (1989), ‘Non- and semi-parametric maximum likelihood estimators and the von Mises method’, *Scandinavian Journal of Statistics* **16**, 97–128.
- Huang, Y.-C. (1994), Empirical distribution function statistics, speed of convergence, and p -variation, PhD thesis, Massachusetts Institute of Technology.
- Huang, Y.-C. (1995), Speed of convergence of classical empirical processes in p -variation norm, preprint, Academica Sinica, Taipei, Taiwan.

- Janson, S. (1984), ‘The asymptotic distributions of incomplete U -statistics’, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **66**, 495–505.
- Kiefer, J. (1959), ‘K-sample analogues of the Kolmogorov-Smirnov and Cramér-v. Mises tests’, *Annals of Mathematical Statistics* **30**, 420–447.
- Kiefer, J. (1970), Deviations between the sample quantile process and the sample df, in M. L. Puri, ed., ‘Nonparametric Techniques in Statistical Inference’, Cambridge University Press, pp. 299–319.
- Komlós, J., Major, P. & Tusnády, G. (1975), ‘An approximation of partial sums of independent RV’s, and the sample DF. I’, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **32**, 111–131.
- Lehmann, E. L. (1951), ‘Consistency and unbiasedness of certain nonparametric tests’, *Annals of Mathematical Statistics* **22**, 165–179.
- Persson, T. (1979), ‘A new way to obtain Watson’s U^2 ’, *Scandinavian Journal of Statistics* **6**, 119–122.
- Randles, R. H. & Wolfe, D. A. (1991), *Introduction to the Theory of Nonparametric Statistics*, Krieger, Malabar, FL. Reprinted with corrections.
- Reeds, J. A. (1976), On the definition of von Mises functionals, PhD thesis, Statistics, Harvard University.
- Rosenblatt, M. (1952), ‘Limit theorems associated with variants of the von Mises statistic’, *Annals of Mathematical Statistics* **23**, 617–623.
- Taylor, S. J. (1972), ‘Exact asymptotic estimates of Brownian path variation’, *Duke Mathematical Journal* **39**, 219–241.
- Watson, G. S. (1962), ‘Goodness-of-fit tests on a circle, II’, *Biometrika* **49**, 57–63.
- Young, L. C. (1936), ‘An inequality of the Hölder type, connected with Stieltjes integration’, *Acta Mathematica (Djursholm)* **67**, 251–282.
- Young, L. C. (1938), ‘General inequalities for Stieltjes integrals and the convergence of Fourier series’, *Mathematische Annalen* **115**, 581–612.

14

A Poisson Fishing Model

Thomas S. Ferguson¹

ABSTRACT A fishing model of Starr, Wardrop, and Woodrooffe is related to the sequential search model of Cozzolino. The latter is generalized to allow an arbitrary joint distribution of capture times and fish sizes. Implications to the foraging models of Oaten and Green and to debugging software are indicated.

14.1 Introduction

The central theme of this paper is a result in sequential analysis that has application to a wide variety of problems. These problems have appeared in papers dealing with sequential estimation in statistics, estimation of the number of species, the fishing problem, the proofreading problem, auditing, foraging, search, etc. Authors in different areas are not always aware of each other, and so often recompute the basic results again. This is partly because the basic assumptions required in the different areas necessarily differ in significant ways. Yet the main result in Section 4 of this paper would be of interest in all these areas. Since the basic assumption of the result is that a certain parameter has a Poisson distribution, it seems appropriate to use the model of Starr, Woodrooffe and others as a fishing problem.

14.2 The Fishing Problem

One of the first papers to deal with the fishing problem was Starr (1974). The main result of this paper is easy to state. There are m fish in a lake, where m is known. The capture time of fish j if one fishes indefinitely is T_j . We assume T_1, \dots, T_m are i.i.d. exponential with mean μ . Let $K(t)$ denote the number of fish caught by time t , so that

$$(1) \quad K(t) = \sum_{j=1}^m I(T_j \leq t),$$

where I denotes the indicator function. There is a constant cost of time and a return of 1 for each fish caught, so the payoff for stopping at time t

¹University of California at Los Angeles

is

$$(2) \quad Y_t = K(t) - ct.$$

The problem is to find a stopping rule τ to maximize EY_τ . Starr shows that the optimal stopping rule is

$$(3) \quad \tau = \inf\{t \geq 0 : K(t) \geq m - c/\mu\}.$$

This is a fixed number of captures rule: Fish until there are at most c/μ fish left.

This was generalized by Starr & Woodroffe (1974). Again there are m fish and the payoff is given by (2), but now the capture times T_1, \dots, T_m , are assumed to be i.i.d. non-negative with absolutely continuous distribution function $F(t)$. There are two basic results of their paper.

The first is as follows. If F has increasing failure rate (IFR), then it is optimal to stop only at catch times. This allows one to discretize the problem and solve it by backward induction. However, the optimal rule will not generally be a fixed number of captures rule; the optimal decision to stop may also depend on time.

In spite of this, the easy case turns out to be the case when F has decreasing failure rate (DFR). In this case, then it may be optimal to stop between capture times, but the infinitesimal look-ahead rule is optimal. The theory for this rule is developed by Ross (1971). Its application to the present problem produces the optimal rule,

$$(4) \quad \tau = \inf\{t \geq 0 : (m - K(t))r_F(t) \leq c\},$$

where $r_F(t)$ is the failure rate of F , namely, $r_F(t) = f(t)/(1 - F(t))$, where $f(t)$ is the density of F .

This result for DFR has been extended by Starr, Wardrop & Woodroffe (1976) to allow the payoff to be of the form

$$(5) \quad Y_t = g(K(t)) - c(t),$$

where g is concave utility function, and c is a concave cost function. The corresponding optimal rule is

$$(6) \quad \tau = \inf\{t \geq 0 : (m - K(t))(g(K(t)) + 1) - g(K(t)))r_F(t) \leq c'(t)\}.$$

The main application of this result, one that motivated generalizing the payoff, is to the statistical problem of estimating the mean of a normal distribution with “delayed” observations. If you start off an experiment with m experimental units and the observations come in sequentially and sporadically, and if you are paying a cost in real time, you might want to stop the experiment early rather than wait for the last observation. If you estimate the mean after $K(t)$ observations, you incur a terminal loss of $\sigma^2/K(t)$. Since $g(k) = -\sigma^2/k$ is concave in k , the rule (6) is optimal.

A further extension is made in Kramer & Starr (1990). In this paper, the fish are allowed to have different sizes, and the time of capture may be

dependent on the size. If we let the size of fish j be denoted by X_j , then the basic assumption of this model is that $(X_1, T_1), \dots, (X_m, T_m)$ are i.i.d. with absolutely continuous distributions with $T_j > 0$ a.s. and $E|X| < \infty$. The payoff for stopping at time t is now the total catch size, $R(t)$, minus a cost of time,

$$(7) \quad Y_t = R(t) - c(t) = \sum_{i=1}^m X_i I(T_i \leq t) - c(t).$$

The infinitesimal look-ahead rule for this problem is

$$(8) \quad \tau = \inf\{t \geq 0 : (m - K(t))E(X | T = t)r_F(t) \leq c'(t)\}$$

where $r_F(t)$ is the failure rate of the marginal distribution of T . Since it is assumed that $c(t)$ is convex, this rule is optimal provided $E(X | T = t)r_F(t)$ is nonincreasing in t . In particular, if F has DFR and if $E(X | T = t)$ is nonincreasing in t , then τ of (8) is optimal. When $E(X | T = t)$ is nonincreasing in t , bigger fish are easier to catch. This is a natural assumption for Kramer and Starr because they are interested in exploration for oil, where the bigger deposits are easier to find. In fact, because of a nice theorem of theirs from an earlier paper, they restrict attention to distributions for which $P(T > t | X = x) = (1 - H(t))^x$ for some distribution function H . The nice theorem states that with appropriate regularity conditions, this is a necessary and sufficient condition that sampling becomes proportional to size. That is, at all times t conditional on the sizes X_1, \dots, X_m of the uncaught fish, the probability that the i th fish is caught next is $X_i/(X_1 + \dots + X_m)$. Sampling proportional to size is a natural assumption for oil exploration. If H has DFR, then Kramer and Starr show that $E(X | T = t)r_F(t)$ is nonincreasing so that (8) is optimal.

14.3 A Search Problem

Let us go back to an earlier version of this problem found in a paper of Cozzolino (1972). This paper is in the area of optimal allocation of search effort initiated by B. O. Koopman in the late 1950's. In the terminology of the fishing problem, the fish are allowed to have different sizes or values that are dependent on their catch times as in Kramer and Starr. However, the number of fish is allowed to be unknown. This is an important generalization since it is rare in applications that m is known exactly. It is useful to express one's uncertainty of m in a prior probability distribution that may then be updated as information is received. Specifically, Cozzolino assumes that the number of fish, M , has a Poisson distribution with parameter λ . Given $M = m$, the sizes of the fish, X_1, \dots, X_m are assumed to be i.i.d. with a gamma distribution, and given M and X_1, \dots, X_m , the capture times are assumed to be independent, with T_i having an exponential

distribution at rate γX_i . Bigger fish are easier to catch. In symbols,

$$(9) \quad \begin{aligned} M &\in \mathcal{P}(\lambda) \\ (X_1, T_1), \dots, (X_M, T_M) | M &\text{ i.i.d.} \\ X_j &\in \mathcal{G}(\alpha, \beta) \\ T_j | X_j &\in \mathcal{G}(1, \gamma X_j) \end{aligned}$$

where $\mathcal{G}(\alpha, \beta)$ is the gamma distribution with density proportional to $\exp(-\beta x)x^{\alpha-1}$. The payoff is the sum of the sizes of the fish caught minus a constant cost per unit time, $Y_t = R(t) - ct$. In addressing the problem of when to stop searching, Cozzolino finds that there is an optimal fixed time rule,

$$(10) \quad \begin{aligned} \tau &= \inf\{t \geq 0 : \lambda E(X | T = t)f(t) \leq c\} \\ &= \frac{\beta}{\gamma} \left[\left(\frac{\lambda\gamma\alpha(\alpha+1)}{c\beta^2} \right)^{1/(\alpha+2)} - 1 \right]^+. \end{aligned}$$

In contrast to the case with a known number of fish when the optimal rule (3) of Starr stops after a fixed number of catches, the optimal rule with a Poisson number of fish stops at a fixed time.

14.4 The Species Problem

Another problem related to this is the species problem. The standard objective in the species problem is to estimate the number of unobserved species or the probability of observing a new species. But our interest is in the stopping problem. When should one abandon the search for new species? This problem was investigated in Rasmussen & Starr (1979) “Optimal and adaptive stopping in the search for a new species”. Their formulation of the problem is a discrete version of the fishing problem outlined above.

Consider an infinite population consisting of m subpopulations, or species. Let p_i denote the proportion of members of the population that belong to species i . Selections are made from the population sequentially at random and the species of the selection is noted. It is assumed that each trial is independent and the probability that the i th species is observed is p_i for each trial. For $i = 1, \dots, m$ and $n = 0, 1, 2, \dots$, let $X_i(n)$ denote the number of times species i is observed among the first n observations. Let $K(n) = \sum_1^m I(X_i(n) > 0)$ denote the number of distinct species observed in the first n trials and let $u(n) = \sum_1^m p_i I(X_i(n) = 0)$ denote the total probability of the unobserved species. The reward for stopping at stage n is $Y_n = g(K(n)) - nc$, where g is a concave function on the integers and $c > 0$ is a constant. The one-stage look-ahead rule (the 1-sla) is optimal

for this problem. It is

$$(11) \quad N = \min \left\{ n \geq 0 : (g(K(n) + 1) - g(K(n))) u(n) \leq c \right\}.$$

The trouble with this analysis for the species problem is that m and the p_i are assumed known, and which species is associated with p_i is also known. However, since the rule N depends on this knowledge only through $u(n)$, Rasmussen and Starr suggest using Turing's nonparametric estimate of this quantity in (11). This estimate is

$$(12) \quad v(n) = \frac{1}{n} \sum_1^m I(X_i(n) = 1).$$

Numerical computation shows that the adaptive strategy (11) with $u(n)$ replaced by $v(n)$ for stopping compares well with the original rule.

I would like to suggest another way to deal with this problem. Namely, the Bayes solution in which the prior distribution of (p_1, \dots, p_m) is taken to be the Dirichlet distribution $D(\alpha_1, \dots, \alpha_m)$. One good feature about this approach is that the one-stage look-ahead rule is still optimal for this adaptive problem. It is (11) with $u(n)$ replaced by the Bayes estimate,

$$(13) \quad w(n) = \frac{\sum_1^m \alpha_i I(X_i(n) = 0)}{n + \sum_1^m \alpha_i}.$$

The drawback of presuming to know which species are associated with which α_i remains, but one may take all α_i equal and estimate parameters from the data, thus providing an adaptive Bayes solution. Alternatively, one may assume the α_i decrease exponentially, (m may even be taken to be infinite).

Banerjee & Sinha (1985) extend Rasmussen and Starr to sampling in batches of size $k > 1$. They also propose a new estimator of the probability of discovering a new species.

Alsmeyer & Irle (1989) put the problem in continuous time, allow stochastic intensities depending on the past, $\lambda_i(t)$ for species i , and allow the reward, r_i to depend on the species, i . As an example, they take constant intensities, and find as the optimal stopping rule, similar to (11),

$$(14) \quad \tau = \inf \left\{ t \geq 0 : \sum_1^m r_i \lambda_i I(X_i(t) = 0) \leq c \right\}.$$

It is assumed that it is known which species are attached to which intensities and rewards. So the fishing model is more appropriate than the species model for this problem.

14.5 The Basic Fishing Model

We generalize Cozzolino's formulation as follows. There are a random

number, M , of fish. The distribution of M is known and $EM < \infty$. Given $M = m$, the sizes and times of capture, $(X_1, T_1), \dots, (X_m, T_m)$, are i.i.d. 2-dimensional random vectors with $E|X_i| < \infty$ and $T_i > 0$ a.s., with known distribution function, $F(x, t)$. As before, we let for fixed t ,

$$(15) \quad \begin{aligned} K(t) &= \sum_{i=1}^M I(T_i \leq t) = \# \text{ fish caught by time } t \\ R(t) &= \sum_{i=1}^M X_i I(T_i \leq t) = \text{total value of fish caught} \end{aligned}$$

The payoff we receive if we stop at time t is $Y_t = R(t) - c(t)$, where $c(t)$ is a given increasing function of t . An optimal (finite-valued) stopping time exists if $E \sup_t Y_t < \infty$. This is true under the assumptions that $EM < \infty$ and $EX^+ < \infty$. The infinitesimal look-ahead rule is

$$(16) \quad \tau = \inf\{t \geq 0 : E(M - K(t) | \mathcal{F}_t) E(X | T = t) r_F(t) \leq c'(t)\},$$

where \mathcal{F}_t denotes the σ -field generated by the observations up to time t . Moreover, $E(M - K(t) | \mathcal{F}_t)$ depends only on $K(t)$. In our model, this rule is optimal if the problem is monotone. The problem is monotone if the validity of the inequality in (16) at $t = t_0$ implies its validity a.s. for all $t > t_0$. For monotonicity, it suffices to have

- (i) $c'(t)$ to be non-decreasing (i.e. c convex.)
- (ii) $E(M - K(t) | \mathcal{F}_t)$ to be non-increasing in t a.s.
- (iii) $E(X | T = t) r_F(t)$ to be non-increasing in t .

The first condition is standard. It was used by Starr, Wardrop and Woodrooffe. The third condition breaks into two conditions. It is satisfied if both $E(X | T = t)$ is non-increasing (bigger fish are easier to catch) and $r_F(t)$ is nonincreasing (T has DFR). However, it is easy to see that it can be satisfied more generally.

The second condition is the critical one. Let us consider some special cases.

- (a) M degenerate at m . Then $E(M - K(t) | \mathcal{F}_t) = m - K(t)$ a.s., which is non-increasing a.s. so that condition (ii) is satisfied.
- (b) M Poisson, $P(\lambda)$. Then $(M - K(t)) | \mathcal{F}_t$ has the Poisson distribution, $P(\lambda P(T > t))$, so that $E(M - K(t) | \mathcal{F}_t) = \lambda P(T > t)$. This is a non-random function, non-increasing in t , so again condition (ii) is satisfied.

In this case, something more interesting is true. Namely, there is an optimal fixed time rule whether or not the problem is monotone, whether or not bigger fish are easier to catch. This is because at time t , the future is independent of the past. Thus, in Cozzolino's problem, there is an optimal fixed time rule whether or not the assumptions on T and X are satisfied. This means that the optimal rule may be found as a simple maximization problem, namely, find t to maximize

$$EY_t = E \sum_1^M X_i I(T_i \leq t) - c(t) = \lambda E(XI(T \leq t)) - c(t).$$

The derivative with respect to t is

$$(17) \quad \frac{d}{dt} EY_t = \lambda E(X | T = t)f(t) - c'(t).$$

There exists a unique root of this expression if and only if the problem is monotone. In this case the optimal rule reduces to

$$(18) \quad \tau = \inf\{t \geq 0 : \lambda E(X | T = t)f(t) \leq c'(t)\}.$$

If the problem is not monotone, we must inspect each of the negative-going roots of (17) to find the value.

(c) *M has the binomial distribution, $\mathcal{B}(W, \pi)$.* Then $M - K(t) | \mathcal{F}_t$ has the binomial distribution, $\mathcal{B}(W - K(t), \pi(t))$, where $\pi(t) = \pi P(T > t) / [1 - \pi + \pi P(T > t)]$. Hence, $E(M - K(t) | \mathcal{F}_t) = (W - K(t))\pi(t)$, and the rule (16) becomes

$$(19) \quad \tau = \inf\{t \geq 0 : (W - K(t))\pi E(X | T = t)f(t) \leq c'(t)[1 - \pi + \pi P(T > t)]\}.$$

(d) *M has the negative binomial distribution, $\mathcal{NB}(\alpha, 1/(\beta+1))$.* This distribution arises when $M | \lambda$ has the Poisson distribution, $\mathcal{P}(\lambda)$ and λ has the gamma distribution, $\mathcal{G}(\alpha, \beta)$. Then $M - K(t) | \mathcal{F}_t \in \mathcal{NB}(K(t) + \alpha, P(T > t) / (\beta + 1))$. We have $E(M - K(t) | \mathcal{F}_t) = (K(t) + \alpha)P(T > t) / (\beta + P(T \leq t))$. Although $P(T > t) / (\beta + P(T \leq t))$ is non-increasing, the other term $K(t) + \alpha$ increases in jumps at the time of each observation. Thus the problem is not monotone. The infinitesimal look-ahead rule can be improved.

(e) *M is beta-binomial, $\mathcal{BB}(W, \alpha, \beta)$.* This distribution arises when $M | \pi$ has the binomial distribution, $\mathcal{B}(W, \pi)$ and π has the beta distribution, $\mathcal{Be}(\alpha, \beta)$. To make the family of distributions closed under prior-to-posterior analysis, it is necessary to add another parameter. The four-parameter beta-binomial $\mathcal{BB}(W, \alpha, \beta, q)$ with $W \geq 0$ integer, $\alpha > 0$, $\beta > 0$, and $q > 0$, is defined as the distribution with mass function proportional to

$$f(m | W, \alpha, \beta, q) \propto \binom{W}{m} B(\alpha + m, \beta + W - m) q^m,$$

where $B(\alpha, \beta)$ represents the beta function. When $q = 1$ this is the beta-binomial distribution, and when $\beta = 1$ and $0 < q < 1$, this is the negative binomial distribution truncated at W . If T has a continuous distribution, this problem is never monotone.

14.6 Proofreading and Testing Computer Software

In proofreading and in testing computer software, the problem is usually to estimate the number of errors remaining after a debugging process.

M represents the number of misprints or bugs in the program. I'll mention two papers. In the software debugging paper of Dalal & Mallows (1988), the model is as follows. M has a negative binomial prior distribution, $M \in \mathcal{NB}(\alpha, 1/(1 + \beta))$, all bugs are equally valuable to detect, and the times of detection are i.i.d. This is an important problem and Dalal and Mallows suggest a method of solving it in a fairly general setting. When M is Poisson with known mean λ , they note that the optimal rule in their setting is a fixed time rule. When λ is large and there is a large number of observations taken before stopping, one may obtain an adaptive approximately optimal solution by replacing λ in the optimal rule by its estimate, $K(t)/P(T \leq t)$.

In the proofreading paper of Ferguson & Hardwick (1989), the basic model is followed but the setting is discrete and the marginal distribution of T is taken to be a mixture of geometrics. If $P(T = t) = E(Q^t(1 - Q))$, then $E(X | T = t)f(t) = E(Q^t(1 - Q)X)$ is decreasing in t , so that in the Poisson case, the 1-sla optimal in general. In the beta-binomial model, the distribution of catch time is discrete, and the 1-sla is optimal in some important cases.

14.7 Foraging

Consider an animal that forages for food in spacially separated patches of prey. He feeds at one patch for awhile and then moves on to another. The problem of when to move to a new patch in order to maximize the rate of energy intake is addressed in the paper of Oaten (1977). His results have been extended in a number of ways by Green (1980, 1987).

For example, take the fisherman who moves from waterhole to waterhole catching fish. Here, M represents the number of fish in a waterhole. Given $M = m$, the distribution of the times of catch T_1, \dots, T_m is known. The expected time to travel from one patch to another is also a known constant, say 1, as is the expected energy required for the trip, a . The problem Oaten and Green consider is to choose a stopping time τ to move to the next waterhole in order to maximize the rate of return, $(EK(\tau) - a)/(E\tau + 1)$.

Green (1987) treats discrete time and considers six cases, when the distribution of M is \{degenerate, Poisson, negative binomial\}, and when the distribution of catch times is \{uniform, exponential\}. The reason for considering the uniform case is interesting. Green is very much attuned to applications, and he is especially interested in birds. It seems that certain birds are systematic foragers, that is they are active in their search for prey and avoid covering the same ground twice. For systematic foragers, the time needed to catch a given prey would be approximately uniform over the time needed to cover the whole patch.

With the observations we have about the Poisson fishing model, we may extend the model of Oaten/Green to allow size of catch to depend on

time. Therefore, our model is given by M with a known distribution, $(X_1, T_1), \dots, (X_M, T_M) | M$ i.i.d. with a known distribution independent of M . We are to maximize $(ER(\tau) - a)/(E\tau + 1)$. This problem may be related to the problem of finding a stopping rule to maximize the return $E(R(\tau) - a - \phi\tau - \phi)$ and then to adjust ϕ so that the optimal return is zero. The resulting ϕ is the optimal rate of return and the optimal rules for the two problems are the same.

In the Poisson case, since the optimal rule is a fixed time rule, we need only compute

$$E(R(t) - a - \phi t - \phi) = EM E(XI(T \leq t)) - a - \phi t - \phi$$

Setting this to zero gives one equation, and setting the derivative to zero gives a second equation, to be solved jointly for t and ϕ . Eliminating ϕ from these two equations gives

$$(20) \quad EM E(XI(T \leq t)) = a + (t + 1)EM E(X | T = t)f(t).$$

As an example, suppose T has the inverse power distribution with density, $f(t) = \theta(1+t)^{-(\theta+1)}$, and suppose that the expectation of X given $T = t$ is $E(X | T = t) = \alpha(1+t)^\gamma$ where $\gamma < \theta$. (This arises, for example, when the distribution of Z given $T = t$ is the gamma, $G(\alpha, (1+t)^{-\gamma})$.) If $\gamma < 0$, bigger fish are easier to catch, and if $\gamma > 0$, smaller fish are easier to catch. In this example, (20) can be solved explicitly for t to give

$$t = (A^{1/(\theta-\gamma)} - 1)^+$$

as the optimal time to stop, where

$$A = \frac{(\theta - \gamma + 1)\lambda\alpha\theta}{\lambda\alpha\theta - a(\theta - \gamma)}.$$

Acknowledgments: Talk given at the XXIV^{es} Journées de Statistique, Brussels, May 20, 1992.

14.8 REFERENCES

- Alsmeyer, G. & Irle, A. (1989), 'Optimal strategies for discovering new species in continuous time', *Journal of Applied Probability* **26**, 695–706.
- Banerjee, P. K. & Sinha, B. K. (1985), 'Optimal and adaptive strategies in discovering new species', *Sequential Analysis* **4**, 111–122.
- Cozzolino, J. M. (1972), 'Sequential search for an unknown number of objects of nonuniform size', *Operations Research* **20**, 293–308.
- Dalal, S. R. & Mallows, C. (1988), 'When should one stop testing software?', *Journal of the American Statistical Association* **83**, 772–779.

- Ferguson, T. S. & Hardwick, J. P. (1989), 'Stopping rules for proofreading', *Journal of Applied Probability* **26**, 304–313.
- Green, R. F. (1980), 'Bayesian birds: A simple example of Oaten's stochastic model of optimal foraging', *Theoretical Population Biology* **18**, 244–256.
- Green, R. F. (1987), Stochastic models of optimal foraging, in A. C. Kamil, J. R. Krebs & H. R. Pullman, eds, 'Foraging Behavior', Plenum Press, New York, pp. 273–302.
- Kramer & Starr, N. (1990), 'Optimal stopping in a size dependent search', *Sequential Analysis* **9**, 59–80.
- Oaten, A. (1977), 'Optimal foraging in patches: A case for stochasticity', *Theoretical Population Biology* **12**, 263–285.
- Rasmussen, S. & Starr, N. (1979), 'Optimal and adaptive search for a new species', *Journal of the American Statistical Association* **74**, 661–667.
- Ross, S. M. (1971), 'Infinitesimal look-ahead stopping rules', *Annals of Mathematical Statistics* **42**, 297–303.
- Starr, N. (1974), 'Optimal and adaptive stopping based on capture times', *Journal of Applied Probability* **11**, 294–301.
- Starr, N. & Woodroffe, M. (1974), 'Gone fishin': Optimal stopping based on catch times, Technical Report 33, Department of Statistics, University of Michigan.
- Starr, N., Wardrop, R. & Woodroffe, M. (1976), 'Estimating a mean from delayed observations', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **35**, 103–113.

15

Lower Bounds for Function Estimation

Catherine Huber¹

15.1 Introduction

Over the last thirty years, it has been recognized that a common abstract framework underlies many basic problems of nonparametric estimation. In that framework, f is an unknown function to be estimated, known to belong to a class \mathcal{F} of smooth functions, and an observation X is available in order to perform the estimation. The X has its values in a measurable space (E, \mathcal{B}) and obeys a probability law indexed by f and denoted P_f . The set of all probabilities P_g where g varies in \mathcal{F} is denoted \mathcal{P} .

This framework applies equally where one is interested in recovering a probability density, or the spectral density of a Gaussian process, or the intensity of a Poisson process, or the hazard rate of a positive random variable. One of the simplest cases is the problem of *nonparametric density estimation*, where f is a probability density with respect to Lebesgue measure λ on \mathbb{R} , and X is an n -sample, $X = (X_1, X_2, \dots, X_n)$, whose component X_i 's are iid P_f , where $dP_f = f \cdot d\lambda$. The smoothness condition that defines \mathcal{F} is generally either a Lipschitz type condition, as in Ibragimov & Has'minskii (1981), or a Sobolev one, as in Bretagnolle & Huber (1979).

Let f be a real-valued function defined on (an interval I of) \mathbb{R} . Then f is said to obey a *Lipschitz condition* if there exists $L > 0$, and an integer $s \geq 0$ and an α in $(0, 1]$ such that $f \in \Sigma(s + \alpha, L)$, where

$$\Sigma(s + \alpha, L) = \{g : |g^{(s)}(x_2) - g^{(s)}(x_1)| \leq L|x_2 - x_1|^\alpha \text{ all } x_1, x_2 \in I\} \quad (1)$$

For Sobolev-type smoothness, let $s \geq 1$ be an integer, $r > 0$ a positive real, and, for $p \in [1, \infty]$ define the L_p norm $\|f\|_p$ on \mathcal{F} by $\|f\|_p^p = \int f^p d\lambda$. Then f is said to obey a *Sobolev condition* $W_p^s(r)$ for an $r > 0$ if $\|f^{(s)}\|_p^p \leq r$.

An *estimator* \hat{f} of f is any measurable function from (E, \mathcal{B}) into $(\mathcal{G}, \mathcal{B}_G)$, with $\mathcal{F} \subset \mathcal{G}$. Let $\hat{\mathcal{F}}$ be the set of all possible estimates for f . (The possibility that estimates may belong to \mathcal{G} but not to \mathcal{F} comes from the fact that when one restricts attention to a concrete setting, for example with kernel

¹Université Paris V

estimates for a density function, it may happen that the best estimate in some sense does not have its values in \mathcal{F} but in some \mathcal{G} strictly greater than \mathcal{F} . Parzen's kernel estimates, which achieve the optimal rates for certain classes of densities (Wahba 1975), are functions which are not positive everywhere, despite the fact that the estimands are, of course, positive).

In order to compare performances of estimators in $\hat{\mathcal{F}}$, a risk R has to be defined for each estimate \hat{f} at each $f \in \mathcal{F}$. For this, we use a discrepancy $D : \mathcal{G} \times \mathcal{G} \rightarrow \mathbb{R}^+$ which vanishes on the diagonal. The function D may be a distance, such as $\|f - g\|_p$, or a nondecreasing function of a distance, such as $\ell(\|f - g\|^p)$, where ℓ is nondecreasing and zero at zero. The risk R is then thus defined as $R(\hat{f}, f) = E_{P_f} D(\hat{f}, f)$. Such an expectation will generally be denoted by $E_f D(\hat{f}, f)$ in what follows. We are particularly interested in estimators \tilde{f} in $\hat{\mathcal{F}}$ which minimize the maximum risk, $\sup_{f \in \mathcal{F}} R(\tilde{f}, f)$: the \tilde{f} is said to be minimax, or optimal in the minimax sense, if

$$\sup_{f \in \mathcal{F}} R(\tilde{f}, f) = \inf_{\hat{f} \in \hat{\mathcal{F}}} \sup_{f \in \mathcal{F}} R(\hat{f}, f).$$

We will speak of the “problem (\mathcal{F}, D) ” as the problem of determining the structure of minimax estimators and the size of the minimax risk for a given choice of \mathcal{F} and D . Solving the problem exactly and non-asymptotically is generally beyond reach. Instead one hopes merely to obtain various sorts of asymptotic information.

15.1.1 OPTIMAL RATES OF CONVERGENCE

The last 20 years have seen the resolution of various questions about the optimal rate of convergence of the minimax risk to zero.

Progress in this direction began by restricting attention to special classes of estimators $\hat{\mathcal{C}} \subset \hat{\mathcal{F}}$. Many papers were written for such special classes. Examples include kernel estimators (Nadaraya 1974, Rosenblatt 1971), orthogonal series estimators (Watson 1969), and maximum likelihood estimators (Weiss & Wolfowitz 1967). See also Wahba (1975). Such results exhibited the existence of a “best known rate of convergence”—the best performance of any known method. More specifically, in the problem of a density f from an n -sample $X = (X_1, X_2, \dots, X_n)$ with law $(f \cdot d\lambda)^{\otimes n}$, typical results were the following. For a given class $\hat{\mathcal{C}}$ of estimators, researchers determined the best possible exponent ρ in an upper bound of the form:

$$\inf_{\hat{f} \in \hat{\mathcal{C}}} \sup_{f \in \mathcal{F}} R(\hat{f}, f) \leq d \cdot n^{-\rho}. \quad (2)$$

The resulting exponent $\rho = \rho(\mathcal{F}, D, \hat{\mathcal{C}})$ depends on \mathcal{F} , or rather on the smoothness condition defining \mathcal{F} , and on $\hat{\mathcal{C}}$, the subclass of estimates under consideration. The exponent ρ is called the *minimax rate of convergence* of the estimates from $\hat{\mathcal{C}}$ for the estimation problem (\mathcal{F}, D) .

Several classes of estimates \hat{C} had been studied by the mid-1970's, for several types of smoothness assumptions, and a remarkable pattern emerged. For a fixed (\mathcal{F}, D) problem, it was frequently found that many classes \hat{C} had exactly the same value ρ . For example, if the density were assumed to lie in the class $\mathcal{F}_{2,2}$ of densities with Sobolev smoothness of type $s = 2$, $p = 2$, and if D is the squared L^2 loss, then appropriate kernel estimates could achieve $\rho(\mathcal{F}, D, \hat{C}_{Ker}) = 4/5$; so could orthogonal series methods: $\rho(\mathcal{F}, D, \hat{C}_{OS}) = 4/5$, etc. On the other hand, with histogram estimates one has $\rho(\mathcal{F}, D, \hat{C}_{Hist}) = 2/3$. Hence, the class of histogram estimators can be outperformed by other classes, and there exist several classes of estimators which all achieve the same rate $4/5$, the best known (at that time) rate behavior over $(\mathcal{F}_{2,2}, D)$.

This type of numerical coincidence led, in the 1970's, to the idea that one might be able to establish an *optimal rate of convergence among all estimators* for a given problem (\mathcal{F}, D) , and one might be able to thereby identify classes of estimators whose minimax rates of convergence were optimal. Such classes would normally be preferable to other classes of estimators whose minimax rates were sub-optimal.

To carry out this program, one could attempt to establish an exponent $\rho^* = \rho^*(\mathcal{F}, D)$ which applied to all estimates, in the sense that

$$\inf_{\hat{f} \in \hat{\mathcal{F}}} \sup_{f \in \mathcal{F}} R(\hat{f}, f) \geq c \cdot n^{-\rho^*}. \quad (3)$$

Then one would be able to identify "preferred" classes of estimates as those achieving the rate $\rho(\mathcal{F}, D, \hat{C}) = \rho^*(\mathcal{F}, D)$. An exciting period of research in the mid 70's through early 80's led to a number of results of exactly this type. I feel fortunate to have participated in this effort, with my colleague Jean Bretagnolle. Our work, and that of others such as Ibragimov & Has'minskii and Stone led to a great consolidation of our understanding of how to select good estimators. For example, we now know that for L^2 -Sobolev smoothness of order $s = 2$ and D the squared L^2 loss, the optimal rate $\rho^*(\mathcal{F}_{2,2}) = 4/5$.

15.1.2 LOWER BOUNDS

A key role played in this consolidation of understanding was the lower bound (3). In this expository paper, I would like to recount the key ideas which were found useful in establishing such lower bound results; while these ideas are nowadays often used in proofs of optimal rates, they are seldom discussed in expository fashion as I do here.

In order to get lower bounds on the minimax risk, the usual device is to build inside \mathcal{F} , for every n , a finite subproblem $\mathcal{F}_{0,n}$ such that $(\mathcal{F}_{0,n}, D)$ is essentially as difficult as the full problem (\mathcal{F}, D) . Two specific constructions of such finite subproblems have been found frequently useful: the hypercube and the pyramid. These constructions are generally chosen because good

bounds are known for the risk over these finite sets (due to Assouad and Fano, respectively) and because they can be easily deployed to give bounds like (3). Accordingly, we often speak of *Assouad's hypercube* (Assouad 1983) and *Fano's pyramid* (Ibragimov & Has'minskii 1981), honoring the key role of certain risk evaluations in motivating the use of these special finite sets.

These risk evaluations have the following key features.

- (i) An observation X is better for testing between two simple hypotheses $H_0 : X \sim P_f$ and $H_1 : X \sim P_g$ when the laws P_f and P_g are “more separated”. The Assouad and Fano approaches quantify the discriminating ability using two specific measures of separation for probability measures.

- Kullback-Leibler information $K(P, Q) = E_p \log dP/dQ$;
- Hellinger distance h defined by: $2h^2(P, Q) = \int (\sqrt{dP} - \sqrt{dQ})^2$.

Note that h is symmetric, $h(P, Q) = h(Q, P)$, but K is not.

- (ii) Using Assouad (respectively Fano), one gets that if a certain finite set—a hypercube (respectively pyramid)—has a large number of elements that are not too “separated” (in either Kullback or Hellinger sense), and if the finite set is sufficiently “massive”, (enough elements), then the estimation problem is difficult in the sense that the minimax risk is large for a certain special loss functions.
- (iii) Hence we can build good lower bounds by constructing hypercubes (or pyramids) consisting of very many vertices, which are not too separated, and for which the special loss function can be directly related to a loss function D in the original problem.

In this paper, we will describe the basic lower bounds over hypercubes and pyramids, and we will construct (or prove the existence of) hypercubes and pyramids for a simple case of density estimation—optimality rates for Sobolev classes.

Upper bounds of type (2) for the L^2 Sobolev problem

$$\mathcal{F}_{s,2} = \{f : \|f^{(s)}\|_2^2 \leq r\}, \quad D(f, g) = \|f - g\|_2^2 \quad (4)$$

have been known for a long time. For example, kernel estimates $\hat{f}_n = K_n * \hat{\mu}_n$ work well, where $\hat{\mu}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ is the empirical measure of the n -sample X , and $K_n(x) = K(x/h_n)/h_n$ is based on a Parzen kernel K of order s , a bounded, continuous, even function such that $\int K(x) dx = 1$, $\int x^j K(x) dx = 0$ for $1 \leq j < s$, and $\int |x|^s K(x) dx < \infty$. It is known that for a bandwidth h_n of order $n^{-1/(2s+1)}$ such kernels achieve the best known rate of convergence for the family $\mathcal{F}_{s,2}$ of densities, so that (2) holds with rate $\rho(\mathcal{F}_{s,2}, D, C_{Ker}) = 2s/(2s+1)$. This rate of convergence will be shown to be the best possible among all estimates as an example of the derivation of (3) by both Assouad's hypercube and Fano's pyramid.

15.2 Assouad's Hypercube

We assume independent observations $X = (X_1, \dots, X_n)$, each with law $dP_f = f \cdot d\lambda$. The unknown density, f , is assumed to lie inside a smooth set of densities $\mathcal{F} = W_p^s(r)$ for some $r > 0$ and $p \in [1, \infty]$.

The discrepancy between two points in \mathcal{F} is measured by $D : \mathcal{G} \times \mathcal{G} \mapsto \mathbb{R}^+$, satisfying the two following conditions:

- (i) Quasi-triangle inequality: $D(f, g) \leq B[D(f, h) + D(h, g)]$ for $f, g, h \in \mathcal{G}$, for a fixed positive constant B ;
- (ii) Superadditivity: $D(f, g) \geq \sum_{j \in J} D(f1_{I_j}, g1_{I_j})$, for each partition $(I_j)_{j \in J}$ of \mathbb{R} , where $f1_{I_j}$ is the restriction of f to the set I_j .

Several types of discrepancy meet these conditions. First, the p th power discrepancy $D(f, g) = \|f - g\|_p^p$ with $p \geq 1$. For $p = 2$ the constant B of the quasi-triangle inequality is equal to 2, and the superadditivity reduces to additivity. Second, the p th power discrepancy on derivatives: $D(f, g) = \|f^{(s')} - g^{(s')}\|_p^p$ where s' is an integer in $[1, s]$. Third, $D(f, g) = \|f^{1/q} - g^{1/q}\|_p^p$ with $(1/p) + (1/q) = 1$, which equals the square of the Hellinger distance when $p = q = 2$.

In this section, we will derive a lower bound for the minimax risk for $(\mathcal{F}_{s,2}, D)$, where D is a discrepancy of the type just described. We proceed by considering a sequence of both abstract and concrete problems (\mathcal{F}, D) ; \mathcal{F} will be successively a two point set, then a hypercube, and finally a set with infinite metric dimension containing the cube.

15.2.1 BASIC INEQUALITY: SEPARATING TWO POINTS

Lemma 1 *Let $(E, \mathcal{B}, (\mathcal{P}, \mathcal{Q}))$ be a statistical experiment with $h^2(P, Q) \leq 1/2$, and let U and V be two positive random variables defined on (E, \mathcal{B}) such that $U + V \geq \Delta > 0$. Then $E_P U + E_Q V \geq \frac{\Delta}{2} \exp(-4h^2(P, Q))$.*

Proof.

$$\int (U dP + V dQ) \geq \int (U + V)(dP \wedge dQ) \geq \Delta \int dP \wedge dQ,$$

where $(a \wedge b) = \min(a, b)$ and $(a \vee b) = \max(a, b)$. Schwarz's inequality applied to $\int \sqrt{dP dQ} = \int \sqrt{dP \wedge dQ} \sqrt{dP \vee dQ}$ shows that

$$2 \int (dP \wedge dQ) \geq \left(\int \sqrt{dP dQ} \right)^2,$$

as $\int (dP \vee dQ) \leq 2$. Moreover,

$$\left(\int \sqrt{dP dQ} \right)^2 = [1 - h^2(P, Q)]^2 \geq \exp(-4h^2(P, Q)),$$

as $\log(1 - u) + 2u \geq 0$ for $u \in (0, 1/2]$, and $h^2 \leq 1/2$. \square

Remark The function $\Phi(t) = -2 \log t - 1 + t$ is positive and convex on $[0, 1]$. By Jensen's inequality,

$$K(P, Q) - h^2(P, Q) = \int \Phi\left(\sqrt{\frac{dQ}{dP}}\right) dP \geq \Phi\left(\int \sqrt{dP dQ}\right) \geq 0.$$

Thus $E_P U + E_Q V \geq \frac{\Delta}{2} \exp(-4K(P, Q))$, an inequality of Assouad (1983) that is slightly weaker than the assertion of Lemma 1, sometimes strictly so, but is sometimes more tractable.

Remark Cartesian products are easy to deal with using Kullback distance because $K(P^{\otimes n}, Q^{\otimes n}) = nK(P, Q)$. Similarly, for Hellinger distance, $h^2(P^{\otimes n}, Q^{\otimes n}) \leq nh^2(P, Q) \wedge 1$, as long as this distance is small enough, which will always be the case in our context.

Remark If we restrict attention to a partition of E , both h^2 and K are additive.

15.2.2 TWO-POINT BOUNDS ON MINIMAX RISK.

Suppose \mathcal{F} is two-element set $\mathcal{F}_0 = \{\theta_1, \theta_2\}$. Lemma 1 gives immediately a bound on minimax risk for (\mathcal{F}_0, D) . Let P_{θ_i} be the law of X when $\theta = \theta_i$, let $\Delta = D(\theta_1, \theta_2)$ be the discrepancy between θ_1 and θ_2 , and let $\delta = h^2(P_{\theta_1}, P_{\theta_2})$ be the distance between the P_{θ_i} . Then

$$\max_{\theta \in \{\theta_1, \theta_2\}} R(\hat{\theta}, \theta) \geq \frac{\Delta}{2B} \exp(-4\delta) \quad \text{for every estimate } \hat{\theta}.$$

This can be seen by the following argument. Let $U = D(\hat{\theta}, \theta_1)$, and $V = D(\hat{\theta}, \theta_2)$. Then $U + V \geq \Delta/B$ by the quasi-triangle inequality for D , and

$$\max_{\theta \in \{\theta_1, \theta_2\}} E_{P_\theta} D(\hat{\theta}, \theta) \geq \frac{1}{2}(E_{P_{\theta_1}} U + E_{P_{\theta_2}} V) \geq \frac{\Delta}{2B} \exp(-4h^2(P_{\theta_1}, P_{\theta_2})),$$

as asserted.

15.2.3 INEQUALITY ON THE CUBE

Suppose now that \mathcal{F} is a finite set \mathcal{F}_0 having 2^N elements, indexed by the cube $C = \{-1, +1\}^N$, whose generic element is $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_N)$. With f_ϵ an element of \mathcal{F}_0 , P_{f_ϵ} defines a corresponding element in a set of probabilities \mathcal{P}_0 . We say that \mathcal{F}_0 and \mathcal{P}_0 are two *hypocubes in correspondence* with each other when they have the following metric properties:

$$D(f_\epsilon, f_{\epsilon'}) = \Delta \sum_j |\epsilon_j - \epsilon'_j|, \quad h^2(P_{f_\epsilon}, P_{f_{\epsilon'}}) = \delta \sum_j |\epsilon_j - \epsilon'_j|, \quad (5)$$

for some $\Delta > 0$ and $\delta > 0$. Then the (\mathcal{F}_0, D) hypercube of side 2Δ corresponds to the (P_0, h^2) hypercube of side 2δ . The following result is similar to a lemma of P. Assouad, but works with Hellinger distance instead of Assouad's original choice.

Lemma 2 *Let $\hat{\mathcal{F}}_0$ be the collection of estimators taking values in \mathcal{F}_0 . Then*

$$\inf_{\hat{f} \in \hat{\mathcal{F}}_0} \sup_{f \in \mathcal{F}_0} R(\hat{f}, f) \geq \frac{N\Delta}{2} \exp(-8\delta).$$

Proof: Let \hat{f} be any element in $\hat{\mathcal{F}}_0$, and let $\epsilon = (\epsilon^j, \epsilon_j)$ where $\epsilon^j = (\epsilon_i)_{i \neq j}$. Let ν be the uniform law over \mathcal{F}_0 , and E_ϵ be the expectation with respect to P_ϵ . Then the Bayes risk $E(\hat{f}, \nu)$ of \hat{f} with respect to ν equals

$$\begin{aligned} E_\nu E_\epsilon D(\hat{f}, f_\epsilon) &= 2^{-N} \sum_{\epsilon} E_\epsilon \left(\sum_j \Delta |\hat{\epsilon}_j - \epsilon_j| \right) \\ &= \Delta 2^{-N} \sum_j \sum_{\epsilon^j} \left(\sum_{\epsilon_j} E_{(\epsilon^j, \epsilon_j)} |\hat{\epsilon}_j - \epsilon_j| \right). \end{aligned}$$

The last summand is of the form $E_P U + E_Q V$ for $P = P_{(\epsilon^j, 1)}$ and $Q = P_{(\epsilon^j, -1)}$, $U = |\hat{\epsilon}_j - 1|$ and $V = |\hat{\epsilon}_j + 1|$. As $\hat{\epsilon}_j$ is either equal to -1 or $+1$, $U + V \geq 2$. Moreover, by the assumption (5), $h^2(P, Q) = 2\delta < 1/2$. The lemma thus follows as $|\sum_j 1| = N$ and $|\sum_{\epsilon^j} 1| = 2^{N-1}$. \square

15.2.4 RISK BOUND IN THE ORIGINAL PROBLEM

Lemma 3 *Let (\mathcal{F}, D) be an abstract estimation problem. If (\mathcal{F}, D) contains a hypercube (\mathcal{F}_0, D) in correspondence with a cube (P_0, h^2) as defined in (5), with respective sidelengths 2Δ and 2δ , then*

$$\inf_{\hat{f} \in \hat{\mathcal{F}}} \sup_{f \in \mathcal{F}} R(\hat{f}, f) \geq \frac{N\Delta}{4B} \exp(-8\delta). \quad (6)$$

This follows from Lemma 2 with respect to the minimax risk on $\mathcal{F}_0 \subset \mathcal{F}$; an extra factor $2/B$ comes in when we extend those results to the case of estimates with values outside \mathcal{F}_0 .

15.2.5 APPLICATION: SOBOLEV CLASSES

We now get a lower bound for the minimax risk in (4). For this discrepancy, the constant B in the triangle inequality is 2. Let us construct inside \mathcal{F} a cube $\mathcal{F}_0 = \left\{ g_0 \left[1 + \sum_{j=1}^N \epsilon_j g_j \right] : \epsilon \in \{-1, +1\}^N \right\}$, where g_0 is a smooth density, with support in $[-1, +1]$, equal to 1 on $[-1/2, 1/2]$. (Compare with

Bretagnolle & Huber 1979.) Each member of the set \mathcal{F}_0 is a perturbation of g_0 in its central part by the g_j 's.

The perturbing functions g_j are obtained through shifting an affine transform \tilde{g} of g_0 : $\tilde{g}(x) = u \cdot [g_0(2Nx + 1) - g_0(2Nx - 1)]$. They are bounded by u , and $\int g_j = 0$.

Note that $\|\tilde{g}\|_2^2 = u^2 N^{2s-1} O(1)$. For the sidelengths:

- $D(g_\epsilon, g_{\epsilon'}) = \sum (\epsilon_j - \epsilon'_j)^2 \|g_j\|_2^2 = u^2 N^{-1} \sum_j |\epsilon_j - \epsilon'_j|$, so that $\Delta = u^2 N^{-1} O(1)$.
- $h^2(g_\epsilon^{\otimes n}, g_{\epsilon'}^{\otimes n}) \leq n/2 \sum_j |\epsilon_j - \epsilon'_j| \int_I (\sqrt{1+g_j} - \sqrt{1-g_j})^2$. As $u = u(n) \rightarrow 0$ as $n \rightarrow \infty$, the integral is controlled by $\|g_j\|_2^2 = u^2/N O(1)$. Thus $\delta = nu^2 N^{-1} O(1)$.

From inequality (6) above, the minimax risk $m(\mathcal{F}_0)$ for (\mathcal{F}_0, D) is greater than

$$\frac{N\Delta}{4B} \exp(-8\delta) = \frac{1}{8} u^2 O(1) \exp(-8nu^2 N^{-1} O(1)).$$

This expression is maximized if we make sure that the exponent is $O(1)$ while keeping u^2 large. The choice $u = N^{-s}$ and $N = n^{1/(2s+1)}$ works nicely and gives: $m(\mathcal{F}_0) \geq n^{-2s/(2s+1)} O(1)$.

The rate of convergence for the squared- L^2 risk cannot be better than $\rho^* = 2s/(2s+1)$, which is of course a well-known result.

15.3 Fano's Pyramid

The framework and goal are the same as previously, except that now the loss function on which the risk is based needs no longer to be a super-additive discrepancy D . A distance d being defined on $\mathcal{F} \times \mathcal{F}$ (or else on $\mathcal{G} \times \mathcal{G}$, $\mathcal{G} \supset \mathcal{F}$), and ℓ being a nondecreasing function, zero at zero, the risk of \hat{f} at f is now defined as $R(\hat{f}, f) = E_{P_f}[\ell(d(\hat{f}, f))]$. The discrepancies D were a special case: for example $d(f, g) = \|f - g\|_p$, $p \geq 1$, $\ell(d) = d^p$ gives $D(f, g) = \|f - g\|_p^p$. We shall still call $D = \ell \circ d$ the loss function though it is no longer a discrepancy as previously defined.

Our lower bounds in this section will be based on the following notion of Δ -Pyramid. Consider inside \mathcal{F} a set Θ consisting of a finite number of points, such that: $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$, with K finite, and $d(\theta_i, \theta_j) \geq \Delta$ for $i \neq j$. Such a pyramid is characterized by the number K of its points and the sidelength Δ : it is called a Δ -separated pyramid, or else a Fano pyramid.

15.3.1 THE KEY QUANTITY: p_e

Let \hat{f} be any estimate. Its maximum risk over a class \mathcal{F} is bounded below by its maximum risk over a Δ -separated pyramid $\Theta \subset \mathcal{F}$. Define a modified

loss, $\ell'(d(\hat{\theta}, \theta)) = 0$ if $\theta = \hat{\theta}$ and $\ell(\Delta/2)$ otherwise. Under our assumptions on d and on Θ , we have $\ell(d(\hat{\theta}, \theta)) \geq \ell'(d(\hat{\theta}, \theta))$, and

$$\begin{aligned} \sup_{f \in \mathcal{F}} R(\hat{f}, f) &\geq \max_{\theta \in \Theta} R(\hat{f}, \theta) \geq \max_{\theta \in \Theta} R'(\hat{f}, \theta) \\ &= \max_{\theta \in \Theta} \ell(\Delta/2) P_\theta(\hat{\theta} \neq \theta) \\ &\geq \ell(\Delta/2) K^{-1} \sum_{\theta \in \Theta} P_\theta(\hat{\theta} \neq \theta). \end{aligned}$$

The average $p_e = K^{-1} \sum_{\theta \in \Theta} P_\theta(\hat{\theta} \neq \theta)$ is the probability of error when using $\hat{\theta}$ to test simultaneously the K hypotheses $H_i : \theta = \theta_i$, under the uniform a priori law on Θ .

15.3.2 FANO'S LEMMA: LOWER BOUND FOR p_e

Let $I(X, \theta)$ denote the mutual information of X and θ , i.e. the Kullback distance of the joint law of (X, θ) to the product of its marginals, or Shannon information, $I(X, \theta) = K(L(X, \theta), L(X) \otimes L(\theta))$.

Lemma 4 (Fano)

$$p_e \geq 1 - \frac{I(X, \theta) + \log 2}{\log(K-1)}$$

Proof: Let U be the uniform probability, and $P = \{p_1, \dots, p_K\}$ be any probability on $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$. By definition, the entropy of P is equal to $H(P) = -\sum_{k=1}^K p_k \log p_k$. We note the key fact that

$$H(P) \leq H(U) = \log K. \quad (7)$$

Indeed, the entropy of P may be expressed in terms of the Kullback divergence of P from the uniform:

$$K(P, U) = \sum p_k \log \frac{p_k}{1/K} = H(U) - H(P) \geq 0,$$

as $K(P, Q) \geq 0$ always.

We now develop an additive decomposition of the entropy $H(P)$. Divide the set Θ into two parts: on one side $\{\theta_i\}$, with probability p_i , and on the other side everything else, $\Theta^{-i} = \Theta \setminus \{\theta_i\} = \{\theta_k\}_{k \neq i}$, with probability $q_i \equiv \sum_{k \neq i} p_k$. Let P^{-i} be the law defined on Θ^{-i} , which is P conditioned by $\theta \in \Theta^{-i}$; $P^{-i}\{\theta = \theta_k\} = p_k/q_i$, $k \neq i$, and let P^i be the Bernoulli law ($p_i, q_i = 1 - p_i$). Then we get the additive decomposition

$$H(P) = H(P^i) + q_i H(P^{-i}). \quad (8)$$

As $H(P^i)$ is the entropy on a two-point set, $H(P^{-i})$ is the entropy on a $(K - 1)$ point set, it follows from (7) and (8) that

$$H(P) \leq \log 2 + q_i \log(K - 1). \quad (9)$$

We now condition on $\{X = x\}$, and suppose that $\hat{\theta} = \theta_i$ on this event. Let us now call $P_x = \{p_{1,x}, \dots, p_{k,x}\}$ the probability law of θ conditional on this value x of X . The error probability, conditionally on $X = x$, is simply $q_{i,x} = P(\theta \neq \theta_i | X = x) = \sum_{k \neq i} p_{k,x}$. Take the mean over X of the inequality (9) applied to $P = P_x$ and call the result $H(\theta|X)$. That is, $H(\theta|X) = E_X\{H(P_x)\}$. From (8) we have

$$H(\theta|X) \leq \log 2 + \log(K - 1)P(\hat{\theta} \neq \theta). \quad (10)$$

We can see in $P(\hat{\theta} \neq \theta)$ the error probability p_e except for the fact that now the law on Θ might not be uniform. By denoting simply $H(\theta)$ the entropy of the marginal law of θ , we have the identity $I(X, \theta) = H(\theta) - H(\theta|X)$, because $I(X, \theta) = E_{P_{X,\theta}} \log(dP(\theta|X)/dP(\theta))$. By replacing in (10) the term $H(\theta|X)$ by $H(\theta) - I(X, \theta)$, we get

$$p_e \geq \frac{1}{\log(K - 1)}[H(\theta) - I(X, \theta) - \log 2].$$

If we now put on Θ the uniform law, $H(\theta) = \log K$ and the asserted inequality follows. \square

15.3.3 UPPER BOUNDS FOR $I(X, \theta)$

To apply Fano's Lemma, we need strong upper bounds on $I(X, \theta)$.

Lemma 5 *Let $X = (X_1, \dots, X_n)$ be an n -sample of $P_\theta^{\otimes n}$. Then $I(X, \theta) \leq nI(X_1, \theta)$.*

Proof: Define

$$\begin{aligned} P &= L(X, \theta) &= \text{joint law of } (X, \theta), \\ \bar{P}_n &= E_\theta(P_\theta^{\otimes n}) &= \text{marginal law of } X, \\ \mu &= L(\theta) &= \text{marginal law of } \theta, \\ \bar{P} &\equiv E_\theta P_\theta = L(X_1) &= \text{marginal law of } X_1. \end{aligned}$$

We notice that \bar{P}_n , the marginal law of X is not in general, a product law, and in particular, is not, in general, equal to $(\bar{P})^{\otimes n}$ so that

$$I(X, \theta) = E_P \log \frac{dP}{d(\bar{P}_n \otimes \mu)} \quad (11)$$

$$= E_P \log \frac{dP}{d(\bar{P}^{\otimes n} \otimes \mu)} + E_P \log \frac{(d\bar{P}^{\otimes n} \otimes \mu)}{d(\bar{P}_n \otimes \mu)}. \quad (12)$$

The second term in the sum above, whose argument inside the logarithm *does not depend on* θ any more, as $\bar{P}^{\otimes n}$ and \bar{P}_n do not depend on θ , may also be written as $E_{\bar{P}_n} \log(d\bar{P}^{\otimes n}/dP_n)$, which is equal to $-K(\bar{P}_n, \bar{P}^{\otimes n})$ and thus is either negative or zero. The first term in the sum involves the logarithm of a product, $\prod_{i=1}^n dP_\theta(x_i)/d\bar{P}(x_i)$, and is equal to the sum of n identical terms—each one $I(X_1, \theta)$ —which gives the Lemma. \square

Lemma 6 *If X has distribution P_0 , then $I(X_1, \theta) \leq E_\theta K(P_\theta, P_0)$, which implies that $\bar{P} = E_\theta P_\theta$ minimizes $E_\theta K(P_\theta, P)$.*

Proof:

$$\begin{aligned} E_\theta E_{P_\theta} \log \frac{dP_\theta}{dP} &= E_\theta K(P_\theta, P_0) + E_{\bar{P}} \log \frac{dP_0}{dP} \\ &= E_\theta K(P_\theta, P_0) - K(\bar{P}, P_0). \square \end{aligned}$$

In consequence, $I(X, \theta) \leq n \max_{\theta \in \Theta} K(P_\theta, P_0)$ for each P_0 .

15.3.4 BOUNDS ON MINIMAX RISK

Combining results from the last two subsections, we get the following: *Let P_0 be any probability, and let $\Theta \subset \mathcal{F}$ be a Δ -separated pyramid of cardinality K . Then*

$$\sup_{f \in \mathcal{F}} R(\hat{f}, f) \geq \ell\left(\frac{\Delta}{2}\right) \left[1 - \frac{n \max_{\theta \in \Theta} K(P_\theta, P_0) + \log 2}{\log(K-1)} \right] \quad (13)$$

This bound is most powerful when we can build a “massive” pyramid, having Δ large, and K large, while keeping $\max_{\theta \in \Theta} K(P_\theta, P_0)$ small.

15.3.5 EXISTENCE OF MASSIVE PYRAMIDS

Actually, we cannot explicitly “build” pyramids. In contrast to the cube, which we have explicitly constructed, we can only prove the existence of appropriate pyramids under certain assumptions, without being able to exhibit one directly. In fact, we first construct massive hypercubes, and then establish the existence of sufficiently massive pyramids within those hypercubes.

Let $\mathcal{F}_0 = \{f : f = f_0 + \sum_{i=1}^N a_i \varphi_i\}$, where f_0 is a basic density (nearly) constant (as in Section 15.2.5), for an example), the φ_i 's are perturbations on disjoint intervals I_i of \mathbb{R} and the a_i 's are either 1 or 0, indicating whether the perturbation is present or not. Let f be one member of \mathcal{F}_0 , and $A(f)$ the set of all subscripts i such that the corresponding perturbation is present:

$$A(f) \equiv \{i \in \{1, 2, \dots, N\} : a_i(f) = 1\} \equiv A \quad A(f') \equiv A'.$$

Let $A \Delta A'$ denote the symmetric difference: $A \Delta A' = \{i : a_i(f) \neq a_i(f')\}$. We say that $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$, a subset of \mathcal{F}_0 , is a k -separated pyramid if for any pair (θ_i, θ_j) , $i \neq j$, at least k of their respective a_i are different, $|A(\theta') \Delta A(\theta)| \geq k$ for $\theta' \neq \theta$; to go from any point of the pyramid to any other one, along the cube, the minimum number of edges is k .

Lemma 7 *The maximal number $S(N, k)$ of vertices of a k -separated pyramid $\Theta(k)$ inside an N -dimensional hypercube \mathcal{F}_0 with 2^N elements obeys*

$$S(N, k) \geq \frac{2^N}{\sum_{j=0}^{k-1} \binom{N}{j}}. \quad (14)$$

Proof: Let f be a point of the cube \mathcal{F}_0 and let

$$V_{k-1}(f) = \{f' \in \mathcal{F}_0 : |A(f) \Delta A(f')| \leq k - 1\}$$

be the $(k-1)$ -neighbourhood of f , the set of points which are at a distance along the cube strictly less than k edges from f . Defining $A \equiv A(f)$ and letting B be its complementary set in $\{1, 2, \dots, N\}$, let us notice that f' and $V_{k-1}(f)$ are of the following type: $A(f') = (A_{-i}) \cup (B_{+j})$ for $i + j \leq k - 1$, where A_{-i} is equal to A minus i of its points and B_{+j} is equal to a subset of B having exactly j points, as $|A(f) \Delta A(f')| = i + j \leq k - 1$. Thus there are $\binom{|A|}{i} \binom{|B|}{j}$ different ways for getting such $A(f')$. It follows that the number of $(k-1)$ -neighbourhoods of f is equal to:

$$|V_{k-1}(f)| = \sum_{i+j \leq k-1} \binom{|A|}{i} \binom{|B|}{j}.$$

In order to compute this sum, let us notice that it can be written by denoting $B(n, p)$ the binomial variable with parameters n and p :

$$\begin{aligned} & \sum_{i+j \leq k-1} \binom{|A|}{i} \binom{|B|}{j} 2^{-|A|} 2^{-|B|} 2^N \\ &= 2^N \sum_{i+j \leq k-1} P(B(|A|, 1/2) = i) P(B(|B|, 1/2) = j) \\ &= 2^N P(B(N, 1/2) \leq k-1) = \sum_{j=0}^{k-1} \binom{N}{j}. \end{aligned}$$

Moreover, the union of all $(k-1)$ -neighbourhoods of the vertices of a maximal k -separated pyramid is necessarily equal to the whole cube \mathcal{F}_0 ; otherwise it would not be maximal: any non covered point would be k -distant from every vertex of the pyramid. Now $\bigcup_{\theta \in \Theta} V_{k-1}(\theta) = \mathcal{F}_0$ implies $\sum_{\theta \in \Theta} |V_{k-1}(\theta)| \geq 2^N$. Assertion (14) follows. \square

The lower bound (14) is the inverse of the probability that a binomial variable $B(N, 1/2)$ be less than or equal to $k-1$. We pick k a fraction of N , a convenient choice being $k = N/4$. A large deviation result (e.g. Hoeffding's Inequality) gives $S(N, N/4) \geq e^{N/8}$ for $N \geq 1$.

15.3.6 APPLICATION: SOBOLEV CLASSES

Recall the inequality (13) giving the lower bound for the risk on a set \mathcal{F} containing a Δ -separated pyramid with K vertices.

We consider the same example as the one used to illustrate the Assouad's cube device, $\mathcal{F} = \{f : \|f^{(s)}\|_p \leq r\}$ and $d(f, g) = \|f - g\|_p$, where $\ell(d) = d$ is the identity.

The cube \mathcal{F}_0 can be roughly the same as in the Assouad construction, f_0 being approximately equal to 1 on an interval, and being C^∞ , and the basic perturbation φ has integral 0 and obeying $\|1 - (\varphi/f_0)\|_\infty \leq c < 1$. each individual perturbation is a scaled and shifted version, $\varphi_{j,N}(x) = u\varphi(n(x-x_j))$, and the typical member of the cube is $f_\theta = f_0 + \sum_{j=1}^N a_j \varphi_{j,N}$. The two parameters N and u have to be chosen in such a way that the regularity conditions defining \mathcal{F} are fulfilled by f_θ . As $f_\theta^{(s)} = f_0^{(s)} + uN^s \varphi^{(s)}$, we have $\|f_\theta^{(s)}\|_p^p = u^p N^{sp} O(1)$, which leads to the condition $uN^s = O(1)$.

In order that the Fano's lower bound be useful, we need that

$$\frac{n \max_{\Theta} K(P_\theta, P_0)}{\log K} = O(1). \quad (15)$$

To see what this means in terms of the perturbation parameters u and N , set $f_\theta(x) = f_0(x)[1 + g_\theta(x)]$ and $\|g_\theta\|_\infty \leq c < 1$ for every θ in Θ . Now

$$K(P_\theta, P_0) = \int f_\theta \left(\log \frac{f_\theta}{f_0} \right) dx = \int (1 + g_\theta) \log(1 + g_\theta) f_0 dx.$$

As $(1 + g_\theta) \log(1 + g_\theta) = g_\theta + \frac{1}{2} g_\theta^2 / (1 - c)$ and $\int g_\theta f_0 dx = 0$,

$$K(P_\theta, P_0) \leq \frac{1}{2(1 - c)} \int g_\theta^2 f_0 dx = \frac{1}{2(1 - c)} \int \left(\frac{f_\theta - f_0}{\sqrt{f_0}} \right)^2 dx.$$

The lower bound can thus be written as

$$m(\mathcal{F}) \geq \ell \left(\frac{\Delta}{2} \right) \left[1 - \frac{n}{\log(K-1)} \sup_{\theta \in \Theta} \left\| \frac{f_\theta - f_0}{\sqrt{f_0}} \right\|_2^2 \frac{1}{2(1 - c)} - \frac{\log 2}{\log(K-1)} \right].$$

We have then $\|(f_\theta - f_0)/\sqrt{f_0}\|_2^2 < u^2 O(1)$, and as $\log K = N/8$, the condition (15) requires: $nu^2/N = O(1)$. Finally, one has to maximize the D -separation Δ corresponding to the $k = N/4$ -spaced pyramid, with $K = e^{N/8}$ points: $D = \|f_\theta - f_{\theta'}\|_p^p = (N/4)(uP\|\varphi\|_p^p/N)$. The separation Δ is thus equal to u .

Then u has to be maximized under the constraints $u = N^{-s}$, $nN^{-2s-1} = O(1)$, which gives $u = n^{-s/(2s+1)}$ and $N = n^{1/(2s+1)}$. For the L^p distance, the optimal rate of the minimax risk is thus equal to $s/(2s+1)$. As a special case, for the L^2 norm and the derivative of order 2 it gives a rate $\rho = 2/5$.

15.4 Discussion

- The pyramid has only $e^{N/8}$ vertices while the cube has many more, 2^N for an $N/4$ pyramid, but, as only the logarithms of cardinalities appear in the lower bound, either approach gives the same rate.
- The pyramid allows for any non decreasing function ℓ of a distance d to define the loss, while the cube needs a function of the distance that is super-additive.
- In the case of censored or truncated observations, the rates achieved by certain estimates have been computed (Mierniczuk 1985), but no lower bound seems to have been derived so far.

15.5 REFERENCES

- Assouad, P. (1983), ‘Deux remarques sur l’estimation’, *Comptes Rendus de l’Academie des Sciences, Paris, Ser. I Math* **296**, 1021–1024.
- Bretagnolle, J. & Huber, C. (1979), ‘Estimation des densites: risque minimax’, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **47**, 119–137.
- Ibragimov, I. A. & Has’minskii, R. Z. (1981), *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.
- Mierniczuk, J. (1985), ‘Some asymptotic properties of kernel estimators of a density function in case of censored data’, *Annals of Statistics* pp. 766–773.
- Nadaraya, E. A. (1974), ‘On the integral mean squared error for some non-parametric estimates for the density function’, *Theory of Probability and Its Applications* **19**, 131–141.
- Rosenblatt, M. R. (1971), ‘Curve estimates’, *Annals of Mathematical Statistics* **42**, 1815–1842.
- Wahba, G. (1975), ‘Optimal convergence properties of variable knot, kernel, and orthogonal series estimators’, *Annals of Statistics* **3**, 15–29.
- Watson, G. S. (1969), ‘Density estimation by orthogonal series’, *Annals of Mathematical Statistics* **40**, 1496–1498.
- Weiss, L. & Wolfowitz, J. (1967), ‘Estimation of a density at a point’, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **7**, 327–335.

16

Some Estimation Problems in Infinite Dimensional Gaussian White Noise

I. Ibragimov¹

R. Khasminskii²

16.1 Introduction

Statistical problems for infinite dimensional Gaussian white noise arise in a natural way when one tries to study statistical questions connected with stochastic partial differential equations (Hübner, Khas'minskii & Rozovskii 1993). In this paper we analyze the simplest situation of estimation of the shift parameter in Gaussian white noise. We considered the one-dimensional case in Ibragimov & Has'minskii (1977). It turns out that the problems for infinite dimensional white noise have a much richer analytical content, as we hope to show in this and other papers that we are planning to write.

Let H be a Hilbert space. Let Q be a symmetric, positive, bounded operator in H . Denote by $W_Q(t)$ a random Gaussian process with independent increments $W_Q(t) - W_Q(s)$ distributed Gaussian with mean zero and correlation operator $\delta(t-s)Q$. That is, W_Q is a Q -Wiener process, in the terminology of DaPrato & Zabczyk (1992).

If $\text{tr}Q < \infty$, the process $W_Q(t)$ is an H -valued continuous random process. If $\text{tr}Q = \infty$, $W_Q(t)$ is a so-called cylindrical Wiener process (DaPrato & Zabczyk 1992).

Let $\mathbf{L}_2(0, 1) = \mathbf{L}_2$ denote the Hilbert space of H -valued functions s with the inner product

$$(s_1, s_2) = \int_0^1 (s_1(t), s_2(t))_H dt$$

and norm $\|s\|^2 = (s, s)$.

We deal with the following observation scheme. We are observing $X_\varepsilon(t)$ where

$$dX_\varepsilon(t) = s(t)dt + \varepsilon dW_Q(t) \quad 0 \leq t \leq 1. \quad (1)$$

¹St.-Petersburg branch of the Steklov Mathematical Institute

²Department of Mathematics Wayne State University

We assume that our unknown shift parameter s belongs to a known set $\Sigma \subseteq \mathbf{L}_2$ and that ε and Q are known to the statistician. Our estimation problem is to estimate the value $\Phi(s)$ of a known function $\Phi : \mathbf{L}_2 \rightarrow U$, for U a Euclidean or Hilbert space. In particular, if Φ is the identity operator in H , our problem is one of estimation s itself. The problem of estimation of a finite dimensional parameter can also be imbedded into this scheme. Namely, let

$$dX_\varepsilon(t) = s(t; \theta)dt + \varepsilon dW_Q(t) \quad \theta \in \Theta \subseteq \mathbb{R}^n.$$

In this situation the parameter set Σ is an n -dimensional manifold in $\mathbf{L}_2(0, 1)$ spanned by $s(\cdot; \theta), \theta \in \Theta$ and $\Phi(s(\cdot; \theta)) = \theta$.

Below we mostly deal with $Q = I$ where I is the identity operator and write $W_I(t) = W(t)$. Formally speaking any problem (1) can be reduced to this one by multiplying both parts of (1) by $Q^{-1/2}$.

Of course if $\text{tr}Q = \infty$ —in particular, if $Q = I$ —we cannot observe $X_\varepsilon(t)$; it makes no sense. The observable elements are random linear functionals

$$\int_0^1 (f(t), dX_\varepsilon(t))_H = \int_0^1 (f(t), s(t))_H dt + \varepsilon \int_0^1 (f(t), dW_Q(t))_H,$$

where f ranges over $Q^{-1/2}\mathbf{L}_2$. The definition and construction of these stochastic integrals can be found in DaPrato & Zabczyk (1992). Statistics are functions of the observable elements.

We consider $W(t)$ as a formal series $\sum_j e_j W_j(t)$ where the W_j , for $j = 1, 2, \dots$, are independent standard Wiener processes and $\{e_j\}$ is an orthonormal basis in H . One can consider the observation

$$dX_\varepsilon(t) = s(t)dt + \varepsilon dW(t)$$

as a sequence of independent observations of the one-dimensional functional shift parameter in a one-dimensional white noise,

$$dY_j(t) = f_j(t)dt + \varepsilon dW_j(t) \quad j = 1, 2, \dots,$$

where

$$Y_j(t) = \int_0^t (e_j, dX_\varepsilon(s))_H, \quad f_j(t) = (s(t), e_j)_H.$$

Instead of (1) one can consider formally a more general observation scheme (Tsirelson 1982). Namely, we are given a linear space with a centered Gaussian measure (E, γ) and its kernel $E_0 \subset E$. (In our initial scheme (1), the kernel E_0 equals $Q^{1/2}\mathbf{L}_2$). The observation is

$$U_\varepsilon = s + \varepsilon \xi \tag{2}$$

where $s \in \sum \subseteq E_0$ and ξ is a Gaussian random element with distribution γ . That is, the observable elements are

$$\langle U_\varepsilon, f \rangle = \langle s, f \rangle + \varepsilon \langle \xi, f \rangle \quad f \in E_0.$$

Here E_0 is a Hilbert space with an inner product $\langle s_1, s_2 \rangle$, and the functional $F(\theta) = \langle \theta, s \rangle$ can be extended to a linear measurable functional $\langle \theta, \xi \rangle$ on (E, γ) with $E\langle \theta, \xi \rangle^2 = \langle \theta, \theta \rangle$.

Sometimes we write (1) in the form (2) as $\dot{X}_\varepsilon = s + \varepsilon \dot{W}_Q$, where, for $f \in Q^{1/2}\mathbf{L}_2$,

$$\langle \dot{X}_\varepsilon, f \rangle = (\dot{X}_\varepsilon, f) = \int_0^1 (f(t), dX_\varepsilon(t))_H = (f, s) + \varepsilon (\dot{W}_Q, f)$$

and

$$Var(\dot{X}_\varepsilon, f) = \varepsilon^2 E(\dot{W}_Q, f)^2 = \varepsilon^2 \|Q^{1/2}f\|^2 = \varepsilon^2 (Qf, f).$$

Let us return to the problem (1). Denote by $P_s^{(\varepsilon)}$ the distribution of X_ε . In particular, $P_0^{(\varepsilon)}$ denotes the distribution of εW_Q . It is well known (Gikhman & Skorokhod 1974, page 490) that if $\sum \subseteq Q^{1/2}\mathbf{L}_2$, then all measures $P_s^{(\varepsilon)}$ are mutually absolutely continuous and

$$\frac{dP_s^{(\varepsilon)}}{dP_0^{(\varepsilon)}}(X_\varepsilon) = \exp\left(\frac{1}{\varepsilon} \int_0^1 (Q^{-1/2}s, dW)_H - \frac{1}{2\varepsilon^2} \|Q^{-1/2}s\|^2\right). \quad (3)$$

In particular, for any $h \in \mathbf{L}_2$,

$$\frac{dP_{s+\varepsilon Q^{1/2}h}^{(\varepsilon)}}{dP_s^{(\varepsilon)}}(X_\varepsilon) = \exp\left(\int_0^1 (h, dW(t))_H - \frac{1}{2} \|h\|^2\right).$$

That is, the family of measures $\{P_s^{(\varepsilon)}, s \in \sum\}$ satisfies the LAN conditions in the sense of Ibragimov & Khas'minskii (1991), with norming operator $A_\varepsilon = \varepsilon Q^{1/2}$ and direction of LAN \mathbf{L}_2 .

Let U be an Euclidean or Hilbert space. Let $\Phi : \mathbf{L}_2 \rightarrow U$ be a Fréchet differentiable function. Suppose that the parameter set \sum is a smooth manifold in \mathbf{L}_2 . Denote by T_s the tangent space of \sum at $s \in \sum$. Let $P(L)$ denote the projector in \mathbf{L}_2 onto L . Consider the operator

$$K = \Phi'(s)Q^{1/2}P(T_s).$$

It follows from Ibragimov & Khas'minskii (1991) that if K is a Hilbert-Schmidt operator, then for any estimator T_ε of $\Phi(s)$ and a wide class of loss functions ℓ ,

$$\lim_{\delta \rightarrow 0} \liminf_{\varepsilon \rightarrow 0} \sup_{t \in U_\delta(s)} E_t \ell(\varepsilon^{-1}(T_2 - \Phi(t))) \geq E \ell(\xi). \quad (4)$$

Here ξ is a U -valued Gaussian random variable with mean zero and correlation operator KK^* and $U_\delta(s)$ denotes the ball in \mathbf{L}_2 with center s and radius δ .

We call an estimator T_ε for which the equality sign is achieved in (4), uniformly for all $s \in \sum$, asymptotically efficient in \sum .

Suppose, for example, that Φ is a linear function and that either Φ or $Q^{1/2}$ is Hilbert-Schmidt operator. Then the statistic $\Phi(\dot{X}_\varepsilon)$ is well defined and if in addition \sum is sufficiently massive (for example, $\sum = Q^{1/2}\mathbf{L}_2$), then $\Phi(\dot{X}_\varepsilon)$ is asymptotically efficient (see Section 4 below).

When neither Φ' nor $Q^{1/2}$ are Hilbert-Schmidt operators one cannot guarantee even the existence of consistent estimators. For example, let

$$dX_\varepsilon(t) = \theta dt + \varepsilon dW_Q(t)$$

and let θ run over the unit ball of H -constants. If $Q^{1/2}$ is a Hilbert-Schmidt operator then the statistic $X_\varepsilon(1)$ is well defined and estimates θ efficiently. If Q is the identity operator and $\dim H = \infty$, then consistent estimators of θ do not exist (see Section 2).

Theorems 1 and 2 in the next section assert, roughly speaking, that consistent estimation of $\Phi(s)$ is possible iff the set $\Phi(Q^{1/2}\sum)$ is compact in U .

Suppose now that the parameter set \sum is bounded in \mathbf{L}_2 . In this case we may and we will suppose that $\sum \subseteq U_1$, where U_r denotes the ball in \mathbf{L}_2 of the radius r and with its center in the origin. Let Φ be a bounded linear operator $L_2 \rightarrow U$. Suppose also that Q is the identity operator. We will distinguish between three cases:

- (i) Φ is a Hilbert-Schmidt operator;
- (ii) Φ is a compact operator;
- (iii) Φ^{-1} is a bounded operator.

This paper concerns mostly the first, simplest case, where we can (as a rule) construct asymptotically efficient estimators (Section 4). In the most delicate case, (ii), consistent estimators always exist but the rate of convergence to $\Phi(s)$ as $\varepsilon \rightarrow 0$ depends on some joint characteristics of Φ and \sum . In the case (iii) consistent estimators exist only if \sum is a compact. Their rate of convergence to $\Phi(s)$ is determined by some metric characteristics of the compact \sum , such as the ε -entropy and the Kolmogorov width.

We illustrate the last two cases by only a few simple examples. Their detailed analysis will be published elsewhere.

Observe that the variant of the LAN conditions formulated above is the development of but one of the many original ideas in Statistics proposed by L. Le Cam. These ideas, which are partially summarized in (Le Cam 1986), have had a strong influence on modern Mathematical Statistics.

16.2 Consistency conditions

From now on we always suppose that $Q = I$, the identity operator in H , and that our observation scheme is the following one

$$dX_\varepsilon(t) = s(t)dt + \varepsilon dW(t), \quad W(t) = W_I(t), \quad 0 \leq t \leq 1, \quad (5)$$

or $\dot{X}_\varepsilon = s + \varepsilon \dot{W}$, with the parameter set $\Sigma \subseteq U_1$.

Let M be a metric space with a metric ρ . Let $\Phi : \Sigma \rightarrow M$. We call estimators T_ε of $\Phi(s)$ M -consistent if for any $\delta > 0$

$$\sup_{s \in \Sigma} P_s^{(\varepsilon)} \{ \rho(\Phi(s), T_\varepsilon) > \delta \}_{\varepsilon \rightarrow 0} \rightarrow 0.$$

Theorem 1 *If M -consistent estimators of $\Phi(s)$ exist, then the set $\Phi(\Sigma)$ is compact in M .*

Remark Recall that the parameter set Σ supposed to be bounded, $\Sigma \subseteq U_1$. For unbounded Σ the result is not true. As a simple counterexample, consider the problem of estimation of the mean $X = s + \varepsilon \xi$. Here $\Sigma = (-\infty, \infty)$, $\Phi(s) = s$, and $\Phi(\Sigma) = (-\infty, \infty)$; and the estimator X is consistent.

Proof: We deduce the theorem from a more general result. To formulate it we need the notions of the Kolmogorov capacity of Σ and the Shannon capacity of the communication channel (5) under the restriction $s \in \Sigma$.

Recall first the definition of the Kolmogorov capacity (Kolmogorov & Tihomirov 1959). Points x_1, \dots, x_n of Σ are called δ -distinguishable (in M) if $\rho(x_i, x_j) > \delta$ for $i \neq j$. Let $N_\delta(\Sigma) = \max\{n : (x_1, \dots, x_n) \text{ are } \delta\text{-distinguishable}\}$. The expression

$$C_\delta(\Sigma, (M, \rho)) = C_\delta(\Sigma) = \ln N_\delta(\Sigma)$$

is called the (Kolmogorov) δ -capacity of Σ .

Notice that if points (x_1, \dots, x_n) are δ -distinguishable, they constitute a δ -net in Σ . A set $\Sigma \subset M$ is compact in M iff $C_\delta(\Sigma) < \infty$ for all $\delta > 0$.

We use also the following concepts, borrowed from information theory. Let ξ and η be two random elements with distributions P_ξ, P_η and joint distribution $P_{\xi\eta}$. The Shannon information (Gallager 1968) in ξ about η is defined as

$$I(\xi, \eta) = E \left(\ln \frac{dP_{\xi\eta}}{dP_\xi \times dP_\eta}(\xi, \eta) \right).$$

Define the Shannon capacity of the channel (5), under the restriction $s \in \Sigma$, as $C_\varepsilon(\Sigma) = \sup I(s, X_\varepsilon)$ where X_ε and s are related by the equation (5), the random elements s do not depend on W , and the upper bound is taken over the set of distributions for s in L_2 whose support belongs to Σ . We deal with only the channel (5), but with different sets Σ . So for us $C_\varepsilon(\Sigma)$ is a characteristic of sets Σ rather than of the channel (5). We shall call it the Shannon capacity of Σ .

Notice that for $\Sigma \subseteq U_r$,

$$C_\varepsilon(\Sigma) \leq \frac{r^2}{2\varepsilon^2}. \quad (6)$$

Indeed, it follows from (3) that $I(s, X_\varepsilon)$ (we omit the upper index ε) equals

$$\begin{aligned} E \left(\ln \frac{dP_{X_\varepsilon|s}}{dP_0}(X_\varepsilon, s) \right) - E \left(\ln \frac{dP_{X_\varepsilon}}{dP_0}(X_\varepsilon, s) \right) = \\ E \left(\frac{1}{2\varepsilon^2} \int_0^1 \|s\|_H^2 dt + \frac{1}{\varepsilon} \int_0^1 (s(t), dW(t))_H \right) + E \left(\ln \frac{dP_0}{dP_{X_\varepsilon}}(X_\varepsilon, s) \right). \end{aligned}$$

Further, $E \int_0^1 (s(t), dW(t))_H = 0$ and by Jensen's inequality

$$E \ln \frac{dP_0}{dP_{X_\varepsilon}}(X_\varepsilon, s) \leq \ln E \frac{dP_0}{dP_{X_\varepsilon}}(X_\varepsilon, s) = 0.$$

The inequality (6) follows.

Theorem 2 *Under the conditions of Theorem 1,*

$$\inf_T \sup_{s \in \Sigma} P_s \{ \rho(T, \Phi(s)) > \delta \} \geq 1 - \frac{\mathcal{C}_\varepsilon(\Sigma) + 1}{C_{2\delta}(\Phi(\Sigma)) - 1}. \quad (7)$$

Proof: Choose in $\Phi(\Sigma)$ a set $\theta = (t_1, \dots, t_n)$ of 2δ -distinguishable points, $\rho(t_i, t_j) > 2\delta$. Let $t_j = \Phi(s_j)$. Evidently, for any T ,

$$\begin{aligned} \sup_{s \in \Sigma} P_s \{ \rho(T, \Phi(s)) > \delta \} &\geq \sup_j P_{s_j} \{ \rho(T, \Phi(s_j)) > \delta \} \\ &\geq \frac{1}{n} \sum_{j=1}^n P_{s_j} \{ \rho(T, \Phi(s_j)) > \delta \}. \end{aligned}$$

Introduce the new estimator

$$\Psi_\varepsilon = \begin{cases} \Phi(s_i) & \text{if } \rho(T_\varepsilon \Phi(s_i)) \leq \delta, \\ T_\varepsilon & \text{if } \min_i \rho(T_\varepsilon, \Phi(s_i)) > \delta. \end{cases}$$

Then

$$\frac{1}{n} \sum_1^n P_{s_j} \{ \rho(T_\varepsilon, \Phi(s_j)) > \delta \} = \frac{1}{n} \sum_1^n P_{s_j} \{ \Psi_\varepsilon \neq t_i \}$$

According to Fano's lemma (Ibragimov & Has'minskii 1981, page 323), the last expression is bigger than $1 - (I(X_\varepsilon, \theta) + \ln 2) / (\ln(n-1))$. Taking the upper bound over all possible n , we find

$$\sup_n \left(1 - \frac{I(X_\varepsilon, \theta) + \ln 2}{\ln(n-1)} \right) \geq 1 - \frac{\mathcal{C}_\varepsilon(\Sigma) + 1}{C_{2\delta}(\Phi(\Sigma)) - 1}.$$

In the last inequality we used the general information theoretic bound $I(\xi, \Phi(\eta)) \leq I(\xi, \eta)$ —see, for example, Gallager (1968). The theorem follows.

Theorem 1 follows immediately from the inequality (7). Indeed if $\Phi(\Sigma)$ is not compact in M , then $C_{2\delta}(\Phi(\Sigma)) = \infty$ for all small $\delta > 0$. At the same time $C_\epsilon(\Sigma) \leq (2\epsilon^2)^{-1}$. Hence for all small $\delta > 0$

$$\sup_s P_s \{ \rho(T_\epsilon, \Phi(s)) > \delta \} = 1.$$

A simple example shows that the general converse of Theorem 1 is not true. Indeed, the function $\Phi(s) = \|s\|^2$ is very smooth, and it maps U_1 into $[0, 1]$, but consistent estimators of $\Phi(s)$ do not exist (Ibragimov, Nemirovskii & Has'minskii 1986). We can prove two partial converses to the theorem. The first one shows that for linear functions we have the full converse theorem.

Theorem 3 *Let B be a normed space. Let $\Phi : \mathbf{L}_2 \rightarrow B$ be a linear compact operator. There exist estimators T_ϵ^* such that*

$$\sup_{s \in U_1} P_s \{ \|\Phi(s) - T_\epsilon^*\|_B > \delta \}_{\epsilon \rightarrow 0} \rightarrow 0$$

for any $\delta > 0$.

Proof: Let at first Φ be an n -dimensional operator. It can be represented as $\Phi(s) = \sum_1^n (f_k, s) \cdot \varphi_k$, where f_k are fixed elements of \mathbf{L}_2 and φ_k are fixed elements of B . Consider the estimator

$$T_\epsilon^* = \sum_1^n (f_k, \dot{X}_\epsilon) \cdot \varphi_k = \sum_1^n \int_0^1 (f_k(t), dX_\epsilon(t))_H \cdot \varphi_k.$$

Evidently

$$\begin{aligned} E \|T_\epsilon^* - \phi(s)\|_B &= \epsilon E \left\| \sum_1^n \int_0^1 (f_k, dw)_H \cdot \varphi_k \right\|_B \\ &\leq \epsilon \sum_1^n \|f_k\| \|\varphi_k\|_B \rightarrow 0 \quad \text{as } \epsilon \rightarrow 0. \end{aligned}$$

Let now Φ be a compact operator. Take an n -dimensional operator $A_n = \sum_{k=1}^n (f_{kn}, \cdot) \varphi_{kn}$ and consider the estimator $T_{\epsilon n}^* = \sum_{k=1}^n (f_{kn}, \dot{X}_\epsilon) \varphi_{kn}$. We find that

$$\sup_s E_s \|\Phi(s) - T_{\epsilon n}^*\|_B \leq \sup_s \|(\Phi - A_n)s\|_B + \epsilon \sum_{k=1}^n \|f_{kn}\| \|\varphi_{kn}\|_B.$$

Hence

$$\inf_T \sup_s E_s \|\Phi(s) - T\|_B \leq \inf_n \inf_{A_n} \left(\|\Phi - A_n\| + \epsilon \sum_{k=1}^n \|f_{kn}\| \|\varphi_{kn}\|_B \right).$$

Every compact operator Φ is the limit in operator norm of a sequence of finite-dimensional operators. The theorem follows.

Remark Let B be a Hilbert space. Denote by s_n the s -numbers of the operator Φ (i.e. the square roots of the eigenvalues of the operator $\Phi^*\Phi$). Then

$$\inf_T \sup_{s \in \Sigma} E_s \|T - \Phi(s)\| \leq \inf_n \left(s_{n+1} + \varepsilon \sqrt{\sum_1^n s_j^2} \right).$$

Indeed, let

$$\Phi(s) = \sum_1^\infty s_j(f_j, s)\varphi_j, \quad s_j \downarrow 0$$

be the Schmidt expansion of the operator Φ , where $\{f_j\}, \{\varphi_j\}$ are orthonormal sequences in L_2 and B respectively. Take

$$T_{\varepsilon n}^* = \sum_1^n s_j(f_j, \dot{X}_\varepsilon)\varphi_j.$$

We find that

$$\begin{aligned} \sup_{s \in \Sigma} E_s \|\Phi(s) - T_{\varepsilon n}^*\|_B &\leq \sup_{s \in \Sigma} \left(\sqrt{\sum_{n+1}^\infty s_j^2(f_j, s)^2} + \varepsilon \sqrt{\sum_1^n s_j^2 E(f_j, \dot{W})^2} \right) \\ &= s_{n+1} + \varepsilon \sqrt{\sum_1^n s_j^2} \end{aligned}$$

Notice also that in this case

$$\inf_{A_n} \|\Phi - A_n\| = s_{n+1},$$

which is a theorem of Allahverdiev (Gohberg & Krein 1969).

In the next theorem we again consider functions $\Phi(s)$ with the values in a metric space (M, ρ) .

Theorem 4 Let the parameter set Σ be compact in L_2 . Let $\Phi : \Sigma \rightarrow M$ be a uniformly continuous function on Σ . Then for any $\delta > 0$

$$\inf_T \sup_s P_s \{\rho(T, \Phi(s)) > \delta\} \rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0.$$

Proof: Define the modulus of continuity of the function $\Phi(s)$ as

$$\omega(\delta) = \sup \{\rho(\Phi(s_1), \Phi(s_2)) : s_1, s_2 \in \Sigma, \|s_1 - s_2\| \leq \delta\}.$$

Since the function $\Phi(s)$ is uniformly continuous, $\omega(\delta) \rightarrow 0$ when $\delta \rightarrow 0$. We deduce our theorem from the following result: there exist estimators T_ε^* such that

$$\sup_s P_s \{\rho(T_\varepsilon^*, \Phi(s)) > \omega(3\delta) + \omega(\delta)\} \leq \sqrt{\frac{\varepsilon}{\delta}} \exp \left(\frac{1}{2} C_\delta(\Sigma) - \frac{\delta^2}{16\varepsilon^2} \right). \quad (8)$$

To construct the estimator T_ε^* choose in \sum the maximal number n of δ -distinguishable (in L_2 -norm) points $s_1 \dots s_n$. We have $n = \exp(C_\delta(\sum)) < \infty$. The points $s_1 \dots s_n$ constitute a δ -net in \sum . Hence for any $s \in \sum$ we can find a point s_j such that $\rho(\Phi(s), \Phi(s_j)) \leq \omega(\delta)$.

Consider the functions

$$\varphi_j(\dot{X}_\varepsilon) = \frac{dP_{s_j}}{dP_0}(\dot{X}_\varepsilon) = \exp\left(\frac{1}{\varepsilon^2} \int_0^1 (s_j(t), dX_\varepsilon(t))_H - \frac{\|s_j\|^2}{2\varepsilon^2}\right)$$

and define the estimator T_ε^* by

$$T_\varepsilon^* = \Phi(s_i) \text{ if } \max_{1 \leq j \leq m} \varphi_j(\dot{X}_\varepsilon) = \varphi_i(\dot{X}_\varepsilon)$$

If the maximum is attained at more than one point, it does not matter which one we choose.

For a given s we can find a point in $\{s_j\}$, say s_1 , such that $\|s - s_1\| \leq \delta$. We can then write that

$$\begin{aligned} P_s\{\rho(T_\varepsilon^*, \Phi(s)) > \omega(3\delta) + \omega(\delta)\} \\ \leq P_s\{\rho(T_\varepsilon^*, \Phi(s_1)) > \omega(3\delta)\} \\ \leq E_{s_1}^{1/2}\left(\frac{dP_s}{dP_{s_1}}(\dot{X}_\varepsilon)\right)^2 \cdot P_{s_1}^{1/2}\{\rho(T_\varepsilon^*, \Phi(s_1)) > \omega(3\delta)\}. \end{aligned} \quad (9)$$

We have

$$\begin{aligned} E_{s_1}\left(\frac{dP_s}{dP_{s_1}}(\dot{X}_\varepsilon)\right)^2 &= E \exp\left(\frac{2}{\varepsilon}(s - s_1, \dot{W}) - \frac{\|s - s_1\|^2}{\varepsilon^2}\right) \\ &= \exp\left(\frac{\|s - s_1\|^2}{\varepsilon^2}\right) \leq \exp\left(\frac{\delta^2}{\varepsilon^2}\right). \end{aligned} \quad (10)$$

Further,

$$P_{s_1}\{\rho(T_\varepsilon^*, \Phi(s_1)) > \omega(3\delta)\} \leq \sum_j P_{s_1}\{\varphi_j(\dot{X}_\varepsilon) \geq \varphi_1(\dot{X}_\varepsilon)\},$$

where the summation is taken over such j that $\rho(\phi(s_j), \phi(s_1)) \geq \omega(3\delta)$ and hence $\|s_j - s_1\| \geq 3\delta$. It follows that

$$\begin{aligned} P_{s_1}\{\rho(T_\varepsilon^*, \Phi(s_1)) > \omega(3\delta)\} \\ \leq \sum_j P_{s_1}\left\{\|s_j - s_1\|^{-1}(s_j - s_1, \dot{W}) \geq \frac{\|s_j - s_1\|}{2\varepsilon}\right\} \\ \leq n \frac{1}{\sqrt{2\pi}} \int_{(3\delta)/(2\varepsilon)}^\infty \exp\left(-\frac{v^2}{2}\right) dv \\ \leq \frac{2}{3\sqrt{2\pi}} \frac{\varepsilon}{\delta} \exp\left(C_\delta(\Sigma) - \frac{9\delta^2}{8\varepsilon^2}\right). \end{aligned} \quad (11)$$

The inequalities (9), (10), and (11) prove the inequality (8), and hence the theorem.

The results of this section generalize slightly some results from Ibragimov & Has'minskii (1977), where we considered observations in one-dimensional white noise. Problems with infinite-dimensional white noise involve parameter sets Σ with a richer analytical structure.

16.3 Some Examples

In what follows we take for H different L_2 -spaces of functions defined on regions G of d -dimensional Euclidean space. Our basic space \mathbf{L}_2 will be the space $L_2([0, 1] \times G)$ of functions $s(t, x_1, \dots, x_d)$. Denote t by x_0 and let $x = (x_1, \dots, x_d)$. As examples of the parameter set Σ we consider certain compact sets of functions that occur frequently in approximation theory as well as in numerical analysis and the theory of partial differential equations.

Example 1 Suppose that G is the unit cube $[0, 1]^d$. The space $L_2([0, 1] \times G)$ consists of periodic square summable functions $s(t, x)$ on $[0, 1] \times G$ with the norm

$$\|s\|^2 = \int_0^1 \int_G s^2(t, x_1, \dots, x_d) dt \dots dx_d.$$

As parameter set Σ we choose a class of smooth functions. We define classes of smooth functions, following Babenko (1960) (see also Mitjagin 1962, Temljakov 1993). Suppose that a polynomial $P(z)$ in $d + 1$ variables $z = (z_0, \dots, z_d)$ is given and let D denote the vector

$$\left(\frac{1}{i} \frac{\partial}{\partial z_0}, \frac{1}{i} \frac{\partial}{\partial z_1}, \dots, \frac{1}{i} \frac{\partial}{\partial z_d} \right) \quad \text{where } i = \sqrt{-1}.$$

The class W_2^P is the set of functions s such that $\|P(D)s\| \leq 1$ and $P(D)s \neq 0$. The set $\Sigma = \Sigma(P)$ equals the intersection of W_2^P with unit ball of $L_2([0, 1] \times G)$.

The special interesting examples we will consider are the following ones:

$$P(z) = z_0^{\beta_0} \dots z_d^{\beta_d}$$

or

$$P(z) = \sum z_j^{2\beta_j}.$$

We consider also the natural generalization of these two classes. Namely, expand s into a Fourier series,

$$s = \sum s_j \exp\{2\pi i(j, x)\}$$

and consider the two sets

$$\Sigma_1(\beta_0, \dots, \beta_d) = \left\{ s : \sum_{j=(j_0, \dots, j_d)} |j_0|^{2\beta_0} \dots |j_d|^{2\beta_d} s_j^2 \leq 1 \right\}, \quad (12)$$

$$\Sigma_2(\beta_0, \dots, \beta_d) = \left\{ s : \sum (|j_0|^{2\beta_0} + \dots + |j_d|^{2\beta_d}) s_j^2 \leq 1 \right\},$$

where the β_j are real positive numbers.

Example 2 Consider a nonperiodic variant of the previous example. Suppose that $G \subset R^d$ is a bounded domain with sufficiently regular boundary (for example, satisfying the restricted cone property—see Agmon 1965). Let $H = L_2(\Omega)$ and $L_2 = L_2([0, 1] \times G)$. Let R be a selfadjoint operator with domain $D_R \subset L_2([0, 1] \times G)$ generated by a formally selfadjoint uniformly elliptic operator

$$A = \sum_{|j| \leq \nu} a_j D^j, j = (j_0, \dots, j_d), |j| = j_0 + \dots + j_d.$$

Define $\Sigma = \{s : (Rs, s) \leq 1\}$.

Example 3 In this example the space H is $L_2(G)$, $G \subseteq R^d$ for $G = [0, 1]^d$ and $L_2 = L_2([0, 1] \times G)$. Let E_k be an ellipse in the complex plane $z_k = x_k + iy_k$ with foci at the points 0, 1 of the real axis. We define Σ as a collection of functions $s(x) = s(x_0, \dots, x_d)$ that are analytic on the $(d+1)$ -dimensional cube $[0, 1]^{d+1}$ and have an analytic continuation to the region $E_0 \times E_1 \times \dots \times E_d$, and which are bounded in this region by some constant L .

One can enlarge the class of such examples. We can also introduce L_p -spaces of functions with the norm

$$\|s\|_p^p = \int_0^1 dt \int_G |s(t; x_1, \dots, x_d)|^p dx_1 \dots dx_d, \quad \text{for } 2 \leq p < \infty,$$

and Σ_q sets defined as $\{s : \|P(D)s\|_q \leq 1\}$.

Let $\Phi : L_2 \rightarrow B$ be a linear operator. The results of the previous section let us decide when (in the framework of these examples) consistent estimation of $\Phi(s)$ is possible. The question then arises: What is the behaviour of

$$\Delta_\varepsilon(\Sigma, \Phi) = \inf_T \sup_{s \in \Sigma} E_s \|T_\varepsilon - \Phi(s)\|_B \quad \text{as } \varepsilon \rightarrow 0?$$

Suppose that Φ^{-1} is bounded. For the crude asymptotics of Δ_ε it is enough to investigate the case $\Phi = I$ only. The answer should depend on the massiveness of Σ . We give some examples. In these examples

$$\Delta_\varepsilon(\Sigma) = \inf_T \sup_{s \in \Sigma} E_s \|T_\varepsilon - s\|$$

- (i) The set $\Sigma_1(\beta_0, \dots, \beta_d)$ of (12). Let $\beta = \min(\beta_0, \dots, \beta_d)$ and $\ell = \text{card}\{j : \beta_j = \beta\}$. Then

$$\Delta_\varepsilon(\Sigma) \asymp \varepsilon^{\frac{2\beta}{2\beta+1}} \left(\ln \frac{1}{\varepsilon} \right)^{\frac{(\ell-1)\beta}{2\beta+1}}.$$

- (ii) The set $\Sigma_2(\beta_0, \dots, \beta_d)$ of (12). Define the number β by the relation $\beta^{-1} = \sum_0^d \beta_j^{-1}$. We have $\Delta_\varepsilon(\Sigma) \asymp \varepsilon^{2\beta/(2\beta+1)}$.

(iii) Let \sum be the set of the Example 2. Then $\Delta_\varepsilon(\Sigma) \asymp \varepsilon^{2\beta/(2\beta+1)}$, where $\beta = \frac{\nu}{2(d+1)}$.

(iv) Let \sum be the set from the Example 3. Then

$$\Delta_\varepsilon(\Sigma) \asymp \varepsilon \ln^{d+1} \left(\frac{1}{\varepsilon} \right).$$

The proofs of these and related results will be published elsewhere.

16.4 Hilbert-Schmidt operators

Let us return to the simplest situation where we estimate the value Φs of a Hilbert-Schmidt linear operator Φ .

For a bounded linear operator $A : \mathbf{L}_2 \rightarrow U$, for U a Hilbert space, define the statistics $A\dot{X}_\varepsilon$ by the following linear functional on U :

$$\langle A\dot{X}_\varepsilon, \varphi \rangle = (\dot{X}_\varepsilon, A^*\varphi) = \int_0^1 (A^*\varphi, dX_\varepsilon(t))_H \quad \text{for } \varphi \in U.$$

If Φ is a Hilbert-Schmidt operator then $\Phi\dot{X}_\varepsilon$ is an ε -consistent estimator of Φs and for sufficiently massive \sum it is even asymptotically efficient. Indeed, the difference

$$\Phi\dot{X}_\varepsilon - \Phi s = \varepsilon\Phi\dot{W}$$

is a Gaussian element in \mathbf{L}_2 with the correlation operator $\varepsilon^2\Phi\Phi^*$. For meager sets \sum we can sometimes find a better estimator. Namely, let $\sum \subseteq L$ where L is a proper subspace of \mathbf{L}_2 . Denote by P_L the projector onto L in \mathbf{L}_2 , and consider the estimator $\Phi P_L\dot{X}_\varepsilon$. Clearly, the difference $\Phi P_L\dot{X}_\varepsilon - \Phi s = \varepsilon\Phi P_L\dot{W}$ is Gaussian with correlation operator $\varepsilon^2\Phi P_L\Phi^*$.

More precisely, given the statistics $\Phi\dot{X}_\varepsilon$ take an orthonormal basis $\{\varphi_j\}$ in U and consider the estimator $\hat{\Phi}_\varepsilon = \sum(\Phi\dot{X}_\varepsilon, \varphi_j)\cdot\varphi_j$. Then

$$\hat{\Phi}_\varepsilon = \sum(\Phi s, \varphi_j)\varphi_j + \varepsilon \sum(\dot{W}, \Phi^*\varphi_j)\cdot\varphi_j = \Phi s + \varepsilon \sum(\dot{W}, \Phi^*\varphi_j)\varphi_j.$$

The difference $\hat{\Phi}_\varepsilon - \Phi s$ is a Gaussian element in U . Indeed

$$\begin{aligned} E\|\sum(\dot{W}, \Phi^*\varphi_j)\varphi_j\|^2 &= \sum E(\dot{W}, \Phi^*\varphi_j)^2 \\ &= \sum \|\Phi^*\varphi_j\|^2 = \text{tr}(\Phi\Phi^*) < \infty. \end{aligned}$$

In the sequel we prefer instead of $\hat{\Phi}_\varepsilon$ to write just $\Phi\dot{X}_\varepsilon$, and so on, which should cause no misunderstanding.

Returning to the estimator $\Phi P_L\dot{X}$, we observe that its quadratic risk is

$$E\|\Phi P_L\dot{X}_\varepsilon - \Phi s\|_U^2 = \varepsilon^2 \text{tr}(\Phi P_L\Phi^*) \tag{13}$$

It follows that, if \sum is a large enough subset of L , the estimator $\Phi P_L(\dot{X}_\varepsilon)$ is asymptotically efficient.

We use this construction to solve the following problem: estimate $s \in \sum \subseteq L$ if we are observing

$$dX_\varepsilon(t) = Asdt + \varepsilon dW(t),$$

where $A : \mathbf{L}_2 \rightarrow \mathbf{L}_2$ is an unbounded linear operator and A^{-1} is a Hilbert-Schmidt operator. We can argue in the following way. If we set $As = v$, our initial problem is reduced to the following one: estimate $A^{-1}v$, where

$$dX_\varepsilon(t) = v(t)dt + \varepsilon dW(t),$$

and v runs the set $A\sum \subseteq AL \subseteq \overline{AL} = V$. We know that the solution of this problem is given by the estimator $A^{-1}P_V\dot{X}_\varepsilon$. Observe that

$$P_V = AP_L(P_LA^*AP_L)^{-1}P_L$$

Indeed, since $P_V^* = P_V$ and $P_V^2 = P_V$, the operator P_V is a projector. Further, if $h = As \in AL$, then $P_Vh = h$, and if $h \perp V$ then $P_Vh = 0$.

We have thus proved that the asymptotically efficient estimator T_ε for s can be written as

$$T_\varepsilon = P_L(P_LA^*AP_L)^{-1}P_LA^*\dot{X}_\varepsilon.$$

It follows from (13) that the quadratic risk of this estimator T_ε is equal to

$$E_s \|T_\varepsilon - s\|^2 = \text{tr}(P_L(P_LA^*AP_L)^{-1}P_L) \quad (14)$$

We illustrate the result by a few examples where A is a differential operator.

Example 4 Take $H = L_2[\alpha, \beta]$, with $-\infty < \alpha < \beta < \infty$ and $\mathbf{L}_2 = L_2([0, 1] \times [\alpha, \beta])$. Let \sum consist of all differentiable functions $s(x)$ on $[\alpha, \beta]$ with $s(\alpha) = 0$. The operator A is defined as $A = a(t, x)\frac{\partial}{\partial x}$, where $a(t, x)$ is a positive continuous and continuously differentiable (with respect to x) function on $[0, 1] \times [\alpha, \beta]$. Clearly, $L = \{s : s \in H, s(\alpha) = 0\}$.

It is easy to check that $A^*s = -\frac{\partial}{\partial x}(a(t, x)s)$ with the domain $D_{A^*} = \{s(t, x) : s(t, \beta) \equiv 0, \frac{\partial s}{\partial x} \in \mathbf{L}_2\}$. Evidently, $P_Ls = \int_0^1 s(t, x)dt$ and

$$\begin{aligned} (P_LA^*AP_L)s &= -\frac{\partial}{\partial x} \int_0^1 a^2(t, x) dt \frac{\partial}{\partial x} \int_0^1 s(t, x) dt \\ &= -\frac{\partial}{\partial x} \left((\bar{a^2}(x)\frac{\partial}{\partial x}\bar{s}(x)) \right), \end{aligned}$$

where $\bar{f}(x)$ denotes the integral $\int_0^1 f(t, x)dt$.

Further,

$$P_L(P_LA^*AP_L)^{-1}P_LA^*s = \int_\alpha^x \frac{\bar{a}s(y)}{\bar{a^2}(y)} dy.$$

Formally our asymptotically efficient estimator is

$$T_\varepsilon^* = \int_{\alpha}^x \frac{\overline{a(\cdot, y) \dot{X}_\varepsilon}}{\overline{a^2(y)}} dy,$$

meaning that

$$T_\varepsilon^*(x) = (\mathbf{1}_{[\alpha, x]} \cdot \frac{\overline{a}}{\overline{a^2}}, \dot{X}_\varepsilon) = \int_0^1 (\mathbf{1}_{[\alpha, x]}(\cdot) \cdot \frac{a(t, \cdot)}{\overline{a^2}(\cdot)}, dX_\varepsilon(t))_H. \quad (15)$$

For this estimator

$$\begin{aligned} E \|T_\varepsilon^* - s\|^2 &= E \|T_\varepsilon^* - s\|_H^2 \\ &= \varepsilon^2 E \left\| \int_0^1 (\mathbf{1}_{[\alpha, x]}(\cdot) \cdot \frac{a(t, \cdot)}{\overline{a^2}(\cdot)}, dw(t)) \right\|_H^2 \\ &= \varepsilon^2 \int_{\alpha}^{\beta} \frac{(\beta - y)}{\overline{a^2}(y)} dy. \end{aligned}$$

Example 5 Let H and Σ be the same as in the previous example, and let

$$As = a(t, x) \frac{\partial s}{\partial x} + b(t, x)s.$$

Clearly

$$A^*s = -\frac{\partial}{\partial x}(as) + bs$$

with the same domain D_{A^*} as in Example 4. Again $B = (P_L A^* A P_L)^{-1}$ is an integral operator on H .

Let us describe the construction of P_L and the efficient estimator for this example. Let $\phi_1 = P_L f$. We can write

$$f = a(t, x)\phi'_1 + b(t, x)\phi_1 + \gamma(t, x) = P_V f + \gamma.$$

Multiplying by a and b , and integrating over t , we obtain the equalities

$$\begin{aligned} \overline{a^2}\phi'_1 + \overline{ab}\phi_1 &= \overline{af} - \overline{a\gamma}, \\ \overline{ab}\phi'_1 + \overline{b^2}\phi_1 &= \overline{bf} - \overline{b\gamma}. \end{aligned}$$

From these equalities and orthogonality γ and $P_V f$ we have (after integration by parts),

$$(\overline{a^2}\phi'_1 - \overline{af}, \phi'_1)_H + ((\overline{b^2} - (\overline{ab})')\phi_1 - \overline{bf}, \phi_1) + \phi_1^2(\beta)\overline{ab}(\beta) = 0.$$

Writing $\phi_2(x) = \overline{a^2}\phi_1 - \overline{af}$, we observe that ϕ_1 is the solution of problem

$$\begin{aligned} \overline{a^2}\phi'_1 - \phi_2 &= \overline{af}; \\ \phi'_2 - (\overline{b^2} - (\overline{ab})')\phi_1 &= -\overline{bf}; \\ \phi_1(\alpha) &= 0; \\ \phi_2(\beta) + \overline{ab}(\beta)\phi_1(\beta) &= 0. \end{aligned} \quad (16)$$

So we have the following result. Let $G(x, y)$ be the Green matrix for the problem (16). Then the efficient estimator \hat{S} for S can be written in the form

$$\hat{S}(x) = \int_0^1 (G_{11}(x, \cdot)a(t, \cdot) - G_{12}(x, \cdot)b(t, \cdot), dX_\epsilon(t, \cdot))_H.$$

It is easy to find the explicit form of G , \hat{S} , and the mean square error of the efficient estimator, if the coefficients a, b depend only on t . The coefficients in (16) are constants for this case.

Observe that the estimator T_ϵ and its quadratic risk can be written in explicit form also for the case

$$As = \sum_{i=0}^n a_i(t) \frac{\partial^i s}{\partial x^i},$$

if H is the same as in the Examples 4 and 5, and \sum is chosen so that A^{-1} is uniquely defined on \sum .

In conclusion, we thank David Pollard for his suggestions that helped us to improve the text of this paper.

Acknowledgments: The research of Ibragimov was partly supported by ISF Grant R36000, Russian National Foundation. The research of Khasminskii was partly supported by ONR Grants N00014-93-1-0936 and N00014-95-1-0793.

16.5 REFERENCES

- Agmon, S. (1965), *Lectures on elliptic boundary value problems*, Van Nostrand.
- Babenko, K. I. (1960), 'On the approximation of periodic functions of many variables by trigonometric polynomials', *Soviet Math. Doklady* **32**, 247–250.
- DaPrato, G. & Zabczyk, J. (1992), *Stochastic equations in infinite dimensions*, Cambridge University Press.
- Gallager, R. (1968), *Information theory and reliable communication*, Wiley.
- Gikhman, I. I. & Skorokhod, A. V. (1974), *The theory of stochastic processes*, Vol. 1, Springer-Verlag.
- Gohberg, I. & Krein, M. (1969), *Introduction to the theory of linear nonselfadjoint operators*, American Mathematical Society.

- Hübner, M. & Rozovskii, B. (1995), 'On asymptotic properties of maximum likelihood estimators in parabolic stochastic PDE's', *Probability Theory and Related Fields*. To appear.
- Hübner, M., Khas'minskii, R. & Rozovskii, B. (1993), Two examples of parameter estimation for stochastic PDE, in 'Stochastic Processes: A festschrift in honor of G. Kallianpur', Springer-Verlag, pp. 149–160.
- Ibragimov, I. A. & Has'minskii, R. Z. (1977), 'On estimation of infinite dimensional parameter in Gaussian white noise', *Soviet Math. Doklady* **236**, 1053–1055.
- Ibragimov, I. A. & Has'minskii, R. Z. (1981), *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.
- Ibragimov, I. A. & Khas'minskii, R. Z. (1991), 'Asymptotically normal families of distributions and efficient estimation', *Annals of Statistics* **19**, 1681–1724.
- Ibragimov, I. A., Nemirovskii, A. & Has'minskii, R. Z. (1986), 'Some problems on nonparametric estimation in Gaussian white noise', *Theory of Probability and Its Applications* **31**, 391–406.
- Kolmogorov, A. N. & Tihomirov, V. M. (1959), ' ε -entropy and ε -capacity of sets in function spaces', *Russian Mathematical Surveys*. American Mathematical Society Translations, Series 2, vol 17, pp. 277–367.
- Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.
- Mitjagin, B. S. (1962), 'Approximation of functions in C and L_p on the torus', *Mathemat. Sbornik* **58**, 397–414.
- Temljakov, V. N. (1993), *Approximation of periodic functions*, Nova Science Publishers, New York.
- Tsirelson, B. S. (1982), 'A geometrical approach to maximum likelihood estimation for infinite dimensional Gaussian shift, I', *Theory of Probability and Its Applications* **27**, 411–418.

On Asymptotic Inference in AR and Cointegrated Models With Unit Roots and Heavy Tailed Errors

P. Jeganathan¹

17.1 Introduction

Consider the AR(q) model

$$X_i = \beta^{(1)} X_{i-1} + \cdots + \beta^{(q)} X_{i-q} + \epsilon_i, \quad \text{for } i = 1, 2, \dots, n, \quad (1)$$

where $\epsilon_i, i \geq 1$, are i.i.d., independent of (X_0, \dots, X_{i-q}) . The *characteristic polynomial* associated with the model (1) is defined by

$$\phi(z) = 1 - \beta^{(1)}z - \beta^{(2)}z^2 - \cdots - \beta^{(q)}z^q. \quad (2)$$

We shall also consider in this paper a related model called cointegrated model. It is however convenient to describe the problems in terms of the model (1), in which context we consider inference problems regarding the unknown parameters $\beta^{(1)}, \dots, \beta^{(q)}$, associated with the situation where

- (C.1) The roots of the polynomial (2) are on, or suitably contiguous to, the unit circle.
- (C.2) The distribution of ϵ_1 is symmetric around zero and has regularly varying tail probabilities: $P(|\epsilon_1| > x) = x^{-\alpha} L(x)$ with $0 < \alpha < 2$ and $L(x)$ a slowly varying function at infinity.
- (C.3) The distribution of ϵ_1 has an absolutely continuous Lebesgue density $p(x)$ such that

$$0 < \lambda^2 = \int_{-\infty}^{\infty} \left(\frac{p^{(1)}(x)}{p(x)} \right)^2 p(x) dx < \infty, \quad (3)$$

where $p^{(1)}(x)$ denotes the derivative, whenever it exists, of $p(x)$.

¹University of Michigan

It will be indicated later that the symmetry requirement in (C.2) can be relaxed in many ways.

An important class of distributions that satisfies (C.2) and (C.3) is the class of symmetric stable laws. It has long been known that such heavy tailed distributions play an important role in modeling time series phenomena in many applied sciences. We refer to Davis & Resnick (1986), DuMouchel (1973), and Zolotarev (1986) for detailed discussions and suitable further references regarding this.

Recently, there have been increasing interest in inference problems in models such as (1) with innovations ϵ_i having heavy tails of the form (C.2). In a basic contribution to the subject, Davis & Resnick (1986) have shown that in a class of *stationary* models which include (1) when the roots of (2) are greater than unity in absolute value, the usual method of moments type estimates, such as Yule-Walker estimates, converge in distribution after suitable normalizations. Under similar stationary situations, Davis, Knight & Liu (1992) have recently established the same type of convergence results for suitable M-estimates with improved order of convergence. For the stationary ARMA model, the asymptotic behavior of Whittle estimates has been more recently studied by Mikosch, Gadrich, Klüppelberg & Adler (1995). The asymptotic distributions involved in these papers are typically non-normal, involving stable distributions.

In the case of first order AR(1) model with the single parameter β assumed to be $\beta = 1$, Chan & Tran (1989) have established the convergence in distribution of least squares (LS) estimate. The corresponding result for a certain class of M-estimates has been established by Knight (1989) with, interestingly, a better order of convergence, that is, with order of convergence of the form $\delta_n = n^{-1/2-1/\alpha}$, where α is as in (C.2), rather than n^{-1} associated with the LS estimate. The preceding results have also been extended to the first order AR(1) models with weakly dependent innovations, in Phillips (1991) and Knight (1991). In the AR(1) case, the convergence in distribution of LS estimates under the alternatives $\beta_n = 1 + n^{-1}h$ has been shown by Chan (1990).

The purpose of the present paper is to demonstrate that using the results given in Jeganathan (1995, first version 1988), a simple and satisfactory solution to the optimal asymptotic inference problems becomes possible within the framework of procedures described in Le Cam (1986) and Le Cam & Yang (1990). According to these procedures, one approaches the inference problems in two stages. In the first stage, one obtains a suitable preliminary estimate. Then in the second stage one obtains in the probable range of possible values of the preliminary estimate a suitable approximation to the likelihood ratios, and the inference problems are then based on the approximating likelihood.

In the situation of (C.1), it will be shown that the approximating likelihoods take a simple form of quadratic approximation, called *locally asymptotically mixed normal* (LAMN) approximation, see for example Le Cam

& Yang (1990, Chapter 5) or Jeganathan (1995) where such approximating structures are discussed with special reference to time series models. In addition, as in the case of LS estimate or M-estimates, the construction of the above procedures themselves will not depend, in view of the special structure of (1), on the nature of the roots of (2), or on the index parameter α involved in (C.2), though the asymptotic "optimality" of the procedures will depend on these facts and on the nature of the distribution of ϵ_1 . In particular, the procedures will be asymptotically optimal under (C.1)–(C.3), in view of LAMN, and in addition, when (C.1) is violated, the procedures can still be used to make at least a rough inference regarding the location of the parameters, which can then be used in the next step to obtain more refined or possibly asymptotically optimal procedures.

Thus, since the nature of the complexity of the roots of (2) themselves do not enter into the construction of preceding procedures, such as the construction of confidence regions, it becomes important to study the problem under as wide generality of the roots as possible, and hence the condition (C.1) which includes all possible unit roots.

It is interesting to note that the preceding LAMN structure is in contrast with the situation where the innovation ϵ_1 has a finite variance since in that case also one has the analogous quadratic approximation to the likelihood ratios (Jeganathan 1995, Section 9), but the LAMN condition fails, and hence the simplifications associated with LAMN are unavailable. In this connection, it may be noted that the important contribution made by Chan & Wei (1988) in studying the asymptotic behavior of LS estimate in the finite variance case has played, as will also in this paper, an important role in dealing with some of the technicalities introduced by the complexity of the unit roots in (C.2).

The present case is also in sharp contrast with the case where the roots of (2) are greater than unity in absolute value (stationary case), in the sense that, roughly speaking, the informations contained in individual conditional densities of X_i given the past are not even uniformly asymptotically negligible, and hence one will not have a quadratic approximation to the likelihood ratios. A study of inference problems associated with such situations will be presented separately elsewhere.

Now, in the present situation the quantities involved in the approximating quadratic of the likelihood ratios will involve the density of the innovation ϵ_1 , which is unfortunately unavailable in general in a usable closed form expression. In fact, as it is well known, even when ϵ_1 is assumed to have a symmetric stable law with characteristic exponent α , a closed form of the density is available only when $\alpha = 1/2$ (Lévy) and $\alpha = 1$ (Cauchy). For this reason it will be indicated that if one uses a general procedure given in Jeganathan (1995, Section 4(c)), one can replace the density of ϵ_1 by a suitable estimate of it based on the entire data, and such a replacement will not affect the limiting behavior of the procedures.

The plan of the rest of the paper is as follows. In Section 2, we recall

some results on the quadratic approximation of the likelihood ratios and the associated construction of inference procedures. Section 3 deals with the model (1) under (C.1)–(C.3). In particular, the requirements of Section 2 are verified, and the consequent results are described. In Section 4, we consider a further related model called cointegrated model that is of much recent interest in econometric literatures, see for instance Engle & Granger (1987). We shall see that the techniques and results associated with the model (1) apply with suitable changes to this related model also, through it has a structure different from (1).

The following notation will be used throughout the paper. For a vector h , its transpose will be denoted by h' . For any real r , the notation $[r]$ stands for the integer part of r . When, for each $n \geq 1$, X_n is a r. v. defined on a probability space with the probability measure P_n , the statement $X_n = o_p(1)$ in P_n means $P_n(|X_n| > \epsilon) \rightarrow 0$ for all $\epsilon > 0$, that is, $\{X_n\}$ converges to 0 with respect to $\{P_n\}$. Similarly, $X_n = O_p(1)$ in P_n means $\{X_n\}$ is bounded in probability with respect to $\{P_n\}$.

17.2 Quadratic approximation

Theorem 1 will describe a specific situation under which one has a *locally asymptotically quadratic* (LAQ) approximation for the model (1), in the sense of Le Cam & Yang (1990, Section 5.2). This result is a special case of the results given in Jeganathan (1995, e.g. Theorem 13). For simplicity and in order to make the ideas transparent, the form of the result given below is intended to be applicable to the model (1) only. Note that no assumption on the nature of the roots of (2), or on the nature of the distribution of ϵ_1 are explicitly involved in the result.

For the process from (1), define $\tilde{X}_i = (X_i, X_{i-1}, \dots, X_{i-q+1})'$. Fix the parameter vector $\beta = (\beta^{(1)}, \dots, \beta^{(q)})'$. Consider the following condition.

(D.1) There are symmetric nonrandom matrices $\delta_n, n \geq 1$, which may depend on β , such that the statements

$$\sup_{1 \leq i \leq n} |\delta_n \tilde{X}_i| = o_p(1) \quad (4)$$

$$\delta_n \Sigma_{i=1}^n \tilde{X}_i \tilde{X}_i' \delta_n = O_p(1) \quad (5)$$

hold under both $P_{\beta,n}$ and $P_{\beta+\delta_n h_n, n}$ for every bounded $\{h_n\}$.

Here $P_{\beta,n}$ is the distribution of (X_1, \dots, X_n) under β . To state the result we define further

$$W_n(\beta) = -\delta_n \Sigma_{i=1}^n \tilde{X}_{i-1} \psi(\epsilon_i(\beta)), \quad (6)$$

where $\epsilon_i(\beta) = X_i - \beta' \tilde{X}_{i-1}$ and $\psi(x) = p^{(1)}(x)/p(x)$ (recall that $p(x)$ is the density of ϵ_1), and

$$A_n = \lambda^2 \delta_n \Sigma_{i=1}^n \tilde{X}_{i-1} \tilde{X}_{i-1}' \delta_n, \quad (7)$$

where λ^2 is as defined in (3). Further define the likelihood ratios

$$\Lambda_n(\beta + \delta_n h, \beta) = \log \frac{dP_{\beta+\delta_n h, n}}{dP_{\beta, n}}.$$

Theorem 1 For the model (1) with the parameter vector β fixed, assume that the condition (3) and the condition (D.1) described above are satisfied. Further assume that the density $p_0(\tilde{X}_0, \beta)$ of the initial vector $\tilde{X}_0 = (X_0, \dots, X_{1-q})$ is such that the difference $p_0(\tilde{X}_0, \beta_n) - p_0(\tilde{X}_0, \beta) = o_p(1)$ under β for every $\beta_n \rightarrow \beta$. Then the following conclusions hold:

- (a) One has $\Lambda_n(\beta + \delta_n h_n, \beta) = h'_n W_n(\beta) - 1/2 h'_n A_n h_n + o_p(1)$ and $W_n(\beta + \delta_n h_n) = W_n(\beta) - A_n h_n + o_p(1)$ in $P_{\beta, n}$ for every bounded $\{h_n\}$.
- (b) For every bounded $\{h_n\}$, the sequences $\{P_{\beta, n}\}$ and $\{P_{\beta + \delta_n h_n, n}\}$ are contiguous.

The next result shows that LAMN structure is obtained when the condition (5) of (D.1) is suitably strengthened as follows.

- (D.2) For every h , the distribution of A_n converges under both $P_{\beta, n}$ and $P_{\beta + \delta_n h, n}$ to the same distribution (that is, not depending on h) of a possibly random, almost surely positive definite, matrix A .

Corollary 1 Assume that the conditions of the preceding Theorem 1 hold, except that (5) of (D.1) is replaced by (D.2). Then, in addition to the conclusion (a) of Theorem 1, the following conclusion (D.3) holds.

- (D.3) The joint distribution of $(W_n(\beta), A_n)$ under $P_{\beta, n}$ converges in distribution to the distribution of $(A^{1/2}Z, A)$ where Z is standard normal independent of A .

Recall that when the conclusions of this corollary hold, one says that the family $\{P_{\beta + \delta_n h, n}; h \in R^q\}$ satisfies the LAMN approximation structure. Further, LAMN implies the conclusion (b) of Theorem 1. Also note that the condition (D.2) is entailed by the LAMN condition. Thus, one verifies (4) and either (D.2) or (D.3) depending on the convenience. In the present context, verification of (D.2) is more direct than that of (D.3).

It may be recalled that when ϵ_1 has finite variance and when the roots of (2) are greater than unity in absolute value (stationary case), LAMN holds with the limiting matrix A non-random. When A is non-random LAMN is called LAN. We refer to Jeganathan (1995) for a detailed discussion of wide ranging time series models where the LAN approximation holds. There, it is also shown that when ϵ_1 has finite variance and when some of the roots of the polynomial (2) are on or near the unit circle, but with no roots inside the unit circle, the conditions of Theorem 1 hold, but (D.2) fails to hold and hence LAMN fails. In a companion paper it is

shown that in many regression models that suitably incorporate the AR structure (1), with the distribution of ϵ_1 being heavy tailed as in (C.2) and with the roots of (2) being greater than unity in absolute value, the condition (4) of (D.1) as well as the LAN itself, fail to hold.

Now, as noted earlier, the density of ϵ_1 is unavailable in a usable form in the heavy tailed case, and hence $(W_n(\beta), A_n)$ will not serve as a basis for the inference procedures that will be described below. However, it is shown in Jeganathan (1995) that one can replace the function $\psi(x) = p^{(1)}(x)/p(x)$, involved in $W_n(\beta)$, and λ^2 by suitable ‘estimates’ of them based on the entire $\epsilon_i(\beta) = X_i - \beta' \tilde{X}_{i-1}$, $i = 1, \dots, n$, and the conclusions of Theorem 1 hold even with this replacement. Since the estimates are based on the usual kernal estimates of the density $p(x)$, we shall not recall the details. It is also mentioned there, though not proved, that in place of kernal estimates one can also use any reasonable estimate, at least in practice. Let us denote such an estimate of $\psi(x)$ by $\psi_n(x, \beta)$ and that of λ by $\lambda_n(\beta)$.

We shall now indicate how the inference procedures are constructed. Let $\tilde{\beta}_n$ be a preliminary estimate such that $\delta_n^{-1}(\tilde{\beta}_n - \beta)$ is bounded in probability under $P_{\beta,n}$. This $\tilde{\beta}_n$ can be obtained in the present context in many ways. For instance, first obtain the LS estimate and use it as a preliminary estimate to obtain an M-estimate through the usual one-step iteration using a convenient bounded influence function. The resulting M-estimate will satisfy the preceding requirement. Such a two stage procedure is necessary since the LS estimate does not have the right rate of convergence as it corresponds to an M-estimate with influence function that has infinite variance.

Now, let β_n^* be a discretized version of $\tilde{\beta}_n$ as described in Le Cam & Yang (1990, page 60) or Jeganathan (1995, Section 2). This means, for instance in the AR(1) case with ϵ_1 having Pareto-like tails, one computes $\tilde{\beta}_n$ only up to an approximation of order a_n where $a_n^{-1} = \Sigma X_i^2$. Note that here $\delta_n = n^{-1/2-1/\alpha}$ and both $a_n^{-1}\delta_n$ and its inverse are bounded in $P_{\beta,n}$ for every β . Further, let $\psi_n(x, \beta)$ and $\lambda_n(\beta)$ be the ‘estimates’ of $\psi(x)$ and λ respectively, as indicated above. Let

$$\hat{\beta}_n = \beta_n^* + \left(\lambda_n^2(\beta_n^*) \Sigma \tilde{X}_{i-1} \tilde{X}'_{i-1} \right)^{-1} \Sigma \tilde{X}_{i-1} \psi_n(\epsilon_i(\beta_n^*), \beta_n^*). \quad (8)$$

This estimate will be such that

$$\delta_n^{-1}(\hat{\beta}_n - \beta) = A_n^{-1} W_n(\beta) + o_p(1) \quad (9)$$

in $P_{\beta,n}$ and hence it will have many asymptotically desirable properties, see Le Cam & Yang (1990, Chapter 5) for the details.

Note that the form of (8) has close resemblance to the form of the one-step Newton-Raphson iteration of the score function using β_n^* as the starting value, but here the usual version (6) of the score function is not available since it depends on the uncomputable density $p(x)$ of ϵ_1 . Also note that

the construction (8) takes into account many simplifications induced by the special linear structure of the model (1). Such structural simplifications may not always be available, unfortunately, for instance when the model is described in terms of a nonlinear structure, and a general procedure of construction of optimal estimates in such situations is described in Le Cam (1986, Chapter 11) and Le Cam & Yang (1990, Sections 5.7–5.9).

Moreover, (9) entails that $2\Lambda_n(\hat{\beta}_n, \beta)$, the analogue of the usual likelihood ratio statistic, can be roughly viewed in the probable range of $\hat{\beta}_n$, to have the approximation

$$(\hat{\beta}_n - \beta)' (\lambda_n^2(\hat{\beta}_n) \Sigma \tilde{X}_i \tilde{X}'_i) (\hat{\beta}_n - \beta), \quad (10)$$

which is the analogue of the usual Wald statistic.

It is important to note that the construction of both (8) and (10) do not explicitly involve the index parameter α or the nature of the roots of (2) or the nature distribution of ϵ_1 . Under LAMN, the asymptotic distribution of (10) under $P_{\beta,n}$ is central Chi-Square with q degrees of freedom, so that the knowledge of the distribution of A is not required, but under the alternatives $P_{\beta+\delta_n h, n}$, the limiting distribution will be the distribution of $(Z + A^{1/2}h)'(Z + A^{1/2}h)$ where Z and A are as in Corollary 1, which is, conditional on A , non-central Chi-Square with non-centrality parameter $h'Ah$. Thus, the knowledge of the distribution of A is required in order to compute the limiting power function of the test procedures based on (10). Note however that in the context of (1), the competing procedures will also usually be of the form (10) with the limiting noncentrality parameters of the form $ch'Ah$ for suitable $c > 1$, for instance when the procedure is constructed using a suitable M-estimate of a bounded influence function, so that the knowledge of the distribution of A is again not required when such procedures are compared among themselves.

17.3 LAMN approximation

17.3.1 PRELIMINARIES

The main purpose of this section is to verify the conditions of Corollary 2 under (C.1)–(C.3). According to the discussions that follow the statement of this corollary it remains to verify (4) of (D.1), and (D.2). Note that the conclusion (D.3) of this corollary can also be verified by other means, as will be indicated below. It is convenient to split (D.2) as follows.

- (D.2a) The distribution of A_n under $P_{\beta,n}$ converges in distribution to the distribution of some almost surely positive definite matrix A .
- (D.2b) The difference between the distributions of A_n under $P_{\beta,n}$ and $P_{\beta+\delta_n h_n, n}$ converges to zero for every bounded $\{h_n\}$.

Throughout this section we fix $\beta = (\beta^{(1)}, \dots, \beta^{(q)})'$ under which the polynomial (2) has roots on the unit circle, and assume, without loss of generality, that $X_i = 0$ for $i = 1 - q, \dots, 0$.

We first describe the procedure of obtaining the normalizing quantities δ_n for which the above statement (**D.2a**) will hold. The procedure employed below for this purpose is the same as the one in Chan & Wei (1988), and is done by noting that the process (1) can be suitably decomposed into components in accordance with the nature of the roots involved, and then reducing the problem to the study of individual components. For this purpose, first note that the polynomial $\phi(z)$, specified in (2), corresponding to the above β can be written in the form

$$\phi(z) = (1 - z)^a (1 + z)^b \prod_{k=1}^m (1 - 2z \cos \theta_k + z^2)^{d_k} \quad (11)$$

for suitable integers $a, b, d_k, k = 1, \dots, m$, representing the multiplicities of the roots, such that $a + b + d_1 + \dots + d_m = q, 0 < \theta_k < \pi, k = 1, \dots, m$. Here we have used the fact that the complex roots of a polynomial with real coefficients occur in conjugate pairs. Thus if we let

$$U_i = (1 - B)^{-a} \phi(B) X_i, \quad (12)$$

where B is the backshift operator, $BX_j = X_{j-1}$, then $(1 - B)^a U_i = \phi(B) X_i = \epsilon_i$, by (1), so that $U_i, i \geq 1$, is an AR(a) process with all roots equal to 1. Similarly, $V_i = (1 + B)^{-b} \phi(B) X_i$ is an AR(b) process with all roots equal to -1 , and for each $k = 1, \dots, m$,

$$Z_{ki} = (1 - 2B \cos \theta_k + B^2)^{-d_k} \phi(B) X_i \quad (13)$$

is an AR($2d_k$) process with all roots equal to $e^{i\theta_k}$ or $e^{-i\theta_k}$.

Note that $U_j = 0, i = 0, \dots, 1 - a$, since it is assumed that $X_i = 0, i = 0, \dots, 1 - q$. Similarly the initial values of the processes $\{V_i\}$ and $\{Z_{ki}\}$ will be zero. Now define

$$\tilde{U}_j = (U_j, \dots, U_{j-a+1})', \quad \tilde{V}_j = (V_j, \dots, V_{j-b+1})' \quad (14)$$

and $\tilde{Z}_{kj} = (Z_{kj}, Z_{k,j-1}, \dots, Z_{k,j-2d_k+2}, Z_{k,j-2d_k+1})'$, for $k = 1, \dots, m$. In view of (11), there is a matrix M such that

$$\tilde{X}_j = M(\tilde{U}'_j, \tilde{V}'_j, \tilde{Z}'_{1j}, \dots, \tilde{Z}'_{mj})'. \quad (15)$$

The idea then is to find normalizing quantities separately for each of the processes (12)–(13) and then suitably combine them together. For this purpose, we need a further reduction. Define

$$U_j(k) = (1 - B)^{a-k} U_j \quad \text{for } k = 1, 2, \dots, a, \quad (16)$$

and note that, for $k = 0, 1, \dots, a - 1$,

$$U_j(k+1) = (1 - B)^{-1} U_j(k) = \sum_{s=1}^j U_s(k). \quad (17)$$

(Here and in what follows, in order to make the operations such as $(1-B)^{-1}$ meaningful, the sequence $\{\epsilon_i\}$ defined previously for $i \geq 1$ is defined for $i = 0, \pm 1, \dots$, such that $\epsilon_j = 0$ for $j \leq 0$.) Since $U_j(1) = \sum_{s=1}^j \epsilon_s$, it follows that the vector

$$\tilde{U}_j^* = (U_j(1), \dots, U_j(a))' \quad (18)$$

is a convenient functional of the partial-sum process $\sum_{j=1}^{[nt]} \epsilon_j$ for $t \in [0, 1]$, that will converge to a process in an appropriate sense after suitable normalization, see Proposition 1 below. This will lead to a suitable normalizing quantities for the process $\{\tilde{U}_j\}$, since there is a matrix M_1 such that $\tilde{U}_j = M_1 \tilde{U}_j^*$.

Similarly, for the process $\{V_j\}$, define $V_j(k) = (1+B)^{b-k} V_j$, for $k = 1, \dots, b$, and

$$\tilde{V}_j^* = (V_j(1), \dots, V_j(b))', \quad (19)$$

where $V_j(k+1) = \sum_{s=1}^j (-1)^s V_s(k)$, for $k = 0, \dots, b-1$, which are functionals of the partial-sum process $\sum_{l=1}^{[nt]} (-1)^l \epsilon_l$ for $t \in [0, 1]$, and there is a matrix M_2 such that $\tilde{V}_j = M_2 \tilde{V}_j^*$.

The analogous representations for the processes $\{\tilde{Z}_{kj}\}$, $1 \leq k \leq m$, are a little less straightforward, though similar, to explain than the preceding ones. Since the following steps are for fixed k , we shall for convenience suppress k in Z_{kj} , so that $\{Z_j\}$ stands for a process with all roots equal to $e^{i\theta}$ or $e^{-i\theta}$, $0 < \theta < \pi$. Define, in analogy with (16) and (18),

$$\begin{aligned} Z_j(r) &= (1 - (2 \cos \theta)B + B^2)^{d-r} Z_j \quad \text{for } r = 1, \dots, d \\ \tilde{Z}_j^* &= (Z_j(1), Z_j(2), \dots, Z_j(d), Z_j(d))'. \end{aligned}$$

Since there is a matrix C such that $\tilde{Z}_j = C \tilde{Z}_j^*$, we need to find the normalizing constant that stabilizes the limiting distribution of $\Sigma \tilde{Z}_j^* \tilde{Z}_j^{*\prime}$. In order to find representations analogous to (17) for this purpose, define

$$R_j(r) = \begin{pmatrix} \sum_{l=1}^j Z_l(r-1) \cos l\theta \\ \sum_{l=1}^j Z_l(r-1) \sin l\theta \end{pmatrix} = (R_{j1}(r), R_{j2}(r))', \quad (20)$$

say, and note that

$$\begin{aligned} Z_j(r) &= (1 - (2 \cos \theta)B + B^2)^{-1} Z_j(r-1) \\ &= \frac{1}{\sin \theta} \sum_{l=1}^j Z_l(r-1) \sin(j-l+1)\theta \\ &= \frac{1}{\sin \theta} [R_{j1}(r) \sin(j+1)\theta - R_{j2}(r) \cos(j+1)\theta]. \end{aligned}$$

Using this, and applying familiar trigonometric identities (Chan & Wei 1988, Lemma 3.3.2), obtain the following identities for the elements of the

matrix $\Sigma \tilde{Z}_j^* \tilde{Z}_j^{*\prime}$:

$$\begin{aligned} 2 \sin^2 \theta \Sigma_{l=1}^j Z_l(s) Z_l(r) &= \Sigma_{l=1}^j [R_{l1}(s) R_{l1}(r) + R_{l2}(s) R_{l2}(r)] \\ &- \Sigma_{l=1}^j [R_{l1}(s) R_{l2}(r) + R_{l2}(s) R_{l1}(r)] \sin(2l+2)\theta \\ &- \Sigma_{l=1}^j [R_{l1}(s) R_{l1}(r) - R_{l2}(s) R_{l2}(r)] \cos(2l+2)\theta \end{aligned} \quad (21)$$

and

$$\begin{aligned} 2 \sin^2 \theta \Sigma_{l=1}^j Z_{l-1}(s) Z_l(r) \\ = \Sigma_{l=1}^j [R_{l-1,2}(s) R_{l2}(r) + R_{l-1,1}(s) R_{l1}(r)] \cos \theta \\ - \Sigma_{l=1}^j [R_{l-1,2}(s) R_{l1}(r) - R_{l-1,1}(s) R_{l2}(r)] \sin \theta \\ - \Sigma_{l=1}^j [R_{l-1,1}(s) R_{l1}(r) - R_{l-1,2}(s) R_{l2}(r)] \cos(2l+1)\theta \\ - \Sigma_{l=1}^j [R_{l-1,1}(s) R_{l2}(r) + R_{l1}(r) R_{l-1,2}(s)] \sin(2l+1)\theta. \end{aligned} \quad (22)$$

In addition, note that, in view of Chan & Wei (1988, Lemma 3.3.1), $R_j(r)$ satisfies the recurrence relationship:

$$R_j(r) = \Sigma_{l=1}^j (\delta + \delta_{l1}) R_l(r-1), \quad (23)$$

where

$$\delta = \frac{1}{2 \sin \theta} \begin{pmatrix} \sin \theta & -\cos \theta \\ \cos \theta & \sin \theta \end{pmatrix}$$

and

$$\delta_{l1} = \frac{1}{2 \sin \theta} \begin{pmatrix} \sin(2l+1)\theta & -\cos(2l+1)\theta \\ -\cos(2l+1)\theta & \sin(2l+1)\theta \end{pmatrix}.$$

Now note that, in view of (20) and (23) and since $Z_j(0) = \epsilon_j$, $R_j(r)$ and hence \tilde{Z}_j^* can be written as functionals of the partial-sum process $(\Sigma_{j=1}^{[nt]} \epsilon_j \cos j\theta, \Sigma_{j=1}^{[nt]} \epsilon_j \sin j\theta)$.

17.3.2 VERIFICATION OF D.2a

The basic assumption we need throughout this section is that there are constants $a_n \uparrow \infty$ such that $a_n^{-1} \Sigma_{j=1}^n \epsilon_j$ converges to a (symmetric) stable r. v. Note that we require that the centering constants b_n that stabilize the limiting distribution of $a_n^{-1} \Sigma_{j=1}^n \epsilon_j - b_n$ are identically or suitably approximately zero. The constants a_n are necessarily of the form $a_n = n^{1/\alpha} L_1(n)$ for a suitable constant $0 < \alpha < 2$ and a slowly varying function $L_1(x)$. Under the condition (C.2) of Section 1, the preceding requirements are satisfied. See Feller (1971) for details. Note that the requirement that distribution of ϵ_1 is symmetric around zero that is imposed in (C.2) can also be replaced by other conditions (Knight 1991, page 202).

For notational convenience, for the rest of the paper we take $a_n = n^{1/\alpha}$.

It is convenient first to state the following results, of which the first one gives the joint limiting process of the partial-sum processes.

Proposition 1 Assume that the condition (C.2) stated in Section 1 is satisfied. Define $\tilde{S}_n(t)$ for $t \in [0, 1]$, as the $(2m + 2)$ -vector process

$$n^{-1/\alpha} \Sigma_{j=1}^{[nt]} (\epsilon_j, (-1)^j \epsilon_j, \epsilon_j \cos j\theta_1, \epsilon_j \sin j\theta_1, \dots, \epsilon_j \cos j\theta_m, \epsilon_j \sin j\theta_m)',$$

where ϵ_j are as in (1). Then the process $\tilde{S}_n(t)$ converges in the Skorokhod space $D([0, 1] \rightarrow R^{2m+2})$ to a symmetric stable process $\tilde{S}(t)$ with index α .

The preceding result serves the purpose similar to that of Chan & Wei (1988, Theorem 2.2). The next result serves the purpose similar to that of Chan & Wei (1988, Theorem 2.1) or Jeganathan (1991, Proposition 8).

Proposition 2 Assume that $U_n(t)$, $n \geq 1$, are processes in the Skorokhod space $D([0, 1] \rightarrow R)$ converging in distribution to the process $U(t)$. Then, $\sup_{0 \leq t \leq 1} |n^{-1} \Sigma_{l=1}^{[nt]} e^{il\theta} U_n(l/n)| = o_p(1)$, for each θ such that $e^{i\theta} \neq 1$.

Proofs of Propositions 1 and 2 will be given later in Subsection 17.3.4. Once these results, together with the preliminaries of Subsection 17.3.1 are given, the verification of (D.2a) is similar to the steps involved for the finite variance case given by Chan & Wei (1988), when the appropriate modifications used in Jeganathan (1991) are taken into account. For this purpose, let us denote the limiting process $\tilde{S}(t)$ of Proposition 1 by

$$\tilde{S}(t) = (S_{01}(t), T_{01}(t), S_{11}(t), T_{11}(t), \dots, S_{m1}(t), T_{m1}(t))'. \quad (24)$$

First consider the process $\{\tilde{U}_j^*\}$, defined in (18). In view of (17), it follows that $n^{-1/\alpha} U_{[nt]}(1) = n^{-1/\alpha} \Sigma_{j=1}^{[nt]} \epsilon_j$, and, for $k = 1, \dots, a - 1$,

$$n^{-k-1/\alpha} U_{[nt]}(k+1) = \int_0^t n^{-k+1-1/\alpha} U_{[ns]}(k) ds,$$

so that the a -vector process $(n^{-k-1/\alpha} U_{[nt]}(k+1), k = 0, 1, \dots, a-1)$ converges in law in $D([0, 1] \rightarrow R^a)$ to the process $(S_{0k}(t), k = 1, \dots, a)$, where $S_{01}(t)$ is as in (24) and, for $k = 2, \dots, a$,

$$S_{0k}(t) = \int_0^t S_{0,k-1}(s) ds = \int_0^t S_{01}(s) \frac{(t-s)^{k-2}}{(k-2)!} ds.$$

Thus, if we choose

$$N_{1n} = \text{diag}\{n^{-1/2-1/\alpha}, \dots, n^{-a+1/2-1/\alpha}\}, \quad (25)$$

then it follows that

$$N_{1n} \Sigma_{j=1}^n \tilde{U}_j^* \tilde{U}_j^{*''} N_{1n} \xrightarrow{\mathcal{L}} \Gamma_1, \quad (26)$$

where (k, i) th element of Γ_1 is given by $\int_0^1 S_{0k}(t) S_{0i}(t) dt$ for $k, i = 1, \dots, a$. It similarly follows that, if we choose

$$N_{2n} = \text{diag}\{n^{-1/2-1/\alpha}, \dots, n^{-b+1/2-1/\alpha}\}, \quad (27)$$

then

$$N_{2n} \Sigma_{j=1}^n \tilde{V}_j^{*} \tilde{V}_j'^* N_{2n} \xrightarrow{\mathcal{L}} \Gamma_2, \quad (28)$$

where (k, i) th element of Γ_2 equals $\int_0^l T_{0k}(t) T_{0i}(t) dt$ for $k, i = 1, \dots, b$, where

$$T_{0i}(t) = \int_0^t T_{01}(s) \frac{(t-s)^{i-2}}{(i-2)!} ds, \quad i = 2, \dots, b.$$

We now consider, for fixed k , the process $\{\tilde{Z}_{kj}\}$. We now denote $R_j(r)$, defined in (20), by $R_{kj}(r)$ when θ involved there is replaced by θ_k . Then, since $Z_{kj}(0) = \epsilon_j$ and in view of (20), $n^{-1/\alpha} R_{k[nt]}(1) \xrightarrow{\mathcal{L}} (S_{k1}(t), T_{k1}(t))$, where the limit is as in (24). Now, suppose that for fixed $i \geq 1$,

$$n^{-i+1-1/\alpha} R_{k[nt]}(i) \xrightarrow{\mathcal{L}} (S_{ki}(t), T_{ki}(t)), \quad (29)$$

say. Then

$$n^{-i-1/\alpha} \Sigma_{j=1}^{[nt]} \delta R_{kj}(i) \xrightarrow{\mathcal{L}} \delta \left(\int_0^t S_{ki}(s) ds, \int_0^t T_{ki}(s) ds \right),$$

and $\sup_{0 \leq t \leq 1} |n^{-i-1/\alpha} \Sigma_{j=1}^{[nt]} \delta_{j1} R_{kj}(i)| = o_p(1)$, by Proposition 2, where the matrices δ and δ_{j1} are as in (23). Thus in view of (23),

$$\begin{aligned} n^{-i-1/\alpha} R_{k[nt]}(i+1) &\xrightarrow{\mathcal{L}} \delta \left(\int_0^t S_{ki}(s) ds, \int_0^t T_{ki}(s) ds \right) \\ &= (S_{k,i+1}(t), T_{k,i+1}(t)), \text{ say.} \end{aligned}$$

Thus, by induction, (29) is true for all $i = 1, \dots, d_k$. This implies that the first of the three sums involved in the right-hand side of (21), for $\theta = \theta_k$, converges in distribution, when multiplied by $(2 \sin^2 \theta_k)^{-1} n^{-s-r+1-2/\alpha}$, to

$$\frac{1}{2 \sin^2 \theta_k} \left[\int_0^1 S_{ks}(t) S_{kr}(t) dt + \int_0^1 T_{ks}(t) T_{kr}(t) dt \right], \quad (30)$$

and, by Proposition 2, the second and third sums of the r.h.s. of (21) converge to zero in probability. Similarly, the l.h.s. of (22), for $\theta = \theta_k$, converges in distribution, when multiplied by the same factor, to

$$\begin{aligned} \frac{1}{2 \sin^2 \theta_k} &\left(\cos \theta_k \left(\int_0^1 T_{ks}(t) T_{kr}(t) dt + \int_0^1 S_{ks}(t) T_{kr}(t) dt \right) \right. \\ &\left. - \sin \theta_k \left(\int_0^1 T_{ks}(t) S_{kr}(t) dt - \int_0^1 S_{ks}(t) T_{kr}(t) dt \right) \right). \quad (31) \end{aligned}$$

Thus, if we choose

$$N_{k+2,n} = \text{diag}\{n^{-1/2-1/\alpha}, n^{-1/2-1/\alpha}, \dots, n^{-d_k+1/2-1/\alpha}, n^{-d_k+1/2-1/\alpha}\},$$

then

$$N_{k+2,n} \Sigma_{j=1}^n \tilde{Z}_{kj}^* \tilde{Z}_{kj}^{*''} N_{k+2,n} \xrightarrow{\mathcal{L}} H_k = (\sigma_{lm}), \quad (32)$$

where $\sigma_{2s-1,2r-1} = \sigma_{2s,2r}$ is given by (30), and $\sigma_{2s-1,2r} = \sigma_{2r,2s-1}$ is given by (31).

Now, we have $\tilde{X}_j = D(\tilde{U}_j^{*''}, \tilde{V}_j^{*''}, \tilde{Z}_{1j}^{*''}, \dots, \tilde{Z}_{mj}^{*'''})' = D\tilde{X}_j^*$, say, for a suitable matrix D . Let $N_n = \text{diag}\{N_{1n}, N_{2n}, \dots, N_{m+2,n}\}$, and choose $\delta_n = N_n D^{-1}$, where $N_{k+2,n}$, for $k \geq 1$, are as above and N_{1n} and N_{2n} are as in (25) and (27). Then

$$\delta_n \Sigma_{j=1}^n \tilde{X}_j \tilde{X}_j' \delta_n = N_n \Sigma_{j=1}^n \tilde{X}_j^* \tilde{X}_j^{*''} N_n. \quad (33)$$

Now, it can be easily seen that the off-diagonal submatrices of (33) that are in the form of sums of the terms such as $\tilde{U}_j^* \tilde{V}_j^{*'}$, $\tilde{U}_j^* \tilde{Z}_{kj}^{*'}$, etc, can be written in the form in which Proposition 2 can be applied, so that such off-diagonal submatrices converge to zero in probability. Thus we have proved the following

Proposition 3 *The quantity $\delta_n \Sigma_{j=1}^n \tilde{X}_j \tilde{X}_j' \delta_n$ converges in distribution under $P_{\beta,n}$ to $\text{diag}\{\Gamma_1, \Gamma_2, H_1, \dots, H_m\}$, where δ_n is as defined above and the submatrices involved in the limit are as defined in (26), (28), and (32). Further, these sub-matrices are positive definite almost surely.*

Thus the limiting matrix A involved in (D.2a) is given by the limit from Proposition 3 multiplied by λ^2 with λ^2 as in (C.3). Here, the fact that the matrices involved are positive definite can be easily shown in many ways. For instance, suppose the contradiction for Γ_1 . Then there is an event B such that $P(B) > 0$ and Gamma_1 is zero on B. This means the vector process $\tilde{S}_0(t) = (S_{01}(t), \dots, S_{0a}(t))$, $0 \leq t \leq 1$, is zero for almost all trajectories on the event B, in particular the process $\tilde{S}_0(t)$ has differentiable trajectories on B. This is a contradiction since $\tilde{S}_0(t)$ is a stable process, so that $P(B) = 0$.

It may also be noted that in the finite variance case the component matrices involved in the limit from Proposition 3 will be mutually independent since the components of the vector (24) are mutually independent. Unfortunately, this is not true for the present infinite variance case (Sub-section 17.3.4).

17.3.3 VERIFICATION OF D.2b

We now consider the verification of (D.2b), the intent of which is to obtain the mixed normal structure of the limit specified in (D.3) of Corollary 1. We first briefly indicate that (D.3) can also be verified by other means. Consider the process $(\tilde{Y}_n(t), \tilde{S}_n(t))$, where $\tilde{S}_n(t)$ is as in Proposition 1, and $\tilde{Y}_n(t)$ is defined in the same way as $\tilde{S}_n(t)$ except that ϵ_j are replaced by $\psi(\epsilon_j)$ with $\psi(x)$ as involved in (6). Note that $\tilde{Y}_n(t)$ is a partial sum process

of independent variables with zero means and finite variances. Then, essentially the same proof of Proposition 1, to be given below, will show that the process $(\tilde{Y}_n(t), \tilde{S}_n(t))$ converges in distribution in $D([0, 1] \rightarrow R^{4m+4})$ to a Lévy process of which the limit $\tilde{Y}(t)$ of $\tilde{Y}_n(t)$ will be Gaussian and $\tilde{S}(t)$, that of $\tilde{S}_n(t)$, will be a stable process composed exclusively of Poisson components, and hence these two limits must be independent, see, for example, Loève (1978, page 533). Note that the components of $\tilde{Y}(t)$ will be independent Brownian motions. It can be seen, using the computations given in Chan & Wei (1988, Lemma 3.3.2) that each component of $W_n(\beta)$, as defined in (6), can be written as a stochastic integral of a suitable functional of $\tilde{S}_n(t)$ with respect to a suitable component of the process $\tilde{Y}_n(t)$. Thus using the ideas of the proof of Proposition 3 together with the results given, for example, in Kurtz & Protter (1991), it will follow that $W_n(\beta)$ converges in distribution jointly with A_n . In addition, using the fact that the limits $\tilde{S}(t)$ and $\tilde{Y}(t)$ are independent, the weak limit of $W_n(\beta)$ can be written in the form $A^{1/2}Z$, where Z is standard q -variate normal independent of the weak limit A of A_n . This will verify the condition (D.3) of Corollary 1.

In what follows, instead of elaborating the preceding sketch, we shall verify (D.2b) directly for many reasons. First, the preceding sketch requires that the process $\tilde{S}_n(t)$ converges jointly with $\tilde{Y}_n(t)$, whereas the requirement (D.2b) might also be of interest in situations where the convergence of $\tilde{Y}_n(t)$ may not be required, for instance when obtaining the limiting distribution of M-estimates under the alternatives.

Second, once (D.2a) is verified, it is usually easy to check whether or not (D.2b) holds at least in specific cases. In fact, in the cointegration model to be treated in Section 4 below, (D.2b) will trivially hold since the distribution of A_n under $P_{\beta,n}$ will be functionally independent of the parameter β .

We first illustrate the verification of (D.2b) in the specific AR(1) case, since the basic idea for the general case will remain the same. In the AR(1) case one has, under the parameter $\beta_n = 1 + h_n n^{-1/2-1/\alpha}$ with $\{h_n\}$ bounded, $X_i = \beta_n X_{i-1} + \epsilon_i$ for $i = 1, \dots, n$, where $X_0 = 0$ and ϵ_i are i.i.d. as in (1). Then

$$X_i = \sum_{l=1}^i \epsilon_l (1 + h_n n^{-1/2-1/\alpha})^{i-l} = \sum_{l=1}^i \epsilon_l \exp(h_n^*(i-l)n^{-1/2-1/\alpha}),$$

for suitable h_n^* such that $h_n^* - h_n \rightarrow 0$. For these X_i 's, we need to show that the asymptotic distribution of the variable $n^{-1-2/\alpha} \sum_{i=1}^n X_i^2$ under h_n will coincide with that of $n^{-1-2/\alpha} \sum_{i=1}^n (\sum_{j=1}^i \epsilon_j)^2$. This will follow if the processes $n^{-1/\alpha} \sum_{i=1}^{[nt]} \epsilon_i$ and $n^{-1/\alpha} X_{[nt]}$ converge in $D([0, 1] \rightarrow R)$ to one and the same limit. To see that this is true, let

$$\epsilon_{nj} = n^{-1/\alpha} \epsilon_j \exp(-h_n^* j n^{-1/2-1/\alpha}). \quad (34)$$

Then note that, by the definition of X_i ,

$$n^{-1/\alpha} X_{[nt]} = \exp(h_n^*[nt]n^{-1/2-1/\alpha}) \Sigma_{j=1}^{[nt]} \epsilon_{nj}, \quad (35)$$

where

$$\sup_{0 \leq t \leq 1} |1 - \exp(h^*[nt]n^{-1/2-1/\alpha})| \rightarrow 0 \quad (36)$$

since $0 < \alpha < 2$. In addition, it is well-known that the process $\Sigma_{j=1}^{[nt]} \epsilon_{nj}$ converges in $D([0, 1] \rightarrow R)$ to a suitable Lévy process with no Gaussian component if and only if

$$\Sigma_{j=1}^{[nt]} P(|\epsilon_{nj}| > x) \quad (37)$$

converges to a suitable limit for all $0 \leq t \leq 1$ and $x > 0$, and

$$\Sigma_{j=1}^n E(|\epsilon_{nj}|^2 I(|\epsilon_{nj}| < \epsilon)) \rightarrow 0 \quad (38)$$

for all $\epsilon > 0$, where $I(F)$ denotes the indicator function of the event F . (Here, we have taken into account of the assumption that the distributions of ϵ_j 's, and hence ϵ_{nj} 's, are symmetric around zero). In addition, the limiting process is determined by the limit of (37).

Note that, by the condition (C.2), the convergences in (37) and (38) hold when $h_n \equiv 0$, that is, when ϵ_{nj} are replaced by $n^{-1/\alpha} \epsilon_j$. Further note that, by (34), the ϵ_{nj} are of the form $n^{-1/\alpha} \epsilon_j a_{nj}$ where a_{nj} are such that $\sup_j |1 - a_{nj}| \rightarrow 0$. Thus, the convergences in (37) and (38) hold with the limits remaining the same for every bounded $\{h_n\}$ on which ϵ_{nj} depends. This, together with (35) and (36), then verifies (D.2b).

Now consider the general case where $\{X_i\}$ is an AR(q) model (1) with the parameter vector given by the form $\beta + \delta_n h_n$ where the characteristic polynomial corresponding to β is of the form (11) with unit roots. In this case, it has been shown in Jeganathan (1991) that the process $\{X_i\}$ has a decomposition analogous to (15) such that the variables U_i, V_i, Z_{ki} , $k = 1, \dots, m$, defined suitably analogous to (12)–(13) are suitable AR processes with the parameters contiguous to those that correspond to the parameter β . For instance, $\{U_i\}$ is an AR(a) process with the parameter of the form $\alpha + N_{1n}^{-1} M_1^{-1} w_n$ for a bounded $\{w_n\}$, where the characteristic polynomial corresponding to the parameter α has all roots equal to +1, N_{1n} is as defined in (25) and M_1 is as defined earlier such that $\tilde{U}_j = M_1 \tilde{U}_j^*$ with \tilde{U}_j and \tilde{U}_j^* are defined respectively as in (14) and (18). Thus

$$\begin{aligned} \epsilon_j &= U_j - (\alpha + N_{1n}^{-1} M_1^{-1} w_n)' \tilde{U}_{j-1} \\ &= (1 - B)^a U_j - w_n' N_{1n}^{-1} \tilde{U}_{j-1}^* \\ &= U_j(0) - n^{-1/2-1/\alpha} w_{1n} U_{j-1} - \dots - n^{-a+1/2-1/\alpha} w_{an} U_{j-1}(a) \end{aligned}$$

with $U_j(k)$ defined as in (16) and $w_n = (w_{1n}, \dots, w_{an})'$.

The procedure then is to verify (D.2b) for each of the above indicated component AR processes of $\{X_i\}$, which will lead to the verification for $\{X_i\}$ itself. In order to simplify the presentation, we shall illustrate the verification for the component process $\{U_i\}$ only. The idea we shall use is the same as that used in AR(1) case above in the sense that we express the sum $n^{-1/\alpha} \sum U_j(0)$ in the form of (35) and apply the same procedure used in the AR(1) case. For this purpose, let $\mathbf{U}_{jn} = (U_j(1), n^{-1}U_{j-1}(2), \dots, n^{-a+1}U_{j-a}(a))'$. Then, it follows from Jeganathan (1991, equation (34)) that $\mathbf{U}_{jn} = (I + n^{-1}A_n(w_n))\mathbf{U}_{j-1,n} + \tilde{\epsilon}_j$, where $\tilde{\epsilon}_j = (\epsilon_j, 0, \dots, 0)'_{1 \times a}$ and

$$A_n(w_n) = \begin{bmatrix} w_{1n}^* & \dots & \dots & \dots & w_{an}^* \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ n^{-1} & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ n^{1-a} & \dots & \dots & n^{-1} & 1 \end{bmatrix}$$

with $w_{in}^* = w_{in}n^{1/2-1/\alpha}$. Then, similar to (35), one has

$$n^{-1/\alpha} \mathbf{U}_{in} = \sum_{j=1}^i n^{-1/\alpha} \tilde{\epsilon}_j \exp(A_n^*(w_n)(i/n - j/n)) \quad (39)$$

for suitable A_n^* . Note that when $w_n = 0$, the first component of this vector is $n^{-1/\alpha} \sum_{j=1}^i \epsilon_j$. From the definition of w_{in}^* and the fact that $0 < \alpha < 2$,

$$\sup_{-n \leq j \leq n} |\exp(A_n^*(w_n)j/n) - \exp(A_n^*(0)j/n)| \rightarrow 0$$

so that the first component of (39) can be written as $n^{-1/\alpha} \sum_{j=1}^i \gamma(i, j)\epsilon_j$ with $\sup_{i,j} |\gamma(i, j) - 1| \rightarrow 0$. This will imply, as in the AR(1) case, that the processes $n^{-1/\alpha} \sum_{j=1}^{[nt]} \epsilon_j$ and $n^{-1/\alpha} \sum_{j=1}^{[nt]} U_j(0)$ will converge in distribution to one and the same process. This completes the verification of (D.2b).

17.3.4 REMARKS ON PROPOSITIONS 1 AND 2

For the proof of Proposition 1, let $\tilde{S}_n^*(t)$ be the $(2m+2)$ -vector process defined in the same way as $\tilde{S}_n(t)$, except that ϵ_j involved in $\tilde{S}_n(t)$ are replaced by ϵ_j^* , where ϵ_j^* are i.i.d. symmetric stable with index α . Then it follows from the arguments of the proof of Klüppelberg & Mikosch (1993, Proposition 2.2) that the convergence of the finite dimensional distributions of one of the processes $\tilde{S}_n(t)$ or $\tilde{S}_n^*(t)$ will entail the convergence for the other with the same limits. In addition, in the present case, the convergence of finite dimensional distributions of these processes will entail (Skorohod 1957) their convergence in the Skorokhod space $D([0, 1] \rightarrow R^{2m+2})$.

Now, the convergence in distribution of $\tilde{S}_n^*(t)$ to $\tilde{S}(t)$ of (24) for each fixed t has essentially been established in Klüppelberg & Mikosch (1993, Theorem 2.4), specifically when $t = 1$ and when the first two components

of the vector $\tilde{S}_n^*(t)$ are omitted, but their arguments hold more generally for $\tilde{S}_n^*(t)$ also. In addition, the explicit representation of the characteristic function of the weak limit $\tilde{S}(t)$ can also be essentially found there, from which it follows that the limit is a stationary R^{2m+2} valued α -stable process whose components are unfortunately not mutually independent, see Definition 2.3 of that paper. Unfortunately the characteristic function of the limiting process takes a rather complicated form depending on whether the θ_j 's involved in the definition of $\tilde{S}_n^*(t)$ are rational or irrational as well as on the “linear dependence” of the irrational θ_j 's.

In the present situation our direct interest is on the distribution of the limit from Proposition 3. Unfortunately it is not clear if any form of explicit representation for this limit distribution is obtainable from the representation of the limit $\tilde{S}(t)$ indicated above. It appears to me that a simpler and more accurate way of approximating the distribution of $\tilde{S}(t)$ or any of its functionals would be to simulate the distribution of $\tilde{S}_k^*(t)$ or its functional for a suitable large k , which is feasible since $\tilde{S}_k^*(t)$ is a partial sum explicitly involving i.i.d. symmetric stable random variables. It is also important to note that the asymptotic distributions of test statistics such as Wald's will be central χ^2 under the null hypothesis, so that the preceding difficulties do not arise. For these reasons, we shall not go into further details of possible representations of the limiting distribution of $\tilde{S}_n^*(t)$.

We next consider the proof of Proposition 2. When the limiting process $U(t)$ is such that almost all trajectories of it are continuous, this result reduces to Jeganathan (1991, Proposition 8), and the proof given there can be adopted to the present more general case as follows.

Proof of proposition 2: First note that for every $0 < u < v < 1$,

$$|\Sigma_{k=[nu]}^{[nv]} e^{ik\theta}| = \left| \left(1 - e^{i([nv]-[nu])\theta} \right) e^{i[nu]\theta} / (1 - e^{i\theta}) \right| \leq C. \quad (40)$$

Now, fix an integer m , and let

$$0 = t_0 < t_1 < \dots < t_r = 1 \quad \text{and } t_p - t_{p-1} > 1/m, \quad p = 1, \dots, r \quad (41)$$

For a given t , let j be such that $t_{j-1} < t \leq t_j$. For the moment, for simplicity of notation, let t be such that $t = t_j$. Then $n^{-1} \Sigma_{k=1}^{[nt]} e^{ik\theta} U_n(k/n)$ equals

$$\begin{aligned} n^{-1} \Sigma_{p=1}^j U_n(t_{p-1}) \Sigma_k \{ e^{ik\theta} : [nt_{p-1}] < k \leq [nt_p] \} \\ + n^{-1} \Sigma_{p=1}^j \Sigma_k \{ e^{ik\theta} (U_n(k/n) - U_n(t_{p-1})) : [nt_{p-1}] < k \leq [nt_p] \}, \end{aligned}$$

which, by (40), is bounded in absolute value by $(Cm/n) \sup_{0 \leq t \leq 1} |U_n(t)| + \inf_{\{t_i\}} \max_{0 < p < r} \sup_{t_{p-1} < s \leq t_p} |U_n(s) - U_n(t_{p-1})| = I_n(n, m) + I_2(n, m)$, say, where the inf extends over all finite sets $\{t_i\}$ that satisfy (41). It is easy to see that this bound remains the same even when $t \neq t_j$ but $t_{j-1} < t < t_j$. Now, since $\sup_t |U_n(t)|$ is bounded in probability, $\lim_{m \rightarrow \infty} \limsup_{n \rightarrow \infty} P(I_1(n, m) > \epsilon) = 0$ for all $\epsilon > 0$. Further, since $\{U_n(t)\}$ converges in law in $D([0, 1] \rightarrow R)$, the same conclusion holds for $I_2(n, m)$ (Billingsley 1968, Theorem 15.2). This completes the proof.

17.4 Regressions with integrated processes

In this section we consider a regression model with regressors forming an AR process with unit roots. Such a model is of much recent interest in econometric literatures. Specifically, in a typical form of this model the observations consist of vectors $(Y_i, X_i), i = 1, \dots, n$, such that

$$Y_i = \beta' X_i + \eta_i \quad (42)$$

and $\{X_i\}$ is an AR model with unit roots, where β is the unknown parameter. For simplicity we take Y_i 's to be real valued and $\{X_i\}$ forming a q -variate first order AR model with unit roots of the simplest form

$$X_i = X_{i-1} + \epsilon_i. \quad (43)$$

In addition, (η_i, ϵ_i) are assumed to be i.i.d. with a $(q+1)$ - variate Lebesgue density $p(u, v), u \in R$ and $v \in R^q$.

Here, the nonstationarity of the multiple time series (Y_i, X_i) is driven by the integrated process $\{X_i\}$, in such a way that a suitable linear combination of the components of (Y_i, X_i) is stationary, that is, cointegrated, represented by (42). Thus the parameter β represents a possible cointegrating relationship. Models of this form are also called cointegrated models, a general discussion of which can be found in Engle & Granger (1987). See also the recent book by Hamilton (1994) and the references contained therein for discussions of models more general than that described by (42) and (43).

When the vectors (η_i, ϵ_i) have finite variances, a study of the asymptotic behavior of the LS estimate based on the single equation (42) can be found in many places, see for instance, Stock (1987), Phillips & Durlauf (1986) and Park & Phillips (1988), with the limiting distribution involved having the form of nonstandard 'unit root distribution'. The asymptotic properties of the Gaussian procedure, that is, the maximum likelihood estimate of β obtained assuming that (η_i, ϵ_i) are i.i.d. Gaussian, has been studied in Johansen (1988) where it is shown that the limiting distribution of such an estimate is mixed normal, analogous to the limit of efficient estimates of the LAMN family of Section 2. In addition, it has also been shown there that the likelihood ratio tests, obtained again under the Gaussian assumption, of various hypotheses regarding the parameter β have limiting χ^2 distributions. Similar results can also be found in Ahn & Reinsel (1990) and Reinsel & Ahn (1992). Independently, it is shown in Jeganathan (1995, Section 6(a)) that the family of distributions generated by the observations $(Y_i, X_i), i = 1, \dots, n$, comes under the LAMN approximation structure, assuming only that the condition (44) given below is satisfied. Since one can take the LS estimate based on the single equation (42) as a preliminary estimate, the method of obtaining asymptotic optimal procedures described in Section 2 is applicable here also. In addition, it is indicated there that

the joint density of (η_1, ϵ_1) that will be involved in this optimal procedure can be replaced by an estimate of it, similar to the situation described in Section 2. In the context of broader cointegrated models and under the Gaussian assumption, a discussion of the LAMN structure for the model described by (42) and (43) can be found in Phillips (1991).

We shall now indicate that the preceding results hold even when the distribution of ϵ_1 is assumed to have a heavy tailed (infinite variance) distribution satisfying (C.2), assuming for simplicity that ϵ_1 is real-valued. In addition, we shall also indicate that the reasons for the LAMN structure are the same both for the finite variance and infinite variance cases. For this purpose we shall assume that the joint density $p(u, v)$ of the vector (η_1, ϵ_1) is such that

$$0 < \lambda^2 = \int |\psi(u, v)|^2 p(u, v) du dv < \infty, \quad (44)$$

where $\psi(u, v) = (p(u, v))^{-1} \frac{\partial p(u, v)}{\partial u}$. As before $P_{\beta, n}$ stands for the joint distribution of the observations (Y_i, X_i) , for $i = 1, \dots, n$.

Theorem 2 Assume that the distribution of ϵ_1 satisfies the assumption (C.2) stated in Section 1. Further assume that the above condition (44) holds. Let $\delta_n = n^{-1/2-1/\alpha}(L_1(n))^{-1}$ with α as in (C.2) and $L_1(n)$ as described at the beginning of Subsection 17.3.2. Further let $W_n(\beta)$ equal $-\delta_n \sum_{i=1}^n X_{i-1} \psi(Y_i - \beta X_i, X_i - X_{i-1})$, with the function ψ as in (44), and A_n equal $\lambda^2 \delta_n^2 \sum_{i=1}^n X_{i-1}^2$. Then the conclusions of Corollary 1 hold for the present situation also.

We shall now briefly indicate the reasonings for the validity of this result. First, consider the model in which, instead of (42), one has $Y_i = \beta X_{i-1} + \eta_i$ with X_i as in (43). Then the process $\{(Y_i, X_i)\}$ is a multivariate first order AR process in the sense that

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \begin{pmatrix} 0 & \beta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} Y_{i-1} \\ X_{i-1} \end{pmatrix} + \begin{pmatrix} \eta_i \\ \epsilon_i \end{pmatrix}. \quad (45)$$

Now one can readily see that all the results of Section 2 that hold for the univariate AR models suitably extend to multivariate AR models. In particular this is true for the model (45) in view of (44), and the quantity analogous to (6) is given by δ_n times the first derivative, whenever it exists, of $\sum_{i=1}^n \log p(Y_i - \beta X_{i-1}, X_i - X_{i-1})$ with respect to β , which is approximately given by W_n , and the quantity analogous to (7) is A_n .

Now, the convergence in distribution of A_n , follows from the results of Section 3, since $X_i = \epsilon_1 + \dots + \epsilon_i$. Further note that the distribution of (X_1, \dots, X_n) and hence of A_n does not depend on the parameter β . Thus the condition (D.2) of Section 2 hold trivially in the present case. Similarly $\sup_{1 \leq i \leq n} |\delta_n X_i| = \sup_{0 \leq t \leq 1} |n^{-1/\alpha} X_{[nt]}| / \sqrt{n} = O_p(1/\sqrt{n})$, since the process $n^{-1/\alpha} X_{[nt]}$ converges in distribution. Hence the requirement (4)

of (D.1) of Section 2 is also satisfied in the present situation. Thus Theorem 2 holds when (Y_i, X_i) satisfies the model (45). Though the model described by (42) and (43) is not exactly the same as (45), it is not difficult to see that the validity of Theorem 2 for one of these models will entail the same for the other one with the same approximation. The details of this statement are implicit in Jeganathan (1995, Section 6a) where a somewhat direct proof of Theorem 2 can be found; the proof given there is restricted to the finite variance case but the proof with suitable modifications holds for the infinite variance case also.

17.5 REFERENCES

- Ahn, S. K. & Reinsel, G. C. (1990), 'Estimation for partially nonstationary multivariate autoregressive models', *Journal of the American Statistical Association* **85**, 813–823.
- Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley, New York.
- Chan, N. H. (1990), 'Inference for near integrated time series with infinite variance', *Journal of the American Statistical Association* **89**, 1069–1074.
- Chan, N. H. & Tran, L. T. (1989), 'On the first order autoregressive process with infinite variance', *Econometric Theory* **5**, 354–362.
- Chan, N. H. & Wei, C. Z. (1988), 'Limiting distributions of least squares estimates of unstable autoregressive process', *Annals of Statistics* **16**, 367–401.
- Davis, R. A. & Resnick, S. (1986), 'Limit theory for sample autocorrelation function of moving averages', *Annals of Statistics* **14**, 533–558.
- Davis, R. A., Knight, K. & Liu, J. (1992), 'M-estimation for autoregressions with infinite variance', *Stochastic Processes and their Applications* **40**, 145–180.
- DuMouchel, W. H. (1973), 'On the asymptotic normality of the maximum likelihood estimate when sampling from a stable distribution', *Annals of Statistics* **1**, 948–957.
- Engle, R. F. & Granger, C. W. J. (1987), 'Co-integration and error correction: Representation, estimation, and testing', *Econometrica* **55**, 251–276.
- Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, Vol. 2, second edn, Wiley, New York.

- Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press.
- Jeganathan, P. (1991), 'On the asymptotic behavior of least-squares estimators in AR time series with roots near the unit circle', *Econometric Theory* 7, 269–306.
- Jeganathan, P. (1995), 'Some aspects of asymptotic theory with applications to time series models', *Econometric Theory* 11, 818–887.
- Johansen, S. (1988), 'Statistical analysis of cointegration vectors', *Journal of Economic Dynamics and Control* 12, 231–254.
- Klüppelberg, C. & Mikosch, T. (1993), 'Spectral estimates and stable process', *Stochastic Processes and their Applications* 47, 323–344.
- Knight, K. (1989), 'Limit theory for autoregressive parameters in an infinite variance random walk', *Canadian Journal of Statistics* 17, 261–278.
- Knight, K. (1991), 'Limit theory of M-estimates in an integrated infinite variance process', *Econometric Theory* 7, 200–212.
- Kurtz, T. G. & Protter, P. (1991), 'Weak limit theorems for stochastic integrals and stochastic differential equations', *Annals of Probability* 19, 1035–1070.
- Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.
- Le Cam, L. & Yang, G. L. (1990), *Asymptotics in Statistics: Some Basic Concepts*, Springer-Verlag.
- Loève (1978), *Probability Theory*, Springer, New York. Fourth Edition, Part II.
- Mikosch, T., Gadrich, T., Klüppelberg, C. & Adler, R. J. (1995), 'Parameter estimation for ARMA models with infinite variance innovations', *Annals of Statistics* 23, 305–326.
- Park, J. Y. & Phillips, P. C. B. (1988), 'Statistical inference in regression with integrated processes: Part I', *Econometric Theory* 4, 468–498.
- Phillips, P. C. B. (1991), 'Optimal inference in cointegrated systems', *Econometrica* 59, 283–306.
- Phillips, P. C. B. & Durlauf, S. N. (1986), 'Multiple time series regression with integrated processes', *Review of Economic Studies* 53, 473–495.
- Reinsel, G. C. & Ahn, S. K. (1992), 'Vector AR models with unit roots and reduced rank structure: Estimation, likelihood ratio test, and forecasting', *Journal of Time Series Analysis* 13, 133–145.

- Skorohod, A. V. (1957), 'Limit theorems for processes with independent increments', *Theory of Probability and Its Applications* **2**, 138–171.
- Stock, J. H. (1987), 'Asymptotic properties of least squares estimators of cointegrating vectors', *Econometrica* **55**, 1035–1056.
- Zolotarev, V. M. (1986), *One-dimensional Stable Distributions*, American Mathematical Society, Providence.

18

Le Cam at Berkeley

E. L. Lehmann¹

Written in appreciation of the pleasure and many benefits I have received from over forty years of friendship and collegiality with Lucien Le Cam.

18.1 Early Life

In 1950 a young Frenchman came to Berkeley for a one-year visit. To his great surprise he stayed on and has remained in Berkeley ever since (though he is still a French citizen).

Le Cam was born in 1924 and spent his early life on a farm leased by his family in Creuse, Central France.

After a somewhat unconventional education, he obtained his License en Sciences (roughly equivalent to a BA) at the University of Paris in 1945. For reasons of convenience he selected as one of his examination topics Statistics, a subject in which he had no background and for which he prepared himself in less than a week by reading the lecture notes of the examiner, George Darmois. With the help of Darmois, he obtained a position as statistician at what was evolving into the Electricité de France. This job enabled him occasionally to attend seminars at the university, and at one of these he met Neyman. The meeting, together with recommendations by Charles Stein (who had spent the preceding year in Paris) and Maurice Fréchet, led to an invitation to spend a year in Berkeley as Lecturer in Statistics.

At Berkeley, he found that his background did not quite match what was expected either in mathematics or statistics. On the one hand, his statistics was applied and included little theory. On the other, the only modern mathematics he knew came from reading the books of Bourbaki. So he accepted a suggestion by Neyman to stay on for another year and complete his training by getting a California Ph.D.

Throughout his school career, Le Cam had problems with examinations, occasionally even managing to fail them. An explanation is offered by a story he tells of himself. In an early intelligence test, he was asked to give the next entry in the sequence 3,5,7,... His answer was 11, and his score on the test dismal. Now, in Berkeley, he managed to fail the Ph.D. Qual-

¹University of California at Berkeley

ifying Examination. The assignment was to give an hour lecture on fixed point theorems and Le Cam decided to present the latest and most general version he was able to find in the literature. This required first providing a considerable amount of preparatory algebraic topology, and before he knew it the hour was over—he had never reached the main result. Repeating the examination a few months later, he took a less abstract approach. However, he included a new proof of his own which was vigorously (and, as it later turned out, incorrectly) disputed by one of the examiners. This time he passed, but only barely.

Le Cam obtained his doctorate in 1952, but this was preceded by another event, his marriage earlier in the year to Louise Romig. In an autobiographical sketch (Albers, Alexanderson & Reid 1990)² he reports that when he told Neyman about the impending marriage, “he flew into a rage. He called my future father-in-law, Harry Romig, who was also a statistician, and told him it was ‘illegal’ for me to get married because I had not finished my thesis. Actually it was finished except for a few pages still to be typed. I went ahead and got married anyway in April 1952. I was awarded my Ph.D. that June”.

The thesis (Le Cam 1953), “On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes Estimates”, has since become a classic. It grew out of the discovery in the preceding year by Joe Hodges of the phenomenon of superefficiency, which had confounded the universally held belief in the efficiency of maximum likelihood estimators. One of the principal results of Le Cam’s thesis now provided damage control by showing that superefficiency can be achieved only on a set of measure zero.

On completing his degree, Le Cam accepted Neyman’s offer of a ladder appointment as Instructor. He obtained tenure in 1958 and the Professorship in 1960. The following year he succeeded David Blackwell as chair of the Statistics Department.

18.2 University and Scientific Administration

Le Cam’s term of office was marked by a phenomenal growth of the Department. In 1961 the Faculty consisted of twelve members; by the end of his term, four years later, the number had increased to twenty. In large measure this success was due to his vigorous and untiring efforts. They were helped by the general popularity of science in the post-Sputnik era, and by continuing growth of enrollment in statistics courses and of the field of statistics in the world at large.

Both as a member of the Department and as its chair Le Cam held strong

²This sketch provides much additional detail on Le Cam’s early life.

opinions which he liked to bring to the attention of his colleagues and the University administration in long, provocative and amusing letters. These were enjoyed by the Faculty but were less popular with the Deans and Chancellors.

Following are excerpts from one such communication (written in 1958) regarding a proposal by his colleague Loève to strengthen the probability part of a one-year introduction to probability and statistics at the senior/first year graduate level. Some of the statisticians in the Department had protested that such additional material was more than what students at this level and in the time available could handle, and that the change would infringe on the statistical content of the course. Here are some excerpts from Le Cam's "Comment on 'Comments on Loève's proposed revision'".

The proposed revision seems to arouse such violent opposition that the inarticulate undersigned feels obligated to communicate his feelings on the subject by the present memo.

Loève's proposal seemed at first so innocuous and so gentle that I did not expect much reaction to it from any side, except possibly that practical people (are statisticians supposed to be practical?) might desire even more emphasis on applicable probability theory. As of now I consider the differences of opinion arising in our midst are irreconcilable and due to such basic differences of understanding of what is and what should be probability theory and how it should be taught that no useful purpose will be derived by arguing any further in the same direction. Therefore the present memo will not propose any solutions but only try to generate some heat on controversial matters.

Our teaching of Probability and consequently Statistics is hopelessly inadequate. As far as I can tell, the people who struggle through [our basic courses] do not even get a knowledge of probability remotely comparable to the knowledge imparted by J. Dubourdieu in his evening lectures to candidates to the Actuarial examination in Paris.

In case one should object that Probability Theory per se is not a worthy subject of study, I shall make my next point: The statisticians who are going to make the headlines in the future are not those who can in a twinkling give an analysis of variance for a Latin Square or a three-way classification. Statistics is basically much more complicated than Probability, would it be only because where the Probabilist has only one measure to cope with, the statistician has a family of measures. There is of course plenty to do in statistics without using complicated machinery. It just requires brains. This last commodity we cannot dispense to the students, but we can give them tools.

If we would give our students a good background in mathematics and probability, we might be able to teach them more, not less, statistics in a shorter time. Otherwise we are bound to produce Ph.D.'s in Statistics who cannot even read the statistical papers of our Symposia, not to mention the sundry applied papers to be found in the same Symposia.

If this memo seems unreasonable to you, I will gladly restore it to its original strength and triple the emphasis on the necessity of teaching probability theory as a major part of our task and not as an undignified accessory to some statistical arguments.

Talking with chairman Le Cam, some time in 1962, Neyman mused on the fact that the following year marked not only the 150th anniversary of Laplace's "Théorie Analytique des Probabilités" but also the 200th anniversary of the posthumous publication of Bayes' "Essay Toward Solving a Problem in the Doctrine of Chances" and the 250th of Jacob Bernoulli's "Ars Conjectandi". To celebrate this remarkable triple anniversary Neyman proposed to Le Cam an international research seminar to be held at Berkeley. It was difficult to obtain financial support for such a meeting in the short time available, and the seminar was replaced by a series of individual lectures over a period of time during 1963, which were published two years later (Neyman & Le Cam 1965).

After plans had been announced and invitations sent, the organizers realized that the first edition of Laplace's book had appeared not in 1813 as claimed but in 1812, and the second edition containing the famous "Essai Philosophique sur les Probabilités" in 1814. The editors got around this difficulty by stating in their foreword that this Essai "must have been written in 1813, 150 years before the Berkeley Seminar of 1963".

The collaboration of Neyman and Le Cam continued with the Fifth Berkeley Symposium (Le Cam & Neyman 1967) scheduled for 1965. The first four symposia, which had taken place at five-year intervals since 1945, had been Neyman's sole responsibility. He now asked Le Cam to chair the Organizing Committee and to serve as Coeditor of the Proceedings which by then had grown to five substantial volumes. Five years later (Le Cam, Neyman & Scott 1973), Le Cam also coedited the Proceedings of the sixth (and last) symposium.

A last conference on which the two men collaborated was devoted to an interdisciplinary meeting on cancer models. Since coming to Berkeley, Le Cam had concentrated on theoretical aspects of statistics, but this changed in 1974 as a result of his daughter Linda's being diagnosed with osteosarcoma. At that time he began to collaborate³ with Vera Byers and her

³Some comments on this collaboration by Vera Byers are given in Albers et al. (1990, p. 171).

husband Al Levin who successfully treated Linda with the immunotherapy they were developing in their laboratory.

Neyman had become interested in models for the mechanism of carcinogenesis in the 1950's, and in 1958 had spent three months at the National Institutes of Health to learn more about the biological background of the problem and to participate in the work being conducted there. In the Spring of 1981, Le Cam told Neyman of experimental work being done at Berkeley that was relevant to his models. In response, Neyman proposed a conference that would bring together researchers approaching this problem from different points of view, and asked Le Cam to undertake its organization. He stressed the urgency of the problem, and the arrangements were in fact completed in less than a month. The conference took place in July 1981, only a few weeks before Neyman's death of heart failure, and the Proceedings—dedicated to his memory—were published the following year (Le Cam & Neyman 1982).

The week following Neyman's death another member of the Statistics Department, Jack Kiefer, who had experienced no previous symptoms, suffered a fatal heart attack at the young age of only 57. Stunned by these two deaths within a few days of each other, the Department hesitated for some time about a suitable commemoration. Eventually, Le Cam and Elizabeth Scott decided to organize a "Conference in Honor of Jerzy Neyman and Jack Kiefer". The conference took place in June 1983 and the Proceedings (Le Cam & Olshen 1985) were published in two volumes.

18.3 Research

Beginning with his thesis, Le Cam's statistical research has centered on large-sample theory. This has included very general treatments of maximum likelihood and Bayes estimation, introduction of contiguity and local asymptotic normality, superefficiency, and his general theory of the (approximate) comparison of experiments. Additional contributions have been made by many of his close to 40 Ph.D. students, and this body of work—which has been referred to (Feigin 1987) as the Le Cam school of asymptotic methods—was given a comprehensive exposition in Le Cam's 1986 book.

A review of this book (Shiryayev & Koshevnik 1989) states that "the endeavor for generality distinguishes Professor Le Cam himself as well as the book under review". Le Cam strongly disclaims any efforts at generality for its own sake, and justifies it as providing better organization of the material⁴. "It was just more convenient", he writes in the preface to his book, "to rewrite the definitions in such a way that the desired theorems are always true, instead of imposing restrictive conditions on a more stan-

⁴Personal communication

dard system". As a result, his work is a powerful resource that takes care of the many special nonregular cases not covered by the simpler classical approaches.

Another reviewer (Dacunha-Castelle 1988) characterizes the book [my translation from the French original] as "the work of a mathematician. It so happens that he is talking about statistics. This is due to fortuitous circumstances, but it is a stroke of good luck for statistics".

Without doubt, Le Cam's work has brought statistics closer to becoming part of mathematics. By the same token, as a consequence of the generality and abstraction of his treatment, it presents considerable difficulty to students and colleagues.⁵ His course in large-sample theory was notoriously hard, and students frequently audited it once or even twice before daring to take it for credit. The somewhat forbidding nature of his work has been greatly alleviated locally by his friendliness and patience. He kept his office door open and never seemed to tire of helping his students (and anyone else) with their problems.

Although the construction of his own general theory of experiments occupies the center of Le Cam's work, it by no means exhausts it. There are for example several papers dealing with the Central Limit Theorem; three expository papers with a historical or philosophical slant; a number of applied papers, particularly on immunotherapy and other aspects of cancer research; and papers on occasional special topics.

Le Cam's asymptotic work has been enormously influential. His theory of contiguity is a principal topic of books by Roussas (1972) and Greenwood & Shiryaev (1985), and it plays a major role in Hájek & Šidák (1967). His general approach to large-sample theory and the comparison of experiments forms the central topic of books by Heyer (1982) and Torgersen (1991).

In acknowledgement of his work, Le Cam received in 1957 a prestigious Sloan Fellowship which enabled him to spend a year in Paris in order "to renew contacts with the Bourbaki group of French mathematicians" and in 1971 a Miller Research Professorship at Berkeley. During 1972/73 he served as Director of the Centre de Recherches Mathématiques of the University of Montreal. The fear that he might accept this as a permanent position was relieved when at the end of the year he decided to return to Berkeley. In 1976 Le Cam was elected to the American Academy of Arts and Sciences. He took early retirement from Berkeley in 1991, but he continues to come to his office daily, with his door remaining open as before.

18.4 REFERENCES

Albers, D., Alexanderson, J. & Reid, C. (1990), *More Mathematical*

⁵To alleviate this difficulty Le Cam & Yang (1990) provide a more accessible introduction to essential parts of the theory. In this connection see also Strasser (1985).

- People*, Harcourt, Brace, Jovanovich, Boston.
- Dacunha-Castelle, D. (1988), 'A propos du livre de Lucien Le Cam: Asymptotic Methods in Statistical Decision Theory', *Metrika* p. 254. (Review of Le Cam 1986.).
- Feigin, P. (1987), 'Review of Greenwood and Shiryaev (1985)', *Journal of the American Statistical Association* **82**, 1195.
- Greenwood, P. & Shiryaev, A. (1985), *Contiguity and the Statistical Invariance Principle*, Gordan and Breach. New York.
- Hájek, J. & Šidák, Z. (1967), *Theory of Rank Tests*, Academic Press. (Also published by Academia, the Publishing House of the Czechoslovak Academy of Sciences, Prague.).
- Heyer, H. (1982), *Theory of Statistical Experiments*, Springer-Verlag, New York. (Second edition of *Mathematische Theorie Statischer Experimente*, Springer-Verlag 1973.).
- Le Cam, L. (1953), 'On some asymptotic properties of maximum likelihood estimates and related Bayes estimates', *University of California Publications in Statistics* **1**, 277–330.
- Le Cam, L. & Neyman, J., eds (1967), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley.
- Le Cam, L. & Neyman, J., eds (1982), *Probability Models and Cancer*, North Holland. Amsterdam.
- Le Cam, L. & Olshen, R., eds (1985), *Proc. Berkeley Confer. in Honor of Jerzy Neyman and Jack Kiefer*, Wadsworth. Monterey.
- Le Cam, L. & Yang, G. L. (1990), *Asymptotics in Statistics: Some Basic Concepts*, Springer-Verlag.
- Le Cam, L., Neyman, J. & Scott, E. L., eds (1973), *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley.
- Neyman, J. & Le Cam, L., eds (1965), *Bernoulli, Bayes, Laplace*, Springer-Verlag, Berlin.
- Roussas, G. (1972), *Contiguity of Probability Measures*, Cambridge University Press.
- Shiryaev, A. N. & Koshevnik, Y. A. (1989), 'Review of Le Cam (1986)', *Bulletin of the Amererican Mathematical Society* **20**, 280–285.

- Strasser, H. (1985), *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*, De Gruyter, Berlin.
- Torgersen, E. (1991), *Comparison of Statistical Experiments*, Cambridge University Press.

19

Another Look at Differentiability in Quadratic Mean

David Pollard¹

ABSTRACT This note revisits the delightfully subtle interconnections between three ideas: differentiability, in an L^2 sense, of the square-root of a probability density; local asymptotic normality; and contiguity.

19.1 A mystery

The traditional regularity conditions for maximum likelihood theory involve existence of two or three derivatives of the density functions, together with domination assumptions to justify differentiation under integral signs. Le Cam (1970) noted that such conditions are unnecessarily stringent. He commented:

Even if one is not interested in the maximum economy of assumptions one cannot escape practical statistical problems in which apparently “slight” violations of the assumptions occur. For instance the derivatives fail to exist at one point x which may depend on θ , or the distributions may not be mutually absolutely continuous or a variety of other difficulties may occur. The existing literature is rather unclear about what may happen in these circumstances. Note also that since the conditions are imposed upon probability densities they may be satisfied for one choice of such densities but not for certain other choices.

Probably Le Cam had in mind examples such as the double exponential density, $1/2 \exp(-|x - \theta|)$, for which differentiability fails at the point $\theta = x$. He showed that the traditional conditions can be replaced by a simpler assumption of differentiability in quadratic mean (DQM): differentiability in norm of the square root of the density as an element of an L^2 space. Much asymptotic theory can be made to work under DQM. In particular, as Le Cam showed, it implies a quadratic approximation property for the log-likelihoods known as local asymptotic normality (LAN).

Le Cam’s idea is simple but subtle. When I first encountered the LAN property I wrongly dismissed it as nothing more than a Taylor expansion

¹Yale University

to quadratic terms of the log-likelihood. Le Cam's DQM result showed otherwise: one appears to get the benefit of the quadratic expansion without paying the twice-differentiability price usually demanded by such a Taylor expansion. How can that happen?

My initial puzzlement was not completely allayed by a study of several careful accounts of LAN, such as those of Le Cam (1970; 1986, Section 17.3), Ibragimov & Has'minskii (1981, page 114), Millar (1983, page 105), Le Cam & Yang (1990, page 101), or Strasser (1985, Chapter 12). None of the proofs left me with the feeling that I really understood why second derivatives are not needed. (No criticism of those authors intended, of course.)

Eventually it dawned on me that I had overlooked a vital ingredient in the proofs: the square root of a density is not just an element of an \mathcal{L}^2 space: *it is an element with norm 1*. By rearranging some of the standard arguments I hope to convince the gentle reader of this note that the fixed norm is the real reason for why an assumption of one-times differentiability (in quadratic mean) can convey the benefits usually associated with two-times differentiability. I claim that the Lemma in the next Section is the key to understanding the role of DQM.

19.2 A lemma

The concept of differentiability makes sense for maps into an arbitrary normed space $(\mathcal{L}, \|\cdot\|)$. For the purposes of my exposition, it suffices to consider the case where the norm is generated by an inner product, $\langle \cdot, \cdot \rangle$. In fact, \mathcal{L} will be $\mathcal{L}^2(\lambda)$, the space of functions square-integrable with respect to some measure λ , but that simplification will play no role for the moment.

A map ξ from \mathbb{R}^k into \mathcal{L} is said to be differentiable at a point θ_0 with derivative Δ , if $\xi(\theta) = \xi(\theta_0) + \Delta(\theta - \theta_0) + r(\theta)$ near θ_0 , where $\|r(\theta)\| = o(|\theta - \theta_0|)$ as θ tends to θ_0 . The derivative Δ is linear; it may be identified with a k -vector of elements from \mathcal{L} .

For a differentiable map, the Cauchy-Schwarz inequality implies that $\langle \xi(\theta_0), r(\theta) \rangle = o(|\theta - \theta_0|)$. It would usually be a blunder to assume naively that the bound must therefore be of order $O(|\theta - \theta_0|^2)$; typically, higher-order differentiability assumptions are needed to derive approximations with smaller errors. However, if $\|\xi(\theta)\|$ is constant—that is, if the function is constrained to take values lying on the surface of a sphere—then the naive assumption turns out to be no blunder. Indeed, in that case, $\langle \xi(\theta_0), r(\theta) \rangle$ can be written as a quadratic in $\theta - \theta_0$ plus an error of order $o(|\theta - \theta_0|^2)$. The sequential form of the assertion is more convenient for my purposes.

(1) **Lemma** *Let $\{\delta_n\}$ be a sequence of constants tending to zero. Let ξ_0, ξ_1, \dots be elements of norm one for which $\xi_n = \xi_0 + \delta_n W + r_n$, with W a fixed element of \mathcal{L} and $\|r_n\| = o(\delta_n)$. Then $\langle \xi_0, W \rangle = 0$ and $\langle \xi_0, r_n \rangle = -\frac{1}{2}\delta_n^2\|W\|^2 + o(\delta_n^2)$.*

Proof. Because both ξ_n and ξ_0 have unit length,

$$\begin{aligned} 0 &= \|\xi_n\|^2 - \|\xi_0\|^2 = 2\delta_n \langle \xi_0, W \rangle && \text{order } O(\delta_n) \\ &\quad + 2\langle \xi_0, r_n \rangle && \text{order } o(\delta_n) \\ &\quad + \delta_n^2 \|W\|^2 && \text{order } O(\delta_n^2) \\ &\quad + 2\delta_n \langle W, r_n \rangle + \|r_n\|^2 && \text{order } o(\delta_n^2). \end{aligned}$$

On the right-hand side I have indicated the order at which the various contributions tend to zero. (The Cauchy-Schwarz inequality delivers the $o(\delta_n)$ and $o(\delta_n^2)$ terms.) The exact zero on the left-hand side leaves the leading $2\delta_n \langle \xi_0, W \rangle$ unhappily exposed as the only $O(\delta_n)$ term. It must be of smaller order, which can happen only if $\langle \xi_0, W \rangle = 0$, leaving

$$0 = 2\langle \xi_0, r_n \rangle + \delta_n^2 \|W\|^2 + o(\delta_n^2),$$

as asserted. \square

Without the fixed length property, the inner product $\langle \xi_0, r_n \rangle$, which inherits $o(\delta_n)$ behaviour from $\|r_n\|$, might not decrease at the $O(\delta_n^2)$ rate.

19.3 A theorem

Let $\{P_\theta : \theta \in \Theta\}$ be a family of probability measures on a space $(\mathcal{X}, \mathcal{A})$, indexed by a subset Θ of \mathbb{R}^k . Suppose P_θ has density $f(x, \theta)$ with respect to a sigma-finite measure λ .

Under the classical regularity conditions—twice continuous differentiability of $\log f(x, \theta)$ with respect to θ , with a dominated second derivative—the likelihood ratio

$$\prod_{i \leq n} \frac{f(x_i, \theta)}{f(x_i, \theta_0)}$$

enjoys the LAN property. Write $L_n(t)$ for the likelihood ratio evaluated at θ equal to $\theta_0 + t/\sqrt{n}$. The property asserts that, if the $\{x_i\}$ are sampled independently from P_{θ_0} , then

$$(2) \quad L_n(t) = \exp(t' S_n - \frac{1}{2} t' \Gamma t + o_p(1)) \quad \text{for each } t,$$

where Γ is a fixed matrix (depending on θ_0) and S_n has a centered asymptotic normal distribution with variance matrix Γ .

Formally, the LAN approximation results from the usual pointwise Taylor expansion of the log density $g(x, \theta) = \log f(x, \theta)$, following a style of argument familiar to most graduate students. For example, in one dimension,

$$\begin{aligned} \log L_n(\theta_0 + t/\sqrt{n}) &= \sum_{i \leq n} (g(x_i, \theta_0 + t/\sqrt{n}) - g(x_i, \theta_0)) \\ &= \frac{t}{\sqrt{n}} \sum_{i \leq n} g'(x_i, \theta_0) + \frac{t^2}{2n} \sum_{i \leq n} g''(x_i, \theta_0) + \dots, \end{aligned}$$

which suggests that S_n be the standardized score function,

$$\frac{1}{\sqrt{n}} \sum_{i \leq n} g'(x_i, \theta_0) \rightsquigarrow N(0, \text{var}_{\theta_0} g'(x, \theta_0)),$$

and Γ should be the information function,

$$-P_{\theta_0} g''(x, \theta_0) = \text{var}_{\theta_0} g'(x, \theta_0).$$

The dual representation for Γ allows one to eliminate all mention of second derivatives from the statement of the LAN approximation, which hints that two derivatives might not really be needed, as Le Cam (1970) showed.

In general, the family of densities is said to be differentiable in quadratic mean at θ_0 if the square root $\xi(x, \theta) = \sqrt{f(x, \theta)}$ is differentiable in the $L^2(\lambda)$ sense: for some k-vector $\Delta(x)$ of functions in $L^2(\lambda)$,

$$(3) \quad \xi(x, \theta) = \xi(x, \theta_0) + (\theta - \theta_0)' \Delta(x) + r(x, \theta),$$

where

$$\lambda |r(x, \theta)|^2 = o(|\theta - \theta_0|^2) \quad \text{as } \theta \rightarrow \theta_0.$$

Let us abbreviate $\xi(x, \theta_0)$ to $\xi_0(x)$ and $\Delta(x)/\xi_0(x)$ to $D(x)$. From (3) one almost gets the LAN property.

(4) **Theorem** Assume the DQM property (3). For each fixed t the likelihood ratio has the approximation, under $\{\mathbb{P}_{n, \theta_0}\}$,

$$L_n(t) = \exp \left(t' S_n - \frac{1}{2} t' \Gamma t + o_p(1) \right),$$

where

$$S_n = \frac{2}{\sqrt{n}} \sum_{i \leq n} D(x_i) \rightsquigarrow N(0, \mathbb{I}_0) \quad \text{and} \quad \Gamma = \frac{1}{2} \mathbb{I}_0 + \frac{1}{2} \mathbb{I},$$

with $\mathbb{I}_0 = 4\lambda(\Delta\Delta' \{\xi_0 > 0\})$ and $\mathbb{I} = 4\lambda(\Delta\Delta')$.

Notice the slight difference between Γ and the limiting variance matrix for S_n . At least formally, $2D(x)$ equals the derivative of $\log f(x, \theta)$: ignoring problems related to division by zero and distinctions between pointwise and $L^2(\lambda)$ differentiability, we have

$$2D(x) = \frac{2}{\sqrt{f(x, \theta_0)}} \frac{\partial}{\partial \theta} \sqrt{f(x, \theta_0)} = \frac{\partial}{\partial \theta} \log f(x, \theta_0).$$

Also, Γ again corresponds to the information matrix, expressed in its variance form, except for the intrusion of the indicator function $\{\xi_0 > 0\}$. The extra indicator is necessary if we wish to be careful about 0/0. Its presence is related to the property called contiguity—another of Le Cam's great ideas—as is explained in Section 5.

At first sight the derivation of Theorem 4 from assumption (3) again appears to be a simple matter of a Taylor expansion to quadratic terms of

the log likelihood ratio. Writing $R_n(x) = r(x, \theta_0 + t/\sqrt{n})/\xi_0(x)$, we have

$$\begin{aligned}\log L_n(t) &= \sum_{i \leq n} 2 \log \frac{\xi(x_i, \theta_0 + t/\sqrt{n})}{\xi(x_i, \theta_0)} \\ &= \sum_{i \leq n} 2 \log \left(1 + \frac{t'}{\sqrt{n}} D(x_i) + R_n(x_i) \right).\end{aligned}$$

From the Taylor expansion of $\log(\cdot)$ about 1, the sum of logarithms can be written as a formal series,

$$(5) \quad \begin{aligned} &2 \sum_{i \leq n} \left(\frac{t}{\sqrt{n}} D(x_i) + R_n(x_i) \right) - \sum_{i \leq n} \left(\frac{t'}{\sqrt{n}} D(x_i) + R_n(x_i) \right)^2 + \dots \\ &= \frac{2t'}{\sqrt{n}} \sum_{i \leq n} D(x_i) + 2 \sum_{i \leq n} R_n(x_i) - \frac{1}{n} \sum_{i \leq n} (t' D(x_i))^2 + \dots\end{aligned}$$

The first sum on the right-hand side gives the $t'S_n$ in Theorem 4. The law of large numbers gives convergence of the third term to $t'P_{\theta_0}DD't$. Mere one-times differentiability might not seem enough to dispose of the second sum. Each summand has standard deviation of order $o(1/\sqrt{n})$, by DQM. A sum of n such terms could crudely be bounded via a triangle inequality, leaving a quantity of order $o(\sqrt{n})$, which clearly would not suffice. In fact the sum of the $R_n(x_i)$ does not go away in the limit; as a consequence of Lemma 1, it contributes a fixed quadratic in t . That contribution is the surprise behind DQM.

19.4 A proof

Let me write \mathbb{P}_n to denote calculations under the assumption that the observations x_1, \dots, x_n are sampled independently from P_{θ_0} . The ratio $f(x_i, \theta_0 + t/\sqrt{n})/f(x_i, \theta_0)$ is not well defined when $f(x_i, \theta_0) = 0$, but under \mathbb{P}_n the problem can be neglected because

$$\mathbb{P}_n\{f(x_i, \theta_0) = 0 \text{ for at least one } i\} = 0.$$

For other probability measures that are not absolutely continuous with respect to \mathbb{P}_n , one should be more careful. It pays to be quite explicit about behaviour when $f(x_i, \theta_0) = 0$ for some i , by including an explicit indicator function $\{\xi_0 > 0\}$ as a factor in any expressions with a ξ_0 in the denominator.

Define D_i to be the random vector $\Delta(x_i)\{\xi_0(x_i) > 0\}/\xi_0(x_i)$, and, for a fixed t , define

$$R_{i,n} = r(\xi_i, \theta_0 + t/\sqrt{n})\{\xi_0(x_i) > 0\}/\xi_0(x_i).$$

Then

$$\frac{\xi(x_i, \theta_0 + t/\sqrt{n})}{\xi_0(x_i)} \{ \xi_0(i) > 0 \} = 1 + t'D_i + R_{i,n}.$$

The random vector D_i has expected value $\lambda(\xi_0 \Delta)$, which, by Lemma 1, is zero, even without the traditional regularity assumptions that justify differentiation under an integral sign. It has variance $\frac{1}{4}\mathbb{I}_0$. It follows by a central limit theorem that

$$S_n = \frac{2}{\sqrt{n}} \sum_{i \leq n} D_i \rightsquigarrow N(0, \mathbb{I}_0).$$

Also, by a (weak) law of large numbers,

$$(6) \quad \frac{1}{n} \sum_{i \leq n} D_i D'_i \rightarrow \mathbb{P}_n(D_1 D'_1) = \frac{1}{4}\mathbb{I}_0 \quad \text{in probability.}$$

To establish rigorously the near-LAN assertion of Theorem 4, it is merely a matter of bounding the error terms in (5) and then justifying the treatment of the sum of the $R_n(x_i)$. Three facts are needed.

(7) **Lemma** Under $\{\mathbb{P}_n\}$, assuming DQM,

- (a) $\max_{i \leq n} |D_i| = o_p(\sqrt{n})$,
- (b) $\max_{i \leq n} |R_{i,n}| = o_p(1)$,
- (c) $\sum_{i \leq n} 2R_{i,n} \rightarrow -\frac{1}{4}t'\mathbb{I}t$ in probability.

Let me first explain how Theorem 4 follows from Lemma 7. Together the two facts (a) and (b) ensure that with high probability $\log L_n(t)$ does not involve infinite values. For $(t'D_i/\sqrt{n}) + R_{i,n} > -1$ we may then appeal to the Taylor expansion

$$\log(1 + y) = y - \frac{1}{2}y^2 + \frac{1}{2}\beta(y),$$

where $\beta(y) = o(y^2)$ as y tends to zero, to deduce that $\log L_n(t)$ equals

$$\frac{2}{\sqrt{n}} \sum_{i \leq n} t'D_i + 2 \sum_{i \leq n} R_{i,n} - \sum_{i \leq n} \left(\frac{t'D_i}{\sqrt{n}} + R_{i,n} \right)^2 + \sum_{i \leq n} \beta \left(\frac{t'D_i}{\sqrt{n}} + R_{i,n} \right),$$

which expands to

$$\begin{aligned} & t'S_n + 2 \sum_{i \leq n} R_{i,n} - \frac{1}{n} \sum_{i \leq n} (t'D_i)^2 \\ & - \frac{2}{\sqrt{n}} \sum_{i \leq n} t'D_i R_{i,n} - \sum_{i \leq n} R_{i,n}^2 + o_p(1) \sum_{i \leq n} \left(\frac{|D_i|^2}{n} + R_{i,n}^2 \right). \end{aligned}$$

Each of the last three sums is of order $o_p(1)$ because $\sum_{i \leq n} |D_i|^2/n = O_p(1)$ and

$$(8) \quad \begin{aligned} \mathbb{P}_n \sum_{i \leq n} R_{i,n}^2 &= n \lambda (\xi_0^2 r(x_1, \theta_0 + t/\sqrt{n}) \{ \xi_0 > 0 \} / \xi_0^2) \\ &\leq n \lambda |r(\cdot, \theta_0 + t/\sqrt{n})|^2 \end{aligned}$$

$$= o(1).$$

By virtue of (6) and (c), the expansion simplifies to

$$t' S_n - \frac{1}{4} t' \mathbb{I}t - \frac{1}{4} t' \mathbb{I}_0 t + o_p(1),$$

as asserted by Theorem 4.

Proof of Lemma 7. Assertion (a) follows from the identical distributions:

$$\begin{aligned} \mathbb{P}_n \left\{ \max_{i \leq n} |D_i| > \epsilon \sqrt{n} \right\} &\leq \sum_{i \leq n} \mathbb{P}_n \{|D_i| > \epsilon \sqrt{n}\} \\ &= n \mathbb{P}_n \{|\Delta_1| > \epsilon \sqrt{n}\} \\ &\leq \epsilon^{-2} \lambda \Delta_1^2 \{|\Delta_1| > \xi_0 \epsilon \sqrt{n}\} \\ &\rightarrow 0 \quad \text{by Dominated Convergence.} \end{aligned}$$

Assertion (b) follows from (8):

$$\mathbb{P}_n \left\{ \max_{i \leq n} |R_{i,n}| > \epsilon \right\} \leq \epsilon^{-2} \mathbb{P}_n \sum_{i \leq n} R_{i,n}^2 \rightarrow 0.$$

Only Assertion (c) involves any subtlety. The variance of the sum is bounded by $4 \sum_{i \leq n} \mathbb{P}_n R_n(x_i)^2$, which tends to zero. The sum of the remainders must lie within $o_p(1)$ of its expected value, which equals

$$2n P_{\theta_0} R_{1,n} = 2n \lambda (\xi_0 r(\cdot, \theta_0 + t/\sqrt{n})),$$

an inner product between two functions in $\mathcal{L}^2(\lambda)$. Notice that the ξ_0 factor makes the indicator $\{\xi_0 > 0\}$ redundant.

It is here that the unit length property becomes important. Specializing Lemma 1 to the case $\delta_n = 1/\sqrt{n}$, with $\xi_n(x) = \xi(x, \theta_0 + t/\sqrt{n})$ and $W = t'\Delta$, we get the approximation to the sum of expected values of the $R_{i,n}$, from which Assertion (c) follows. \square

A slight generalization of the LAN assertion is possible. It is not necessary that we consider only parameters of the form $\theta_0 + t/\sqrt{n}$ for a fixed t . By arguing almost as above along convergent subsequences of $\{t_n\}$ we could prove an analog of Theorem 4 if t were replaced by a bounded sequence $\{t_n\}$ such that $\theta_0 + t_n/\sqrt{n} \in \Theta$. The extension is significant because (Le Cam 1986, page 584) the slightly stronger result forces a form of differentiability in quadratic mean.

19.5 Contiguity and disappearance of mass

For notational simplicity, consider only the one-dimensional case with the typical value $t = 1$. Let ξ_0^2 be the marginal density, and \mathbb{Q}_n be the joint distribution, for x_1, \dots, x_n sampled with parameter value $\theta_0 + 1/\sqrt{n}$. As before, ξ_0^2 and \mathbb{P}_n correspond to θ_0 . The measure \mathbb{Q}_n is absolutely contin-

uous with respect to \mathbb{P}_n if and only if it puts zero mass in the set

$$A_n = \{\xi_0(x_i) = 0 \text{ for at least one } i \leq n\}.$$

Writing α_n for $\lambda\xi_n^2\{\xi_0 = 0\}$, we have

$$\mathbb{Q}_n A_n = 1 - \prod_{i \leq n} (1 - \mathbb{Q}_n\{\xi_0(x_i) = 0\}) = 1 - (1 - \alpha_n)^n.$$

By direct calculation,

$$\alpha_n = \lambda(r_n + \Delta/\sqrt{n})^2\{\xi_0 = 0\} = \lambda\Delta^2\{\xi_0 = 0\}/n + o(1/n).$$

The quantity $\tau = \lambda\Delta^2\{\xi_0 = 0\}$ has the following significance. Under \mathbb{Q}_n , the number of observations landing in A_n has approximately a Poisson(τ) distribution; and $\mathbb{Q}_n A_n \rightarrow 1 - e^{-\tau}$.

In some asymptotic sense, the measure \mathbb{Q}_n becomes more nearly absolutely continuous with respect to \mathbb{P}_n if and only if $\tau = 0$. The precise sense is called contiguity: *the sequence of measures $\{\mathbb{Q}_n\}$ is said to be contiguous with respect to $\{\mathbb{P}_n\}$ if $\mathbb{Q}_n B_n \rightarrow 0$ for each sequence of sets $\{B_n\}$ such that $\mathbb{P}_n B_n \rightarrow 0$.* Because $\mathbb{P}_n A_n = 0$ for every n , the condition $\tau = 0$ is clearly necessary for contiguity. It is also sufficient.

Contiguity follows from the assertion that L , the limit in distribution under $\{\mathbb{P}_n\}$ of the likelihood ratios $\{L_n(1)\}$, have expected value one. (“Le Cam’s first lemma”—see the theorem on page 20 of Le Cam and Yang, 1990.) The argument is simple: If $\mathbb{P}L = 1$ then, to each $\epsilon > 0$ there exists a finite constant C such that $\mathbb{P}L\{L < C\} > 1 - \epsilon$. From the convergence in distribution, $\mathbb{P}_n L_n\{L_n < C\} > 1 - \epsilon$ eventually. If $\mathbb{P}_n B_n \rightarrow 0$ then

$$\begin{aligned} \mathbb{Q}_n B_n &\leq \mathbb{P}_n B_n L_n\{L_n < C\} + \mathbb{Q}_n\{L_n \geq C\} \\ &\leq C\mathbb{P}_n B_n + 1 - \mathbb{P}_n L_n\{L_n < C\} \\ &< 2\epsilon \quad \text{eventually.} \end{aligned}$$

For the special case of the limiting $\exp(N(\mu, \sigma^2))$ distribution, where $\mu = -\frac{1}{4}\mathbb{I}_0 - \frac{1}{4}\mathbb{I}$ and $\sigma^2 = \mathbb{I}_0$, the requirement becomes

$$1 = \mathbb{P} \exp(N(\mu, \sigma^2)) = \exp(\mu + \frac{1}{2}\sigma^2).$$

That is, contiguity obtains when $\mathbb{I}_0 = \mathbb{I}$ (or equivalently, $\lambda(\Delta^2\{\xi_0 = 0\}) = 0$), in which case, the limiting variance of S_n equals Γ . This conclusion plays the same role as the traditional dual representation for the information function. As Le Cam & Yang (1990, page 23) commented, “The equality ... is the classical one. One finds it for instance in the standard treatment of maximum likelihood estimation under Cramér’s conditions. There it is derived from conditions of differentiability under the integral sign.” The fortuitous equality is nothing more than contiguity in disguise.

From the literature one sometimes gets the impression that $\lambda\Delta^2\{\xi_0 = 0\}$ is always zero. It is not.

(9) **Example** Let λ be Lebesgue measure on the real line. Define

$$f_0(x) = x\{0 \leq x \leq 1\} + (2-x)\{1 < x \leq 2\}.$$

For $0 \leq \theta \leq 1$ define densities

$$f(x, \theta) = (1 - \theta^2)f_0(x) + \theta^2f_0(x - 2).$$

Notice that

$$(10) \quad \lambda \left| \sqrt{f(x, \theta)} - \sqrt{f(x, 0)} - \theta \sqrt{f(x, 1)} \right|^2 = (\sqrt{1 - \theta^2} - 1)^2 = O(\theta^4).$$

The family of densities is differentiable in quadratic mean at $\theta = 0$ with derivative $\Delta(x) = \sqrt{f(x, 1)}$. For this family, $\lambda\Delta^2\{\xi_0 = 0\} = 1$.

The near-LAN assertion of Theorem 4 degenerates: $\mathbb{I}_0 = 0$ and $\mathbb{I} = 4$, giving $L_n(t) \rightarrow \exp(-t^2)$ in probability, under $\{\mathbb{P}_{n, \theta_0}\}$. Indeed, as Aad van der Vaart has pointed out to me, the limiting experiment (in Le Cam's sense) for the models $\{\mathbb{P}_{n, t/\sqrt{n}} : 0 \leq t \leq \sqrt{n}\}$ is not the Gaussian translation model corresponding to the LAN condition. Instead, the limit experiment is $\{\mathbb{Q}_t : t \geq 0\}$, with \mathbb{Q}_t equal to the Poisson(t^2) distribution. That is, for each finite set T and each h , under $\{\mathbb{P}_{n, h/\sqrt{n}}\}$ the random vectors

$$\left(\frac{d\mathbb{P}_{n, t/\sqrt{n}}}{d\mathbb{P}_{n, h/\sqrt{n}}} : t \in T \right)$$

converge in distribution to

$$\left(\frac{d\mathbb{Q}_t}{d\mathbb{Q}_h} : t \in T \right),$$

as a random vector under the \mathbb{Q}_h distribution. \square

The counterexample would not work if θ were allowed to take on negative values; one would need $\Delta(x) = -\sqrt{f(x, 1)}$ to get the analog of (10) for negative θ . The failure of contiguity is directly related to the fact that $\theta = 0$ lies on boundary of the parameter interval.

In general, $\lambda\Delta\Delta'\{\xi_0 = 0\}$ must be zero at all interior points of the parameter space where DQM holds. On the set $\{\xi_0 = 0\}$ we have $0 \leq \sqrt{n}\xi(x, \theta_0 + t/\sqrt{n}) = t'\Delta + \sqrt{nr_n}$, where $\|\sqrt{nr_n}\| \rightarrow 0$. Along a subsequence, $\sqrt{nr_n} \rightarrow 0$, leaving the conclusion that $t'\Delta \geq 0$ almost everywhere on the set $\{\xi_0 = 0\}$. At an interior point, t can range over all directions, which forces $\Delta = 0$ almost everywhere on $\{\xi = 0\}$; at an interior point, $\Delta\Delta'\{\xi = 0\} = 0$ almost everywhere. More generally, one needs only to be able to approach θ_0 from enough different directions to force $\Delta = 0$ on $\{\xi_0 = 0\}$ —as in the concept of a contingent in Le Cam & Yang (1990, Section 6.2).

The assumption that θ_0 lies in the interior of the parameter space is not always easy to spot in the literature.

Some authors, such as Le Cam & Yang (1990, page 101), prefer to dispense with the dominating measure λ , by recasting differentiability in

quadratic mean as a property of the densities $d\mathbb{P}_\theta/d\mathbb{P}_{\theta_0}$, whose square roots correspond to the ratios $\xi(x, \theta)\{\xi_0 > 0\}/\xi_0(x)$. With that approach, the behaviour of Δ on the set $\{\xi_0 = 0\}$ must be specified explicitly. The contiguity requirement—that P_θ puts, at worst, mass of order $o(|\theta - \theta_0|^2)$ in the set $\{\xi_0 = 0\}$ —is then made part of the definition of differentiability in quadratic mean.

19.6 REFERENCES

- Ibragimov, I. A. & Has'minskii, R. Z. (1981), *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.
- Le Cam, L. (1970), 'On the assumptions used to prove asymptotic normality of maximum likelihood estimators', *Annals of Mathematical Statistics* 41, 802–828.
- Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.
- Le Cam, L. & Yang, G. L. (1990), *Asymptotics in Statistics: Some Basic Concepts*, Springer-Verlag.
- Millar, P. W. (1983), 'The minimax principle in asymptotic statistical theory', *Springer Lecture Notes in Mathematics* pp. 75–265.
- Strasser, H. (1985), *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*, De Gruyter, Berlin.

20

On a Set of the First Category

Hein Putter¹
Willem R. van Zwet²

ABSTRACT In an analysis of the bootstrap Putter & van Zwet (1993) showed that under quite general circumstances, the bootstrap will work for “most” underlying distributions. In fact, the set of exceptional distributions for which the bootstrap does not work was shown to be a set D of the first category in the space \mathcal{P} of all possible underlying distributions, equipped with a topology Π . Such a set of the first category is usually “small” in a topological sense. However, it is known that this concept of smallness may sometimes be deceptive and in unpleasant cases such “small” sets may in fact be quite large.

Here we present a striking and hopefully amusing example of this phenomenon, where the “small” subset D equals all of \mathcal{P} . We show that as a result, a particular version of the bootstrap for the sample minimum will never work, even though our earlier results tell us that it can only fail for a “small” subset of underlying distributions. We also show that when we change the topology on \mathcal{P} —and as a consequence employ a different resampling distribution—this paradox vanishes and a satisfactory version of the bootstrap is obtained. This demonstrates the importance of a proper choice of the resampling distribution when using the bootstrap.

20.1 Introduction

Many of the results of asymptotic statistics cannot be established in complete generality. One often has to allow the possibility that the result will not hold if the underlying probability distribution belongs to a small subset D of the collection of all possible underlying probability distributions \mathcal{P} . In many concrete examples, D will turn out to be empty, but in general one has to take the existence of such an exceptional set into account.

If \mathcal{P} is a parametric model, the exceptional set D will typically be small in the sense that it is indexed by a set of Lebesgue measure zero in the Euclidean parameter space. From a technical point of view, its occurrence is caused by an application of a result like Egorov’s or Lusin’s theorem where exceptional sets of arbitrarily small Lebesgue measure occur. In more general models one could conceivably use similar tools for more general

¹University of Leiden

²University of Leiden and University of North Carolina.

measures, but it is difficult to think of a measure on \mathcal{P} which is such that we can agree that a set of measure zero is indeed small in a relevant sense.

In a recent study of resampling, we have followed a different path and established asymptotic results where the exceptional set is small in a topological rather than a measure-theoretic sense (Putter & van Zwet 1993). If we equip the set \mathcal{P} with a metric p , the exceptional set D in these results is a set of the first category in the metric space (\mathcal{P}, p) . We recall that a set of the first category is a countable union of nowhere dense sets, and that a set is nowhere dense if its closure has empty interior. Equivalently, a set is of the first category if it can be covered by a countable union of closed sets, each of which has empty interior.

This concept of a small set was used by Le Cam as early as Le Cam (1953), where it is shown that superefficiency can only occur on a set of the first category. In a parametric setting, Le Cam was careful to point out that under the right conditions the exceptional set also corresponds to a set of Lebesgue measure zero in the parameter space. The same is true for the results in Putter & van Zwet (1993), as shown by Putter (1994).

Of course the question remains whether a set of the first category is indeed small in any accepted sense. If (\mathcal{P}, p) is complete, we know that a set of the first category is small, for example in the sense that it has a dense complement (cf. Dudley 1989, pp. 43–44). If (\mathcal{P}, p) is not complete, then a set of the first category can be uncomfortably large: in fact we shall see that the entire space \mathcal{P} may be of the first category itself.

In this note we discuss a particular statistical model \mathcal{P}_0 equipped with Hellinger metric H , such that (\mathcal{P}_0, H) is not complete and \mathcal{P}_0 is of the first category in (\mathcal{P}_0, H) . An application of our results on resampling shows that a particular version of the bootstrap will work except if the underlying distribution belongs to a set D of the first category. Unfortunately, it turns out that $D = \mathcal{P}_0$ so that we have no guarantee that this version of the bootstrap will ever work, and indeed it may never do. Luckily, our analysis also shows that we need not despair. It turns out that our problems are not caused by any inherent pathology of the model \mathcal{P}_0 , but by a wrong choice of metric on \mathcal{P}_0 . If we replace H by a different, complete, metric and modify the construction of the bootstrap accordingly, the pathology disappears and we obtain a version of the bootstrap that will work for any $P \in \mathcal{P}_0$. In fact the example may serve to clarify the importance of a correct choice of the resampling distribution when using the bootstrap.

In Section 2 we exhibit the particular class of distributions \mathcal{P}_0 which is of the first category in (\mathcal{P}_0, H) . In section 3 we show that this class is not an artificial construct, but that it is the natural model for a statistical situation of interest. We then proceed to make the connection with a result on the bootstrap in Putter & van Zwet (1993) and show that this result doesn't produce a satisfactory version of the bootstrap for this model. Finally we show that a different choice of metric on \mathcal{P}_0 will resolve our problems.

20.2 A set of the first category

Let us consider the class \mathcal{P}_0 of probability distributions P on $(0, \infty)$ which have distribution functions F satisfying

$$(1) \quad \lim_{x \downarrow 0} \frac{F(x)}{x} = a(P) \in (0, \infty).$$

We equip \mathcal{P}_0 with Hellinger metric H . For distributions $P, Q \in \mathcal{P}_0$, with densities f and g with respect to a common σ -finite measure ν , this is defined by

$$H(P, Q)^2 = \int (\sqrt{f} - \sqrt{g})^2 d\nu.$$

(2) Proposition *The set \mathcal{P}_0 is of the first category in (\mathcal{P}_0, H) .*

Proof. For $k = 1, 2, \dots$, let $\delta_k = 1/k$ and

$$B_k = \{P \in \mathcal{P}_0 : \left| \frac{F(x)}{x} - \frac{F(\delta_k)}{\delta_k} \right| \leq 1 \text{ for } 0 < x \leq \delta_k\}.$$

Clearly, $\mathcal{P}_0 \subset \bigcup_{k=1}^{\infty} B_k$, and since convergence in Hellinger metric implies pointwise convergence of distribution functions, we see that each B_k is closed in (\mathcal{P}_0, H) . It remains to be shown that no B_k contains an open set.

Fix k and choose a distribution $P \in B_k$ with distribution function F and with $a(P) = \alpha$. Define $G_n(x) = \min(n^{-1}, (3 + \alpha)x)$, $F_n = \max(G_n, F)$, and let P_n be the distribution with distribution function F_n . Then $a(P_n) = 3 + \alpha$ but, for n large enough, $F_n(\delta_k)/\delta_k = F(\delta_k)/\delta_k \leq 1 + \alpha$ because $P \in B_k$. It follows that $P_n \notin B_k$ for large n , even though P_n converges to P in Hellinger metric. \square

20.3 A bootstrap fiasco

Let \mathcal{P} be a class of probability distributions on \mathbb{R} . We equip \mathcal{P} with a metric ρ . Let X_1, X_2, \dots be independent and identically distributed (i.i.d.) random variables with (unknown) common distribution $P \in \mathcal{P}$. We are interested in the large sample behavior of a random variable

$$(3) \quad Y_N = y_N(X_1, \dots, X_N; P).$$

Let $\tau_N(P)$ denote the distribution of Y_N under $P \in \mathcal{P}$, and suppose that, for every $P \in \mathcal{P}$, $\tau_N(P)$ converges weakly to a limit distribution $\tau(P)$. If $\widehat{P}_N = P_N(X_1, \dots, X_N)$ is an estimator of P taking values in \mathcal{P} , then $\tau_N(\widehat{P}_N)$ is called a bootstrap estimator of $\tau_N(P)$, or of $\tau(P)$, with resampling distribution \widehat{P}_N . For all P and \widehat{P}_N , the distributions $\tau_N(P)$, $\tau(P)$, and $\tau_N(\widehat{P}_N)$ are elements of the class \mathcal{R} of all probability measures on \mathbb{R} . We equip this class with Lévy distance ℓ , or any other metric which

metrizes weak convergence. The bootstrap is said to work for a particular $P \in \mathcal{P}$ if it is an ℓ -consistent estimator of $\tau_N(P)$, i.e. if $\ell(\tau_N(\hat{P}_N), \tau_N(P))$ converges to zero in probability under P . As $\ell(\tau_N(P), \tau(P)) \rightarrow 0$, this is the same as ℓ -consistency for estimating the limit distribution $\tau(P)$.

The following two propositions are taken from Putter & van Zwet (1993).

(4) Proposition Suppose that

- (i) The sequence of maps $\tau_N : (\mathcal{P}, p) \rightarrow (\mathcal{R}, \ell)$ is equicontinuous on \mathcal{P} ;
- (ii) \hat{P}_N takes values in \mathcal{P} and is a p -consistent estimator of P , i.e. $p(\hat{P}_N, P) \xrightarrow{P} 0$ for every $P \in \mathcal{P}$.

Then the bootstrap $\tau_N(\hat{P}_N)$ works for every $P \in \mathcal{P}$.

(5) Proposition Suppose that

- (i) $\tau_N : (\mathcal{P}, p) \rightarrow (\mathcal{R}, \ell)$ is continuous for every N ;
- (ii) For every $P \in \mathcal{P}$, $\tau_N(P)$ converges weakly to a limit $\tau(P)$;
- (iii) \hat{P}_N takes values in \mathcal{P} and is a p -consistent estimator of P , i.e. $p(\hat{P}_N, P) \xrightarrow{P} 0$ for every $P \in \mathcal{P}$.

Then there exists a set D of the first category in (\mathcal{P}, p) such that the sequence τ_N is equicontinuous at every $P \in \mathcal{P} \setminus D$ and hence the bootstrap $\tau_N(\hat{P}_N)$ works for every $P \in \mathcal{P} \setminus D$.

Usually these results are used with Hellinger distance H for p , and on closer inspection it often turns out that the exceptional set D may be taken to be empty.

In the remainder of this paper we shall consider a specific example of this situation. We choose $\mathcal{P} = \mathcal{P}_0$, the class of distributions defined in (1). For i.i.d. random variables X_1, \dots, X_N taking values in $(0, \infty)$ with common distribution $P \in \mathcal{P}_0$, we define

$$(6) \quad Y_N^0 = N \min\{X_1, \dots, X_N\}.$$

Note that \mathcal{P}_0 is a natural model for studying the large sample behavior of Y_N^0 , since it is precisely the class of underlying distributions for which the distributions $\tau_N(P)$ of Y_N^0 under P converge weakly to a non-degenerate limit, which is an exponential distribution with parameter $a(P)$.

Bootstrapping the sample minimum is a problem of some notoriety as it is an early example where the usual choice of the empirical distribution P_N for the resampling distribution \hat{P}_N does not work. To check whether the bootstrap with a different choice of \hat{P}_N will work for “most” $P \in \mathcal{P}_0$, we may appeal to Proposition 5. In doing so, we are still free to choose a metric p on \mathcal{P}_0 and we shall make the usual choice by taking p to be Hellinger distance H . Since Y_N^0 is a function of X_1, \dots, X_N only, and not of P , it is easy to see that $\tau_N : (\mathcal{P}_0, H) \rightarrow (\mathcal{R}, \ell)$ is continuous for each N . As $\tau_N(P)$ converges weakly to a limit $\tau(P)$ for every $P \in \mathcal{P}_0$, Proposition 5 asserts that if \hat{P}_N is a Hellinger consistent estimator with values in \mathcal{P}_0 , then the bootstrap $\tau_N(\hat{P}_N)$ will work except for P in a set

D of the first category in (\mathcal{P}_0, H) . The content of Proposition 2 having made us somewhat suspicious, we may want to investigate the nature of the exceptional set D where the functions $\tau_N : (\mathcal{P}_0, H) \rightarrow (\mathcal{R}, \ell)$ are not equicontinuous. Since $\tau(P)$ depends on P only through $a(P)$, and any $P \in \mathcal{P}_0$ may be approximated arbitrarily well in Hellinger distance by a sequence $P_r \in \mathcal{P}_0$ with a constant value of $a(P_r)$ different from $a(P)$, we know that the limit distribution τ is nowhere continuous in P . This implies that the functions τ_N are not equicontinuous at any $P \in \mathcal{P}_0$, so that our worst suspicions are confirmed: the exceptional set D equals the entire set of possible distributions in this case. Our application of Proposition 5 with $p = H$ has therefore produced no positive information concerning this example at all.

Even though Proposition 5 is vacuous in this case, it might still by a stroke of luck be true that the bootstrap estimate $\tau_N(\hat{P}_N)$ would work for most reasonable Hellinger-consistent estimators \hat{P}_N of P . First of all we note that it is indeed possible to construct an estimator of P which is Hellinger-consistent for every distribution P on \mathbb{R} which has no singular part (cf. Devroye & Győrfi 1990, p. 1497). All we have to do is to assign probability k/N to all values which were observed $k > 1$ times, and add a kernel density estimator based on the remaining values which have only been observed once. Using the normal kernel we arrive at an estimator F_N^* for the distribution function F of P which is given by

$$(7) \quad F_N^*(x) = \frac{1}{N} \sum_{i=1}^N \left(\delta_i 1_{(0,x]}(X_i) + (1 - \delta_i) \Phi \left(\frac{x - X_i}{h_N} \right) \right)$$

where Φ is the standard normal distribution function,

$$(8) \quad \delta_i = \begin{cases} 0 & \text{if } X_j \neq X_i \text{ for } j \neq i, \\ 1 & \text{otherwise,} \end{cases}$$

and $h_N \rightarrow 0$ but $Nh_N \rightarrow \infty$. Admittedly, F_N^* does not satisfy (1) and hence the corresponding estimator P_N^* of P does not take its values in \mathcal{P}_0 as is required in Proposition 5. However this defect is easily cured by considering the following slight modification of F_N^* ,

$$(9) \quad \hat{F}_N(x) = \begin{cases} x F_N^*(M_N)/M_N & \text{for } 0 \leq x < M_N, \\ F_N^*(x) & \text{for } x \geq M_N, \end{cases}$$

where $M_N = \min(X_1, \dots, X_N)$. Clearly \hat{F}_N satisfies (1), and hence the corresponding estimator \hat{P}_N of P takes its values in \mathcal{P}_0 and is Hellinger consistent for every $P \in \mathcal{P}_0$ which has no singular part. Nevertheless we shall see that the bootstrap $\tau_N(\hat{P}_N)$ based on this estimator does not work for any $P \in \mathcal{P}_0$.

The bootstrap $\tau_N(\hat{P}_N)$ has distribution function

$$\hat{H}_N(y) = 1 - \left(1 - \hat{F}_N(y/N) \right)^N.$$

For $P \in \mathcal{P}_0$, the limit distribution $\tau(P)$ is exponential with parameter $a(P) \in (0, \infty)$, and hence the bootstrap $\tau_N(\widehat{P}_N)$ will work for a particular $P \in \mathcal{P}_0$ if and only if

$$\sup_{y>0} |\widehat{H}_N(y) - [1 - \exp\{-a(P)y\}]| \xrightarrow{P} 0.$$

This is easily seen to be equivalent to

$$(10) \quad N\widehat{F}_N\left(\frac{y}{N}\right) - a(P)y \xrightarrow{P} 0,$$

for every $y > 0$.

However, (10) cannot hold. If F_N denotes the empirical distribution function, (7) implies that for all x ,

$$F_N^*(x) \geq \frac{1}{N} \sum_{i=1}^N 1_{(0,x]}(X_i) \left[\delta_i + \frac{1 - \delta_i}{2} \right] \geq 1/2 F_N(x).$$

As $F_N(x) = 0$ for $x < M_N$, we also find that for all x ,

$$\widehat{F}_N(x) \geq 1/2 F_N(x).$$

Hence, for every $y > 0$, the definition (1) ensures that as $N \rightarrow \infty$,

$$\begin{aligned} P\left(|N\widehat{F}_N\left(\frac{y}{N}\right) - a(P)y| \geq a(P)y\right) &\geq P\left(N\widehat{F}_N\left(\frac{y}{N}\right) \geq 2a(P)y\right) \\ &\geq P\left(N\widehat{F}_N\left(\frac{y}{N}\right) \geq 4a(P)y\right) = P\left(Z \geq 4a(P)y\right) + o(1) \not\rightarrow 0, \end{aligned}$$

where Z has a Poisson distribution with expectation $a(P)y > 0$. This shows that (10) is false, and as a consequence, the bootstrap based on \widehat{P}_N does not work for any $P \in \mathcal{P}_0$, and the fiasco is indeed complete.

20.4 A bootstrap success

Luckily, the disastrous results of the previous section also indicate quite clearly how the damage may be repaired. Our problems in the previous section originate from the fact that the parameter of the exponential limit distribution $a(P)$ is not a continuous function of the underlying distribution $P \in \mathcal{P}_0$ with respect to Hellinger distance on \mathcal{P}_0 . Hence we should look for a different metric on \mathcal{P}_0 , and in view of the definition of $a(P)$ in (1), one obvious candidate is a metric π defined by

$$(11) \quad \pi(P, Q) = \sup_{x>0} \frac{|F(x) - G(x)|}{x}$$

where F and G denote the distribution functions corresponding to P and Q .

With this new metric π , things immediately fall into place. The metric space (\mathcal{P}_0, π) is easily seen to be complete and hence sets of the first category have dense complements. Clearly the exponential limit distribution

$\tau(P)$ is continuous when viewed as a map $\tau : (\mathcal{P}_0, \pi) \rightarrow (\mathcal{R}, \ell)$. Also, the sequence of distributions $\tau_N(P)$ of Y_N is equicontinuous on \mathcal{P}_0 . To see this, note that for underlying distributions P and Q with distribution functions F and G , Y_N has distribution functions

$$H_{N,P}(y) = 1 - (1 - F(y/N))^N$$

and

$$H_{N,Q}(y) = 1 - (1 - G(y/N))^N.$$

Fix $P \in \mathcal{P}_0$ and $0 < \epsilon < 1$. Choose positive numbers y_0 and z_0 such that

$$y_0 = \frac{4 \log(4/\epsilon)}{a(P)} \quad \text{and} \quad F(z) \geq 1/2 a(P) z \quad \text{for } 0 \leq z \leq z_0$$

and note that $|a^N - b^N| \leq N|a - b|$ if $0 \leq a, b \leq 1$. If $N \geq y_0/z_0$ we choose $\pi(P, Q) \leq \frac{\epsilon}{2y_0} \leq \frac{a(P)}{4}$ and find

$$\begin{aligned} \sup_y |H_{N,P}(y) - H_{N,Q}(y)| &= \sup_y \left| (1 - F(y/N))^N - (1 - G(y/N))^N \right| \\ &\leq N \sup_{y \leq y_0} |F(y/N) - G(y/N)| + (1 - F(y_0/N))^N + (1 - G(y_0/N))^N \\ &\leq y_0 \pi(P, Q) + \exp\{-NF(y_0/N)\} + \exp\{-NG(y_0/N)\} \\ &\leq \frac{\epsilon}{2} + \exp\{-1/2 a(P) y_0\} + \exp\{-1/2 a(P) y_0 + y_0 \pi(P, Q)\} \\ &\leq \frac{\epsilon}{2} + \left(\frac{\epsilon}{4}\right)^2 + \exp\{-1/4 a(P) y_0\} = \frac{\epsilon}{2} + \frac{\epsilon^2}{16} + \frac{\epsilon}{4} < \epsilon. \end{aligned}$$

On the other hand, if $1 \leq N < y_0/z_0$, we choose y_1 such that

$$1 - F\left(\frac{z_0 y_1}{y_0}\right) \leq \frac{\epsilon}{4}.$$

For $\pi(P, Q) \leq \frac{\epsilon}{4y_1}$ we find

$$\begin{aligned} \sup_y |H_{N,P}(y) - H_{N,Q}(y)| &\leq y_1 \pi(P, Q) + (1 - F(y_1/N))^N + (1 - G(y_1/N))^N \\ &\leq \frac{\epsilon}{4} + (1 - F(y_1/N)) + \left(1 - F(y_1/N) + \frac{y_1}{N} \pi(P, Q)\right) \\ &\leq \frac{\epsilon}{4} + \left(1 - F\left(\frac{z_0 y_1}{y_0}\right)\right) + \left(1 - F\left(\frac{z_0 y_1}{y_0}\right) + \frac{\epsilon}{4}\right) \leq \epsilon. \end{aligned}$$

Hence for every $0 < \epsilon < 1$ there exists $\delta > 0$ depending on P but not on Q , such that $\pi(P, Q) \leq \delta$ implies $\sup_y |H_{N,P}(y) - H_{N,Q}(y)| \leq \epsilon$ for all N , which establishes the equicontinuity of $\{\tau_N\}$ on \mathcal{P}_0 .

By Proposition 4 the equicontinuity of $\tau_N : (\mathcal{P}_0, \pi) \rightarrow (\mathcal{R}, \ell)$ implies that the bootstrap $\tau_N(\tilde{P}_N)$ will work for all $P \in \mathcal{P}_0$ if \tilde{P}_N is a π -consistent estimator of P . An example of such an estimator is the random distribution

\tilde{P}_N with distribution function

$$(12) \quad \tilde{F}_N(x) = \begin{cases} xF_N(\xi_N)/\xi_N & \text{for } 0 \leq x < \xi_N \\ F_N(x) & \text{for } x \geq \xi_N \end{cases},$$

where F_N denotes the empirical distribution function, $\xi_N \rightarrow 0$ but $N\xi_N \rightarrow \infty$. To see this, let F denote the distribution function corresponding to the underlying distribution $P \in \mathcal{P}_0$. Then $F_N(\xi_N)/\xi_N \xrightarrow{P} a(P)$ if $\xi_N \rightarrow 0$ and $N\xi_N \rightarrow \infty$, since

$$E \frac{F_N(\xi_N)}{\xi_N} = \frac{F(\xi_N)}{\xi_N} \rightarrow a(P)$$

and

$$\sigma^2 \left(\frac{F_N(\xi_N)}{\xi_N} \right) = \frac{F(\xi_N)(1 - F(\xi_N))}{N\xi_N^2} \rightarrow 0.$$

It follows that, if $\xi_N \rightarrow 0$ and $N\xi_N \rightarrow \infty$ as $N \rightarrow \infty$, then

$$\sup_{x \leq \xi_N} \frac{|\tilde{F}_N(x) - F(x)|}{x} \leq \left| \frac{F_N(\xi_N)}{\xi_N} - a(P) \right| + \sup_{x \leq \xi_N} \left| \frac{F(x)}{x} - a(P) \right| \xrightarrow{P} 0.$$

Also, for every sequence $\eta_N \rightarrow 0$ with $\eta_N > \xi_N$,

$$\sup_{\xi_N \leq x \leq \eta_N} \frac{|\tilde{F}_N(x) - F(x)|}{x} = \sup_{\xi_N \leq x \leq \eta_N} \frac{|F_N(x) - F(x)|}{F(x)/a(P)} + o(1)$$

and this tends to zero in probability if $N\xi_N \rightarrow \infty$ (cf. Chang 1955, Theorem 1; see also Shorack & Wellner 1986, p. 424). Finally, taking η_N such that $N^{1/2}\eta_N \rightarrow \infty$, we have

$$\sup_{x > \eta_N} \frac{|F_N(x) - F(x)|}{x} \leq \sup_x \frac{|F_N(x) - F(x)|}{\eta_N} = O_P(N^{-1/2}\eta_N^{-1}).$$

Thus the π -consistency of \tilde{F}_N follows and we have shown

(13) **Proposition** *If \tilde{P}_N is an estimator of P with distribution function \tilde{F}_N given by (12), then the bootstrap estimator $\tau_N(\tilde{P}_N)$ of the distribution $\tau_N(P)$ of Y_N^0 is consistent for all $P \in \mathcal{P}_0$.*

Acknowledgments: This research was supported by the Netherlands' Organization for Scientific Research (NWO) and by the Sonderforschungsbereich 343 "Diskrete Strukturen in der Mathematik" at the University of Bielefeld, German Federal Republic.

We are grateful to David Pollard for his careful reading of this paper which led to many improvements.

20.5 REFERENCES

- Chang, L.-C. (1955), ‘On the ratio of the empirical distribution to the theoretical distribution function’, *Acta Math. Sinica* 5, 347–368. (English Translation in: *Selected Translations in Mathematical Statistics and Probability*, 4, 17–38 (1964).).
- Devroye, L. & Györfi, L. (1990), ‘No empirical probability measure can converge in the total variation sense for all distributions’, *Annals of Statistics* 18, 1496–1499.
- Dudley, R. M. (1989), *Real Analysis and Probability*, Wadsworth, Belmont, California.
- Le Cam, L. (1953), ‘On some asymptotic properties of maximum likelihood estimates and related Bayes estimates’, *University of California Publications in Statistics* 1, 277–330.
- Putter, H. (1994), Consistency of Resampling Methods, PhD thesis, University of Leiden.
- Putter, H. & van Zwet, W. R. (1993), Consistency of plug-in estimators with application to the bootstrap, Technical report, University of Leiden.
- Shorack, G. R. & Wellner, J. A. (1986), *Empirical Processes with Applications to Statistics*, Wiley, New York.

21

A Limiting Distribution Theorem

C. R. Rao¹
L. C. Zhao²

ABSTRACT Under some mild conditions, we establish the limiting distribution of a test statistic proposed by Ebrahimi & Habibullah (1992) for testing exponentiality based on sample entropy.

21.1 Introduction and main results

In survival studies it is frequently assumed that the life of a product has an exponential distribution, and the justification of such a distribution based on observed data would be of some practical interest. In the literature, considerable attention has been paid to testing the hypothesis of exponentiality, and some important results are reported in Ebrahimi & Habibullah (1992).

Now let X_1, \dots, X_n be an iid sample from a distribution F with a density $f(x)$ over a non-negative support and with mean $\mu < \infty$, and let $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ be the order statistics. The hypothesis of interest is

$$H_0 : f(x) = f_0(x, \lambda) \equiv \lambda e^{-\lambda x} \quad \text{against} \quad H_1 : f(x) \neq f_0(x, \lambda) \quad (1)$$

where $\lambda = 1/\mu$ is specified or unspecified. To test H_0 , Ebrahimi & Habibullah use the Kullback-Leibler information function between two distributions given by

$$I(F; F_0) = \int_0^\infty f(x) \log (f(x)/f_0(x)) dx = -H(F) - \log \lambda + 1 \quad (2)$$

with

$$H(F) = - \int_0^\infty f(x) \log f(x) dx, \quad (3)$$

¹The Pennsylvania State University

²The Pennsylvania State University and University of Science and Technology of China, China.

the entropy of F . The suggested test criterion is

$$I_{mn} = -H_{mn} + \log \bar{X} + 1 \quad \text{with} \quad \bar{X} = n^{-1} \sum_{i=1}^n X_i, \quad (4)$$

or

$$KL_{mn} = \exp(-I_{mn}) = e^{-1} \hat{\lambda} \exp(H_{mn}), \quad (5)$$

where $\hat{\lambda} = 1/\bar{X}$ is an estimate of λ and

$$H_{mn} = n^{-1} \sum_{i=1}^n \log \left(\frac{n}{2m} (X_{(i+m)} - X_{(i-m)}) \right) \quad (6)$$

is the sample entropy, proposed by Vasicek (1976). Here, the window width m is a positive integer smaller than $n/2$, $X_{(j)} = X_{(1)}$ if $j < 1$ and $X_{(j)} = X_{(n)}$ if $j > n$. Note that large values of I_{mn} or small values of KL_{mn} favour H_1 . As pointed out in Ebrahimi & Habibullah (1992), the sampling distribution of I_{mn} is intractable, so they determine the critical points of the test by means of Monte Carlo simulations. No limiting distribution is available. Also, by means of Monte Carlo simulations, the power of the proposed test under various alternatives, say, Weibull, gamma and log-normal distributions, is compared with that of other leading tests suggested by Van-Soest (1969) and Finkelstein & Schafer (1971). The simulations with small sample-sizes showed that the test (4) performs very well.

In this paper, we establish the limiting distribution of the statistic I_{mn} . We have the following main theorem.

Theorem 1 *Assume that*

$$m/\log n \rightarrow \infty \quad (7)$$

and

$$m(\log n)^{2/3}/n^{1/3} \rightarrow 0 \quad (8)$$

as $n \rightarrow \infty$. Then, under H_0

$$(6mn)^{1/2} (I_{mn} - \log(2m) - \gamma + R_{2m-1}) \xrightarrow{\mathcal{L}} N(0, 1), \quad (9)$$

where

$$R_m = \sum_{j=1}^m 1/j, \quad (10)$$

and

$$\gamma = \lim_{n \rightarrow \infty} (R_n - \log n) \quad (11)$$

is the Euler constant.

The proof of the theorem is given in Section 3 and some necessary lemmas are given in Section 2.

21.2 Some Lemmas

In order to prove the theorem we establish some lemmas. Throughout this paper, we denote by c a constant which may take different values in different places.

Let U_1, \dots, U_n be iid random variables uniformly distributed in the interval $(0,1)$ and let W_1, \dots, W_{n+1} be iid standard exponential random variables with the mean 1. For convenience we put $U_{(i)}^*$ equal to 0, $U_{(i)}$, or 1 according to $i = 0$, $0 \leq i \leq n$, or $i = n+1$. The following facts are well known.

$$-\log(1 - U_i) \stackrel{d}{=} W_i \quad \text{for } 1 \leq i \leq n. \quad (12)$$

$$\left(U_{(i)}^* - U_{(i-1)}^* \right)_{1 \leq i \leq n+1} \stackrel{d}{=} \left(\frac{W_i}{S} \right)_{1 \leq i \leq n+1} \text{ where } S = \sum_{i=1}^{n+1} W_i. \quad (13)$$

$$\begin{aligned} (-\log(1 - U_{(r)}))_{1 \leq r \leq n} &\stackrel{d}{=} (-\log U_{(n-r+1)})_{1 \leq r \leq n} \\ &\stackrel{d}{=} \left(\sum_{i=1}^r \frac{W_i}{n-i+1} \right)_{1 \leq r \leq n} \end{aligned} \quad (14)$$

For (13) and (14), reference may be made to Reiss (1989, pages 40 and 37).

We have the following lemmas.

Lemma 2 . Assume that (7) holds. Then

$$T_n = \max^* \left| \left((n+1)(U_{(j)}^* - U_{(i)}^*) / (j-i) \right) - 1 \right| \rightarrow 0 \quad a.s..$$

Hereafter, \max^* and \sum^* are taken for all (i,j) with $0 \leq i < j \leq n+1$ and $j-i \geq cm$.

Proof: By (13),

$$T_n \stackrel{d}{=} \max^* \left| \frac{(W_{i+1} + \dots + W_j) / (j-i)}{S/(n+1)} - 1 \right|.$$

First, we show that for any $\epsilon \in (0, 1/2)$,

$$\begin{aligned} \sum_n P(T_n \geq \epsilon) &\leq \sum_n P \left(\left| \frac{S}{n+1} - 1 \right| \geq \frac{\epsilon}{3} \right) \\ &\quad + \sum_n \sum^* P \left(\left| \frac{W_{i+1} + \dots + W_j}{j-i} - 1 \right| \geq \frac{\epsilon}{3} \right). \end{aligned} \quad (15)$$

For simplicity, we write $(i,j) \in N^*$ iff $0 \leq i < j \leq n+1$ and $j-i \geq cm$. Write $(W_{i+1} + \dots + W_j) / (j-i) = \bar{W}(i,j)$. To prove (15), we only need to show that for any $\epsilon \in (0, 1/2)$,

$$\{T_n \geq \epsilon\} \subset \left\{ \left| \frac{S}{n+1} - 1 \right| \geq \frac{\epsilon}{3} \right\} \cup \left\{ \max^* |\bar{W}(i,j) - 1| \geq \frac{\epsilon}{3} \right\}. \quad (16)$$

In fact, if the event on the right hand side of (16) does not happen, then $1 - \epsilon/3 < S/(n+1) < 1 + \epsilon/3$ and for any $(i, j) \in N^*$,

$$1 - \epsilon/3 < \bar{W}(i, j) < 1 + \epsilon/3$$

and

$$1 - \epsilon < \frac{1 - \epsilon/3}{1 + \epsilon/3} < \frac{\bar{W}(i, j)}{S/(n+1)} < \frac{1 + \epsilon/3}{1 - \epsilon/3} < 1 + \epsilon.$$

This means that $T_n < \epsilon$, and (15) is proved.

With $t = \epsilon/(1 + \epsilon)$ and $c = \epsilon - \log(1 + \epsilon) > 0$, we have

$$\begin{aligned} P\left\{\frac{W_1 + \dots + W_k}{k} - 1 \geq \epsilon\right\} &\leq \exp(-tk(1 + \epsilon))(Ee^{tW_1})^k \\ &= \exp(-tk(1 + \epsilon))(1 - t)^{-k} \\ &= \exp(-tk(1 + \epsilon) - k \log(1 - t)) \\ &= \exp(-ck). \end{aligned}$$

Similarly,

$$P\left\{\frac{W_1 + \dots + W_k}{k} - 1 \leq -\epsilon\right\} \leq \exp(-ck).$$

From these, (15) and (7), it follows that

$$\sum_n P(T_n \geq \epsilon) \leq \sum_n 2e^{-cn} + \sum_n \sum^* 2e^{-cm} \leq C \sum_n n^2 e^{-cm} < \infty.$$

By the Borel-Cantelli lemma, Lemma 2 follows.

Write

$$H_{mn}(U) = n^{-1} \sum_{i=1}^n \log \left(\frac{n}{2m} (U_{(i+m)} - U_{(i-m)}) \right). \quad (17)$$

where $U_{(j)} = U_{(1)}$ if $j < 1$ and $U_{(j)} = U_{(n)}$ if $j > n$.

Lemma 3 Assume that $m \rightarrow \infty$ and (8) holds when $n \rightarrow \infty$. Then

$$(6mn)^{1/2} (-H_{mn}(U) - \log(2m) - \gamma + R_{2m-1}) \xrightarrow{\mathcal{L}} N(0, 1). \quad (18)$$

Note that Dudewicz & Van der Meulen (1981) obtained this result by assuming that $m \rightarrow \infty$ and $m = O(n^{(1/3)-\delta})$ for some $0 < \delta < 1/3$ as $n \rightarrow \infty$. In such a case the assumption (8) is satisfied. Here we give a finer analysis.

Proof: Write

$$L_1(m, n) = \sum_{i=0}^{n-m+1} \log(U_{(i+m)}^* - U_{(i)}^*), \quad (19)$$

$$L_2(2m, n) = \sum_{i=1}^n \log(U_{(i+m)} - U_{(i-m)}),$$

$$\Delta(m, n) = n^{-1} (L_2(2m, n) - L_1(2m, n)). \quad (20)$$

By Cressie (1976, Corollary 4.1), if $m \rightarrow \infty$ and $m/n^{1/3} \rightarrow 0$ as $n \rightarrow \infty$, then

$$3mn^{1/2}n^{-1} \left(-L_1(m, n) - (n+2-m)(\log(n+1) + \gamma - R_{m-1}) \right) \xrightarrow{\mathcal{L}} N(0, 1),$$

by which, we have

$$(6mn)^{1/2} \left(-\frac{L_1(2m, n)}{n} - \left(1 - \frac{2m-2}{n} \right) (\log(n+1) - \gamma - R_{2m-1}) \right) \xrightarrow{\mathcal{L}} N(0, 1) \quad (21)$$

Using (8),

$$m^{3/2}n^{-1/2} (\log(n+1) - \gamma - R_{2m-1}) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Then from (21), we have

$$(6mn)^{1/2} (-n^{-1}L_1(2m, n) - \log n - \gamma + R_{2m-1}) \xrightarrow{\mathcal{L}} N(0, 1). \quad (22)$$

It is easily shown that

$$-n^{-1}L_2(2m, n) = -H_{mn}(U) + \log n - \log(2m), \quad (23)$$

and

$$\begin{aligned} n\Delta(m, n) &= \sum_{i=1}^m \log(U_{(i+m)} - U_{(1)}) + \sum_{i=n-m-1}^n \log(U_{(n)} - U_{(i-m)}) \\ &\quad - \log U_{(2m)} - \log(1 - U_{(n-2m+1)}). \end{aligned} \quad (24)$$

By (14) and the condition $m(\log n)^2/n \rightarrow 0$, we get

$$\begin{aligned} E(mn)^{1/2}|n^{-1}\log U_{(2m)}| \\ &= E(mn)^{1/2}|n^{-1}\log(1 - U_{(n-2m+1)})| \\ &= (m/n)^{1/2}E \sum_{i=1}^{n-2m+1} \frac{W_i}{n-i+1} = (m/n)^{1/2} \sum_{i=1}^{n-2m+1} \frac{1}{n-i+1} \\ &\leq c(m/n)^{1/2} \log n \rightarrow 0. \end{aligned} \quad (25)$$

By (14) and (8),

$$\begin{aligned} E(mn)^{1/2} \left| n^{-1} \sum_{i=n-m-1}^n \log(U_{(n)} - U_{(i-m)}) \right| \\ &= E(m/n)^{1/2} \left| \sum_{i=n-m-1}^n \log(1 - U_{(i-m+1)}) \right| \\ &= (m/n)^{1/2} \sum_{i=n-m-1}^n \sum_{j=1}^i \frac{1}{n-j+1} \leq cm^{3/2}n^{-1/2} \log n \\ &\rightarrow 0. \end{aligned} \quad (26)$$

In the same way,

$$E(mn)^{1/2} |n^{-1} \sum_{i=1}^m \log(U_{(i+m)} - U_{(1)})| \leq cm^{3/2} n^{-1/2} \log n \rightarrow 0. \quad (27)$$

By (24)–(27),

$$(mn)^{1/2} \Delta(m, n) \rightarrow 0 \quad \text{in pr. as } n \rightarrow \infty. \quad (28)$$

By (20), (22), (23) and (28), we get (18), and Lemma 3 is proved.

21.3 Proof of the Theorem

In this section it is assumed that H_0 is true. By (4), without loss of generality we can assume that $\lambda = 1$. Write

$$I_{mn} = \log \bar{X} - (\bar{X} - 1) - H_{mn}(U) - G_n, \quad (29)$$

where

$$G_n = n^{-1} \sum_{i=1}^n \log \left(\frac{1 - U_{(i)}}{U_{(i+m)} - U_{(i-m)}} \log \frac{1 - U_{(i-m)}}{1 - U_{(i+m)}} \right).$$

By the SLLN, $\bar{X} - 1 \rightarrow 0$ a.s.. By the Taylor expansion and (8), with probability one for n large,

$$(mn)^{1/2} |\log \bar{X} - (\bar{X} - 1)| \leq c(mn)^{1/2} (\bar{X} - 1)^2 \xrightarrow{P} 0 \quad \text{as } n \rightarrow \infty. \quad (30)$$

By (30) and Lemma 3, in order to prove the theorem, it is enough to prove that

$$(mn)^{1/2} G_n \xrightarrow{P} 0 \quad \text{as } n \rightarrow \infty. \quad (31)$$

In the following, we will write $G_n = G_{n0} + G_{n1} + G_{n2}$ and establish the convergence of G_n via that of the summands. Let

$$\begin{aligned} Z &= Z_{mn} = \{3m, 3m+1, \dots, n-3m\}, \\ G_{n0} &= n^{-1} \sum_{i \in Z} \log(1 + \Delta_i), \end{aligned} \quad (32)$$

where

$$\Delta_i = \frac{1 - U_{(i)}}{U_{(i+m)} - U_{(i-m)}} \log \frac{1 - U_{(i-m)}}{1 - U_{(i+m)}} - 1. \quad (33)$$

By Lemma 2, for $i \in Z$, there exist random variables ζ_{mn} , θ_{i1} and θ_{i2} such that for $|\theta_{i1}| + |\theta_{i2}| \leq 1$,

$$(U_{(i)} - U_{(i-m)})/(1 - U_{(i)}) = \frac{m}{n-i+1} (1 + \theta_{i1} \zeta_{mn}), \quad (34)$$

$$(U_{(i+m)} - U_{(i)})/(1 - U_{(i)}) = \frac{m}{n-i+1} (1 + \theta_{i2} \zeta_{mn}), \quad (35)$$

and

$$0 \leq \zeta_{mn} \rightarrow 0 \quad \text{a.s. as } n \rightarrow \infty. \quad (36)$$

Since $m/n < 1/2$, by (34)–(36), with probability one for n large,

$$0 \leq \frac{U_{(i)} - U_{(i-m)}}{1 - U_{(i)}} \leq 1/2, \quad 0 \leq \frac{U_{(i+m)} - U_{(i)}}{1 - U_{(i)}} \leq 1/2 \quad \text{for all } i \in Z.$$

By the Taylor expansion, there exist a constant $c > 0$ and random variables θ_i with $|\theta_i| \leq 1$ for all $i \in Z$ such that with probability one for n large, for all $i \in Z$,

$$\begin{aligned} & \log(1 - U_{(i-m)}) - \log(1 - U_{(i+m)}) \\ &= \log\left(1 + \frac{U_{(i)} - U_{(i-m)}}{1 - U_{(i)}}\right) - \log\left(1 - \frac{U_{(i+m)} - U_{(i)}}{1 - U_{(i)}}\right) \\ &= \frac{U_{(i)} - U_{(i-m)}}{1 - U_{(i)}} - \frac{1}{2} \left(\frac{U_{(i)} - U_{(i-m)}}{1 - U_{(i)}}\right)^2 + \frac{U_{(i+m)} - U_{(i)}}{1 - U_{(i)}} \\ &\quad + \frac{1}{2} \left(\frac{U_{(i+m)} - U_{(i)}}{1 - U_{(i)}}\right)^2 + c\theta_i \left(\frac{U_{(i+m)} - U_{(i-m)}}{1 - U_{(i)}}\right)^3 \\ &= \frac{U_{(i+m)} - U_{(i-m)}}{1 - U_{(i)}} \left(1 + \frac{U_{(i+m)} + U_{(i-m)} - 2U_{(i)}}{2(1 - U_{(i)})}\right) \\ &\quad + c\theta_i \left(\frac{U_{(i+m)} - U_{(i-m)}}{1 - U_{(i)}}\right)^3. \end{aligned} \quad (37)$$

and

$$\Delta_i = \frac{U_{(i+m)} + U_{(i-m)} - 2U_{(i)}}{2(1 - U_{(i)})} + c\theta_i \left(\frac{U_{(i+m)} - U_{(i-m)}}{1 - U_{(i)}}\right)^2. \quad (38)$$

By (38) and Lemma 2,

$$G_{n0} = (2n)^{-1} \sum_{i \in Z} \frac{U_{(i+m)} + U_{(i-m)} - 2U_{(i)}}{1 - U_{(i)}} + R_n \stackrel{\Delta}{=} Q_n + R_n, \quad (39)$$

and with probability one for n large,

$$\begin{aligned} (mn)^{1/2} R_n &\leq c(mn)^{1/2} n^{-1} \sum_{i \in Z} \left(\frac{U_{(i+m)} - U_{(i-m)}}{1 - U_{(i)}}\right)^2 \\ &\leq c(m/n)^{1/2} \sum_{i \in Z} \left(\frac{2m}{n - i + 1}\right)^2 \leq cm^{3/2} n^{-1/2} \rightarrow 0. \end{aligned} \quad (40)$$

Here we used (8). By (39) and (13),

$$Q_n \stackrel{d}{=} (2n)^{-1} \sum_{i=3m}^{n-3m} \frac{(W_{i+1} + \dots + W_{i+m}) - (W_i + \dots + W_{i-m+1})}{W_{n+1} + \dots + W_{i+1}}$$

$$\begin{aligned}
&= (2n)^{-1} \sum_{j=4m}^{n-2m} \frac{W_j + \dots + W_{j-m+1}}{W_{n+1} + \dots + W_{j-m+1}} \\
&\quad - (2n)^{-1} \sum_{j=3m}^{n-3m} \frac{W_j + \dots + W_{j-m+1}}{W_{n+1} + \dots + W_{j+1}} \\
&\stackrel{\Delta}{=} J_{n1} + J_{n2} + J_{n3}.
\end{aligned} \tag{41}$$

where

$$\begin{aligned}
J_{n1} &= -(2n)^{-1} \sum_{j=4m}^{n-3m} \frac{(W_j + \dots + W_{j-m+1})^2}{(W_{n+1} + \dots + W_{j-m+1})(W_{n+1} + \dots + W_{j+1})} \\
J_{n2} &= (2n)^{-1} \sum_{j=n-3m+1}^{n-2m} \frac{W_j + \dots + W_{j-m+1}}{W_{n+1} + \dots + W_{j-m+1}} \\
J_{n3} &= -(2n)^{-1} \sum_{j=3m}^{4m-1} \frac{W_j + \dots + W_{j-m+1}}{W_{n+1} + \dots + W_{j+1}}.
\end{aligned}$$

Note that if ξ and η are independent gamma random variables with densities

$$g(x) = x^{u-1} e^{-x}/\Gamma(u), h(x) = x^{v-1} e^{-x}/\Gamma(v)$$

respectively, where $u > 0$ and $v > 2$, then

$$E(\xi/\eta) = E(\xi)E(1/\eta) = u/(v-1), \tag{42}$$

$$E(\xi/\eta)^2 = E(\xi^2)E(1/\eta^2) = \frac{u(u+1)}{(v-1)(v-2)}. \tag{43}$$

By (43) and (8),

$$\begin{aligned}
E|(mn)^{1/2}J_{n1}| &\leq (m/n)^{1/2} \sum_{j=4m}^{n-3m} E\left(\frac{W_j + \dots + W_{j-m+1}}{W_{n+1} + \dots + W_{j+1}}\right)^2 \\
&= (m/n)^{1/2} \sum_{j=4m}^{n-3m} \frac{m(m+1)}{(n-j)(n-j-1)} \\
&\leq cm^{3/2}n^{-1/2} \rightarrow 0.
\end{aligned} \tag{44}$$

By (42) and (8),

$$\begin{aligned}
E|(mn)^{1/2}J_{n2}| &\leq (m/n)^{1/2} \sum_{j=n-3m+1}^{n-2m} E\frac{W_j + \dots + W_{j-m+1}}{W_{n+1} + \dots + W_{j+1}} \\
&= (m/n)^{1/2} \sum_{j=n-3m+1}^{n-2m} \frac{m}{n-j} \\
&\leq cm^{3/2}n^{-1/2} \rightarrow 0.
\end{aligned} \tag{45}$$

and

$$E|(mn)^{1/2}J_{n3}| \leq cm^{5/2}n^{-3/2} \rightarrow 0. \quad (46)$$

From (39)–(41) and (44)–(46), it follows that

$$(mn)^{1/2}G_{n0} \xrightarrow{P} 0 \quad \text{as } n \rightarrow \infty. \quad (47)$$

For $i \leq 3m - 1$, it is easily seen that

$$\begin{aligned} 0 &\leq \frac{U_{(i)} - U_{(i-m)}}{1 - U_{(i)}} \leq \frac{U_{(3m)}}{1 - U_{(3m)}} \rightarrow 0 \quad \text{a.s.} \\ 0 &\leq \frac{U_{(i+m)} - U_{(i)}}{1 - U_{(i)}} \leq \frac{U_{(4m)}}{1 - U_{(3m)}} \rightarrow 0 \quad \text{a.s..} \end{aligned}$$

Using an argument similar to that for getting (37), with probability one for n large we have for all $i \leq 3m - 1$,

$$\begin{aligned} |\Delta_i| &\leq c(U_{(i+m)} - U_{(i-m)})/(1 - U_{(i)}) \\ &\leq cU_{(4m)}/(1 - U_{(3m)}) \rightarrow 0 \quad \text{a.s.,} \end{aligned} \quad (48)$$

and by (8),

$$\begin{aligned} (mn)^{1/2}|G_{n1}| &\triangleq (mn)^{1/2}|n^{-1} \sum_{i=1}^{3m-1} \log(1 + \Delta_i)| \\ &\leq c(m/n)^{1/2}3mU_{(4m)}/(1 - U_{(3m)}) \rightarrow 0 \quad \text{a.s.} \end{aligned} \quad (49)$$

Write

$$G_{n2} = n^{-1} \sum_{i=n-3m+1}^n \log \left(\frac{1 - U_{(i)}}{U_{(i+m)} - U_{(i-m)}} \log \frac{1 - U_{(i-m)}}{1 - U_{(i+m)}} \right). \quad (50)$$

We have

$$\begin{aligned} (mn)^{1/2}G_{n2} &\leq (m/n)^{1/2}3m \log \log \frac{1 - U_{(n-4m)}}{1 - U_{(n)}} \\ &\quad + (m/n)^{1/2} \sum_{i=n-3m+1}^n \log \frac{1 - U_{(n-3m)}}{U_{(i)} - U_{(i-m)}}. \end{aligned} \quad (51)$$

By (13), for any $\epsilon > 0$,

$$\begin{aligned} P\left(\frac{1 - U_{(n-4m)}}{1 - U_{(n)}} \geq \epsilon m^3\right) \\ = P\left(\frac{W_{(n+1)} + \dots + W_{n-4m+1}}{W_{n+1}} \geq \epsilon m^3\right) \\ \leq P\left(W_{n+1} \leq \frac{1}{m}\right) + P\left(\sum_{j=0}^{4m} W_{n-j+1} \geq \epsilon m^2\right) \\ \rightarrow 0. \end{aligned} \quad (52)$$

By Lemma 2,

$$\frac{1 - U_{(n-3m)}}{U_{(i)} - U_{(i-m)}} \rightarrow 3 \quad \text{a.s.} \quad (53)$$

uniformly for $n - 3m < i \leq n$. By (51)–(53) and (8),

$$(mn)^{1/2}G_{n2} \leq m^{3/2}n^{-1/2}o_p(\log \log m) + m^{3/2}n^{-1/2}O_p(1) \rightarrow 0. \quad (54)$$

Now we have

$$\begin{aligned} -(mn)^{1/2}G_{n2} &\leq (m/n)^{1/2}3m \log \frac{1 - U_{(n-4m)}}{1 - U_{(n)}} \\ &\quad -(m/n)^{1/2} \sum_{i=n-3m+1}^n \log \log \frac{1 - U_{(i-m)}}{1 - U_{(i)}}. \end{aligned} \quad (55)$$

By Lemma 2,

$$\frac{1 - U_{(i-m)}}{1 - U_{(i)}} = 1 + \frac{U_{(i)} - U_{(i-m)}}{1 - U_{(i)}} \geq 1 + \frac{U_{(i)} - U_{(i-m)}}{1 - U_{(n-3m)}} \rightarrow 1 + 1/3 \quad \text{a.s.} \quad (56)$$

uniformly for $n - 3m < i \leq n$.

By (55), (52), (56) and (8), there exists a constant $\delta \in (0, 1/3)$ such that with probability one for n large

$$\begin{aligned} -(mn)^{1/2}G_{n2} &\leq 3m^{3/2}n^{-1/2} \log \frac{1 - U_{(n-4m)}}{1 - U_{(n)}} \\ &\quad +(m/n)^{1/2}3m \log \left(\log \left(\frac{4}{3} - \delta \right) \right)^{-1} \\ &= o_p(m^{3/2}n^{-1/2} \log m) + O_p(m^{3/2}n^{-1/2}) \rightarrow 0. \end{aligned} \quad (57)$$

By (54) and (57),

$$(mn)^{1/2}G_{n2} \xrightarrow{P} 0 \quad \text{as } n \rightarrow \infty. \quad (58)$$

Now (31) follows from (47), (49), (58) and the fact that $G_n = G_{n0} + G_{n1} + G_{n2}$, and the proof of the theorem is completed.

Acknowledgments: The research work of this paper is sponsored by the Army Research office under the Grant DAAH04-93-G-0030.

21.4 REFERENCES

- Cressie, N. (1976), 'On the logarithms of high-order spacings', *Biometrika* **63**, 343–355.

- Dudewicz, E. & Van der Meulen, E. (1981), 'Entropy based tests of uniformity', *Journal of the American Statistical Association* **76**, 967–974.
- Ebrahimi, N. & Habibullah, M. (1992), 'Testing exponentiality based on kullback-leibler information', *Journal of the Royal Statistical Society, Series B* **54**, 739–748.
- Finkelstein, J. & Schafer, R. E. (1971), 'Improved goodness of fit tests', *Biometrika* **8**, 641–645.
- Reiss, R. D. (1989), *Approximate Distributions of Order Statistics with Applications to Nonparametric Statistics*, Springer-Verlag.
- Van-Soest, J. (1969), 'Some goodness of fit tests for the exponential distribution', *Statistica Neerlandica* **23**, 41–51.
- Vasicek, O. (1976), 'A test for normality based on sample entropy', *Journal of the Royal Statistical Society, Series B* **38**, 54–59.

22

Minimum Distance Estimates with Rates under ϕ -mixing

George G. Roussas¹
Yannis G. Yatracos²

ABSTRACT On the basis of the segment of observations X_1, \dots, X_n from a ϕ -mixing sequence of random variables, a minimum distance estimate \hat{P}_n of the probability measure P , governing the process, is constructed. Under suitable regularity conditions, it is shown that \hat{P}_n is weakly uniformly consistent, within the class \mathcal{P} of assumed probability measures, at the same rate as in the independent identically distributed case. Strengthening of the underlying assumptions provides for strong consistency.

22.1 Introduction

Most statistical inference in the literature is carried out under the basic assumption that the underlying observations are independent and identically distributed (i.i.d.). Practical considerations, however, dictate to suppress at least the independence assumption and replace it by a suitable mode of dependence. The kind of dependence to be adopted should accommodate a substantial class of problems of practical importance, and also be mathematically tractable. It seems that these two requirements are satisfied by mixing conditions, which have received considerable attention the last few years. The specific mode of mixing to be employed in this paper is ϕ -mixing.

Let \mathcal{P} be a family of probability measures defined on a measurable space $(\mathcal{X}, \mathcal{A})$, and for each $P \in \mathcal{P}$, let $\{X_n : n \geq 1\}$ be a ϕ -mixing sequence of random variables (*r.v.'s*) with mixing coefficient $\phi(\cdot)$. It is assumed that \mathcal{P} is L_1 -totally bounded, and the objective is to construct a minimum distance estimate \hat{P}_n of P , based on the segment of observations X_1, \dots, X_n of the underlying sequence of *r.v.'s*. Such an estimate is actually constructed, and is shown to be uniformly consistent, in the probability sense, at the same rate as that occurring in the i.i.d. case which was investigated by Yatracos (1985). An upper bound in probability is also established for the L_1 -distance, and this bound depends on the Kolmogorov entropy as well

¹ University of California, Davis

² University of California, Santa Barbara, and Université de Montréal

as the mixing coefficient. This is the main result of the paper stated in Theorem 23. Its proof hinges upon an exponential inequality for ϕ -mixing sequences, formulated here as Proposition 7 (see also Corollary 17). A version of these results under stationarity may be found, e.g., in Roussas & Ioannides (1988).

The basic idea involved in the minimum distance approach is that of discretizing \mathcal{P} . This technique was introduced by Le Cam (1960) in a fundamental paper on Locally Asymptotically Normal (LAN) Families of Distributions. It was used to construct a preliminary $n^{1/2}$ -consistent estimate (Le Cam 1969, pages 103–107); also, in abstract estimation theory, by using multiple testing procedures in order to derive estimates of a measure (Le Cam 1973a). In this latter case, the performance of estimates was related, for the first time, to the notion of the metric dimension of \mathcal{P} , when the Hellinger distance between measures was employed. A modification of Le Cam's 1973 method was utilized by Birgé (1983) to show that the rates of convergence obtained in the framework of the Hellinger and L_p -distances were optimal in several instances in density estimation problems. Also, in a series of papers on related work, Yatracos (1985, 1989, 1992) has shown that the rates of convergence in L_1 -distance of minimum distance estimates are related to the Kolmogorov entropy. The estimated quantities are either probability density functions or regression-type functions. In closing this section, it is mentioned that all limits are taken as $n \rightarrow \infty$.

22.2 Some definitions and preliminary results

In all that follows, all *r.v.*'s involved are defined on an underlying probability space $(\mathcal{X}, \mathcal{A}, P)$, which will not be referred to explicitly hereafter. Thus, let Z_n , for $n = 1, 2, \dots$, be *r.v.*'s, and for $1 \leq i < j \leq \infty$, denote by \mathcal{F}_i^j the σ -field induced by the *r.v.*'s $\{Z_n : n = i, i+1, \dots, j\}$.

(1) Definition *The sequence of r.v.'s $\{Z_n : n \geq 1\}$ is said to be ϕ -mixing with mixing coefficient $\phi(n)$ if, as $n \rightarrow \infty$,*

$$\sup \left\{ \frac{|P(A \cap B) - P(A)P(B)|}{P(A)} ; A \in \mathcal{F}_1^k, B \in \mathcal{F}_{k+n}^\infty, k \geq 1 \right\} \leq \phi(n) \downarrow 0;$$

or, equivalently, for all $A \in \mathcal{F}_1^k, B \in \mathcal{F}_{k+n}^\infty$, and $k \geq 1$:

$$(2) \quad |P(A \cap B) - P(A)P(B)| \leq \phi(n)P(A) \quad \text{with } \phi(n) \downarrow 0.$$

It is known that ϕ -mixing is also characterized as in the following proposition. This is a convenient characterization for the purposes of this paper.

(3) Proposition *The sequence of r.v.'s $\{Z_n : n \geq 1\}$ is ϕ -mixing with mixing coefficient $\phi(n)$, if and only if,*

$$(4) \quad \sup\{|P(B | \mathcal{F}_1^k) - P(B)| ; B \in \mathcal{F}_{k+n}^\infty, k \geq 1\} \leq \phi(n) \downarrow 0 \quad \text{a.s.}$$

Proof. A brief justification of this result is as follows: Suppose that the inequality in (4), or its equivalent expression below, holds true; namely,

$$(5) \quad -\phi(n) \leq P(B | \mathcal{F}_1^k) - P(B) \leq \phi(n) \quad \text{a.s.}$$

for all $B \in \mathcal{F}_{k+n}^\infty$ and $k \geq 1$. Since $P(B | \mathcal{F}_1^k)$ is \mathcal{F}_1^k -measurable, integration over any $A \in \mathcal{F}_1^k$ of all three parts in (5) yields:

$$-\phi(n)P(A) \leq P(A \cap B) - P(A)P(B) \leq \phi(n)P(A)$$

for all $B \in \mathcal{F}_{k+n}^\infty$, $k \geq 1$, and this is equivalent to the inequality in (2). Next, suppose that (2) prevails, whereas the inequality in (4) fails. Then there must exist $k_0 \geq 1$ and $B_0 \in \mathcal{F}_{k_0+n}^\infty$ for which:

$$(6) \quad |P(B_0 | \mathcal{F}_1^{k_0}) - P(B_0)| > \phi(n) \quad \text{on a set } A \in \mathcal{F}_1^{k_0} \text{ with } P(A) > 0.$$

Let A^+ and A^- in $\mathcal{F}_1^{k_0}$ be defined by

$$\begin{aligned} A^+ &= A \cap \{P(B_0 | \mathcal{F}_1^{k_0}) - P(B_0) > \phi(n)\}, \\ A^- &= A \cap \{P(B_0 | \mathcal{F}_1^{k_0}) - P(B_0) < -\phi(n)\}. \end{aligned}$$

Observe that at least one of $P(A^+)$ and $P(A^-)$ is positive. Suppose that $P(A^+) > 0$. Then, by (6),

$$P(B_0 | \mathcal{F}_1^{k_0})I_{A^+} - P(B_0)I_{A^+} > \phi(n)I_{A^+},$$

and, by integrating over A^+ ,

$$P(A^+ \cap B_0) - P(A^+)P(B_0) > \phi(n)P(A^+).$$

This, however, violates inequality (2). Similarly, if $P(A^-) > 0$. \square

The following proposition is instrumental in establishing the main result of this paper which is Theorem 23. For the formulation of the proposition, choose positive integers $p = p(n)$ such that $p \rightarrow \infty$ and $r = r(n)$ the largest (positive) integers for which $2pr \leq n$, so that $n/2pr \rightarrow 1$.

(7) Proposition Suppose the r.v.'s $\{Z_n : n \geq 1\}$ are centered at expectation and bounded by M , and they form a ϕ -mixing sequence with ϕ -mixing coefficient $\phi(n)$ such that $\phi^* := \sum_{n=1}^\infty \phi(n) < \infty$. Set $S_n = \sum_{i=1}^n Z_i$, and let p and r be as above. Then, for $C_0 = 1 + 4\phi^*$, and any $0 < \epsilon_n \leq C_0Mr$:

$$(8) \quad P(|S_n| \geq \epsilon_n) \leq 6(1 + 2\sqrt{\epsilon_n}\phi(p))^r \exp\left(-\frac{\epsilon_n^2}{4C_0M^2pr}\right), \quad n \geq 1.$$

Proof. For $k = 2, \dots, r$, and $i, j = 1, \dots, k$, set

$$\begin{aligned} U_i &= Z_{2(i-1)p+1} + \dots + Z_{(2i-1)p}, & V_j &= Z_{(2j-1)p+1} + \dots + Z_{2jp}, \\ U_k^* &= \sum_{i=1}^k U_i, & V_k^* &= \sum_{j=1}^k V_j, & W_r &= Z_{2pr+1} + \dots + Z_n, \end{aligned}$$

so that

$$(9) \quad S_n = U_r^* + V_r^* + W_r.$$

Let $\lambda > 0$ and $2 \leq k \leq r$, and use the $\mathcal{F}_1^{(2k-3)p}$ -measurability of $\sum_{i=1}^{k-1} U_i$ to obtain

$$\begin{aligned} (10) \quad \mathbb{E}e^{\lambda U_k^*} &= \mathbb{E}\left(e^{\lambda \sum_{i=1}^{k-1} U_i} e^{\lambda U_k}\right) \\ &= \mathbb{E}\left(e^{\lambda \sum_{i=1}^{k-1} U_i} \mathbb{E}[e^{\lambda U_k} \mid \mathcal{F}_1^{(2k-3)p}]\right) \\ &= \mathbb{E}\left(e^{\lambda U_{k-1}^*} \left(\mathbb{E}[e^{\lambda U_k} \mid \mathcal{F}_1^{(2k-3)p}] - \mathbb{E}e^{\lambda U_k}\right)\right) \\ &\quad + \mathbb{E}e^{\lambda U_{k-1}^*} \mathbb{E}e^{\lambda U_k}. \end{aligned}$$

Now the *r.v.* U_k is $\mathcal{F}_{2(k-1)p+1}^\infty$ -measurable, $|\lambda U_k| \leq \lambda p M$, and the σ -fields $\mathcal{F}_1^{(2k-3)p}$ and $\mathcal{F}_{2(k-1)p+1}^\infty$ are separated by p units. Therefore, by way of Proposition 3 (for details, see Roussas & Ioannides 1987, Theorem 5.5, page 101), it follows that

$$(11) \quad |\mathbb{E}(e^{\lambda U_k} \mid \mathcal{F}_1^{(2k-3)p}) - \mathbb{E}e^{\lambda U_k}| \leq 2e^{\lambda M p} \phi(p) \quad \text{a.s.}$$

At this point, utilize familiar inequalities about the exponential function, $\exp(t)$, the boundedness of the Z'_n 's, the assumption that they are centered at expectation, and the fact that $\mathbb{E}U_k^2 \leq C_0 M^2 p$ to obtain

$$(12) \quad \mathbb{E}e^{\lambda U_k} \leq e^{C_0 \lambda^2 M^2 p} \quad \text{provided } |\lambda U_k| \leq 1/2,$$

which follows from $\lambda M p \leq 1/2$. By means of (11) and (12), relation (10) becomes:

$$(13) \quad \mathbb{E}e^{\lambda U_k^*} \leq e^{C_0 \lambda^2 M^2 p} (1 + 2\sqrt{e} \phi(p)) \mathbb{E}e^{\lambda U_{k-1}^*} \quad \text{for } 0 < \lambda \leq 1/2 M p.$$

Applying (13) for $k = r, r-1, \dots, 2$, multiplying out the resulting inequalities, and utilizing (12), we get:

$$(14) \quad \mathbb{E}e^{\lambda U_r^*} \leq e^{C_0 \lambda^2 M^2 p r} (1 + 2\sqrt{e} \phi(p))^r \quad \text{for } 0 < \lambda \leq 1/2 M p.$$

For $\epsilon_n > 0$, employ the usual inequalities and minimize the resulting upper bound in order to obtain:

$$(15) \quad P(|U_r^*| \geq \epsilon_n) \leq 2 (1 + 2\sqrt{e} \phi(p))^r e^{-\epsilon_n^2 / 4 C_0 M^2 p r} \quad \text{for } \epsilon_n \leq C_0 M r.$$

Clearly, V_r^* and W_r satisfy the inequality in (15), so that, by (9),

(16)

$$P(|S_n| \geq \epsilon_n) \leq 6 (1 + 2\sqrt{e} \phi(p))^r e^{-\epsilon_n^2 / 4 C_0 M^2 p r} \quad \text{for } 0 < \epsilon_n \leq C_0 M r,$$

as was to be shown. \square

(17) **Corollary** Set $\bar{S}_n = S_n/n$. Then, for $\epsilon_n \leq C_0 M r/n$,

$$(18) \quad P(|\bar{S}_n| \geq \epsilon_n) \leq 6 (1 + 2\sqrt{e} \phi(p))^r \exp\left(-\frac{n \epsilon_n^2}{2 C_0 M^2}\right), \quad n \geq 1.$$

Proof. Follows from (16), replacing ϵ_n by $n\epsilon_n$ and utilizing the fact that $2pr \leq n$. \square

22.3 Minimum distance estimate of a probability measure

Let \mathcal{P} be the family of probability measures described in Section 1. Endow \mathcal{P} with the total variation distance d defined by:

$$(19) \quad d(P, Q) = \|P - Q\| = 2 \sup\{|P(A) - Q(A)|; A \in \mathcal{A}\}, \quad P, Q \in \mathcal{P}.$$

The following result will be used in the sequel.

(20) **Lemma** (Yatracos 1985) Assume that the metric space (\mathcal{P}, d) is totally bounded, let μ be a dominating measure, and for $a > 0$, let $N(a)$ be the number of radius a balls with centers $P_i, i = 1, \dots, N(a)$ needed to cover \mathcal{P} . Then there exists a class $\mathcal{F}_{N(a)} \subseteq \mathcal{A}$ with cardinality $|\mathcal{F}_{N(a)}| \leq N(a)^2$ such that, for any P, Q in \mathcal{P} :

$$(21) \quad \|P - Q\| \leq 4a + 2 \max\{|P(A) - Q(A)|; A \in \mathcal{F}_{N(a)}\},$$

where $\mathcal{F}_{N(a)}$ is the collection of the sets:

$$\{x \in \mathcal{X}; \frac{dP_i}{d\mu}(x) > \frac{dP_j}{d\mu}(x), 1 \leq i < j \leq N(a)\}.$$

On the basis of the segment X_1, \dots, X_n from the underlying process, define the empirical measure μ_n on \mathcal{A} in the usual way: $\mu_n(A) = n^{-1} \sum_{i=1}^n I_A(X_i)$, where $I_A(X_i)$ is the indicator of the event $\{X_i \in A\}$. Next, for $a_n > 0$, let $P_i, i = 1, \dots, N(a_n)$, to be denoted by N_n for simplicity, and \mathcal{F}_{N_n} be the quantities cited in Lemma 20. Then estimate the (unknown) probability measure P , governing the process, by the probability measure \hat{P}_n which is that one among the $P_i, i = 1, \dots, N_n$, which minimizes the quantities: $\max\{|\mu_n(A) - P_i(A)|; A \in \mathcal{F}_{N_n}, i = 1, \dots, N_n\}$. To put it differently, \hat{P}_n satisfies the following relationship:

$$(22) \quad \max_{A \in \mathcal{F}_{N_n}} |\mu_n(A) - \hat{P}_n(A)| = \min_{i=1, \dots, N_n} \max_{A \in \mathcal{F}_{N_n}} |\mu_n(A) - P_i(A)|.$$

If the $N(a)$ is the most economical number of the balls as described in the lemma, then the function $\log_2 N(a)$ is called Kolmogorov's *entropy* of the space (\mathcal{P}, d) . At this point, recall that the notation $x_n \sim y_n$ signifies that: $x_n = O(y_n)$ and $y_n = O(x_n)$.

We may now formulate the main result of this paper.

(23) **Theorem** With the metric d defined by (19), suppose that the metric space (\mathcal{P}, d) is totally bounded, and that for each $P \in \mathcal{P}$, the process $\{X_n : n \geq 1\}$ is ϕ -mixing with mixing coefficient $\phi(n)$ as in Proposition 7. Furthermore, suppose that the entropy of the space (\mathcal{P}, d) is such that the relation $a_n \sim (\log_2 N_n/n)^{1/2}$ holds with $0 < a_n \rightarrow 0$. Then, the minimum distance estimate \hat{P}_n of P , defined by (22), is uniformly weakly consistent at the rate a_n , as in the i.i.d. case. That is to say, for every $\epsilon > 0$, there exists $b(\epsilon) > 0$ such that $\sup_{P \in \mathcal{P}} P\{\|\hat{P}_n - P\| \geq b(\epsilon)a_n\} < \epsilon$ for $n \geq 1$. In

addition,

$$\|\hat{P}_n - P\| \leq C \{a_n + (\log_2 N_n/n)^{1/2} + (\phi(p_n)/p_n)^{1/2}\}$$

in probability, where C is a positive constant and $\{p_n\}$ is the sequence defined in (32) below.

Proof. The approach used in Yatracos (1985) applies here. By means of the triangular inequality for d , as defined in (19), and relations (21) and (22), one arrives at the following relation, after some manipulations:

$$(24) \quad \|\hat{P}_n - P\| \leq 5a_n + 4 \max\{|\mu_n(A) - P(A)|; A \in \mathcal{F}_{N_n}\}.$$

For an arbitrary $A \in \mathcal{A}$, put $Y_i = I_A(X_i)$ for $i = 1, \dots, n$, and

$$(25) \quad \bar{S}_n(A) = \frac{1}{n} \sum_{i=1}^n (Y_i - \mathbb{E}Y_i) = \mu_n(A) - P(A).$$

Applying (18) to $\bar{S}_n(A)$ in (25), we obtain, for all n :

$$P(|\bar{S}_n(A)| \geq \epsilon_n) \leq 6(1 + 2\sqrt{e}\phi(p))^r \exp\left(-\frac{n\epsilon_n^2}{2C_0}\right) \quad \text{for } \epsilon_n \leq \frac{C_0 r}{n}.$$

By the fact that $|\mathcal{F}_{N_n}| \leq N_n^2$, one then has, for all n ,

$$(26) \quad \begin{aligned} P(\max\{|\mu_n(A) - P(A)|; A \in \mathcal{F}_{N_n}\} \geq \epsilon_n) \\ \leq 6N_n^2 (1 + 2\sqrt{e}\phi(p))^r \exp\left(-\frac{n\epsilon_n^2}{2C_0}\right) \quad \text{for } \epsilon_n \leq \frac{C_0 r}{n}, \end{aligned}$$

since here M may be taken to be equal to 1. Inequalities (24) and (26) yield, for all n and with $\epsilon_n \leq C_0 r/n$,

$$P(\|\hat{P}_n - P\| \geq \epsilon_n) \leq 6N_n^2 (1 + 2\sqrt{e}\phi(p))^r \exp\left(-\frac{n}{32C_0}(\epsilon_n - 5a_n)^2\right),$$

for all n and $5a_n < \epsilon_n \leq C_0 r/n$. Furthermore, since $C_0/3p < C_0 r/n$, it suffices to require that: $5a_n < \epsilon_n \leq C_0/3p$. Combining these observations, we have then:

(27)

$$P(\|\hat{P}_n - P\| \geq \epsilon_n) \leq 6N_n^2 (1 + 2\sqrt{e}\phi(p))^{n/2p} \exp\left(-\frac{n}{32C_0}(\epsilon_n - 5a_n)^2\right),$$

for all n and $5a_n < \epsilon_n \leq C_0/3p$. By the fact that the expression on the right-hand side of (27) is independent of $P \in \mathcal{P}$, it suffices to show that this expression tends to 0. For simplicity, set $2\sqrt{e} = C_1$, $(1/32C_0) = C_2$, and use similar notation for subsequent constants. From (27), we have then to determine ϵ_n to satisfy (28), subject to restriction (29):

$$(28) \quad C_2 n(\epsilon_n - 5a_n)^2 - 2 \log N_n - \frac{n}{2p} \log(1 + C_1 \phi(p)) \rightarrow \infty,$$

$$(29) \quad 5a_n < \epsilon_n \leq C_0/3p.$$

By taking $C_2 n(\epsilon_n - 5a_n)^2 = 3 \log N_n + (n/p) \log(1 + C_1 \phi(p))$, or after solving for ϵ_n and renaming constants,

$$(30) \quad \epsilon_n = 5a_n + \left(C_3 \frac{\log N_n}{n} + C_4 \frac{1}{p} \log(1 + C_1 \phi(p)) \right)^{1/2},$$

the convergence in (28) is satisfied, and so is the left-hand side inequality in (29). By means of various inequalities and with a view of simplifying the choice of ϵ_n , we are led to choosing ϵ_n as follows:

$$(31) \quad \epsilon_n = 5a_n + C_5 (\log N_n/n)^{1/2} + C_6/p.$$

Relation (31) suggests choosing

$$(32) \quad p = C_7(n/\log N_n)^{1/2}$$

with a suitable choice of a positive integer C_7 , which then gives

$$(33) \quad \epsilon_n = C_8 (\log N_n/n)^{1/2}.$$

This choice of ϵ_n satisfies the convergence in (28) and both inequalities in (29). Furthermore, it is consistent with its prior choice as shown in (31), provided $a_n \sim (\log_2 N_n/n)^{1/2}$, which is part of the assumptions in the theorem. For the second conclusion of the theorem, observe that relations (24) and (26) imply that

$$(34) \quad \|\hat{P}_n - P\| \leq 5a_n + 4\epsilon_n \quad \text{with probability tending to 1.}$$

Relation (30) and elementary inequalities yield

$$(35) \quad \epsilon_n \leq 5a_n + C_9 (\log_2 N_n/n)^{1/2} + C_{10} (\phi(p)/p)^{1/2}.$$

Relations (34) and (35) complete the proof. \square

(36) **Corollary** Suppose the conditions of Theorem 23 are satisfied, let $\phi(n) = cn^{-(1+\delta)}$ some $c > 0$ constant and $\delta > 0$, and assume further that, for every n , X_n takes values in $[0, 1]^d$ and the collection of densities Θ corresponding to \mathcal{P} are the q -“smooth” densities defined as in Yatracos (1989, page 1601). Then the rate of convergence is $a_n \sim n^{-q/(2q+d)}$ and the convergence holds almost surely.

Proof. By $N_n \sim 2^{(1/a_n)^{d/q}}$, and the assumption $a_n \sim (\log_2 N_n/n)^{1/2}$ it follows that the rate of convergence is given by $a_n \sim n^{-q/(2q+d)}$. This rate is the same as that in the i.i.d. case. Almost sure convergence follows by the Borel-Cantelli lemma. \square

Acknowledgments: The research of Yatracos was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

22.4 REFERENCES

- Birgé, L. (1983), ‘Approximation dans les espaces métriques et théorie de l'estimation’, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **65**, 181–237.
- Le Cam, L. (1960), ‘Locally asymptotically normal families of distributions’, *University of California Publications in Statistics* **3**, 37–98.
- Le Cam, L. (1969), *Théorie Asymptotique de la Décision Statistique*, Les Presses de l’ Université de Montréal.
- Le Cam, L. (1973), ‘Convergence of estimates under dimensionality restrictions’, *Annals of Statistics* **1**, 38–53.
- Roussas, G. & Ioannides, D. (1988), Probability bounds for sums in triangular arrays of random variables under mixing conditions, in K. Matusita, ed., ‘Statistical Theory and Data Analysis II’, Elsevier Science Publishers B.V. (North-Holland), pp. 293–308.
- Roussas, G. G. & Ioannides, D. (1987), ‘Moment inequalities for mixing sequences of random variables’, *Stochastic Analysis and Applications* **5**(1), 61–120.
- Yatracos, Y. G. (1985), ‘Rates of convergence of minimum distance estimators and Kolmogorov’s entropy’, *Annals of Statistics* **13**, 768–774.
- Yatracos, Y. G. (1989), ‘A regression type problem’, *Annals of Statistics* **17**, 1597–1607.
- Yatracos, Y. G. (1992), ‘ L_1 -optimal estimates for a regression type function in R^d ’, *Journal of Multivariate Analysis* **40**, 213–220.

23

Daniel Bernoulli, Leonhard Euler, and Maximum Likelihood

Stephen M. Stigler¹

23.1 Introduction

The history of statistical concepts usually hinges on subtle questions of definition, on what one sees as a crucial element in the concept. Is the simple statement of a goal crucial? Or do we require the investigation of the implications of pursuing that goal, perhaps including the discovery of anomalies that require specification of conditions under which a claimed property holds? Or the detailed successful exploration of those conditions? Such considerations certainly arise in the case of the method of maximum likelihood. If the object of study is the modern theory of maximum likelihood, of its efficiency in large samples in a parametric setting, then an argument could be made for beginning with Edgeworth (1908–1909) (see Pratt (1976)), or Fisher (1912 or 1922 or 1935) (see Edwards, 1974), or even Wald (1949) or Le Cam (1953). It might be thought that the question would be easy to resolve if instead of worrying about mathematical rigor and the deeper questions of inference, including the interpretation of statistical information, we only asked about the introduction of the idea of choosing, as an estimate, that value which maximizes the likelihood function, but that is not the case. Even at that level difficulties of interpretation arise. Was Gauss employing maximum likelihood in 1809 when he arrived at the method of least squares, or, as some of his development would lead you to believe, was he maximizing a posterior density with a uniform prior?

One of the earliest works that has been cited as the first to formulate the method of maximum likelihood is a short paper by Daniel Bernoulli, published in 1778 in St. Petersburg. Bernoulli's paper is not the only early claimant; Sheynin (1971) discusses a section of a 1760 book by Johann Lambert that certainly predates Bernoulli and can arguably be interpreted as presenting the method. But Bernoulli's paper has been accorded the lion's share of attention, being clearly expressed, accessible, and, by virtue

¹University of Chicago

of being widely known, more influential on contemporaries. Nonetheless, despite Bernoulli's characteristic clarity of expression and a large amount of subsequent historical discussion, there remains an element of uncertainty or confusion about the paper. It is the goal of this present paper to try to dispel this confusion and shed light on the understanding of the topic by Bernoulli and his contemporary Leonhard Euler, by reexamining his paper in the light of an unpublished manuscript treatment of the topic. The manuscript, by Daniel Bernoulli and plausibly dated to 1769, is presented in translation.

23.2 Bernoulli's 1778 Paper

Bernoulli's published paper consists of 20 sections spread over 30 pages. The emphasis of this somewhat discursive paper is on what might be called mathematical philosophy, on the nature of astronomical error and how it might best be handled by mathematical methods—if it is amenable to those methods at all. Formal analysis is not introduced until section 11, and most of sections 15–20 is given to examples and further philosophical discussion. Yet historical commentary has focused on the analysis of sections 11–14, the portion that has the closest apparent brush with modern ideas. See Todhunter (1865); Pearson (1978); Sheynin (1972a,b); and Edwards (1974).

In brief outline, Bernoulli's argument runs like this: The common practice of taking the arithmetic mean of a set of observations cannot be correct, for it weights all observations equally, while they are surely not all of equal validity. Indeed, he noted, astronomers seem to acknowledge this by discarding extreme observations—those too far from the rest to be considered as plausible measurements—before taking the mean. Bernoulli did not claim that the probabilities of errors of various magnitudes could be specified precisely, but he felt that some of their characteristics could be stated qualitatively. In this, he seems to follow Laplace's 1774 practice (Stigler, 1986b) of starting with a list of properties an error curve should have, but he cited neither Laplace nor any other writer.

Bernoulli regarded it as axiomatic that errors above and below the true point may be taken as equally possible, so the scale of the probabilities of the errors will be symmetrical about that point. Furthermore, observations nearer the true point will be more probable than those more distant, and errors beyond some limit of maximum error will not be possible. This of course does not specify the scale of probabilities exactly, but Bernoulli suggested that it could be approximated by a semi-ellipse, or even a semi-circle. Without claiming more than that the results should improve over those derived by taking a simple mean, he then proceeded to analyze the problem assuming the curve was a semi-circle of radius r .

Bernoulli would have the astronomer determine the radius r of the semi-circle subjectively, as the greatest error he would never exceed. Then if x

represents the distance of the true point from the smallest observation A , and if the observations are recorded as $A, A+a, A+b, A+c, \dots$, Bernoulli proposed to select as the correction x to be applied to the observation A , in order to best determine the true point, the value that maximizes the probability of the entire complex of observations, namely

$$\sqrt{(r^2 - x^2)} \times \sqrt{(r^2 - (x-a)^2)} \times \sqrt{(r^2 - (x-b)^2)} \times \sqrt{(r^2 - (x-c)^2)} \times \text{etc.}$$

He noted that this was equivalent to maximizing

$$(r^2 - x^2) \times (r^2 - (x-a)^2) \times (r^2 - (x-b)^2) \times (r^2 - (x-c)^2) \times \text{etc.}$$

To accomplish this, he took the derivative, set it equal to zero, and solved for x . The true point was then taken as $A+x$.

Naturally, not all went as smoothly as this outline might imply. Bernoulli was not only a pioneer in asking for the maximum likelihood estimate, he was a pioneer in discovering problems in actually finding the estimate. For the case of a single observation, all was well: Then the procedure returned the observation itself as the best choice for the true point. But already with only two observations, difficulties appeared—what we now call the likelihood equation had, in that case, not one but three roots, three prospective solutions. In that case Bernoulli simply chose the root in the middle, stating that it was the only “useful root.” For three observations, the situation was even worse—the equation to be solved was of the fifth degree, and hence could admit as many as five roots. Again, without explanation, Bernoulli would select as the estimate the only “useful root,” and he gave a rational approximation for that root. He noted that his solution differed from the common mean. He did not apologize for the complicated equations, although he admitted that a very long calculation was involved, and he described the equation for three observations as “enormous” and “monstrous.” Rather, he thought the implications “metaphysical rather than mathematical,” and he said the calculations suggested that observations should seldom if ever be discarded, and then only after careful attention. He did note that several other error curves, such as the parabola

$$\frac{\rho}{r^2} (r^2 - x^2),$$

would give exactly the same results.

23.3 Euler as Discussant

Bernoulli had sent his paper to St. Petersburg for publication, and he would presumably not have been surprised that his friend, former colleague, and fellow son of Basel, Leonhard Euler appended his own commentary to the published version. Indeed, when Bernoulli found out before publication that

Euler had done so, he was at least initially quite happy about it (Sheynin, 1972a, p. 51). Euler, Bernoulli's junior by seven years, was perhaps the greatest analyst of that century and the most prolific mathematician of all time. He was not silent on any mathematical subject, and frequently commented in print on other mathematicians' work. The two had been together briefly at St. Petersburg in the early 1730s and had maintained contact since. Their long acquaintanceship did not keep Euler from being critical of Bernoulli's approach, although his criticism was politely expressed.

Euler recapitulated Bernoulli's approach in a few short paragraphs, omitting the square root and going directly to $r^2 - (x - a)^2$ as the "degree of goodness" of an observation at $\Pi + a$ when the true value is $\Pi + x$. Euler saw no supporting argument for Bernoulli's procedure of multiplying these expressions and seeking a maximum; indeed, he presented a clever mathematical objection. What, Euler asked, if one of the observations differed from x by the amount r (or an amount just short of r). Then $r^2 - (x - a)^2$ would be zero for that observation, and the product of terms would similarly be reduced to zero—a value that could not possibly be considered a maximum. Yet that, said Euler, runs counter to the principles of the theory of probability; an observation of no perceived value when introduced into an analysis should have no effect at all, but with Bernoulli's criterion, it destroys the worth of the whole complex!

As an alternative, Euler suggested considering the estimate $\Pi + x$ based on observations at $\Pi + a$, $\Pi + b$, $\Pi + c$, $\Pi + d$, etc. as given by a weighted average,

$$x = \frac{a\alpha + b\beta + c\gamma + d\delta + \text{etc.}}{\alpha + \beta + \gamma + \delta + \text{etc.}},$$

where the weights were given by $\alpha = r^2 - (x - a)^2$, etc. Euler explained how to solve this for x , gave a continued fraction expression for the solution, and went on to explain that it could be viewed as the solution to a maximum problem different from Bernoulli's, namely, choose x to maximize

$$\{r^2 - (x - a)^2\}^2 + \{r^2 - (x - b)^2\}^2 + \{r^2 - (x - c)^2\}^2 + \text{etc.}$$

A modern statistician would recognize Euler's suggestion as mathematically equivalent to a form of M-estimator; we will discuss this connection more fully later.

Historical discussion of Euler's commentary has not been kind. Todhunter (1865, p. 238) wrote "Euler seems to have objected to the wrong part of Daniel Bernoulli's method; the particular law of probability is really the arbitrary part, the principle of making the product of the probabilities a maximum is suggested by the Theory of Probability." Sheynin (1972a,b) stated bluntly that Euler "misunderstood" Bernoulli, and characterized his reasoning as erroneous. Pearson (1978, pp. 268-9) described Euler's commentary as "very wide of the point," and stated "Both Bernoulli and Euler

seem to me to reach false conclusions by starting from arbitrary assumptions, but Euler more completely so than Bernoulli." What was it that led Euler to take an approach that would elicit such a reaction? Could he really have been ignorant of the fundamental rule for multiplying probabilities of independent events? I will argue to the contrary, but first we need to examine the history of Bernoulli's paper a bit more closely.

23.4 Bernoulli's 1769 Manuscript

Bernoulli's 1778 paper was not his first work under that title. It has been known for some time that he had circulated a short paper on this topic several years earlier. In 1772 his nephew Jean Bernoulli III made reference in a footnote published in an astronomy journal to "a small memoir" of Daniel's on this problem "that is not yet printed." Laplace, in his 1774 paper (Stigler, 1986b) refers to this footnote, adding that he had not seen Daniel's memoir. (Both Jean III's footnote and Laplace's statement are reprinted in Stigler (1978).) In an encyclopedia article not published until 1785 (but presumably written in the early or middle 1770s), Jean III went much further. He referred to Daniel's "short paper" as having been sent to him in 1769, and he gave an outline of the approach it presented (Bernoulli, 1785). The 1785 article has formed the basis for subsequent comment by Todhunter (1865, pp. 442–3), who noted the difference between the two versions, and Sheynin (1972a). It is not widely known that the 1769 manuscript survives today in the archives at the University of Basel, although its existence has been previously noted (Stigler, 1986a, p. 110).

Bernoulli's 1769 Latin manuscript is superficially similar to the 1778 version: They have the same title, the same first sentence, some of the same language; the semi-circle even makes its appearance in the manuscript. But the use of the semi-circle, and the procedures Bernoulli derived from it, are strikingly different. In 1769 the semi-circle centered at the true value was variously described as giving the value of an observation, the frequency of the observations, or the probability of an observation, but it was employed as a weight function! In much the same way that Euler was to use $r^2 - (x-a)^2$ (the square of Bernoulli's $\sqrt{r^2 - (x-a)^2}$), Bernoulli would choose as his estimate that value which agreed with the weighted average of the observations, weighted by his semi-circle weight function centered at the estimate. He would calculate the estimate iteratively. He would start with the mean, position the semi-circle at the mean, determine the weights to be applied to the observations, find the weighted average, move the semi-circle to the new position; then repeat the process until the weighted average ceased to change by a significant amount. Figure 1 shows Bernoulli's own drawing of the semi-circle at the first two iterations.

Bernoulli's procedure is mathematically exactly that for calculating the

robust M-estimate by iterative reweighting, using the weight function $\sqrt{r^2 - u^2}$. (Thisted, 1986, pp. 141, 151.) Conceptually, Bernoulli could not be said to be engaged in robust estimation—his paper shows not so much as a whisper of the modern concept of robustness. But mathematically, his estimate differs from modern procedures only in the informal way he specifies his scale factor, r . Euler's procedure is equivalent, except for the use of the weight function $r^2 - u^2$ and Euler's solution by continued fractions rather than iteration. A popular modern cousin of this would be the corresponding version of Tukey's biweight weight function, $(r^2 - u^2)^2$. It may seem ironic that in 1769 Bernoulli, based on considerations superficially similar to those of his 1778 paper, propounded an estimate that can be computed rather easily for any number of observations, while in 1778, based on the same semi-circle, he derives an estimate that is all but beyond computation even for a mere three observations. Clearly his reconsideration of the question was not inspired by computational expediency. And, since he showed no new qualms in his choice of the semi-circle in 1778, that element, taken as the most questionable portion of the work by some commentators, was not for him an issue. What, then, did lead him to revise the paper in such a fundamental way?

23.5 The 1778 Papers Reconsidered in Light of the Manuscript

There are two substantial differences between Bernoulli's treatments of 1769 and 1778. The first involves a change in the conceptual interpretation of a vaguely, even ambiguously defined quantity; the second a step forward in Bernoulli's understanding of the use of probability with observational evidence. The first concerned the interpretation of the curve of errors, the second the combination of curves.

Mathematicians are intellectually greedy. When an extra property can be obtained for no additional cost, they seize it. When a conclusion stronger than that required can be had for free, they grab for it. Some are insatiable; all are at least opportunistic. Unfortunately, not all that appears costless actually is. In mathematics as in economic life, there may be no free lunch, even if the price of lunch is only caution in interpretation. Bernoulli's greed was natural enough; it involved the way he took advantage of the unusual opportunity his problem presented for a confusion of two very different properties. These were, (a) that larger errors of observation could reasonably be supposed to be less frequent than smaller, and so a curve that describes the frequency or probability of observations may be reasonably supposed to decrease from a maximum at the true point. And (b) observations that are less accurate—are further from the true point—are hence less valuable for the purposes of estimation than those closer to the truth.

Bernoulli yielded to the obvious temptation: He identified worth or value with frequency or probability, and let the same curve carry the burden of simultaneously describing both. And yet the two properties, while happily coincident as far as the qualitative shape of the curve goes, bear no necessary relation to one another—they turn out to be actually antagonistic in Bernoulli's problem. As we now know, with Bernoulli's choice of a bounded error density, extreme observations may miss the mark by more, but they have greater—not less—*inferential* value than those near the center!

The signs of this confusion are evident in 1769 and had not disappeared in 1778. In 1769 he several times referred to “the value or probability” of an observation, as if these are one and the same. In 1778 the emphasis is more on the curve as a curve of probability, but there it is clear throughout that Bernoulli believed the curve decreases from its center because the more discrepant observations correspond to a diminished level of skill; that is, that extreme observations are both less frequent and less valuable. The notion that more extreme observations could be *more* valuable would have astonished Bernoulli equally in 1769 and 1778, even if it were pointed out to him that there are traces of this behavior in his 1778 analysis.

If Bernoulli's ambivalence about the interpretations of the semi-circle changed but little over the decade, what was the change in his thinking that led him to derive such a radically different procedure in 1778? The key was that in 1769 he considered the curve only as describing the characteristics of the observations separately, as individuals, and this encouraged him to emphasize the “value” side, treating the curve as a weighting function. But by 1778 he had come to be concerned with achieving “the highest degree of probability for the *complex of observations as a whole*.” [italics added] He no longer sought to treat the individual observations first by weighting them according to the “value” as determined from his curve and only then aggregating them in the traditional manner by taking the mean of the observations so weighted. Instead he wanted to optimize the probability of the aggregate of observations! The first of these had led him to the estimate which agreed with the expectation based on the probability-value weights, the second to the maximum likelihood estimate. By thinking of the “complex of observations” he came to multiply the probabilities and maximize. While focusing on probabilities for single observations separately he had thought only of combining them through what he called the law of probability—an expectation, or probability-weighted mean.

Bernoulli had thus in 1778 come to an important realization, one crucial to the development of theoretical statistics, the realization that one should treat the data as a set and not simply as individuals to be weighted, discarded, combined. He went further and explored one of the consequences of this approach, when he asked for three observations what relationship his maximum likelihood estimate bore to the simple arithmetic mean (the “common rule”). He gave the approximation to the “useful root” for the

correction x to be used for three observations at A , $A+a$, and $A+b$, to be

$$x = \frac{a+b}{3} + \frac{2a^3 - 3a^2b - 3ab^2 + 2b^3}{27r^2}.$$

The first term gave the correction that would correspond to the arithmetic mean, and the second permitted Bernoulli to make this remark: "The common rule for three observations gives somewhat too small a result when $a < \frac{1}{2}b$ and too large a result if $a > \frac{1}{2}b$, and cannot ever be applied with greater certainty than when the intermediate observation is approximately equidistant from the two extremes." Had Bernoulli developed this line further, he might have realized the consequence, that the condition $a < \frac{1}{2}b$ corresponds to b being an extremely large observation relative to the first two, and if the arithmetic mean was too small in that situation, it was because the maximum likelihood estimate was pulled closer to the extreme than was the simple mean. That is, the extreme was more influential than either of the other observations, and hence its inferential value was larger, not less, by its extremity! But his concern was only to demonstrate the difference between his procedure and the common one, not to analyze the weighting of observations implicit in the maximum likelihood procedure. The latter, more subtle, question seems not to have been raised until the twentieth century.

What led to Bernoulli's change of perspective between 1769 and 1778? With no direct evidence we can only speculate. One possibility is that he had encountered Lambert's 1760 work in the meantime. In support of this it could be argued that, first, he had sent the 1769 manuscript to his nephew Jean III who was at that time in Berlin, as was Lambert, and that Jean III had even made specific reference to Lambert's work on taking a mean in the 1772 footnote he published in his astronomical journal. On the other hand, Lambert's work was well buried in a 1760 book that Bernoulli had evidently not seen by 1769, and Lambert's own approach was via a multinomial distribution of errors—very different from Bernoulli's analytical approach. Furthermore, Bernoulli in 1778 specifically stated that his estimate "is deduced from principles never used before," an explicit priority claim that would have been highly unlikely had he been made aware of Lambert by his nephew. Even with the lax citation standards of the time, to make such a claim when knowledgeable people were aware it was untrue would be a flagrant transgression of the norms of scientific discourse. I tend to believe that either Bernoulli had not seen Lambert, or he had seen it but thought it fundamentally different and not relevant to his work.

Another, to me more likely, potential inspiration to Bernoulli's 1778 reconceptualization was Laplace's 1774 paper on the application of inverse probability to estimation, a paper that had been prominently published in one of the most prestigious and widely circulated scientific journals of the age, the French Académie des Sciences' *Mémoires présentés par divers*

Savans. Bernoulli was a Foreign Associate of the Académie and would presumably have seen the volume. Laplace did not discuss maximum likelihood there, so Bernoulli could have made his priority claim in good conscience with Laplace in mind. But Laplace most definitely did treat the complex of all observations as a whole, multiplying the densities together just as Bernoulli had. Laplace, like Bernoulli in 1778, was only able to handle three observations, although his procedure, finding the posterior median with uniform prior distributions on both scale and location parameters, was quite different from Bernoulli's (Stigler, 1986b).

23.6 Euler, Reconsidered

The 1769 manuscript also sheds interesting light on Euler's 1778 discussion. Now, Kendall (1961) in his commentary on Euler made excuses for him, stating that perhaps he misunderstood Bernoulli because he was 71 years old (but he was seven years Bernoulli's junior), or because he was partially blind (but in mathematics he still saw more clearly than anyone else). The manuscript suggests another possibility, namely that Euler was, in fact, right.

In examining Euler's discussion, several points should be born in mind. First, despite Bernoulli's change of emphasis on the interpretation of the semi-circle, the prose of 1778 still supported interpreting it as reflecting a narrow conception of the value of an observation, where decreasing value simply meant decreasing accuracy. Indeed, Bernoulli opened his discussion in both 1769 and 1778 with the rationale that the "common rule" of taking the arithmetic mean cannot be generally best because it treats all observations equally, with the same weight. We know now that treating the observations equally can be consistent with their being unequally probable—Gauss showed us that for the normal distribution in 1809, in showing how the mean could be derived from probability, even maximum likelihood, considerations. Having unequal probabilities does not, as Bernoulli can be read as claiming, rule out the arithmetic mean. Euler in fact seized upon the "value" interpretation that Bernoulli, even in 1778, emphasized in his rationale, and he then reasoned entirely in its terms. He wrote of "degree of goodness," not probability, and his development of a heuristically weighted mean within that context is entirely reasonable.

Second, even accepting the interpretation of the curve as a probability curve, it did not fully model the data-generating mechanism. Bernoulli was explicit in recognizing the possibility of observations beyond the limits of the curve, but vague about how to decide which fit this description and should be summarily rejected. Euler's clever example showed how an outlier could make a joker of Bernoulli's likelihood function. Bernoulli's maximum likelihood procedure is thus incompletely specified. It requires both the subjective determination of r , and the subjective discarding of

observations considered impossibly extreme. Euler in effect chided Bernoulli on this, by showing how a single observation at the boundary of feasibility could destroy the procedure.

And third, Bernoulli's principle of choosing the value that maximizes the probability of the data does not in fact clearly provide a solution in principle to the estimation problem. The extensive literature (including many debates) on the topic of likelihood in the present century testifies to the elusiveness of the concept when considered outside the frame of Bayesian analysis, and both Bernoulli and Euler were writing in an era when only Laplace would be placed within that frame. Within the probability framework of Bernoulli's time, maximum likelihood is not *prima facie* an optimal procedure; it is not, in that historical context, the solution of an optimization problem. As was recognized by Fisher, in the frequency approach to inference it requires proof that the maximizing of the likelihood function will yield the most accurate estimate. The provision of such a proof has been a major research agenda, and it can only be said to be satisfactorily accomplished in terms of asymptotic theory in regular cases, indeed in cases that would exclude the semi-circle of Bernoulli! In this sense Euler was absolutely correct when he stated that "the distinguished author has not supported this principle of the maximum by any proof." And by showing that without some care the function being maximized would be destroyed by a well placed observation, and offering in its place an alternative estimate, Euler was acting in the grand tradition of mathematical statistics.

Could Euler have seen the earlier manuscript? He was in St. Petersburg from 1766 through 1783, but he had been in Berlin (where Jean Bernoulli III had seen it in 1769) before that, and he kept in frequent correspondence with the mathematicians of Europe. The fact that the procedure he suggested was a close cousin to one Bernoulli had suggested in 1769, that his analysis is quite different from that Bernoulli had employed in the earlier work, and that it was presented as if new, suggests that Euler had not seen the earlier work. Great minds, so it is said, think alike.

23.7 Conclusion

Taken together, the two versions of Bernoulli's paper and Euler's commentary provide a revealing glimpse of the development of mathematical statistics in the late eighteenth century. We see Bernoulli struggling with one of the fundamental problems of statistics, the combination of observations to form an estimate. Using a combination of vague notions of weight, value, and probability, without distinguishing among these adequately to avoid confusion, he first emphasized weight/value and concocted one estimate in an ingenious fashion, producing what we might now recognize as an early version of a mathematical algorithm used to compute M-estimates. After reconsideration, for reasons that can only be guessed at, he attacked

the problem anew and, emphasizing probability this time, came up with an early version of what we now call a maximum likelihood estimate. Bernoulli never fully resolved the tension between the two interpretations—value and probability—that he wrestled with. Euler's commentary recognized inadequacies in Bernoulli's second paper more perceptively than did many later readers, who saw there only an early version of a successful modern concept. Euler returned to the emphasis on value and, presumably independently of Bernoulli, suggested a procedure very much like Bernoulli's first solution.

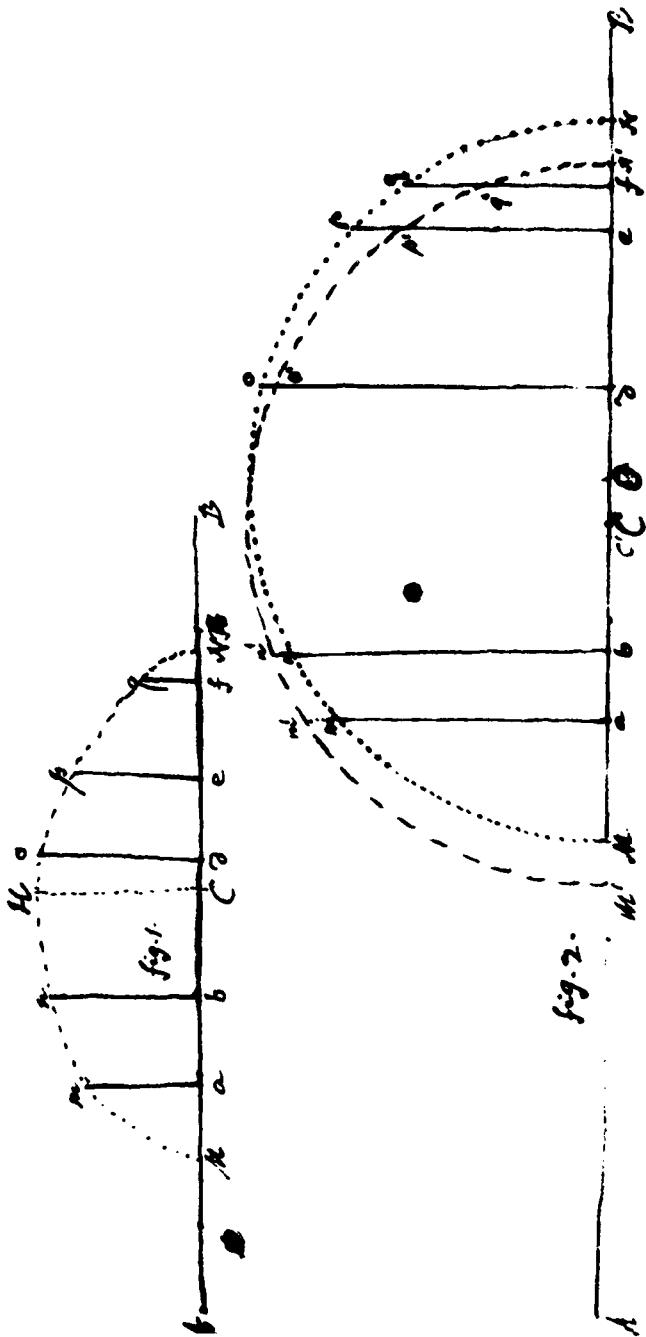
The dialogue here, both within and between great mathematical scientists, helps us appreciate the difficulty in formulating new concepts. Perhaps maximum likelihood is not a concept after all, but a theme whose early notes may have been struck in the last half of the eighteenth century, but whose fuller development came only 200 years later. Many modern variations on this theme, some rhapsodic, others discordant, have enriched our statistical symphony and gone far, far beyond the simple early structure, but all can still be recognized as echoes of the early notes of Bernoulli and Euler.

23.8 REFERENCES

- Bernoulli, D. (1769), 'Dijudicatio maxime probabilis plurium observationum discrepantium atque verisimillima inductio inde formanda', Manuscript; Bernoulli MSS f.299-305, University of Basel.
- Bernoulli, D. (1778), 'Dijudicatio maxime probabilis plurium observationum discrepantium atque verisimillima inductio inde formanda', *Acta Academiæ Scientiarum Imperialis Petropolitanæ* for 1777, pars prior pp. 3–23. Reprinted in Bernoulli (1982). English translation in Kendall (1961) pp. 3–13, Translation reprinted in Pearson and Kendall (1970) pp. 157–167.
- Bernoulli, D. (1982), *Die Werke von Daniel Bernoulli: Band 2, Analysis, Wahrscheinlichkeitsrechnung.*, Birkhäuser, Basel.
- Bernoulli, J. III. (1785), 'Milieu', *Encyclopédie Méthodique; Mathématiques* 2, 404–409. Figs. 2 and 3, Plate 1 of 'Géométrie' plates.
- Edgeworth, F. Y. (1908–09), 'On the probable errors of frequency constants', *Journal of the Royal Statistical Society* 71, 381–397, 499–512, 651–678; 72, 81–90.
- Edwards, A. (1974), 'The history of likelihood', *International Statistical Review* 42, 9–15.
- Euler, L. (1778), 'Observationes in præcedentem dissertationem illustr. Bernoulli', *Acta Academiæ Scientiarum Imperialis Petropolitanæ* for 1777, pars prior, pages 24–33. Reprinted in Euler's *Opera Omnia*,

- Ser. 1, 7, 280–290. English translation in Kendall (1961), pages 13–18. Translation reprinted in Pearson and Kendall (1970) pages 167–172.
- Fisher, R. A. (1912), ‘On an absolute criterion for fitting frequency curves’, *Messenger of Mathematics* 41, 155–160.
- Fisher, R. A. (1922), ‘On the mathematical foundations of theoretical statistics’, *Philosophical Transactions of the Royal Society of London, Series A* 222, 309–368.
- Fisher, R. A. (1935), ‘The logic of inductive inference’, *Journal of the Royal Statistical Society* 98, 39–54.
- Gauss, C. F. (1809), *Theoria Motus Corporum Celestium*, Perthes et Besser, Hamburg. Translated, 1857, as *Theory of Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*, trans. C. H. Davis. Boston, Little, Brown. Reprinted, 1963; New York, Dover.
- Kendall, M. G. (1961), ‘Daniel Bernoulli on maximum likelihood’, *Biometrika* 48, 1–18. Reprinted in Pearson and Kendall (1970), pages 155–172.
- Kendall, M. G. & Plackett, R. L. (1977), *Studies in the History of Statistics and Probability*, Vol. II, Griffin, London.
- Le Cam, L. (1953), ‘On some asymptotic properties of maximum likelihood estimates and related Bayes estimates’, *University of California Publications in Statistics* 1, 277–330.
- Pearson, E. S. & Kendall, M. G. (1970), *Studies in the History of Statistics and Probability*, Griffin, London.
- Pearson, K. (1978), *The History of Statistics in the 17th and 18th Centuries, Against the Changing Background of Intellectual, Scientific and Religious Thought*, Griffin, London. Lectures from 1921–1933, Ed. by E. S. Pearson.
- Pratt, J. W. (1976), ‘F. Y. Edgeworth and R. A. Fisher on the efficiency of maximum likelihood estimation’, *Annals of Statistics* 4, 501–514.
- Sheynin, O. B. (1971), ‘J. H. Lambert’s work on probability’, *Archive for History of Exact Sciences* 7, 244–256.
- Sheynin, O. B. (1972a), ‘On the mathematical treatment of observations by L. Euler’, *Archive for History of Exact Sciences* 9, 45–56.
- Sheynin, O. B. (1972b), ‘Daniel Bernoulli’s work on probability’, *RETE* 1, 273–300. Reprinted in Kendall and Plackett (1977), pages 104–300.

- Stigler, S. M. (1978), 'Laplace's early work: Chronology and citations', *Isis* **69**, 234–254.
- Stigler, S. M. (1980), 'R. H. Smith, a Victorian interested in robustness', *Biometrika* **67**, 217–221.
- Stigler, S. M. (1986a), *The History of Statistics: The Measurement of Uncertainty Before 1900*, Harvard University Press, Cambridge, Massachusetts.
- Stigler, S. M. (1986b), 'Laplace's 1774 memoir on inverse probability', *Statistical Science* **1**, 359–378.
- Thisted, R. A. (1986), *Elements of Statistical Computing: Numerical Computation*, Chapman Hall, London.
- Todhunter, I. (1865), *A History of the Mathematical Theory of Probability From the Time of Pascal to That of Laplace*, Macmillan, London. Reprinted 1949, 1965, New York, Chelsea.
- Wald, A. (1949), 'Note on the consistency of the maximum likelihood estimate', *Annals of Mathematical Statistics* **20**, 595–601.



Daniel Bernoulli's original drawing, from the 1769 manuscript. The first portion (referred to as "fig. 1" in Bernoulli's paper) shows his semi-circle, the second ("fig. 2") illustrates the first two steps in his iteration.

**The Most Probable Choice Between Several Discrepant
Observations and the Formation Therefrom of the Most Likely
Induction [1769]**

[Dijudicatio maxime Probabilis Plurium observationum discrepantium
atque

Verisimillima inductio inde formanda]

By Daniel Bernoulli

[Bernoulli MSS f. 299–305, Translated by Seth Lerer, Michael McGrade,
and Stephen Stigler]

1. It is to the Astronomers, the most scrupulously sagacious of men, that I offer these doubts which have, at times, occurred to me about the universally accepted rule for handling several discrepant observations of the same event. By this rule, all observations are combined into one sum, which afterwards is divided by the number of observations. The result of this division is accepted as the true, sought quantity, until another better and more accurate one is deduced. In this way, if the observations are considered as having, in effect, the same weight, then the center of gravity is chosen as the single point that would be in agreement with the most careful of all possible observations. This rule agrees with that used in the theory of probability, when all deviations from the truth are considered equally likely: thus indeed, if these same observations are considered to have the stipulated equal probabilities, then, according to the second rule of probability, the estimated value will be given by the sum of the observations divided by their number.

2. But is it right to consider the several observations to be of the same weight or importance, or equally prone to any and every error? Are errors of a certain number of minutes as easy to make as errors of the same number of degrees? Is probability everywhere the same? Such an assertion would clearly be ridiculous. This is the reason, at any rate, that Astronomers prefer to reject entirely observations which they judge to be too far from the truth, while otherwise retaining the rest, and, indeed, assigning them the same reliability. That is, after making enough observations, they show some of them to be inconsistent, and they consider the ones they retain to have equal worth. Still, I do not see, in the end, what rule they propose which will tell them when to fully admit nearby observations or to clearly reject those that are way off. Why should it never happen that you restore a rejected observation, which, if it is retained, gives the best answer? Nevertheless, if they find their observations to be different (even though they may have been made under the same excellent observational conditions and with the same attention), and if they have to choose between one or another as of greater or lesser worth, then they will try to determine correctly which measurement they will value and which single observation they will retain. Certainly, they will see that it is impossible to define anything accurately following this analysis. For the time being, let us say this: their mode of inquiry makes it impossible to determine anything accurately, although it seems to me that the Astronomers follow this complicated method just as devotedly as if they had followed the simplest one.

3. Now, let there be a straight line *AB* (fig. 1), on which is imagined a certain point *C*, from which position a certain set of observations is made. Let these observations fall on points *a*, *b*, *d*, *e*, and *f*. Thus the following errors are com-

mitted: Ca , Cb , Cd , Ce , Cf . I do not doubt for a moment that minor errors are committed frequently, and that major ones are committed rarely. Let us compare the observer with an archer who, on several occasions, aims his arrow at a vertical line through the point C . Let us also imagine this line as within a rectangular plane of a certain height and width, so that the archer himself is uncertain about the intended plane of his observations. Furthermore, suppose we have only to consider horizontal errors, so that all points can refer to our horizontal line. Let us imagine this line AB to be divided into small equal portions; in this way, we suppose that the arrow's flight may be measured as accurately from one single observation as from many, regardless of how much it may have missed the vicinity of C . We will suppose that that point is as good a center as any of the chance wanderings which may be offered in its place. Thus, let there be a number of possible flight paths, namely a , b , d , e , f , from which the perpendicular lines proportional to their probabilities are am , bn , do , ep , fq . In this way, by the second law of probability, there can be posited the following equation (if taken from an arbitrary point A):

$$AC = \frac{Aa \times am + Ab \times bn + Ad \times do + Ae \times ep + Af \times fq}{am + bn + do + ep + fq}.$$

Clearly, then, this equation shows now that the point C then would fall at the center of gravity of the lines am , bn , do , ep , fq , erected perpendicular from the points a , b , d , e , f . Our earlier supposition had had point C as the center of gravity of the points themselves. If it happens that the single lines am , bn , do , ep , fq , are equal to each other, then both laws of probability are satisfied at once.

4. Errors may be expressed as the sections of the plane defined by the line Ca , and, respectively, by the lines Cb , Cf , Ce , and Cd . The degree of error may be judged from the products of $Ca \times am$, and from $Cb \times bn$, for these sections, and—for the other amounts—the products of $Cf \times fq$, $Ce \times ep$, and $Cd \times do$. By this general rule, an equality of the degrees of errors is assumed between the two sides, so that there is no reason why the sum of degrees of error for one side should be larger than the sum of the other; by our other, earlier hypothesis, an equality of amounts of error is assumed for the two sides, so that there is no reason one ought to consider the sum of one side's amounts of error to be smaller than that of the other. If the number of observations is imagined to be very great, the result will be found adequately by either hypothesis. But if an error is sought among a small number of observations, one ought to establish a more probable and more likely result *a posteriori* under the general rule, because the probable weight belonging to a single observation may only be considered *a posteriori*. This is because an observation is *a priori* uncertain, and by any other method the result is subject to the lot of chance. Therefore, a safer means of investigation must be developed when we have explored as much as possible the nature of chance. I will illustrate this remark with another example, which has no small relevance to our argument.

5. Let the true value of a sought quantity be 1007; and let us arrange this value to be determined in part by chance [literally: by uncertain outcome], but for the most part by certainty. Then let us establish that, however often the thing might be newly considered, the value never fell below 1002 nor exceeded 1012. Now, as for the uncertain observations, let us substitute the everyday throws of two dice, and let us add the number resulting from any throw to one-thousand. In this

way there is no doubt that, if the number of throws is repeated very often, the computed value should approach most nearly to the true value of 1007, whether we compute the probabilities for various throws or not. But on the other hand if there should be a small number of throws for us to observe, it is probable that we will approach the value 1007 more closely if any throw is multiplied by its own probability, than if this multiplication were omitted. Let us suppose, for instance, four tosses which produced the numbers 2, 3, 6, 8; consequently the commonly accepted rule will give us $4 \frac{3}{4}$ as the mean and for a true value we have 1004 $\frac{3}{4}$. If, indeed, these same numbers—2, 3, 6, 8—are multiplied by the number of combinations that can produce them, that is, by 1, 2, 5, 5, then our rule will give $78/13 = 6$ which is much nearer the true value. If the tossed numbers are 2, 3, 7, 8, the first rule will give us 5; the second, $6 \frac{3}{7}$; for throws 4, 5, 10, 11 we obtain the numbers $7 \frac{1}{2}$ and 7. Throws of 5, 6, 8, 9 give the true number 7 by both rules. Nevertheless, I will not proceed lest you infer that combinations are impossible to obtain, for which the commonly accepted rule approaches the true value better than the other. Truly, therefore, that will rarely happen—and whenever it does happen the numbers obtained by both methods differ by quite little from one another. Thus the four throws 5, 6, 7, 10 computed by the commonly accepted rule produce the mean number 7; the other rule however gives only $6 \frac{7}{9}$; it is obvious from these things that one could put forth many other examples of these dice throws. Indeed, it might have been possible to develop this line of argument more clearly from that method of conjecturing, if we were free to dwell upon these particular inquiries. I will point out this general distinction between the two rules: if the commonly accepted rule is used, and the number of throws is increased, note that to whatever extent certainty is determined to occur, to that extent it is only determined by the frequencies of the throws.

6. What will happen, then, if the value or probability of a certain observation is considered and that quantity can be determined? To my mind, anyway, determining that value would not be enough to contend that we will necessarily reach a more reliable judgment. Of course, I willingly confess that this inquiry is both uncertain and indeterminate. Indeed, how could we expect to hold any hypothesis which is so manifestly incongruous? But, I do not think it is possible to expect anything better, and I believe that, failing all alternatives, that particular method is better which agrees with the actual condition or state of something which is itself manifestly revealed by the nature of the argument. Indeed, I will speak frankly about this very approach, which seems most suitable to me. Looking back at paragraph three, the position of point *C* was determined from the observed points *a*, *b*, *d*, *e*, *f*; now, I will inquire into the nature of the curve *AmnopqN* (from which it will be possible to read a scale of probabilities for any observation), the curve which connects the lines *am*, *bn*, *do*, *ep*, *fq*, corresponding to the various probabilities of the points *a*, *b*, *d*, *e*, *f*. Now for this method of analysis to be accurate, it must be subject to certain various axioms:

- a) In as much as deviations are equally easy in both directions, it follows that the scale should have two perfectly similar and equal branches, *HM* and *HN*;
- b) Because all observations tend towards the true point *C*, the deviant ones closer to *C* will certainly be more frequent than the rarer ones which stray

to a greater extent; furthermore, the former are the more probable, the latter the less so. Let this be compared to that place in paragraph three where I spoke of the archer aiming arrows at point C , and concerning this—if it is agreeable—let an experiment be chosen that will physically demonstrate this point.

- c) The degree of probability thus will be greatest at the point C and where the tangent to the scale for the point H will be parallel with line AB .
- d) There is no doubt that these observations from both sides have their own limits, beyond which an observer may never pass; these limits are only as narrow as the dexterity of the observer. The less careful the observer, the more inaccurate his observations. If the limits defined by the points M and N are assumed to be equidistant from point C , then the probabilities of these points are zero, and the usefulness of these points vanishes.
- e) Finally, since points M and N form in themselves the limits between what can happen and what cannot, it is fitting that the last point of the scale on either side should approach closely the line AB in such a way that the tangents at the endpoints M and N become perpendicular to the line AB ; similarly, moreover, the curve itself will indicate that it is impossible to pass beyond points M and N .

7. Let us construct a semi-ellipse of any parameter along the axis MN ; this will plainly satisfy all the conditions. The parameter of the ellipse is arbitrary, because we are concerned only with the proportions between probabilities. Indeed, we may take the semicircle, mapped as the curve MHN . For my part, I am not going to accept as certain one hypothesis or another. I just want to affirm this one thing: it is far better to prefer a commonly held hypothesis which alone, *easily* and *simply*, seems more probable both from all observations and from a single one than accepting an infinite straight line—parallel to and near the line AB —as a scale of probabilities. I would note that if a different semicircle is accepted as the scale, and if it too satisfies the given conditions, then for our purposes it would be pointless to discriminate excessively between them. Now, let us agree that it is the center of this “controlling semicircle” which is sought, so that center will be found located at the point C . For indeed, if the location of this center was arbitrary, then there could be no well-defined relationship among the probabilities of the observations. An observation’s distance from the true point C is all that is needed to describe together the greater and lesser degrees of probability, in such a way that, within the controlling semicircle, this distance is all you need to define the degree of probability; therefore, it is essential that both points [i.e. the center of an indefinitely large number of observations and the true point C] coincide as one and the same point. Certainly this coincidence gives an equation whose value depends on the position of the point C as defined by the second law of probability. [i.e. the observations are unbiased.]

8. Let us now consider the way in which the radius of the controlling semicircle may be established. Clearly, it does not seem advisable to me to choose the greatest error which can be committed as our radius, however often the observations have been undertaken; nor, on the other hand, should one be able to represent a larger line by the smallest possible limiting error. If a justly small radius is held

to be large, then it could be done in such a way that the larger error would perhaps result on the other side, as if no correction were made at all. In fact, I think this should be well noted: whether we set the radius at infinity, or whether we set infinity larger than the greatest error committed (that is, either Cf or Ca), then we will show by the generally accepted law that for every single observation there is a value which is equal to it or of the same probability. In this manner, if am , bn , do , etc., are accurately set around the center C , then they all but become equal to each other. From these remarks, MN is understood to represent double or two times the single value of the radius of the entire field of ambiguity, while the radius CM or CN remains the one which can approach the greatest deviation. Behold now, how I lay bare for all inspection all the achievements of the Astronomers!

9. I would prefer it if, above all else, an observer carefully estimates the greatest error (which he may be certain he never exceeded himself), lest he offend all the gods and angry goddesses with all those observations he *thinks* he determined. That man may himself be neither a severe nor a lenient judge; but, if he becomes trained in the ways of moderation, then he will need to consult no other oracle but proper experience. Nonetheless, he may draw upon many methods, whether to agree with the judgment of one, or to dishonor the method of another. Let it suffice that observations, having already been made, should all be able to account for those which lie in the *field of ambiguity*. Let the radius of the controlling semicircle be r , and thus the width of the whole field of ambiguity equals $2r$.

The sought position of the controlling semicircle's center, C , is then found by the rule that assumes that the sum of all the values

$$Ca \times am + Cb \times bm, \text{etc.} = Cd \times do + Ce \times ep + Cf \times fg, \text{etc.}$$

Then there will be the greatest chance that this very point will be considerably closer to the sought true position than is permitted by the common rule that simply takes $Ca + Cb, \text{etc.} = Cd + Ce + Cf, \text{etc.}$ Therefore, the discrete points a, b, d, e, f , are data of position, and the point C alone is determined by the previous equation.

10. Let us therefore set $Ca = x$, and by the letters $a, b, C, d, e, \text{etc.}$ let us understand quantities to be denoted, which have been determined from the data positions of the points a, b, d, e, f . Then our equation will look like this:

$$\begin{aligned} &x\sqrt{rr - xx} + \overline{x-a}\sqrt{rr - (x-a)^2} + \text{etc.} \\ &= \overline{b-x}\sqrt{rr - (b-x)^2} + \overline{c-x}\sqrt{rr - (c-x)^2} + \text{etc.} \end{aligned}$$

[Note: Bernoulli has changed notation here, so that x = the correction to be applied to the smallest observation to obtain C , and $x - a$ the correction to be applied to the second smallest, $b - x$ to the third smallest, etc.]

Of course, an equation like this is virtually unmanageable, and has the greatest number of useless roots. This is so to such an extent that this analytic expression, once created, tells us nothing, for the useful and useless roots are, in fact, inextricably mixed together. We may prefer a simpler calculation, if the remarks in our treatise are going to be of any use at all. Therefore, to approach another topic, we must try that method which I will present now.

11. Let AB (fig. 2) be a line on which observations are conjectured and from which a certain point A is established. Then we may propose to mark these

observations by the points a, b, d, e, f , etc. One may seek, by the usual rule, a point O among the observed points a, b, d, e, f , etc. such that

$$AO = \frac{Aa + Ab + Ad + Ae + Af + \text{etc.}}{n}$$

where n is considered the number of observations. Thus a semicircle $MmnopqN$ is determined, with the center at O and the radius as r (which we had first accepted for our controlling semicircle), and where ma, bn, do, ep, fq lay perpendicular to MN as various degrees of probability, which had agreed with analogous observations. Then one may seek the center of gravity for all the lines am, bn, do, ep, fq which will have the form of the equation:

$$\frac{Aa \times am + Ab \times bn + Ad \times do + Ae \times ep + Af \times fq}{am + bn + do + ep + fq}.$$

AC may be taken to be equal to that same value. If, indeed, the center C and radius r are altered by turns, then the second controlling semicircle may be described by $M'm'n'o'p'q'N'$, and one can deduce from the following equation the corrected point C' . And *a priori* point C scarcely varies. Because the line Aa is arbitrary and unchanging even through all calculations, then Aa can be supposed to $= 0$, and thus this same endpoint a can be taken as the origin. By this means is revealed:

$$aC = \frac{ab \times bn + ad \times do + ae \times ep + af \times fq}{am + bn + do + ep + fq}.$$

12. If a single observation is made, then both rules agree that the true point may be supposed to be the same as the point of observation, because no correction is available without other observations. Nevertheless, any single point could be taken by any judgment within the whole field of ambiguity, by any commonly accepted hypothesis. In just such a way, any of a group of equally probable values can be chosen. So to make more firm the fundamental nature of our new hypothesis, as giving the greatest amount of probability, requires more observations.

Now, let there be observations of two events, where both general rules agree in prescribing that a true position be accepted at a point midway between both observations, because undoubtedly the points O and C coincide. But, looking back, this assertion must be made more firm according to the second new argument of that second, general hypothesis. But in fact, our rule *can* be just as certain, even if both observations are as infelicitous as possible (i.e., on the condition that the rarest chance happens at opposite ends of the field of ambiguity). This, of course, is impossible to happen, at a time when our observations are undertaken with our greatest attention.

Let us suppose that three observations have taken place, and that they occurred at points b, d , and e . It will help to compare the individual distances with the accepted radius of the controlling semicircle. We have accepted the value of this radius as uniform for the entire curve. We now propose that it be divided into 1000 parts. In this way, if the greatest degree of error is estimated at $160''$ and the distance bd is found to be (for the sake of argument) $120''$ or $200''$, then bd will be either 750 units or 1250 units respectively. Thus is given the distance of any point from the center of the controlling semicircle, and as long as this distance

is sought in the table of sines, an equal cosine will be corresponding along the imposed diagonal. These values will become apparent without any calculation.

13. Example 1: from three observations. Let $bd = 900$ units, and let $be = 1200$ units. Then (by paragraph 10 [sic, 11]), bO will = 700 units, when (by the calculations of the general rule) it is measured as the distance between the observed point b and the true position. Moreover, Od will = 200 units and Oe will = 500 units. Hence, $bn = 714$ units, $do = 980$ units, and $ep = 866$ units. Thus (according to paragraph 10):

$$bC = \frac{900 \times 980 + 1200 \times 866}{714 + 980 + 866} = 750 \text{ units.}$$

Therefore, when bC is greater than bO , the point C will be measured or rather located on the opposite side between O and d , and OC will = -50 units. This correction is thus the one which arises from our hypothesis after the first operation.

After the first correction, we may proceed to the second, in order to locate point C' . Therefore, we may posit the center of the controlling semicircle at the point C at we have determined it; thus it follows that $bC = 750$ units, $bn' = 661$ units, $Cd = 150$ units, $do' = 989$, $Ce = 450$, and $ep' = 893$. From these values, now,

$$bC' = \frac{900 \times 989 + 1200 \times 893}{661 + 989 + 893} = 771.$$

This second correction can be used as a basis for a third correction, by location the center of the controlling semicircle at C' . If the operations are repeated they lead to $bC'' = 780$, which differs from the preceding smaller value of 771.

After a fourth correction we get 784, after a fifth, 787, and thus we continue until 792 is reached. There are, by the way, many devices that can be used to shorten this operation.

From this example, the true distance from the observed point b may be estimated, either by the new method at 792 units or by the commonly accepted one at 700 units. But the new method seems to me to generally give a more probable estimate. Still, if both give identical results that does not detract from our new rule. In any event, the true position may be estimated as midway between 700 and 792; this is found to be 748 [sic]. Perhaps 92 will seem a rather large correction to be made to the common method as corresponding to the greatest possible degree of error [ie. maximum of the curve]. But it does not seem to me, on that account, that successive observations might have then been made inaccurately. You see, if the observation in the middle at d is three times as far away from the endpoint b as from the endpoint e , then certainly the degree of error of that observation b cannot be great with respect to the greatest possible degree of error, for indeed, among nine observations, eight will probably come out better.

14. Now, then, let us bring forward another example, in which I see a small correction present which develops from our new method.

Example 2: from three observations. Now, let $bd = 400$ and $be = 860$; thus

$$bO = \frac{400 + 860}{3}.$$

Let us put the center of the controlling semicircle at the point O . Let $bn = 907$; $Od = 20$; $do = 1000$; $Oe = 440$; $ep = 898$; therefore,

$$bC = \frac{400 \times 1000 + 860 \times 898}{907 + 1000 + 898} = 418.$$

If in turn the center of the controlling semicircle is located such that bC' would = 417, and if this second correction therefore is able to suffice, then it will come to no notable variation. Thus for this reason, both answers now differ from one another by nothing except three units, whereas in the previous example they differed by 92 units.

15. Now let me choose a third example which this controlling semicircle of ours may explain. I had correctly warned earlier that the larger the value which the radius is assumed to have, then the less useful the semicircle becomes. Let us therefore imagine a radius which, as in the first example, measures 1000 units. When it is increased by half again as much, it equals 1500 of the same units. Now, let the distance $bd = 900$ units, and the distance $be = 1200$ units. We will change these numbers 1500, 900, 1200 into 1000, 600, 800; that is, from those one and a half times as great. Thus, $bO = \frac{2}{3} \times 700 = 466.6$; and by this, $bn = 884.4$; $Od = 133.3$; $dO = 991.0$; $Oe = 333.3$; $ep = 942.8$. Hence after the first correction, $bC = 478.6$. Therefore, in this way the first correction gives as much as 12 units greater or 18 units less, whereas in the first example, 50 of these units were up for grabs. The second correction has $bC' = 481$, which we can be pleased with, because nothing at all affects this value. Thus, those 481 units in this example will be worth the $721\frac{1}{2}$ units of example one.

[no paragraph 16]

17. This example of ours—compared to the first one—shows how much it matters that any given observer brings to the process of observing a judgment of his own dexterity, not because he considers the controlling semicircle remarkably small, but because his observation uniquely reflects the abilities of each unique man. It will be considered worse to err on the side of excess, rather than on the side of weakness; for one to judge cautiously is better than to judge boldly. Perhaps the need for this kind of circumspection will not suit everyone, but it suits me fine. In whatever way therefore, the subject is approached, there will be always, by whatever method, an eventual uncertainty. That which is most likely, or most probable, must be sought. Everything which abounds in uncertainty should not be considered any further until it is examined through strict reflection.

It certainly follows that these same observations cannot be made without discrimination, especially when they have been established by two separate Astronomers, either one of whom may excel in experience, skill, or ability. Consider the first observer to be overly certain that he is never in error by more than precisely 1000 seconds from the true position, and suppose that the other observer has been obviously less careful in his observations, such that he can only be precise within 1500 seconds. For both observers, then, chance powerfully affects these same observations, as we have shown in paragraph 13. Now, if I myself will estimate that the true position is a distance of precisely 792 seconds from point b —according to the first observer—and, according to the second, that the distance is nothing else but $721\frac{1}{2}$; while, by the commonly accepted rule, from the manner of both observers, the preceding distance is assigned the value of precisely 700 seconds. This division of mine, even if the paradoxes will not seem

too violently absurd, will have differed from the generally accepted rule. This is so because a more intelligent observer will be better at gauging his own response than a duller one. This, however, seems to me little to fight about. I see the success of such labors as these to be partly the result of labor and partly the result of luck. In one case, the first observer may simply have had bad luck, whereas the other may have succeeded favorably enough in everything (it may be said) even if both of these same observers will have had expert experience.

Taking my consideration into account, what I perceive to be the issue given greatest attention by observant men, is when one is concerned with estimating the median quantity from the greatest number of observations. I do *not* perceive it to be the way of determining the median value equally for everything by using that same second law of probability.

18. Let four or more observations be attempted, all plainly using the same method. Let four observations be placed at points a , b , c [sic], and e . Let $ab = 200$; $ad = 800$; $ae = 1200$. Thus, the general rule gives $aO = 550$. Then, imagine these points inscribed on the semicircle $MmnopN$, with center at O and a radius of 1000. Thus, $aO = 550$ and $am = 835$; $bO = 350$; $bn = 937$; $Od = 250$; $do = 969$; $Oe = 650$; $ep = 760$. From these assertions may be deduced the first correction:

$$\begin{aligned} aC &= \frac{ab \times bn + ad \times do + ae \times ep}{am + bn + do + ep} \\ &= \frac{200 \times 937 + 800 \times 969 + 1200 \times 760}{835 + 937 + 969 + 760} \\ &= 535. \end{aligned}$$

Then, if the center of the controlling semicircle is located at point C , and if the correction is repeated, then one may find $aC' = 531$. Thus, aC' certainly appears to approximate 530, or a little less. To me, anyway, one should not be afraid to say that this last estimate is more probable, than when the common rule is applied.

18. [sic] If the semicircle which I had used as my scale of moderation does not come close to pleasing someone, I permit him to use another. But from that same method the Astronomers will see that the issue returns to exactly the same point; in other words, that individual observations may be established with the same value or weight. For, if this is in fact established, they will have to come back to the usual rule; if this is not realized, then from now on intermediate observations will be considered as worth believing in as the most extreme ones, and further, if this be allowed to happen, I may wish (at any rate) that one look very carefully at one of them, and that he especially compare the value of any one observation with the rest, by which it will itself appear narrower or diminished. If the true value of another observation cannot be estimated along the scale of probabilities, one ought, when possible, to multiply (at least) however many observations by the same integral value, and divide the resulting sum by the sum of the values or the probabilities. These things having been posited, if the method of estimating is taken with a grain of salt, it may be agreed without a doubt and with reason that the result (as has been said)—which is really nothing but the mean—is found by mere arithmetic.

24

Asymptotic Admissibility and Uniqueness of Efficient Estimates in Semiparametric Models

Helmut Strasser¹

ABSTRACT The concept of local asymptotic efficiency of estimators can be made precise in several ways. In semiparametric theory most authors are using local asymptotic minimaxity or asymptotic convolution theorems. We will show how Le Cam's asymptotic admissibility theorem and Hájek's asymptotic uniqueness result can be applied to semiparametric problems.

24.1 Introduction

Assertions about efficiency bounds for estimator sequences can be stated in several ways. Frequently they are expressed as results on asymptotic minimaxity. Another possibility is to restrict attention to so-called regular sequences of estimators and to apply Hájek's (1970) convolution theorem. However, already the early and fundamental paper of Le Cam (1953, Theorems 12 and 13) gives efficiency bounds in terms of an asymptotic admissibility assertion. This result has been rediscovered and improved towards a stochastic uniqueness assertion by Hájek (1972), and its mathematical essence has been revealed by Le Cam (1972, 1979). An abstract presentation of the result is contained in Le Cam (1986, page 112 ff.).

At present most papers on semiparametrics state efficiency bounds for estimator sequences in terms of assertions on regular or median unbiased sequences (cf. Pfanzagl & Wefelmeyer 1982, and Bickel & Klaassen 1986). The general result by Le Cam and Hájek concerning asymptotic admissibility and stochastic uniqueness is hardly ever noticed.

In this paper we will show how Le Cam's asymptotic admissibility and Hájek's uniqueness result can be applied to semiparametric problems. In the abstract context of local asymptotic normality we will use an extended concept of the canonical gradient of a functional (Pfanzagl & Wefelmeyer

¹University of Economics and Business Administration, Vienna

1982). In section 2 we are not going to present any new results but are trying to embed recent developments into the unifying ideas by Le Cam.

However, in section 3 we will give an application of the abstract semiparametric admissibility and uniqueness results to problems with stochastic nuisance parameters. In fact, we show that estimator sequences attaining the risk bounds established for the S-model by several authors (e.g. Pfanzagl & Wefelmeyer 1982, page 226 ff.) are asymptotically uniquely determined with respect to stochastic convergence. This uniqueness property seems not having been noticed in the semiparametric theory. As is shown in Strasser (1996, section 3, Theorem 3.4), it is just this semiparametric uniqueness assertion which makes it possible to extend efficiency bounds for the S-model to the F-model with incidental nuisance parameters.

24.2 The general case

Let H be a linear space and let $H_n \subseteq H$, $n \in \mathbb{N}$, be convex subsets satisfying $H_n \subseteq H_{n+1}$, $n \in \mathbb{N}$. For every $n \in \mathbb{N}$ let $(\Omega_n, \mathcal{A}_n)$ be a measurable space and $(Q_{nh} | \mathcal{A}_n : h \in H_n)$ a family of probability measures. We assume that the sequence of experiments $E_n = (\Omega_n, \mathcal{A}_n, (Q_{nh} : h \in H_n))$ is asymptotically normal, that is, it converges weakly to a Gaussian shift experiment. This is made precise by the following definition. (Let ν_{μ, σ^2} be the normal distribution with mean μ and variance σ^2 .)

Definition 1 (Local Asymptotic Normality) *There exists a continuous positive definite quadratic form $B(\cdot, \cdot)$ on H such that the processes $(L_n(\cdot))$ defined by*

$$\frac{dQ_{nh}}{dQ_{n0}} = \exp \left(L_n(h) - \frac{1}{2}B(h, h) \right), \quad h \in H_n,$$

satisfy the following conditions:

(i) $L(L_n(h)|Q_{n0}) \rightarrow \nu_{0, B(h, h)}$ weakly, if $h \in \bigcup H_n$.

(ii) $L_n(\alpha h_1 + \beta h_2) - \alpha L_n(h_1) - \beta L_n(h_2) \xrightarrow{Q_{n0}} 0$, if $h_1, h_2 \in \bigcup H_n$ and $\alpha, \beta \in \mathbb{R}$.

It should be noted that $(L_n(\cdot))$ is simply a notation for the loglikelihood processes. In most examples these likelihood processes can be expanded by *linear* processes $(X_n(\cdot))$ satisfying $L_n(h) - X_n(h) \rightarrow 0$ (Q_{n0}) for all $h \in \bigcup H_n$. Let us call such sequences of linear processes *central sequences*.

Example 1 A common special case is as follows. Let $\mathcal{P} = \{P_h : h \in H_1\}$ be a parametrized family of probability measures. Define $H_n := \sqrt{n}H_1$ and $Q_{nh} := P_{h/\sqrt{n}}^n$. Assume that the family \mathcal{P} is differentiable in quadratic mean at $h = 0$ with injective derivative $D : H \rightarrow L^2(P_0)$. In this case the

form $B(h_1, h_2) := P_0(Dh_1 \cdot Dh_2)$ is such that the conditions of Definition 1 are satisfied. A sequence of central linear processes is given by

$$X_n(h)(\omega) = \frac{1}{\sqrt{n}} \sum_{i=1}^n Dh(\omega_i).$$

Endowed with B as inner product, the linear space H is an inner product space. Without loss of generality we assume that H is complete and $\bigcup H_n = H$. If the families (Q_{nh}) arise from a local parametrization as in the preceding example and if the manifold \mathcal{P} is of finite dimension then the linear space H is finite dimensional too and the linear processes $(X_n(\cdot))$ can be represented as inner products $X_n(h) = B(h, Z_n)$ with H -valued random variables Z_n . However, in semiparametric and in nonparametric statistics one is dealing with situations where H is an infinite dimensional linear space. In such a case it is not always possible to represent the linear processes $(X_n(\cdot))$ as inner products with H -valued random variables.

Discussion In order to explain the local asymptotic approach to estimation let us have another look to the case of Example 1. Let $\kappa : \mathcal{P} \rightarrow \mathbb{R}$ be a (possibly nonlinear) function admitting an expansion

$$\kappa(P_h) = \kappa(P_0) + \int Dh g \, dP_0 + o(P_0(h^2)), \quad h \in H_1.$$

The linear function $f(h) = \int Dh g \, dP_0$ is the derivative of κ at P_0 . (In general, the gradient g is not uniquely determined.) In the local asymptotic framework we try to find an estimator sequence (κ_n) such that the distributions of $\sqrt{n}(\kappa_n - \kappa(P_h/\sqrt{n}))$ under $P_{h/\sqrt{n}}^n$ are concentrated around zero as much as possible. In view of the expansion of the function κ , this amounts to saying that the distributions of the random variables $T_n = \sqrt{n}(\kappa_n - \kappa(P_0))$ under Q_{nh} are concentrated around $f(h)$ as much as possible. Thus, from the local asymptotic point of view we are dealing with the estimation of the *linear* function f by random variables ('estimators') of the form T_n . An efficient sequence of estimators can be characterized in the following way: Define the canonical gradient g^* in the sense of Pfanzagl & Wefelmeyer (1982, page 72) of f (resp. κ) to be the orthogonal projection of any gradient g onto the tangent space $\overline{\{Dh : h \in H\}}$. Then an estimator sequence is (locally asymptotically) efficient (at P_0) if it satisfies

$$T_n(\omega) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g^*(\omega_i) \xrightarrow{Q_{n0}} 0.$$

Let $f : H \rightarrow \mathbb{R}^p$ be a continuous linear function. We are going to study the semiparametric problem of estimating f . In the general case the description of efficient estimator sequences can be done as follows: Let each component f_i of f be represented in terms of the quadratic form B by

$$f_i(h) = B(h_i^*, h), \quad h \in H, \quad i = 1, 2, \dots, p.$$

Define T_n to be

$$T_n = (X_n(h_1^*), \dots, X_n(h_p^*)), n \in \mathbb{N}.$$

This type of construction extends the notion of canonical gradients to general (e.g. non-i.i.d.) cases where tangent spaces in the sense of Pfanzagl & Wefelmeyer (1982) are not available. By abuse of language we will refer to h_i^* as canonical gradients.

In the following let $A := (B(h_i^*, h_j^*))$ denote the Gram matrix of the canonical gradients of f . Let \mathcal{W} be the set of all loss functions $W : \mathbb{R}^p \rightarrow [0, 1]$ satisfying the following conditions:

- (i) The level sets $\{W \leq \alpha\}$ are convex.
- (ii) W is centrally symmetric, i.e. $W(x) = W(-x)$, $x \in \mathbb{R}^p$.
- (iii) There is at least one $\alpha < \sup W$ such that the level set $\{W < \alpha\}$ contains the origin 0 as inner point.

Parts (ii) and (iii) of the following theorem contains the main admissibility and uniqueness assertions in a general semiparametric framework.

Theorem 1

(i) For all bounded continuous functions $W \in \mathcal{W}$ and all $h \in \bigcup H_n$ we have

$$\lim_{n \rightarrow \infty} \int W(T_n - f(h)) dQ_{nh} = \int W d\nu_{0,A}.$$

(ii) If (S_n) is an arbitrary estimator sequence and if there is some $W_1 \in \mathcal{W}$ and some $h_1 \in \bigcup H_n$ such that

$$\liminf_{n \rightarrow \infty} \int W_1(S_n - f(h_1)) dQ_{nh_1} < \int W_1 d\nu_{0,A},$$

then there is some $W_2 \in \mathcal{W}$ and some $h_2 \in \bigcup H_n$ such that

$$\limsup_{n \rightarrow \infty} \int W_2(S_n - f(h_2)) dQ_{nh_2} > \int W_2 d\nu_{0,A},$$

(iii) If (S_n) is an arbitrary estimator sequence and if for all $W \in \mathcal{W}$ and all $h \in \bigcup H_n$

$$\limsup_{n \rightarrow \infty} \int W(S_n - f(h)) dQ_{nh} \leq \int W d\nu_{0,A},$$

then we have

$$S_n - T_n \xrightarrow{Q_{n0}} 0.$$

Assertion (i) of Theorem 1 is obvious and stated only for matters of completeness. Assertion (ii) contains the admissibility statement, and (iii) is the result on the stochastic uniqueness.

Our formulation of admissibility and uniqueness differs from more usual ones in that we are admitting cases where $p > 1$. The price to be paid is that we have to compare distributions by families of loss functions instead of considering one fixed loss function. It is remarkable that for $p = 1$ assertions (ii) and (iii) are valid even if the conditions are claimed only for one single loss function in \mathcal{W} . This property does not extend to $p > 2$ due to the Stein inadmissibility phenomenon.

The difficult part of Theorem 1 to be proved is part (iii) for the case $p = 1$. This being proved, the complete assertion of part (iii) for $p > 1$ follows since the set \mathcal{W} contains sufficiently many loss functions which depend only on one single component of f . Part (ii) is an immediate consequence of part (iii).

The proof of assertion (iii) for $p = 1$ can be organized in different ways. First, instead of H we need only consider the one-dimensional orthogonal complement of the null space of f . Thus, the assertion being reduced to a one-dimensional parameter space it can be proved along the lines Le Cam or Hájek have shown to us. One possibility is to prove admissibility by a Bayesian argument while passing to the limit $n \rightarrow \infty$. This is Le Cam's (1953) original method. Another course is followed by Hájek (1972). There the limiting process $n \rightarrow \infty$ is performed first and then the non-asymptotic version of Blyth's (1951) admissibility theorem is applied to the situation in the limit. This idea is basic for the general asymptotic decision theory which has been developed by Le Cam in his papers (1972, 1979), and is presented in his monograph (Le Cam 1986). A proof of assertion (iii) for $p = 1$ along the lines of Hájek's and Le Cam's ideas could be found in Strasser (1985, Theorem 83.5).

Let us finish this section with some remarks concerning the relation of our results to asymptotic minimax theorems. A typical assertion of this type is Le Cam & Yang (1990, Theorem 2 on p. 84). In case of $p = 1$ when admissibility theorems are available based on one fixed loss function then the admissibility assertion implies the minimax assertion. In this respect, admissibility is stronger than minimaxity (Le Cam & Yang 1990, p. 87). In our case, however, since we are dealing with families of loss functions in order to cover $p > 1$ the asymptotic minimax theorem is not an implication of our Theorem 1.

24.3 An application

Let $\Theta \subseteq \mathbb{R}^p$ and $\Lambda \subseteq \mathbb{R}^q$ be open sets. For every pair $(\theta, \eta) \in \Theta \times \Lambda$ let $P_{\theta, \eta}$ be a probability measure on a measurable space (Ω, \mathcal{A}) . We are going to estimate the parameter θ in a model where the nuisance parameter η

is governed by independent and identically distributed random variables. This model is called the S-model. As references for this model see Pfanzagl & Wefelmeyer (1982, page 226 ff.), Bickel & Klaassen (1986), and Pfanzagl (1993). Some technical details, quoted in this section, are proved in Strasser (1996).

Let $\Gamma|\mathcal{B}(\Lambda)$ be a probability measure and define

$$Q_{\theta,\Gamma}(A \times B) = \int_B P_{\theta,\eta}(A) \Gamma(d\eta).$$

Let us denote $Q'_{\theta,\Gamma}(A) := Q_{\theta,\Gamma}(A \times \Lambda)$, $A \in \mathcal{A}$.

In order to investigate the local asymptotic structure of the family $(Q_{\theta,\Gamma})$ we fix a pair (θ, Γ) and define the linear space

$$H := \left\{ (s, k) \in \mathbb{R}^p \times L^2(\Gamma) : \int k d\Gamma = 0 \right\}.$$

Let $\mathcal{C}_{\infty}(\Lambda)$ be the set of all continuous functions with compact support on Λ . For every $n \in \mathbb{N}$ let

$$H_n := \left\{ (s, k) \in \mathbb{R}^p \times \mathcal{C}_{\infty}(\Lambda) : \int k d\Gamma = 0, \theta + \frac{s}{\sqrt{n}} \in \Theta, 1 + \frac{k}{\sqrt{n}} \geq 0 \right\}.$$

For $n \in \mathbb{N}$ and $h := (s, k) \in H_n$ let

$$Q_{nh} := Q_{\theta+s/\sqrt{n}, (1+k/\sqrt{n})\Gamma}^n$$

where $(1 + k/\sqrt{n})\Gamma$ denotes the measure with density $(1 + k/\sqrt{n})$ with respect to Γ .

Let us consider the sequence of experiments

$$E_n := \left(\Omega^n, \mathcal{A}^n, (Q_{nh} : h \in H_n) \right), n \in \mathbb{N}.$$

Assumption 1 *The family $(P_{\theta,\eta})$ of probability measures is continuously differentiable in quadratic mean.*

Denoting the first partial loglikelihood derivative of $(P_{\theta,\eta})$ by $l_{1\theta}$ we define a quadratic form on the linear space H by

$$B_{\theta,\Gamma}(h_1, h_2) := \int Q_{\theta,\Gamma}(l_{1\theta} \cdot s_1 + k_1 | \mathcal{A}) Q_{\theta,\Gamma}(l_{1\theta} \cdot s_2 + k_2 | \mathcal{A}) dQ_{\theta,\Gamma},$$

where $h_1 := (s_1, k_1)$, $h_2 := (s_2, k_2)$. (By $Q_{\theta,\Gamma}(\cdot | \mathcal{A})$ we denote the conditional expectation operator.)

Assumption 2 *The quadratic form $B_{\theta,\Gamma}(\cdot, \cdot)$ is positive definite on H .*

It can be shown that under the Assumptions 1 and 2 the sequence of experiments (E_n) is asymptotically normal in the sense of Definition 1 with the quadratic form $B_{\theta,\Gamma}$. (Cf. Pfanzagl & Wefelmeyer 1982, p. 50 and chapter 14, or applying our notation, Strasser 1993, section 2.) One obtains as central sequence the sequence of linear processes

$$X_n(h, \omega) = \frac{1}{n} \sum_{i=1}^n Q_{\theta,\Gamma}(l_{1\theta} \cdot s + k|\mathcal{A})(\omega_i), \quad h = (s, k) \in H.$$

Let (κ_n) be a \sqrt{n} -consistent sequence of estimators for the parameter θ . This sequence is locally asymptotically efficient at the point (θ, Γ) if for the sequence of experiments (E_n) the estimator sequence $(\sqrt{n}(\kappa_n - \theta))$ is an efficient sequence of the linear function $f(s, k) = s$. An efficient sequence for f can be characterized along the lines of section 2 by means of the canonical gradients of f . Assumption 2 implies the existence of $h_i^* := (s_i^*, k_i^*) \in H$ such that

$$B_{\theta,\Gamma}(h, h_i^*) = f_i(h) = s_i, \quad h \in H, i = 1, \dots, p.$$

These elements h_i^* are the canonical gradients of the components f_i of f . Thus, we obtain as an efficient estimator sequence for f

$$T_n = (X_n(h_1^*), \dots, X_n(h_p^*)).$$

If $A_{\theta,\Gamma} := (Q_{\theta,\Gamma}(h_i^*, h_j^*))$ then we have by Theorem 1, part (i),

$$\mathcal{L}(T_n | Q_{nh}) \rightarrow \nu_{f(h), A_{\theta,\Gamma}} \text{ weakly, } h \in \bigcup H_n.$$

With these preparations we may specialize part(ii) and (iii) of Theorem 1 to the situation considered in this section. Thus, we arrive at an asymptotic admissibility and uniqueness assertion for estimation problems with nuisance parameters.

Theorem 2

(i) If (κ_n) is an arbitrary estimator sequence for the parameter θ and if there are some $W_1 \in \mathcal{W}$ and $(s_1, k_1) \in \bigcup H_n$ such that

$$\liminf_{n \rightarrow \infty} \int W_1 \left(\sqrt{n}(\kappa_n - \theta) - s_1 \right) dQ_{\theta+s_1/\sqrt{n}, (1+k_1/\sqrt{n})\Gamma}^n < \int W_1 d\nu_{0, A_{\theta,\Gamma}},$$

then there are some $W_2 \in \mathcal{W}$ and $(s_2, k_2) \in \bigcup H_n$ such that

$$\limsup_{n \rightarrow \infty} \int W_2 \left(\sqrt{n}(\kappa_n - \theta) - s_2 \right) dQ_{\theta+s_2/\sqrt{n}, (1+k_2/\sqrt{n})\Gamma}^n > \int W_2 d\nu_{0, A_{\theta,\Gamma}}.$$

(ii) If (κ_n) is an arbitrary estimator sequence for the parameter θ and if for all $W \in \mathcal{W}$ and all $(s, k) \in \bigcup H_n$

$$\limsup_{n \rightarrow \infty} \int W \left(\sqrt{n}(\kappa_n - \theta) - s \right) dQ_{\theta+s/\sqrt{n}, (1+k/\sqrt{n})\Gamma}^n \leq \int W d\nu_{0, A_{\theta,\Gamma}},$$

then

$$\sqrt{n}(\kappa_n - \theta) - T_n \xrightarrow{Q_{\theta,\Gamma}^n} 0.$$

24.4 REFERENCES

- Bickel, P. J. & Klaassen, C. A. J. (1986), 'Empirical Bayes estimation in functional and structural models, and uniformly adaptive estimation of location', *Advances in Applied Mathematics* 7, 55–69.
- Blyth, C. R. (1951), 'On minimax statistical procedures and their admissibility.', *Annals of Mathematical Statistics* 22, 22–42.
- Hájek, J. (1970), 'A characterization of limiting distributions of regular estimators', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 14, 323–330.
- Hájek, J. (1972), Local asymptotic minimax and admissibility in estimation, in L. Le Cam, J. Neyman & E. L. Scott, eds, 'Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 175–194.
- Le Cam, L. (1953), 'On some asymptotic properties of maximum likelihood estimates and related Bayes estimates', *University of California Publications in Statistics* 1, 277–330.
- Le Cam, L. (1972), Limits of experiments, in L. Le Cam, J. Neyman & E. L. Scott, eds, 'Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 245–261.
- Le Cam, L. (1979), On a theorem of J. Hájek, in J. Jurečková, ed., 'Contributions to Statistics—Hájek Memorial Volume', Akadémia, Prague, pp. 119–135.
- Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.
- Le Cam, L. & Yang, G. L. (1990), *Asymptotics in Statistics: Some Basic Concepts*, Springer-Verlag.
- Pfanzagl, J. (1993), 'Incidental versus random nuisance parameters', *Annals of Statistics* 21, 1663–1691.
- Pfanzagl, J. & Wefelmeyer, W. (1982), *Contributions to a General Asymptotic Statistical Theory*, Vol. 13 of *Lecture Notes in Statistics*, Springer-Verlag, New York.
- Strasser, H. (1985), *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*, De Gruyter, Berlin.
- Strasser, H. (1996), 'Asymptotic efficiency of estimates for models with incidental nuisance parameters', *Annals of Statistics* 24, 879–901.

25

Contiguity in Nonstationary Time Series

A. R. Swensen¹

ABSTRACT We show how a result of Cox & Llatas (1991) can be derived using contiguity arguments. Also we compare the asymptotic power functions of three tests of the characteristic polynomial of an AR(1) process having a root at unity.

25.1 Introduction

In this note we shall deal with the use of contiguity arguments in a non-standard situation. In particular we shall consider local neighbourhoods of an autoregressive process having a root at unity. To be specific, suppose that the data are generated by the difference equation

$$X_t - (1 + \phi/T)X_{t-1} = \epsilon_t \quad t = 1, \dots, T, \quad (1)$$

where X_0 is assumed to be fixed, and $\epsilon_1, \epsilon_2, \dots$ are independent and identically distributed with mean zero and variance σ^2 and ϕ is a local parameter. Let $P_{\phi,T}$ be the joint distribution of the observations from (1).

Models with a root at unity have received considerable interest in econometrics during the last fifteen years. Whether a model of this type is fitted or a competitor of the form of a linear trend plus a stationary term is chosen, has substantial consequences for the economic interpretation. In the first case the effect of a random shock is permanent and has a lasting effect, while in the second the effect of a random shock is temporary.

In a previous note, Swensen (1993), we explained how contiguity arguments similar to those of Le Cam's third lemma (Hájek & Šidák 1967, page 208), can be used to find the asymptotic distribution of various statistics under $P_{\phi,T}$. What is needed is that the asymptotic distribution of the statistics under the hypothesis can be represented as the distribution of a functional of the Brownian motion W , which is the weak limit when the observations of (1) are suitably rescaled, and $\phi = 0$. The distribution of such a statistics is, under the local alternatives ϕ/T , given by evaluating the

¹Central Bureau of Statistics of Norway and University of Oslo

same functional at K , where K is a solution of the stochastic differential equation

$$dK_u = \phi K_u dr + dW_u \quad \text{with } K_0 = 0. \quad (2)$$

We shall show below that a similar result holds for a more general class of statistics, the M estimators considered recently by Cox & Llatas (1991). They used a direct derivation under the alternatives and obtained the result without the smoothness conditions on the distribution of the errors that we shall require.

Jeganathan (1995) has also considered the M estimators in a more general setup and demonstrated the same result. The novelty of the present exposition, as also was the case in Swensen (1993), is thus to furnish the details of how some simple calculations related to the distribution of the Radon-Nikodym derivative of the strong solutions of (2) for different values of ϕ , together with contiguity arguments along the lines of Le Cam's third lemma, give the result. We think that the simplicity of the arguments in the derivation justifies another exposition even though the results we end up with are well known under weaker assumptions.

A number of tests has been suggested for the hypothesis $H: \phi = 0$. Drawing on the asymptotic results mentioned above we shall study the power of three of them: The T-test of Dickey & Fuller (1979), the one based on the least squares estimator proposed by the same authors and a test recently suggested by Kahn & Ogaki (1990).

In the next section we derive the distribution of the M estimator as described above and in the final section we compare the asymptotic power functions of the three tests.

25.2 The distribution of M-estimators in local neighbourhoods

Consider a statistics having a representation of the form

$$\hat{\phi}_T = \frac{\sum_{t=1}^T X_{t-1}\psi(\epsilon_t)}{\sum_{t=1}^T X_{t-1}^2} + o_{P_0,T}(T^{-1}) \quad (3)$$

where ψ is a differentiable function with derivative $\dot{\psi}$. Furthermore, we shall assume that $E\psi(\epsilon_1) = 0$ and $E\dot{\psi}(\epsilon_1) = 1$. As explained in Cox & Llatas (1991), the last condition is merely a convenient normalization as long as $E\dot{\psi}(\epsilon_1) \neq 0$.

We shall require that the distribution of the errors is absolutely continuous with respect to Lebesgue measure on the real line. Let f be the common density. Furthermore we assume:

Assumption 3 *The density f is absolutely continuous with respect to the*

Lebesgue measure on the real line with derivative \dot{f} and satisfies

$$0 < I(f) = \int_{-\infty}^{\infty} \left[\dot{f}(x)/f(x) \right]^2 f(x) dx < \infty.$$

Then it follows from the results in Jeganathan (1995) that the logarithm of the likelihood ratio $\Lambda_{1,T}$ can be expanded as

$$\Lambda_{1,T} = 2 \sum_{t=1}^T Z_{t,T} - 2 \sum_{t=1}^T Z_{t,T}^2 + o_{P_0,T}(1) \quad (4)$$

where $Z_{t,T} = -(\phi X_{t-1}/T) \dot{s}(\epsilon_t)$ and $\dot{s}(x) = \frac{1}{2} \left(\dot{f}(x)/f(x) \right) I[f(x) > 0]$ is the quadratic mean differential of $\sqrt{f(x)}$.

The essential part of a contiguity argument along the lines of Le Cam's third lemma is now to show the joint convergence under $P_{0,T}$ of the likelihood ratio and the sequence of statistics which is the focus of interest.

This can be deduced from the weak convergence of the partial sums

$$\frac{1}{\sqrt{T}} \left(\sum_{i=1}^{[rT]} \epsilon_i, 2 \sum_{i=1}^{[rT]} \dot{s}(\epsilon_i), \sum_{i=1}^{[rT]} \psi(\epsilon_i) \right)', \quad 0 \leq r \leq 1$$

towards the Brownian motion $(W_{1r}, W_{2r}, W_{3r})'$ with covariance matrix

$$r \begin{pmatrix} \sigma_1^2 & 1 & \rho\sigma_1\sigma_3 \\ 1 & \sigma_2^2 & 1 \\ \rho\sigma_1\sigma_3 & 1 & \sigma_3^2 \end{pmatrix}.$$

Here $[rT]$ denotes the integer value of rT and $\sigma_1^2 = E\epsilon_1^2 = \sigma^2$, $\sigma_2^2 = I(f) = 4E\dot{s}(\epsilon_1)^2$, $\sigma_3^2 = E\psi(\epsilon_1)^2$. Furthermore $\rho\sigma_1\sigma_3 = E(\epsilon_1\psi(\epsilon_1))$. Integrating by parts and using that $E|\epsilon_1| < \infty$ and $E|\psi(\epsilon_1)| < \infty$, one has

$$2E\epsilon_1\dot{s}(\epsilon_1) = \int_{-\infty}^{\infty} x\dot{f}(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1$$

and

$$2E\psi(\epsilon_1)\dot{s}(\epsilon_1) = \int_{-\infty}^{\infty} \psi(x)\dot{f}(x)dx = \int_{-\infty}^{\infty} \dot{\psi}(x)f(x)dx = E\dot{\psi}(\epsilon_1) = 1$$

where the last equality follows from the assumptions on ψ referred to above. This explains the entries equal to one in the covariance matrix above.

Under some regularity conditions, which are satisfied in the present situation, the weak convergence of a sequence of martingales towards a martingale implies the simultaneous convergence of the sequence of martingales

and their associated predictable processes towards the corresponding processes of the limiting martingale (Jacod & Shiryaev 1987, Theorem VI.6.1). This implies the weak convergence of

$$\frac{1}{T} \left(\sum_{t=1}^{[rT]} X_{t-1} \epsilon_t, 2 \sum_{t=1}^{[rT]} X_{t-1} \dot{s}(\epsilon_t), \sum_{t=1}^{[rT]} X_{t-1} \psi(\epsilon_t), \frac{1}{T} \sum_{t=1}^{[rT]} X_{t-1}^2 \right)$$

towards

$$\left(\int_0^r W_{1u} dW_{1u}, \int_0^r W_{1u} dW_{2u}, \int_0^r W_{1u} dW_{3u}, \int_0^r W_{1u}^2 du \right).$$

Using the asymptotic expansion of $\Lambda_{1,T}$ in (4), the joint convergence of the distribution of

$$\left(T \frac{\sum_{t=1}^T X_{t-1} \psi(\epsilon_t)}{\sum_{t=1}^T X_{t-1}^2}, \Lambda_{1,T} \right)$$

towards the distribution of

$$\left(\frac{\int_0^1 W_{1u} dW_{3u}}{\int_0^1 W_{1u}^2 du}, \phi \int_0^1 W_{1u} dW_{2u} - \frac{1}{2} \phi^2 \sigma_2^2 \int_0^1 W_{1u}^2 du \right)$$

is thus established.

As in our previous note on the subject (Swensen 1993), we now consider the distribution of a conditional version of the limiting likelihood ratio. It turns out that the distribution is the same as the distribution of the Radon-Nikodym derivative of the distributions of the solutions arising in a stochastic differential equation. Denote the random variable

$$\phi \int_0^r W_{1u} dW_{2u} - \frac{1}{2} \phi^2 \sigma_2^2 \int_0^r W_{1u}^2 du$$

by Λ_r , for $0 \leq r \leq 1$ and let $\mathcal{F}_{1,3}$ be the σ -field induced by $\{\mathbf{W}_u = (W_{1u}, W_{3u}), 0 \leq u \leq 1\}$. Define L_r by $L_r = \exp(\Lambda_r)$, and consider the distribution of $E(L_1 | \mathcal{F}_{1,3})$. Denote the matrix

$$\begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_3 \\ \rho \sigma_1 \sigma_3 & \sigma_3^2 \end{pmatrix}$$

by Σ and let $\mathbf{e} = (1, 1)'$. From Phillips (1989, Lemma 3.1), the conditional distribution of W_2 given $\mathcal{F}_{1,3}$ is equal to the distribution of

$$\mathbf{e}' \Sigma^{-1} \mathbf{W} + (\sigma_2^2 - \mathbf{e}' \Sigma^{-1} \mathbf{e})^{1/2} \tilde{B}$$

where \tilde{B} is a standard Brownian motion independent of $\mathbf{W} = (W_1, W_3)'$. Using the formula for the moment generating function of a Gaussian variable, the distribution of $E(L_1 | \mathcal{F}_{1,3})$ equals the distribution of

$$\exp \left(\phi \int_0^1 W_{1u} d(\mathbf{e}' \Sigma^{-1} \mathbf{W}_u) - \frac{1}{2} \phi^2 \mathbf{e}' \Sigma^{-1} \mathbf{e} \int_0^1 W_{1u}^2 du \right). \quad (5)$$

Let

$$C = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_3 & \sigma_3\sqrt{1-\rho^2} \end{pmatrix}$$

so that $\mathbf{W}_u = C\mathbf{B}_u$ where $\mathbf{B}_u = (B_{1u}, B_{3u})'$ and B_1 and B_3 are independent standard Brownian motions; that is, $EB_{1u}^2 = EB_{3u}^2 = u$ and $EB_{1u}B_{3u} = 0$. Then $\Sigma = CC'$ and by a straightforward application of Ito's lemma,

$$\int_0^1 W_{1u} d(\mathbf{e}' \Sigma^{-1} \mathbf{W}_u) = \int_0^1 \mathbf{B}'_u C' \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \Sigma^{-1} C d\mathbf{B}_u.$$

Denote by $\mathbf{a}(\mathbf{B}_u)'$ the vector valued variable

$$\mathbf{B}'_u C' \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \Sigma^{-1} C.$$

Then

$$\begin{aligned} \mathbf{a}(\mathbf{B}_u)' \mathbf{a}(\mathbf{B}_u) &= \mathbf{B}'_u C' \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \Sigma^{-1} C C' \Sigma^{-1} \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} C \mathbf{B}_u \\ &= \mathbf{B}'_u C' \begin{pmatrix} \mathbf{e}' \Sigma^{-1} \mathbf{e} & 0 \\ 0 & 0 \end{pmatrix} C \mathbf{B}_u = \mathbf{e}' \Sigma^{-1} \mathbf{e} W_{1u}^2, \end{aligned}$$

so that (5) may be written

$$\exp \left(\phi \int_0^1 \mathbf{a}(\mathbf{B}_u)' d\mathbf{B}_u - \frac{1}{2} \phi^2 \int_0^1 \mathbf{a}(\mathbf{B}_u)' \mathbf{a}(\mathbf{B}_u) du \right),$$

which has the same distribution as the Radon-Nikodym derivative of the distribution of \mathbf{B} and the distribution of the solution of the stochastic differential equation

$$\begin{aligned} d\mathbf{F}_u &= \phi \mathbf{a}(\mathbf{F}_u) du + d\mathbf{B}_u \\ &= \phi C' \Sigma^{-1} \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} C \mathbf{F}_u du + d\mathbf{B}_u, \mathbf{F}_0 = 0. \end{aligned}$$

In particular this implies that $EL_1 = 1$, so that the contiguity of $P_{\phi,T}$ and $P_{0,T}$ follows.

Setting $\mathbf{K} = CF$, and using Ito's lemma once more the stochastic differential equation takes the form

$$d\mathbf{K}_u = \phi \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \mathbf{K}_u du + d\mathbf{W}_u$$

or

$$\begin{aligned} dK_{1u} &= \phi K_{1u} dr + dW_{1u} \\ dK_{3u} &= \phi K_{1u} dr + dW_{3u}. \end{aligned}$$

Now we can conclude using arguments similar to those used in Le Cam's third lemma, that the asymptotic distribution of $T(\hat{\phi}_T - 1)$ under $P_{\phi,T}$ is equal to the distribution of

$$\frac{\int_0^1 K_{1u} dK_{3u}}{\int_0^1 K_{1u}^2 du}.$$

By the last of the previous equations this may be written

$$\phi + \frac{\int_0^1 K_{1u} dW_{3u}}{\int_0^1 K_{1u}^2 du}.$$

We have thus the following result.

Proposition 1 (*Cox and Llatas*). *Assume that the regularity conditions on ψ and f stated above are satisfied. Let $\hat{\phi}_T$ be as defined in (3). Then the distribution of $T(\hat{\phi}_T - (1 + \phi/T))$ under $P_{\phi,T}$ converges towards the distribution of $\int_0^1 K_{1u} dW_{3u} / \int_0^1 K_{1u}^2 du$.*

Remark that only the second moment of the error distribution is assumed to exist, not the $(2+\delta)$, for some $\delta > 0$, as assumed by Cox & Llatas (1991). However, as is mentioned in the Introduction, the present approach requires stronger smoothness conditions on the errors.

25.3 Comparison of the local power of three alternative tests for a unit root in AR(1) processes

We still assume that the observations are generated by (1), and that the regularity conditions introduced in the previous section continue to hold. A hypothesis of considerable interest is $H : \phi = 0$ against $A : \phi < 0$.

The first test we shall consider is based on the least squares estimator $\hat{\alpha}_T = \sum_{t=1}^T X_t X_{t-1} / \sum_{t=1}^T X_{t-1}^2$. It consists of rejecting the hypothesis H when $(\sum_{t=1}^T X_{t-1}^2)^{1/2}(\hat{\alpha}_T - 1)/\hat{\sigma}$ is small, where $\hat{\sigma}$ is a consistent estimator of σ . This is a test which is asymptotically equivalent to the test based on the statistic $\hat{\tau}$ of Dickey & Fuller (1979). Using the convergence results of the previous section, the asymptotic distribution of the test statistics equals $\int_0^1 W_{1u} dW_{1u} / \sigma(\int_0^1 W_{1u}^2 du)^{1/2}$ if $\phi = 0$. Let $t_{1,\alpha}$ be the α -fractile of this distribution.

An alternative test is to reject the hypothesis if $\hat{\alpha}_T - 1$ is small. This is asymptotically equivalent to the test based on the statistic $\hat{\rho}$ of Dickey & Fuller (1979). The asymptotic distribution if $\phi = 0$ of $T(\hat{\alpha}_T - 1)$ is $\int_0^1 W_{1u} dW_{1u} / \int_0^1 W_{1u}^2 du$. Let $t_{2,\alpha}$ be the α fractile of this distribution.

A third test recently proposed by Kahn & Ogaki (1990) is to reject the hypothesis if $2(\hat{b}_T - 0.5) < z_\alpha$, where $\hat{b}_T = \sum_{t=1}^T X_t \Delta X_t / \sum_{t=1}^T \Delta X_t$ and z_α is the α -fractile of the χ^2 -distribution with one degree of freedom. Since

$$\sum_{t=1}^T X_t \Delta X_t / \sum_{t=1}^T \Delta X_t^2 = 1 + \sum_{t=1}^T X_{t-1} \Delta X_t / \sum_{t=1}^T \Delta X_t^2,$$

$(\hat{b}_T - 1)$ converges in distribution under H towards the distribution of $\int_0^1 W_{1u} dW_{1u} / \sigma^2$, which equals $\frac{1}{2}((W_{11}/\sigma)^2 - 1)$ by an easy application of Ito's lemma. Therefore $2(\hat{b}_T - 0.5)$ converges in distribution towards a random variable having a χ^2 -distribution with one degree of freedom.

All the tests are thus functionals of W_{1t} . Hence, as explained above and in some more detail in Swensen (1993), the asymptotic distribution under $P_{\phi,T}$ may be found by evaluating the functionals at K_1 where

$$dK_{1u} = \phi K_{1u} dt + dW_{1u} \quad \text{with } K_{10} = 0. \quad (6)$$

Thus, the three asymptotic power functions are

$$\begin{aligned} \beta_1(\phi) &= P \left\{ \frac{\int_0^1 K_{1u} dK_{1u}}{\sigma(\int_0^1 K_{1u}^2 du)^{1/2}} < t_{1,\alpha} \right\} \\ \beta_2(\phi) &= P \left\{ \frac{\int_0^1 K_{1u} dK_{1u}}{\int_0^1 K_{1u}^2 du} < t_{2,\alpha} \right\} \\ \beta_3(\phi) &= P \left\{ \left(\frac{K_{11}}{\sigma} \right)^2 < z_\alpha \right\}. \end{aligned}$$

From Phillips (1987, Lemma 2) it follows that $2|\phi| \int_0^1 K_{1u}^2 du \rightarrow \sigma^2$ in probability and that $\int_0^1 K_{1u} dW_{1u} / \sigma \left(\int_0^1 K_{1u}^2 du \right)^{1/2}$ converges to a standard Gaussian distribution as $\phi \rightarrow -\infty$. Writing

$$\beta_2(\phi) = P \left\{ \frac{\int_0^1 K_{1u} dK_{1u}}{(\int_0^1 K_{1u}^2 du)^{1/2}} < t_{2,\alpha} \left(\int_0^1 K_{1u}^2 du \right)^{1/2} \right\}$$

and

$$\beta_3(\phi) = P \left\{ \frac{\int_0^1 K_{1u} dK_{1u}}{(\int_0^1 K_{1u}^2 du)^{1/2}} < \frac{(\sigma^2/2)(z_\alpha - 1)}{\left(\int_0^1 K_{1u}^2 du \right)^{1/2}} \right\},$$

it follows that as $\phi \rightarrow -\infty$ and $\alpha < \frac{1}{2}$

$$\beta_3(\phi) < \beta_1(\phi) < \beta_2(\phi) \quad (7)$$

since $t_{1,\alpha} < 0$ and $z_\alpha - 1 < 0$ when $\alpha < \frac{1}{2}$. By a similar argument one also shows that $\beta_i(\phi) \rightarrow 1$ as $\phi \rightarrow -\infty$, for $i = 1, 2, 3$.

Finally, we remark that in a family of probability measures of the form

$$dP_\phi = \exp\left(\phi x - \frac{1}{2}\phi^2 y\right) d\mu \quad (8)$$

where μ is a bivariate σ -finite measure, the locally most powerful test for $H: \phi = 0$ against $\phi < 0$ is $I[x < k]$ where k is chosen to ensure level α . This corresponds to the third test mentioned above. Since also β_1 and β_2 belong to the tests whose power functions under $\phi < 0$ may be evaluated by measures of the form (8),

$$\beta_i(\phi) < \beta_3(\phi) \quad \text{for } i = 1, 2 \quad (9)$$

when ϕ is close enough to zero. In terms of the sequence of tests, $\delta_{i,T}$ for $i = 1, 2, 3$, (7) and (9) take, of course, the form:

Proposition 2 *Let δ_{1T}, δ_{2T} and δ_{3T} denote the tests described above such that $\lim E_{0,T}\delta_{iT} = \alpha$ for $i = 1, 2, 3$ and $\alpha < 1/2$. Then there exists $\phi_1 < \phi_0 < 0$ such that $\lim E_{\phi,T}\delta_{3T} < \lim E_{\phi,T}\delta_{1T} < \lim E_{\phi,T}\delta_{2T}$ when $\phi < \phi_1$ and $\lim E_{\phi,T}\delta_{iT} < \lim E_{\phi,T}\delta_{3T}$ for $i = 1, 2$ when $\phi_0 < \phi < 0$.*

25.4 REFERENCES

- Cox, D. D. & Llatas, I. (1991), 'Maximum likelihood type estimation for nearly autoregressive time series', *Annals of Statistics* **19**, 1109–1128.
- Dickey, D. A. & Fuller, W. A. (1979), 'Distribution of the estimators for autoregressive time series with a unit root', *Journal of the American Statistical Association* **74**, 427–431.
- Hájek, J. & Šidák, Z. (1967), *Theory of Rank Tests*, Academic Press.
(Also published by Academia, the Publishing House of the Czechoslovak Academy of Sciences, Prague.).
- Jacod, J. & Shiryaev, A. N. (1987), *Limit Theorems for Stochastic Processes*, Springer-Verlag, Berlin.
- Jeganathan, P. (1995), 'Some aspects of asymptotic theory with applications to time series models', *Econometric Theory* **11**, 818–887.
- Kahn, J. A. & Ogaki, M. (1990), 'A chi-square test for a unit root', *Economics Letters* **34**, 37–42.
- Phillips, P. C. B. (1987), 'Towards a unified theory for autoregression', *Biometrika* **74**, 535–547.
- Phillips, P. C. B. (1989), 'Partially identified econometric models', *Econometric Theory* **55**, 181–240.
- Swensen, A. R. (1993), 'A note on asymptotic power calculations in nearly nonstationary time series', *Econometric Theory* **9**, 659–667.

26

More Optimality Properties of the Sequential Probability Ratio Test

E. Torgersen¹

ABSTRACT Consider the problem of testing sequentially the null hypothesis “ $\theta = 0$ ” against the alternative “ $\theta = 1$ ” on the basis of i.i.d. potentially observable variables X_1, X_2, \dots . Let N be a stopping rule admitting a test based on (X_1, \dots, X_N) having probabilities of errors α_0 and α_1 . Then the Hellinger transform of (X_1, \dots, X_N) is at most equal to that of (X_1, \dots, X_{N^*}) where N^* is the stopping rule of a sequential probability ratio test δ^* having the same probabilities of errors. In particular the Hellinger distance between the distributions of (X_1, \dots, X_N) under $\theta = 0$ and $\theta = 1$ is at least equal to the same distance for (X_1, \dots, X_{N^*}) . This remains so if the Hellinger distance is replaced by the statistical distance and provided the number 1 is not outside the stopping bounds.

26.1 Introduction

Exact results in sequential analysis are often difficult to obtain due to the possibilities of excesses over the stopping boundaries. However arguing as if there were no excesses not only yields useful approximations but also indicates sharp results which may be fully established by more refined methods. Thus the approximations to the stopping boundaries given in Wald (1947) would have been exact, which they are not, and the optimality property of the SPRT would have been a simple consequence of the fact that the expected number of observations is a monotonically increasing functional of the information provided by the sample.

Of course there are excesses over the stopping boundaries and finding a generally acceptable proof of Wald’s conjecture was considered a challenge at that time. Among the several contributors were Arrow, Blackwell & Girshick (1949), Wald & Wolfowitz (1948), Matthes (1963) and Le Cam (material appearing in Section 3.12 of the first edition of Lehmann 1959), who provided an important link in the admissibility part of the proof.

¹University of Oslo

Another simple consequence of disregarding the excesses over the stopping boundary would have been that the statistical distance between the distributions of the sample obtained by the stopping rule of a SPRT minimizes the same distance over all stopping rules permitting tests having the same (or smaller) probabilities of errors.

However, in this case it is fairly simple to see that it actually is permissible to disregard these excesses, provided the number 1 is contained in the continuation interval. By the bimodality of the laws of likelihoods determined by the sample of the stopping rule of a SPRT and by the Neyman-Pearson lemma it follows that *the statistical distance between the distributions of the sample is exactly as it would have been if there were no excesses over the boundaries*. If another stopping rule permits a test with the same probabilities of errors, then the statistical distance between the distributions of the sample is at least the same as the one given by the double dichotomy provided by the testing criterion. The latter quantity however is also equal to the one provided by the probabilities of errors and thus, in turn, equal to the same distance for the SPRT.

The similar conjecture for the Hellinger distance is also true and, in fact, is valid without any restrictions on the stopping boundaries.

We shall derive this as a consequence of the more general fact that if a sequential stopping rule admits a test with prescribed probabilities of errors then the Hellinger transform of the sample never exceeds the Hellinger transform of the sample provided by any SPRT with at least as large probabilities of errors.

Now the expected number of observations for any stopping rule which does not proceed beyond the appearance of a vanishing likelihood is monotonically determined by the local behaviour of the Hellinger transform (of the observed sample) at the one point distribution at the parameter under consideration. (A proof is provided in the next section). Thus *optimality* in terms of *minimality* of expected number of observations is implied by *optimality* in terms of *maximality* of Hellinger transforms.

We shall argue this quite traditionally by dynamic programming. By Weiss (1953) the asserted optimality is equivalent to the assertion that the SPRT's are the solutions to an appropriate dynamic programming problem. This paper of Weiss was followed by Le Cam (1954) who simplified and generalized the admissibility arguments.

We shall here follow the path of Weiss (1953) who showed that it suffices to consider the expected number of observations for a fixed underlying distribution. The corresponding programming problem is neatly described in Lorden (1980). Attaching weights to the probabilities of errors this setup is more in terms of Lagrange multipliers than in terms of prior distributions. Several of the arguments in Lorden's proof carry over to situations involving other cost functions than the one given by constant cost per observation.

Before proceeding to the mathematics it should be noted that throughout we shall omit the qualifications "almost surely" and "almost everywhere".

26.2 Preliminaries

Let X_1, X_2, \dots be independent, identically distributed (i.i.d) variables each having distribution P_θ , where P_θ is a probability measure on some measurable space $(\mathcal{X}, \mathcal{A})$. We shall here assume that $\theta = 0$ or $\theta = 1$ and permit ourselves to use P_θ and E_θ as general notations for, respectively, probability under θ and expectation under θ . We shall assume throughout that $P_0 \neq P_1$.

Choose $P_0 + P_1$ maximal versions $g = dP_1/dP_0$ and $f = dP_0/dP_1$. Then $f = 1/g$ and $g = 1/f$. Assertions involving g become assertions involving f , by symmetry.

Define the likelihood process $(Z_0, Z_1, \dots, Z_\infty)$ by putting $Z_0 = 1$, $Z_n = g(X_1) \dots g(X_n)$, for $n = 1, 2, \dots$, and $Z_\infty = 0$. Thus Z_n converges under P_0 to Z_∞ as $n \rightarrow \infty$.

We shall need a few facts about Hellinger transforms of samples obtained by stopping rules. (These are particular cases of formulas in Greenshtein & Torgersen 1993).

Let N be a stopping rule and let H_N be the Hellinger transform of the sample (X_1, \dots, X_N) . We shall relate H_N to the Hellinger transform $H = H_1$ of a single observation. Thus, for $0 \leq t \leq 1$:

$$H(t) = E_0 Z_1^t \quad (1)$$

and

$$H_N(t) = E_0 Z_N^t. \quad (2)$$

Using the fact that $Z_{n+1} = Z_n g(X_{n+1})$ we find:

$$\begin{aligned} E_0(Z_n^t - Z_{n+1}^t)I_{[N>n]} &= E_0(Z_n^t - E_0(Z_{n+1}^t | X_1, \dots, X_n))I_{[N>n]} \\ &= E_0(Z_n^t - Z_n^t H(t))I_{[N>n]} \\ &= (1 - H(t))E_0 Z_n^t I_{[N>n]}. \end{aligned}$$

It follows that, for any $t \in [0, 1]$ and any bounded stopping time N ,

$$\begin{aligned} 1 - H_N(t) &= E_0(1 - Z_N^t) \\ &= E_0 \sum_n (1 - Z_n^t)(I_{[N>n-1]} - I_{[N>n]}) \\ &= E_0 \sum_n (Z_n^t - Z_{n+1}^t)I_{[N>n]} \\ &= (1 - H(t))E_0(1 + Z_1^t + \dots + Z_{N-1}^t). \end{aligned}$$

By uniform integrability of Z_N^t , for varying stopping rules and for fixed $t \in]0, 1[$, and by Fatou's Lemma we find by approximation with bounded stopping rules that:

$$\frac{1 - H_N(t)}{1 - H(t)} = E_0(1 + Z_1^t + \dots + Z_{N-1}^t)$$

for any stopping rule and any $t \in]0, 1[$. Letting $t \rightarrow 0$ this yields:

$$E_0(N \wedge N_0) = \lim_{t \rightarrow 0} (1 - H_N(t))/(1 - H(t)) \quad (3)$$

where N_0 is the smallest n such that $g(X_n) = 0$. By symmetry

$$E_1(N \wedge N_1) = \lim_{t \rightarrow 1} (1 - H_N(t))/(1 - H(t)) = E_0(1 + \dots + Z_{N-1}) \quad (4)$$

where N_1 is the smallest n such that $f(X_n) = 0$.

26.3 A dynamic programming problem

The dynamic programming problem to be considered involves constants t , u and v where t may be any number in $[0, 1]$ while u and v may be any positive numbers. It is defined as follows: If you stop without observing, then your payoff is $Y_0 = u \wedge v$. If you stop at the n th stage then your payoff is $Y_n = 1 + Z_1^t + \dots + Z_{n-1}^t + (u \wedge (vZ_n))$; $n = 1, 2, \dots$. If you proceed to infinity then your payoff is $Y_\infty = 1 + Z_1^t + Z_2^t + \dots$.

The payoff Y_n at the n th stage consists of two parts, a cost part due to a cost of Z_{i-1}^t of observing X_i , $i = 0, 1, \dots$, and a “risk” part $u \wedge (vZ_n)$ derived from an opportunity to choose the smallest of the constant “risk” u and the stochastic “risk” vZ_n .

If $t = 0$ then we have, except for a sufficiency reduction, the setup used by Lorden (1980).

Our task is, for varying u and v , to characterize the optimal stopping rules N for minimizing expected payoff $E_0 Y_N$. Say that a stopping rule N is a Wald stopping rule if there are constants $0 < B \leq A < \infty$ so that for $n = 0, 1, \dots, \infty$:

$$[N > n] \subseteq [B \leq Z_n \leq A] \text{ and } [N = n] \subseteq [Z_n \leq B] \cup [Z_n \geq A]$$

If $t = 0$ then by Lorden (1980) a stopping rule is optimal for some $u, v > 0$ if and only if it is a Wald stopping rule for some (B, A) such that $0 < B \leq 1 \leq A < \infty$. If still $t = 0$ but stopping rules are required to provide at least one observation then a stopping rule is optimal for some $u, v > 0$ if and only if it is a Wald stopping rule.

This statement is unaltered if we replace $E_0 N$ with $E_0(N \wedge N_0)$. Invoking symmetry we obtain the same statements for $t = 1$.

The contribution of this paper is to show that the above results are valid for all $t \in [0, 1]$, and we state this as the following theorem:

Theorem 1 *The above assertions holds for all $t \in [0, 1]$.*

Proof: By the general theory of optimal stopping, see Chow, Robbins & Siegmund (1971) or Ferguson (1992), optimal stopping rules exist and may

conveniently be described in terms of the variables:

$$V_n = \text{essinf}_{N \geq n} E_0(Y_N | X_1, \dots, X_n), n = 0, 1, \dots$$

It may be checked that the versions of V_n , for varying u and v , may be specified jointly measurable in (X_1, \dots, X_n, u, v) . Put $R_n(u, v) = E_0 V_n$, $n = 0, 1, \dots$ so that, by the existence of a stopping rule achieving the essinf above:

$$R_n(u, v) = \inf_{N \geq n} E_0 Y_N.$$

The fundamental equation of this programming problem is:

$$V_n = Y_n \wedge E_0(V_{n+1} | X_1, \dots, X_n).$$

In particular

$$R_0(u, v) = u \wedge v \wedge R_1(u, v) \quad (5)$$

A stopping rule N is optimal if and only if, for $n = 0, 1, \dots$,

$$[N > n] \subseteq [Y_n \geq E_0(V_{n+1} | X_1, \dots, X_n)] \quad (6)$$

and

$$[N = n] \subseteq [Y_n \leq E_0(V_{n+1} | X_1, \dots, X_n)] \quad (7)$$

In particular we may stop at the smallest $N^* = n$ such that $V_n = Y_n$. If the V_n are specified jointly measurable, as mentioned above, then N^* becomes jointly measurable in the same sense.

In our case we find for $N \geq n$ that

$$\begin{aligned} Y_N &= 1 + Z_1^t + \dots + Z_{n-1}^t \\ &\quad + Z_n^t \left(1 + X_{n+1}^t + \dots + X_N^t + \frac{u}{Z_n^t} \wedge \left(\frac{vZ_n}{Z_n^t} X_{n+1}, \dots, X_N \right) \right). \end{aligned}$$

Applying N^* with u and v replaced by, respectively u/Z_n^t and vZ_n/Z_n^t and with X_1, X_2, \dots replaced by X_{n+1}, X_{n+2}, \dots we find that:

$$V_n = 1 + Z_1^t + \dots + Z_{n-1}^t + Z_n^t R_0(u/Z_n^t, vZ_n/Z_n^t). \quad (8)$$

Similarly, since optimal rules exist,

$$\begin{aligned} E_0(V_{n+1} | X_1, \dots, X_n) &= \text{essinf}_{N \geq n+1} E_0(Y_N | X_1, \dots, X_n) \\ &= 1 + Z_1^t + \dots + Z_{n-1}^t + Z_n^t R_1(u/Z_n^t, vZ_n/Z_n^t) \end{aligned} \quad (9)$$

In particular $EV_1 = \inf_{N \geq 1} EY_N = R_1(u, v)$.

From (5)–(7) we find for an optimal rule that

$$[N > 0] \subseteq [u \wedge v \geq R_1(u, v)] \quad (10)$$

and

$$[N = 0] \subseteq [u \wedge v \leq R_1(u, v)] \quad (11)$$

By (5)–(9) we find for $n = 0, 1, 2, \dots$ that

$$[N > n] \subseteq \left[\frac{u}{Z_n^t} \wedge \frac{vZ_n}{Z_n^t} \geq R_1\left(\frac{u}{Z_n^t}, \frac{vZ_n}{Z_n^t}\right) \right] \quad (12)$$

and

$$[N = n] \subseteq \left[\frac{u}{Z_n^t} \wedge \frac{vZ_n}{Z_n^t} \leq R_1\left(\frac{u}{Z_n^t}, \frac{vZ_n}{Z_n^t}\right) \right] \quad (13)$$

Thus we are led to consider the complement $I_{u,v}$ of the “forced stopping region” i.e. the set:

$$I_{u,v} = \left\{ z : \frac{u}{z^t} \wedge \frac{vz}{z^t} \geq R_1\left(\frac{u}{z^t}, \frac{vz}{z^t}\right) \right\} = \{z : u \wedge (vz) \geq F(z)\}$$

where $F(z) = z^t R_1(u/z^t, vz/z^t)$.

Observe next, by considering stopping at $N \equiv 1$ that

$$1 \leq R_1(u, v) \leq 1 + E_0(u \wedge (vZ_1)) \leq 1 + (u \wedge v) \quad (14)$$

Thus $z^t \leq F(z) \leq z^t + [u \wedge (vz)]$. It follows that $\lim_{z \rightarrow 0} F(z) = 0$ or = 1 as $t > 0$ or $t = 0$. Furthermore, by the identity

$$F(z) \equiv \inf_{N \geq 1} E_0((1 + \dots + Z_{N-1}^t)z^t + (u \wedge (vzZ_N))) ,$$

we see that F is concave, monotonically increasing (strictly increasing if $t > 0$) and that $\lim_{z \rightarrow \infty} F(z) = \sup_z F(z) = \infty$ when $t > 0$ while $\lim_{z \rightarrow \infty} F(z) = \sup_z F(z) \in [1, 1 + u]$ when $t = 0$.

Considering the graphs of the concave functions F and $z \rightarrow u \wedge (vz)$ we infer that $u/v \in I_{u,v}$ whenever this set is not empty and if so then $I_{u,v}$ is the closed interval $[B, A]$ for uniquely determined numbers $B \leq A$.

Consider next the function $w \rightarrow R_1(w, w) - w$. As this function is concave and $\rightarrow 1$ or $\rightarrow -\infty$ as $w \rightarrow 0$ or $w \rightarrow \infty$ we conclude from (14) that there is a unique $w_0 \geq 1$ (> 1 if P_0 and P_1 are not disjoint) so that $R_1(w_0, w_0) = w_0$. (Although most of the other quantities here depend on t the number $w_0 = 2/\|P_1 - P_0\|$ does not depend on t and is the same number as in Lorden's paper.) The condition for obtaining a non empty set $I_{u,v}$ was that $F(u/v) \leq u$ and this may be written $u^{1-t}v^t \geq R_1(u^{1-t}v^t, u^{1-t}v^t)$. Thus $I_{u,v}$ is nonempty if and only if (u, v) belongs to the convex set

$$H = \{(u, v) : u^{1-t}v^t \geq w_0\}.$$

By (10)–(11) the condition for being allowed to observe X_1 is that $u \wedge v \geq R_1(u, v)$ i.e. that $F(1) \leq u \wedge v$ which is just the condition that $1 \in I_{u,v}$ i.e. that $B \leq 1 \leq A$

Altogether this proves that optimal solutions are of the described form. It remains to show that, by varying u and v in the convex set H , we may obtain any Wald stopping rule as an optimal solution. This may be accomplished by first showing that theorem 1 in Lorden (1980) without much ado carries over to our situations. This result ensures the existence of positive, continuous, concave and nondecreasing functions U and V such that

$$\operatorname{sgn}(R_1(u, v) - v) = \operatorname{sgn}(V(u) - v) \quad (15)$$

$$\operatorname{sgn}(R_1(u, v) - u) = \operatorname{sgn}(U(v) - u) \quad (16)$$

and

$$\begin{aligned} \operatorname{sgn}(R_1(w, w) - w) &= \operatorname{sgn}(w_0 - w) \\ &= \operatorname{sgn}(V(w) - w) = \operatorname{sgn}(U(w) - w) \end{aligned} \quad (17)$$

where the function V is bounded above when $t < 1$ while U is bounded above when $t > 0$. Note that the first equality in (17) is a direct consequence of our definition of w_0 . The remaining two equalities in (17) are clearly satisfied by any functions U and V satisfying (15) and (16). Noting, for any given u , that the function $v \rightarrow R_1(u, v) - v$ is concave and $\rightarrow 1$ or $\rightarrow -\infty$ as $v \rightarrow 0$ or $v \rightarrow \infty$ we see that there is a unique number $v = V(u)$ satisfying (15) for any u . In other words V is the unique function satisfying the identity $R_1(u, V(u)) \equiv u$. Likewise there is a unique function U satisfying the identity $R_1(U(v), v) \equiv v$ and this function satisfies (16). The concavity of R_1 implies that the set

$$\{(u, v) : v \leq V(u)\} = \{v : v \leq R_1(u, v)\}$$

is convex and thus that V is concave. It follows that if we put $V(0) = 1$ then V becomes a continuous monotonically increasing concave function on $[0, \infty[$. The assertions for U is argued likewise. If $0 < t < 1$ then

$$R_1(u, v) \leq E_0(1 + Z_1^t + Z_2^t + \dots) = 1 + H(t) + H(t)^2 + \dots = (1 - H(t))^{-1},$$

so that V and U are both bounded above by the constant $(1 - H(t))^{-1}$.

Let \tilde{N} be the smallest n such that $Z_n < 1/2$. Then $E_0 \tilde{N} < \infty$. If $t = 0$ this shows that

$$R_1(u, v) \leq E_0 \tilde{N} + E(u \wedge (vZ_{\tilde{N}})) \leq E_0 \tilde{N} + \frac{1}{2}v \leq v$$

when $v \geq 2E_0 \tilde{N}$ so that V is bounded by $2E_0 \tilde{N}$ in this case. Likewise, or by symmetry, it may be argued that U is bounded when $t = 1$. From (14) we infer that:

$$1 \leq U(v) \leq 1 + v \quad \text{and} \quad 1 \leq V(u) \leq 1 + u \quad (18)$$

Equations (15)–(17) were the basis for establishing the existence of numbers u, v yielding a given interval $[B, A]$. If $0 < t < 1$ additional arguments are needed. The task is to show that there to any numbers B, A such that $0 < B \leq A < \infty$ there is a point $(u, v) \in H$ such that

$$R_1\left(\frac{u}{z^t}, \frac{vz}{z^t}\right) \leq \frac{u}{z^t} \wedge \frac{vz}{z^t} \quad (19)$$

if and only if $z \in [B, A]$. In view of (15)–(16) and (19) this is equivalent to the simultaneous validity of the inequalities:

$$v \geq z^{t-1}V\left(\frac{u}{z^t}\right) \quad (20)$$

and

$$u \geq z^tU\left(\frac{vz}{z^t}\right) \quad (21)$$

As the functions $z \rightarrow z^{t-1}V(u/z)$ and $z \rightarrow z^tU(vz^{1-t})$ are, respectively, monotonically decreasing (strict if $t < 1$) and monotonically increasing (strict if $t > 0$) we infer that (20) is equivalent to

$$z \geq B \quad (22)$$

and that (21) is equivalent to

$$z \leq A. \quad (23)$$

Furthermore A and B are the unique solutions of the equations:

$$v = B^{t-1}V\left(\frac{u}{B^t}\right) \quad (24)$$

and

$$u = A^tU(vA/A^t). \quad (25)$$

Choose now a fixed number $B > 0$. That imposes the condition (24) on v . Thus $A = A(u)$ becomes a function of u . What are the possible values of this function? Putting $u = u_* = B^t w_0$ we find that

$$\begin{aligned} v = v_* &= B^{t-1}V(u_*/B^t) = B^{t-1}V(w_0) \\ &= B^{t-1}w_0. \end{aligned}$$

In particular $u_*^{1-t}v_*^t = w_0$ so that $(u_*, v_*) \in H$ and $B^tU(v_*B/B^t) = B^tU(w_0) = B^t w_0 = u_*$ so that $A = B$ satisfies (18). Thus $A(u_*) = B$ so that B is a possible value of the function $u \rightarrow A(u)$. Note further that $A(u)$ is well defined whenever $u \geq u_*$ since then $u^{1-t}v^t \geq w_0$ i.e. $(u, v) \in H$. By (25) this implies that U is unbounded when $t = 0$.

If $u_0 > u_*$ and $u \in [u_*, u_0]$ then, by (25) and (18), $u_0 \geq A^t$ for all t while $u_0 \geq U(vA)$ when $t = 0$. It follows, since U is unbounded for $t = 0$,

that A is bounded on bounded sets in any case. Thus, by the determining equation (25), the function A is continuous on $[u_*, \infty[$.

Using (25) and (18) once more we find that $u \leq A^t + vA \leq A^t + \bar{v}A$ where $\bar{v} = B^{t-1} \sup_u V(u)$. If $t < 1$ then $\bar{v} < \infty$ and we conclude that $A(u) \rightarrow \infty$ as $u \rightarrow \infty$. If $t = 1$ then $u = AU(v) \leq A \sup_v U(v) < \infty$ so that $\lim_{u \rightarrow \infty} A(u) = \infty$ in any case.

By the intermediate value theorem for continuous functions we conclude that the function $u \rightarrow A(u)$ passes through all values in $[B, \infty[$ as u passes from u_* to ∞ .

Altogether this shows that if we restrict attention to stopping rules $N \geq 1$ then any Wald stopping rule (B, A) is obtainable. Omitting this restriction on N we find that any stopping rule (B, A) with $0 < B \leq 1 \leq A < \infty$ is obtainable. \square

26.4 Optimality of the sequential probability ratio test

Consider the problem of testing the null hypotheses “ $\theta = 0$ ” against the alternative “ $\theta = 1$ ”. The performance function of a test δ is given by the probabilities of errors:

$$\alpha_0(\delta) = P_0(\text{claiming } \theta = 1)$$

and

$$\alpha_1(\delta) = P_1(\text{not claiming } \theta = 1).$$

Let N^* be a Wald stopping rule with stopping boundaries B and A . A *sequential probability ratio test (SPRT)* based on (X_1, \dots, X_{N^*}) is a test δ^* which rejects the null hypothesis whenever $Z_{N^*} > B$ and which do not reject the null hypothesis when $Z_{N^*} < A$. (Thus if $B = A$ then any test based on X_1, \dots, X_{N^*} qualifies). It follows then by the Neyman-Pearson Lemma that δ^* is the most powerful test based on (X_1, \dots, X_{N^*}) and has significance level $\alpha_0(\delta^*)$.

The deduction of the asserted optimality of a SPRT may now be argued from Theorem 1 and for general $t \in [0, 1]$ just as it is usually done when $t = 0$.

Theorem 2 *Let δ^* be a SPRT based on a Wald stopping rule N^* . Assume that the stopping rule N admits a test δ based on (X_1, \dots, X_N) such that $\alpha_0(\delta) \leq \alpha_0(\delta^*)$ and $\alpha_1(\delta) \leq \alpha_1(\delta^*)$. Then $H_N(t) \leq H_{N^*}(t)$, $0 < t < 1$, $E_\theta N \geq E_\theta N^*$; $\theta = 0, 1$ with strict inequalities everywhere provided that either $\alpha_0(\delta) < \alpha_0(\delta^*)$ or $\alpha_1(\delta) < \alpha_1(\delta^*)$.*

Remark Interpretations in terms of parameter dependent cost per observation follow by writing:

$$\begin{aligned} \frac{1 - H_N(t)}{1 - H(t)} &= E_0(1 + g(X_1)^t + \dots + g(X_1)^t \dots g(X_{N-1})^t) \\ &= E_1(1 + f(X_1)^{1-t} + \dots + f(X_1)^{1-t} \dots f(X_{N-1})^{1-t}) \end{aligned}$$

Although these cost functions increase with the number of observations, they can, for $0 < t < 1$, never exceed an integrable, and thus finite, bound.

Remark From the point of view of this paper it might be more satisfactory to replace the inequalities for the probabilities of errors by the more general assertion that the double dichotomy

$$\begin{pmatrix} 1 - \alpha_0(\delta), & \alpha_0(\delta) \\ \alpha_1(\delta), & 1 - \alpha_1(\delta) \end{pmatrix}$$

is at least as informative as the double dichotomy

$$\begin{pmatrix} 1 - \alpha_0(\delta^*), & \alpha_0(\delta^*) \\ \alpha_1(\delta^*), & 1 - \alpha_1(\delta^*) \end{pmatrix}.$$

However if this is so then the test δ may be replaced by another satisfying the required inequalities.

Proof: Let $t \in [0, 1]$ and let B and A be the stopping boundaries of N^* . By Theorem 1 there are positive constants u and v such that

$$E_0(1 + \dots + Z_{N-1}^t + u \wedge (vZ_N)) \geq E_0(1 + \dots + Z_{N^*-1}^t + u \wedge (vZ_{N^*})).$$

As

$$E_0(u \wedge (vZ_N)) \leq E_0(u\delta + vZ_N(1 - \delta)) \leq u\alpha_0(\delta) + v\alpha_1(\delta)$$

while, since $B \leq u/v \leq A$,

$$E_0(u \wedge (vZ_{N^*})) = E_0(u\delta^* + vZ_{N^*}(1 - \delta^*)) = u\alpha_0(\delta^*) + v\alpha_1(\delta^*)$$

it follows that $E_0(1 + \dots + Z_{N-1}^t) + u\alpha_0(\delta) + v\alpha_1(\delta) \geq$ the same expression for δ^* . Hence $E_0(1 + \dots + Z_{N-1}^t) \geq E_0(1 + \dots + Z_{N^*-1}^t)$. \square

Say that an experiment \mathcal{E} is at least as informative as another experiment \mathcal{F} for the Hellinger ordering if the Hellinger transform of \mathcal{E} is everywhere at most equal to the Hellinger transform of \mathcal{F} . In terms of this ordering we obtain the following corollary.

Corollary 2 Consider for any stopping rule N the experiment \mathcal{D}_N obtained by observing X_1, \dots, X_N . Let δ^* be a SPRT based on a Wald stopping rule N^* . Then \mathcal{D}_{N^*} is the greatest lower bound for the Hellinger ordering among all experiments \mathcal{D}_N which are (over all) at least as informative as the double dichotomy

$$\begin{pmatrix} 1 - \alpha_0(\delta^*), & \alpha_0(\delta^*) \\ \alpha_1(\delta^*), & 1 - \alpha_1(\delta^*) \end{pmatrix}.$$

It would be most interesting to know whether the minimality in the Corollary extends to the overall comparison of experiments.

Whether or not this is so it might be interesting to know the deficiencies, in the sense of Le Cam (1964), of a double dichotomy

$$\begin{pmatrix} 1 - \alpha_0, & \alpha_0 \\ \alpha_1, & 1 - \alpha_1 \end{pmatrix}$$

with respect to the dichotomy provided by the Wald stopping rule admitting a SPRT with probabilities of errors α_0 and α_1 . By Theorem 2 this dichotomy is unique up to equivalence. If feasible these quantities might be the statistically most meaningful way of measuring the amount of excesses beyond the stopping boundaries.

26.5 REFERENCES

- Arrow, N. D., Blackwell, D. & Girshick, M. A. (1949), 'Bayes and minimax solutions of sequential decision problems', *Econometrica* **17**, 213–244.
- Burkholder, D. L. & Wijsman, R. A. (1963), 'Optimum properties and admissibility of sequential tests', *Annals of Mathematical Statistics* **34**, 1–17.
- Chow, Y. S., Robbins, H. & Siegmund, D. (1971), *Great Expectations: The Theory of Optimal Stopping*, Houghton-Mifflin, Boston.
- Ferguson, T. S. (1967), *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press, Boston.
- Ferguson, T. S. (1992), Optimal stopping, Technical report, Mathematics Department, UCLA.
- Greenshtein, E. & Torgersen, E. (1993), Statistical information and expected number of observations for sequential experiments, Technical report, Oslo University.
- Kiefer, J. & Weiss, L. (1957), 'Some properties of generalized sequential probability ratio tests', *Annals of Mathematical Statistics* **28**, 57–75.
- Le Cam, L. (1954), 'Note on a theorem of Lionel Weiss', *Annals of Mathematical Statistics* **25**, 791–794.
- Le Cam, L. (1964), 'Sufficiency and approximate sufficiency', *Annals of Mathematical Statistics* **35**, 1419–1455.
- Lehmann, E. L. (1959), *Testing Statistical Hypotheses*, Wiley, New York.
(The cited material, in Section 13.2, does not appear in later editions.).

- Lorden, G. (1980), 'Structure of sequential tests minimizing expected sample size', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **51**, 291–302.
- Matthes, T. K. (1963), 'On the optimality of sequential probability ratio tests', *Annals of Mathematical Statistics* **34**, 18–21.
- Wald, A. (1947), *Sequential Analysis*, Wiley, New York.
- Wald, A. & Wolfowitz, J. (1948), 'Optimum character of the sequential probability ratio test', *Annals of Mathematical Statistics* **19**, 326–339.
- Weiss, L. (1953), 'Testing one simple hypothesis against another', *Annals of Mathematical Statistics* **24**, 273–281.

27

Superefficiency

A. W. van der Vaart¹

ABSTRACT We review the history and several proofs of the famous result of Le Cam that a sequence of estimators can be superefficient on at most a Lebesgue null set.

27.1 Introduction

The method of maximum likelihood as a general method of estimation in statistics was introduced and developed by Fisher (1912, 1922, 1925, 1934). It gained popularity as it appeared that the method automatically produces efficient estimators if the number of observations is large. The concept of asymptotic efficiency was invented by Fisher as early as 1922 roughly in the form as we use it for regular models today: a sequence of statistics is efficient if it tends to a normal distribution with the least possible standard deviation. In the 1930s and 1940s there were many steps in the direction of a rigorous foundation of Fisher's remarkable insights. These consisted both of proofs of the asymptotic normality of maximum likelihood estimators and of obtaining lower bounds for the variance of estimators.

Chapters 32 and 33 of Cramér (1946) give a summary of the state of affairs in the mid 1940s, even though some work carried out in the early war years, notably Wald's, had been unavailable to him. Chapter 32 gives a rigorous proof of what we now know as the Cramér-Rao inequality and next goes on to define the asymptotic efficiency of an estimator as the quotient of the inverse Fisher information and the asymptotic variance. Next Chapter 33 gives a rigorous proof of asymptotic normality of the maximum likelihood estimator, based on work by Dugué (1937).

Cramér defines an estimator sequence to be *asymptotically efficient* if its asymptotic efficiency (the quotient mentioned previously) equals one. Thus combination of the results of the two chapters leads to the correct conclusion that the method of maximum likelihood produces asymptotically efficient estimators, under some regularity conditions on the underlying densities. Apparently the conceptual hole in the definition was not fully recognized until 1951, even though the difficulty must have been clear to

¹ Vrije Universiteit

several authors who had worked on establishing efficiency within restricted classes of estimators.

In 1951 Hodges produced the first example of a *superefficient* estimator sequence: an estimator sequence with efficiency at least one for all θ and more than one for some θ . An abstraction of Hodges' example is the following. Let T_n be a sequence of estimators of a real parameter θ such that the sequence $\sqrt{n}(T_n - \theta)$ converges to some limit distribution if θ is the true parameter, under every θ . If $S_n = T_n 1\{|T_n| > n^{-1/4}\}$, then the probability of the sequence of events $\{T_n = S_n\}$ converges to one under every $\theta \neq 0$, while under $\theta = 0$ the probability of the event $\{S_n = 0\}$ converges to one. In particular, if the first sequence of estimators T_n is asymptotically efficient in the sense of Cramér, then the sequence S_n is superefficient at $\theta = 0$.

Hodges' example revealed a difficulty with the definition of asymptotic efficiency and threw doubt on Fisher's assertion that the maximum likelihood estimator is asymptotically efficient. In this paper we review three lines of approach addressing the matter. They were all initiated by Le Cam. Already in 1952 Le Cam had announced in an abstract to the Annals of Mathematical Statistics that the set of superefficiency can never be larger than a Lebesgue null set. In the next section we review his proof, which appeared in Le Cam (1953). Le Cam's second approach is present in Le Cam (1973) and is based on automatic invariance. We discuss it in Section 3. The third approach combines elements of both papers and is given in Section 4. Particularly in this last section we do not strive for the utmost generality. We hope that simple proofs may help these beautiful results finally find their way into text books and lecture notes.

In the following, *superefficiency* of a sequence of estimators in the locally asymptotically normal case will be understood in the sense that

$$\limsup_{n \rightarrow \infty} E_\theta \ell(\sqrt{n}(T_n - \theta)) \leq \int \ell dN_{0, I_\theta^{-1}},$$

for every θ , with strict inequality for some θ . Here I_θ is the Fisher information matrix and ℓ a given loss function.

Convergence in distribution is denoted \rightsquigarrow and convergence in distribution under a law given by a parameter θ by ${}^\theta \rightsquigarrow$.

27.2 The 1953 proof

Le Cam (1953) started his paper with examples of superefficient estimator sequences. These include Hodges' example, but also estimators that are superefficient on a dense set of parameters. Next he went on to prove that superefficiency can occur only on Lebesgue null sets. The main idea is that the sequence of maximum likelihood estimators is asymptotically Bayes with respect to Lebesgue absolutely continuous priors on the parameter

set. Specifically, let $\hat{\theta}_n$ be the maximum likelihood estimator based on a sample of size n from a density p_θ . Under smoothness conditions on the map $\theta \mapsto p_\theta$ Le Cam showed that

$$(1) \quad \limsup_{n \rightarrow \infty} \int E_\theta \ell(\sqrt{n}(\hat{\theta}_n - \theta)) \pi(\theta) d\theta \leq \liminf_{n \rightarrow \infty} \int E_\theta \ell(\sqrt{n}(T_n - \theta)) \pi(\theta) d\theta,$$

for every sequence of estimators T_n , most prior densities $\pi(\theta)$ and most symmetric, bounded, continuous loss functions ℓ . Since the standardized sequence of maximum likelihood estimators $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal $N(0, I_\theta^{-1})$, the first limsup exists as an ordinary limit. Application of Fatou's lemma immediately yields that

$$\int \left(\int \ell dN_{0, I_\theta^{-1}} - \limsup_{n \rightarrow \infty} E_\theta \ell(\sqrt{n}(T_n - \theta)) \right) \pi(\theta) d\theta \leq 0.$$

Superefficiency of the sequence T_n would imply that the integrand is non-negative. Since it integrates nonpositive it must be zero almost surely under π .

Rigorous proofs of the asymptotic normality of the sequence of maximum likelihood estimators were available, for instance from Cramér (1946). The essential part of the preceding argument was therefore the proof of (1). Le Cam based his proof on a version of the Bernstein-von Mises theorem. Let Θ be a random variable with Lebesgue density π on the parameter set and consider $\prod_{i=1}^n p_\theta(x_i)$ as the conditional density of (X_1, \dots, X_n) given $\Theta = \theta$. Le Cam (1953) proved (under regularity conditions) that, for every θ , with $\|\cdot\|$ denoting the total variation norm,

$$\left\| \mathcal{L}(\sqrt{n}(\Theta - \hat{\theta}_n) | X_1, \dots, X_n) - N(0, I_\theta^{-1}) \right\| \rightarrow 0, \quad \text{a.s. } [P_\theta].$$

This strengthened earlier results by Bernstein (1934) and von Mises (1931) to the point where application towards proving (1) is possible. In the present notation the Bayes risk of T_n can be written

$$\begin{aligned} & \int E_\theta \ell(\sqrt{n}(T_n - \theta)) \pi(\theta) d\theta \\ &= E E \left(\ell(\sqrt{n}(T_n - \hat{\theta}_n)) - \sqrt{n}(\Theta - \hat{\theta}_n) \mid X_1, \dots, X_n \right). \end{aligned}$$

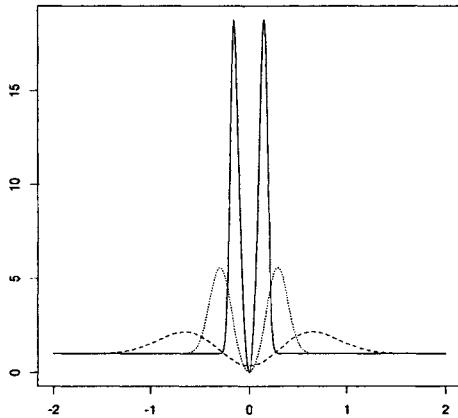
According to the Bernstein-von Mises theorem the conditional expectation in this expression satisfies, setting $\mu_n = \sqrt{n}(T_n - \hat{\theta}_n)$,

$$E \left(\ell(\mu_n - \sqrt{n}(\Theta - \hat{\theta}_n)) \mid X_1, \dots, X_n \right) - \int \ell dN_{-\mu_n, I_{\hat{\theta}_n}^{-1}} \rightarrow 0,$$

almost surely under every θ , hence also under the mixtures $\int P_\theta^\infty \pi(\theta) d\theta$. It is assumed at this point that the loss function ℓ is bounded; otherwise a stronger version of the Bernstein-von Mises theorem would be necessary. For the usual symmetric loss functions the normal expectation in the preceding display decreases if μ_n is replaced by zero. This readily leads to (1).

From today's perspective the references to asymptotic normality and the maximum likelihood estimator are striking. As Le Cam was later to point out neither of the two are essential for the principle of superefficiency. The special role of maximum likelihood estimators was removed in Le Cam (1956), where they were replaced by one-step estimators. Next Le Cam (1960, 1964) abstracted the structure of asymptotically normal problems into the 'local asymptotic normality' condition and finally removed even the latter in Le Cam (1972, 1973).

The use of Bayes estimators is in tune with the statistical paradigm of the 1940s and 1950s. Wald (1950)'s book firmly established statistical decision theory as the basis of statistical reasoning. A main result was that Bayes estimators (or rather their limit points) form a complete class. Wolfowitz (1953) exploited this to explain the impossibility of superefficiency in an informal manner. The preceding argument shows that the risk functions of Bayes estimators are asymptotically equivalent to the risk function of the maximum likelihood estimator. Thus asymptotically the maximum likelihood estimator is the only Bayes estimator. This would establish its asymptotic admissibility, and also its optimality. Only a more precise argument would properly explain the role of sets of Lebesgue measure zero.



Quadratic risk function of the Hodges estimator based on a sample of size 10 (dashed), 100 (dotted) and 1000 (solid) observations from the $N(\theta, 1)$ -distribution.

In the final section of his paper Le Cam (1953) also showed that in the case of one-dimensional parameters superefficient estimators necessarily have undesirable properties. For the Hodges' estimator $T = \bar{X}1\{\lvert\bar{X}\rvert > n^{-1/4}\}$ based on a sample of size n from the $N(\theta, 1)$ -distribution this is illustrated in the Figure, which shows the risk function $\theta \mapsto nE_\theta(T - \theta)^2$ for three different values of n . Le Cam shows that this behaviour is typical: superefficiency at the point θ for a loss function ℓ implies the existence of a sequence $\theta_n \rightarrow \theta$ such that $\liminf E_{\theta_n} \ell(\sqrt{n}(T_n - \theta_n))$ is strictly larger

than $\int \ell dN_{0,1/I_\theta}$. For the extreme case where the asymptotic risk at θ is zero, the liminf is even infinite for a sequence $\theta_n \rightarrow \theta$.

This result may be considered a forerunner of the local asymptotic minimax theorem of Hájek (1972), which states that the maximum risk over a shrinking neighbourhood of θ is asymptotically bounded below by $\int \ell dN_{0,1/I_\theta}$. A much earlier result of this type was obtained by Chernoff (1956), who essentially showed that

$$\lim_{c \rightarrow \infty} \liminf_{n \rightarrow \infty} \sup_{-c < h < c} E_{\theta+h/\sqrt{n}} \left(\sqrt{n}|T_n - \theta - h/\sqrt{n}| \wedge c \right)^2 \geq \frac{1}{I_\theta}.$$

Chernoff's proof is based on a version of the Cramér-Rao inequality, which he attributed to Stein and Rubin. The theorem may have looked somewhat too complicated to gain popularity. Nevertheless Hájek's result, for general locally asymptotically normal models and general loss functions, is now considered the final result in this direction. Hájek wrote:

The proof that local asymptotic minimax implies local asymptotic admissibility was first given by LeCam (1953, Theorem 14). . . . Apparently not many people have studied Le Cam's paper so far as to read this very last theorem, and the present author is indebted to Professor LeCam for giving him the reference.

Not reading to the end of Le Cam's papers became not uncommon in later years. His ideas have been regularly rediscovered.

Le Cam's Theorem 14 about the bad properties of superefficient estimators only applies to one-dimensional estimators. Le Cam commented:

In the case of an r -dimensional parameter the problem becomes more complicated. The difficulties involved are conceptual as well as mathematical.

This is very true: we now know that a similar result is false in dimensions three and up. Only three years after Le Cam's paper, Stein (1956) published his famous paper on estimating a multivariate normal mean. The James-Stein estimator

$$T_n = \bar{X} - (d-2) \frac{\bar{X}}{n\|\bar{X}\|^2}$$

is superefficient at $\theta = 0$ for the loss function $\ell(x) = \|x\|^2$, and it does not behave badly in a neighbourhood of this point (for this loss function).

27.3 Automatic invariance

The second approach to proving the impossibility of superefficiency is based on the remarkable fact that the usual rescaling of the parameter (for in-

stance considering $h = \sqrt{n}(\theta - \theta_0)$ as the parameter, rather than θ) automatically leads to asymptotic equivariance at almost every θ_0 . This idea is put forward in Le Cam (1973) and leads to a completely different proof than the proof in Le Cam (1953). On comparing the two papers the difference in style is also apparent. The main result of Le Cam (1973) asserts shift invariance of limit experiments and is stated within the abstract framework of L - and M -spaces. The important application is only indicated in the second last paragraph:

Toutefois, et pour conclure, mentionnons que la démonstration du résultat de convolution de Hájek (1970) s'étend à tous les cas considérés ici, pourvu qu'elle soit faite par la méthode décrite dans Le Cam (1972).

The application must have been obvious to Le Cam. This would explain that Section 8.4 of Le Cam (1986), which is concerned with the same subject, seems to end without a conclusion regarding superefficiency as well. In this section we present a simplified version of Le Cam's (1973) result, suited to superefficiency.

Asymptotic equivariance subsumes the Hájek regularity property, which was the key requirement for Hájek (1970)'s convolution theorem. He defined an estimator sequence T_n based on n observations from a smooth parametric model to be *regular* at the parameter θ if

$$(2) \quad \sqrt{n}(T_n - \theta - h/\sqrt{n}) \xrightarrow{\theta+h/\sqrt{n}} L_\theta, \quad \text{every } h,$$

for some fixed probability distribution L_θ . The independence of L_θ of h is the crucial feature of regularity. Hájek's (1970) convolution theorem states that in this situation the limiting distribution L_θ is a convolution of the type

$$L_\theta = N(0, I_\theta^{-1}) * M_\theta.$$

This certainly implies that the covariance matrix of L_θ is bounded below by the inverse I_θ^{-1} of the Fisher information matrix.

The 'local uniformity' in the weak convergence required by Hájek regularity looks not too unnatural, though on closer inspection not all interesting estimators turn out to be regular. Shrinkage estimators are not regular at the shrinkage point; estimators that are truncated to a parameter set (such as $\bar{X} \vee 0$ if a mean is known to be positive) are not regular at the boundary of the parameter set. In these examples the set of points of irregularity is very small. It turns out that this is necessarily the case. Below we shall show that the regularity holds automatically at almost all parameter points for any estimator sequence such that $\sqrt{n}(T_n - \theta)$ has a limiting distribution under every θ .

This 'automatic regularity' is the key connection to superefficiency, for it follows that the limit distributions L_θ are convolutions for almost every θ . In particular the asymptotic covariance matrix is bounded below by the inverse Fisher information matrix for almost every θ . This constitutes a

modern proof of the fact that superefficiency can occur only on Lebesgue null sets.

Hájek's proof of the convolution theorem is based on delicate calculations using the special character of local asymptotic normality. Le Cam (1972)'s theory of limiting experiments puts the result in a very general framework. Not only does it offer much insight in the Gaussian situation, it also allows superefficiency statements in many other situations. We shall now carry out the preceding steps in more detail and in much greater generality.

From the more general point of view regularity is better described as (local) asymptotic equivariance. My favourite version (Van der Vaart (1991)) of Le Cam's result is as follows. A sequence of experiments (or statistical models) is said to *converge* to a limit if the marginals of the likelihood ratio processes converge in distribution to the corresponding marginals of the likelihood ratio processes in the limit experiment. The precise definition of convergence is not important for this paper: convergence of experiments only enters as a condition of the following theorem and through statements regarding concrete examples, which are not proven here.

(3) **Proposition** *Let the experiments $(\mathcal{X}_n, \mathcal{A}_n, P_{n,h}: h \in H)$ converge to a dominated experiment $(P_h: h \in H)$. Let $\kappa_n: H \mapsto \mathbb{D}$ be maps with values in a Banach space \mathbb{D} such that*

$$r_n(\kappa_n(h) - \kappa_n(h_0)) \rightarrow \dot{\kappa}h - \dot{\kappa}h_0,$$

for some map $\dot{\kappa}: H \mapsto \mathbb{D}$ and linear maps $r_n: \mathbb{D} \mapsto \mathbb{D}$. Let $T_n: \mathcal{X}_n \mapsto \mathbb{D}$ be arbitrary maps with values in \mathbb{D} such that the sequence $r_n(T_n - \kappa_n(h))$ converges in distribution under h , for every h , to a probability distribution that is supported on a fixed separable, Borel measurable subset of \mathbb{D} . Then there exists a randomized estimator T in the limit experiment such that $r_n(T_n - \kappa_n(h))$ converges under h to $T - \dot{\kappa}h$, for every h .

In this proposition a randomized estimator is a measurable map $T: \mathcal{X} \times [0, 1] \mapsto \mathbb{D}$ whose law is to be calculated under the product of P_h and the uniform law. Thus $T = T(X, U)$ is based on an observation X in the limit experiment and an independent uniform variable U .

We could call the estimator sequence T_n in the preceding proposition *regular* if the limiting distribution under h of the sequence $r_n(T_n - \kappa_n(h))$ is the same for every h . Then the matching randomized estimator in the limit experiment satisfies

$$\mathcal{L}_h(T - \dot{\kappa}h) = \mathcal{L}_0(T), \quad \text{every } h.$$

This may be expressed as: T is *equivariant-in-law* for estimating the parameter $\dot{\kappa}h$.

Within the context of the proposition 'regularity' has lost its interpretation as a local uniformity requirement. This is recovered when the proposition is applied to 'localized' experiments. For instance, local asymptotic normality of the sequence of experiments $(\mathcal{X}_n, \mathcal{A}_n, P_\theta^n: \theta \in \Theta)$ at θ entails

convergence of the sequence of local experiments to a Gaussian experiment:

$$(\mathcal{X}_n, \mathcal{A}_n, P_{\theta+h/\sqrt{n}}^n : h \in \mathbb{R}^d) \rightarrow (N_d(h, I_\theta^{-1}) : h \in \mathbb{R}^d).$$

Regularity of a given estimator sequence for the functionals $\kappa_n(h) = \theta + h/\sqrt{n}$ in this sequence of localized experiments means exactly Hájek regularity as in (2). The proposition shows that the limit distribution L_θ is the distribution under $h = 0$ of an equivariant-in-law estimator for h in the Gaussian experiment. Hájek's convolution theorem is reproved once it has been shown that every such equivariant-in-law estimator can be decomposed as a sum $X + W$ of two independent variables, where X is $N_d(0, I_\theta^{-1})$ -distributed. In any case the 'best' equivariant-in-law estimator is X itself, so that the covariance of L_θ is not smaller than I_θ^{-1} .

The following theorem shows that estimator sequences in rescaled experiments are automatically (almost) regular, at almost every parameter. The proof of the theorem is based on an extension of a lemma by Bahadur (1964), who used his lemma to rederive Le Cam (1953)'s result for one-dimensional parameters. Denote the d -dimensional Lebesgue measure by λ^d .

(4) **Theorem** Let $(\mathcal{X}_n, \mathcal{A}_n, P_{n,\theta} : \theta \in \Theta)$ be experiments indexed by a measurable subset Θ of \mathbb{R}^d . Let $T_n : \mathcal{X}_n \mapsto \mathbb{D}$ and $\kappa_n : \Theta \mapsto \mathbb{D}$ be maps into a complete metric space such that the maps $\theta \mapsto E_\theta^* f(r_{n,\theta}(T_n - \kappa_n(\theta)))$ are measurable for every Lipschitz function $f : \mathbb{D} \mapsto [0, 1]$. Suppose that

$$r_{n,\theta}(T_n - \kappa_n(\theta)) \xrightarrow{\theta} L_\theta, \quad \lambda^d - \text{a.e. } \theta,$$

for probability distributions L_θ supported on a fixed separable, Borel measurable subset of \mathbb{D} . Then for any matrices $\Gamma_n \rightarrow 0$ there exists a subsequence of $\{n\}$ such that

$$r_{n,\theta+\Gamma_n h}(T_n - \kappa_n(\theta + \Gamma_n h)) \xrightarrow{\theta+\Gamma_n h} L_\theta, \quad \lambda^{2d} - \text{a.e. } (\theta, h),$$

along the subsequence.

(5) **Lemma** For $n \in \mathbf{N}$ let $g_n, g : \mathbb{R}^d \mapsto [0, 1]$ be arbitrary measurable real functions such that

$$g_n \rightarrow g, \quad \lambda^d - \text{a.e.}$$

Then given any sequences of vectors $\gamma_n \rightarrow 0$ and matrices $\Gamma_n \rightarrow 0$ there exists a subsequence of $\{n\}$ such that

$$\begin{aligned} (i) \quad & g_n(\theta + \gamma_n) \rightarrow g(\theta), & \lambda^d - \text{a.e. } \theta, \\ (ii) \quad & g_n(\theta + \Gamma_n h) \rightarrow g(\theta), & \lambda^{2d} - \text{a.e. } (\theta, h), \end{aligned}$$

along the subsequence. If $g_n(\theta + \Gamma_n h_n) - g_n(\theta + \Gamma_n h) \rightarrow 0$ for every sequence $h_n \rightarrow h$, then (ii) is true for every $h \in \mathbb{R}^d$, for almost every θ .

Proof. We prove statement (ii), the proof of (i) being slightly simpler. We may assume without loss of generality that the function g is integrable;

otherwise we first multiply each g_n and g with a suitable, fixed, positive, continuous function. Write p for the standard normal density on \mathbb{R}^d . Then

$$\iint |g(\theta + \Gamma_n h) - g(\theta)| p(\theta) d\theta p(h) dh \rightarrow 0.$$

This follows since the inner integral converges to zero for every h by the L_1 -continuity theorem (e.g. Theorem 8.19 of Wheeden and Zygmund) and next the outer integral converges to zero by the dominated convergence theorem.

If p_n is the density of the $N_d(0, I + \Gamma'_n \Gamma_n)$ -distribution, then

$$\iint |g_n(\theta + \Gamma_n h) - g(\theta + \Gamma_n h)| p(\theta) p(h) d\theta dh = \int |g_n(u) - g(u)| p_n(u) du.$$

The sequence p_n converges in L_1 to the standard normal density. Thus the integral on the right converges to zero by the dominated convergence theorem. Combination with the preceding display shows that the sequence of functions $(\theta, h) \mapsto g_n(\theta + \Gamma_n h) - g(\theta)$ converges to zero in mean, and hence in probability, under the standard normal measure. There exists a subsequence along which it converges to zero almost surely.

In the proof of the theorem abbreviate $r_{n,\theta}(T_n - \kappa_n(\theta))$ to $T_{n,\theta}$. Assume without loss of generality that $\Theta = \mathbb{R}^d$; otherwise fix θ_0 such that $T_{n,\theta_0} \xrightarrow{\theta_0 \rightsquigarrow} L_{\theta_0}$ and let $P_{n,\theta} = P_{n,\theta_0}$ for every θ not in Θ . Let \mathbb{D}_0 be the separable Borel subset of \mathbb{D} on which the limit distributions L_θ concentrate. There exists a countable collection \mathcal{F} of Lipschitz functions $f: \mathbb{D} \mapsto [0, 1]$, depending only on \mathbb{D}_0 , such that weak convergence of a sequence of maps $T_n: \mathcal{X}_n \mapsto \mathbb{D}$ to a Borel measure L on \mathbb{D}_0 is equivalent to $E^* f(T_n) \rightarrow \int f dL$ for every $f \in \mathcal{F}$. Consider the functions

$$g_n(\theta; f) = E_\theta^* f(T_{n,\theta}); \quad g(\theta; f) = \int f dL_\theta.$$

For every fixed f these functions are measurable by assumption and $g_n \rightarrow g$ pointwise. By the lemma there exists a subsequence of $\{n\}$ along which

$$E_{\theta + \Gamma_n h}^* f(T_{n,\theta + \Gamma_n h}) \rightarrow \int f dL_\theta, \quad \lambda^{2d} - \text{a.e.}.$$

The subsequence depends on f , but by a diagonalization scheme we can construct a subsequence for which this is valid for every f in the countable set \mathcal{F} . \square

(6) Example Suppose that for numbers $r_n \rightarrow \infty$ the sequence of variables $r_n(T_n - \theta)$ converges under θ to a limit distribution L_θ for almost every θ . The preceding theorem asserts that for almost every θ there exists a set H_θ with $\lambda^d(H_\theta^c) = 0$ such that

$$r_n(T_n - \theta - h/r_n) \xrightarrow{\theta + h/r_n} L_\theta, \quad \text{every } h \in H_\theta.$$

Thus the sequence T_n is almost Hájek regular, at almost every θ . The first ‘almost’ refers to the set H_θ which is almost all of \mathbb{R}^d . In most situations

this ‘almost’ can be removed from the statement. It is often the case that

$$\|P_{n,\theta+h_n/r_n} - P_{n,\theta+h/r_n}\| \rightarrow 0, \quad \text{every } h_n \rightarrow h.$$

Then the total variation distance between the laws of $T_n - \theta - h/r_n$ under $\theta + h_n/r_n$ and $\theta + h/r_n$ converges to zero as well and in view of the last assertion of Lemma 5 the set H_θ can be taken equal to \mathbb{R}^d . \square

The combination of Theorem 4 and the proposition gives interesting applications far beyond the Gaussian case. The key is that for almost all θ the limit distribution L_θ of an estimator sequence is not ‘better’ than the null distribution of the best equivariant estimator in the limit experiment. Also when the latter cannot be characterized as a convolution, the equivariance implies a lower bound on the risk. The locally asymptotically mixed normal case is well-documented in Jeganathan (1982, 1983). We give two other examples.

(7) **Example** Suppose the problem is to estimate θ based on a sample of size n from the uniform distribution P_θ on the interval $[\theta, \theta + 1]$. The sequence of experiments $(P_{\theta+h/n}^n : h \in \mathbb{R})$ converges for each θ to the experiment consisting of observing a pair with the same distribution as $(V + h, h - W)$ for independent standard exponential variables V and W . If the sequence $n(T_n - \theta)$ converges in distribution to a limit for every θ , then for almost every θ the limit distribution L_θ is the distribution of an equivariant-in-law estimator T based on $(V + h, h - W)$. The best such estimator in terms of bowl-shaped loss functions is $\frac{1}{2}((V + h) + (h - W)) = \frac{1}{2}(V - W) + h$. Its invariant standardized law is the Laplace distribution, so that we may conclude

$$\int \ell dL_\theta \geq \int \ell(x) e^{-2|x|} dx, \quad \lambda^d - \text{a.e. } \theta.$$

In this problem a characterization as a convolution is impossible. This can be seen from the fact that $V + h$ and $\frac{1}{2}(V - W)$ are both equivariant estimators, but their laws have no convolution factor in common. \square

(8) **Example** Suppose the problem is to estimate θ based on a sample of size n from the distribution P_θ with density $p(\cdot - \theta)$ on the real line, where $p(x)$ is differentiable at every $x \notin \{a_1, \dots, a_m\}$ with $\int |p'(x)| dx < \infty$ and has discontinuities at each a_i with $p(a_i-) = 0 < p(a_i+)$. Then the sequence $(P_{\theta-h/n}^n : h \in \mathbb{R})$ converges for each θ to the experiment consisting of observing a single random variable with the same distribution as $V + h$ for a standard exponential variable V with mean $1/\sum p(a_i+)$. Since the limit experiment is a full shift experiment, it admits a convolution theorem. If the sequence $n(T_n - \theta)$ converges in distribution to a limit for every θ , then for almost every θ the limit distribution L_θ contains the distribution of V as a convolution factor. \square

27.4 Superefficiency and loss functions

The combined results of the preceding section give a deep characterization of the limiting distributions of a sequence of estimators, valid at almost every θ . Apart from measurability the only assumption is the mere existence of limiting distributions.

The latter is a fair assumption for this type of result, but what can be said without it? Equation (1) and asymptotic normality of the maximum likelihood estimator show that for any estimator sequence T_n

$$(9) \quad \liminf_{n \rightarrow \infty} \int E_\theta \ell(\sqrt{n}(T_n - \theta)) \pi(\theta) d\theta \geq \iint \ell dN_{0, I_\theta^{-1}} \pi(\theta) d\theta,$$

for most prior densities π . In view of Fatou's lemma this readily gives

$$\limsup_{n \rightarrow \infty} E_\theta \ell(\sqrt{n}(T_n - \theta)) \geq \int \ell dN_{0, I_\theta^{-1}}, \quad \lambda^d - \text{a.e.}$$

This cannot be strengthened by replacing the limsup by a liminf, basically because the sequence $\{n\}$ has too many subsequences.

(10) **Example** For the parameter set equal to the unit interval and $k \in \mathbb{N}$ define estimators $T_{2^k+i} = i2^{-k}$ for $i = 1, \dots, 2^k$. Given a parameter θ define a subsequence of $\{n\}$ by $n_k = 2^k + i_k$, for i_k the integer such that $(i_k - 1)2^{-k} < \theta \leq i_k 2^{-k}$. Then $\sqrt{n_k}|T_{n_k} - \theta| \leq \sqrt{2}2^{-k/2}$, whence $\liminf E_\theta \ell(\sqrt{n}(T_n - \theta)) = \ell(0)$ for every symmetric loss function which is continuous at zero, and every θ . \square

Le Cam (1953) established (9) under regularity conditions somewhat better than those given in Cramér's book. His result was improved in his later papers and also in Strasser (1978). The integral $\int \ell dN_{0, I_\theta^{-1}}$ is the minimax risk for estimating h based on a single observation from the $N(h, I_\theta^{-1})$ -distribution, the limit of the local experiments around θ . The following theorem establishes a relationship between the limiting pointwise risk and the minimax risk in the local limit experiments in great generality, under very mild conditions. It is assumed that the sequence of experiments $(P_{n,\theta+h/r_n}: h \in \mathbb{R}^d)$ converges for almost every θ to a dominated experiment \mathcal{E}_θ in which the minimax theorem is valid in the form

$$\sup_P \inf_T \int E_{\theta,h} \ell(T - h) dP(h) = \inf_T \sup_h E_{\theta,h} \ell(T - h).$$

Here the first supremum is taken over all probability measures with compact support and the infimum over all randomized estimators in \mathcal{E}_θ . This is generally the case, perhaps under some regularity condition on the loss function. Le Cam has broadened the definition of estimators to ensure that the minimax theorem is always true, but we wish to keep the statement simple. According to Le Cam (1973) the local limit experiments are for almost all θ shift-invariant. In Euclidean shift experiments the minimax risk

is typically obtained as the limit of Bayes risks for a sequence of uniform priors that approach the improper Lebesgue prior.

(11) **Theorem** Let $(P_{n,\theta}: \theta \in \Theta)$ be measurable experiments indexed by an open subset $\Theta \subset \mathbb{R}^d$. Suppose that for almost every θ the local experiments $(P_{n,\theta+h/\tau_n}: h \in \mathbb{R}^d)$ converge to dominated experiments \mathcal{E}_θ in which the minimax theorem holds as mentioned for a given bounded subcompact loss function ℓ . Then for every estimator sequence T_n

$$\limsup_{n \rightarrow \infty} E_\theta \ell(r_n(T_n - \theta)) \geq \inf_T \sup_h E_{\theta,h} \ell(T - h), \quad \lambda^d - \text{a.e. } \theta,$$

where the infimum is taken over all randomized estimators in \mathcal{E}_θ .

Proof. Let π be a probability density on a subset of Θ that is bounded away from the boundary of Θ . Let P be a probability measure with compact support and set $\gamma_n = r_n^{-1}$. Abbreviate $R_n(\theta) = E_\theta \ell(r_n(T_n - \theta))$. By a change of variables

$$\begin{aligned} & \left| \int R_n(\theta) \pi(\theta) d\theta - \iint R_n(\theta + \gamma_n h) dP(h) \pi(\theta) d\theta \right| \\ & \leq \|\ell\|_\infty \iint |\pi(\theta) - \pi(\theta - \gamma_n h)| d\theta dP(h) \rightarrow 0, \end{aligned}$$

by the L_1 -continuity theorem and the dominated convergence theorem. Essentially as a consequence of Proposition 3, applied to a compactification of \mathbb{R}^d ,

$$\liminf_{n \rightarrow \infty} \int E_{\theta+\gamma_n h} \ell(r_n(T_n - \theta - \gamma_n h)) dP(h) \geq \inf_T \int E_{\theta,h} \ell(T - h) dP(h),$$

where the infimum is taken over all randomized estimators T for the parameter $h \in \mathbb{R}^d$ in \mathcal{E}_θ . This is valid for almost every θ . Combination of the preceding displays and Fatou's lemma gives

$$\liminf_{n \rightarrow \infty} \int E_\theta \ell(r_n(T_n - \theta)) \pi(\theta) d\theta \geq \int \inf_T \int E_{\theta,h} \ell(T - h) dP(h) \pi(\theta) d\theta.$$

By assumption there exists for each θ and m a probability measure $P_{\theta,m}$ such that the inner integral is within distance $1/m$ of the minimax risk in \mathcal{E}_θ . The desired conclusion follows by the monotone convergence theorem as $m \rightarrow \infty$. \square

27.5 REFERENCES

- Bahadur, R. R. (1964), 'On Fisher's bound for asymptotic variances', *Annals of Mathematical Statistics* 35, 1545–1552.
 Bernstein, S. (1934), *Theory of Probability*, GTTI, Moscow. (Russian).
 Chernoff, H. (1956), 'Large sample theory: parametric case', *Annals of Mathematical Statistics* 27, 1–22.

- Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press.
- Fisher, R. A. (1912), 'On an absolute criterion for fitting frequency curves', *Messenger of Mathematics* **41**, 155–160.
- Fisher, R. A. (1922), 'On the mathematical foundations of theoretical statistics', *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309–368.
- Fisher, R. A. (1925), 'Theory of statistical estimation', *Proceedings of the Cambridge Philosophical Society* **22**, 700–725.
- Fisher, R. A. (1934), 'Two new properties of mathematical likelihood', *Proceedings of the Royal Society of London, Series A* **144**, 285–307.
- Hájek, J. (1970), 'A characterization of limiting distributions of regular estimators', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **14**, 323–330.
- Hájek, J. (1972), Local asymptotic minimax and admissibility in estimation, in L. Le Cam, J. Neyman & E. L. Scott, eds, 'Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 175–194.
- Jeganathan, P. (1982), 'On the asymptotic theory of estimation when the limit of the log-likelihood ratios is mixed normal', *Sankhyā: The Indian Journal of Statistics, Series A* **44**, 173–212.
- Jeganathan, P. (1983), 'Some asymptotic properties of risk functions when the limit of the experiment is mixed normal', *Sankhyā: The Indian Journal of Statistics, Series A* **45**, 66–87.
- Le Cam, L. (1953), 'On some asymptotic properties of maximum likelihood estimates and related Bayes estimates', *University of California Publications in Statistics* **1**, 277–330.
- Le Cam, L. (1956), On the asymptotic theory of estimation and testing hypotheses, in J. Neyman, ed., 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 129–156.
- Le Cam, L. (1960), 'Locally asymptotically normal families of distributions', *University of California Publications in Statistics* **3**, 37–98.
- Le Cam, L. (1972), Limits of experiments, in L. Le Cam, J. Neyman & E. L. Scott, eds, 'Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 245–261.
- Le Cam, L. (1973), 'Sur les contraintes imposées par les passages à la limite usuels en statistique', *Proceedings 39th Session of the International Statistical Institute XLV*, 169–177.
- Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.
- Stein, C. (1956), Inadmissibility of the usual estimator for the mean of a normal distribution, in J. Neyman, ed., 'Proceedings of the Third

- Berkeley Symposium on Mathematical Statistics and Probability', Vol. I, University of California Press, Berkeley, pp. 197–206.
- Strasser, H. (1978), 'Global asymptotic properties of risk functions in estimation', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **45**, 35–48.
- van der Vaart, A. (1991), 'An asymptotic representation theorem', *International Statistical Review* **59**, 97–121.
- von Mises, R. (1931), *Wahrscheinlichkeitsrechnung*, Springer-Verlag, Berlin.
- Wald, A. (1950), *Statistical Decision Functions*, Wiley, New York.
- Wolfowitz, J. (1953), 'The method of maximum likelihood and the Wald theory of decision functions', *Indagationes Mathematicae* **15**, 114–119.

28

Le Cam's Procedure and Sodium Channel Experiments

Grace L. Yang¹

28.1 Introduction

Consider a random variable U which has either a discrete distribution,

$$\begin{aligned} P[U = 0] &= 1 - p, \\ P[U = u] &= p(1 - \lambda)\lambda^u, \end{aligned} \quad (1)$$

for $u = 1, 2, \dots$, with parameters $0 < p < 1$ and $0 < \lambda < 1$, or a mixture distribution with an atom at zero and an exponential density for $u > 0$,

$$\begin{aligned} P[U = 0] &= 1 - p, \\ p\lambda \exp(-\lambda u), \end{aligned} \quad (2)$$

for $\lambda > 0$.

In either case, the distribution of the sum $W = \sum_{r=1}^m U_r$ of m i.i.d. replicates of U can be represented by

$$f(w, \delta; p, \lambda) = (1 - p)^{m(1-\delta)} \left[\sum_{k=0}^m \binom{m}{k} p^k (1 - p)^{m-k} g_k(w, \lambda) \right]^{\delta}, \quad (3)$$

where $\delta = I[W > 0]$ is the indicator of the event $[W > 0]$, and the $g_k(w, \lambda)$ are determined by the underlying model (1) or (2). If U has the truncated geometric distribution (model (1)), then (3) is the probability $P[W = w], w = 0, 1, 2, \dots$, with

$$g_k(w, \lambda) = \binom{w-1}{w-k} \lambda^{w-k} (1 - \lambda)^k. \quad (4)$$

We set the binomial coefficient $\binom{w-1}{w-k}$ equal to 0 for $w-1 < w-k$. If U has the truncated exponential distribution (model (2)), then the distribution of W has an atom at $W = 0$ and a density on $W > 0$, with

$$g_k(w, \lambda) = \lambda(\lambda w)^{k-1} \exp(-\lambda w)/(k-1)!. \quad (5)$$

¹University of Maryland

Our problem is to estimate the vector parameter $\theta = (p, \lambda)'$ with n i.i.d observations, $W_i, i = 1, \dots, n$, where W_i follows the distribution given by (3).

Model (3) is constructed for analyzing sodium channel experiments in neurophysiology in which the sum W are observable but not the individual components U_r , hence the mixture model.

The characteristics of the W and the parameters p and λ are intimately related to the generation of nerve impulses. Further explanation is given in Section 3. For now, we address the estimation of the parameters (p, λ) .

Usually about 500 independent observations of W can be recorded from the experiment. That sample size would make the likelihood function, $\prod_{j=1}^n [f(w_j, \delta_j); (p, \lambda)]$, a polynomial in p and λ of degrees 2000 if the underlying model is (4) and $m = 4$ (a typical experimental value), where f is defined by (3) with (w, δ) replaced by the observed (w_j, δ_j) for $j = 1, \dots, n$. The likelihood function for model (5) is equally complicated. Finding maximum likelihood estimates (MLE), even with the state-of-art computing power, is a difficult job. For the mixture model in general and model (3) in particular the method of moment estimates are readily available. However, they are known to be inefficient. A suitable and easily implementable estimation procedure is the Le Cam procedure. The LAN type of condition are satisfied for model (3) which ensures the asymptotic optimality of the estimates.

Since the publication of the fundamental paper “Locally Asymptotically Normal Families of Distributions” (Le Cam 1960), the theory of LAN has been studied extensively and generalized by many authors.

While this elegant theory has received a lot of attention, but its practical application has been by and large ignored. The Le Cam procedure has yet to make its way to a software package.

In this note, we look at the LAN theory from the view of applications. For clarity of exposition, we consider only finite dimensional Euclidean parameter space Θ and a very simple set of LAN conditions. We highlight several of Le Cam’s results and hope to motivate the reader to apply this powerful procedure.

28.2 The Le Cam Procedure

Let $\mathcal{E}_n = \{P_{\theta,n}; \theta \in \Theta\}$ be a family of probability measures on a σ -field \mathcal{A}_n of the sample space \mathcal{X}_n and Θ a subset of a finite-dimensional Euclidean space. Let X_1, \dots, X_n be random vectors with joint distribution $P_{\theta,n}$. The Le Cam procedure for estimating θ calls for global search for a “preliminary” estimates θ_n^* in the whole space Θ then a refinement of θ_n^* by adding to it a term $\delta_n(M_n^{-1}Y_n)$ to obtain the final estimate

$$\hat{\theta}_n = \theta_n^* + \delta_n(M_n^{-1}Y_n), \quad (6)$$

where δ_n is a nonrandom scale factor, $M_n^{-1}Y_n$ a random vector and M_n^{-1} the estimate of the variance of $\hat{\theta}_n$.

Despite the familiar appearance, $\hat{\theta}_n$ is not a one-step Newton-Raphson. For the construction does not call for unlimited iterations and $M_n^{-1}Y_n$ need not involve the first two derivatives of the loglikelihood. In fact, the construction does not even require the existence of the derivatives. Furthermore, since $\hat{\theta}_n$ is not obtained by global maximization over Θ , it is not an MLE except incidentally. The estimate $\hat{\theta}_n$ is composed of a δ_n -consistent estimate θ_n^* and a correction term $M_n^{-1}Y_n$ whose computational formula is provided by the theory. There are different ways of computing the correction term, the form $M_n^{-1}Y_n$ is used for comparisons with other familiar methods, e.g. the delta method.. Since in general it is considerably easier to find a consistent estimator than an MLE, the computational advantage of (6) over MLE in a complicated model like (3) is clear.

An estimate θ_n^* is called δ_n -consistent if there is a sequence of positive numbers δ_n tending to zero as n tends to infinity and for every $\epsilon > 0$, there exist b (may depend on ϵ and θ) and $n_\epsilon(\theta)$ such that for $n \geq n_\epsilon(\theta)$

$$P_{\theta,n}[|\theta_n^* - \theta| \leq \delta_n b] \geq 1 - \epsilon, \quad \theta \in \Theta,$$

where $|\cdot|$ is the usual Euclidean distance. A restatement of this says that with probability $1 - \epsilon$ the true parameter θ equals $\theta_n^* + \delta_n \tau$ for some τ in the set $\{t : |t| \leq b\}$. Since we have a large sample size n at our disposal, the information of the preliminary estimate θ_n^* will bring us to local neighborhoods of radius $\delta_n b$, $\Theta_n = \{\eta : |\eta - \theta| \leq \delta_n b\}$, of the true θ . Estimation of θ becomes that of estimating the local parameter τ .

Now suppose θ_n^* is given. We proceed to estimate τ by looking at the loglikelihood ratio

$$\Lambda_n(\theta_n^* + \delta_n \tau; \theta_n^*) = \log \frac{dP_{\theta_n^* + \delta_n \tau, n}}{dP_{\theta_n^*, n}} \quad \text{for } \tau \in \{t : |t| \leq b\} \quad (7)$$

in the vicinity of θ_n^* .

Localization paves the way for local approximation of $\Lambda_n(\theta_n^* + \delta_n \tau; \theta_n^*)$ by simple forms. Namely, with θ_n^* fixed, we approximate $\Lambda_n(\theta_n^* + \delta_n \tau; \theta_n^*)$ locally by a linear-quadratic form in τ (discussed below). The estimate $\hat{\tau}_n$ is the value that maximizes the linear quadratic approximation. Asymptotic optimal properties of the estimates $\hat{\tau}_n$ and $\hat{\theta}_n$ have been established under the LAN conditions and many of its generalizations. To keep the exposition simple, we restrict ourselves to a version of the LAN conditions and the product experiment where $P_{\theta,n}$ is a product measure $\prod_{j=1}^n p_{j,n}(\theta)$ with $p_{j,n}(\theta)$ the probability distribution of X_j for $j = 1, \dots, n$. Formally, the loglikelihood ratio is defined by $\Lambda_n(\eta; \theta) = \log(dP_{\eta,n}/dP_{\theta,n})$, which equals $+\infty, -\infty$, respectively on sets such that $P_{\eta,n}$ is $P_{\theta,n}$ singular, and $P_{\theta,n}$ is $P_{\eta,n}$ singular. In the following all the limits are taken as $n \rightarrow \infty$.

Definition 2 The family $\mathcal{E}_n = \{P_{\eta,n} : \eta \in \Theta\}$ is called locally asymptotically normal at $\theta \in \Theta$ if the following conditions A1-A4 hold,

- A1. Θ is an open subset of R^k .
- A2. The sequence $\{P_{\theta+\delta_n t_n, n}\}$ and $\{P_{\theta, n}\}$ are contiguous for all bounded $|t_n|$ sequences.
- A3. There exist random vectors $S_n(\theta)$ and random matrices $K_n(\theta)$ such that for any bounded sequences $|t_n|$, the difference

$$\Lambda_n(\theta + \delta_n t_n, \theta) - \left(t'_n S_n(\theta) - \frac{1}{2} t'_n K_n(\theta) t_n \right)$$

tends to zero in $P_{\theta, n}$ probability.

- A4. $K_n(\theta)$ tends to a nonrandom positive definite matrix $K(\theta)$ in $P_{\theta, n}$ probability.

To apply the linear-quadratic approximation in A3 to (7), one needs to justify the substitution of θ in A3 by the random variable θ_n^* . Le Cam (1960) legalizes the substitution by introducing a discretizing trick on θ_n^* . Or if the functions $s \rightsquigarrow \Lambda_n(\theta + \delta_n s, \theta)$ are equi-continuous on bounded sets of values of s , then discretization of θ_n^* is unnecessary. In terms of applications, discretization amounts to computing θ_n^* only to $|\log(\delta_n)|$ decimals in each coordinate.

Note also if there exists one pair $(S_n(\theta), K_n(\theta))$ that satisfies A3, then one can find many other pairs $(\tilde{S}_n(\theta), \tilde{K}_n(\theta))$ that satisfies A3. This fact leads to a variety of linear-quadratic approximations of $\Lambda(\theta_n^* + \delta_n \tau, \theta_n^*)$. We shall illustrate some of them below.

Note that the substitution is of the kind $\theta_n^* = \theta + \delta_n \beta$ with β a random variable bounded in $P_{\theta, n}$ probability. Let us first consider the substitution by a non random quantity in the log likelihood ratios, i.e. if θ is replaced by $\theta + \delta_n a_n$ for a bounded non random a_n . Then the log likelihood ratio becomes

$$\begin{aligned} H_n(a_n, t_n) &= \Lambda_n(\theta + \delta_n(t_n + a_n); \theta + \delta_n a_n) \\ &= \Lambda_n(\theta + \delta_n(t_n + a_n); \theta) - \Lambda_n(\theta + \delta_n a_n; \theta), \end{aligned} \quad (8)$$

where t_n and a_n are any bounded sequences. Applying A2 and A3 to the log likelihood ratios in the second equality yields a linear-quadratic approximation to $H_n(a_n, t_n)$ in t_n as follows,

$$H_n(a_n, t_n) - [t'_n(S_n - K_n a_n) - \frac{1}{2} t'_n K_n t_n] \quad (9)$$

tends to zero in $P_{\theta, n}$ probability. Here and in the sequel, for notational convenience, we shall suppress θ and write S_n for $S_n(\theta)$ and K_n for $K_n(\theta)$

whenever there is no danger of ambiguity. We will use the notation $A_n \sim B_n$ to denote that the difference $A_n - B_n$ tends to zero in $P_{\theta,n}$ probability. Or we say A_n approximates B_n . The convergence in probability will always mean in $P_{\theta,n}$ probability unless specified otherwise. Of course the contiguity condition A2 implies the convergence under $P_{\theta+\delta_n t_n, n}$ as well.

Notice that an effect of the substitution is the change of the linear term from S_n of A3 to $S_n - K_n a_n$ of equation (9).

In the next step, we take a basis $\{u_1, \dots, u_k\}$ of R^k , for instance, the natural basis. Set $u_0 = 0$. Then the second differences of the likelihood ratios H_n provide an approximation to K_n , namely,

$$-\{H_n(a_n, u_i + u_j) - H_n(a_n, u_i) - H_n(a_n, u_j)\} \sim u_i' K u_j,$$

for $i, j = 0, \dots, k$.

Denote the second differences by $u_i' K_n^* u_j$. It follows from equation (9) that for each vector u_i ,

$$H_n(a_n, u_i) \sim [u_i'(S_n - K_n a_n) - \frac{1}{2} u_i' K_n^* u_i],$$

and

$$H_n(a_n, u_i) \sim [u_i'(S_n - K_n^* a_n) - \frac{1}{2} u_i' K_n^* u_i].$$

All of this assumes that $\{a_n\}$ is a bounded non random sequence. But if θ_n^* has been discretized so that the number of possible values in spheres of diameter $\delta_n b$, $b > 0$, remain bounded, then the above approximations remain valid if a_n is replaced by $\delta_n^{-1}(\theta_n^* - \theta)$, i.e. by the random variable β .

We could then proceed as in Le Cam & Yang (1990, page 58) to construct estimates. There are other possible ways of approximating K_n which we shall now address. For each component probability measure p_{jn} , set

$$\begin{aligned} \Lambda_{n,ji} &= \log \frac{dp_{jn}(\theta + \delta_n(a_n + u_i))}{dp_{jn}(\theta + \delta_n a_n)} \\ Z_{n,ji} &= \frac{1}{\alpha} \left[\left(\frac{dp_{jn}(\theta + \delta_n(a_n + u_i))}{dp_{jn}(\theta + \delta_n a_n)} \right)^\alpha - 1 \right] = \frac{1}{\alpha} \left[\exp(\alpha \Lambda_{n,ji}) - 1 \right] \end{aligned} \quad (10)$$

for $i = 1, \dots, k$, $j = 1, \dots, n$ and $\alpha \in (0, 1]$. Then $H_n(a_n, u_i) = \sum_{j=1}^n \Lambda_{n,ji}$. By Taylor's expansion, we write

$$\sum_{j=1}^n Z_{ji} = \sum_{j=1}^n \Lambda_{ji} + \frac{\alpha}{2} \sum_{j=1}^n \Lambda_{ji}^2 + \epsilon_n \quad (11)$$

for a random remainder term ϵ_n . The $Z_{n,ji}$ are called centering variables.

Instead of using the likelihood ratios $\Lambda_{n,ji}$, we can obtain linear-quadratic approximations in terms of the user friendly $Z_{n,ji}$. For this purpose, let us assume that the random variables $\Lambda_{n,ji}$ satisfy the uan condition, that is,

$$\sup_{1 \leq j \leq n} h_j^2(\theta + \delta_n(a_n + u_i), \theta + \delta_n u_i) \rightarrow 0,$$

where $h_j(s, t)^2 = \int (\sqrt{dp_{jn}(s)} - \sqrt{dp_{jn}(t)})^2 / 2$, squared Hellinger distance.

Under the LAN and the uan conditions, Le Cam (1966) showed that ϵ_n tends to zero in $P_{\theta,n}$ probability.

It also can be shown that the asymptotic normality of $H_n(a_n, u_i)$ implies that

- (a) $\sum_j \Lambda_{n,ji}^2 \sim \sum_j Z_{n,ji}^2$
- (b) $\sum_j Z_{n,ji}^2 \sim \text{Var}(\sum_j Z_{n,ji})$
- (c) $\text{Var}(\sum_j Z_{n,ji}) \sim \text{Var}(\sum_j \Lambda_{n,ji})$
- (d) $\text{Var}(\sum_j \Lambda_{n,ji}) \sim u'_i K_n(\theta) u_i$
- (d) if $\widetilde{M}_n(\theta)$ is the matrix with (i, l) th entry $\sum_j Z_{n,ji} Z_{n,jl}$, then $K_n(\theta) \sim \widetilde{M}_n(\theta)$.

Just as before these approximations of the quadratic terms remain correct if a_n is replaced by $\delta_n^{-1}(\theta_n^* - \theta)$.

A crucial step in proof of (a), (b), (c), (d) and (e) is to show that, for any truncated $Z_{n,ji}^c$ of $Z_{n,ji}$, where $Z_{n,ji}^c = Z_{n,ji}$ for $|Z_{n,ji}| \leq c$ and zero otherwise, with c an arbitrary positive constant, $\sum_j (EZ_{n,ji}^c)^2 \rightarrow 0$. See Le Cam & Yang (1990, Lemmas 1 and 2, p 34-36).

When $\alpha = 1$, the random variable Z_{ji} is a difference quotient. An important application of $\alpha = 1$ can be found in Section 3. It is used to show that the LAN conditions are preserved under information loss. It is worth noting that in Section 3, the observable variables are identically distributed. The proof of the sum of the squared truncated expectations tending to zero simplifies drastically. By far the most important special case is $\alpha = \frac{1}{2}$. It gives the famous Le Cam's Second Lemma. (See Pollard (1996)—in this volume—for an alternative proof).

To continue, telescoping (a)–(d) we see that both terms $u'_i K_n(\theta) u_i$ and $u'_i \widetilde{M}_n(\theta) u_i$ approximate $\sum_j \Lambda_{n,ji}^2$. Applying these results to (11), we obtain for each u_i the following approximation in terms of the $Z_{n,ji}$,

$$H_n(a_n, u_i) \sim \left\{ \sum_j Z_{n,ji} - \frac{\alpha}{2} u'_i \widetilde{M}_n(\theta) u_i \right\}. \quad (12)$$

Since this is computed in terms of the basis $\{u_1, \dots, u_k\}$, we need to reparametrize $\eta \in \Theta_n$ in terms of the same basis as

$$\eta - \theta = \delta_n \sum_{i=1}^k v_i u_i = \delta_n \tau \quad (13)$$

Thus, when θ is fixed, estimating τ is to estimate the scalars v_i for $i = 1, \dots, k$.

Comparing equation (12) with equation (9), we see that

$$\tau'(S_n - K_n a_n) \sim \sum_{i=1}^k v_i \left[\sum_j Z_{n,ji} + \frac{1}{2}(1-\alpha) u_i' \tilde{M}_n(\theta) u_i \right] = \tau' \tilde{Y}_n(\theta), \quad (14)$$

where $\tilde{Y}_n(\theta)$ is a $k \times 1$ column vector whose i th component is

$$\tilde{Y}_{ni}(\theta) = \sum_j Z_{ji} + \frac{1}{2}(1-\alpha) u_i' \tilde{M}_n(\theta) u_i, \quad \alpha \in (0, 1]$$

Then the following linear-quadratic approximation holds,

$$H_n(a_n, \tau) \sim \{\tau' \tilde{Y}_n(\theta) - \frac{1}{2} \tau' \tilde{M}_n(\theta) \tau\}. \quad (15)$$

Equation (15) is almost what we need. Now replacing $\theta + \delta_n a_n$ by θ_n^* in the definition of $Z_{n,ji}$, the resulting $\tilde{Y}_n(\theta_n^*)$ and $\tilde{M}_n(\theta_n^*)$ are computable (see (d)) and will be denoted by Y_n and M_n respectively. We may therefore approximate $\Lambda_n(\theta_n^* + \delta\tau; \theta_n^*)$ (defined by (7)) by

$$L^{(1)}(\tau) = \tau' Y_n - \frac{1}{2} \tau' M_n \tau. \quad (16)$$

The estimate $\hat{\tau}_n$ is the value τ that maximizes $L^{(1)}(\tau)$. Assuming M_n^{-1} exists, then $\hat{\tau}_n = \sum_{i=1}^k \hat{v}_i u_i = M_n^{-1} Y_n$. It follows by (9) that η can be estimated by

$$\hat{\eta} = \theta_n^* + \delta_n \hat{\tau}_n = \theta_n^* + \delta_n M_n^{-1} Y_n,$$

which is our final estimate of θ as given by (6).

Any $\alpha \in (0, 1]$ can be used to construct (Y_n, M_n) . For other fitting methods see (Le Cam 1977, Le Cam 1986).

Under the LAN conditions and the discretization condition on θ_n^* , for any of the pairs $(\hat{\theta}_n, M_n)$ discussed above the differences $\delta_n^{-1}(\hat{\theta}_n - \theta) - K_n^{-1}(\theta) S_n(\theta)$ and $M_n - K_n(\theta)$ tends to zero in $P_{\theta,n}$ probability. The inverse M_n^{-1} serves as an estimate of the variance of $\hat{\theta}_n$. The following asymptotic optimal properties hold:

- (i) By the Hájek-Le Cam minimax theorem, the local estimator $\hat{\tau}_n$ attains the local asymptotic minimax lower bound for any nonnegative bowl shape loss function.
- (ii) The standardized sequence $\delta_n^{-1}(\hat{\theta}_n - \theta)$ has an asymptotic normal distribution $N(0, K^{-1}(\theta))$, $\theta \in \Theta$. It follows by the Hájek-Le Cam convolution theorem that $\hat{\theta}_n$ are asymptotically most concentrated estimates among all sequences of estimates that are δ_n -consistent and regular.

Recall that under the LAN conditions, any δ_n -consistent sequence is δ_n -regular at Lebesgue almost all θ .

Several remarks are in order. The construction relies on the existence of a k -dimensional basis $\{\delta_n u_1 + \theta, \dots, \delta_n u_k + \theta\} \in \Theta_n \subset \Theta \subset R^k$. If Θ_n lies in a lower dimensional subspace of R^k , then it is not possible to find a k -dimensional basis although the LAN conditions still hold. See Le Cam & Yang (1988, p. 105) for the fast spiral example in which Θ_n (with the true $\theta = 0$) looks like a straight line in R^2 for n sufficiently large.

The LAN condition A3 is a statement of existence of $(S_n(\theta), K_n(\theta))$. A lot more is known about this condition when $P_{\theta,n}$ is the distribution of iid random vectors, $X_j, j = 1, \dots, n$. Let p_θ be the distribution of X_j and $Z(\tau) = \sqrt{dp_\tau/dp_\theta}, \tau \in \Theta$. If $Z(\tau)$ satisfies the differentiability in quadratic mean condition (DQM_0) at θ , then LAN conditions are satisfied with $\delta_n = 1/\sqrt{n}$.

In this connection, we state a very interesting condition obtained by Le Cam (1970; 1986, p. 588). Let $h(s, t)$ denote the Hellinger distance between p_s and p_t such that $h^2(s, t) = \frac{1}{2} \int (\sqrt{dp_s} - \sqrt{dp_t})^2$ for $s, t \in \Theta$. If $\sigma(t) = \limsup_{s \rightarrow t} h(s, t)/|s - t| < \infty$ for Lebesgue almost t , then the DQM_0 conditions hold for Lebesgue almost all t . A strengthened version of this condition is available for one dimensional parameters. It says that if $\sigma(t) < \infty$ and at a point θ interior to Θ (which is satisfied since our Θ is assumed to be open) such that $\lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \int_{-\epsilon}^{\epsilon} |\sigma(\theta + t) - \sigma(\theta)| dt = 0$ then DQM_0 holds at θ .

So if Le Cam's condition holds, there is no need in this case to explicitly calculate $(S_n(\theta), K_n(\theta))$ in order to construct statistics (Y_n, M_n) .

Finally we note that the success of $\hat{\theta}_n$ depends on both the existence of a δ_n -consistent estimate of θ and the fulfillment of the LAN conditions. The same rate δ_n is required for the consistent estimate and the LAN conditions. However, a family of probability measures $\{P_{\theta,n}; \theta \in \Theta\}$ that satisfies the local LAN conditions at rate δ_n does not imply the existence of a global δ_n -consistent estimate. For the iid case where $P_{\theta,n} = \prod_{j=1}^n p_{j\theta}$ and the rate $\delta_n = 1/\sqrt{n}$, Le Cam (1986, p. 608) showed that under general conditions there are \sqrt{n} -consistent estimates θ_n^* such that $\sqrt{n}|\theta_n^* - \theta|$ is bounded in $P_{\theta,n}$ probability at each θ that satisfies DQM_0 with a non singular covariance matrix for derivative in quadratic mean of $Z(\tau)$.

For more general δ_n and independent but not identically distributed variables, "metric dimension" (Le Cam 1986, Section 16.5) restrictions may be needed for determining the achievable rates of the estimates.

28.3 Sodium Channel Experiments

In a major technical break through using the patch-clamp method (Sakman & Neher 1983), it becomes possible to record microscopic current (in pico

Ampères) that flows through single ion channel located across cell membranes. These membrane channels are highly selective allowing only the passages of specific ions in and out of cells. In the experiment considered here, a patch of cell membrane that contains several sodium channels (only sodium ions can pass through) are isolated for studying channel responses under electrical stimulation. A stimulus (a voltage pulse) is applied to the patch to elicit responses from the sodium channels.

In the resting state, these channels are closed. A response means that a channel opens under stimulation. A known response pattern is that a channel may either not respond to the stimulus or if it responds, it will open and close repetitively a random number of times before it eventually tires out and stays closed. In a patch-clamp experiment, the number of times the channels open and the amount of current that flows through the channels in the patch are recorded. In model (3), m denotes the number of sodium channels in the patch and p the single channel response probability. Let U_r for $r = 1, \dots, m$ be the response variable of individual channels in the patch. For instance, if U_r measures the frequency of a single channel response, then under certain assumptions the distribution of U_r follows model (1). The nonresponse event is $[U_r = 0]$, the positive U represents the frequency of channel openings.

Now if the stimulation experiment is repeated n times, we denote the response of the r th channel to the j th stimulus by X_{rj} , $j = 1, \dots, n, r = 1, \dots, m$. That is, the X_{rj} are replicates of U_r . In each of these experiments however, the recorded response is the sum $W_j = \sum_{r=1}^m X_{rj}$ and not the individual X_{rj} . Under the iid assumption of the X_{rj} , the distribution of W_j is given by model (3) which is to be used to estimate the parameter $\theta = (p, \lambda)'$.

To construct the estimate $\hat{\theta}_n$, let, for every fixed j let $\mathcal{A}_{j,n}$ be the σ -field generated by X_{rj} for $r = 1, \dots, m$ and $\mathcal{B}_{j,n}$ be the σ -field generated by W_j . Thus $\mathcal{B}_{j,n} \subset \mathcal{A}_{j,n}$. Let $p_{j,n}(\theta)$ be the product probability measure on $\mathcal{A}_{j,n}$ of the observations X_{rj} for $r = 1, \dots, m$. The family of the probability measures $\mathcal{E}_n = \{P_{\theta,n}; \theta \in \Theta\}$ of the refined (but not observable) data $X_{rj}, r = 1, \dots, m; j = 1, \dots, n$ consists of product measure $P_{\theta,n} = \prod_{j=1}^n p_{jn}(\theta)$. Let us denote the probability measure of the coarser observation W_j by $q_{j,n}(\theta)$. It is the measure determined by model (3). Then the family of probability measures $\mathcal{F}_n = \{Q_{\theta,n}; \theta \in \Theta\}$ of the coarser data $\{W_j, j = 1, \dots, n\}$ consists of the measures $Q_{\theta,n} = \prod_{j=1}^n q_{jn}(\theta)$. Since $q_{jn}(\theta)$ is the restriction of $p_{jn}(\theta)$ to the sub σ -field $\mathcal{B}_{j,n}$ of $\mathcal{A}_{j,n}$, the information loss from the X to the W -data is componentwise, i.e., from X_j to W_j , for $j = 1, \dots, n$, here X_j denotes the vector $(X_{1,j}, \dots, X_{m,j})$. In this case, it can be shown that if the \mathcal{E}_n satisfy the LAN conditions, so do the coarser experiments \mathcal{F}_n (Le Cam & Yang 1988). The proof that \mathcal{E}_n satisfy the LAN conditions is straightforward. Therefore, we conclude that the \mathcal{F}_n statify the LAN conditions and the estimation procedure described in the last section applies.

We need the centering variables for the experiments \mathcal{F}_n similar to the $Z_{n,ji}$ in equation (10). There is a useful connection here. The probability measures $q_{jn}(\theta)$ are the restrictions of the corresponding $p_{jn}(\theta)$ to the σ -fields \mathcal{B}_{jn} . It follows that the conditional expectation of $Z_{n,ji}$ (for $\alpha = 1$) under $P_{\theta + \delta_n a_n, n}$ is

$$E[Z_{n,ji} | \mathcal{B}_{jn}] = \frac{dq_{jn}(\theta + \delta_n(a_n + u_i))}{dq_{jn}(\theta + \delta_n a_n)} - 1 \quad \text{almost surely.}$$

The right hand side is precisely the $Z_{n,ji}$ with p_{jn} replaced by q_{jn} . The value $\alpha = 1$ is in the interval $(0,1]$ that permits approximations in the sense of eqs. (10) - (12). Of course, we are applying (10) - (12) to the probability measures q_{jn} . Here the conditional expectation plays an important role in proving the preservation of the LAN conditions under information loss.

Thus the construction of the estimate follows the same procedure as discussed in Section 2. We need only to replace the measure $p_{jn}(\theta)$ by $q_{jn}(\theta)$ throughout (Le Cam & Yang 1988, Yang & Swenberg 1992). The quantity $dq_{jn}(\theta)$ is given by equation (3) with $g_k(w, \lambda)$ replaced by either (4) or (5) as the case may be. For the preliminary estimates, $\theta_n^* = (p_n^*, \lambda_n^*)$, we use the method of moments estimates for p and λ . They can be easily computed from the relationship $W = \sum_{r=1}^m U_r$.

28.4 Concluding Remarks

The LAN type of conditions are much weaker than those used for the maximum likelihood estimates.

In the Le Cam procedure, there is a considerable amount of freedom in finding a local approximation for $\Lambda_n(\theta_n^* + \delta_n \tau; \theta_n^*)$. As a practical guideline for finite sample applications, one could try out different approximations and retain those that fit the data. This can be fairly easily carried out by computer. It would be interesting to investigate analytically the implication of different approximations to the finite sample procedure.

Acknowledgments: The research was supported by ONR Grant N00014-93-J-1097.

28.5 REFERENCES

- Le Cam, L. (1960), 'Locally asymptotically normal families of distributions', *University of California Publications in Statistics* **3**, 37–98.
- Le Cam, L. (1966), Likelihood functions for large numbers of independent observations, in F. N. David, ed., 'Research Papers in Statistics',

- Wiley, New York, pp. 167–187. Festschrift for J. Neyman.
- Le Cam, L. (1970), ‘On the assumptions used to prove asymptotic normality of maximum likelihood estimators’, *Annals of Mathematical Statistics* **41**, 802–828.
- Le Cam, L. (1977), On the asymptotic normality of estimates, in ‘Proceedings of the IASPS Symposium’, Polish Scientific Publishers, Warsaw, pp. 203–217. Symposium to honor Jerzy Neyman.
- Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.
- Le Cam, L. & Yang, G. L. (1988), ‘On the preservation of local asymptotic normality under information loss’, *Annals of Statistics* **16**, 483–520.
- Le Cam, L. & Yang, G. L. (1990), *Asymptotics in Statistics: Some Basic Concepts*, Springer-Verlag.
- Pollard, D. (1996), Another look at differentiability in quadratic mean, in ‘Research Papers in Probability and Statistics: Festschrift for Lucien Le Cam’, Springer-Verlag.
- Sakman, N. B. & Neher, E., eds (1983), *Single Channel Recording*, Plenum Press, New York.
- Yang, G. L. & Swenberg, C. E. (1992), ‘Estimation of open dwell time and problems of identifiability in channel experiments’, *Journal of Statistical Planning and Inference* **33**, 107–119.

Assouad, Fano, and Le Cam

Bin Yu¹

ABSTRACT This note explores the connections and differences between three commonly used methods for constructing minimax lower bounds in nonparametric estimation problems: Le Cam's, Assouad's and Fano's. Two connections are established between Le Cam's and Assouad's and between Assouad's and Fano's. The three methods are then compared in the context of two estimation problems for a smooth class of densities on [0,1]. The two estimation problems are for the integrated squared first derivatives and for the density function itself.

29.1 Introduction

In nonparametric estimation problems, minimax is a commonly used risk criterion. An optimal minimax rate is often obtained by first deriving a minimax lower bound, often nonasymptotic, on the risk and then constructing an explicit estimator which achieves the rate in the lower bound. See for example, Has'minskii & Ibragimov (1978), Bretagnolle & Huber (1979), Stone (1984), Birgé (1986), Bickel & Ritov (1988), Donoho & Nussbaum (1990), Fan (1991), Donoho & Liu (1991), Birgé & Massart (1992), and Pollard (1993). Depending on the class considered and the function or functional estimated, techniques used to derive lower bounds differ. Three general methods have been widely employed: one formalizes some arguments of Le Cam (1973), and the other two are based on inequalities of Assouad (1983) and Fano (compare with Cover & Thomas 1991, p. 39). The first method will be referred to in this note as Le Cam's method. (Note that Le Cam (1986, Chapter 16) has developed a much more general theory.) It deals with two sets of hypotheses, while the Assouad and Fano methods deal with multiple hypotheses, indexed by the vertices of a hypercube and those of a simplex, respectively.

In this note, we explore the connections and differences between these three lemmas with the hope of shedding light on other more general problems. Section 2 contains two results (Lemma 2 and Lemma 5), which relate the three methods. It is known that Assouad's lemma (Lemma 2) gives

¹University of California at Berkeley.

very effective lower bounds for many global estimation problems. One way to understand that lemma is through Le Cam's method: the global estimation problem can be decomposed into several sub-estimation problems, and Assouad's Lemma is obtained by applying Le Cam's method to the sub-problems. In Lemma 5 we use a simple packing number result to extract a subset of the vertices of a hypercube to which the Fano method is applied, thereby obtaining a lower bound similar to that of Assouad. Hence in this sense, Fano's method is stronger, as observed by Birgé (1986).

From the examples worked out in the literature, it appears that Le Cam's method often gives the optimal rate when a real functional is estimated, but it can be non-straightforward to find the appropriate two sets of hypotheses in some problems. On the other hand, the other two lemmas seem to be effective when the whole unknown function is being estimated, although Assouad's Lemma seems easier to use and therefore more popular than Fano's. In Section 3, we demonstrate this point in the context of a particular smooth class of densities on $[0,1]$ and with two estimation problems, one for a real functional and one for the whole density. For the functional, Le Cam's method gives the optimal rate of convergence, while Assouad's and Fano's provide the optimal rate for the whole density.

After the completion of this note, closely related work by C. Huber was brought to my attention. In her article, which appears in this volume, she explores the connection between Assouad's and Fano's methods.

29.2 The three methods

Assume that \mathcal{P} is a family of probability measures and $\theta(P)$ is the parameter of interest with values in a pseudo-metric space (\mathcal{D}, d) . (It would be inconvenient to require that $d(\theta, \theta') = 0$ implies that $\theta = \theta'$.) Let $\hat{\theta} = \hat{\theta}(X)$ be an estimator of $\theta(P)$ based on an X with distribution P , and denote by $co(\mathcal{P})$ the convex hull of \mathcal{P} .

Le Cam (1973) relates the testing problem of two sets of hypotheses to the L^1 distance of the convex hulls of the two hypothesis sets. Roughly speaking, if one wants to test between these two sets well, then their convex hulls have to be well separated. Since estimators also define tests between subsets of \mathcal{D} , Le Cam's testing bound also provides a lower bound for the accuracy of an estimator.

Lemma 1 (Le Cam's method) *Let $\hat{\theta}$ be an estimator of $\theta(P)$ on \mathcal{P} taking values in a metric space (\mathcal{D}, d) . Suppose that there are subsets D_1 and D_2 of \mathcal{D} that are 2δ -separated, in the sense that, $d(s_1, s_2) \geq 2\delta$ for all $s_1 \in D_1$ and $s_2 \in D_2$. Suppose also that \mathcal{P}_1 and \mathcal{P}_2 are subsets of \mathcal{P} for which $\theta(P) \in D_1$ for $P \in \mathcal{P}_1$ and $\theta(P) \in D_2$ for $P \in \mathcal{P}_2$. Then*

$$\sup_{P \in \mathcal{P}} E_P d(\hat{\theta}, \theta(P)) \geq \delta \cdot \sup_{P_i \in co(\mathcal{P}_i)} \|P_1 \wedge P_2\|,$$

where the affinity $\|P_1 \wedge P_2\|$ is defined through

$$\|P_1 - P_2\|_1 = 2(1 - \|P_1 \wedge P_2\|).$$

Proof: For $P_i \in \mathcal{P}_i$,

$$\begin{aligned} M &:= 2 \sup_{P \in \mathcal{P}} E_P d(\hat{\theta}, \theta(P)) \\ &\geq E_{P_1} d(\hat{\theta}, \theta(P_1)) + E_{P_2} d(\hat{\theta}, \theta(P_2)) \\ &\geq E_{P_1} d(\hat{\theta}, \mathcal{D}_1) + E_{P_2} d(\hat{\theta}, \mathcal{D}_2), \end{aligned}$$

which implies

$$E_{P_1} d(\hat{\theta}, \mathcal{D}_1) + E_{P_2} d(\hat{\theta}, \mathcal{D}_2) \leq M \quad \text{for all } P_i \in co(\mathcal{P}_i).$$

Since $d(\hat{\theta}, \mathcal{D}_1) + d(\hat{\theta}, \mathcal{D}_2) \geq d(\mathcal{D}_1, \mathcal{D}_2) \geq 2\delta$, then for any $P_i \in co(\mathcal{P}_i)$,

$$M \geq 2\delta \inf_{f_1 \geq 0; f_1 + f_2 = 1} (E_{P_1} f_1 + E_{P_2} f_2) = 2\delta \|P_1 \wedge P_2\|.$$

Hence

$$M := 2 \sup_{P \in \mathcal{P}} E_P d(\hat{\theta}, \theta(P)) \geq 2\delta \cdot \sup_{P_i \in co(\mathcal{P}_i)} \|P_1 \wedge P_2\|.$$

□

Remark

(i) If d is not a pseudo-metric, but a non-negative symmetric function satisfying the following “weak” triangle inequality, that is, for some constant $A \in (0, 1)$,

$$d(x, z) + d(z, y) \geq Ad(x, y),$$

then the lower bound holds with an extra factor A . This observation is very useful in Example 2 in Section 1.3

(ii) In many cases, better lower bounds are obtained by considering the convex hulls of the \mathcal{P}_i , because the supremum of $\|P_1 \wedge P_2\|$ over the convex hulls can be much larger than the supremum over the \mathcal{P}_i themselves.

Assouad’s lemma gives a minimax lower bound over a class of 2^m hypotheses (probability measures) indexed by vertices of a m -dimensional hypercube. The following form of Assouad’s lemma is reworded from Devroye (1987, p. 60) (compare with Le Cam 1986, p. 524) to emphasize the decomposability of the (pseudo) distance d into a sum of m (pseudo) distances, which correspond to m estimation subproblems. The proof is rewritten to make the point that each subproblem is like testing the hypotheses indexed by neighboring vertices on the hypercube along the direction determined by the particular subproblem and the argument used in Le Cam’s method (Lemma 1) can be applied to each of the subproblems.

Lemma 2 (Assouad's Lemma) Let $m \geq 1$ be an integer and let $\mathcal{F}_m = \{P_\tau : \tau \in \{-1, 1\}^m\}$ contain 2^m probability measures. Write $\tau \sim \tau'$ if τ and τ' differ in only one coordinate, and write $\tau \sim_j \tau'$ when that coordinate is the j th. Suppose that there are m pseudo-distances on \mathcal{D} such that for any $x, y \in \mathcal{D}$

$$d(x, y) = \sum_{j=1}^m d_j(x, y), \quad (1)$$

and further that, if $\tau \sim_j \tau'$,

$$d_j(\theta(P_\tau), \theta(P_{\tau'})) \geq \alpha_m. \quad (2)$$

Then

$$\max_{P_\tau \in \mathcal{F}_m} E_\tau d(\hat{\theta}, \theta(P_\tau)) \geq m \cdot \frac{\alpha_m}{2} \min\{\|P_\tau \wedge P_{\tau'}\| : \tau \sim \tau'\}.$$

Proof: For any given $\tau = (\tau_1, \dots, \tau_m)$, let τ^j denote the m -tuple that differs from it in only the j th position. Then $d(\theta(P_\tau), \theta(P_{\tau^j})) \geq \alpha_m$.

$$\begin{aligned} & \max_{\tau} E_\tau d(\theta(P_\tau), \hat{\theta}) \\ &= \max_{\tau} \sum_{j=1}^m E_\tau d_j(\theta(P_\tau), \hat{\theta}) \\ &\geq 2^{-m} \sum_{\tau} \sum_{j=1}^m E_\tau d_j(\theta(P_\tau), \hat{\theta}) \\ &= \sum_{j=1}^m 2^{-m} \sum_{\tau} E_\tau d_j(\theta(P_\tau), \hat{\theta}) \\ &= \sum_{j=1}^m 2^{-(m+1)} \sum_{\tau} (E_\tau d_j(\theta(P_\tau), \hat{\theta}) + E_{\tau^j} d_j(\theta(P_{\tau^j}), \hat{\theta})) \end{aligned}$$

For each fixed τ and j , we have a pair of hypotheses P_τ and P_{τ^j} sitting on the neighboring vertices of the hypercube along direction j . Therefore, as in Le Cam's method (Lemma 1), the average estimation error over these two hypotheses can be bounded from below by $\frac{1}{2}\alpha_m\|P_\tau \wedge P_{\tau^j}\|$. Thus,

$$\begin{aligned} \max_{\tau} E_\tau d(\theta(P_\tau), \hat{\theta}) &\geq \sum_{j=1}^m 2^{-m} \sum_{\tau} \alpha_m \|P_\tau \wedge P_{\tau^j}\| \\ &\geq m \frac{\alpha_m}{2} \min\{\|P_\tau \wedge P_{\tau'}\| : \tau \sim \tau'\} \end{aligned}$$

□

The relation $\tau \sim \tau'$ can also be written $W(\tau, \tau') = 1$, where W denotes the *Hamming distance*,

$$W(\tau, \tau') = \frac{1}{2} \sum_{j=1}^m |\tau_j - \tau'_j|,$$

the number of places where τ and τ' differ.

From Remark (i) after Lemma 1, Assouad's lower bound holds with an extra factor A if d_j are non-negative symmetric functions satisfying the weak triangle inequality with the same constant A .

Devroye (1987, p. 77) (compare with Le Cam 1986, p. 524) contains a generalized Fano's lemma in the case that $\theta(P)$ is the density of P and d is the L^1 norm. We present here a slightly stronger version whose proof is based on ideas from Han and Verdú (1994). We find their proof less involved than those in the statistics literature. It is based on information theory concepts and Fano's original inequality (compare with Cover & Thomas 1991, p. 39).

Lemma 3 (Generalized Fano method) *Let $r \geq 2$ be an integer and let $\mathcal{M}_r \subset \mathcal{P}$ contain r probability measures indexed by $j = 1, 2, \dots, r$ such that for all $j \neq j'$*

$$d(\theta(P_j), \theta(P_{j'})) \geq \alpha_r,$$

and

$$K(P_j, P_{j'}) = \int \log(P_j/P_{j'}) dP_j \leq \beta_r.$$

Then

$$\max_j E_j d(\hat{\theta}, \theta(P_j)) \geq \frac{\alpha_r}{2} \left(1 - \frac{\beta_r + \log 2}{\log r}\right).$$

Proof: Write θ_j for $\theta(P_j)$. Let Y be a random variable uniformly distributed on the hypothesis set $\{1, 2, \dots, r\}$, and X be a random variable with the conditional distribution P_j given $Y = j$. Define Z as the value of j for which $d(\hat{\theta}(X), \theta_j)$ is a minimum. (It does not matter how we handle ties.) Because $d(\theta_j, \theta_{j'}) \geq \alpha_r$ for $j \neq j'$, we certainly have $Z = j$ when $d(\hat{\theta}(X), \theta_j) < \alpha_r/2$. It follows that

$$\begin{aligned} \max_j E_j d(\hat{\theta}, \theta(P_j)) &\geq \frac{\alpha_r}{2} \max_j P(d(\hat{\theta}(X), \theta_j)) \geq \frac{\alpha_r}{2} | Y = j \\ &\geq \frac{\alpha_r}{2r} \sum_{j=1}^r P(Z \neq j | Y = j) \\ &= \frac{\alpha_r}{2} P(Z \neq Y). \end{aligned}$$

Let h be the entropy function with the natural log,

$$h(p) = -p \log p - (1-p) \log(1-p) \quad \text{for } p \in (0, 1).$$

Then $0 \leq h(\cdot) \leq \log 2$. Denote by $I(Y; Z) = K(P_{(Y,Z)}, P_Y \times P_Z)$ the mutual information between Y and Z , and by $H(Y|Z)$ the equivocation or the average posterior entropy of Z given Y . Then

$$I(Y; Z) = H(Y) - H(Y|Z) = \log r - H(Y|Z).$$

Furthermore, by a property of mutual information and the convexity of the Kullback-Leibler divergence (Cover & Thomas 1991, pp. 30, 33), we have

$$\begin{aligned} I(Y; Z) = I(Y; \hat{\theta}(X)) &\leq I(Y; X) = \frac{1}{r} \sum_{i=1}^r K\left(P_i, \frac{1}{r} \sum_{j=1}^r P_j\right) \\ &\leq \frac{1}{r^2} \sum_{i,j} K(P_i, P_j). \end{aligned}$$

It follows from Fano's inequality (Cover & Thomas 1991, p. 39),

$$H(Y|Z) \leq P(Z \neq Y) \log(r-1) + h(P(Z=Y)),$$

that

$$\begin{aligned} P(Z \neq Y) \log(r-1) &\geq H(Y|Z) - h(1/2) \\ &= H(Y) - I(Y; Z) - \log 2 \\ &\geq \log r - \frac{1}{r^2} \sum_{i,j} K(P_i, P_j) - \log 2. \end{aligned}$$

Increase the $\log(r-1)$ to $\log r$ and replace $K(P_i, P_j)$ by its upper bound β_r , then substitute the resulting lower bound for $P(Z \neq Y)$ into the minimax inequality to get the asserted bound. \square

As remarked by Birgé (1986, p. 279), “[Fano’s Lemma] is in a sense more general because it applies in more general situations. It could also replace Assouad’s Lemma in almost any practical case ...”. Indeed, Lemma 3 implies a result similar to Assouad’s Lemma. The idea is to select the maximal subset of vertices from the m -dimensional hypercube which are $m/3$ apart in Hamming distance and apply Fano’s Lemma to the selected set of vertices.

Lemma 4 *For a universal positive constant c_0 , and each $m \geq 6$, there exists a subset A of $\{-1, +1\}^m$ consisting of at least $\exp(c_0 m)$ vertices, each pair greater than $m/3$ apart in Hamming distance.*

Proof: Let k be the integer part of $m/6$. Let A be a maximal set of vertices, each pair at least $2k+1$ apart in Hamming distance. The Hamming

ball $B(\tau, 2k)$ of radius $2k$ and center τ contains $N(m, 2k) = \sum_{r=0}^{2k} \frac{m!}{r!(m-r)!}$ vertices, corresponding to the subsets of $2k$ or fewer coordinates at which a vertex in the ball can differ from τ .

Because A is maximal, no vertex in the cube $\{-1, +1\}^m$ can lie further than $2k$ from A ; the whole cube is covered by a union of balls $B(\tau, 2k)$, with τ ranging over A . This union contains at most $|A|N(m, 2k)$ vertices, which is therefore an upper bound for 2^m .

It remains to calculate an upper bound for $N(m, 2k)$ by means of the usual generating function argument. Let Z denote a random variable with a $\text{Bin}(m, 1/2)$ distribution. Put $s = (m - 2k)/2k$, which is greater than 1. Then

$$N(m, 2k) = 2^m P(Z \leq 2k) \leq 2^m E s^{2k-Z} = 2^m s^{2k} \left(\frac{1}{2} + \frac{1}{2s} \right)^m.$$

The bound simplifies to $\exp(mh(2k/m))$, which leads to the asserted lower bound for A if we take $k = [m/6]$. \square

Lemma 5 *Let $m \geq 1$ be an integer and let $\mathcal{F}_m = \{P_\tau : \tau \in \{-1, +1\}^m\}$ contain 2^m probability measures, and let W be the Hamming distance. Suppose that there are constants α_m and γ_m such that*

$$d(\theta(P_\tau), \theta(P_{\tau'})) \geq \alpha_m W(\tau, \tau') \quad (3)$$

$$K(P_\tau, P_{\tau'}) \leq m\gamma_m. \quad (4)$$

Then

$$\max_{P_\tau \in \mathcal{F}_m} E_\tau d(\hat{\theta}_n, \theta(P_\tau)) \geq m \cdot \frac{\alpha_m}{6} \left(1 - \frac{1}{c_0} (\gamma_m + \log 2/m) \right).$$

Proof: Apply Lemma 3 to the set A of $r = \exp(c_0 m)$ vertices given by Lemma 4. From (3) and the $m/3$ separation of vertices in A , we have $d(\theta(P_\tau), \theta(P_{\tau'})) \geq \alpha_m W(\tau, \tau') \geq m\alpha_m/3$ for distinct vertices in A . \square

Let us now compare the conditions in Assouad's lemma and those in Lemma 5. The first two conditions (1) and (2) in Assouad's Lemma do not quite imply condition (3) in Lemma 5, but condition (2) together with the following stronger condition

$$\min_j \{d_j(\theta(P_\tau), \theta(P_{\tau'})) : \tau_j \neq \tau'_j\} \geq \alpha_m.$$

would imply condition (3). Note that this new condition is satisfied by hypercube classes constructed through perturbations of a fixed density over a partition, as in the next section. Moreover, note that condition (4) implies a lower bound on the affinity $\|P_\tau \wedge P_{\tau'}\|$ through the Kullback-Csiszár-Kemperman inequality (Devroye 1987, p. 10):

$$\|P_\tau - P_{\tau'}\|_1 \leq \sqrt{2K(P_\tau, P_{\tau'})}.$$

Since $\|P_\tau - P_{\tau'}\|_1 = 2(1 - \|P_\tau \wedge P_{\tau'}\|)$, $K(P_\tau, P_{\tau'}) \leq \gamma_m$ implies

$$\|P_\tau \wedge P_{\tau'}\| \geq 1 - \sqrt{\gamma_m/2}.$$

This seems to suggest that condition (4) in Lemma 5 is stronger than the affinity condition in Assouad's lemma. However, as is the case in many, if not all, hypercube classes one actually constructs, the probability measures in the hypercube class often have densities bounded from below by $c > 0$. In this case, a lower bound β_m on the affinity implies an upper bound on the Kullback divergence:

$$\begin{aligned} K(P_\tau, P_{\tau'}) &\leq c^{-1} \|P_\tau - P_{\tau'}\|_1 \\ &= 2c^{-1}(1 - \|P_\tau \wedge P_{\tau'}\|) \\ &\leq 2c^{-1}(1 - \beta_m) \end{aligned}$$

provided that $\|P_\tau \wedge P_{\tau'}\| \geq \beta_m$.

So far we have connected Le Cam's method with Assouad's in Lemma 2 and Fano's with Assouad's in Lemma 5. Comparing Le Cam's with Fano's would complete the circle. In a way, they are similar in that they both deal with hypothesis testing: Le Cam's for two sets of hypotheses, and Fano's for multiple hypotheses. Fano's method, however, does not cover the case of testing two simple hypotheses since the lower bound it gives when $r = 2$ is non-positive.

In the next section, we apply the above lemmas to a concrete class where Le Cam's method provides the optimal rate of convergence for a quadratic functional estimation problem and both Fano's and Assouad's lemmas provide the optimal rate for a global density estimation problem. These results are known. See for example Bickel & Ritov (1988) and Devroye (1987) respectively. The common feature of these two lower bound problems is that they both rely on the same hypercube class.

29.3 Two examples

Let \mathcal{M} denote the class of smooth densities f 's on $[0, 1]$ for which

$$0 < c_0 \leq f(x) \leq c_1 < \infty, \quad |f^{(2)}(x)| \leq c_2 < \infty, \quad \int_0^1 f(x)dx = 1.$$

Let us apply the results from Section 2 to derive lower bounds for minimax rates in two cases: a quadratic functional of f , with errors measured by the usual Euclidean distance; and the whole density f , with errors measured by Hellinger distance.

In both cases assume the estimators are based on a sample of n independent observations from some f in \mathcal{M} . Write f^n for the joint density, and \mathcal{P} for the corresponding class of product measures, with f in \mathcal{M} .

The lemmas will be applied to small perturbation of the uniform density, u , on $[0, 1]$. Take g a fixed twice differentiable function on $[0, 1]$ for which

$$\int_0^1 g(x)dx = 0, \int_0^1 g^2(x)dx = a > 0 \text{ and } \int_0^1 (g'(x))^2 dx = b > 0.$$

Divide $[0, 1]$ into m disjoint intervals of size $1/m$ and denote their centers by x_1, \dots, x_m . For $j = 1, 2, \dots, m$, let

$$g_j(x) = cm^{-2}g(mx - x_j)$$

with c small enough so that $|g_j| < 1$. Let

$$\mathcal{M}_m = \{f_\tau = 1 + \sum_{j=1}^m \tau_j g_j(x) : \tau = (\tau_1, \dots, \tau_m) \in \{-1, +1\}^m\},$$

and define the *hypercube class*

$$\mathcal{F}_m = \mathcal{M}_m^n = \{f_\tau^n : f_\tau \in \mathcal{M}_m\}.$$

Note that \mathcal{M}_m is simply the class of perturbed uniform densities with a rescaled g as the perturbation.

Example 1 Consider the quadratic functional

$$T(f) = \int_0^1 (f'(x))^2 dx$$

on \mathcal{F}_m . That is, $\theta(f^n) = \theta(f) = T(f)$, which takes values in the real line equipped with its metric $d(\theta, \theta') = |\theta - \theta'|$.

To obtain a minimax lower bound, we might try to use Assouad's lemma or Fano's method. Unfortunately, the functional $\theta(f) = T(f)$ takes the same value on the vertices of the hypercube and therefore the two results give only the trivial lower bound zero. However $\theta(u) = 0$, which differs from $\theta(f_\tau)$ for every τ , which lets us apply Le Cam's method to u^n and f_τ^n . For any fixed τ on the hypercube, it is easy to check that

$$\begin{aligned} H^2(u, f_\tau) &= O\left(\sum_j \int g_j^2\right) = O(m \cdot c^2 \cdot a \cdot m^{-5}) = O(m^{-4}), \\ \|u^n - f_\tau^n\|_1 &\leq 2H^2(u^n, f_\tau^n) \\ &= 2(1 - (1 - 2^{-1}H^2(u, f_\tau))^n) \\ &= 2(1 - (1 - O(m^{-4}))^n), \end{aligned}$$

and

$$T(u) = 0 \neq T(f_\tau) = \sum_j \left(\int g'_j\right)^2 = c^2 b m^{-2}.$$

If we choose $m = O(n^{1/4})$, then

$$\|u^n \wedge f_\tau^n\| = 1 - \|u^n - f_\tau^n\|_1/2 \geq (1 - O(m^{-4}))^n > 0,$$

and by Le Cam's method (Lemma 1), a lower bound on the minimax estimation rate is

$$|T(u) - T(f_\tau)| = |T(f_\tau)| = O(n^{-2/4}) = O(n^{-1/2}).$$

Unfortunately, this rate is not optimal, but the minimax optimal rate can be obtained (Bickel & Ritov 1988, Birgé & Massart 1992, Pollard 1993), by Le Cam's method applied to $\mathcal{P}_1 = \{u^n\}$ and $\mathcal{P}_2 = \mathcal{F}_m$. To be precise, an upper bound on the L^1 distance is obtained between u^n and the mixture of the product measures of the densities indexed by the vertices of the hypercube. Hence we can derive a lower bound on $\sup_{P_i \in co(\mathcal{P}_1)} \|P_1 \wedge P_2\|$. Denote by $h_n(x^n) = 2^{-m} \sum_\tau \prod_{i=1}^n f_\tau(x_i)$ the mixture of the product measures. Then the L^1 distance between u^n and h_n can be bounded, for example, by Pollard (1993) or by Birgé & Massart (1992) as

$$\|u^n - h_n\|_1^2 \leq \exp(2^{-1} n^2 \sum_j (\int g_j^2)^2) - 1. \quad (5)$$

Note that

$$n^2 \sum_j (\int g_j^2)^2 = c^4 a^2 n^2 m^{-9},$$

and if we choose $m = O(n^{2/9})$ and c small, then there is an $\epsilon > 0$ such that

$$\|u^n - h_n\|_1^2 \leq \exp(2^{-1} n^2 \sum_j (\int g_j^2)^2) - 1 < (2(1 - 2\epsilon))^2.$$

Hence

$$\|u^n \wedge h_n\| = 1 - \|u^n - h_n\|_1/2 \geq 1 - (2 - 2\epsilon)/2 = \epsilon > 0.$$

Thus by Le Cam's method (Lemma 1) and because

$$T(u) = 0, \quad T(f_\tau) = c^2 b m^{-2} = O(n^{-4/9}),$$

we have a lower bound decreasing at the slower $n^{-4/9}$ rate, which turns out to be the achievable rate (Bickel & Ritov 1988).

Example 2 Consider estimation of the whole density $\theta(f^n) = \theta(f) = f$ as an element of the space $\mathcal{D} = \{ \text{densities on } [0,1] \}$ equipped with the Hellinger metric, defined by

$$H^2(f, g) = \int_0^1 \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 dx.$$

That is, $d(f, g) = H(f, g)$.

Denote by A_j the j th subinterval of size $1/m$ of $[0,1]$ and let

$$d_j(f, g) = \int_{A_j} (\sqrt{f} - \sqrt{g})^2, \text{ then } d(f, g) = \sum_j d_j(f, g).$$

Note that d_j are not pseudo-metrics, but non-negative symmetric functions satisfying the weaker triangle inequality with a universal constant $A = 1/2$. Therefore Assouad's method (Lemma 2) applies with an extra factor $1/2$ in the lower bound.

Since on A_j

$$(\sqrt{f_\tau} + \sqrt{f_{\tau^j}})^2 = f_\tau + f_{\tau^j} + 2\sqrt{f_\tau f_{\tau^j}} \leq 2(f_\tau + f_{\tau^j}) = 4.$$

$$\begin{aligned} d_j(f_\tau, f_{\tau^j}) &= \int_{A_j} (\sqrt{f_\tau} - \sqrt{f_{\tau^j}})^2 \geq 4^{-1} \int_{A_j} (2g_j(x))^2 dx \\ &= c^2 am^{-5} \equiv \alpha_m. \end{aligned}$$

Note that for $\tau \sim \tau'$, $\|P_\tau \wedge P_{\tau'}\| = \|f_\tau^n \wedge f_{\tau'}^n\| \geq O((1 - O(m^{-5}))^n)$. Plugging this and α_m into the expression in Assouad's Lemma and maximizing the lower bound by choosing $m = O(n^{1/5})$, we obtain a lower bound of order $O(n^{-4/5})$, which is achieved by a kernel estimator with binwidth $O(n^{-1/5})$; hence the rate is optimal.

Since all the densities in \mathcal{M}_m are bounded from below by $1 - c_g$ for $c_g = c \cdot \sup_x |g(x)|$, one can bound the K-L divergence from above as follows

$$\begin{aligned} K(f_\tau^n, f_{\tau'}^n) = nK(f_\tau, f_{\tau'}) &\leq n \int_0^1 \frac{(\sqrt{f_\tau} - \sqrt{f_{\tau'}})^2}{f_\tau} dx \\ &\leq n(1 - c_g)^{-1} \int_0^1 (\sqrt{f_\tau} - \sqrt{f_{\tau'}})^2 dx \\ &\leq 2n(1 - c_g)^{-1} m \alpha_m \equiv m \gamma_m \end{aligned}$$

where $\gamma_m = 2(1 - c_g)^{-1} n \alpha_m$. Note also that

$$d(\theta(f_\tau), \theta(f_{\tau'})) = H^2(f_\tau, f_{\tau'}) = \sum_j \int_{A_j} (\sqrt{f_\tau} - \sqrt{f_{\tau'}})^2 dx \geq \alpha_m W(\tau, \tau').$$

Recalling Lemma 4, and choosing $m = O(n^{1/5})$ we obtain a lower bound of the optimal order $O(n^{-4/5})$. Therefore, both Fano's method and Assouad's lemma give the optimal rate lower bound for this problem.

Acknowledgments: This work began when I was visiting Yale University in the spring of 1993. I would like to thank members of the Statistics Department for a friendly working environment and Professor David Pollard in

particular for many stimulating discussions on related topics and for many helpful comments on the draft. Thanks are also due to Professor Sergio Verdú for commenting on the draft.

Research supported in part by ARO Grant DAAL03-91-G-007.

29.4 REFERENCES

- Assouad, P. (1983), 'Deux remarques sur l'estimation', *Comptes Rendus de l'Academie des Sciences, Paris, Ser. I Math* **296**, 1021–1024.
- Bickel, P. J. & Ritov, Y. (1988), 'Estimating integrated squared density derivatives: sharp best order of convergence estimates', *Sankhyā: The Indian Journal of Statistics, Series A* **50**, 381–393.
- Birgé, L. (1986), 'On estimating a density using Hellinger distance and some other strange facts', *Probability Theory and Related Fields* **71**, 271–291.
- Birgé, L. & Massart, P. (1992), Estimation of integral functionals of a density, Technical Report 024-92, Mathematical Sciences Research Institute, Berkeley.
- Bretagnolle, J. & Huber, C. (1979), 'Estimation des densites: risque minimax', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **47**, 119–137.
- Cover, T. M. & Thomas, J. A. (1991), *Elements of Information Theory*, Wiley, New York.
- Devroye, L. (1987), *A Course in Density Estimation*, Birkhäuser, Boston.
- Donoho, D. L. & Liu, R. C. (1991), 'Geometrizing rates of convergence, II', *Annals of Statistics* **19**, 633–667.
- Donoho, D. L. & Nussbaum, M. (1990), 'Minimax quadratic estimation of a quadratic functional', *Journal of Complexity* **6**, 290–323.
- Fan, J. (1991), 'On the estimation of quadratic functionals', *Annals of Statistics* **19**, 1273–1294.
- Gilbert, E. N. (1952), 'A comparison of signaling alphabets', *Bell System Technical Journal* **31**, 504–522.
- Han, T. S. & Verdú, S. (1994), 'Generalizing the Fano inequality', *IEEE Transactions on Information Theory* **40**, 1247–1251.
- Has'minskii, R. & Ibragimov, I. (1978), On the non-parametric estimation of functionals, in P. Mandl & M. Hušková, eds, 'Prague Symposium on Asymptotic Statistics', North Holland, Amsterdam, pp. 41–52.

- Ibragimov, I. A. & Has'minskii, R. Z. (1981), *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.
- Le Cam, L. (1973), 'Convergence of estimates under dimensionality restrictions', *Annals of Statistics* 1, 38–53.
- Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.
- Pollard, D. (1993), Hypercubes and minimax rates of convergence, Technical report, Yale University.
- Stone, C. (1984), 'An asymptotically optimal window selection rule for kernel density estimates', *Annals of Statistics* 12, 1285–1297.