ISSN 0873-4275

INSTITUTO NACIONAL DE ESTATÍSTICA

PORTUGAL

REVISTA DE ESTATÍSTICA

STATISTICAL REVIEW

23rd European Meeting of Statisticians
Funchal, August 2001
Invited Papers

VOLUME II
2º QUADRIMESTRE 2001

EDIÇÃO ESPECIAL

REVISTA DE ESTATÍSTICA

STATISTICAL REVIEW

Acknowledgements

- Fundação Calouste Gulbenkian

and

- Ministério da Ciência e da Tecnologia
- FCT — Fundação Para a Ciência e Tecnologia
  Apoio do Programa Operacional Ciência, Tecnologia, Inovação do Quadro Comunitário de Apoio III

have sponsored the publishing process of this special issue of
Revista de Estatística – Statistical Review,
proceedings of the
23rd European Meeting of Statisticians
Tecnopolo Funchal, Madeira, Portugal
2001 August 13-18

The organizers of the *23rd European Meeting of Statistics* express their gratitude to the sponsors:

- Universidade de Lisboa
  - FCUL – Faculdade de Ciências da Universidade de Lisboa
  - DEIO – Departamento de Estatística e Investigação Operacional da FCUL
  - CEAUL – Centro de Estatística e Aplicações da Universidade de Lsiboa
- Universidade da Madeira
  - DMUma – Departamento de Matemática da Universidade da Madeira
  - CITMA – Centro de Inovação Tecnológica da Madeira
- INE – Instituto Nacional de Estatística
  - Revista de Estatística – Statistical Review
- SPE – Sociedade Portuguesa de Estatística
- Ministério da Ciência e da Tecnologia
  - FCT – Fundação Para a Ciência e Tecnologia

    Apoio do Programa Operacional Ciência, Tecnologia, Inovação do Quadro Comunitário de Apoio III
- Presidência do Governo Regional, Região Autónoma da Madeira
  - Secretaria Regional de Educação, Região Autónoma da Madeira
  - Secretaria Regional do Plano e Coordenação, Região Autónoma da Madeira
  - Secretaria Regional do Turismo e Cultura, Região Autónoma da Madeira
- Fundação Calouste Gulbenkian
- Fundação Berardo
- Câmara Municipal do Funchal
- Bitranlis - Agentes Transitários, Lda
- CTT – Correios de Portugal
- Cimentos Madeira, Lda
- Caixa Geral de Depósitos
- B.I.C. - Banco Internacional de Crédito
- TAP Air Portugal
- Livraria Escolar Editora
- Timberlake Consultants

Lisboa, 2001 March 30th

WELCOME TO MADEIRA

Survival is the art of taking the appropriate decisions at the right opportunities, using the available information in a sensible way. From that point of view, it was arguably its superior ability to evaluate risks and probabilities that gave the human species a break in the evolutionary process.

Scientific research is now widely recognized as the most rewarding single investment of mankind, and I feel proud that the Government of Madeira, under my leadership, gave such a high priority to education and to research, realizing that this is a major asset for our future welfare.

Probability and Statistics, being exciting sciences in themselves, are now unavoidable tools for the development of all sciences. We may indeed claim that the modern paradigm of Science changed with the development of statistical inference, and that experimental design was a major revolution in scientific methodology. New developments in Statistics will certainly change the methodology of Science further, as well as our views of Society, improving our capacity of inventing order out of apparent surface chaos.

As a politician, and as a former journalist, I am well aware of the role of Probability and Statistics in the process of gathering and evaluating information, of extracting knowledge from information, of decision making based on the sensible use of incomplete (and sometimes messy) information. Statistics is a cornerstone of democracy and of full citizenship, and we feel grateful to those who develop the right tools and models we all need and use.

We feel therefore happy to host the *23rd European Meeting of Statisticians*. Welcome to Madeira. I wish you a constructive Meeting, and I expect that participants and their families will have a memorable time off in Madeira.

Alberto João Jardim,
President of the Regional Government of Madeira

Foreword

This special issue of *Revista de Estatística – Statistical Review* contains the extended abstracts of the invited papers for the *23rd European Meeting of Statisticians*, Funchal, 13-18 August 2001, and the papers submitted to the jury of the Berardo Foundation Prize "Young Statistician 2001" We wish to express our gratitude to the members of the Jury,.

Laurens de Haan (Erasmus University, Rotterdam)
Arnoldo Frigessi (Oslo Un. and Norwegian Computing Centre)
Nils Lid Hjört (Oslo University)
Isaac Meilijson (Tel-Aviv University)
Michael Sørensen (Copenhagen University)
M. Antónia Amaral Turkman (Lisbon University)

for the hard work involved.

FCT – Fundação Para a Ciência e Tecnologia and *Calouste Gulbenkian Foundation* sponsored the *EMS 2001*, and so underpinned the preparation of this special issue of *Revista de Estatística – Statistical Review.*

The *23rd European Meeting of Statisticians*, organized under the auspices of the European Regional Committee of the Bernoulli Society, has been a joint venture of the University of Lisbon, the University of Madeira and INE – Instituto Nacional de Estatística. The invited programme was organized by the Programme Committee:

Anthony Davison (Lausanne), Chairman
Isaac Meilijson (Tel-Aviv)
Mauro Piccioni (L'Acquila, Rome)
Nils Lid Hjört (Oslo)
Olle Häggström (Göteborg)
Teresa Alpuim (Lisboa)

with help from the *Bernoulli Society* representative Arnoldo Frigessi (Oslo).

We wish to express our warm thanks to all the invited speakers and session organizers for their contribution to the high scientific standard of EMS 2001.

We wish to express our gratitude to Mr. Nuno Barreto for his careful retyping and editing part of the papers, and to Mrs. Liliana Martins for her skill in desktop publishing.

Anthony Davison
Adrião Ferreira da Cunha
Isabel Fraga Alves
Dinis Duarte Pestana

Lisboa, 2001 May 12th

GLOBAL INDEX
ÍNDICE GLOBAL

PROBABILITY AND STATISTICS IN BIOINFORMATICS

PROBABILITY APPROXIMATIONS FOR RARE EVENTS

QUANTUM PROBABILITY AND STATISTICS

RECENT DEVELOPMENTS IN TIME SERIES

INFORMATIONS ON STATISTICAL REVIEW
INFORMAÇÕES SOBRE A REVISTA DE ESTATÍSTICA

REVISTA DE ESTATÍSTICA
13ª PÁGINA
2º QUADRIMESTRE DE 2001

PART I
INVITED LECTURES

REVISTA DE ESTATÍSTICA
15ª PÁGINA
2º QUADRIMESTRE DE 2001

FORUM LECTURE

Peter J. Donnelly

OPENING LECTURE

Frank den Hollander

CLOSING LECTURE

Andrew Barron

REVISTA DE ESTATÍSTICA
17ª PÁGINA
2º QUADRIMESTRE DE 2001

Some Statistical Challenges in Modern Genetics

Peter Donnelly
University of Oxford
Department of Statistics,
1 South Parks Road, Oxford OX1 3TG
UK
donnelly@stats.ox.ac.uk

Driven by the genome projects, the advent of high-throughput experimental techniques means that there are growing amounts of data which document DNA sequence variation between individuals within populations. In principle, such data shed light on the evolutionary processes themselves, and on the past history of the relevant populations. Such data from humans is often augmented by information about the disease status of the individuals involved (for one or sometimes several diseases). In this case there is also information about the genetic basis of the diseases in question.

Sensible interpretation of this kind of data represents a considerable statistical challenge. Natural probability models for observed data arise from stochastic processes which model the evolution of the population. Such processes have been the subject of intensive study over several decades, either forward in time, typically as measure-valued diffusions, or backwards in time, via the coalescent, a random tree which describes the ancestral relationships amongst sampled sequences.

Although the structure of the stochastic models is rather well understood, there are no explicit expressions for probabilities of interest, and hence for likelihoods. On the other hand, it turns out that at a single chromosomal location, DNA sequences from distinct individuals are highly positively correlated. This severely limits the information in such data, and puts a premium on the use of efficient, ideally likelihood-based, inference methods.

We will give an informal introduction into the stochastic models which arise in this context, and survey recent computationally-intensive statistical procedures for approximating likelihood surfaces for population genetics data. Several substantive applications will be described with the aim of illustrating the vital role played by modern statistical science in answering key scientific questions.

Random Polymers

Frank den Hollander
EURANDOM
P.O. Box 513
5600 MB EINDHOVEN
The Netherlands
DENHOLLANDER@EURANDOM.TUE.NL

Polymer chains can be modelled as random processes with a self-interaction. Typically this self-interaction is long-ranged in space and/or time. As such, polymers represent a new area in probability theory, different from the more classical areas of Markov chains, Gibbs random fields, spin-flip systems and the like, where the interaction is short-ranged.

In this talk we present a mini-review describing the progress that has been made over the past 15 years in the mathematical analysis of polymer chains. We discuss five examples:

1. soft polymers
2. elastic polymers
3. charged polymers
4. heteropolymers
5. branching polymers

For each of these examples we present the main theorems and the main conjectures, focussing on the scaling behaviour of the polymer in the limit as its length becomes large. It turns out that the long-range nature of the self-interaction leads to a remarkable dependence on the dimension and on the interaction parameters.

Information Theory in Probability and Statistics

Andrew Barron
Yale University
Department of Statistics
New Haven, CT 06520-8290
andrew.barron@yale.edu

The role of information theory in probability limit theorems and mathematical statistics is reviewed. In probability theory, I discuss three themes to the use of information theory. The first concerns use of a simple chain rule to identify and characterize limits of Markov chains, martingales, and information projections and associated Pythagorean inequalities for convex sets of distributions. The second concerns characterization of large deviation exponents for sample averages and empirical distributions and its relationship to conditional limit theorems and concentration inequalities. The third concerns central limit theorems in which measures of information and their derivatives provide natural proofs of convergence to the normal distribution. The thread binding these areas of probability is the use of increments of information to establish convergence and characterize the limit.

Information theory and, in particular, data compression theory provide equally important tools for mathematical statistics. Efficiency, minimax rates, Bayes asymptotics, and model selection criteria are some of the statistical topics fruitfully addressed from this perspective. We briefly discuss some results for exponential families and recent results for mixture model estimation made possible by examining increments of information. In particular, each new component in a mixture sufficiently increases the likelihood (and decreases the information divergence from a target density) to provide information divergence of order 1/K using a K component mixture.

References

Andrew Barron (1986), ``Entropy and the Central Limit Theorem,'' *Annals of Probability*, **14**, p.336-342.

Andrew Barron (1998), ``Information-theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems,'' in *Bayesian Statistics*, **6**, p.27-52.

Andrew Barron (2000), ``Limits of Information, Markov Chains, and Projection,'' Proceedings of the *IEEE International Symposium on Information Theory*, Sorrento, Italy, p.25.

Andrew Barron, Lucien Birge, and Pascal Massart (1999), ``Risk Bounds for Model Selection via Penalization,'' Probability Theory and Related Fields, 113, p.301-413.

Andrew Barron and Nicolas Hengartner (1998), ``Information Theory and Superefficiency,'' *Annals of Statistics*, **26**, p.1800-1825.

Andrew Barron, Jorma Rissanen, and Bin Yu (1998), ``The Minimum Description Length Principle in Coding and Modeling,'' *IEEE Transactions on Information Theory*, **44**, p.2743-2760.

Andrew Barron, Mark Schervish, and Larry Wasserman (1999), ``The Consistency of Posterior Distributions in Nonparametric Problems,'' *Annals of Statistics*, **27**, p.536-561.

Andrew Barron and Chyong-Hwa Sheu (1991), ``Approximation of Density Functions by Sequences of Exponential Families,'' *Annals of Statistics*, **19**, p.1347-1369.

R. Bell and Thomas Cover (1980) ``Competitive Optimality of Logarithmic Investment,'' *Mathematics of Operations Research*, **5**, p.161-166.

Thomas Cover and Joy Thomas (1991). *Elements of Information Theory*, Wiley.

Imre Csiszar (1984), ``Sanov Property, Generalized I-Projection and a Conditional Limit Theorem,'' *Annals of Probability*, **12**, p.768-793.

J. Fritz (1973), ``An Information-Theoretical Proof of Limit Theorems for Reversible Markov Processes,'' in *Transactions of the Sixth Prague Conference on Information Theory*, Statistical Decision Functions, and Random Processes, Prague, Sept. 1971, Academia Publ., Czech Acad. Science.

Oliver Johnson (2000), ``Entropy Inequalities and the Central Limit Theorem,'' *Stochastic Processes and their Application*, **88**, p.291-304.

John Kieffer (1991), ``An Almost Sure Convergence Theorem for Sequences of Random Variables Selected from Log-Convex Sets, in Almost Everywhere Convergence II, Academic Press.

S. Kullback, J. C. Keegal, and J. H. Kullback (1980), *Topics in Statistical Information Theory*, Springer-Verlag.

Jonathan Li (1999), Estimation of Mixture Models, Ph.D. Dissertation, Department of Statistics, Yale University.

Jonathan Li and Andrew Barron (2000), ``Mixture Density Estimation,'' in *Advances in Neural Information Processing Systems*, **12**, MIT Press, p.279-285.

Yu. V. Linnik (1959), ``An Information-Theoretic Proof of the Central Limit Theorem with the Lindeberg Condition,'' *Theory of Probability and Its Applications*, **4**, p.288-299.

Pascal Massart (2000), ``Some Applications of Concentration Inequalities to Statistics,'' Universite Paris-Sud, Preprint.

Fleming Topsoe (1979). ``Information Theoretical Optimization Techniques,'' *Kybernetika*, **15**, p.8-27.

Yuhong Yang and Andrew Barron (1999), ``Information-Theoretic Determination of Minimax Rates of Convergence,'' *Annals of Statistics*, **27**, p.1564-1599.

REVISTA DE ESTATÍSTICA
21ª PÁGINA
2º QUADRIMESTRE DE 2001

SPECIAL INVITED LECTURES

Søren Asmussen
Veronique Gayrard
Irène Gijbels

REVISTA DE ESTATÍSTICA
23ª PÁGINA
2º QUADRIMESTRE DE 2001

Large Deviations and Exponential Tilting in Rare Events Simulation

Søren Asmussen
Lund University
Sweden
asmus@maths.lth.se

The evaluation of small probabilities, of order from 10^{-2} to 10^{-10}, come up in a number of application areas like insurance risk, telecommunications and reliability. The difficulty in evaluating them by simulation lies in controlling the relative error. Formally, let $A(x)$ be a family of events indexed by a parameter x and satisfying $z(x)=P(A(x)) \to 0$, $x \to \infty$. A simulation estimation scheme is then a family of r.v.'s $Z(x)$ which can be generated by simulation and has $EZ(x) = z(x)$ (in practice, the simulation of $z(x)$ is performed by producing N i.i.d. replications of $Z(x)$, using the average as estimator of $z(x)$ and assessing the statistical error by a confidence interval based upon the empirical variance). If we let $\sigma^2(x) = Var(Z(x))$, the relative error is defined as $\sigma(x)/z(x)$, and ideally, one looks for schemes having the property that this quantity remains bounded as $x \to \infty$ or at least grows slower than any negative power of $z(x)$ (this property is referred to as logarithmic efficiency). This fails in particular for the crude Monte Carlo method, where $Z(x)$ is the indicator of $A(x)$ and $\sigma^2(x)$is of order $z(x)$, implying that N must be chosen very large as $z(x)$ becomes small.

The prototype of an algorithm with bounded relative error is Siegmund's 1976 algorithm for estimating the probability $z(x)$ that a random walk with negative drift ever exceeds level x.

If F is the increment distribution, the algorithm uses importance sampling, where F is exponentially twisted with a certain parameter familiar from work of Cramér and Feller.

A similar simple example of exponential tilting is the estimation of the probability that a sum of n i.i.d. terms is much bigger than its mean, where the choice of the tilting parameter is based upon the familiar saddlepoint argument.

In more complex situations, it is usually not obvious how to choose the importance sampling distribution P^*. A general approach is based upon the observation that choosing P^* as P_x, the conditional P-distribution given $A(x)$, would lead to $\sigma^2 = 0$. This choice is not practicable because the likelihood ratio involves $z(x)$ which is unknown, but suggest to try to make P^* as close to P_x as possible. This necessitates a study of the asymptotic form of P_x, in particular of describing the most likely path leading to the rare event, and is often performed using large deviations techniques. Indeed this approach explains the particular form of the exponential tilting in the above two simple examples, as well as it applies to a number of other problems.

Rather recent counterexamples indicate, however, that the idea of involving asymptotics of P_x has its limitations. For example, Glasserman & Wang (1997) found an example in tandem queues (mathematically, a two-dimensional reflected random walk) where a path different form the most likely one gives so large a contribution to σ^2 that the relative error blows up. The problem is associated with the role of the reflecting boundary, which was further investigated by Asmussen, Frantz, Jobmann & Schwefel (2000). They provided further counterexamples, but also an algorithm which deals with the boundary problem in some simple cases.

Exponential tilting requires the existence of sufficiently many exponential moments and is therefore intrinsically impossible in problems involving heavy-tailed distributions like the Pareto. Rare events simulation in this setting was investigated by Asmussen, Binswanger & Højgaard (2000). The most likely path can still be described in many cases (typically, it involves one large jump rather than many slightly biased ones as in the light-tailed case) but it was found that simulating using this asymptotical description typically yields an infinite variance. However, some logarithmically algorithms were exhibited, one based upon order statistics and conditional Monte Carlo, and one upon a different importance sampling scheme. Unfortunately, the class of problems where the algorithms apply is rather limited, and the area of rare events simulation in heavy-tailed settings is still largely open.

References

S. Asmussen & R.Y. Rubinstein (1995) Steady—state rare events simulation and its complexity properties. In Advances in Queueing: Models, Methods & Problems (J. Dshalalow ed.), 429-466. CRC Press.

S. Asmussen, K. Binswanger & B. Højgaard (2000) Rare events simulation for heavy--tailed distributions. *Bernoulli* **6**, 303-322.

S. Asmussen, P. Frantz, M. Jobmann & H.P. Schwefel (2000) Large deviations and fast simulation in the presence of boundaries. Submitted.

P. Glasserman & Y. Wang (1997) Counterexamples in importance sampling for large deviations probabilities. *Ann. Appl. Probab.* **7**, 731-746.

D. Siegmund (1976) Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.* **4**, 673-684.

Statistical Mechanics of the Hopfield Model: Equilibrium and Dynamical Aspects

Véronique Gayrard
Univ. Lausanne
gayrard@masg1.epfl.ch

We intend to give a survey of the statistical mechanics spproach to the analysis of the Hopfield model of neural networks. This model is intended to describe in a very simplified way the the process of retrieval of information in an associative memory. The functioning of the memory will be seen to be described by a family of reversible Markov chains in random environment, indexed by a large parameter. We will first focus on the study of the invariant measure of the chains - the so-called Gibbs measure -: this corresponds to the analysis of the equilibrium statistical mechanics of the model.The second part of the talk will be devoted to an investigation of dynamical phenomena such as metastability and, depending on time, the aging phenomenon.

This talk will be based on joint works with G. Ben Arous, A. Bovier, M. Eckhoff, and M. Klein.

REVISTA DE ESTATÍSTICA
25ª PÁGINA
2º QUADRIMESTRE DE 2001

Nonparametric Function Estimation and Discontinuities

Irène Gijbels
Université catholique de Louvain, Institut de Statistique
Voie du Roman Pays 20
B-1348 Louvain-la-Neuve, Belgium
gijbels@stat.ucl.ac.be

The literature on nonparametric estimation of functions, as for example densities, hazard and regression functions, is very vast. It is often assumed that the functions of interest are smooth, i.e. are continuous, or have a certain number of continuous derivatives. Therefore, most statistical methods are designed for such smooth functions, and theoretical results also focus on classes of smooth functions.

In a number of important applications though it is more realistic to allow for a model in which the function of interest is smooth except at a finite number of locations where it shows, for example, a jump discontinuity. Examples of regression functions with jump discontinuities (in the function itself or in its derivatives) are the crown-heel lengths growth data (Lampl, Veldhuis and Johnson (1992)) and the stock market data (Wang (1995)). Examples of hazard functions with discontinuities are the heart transplant data (Miller and Halpern (1982)) and the leukaemia data (Brochstein et al. (1987)). Regression or hazard analysis of such data require special techniques. Indeed, application of common smoothing methods would lead to smooth curve estimates. When dealing with unsmooth regression or hazard functions there are various aspects of inference that are of importance: estimation of the locations of the discontinuities, estimation of the "jump sizes", estimation of the entire function of interest, testing for smooth versus unsmooth functions, confidence bands for the unsmooth function, etc. It has been shown in the past years, how to adapt the classical nonparametric estimation techniques, such as for example kernel smoothing methods, spline methods and wavelet methods, to the situation of possible unsmooth functions. In this talk we mainly focus on kernel methods and on wavelet deconvolutions, the latter to a lesser extent. An overview of recent developments in this area will be presented. See for example Antoniadis and Gijbels (2001) and Antoniadis, Gijbels and MacGibbon (2000) for estimation of change-points in a regression function or a hazard function respectively, using wavelet decomposition techniques.

In particular we mention the problem of practical choices of `smoothing parameters' involved in inference problems for unsmooth curves. For a kernel-based procedure proposed by Gijbels, Hall and Kneip (1999) we discuss a bootstrap procedure that allows to deal with the crucial issue of choosing the smoothing parameters in estimation as well as testing problems. The performance of the bootstrap-based testing procedure has been compared with that of asymptotic testing procedures such as those proposed by Müller and Stadtmüller (1999) and Grégoire and Hamrouni (2001). See also Gijbels and Goderniaux (2000, 2001) and references therein.

References

Antoniadis, A. and Gijbels, I. (2001). Detecting abrupt changes by wavelet methods. *Journal of Nonparametric Statistics*, to appear.

Antoniadis, A., Gijbels, I. and MacGibbon, B. (2000). Nonparametric estimation for the location of a change-point in an otherwise smooth hazard function under random censoring. *Scandinavian Journal of Statistics*, **27**, 501-519.

Brochstein, J.A., Kernan, N.A., Groshen, S., Cirrincione, C., Shank, B., Emanual, D., Laver, J. and O'Reilly, R.J. (1987). Allogenic bone marrow transplantation after hyperfractionated total body irradiation and cyclophosphamide in children with acute leukaemia. *New England Journal of Medicine*, **317**, 1618-1624.

Gijbels, I. and Goderniaux, A.-C. (2000). Bandwidth selection for change point estimation in nonparametric regression. Institut de Statistique, Université catholique de Louvain, Discussion Paper 0024.

Gijbels, I. and Goderniaux, A.-C. (2001). Bootstrap tests for detecting change points in nonparametric regression. Manuscript.

Gijbels, I., Hall, P. and Kneip, A. (1999). On the estimation of jump points in smooth curves. *The Annals of the Institute of Statistical Mathematics*, **51**, 231-251.

Grégoire, G. Z. and Hamrouni, Z. (2001). Two nonparametric tests for change-point problems. *Journal of Nonparametric Statistics*, to appear.

Lampl, M., Veldhuis, J.D. and Johnson, M.L. (1992). Saltation and stasis: a model of human growth. *Science*, **258**, 801-803.

Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika*, **69**, 521-531.

Müller, H.-G. and Stadtmüller, U. (1999). Discontinuous versus smooth regression. *The Annals of Statistics*, **27**, 299-337.

Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika*, **82**, 385-397.

REVISTA DE ESTATÍSTICA
27ª PÁGINA
2º QUADRIMESTRE DE 2001

INVITED THEMATIC SESSIONS

REVISTA DE ESTATÍSTICA
29ª PÁGINA
2º QUADRIMESTRE DE 2001

ASYMPTOTIC STATISTICS

Organizer: Aad van der Vaart

Invited Speakers: Piet Groeneboom
Guenther Walther
Lucien Birgé

REVISTA DE ESTATÍSTICA
31ª PÁGINA
2º QUADRIMESTRE DE 2001

Vertex Direction Algorithms for Computing Nonparametric Function Estimates

Piet Groeneboom
Delft University
Department of Mathematics
2628 CD Delft
The Netherlands
p.groeneboom@its.tudelft.nl

Geurt Jongbloed
Vrije Universiteit
Department of Mathematics
1081 HV Amsterdam
The Netherlands
geurt@cs.vu.nl

In many situations where one wants to estimate a function nonparametrically, the estimate is in fact determined by a relatively small number of parameters. Well-known examples are the nonparametric maximum likelihood estimator of a density under monotonicity or convexity constraints and the least squares estimator of a monotone or convex regression function.

In the case of a monotonicity constraint, we know from the asymptotic distribution theory that the estimator will generally have a number of jumps of order $n^{1/3}$. This means that the number of parameters that has to be estimated is also of order $n^{1/3}$, if n is the sample size. Similarly, in estimating a function under convexity constraints the number of parameters that has to be estimated generally is of order $n^{1/5}$.

This suggests that an efficient algorithm should use this information in such a way that it should try to systematically search for these parameters, possibly starting from scratch with zero parameters. However, the usual algorithms (like the EM algorithm) work with a number of parameters that is at least as large as the sample size. As an example, the EM algorithm will spend most of the time in reducing the "nonrelevant" parameters to zero during its iterations. A similar remark holds for interior point methods (both direct and primal-dual versions), since, for example in the case of monotonicity constraints, the inequality constraints have to be strict during the whole iteration process.

The need for an efficient algorithm is even more pressing in a situation where one wants to estimate nonparametrically a multivariate function. As an example, in the estimation of the nonparametric maximum likelihood estimator of the multivariate distribution function for k-variate interval censored data one essentially has to estimate again a number of parameters of order $n^{1/3}$ (according to recently developed theory), but the usual algorithms (like the EM algorithm again) try to estimate in fact n^k parameters.

This means that even for bivariate interval censored data the computing time of the MLE for a sample size of 500 becomes prohibitive: the EM algorithm did not converge to the solution at an accuracy of 10^{-5} in two full days for this sample size on a Compaq AlphaServer 800 5/500, with a CPU of 21164/500 MHz Alpha, see Song

(2001). Similar findings are reported in Song (2001) w.r.t. the vertex exchange algorithm of Gentleman and Vandal (1999a-b).

It turns out to be possible to develop an algorithm which takes advantage of the fact that the order of the number of relevant of parameters, given by the asymptotic theory, is small in comparison to sample size. In this algorithm (Groeneboom, Jongbloed and Wellner (2001)) we start "from scratch" with zero parameters and do a systematic search for the relevant parameters, keeping the number of parameters during the iterations small. This method can be used both for least squares estimators and maximum likelihood estimators. This algorithm is what we call *"of vertex direction type"* . It is related to the algorithm studied in Simar (1976) and Boehning (1982).

The essence of our algorithm is that we add the parameters one by one, alternating between a search without the constraints and a check whether the solution without the constraints actually has a constraint violation. At each step we check whether we can improve our solution by adding a parameter, corresponding to a certain "direction". For this direction we first compute an unconstrained solution, involving all parameters we have at that iteration step. If there is a constraint violation we remove a parameter which, according to a certain criterion, is the "worst violator". We then compute the unconstrained solution without this violator and make a step in the direction of this new solution. We call this step the *support reduction step* and it is a very essential step of our procedure. So the whole procedure proceeds by alternating between adding new parameters and (if necessary) a parameter reduction step. It is proved that this algorithm cannot run into a loop in the sense that newly added parameters would be removed at the next iteration step.

An initial version of our algorithm was inspired by the "hinge algorithm", described in Meyer (1997), and used in Groeneboom, Jongbloed and Wellner (2000 a-b). However, convergence of the latter algorithm to the solution of the optimization problem has not been proved. For this reason we work with a modification of the support reduction step as used in Meyer (1997) for which we can actually prove convergence.

The theory will in particular be demonstrated for the example of bivariate interval censoring where we use this algorithm in an iterative way by reducing the problem of computing the MLE to an iterative least squares problem, although it is also possible to use the algorithm in a version where a nonlinear optimization problem would be used at each step. But it turns out that a "safe" nonlinear optimization step at each iteration takes more time than using a complete loop of the vertex direction method in solving a least squares problem at each iteration.

References

Boehning, D. (1982). Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process. *Ann. Statist.* **10**, 1006-1008.

Gentleman, R. and Vandal, A.C. (1999a). Computational algorithms for censored data problems using intersection graphs. Submitted.

Gentleman, R. and Vandal, A.C. (1999b). Graph-theoretical aspects of bivariate censored data. Submitted.

Groeneboom, P., Jongbloed, G., and Wellner, J. A. (2000a). A canonical process for estimation of convex functions: the ``invelope" of integrated Brownian motion + t^4. Tentatively accepted by the Annals of Statistics.

Groeneboom, P., Jongbloed, G., and Wellner, J.A. (2000b). Estimation of convex functions: characterizations and asymptotic theory. Tentatively accepted by the Annals of Statistics.

Groeneboom, P., Jongbloed, G., and Wellner, J. A. (2001). Vertex direction algorithms for computing nonparametric functions estimates. *Manuscript in preparation.*

Meyer, M. C. (1997). Shape Restricted Inference with Applications to Nonparametric Regression, Smooth Nonparametric Function estimation, and Density Estimation.
Ph.D. Dissertation, Department of Statistics, University of Michigan, Ann Arbor, Michigan.

Simar, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Ann. Statist.* **4**, 1200-1209.

Song, S. (2001). Estimation with univariate "mixed case" interval-censored data and bivariate interval-censored data: bivariate current status data. Pfh. D. dissertation, University of Washington, Seattle, U.S.A.

An Overview of Some Recent Results in Model Selection via Penalization

Lucien Birgé
Université Paris VI and UMR CNRS 7599 "Probabilités et modèles aléatoires"
Laboratoire de Probabilités
boîte 188 Université Paris VT
4 Place Jussieu F-75252 Paris Cedex
05 France
LB@CCR.JUSSIEU.FR

The purpose of this lecture is to give an account of the modern theory of model selection via penalization, explain the main ideas, some recent results and how to practically implement the method.

References

BARAUD, Y. (2000). Model selection for regression on a fixed design. Probab. *Theory Relat. Fields* **117**, 467-493.

BARAUD, Y., COMTE, F. and VIENNET, G. (1997). Adaptive estimation in an autoregression and geometrical β -mixing framework. *Technical Report* No. **97.75**. Mathématiques, Université Paris-Sud, Orsay.

BARAUD, Y., COMTE, F. and VIENNET, G. (1999). Model selection for (auto-)regression with dependent data. Technical Report LMENS-99-12. Ecole Normale Supérieure, Paris.

BARRON, A.R., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113**, 301-415.

BIRGÉ, L. and MASSART, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics* (D. Pollard, E. Torgersen and G. Yang, eds.), 55-87. Springer-Verlag, New York.

BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. To appear in *JEMS.*

BIRGÉ, L. and MASSART, P. (2001). A generalized C_p criterion for Gaussian model selection. *Technical Report. Lab. de Probabilités*, Université Paris VI.

Bikernel Oscillation Analysis for the Mixture Complexity

Guenther Walther
Dept. of Statistics, Stanford University
Stanford, CA 94305, USA
walther@stat.stanford.edu

The problem under consideration is to determine the number of components in a location mixture. Some prominent examples in the literature are the number of kinds of chondrite in meteorites, the number of different paper types used in the production of certain stamps, or the number of genetic components determining blood pressure.

Powerful theoretical results are available in the case where the class of component distributions is parametric. There is also a large literature on an approach where class of component distributions is nonparametric, motivated by the fact that the parametric approach is sensitive to the structure imposed on the component distributions. For example, if the mixture distribution is skewed and the class of component distributions is the normal family, then many normal components are required in the mixture to pick up the skewness, which can result in a considerable overestimate of the mixture complexity. The nonparametric approaches usually proceed by mode- or bump-hunting, i.e. by establishing a lower confidence bound on the number of modes of f or of `bumps' (maxima of the density derivative). One disadvantage of such an approach is that it is not very sensitive to detect mixing. For example, the means of two homoscedastic normal distributions need to be separated by at least two standard deviations before a mixture becomes bimodal.

Comparing the refined theory for the parametric case with the nonparametric approach suggests that progress in the latter case hinges on a better understanding of the properties of mixtures in the nonparametric setting. This paper develops the requisite theory for the nonparametric approach and the accompanying statistical procedure. It is shown that a simple but powerful criterion is obtained by studying the qualitative behavior of the sum of two certain convolutions, in terms of counting certain upcrossings of this sum. The resulting procedure has the advantage over mode-hunting approaches that it is sensitive to detect mixing in more general unimodal situations, yet it cannot be improved upon in the more restricted situation where mixing manifests itself in multimodality, even if one is allowed to use that knowledge a priori. This is explained heuristically and made precise in the asymptotic minimax framework.

REVISTA DE ESTATÍSTICA
35ª PÁGINA
2º QUADRIMESTRE DE 2001

BAYESIAN NONPARAMETRICS

Organizer: Stephen Walker

Invited Speakers: Subhashis Ghosal
Donato M. Cifarelli, Pietro Mulieri and Sonia Petrone
Paul Damien

REVISTA DE ESTATÍSTICA
37ª PÁGINA
2º QUADRIMESTRE DE 2001

Bayesian Nonparametric Inference

Stephen Walker
Imperial College
London
massgw@maths.bath.ac.uk

Bayesian nonparametrics is concerned with constructing prior probabilities on spaces of densities or distributions which cover more of these type of functions than those provided by a prior probability which puts all mass on distributions indexed by a nite dimensional parameter. One interpretation of the work of de Finetti (1938) is that we should put priors which have all distributions in the support. Due to the previous technical diÆculties involved in doing this, common practice was to solve the problem by limiting the support to a set of parametric distributions, and to induce a prior on this set of distributions by constructing the prior on the parameter space. Not doing this, we enter the realms of Bayesian nonparametrics.

Modern Bayesian nonparametrics took off following the paper of Ferguson (1973) which formalised the notion of a Dirichlet process. These processes have sample paths which are distribution functions and hence the prior is guaranteed to exist because the process exists. This generated a lot of work on using stochastic processes for characterising nonparametric priors; Doksum (1974), Dykstra and Laud (1981), Hjort (1990), for example.

A lot of applied work has concentrated on the Mixture of Dirichlet Process (MDP) model following the introduction of Markov chain Monte Carlo methods. Pioneer work was done by Escobar (1994), Escobar and West (1995), and MacEachern (1994). It is fair to say that a lot of the current work being done in Bayesian nonparametrics is concerned with developing sampling algorithms for estimating nonparametric models.

Bayesian nonparametrics offers modelling which, these days, are no more dfficult to work with or understand than parametric models. A recent review and comprehensive list of references is given in Walker et al. (1999). The key is that Bayesian nonparametrics can be done and hence the need to look for suitable parametric models is obviated. Also avoided is the need to keep reassigning probability 1 to parametric models until a suitable model is found, if ever.

A topic receiving a lot of attention in recent years is the notion of Bayesian consistency. This is a tough area, the mathematics being quite difficult. The task is to construct priors which lead to posterior distributions accumulating in neighbourhoods of the true model, as the sample size grows. Bayesians are divided as to whether this is an important property or not. A good review with motivation is given by Wasserman (1998).

References

de Finetti, B. (1938). Sur la condition d'equivalence partielle. VI Colloque Geneve. *Act. Sci. Ind.* **739**. Herman, Paris.

Doksum, K.A. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2**, 183-201.

Dykstra, R.L. and Laud, P.W. (1981). A Bayesian nonparametric approach to reliability. *Ann. Statist.* 9, 356-367.

Escobar, M.D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89**, 268-277.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577-588.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209-230.

Hjort, N.L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18**, 1259-1294.

MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communication in Statistics: Simulation and Computation.* **23**, 727-741.

Walker, S.G., Damien P., Laud, P.W. & Smith, A.F.M. (1999). Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society Series B* **61**, 485-527.

Wasserman, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey, P. Müller and D. Sinha, eds.), 293-304. *Lecture Notes in Statistics, Springer.*

REVISTA DE ESTATÍSTICA
39ª PÁGINA
2º QUADRIMESTRE DE 2001

Asymptotics for Bayesian Density Estimation with Mixtures

Subhashis Ghosal*
University of Minnesota
ghosal@stat.umn.edu

Mixtures parametric family of densities give us simple flexible non parametric classes of densities, and are very useful for model based inference such as the Bayesian method. One can. then induce a prior distribution on the densities simply by specifying a prior distribution on the mixing distribution. This approach was first used by Ferguson (1983) and Lo (1984), who used a Dirichlet process prior for the mixing distribution and gave expressions for posterior expectations of functions. These expressions however involve too many terms, and are unsuitable for computations. To address computational issues, Gibbs sampling techniques to compute the posterior mean and other posterior characteristics have been developed, see, for example, Escobar and West (1995) and the references therein.

In this article, we are concerned with consistency and the rate of convergence of posterior distribution of the density in mixture models. Informally, the posterior is said to be consistent at a true density p_0 if the posterior distribution on the density p, as a random distribution, converges to the degenerate measure as more and more data are generated from the density p_0. Equivalently, the posterior probability of any neighborhood of p_0 tends to one. To measure the, rate of convergence, the neighborhood is let shrink with the increasing sample size. The minimum size of the neighborhood that holds most of the posterior probability is defined as the rate.

A general theorem of Schwartz (1965) asserts that posterior is consistent at a given point if the complement of any neighborhood of the true parameter can be tested against the point null with exponentially small probability of errors and every Kullback-Leibler neighborhood of the true parameter has positive prior probability. While Schwartz's theorem is very useful for the verification of consistency in many nonparametric problems, a test with uniformly small probability of errors does not exist for densities with variation distance on it due to the lack of compactness. Barron (1986) pointed out that the condition could be relaxed by intersecting the complement of the neighborhood of the true density with a set whose complement has exponentially small prior probability. Barron, Schervish and Wasserman (1999) showed that for densities with the Hellinger distance on it, suitable tests may be constructed from bounds for the bracketing entropy. Ghosal, Ghosh and Ramamoorthi (1999) constructed the desired tests based on merely a similar bound for the metric entropy with respect to the variation distance. Ghosal, Ghosh and van der Vaart (2000) studied the convergence rate of posterior distribution. They showed that the rate is obtained as the maximum of the solution of the entropy .equation that gives the "best rate" for estimators and the concentration rate of the prior probability at the true density.

---

* Most of the results reviewed here are collected from several papers written in collaboration with J. K. Ghosh, R. V. Ramamoorthi and Aad van der Vaart.

For density estimation on the real line, mixtures of normals form a flexible class. Ghosal, Ghosh and Ramamoorthi (1999) showed that Dirichlet mixture of normal prior gives rise to a consistent posterior under general conditions for the weak topology and the variation distance. If the true density is a compact location mixture of normal, or if it is itself compactly supported and satisfies Kullback-Leibler continuity, then the posterior is weakly consistent. Compactness of the support can be dispensed by bounding the tail of the mixture density using Doss and Sellke's (1982) bounds for the Dirichlet probability. For consistency in the variation distance, one has to consider suitable sieves. If the base measure has only thin tails and the prior density for the scale of normal has high degree of contact at 0, then consistency in variation holds. For instance, this holds if the base measure is normal and σ^2 of the normal kernel has a truncated inverse gamma prior.

Ghosal and van der Vaart (2001) showed that the posterior converges at a nearly parametric rate if the true density generating the observations is also a mixture of normals with standard deviations bounded away from zero and infinity. They considered both location and location-scale mixtures of normals. For the location mixture, the rate turns out to be $(\log n)^k/\sqrt{n}$, where $k \geq 1$ is a constant depending on the tail of the base measure.

On the real line, the favorite kernel is the normal density, although other kernels have also been used. Gasparini (1992) considered random histograms obtained as a particular type of mixtures of the uniform kernel. He called this the Dirichlet density process and showed consistency under finiteness of variance and a "Kullback-Leibler continuity" condition. Special kernels may also be used to yield special shape of mixtures. A scale mixture of uniforms will give rise to a decreasing density on the positive half line. Symmetrization of this is a symmetric strongly unimodal density, and is often a reasonable model for error distribution. Brunner and Lo (1989) used this idea and attached a prior by considering a Dirichlet process prior for the mixing distribution.

If the density is defined on a bounded interval, normal mixtures are no longer appropriate. On the unit interval, beta densities form a flexible class. Indeed, by Bernstein polynomial approximation, mixtures of only some special beta's will be able to approximate any continuous density on the unit interval. Petrone (1999) used this idea to construct a prior based on the Bernstein polynomials by putting an appropriate prior on the index and the coefficients. Petrone and Wasserman (2000) showed that the Bernstein prior is weakly consistent provided it has full support. The conclusion can be easily strengthened to consistency for the variation distance if the prior for the index decays rapidly. Ghosal (2001) obtained the rates of convergence of posterior for the Bernstein polynomial prior: If the true density is itself Bernstein polynomial and the prior for the index has exponentially decaying tails, then the posterior converges at a rate $(\log n)/\sqrt{n}$. In general, if the true density is bounded below and twice continuously differentiable, then the posterior converges at a rate $n^{-1/3}(\log n)^{5/6}$.

Mixtures of other kernels may be considered. On the positive half line, mixtures of gamma, Weibull or lognormal densities form flexible families. Densities on the half line are of interest in many situations. In particular, survival functions may be modeled using such mixtures. This gives a convenient method for Bayesian survival analysis, where the mixing distribution could be given a Dirichlet prior. The study of consistency and rate of convergence for these priors will be of substantial interest.

References

BARRON, A. R. (1986). On uniformly consistent tests and Bayes consistency. Unpublished manuscript.

BARRON, A. R., SCHERVISH, M. and WASSERMAN, L. (1999). The consistency of posterior distributions in non parametric problems. *Ann. Statist.* **27** 536-561.

BRUNNER, L. J. and Lo, A. Y. (1989). Bayes methods for a symmetric unimodal density and its mode. *Ann. Statist.* **17** 1550-1566.

Doss, H. and SELLKE, T. (1982). The tails of probabilities chosen from a Dirichlet prior. *Ann. Statist.* **10** 1302-1305.

ESCOBAR, M. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577-588.

FERGUSON, T. S. (1983). Bayesian density estimation by mixtures of Normal distributions. In *Recent Advances in Statistics* (Rizvi M., Rustagi, J. and Siegmund, D., Eds.) 287-302.

GASPARINI, M. (1992). Bayes Nonparametrics for biased sampling and density estimation. *Ph. D. thesis*, University of Michigan.

GHOSAL, S. (2001). Convergence rates for density estimation with, Bernstein polynomials. *Ann. Statist.* (to appear).

GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27** 143-158.

GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). -Convergence rates of posterior distributions. *Ann. Statist.* **28** 500-531.

GHOSAL, S. and VAN DER VAART, A. W. (2001). Entropies and rates of convergence of maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* (To appear).

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates I: Density estimates. *Ann. Statist.* **12** 351-357.

PETRONE, S. (1999). Bayesian density estimation using Bernstein polynomials. *Canad. J. Statist.* **26** 373-393.

PETRONE, S. and WASSERMAN, L. (2000). Consistency of Bernstein polynomial posteriors. *Preprint*.

SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10-26.

Bayesian Nonparametric Inference for Survival Data

Paul Damien
University of Michigan Business School
pdamien@bus.umich.edu

Nonparametric Bayesian methods will be discussed within the context of survival data. Full Bayesian inference with and without covariates will be exemplified for hazard rate models.

REVISTA DE ESTATÍSTICA
43ª PÁGINA
2º QUADRIMESTRE DE 2001

Predictive Inference: A Review and New Developments

Donato M. Cifarelli, P. Muliere
Bocconi University
Viale Isonzo 25, 20135 Milano, Italy
michele.cifarelli@uni-bocconi.it, pietro.muliere@uni-bocconi.it

Petrone, S.
University of Insubria
Via Ravasi 2, 21100 Varese, Italy
spetrone@eco.uninsubria.it

1. Introduction

The general predictive problem related to a sequence $\{X_n\}$ of random variables involves the evaluation of the probability of an event, dependent on the future realisations of some of the variables of the sequence, when the outcomes of a finite number of variables of the same sequence are assumed to be known. The treatment of this problem which takes into account only strictly observable events has been called *completely predictive.* The recourse to a parametric model, even if it is not necessary, generally simplifies the mathematical aspects of a predictive problem; this approach, which is referred to as *hypothetical*, is the one prevalent in the traditional Bayesian literature. However caution must be adopted when following such an approach; for instance, consistently with de Finetti for whom only observable facts are subject to probabilistic evaluation, it is possible to question the adoption of the hypothetical approach whenever one is not in the position to elicit a prior distribution for the parameter appearing in the model.

2. Parametric versus Nonparametric

The basic predictive assumption for a sequence of random variables is that of exchangeability. De Finetti style theorems characterise models in terms of invariance. The idea is that the statistician begins the model building phase by postulating reasonable symmetries for the distribution of the observable facts.

Let $X_1, X_2, \ldots$ be an exchangeable sequence of random variables defined on $X \subseteq R.$. From de Finetti's representation theorem (de Finetti, 1937) there exists a random distribution function F conditional on which $X_1, X_2, \ldots$ are i.i.d. from F. That is, there exists a probability measure, defined on the space of probability measures on X, such that the joint distribution of $X_1, X_2, \ldots X_n$, for any n, can be written as

$$P(X_1 \in A_1, \ldots X_n \in A_n) = \int \left\{ \prod_{i=1}^{n} F(A_i) \right\} \mu(dF)$$

where μ is the de Finetti (or prior) measure. Therefore, if we assume only the exchangeability, the representation theorem involves an infinite dimensional parameter. This parameter is the weak limit (with probability P one) of the sequence of the empirical distribution functions. In order to justify the dependence of this limit to a finite-

dimensional parameter (hence a parametric approach) further assumptions must be introduced on the observables. For example, as we shall see, the existence of a predictive sufficient statistics.

3. Characterisation of Priors using Predictive Assumptions

The general predictive problem reduces to the computation of the conditional probability

$$P(X_{n+1} \in A \mid X_1, X_2, \cdots X_n)$$

for measurable sets A. The assumption of exchangeability and the representation theorem imply that

$$P(X_{n+1} \in A \mid X_1, X_2, \ldots, X_n) = E(F(A) \mid X_1, X_2, \ldots, X_n).$$

Without further assumptions, we need a prior on the infinite-dimensional parameter F. Unfortunately, when de Finetti (1935) suggested the general predictive approach, non-parametric priors were not known yet. Today, many proposals can be found in the literature, but there remains the problem of how to select the prior. One approach is to select μ by appealing to prior information about F and attempting to incorporate this information into μ. This is often a difficult task for non-parametric priors. Alternatively, we may try to describe our state of knowledge in terms of probabilistic assumptions on X_{n+1} given $X_1, X_2, \cdots X_n$, for n=1,2,..., and consequently characterise the prior μ.

Indeed, when $\{X_n\}$ is an infinite sequence of random variables, the completely predictive approach to the construction of the law of the sequence is based on the specification of the distribution F_1 of X_1 and of the predictive distribution F_{n+1} of X_{n+1} given $X_1, X_2, \cdots X_n$ for all $n \geq 1$. Whereas the Ionescu-Tulcea extension theorem states consistency conditions which guarantee the existence of a unique law for $\{X_n\}$ determined by the sequence $\{F_n\}$, Fortini, Ladelli and Regazzini (2000) give necessary and sufficient conditions on the sequence $\{F_n\}$ for the exchangeability of the law of $\{X_n\}$. This result characterises exchangeability in purely predictive terms; the de Finetti measure of the sequence $\{X_n\}$ is then obtained by means of de Finetti's Representation Theorem.

Many priors used in Bayesian nonparametrics can easily be constructed following this approach; for example, the Dirichlet process (Regazzini, 1978; Lo, 1991), the Polya trees (Walker and Muliere, 1997a), the beta-Stacy (Walker and Muliere, 1997b), the Neutral to the right processes (Walker and Muliere (1999).

4. Predictive Sufficiency

For justifying a parametric approach , we need further assumptions on the observables, in addition to exchangeability. From a predictive point of view, an example of such assumptions is predictive sufficiency. Predictive sufficiency and its properties have been investigated in a number of papers among which: Campanino and Spizzichino (1981), Cifarelli and Regazzini (1980,1981,1982), Dawid (1982), Secchi (1987), Muliere

and Secchi (1992), Fortini, Ladelli and Regazzini (2000). Related notions of sufficiency have been studied by Lauritzen (1984,1988) and Diaconis and Freedman (1984).

In many practical situations, the researcher can assume, in addition to exchangeability, that a statistic T_n summarises all the information provided by $X_1, X_2, \ldots X_n$ for predicting X_{n+1}. Then T_n is called *predictive sufficient statistic.*

When T_n is a linear function, Cifarelli and Regazzini (1982) have shown that, under some hypotheses, the probability law of $X_1, X_2, \ldots X_n$ can be represented by means of a parametric model, where the model F is the limit of the sequence of predictive distributions of X_{n+1} and the prior on the parametrer Θ is the limit law of the sequence $\{T_n\}$. Fortini, Ladelli and Regazzini (2000) relax the hypotheses required.

This result does not say how to select the prior on Θ. Anyway, Muliere and Secchi (1992) show that it is often reasonable to approximate the posterior distribution of Θ given $X_1, X_2, \ldots X_n$ by means of the distribution of $\{T_n\}$ obtained by using a bootstrap procedure. This procedure results equivalent, from the completely predictive point of view, to those obtained by a Bayesian who decides to adopt a suitable improper prior distribution for Θ.

5. Urn Schemes for Constructing Priors

In the context of Bayesian non-parametric inference, the importance of Blackwell and Mac-Queen's result (Blackwell and Mac-Queen, 1973) is that it gives a simple and concrete procedure for constructing an infinite sequence of random variables with Dirichlet process as de Finetti measure. The procedure has the additional advantage of making intuitively clear some of the mathematical properties of the Dirichlet process, like its conjugate property or the form of the predictive distribution of the $(n+1)$-th random variables generated by a Dirichlet process conditionally on the values of the first n variables.

In the spirit of Blackwell and MacQueen we present in this section a class of stochastic processes defined on a countable space of Polya urns which will be convenient for constructing more general classes of priors commonly used in Bayesian non-parametric inference, such as Polya trees and beta-Stacy processes. There are situations where the assumption of exchangeability for the sequence of observations is too restrictive or does not incorporate all the relevant information about the data. A weaker assumption is that of partial exchangeability, introduced by de Finetti (1938) and considered also by Diaconis and Freedman (1980). For the connections between the two ideas of partial exchangeability see Fortini, Ladelli, Petris and Regazzini (1999). When $\{X_n\}$ is an infinite sequence of random variables with values in a discrete space, partial exchangeability (in the sense of Diaconis and Freedman)and recurrence imply that the law of the sequence is that of a mixture of Markov chains (Diaconis and Freedman, 1980); that is, conditionally on a random transition matrix Π, $\{X_n\}$ is a Markov chain with transition matrix Π. The prior distribution for Π may often be characterised in purely predictive terms; for example, Muliere, Secchi and Walker (2000) introduce an urn scheme called *reinforced urn process* which generates mixtures of Markov chains such that the law of Π is the product of Dirichlet processes. Reinforced urn processes have applications to survival analysis whenever individual specific data is modelled by a Markov chain and individuals from the population are assumed to be exchangeable.

6. Consistency

For subjective Bayesians like de Finetti and Savage probabilities represent degree of belief and there are no objective probability models. Bayesians learn from experience, so opinions based on very different priors will *merge* as data accumulate. A general result of this type was provided by Blackwell and Dubins (1962). A result relating *merging of opinions* and *posterior consistency* is discussed in Diaconis and Freedman (1986).

Roughly speaking, the posterior is consistent for F_0 if it cumulates around F_0 as the sample size increases, almost surely with respect the product measure F_0^{∞}. Bayesian nonparametric methods have only recently started to undergo asymptotic studies. Much of the papers are influenced by the paper of Diaconis and Freedman (1986). For a comprehensive review of this area see Wasserman (1998) and Ghoshal,Gosh and Ramamoorthi (1997). Recent results are in Walker and Hjort (2001) and Petrone and Wasserman (2001).

Our aim is to discuss consistency from a predictive point of view. In particular, we shall focus on the asymptotic behaviour of the predictive distribution. As shown in the previous sections, the representation theorem for exchangeable sequences and the results about predictive sufficiency ensure that the sequence of predictive distribution functions converges to the random distribution function F conditionally on which the observables are i.i.d. Our aim is to discuss how these results, which involve a *random* limit distribution F, are related to the notion of consistency of Doob (1948) or of Diaconis and Freedman (1986), in which F is the (fixed) true distribution.

On the other hand, starting from a paper by Diaconis and Freedman (1990) , we might replace the true distribution function with the empirical distribution function. In particular it is of interest to study the asymptotic distance (in some sense) between the predictive distribution function and the empirical distribution function; results in this direction are proved in Berti and Rigo (1997).

References

Berti, P. and Rigo,P., (1997), A Glivenko-Cantelli theorem for exchangeable random variables, *Statistics and Probability Letters*, **32**, pp.385-391.

Cifarelli, D:M. and Regazzini, E., (1982), Some considerations about mathematical statistics teaching methodology sugested by the concept of exchangeability. *Exchangeability in Probability and Statistics* (G. Koch and F. Spizzichino, eds.), pp. 217-232, North-Holland, Amsterdam.

Cifarelli, D.M. , Muliere, P. and Secchi, P. (1999), Prior processes for Bayesian nonparametrics, *Technical Report* n.377/P, Politecnico di Milano.

Cifarelli, D.M. , Muliere, P. and Secchi, P. (2000), Urn schemes for constructing priors, *Technical Report* n.**429**/P, Politecnico di Milano.

Fortini, S., Ladelli, L. and Regazzini, E., (2000), Exchangeability, Predictive distributions and Parametric Models, *Sankya*, vol. **62**, Ser. A., pp. 86 –109.

Muliere, P. and Secchi, P., (1992), Exchangeability, Predictive Sufficiency and Bayesian Bootstrap, *Journal of the Italian Statistical Society*, **3**, pp.377-404.

Muliere, P., Secchi, P. and Walker, S.G., (2000), Urn schemes and reinforced random walks, *Stochastic Processes and their Applications*, **88**, pp.59-78.

Petrone, S. and Wasserman, L. (2001), Consistency of Bernstein posterior (under revision for *Journal of the Royal Statistical Society, Ser. B*).

Walker, S. and Hjort, N.L. (2001), On Bayesian consistency. *Journal of the Royal Statistical Society*, Ser. B, to appear.

Walker, S. and Muliere, P., (1999), A characterization of a Neutral to the right prior via an extension of Jhonson's sufficientness postulate, *The Annals of Statistics*, **27**, pp.589-599.

REVISTA DE ESTATÍSTICA
47ª PÁGINA
2º QUADRIMESTRE DE 2001

CAUSAL INFERENCE AND GRAPHICAL MODELLING

Organizer: Nanny Wermuth

Invited Speakers: David R. Cox
Sarah Darby
Odd Aalen

REVISTA DE ESTATÍSTICA
49ª PÁGINA
2º QUADRIMESTRE DE 2001

Some Statistical Implications of Causality

D. R. Cox
Nuffield College and Department of Statistics
Oxford, UK
david.cox@nuffield.oxford.ac.uk

The aethos of statistics is on the whole healthily empirical and impatient of philosophical discussion. This attitude has, however, some negative consequences, notably a tendency to separate statistical analysis from deeper issues of interpretation. Yet many if not the majority of applications of statistical methods have as their objective the aiding of understanding of some phenomenon and this can be regarded as involving some notion of causality; in particular, in a technological context we may wish to assess the consequences of an intervention, medical or social for example. How would the real world be if some change were implemented, medical treatment A used rather than B, for instance?

Some recent discussions have tended to claim that causality can be relatively readily established, sometimes with rather minimal subject-matter input and even with cross-sectional observational studies. The ideas involved are interesting but observation suggests that the conclusion is dangerous, in a medical context at least.

The implications for statistical analysis are not particularly controversial but include the following, some of which will be illustrated by examples:

(i) to incorporate background knowledge into statistical models
(ii) to synthesize information from various sources, in line with Fisher's dictum
(iii) to check for the absence of qualitative interaction of effects under study with background variables
(iv) to formulate regression-type models that are potentially causal
(v) to be careful that the inclusion of explanatory variables in the models (iv) is consistent with a causal interpretation of the parameters of primary interest
(vi) to check in hierarchical systems that the regression coefficients used for interpretation are at an appropriate level in the hierarchy.

Does Diet Affect Risk of Lung Cancer?

Sarah Darby, Richard Doll
University of Oxford, CTSU,
Harkness Building, Radcliffe Infirmary,
Oxford OX2 6HE, UK
sarah.darby@ctsu.ox.ac.uk

Elise Whitley
University of Bristol
Department of Social Medicine,
Bristol BS8 2PR, UK

Timothy Key
ICRF Cancer Epidemiology Unit,
Gibson Building, Radcliffe Infirmary,
Oxford OX2 6HE, UK

Paul Silcocks
Trent Institute for Health Service Research,
Queens Medical Centre,
Nottingham NG7 2UH, UK

In 1975, it was reported on the basis of an observational study that the risk for lung cancer was about 60% lower in subjects with a high intake of vitamin A than in subjects with a low intake. Subsequent observational studies confirmed this negative association but suggested that the risk factor was provitamin A carotenoids, such as β-carotene, rather than vitamin A (i.e. pre-formed retinol) itself. Negative associations with lung cancer risk were also observed for a group of related dietary factors including intake of several non-provitamin A carotenoids and total intake of fruit and vegetables.

The consistent findings from observational studies led to the proposal that dietary β-carotene might reduce lung cancer risk and to the establishment of randomized controlled trials to test this hypothesis using supplements of β-carotene in human populations. However, the results of the trials were surprising and have given no support to the hypothesis: the two trials with the largest numbers of cases of lung cancer found that risk was significantly higher in subjects who took β-carotene than in those who did not, while other trials in lower risk subjects with smaller numbers of lung cancers reported no significant effect.

Despite the results of the trials, which indicate that β-carotene itself almost certainly does not protect against lung cancer, recently published observational studies continue to show an inverse association of fruit and vegetable intake with lung cancer risk, as do related indices such as carotene intake. The apparent protective effect of fruits and vegetables could be due to a biological effect of one or several of the thousands of chemicals naturally present in these foods, but it also remains possible that the observed association with fruits and vegetables may be partly due to confounding by other dietary factors and perhaps by smoking and non-dietary factors.

We have examined the relationship between diet and lung cancer in a case-control study of 982 cases of lung cancer and 1486 population controls in south-west

England in which subjects were interviewed personally about their smoking habits and their consumption of foods and supplements rich in retinol or carotene. Analyses were performed for 15 dietary variables, including intake of pre-formed retinol and carotene. When these were considered individually there were significant associations ($p<0.01$) with lung cancer risk for 8 of them, after adjustment for smoking. When the 15 variables were considered simultaneously, significant associations after adjustment for smoking remained for 5: pre-formed retinol (increased risk), and fish liver oil, vitamin pills, carrots and tomato sauce (decreased risk).

It is unlikely that all 5 associations represent biological effects, or that they can all be explained by residual confounding by smoking, or by biases. We conclude that there is at least one as yet unidentified factor that is causally related to lung cancer risk and of considerable importance in this population in terms of the number of cases of lung cancer that can be attributed to it.

Reference

Darby SC, Whitley E, Doll R, Key T, Silcocks. Diet, Smoking and Lung Cancer: a Case-control Study of 1000 Cases and 1500 Controls in South-west England. *British Journal of Cancer*. In press.

Analyzing Clustering of Deaths in Criminal Cases. Can Statistics Throw Light on the Causality?

Odd O. Aalen
University of Oslo
Section of Medical Statistics
P.O.Box 1122 Blindern
N-0317 Oslo, Norway
o.o.aalen@basalmed.uio.no

Clustering of deaths, particularly in health institutions, may lead to suspicion that the deaths are not natural. For instance, there have been cases where an inordinate number of deaths happened when a particular nurse was on duty. This has sometimes lead to charges of manslaughter. The task of a statistical expert witness in such cases will be discussed. To which extent can a statistical analysis throw light on the evidence that the clustering represents. This is a challenge to the power of statistics to assess causality.

Specific issues that arise in studies of this nature are the following:

- The charge against the defendant is raised after the statistical data appears. There may be no prior suspicion, and so one is faced with the issue of evaluating a hypothesis established aposteriori on the basis of an observed cluster. This is a well-known situation in medicine. There are for instance many reports of increased numbers of cancers in neighborhoods, work places, schools etc. When evaluating such clusters one has to estimate the probability that clusters of the kind shall arise by chance.
- The statistical analysis of clusters is difficult because one must judge the number of similar possibilities to consider. For, instance, when evaluating a cluster of deaths in a nursing home, how many nursing homes should one consider when evaluating the probability of the cluster arising occasionally by chance. How large a geographical area, and which time period should be considered?
- A *specific cause* of the clustering of deaths is hypothesized, namely that for which the defendant is charged. This is different from the usual statistical study, say in epidemiology, where the causal connection is of a more diffuse nature. For instance, when asserting that smoking causes lung cancer, one implies that the total exposure to smoking over many years causes cancer in a subset of those exposed to the risk. However, one can certainly not pinpoint a single decisive event, and also the biological mechanism may be only partially known.
- If one concludes, when judging a cluster, that it can hardly have arisen by chance, then the question of competing causal explanations arises. In addition to the specific cause suggested in the criminal indictment, there may be other, less specific causes that should also be considered. For instance, a cluster of deaths in a nursing home could possibly be due to non-criminal neglect on the part of a nurse. There is hardly much knowledge of whether such neglect could influence the death rate, but one may have to consider the possibility. Statistically, one can show that only small rises in the death rate may have a big effect on the likelihood of extreme clusters.

- An important issue is which statistical approach to use. Should one compute P-values or likelihood ratios, or should one use a Bayesian approach? P-values (adjusted for an aposteriori hypothesis) may be reasonably simple to compute under the null hypothesis of no criminal action. Likelihood ratios presume that one can also compute probabilities under an alternative hypothesis, and this is far more difficult. For instance, it might not be clear what is a reasonable alternative hypothesis. It could be the specific charge raised against the defendant, but this is dictated by the course of the police investigation, and may not be a useful alternative hypothesis from a statistical point of view. A Bayesian approach would in addition require knowledge of the prior probability of guilt, which may be impossible to assess. Also, courts don't want posterior probabilities of guilt from an expert witness.
- A statistical analysis may necessarily be somewhat technical, and courts have been known to object to this. The task of the expert witness is to make the analysis as accessible as possible. But there are difficult statistical issues involved, and the question of layman understanding is certainly an issue.
- To which extent can statistical analysis constitute proof in the court? This might differ between different countries, but statistics does not appear to have been used as single or main evidence in a criminal case involving such a serious charge as serial murder. The attitude seems to be that one must have specific proof that at least one person has been murdered. Then the statistical analysis can constitute important supplementary evidence.

A case that occurred in Norway a few years ago shall be used to illustrate some of the above points. We shall also refer to similar cases that have been described in the literature, see the references.

References

Buehler, J.W. et al. (1985). Unexplained deaths in a children's hospital: An epidemiological assessment. *The New England Journal of Medicine*, **313**, 211-216

Bølviken, E. and Egeland, T. (1995). Arson, statistics and the law: can the defendant's proximity to a large number of fires be explained by chance? *Science & Justice*, **35**, 97-104.

Fienberg, S. and Kaye, D. H. (1991). Legal and statistical aspects of some mysterious clusters. *Journal of the Royal Statistical Society* (Series A), **154**, 61-74.

Istre, G.R. et al. (1985). A mysterious cluster of deaths and cardiopulmonary arrests in a pediatric intensive care unit. *The New England Journal of Medicine*, **313**, 205-211.

Jones, C. (1998). Is a murder charge an occupational hazard of intensive care nursing? *Intensive and Critical Care Nursing*, **14**, 208-212.

Stross, J.K. et al. (1976). An epidemic of mysterious cardiopulmonary arrests. *The New England Journal of Medicine*, **295**, 1107-1110.

REVISTA DE ESTATÍSTICA
55ª PÁGINA
2º QUADRIMESTRE DE 2001

CONCENTRATION OF MEASURES

Organizer: Pascal Massart

Invited Speakers: Sergey G. Bobkov
Patricia Reynaud-Bouret
Paul-Marie Samson

REVISTA DE ESTATÍSTICA
57ª PÁGINA
2º QUADRIMESTRE DE 2001

On Concentration of Distributions of Weighted Sums

Sergey Bobkov
University of Minnesota
School of Mathematics
127 Vincent Hall, 206 Church Street S.E.
Minneapolis, MN 55455, USA
bobkov@math.umn.edu

Given a vector X in R^n of non-correlated random variables, with unit variances, denote by S(t) their wighted sum, with vector t of coefficients taken from the unit euclidean sphere. It is known as an application of the concentration phenomenon on the sphere [1] that, for most t's, the distributions of S(t) are very close to a certain "typical" distribution (on the real line).In general, it depends on X, but in many cases of interest, this typical distribution is standard normal. In one special situation, when X is uniformly distributed over isotropic compact body in R^n, a quantitative discription of this concentration property was resently obtained in [2]. We are discussing closeness to the typical distribution in the general situation, as well as some refinements for the class of log-concave probability distributions on the euclidean space.

References

[1] Sudakov, V.N. (1978). Typical distributions of linear functionals in finite-dimensional spaces of higher dimension. *Soviet Math. Dokl.*, 19, No. 6, 1578-1582.

[2] Antilla, M., Ball, K., and Perissinaki, I (1998). The central limit theorem for convex bodies. Preprint.

Concentration Inequalities for Poisson Processes and Applications in Statistics

Patricia Reynaud-Bouret
Laboratoire de Mathématiques
Département Probabilité et Statistiques
Bât 425, Université Paris-Sud
91405 Orsay Cedex
Patricia.Reynaud@math.u-psud.fr

We establish concentration inequalities for suprema of integral functionals of Poisson processes which are analogous to Talagrand's inequalities for empirical processes. These inequalities are used as crucial tools to construct penalized projection estimators of the intensity of an inhomogeneous Poisson process.

1. Probabilistic Results

A concentration inequality can be written in the following form:

$$\forall u > 0, \mathbb{P}\left(Z \geq E(Z) + f(u)\right) \leq \exp[-u]$$

where Z is a random variable, and f a proper function.

Concentration inequalities were proved by B.S. Cirel'son, I.A. Ibragimov and V.N. Sudakov for Z a 1-Lipschitz function of a Gaussian vector and $f(u) = \sqrt{2u}$ (see [1]).

M. Talagrand (see [2]) proved that such inequalities can be written in the n-sample framework. More precisely, let $(X_1, \ldots, X_n)$ be n random variables, i.i.d., with law $d\mathbb{P} = sd\mu$. Let $\mathbb{P}_n$ be the associated empirical measure., Let $\{\psi_a, a \in A\}$ be a countable family of functions bounded by 1. Then a concentration inequality holds with

$$Z = \sup_{a \in A}\left(\mathbb{P}_n(\psi_a) - \mathbb{P}(\psi_a)\right),$$

$$f(u) = c_1\sqrt{v_n u} + c_2 u \text{ and } v_n = \mathbb{E}\left(\sup_{a \in A}\sum_{i=1}^{n}\left(\psi_a(X_i) - \mathbb{E}(\psi_a(X_i))\right)^2\right),$$

where c_1, c_2 are proper constants. The constants c_1 and c_2 are computed via M. Ledoux's methods in a paper of P. Massart [3].

We prove that the same kind of inequalities hold for Poisson processes. More precisely, let us give the definition of Poisson processes to fix the notations.

Definition 1 Let $(\mathbb{X}, \mathcal{X})$ be a measurable space. Let N be a random countable subset of $\mathbb{X}$. n is said to be a Poisson process on $(\mathbb{X}, \mathcal{X})$ if

- for all $A \in \mathcal{X}$, the number of points of N lying in A is a random variable N_A which obeys a Poisson law with parameter denoted by $\nu(A)$,

- for all finite family of disjoint sets $A_1,\ldots,A_n$ of $\mathcal{X}$, $N_{A_1},\ldots,N_{A_n}$ are independent random variables.

The so defined function $\nu : \mathcal{X} \to \mathbb{R}_+$ is a measure without atom and is called the "mean measure" of N. This measure is supposed here to be finite to obtain almost surely a finite set of points for N. We denote by dN the random discrete measure $\sum_{T\in N}\delta_T$.

Our main probabilistic result is the following:

Theorem 1 Let N be an inhomogeneous Poisson process on $(\mathbb{X},\mathcal{X})$ with finite mean measure ν. Let $\{\psi_a, a\in A\}$ be a countable family of functions with values in $[-b,b]$. One considers

$$Z=\sup_{a\in A}\int_{\mathbb{X}}\psi_a(x)(dN_x-d\nu_x) \text{ or } \sup_{a\in A}\left|\int_{\mathbb{X}}\psi_a(x)(dN_x-d\nu_x)\right|.$$

Then for any positive number u

$$\mathbb{P}\left(Z\geq \mathbb{E}(Z)+2\sqrt{\nu u}+cbu\right)\leq \exp(-u)$$

where

$$\nu=\frac{1}{2}\left[\mathbb{E}\left(\sup_{a\in A}\int\psi_a^2(x)dN_x\right)+\sup_{a\in A}\int\psi_a^2(x)d\nu_x\right]$$

and where c can be taken equal to 5/4.

We can remark the similarity between Talagrand's inequality and this result with the correspondence $nd\mathbb{P}_n\approx dN$ and $d\mathbb{P}\approx d\nu$, which can be interpreted through the following property: the set of points of N, conditionally to the event $\{N_{\mathbb{X}}=n\}$ has the same law as a n-sample of variables with law $\nu/\nu(\mathbb{X})$.

The proof of this inequality is in two steps as the ones of M. Ledoux [4] and P. Massart [3] in the empirical case. The first step provides a concentration inequality for $Z=\sup_{a\in A}\int_{\mathbb{X}}\psi_a dN$ where the ψ_a's are bounded positive functions which is of independent interest.

L. Wu [5] recently proves analogous results for $Z=f(N)$ where f is a 1-Lipschitz function (in some sense) of the Poisson process. His results can lead to concentration formula for i.i.d. vectors of Poisson variables, already proved by S.G. Bobkov and M. Ledoux [6]. Very general results about concentration inequality for infinitely divisible vectors were also proved by C. Houdré [7]. The results of L. Wu and C. Houdré are very general but provide weaker results concerning the variance term v in the special case of suprema that we consider here.

2. Statistical Applications

The reason for focusing on suprema is that our aim is mainly to estimate adaptively the intensity of an inhomogeneous Poisson process, i.e. the Radon-Nikodym derivative $s=\frac{d\nu}{d\mu}$, where ν is the mean measure of the Poisson

process and μ is a known measure on $\mathbb{X}$; these functionals will appear naturally in the construction of our estimators. We apply penalized model selection methods introduced by L. Birgé and P. Massart (see for instance [9]). These methods are known to use concentration inequalities in several different frameworks like n-sample framework or white noise framework (see for instance [8]). Let us give a simple example in the Poisson framework: the histograms.

If m is a partition of the set $\mathbb{X}$, we can construct the projection estimator on m by

$$\hat{s}_m = \sum \frac{N_I}{\mu(I)} 1\!\!1_I .$$

We wish to select the best partition, i.e. the one such that the risk $R_m = \mathbb{E}\|s - \hat{s}_m\|_2^2$ is minimal. Obviously, the optimal partition depends on s, and we cannot find it without knowing s. The unbiased estimation of the risk leads to the model selection criterion:

$$\hat{m} = \arg\min_{m \in \mathcal{M}_{\mathbb{X}}} \left\{ -\int_{\mathbb{X}} \hat{s}_m^2(x) d\mu_x + 2 \sum_{I \in m} \frac{N_I}{\mu(I)} \right\}$$

where $\mathcal{M}_{\mathbb{X}}$ is a given set of partitions. The penalized histogram projection estimator will be $\hat{s}_{\hat{m}}$.

The unbiased risk principle relies heavily on the fact that:

$$\mathbb{E}\left(\sum_{I \in m} \frac{N_I}{\mu(I)} \right) = \mathbb{E}\left(\left\| \hat{s}_m - s_m \right\|^2 \right)$$

where s_m is the projection of s on the partition m i.e.

$$s_m = \sum_{I \in m} \frac{\int_I s d\mu}{\mu(I)} 1\!\!1_I .$$

Hence to understand the performances of the selecting estimator $\hat{s}_{\hat{m}}$, we have to understand how $\mathcal{X}\hat{m} = \left\| s_{\hat{m}} - \hat{s}_{\hat{m}} \right\|$ is small. This quantity is doubly random since the index $\hat{m}$ is chosen randomly. Consequently, we have to control all the $\mathcal{X}m$'s, to control this one. This explains why we need concentration inequalities. Since

$$\mathcal{X}m = \sup_{\|a\|_2 \le 1} \int_{\mathbb{X}} \sum_{I \in m} a_I \frac{1\!\!1_I}{\mu(I)} \left(dN_x - s(x) d\mu_x \right) = \sup_{\|a\|_2 \le 1} \sum_{I \in m} a_I \frac{N_I - \int_I s d\mu}{\mu(I)}$$

Theorem 1 can be applied as well as the results of S.G. Bobkov and M. Ledoux [6], for instance. If we want to look at more smooth functions than histograms, the quantity $\mathcal{X}m$ still exists (with a more general function instead of $1\!\!1_I / \mu(I)$), but we cannot assume that it comes from independent variables (the N_I's). That is the reason why we need concentration inequality for suprema of integral functionals of Poisson processes.

Moreover we wish to have a concentration inequality where the variance term v does not depend any more on the dimension of space (the cardinality of the partition

for histograms, for instance). Without such inequality, the rate of convergence of our estimator would not be optimal and this is precisely why we cannot use Wu's or Houdré's inequalities. We can derive such an inequality from Theorem 1:

Corollary 1 Let N be a Poisson process on $(\mathbb{X},\mathcal{X})$ with finite mean measure ν. Let $\{\psi_a, a \in A\}$ be a countable family of functions with values in $[-b,b]$. One considers

$$Z = \sup_{a\in A}\left|\int_{\mathbb{X}} \psi_a(x)(dN_x - d\nu_x)\right| \text{ and } v_0 = \sup_{a\in A}\int_{\mathbb{X}} \psi_a^2(x)d\nu_x.$$

Then for any positive numbers u and ε:

$$P\left(Z \geq (1+\varepsilon)E(Z) + \sqrt{2\cdot v_0 u} + \cdot(\varepsilon)bu\right) \leq \exp(-u),$$

where $\cdot = 6$ and $\cdot(\varepsilon) = 1.25 + 32/\varepsilon$.

All the results we have presented here can be found in [10].

References

[1] Cirel'son, B.S., Ibragimov, I.A: and Sudakov, V.N. (1975) Norms of Gaussian sample functions. Proc. 3rd Japan-USSR Symp. Probab. Theory, Taschkent 1975, *Lect. Notes Math.* **550**, 20-41.

[2] Talagrand, M. (1996) New concentration inequalities in product spaces. *Invent. Math.*, **126**, 3, 505-563.

[3] Massart, P. (2000) About the constants in Talagrand's concentration inequalities for empirical processes. *Ann Proba.*

[4] Ledoux, M. (1996) On Talagrand deviation inequalities for product measures. *ESAIM: Probability and statistics* **1**.

[5] Wu, L. (2000) A new modified logarithmic Sobolev inequality for Poisson point process and several applications. Probability Theory and Related Fields.

[6] Bobkov, S.G. and Ledoux, M. (1998) On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures. *J. Funct. Anal.*, **156**, 2, 347365.

[7] Houdré, C. (2000) Remarks on Deviation Inequalities for Functions of Infinitely Divisible Random Vectors. Conf. AMS-Bernoulli, Guanajuato, Mexico. *http://www.math.gatech.edu/ houdre.*

[8] Massart, P. (2000) Some applications of concentration inequalities. *Ann. de Toulouse.*

[9] Birgé, L. and Massart, P. (1997) From model selection to adaptive estimation. Festschrift for Lucien Le Cam, Springer, 55-87.

[10] Reynaud-Bouret, P. Concentration inequalities for inhomogeneous Poisson processes and adaptive estimation of the intensity. *Unpublished manuscript.*

Concentration Inequalities on Product Spaces for Mixing Processes, Coupling Methods

Paul-Marie Samson
Université Marne la Vallée
Cité Descartes - 5, Boulevard Descartes
Champs-sur-Marne - 77454 MARNE LA VALLEE CEDEX 2
samson@math.univ-mlv.fr

The concentration of measure phenomenon has been deeply investigated by M. Talagrand as a means to obtain new exponential deviation inequalities. One of the first result of the starting point of his developments is the following simple inegality for arbitrary product measure Talagrand. For $i=1,\ldots,n$, let μ_i be probability measures on $[0,1]$ and denote by P the product measure $\mu_1 \otimes \ldots \otimes \mu_n$. Then for every convex Lipschitz function f on $\mathbb{R}^n$ with $\|f\|_{\text{Lip}} \leq 1$, for every $t \geq 0$,

$$P(f \geq M + t) \leq 2e^{-t^2/4}, \tag{1}$$

where M is a median of f. Following M. Talagrand, M. Ledoux has proposed a simple method based on logarithmic Sobolev inequality for product measures to reach the deviation of f from the exact mean of f with the best possible constants. For every separately convex functions f on $\mathbb{R}^n$ with $\|f\|_{\text{Lip}} \leq 1$, for every $t \geq 0$,

$$P\left(f \geq \int f dP + t\right) \leq 2e^{-t^2/2}.$$

The deviation inequality (1) has been extended to some measures of contracting Markov Chains by K.Maxton as a consequence of an information inequality.

The main purpose of this presentation is to extend Marton's information theoretic approach to larger classes of dependent sequences such as Doeblin recurrent Markov chains and Φ-mixing processes. Let for example, $(X_i)_{i\in\mathbb{Z}}$ be a Markov chains or a Φ-mixing processes. Denote by P the law on $\mathbb{R}^n$ of a sample X of size n taken from $(X_i)_{i\in\mathbb{Z}}$. We will introduce a matrix Γ of dimension n, with coefficients that will measure the dependence between the random variables $(X_1,\ldots,X_n)$ of the sample X. In the interesting cases, the operator norm $\|\Gamma\|$ of the matrix Γ will be bounded independently of the size of the sample. This condition is satisfied for uniformly ergodic Markov chains satisfying a so-called Doeblin condition. Other examples are the Φ-mixing processes for which the sequence of Φ-mixing coefficients is summable.

Let now P denote the law of the sample X on $\mathbb{R}^n$. Let f be a real function on $\mathbb{R}^n$ such that for every x, y in $\mathbb{R}^n$,

$$f(y) - f(x) \leq \sum_{i=1}^{n} \alpha_k(y)_{x_k \neq y_k}.$$

As a main result, we show that for every probability measures Q with Radon-Nikodym derivative dQ/dP with respect to the measure P,

$$\int f dQ - \int f dP \le \|\Gamma\| \sqrt{2K(Q \mid P)} \sqrt{\int \sum_{i=1}^{n} \alpha_k^2 dQ},$$

where $K(Q \mid P)$ is the Kullback distance between the measure Q and P,

$$K(Q \mid P) = \int \log\left(\frac{dQ}{dP}\right) dQ.$$

Furthermore,

$$\int f dP - \int f dQ \le \|\Gamma\| \sqrt{2K(Q \mid P)} \sqrt{\int \sum_{i=1}^{n} \alpha_k^2 dP},$$

Such Pinsker type inequalities (or measure transportation inequalities) are obtained by coupling measures methods. We recently improve these methods to get better results in the independent case.

One of the main interest of these inequalities is to provide exponential concentration inequalities for the suprema of sums of random variables. In particular, we simply obtain the following famous result of P. Massart for suprema of empirical processes in the independent case. Let $\mathcal{T}$ be a countable set and let $(X_{1,t})_{t\in\mathcal{T}},\ldots,(X_{n,t})_{t\in\mathcal{T}}$ be n independent processes. Assume that for every t and k the values of $X_{k,t}$ are in $[0,1]$. Let us consider $Z = \sup_{t\in\mathcal{T}} \sum_{k=1}^{n} X_{k,t}$. Then for every $\lambda \ge 0$, setting $\phi(\lambda) = \exp(\lambda) - \lambda - 1$,

$$\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] \le e^{\mathbb{E}(Z)\phi(\lambda)},$$

which implies that for every $t \ge 0$,

$$\mathbb{P}\left[Z \ge \mathbb{E}[Z] + t\right] \le e^{-\mathbb{E}[Z]h\left(\frac{t}{\mathbb{E}[Z]}\right)},$$

where $h(u) = (1+u)\log(1+u) - u$, for $u \ge 0$.

This concentration inequality can be viewed as a functional version of Benett's or Bernstein's inequalities for sums of independent and bounded real-valued random variables. We will present some extentions of this concentration type inequalities in our context of dependence.

References

Ledoux, M. (1996). Talagrand deviation inequalities for product measures. ESAIM: *Probab. Statist.* **1** 63-8'7.

Marton, K. (1996). Bounding $\bar{d}$-distance by information divergence: a method to prove measure concentration. *Ann. Proba.* **24** 927-939.28, No. 1, 416-461.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Proba.* **28**, 863-884.

Maurey, B. (1991). Some deviation inequalities, *Geometric and Func. Anal.* **1** 188-197.

Samson, P.M. (2000). Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Proba.* **28** 416-461.

Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S.* **81** 73-205.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Ivent. Math.* **126** 505-563.

REVISTA DE ESTATÍSTICA
65ª PÁGINA
2º QUADRIMESTRE DE 2001

DISEASE MAPPING AND SPATIAL EPIDEMIOLOGY

Organizer: Sylvia Richardson

Invited Speakers: Peter J. Green
Leo Knorr-Held
Robert Wolpert

REVISTA DE ESTATÍSTICA
67ª PÁGINA
2º QUADRIMESTRE DE 2001

Disease Mapping and Spatial Epidemiology

Sylvia Richardson
Department of Epidemiology and Public Health
Imperial College
Norfolk Plac,
London, W2 1PG, UK.
sylvia.richardson@ic.ac.uk

1. An Introduction to the Session

Spatial analyses abound in the epidemiological literature. Indeed, analysing sources of heterogeneity in diseases or health data provides valuable knowledge in epidemiology. Geographical variations of chronic diseases have long been recognised as been able to suggest important aetiological clues. Recently, the availability of geographically indexed health and population data at a small scale, with advances in computing and geographical information systems have opened the way for serious exploration of small area health statistics based on routinely collected data, the recent book by Elliott et al (2000) gives a series of interesting examples.

The statistical models developed are closely linked to the declared aims of small area analyses and the type of data available. **Disease mapping** is carried out to summarise spatial and spatio-temporal variation in risk. **Geographic correlation studies** prolong disease mapping studies and are aimed at exploiting geographical variations in exposure to environmental variables (such as air pollution, background radiation, water quality) and lifestyle factors (such as smoking and diet), again in order to gain clues as to disease aetiology. Finally, **Point source type studies** are carried out when an increased risk close to a "source" is suspected. Morris and Wakefield (2000) provide a review of the assessment of disease risk in relation to point/line sources.

Disease mapping exercises require an exhaustive recording of cases and an assessment of the population at risk. One can distinguish between point data where the "exact location" of the case is known and area-referenced count data which correspond to number of cases aggregated over geographically defined areas. In what form disease data are made available follows usually from the public health procedures and the confidentiality rules adopted in each country. Similarly, exposure characteristics may be available at the individual level, at a continuum of locations, or as aggregated summaries. Corresponding to these types of data, spatial models have been defined using either a point process framework or directly at the aggregated level (see Richardson, 2001, for a review). Most current approaches use the generic methodology of Bayesian hierarchical models.

Two presentations (Green and Knorr-Held) are addressing disease mapping issues. Both extend in an innovative manner the current family of models used for the spatial analysis of disease counts and use data collected at the aggregated level. Issues of model choice will be highlighted by Green, whilst Knorr-Held introduces further stratification in disease mapping models based on classification of the disease outcome into stages of severity. Wolpert's presentation discusses the point process framework and the problem of formulating appropriate models for data (outcome and covariates) collected at different spatial scales. This is linked to the delicate question

of how covariates are introduced in spatial models and the consequent ecological bias issue (Greenland and Robins, 1994).

References

Elliott, P., Wakefield, J.C., Best, N.G. and Briggs, D.J. (2000). Spatial Epidemiology: Methods and Applications. Oxford University Press, Oxford.

Greenland, S. and Robins, J. (1994). Ecological studies: biases, misconceptions, and counterexamples. *American Journal of Epidemiology*, **139**, 747--760.

Morris, S.E. and Wakefield, J.C. (2000). Assessment of disease risk in relation to a pre-specified source. In *Spatial Epidemiology: Methods and Applications*
Elliott, P. et al (eds). Oxford University Press, Oxford.

Richardson, S (2001) Spatial models in epidemiological applications. To appear in *Highly Structured Stochastic Systems*, Green, P.J., Hjort, N.L. and Richardson S. (eds),
Oxford University Press, available at
http://www.stats.bris.ac.uk/L2000/Chapter/srichardson.ps

REVISTA DE ESTATÍSTICA
69ª PÁGINA
2º QUADRIMESTRE DE 2001

Spatial Mixtures and Model Choice in Disease Mapping

Peter J. Green
University of Bristol
Department of Mathematics,
University Walk,
Bristol, BS8 1TW, UK.
P.J.Green@bristol.ac.uk

1. Introduction

We consider the modelling of spatial heterogeneity for count data on a rare phenomenon, observed in a pre-defined set of areas. This is a general set-up which arises in many domains of application, for example in ecology or agricultural science. The motivating situation that we have in mind throughout, that of observed disease counts with few observed events in each area, belongs to the domain of epidemiology. There are many reasons for suspecting heterogeneity in the underlying event rate and wanting to characterise it. In disease mapping, for example, it is hardly plausible that all the relevant factors acting on the underlying disease risk can be identified or measured at the area level. Thus there remains residual heterogeneity, which is likely to have a spatial structure inherited partly from that of risk factors for the disease. Spatially structured heterogeneity also arises naturally in agricultural field trials and other applications.

It is interesting to characterise this spatial structure further, as discovery of either local discontinuities or smooth gradients can be exploited for further study or action. In epidemiology, current aetiological hypotheses made at the individual level can be usefully confronted with their aggregated counterparts, keeping in mind the delicate issue of ecological bias. The suspicion of a local excess in disease occurrence or the highlighting of geographical inequalities in medical treatment are important public health concerns that can be addressed by an analysis of the spatial heterogeneity. Note that, most often, we are in an observational framework where there is little or no control over the sources of variability. Our aim in this paper is thus to propose a new class of spatial models for the heterogeneity of count data and to demonstrate its flexibility.

2. Models

Spatial heterogeneity of count data on a rare phenomenon occurs commonly in many domains of application, in particularly in disease mapping. We present new methodology to analyse such data, based on a hierarchical allocation model. We assume that the counts follow a Poisson model at the lowest level of the hierarchy, and introduce a finite mixture model for the Poisson rates at the next level. The novelty lies in the modelling of allocations to the mixture components, where we consider three possibilities, in each of which the number of components of the spatial mixture is treated as unknown.

One follows a spatially correlated process, the Potts model (Green and Richardson, 2000), and the others are based on multinomial draws from correlated weights processes defined using gaussian fields (Fernandez and Green, 2000).

Inference is performed in a Bayesian framework using reversible jump MCMC (Green, 1995). The models introduced can be viewed as Bayesian semiparametric approaches to specifying flexible spatial distribution in hierarchical models. They could also be used in contexts where the spatial mixture subgroups are themselves of interest, as in health care monitoring.

Performance of the models and comparison with an alternative well-known Markov random field model specification for the Poisson rates (Besag, York and Mollié, 1991) are demonstrated on synthetic data sets. We found that our allocation model avoids the problem of oversmoothing in cases where the underlying rates exhibit discontinuities, while giving equally good results in cases of smooth gradient-like or highly autocorrelated rates. The methodology is illustrated on epidemiological applications to data on rare disease and health outcome in France.

3. Model Choice

The final part of the talk will discuss choice between model specifications in this area. In principle, model choice can itself be treated in a fully Bayesian way (as was done, for example, in the context of ion channel data by Hodgson and Green, 1999), but there are obvious difficulties in relying on posterior model probabilities alone in model choice decisions, not least the arbitrariness of prior model probabilities. I will discuss the use of various decision-theoretic criteria in this context, including the Deviance Information Criterion of Spiegelhalter, Best and Carlin (1998).

References

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration with applications in spatial statistics (with discussion). *Annals of the Institute of Mathematical Statistics*, **43**, 1--59.

Fernandez, C. and Green, P.J. (2000) Modelling spatially correlated data via mixtures: a Bayesian approach. Technical report, University of Bristol, Department of Mathematics.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82**, 711-732 .

Green, P. J. and Richardson, S. (2000) Spatially correlated models for count data. Technical report, University of Bristol, Department of Mathematics.

Hodgson, M. E. A. and Green, P. J. (1999) Bayesian choice among Markov models of ion channels using Markov chain Monte Carlo, *Proc. R. Soc. Lond.* A **455**, 3425-3448.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). Read to the Royal Statistical Society on 15 January 1997. *J. Roy. Statist. Soc.* (B), **59**, 731-792.

Spiegelhalter, D. J., Best, N. G. and Carlin, B. P. (1998) Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Unpublished manuscript.

REVISTA DE ESTATÍSTICA
71ª PÁGINA
2º QUADRIMESTRE DE 2001

Disease Mapping of Stage-Specific Cancer Incidence Data

Leonhard Knorr-Held, Günter Raßer
Ludwig-Maximilians-University Munich
Department of Statistics
Ludwigstrasse 33
80539 Munich, Germany
leo@stat.uni-muenchen.de, rasser@stat.uni-muenchen.de

Nikolaus Becker
German Cancer Research Center, Department of Biostatistics
Im Neuenheimer Feld 280
69120 Heidelberg, Germany
n.becker@dkfz.de

1. Introduction

There has been much recent interest into the spatial analysis of observational disease data. The work can be categorized into two groups, methodology for data where the exact location of each case is known, and methodology for aggregated data, where the total number of cases is given in predefined administrative areas. Bayesian approaches for the second type of data includes the seminal work by Besag, York and Mollie (1991) who propose a Markov random field model for the spatial smoothing of disease rates. This model is nowadays widely used for „disease mapping", the study of spatial variation in disease risk, for reviews see for example Clayton and Bernardinelli (1992), Knorr-Held and Becker (2000) or Wakefield *et al.* (2000).

Probably the most prominent application is the statistical analysis of age-standardized cancer mortality rates, as such data are routinely collected throughout the world. A spatial analysis may help to identify a „spatial signal", which is particular important for rare diseases, where the raw data exhibit too much variation and are not particularly helpful in order to judge the variation of the underlying disease risk. The estimated spatial pattern may give hints to relevant unobserved risk factors, although some general problems of interpretation may remain due to the observational type of the data.

In this paper we extend the methodology to the analysis of cancer incidence data with additional knowledge on the stage of disease at time of diagnosis. Our aim can be described as to adjust the crude observed data for effects which can be attributed to age, and to assess whether there is any spatial variation left in the (adjusted) stage proportions. This is of clear public health importance for diseases where screening programs increase the probability of a cancer diagnosis in an early stage of the disease and hence the probability of survival.

2. Model

Let n_{ij} denote the number of person-years at risk in district $i = 1,...,I$ and age group $j = 1,...,J$. For each cell (i,j) let y_{ijs} denote the number of diagnosed cases of

disease in stage $s = 1,...,S$. We assume that the stages are ordered by severity of the disease with stage S being the most severe. Finally let y_{ij0} be the number of all person-years under risk, which have not being diagnosed with the disease. We now assume that $y_{ij} = (y_{ij0}, y_{ij1}, ..., y_{ijS})'$ follows a multinomial distribution with parameters n_{ij} and probability vector $\pi_{ij} = (\pi_{ij0}, \pi_{ij1}, ..., \pi_{ijS})'$ where $\sum_{s=0}^{S} \pi_{ijs} = 1$.

We propose two formulations based on regression models for categorical data on an ordered scale (for a recent review see Fahrmeir and Tutz, 2001, Chapter 3). In the first approach we model *cumulative* probabilities of disease risk, whereas in the second we model *conditional* probabilities. More specifically, in the *cumulative* model we factorize the log odds of the *cumulative probabilities* $\pi_{ij0} + \pi_{ij1} + ... + \pi_{ijs}$, $s = 0,...,S-1$, into spatial and age group effects. In the second formulation, the so-called *sequential* model, we consider the probability that a person is diagnosed with the disease in a specific stage, given that she is diagnosed in this or in a higher stage.

Hence we decompose the log-odds of the *conditional* probabilities $\pi_{ijs}/(\pi_{ijs} + ... + \pi_{ijS})$ into spatial and age groups effects. The spatial and age group effects are assumed to be stage-specific in both formulations, i.e. for each stage s there is a separate set of parameters.

Note that we work directly on data stratified by age, which is in contrast to commonly used disease mapping methods, where the data are already standarized by age in advance.

The two alternative models proposed above are now completed by assinging prior distributions to all unknown parameters. For both the spatial and the age group parameters we will use priors which favour a nearly constant pattern, implied by a high prior mass on very small values of the corresponding variance parameter. However, the priors we use for these variance parameters are highly dispersed, hence the formulation will be flexible enough to capture spatial or temporal gradients or trends if there is evidence in the data for it.

More specifically we use Gaussian pairwise difference priors (Besag *et al.*, 1995) for the district and age group-specific parameters. These models neither impose stationarity nor assume a specific parametric form; in fact they are closely related to non- and semiparametric smoothing methods as reviewed by Fahrmeir and Knorr-Held (2000).

Inference has been carried out using C++ routines developed by the first author. We have used Markov chain Monte Carlo (MCMC) to sample from the two posterior distributions, applying univariate Gaussian Metropolis random walk proposals for all age group and spatial parameters, while Gibbs steps have been used for the remaining precision parameter. The spread of each Metropolis proposal was tuned in an automatic fashion - prior to the collection of the posterior samples - so that the corresponding acceptance rate for each parameter was between 35 and 45%. Alternatively one could employ a block sampling algorithm as recently proposed by Rue (2001) and Knorr-Held and Rue (2001) to improve mixing and convergence properties of the simulated Markov chain. Problems with single-site updating typically arise for sparse data. However, the data we considered in our application is not particularly sparse and MCMC mixing was fine for the single-site scheme we have implemented.

3. Application

We apply the two approaches to incidence data on cervical cancer in the former German Democratic Republic (GDR), 1980-1989. The data is stratified by 216 administrative districts, 15 age groups (15-19, 20-24, 25-29, ..., 80-84 and 85+) and S=5 stages. Of particular interest is the first stage, which denotes an asymptomatic, not yet malignant pre-stage of cervical cancer, typically diagnosed in screening programs.

For a first assessment of the model fit, we routinely monitor the multinomial posterior deviance. There seems to be some evidence that the sequential model fits the data better than the cumulative model, with a smaller posterior deviance.

The results obtained with the sequential model suggest that there are large spatial differences in the (age-adjusted) proportions of the first stage, which indicates spatial variability in the time of introduction and effectiveness of screening programs.

Acknowledgements

Most of this work was undertaken while the first author was at Imperial College School of Medicine, London.

References

Besag, J. E., Green, P. J., Higdon, D. M. and Mengersen, K. L. (1995). Bayesian computation and stochastic systems (with discussion), *Statistical Science*, 10, 3-66.

Besag, J. E., York, J. C. and Mollie, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

Clayton, D. G. and Bernardinelli, L. (1992). Bayesian methods for mapping disease risks, In Small Area Studies in *Geographical and Environmental Epidemiology* (eds. J. Cuzick and P. Elliot), 205-220. Oxford University Press, Oxford.

Fahrmeir, L. and Knorr-Held, L. (2000). Dynamic and semiparametric models. In Smoothing and Regression: Approaches, Computation and Applications (ed. M. Schimek), Ch. 18, 513-544, Wiley & Sons, New York.

Fahrmeir, L. and Tutz, G. (2001). Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd Edition, Springer-Verlag, New York.

Knorr-Held, L. and Becker, N. (2000). Bayesian modelling of spatial heterogeneity in disease maps with application to German cancer mortality data, *Allgemeines Statistisches Archiv*, 84, 121-140.

Knorr-Held, L. and Rue, H. (2001). On block updating in Markov random field models for disease mapping. Scandinavian Journal of Statistics, in revision.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields with applications. *Journal of the Royal Statistical Society*, Series B, in press.

Wakefield, J. C., Best, N. G. and Waller, L. A. (2000). Bayesian approaches to disease mapping. In Spatial Epidemiology: Methods and Applications (eds. P. Elliot, J. C. Wakefield, N. G. Best and D. J. Briggs). Oxford University Press, Oxford.

Disease mapping and Small area Statistics

Robert L. Wolpert
Institute of Statistics and Decision Theory
Duke University
USA
wolpert@stat.Duke.EDU

Spatial detail is lost whenever data are aggregated in small area statistical problems, sometimes distorting evidence due to the well-known phenomenon variously known as the "modifiable area unit problem" (MAUP) or "ecological bias". When exposure data, covariates, and health effects are available at different spatial resolutions, some or all of which may differ from the resolution needed for the inferential goals, it is common to pick some fixed partition of the region of interest and begin by aggregating or interpolating all the data to a common resolution, just to simplify the statistical analysis. This standard approach only exacerbates the problem of ecological bias.

We use marked point process models to offer an alternative: spatially continuous underlying random-field models that allow us to use each variable at its natural spatial resolution, without any further aggregation or interpolation, and to support inference at any level of spatial detail (or simultaneously at different levels). There is still no way to restore any information lost in aggregation before we receive the data, but by avoiding unnecessary further aggregation we reduce the effects of ecological bias when compared with the standard approach.

The methods, based on a new computational algorithm for simulating point processes, are illustrated with examples studying the incidence of respiratory disease in Huddersfield, UK and in London. Most of the work was developed jointly with Katja Ickstadt of Dortmund University, DE and Nicola Best of Imperial College, UK.

REVISTA DE ESTATÍSTICA
75ª PÁGINA
2º QUADRIMESTRE DE 2001

PERFECT SIMULATION

Organizer: Jesper Møller

Invited Speakers: Duncan Murdoch
Elke Thönnes
Antonietta Mira

Perfect Simulation Session

Jesper Møller
Aalborg University
Department of Mathematical Sciences
jm@math.auc.dk

Over the last decade there has been an explosion of interest in developing and applying Markov chain Monte Carlo (MCMC) methods in statistics. Regular MCMC methods are only correct in the limit where an infinite number of steps in the simulations have been performed. A recent topic, which has drawn great attention after the seminal work of Propp and Wilson (1996), is "exact" or "perfect" simulation. A simulation algorithm is said to be perfect when one is assured that equilibrium has been attained when the algorithm finishes; here its running time is allowed to be random. This is obviously appealing and useful for several reasons: One needs not to worry about whether one has used an appropriate burn in before sampling; iid sampling is possible so that e.g. asymptotic variances of Monte Carlo estimates can be straightforwardly calculated; and one can compare other "non-perfect" algorithms with perfect simulations.

Propp and Wilson consider MCMC algorithms (especially the Gibbs sampler) for simulating lattice models from statistical physics satisfying a certain monotonicity condition and with a finite but large state space such as the Ising model and the accompanying random cluster model. The idea is to use possibly several runs of the algorithm backwards in time (started from time 0) and by monotonicity and coupling dominate these by some lower and upper chains until there is coalescence at time 0. A drawback of ProppWilson type algorithms is their sensitivity to user impatience: stopping very long runs of the algorithm before termination can cause significantly biased output. An alternative perfect simulation algorithm by Fill (1998), based on rejection sampling, overcomes this problem.

These ideas have now been extended in many ways, and today perfect simulation techniques have proven to be particular useful in spatial statistics, stochastic geometry and statistical physics. The range of applications for more mainstream statistical problems, particularly in Bayesian statistics, is so far more limited, but this view may quickly change as perfect simulation is an active area of current research.

For references, including survey papers, see David Wilson's webpage on Perfectly Random Sampling with Markov Chains:

http://dimacs.rutgers.edu/~dbwilson/exact.html/

References

J.A. Fill (1998). An interruptible algorithm for perfect sampling via Markov chains. *Annals of Applied Probability* **8**, 131-162.

J. G. Prbpp and D. B. Wilson (1996. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223-252.

Perfect Sampling Algorithms: Descendants of CFTP

Duncan Murdoch
University of Western Ontario
Dept. of Statistical and Actuarial Sciences
Western Science Centre
London, Ontario, Canada
murdoch@stats.uwo.ca

1. Introduction

Propp and Wilson (1996) introduced a simulation algorithm called *coupling from the past* (CFTP). This algorithm did something which on the face of it appeared to be impossible: starting with a Markov chain simulation, it constructed a sample which was distributed *exactly* according to the limiting distribution of the chain. The recipe "run for an infinite length of time, then take a sample" was not needed, a finite amount of computation was sufficient. This generated a great deal of excitement among people doing Markov chain Monte Carlo, because it appeared to work around the difficult problem of determining the "burn-in" period of a chain, after which samples can be treated as though drawn from the limiting distribution. Though progress has been made in applying CFTP to wider ranges of models (see David Wilson's web page *<http://dimacs.rutgers.edu/~dbwilson/exact.html>* for a comprehensive bibliography), it has not yet turned out to be a panacea: work is still required to apply it to most problems, and many problems are still intractable. It is, however, very effective on some Markov chains and other Markov processes, particularly in stochastic geometry (Møller, 2000).

This raises the question: if CFTP is so effective on certain Markov chains and processes, might there not be related algorithms that are particularly effective in other situations? In this paper we discuss two attempts to find such algorithms. Section 2 briefly reviews CFTP, and extracts a central idea which motivates our attempts. Section 3 describes the first attempt: estimation of the limit of an alternating series. Section 4 describes the second: simulation of stochastic differential equations.

2. Review of CFTP

A standard description of CFTP goes like this. Suppose we have a Markov chain with a stochastic recursive sequence representation

$$X_{t+1} = \varphi(X_t, U_{t+1}) \tag{1}$$

Here U_{t+1} is a sample from a sequence of independent and identically distributed random variables, and $\varphi(X_t, U_{t+1})$ represents the algorithm that makes use of this source of randomness to generate a new sample X_{t+1} from an old one X_t. Our aim is to arrange that a particular simulated value (by convention, X_0) has the limiting distribution of the chain. To do this, imagine simulating an infinite sequence of U_t values running back in time. (We will actually only generate them as needed, but once generated, the value is fixed, so if we need it again, we see the same value.) We then search backwards for a time $T < 0$ such that for *every* possible value of X_T, the resulting value of X_0 is the same. Since a chain that was run from the indefinite past must have passed through *some* state at time T, the distribution of X_0 must be the same as if we had performed the infinite simulation, i.e. it must be distributed according to

the limiting distribution of the chain. Papers such as Murdoch and Green (1998) illustrate that CFTP can be practical even when the state space is a continuum.

What can we abstract from this algorithm? One nonstandard description of CFTP is as follows. Let X^T be the value of X_0 that would be obtained by running (1) forwards from some particular state at time $T < 0$ (e.g. $X_T = 0$). Then the distribution of X^T clearly converges to the target distribution as $T \to -\infty$, and (assuming that CFTP will succeed) we also know that the sequence X^T, $T=0, -1, \ldots$ will eventually become constant. Moreover, we have a way to determine when this occurs. Once it has become constant we need simulate no more, we have our perfect sample from the target.

There are many situations where we have a sequence of approximations converging to a target distribution. In the next two sections we describe two of them. Our aim will be to couple the terms in the sum in such a way that the simulated values eventually become constant, and we have a way to detect this event. We will then have draws from the limiting distribution of the sequence.

3. Approximation of the Limit of a Series

Suppose that a_n is a non-random decreasing sequence which converges to 0. Then the sequence of alternating sums $S_n = \Sigma_0^n (-1)^i a_i$ converges to a limit S; our aim is to approximate this limit. We will do this by generating a random variable X with $X \sim N(S, \sigma^2)$, for a specified variance σ^2.

We base our sampler on either the layered multishift coupler (Wilson, 2000) or the bisection coupler (Green and Murdoch, 1998). Both of these couplers allow us to generate a random function $X(\mu)$ such that for each μ, $X(\mu) \sim N(\mu, \sigma^2)$. Both couplers give piecewise constant step functions with the steps at random locations; Wilson's coupler is non-decreasing in μ.

We start by generating a single realization of $X(\mu)$. We then evaluate a sufficient number of terms S_n so that we can determine which step of $X(\mu)$ contains $X(S)$; this is straightforward, since the even partial sums $S_0, S_2, \ldots$ form a sequence decreasing to S, and the odd partial sums $S_1, S_3, \ldots$ form a sequence increasing to S. We simply seek an interval $[S_{2n-1}, S_{2n}]$ which lies entirely within one step of $X(\mu)$; then $X(S) = X(S_{2n})$.

Once we have drawn $X(S)$, we can form a confidence interval for S by elementary means. For example, a 95% confidence interval is $X(S) \pm 1.96\sigma$. Since we were free to choose σ, we can choose to make this confidence interval as short as we like. The cost of choosing a small variance is that the steps in $X(\mu)$ will be very short, so n will need to be large (and its distribution is highly skewed to the right). For example, in one series of 1000 simulations approximating $\log 2$ to one decimal place as the limit of $1 - 1/2 + 1/3 - \ldots$, the median number of terms in the sum was 31, the mean was 238, and the maximum was over 140000.

All of this computation is not very useful. In all cases in this series of simulations, the confidence interval was longer than the length of the calculated bounding interval $[S_{2n-1}, S_{2n}]$. Since the intersection of the confidence interval and the bounding interval is also a valid 95% confidence interval, we might expect that to do better than either, but in the majority of cases (83%), the confidence interval generated by this method completely enclosed the bounds. In other words, most of the time we might as well not have generated it, because it gives no more information about S than we had already. Since the overall coverage probability is 95%, the coverage in cases

where the intersection is shorter than the bounds must be lower. In this simulation it was 67%.

4. Stochastic Differential Equations Stochastic differential equations (SDEs) represent continuous time processes, or diffusions. For example,

(2) $$dx_t = \mu(x_t)\, dt + \sigma(x_t)\, db_t$$

denotes that the process x_t tends to drift upwards at the rate $\mu(x_t)$, and diffuses randomly with a variance of $\sigma(x_t)^2$ per unit of time. Our aim will be to simulate x_t at one or more fixed time points, given the value at $t = 0$. If both the drift and diffusion functions are constant, then the solution to this SDE is known: $x_t - x_0 \sim N(\mu t, \sigma^2 t)$. However, most other SDEs have no explicit solution, and approximations are necessary for simulation.

A simple approximation is Euler's method. Here one acts as though the drift and diffusion functions are constant over small time intervals, and simulates using the distribution given above. At the end of the interval the drift and diffusion rates are recalculated, and the simulation is repeated.

This looks like a situation where the general principle of CFTP might apply. One can improve the accuracy of the Euler approximations to any degree desired by reducing the step size. Is it possible to couple together Euler approximations with different step sizes, in such a way that as the step size tends to zero, the result of the simulation eventually becomes constant?

5. Conclusion

It is clear that CFTP is not the only algorithm that can simulate from a limit using only approximations to it. However, whether other algorithms exist that will turn out to be as useful as CFTP remains to be seen.

Acknowledgements

The work in section 3 is closely based on joint work with Jeffrey Rosenthal of the University of Toronto.

This work was partially supported by an NSERC research grant.

References

Green, P.J. and Murdoch, D.J. (1998). Exact sampling for Bayesian inference: towards general purpose algorithms. In: *Bayesian Statistics* **6**, J.M. Bernardo, J.O. Berger, A.P. Dawid and A. F. M. Smith, eds.

Møller, J. (2000). A review of perfect simulation in stochastic geometry. Research Report R-00-2016, Department of Mathematical Sciences, Aalborg University. To appear in *Selected Proceedings of the Symposium on Inference for Stochastic Processes*, Eds. I.V. Basawa, C.C. Heyde and R.L. Taylor, IMS Lecture Notes & Monographs Series, 2001.

Murdoch, D.J. and Green, P.J. (1998). Exact sampling from a continuous state space. *Scandinavian Journal of Statistics* **25**, 483--502.

Propp, J.G. and Wilson, D.B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223-252.

Wilson, D. (2000). Layered multishift coupling for use in perfect sampling algorithms (with a primer on CFTP). In: *Monte Carlo Methods—Fields Institute Communications, Vol.* **26**, N. Madras, ed.

REVISTA DE ESTATÍSTICA
81ª PÁGINA
2º QUADRIMESTRE DE 2001

Generic CFTP in Stochastic Geometry and Beyond

Elke Thönnes
University of Warwick
Department of Statistics
Coventry, CV4 7AL
United Kingdom
elke@stats.warwick.ac.uk

1. Introduction

While having established itself as an important statistical tool, Markov Chain Monte Carlo (MCMC) needs careful implementation, in particular when deciding on the length of the Markov chain run. Ideally one would like to run the chain until it is in or close to equilibrium, but how do we decide on this? When the seminal paper by Propp and Wilson appeared in 1996, it promised a solution to the notorious problem of determining convergence. Their *coupling from the past* (CFTP) algorithm dynamically extends the run-time of the Markov chain until it has reached equilibrium. The diagnosis of convergence is based on a clever coupling construction backwards in time – note that a corresponding forwards-time construction would lead to bias.

The method was enthusiastically received by the MCMC community, in particular by researchers in stochastic geometry and spatial statistics. Important early papers outside of the spatial statistics context include Foss and Tweedie (1996) and Green and Murdoch (1996). Foss and Tweedie show that the CFTP algorithm by Propp and Wilson is only applicable to uniformly ergodic Markov chains. But many interesting Markov chains, especially in stochastic geometry, are only geometrically ergodic. However, an extension of the original algorithm due to Kendall (1998), *dominated CFTP,* makes the method also amenable to these chains. While this limited space only allows for a very vague description of dominated coupling from the past, an introduction to both classic and dominated CFTP can be found in Thönnes (2000).

2. Generic CFTP

Dominated coupling from the past was first developed for problems in stochastic geometry where it has lead to successful developments, see for example Kendall and Thönnes (1997), Kendall and Møller (2000) and Møller et al (1997). Classic CFTP is in fact a special case of dominated CFTP and so we speak of generic CFTP when we refer to either classic or dominated CFTP without distinguishing between the two methods, see also Kendall and Thönnes (2001).

Generic CFTP is based on two building blocks: a dominating process and an envelope process. The dominating process is an easy to simulate, stationary process that stochastically dominates the target Markov chain. Its function is twofold: it supplies the randomness to evolve the target chain in form of random numbers or marks and it provides stochastically varying random bounds on the values that the target chain may take at a specific time. Naturally we need to specify a state space for the dominating process which generally is an augmentation of the state space of the target chain. The random bounds that the dominating process supplies are a subset of this state space and, together with the marks, are passed on to the envelope process.

We need to start the dominating process in equilibrium and extend it backwards in time to produce a consistent stationary path on a time interval [-T,0] with increasing T. If the dominating process is time-reversible then this is easily done by simulating it forwards in time and then reversing it.

The purpose of the envelope process is to encompass the relevant paths of the target chain and provide a sufficient criterion for the convergence of the target chain based on the coalescence of the relevant paths. It uses the random bounds supplied to the dominating process to choose an initial state at the appropriate time –T and the marks to evolve forwards in time. It shares the state space of the dominating process and so takes values that are subsets of the state space of the dominating process. Once its state at time 0 is a singleton, this singleton constitutes a perfect sample.

This generic description lends itself to a computer implementation using an object-oriented approach, see Figure 1. This implementation consists of two classes. The first class is a dominating process which has a start method that determines the starting time and an extend method which extends the current realisation backwards in time. It has two attributes, the marks and the states, which are inherited by the second class, the envelope process. The methods of the second class are an evolve method and the coalesce method that tests whether the state of the process at time 0 is a singleton. Of course, these generic methods and attributes need further specification adapted to the specific problem at hand.

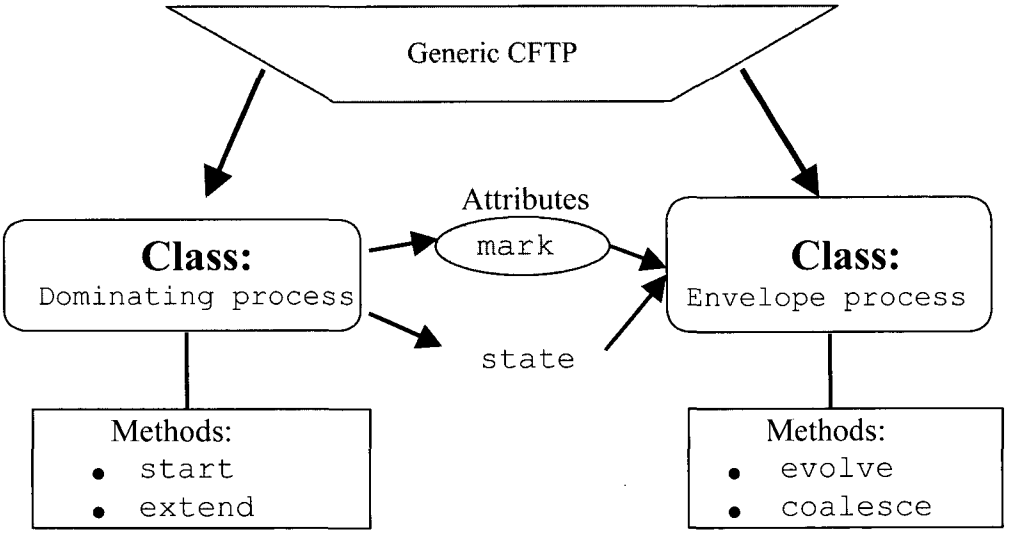

Figure 1. Generic CFTP in an object-oriented approach.

For the simulation of Markov point processes using spatial birth-and-death processes the two building blocks, dominating process and envelope process, appear as quite natural constructions. The dominating process is a stationary birth-and-death process D whose birth rate is higher and whose death rate is lower than the corresponding rates in the target Markov chain. The envelope process uses as its starting state at time –T the set of all subsets of D_{-T}. Its evolution can then be derived from the dominating process by modification of transitions.

Classic CFTP as originally developed by Propp and Wilson assumes that the state space of the target chain is equipped with a partial order and a maximal and minimal element with respect to that partial order. All states of the target chain are thus bounded between the maximal and minimal element. The dominating process is

now simply a marked, set-valued process which is actually constant, taking as value the set of states that lie between the maximal and the minimal element, and thus the state space of the target chain. However its marks vary so as to allow the evolution of a coupled target chain. Using the randomness supplied by the dominating process, the envelope process then describes the values that the paths of the target chain started in all initial states may take at a specific time if coupled to the dominating process.

3. Going Beyond Stochastic Geometry

While dominated coupling from the past has been a success story in spatial statistics and stochastic geometry, its application in non-spatial statistical problems has been limited. However, the extension of perfect simulation techniques from stochastic geometry to non-spatial applications promises to be rewarding.

On one hand, dominating coupling from the past may produce perfect samples in shorter runtime (in terms of number of iterations) than classic coupling from the past. On the other hand, and more importantly, dominated coupling from the past methods for stochastic geometry are bound to have useful analogues in non-spatial applications.

One example of a non-spatial application that can be addressed using dominated coupling from the past is the exact simulation of solutions to stochastic difference equations, see Kendall and Thönnes (2001). Let the distribution of X be defined by the equation

$$L(X) \quad = \quad L(B\,(X + C)),$$

where $L(X)$ denotes the law of X, and B and C are non-negative random variables which satisfy certain sufficient criteria. We can use this equation to define a Markov chain whose equilibrium, given ergodicity, is a solution to the above stochastic difference equation. Dominated CFTP for this example is based on a dominating process in form of a random walk with negative drift. By an appropriate reflection at zero, which is inspired by the associated random walk, the process not only stochastically dominates the target chain but also delivers an equilibrium distribution that is standard and easy to simulate. This dominating process is not time-reversible but its extension backwards is straightforward as its time-reversal is easily computed. The evolution of the envelope process is based on a γ-coupling (see Lindvall, 1992, pp 18-20) using marks supplied by the dominating process as well as marks that are generated when necessary.

The scope of the developed CFTP algorithm is quite general. Examples of distributions that fit into this context are perpetuities as well as simple ARCH and GARCH models. Moreover, in amenable cases, the algorithm allows for "omnithermal" CFTP, that is it produces perfect samples for a whole distribution family of B and C.

4. Conclusions

While perfect simulation has been successfully implemented in many applications, in particular in stochastic geometry, one challenge is to widen the scope of these methods. Generic CFTP formalises the methods used in both dominated and classic CFTP and may thus clarify the use of these methods in other contexts. One area of interest is whether the methods that have been so successful in stochastic

geometry can be generalized to non-spatial applications. For an overview on what has been achieved so far, the interested reader is referred to David Wilson's webpage at

http://dimacs.rutgers.edu/~dbwilson/exact.html.

References

Foss, S.G and Tweedie, R.L. (1998) Perfect simulation and backward coupling. *Stochastic models* **14**, 187-203.

Häggström, O., van Lieshout, M.N.M and Møller, J. (1999) Characterisation results and Markov Chain Monte Carlo algorithms including exact simulation for some spatial point processes, *Bernoulli* **5**, 641-659.

Kendall, W.S. (1998) Perfect simulation for the area-interaction point process. In *Probability Towards 2000*, (eds L.Accardi and C.C. Heyde), 218-234, NewYork, Springer.

Kendall, W.S. and Møller, J. (2000) Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes, *Advances in Applied Probability* **32**, 844-865.

Kendall, W.S. and Thönnes, E. (1999) Perfect simulation in stochastic geometry, *Pattern Recognition* 32, 1569-1586.

Kendall, W.S and Thönnes, E. (2001) A general formulation of CFTP for Markov chains (working title), work in progess.

Lindvall, T. (1992) *Lectures on the Coupling Method,* New York, John Wiley & Sons.

Murdoch, D.J. and Green, P.G. (1998) Exact sampling from a continuous state space, *Scandinavian Journal of Statistics* **25**, 483-502.

Propp, J.G and Wilson, D.B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223-252.

Thönnes, E (2000) A primer on perfect simulation. In *Statistical Physics and Spatial Statistics,* (eds K.R. Mecke and D. Stoyan), Lecture Notes in Physics, New York, Springer.

REVISTA DE ESTATÍSTICA
85ª PÁGINA
2º QUADRIMESTRE DE 2001

Perfect Simulation for Bounded Distributions via Slice Sampling

Antonietta Mira
University of Insubria
Department of Economics
Via Ravasi 2, 21100 Varese
Italy
anto@aim.nnipv.it

Jesper Møller
Aalborg University
Department of Mathematical Sciences
Fredrik Bajers Vej 7G, DK-9220 Aalborg
Denmark
jm@math.auc.dk

Gareth Roberts
Lancaster University
Department of Mathematics and Statistics
Lancaster, LA1 4AF
UK
G.O.Roberts@lancaster.ac.uk

1. Introduction

The goal is to obtain a random sample from a complicated target distribution. To this aim we combine two quite powerful ideas that have recently appeared in the Markov chain Monte Carlo (MCMC) literature: the slice sampler (SS) and perfect simulation (PS).

The SS is a method of constructing, with the aid of auxiliary variables, a reversible Markov chain with a specified stationary distribution (Swendsen and Wong, 1987). The *simple SS* (SSS) is a special case where a single auxiliary variable is used. As for every MCMC method, draws from the stationary distribution, the target distribution, are obtained only after a "sufficiently long" run of the simulation. It is typically impossible to determine how long is sufficiently long.

Perfect simulation is a clever way of running a Markov chain which ensures that the terminal value of the implementation is an exact draw from the stationary distribution of the chain (Propp and Wilson, 1996).

By exploiting monotonicity properties of the SS we show that a perfect version of the algorithm can be easily implemented, at least when the target distribution is bounded. This eliminates the problem of determining how long the SS Markov chain must be run before it has reached stationarity.

Numerous practical extensions and real applications of the perfect SS are presented in Mira et al. (2001), including a prefect SS for unbounded distributions.

2. Slice Sampler

Here we will briefly introduce the SS. More extensive descriptions of the algorithm can be found in Mira and Tierney (1998) and Roberts and Rosenthal (1999).

Suppose $\pi(x)$, $x \in \mathcal{X}$ is an un-normalised integrable density with respect to the measure μ and let ν_π be the corresponding probability measure. In the SSS we introduce an auxiliary variable, $u \in \mathcal{U}$, and construct the joint distribution of u and x by taking the marginal for x unchanged and defining the conditional distribution of u given x to be uniform on $(0,\pi(x))$.

An irreducible and aperiodic Markov chain $\{(X_n,U_n)\}_{n=0}^{\infty}$ is then set up over the enlarged state space $\mathcal{X}\times\mathcal{U}$ having the probability measure corresponding to $\pi(x,u) \propto I_{\{u<\pi(x)\}}(x,u)$, as its unique stationary distribution, where $I_A(x)$ is the indicator function of the set A. In particular we will perform a Gibbs sampler: in the vertical update $u \mid x$ is sampled uniformly on $(0,\pi(x))$; in the horizontal update $x \mid u$ is sampled from the normalisation of the restriction of μ to the set $A_\pi(u) = \{x : \pi(x) > u\}$. The marginal X-chain $\{X_n\}_{n=0}^{\infty}$ has ν_π as its stationary distribution and can thus be used to estimate integrals with respect to π.

When $\mathcal{X} = \mathbf{R}^d$ and μ is the d-dimensional Lebesgue measure, the SSS is uniformly ergodic if π is bounded and the support of π has finite Lebesgue measure (Mica and Tierney, 2001).

The simple SS is stochastically monotone with respect to this ordering (Roberts and Rosenthal, 1999):

(1) $$x \prec x' \text{ if and only if } \pi(x) \le \pi(x')$$

3. Perfect Slice Sampler

Consider a positive recurrent Markov chain with invariant distribution ν_π, specified by a stochastic recursive sequence (SRS): $X_{n+1} = f(X_n,\gamma_n), n > 0$, where $\{\gamma_n\}_{n=-\infty}^{\infty}$ is a sequence of iid random variables, $n \in \mathbf{Z}$. Below we briefly describe the Propp and Wilson (1996) coupling from the past (CFTP) algorithm. Let $X_n^{(x,t)}$ be the value at time n of the chain started in x at time $-t$. The CFTP algorithm applies, in theory, provided the *vertical backward coupling time*, $T = \inf\{t > 0 : X_0^{(x,t)} = X_0^{(y,t)} \text{ for all } x,y\}$, is almost surely finite. In fact $X_0^{(x,t)} = X_0^{(y,t)} = X_0^{(\cdot,t)} \sim \nu_\pi$ for all states x,y, and times $t \ge T$. PS becomes feasible, in practice, if there is a partial ordering $\prec$ on the state space such that $f(x,\gamma) \prec f(y,\gamma)$ if $x \prec y$ and if there exist a maximal, $x_{\max}$, and a minimal, $x_{\min}$, state (i.e. $x_{\max} \prec x \prec x_{\min}, \forall x$). Usually these states are assumed to be unique, but for the perfect SSS more than one maximal or minimal state exist, furthermore the existence of $x_{\max}$

can be eliminated (Mira et al. 2001). CFTP works as follows. Choose a time $T_1 > 0$, generate $\{\gamma_n\}_{n=-T_1}^{-1}$ and set,

$$X_{-T_1}^{(x_{\max},T_1)} = x_{\max},\ X_{n+1}^{(x_{\max},T_1)} = f\left(X_n^{(x_{\max},T_1)},\gamma_n\right),\quad -T_1 \le n < 0,$$
$$X_{-T_1}^{(x_{\min},T_1)} = x_{\min},\ X_{n+1}^{(x_{\min},T_1)} = f\left(X_n^{(x_{\min},T_1)},\gamma_n\right),\quad -T_1 \le n < 0.$$

If $X_0^{(x_{\max},T_1)} = X_0^{(x_{\min},T_1)}$ that is if coalescence occurs, then $X_0^{(x_{\min},T_1)} = X_0^{(x,T_1)}$ far all x, and the common value, $X_0^{(\cdot,T_1)}$, is necessarily distributed as ν_π. Otherwise choose a new value $T_2 > T_1 > 0$ and restart the backward simulation from time $-T_2$. When running the simulation over the time range $[-T_1,0]$, we need to reuse the same random numbers, $\{\gamma_n\}_{n=-T_1}^{-1}$, used in the first stage of the simulation. The procedure is repeated for $k = 1,2,\ldots$ until $X_0^{(x_{\max},T_k)} = X_0^{(x_{\min},T_k)}$, whereby $X_0^{(x_{\min},T_k)} = X_0^{(\cdot,T_k)}$ and so we return $X_0^{(x_{\min},T_k)} \sim \nu_\pi$. Notice that sample paths of maximal and minimal chains started at time point further back will be sandwiched in between paths started at earlier times (funneling property).

We now give an explicit SRS for the SSS, which preserves monotonicity with respect to the order given in (1). We assume for simplicity that maximal and minimal states exist and are unique; in fact, as explained in Mira et al. (2001), all we need to assume is that $\mu(\mathcal{X})$ and $\sup\pi$ are both finite. The SRS construction allows the continuum of chains implicitly defined in the PS, to be mapped to a countable collection of images in any particular iteration, of which only a finite number need ever be explicitly calculated. We shall carry out the vertical slice first, followed by the horizontal slice. For all $t < 0$ define a vertical slice variable, $\varepsilon_t \sim U[(0,1)]$. Then, for the chain that, at time t, is in state x, set $U_t(x) = \varepsilon_t \pi(x)$. The horizontal slice is more complicated. At each time $t < 0$ construct an infinite sequence of random variables, $\mathbf{W}_t = \{W_{t,j} : j = 1,2,\ldots\}$ by $W_{t,1} \sim U[A_\pi(U_t(x_{\min}))]$ and $W_{t,j} \sim U[A_\pi(\pi(W_{t,j-1}))]$. Let $\sigma_t(x) = \inf\{j \ge 1 : \pi(W_{t,j}) \ge U_t(x)\}$, and set

$$f\left(x,(\varepsilon_t,\mathbf{W}_t)\right) = W_{t,\sigma_t(x)}. \qquad (2)$$

It is easy to check that $\sigma_t(x)$ is almost surely finite for all $x \in \mathcal{X}$. Since $\gamma_t = (\varepsilon_t,\mathbf{W}_t)$, $t \in \mathbf{Z}$, is independent of x, (2) is indeed an SRS representation for some Markov chain. The chain simulated is in fact a SSS because $W_{t,\sigma_t(X_{t-1})}$ given $U_t(X_{t-1}) = u$ is distributed as P_u: this is just an adaptive rejection sampling scheme where the rejection region becomes more and more refined as the simulation proceeds. The function in (2) is monotone in its first argument since, for all t, $W_{t,\cdot}$ and $\sigma_t(\cdot)$ are non-decreasing sequences (the W's with respect to $\prec$) by construction. Hence, using (2), we have a CFTP algorithm for simulating from ν_π. Because of the funnelling

property and since $\sigma_t(\cdot)$ is non-decreasing, it will never be necessary to increase the number of simulated $W_{t,j}$'s.

4. Extensions

Even if an ordering has been defined on the state space it can be hard or impossible to find a maximal and/or a minimal state. Following ideas well summarized in Kendall and Møller (2000), it is possible to construct upper and lower bounding stationary processes which allow the identification of upper and lower starting values for the sandwiching algorithm algorithm outlined in the previous section. A way of exhibiting a bounding process for the SS is described in Mira at al. (2001b). This is probabilistically the most natural construction since it is based entirely on $Q_\pi(u) = \mu\left[A_\pi(u)\right]$ a function that completely characterizes the SS (Roberts and Rosenthal, 1999). Unfortunately implementation of this idea in real applications is hampered by the fact that explicit information about Q'/Q is needed, and this is unlikely to be available. A more practical construction appears in Mira et al. (2001a) where various other extensions of the perfect SS are also discussed including the perfect product SS with multiple auxiliary variables. Examples of applications are the Ising model on a two dimensional grid at the critical temperature and various other auto-models.

References

Kendall, W. and Moller, J. (2000). Perfect Metropolis-Hastings simulation of locally stable point processes. *Adv. Appl. Prob.* **32**, to appear.

Mira, A. and Tierney, L. (1998). Efficiency and Convergence Properties of Slice Samplers. *Scandinavian Journal of Statistics*, to appear.

Mira, A., Moller, J. and Roberts, G. (2001x). Perfect slice samplers. *J. Royal Stat. Soc.* B. To appear.

Mira, A., Moller, J. and Roberts, G. (2001b). Perfect simple slice sampler. *Proceedings of the ISI conference 2001*. To appear.

Propp, J. and Wilson, D. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223-252.

Roberts, G. and Rosenthal, J. (1999). Convergence of slice samper Markov chains. *J of Royal Statistical Society* B **61**, 643-660.

Wilson, D_ (1999). How to couple from the past using a read-once source of randomness. *Random Structures and Algorithms*.

REVISTA DE ESTATÍSTICA
89ª PÁGINA
2° QUADRIMESTRE DE 2001

PROBABILITY AND STATISTICS IN BIOINFORMATICS

Organizer: Timo Koshi

Invited Speakers: Jotun Hein
Sophie Schbath

REVISTA DE ESTATÍSTICA
91ª PÁGINA
2º QUADRIMESTRE DE 2001

Probability and Statistics in Bioinformatics

Timo Koski
Linkoping University
Department of mathematics
S-581 83 Linkoping
Sweden
tikos@mai.liu.se

With the completion of the human genome and the genomes of many other organisms the task of organizing and understanding the generated data becomes more and more important. In the past decade computational approaches to molecular and structural biology have attracted attention from computer scientists, mathematicians and statisticians (Salzberg et.al. 1999, Waterman 1995). Among available computational methods those that are based on explicit probabilistic and statistical models are the focus of this session. Since bioinformatics explicitly or implicitly concerns the analysis of biological data that are intrinsically probabilistic, such models should be also at the core of bioinformatics.

In the past decades we have witnessed the likelihood approach to pairwise alignments and to construction of phylogenetic trees, probabilistic methods to RNA secondary structure, the EM –algorithm for finding regulatory binding motifs and the Markov and the hidden Markov models for for DNA base composition analysis and gene prediction and analysis of DNA words.

Probabilistic models in bioinformatics apply a notion of modularity: complex systems are built by combining simpler parts. Probability theory serves as the foundation whereby the parts are combined , and ensuring that the system as a whole is coherent (in a Bayesian sense) and providing ways to interface data. (Durbin et.al. 1998).

References

M.Durbin, S. Eddy, A.Krogh and G. Mitchison (1998): *Biological Sequence Analysis. Probabilistic models for proteins and nucleic acids*. Cambridge University Press.

T. Koski (2001): *Hidden Markov Models in Bioinformatics*. Kluwer Series in *Computational Biology*, forthcoming.

S.L. Salzberg, D.B. Searls, S. Kasif (ed).s) (1999): *Computational Methods in Molecular Biology*. Elsevier.

M.Waterman (1995): *Introduction to computational biology*. Chapman and Hall.

Algorithms for Statistical Multiple Alignment

J. Hein, J. L. Jensen, K. Mouridsen, C. S. N. Pedersen
Aarhus University
Denmark
jotun.hein@biology.au.dk

Statistical approaches to alignment based on an explicit evolutionary model of insertions and deletions have great potential for real data analysis. Alternative algorithms are here presented that allows the calculation of the probability of a set of sequences related by a binary tree that has evolved according to the Thorne-Kishino-Felsenstein model (1991) for a fixed set of parameters. One central idea is to define a Markov chain that generates ancestral sequences and their alignment at two neighboring nodes in a tree. The running time of these algorithms are k'th power (l - sequence length, k -number of sequences).

A Gibbs sampling approach is also presented that should be applicable for larger number of sequences.

Finally, open problems extending the basic statistical alignment problem are discussed, such as combining statistical alignment with comparative genefinding and the relationship between hidden Markov model alignments and statistical alignment.

References

Thorne, J. L., H. Kishino and J. Felsenstein (1991). An Evolutionary Model for Maximum Likelihood Alignment of DNA Sequences. *J. Mol. Evol.* 33: 114-124.

Steel, M. & J.J. Hein (2001): A generalisation of the Thorne-Kishino-Felsenstein model of Statistical Alignment to k sequences related by a star tree. (In Press, *Letters in Applied Mathematics*)

Hein, J., C. Wiuf, B. Knudsen, Møller, M., and G. Wibling (2000): Statistical Alignment: Computational Properties, Homology Testing and Goodness-of-Fit. (*J. Molecular Biology* 302: 265-279)

J.J.Hein (2001): A generalisation of the Thorne-Kishino-Felsenstein model of Statistical Alignment to k sequences related by a binary tree. (p179-190 Pac. Symp. Biocompu. 2001)

REVISTA DE ESTATÍSTICA
93ª PÁGINA
2º QUADRIMESTRE DE 2001

Distribution of Word Counts in DNA Sequences and Quality of Approximations

Sophie Schbath
INRA, Unité Mathématique Informatique & Génome
78352, Jouy-en-Josas
France
Sophie.Schbath@jouy.inra.fr

Counts of words are part of the elementary statistics used in biological sequence analysis, for example to find significantly under- or over-represented words. These words having a frequency either too high or too low might point out unknown biological constraints. Recognized by specific proteins, exceptional words may be involved in the DNA protection, the replication, the gene expression etc.

The knowledge of the statistical distribution of these counts is necessary to assess the significance of the observed results. Because a DNA sequence is naturally represented as a finite but long sequence of letters among the 4-letter DNA alphabet {A, C, G, T}, Markovian models are widely considered. As we will see, the choice of the order of Markov model has an important influence in the interpretation of the results.

The exact distribution of the count of a word is known under the hypothesis that the letters are independent or under a Markov model (see Regnier (1999) or Robin and Daudin (1999)). It is given through its probability generating function which is a rational function. The Taylor expansion of this generating function can then be obtained with a finite recurrence. It is theoretically possible to calculate this distribution for any word in any sequence. In practice, it is impossible to compute it in a reasonable time for long sequences or for very frequent words. On the other hand, two kinds of approximations exist: Gaussian approximations (Kleffe and Borodovsky (1992), Prum et al. (1995)) and compound Poisson approximations (Schbath (1995), Geske et al. (1995)). The asymptotic framework in which these approximations are valid are different but they both require that the length of the sequence tends to infinity. Their main advantage is that, in most cases, they require very lower computation times.

A special attention has to be given to the influence of the estimation of the transition probabilities according to the observed DNA sequence.

Moreover, exact and approximate distributions of word counts depend on the overlapping structure of the word; indeed, two occurrences of a word can or cannot overlap in a sequence.

After describing these two approximations, we will discuss their quality with respect to the word frequency and the sequence length. We will also present the rules suggested by Robin and Schbath (2000) for choosing between the Gaussian distribution, the compound Poisson distribution and the exact distribution when finding exceptional motifs in DNA sequences.

This talk will be illustrated by the E. coli genome analysis, in particular, the significantly high frequency of the so-called Chi motif and the significantly low frequency of the 6-letter biological palindromes will be presented.

Here are some additional related references : El Karoui et al. (1999), Erhardsson (2000), Reinert and Schbath (1998), Reinert et al. (2000), Rocha et al. (1998), Schbath et al. (1995).

References

El Karoui, M., Biaudet, V., Schbath, S. and Gruss, A. (1999). Characteristics of Chi distribution on different bacterial genomes, Research in *Microbiology* **150**, 579-587.

Erhardsson, T. (2000). Compound Poisson approximation for counts of rare patterns in Markov chains and extreme sojourns in birth-death chains. *Ann. Appl. Probab.* **10**, 573-591.

Geske, M.X., Godbole, A.P., Schaffner, A.A., Skolnick, A.M., Wallstrom, G.L. (1995). Compound Poisson approximations for word patterns under Markovian hypotheses. *J. Appl. Prob.* **32**, 877-892.

Kleffe, J. and Borodovsky, M. (1992). First and second moment of counts of words in random texts generated by Markov chains. *Comp. Applic. Biosci.* **8**, 433-441.

Prum, B., Rodolphe, F. and Turckheim, E. de. (1995). Finding words with unexpected frequencies in DNA sequences. *J. R. Statist. Soc.* B. **57**, 205-220.

Regnier, M. (1999). A unified approach to word occurrence probabilities. To appear in *Discrete Applied Mathematics*, Special Issue on Computational Biology, preliminary version at RECOMB'98.

Reinert, G. and Schbath, S. (1998). Compound Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comp. Biol.* **5**, 223-253.

Reinert, G., Schbath, S. and Waterman, M.S. (2000). Probabilistic and Statistical Properties of Words: An Overview. *J. Comp. Biol.* **7**, 1-46.

Robin, S. Daudin, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.* **36**, 179-193.

Robin, S. and Schbath, S. (2000). Numerical comparison of several approximations of the word count distribution in random sequences. Submitted.

Rocha, E., Viari, A. and Danchin, A. (1998). Oligonucleotide bias in Bacillus subtilis: general trends and taxonomic comparisons. *Nuc. Acids. Res.* **26**, 2971-2980.

Schbath, S. (1995). Compound Poisson approximation of word counts in DNA sequences. ESAIM: *Probability and Statistics.* **1**, 1-16.

Schbath, S. , Prum, B. and de Turckheim E. (1995). Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences, *J. Comp. Biol.* **2**, 417-437.

REVISTA DE ESTATÍSTICA
95ª PÁGINA
2º QUADRIMESTRE DE 2001

PROBABILITY APPROXIMATIONS FOR RARE EVENTS

Organizer: Holger Rootzén

Invited Speakers: Patrik Albin
Tailen Hsing
Igor Rychlik

REVISTA DE ESTATÍSTICA
97ª PÁGINA
2º QUADRIMESTRE DE 2001

Probability Approximations for Rare Events

Holger Rootzén
Chalmers University of Technology
Dept. Mathematics
S-41296 Gothenburg
Sweden
rootzen@math.chalmers.se

Probability approximations for rare events form an interesting and beautiful body of mathematical theory. However, the driving force behind the large effort spent on such approximations comes from applications – often connected with serious or catastrophic occurences – to many different parts of science and technology.

A main line of development was started by Fischer and Tippet in 1928 with their delineation of the possible limits for linearly normalized maxima of i.i.d random variables. This was largely completed by Gnedenko in 1945, and in later work by de Haan. Through contributions of many researchers, including Berman and Leadbetter, this theory has been vastly extended to large classes of stochastic processes. The proofs often consist of quite sophisticated use of rather elementary probability calculations. They also include an important and broadly useful set of analytical inequalities for multivariate normal distributions, in various guises going under the names Slepians inequality and the Normal Comparisons Lemma. For a good exposition of this, and many other useful facts about normal distributions, see Tong (1990).

It was early realized that asymptotic results for maxima are immediate consequences of Poisson approximations for indicator functions. The connection is obvious: the maximum of a set of variables is less than a level u if and only if the indicators that the variables are greater than u sum to zero. It was also noted that standard proofs for maxima often, with minor changes only, in fact proved Poisson approximations, and applied to much more general rare occurences than maxima. In particular this leads to the asymptotic distribution for all extreme order statistics and not just for maxima. Books reviewing parts of this development include Leadbetter et al. (1983), Resnick (1987) and more recently Piterbarg (1996) and Embrechts et al. (1999). The last mentioned reference is a good place to start if one wants to learn more about the subject.

The first talk of this session is a part of the still very vigorous development of this theory and also links with upcrossings as discussed below. In another direction, the results are currently being extended to more complicated situations, often coming from spatial or geometric models. This includes many challenging and difficult mathematical problems. The second talk of this section is concerned with one such problem. The so-called Stein-Chen method is used as a part of the proofs in this talk. This method is a very important basic tool which is useful a very wide range of problems and in current research perhaps is *the* main tool for deriving probability approximations for rare events. The basic reference for the Stein-Chen method is the book by Barbour et al (1992).

A somewhat different line of development was started in the 1940-ies with celebrated papers by Rice and by Kac where they, using different methods, found explicit formulas for the numbers of upcrossings of a level by continuous time

processes. Cramér later noted the close relation between upcrossings and maxima. This theory has since been refined and generalized through the efforts of many researchers, and also has developed into a much used tool for many engineering problems, including metal fatigue, structural safety and ocean engineering. Leadbetter et al (1983) contains a review of the state of this theory in the mid 1980-ies, and more recent references may be found e.g. in Albin (2000) . The third talk of this session derives a very general version of these results from a theorem by Banach, and shows how they may be used in important applications.

References

P. Albin, Extremes and upcrossing intensities for **P**-differentiable stationary processes. *Stoch. Proc. Appl.* **87**, 199-234 (2000).

A. D. Barbour, L. Holst and S. Jansson, *Poisson approximation.* Oxford University Press, New York (1992).

P. Embrechts, C. Klueppelberg, and T. Mikosch, *Modelling extremal events for insurance and finance.* Springer, Berlin (1999, second corrected printing).

G. Lindgren, M.R. Leadbetter and H. Rootzén, *Extremes and related properties of stationary sequences and processes*. Springer, New York (1983).

V. I. Piterbarg, *Asymptotic methods in the theory of Gaussian processes and fields.* AMS translations of Mathematical Monographs, vol. 148, Providence, Rhode Island (1996).

S.I. Resnick *Extreme values, regular variation, and point processes.* Springer, New York, (1987).

Y.L. Tong, *The multivariate normal distribution.* Springer, Berlin (1990).

REVISTA DE ESTATÍSTICA
99ª PÁGINA
2º QUADRIMESTRE DE 2001

Extremes of Infinitely Divisible Stationary Processes

Patrik Albin
Department of Mathematics
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden
palbin@math.chalmers.se

A unified theory for local and global extremes (rare large values), of infinitely divisible (i.d.) stationary stochastic processes, with (at least) exponentially light tails, is presented. The case of i.d. processes with subexponential heavy tails is already well-understood, at least locally, by completely different methods, from work by Rosinski and Samorodnitsky (Ann. Probab., 1993).

In addition to more or less usual tricks and estimates in the field of extreme value theory, we use Esscher transforms and stochastic comparison of i.d. random vectors, using their spectral (dynamical) representation (model), thus converting qualitative and quantitative issues for multivariate "rare i.d. probabilities", to simpler such, yielding to the usual tricks.

Locally, the theory extends to non-stationary processes in different ways. Not suprisingly, stationary increment processes cause a little extra trouble, and similarly with self-similar processes.

The Longest Edge of Certain Graphs

Tailen Hsing
Texas University
thsing@stat.tamu.edu

Holger Rootzén
Chalmers University of Technology
Dept. Mathematics
S-41296 Gothenburg
Sweden
rootzen@math.chalmers.se

For any finite set A of points in $\Re^d$, denote by NNG(A) and MST(A) the nearest neighbor graph and minimal spanning tree, respectively, generated by the points in A. Let $M_{\text{NNG(A)}}$ and $M_{\text{MST(A)}}$ be the longest edge length of NNG(A) and MST(A), respectively. For a given distribution function F on $\Re^d$, let $\mathcal{P}_n$, be the set of points of a Poisson process with intensity measure equal to nF and $\mathcal{I}_n$ the set of points comprised of n iid. random variables distributed according to F. This paper considers the asymptotic distributions of $M_{\text{NNG}(\mathcal{P}_n)}$, $M_{\text{MST}(\mathcal{P}_n)}$, $M_{\text{NNG}(\mathcal{I}_n)}$ and $M_{\text{MST}(\mathcal{I}_n)}$ for a class of distribution functions F. Such problems have been investigated in Penrose (1997, 1998), in which F is assumed to be uniform in the unit cube or symmetric normal. Here, we are primarily interested in the case where F has an unbounded support.

We will assume that $d=2$ for convenience. Write $u(\mathbf{x})=-\log\big(f(\mathbf{x})\big)$ where f is the density of F. Let $x(\cdot,u)$ denote the level curve $u(\mathbf{x})=u$. We first focus on the case where the level curves are parallel and are given by the following:

$$\mathbf{x}(t,u)=\mathbf{c}(t)+\omega(u)\mathbf{n}(t) \text{ for } t\in I=[0,L) \text{ and } u\geq \text{some } u_0>0$$

where $\mathbf{c}(t)$ is a closed, strictly convex curve parameterized by length, $\omega(u)$ is an increasing function with $\omega(u_0)=0$ and $\omega(\infty)=\infty$, and $\mathbf{n}(t)$ is the unit normal of $\mathbf{c}$ at the parameter t. Assume that $\mathbf{c}(t)$ is twice-differentiable with $|\ddot{\mathbf{c}}(t)|\in(0,\infty)$ and $\omega(u)=u^{\alpha}\exp\Big(\int_{y_0}^{u}a(y)y\,dy\Big)$ where $\alpha>0$, $a(y)\to 0$ and $ya'(y)\to 0$ as $y\to\infty$. Then for each $\tau>0$,

$$\lim_{n\to\infty}P\Big(M_{\text{NNG}(\mathcal{P}_n)}\leq r_n\Big)=\lim_{n\to\infty}P\Big(M_{\text{MST}(\mathcal{P}_n)}\leq r_n\Big)$$
$$=\lim_{n\to\infty}P\Big(M_{\text{NNG}(\mathcal{P}_n)}\leq r_n\Big)=\lim_{n\to\infty}P\Big(M_{\text{MST}(\mathcal{P}_n)}\leq r_n\Big)=e^{-\tau}$$

where, with $\xi(u)=1/\omega'(u)$ and $\lambda(u)=$ length of the level curve $\mathbf{x}(\cdot,u)$,

$$r_n=1\xi(\log n)\Big(\log\big[\xi(\log n)\lambda(\log n)\big]-12\log_2\big[\xi(\log n)\lambda(\log n)\big]-\log\big(\tau\sqrt{2\pi}\big)\Big).$$

This generalizes the main results in Penrose (1998). We also consider a few cases where the level curves $x(\cdot,u)$ are not parallel, including skewed normal distributions.

References

Penrose, M. D. (1997). The longest edge of the random minimal spanning tree. *Ann. Appl. Probab.* **7** 340-361.

Penrose, M. D. (1998). Extremes for the minimal spanning tree on normally distributed points. *Adv. in Appl. Probab.* **30** 628-639.

Description of Ocean Waves: Applications of the Generalized Rice's Formula

Igor Rychlik
University of Lund
Centre for Mathematical Sciences
Box 118
S-221 00 Lund, Sweden
igor@maths.lth.se

The celebrated Rice formula, Kac (1943) and Rice (1944, 1945), for the expected number of times a stationary process X(t), 0<t<1, "crosses" a fixed level u has found application in various engineering problems, especially in safety analysis of structures interacting with the environment, for example through wind pressure, ocean waves or temperature variations. The safety of a structure may depend on extreme and rare events such as loads which exceed the strength of a component, or on everyday load variability that may cause changes in the properties of the material, e.g. cracking (fatigue) or other types of aging processes. In the first case, the number of rare events that occur in time or in space is often modeled as a Poisson process. Then, the Rice formula is used to compute the intensity of events, and hence gives the parameters in the Poisson model. In the second case, the aging process may depend both on the event frequencies and of their magnitudes. A magnitude of an event is called a ``mark". Such questions have resulted in the so called Slepian model processes, see Kac and Slepian (1959), describing process after an arbitrary crossing of level u.

In this talk we shall restrict our attention to the environmental loads originating from motion of ocean waves. Although, sea surface elevation for a developed sea W(x,y;t), say, is usually described by a two-dimensional field evolving in time, the loads acting on structures, like oil platforms or ships, are often related to the size and shape of individual waves. Consequently a sea surface is seen as a collection of moving "apparent waves". An observer easily recognizes individual waves. However, his impressions are difficult to formalize mathematically. Consequently, first we need to define the concept of an apparent wave, and then we specify some measurable properties of such a wave, called the wave characteristics. Wave height, wavelength, period, steepness or wave speed, are all examples of such characteristics. The most basic statistical problem in this context is to determine how many, in average, waves have some specific property, e.g. how many waves have crest above 10 meters. Answers to such questions depend both on what kind of data that are available and also on the way we measure "how many". For example, measurements of a buoy, sea level recorded by a moving vessel, photography or a movie of a sea, represent different data sets. In the case of a movie, where time dynamics is involved, there are, mathematically strictly speaking, uncountable many waves to analyze; giving a clear meaning to "how many" becomes crucial.

The first results of this types where obtained by Longuet-Higgins (1957) who derived formulas for the velocity of a point on a contour line defined as {(x,y): W(x,y;0)=u} for a fixed level u. (The contour line represents the front of a wave.) Here the term "how many" means the average length of the contour where the velocity satisfies some restrictions. The normalized length becomes a probability measure describing motion of waves. In the case of sea elevation a contour is a line in two-dimensional space. However the object can be generalized to higher dimensions. Then length of contour

becomes a suitable Hausdorf integral. The formula for the average "length" of the contour line is called a generalized Rice's formula, see work by Zähle (1984). The velocity of points on a contour, considered by Longuet-Higgins, was, maybe, the first application of this formula. Another example taken from this pioneering work is the velocity of local extremes of the sea surface. In this case the contour consists of isolated points and one wishes to compute the intensity of local extremes for which the velocity satisfies some restrictions, e.g. has speed higher than some threshold. The formula derived by Longuet-Higgins has been generalized (and formally proved) independently by Belyaev (1972) and Brillinger (1972). The result is a special case of the generalized Rice's formula.

In this talk we shall introduce the generalized Rice's formula and demonstrate its relation to the classical results from analysis such as; Banach theorem, see Banach (1925), and area and co-area theorem; see Federer (1969). The applications of the formula for computation of distributions of different wave characteristics will be given. Examples are taken from Podgorski et al (2000a), (2000b) and Rychlik (2001).

The formulas themselves would be worthless if they would not allow for effective computations. This is an issue here as the generalized Rice's formulas are complicated multidimensional integrals (often infinite dimensional). However for Gaussian fields refined numerical techniques have been implemented. They are available in the form of a MATLAB toolbox WAFO making the theory available for practitioners. The toolbox is available, free of charge, at:

http://www.maths.lth.se/matstat/wafo/.

References

Banach, S. (1925). Sur les lignes rectifiables et les surfaces dont l'aire est finie. *Fund. Math.*, 7, 225-237.

Belyaev, Yu.K. (1972) The general formula for the mean number of crossings for random processes and fields, in Yu. K. Belyaev, ed. *Bursts of Random Fields* Moscow University Press (in Russian).

Brillinger, D.R. (1972) On the number of solutions of systems of random equations. *Ann. Math. Statist.*, **43**, 534-540.

Federer, H. (1969). Geometric Measure Theory. Springer Verlag, New York.

Kac, M. (1943). On the average number of real roots of a random algebraic equation. *Bull. Amer. Math. Soc.*, **30**, 1215-1228.

Kac, M., Slepian, D. (1959) Large excursions of Gaussian processes. *Ann. Math. Statist.*, **30**, 1215--1228.

Longuet-Higgins, M. S. (1957) The statistical analysis of a random, moving surface. *Phil. Trans. Roy. Soc.* A, **249**, 321-387.

Podgorski, K., Rychlik, I., Sjö, E. (2000a) Statistics for velocities of Gaussian waves. *Inter. J. of Offshore and Polar Engineering*, **10**, 91-98.

Podgorski, K., Rychlik, I., Ryden, J., Sjö, E. (2000b). How big are the big waves in Gaussian sea. *Inter. J. of Offshore and Polar Engineering*, **10**, 161-169.

Rice, S.(1944). The mathematical analysis of random noise. *Bell Syst. Techn. J.*, **23**, 282-332.

Rice, S. (1945). The mathematical analysis of random noise. *Bell Syst. Techn. J.*, **24**, 46-156.

Rychlik, I. (2001) On some reliability applications of Rice's formula for the intensity of level crossings. To appear in *Extremes*.

Zähle, U. (1984). A general Rice formula, Palm measures, and horizontal-window conditioning for random fields. *Stochastic Processes and their Applications*, **17**, 265-283.

REVISTA DE ESTATÍSTICA
105ª PÁGINA
2º QUADRIMESTRE DE 2001

QUANTUM PROBABILITY AND STATISTICS

Organizer: Inge Helland

Invited Speakers: Luigi Accardi
Viachaslav P. Belavkin
Richard D. Gill

REVISTA DE ESTATÍSTICA
107ª PÁGINA
2º QUADRIMESTRE DE 2001

Quantum Probability and Statistics

Inge S. Helland
Department of Mathematics, University of Oslo,
P.O.Box 1053 Blindern, N-0316 Oslo, Norway.
ingeh@math.uio.no

Both quantum theory and modern statistics had its initial development in the beginning of the last century. Both are based on probability theory, and both are concerned with the prediction of new observation on the basis of data. Nevertheless, the lack of scientific contact between the two disciplines has been striking. The mathematical foundation of quantum mechanics was laid by von Neumann in the 1930th at about the same time as Kolmororov gave us the foundation of ordinary probability theory. At about the same time as Dirac was developing relativistic quantum theory in Cambridge, R.A. Fisher completely independently founded modern statistical inference in Rothamsted and London. Similarly, while modern quantum field theory was being developed by Feynman in Princeton and Schwinger in Harvard, J. Neyman and coworkers developed the statistical theory as we know it today. One of the few early contacts is Feynmann's (1951) Berkeley Symposium paper on the interpretation of probabilities in quantum mechanics.

The lack of contact between the two disciplines is of course closely related to the difference in foundation. In statistics, the state of a given system is given simply by a probability measure on some measurable space. In quantum theory in its most common formulation the state of a system is given by a vector v in some abstract Hilbert space, each observator is associated with a selfadjoint operator A on the same Hilbert space in such a way that the expectation of this observator in the state v is given by (v,Av). A consequence of this is that one gets transition probabilities of the form $|(v,u)|^2$. Also, in the absence of socalled superselection rules, linear combination of statevectores form new statevectors, which lead to interference phenomena unknown to classical statistics. Related to this are several apparent paradoxes of quantum mechanics, which still are much discussed in the literature.

The quantum formalism as such is the result of a long development within physics, starting with discoveries by Max Planck, and where contributions have been made by Bohr, Pauli, Schrödinger, Heisenberg and many others. There are many good books on quantum theory. Two of these, which can be recommended because they also include discussions of philosophical aspects, are Peres (1993) and Isham (1995).

Many authors have tried to find deeper foundations leading to the formalism of quantum theory. Several mathematical approaches are discussed in Wightman (1976). One such approach is quantum logic, treated in detail by Beltrametti and Cassinelli (1981).

The earliest book on the mathematical foundation of quantum mechanics is von Neumann (1932); in English translation: von Neumann (1955). This book has had great influence, and it can be considered to be a forerunner of quantum probability. For physicists, von Neuman's book is supplemented by the book of Dirac (1930), which also may be looked upon as a forerunner for modern quantum field theory.

The development of quantum probability as a mathematical discipline was started in the 1970's. A first important topic was to develop a noncommutative analogue of the notion of stochastic processes; see Accardi (1976) and references there. Other topics were noncommutative conditional expectations and quantum filtering and prediction theory (Belavkin, 1985).

Quantum probability was made popular among ordinary probabilists by Meyer (1995). A related book is Parthasarathy (1992), which discusses the quantum stochastic calculus founded by Hudson and Parthasarathy, but also many other themes

related to the mathematics of current quantum theory. An account of several differents topics in quantum probability may be found in the series Quantum Probability and Applications edited by L. Accardi and co-workers. An example of a symposium proceeding aiming at covering both conventional probability theory and quantum probability is Accardi and Heyde (1998).

There are also links between quantum theory and statistical inference theory. A systematic treatment of quantum hypothesis testing and qantum estimation theory was first given by Helstrom (1976). In Holevo (1982) several aspects of quantum inference are discussed in depth; among other things the book contains a chapter on symmetry groups. A survey paper on quantum inference is Malley and Hornstein (1993).

As an example of a particular topic of interest, consider that of Fisher information. Since a quantum state ordinarily allows several measurements, this concept can be generalized in a natural way. A quantum information measure due to Helstrom can be shown to give the maximal Fisher information over all possible measurements; for a recent discussion see Barndorff-Nielsen and Gill (2000).

In this way one can point at several links between ordinary probability and statistics on the one hand and their quantum counterparts on the other hand. However, a general theory encompassing both sides, based on a reasonably intuitive foundation, is still lacking; many will say that it is impossible to find such a theory.

References

Accardi, L. (1976). Non relativistic quantum mechanics as a noncommutative Markov process. *Adv. in Math.* **20**, 329-366.

Accardi, L. and C.C. Heyde (1998). *Probability Towards 2000*. Springer-Verlag, New York.

Barndorff-Nielsen, O.E. and R.D. Gill (2000). Fisher Information in Quantum Statistics. To appear in J. Phys. A. Available on *http://www.math.uu.nl/people/gill/pub.html*.

Belavkin, V.P. (1985). Non-demolition measurement and control in quantum dynamical systems. Proc. Of the conf. "Information complexity and control in quantum physics", Udine 1985. Springer-Verlag, 331-336.

Beltrametti, E.G. and G. Cassinelli (1981). *The Logic of Quantum Mechanics*. Addison-Wesley Publishing Company, London.

Dirac, P.A.M. (1930). *The Principles of Quantum Mechanics*. The Claredon Press, Oxford.

Feynman, R.P. (1951). The Concept of Probability in Quantum Mechanics. In: *Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley.

Helstrom, C.W. (1976). *Quantum Detection and Estimation Theory*. Academic Press, New York.

Holevo, A.S. (1982). *Probabilistic and Statistical Aspects of Quantum Theory*. North-Holland Publishing Company, Amsterdam.

Isham, C.J. (1995). *Lectures on Quantum Theory*. Imperial College Press, London.

Meyer, P.-A. (1995). *Quantum Probability for Probabilists*. Springer-Verlag, Berlin.

Malley, J.D. and J. Hornstein (1993). Quantum Statistical Inference. *Statistical Science* **8**, 433-457.

Parthasarathy, K.R. (1992). *An Introduction to Quantum Stochastic Calculus*. Birkhäuser Verlag, Basel.

Peres, A. (1993). *Quantum Theory: Concepts and Methods*. Kluwer Academic Publishers, Dordrecht.

von Neumann, J. (1932). *Mathematische Grundlagen der Quantenmechanik*. Springer, Berlin.

von Neumann, J. (1955). *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, Princeton.

Wightman, A.S. (1976). Hilbert's Sixth Problem: Mathematical Treatment of the Axioms of Physics. In: *Proceeding of the symposium in pure mathematics of the AMS*, Northern Illinois University, 1974. Mathematical developments arising from Hilbert problems. American Mathematical Society, Providence, Rhode Island.

REVISTA DE ESTATÍSTICA
109ª PÁGINA
2º QUADRIMESTRE DE 2001

Quantum Theory, Chameleons and the Statistics of Adaptive Systems

Luigi Accardi
Centro Vito Volterra
Università degli Studi di Roma "Tor Vergata"
Facoltà di Economia
Via di Tor Vergata, snc
00133 Rome, Italy
accardi@volterra.mat.uniroma2.it

The fact that the mathematical formalism, used to make predictions on quantum phenomena, is completely different from the mathematical formalism used in classical probability, was recognized since the early days of quantum theory in the late 1920's.

However only in the early 1980's it was realized that this difference of the mathematical formalisms reflects a deeper probabilistic phenomenon: namely the fact that, in general, the Bayes definition of "conditional probability" cannot be applied to the conditional probabilities that are considered in quantum theory (cf. [Ac81a], [Ac84] and [Ac86]) for a more systematic exposition.

This discovery lead to a rethinking of the basic axioms of probability theory in the light of the new probabilistic ideas emerged from quantum physics.

The new elements that quantum theory bought into the probabilistic thought are essentially two:

1. the existence of incompatible events
2. the "chameleon effect"

Classical probability is based on Boolean logic, where it is postulated that the "join" of two meaningful proposition P and $Q(P \cap Q)$ is always a meaningful proposition. However, if we replace the platonic notion of "proposition" by the empirical notion of "experimentally verifiable proposition", we see that the existence of incompatible events implies the existence of statements which, when considered individually, are empirically verifiable, but whose "joint event" is not. The most famous example, first pointed out by Heisenberg, is given by the two statements:

- $Q(t)$: the position of a particle at time t is in an interval I
- $P(t)$: the momentum of a particle at time t is in an interval J

It is known that, if the intervals I and J are small enough, the joint event $Q(t) \cap P(t)$ is not empirically verifiable because of the Heisenberg principle. Notice that this principle concerns a single particle and no probability is involved. A few years later H. Weyl formulated a statistical variant of the Heisenberg principle, expressing the fact that the product of the covariances of position and momentum in a given quantum state cannot be smaller than a certain quantity $(h/2\pi)$ which is independent of the state considered.

The Heisenberg principle, in its original formulation, expresses a physical limitation, not a logical impossibility. There is no logical reason why we should not be able to measure with arbitrary precision, on the same particle and at the same time,

position and momentum. In fact classical mechanics is a perfectly coherent theory from the logical point of view, but does not contemplate such an impossibility.

The development of quantum probability has brought to light new forms of incompatibility between pairs of events, which have a logic rather than physical root. This type of incompatibility is not peculiar to quantum theory.

Consider, for example, the response of an iron ball, rotating along fixed axis A to the action of a constant magnetic field, directed along a different axis B. For example one can imagine that B is the z-axis of a frame, that the rotating particle is "fired" in a direction perpendicular to the B-axis, and that one measures if it deviates on the left or on the right of the B-axis. This is a perfectly measurable event and the same is true if we replace the B-axis by another axis C, not parallel to B.

Let us denote $[A,B]$ the former event and $[A,C]$ the letter.

A little thought shows that the joint event $[A,B]\cap[A,C]$ cannot be experimentally realized because a single particle cannot be simultaneously subject to the "only action of the B-field" and to the "only action of the C-field": two different magnets cannot be put in the same space point.

Another example can be taken from medicine: suppose that two different medicines B and C are proposed as a cure of the same illness A. We can separately experiment medicine B or medicine B on a patient, but again the point event $[A,B]\cap[A,C]$ makes no sense because the same patient, at the same time, cannot be cured "only by medicine B" and "only by medicine C".

A third example is given by the colour of a chameleon: the chameleon is in a box and an experimentalists can measure either its colour on a leaf C_F or its colour on the wood C_W. It is clear that the simultaneous point event $C_F\cap C_W$ makes no sense because the same chameleon cannot be at the same time both "only on the leaf" and "only on the wood".

Notice the physical difference between the three examples described above and the following, more classical situation: in a box there are many balls whose colour can be either green or brown. Moreover each ball is either made of glass or of wood. Clearly in this case we can make an experimental analysis of the joint statistics "colour-material".

The difference between the two types of examples considered above is that, in the former case one measures the "response" to an action; in the latter one measures a property which is pre-existing to the measurement and independent of it.

Notice however that in all the above situations the results of the measurements are "pre-determined": the laws of classical electromagnetism allow (in principle) to predict exactly the deviation of the particle once the initial data (velocity, angular velocity, strength of the magnetic field,...) are known; the laws of chemistry and biology allow a similar prediction for the medicines; the knowledge that the chameleon is a usual one and not a mutant which becomes brown on a leaf and green on a piece of wood, allows to predict deterministically its colour; finally for balls the situation is even simpler.

However it is also clear that the word "predetermination" is used with a different meaning in the two contexts: in the case of balls "we measure what it was"; in the case of chameleons "we measure what it happens".

In the former case we will speak of "passive systems"; in the latter we will speak of "adaptive systems".

The difference between the two cases has rather deep consequences on the type of inductions that we can make in the two situations. Since statistical inference is a generalization of inductive inference, it is clear that these consequences will also have non trivial implications on the kind of statistical inference we can make on the two types of systems.

To illustrate the implications for statistics of the difference between active and passive systems let us consider the following variant of the above experiments: consider two boxes; in one there are pairs of balls; each pair contains 1 green ball and 1 brown ball, moreover one ball of the pair weights 10 grams and the other one 20 grams. The other box contains pairs of chameleons: exactly one, in each pair weights 10g and the other 20g, moreover, in each pair, exactly one is healthy (becomes green on a leaf, brown on a piece of wood) and the other one is a mutant (green on wood, brown on leaf). In both cases we do not know the statistics of the joint distributions colour-weight.

Suppose you are interested in such a statistics at time t, but the rules of the game are such that you can do only one measurement at a time, both on balls or on chameleons (incompatibility). A reasonable strategy for balls would be the following: at time t you measure the colour of one ball and the weight of the other one. Suppose you find: "brown" in ball 1 and "10g" in ball 2. Then you conclude that ball 1, at time t, is brown and weights 20g and ball 2, at time t, is green and weights 10g.

Suppose now you want to apply the same strategy of measurement to chameleons. Can you draw the same conclusion? No of course! In fact suppose that, at time t, you measure the weight of chameleon 2 while he was in the box and you find 10g. Suppose you measure, at time t, the colour of chameleon 1 on the leaf and you find "brown". You can only be sure that "if you had measured the colour of chameleon 2 on the leaf you would have found green". However nothing prevents the possibility that the chameleon in the box, i.e. at time t is brown. Since you are interested in the joint statistics colour-weight at time t, you are not allowed to make, for chameleons, the same inference you made for balls.

Can we push this difference further? The answer is "yes". In my talk a simple experimental situation will be described in which, by exploiting the chameleon effect, one can reproduce exactly, by local independent binary choices of individuals who are situated far away of each other and do not communicate with each other, exactly the same empirical correlations that are experimentally obtained in the well known Einstein-Podolsky-Rosen experiments. In the past 37 years the possibility of realizing such an experiment has been firmly denied by the entire community of physicists.

Without using the Chameleon effect, the possibility of such a reproduction is excluded by a mathematical constraint: an inequality among these correlations, discovered by Bell and violated by some quantum mechanical system. The probabilistic meaning of this inequality, first pointed out in [Ac81a] will be shortly reviewed.

References

Accardi, L. (1981). Topics in Quantum Probability, *Physics Reports*, **77**, 169-192.

Accardi, L. (1984). The probabilistic roots of the quantum mechanical paradoxes. In: *The wave-particle dualismm*, ed. S.Diner et al.:, Reidel 297-330.

Accardi, L. (1988). Foundations of Quantum Mechanics : a quantum probabilistic approach. In *The Nature of Quantum Paradoxes* ; eds. G.Tarozzi, A. van der Merwe Reidel, 257-323.

Accardi, L., Regoli, M. (2000). Experimental violation of Bell's inequality by local classical variables. Poster discussed at the Towa Statphys conference, Fukuoka 8-11 November (1999), in: *Statistical Physics*, M. Tokuyama, H.E. Stanley (eds.), American Institute of Physics, AIP Proceedings 519, 645-648.

Accardi, L., Regoli M. (2000). Non-locality and quantum theory: new experimental evidence, Paper: quant-ph/0007019, Comments: 21 pages, Plain TeX, 9 eps pictures. A talk given at Nottingham conference, 5 July 2000.

Accardi L. (2000). Locality and Bell's inequality. Preprint *Volterra*, N. **427**, quant-ph/0007005.

REVISTA DE ESTATÍSTICA
113ª PÁGINA
2º QUADRIMESTRE DE 2001

Quantum Filtering and Bayesian Quantum Mechanics

Viacheslav P. Belavkin
School of Mathematics
University of Nottingham
NG7 2RD, UK
vpb@maths.nott.ac.uk

The Bayesian method in quantum signal processing was initiated by C W Helstrom and developed in quantum detection [1], estimation [2,3] and hypothesis testing [4] theory in the 70th. The aim of this theory was to find an optimal quantum measurement, minimizing a cost function of the quantum state estimation under the given probabilities of the possible states. The usefulness of the entropy restrictions for finding of quantum estimation was shown in [2], where the quantum regression problem was solved under the condition of fixed entropy of quantum measurement. This corresponds to the maximal entropy principle in classical statistics.

In the 80th we developed these methods into the dynamical theory of quantum estimation, prediction and filtering using the analogy with the classical filtering theory for Markov processes. The main result of this development was the derivation of new stochastic wave equation for quantum posterior states in quantum mechanics with continuous nondemolition observation. This equation in its non-normalised and normalised forms plays a similar role in quantum statistics as the Zakai or Stratonovich equation in the classical Bayesian statistics of stochastic processes. In the beginning of 80th this equation was taken by physicists as the stochastic quantum Master equation of the Modern Quantum Theory which treats quantum dynamics together with classical trajectories of quantum experimental events such as jumps, diffusions, reductions and localizations. We show how the paradoxes of old quantum mechanics can be reduced to the statistical problems of quantum state estimation in the framework of the modern quantum mechanics which we call here the Bayesian quantum mechanics. The main purpose of this quantum mechanics is the Bayessian predicrion of the statistics of quantum events, and the main mathematical tool is the quantum filtering equation for the posterior quantum states.

Quantum entanglements give seemingly a possibility to gain more information in a quantum system than in the corresponding classical one, described by the same rank algebra. We show it by analysing the mathematical notion of quantum entanglement and comparing the mutual information achieved via quantum entanglements with the semi-classical one, achieved via encodings, corresponding to the nondemolition measurements. In order to prove that it is a real achievement, we have to introduce the notion of value of quantum information in the sense of valuableness of one bit of information for the purpose to achieve a certain aim. We find the value of so called q-bit, achieved via quantum entanglements, in comparison with the value of a classical c-bit, achieved via the semi-classical entanglements. This makes a link of quantum information theory with the theory of quantum estimation and hypothesis testing.

We consider the dual maximum entropy problem for the quantum measurements under the condition of a fixed mean error of quantum estimation. The corresponding classical problem, well-known as the Kolmogorov epsilon-entropy problem, was elaborated into the theory of value of (classical) information by R

Stratonovich [5]. So we give the formulation and preliminary results for a new branch of quantum measurement and information theory, the quantum epsilon-entropy theory, related with classical problem of optimal quantization [6]. We restrict ourselves to the simplest quantum systems, described by the algebra of all operators in a Hilbert space. The general case will be developed elsewhere by use of more general notion of the entropy [7] and relative entropy [8] within the C*-algebraic approach to quantum information.

References

[1] C W Helstrom, Quantum detection and estimation theory, Academic Press, 1976.

[2] V P Belavkin, Optimal quantum randomized filtration, *Problems of Control and Information Theory*, 3--25, 1974.

[3] A S Holevo, Probabilistic aspects of quantum theory, Kluwer Publisher, 1980.

[4] V P Belavkin, Optimal multiple quantum hypothesis testing, *Stochastics*, 3--40, 1975.

[5] R L Stratonovich, Theory of information, Sov Radio, Moscow, 1976.

[6] V P Belavkin, Optimal quantization of random vectors, Isvestia A N USSR, *Techn Kibernetika*, 1--20, 1970.

[7] F Hiai, Ohya M, Tsukada M, Sufficiency, KMS condition and relative entropy in von Neumann algebras, *Pacific J Math*, 93--99, 1981.

[8] V P Belavkin, Staszewski P, C*-algebraic generalization of relative entropy and entropy, *Ann Inst H Poincar'e, Sect* A **37**, 1982.

[9] V P Belavkin, M Ohya, Quantum Entanglements and Entangled Mutual Entropy. *Los Alamos Archive*, Quant.-Ph/9812082, 1--16, 1998.

[10] V P Belavkin, Measurement, Filtering and Control in Quantum Open Systems. Report in *Mathematical Physics*, Vol. **43**, No 3, 405--425, 1999.

Teleportation into Quantum Statistics

Richard Gill
Mathematical Institute
University Utrecht
Box 80010 3508 TA Utrecht
Netherlands
gill@math.uu.nl

I first discuss an experiment recently carried out in Delft, the Netherlands, by the group of Hans Mooij, see *http://vortex.tn.tudelft.nl/*. The experiment was reported in *Science* in 1999. Switching on a magnetic field causes electric current to flow around a superconducting aluminium ring. The aluminium ring is a thousandth of a millimeter in diameter, and a billion electrons are involved in the current flow. From a classical physical viewpoint one can imagine just two kinds of current flow of a given size in this little circuit: *clockwise*, and *anti-clockwise*. The claim of the experimenters was that they produced an electric current in the state $\alpha|\text{clockwise}\rangle+\beta|\text{anticlockwise}\rangle$, where α and β are two complex numbers, with $|\alpha|^2+|\beta|^2=1$. $|\text{clockwise}\rangle$ and $|\text{anticlockwise}\rangle$ stand for two orthogonal unit vectors in a complex vector space, we can think of them as two-dimensional complex column vectors, say the unit vectors, (1 0) and (0 1), transpose. This object has been called *The Delft Qubit*; a qubit being a single bit in the memory of a future quantum computer. A classical computer works with a memory, the bits of which can register only 0 or 1, however a quantum computer allows *coherent superpositions* of 0 and 1, such as the state I have just talked about. Another description is *The Schrödinger Squid*; this name refers to the device: a Superconducting Quantum Interference Device; and to the infamous Schödinger cat. Now one might ask, how could the experimenters know that this state has been produced? Well, by repeating the experiment about ten thousand times, and each time measuring the current. This is done by a second squid, surrounding the first, and connected to the outside world by a lot of circuitry. It does not directly give us estimates of α and β. In fact, in first instance, it does nothing interesting at all: the measurement essentially looks to see whether the current is flowing clockwise or anticlockwise. This forces the quantum state to jump into either of the states $|\text{clockwise}\rangle$ or $|\text{anticlockwise}\rangle$, and it makes this choice with probabilities $|\alpha|^2$ and $|\beta|^2$. The experimentalists find the same values of these probabilities (relative frequencies), as are predicted by an elaborate theoretical physical calculation concerning the whole system.

So this does not *prove* anything at all: one would have seen the same relative frequencies, if the qubit had from the start been, in a fraction $|\alpha|^2$ of the times, in state $|\text{clockwise}\rangle$, and in a fraction $|\beta|^2$ of the times, in state $|\text{anticlockwise}\rangle$. However, small developments in the technology of this experiment will make the finding more secure. The aim is not just to create qubits but to manipulate them. In particular, it should be possible to implement the following linear (unitary) transformation of the state, sending the basis vectors (1 0) and (0 1) (transposed), into (1 1) and (1 -1) (divided by root 2, and transposed). The result of this *unitary transformation* is to convert the original qubit into the state $(\alpha+\beta)|\text{clockwise}\rangle+(\alpha-\beta)|\text{anticlockwise}\rangle$ (divided by root 2). If we now measure, we will find relative frequencies of $|\alpha+\beta|^2/2$ and $|\alpha-\beta|^2/2$, different from the relative frequencies had the state been inititally in a fraction $|\alpha|^2$ of the times,

|clockwise>, and in a fraction $|\beta|^2$ of the times, anticlockwise (as the reader may compute, one would then observe 50:50 clockwise, anticlockwise).

The idea of *quantum computation* is to store programme and input of some algorithm, coded in a sequence of 0's and 1's, into the basis states |0> and |1> of a large number of qubits. The whole system next evolves unitarily, and at the end of the computation, a series of (possibly random) zeros and ones are read off by measuring each qubit separately. The possibilities allowed by the basic model of quantum mechanics allow, for instance, (with an algorithm of Peter Shor) to factor large integers in polynomial time, which will make all currently used cryptography methods obsolete! Fortunately quantum cryptography promises a secure alternative. One cannot look at a qubit without disturbing it, and if this idea is cleverly exploited, it becomes possible to transmit messages coded in qubit states, such that the action of any eavesdropper would be detected by the recipient.

What is the basic mathematical model behind all this, what then are the statistical problems, and what do we know about the solutions? We have seen the notion of *states* (more precisely, pure state), mathematically formalized as unit vectors in a complex vector space. States can be *unitarily transformed*, that is to say, one may implement an orthogonal transformation (change of basis) and get a new state. In principle, any desired unitary transformation could be implemented by setting up appropriate external fields. It is a manipulation of the state of the quantum system, involving, for instance, magnetic fields, which one can control, but without back-action on the real world outside. No information passes from the quantum system into the real world. What I have not yet told you is the mathematical model for bringing initially separate quantum systems into (potential) interaction with one another. This is the essential ingredient of the quantum computer: one should not have N separate qubits, but one quantum system of the N qubits together. The appropriate model for this is the *formation of tensor products*. In words, two separate systems brought together have as state, a vector in a space of dimension equal to the product of the two original dimensions; and the new state vector has as components, all the products of a component of each of the two original state vectors. The N qubits of a quantum computer live in a 2^N dimensional state space. The initial state is a product state, but a unitary evolution can bring the joint system into a state, which cannot be represented as a product state. This phenomenon is called entanglement.

The last ingredient has already been touched upon, and that is *measurement.* At this stage, and only at this stage, is information passed from the quantum system into the real world. The information is random, but its probability distribution depends on the state of the system. The system makes itself makes a random jump. The basic measurement is characterized by a collection of orthogonal subspaces of the whole state space, together spanning the whole space; together with a real number or label, associated to each subspace. This collection of subspaces and numbers somehow corresponds to an experiment one might do in the laboratory. When the experiment is carried out, the state vector of the quantum system gets projected into one of the subspaces (and renormalized to have length one); the corresponding real number or label becomes known in the real world; and that happens with probability equal to the squared length of the projection of the original state vector into the subspace. By Pythagoras, these squared lengths add up to 1.

These are all the ingredients: state vectors (also called pure states), unitary evolution, entanglement (formation of product systems), and (simple) measurement. In my talk I will illustrate them by the beautiful example of quantum teleportation, discovered by Charles Bennett (IBM) *et al.* in the mid nineties, and done in the laboratory, just a couple of years later, by Anton Zeilinger, in Innsbruck. The experiment is done with polarized

photons, and the basic states can be thought of as |horizontal>, |vertical>. The problem is as follows. Alice is given a qubit (polarized photon) in an unknown state. She wants to transmit it to Bob, and can only communicate with Bob by email. What can she do? She could measure the qubit, e.g., look to see if the photon is polarized horizontally or vertically. She gets the answer: yes or no; it is random, with probabilities depending on the unknown α, β. The photon's original state is now destroyed, we cannot learn anything more about it. So all she could do is email to Bob: it was (e.g.) horizontal. He makes a horizontally polarized photon. This is a poor, random, copy of the original one, and the original one has gone. Can they do better? Well, there are many other measurements Alice could make, but they all have the same property, of only providing a small, random, amount of information about the original state, and destroying it in the process. In fact it is an old result from the theory of quantum statistical inference, that whatever measurement is carried out by Alice, the Fisher information matrix based on the probability distribution of the outcome of the experiment, concerning the unknown parameters α, β, has a strictly positive lower bound.

In order to succeed, Alice and Bob need a further resource. What they do is arrange that each of them has another photon, these two (extra) photons in the joint state |0> tensor-product |1> - |1> tensor-product |0> (divided by root 2). This is nowadays a routine matter. Now we have three qubits, living together in an eight-dimensional space, of which four of the dimensions - two of the qubits - are on Alice's desk, the other two dimensions - one qubit - on Bob's desk. In my talk I will show you three lines of elementary algebra, with the astounding implication that Alice can carry out a measurement on her desk, get one of 4 random outcomes, each with probability 1/4, then email to Bob which outcome she obtained; he correspondingly carries out one of 4 different, prescribed, unitary operations, and now his photon has magically transformed into an identical copy of the original, unknown, qubit which was given to Alice. Two (unknown) complex numbers α and β have been transmitted, with complete accuracy, by transmitting two bits of classical information .

Now it is worth asking: how can we know that a certain experiment has actually succeeded? The answer is of course by statistics. One needs, many times, to provide Alice with qubits in various states. Some of these times, the qubits are not teleported, but are measured in Alice's laboratory. On the other occasions, the qubits are teleported to Bob, and then measured in Bob's laboratory. The predictions of quantum theory are that the *statistics* of the measurements at Alice's place, are the same as the statistics of the measurements at Bob's place.

Now I am close to describing new and interesting statistical problems. For instance, suppose I am given N qubits in an identical, unknown, state, what is the best way to determine that state? It is known that whatever one does, one cannot do better than a certain inaccuracy, of the order of size of 1 over root N. It is not known what *constant* over root N, is best. And a most intriguing question, only partially solved, is: does it pay off to consider the N qubits as one joint system, having a state of rather special form in a 2^N dimensional state space, or can one just as well measure them separately? Note that considered collectively, we have a much vaster repertoire of possible measurements, so from a mathematical point of view, the answer should surely be that joint measurements pay off. However physical intuition would perhaps say the opposite. I have worked on asymptotic versions of this problem. So far the physicists have hardly considered this route, and the literature has largely seen calculations in rather special situations (N=2, for instance), with conclusions which depend on all kinds of features of the problem - prior distributions if you are a Bayesian, loss functions in any case - which are really arbitrary.

The advantage of my approach is that these extraneous and arbitrary features become irrelevant for large, but finite N; the problem *localizes*, second order approximations are good, loss functions might as well be quadratic, prior distributions are irrelevant. Using the van Trees inequality (a Bayesian Cramér-Rao bound) I have, together with Serge Massar, derived *frequentist* large sample results on what is asymptotically best, under various measurement resource scenarios. At *http://www.math.uu.nl/people/gill* are reprints of a survey paper on quantum asymptotic statistics, Gill (2001, *IMS Monograph* 36, 255-285), the original work Gill & Massar (2000; *Phys. Rev. A* 61, 2312-2327), other reprints, and work in progress, including a new survey paper on quantum statistical inference, Barndorff-Nielsen, Gill & Jupp (2002, to appear).

Before describing our work, I must extend the notion of state used so far. Above, a so-called pure state is described by two complex numbers α , β ; $|\alpha|^2+|\beta|^2=1$. I cheated a little: one can multiply these two numbers by the same, but arbitrary, complex number of absolute value 1, and we are still talking about the same state (the same future predictions, same statistics, for whatever measurements). So I can renormalize the state so that $\alpha = \cos \theta/2$, $\beta = e^{i\phi} \sin \theta/2$, where θ, ϕ are real angles, and we are still talking about the same state. Now it turns out that one can usefully consider the angles θ, ϕ as polar coordinates of a point on the surface of the unit sphere in real three-dimensional space. This geometric picture corresponds sometimes to directions in the real world, for instance if we had been talking about spin of an electron. Moreover, the most simple measurement devices also correspond to real directions, and the probabilities of different outcomes can be read off from the joint geometric picture of state and measurement. In particular, a measurement of spin of an electron in a (real, physical) direction, produces spin-up, and spin-down (relative to that direction), with probabilities proportional to the lengths by which the projection of the state (point on the surface of the sphere), onto the diameter of the sphere in the direction of the measurement, divides the diameter.

Suppose we had not been given particles in a given, pure state, but particles in various states according to a probability distribution over the surface of the sphere. Then it turns out that all that counts, for predictions and measurement statistics, is the centre of gravity of that mass distribution over the surface of the sphere: namely, a point inside the unit ball. Such a state is called a mixed state. Not just for qubits, but in complete generality, one can extend the notion of pure, to mixed states, which are in general represented by certain complex matrices called density matrices. We must also consider more general measurements. One might for instance take a particle in an unknown state, bring it into interaction with another, in some known state; after a unitary evolution, measure the auxiliary particle, discard it, bring in a third, carry out another unitary evolution and do another simple measurement, all depending on the results of the first. Considering *all* possibilities together we have a vast spectrum of possibilities for getting real data out of measurements of a quantum system. Surprisingly, there is a compact mathematical representation of every possible way (using only the ingredients above) to measure a quantum system, using the notion of Operator-valued Probability Measure. It beautifully meshes with the notion of mixed states, and provides us with a clean mathematical framework which can be used as a starting point for constructing optimal experiments.

The most exciting result we have found is as follows: if the unknown state is known to be pure, then a certain very simple but adaptive strategy of basic yes/no measurements on the qubits, achieves the maximal achievable accuracy. If however the state is mixed, then we do not know the best strategy. Limited to separate measurements, we do know what can be achieved. We know that joint measurements can achieve startling increases in accuracy. But we do not know how much can be maximally achieved (there are known bounds, but they are known to be inachievable). This seems to be a promising future research direction.

REVISTA DE ESTATÍSTICA
119ª PÁGINA
2° QUADRIMESTRE DE 2001

RECENT DEVELOPMENTS IN TIME SERIES

Organizer: Jan Beran

Invited Speakers: Liudas Giraitis
Yuanhua Feng
Donald B. Percival

VOLUME II

2º QUADRIMESTRE DE 2001

REVISTA DE ESTATÍSTICA
121ª PÁGINA
2º QUADRIMESTRE DE 2001

Recent Developments in Time Series

Jan Beran
University of Konstanz
Germany
JanBeran@compuserve.com

Some of the currently most active areas in time series analysis are nonlinear processes (in particular modelling of volatility), long-range dependence, non- and semiparametric inference, distinction between various types of stationarity and nonstationarity, and wavelets. The variety of topics and methods in these areas is illustrated by the following three talks.

Liudas Giraitis (joint work with Kokoszka, Leipus, Robinson, Surgailis) considers nonlinear time series models with long memory, with special emphasis on temporal dependence in volatility, leverage effect and applications to financial time series. Many time series in finance exhibit almost no correlations in the returns but strong and possibly long-range dependence in volatility. A second effect is a negative cross-correlation between volatility and levels, the so-called leverage effect. In this talk several classes of nonlinear models that are suitable modifications of ARCH models (in particular the so-called LARCH model) are discussed. In particular, conditions are derived under which long memory in levels and volatility, and the leverage effect occur. These results are of fundamental importance for modeling volatility and leverage effect, with short- or long-range dependence.

Yuanhua Feng (joint work with Jan Beran) discusses semiparametric time series models, in particular the so-called SEMIFAR model.

The SEMIFAR model incorporates stationarity, difference stationarity, nonparametric trends and a fractional dependence structure, including short memory, antipersistence and long memory. An important problem addressed in this context is how to distinguish between long-range dependence, strong short-range dependence, stochastic nonstationarity and deterministic trend. A key issue is optimal data-driven nonparametric smoothing. Here, recent methods, algorithms, asymptotic results, simulations and applications to observed series (in particular from finance) are discussed. Most of the algorithms are based on an iterative procedures using asymptotic expressions for the integrated mean squared error of the estimated trend function. A distinction between the various components is possible sample sizes of about 200 and above. Further improvements are to be expected by using more accurate finite sample criteria.

An alternative way to dealing with possible trend components is outlined in the talk by Peter Craigmile (joint work with Donald Percival).

The wavelet transfom is particularily suited to separate a deterministic trend from random, and possibly nonstationary noise, and for estimating the dependence parameters of fractional processes. This is demonstrated in the specific context of a polynomial trend plus a fractionally differenced process. The trend is separated from the stochastic component using the discrete wavelet transform (DWT) and a maximum likelihood approach.

These results are a further important step towards building a wavelets-toolkit for the analysis of time series that may be nonstationary and may exhibit a wide range of dependence structures, including antipersistence, short- and long-memory.

Wavelet-Based Maximum Likelihood Estimation for Trend Contaminated Long Memory Processes

Peter F. Craigmile
Department of Statistics
University of Washington
Box 354322
Seattle, USA
WA 98195-4322
pfc@stats.washington.edu

Donald B. Percival
Applied Physics Laboratory
Box 355640
University of Washington
Seattle, USA
WA 98195-5640.
Insightful Corporation
1700 Westlake Avenue North
Suite 500
Seattle, USA.
WA 98109-9891
dbp@apl.washington.edu

A common problem in the analysis of time series is how to deal with a possible trend component, which is usually thought of as large scale (or low frequency) variations or patterns in the series that might be best modelled separately from the rest of the series. Trend is often confounded with low frequency stochastic fluctuations, particularly in the case of models such as fractionally differenced (FD) processes, which can account for long memory dependence (slowly decaying auto-correlation) and can be extended to encompass non-stationary processes exhibiting quite significant low frequency components. In this talk we assume a model of polynomial trend plus FD noise and apply the discrete wavelet transform (DWT) to separate a time series into pieces that can be used to estimate both the FD process parameters and the trend. The estimation of the process parameters is based on an approximative maximum likelihood approach that is made possible by the fact that the DWT decorrelates FD process approximately. We discuss the large sample theory for estimators based upon this approach.

ARCH Models with Long Memory

Liudas Giraitis
London School of Economics
Department of Economics
Houghton Street, London, WC2A 2AE, U.K.
L.Giraitis@lse.ac.uk

Donatas Surgailis
Institute of Mathematics and Informatics
Akademijos 4, 2600 Vilnius, Lithuania
sdonatas@ktl.mii.lt

The interest in models of heteroskedastic time series with long memory exists in econometrics and finance, where empirical facts about asset returns motivated the necessity to study stationary processes which exhibit long memory in conditional variance. A number of such models were proposed in the ARCH literature; however, long memory properties of some these models have not been so far theoretically established, and even the existence of a stationary solution remains controversial (Mikosch and Stărică (2000)). The classical GARCH(p,q) and ARCH (∞) models are known to be short memory (Giraitis, Kokoszka and Leipus (2000)).

Robinson (1991) introduced the *Linear ARCH (LARCH)* model, in which the conditional variance σ_t^2 of observable sequence, r_t, is the square of a linear combination of $r_s, s<t$ with square summable weights a_j, $j \geq 1$. The LARCH model specializes, when σ_t depends only on r_{t-1}, to the asymmetric ARCH model of Engle (1990), and, when σ_t depends on finitely many $r_s, s<t$, to the Quadratic ARCH model of Sentana (1995). The LARCH model was recently studied in Giraitis, Robinson and Surgailis (2000), Giraitis et al. (2001). As shown in Giraitis, Robinson and Surgailis (2000), integer powers r_t^ℓ, $\ell \geq 2$, can have long memory autocorrelations. The cross-covariance function between future volatility and levels, $h_t = Cov(\sigma_t^2, r_0)$, was studied in Giraitis et al. (2001), and a linear inhomegeneous equation for h_t derived. It was shown that the LARCH model (unlike GARCH (p,q) models) incorporates the leverage property such that $h_t < 0$ for $0 < t \leq k$, where the value of k may be infinite. As shown in Giraitis et al. (2001), the h_t decay in the manner of the moving average weights a_j which may be chosen as in long memory ARFIMA process.

As far as ARCH models have zero conditional mean, attempts have been made to generalize them to include non-zero conditional mean (Ling and Li (1997), Teyssièrre (2000)). Giraitis and Surgailis (2001) introduce the following generalization of the LARCH model:

(1) $$r_t = \xi_t \sigma_t + m_t,$$

where ξ_t is an i.i.d. noise, and σ_t, m_t are moving averages in $r_s, s<t$, with square summable weights a_j, b_j, respectively. In (1), $m_t = E[r_t \mid r_s, s<t]$ is the conditional

mean and $\sigma_t^2 = \text{var}[r_t \mid r_s, s < t]$ is the conditional variance. Stationary solution of (1) is obtained as an orthogonal Volterra series. In the case $\sigma_t \equiv 1$, (1) is the classical AR(∞) model, while $m_t \equiv 0$ gives the LARCH model. Another particular case of (1) is the ARCH(∞) model (Giraitis, Kokoszka and Leipus (2000)). In the general case, (1) may exhibit long memory both in conditional mean and in conditional variance, with arbitrary memory parameters $0 < d_1, d_2 < 1/2$.

References

Engle, R.F. (1990). Stock volatility and the crash of '87. Discussion. *The Review of Financial Studies*, **3**, 103-106.

Giraitis, L., Kokoszka, P. and Leipus, R. (2000). Stationary ARCH models: dependence structure and central limit theorem. *Econometric Theory*, **16**, 3-22.

Giraitis, L., Robinson, P. M. and Surgailis, D. (2000). A model for long memory conditional heteroskedasticity. *Annals of Applied Probability*, **10**, 1002-1024.

Giraitis, L., Leipus R., Robinson, P. M. and Surgailis, D. (2001). LARCH, leverage and long memory. *Preprint.*

Giraitis, L and Surgailis, D. (2001). A class of bilinear models with long memory. Preprint.

Ling, S. and Li, W.K. (1997). On fractionally integrated autoregressive moving average time series models with conditional heteroskedasticity. *JASA*, **92,** 1184-1193.

Mikosch, T. and Stărică, C. (1999). Change of structure in financial time series, long range dependence and the GARCH model. Preprint.

Robinson, P. M. (1991). Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *Journal of Econometrics*, **47**, 67-84.

Sentana, E. (1995). Quadratic ARCH models. *Review of Economic Studies*, **62**, 639- 661.

Teyssièrre, G. (2000). Double long-memory financial time series. Preprint.

Semiparametric Fractional Autoregressive Model

Jan Beran, Yuanhua Feng
Department of Mathematics and Statistics
University of Konstanz
78457 Konstanz, Germany
Jan.Beran@uni-konstanz.de, Yuanhua.Feng@uni-konstanz.de

1. The *SEMIFAR* Models

SEMIFAR (semiparametric fractional autoregressive) models, introduced by Beran (1999), provide a unified approach for simultaneous modelling of deterministic trends, stochastic trends as well as short-memory, long-memory and antipersistent components in an observed time series. This paper summrizes recent developments on *SEMIFAR* models. The focus is on a data-driven algorithm for estimating such a model, which combines the nonparametric estimation of the deterministic trend and the maximum likelihood estimation of the parameters characterizing the model.

A *SEMIFAR* model is a Gaussian process $\tilde{Y}_i$ with an existing smallest integer $m \in \{0,1\}$ such that

(1) $$\phi(B)(1-B)^{\delta}\left\{(1-B)^{m}\tilde{Y}_i - g(t_i)\right\} = \varepsilon_i,$$

where $t_i = i/n$, ε_i $(i = \ldots -1,0,1,\ldots)$ iid with $E(\varepsilon_i) = 0$, $\mathrm{var}(\varepsilon_i) = \sigma_\varepsilon^2 < \infty$, g is a smooth function, $B\tilde{Y}_i = \tilde{Y}_{i-1}$ and $\phi(B) = 1 - \sum_{j=1}^{r}\phi_j B^j$, $\phi(z) \neq 0$ for $|z| \leq 1$, $-1/2 < \delta < 1/2$. Here the fractional difference $(1-B)^{\delta}$ introduced by Granger and Joyeux (1980) and Hosking (1981) is defined by

(2) $$(1-B)^{\delta} = \sum_{k=0}^{\infty} b_k(\delta)B^k$$

with

(3) $$b_k = (-1)^k \frac{\Gamma(\delta+1)}{\Gamma(k+1)\Gamma(\delta-k+1)}.$$

Let $Y_i = (1-B)^m \tilde{Y}_i$ (with $Y_1 = 0$ for $m = 1$). Then we have

(4) $$Y_i = g(t_i) + X_i, \quad i = 1, 2, \ldots, n,$$

where $X_i = \phi^{-1}(B)(1-B)^{-\delta}\varepsilon_i$. Equation (4) is a nonparametric regression model with

- long memory (for $\delta > 0$),
- short memory (for $\delta = 0$) and
- antipersistence (for $\delta < 0$).

2. Estimation of the *SEMIFAR* Models

The trend g has to be estimated nonparametrically. A kernel estimate of g (Hall and Hart, 1990 and Beran, 1999) is given by

$$\hat{g}_k(t) = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{t-t_i}{b}\right) Y_i \,, \tag{5}$$

where K is a k-th order kernel function (with k even) and b is the bandwidth.

Local polynomial estimates $\hat{g}_p(t)$ of g are considered in Beran and Feng (1999a). They are defined as the solution of the local least squares problem

$$Q = \sum_{i=1}^{n}\left\{ Y_i - \sum_{j=0}^{p} \beta_j (t_i - t)^j \right\}^2 K\left(\frac{t-t_i}{b}\right) \Rightarrow \min \,, \tag{6}$$

where K is a second order kernel and p (odd) is the order of the local polynomial. It can be shown that $\hat{g}_p(t)$ is asymptotically equivalent to a certain k-th order kernel estimate with boundary correction, where $k = p+1$.

Let $\theta^0 = (\sigma^2_{\varepsilon,0}, d^0, \phi_1^0, ..., \phi_r^0)' = (\sigma^2_{\varepsilon,0}, \eta^0)'$ be the true unknown parameter vector in (1), where $d^0 = m^0 + \delta^0$, $-1/2 < \delta^0 < 1/2$ and $m^0 \in \{0,1\}$. Then $\tilde{Y}_i$ in (1) admits an infinitive autoregressive representation

$$\sum_{j=0}^{\infty} a_j(\eta^0)[c_j(\eta^0)\tilde{Y}_{i-j} - g(t_{i-j})] \,, \tag{7}$$

where the coefficients a_j and $a_j c_j$ are obtained by matching the powers in B. For a chosen value of $\theta = (\sigma^2_\varepsilon, m+\delta, \phi_1, ..., \phi_r)' = (\sigma^2_\varepsilon, \eta)'$ denote by

$$e_i(\eta) = \sum_{j=0}^{i-m-2} a_j(\eta)[c_j(\eta)\tilde{Y}_{i-j} - \hat{g}(t_{i-j}; m)] \,, \tag{8}$$

the (approximate) residuals. Following the idea in Beran (1995), Beran (1999) considered an approximate maximum likelihood estimate $\hat{\theta} = (\hat{\sigma}^2_\varepsilon, \hat{\eta})$ of θ^0, where $\hat{\eta}$ is obtained by minimizing

$$S_n(\eta) = \frac{1}{n}\sum_{i=m+2}^{n} e_i^2(\eta) \tag{9}$$

with respect to η and

$$\hat{\sigma}^2_\varepsilon = \frac{1}{n}\sum_{i=m+2}^{n} e_i^2(\eta) \,. \tag{10}$$

Under regularity conditions the following asymptotic results can be obtained:

1. The asymptotically optimal bandwidth for estimating g is: $b_{opt} = C_{opt} n^{(2\delta-1)/(2k+1-2\delta)}$, where C_{opt} is an unknown constant.
2. The rate of convergence of $\hat{g}_p(t)$ (or $\hat{g}_k(t)$) with b_{opt} is of order $O(n^{(2\delta-1)k/(2k+1-2\delta)})$.

3. $\hat{\theta}$ is root n consistent under some condition on the rate of convergence of $\hat{g}$.

   In particular, the condition on $\hat{g}$ in 3 is fulfilled for all $-1/2 < \delta < 1/2$, if $k = 4$ is used.

3. A Data-Driven Algorithm

The iterative plug-in idea in Gasser et al. (1991) is adapted by Herrmann et al. (1992) and Ray and Tsay (1997) to select the bandwidth in nonparametric regression with short or long memory, respectively. Another variant of this idea is proposed by Beran and Feng (2000). Different algorithms for estimating *SEMIFAR* models are developed by combining the data-driven estimation of g and the maximum likelihood estimation of θ^0 (see Beran and Feng, 2000). The algorithm consists of three steps. In step 1, a bandwidth for estimating m^0 is obtained. The BIC is used for determining the autoregressive order and for deciding between $m=0$ and $m=1$ (see e.g. Beran et al., 1998). In step 2, m^0 is estimated. In step 3, iterations are carried out for obtaining an optimal bandwidth for the trend function, alternating between estimation of θ^0 and g.

4. Simulation and Applications

To study the practical performance of *SEMIFAR* models and the proposed algorithms, a large simulation study was carried out, including three regression functions and the parameter combinations with $m^0 \in \{0,1\}$, $\delta^0 \in \{-0.4, -0.2, 0, 0.2, 0.4\}$, $\phi_1^0 \in \{-0.7, -0.3, 0, 0.3, 0.7\}$. Here we have $p_0 = 0$ for $\phi_1^0 = 0$ and $p_0 = 1$ otherwise. 200 replications were simulated for each parameter combination with two sample sizes $n = 500$ and $n = 1000$. The simulation results show that the proposed data-driven algorithm works well in all cases, although the performance differs from case to case. For a detailed report of the simulation study see Beran and Feng (2000, 2001a).

SEMIFAR models were applied to analyze data from different areas, such as financial markets (see e.g. Beran and Ocker, 1999, 2001). Forecasting with *SEMIFAR* models is discussed in Beran and Ocker (1999). Beran and Ocker (2001) apply *SEMIFAR* models to study volatility in financial data. Climatological data are considered in Beran (1999). Modelling of exchange rates and commodity price series is discussed in Beran et al. (2000).

5. Concluding Remarks

The *SEMIFAR* models may be extended in different ways. One simple extension is to introduce an MA part on the right hand side of (1), thus defining a *SEMIFARMA* model. Another useful extension is obtained by using non-iid innovations ε_i. Beran and Feng (1999b) propose the *SEMIFAR-GARCH* model, where ε_i is assumed to be a GARCH (generalized autoregressive conditional heteroscedastic) process and derive asymptotic results, such as the asymptotic normality of $\hat{g}_p(t)$ and $\hat{g}_k(t)$.

Another important field of current research is the design of optimal algorithms. The data-driven algorithm discussed in section 3 is based on the iterative plug-in

bandwidth selection rule. Such an algorithm requires the estimation of the k -th order derivative of g based on the assumption that $g^{(k)}$ exists. Another well known bandwidth selection rule is the double-smoothing procedure based on bootstrap idea (see e.g. Müller, 1985, Härdle et al., 1992 and Heiler and Feng, 1998). Beran et al. (2000) show that the double-smoothing method is superior to plug-in algorithms, both, theoretically and in practice, at least for nonparametric regression with iid errors. An adaptation of the proposal in Beran et al. (2000) to nonparametric regression with long-range dependence is currently being developed.

References

Beran, J. (1995). Maximum likelihood of estimation of the differencing parameter for invertible short- and long-memory autoregressive integrated moving average models, *J. Roy. Statist. Soc. Ser.* B, **57**, 659-672.

Beran, J. (1999). SEMIFAR models -- A semiparametric framework for modelling trends, long range dependence and nonstationarity, Discussion paper No. 99/16, Center of Finance and Econometrics (CoFE), University of Konstanz.

Beran, J., Bhansali, R.J. and Ocker, D. (1998). On unified model selection for stationary and nonstationary short- and long-memory autoregressive processes, *Biometrika*, **85**, 921-934.

Beran, J. and Feng, Y. (1999a). Locally polynomial fitting with long-range dependent errors, Discussion paper No. 99/07, CoFE, University of Konstanz.

Beran, J. and Feng, Y. (1999b). Local polynomial estimation with a FARIMA-GARCH error process, Discussion paper No. 99/08, CoFE, University of Konstanz..

Beran, J. and Feng, Y. (2000). Data-driven estimation of the semiparametric fractional autoregressive models. Discussion Paper, CoFE, 00/16, University of Konstanz.

Beran, J. and Feng, Y. (2001). Supplement to 'Data-driven estimation of semiparametric fractional autoregressive models' -- Detailed simulation results, Preprint, University of Konstanz.

Beran, J., Feng, Y., Franke, G., Hess, D. and Ocker, D. (2000). Modelling trends, stationarity and difference stationarity in finance and economics, Discussion paper, CoFE 99/18, University of Konstanz.

Beran, J., Feng, Y. and Heiler, S. (2000). Modifying the double smoothing bandwidth selection in nonparametric regression, Discussion Paper, CoFE, 00/37, University of Konstanz.

Beran, J. and Ocker, D. (1999). SEMIFAR forecasts, with applications to foreign exchange rates, *J. Statistical Planning and Inference*, **80**, 137-153.

Beran, J. and Ocker, D. (2001). Volatility of stock market indices -- An analysis based on SEMIFAR models, *J. Busin. Econ. Statist.* **19**, 103-116.

Gasser, T., Kneip, A. and Köhler, W. (1991). A flexible and fast method for automatic smoothing, *J. Amer. Statist. Assoc.*, **86**, 643-652.

Granger, C.W.J. and Joyeux, R. (1980). An introduction to long-range time series models and fractional differencing, *J. Time Ser. Anal.*, **1**, 15-30.

Härdle, W., Hall, P. and Marron, J.S. (1992). Regression smoothing parameters that are not far from their optimum, *J. Amer. Statist. Assoc.*, **87**, 227-233.

Hall, P. and Hart, J.D. (1990). Nonparametric regression with long-range dependence, *Stochastic Process. Appl.*, **36**, 339-351.

Heiler, S. and Feng, Y. (1998). A root n bandwidth selector for nonparametric regression, *J. Nonparametric Statist.*, **9**, 1-21.

Herrmann, E., Gasser, T. and Kneip, A. (1992). Choice of bandwidth for kernel regression when residuals are correlated, *Biometrika*, **79**, 783-795.

Hosking, J.R.M. (1981). Fractional differencing, *Biometrika,* **68**, 165-176.

Müller, H.G. (1985). Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators, *Statist. Decisions*, Supp. Issue **2**, 193-206.

Ray, B.K. and Tsay, R.S. (1997). Bandwidth selection for kernel regression with long-range dependence, *Biometrika*, **84**, 791-802.

REVISTA DE ESTATÍSTICA
129ª PÁGINA
2º QUADRIMESTRE DE 2001

STATISTICS IN THE ENVIRONMENTAL SCIENCES

Organizer: K. Feridun Turkman

Invited Speakers: Jonathan Tawn
Amilcar Soares
James V. Zidek

REVISTA DE ESTATÍSTICA
131ª PÁGINA
2º QUADRIMESTRE DE 2001

Statistics in the Environmental Sciences

K. F. Turkman
University of Lisbon
Department of Statistics and Operations Research
DEIO, Faculdade de Ciências
Bloco C2, Campo Grande
1749-016 Lisboa, Portugal
kamil.turkman@fc.ul.pt

Environmental statistics has developed rapidly over the last decade becoming an important, high profile specific theme. Many organizations, from universities to specialised institutions are signalling their commitment to this theme by including it in their research and education portfolios, as sections in their societies and as sessions in their conferences. Environmental statistics differs from other statistical topics not only in the areas of application in such diverse fields as conservation, pollution control, monitoring of ecosystems, management of resources but also on the use of specific statistical methodologies and models, demanding new approaches and methods. Environmental statistics is thus becoming one of the most active fields and particularly modelling spatial-temporal dependence structures has been one of the leading areas of research.

Almost all environmental processes show variability over space and time involving complicated spatio-temporal structures and interactions. The modelling of space-time variability is inevitably complicated by the fact that we need to capture the space and time dependence structures as well as the structure that creates space-time interactions. This quest is further complicated by the fact that most environmental series are non-linear and nonstationary, thus suggesting that we should model the dependence structures by methods other than the classical use of second order moments for linear models. Related problem is the insufficiency of the Gaussian processes for most of the environmental series. These difficulties together with the large data sets typical of many environmental problems, often require the practitioners to apply many unrealistic assumptions in their models. Typically, a practitioner will need to assume that there are no space-time interactions, in the sense that the spatio-temporal dependence structure separates through additive or multiplicative models into two parts and that a linear or better yet a gaussian structure govern the spatial and temporal variation. Even then the problems are complicated due to the curse of dimension; one needs to work with very high dimensional multivariate normal distributions whose covariance structure in principle needs to be estimated. Bayesian hierarchical models (*Wikle, Berliner and Cressie,* 1998 and *Zidek, White, Sun and Burnett*, 1998) seem to be particularly suited for the modelling of spatio-temporal processes. Such models provide simple strategies for incorporating complicated space-time interactions at different stages of the hierarchy, thus making it relatively feasible to implement in high dimensions. However, there is still much work to be done on Bayesian hierarchical models, as the problem of model validation has not been addressed sufficiently, and models other than the ones based on Gaussian structures are very difficult to implement.

One strategy to model spatio-temporal variation is to model this complicated structure through the mean function of the process at various stages, assuming

conditional independence. See for example, *Wikle, Berliner and Cressie* (1998). Other possible alternative is to model this variation through the estimation of the massive covariance structure as it is advocated by *Brown, Le and Zidek* (1994). In any case, all these methods are based on the assumption that the space-time dependence structures do not have interactions, in other words it is assumed that the underlying Gaussian process has a separable spatio-temporal covariance structure. Recently *Cressie and Huang* (1999), *De Cesare, Myers and D. Posa,* (2000) and *De Iaco, Myers and Posa* (2001) gave new methodologies for developing classes of nonseparable spatio-temporal stationary covariance functions in closed forms. Once statistical methods of identifying and estimating these covariance structures become available, we should be able to use more realistic models for the environmental series, nevertheless still within the restriction of normality and linearity.

Related and equally important subject is the modelling of extremes of spatial processes. Standard measures of dependence such as variogram and covariogram are second order properties of the spatial processes and for processes other than the Gaussian, they do not give information regarding the dependence on the tail of joint distributions. As such, they are not very useful tools in extrapolating extreme events in space and indeed they can be quite misleading. Therefore, there is an urgent need to define new concepts of spatial dependence of extremes. The pioneering work on this area can be found in *Ancona-Navarrete and Tawn* (2001).

This session on Statistics in the Environmental Sciences will focus on modelling spatio-temporal dependence structures. Specifically, the invited papers will focus on the geostatistical perspective of space-time models, space-time interaction issues in spatial prediction of pollution levels and modelling extreme values of spatial environmental processes.

References

Brown, P.J., Le, N.D. and Zidek, J.V. (1994) Multivariate spatial interpolation and exposure to air pollutants. *Canadian Journal of Statistics*, **22**, 489-509.

Cressie, N. A. C., Huang, H-C. (1999) Classes of nonseparable spatio-temporal stationary covariance functions. *Journal of American Statistical Association*, **94**, 1330-1340.

De Cesare, D.E. Myers and D. Posa (2000), Product-sum covariance for space-time modelling: an environmental application. *Environmetrics* 12, 11-23

De Iaco, D.E. Myers and D. Posa (2001), Nonseparable space-time covariance models: some parametric families. To appear in *Mathematical Geology*

Wikle, C. K., Berliner, L. M., Cressie, N. A. C. (1998) Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, **5**, 117-154.

Zidek, J.V., R. White, N.D. Le. W. Sun and R.T. Burnett (1998) Imputing unmeasured explanatory variables in environmental epidemiology with application to health impact analysis of air pollution. *Environmental and Ecological Statistics*, **5**, 99-115.

REVISTA DE ESTATÍSTICA
133ª PÁGINA
2º QUADRIMESTRE DE 2001

Modelling Extreme Values of Spatial Environmental Processes

Jonathan Tawn
Department of Mathematics and Statistics
Lancaster University
Lancaster
LA1 4YF
UK
j.tawn@lancaster.ac.uk

1. Introduction

Obtaining an understanding of the spatial behaviour of the extreme values of various environmental processes is important for the protection and insurance of property and life. There are two components of any description of the extreme values of a spatial process: a model for how the marginal characteristics change over space and a model for the spatial dependence structure. Asymptotically motivated models for the extremes of univariate series are widely used, and can easily be adapted for the marginal component of the spatial model. The focus here is on discussing ways of measuring, and modelling, spatial dependence among extreme values, for which much less is known. The dependence structure and the spatial nature of the process are often ignored in analyses of environmental extremes. The aim here is to show that dependence modelling is a fundamental component of some analyses. The talk will be illustrated with applications to rainfalls and sea-levels.

2. Measuring Pairwise Extremal Dependence

Standard measures of spatial dependence, such as the variogram, describe the dependence in the body of the process and depend on the marginal properties of the process. When a process may have different characteristics for extreme and non-extreme events or when it is heavy tailed, the variogram can give a false impression of the spatial dependence in extreme events. Here a measure of pairwise extremal dependence for spatial processes, that is marginally invariant, will be introduced.

There are three forms of extremal dependence: asymptotic dependence, asymptotic independence and exact independence. If the process is asymptotically dependent then large values at different sites tend to occur in the same event however large the values are, whereas for an asymptotically independent process large values can occur in the same event, but increasingly they tend to occur in different events for the larger values. For any pair of locations, the measure enables us to distinguish between the different classes of extremal dependence and provides us with an estimate of the associated degree of dependence in the selected class. This is valuable diagnostic information for modelling the extreme values of a spatial process. In the few spatial extreme value studies previously undertaken the process has been assumed to be asymptotically dependent (e.g. Coles and Tawn, 1996) which, if false, leads to bias in estimates of the spatial risk. The measure of dependence that will be introduced is motivated by results of Ledford and Tawn (1996, 1997, 2001) for

bivariate dependence. The measure has many properties similar to the variogram, for further details see Ancona-Navarrete and Tawn (2001).

3. Higher Order Dependence

Pairwise dependence does not fully characterise the dependence structure of the extreme values of a process; so measures of higher order dependence need to be studied. For the class of asymptotically dependent processes I will describe work of Schlather and Tawn (2001) on the higher order structure. These results are required for the construction of self-consistent dependence models and to enable the construction of self-consistent estimators for higher order dependence measures.

References

Ancona-Navarrete, M.A. and Tawn, J.A. (2001). Diagnostics for extremal dependence in spatial processes. Submitted.

Coles, S.G. and Tawn, J.A. (1996). Modelling extremes of the areal rainfall process, *J. R. Statist. Soc.*, B, **58**, 329-347.

Ledford, A.W. and Tawn, J.A. (1996). Statistics for near independence in multivariate extreme values, *Biometrika*, **83**, 169-187.

Ledford, A.W. and Tawn, J.A. (1997). Modelling dependence within joint tail regions, J. R. Statist. Soc., B, 59, 475-499.

Ledford, A.W. and Tawn, J.A. (2001). Diagnostics for dependence within time-series extremes. Under revision.

Schlather, M. and Tawn, J.A. (2001). A dependence measure for multivariate and spatial extreme values: properties and inference. In preparation.

REVISTA DE ESTATÍSTICA
135ª PÁGINA
2º QUADRIMESTRE DE 2001

Space-Time Models in Environmental Sciences: A Geostatistical Perspective

Amilcar Soares
CMRP-Instituto Superior Técnico
Av. Rovisco Pais
1049-001 Lisboa
Portugal
ncmrp@alfa.ist.utl.pt

1.Introduction

The modeling of space-time processes has been recognized as a critical and fundamental issue in the environmental field: air, soil and groundwater pollution, climate changes, forest and other biopopulations evolution are just a few examples.

This paper aims at presenting a geostatistical perspective of spatio-temporal modeling – a probabilistic framework for data analysis, inference and prediction based on the joint space-time correlation between observations. Based on two decades of applications, the geostatistical space-time modeling can be viewed as an extension of spatial analysis to include the additional time dimension.

Among the different families of stochastic models of natural resources in earth sciences, this paper will focus on those that can be defined as spatial models, usually the scope of geostatistics or spatial statistics, and that incorporate a temporal component. With these models the objective is to predict, at a given spatial location and in a fixed period of time, the distribution of the attributes of a natural resource or a property of a spatial phenomena, or to access the uncertainty about the knowledge of that attribute or property. The characterization of a plume of contaminants in soil or water, which is sampled or monitored in several periods of time in some spatial spots, the analysis of the air quality of a region which is systematically monitored over time, the planning and control of a ecological resource observed in a given sample pattern at different periods of time, are just a few examples of problems that can be approached by such models.

There are as many approaches to space-time modelling as there are specificities of each case study regarding the amount of available information and the final objectives of the study. The objectives of the models treated in this study can be summarised according to the purpose of the use of the time data:

- Data collected in the past, at different periods of time, is used in a joint space-time framework to infer the spatial distribution of a given attribute at the present time or in a period in the very near future.
- Historical data is used to build a spatial and time trend. These trends are interpreted as spatio-temporal random fields and are inferred in space for fixed periods of time.
- Spatio-temporal uncertainty assessment is the aim of the third type of models presented in this paper. Deterministic models that mimic the complexity of some dynamic phenomena can be used, together with spatial stochastic simulation models, for uncertainty assessment and to visualized extreme scenarios of the attribute.

Examples of air pollution characterization and ecological applications are presented to detail and illustrate some of the methods that are presented.

2. Joint Space - Time Models

Consider an attribute Z(x,t) defined at a spatial location x, $x \in D$, and at instant of time $t \in T$, two different conceptual models can be adopted, regarding the decisions of stationarity of the random function RF Z(x,t):

1. Z(x,t) can be considered a second order stationary RF, which means considering a constant mean in spatial and time domain: $E\{Z(x_1,t_1)\} = E\{Z(x_2,t_2)\} = m$ and considering the space and time covariance independent of the space-time location (x,t): $C(Z(x_1,t_1).\ Z(x_2,t_2)\} = C(h,t)$ where $h=x_1-x_2$ and $t=t_1-t_2$.
2. When a time, space or space-time trend is evident in the physical phenomenon, a non-stationarity of Z(x,t) can be assumed. The RF Z(x,t) is decomposed into a mean component (the trend) M(x,t) and a residual component of zero mean R(x,t): Z(x,t)=M(x,t)+R(x,t). The mean M(x,t) and R(x,t) can also be decomposed into space and time components.

Stationary Joint Space-Time Models

Assuming the second order stationarity of Z(x,t), the problem consists in predicting in any spatial location $x0 \in D$ and in a time period $t \in T$ the unsampled value $Z(x_0,t_0)$. This can be done by an ordinary kriging predictor (Isaaks and Srivastava, 1989, Cressie, 1993,Goovaerts, 1997):

$$Z(x_0,t_0)^* = \sum_{\alpha} \lambda(x_\alpha) Z(x_\alpha,t_\alpha)$$

where $Z(x\alpha,t\alpha)$ are the neighbourhood observations at $(x\alpha,t\alpha)$.

Space-time covariance C(h,t) is required to calculate the above predictor. Some possible approaches can be used to estimate C(h,t) from the experimental observations Z(x,t) (Kyriakidis et al, 1999):

The space-time covariance can be decomposed into a space and a time component (Rouhani and Myers 1990, Cressie and Huang 1999, De Cesare et al, 1997), or assuming the same parametric form in space and time (Armstrong et al, 1993) or even defined in a metric across the space and time domains (Buxton and Plate, 1994).

Example 1: This is a case study of air quality control in an industrial area (Barreiro-Seixal) located south of Lisbon. Values of SO_2 are systematically measured on a daily basis in a series of monitoring stations. This is a typical situation of high density measurements in the time domain in just a few monitoring stations. Maps of predicted values of SO2 with space-time model will be presented (Fig 1.a).

Spatial Models with Time and Space Trends

The second type of model treated in this paper can be used to deal with non-stationary situations. The attribute value z is decomposed into a trend and a residual: Z(x,t)=M(x,t)+R(x,t). Two similar models that include historical data in a time and spatial trend are presented below. Host, More and Switzer (1995) proposed the decomposition of Z(x,t) into the following terms : $Z(x,t)=M(x)+R_2(t)+S(x).S(t).R(x,t)$

M(x) is a purely spatial component; $R_2(t)$ is a temporal modulation of the field at discrete times t. It is a zero mean residual corresponding to a correction of M(x) for time t; S(x).S(t) can be viewed as the standard deviation of Z(x,t) which is decomposed into a spatial and a time component S(x) and S(t) respectively; R(x,t) is the spatio-temporal residual with zero mean and unitary variance. All these components are interpreted as spatio-temporal random fields.

Example 2: This model was applied to the prediction of airborne solid particles in Setubal peninsula (south of Lisbon). The particulate pollutant basically comes from a

cement plant and is captured at some monitoring stations in regular time intervals (Fig. 1.b).

Another model is proposed by Santos et al (2000) with a similar objective of incorporating historical data in a space and time trend. While in the previous model the data is collected at the same sampling stations, here the data is collected at different spatial locations in each period. As the data is not collected in all time periods at the same sampling plots, the idea introduced in this algorithm is to make a weighted average of the Nt predicted maps, in which one predicted point value is weighted by a proximity measure to the neighbourhood experimental data of each year. In each time period ti the values $Z(x_0,t_i)$ are predicted for the entire region. The spatial trend $M(x,t)$ is then obtained by a linear combination of predicted values $[Z(x_0,t_i)]^*$ at different periods of time t_i, i=1, N_t., where the weights are given by a the kriging variance.

Example 3: This model was applied to predict the abundance of a migratory bird – the wood pigeon - in order to control and plan this ecological resource in time and space. Prediction of wood pigeon abundance should take into account the migration pattern of this species and local abundance of individuals measured in each sampling resort. Predicted pattern of wood pigeon abundance for the period 1992/1996 and the predicted map of 1996 pigeon abundance will be the final output of this study (Fig. 1 c).

3. Space-time Uncertainty Assessment

The characterization of spatial uncertainty using only spatial models has been addressed, in earth sciences, through the use of stochastic simulation algorithms (Deutch, Journel, 1998, Goovaerts, 1997). In dynamic processes, the two components – space and time – usually have quite different levels of uncertainty: the heterogeneity of the static component – normally related to the space – sometimes cannot be compared with the complexity of the dynamic part of the process; on the other hand the knowledge that one have about both components is usually quite different. This is possibly the main reason why simulation algorithms of spatial processes with a time component are still at an early stage.

However, according to how the dynamic and static components are combined in the same model, two different type of approaches are presented in this study:

i) the first type of approach uses simulation as a tool for spatial uncertainty assessment at a fixed period of time, taking into account the spatio-temporal model and available data in space and time. In these models the time component is used to assess the spatial uncertainty of the static component of the system. An example of simulation of particulate emissions at Arrabida peninsula is presented.

ii) Another type of approach has a completely different objective, which is to preview extreme scenarios regarding the dynamic of the physical phenomenon. Hence a deterministic model, which mimics the dynamics of the physical phenomenon, is added to a stochastic model in order to give the uncertainty related to different possible situations of the static and dynamic components of the reality. This approach is illustrated with the prediction of extreme scenarios of air pollution impacts by simulating a solid particulate contamination with a deterministic model - Gaussian dispersion plume Pereira et al. (1997).

References

Armstrong M., Chetboun G., Hubert P. (1993). Kriging the Rainfall in Lesotho. *Geostatistics Troia'92*. Soares Ed. Kluwer Academic Pub, 661-672.

Buxton B., Plate A., (1994). Joint Temporal-Spatial Modeling of Concentrations of Hazardous Pollutants in Urban Air. *Geostatistics for the Next Century*. Dimitrakopoulos Ed.. Kluwer Academic Pub, 75-87.

Cressie N., (1993), Statistics for Spatial Data. New York, Wiley.

Cressie N. and Huang H.C., 1999. Classes of Nonseparable Spatio-Temporal Stationary Covariance Functions. *Journal of A. Statistical Association*, **94**, 1330-1340.
De Cesare L., Myers D., Posa D. (1997). Spatio-Temporal Modeling of SO_2 in Milan District. *Geostatistics Wollongong'96* . Baafi Ed.. Kluwer Academic Pub, 1031-1042.
Goovaerts P., (1997) – Geostatistics for Natural Resources Characterization. Oxford University Press., 483 pp.
Host G., Omre H., Switzer P. (1995). Spatial Interpolation Errors for Monitoring Data. *Journal American Statistical Association*, Vol **90**, N-431, p. 853-861.
Isaaks, E., Srivastava, M., 1989, Applied Geostatistics. Oxford U. Press.
Journel A. (1986). Geostatistics: Models and tools for the earth sciences. *Mathematical Geology*, **18** (1), 119-140.
Kyriakidis P., Journel A., (1999). Geostatistical Space-Time Models: A Review. *Mathematical Geology*, **31** (6), 651-685.
Pereira M., Soares A., Branquinho C. (1997) Stochastic Simulation of Fugitive Dust Emissions. *Geostatistics Wollongong'96* . Baafi Ed.. Kluwer Academic Pub, 1055-1065.
Rouhani S., Myers D. (1990). Problems in Space-time Kriging of Geohydrological Data. *Mathematical Geology*, **22**(5), 611-623.
Santos E., Almeida J., Soares A. (2000). Geostatistical Characterization of the Migration Patterns and Pathways of the Wood Pigeon in Portugal. Proceedings of the 6th International Geostatistics Congress. Cape Town.

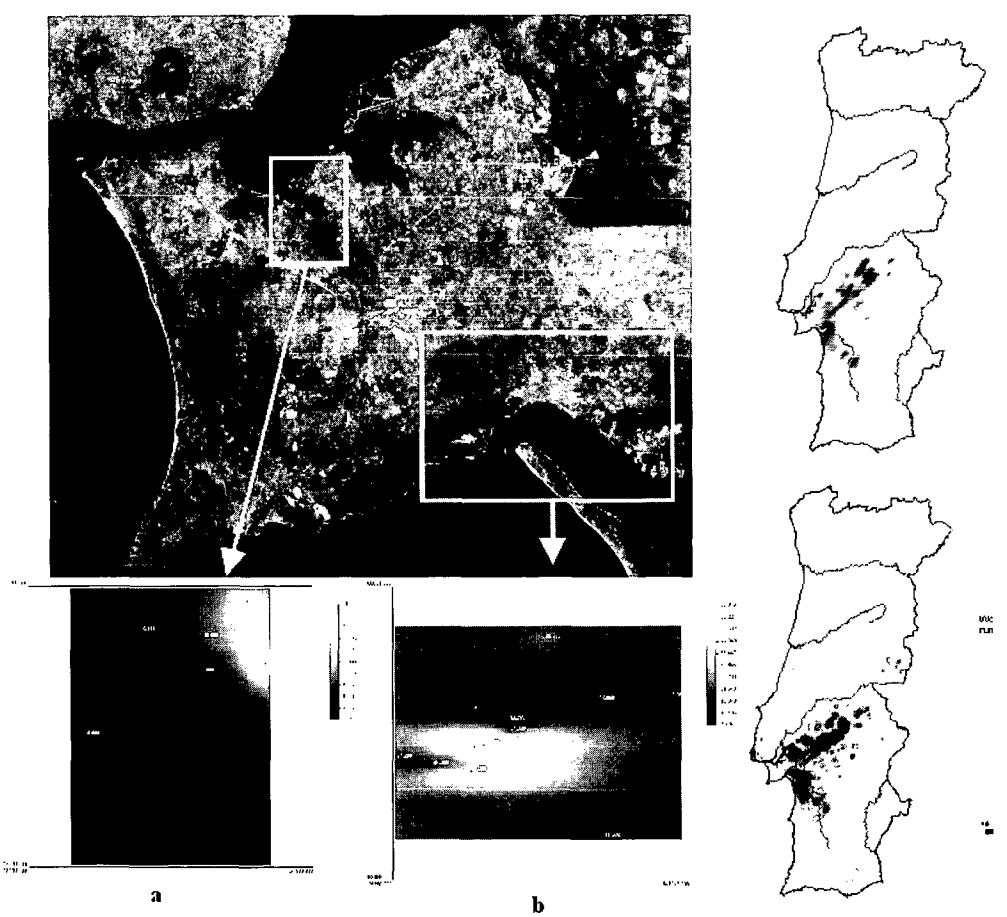

Fig 1- a)-predicted values of SO2 at Barreiro-Seixal; b) Prediction of airborn solide particles at Arrábida peninsula ; c) Predicted pattern of wood pigeon (1992/96) and pigen abundance (1996) in Portugal.

REVISTA DE ESTATÍSTICA
139ª PÁGINA
2º QUADRIMESTRE DE 2001

Space-Time Interaction Issues in Spatial Prediction of Pollution Fields

James V Zidek, Nhu D Le, Li Sun
Department of Statistics
University of British Columbia
6356 Agricultural Road – Room 333b
Vancouver, B.C. V6T 1Z2
Canada
jim@stat.ubc.ca

Our focus in this paper will be the ambient hourly P10 field over Vancouver, Canada. Interest in this pollutant (and its relative PM2.5) derives from its established association with acute morbidity and mortality, in particular cardiovascular mortality. Controversy about the role of measurement error in this association has led to the need for better estimates of human exposure and in turn for reliable methods for predicting local levels of pollution from ambient levels.

One of the well known computational models for predicting population exposure levels, pNEM, requires hourly levels as an input even though most epidemiological studies rely on daily or longer term concentration aggregates. Regulatory criteria may also require hourly levels even though they may eventually be expressed in terms of aggregates such as daily maxima, daily averages or the maximum of moving averages over the day.

The need for spatial prediction arises in diverse fields, in particular geostatistics wherein the method of kriging was introduced in the 1960s by Matheron (Cressie, 1991). At any given unmonitored location, kriging uses a best linear predictor based on concentration levels observed at existing monitoring stations. The coefficients of that predictor are inversely proportional to the distances of the unmonitored location from the ``gauged'' stations. However, optimality of the predictor assumes a known spatial covariance or equivalently variogram. The method has been greatly refined over the years and can now incorporate covariables as well as deal with multivariate responses through co-kriging (see Haas 1990).

A limitation of kriging methods for use in the spatial prediction of air pollution concentrations stems from their dependence on spatial stationarity or even *isotropy* in the spatial covariance structure for the response field being predicted. In our experience that covariance does not depend on just the vector difference between the locations of sites and certainly not on the Euclidean distance between them.. Other factors such as elevation or local climate can play an important role in determining concentration levels so that two stations in close proximity can have very different concentration series.

A potentially serious problem may arise from the failure in kriging to incorporate uncertainty about the spatial covariance. The resulting impact on prediction error intervals can be considerable and in the wrong direction (see Sun 1998). That is, decision makers my be imbued with unrealistically high levels of confidence due to unduly short intervals. Attempts have been made to compensate for this limitation (e.g. Handcock and Stein 1993, De Oliveria et al. 1997). However, these modifications still require isotropy for the response field.

Developed as a geostatistical tool where the response field is fixed, kriging does not seem well suited to the analysis of dynamically evolving space-time fields. In particular, much data from previous realizations of the response field are ignored when applied to the analysis of pollution fields. In turn, this means that an unrealistically large number of stations may be required to make application feasible. However, in practice there may just a few such stations in an urban area.

Limitations like this led the authors and their co-investigators in earlier work, to develop an alternative approach to spatial prediction. (Le and Zidek 1992, Brown, Le and Zidek 1994a). Our hierarchical Bayesian approach leaves the spatial covariance function completely unspecified. Temporal and spatial modeling can be done in a convenient and flexible way. Uncertainty about the covariance structure is incorporated through the second level prior; unrealistically small credible regions for the interpolants are thereby avoided. Moreover, the isotropy assumption is avoided through the use of appealing non-parametric approach of Sampson and Guttorp (1992).

Finally, the usual Bayesian updating process will correct model mis-specifications as new data become available.

The orginal theory of Le and Zidek (1992) has been substantially extended through a succession of improvements that take account of practical difficulties which can arise in application to pollution fields. The need to deal with a multiplicity of pollutants led to a multivariate version of the theory (Brown, Le and Zidek 1994a). Networks can be a union of earlier networks developed for a variety of purposes. As a result different gauged stations may systematically measure different suites of pollutants and this led to a further extension of the theory (Le, Sun and Zidek 1997). Stations may also start operations at different times so that the combined dataset has the structure of a "staircase" . To take full advantage of such data within the Bayesian framework requires a further extension. That extension was made for a single pollutant in the first instance (Le, Sun and Zidek 2001). It was subsequently extended to handle multiple pollutants (Kibria, Sun, Zidek and Le 2001). Both of these latter works take advantage of the added flexibility in prior modelling of spatial covariance structures afforded by the Generalized Inverted Wishart distribution developed by Brown, Le and Zidek (1994b). Unlike the classical Inverted Wishart with has a single degrees of freedom parameter to represent the level of uncertainty in the model, its more general cousin has more such parameters as well as a more flexible hypercovariance structure. In particular, different degrees of freedom can be assigned to the different steps in the data staircase.

Validation studies (Sun et al. 1998) indicate that the method performs very well. In particular, based on cross-validatory assessment, prediction intervals derived from the Bayesian posterior predictive distribution seem quite well calibrated so that 95% intervals do cover the true concentration about 95% of the time. As well, the method has been used successfully in several health impact studies of air pollution (Duddek et al. 1995, Zidek et al. 1998a), including one in BC.

However, in the work cited above, applications involve daily or even monthly levels of aggregration over time. The resulting response fields seem fairly uncomplicated. In particular, temporal and spatial correlations are "separable". That is, the space-time series can be "whitened" simply by fitting the same autoregressive model to all individual series so as to remove the temporal correlation. The theory above then gets applied to the residual series. After the residuals have been spatially predicted, the AR components can be re-installed to get back onto the original data scale.

The space-time correlation structure for short time (for example, hourly) temporal aggregate series proves to be inseparable and generally more complicated. Li et al (1998) analyze hourly ambient log PM10 concentrations collected in the Vancouver and find that the series at each monitoring site follow an AR(3) series pretty well after removing the same trend model T = mean + hour +day +week at each site. (Although some evidence of correlation at lag 40 hours is seen by these authors.) Initially, we expected to be able to whiten the residual series in the same manner as in the earlier studies, by fitting an AR(3) model and then applying the spatial predictor the resulting residuals.

Our expectation proved to be ill-founded. For hourly (and other short term temporal) aggregates of Vancouver's log PM10 series removing temporal structure in this way also removes the spatial structure on which spatial prediction must inevitably rely. Thus for inseparable processes like that for log PM10 a new approach was called for. That approach will be the subject of the presentation which will be based on a paper of Zidek, Sun, Le and Özkaynak (2001).

The approach we use "blocks" the temporal series by day so that responses are 24 tuples of hourly levels. It turns out for the series we have examined, including log PM10, these 24 tuples prove to have a multivariate AR(1) structure. By fitting the same multivariate AR(1) model to all the series, we successfully whiten them while losing little of the spatial correlation. The multivariate predictor can now be used to spatially predict the residual 24 tuples in the manner described above.

We will demonstrate the approach using Vancouver data from 10 monitoring stations using TEOM monitors over the period, 1994 to 1996. In doing so we will discuss both meteorological and structural trends. Such trends will be removed before going to the MAR (1) analysis.

We believe the approach in the paper will work with other such series as well, but further analysis is currently underway. The method can easily be extended to multiple pollutants by taking vectors of pollutants within hours. In a current investigation we are conidering 5 pollutants meaning that our approach will have to address 5*24 = 120 tuples of responses at the various sites.

References

Brown, PJ, Le, ND and Zidek, JV (1994a). Multivariate spatial interpolation and exposure to air pollutants. *Canadian J of Statist*, **22**, 489-509.

Brown, PJ, Le, ND and Zidek, JV (1994b). Inference for A Covariance Matrix, Aspect of Uncertainty: A Tribute to D.V. Lindley AFM Smith and PR Freeman (Eds). New York: Wiley.

Burnett RT, Dales RE, Raizenne MR, Krewski D, Summers PW, Roberts GR, Raad-Young M, Dann T, Brook J (1994). Effects of low ambient levels of ozone and sulfates on the frequency of respiratory admissions to Ontario hospitals *Environ Research*, **65**, 172-94.

Cressie N (1991). *Statistics for Spatial Data*. New York: Wiley.

De Oliveria V, Kedem B and Short DA (1997). Bayesian prediction of transformed Gaussian random fields. *J Am Statist Assoc*, **92**, 1422-1433.

Haas T. (1990). Lognormal and moving window methods of estimating acid deposition *J Amer Statist Assoc*, **85**, 950-963.

Handcock MS, Stein ML(1993). A Bayesian analysis of Kriging." *Technometrics*, **35**, 403-410.

Kibria, BMG, Sun, L, Zidek, JV and Le, ND (2001). A Bayesian approach to backcasting and spatially predicting unmeasured multivariate random space-time fields with application to PM2.5. *J Amer Statist Assoc*. Under revision.

Le, ND and Zidek, JV (1992). Interpolation with Uncertain Spatial Covariance: a Bayesian Alternative to Kriging. J Multivariate Analysis., 43, 351-374.

Le, ND, Sun, W and Zidek, JV (1997). Bayesian multivariate interpolation with data missing by design. *J. R. Statist. Soc.* B, **59**, 501-510.

Le ND, Sun L, Zidek JV (2001). Bayesian Spatial Interpolation and Backcasting Using Gaussian-Generalized Inverted Wishart Model." Can J of Statist. Under Revision.

Li, K, Le, ND, Sun, L and Zidek, JV (1998). Spatial-temporal Models for ambient hourly PM10 in Vancouver. *Environmetrics*, **10**, 321-338.

Sampson, P and Guttorp, P (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J Amer Statist Assoc*, **87**, 108-119.

Sun, W.(1998). Comparison of a co-kriging method with a Bayesian alternative. *Environmetrics*, **9**, 445-457.

Sun, W, Le, ND, Zidek, JV and Burnett, R (1998) Assessment of a Bayesian multivariate spatial interpolation approach for health impact studies. *Environmetrics*, **9**, 565-586.

Zidek, JV, White, R, Le, ND, Sun, W and Burnett, (1998). Imputing Unmeasured Explanatory Variables in Environmental Epidemiology With Application To Health Impact Analysis of Air Pollution. *Environmental and Ecological Statistics*, **5**, 99-115.

Zidek, JV, Sun, L, Le, ND and Özkaynak, H (2001). Contending with space-time interaction in the spatial prediction of pollution: Vancouver's hourly ambient PM10 field. Environmetrics. Under revision.

REVISTA DE ESTATÍSTICA
143ª PÁGINA
2º QUADRIMESTRE DE 2001

STATISTICS OF EXTREMES

Organizer: M. Ivette Gomes

Invited Speakers: Jan Beirlant
Andreia Hall
Anthony Ledford

REVISTA DE ESTATÍSTICA
145ª PÁGINA
2º QUADRIMESTRE DE 2001

Statistics of Extremes

M. Ivette Gomes*
Universidade de Lisboa (F.C.U.L.), D.E.I.O. and C.E.A.U.L.
Bloco C2, Piso 2, Campo Grande
1749-016 Lisboa Codex, Portugal
ivette.gomes@fc.ul.pt

For a long time, the *extremal limit theorem* attracted far less attention than the *central limit theorem*, and the work of Fisher and Tippett (1928), Fréchet (1936), von Mises (1936), Gnedenko (1943) and Mejzler (1956) on limit distributions for suitably normed extremes of independent random variables had much lesser impact in probability theory and statistics than their counterparts for sums.

But the world is not normal, and the extreme part of the sample has an outstanding relevance in many applied areas, ranging from hydrological, climatological and environmental problems to risk management in the financial industry, extreme claim sizes and structural reliability. Thus, the *Statistics of Extremes* soon caught the interest of statisticians as may be seen in the pioneering book by Gumbel (1958). From 1960 onwards *Order Statistics* and *Extreme Value Theory* (*EVT*) had an enormous development, and the books by David (1970), Galambos (1978), Leadbetter *et al.* (1983), Resnick (1987) and Reiss (1989) are just a few providing excellent overviews of the field. The more recent books by Beirlant *et al.* (1996), Embrechts *et al.* (1997) and Reiss and Thomas (1997) have a more applied scope, dealing with the statistical modelling and analysis of extremes. In fact, this latter one has even a broader scope, since the view of the authors, with which we entirely agree, is that "*the analysis of extreme values must be embedded in other various approaches of main stream statistics*".

Topics like parametric modelling of extremes based on "exploratory" diagnostic tools, inference methodology associated to *max-stable* distributions and to *generalized Pareto* distributions (which are stable on their own right in the *POT* scheme), random censoring, the *tail* and the *extremal index*, *rates of convergence* and *penultimate approximations*, *multivariate extremes* are a list of what extreme value can offer to *Insurance* — large claims in actuarial decisions, probable maximum loss, reserve questions and ruin theory —, to *Finance* — extreme returns in asset prices, extremes of time series with *ARCH* or *GARCH* structure —, to *Environment* — site-specific flood frequency analysis, sea levels, pollution data analysis, clustering of extremes and global warming, to mention a few applications of the field of *Statistical Extremes*. We would dare to say that nowadays the fashion of Extremes lies essentially in its application to the field of Finance. Indeed the *VaR* (*Value at Risk*) is a high quantile, and consequently, a parameter of rare events, being then adequate the use of *Statistical Extremes'* methodology for its adequate estimation and validation. Speaking of *EVT*, Embrechts *et al.* (1997) claim the following: "*though not providing a risk manager in a bank with the financial product he or she can use for monitoring financial risk on a global scale, we will provide that manager with the stochastic methodology needed for the construction of various components of such a global tool*".

* Research partially supported by FCT / POCTI / FEDER.

Till the early seventies, parametric methodologies were the most commonly used among the practitioners in the field of *Extremes*, and those methodologies were essentially based on the asymptotic results derived first for maximum values, later for the other extreme order statistics, and still later for the asymptotic behaviour of the excesses over a high level u. Indeed, the push on *Statistical Extremes* was due to several criticisms put forward to the classical Gumbel's method, also called the method of *Annual Maxima*, where maximum values of arbitrary large subsamples were recorded and assumed to be a random sample from an *Extreme Value* model:

1. The classical asymptotic results related to maximum and minimum values were associated to an original i.i.d. set-up. But the dependence inherent to real data led immediately to the question of whether it was possible to generalize those asymptotic results to dependent schemes, either stationary or non-stationary.
2. There seemed to be no reason for not to consider further top order statistics of the sample, whenever available. Surely they would provide additional relevant information on the tail.
3. And why should we consider arbitrary sub-samples, and not the excesses of the whole set of data over a high deterministic level u, or equivalently a sample of the high top order statistics of the whole set of data?

Item 1. led to the development of a nice asymptotic theory of extreme values in dependent structures, summarized in the nowadays classical book of Leadbetter, Lindgren and Rootzen (1983), but further developed in a great variety of papers, of which we shall have here an application. Item 2. led to multidimensional parametric models, studied first by Weissman (1978), Gomes (1978, 1981), Smith (1986), and largely applied in practice, mainly in the fields of Climatology and Hidrology. Item 3. led to the *Peaks Over Threshold* (*POT*) methodology, initiated by Todorovic and Zelenhasic (1970), and further studied by Davison and Smith (1989) among others, and was in a certain sense the root to the development of the whole lot of semi-parametric estimators, dating from Hill (1975) and Pickands (1975) work, where the high level instead of deterministic is random.

These semi-parametric methodologies have received an enormous attention in the last two decades. Under a semi-parametric approach there is always the need to choose the adequate number k of top order statistics to consider, or equivalently to choose the adequate threshold u above which we have relevant information on the tail. The choice of the threshold has been an open problem for a long time, but in the last decade asymptotic methodologies together with computer intensive methods, made possible to find solutions to this problem. And we have several interesting ways to approach the problem, among which I would like to mention

- the *regression diagnostics* technique of Beirlant *et al.* (in a series of papers from 1996 onwards), of which we will hear more in this session;
- the *bias estimation* technique of Drees and Kaufmann, 1998;
- the *bootstrap methodology* of Draisma *et al.* (1999) and Danielsson *et al.* (2001).

It is also worth mentioning the effort in reducing the bias of the most common semi-parametric estimators of parameters of rare events. Here *the Jackknife methodology* has played an interesting role, and reference should be made to Gomes *et al.* (1998), but we cannot forget also the revival of the maximum likelihood and least-squares estimation under a semi-parametric context, undertaken by Feuerverger and

Hall (1999), as well as Drees (1996) and Beirlant *et al.*'s (1999) techniques of bias reduction.

The new developments in the field of multivariate extremes and spatial extremes, with applications to Environmental Science and Structural Engineering, using mainly rich parametric models, with all sorts of covariates have increased the importance of *EVT*, and have a pioneer in Tiago de Oliveira. Intrincate dependent structures have been put forward, and interesting models have been developed. Here, the ability to test for independence in multivariate extremes is important for applied statistical modelling (see, for instance, Tawn (1988) and Ledford and Tawn (1996)), and we shall also hear more of it in this session on *Statistics of Extremes*.

References

Beirlant, J., Teugels, J.L. and Vynckier, P. (1996). *Practical Analysis of Extreme Values*. Leuven University Press, Leuven (Belgium).

Beirlant, J., Vynckier, P. and Teugels, J.L. (1996a). Excess function and estimation of the extreme-value index. *Bernoulli* **2**, 293-318.

Beirlant, J., Vynckier, P. and Teugels, J.L. (1996b). Tail index estimation, Pareto quantile plots, and regression diagnostics. *J. Amer. Statist. Assoc.* **91**, 1659-1667.

Beirlant, J., Dierckx, G, Goegebeur, Y. and Matthys, G. (1999). Tail index estimation and an exponential regression model. *Extremes* **2**, 177-200.

Danielsson, J., L. de Haan, L. Peng and de Vries, C. (2001). Using a bootstrap method to choose the sample fraction in the tail index estimation. *J. Multivariate Analysis* **76**, 226-248.

David, H.A. (1981). *Order Statistics*, 2nd ed., Wiley, New York.

Davison, A.C. and Smith, R.L. (1989). Models for exceedances over high thresholds. *J. Royal Statist. Soc.* B **52**, 393-442.

Draisma, G., L. de Haan, L. Peng and Pereira, T.T. (1999). A bootstrap-based method to achieve optimality in estimating the extreme value index. *Extremes* **2**, 367-404.

Drees, H. (1996). Refined Pickands estimators with bias correction. *Comm. Statist. Theory Meth.* **25**, 837-851.

Drees, H. and Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stoch. Proc. and Appl.* **75**, 149-172.

Embrechts, P., Kluppelberg, C. and Mikosh, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin.

Fisher, R.A. and Tippett, L.H.C. (1928). Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proc. Cambridge Phil. Soc.* **24**, 180-190.

Feuerverger, A. and Hall, P. (1999). Estimating a tail exponent by modelling departure from a Pareto distribution. *Ann. Statist.* **27**, 760-781.

Fréchet, M. (1927). Sur la loi de probabilité de lécart maximum. *Ann. Soc. Pol. Math.* **6**, 93-116.

Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics* (2nd edition). Krieger.

Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.* **44**, 423-453.

Gomes, M.I. (1981). An *i*-dimensional limiting distribution function of largest values and its relevance to the statistical theory of extremes. *Statistical Distributions in Scientific Work* **6**, 389-410, (C. Taillie et al., eds.) D.Reidel, Dordrecht.

Gomes, M.I., Martins, M.J. and Neves, M. (1998). Alternatives to a semi-parametric estimator of parameters of rare events — the Jackknife methodology. Notas e Comunicações CEAUL18/98. *Extremes*, to appear.

Gumbel, E.J. (1958). *Statistics of Extremes*. Columbia Univ. Press, New York.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3**, 1163-1174.

Leadbetter, M.R., Lindgren, G. and Rootzen, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York.

Ledford, A.W. and J.A. Tawn (1996). Statistics for near independence in multivariate extreme values. *Biometrika* **83**, 169-187.

Mejzler, D. (1956). On the problem of the limit distribution for the maximal term of a variational series. *L'vov Politechn. Inst. Naucn Zp. (Fiz.-Mat.)* **38**, 90-109. (in russian).

von Mises, R. (1936). La distribution la plus grande de *n* valeurs. *Rev. Math. Union* **1**, 141-160. Reprinted in *Selected Papers of Richard von Mises, Amer. Soc.* **2** (1964), 271-294.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119-131.

Reiss, R.-D. (1989). *Approximate Distributions of Order Statistics*. Springer, New York.

Reiss, R.-D. and Thomas, M. (1997). *Statistical Analysis of Extreme Values, with Applications to Insurance, Finance, Hydrology and Other Fields* (with Xtremes on CD-ROM). Birkhauser, Basel.

Resnick, S. (1987). *Extreme Values, Regular Variation and Point Processes*. Springer, New York.

Smith, R.L. (1986). Extreme value theory based on the *r* largest annual events. *J. Hydrology* **86**, 27-43.

Tawn, J.A. (1988). Bivariate extreme value theory: models and estimation. *Biometrika* **75**, 397-415.

Todorovic, P. and E. Zelenhasic (1970). A stochastic model for flood analysis. *Water Resour. Res* **6**, 1641-1648.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the *k* largest observations. *J. Amer. Statist. Assoc.* **73**, 812-815.

Tail Estimation and Regression Models

Jan Beirlant
Katholieke Universiteit Leuven
University Center for Statistics
De Croylaan 52B
3001 Leuven, Belgium
jan.beirlant@wis.kuleuven.ac.be

Gunther Matthys
Katholieke Universiteit Leuven
University Center for Statistics
De Croylaan 52B
3001 Leuven, Belgium
gunther.matthys@ucs.kuleuven.ac.be

1. Introduction

We consider the classical problem of univariate tail estimation under the maximum domain of attraction condition. To this end, let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with distribution function F and with tail quantile function U defined by $U(x) = \inf\{y : F(y) \geq 1 - 1/x\}$. Further we denote the order statistics by $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$. Then we assume that the properly centred and normed sample maxima $X_{n,n} = \max\{X_1, \ldots, X_n\}$ converge in distribution to a non-degenerate limit, or that for some $\gamma \in \mathbb{R}$ there exist sequences of constants $a_n > 0$, $b_n \in \mathbb{R}$ such that

(1)
$$\lim_{n\to\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = H_\gamma(x),$$

where H_γ denotes the extreme value distribution defined as

$$H_\gamma(x) = \begin{cases} \exp(-(1+\gamma x)^{-1/\gamma}) & \text{for } \gamma \neq 0, \\ \exp(-e^{-x}) & \text{for } \gamma = 0. \end{cases}$$

The main objective of an extreme value analysis is the estimation of extreme quantiles $x_p = U(1/p)$ or small tail probabilities $1 - F(x)$, where a necessary intermediate step is to estimate the extreme value index γ. In such an analysis several problems occur, e.g. the adaptive choice of the number of largest observations (denoted here by k) that will be used in the estimation, the substantial bias of most estimators at some instances and, correspondingly, the use of non-reliable asymptotic confidence intervals, the non-invariance of several estimators with respect to shifts, etc. Our purpose is to indicate that on the basis of certain regression models for different kinds of (generalized) spacings between the largest observations, methods can be constructed that help to relax the problems mentioned above.

In the case $\gamma > 0$, which corresponds to the Pareto-type distributions with $U(x) = x^\gamma \ell(x)$ for some slowly varying function ℓ, a first generalized regression model

of this kind was proposed by Feuerverger and Hall (1999) and Beirlant *et al.* (1999) for the spacings $Z_j = j(\log X_{n-j+1,n} - \log X_{n-j,n})$:

$$\textbf{(2)} \qquad Z_j = \left(\gamma + b \cdot \left(j/k+1\right)^{-\rho}\right) f_j, \quad 1 \le j \le k,$$

with $b \in \mathbb{R}$, $\rho < 0$ and $f_1, f_2, \ldots$ denoting an i.i.d. sequence of standard exponential rv's. In the next section we indicate how this regression model, which arises from asymptotic considerations with respect to the Hill (1975) estimator for $\gamma > 0$, can help to reduce the bias of this estimator for $\gamma > 0$ and of Weissman's (1978) estimator for x_p, to obtain estimates of γ and x_p which are much more stable as a function of k, and to create methods to choose k both for the Hill and for the Weissman estimator.

In a final section we generalize this approach to all maximum domains of attractions. Here regression models can be constructed with, for instance, the following generalized spacings quantities as response variables:

$$V_j = j \log \frac{X_{n-j+1,n} H_{j-1,n}}{X_{n-j,n} H_{j,n}} - j \log \frac{j}{j-1} + \frac{j}{j-1}, \quad 2 \le j \le k,$$

where $H_{j,n}$ denotes the Hill estimator based on the $j+1$ largest observations, or alternatively, with

$$W_j = j \log \frac{X_{n-j+1,n} - X_{n-k,n}}{X_{n-j,n} - X_{n-k,n}}, \quad 1 \le j \le k.$$

We will concentrate on the variables W_j, while representations of V_j are discussed in the contribution presented by A. Guillou in the present volume.

2. The Pareto-Type Case

In this case the Hill estimator still plays a central role in extreme value literature, partly due to its variance-optimality in asymptotic sense. The Hill estimator can be viewed as the maximum likelihood estimator for γ under the reduced model (2) setting $b = 0$:

$$H_{k,n} = \tfrac{1}{k} \sum_{i=1}^{k} i(\log X_{n-i+1,n} - \log X_{n-i,n}).$$

Turning to high quantiles, it is natural to construct estimates based on the regular variation of the tail quantile function U, relating the quantile $x_p = U(1/p)$ of interest to $U(n/k)$, which is then estimated by the empirical quantile $X_{n-k,n}$. This leads to the Weissman estimator

$$\hat{x}_{p,k}^{(0)} = X_{n-k,n} \left(\frac{k+1}{(n+1)p}\right)^{H_{k,n}}.$$

One can reduce the bias of these estimators using maximum likelihood estimators $\hat{\gamma}_k$, $\hat{b}_k$ and $\hat{\rho}_k$ for the parameters γ, b and ρ under model (2) and this for each $k = 3, \ldots, n-1$. As (2) is based on the classical slow variation with remainder condition on ℓ, given by

$$\log \frac{\ell(\lambda x)}{\ell(x)} \sim b(x) \frac{\lambda^{\rho} - 1}{\rho} \quad \text{as } x \to \infty \text{ for all } \lambda \ge 1$$

with $b(x) \to 0$, one can adapt the Weissman estimator to

$$\hat{x}_{p,k}^{(1)} = X_{n-k,n}\left(\frac{k+1}{(n+1)p}\right)^{\hat{\gamma}_k} \exp\left(\hat{b}_k \frac{1-\left(\frac{(n+1)p}{k+1}\right)^{-\hat{\rho}_k}}{-\hat{\rho}_k}\right),$$

which yields estimates that are much more stable as a function of k.

Concerning the selection of tail sample fractions when using a classical tail index estimator such as Hill's estimator, it is intuitively clear that estimates for the bias parameter b in (2) provide help when locating the values of k for which the bias of the estimator is too large, or the value of k for which the mean squared error of the estimator is minimized.

i) Guillou and Hall (2001) propose to choose $H_{\hat{k},n}$ where $\hat{k}$ is the smallest value of k for which

$$\sqrt{\frac{k}{12}}\frac{\left|\hat{b}_{LS,k}\right|}{H_{k,n}} > c_{crit}$$

for all $k > \hat{k}$. Here c_{crit} denotes a critical value such as 1.25 or 1.5, while $\hat{b}_{LS,k}$ is the least squares estimator of b in (2), obtained after setting $\rho = -1$.

ii) When using the mean squared error criterium, one can substitute the estimates of γ, b and ρ in the expressions for the value of k_{opt} which minimizes the *asymptotic* mean squared error.

In case of the Hill estimator this leads to simulation results which compare favorably with other methods such as the bootstrap procedure of Danielsson *et al.* (1997), or the sequential procedure of Drees and Kaufmann (1998). A similar mean squared error method can also be applied to the Weissman's estimator.

3. The General Case $\gamma \in \mathbb{R}$

In the general case the concept of second order regular variation can be incorporated from de Haan's (1970) formulation of the maximum domain of attraction condition (1): for some measurable function a_U and all $t > 0$

$$\lim_{x\to\infty} \frac{U(tx)-U(x)}{a_U(x)} = \begin{cases} (t^\gamma - 1)/\gamma & \text{for } \gamma \neq 0, \\ \log t & \text{for } \gamma = 0. \end{cases} \tag{3}$$

Based on (3), one can derive the following approximate representation of the random variables W_j:

$$W_j = \frac{\gamma}{1-\left(j/k+1\right)^\gamma} f_j, \quad 1 \le j \le k. \tag{4}$$

By construction the maximum likelihood estimator $\hat{\gamma}_k^W$ of γ under (4) is shift and scale invariant. Moreover, it extends the qualities of the maximum likelihood estimator from the peaks-over-thresholds method for $\gamma > -1/2$ (Smith, 1987) to all real γ. In comparison with other estimators, such as the moment estimator, this novel estimator performs well.

Continuing this line of research, an extreme quantile estimator can be constructed on the basis of (3), estimating $a_U(n/k)$ from the following (approximate) exponential representation of the simple spacings $S_j = X_{n-j+1,n} - X_{n-j,n}$:

$$jS_j = a_U\left(n/k\right)\left(\frac{j}{k+1}\right)^{-\gamma} f_j, \quad 1 \le j \le k.$$

This leads to the estimator

$$\hat{a}_U\left(n/k\right) = \tfrac{1}{k}\sum_{j=1}^{k} jS_j\left(\frac{j}{k+1}\right)^{-\hat{\gamma}_k^W}$$

for $a_U(n/k)$, and the extreme quantile estimator

$$\hat{x}_{p,k}^W = X_{n-k,n} + \hat{a}_U\left(n/k\right)\frac{\left(\frac{k+1}{(n+1)p}\right)^{\hat{\gamma}_k^W} - 1}{\hat{\gamma}_k^W}.$$

Introducing slow variation with remainder conditions, representation (4) can be refined to

$$W_j = \frac{\gamma + b\left(j/k+1\right)^{-\rho}}{1-\left(j/k+1\right)^{\gamma}\exp\left\{b\frac{\left(j/k+1\right)^{-\rho}-1}{-\rho}\right\}} f_j, \quad 1 \le j \le k,$$

which makes it again possible to calculate joint estimates $\hat{\gamma}_k^B$, $\hat{b}_k^B$ and $\hat{\rho}_k^B$ for the parameters γ, b and ρ by maximization of the corresponding loglikelihood. The bias-reduced estimator $\hat{\gamma}_k^B$ can be used as an interesting data analytical tool complementary to $\hat{\gamma}_k^W$. When plotting both extreme value index estimators for a particular data set, $\hat{\gamma}_k^B$ will inform the analyst of the quality, and especially the bias of $\hat{\gamma}_k^W$. If the estimators show a similar pattern over a sizable range of k-values, then one can rely on $\hat{\gamma}_k^W$ with a proper choice for the position of the threshold $X_{n-k,n}$, ideally situated in or just beyond the region of congruence. If the patterns diverge rapidly one should be cautious. Further, an adaptive selection method as the one given by Hall and Guillou for Pareto-type tails is also feasible here. Finally, with similar techniques the extreme quantile estimator $\hat{x}_{p,k}^W$ can be corrected for bias.

References

Beirlant, J., Dierckx, G., Goegebeur, Y. and Matthys, G. (1999). Tail index estimation and an exponential regression model, *Extremes* **2**, 177-200.

Danielsson, J.L., de Haan, L., Peng, L. and de Vries, C.G. (1997). A bootstrap based method to choose the sample fraction in tail index estimation, *J. Multivariate Analysis*, to appear.

Drees, H. and Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation, *Stoch. Proc. Applications* **75**, 149-172.

Feuerverger, A. and Hall, P. (1999). Estimating a tail exponent by modelling departure from a Pareto distribution, *Ann. Statist.* **27**, 760-781.

Guillou, A. and Hall, P. (2001). A diagnostic for selecting the threshold in an extreme value analysis, *J. Roy. Statist. Soc. Ser. B*, to appear.

de Haan, L. (1970). On Regular Variation and the Weak Convergence of Sample Extremes, *Math. Centre Tract* **32**, Amsterdam.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Ann. Statist.* **3**, 1163-1174.

Smith, R.L. (1987). Estimating tails of probability distributions, *Ann. Statist.* **15**, 1174-1207.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations, *J. Amer. Statist. Assoc.* **73**, 812-815.

REVISTA DE ESTATÍSTICA
153ª PÁGINA
2º QUADRIMESTRE DE 2001

Using Statistics to Assess the Performance of Stochastic Optimizers

Andreia Hall
University of Aveiro
Department of Mathematics
Campus Universitário
3810 Aveiro, Portugal
andreia@mat.ua.pt

Helena Ferreira
University of Beira Interior
Department of Mathematics
6200 Covilhã, Portugal
ferreira@ubista.ubi.pt

João Pedro Cruz
University of Aveiro
Department of Mathematics
Campus Universitário
3810 Aveiro, Portugal
jpedro@mat.ua.pt

Adelaide Freitas
University of Aveiro
Department of Mathematics
Campus Universitário
3810 Aveiro, Portugal
adelaide@mat.ua.pt

In this talk we use extreme value methodologies to characterise the performance of single-objective stochastic optimizers. We consider the performance of an optimizer in terms of the quality of the solutions produced within a given time of execution.

Single-objective optimizers produce one scalar outcome per optimization run, which is the best objective value found within the run. Optimizer performance depends greatly on the objective function and on the optimization algorithm. The same algorithm may produce very different results if applied to different objective functions. Also, the initial conditions (starting point) and the parameters of the algorithm have a strong influence on the quality of the solutions obtained. In this work we consider two specific objective functions and an evolutionary algorithm described in Baeck et al. (1991).

Stochastic optimizers can be seen as estimators of the optimum value of the objective function. By describing the distribution of optimization outcomes, the performance of the optimizer can be assessed, for instance through the usual measures of estimator qualities such as bias or mean square error. The optimum value is usually the endpoint of the distribution and we consider its estimation both through a parametric and a semi-parametric approach.

Given that the result of an optimization run is the best value (minimum value) among all generations within itself , it is natural to expect that, for a large number of generations, its distribution is close to an extreme-value distribution. Estimation of the tail index of this distribution clearly suggests a Weibull fit. We apply different goodness-of-fit tests for the Weibull distribution described in Lockhart et al. (1994), not only to the outcomes but also to the minimum value of groups of outcomes, since the tests tends to fail when applied directly to the outcomes.

In order to produce reliable parameter estimates, in particular of the endpoint, a large number of optimization runs is required and the procedure may become unrealistic in terms of required time. Ideally it would be desirable to use the intermediate steps of each run in the estimation process.

In most optimization algorithms, the intermediate steps produce dependent (and in many cases non-identically distributed) solutions. In order to be able to use classical extreme-value theory methodologies to estimate the endpoint, certain dependence conditions must be verified. One of these conditions is the D condition for asymptotic independence introduced by Leadbetter (1974). Under this condition the Extremal Types Theorem is generalised for a large number of stationary random sequences. Theoretical validation of D condition is generally technically difficult and has been done for several models such as Autoregressive Moving Average models and certain Markov chains. However, practical validation of the condition raises several problems, which are difficult to deal with. In some situations the condition is indirectly validated through fitting a model to the data for which the condition is known to hold. We propose a non-parametric hypothesis test of *quasi*-independence, which may be used to determine the consistency of a sample with the hypothesis of validation of D condition, without any assumptions on the underlying model. The test is then applied to the values within optimization runs.

References

Baeck, T., Hoffmeister, F. and Schwefel, H. (1991) A survey of evolution strategies. Genetic Algorithms: Proceedings of the Fourth International Conference, 2-9.

Leadbetter, R. (1945). On extreme values in stationary sequences, *Z.Wahrsch.zerw.Gebiet.* **28**, 289-303.

Lockhart, R. and Stephens, M. (1994) Estimation and tests for the three-parameter Weibull Distribution. *Journal of the Royal Statistical Society* B, No. **3**, 491-500.

REVISTA DE ESTATÍSTICA
155ª PÁGINA
2º QUADRIMESTRE DE 2001

Regular Score Tests of Independence for Multivariate Extreme Values

Anthony Ledford, Alexandra Ramos
Department of Mathematics and Statistics
University of Surrey
Guildford, Surrey, GU2 7HX, UK
a.ledford@eim.surrey.ac.uk

The ability to test for independence in multivariate extremes is very important for applied statistical modelling, allowing simplification where appropriate and thus the selection of parsimonious models. Various regimes for testing independence in multivariate extremes have been proposed in the literature. Our focus here is on those based on score statistics, and our starting points will be the test given by Tawn (1988), and additionally, the related Ledford and Tawn (1996) test that is derived under a more general modelling framework. Both of these approaches yield nonregular test statistics that are asymptotically standard normally distributed. However, simulation shows that impractically large data sets are required for normality to hold acceptably. Thus for sample sizes typical of practical applications, the sampling distribution of the score statistic under independence needs to be explored, typically via large scale simulation, in order to determine appropriate critical points. Exploiting these existing tests in practical work therefore remains problematic and time consuming. In this talk we will focus on regularising these existing results in order to obtain score tests that converge rapidly to normal distributions.

1. Introduction

We start by examining the Tawn (1988) and Ledford and Tawn (1996) independence testing frameworks. For algebraic simplicity we focus on the bivariate case, and initially restrict attention to unit Fréchet marginal distributions and the logistic dependence structure. Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ denote independent and identically distributed observations from the bivariate extreme value distribution $F(x,y) = \exp\{-(x^{-1/\alpha} + y^{-1/\alpha})^{\alpha}\}$ where $0 < \alpha \leq 1$ is a dependence parameter. Note that $\alpha = 1$ corresponds to independence.

The above set up is essentially that of Tawn (1988), except that unit Fréchet marginals are used here rather than unit exponential. For our purposes, the important point is that the joint distribution $F(x,y) = \exp\{-(x^{-1/\alpha} + y^{-1/\alpha})^{\alpha}\}$ is assumed to hold over the entire (X,Y) domain. Thus, under this framework, the likelihood contribution of each observation is precisely the joint density $f(x,y) = \partial^2 F(x',y') / \partial x' \partial y'|_{(x,y)}$. Ledford and Tawn (1996) relaxed this approach by assuming that this joint distribution holds explicitly only in the region $R_{11} = \{(x,y): x > u, y > u\}$ where u denotes a fixed high threshold. In order to construct the likelihood under this assumption they used censoring. They took the likelihood contribution of a point falling in R_{11} to be the joint density, as above; that of a point with coordinates (x,y) in region $R_{01} = \{(x,y): x \leq u, y > u\}$ they took to be $\Pr(X \leq u, Y = y) = \partial F(x',y') / \partial y'|_{(u,y)}$. A corresponding expression holds for points in region $R_{10} = \{(x,y): x > u, y \leq u\}$, whereas the likelihood

contribution of a point in region $R_{00} = \{(x,y): x \le u, y \le u\}$ was given by $\Pr(X \le u, Y \le u) = F(u,u)$.

Under either approach, the score contribution for the point (X_i, Y_i) for testing independence is given by

$$U_i = \left. \frac{d \log L(\alpha; X_i, Y_i)}{d\alpha} \right|_{\alpha=1}$$

where $L(\alpha; X_i, Y_i)$ denotes the corresponding likelihood expression. It can be shown that U_i has mean zero and infinite variance in both cases. Hence the usual $\sqrt{n}$ normalisation in the Central Limit Theorem is not powerful enough to yield a non-degenerate limit distribution for the total score $U_{(n)} = \sum_{i=1}^{n} U_i$ Instead, a more powerful normalisation is required, and it may be shown that $U_{(n)} / \sqrt{(n/2)\log n}$ converges in distribution to a standard normal random variable in the limit as $n \to \infty$. It is well known that extremely large values of n are required for this limit result to be acceptable as the basis for an approximation, so the resulting tests are not straightforward to implement practically.

2. Regular Testing

In both the Tawn (1988) and Ledford and Tawn (1996) frameworks, it is the infinite variance of U_i that leads to the nonregularity. Careful examination reveals that this is due, in both cases, to the joint density being used as the likelihood contribution over a region that extends to $(\times, \times)$. Our approach to overcoming this drawback in both approaches is to modify them by applying censoring to region R_{11}. Specifically, we take the likelihood contribution of a point falling in region R_{11} to be the joint survivor probability of region R_{11}, which, for the logistic model specified above, is given by

$$\Pr(X_i > u, Y_i > u) = 1 - 2\exp(-1/u) + \exp(-2^{\alpha} u^{-1}).$$

It is straightforward to show that the resulting likelihoods yield score contributions with zero mean and finite variance. Hence the total score may be normalised by $\sqrt{n}$ in order to obtain a normally distributed nondegenerate limit. This simple procedure therefore yields a regular test of independence. Assessing the impact of the additional censoring is clearly an important issue.

During the presentation we will derive modified score tests as described above and will show results that compare their performance against those of the existing tests for a range of sample sizes. The rapid convergence of the modified tests to normality will be demonstrated. The power functions of the original and modified tests will be discussed, and we will show that any loss of power introduced by the additional censoring is small. We will also consider the associated likelihood ratio tests and will examine how these perform in comparison to those from the unmodified approaches. Robustness of the various tests will be examined, and alternative dependence structures, such as the mixed model, will be considered.

References

Tawn JA (1988). Bivariate extreme value theory: models and estimation. *Biometrika*, **75**, 397-415.

Ledford AW and Tawn JA (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, **83**, 169-187.

REVISTA DE ESTATÍSTICA
157ª PÁGINA
2º QUADRIMESTRE DE 2001

STOCHASTIC MODELS IN FINANCE

Organizer: Wolfgang Runngaldier

Invited Speakers: Ernst Eberlein
Monique Jeanblanc
Kristian R. Miltersen
Marek Rutkowski

REVISTA DE ESTATÍSTICA
159ª PÁGINA
2º QUADRIMESTRE DE 2001

Stochastic Models in Finance

Wolfgang J. Runggaldier
Univ. Padora
Italy
runggal@math.unipd.it

The purpose of this session is to make the audience aware of what kind of stochastic models may arise in the applications to current problems in Finance.

Traditionally, Brownian motion was used for modeling risk factors such as equity processes, interest and foreign exchange rates, etc. Based on empirical evidence on stock return distributions, researchers have started to use instead Levy processes in order to increase accuracy of the models. The presentation by ERNST EBERLEIN deals with this issue by considering models driven by a generalized hyperbolic Levy motion. The presenter will discuss various aspects related to such models, that can be used to describe not only equity prices, but prices of zero coupon bonds as well.

A broad area in Finance, where the modeling framework is still under discusison, is the term structure of interest rates. Traditionally, the primary elements were taken to be the instantaneous interest rates, that are highly theoretical objects and do not correspond in any simple way to real interest rates. More recently, researchers and especially practitioners have therefore started to consider interest rates with finite compounding periods assuming, furthermore, a lognormal structure to justify the widespread use of the

Black and Scholes formula to price interest rate derivatives. Despite this further evolution, there are still some open problems, e.g. in the analysis of futures prices. The presentation by KRISTIAN MILTERSEN is devoted to a very recent approach that is intended to overcome the latter problems by using futures prices as the basic building blocks.

The traditional pricing approaches in bond markets did not take the possibility of default of bond issuing companies into account. As this is an important reality, it has led to recent approaches to the modeling of credit risk. Both the presentations by M. Jeanblanc and M. Rutkowski fall into this area, but their aims are quite different thus allowing for a more extensive overview of modeling issues in this important field.

Concentrating mainly on what is called the reduced form approach to credit risk modeling, MONIQUE JEANBLANC deals with the role that various levels of information may play. Despite its considerable impact in practice, the role of incomplete or only partial knowledge has been mostly ignored in the past. It has however become a topic of major current interest and M. Jeanblanc is going to show some of its aspects in the context of credit risk. For the case when defaultable zero-coupon bonds are traded, she will furthermore show a representation theorem that links hedging strategies in default-free and defaultable markets.

Default need not come as a surprise. Bond issuing institutions may be more or less solid and their rating level may change over time. The riskiness, from the point of view of credit risk, of an institution can therefore be linked to its rating and the presentation by MAREK RUTKOWSKI is intended to model credit migrations. One of the goals is to model the price process of a defaultable bond, for a given initial credit state. To this effect one has to take into account not only the fluctuations of the price due to the presence of a Wiener noise, but also to the sudden jumps due to rating upgrades or downgrades.

For the key references please consult the individual abstracts.

More Realistic Modelling of Risk in Finance

Ernst Eberlein
Universität Freiburg
Institut für Mathematische Stochastik
Eckerstrasse 1
D-79104 Freiburg
eberlein@stochastik.uni-freiburg.de

The shape of the distribution of returns is a key assumption in modelling asset prices. Until nowadays the models used by practitioners such as the classical geometric Brownian motion

$$dS_t = S_t(\mu dt + \sigma dB_t)$$

are driven by a Brownian motion $(B_t)_{t\geq 0}$. This means that they assume normally distributed returns. Analysis of data from the financial markets shows instead that empirical return distributions derived from daily or intraday asset prices are far from being normal. Typically there is a higher peak at the origin, less mass in the flanks and much more mass in the tails. This observation is not restricted to empirical distributions from stock price data. The same characteristics can be observed for bond prices, i.e. interest rate data, as well as for foreign exchange rates. If one considers not only the changes in value of a single instrument, but of a large portfolio of instruments including derivatives, a further phenomenon arises. The return distribution is often skewed in a way which can no longer be neglected. This rules out the assumption of normal returns from the beginning.

In order to increase accuracy of the statistical model one has to look for a more flexible class of probability distributions. Generalized hyperbolic distributions turned out to be tailor-made to fit the returns from financial time series. They contain as subclasses hyperbolic as well as normal inverse Gaussian and many other well-known distributions. Their density is given by

$$d_{GH}(x) = a(\lambda,\alpha,\beta,\delta)\left(\delta^2 + (x-\mu)^2\right)^{(\lambda-1/2)/2} \times K_{\lambda-1/2}\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right)\exp(\beta(x-\mu))$$

where

$$a(\lambda,\alpha,\beta,\delta) = \frac{\left(\alpha^2-\beta^2\right)^{\lambda/2}}{\sqrt{2\pi}\alpha^{\lambda-1/2}\delta^{\lambda}K_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)}$$

K_λ is a modified Bessel function of the third kind with index λ and λ, α, β, δ, μ are the parameters. Although available in the literature since 1977, when they were introduced by Ole E. Barndorff-Nielsen as a variance-mean mixture of normal and generalized inverse Gaussian distributions, in the context of finance generalized hyperbolic distributions were only used in recent years (Eberlein and Keller (1995), Eberlein, Keller, and Prause (1998), Barndorff-Nielsen (1998), Eberlein and Prause (1998)). Since they are infinitely divisible, they generate in a canonical way a process

$(X_t)_{t\geq 0}$ with stationary, independent increments, the generalized hyperbolic Lévy motion. The asset price model derived from this is

$$S_t = S_0 \exp(X_t).$$

This model is natural since along time intervals of length 1 it produces returns which have a generalized hyperbolic distribution. Thus one gets exactly those distributions which one sees in the data.

$$-rf(y,t)+(\partial_2 f)(y,t)+(\partial_1 f)(y,t)b+\frac{1}{2}(\partial_{11} f)(y,t)c$$
$$+\int\left(f(y+x,t)-f(y,t)-(\partial_1 f)(y,t)x\right)F(dx)=0$$

As the classical Brownian motion, this new model can be used for a number of problems in finance: pricing of derivatives and structured products, modelling of term structures, risk management as well as portfolio optimisation. For risk-neutral valuation of derivatives one has to identify equi-valent martingale measures first. Those can be characterised by the triplet of characteristics (b,c,F) of the Lévy process, where b describes the drift term, c the Gaussian part and F the Lévy measure. It is only this triplet together with the interest rate r which determines the value $f(\zeta_t,t)$ of an option, written as a function of the log forward price $\zeta_t = \ln(e^{r(T-t)}S_t)$ and time t. f is the solution of the following partial integro-differential equation

with boundary condition $f(y,T)=w(e^y)$, where w is the payoff of the option. Since the risk-neutral value is given by the conditional expectation of the discounted payoff w, this result extends the Feyn-man-Kac formula. Via duality theory one can show that the choice of a minimum distance martingale measure corresponds to maximizing expected utility with respect to a unique utility function (Goll and Rüschendorf (2001)).

By writing X_t in the form

$$dX_t = \sigma\, dL_t$$

where $(L_t)_{t\geq 0}$ is a standardized Lévy process and replacing σ by a random process $(\sigma_t)_{t\geq 0}$ one can improve the model further to take stochastic volatility into account (Eberlein, Kallsen, and Kristen (2001)). Substantial work in modelling $(\sigma_t)_{t\geq 0}$ has been done recently by Barndorff-Nielsen and Shephard (2001).

Under weak assumptions on the moment generating function, Lévy processes can also be used to model prices of zero coupon bonds. Starting from the standard Heath-Jarrow-Morton diffusion model in the risk-neutral setting one can replace the driving Brownian motion by an appropriate Lévy process. The result obtained is the family of processes

$$P(t,T)=P(0,T)\exp\left[\int_0 r(s)ds-\int_0\theta(\sigma(s,T))ds+\int_0\sigma(s,T)dX_s\right]$$

where $\theta(u)=\log(E[\exp(uX_1)])$ denotes the log of the moment generating function. The processes $P(t,T)$ given above describe a subclass of general semimartingale interest rate models developed by Björk, di Masi, Kabanov, and Runggaldier. It has been shown recently by S. Raible that the martingale measure for this model is unique.

References

Barndorff-Nielsen, O.E. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society London*, ser. A, **353**, 401-419.

Barndorff-Nielsen, O.E. (1998). Processes of normal inverse Gaussian type. *Finance and Stochastics* **2**, 41-68.

Barndorff-Nielsen, O.E. and Shephard, N. (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society* **63**.

Björk, T., di Masi, G., Kabanov, M., and Runggaldier, W. (1997). Towards a general theory of bond markets. *Finance and Stochastics* **2**, 141-174.

Eberlein, E. and Keller, U. (1995). Hyperbolic distributions in finance. *Bernoulli* **1**, 281-299.

Eberlein, E., Kallsen, J., and Kristen, J. (2001). Risk management based on stochastic volatility. FDM-Preprint. University of Freiburg.

Eberlein, E., Keller, U., and Prause, K. (1998). New insights into smile, mispricing, and value at risk: The hyperbolic model. *Journal of Business* **71**, 371-406.

Eberlein, E. and Prause, K. (1998). The generalized hyperbolic model: Financial derivatives and risk measures. FDM-Preprint No. 56, University of Freiburg.

Eberlein, E. and Raible, S. (1999). Term structure models driven by general Lévy processes. *Mathematical Finance* **9**, 31-53.

Goll, Th. and Rüschendorf, L. (2001). Minimax and minimal distance martingale measures and their relationship to portfolio optimisation. To appear in *Finance and Stochastics*.

Osborne, M.F.M. (1959). Brownian motion in the stock market. *Operations Research* **7**, 145-173.

Raible, S. (2000). Lévy processes in finance: theory, numerics, and empirical facts. PhD Thesis. Mathematics Faculty, University of Freiburg.

Samuelson, P. (1965). Rational theory of warrant pricing. *Industrial Management Review* 6, 13-32.

Stochastic Models in Finance

Monique Jeanblanc
Université d'Evry
Boulevard François Mitterand
91025 Evry Cedex France
jeanbl@grozny.lami.univ_evry.fr

The aim of these lecture is to provide a relatively concise - but still self-contained - overview of mathematical notions and results which underpin the valuation of defaultable claims.

Our goal is to furnish results which cover both the classic *value-of-the-firm* (or *structural*) approach, as well as the more recent *intensity-based* methodology.

We study in particular the case when an information flow - formally represented by some filtration $(\mathcal{F}_t, t \geq 0)$ - is present. At the intuitive level, $(\mathcal{F}_t, t \geq 0)$ is generated by prices of some assets, or by other economic factors (e.g., interest rates). We establish a representation theorem, in order to understand the meaning of complete market in a defaultable world and determine the links between the hedging strategy in the default free world and the defaultable one. Whereas the default time is mainly the time where a stochastic boundary is reached, the role of the information plays an important role, especially in the context of partial information. We developp this last point in a structural approach, which leads us to uses the reduced form tools to solve the problem.

REVISTA DE ESTATÍSTICA
163ª PÁGINA
2º QUADRIMESTRE DE 2001

The Market Model of Future Rates*

Kristian R. Miltersen
University of Southern Denmark
Department of Accounting, Finance and Law
Campusvej 55, DK-5230 Odense M
Denmark
krm@sam. sdu. dk

J. Aase Nielsen
University of Aarhus
Department of Operations Research
Bldg. 530, Ny Munkegade, Dk-8000
Aarhus, Denmark
atsjan@imf. au. dk

Klaus Sandmann
University of Mainz
Department of Banking
Jakob-Welder-Weg 9, D-55099 Mainz
Germany
sandmann@forex. bwl. uni-mainz. de

Interest rate futures are basic securities and at the same time highly liquid traded objects. Despite this observation most models of the term structure of interest rate assume forward rates as primary elements. The processes of futures prices and rates are therefore endogenously determined in these models. In addition, in these models hedging strategies are based on forward and/or spot contracts and only to a limited extend on futures contracts.

Inspired by the market model approach of forward rates by Miltersen, Sandmann, and Sondermann (1997), the starting point of this paper is a model of future prices. Moreover, we show that the futures model is an extension of the forward LIBOR model. In addition to the pricing of caps and floors with the Black formula, this new approach allows for the pricing of future style options in closed form. As an important example we price options on the Eurodollar futures as closed form solutions as well.

1. Introduction

For the last 25 years the modeling of the term structure of interest rates has been one of the most deeply studied subjects within the arbitrage theory of financial markets: In general two questions have been raised. Firstly, the question of the proper model: Contrary to the dynamic modeling of one stock price process, the modeling of the term structure of interest rates cannot concentrate only on one zero coupon bond price process. Instead, the price processes of bonds (both coupond bearing and zero coupon) with different maturities as well as interest rates (both spot and forward for

* Financial surport from the Danish Social Science Research Council is gratefully acknowledged by all three authors.

different compounding intervals) have to be integrated within a single consistent arbitrage free framework. Secondly, the question of the pricing of different interest rate derivative, contracts: This includes options on bonds, nominal interest rates, and interest rate swaps among others. Since the basic methodology of no-arbitrage pricing is a relative pricing approach, the derived pricing formulas for these contracts are only valid relative to the assumed modeling framework.

In contrast to the generally accepted Black-Scholes framework for the pricing of options on stocks, the modeling framework of the term structure of interest rates is for practitioners as well as for academics under principle discussion. Without going into a detailed discussion the breakthrough result in modeling the term structure of interest rates has been given by Heath, Jarrow, and Morton (1992). The so-called Heath-Jarrow-Morton model is based on the dynamics of instantaneous forward rates. The dynamics of these forward rate processes are the exogenous elements to the model. The no-arbitrage condition implies an important constraint to the process specification, i.e. the specification of the drift process, whereas the volatility processes can be specified freely as modeling parameters. The modeling framework of Heath, Jarrow, and Morton (1992) highlights the dynamic relationship between different interest rate depending objects like bonds of different maturities, yields, forward rates, etc., which has to be satisfied in a continuous time dynamic setting without arbitrage.

The strength and elegance of the Heath-Jarrow-Morton model comes from the exogenous modeling of the instantaneous forward rate processes, however, this is also the most critical aspect of the model: The instantaneous interest rates are only highly theoretical objects definde by taking the limit as the compounding interval approach zero. These rates does not correspond in any simple way to interest rates observed in real financial markets. It seems to be at least an interesting question whether the disregard of the difference between the dynamics of real observed interest rates with finite compounding periods and instantaneously compounded interest rates lead to precipitate conclusions. A first hint in this direction has been given by Sandmann and Sondermann (1997). One way to exclude negative forward interest rates within the Heath-Jarrow-Morton framework is to assume a lognormal volatility structure. As pointed out by Hogen arid Weintraub (1993) this modeling assumption imposed on instantaneous interest rates implies that rollover returns are infinite. Furthermore, Eurodollar futures cannot be evaluated within this model specification. It was further argued that this negative result about lognormal term structure modeling takes over to the caplet formula by Black (1976). Therefore, Black's formula was thought to be inconsistent with an arbitrage free model of the term structure of interest rates. As shown by Sandmann and Sondermann (1997) the negative result of lognormal interest rates is a result of modeling the wrong rates. If one, instead, imposes the lognormality assumption on a more realistic interest rate notion, namely the effective interest rate, then we get a finite expected value of the rollover return. Although the notion of interest rates was changed, the paper was still based on the instantaneous concept of interest rates. The next step was to consider finite compounding periods. Assuming a lognormal structure on nominal or effective interest rates Miltersen, Sandmann, and Sondermann (1995,1997) justified Black's formula for caps and floors and derived the relationship between what has later been termed the market model approach and the Heath-Jarrow-Morton framework. Hence, it was -shown that Black's formula is indeed consistent, i.e. it is justified that for given assumptions on the dynamics of the nominal forward rates, Black's formula gives arbitrage free prices for caps and floors in a full fledged term structure of interest rates model. Further insights including approximation and pricing of swaptions were then subsequently derived by Brace,

Gatarek, and Musiela (1997) and Jamshidian (1997). Recently the modeling assumption has been extended to include other volatility structures, cf. e.g. Andersen and Andreasen (2000) as well as Zühlsdorf (2000).

From the modeling point of view the main impact of the market model approach is to shift the objective to nominal forward rates. Instantaneous forward rates are within this context endogenously determined; and therefore, the modeling assumptions are more closely related to observed market data. On the other hand the analysis of futures prices has not been addressed in this context. This paper tries to approach this problem. Contrary to the traditional setup this model uses futures prices as the basic building block. In addition to the pricing of caps and floors with the Black formula, this new approach allows for the pricing of future style options in closed form. As an important example we price options on the Eurodollar futures as closed form solutions as well.

The paper is organized as follows: In Section 1 we recall some known results of the relationship between forward and futures prices and rates. Section 2 contains the main model of the term structure of interest rates. This model is within the Heath-Jarrow-Morton framework. Using the insight from the Heath-Jarrow-Morton model clarifies the relationship between the volatility structure and the initial future rate curve. This section also survays-in a probabilistic fasion-the important steps of the convensional market model. Inspired by this, Section 3 introduces a similar structure based on futures rates that the convensional market model does based on forward rates. This structure includes the forward rate market model as a special case. The pricing of futures style options on futures prices and future rates is adressed in Section 3.1. Within the market model approach closed form solutions for these options are derived. Finally in Section 3.2, we derive several pricing result for exotic interest rate option in the context of the market model.

References

Andersen, L. and J. Andreasen (2000): Volatility Skews and Extensions of the LIBOR Market Model; working paper.

Black, and M. Scholes (1973): The Pricing of Options and Corporate Liabilities; *Journal of Political Economy* **81**, 637-654.

Brace, A., D. Gatarec and M. Musiela (1997): The Market Model of Interest Rate Dynamics; *Mathematical Finance* **7**(2), 127-154.

Breeden, D. and R. Litzenberger (1978): Prices of State-Contingent Claims Implicit in Option Prices; *Journal of Business* **51**, 621-651.

Cox, J. C., J. E. Ingersoll jr. and S. A. Ross (1981): The Relation between Forward Prices and Futures Prices; *Journal of Financial Economics* **9**, 321-346.

Geman, H., N. El Karoui and J. -C. Rochet (1995): Changes of Numeraire, Changes of Probability Measure and Option Pricing; *Journal of Applied Probability* **32**, 443-458.

Heath, D., R. Jarrow and A. Morton (1992): Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claim Valuation; *Econometrica* **60** (1), 77-105.

Hogan, M. and K. Weintraub (1993): The Lognormal Interest Rate Model and Eurodollar Futures; Citybank, New York, working paper.

Ingersoll jr., J. E. (1987): Theory of Financial Decision Making; Littlefield Publishers, Inc.

Jamshidian, F. (1997): LIBOR and Swap Market Models and Measures; *Finance and Stochastics* **1** (4), 293-330.

Miltersen, K., K. Sandmann and D. Sondermann (1997): Closed Form Solutions for Term Structure Derivatives with Log-Normal Interest Rates; *Journal of Finance* **52** (1), 409-430.

Sandmann, K., D. Sondermann and K. Miltersen (1995): Closed Form Term Structure Derivatives in a Heath-Jarrow-Morton Model with Log-Normal Annually Compounded Interest Rates; Proceedings of the *Seventh Annual European Research Symposium*. Bonn, September 1994. Chicago Board of Trade, 145-164.

Sandmann, K. and D. Sondermann (1997): A Note on the Stability of Lognormal Interest Rate Models; *Mathematical Finance* **7** (2), 119-125.

Zühlsdorff, C. (2000): Extended Market Models with Affine and Quadratic Volatility; working paper, University of Bonn.

Defaultable Term Structure: Conditionally Markov Approach

Tomasz R. Bielecki
The Northeastern Illinois University
Department of Mathematics
5500 N. St. Louis Ave., Chicago, IL 60625, USA
T-Bielecki@neiu.edu

Marek Rutkowski
Warsaw University of Technology
Faculty of Mathematics and Information Science
pl. Politechniki 1, 00-661 Warszawa, Poland
markrut@mini.pw.edu.pl

A new approach to modelling of credit risk, to valuation of defaultable debt, and to pricing of credit derivatives is presented. The model - based on the Heath, Jarrow & Morton (1992) methodology - uses the available information about the credit spreads combined with the available information about the recovery rates to derive the intensities of credit migrations between various credit ratings classes. Our results complement previous work of Arvanitis, Gregory & Laurent (1999), Duffle & Singleton (1998), Jarrow, Lando & Turnbull (1997), or Schönbucher (1998), among others. Let us finally mention the recent papers by Maksymiuk & Gatarek (1999) and Pugachewsky (1999), who also deal with various generalizations of the HUM framework that cover the credit risk; they do not allow for the credit migrations, though.

1. Default-Free Bonds

Let $B(t,T)$ and $D_C(t,T)$ denote time t prices of *default-free* and *default-risksy* (or *defaultable*) zero coupon bonds maturing at time T, respectively. The default-free bond pays \$1 at time T. The recovery payment for the default-risky bond needs to be modeled. The meaning of the subscript C_t in the notation $D_C(t,T)$ will be explained later in the text. For simplicity, we focus on the recovery scheme in which the recovery payment is received by the holder of the defaultable bond at the maturity time of the bond (this is referred to as the fractional recovery of treasury). Of course, if the defaultable bond does not default prior to or on the maturity date, then it pays \$1 at maturity.

For a fixed horizon date $T^* > 0$, let $(\Omega, \mathbb{F}, \mathbf{P})$ denote the underlying probability space, endowed with the filtration $\mathbb{F} = (\mathcal{F}_t)_{t \in [0,T^*]}$. The process r_t represents the short-term interest rate, and $B_t = \exp\left(\int_0^t r_u du\right)$ is the savings account, as usually. In addition, let the default-free instantaneous forward rate be $f(t,T)$, so that the price $B(t,T)$ of a unit default-free zero coupon bond equals

$$B(t,T) = \exp\left(-\int_t^T f(t,u)du\right). \tag{1}$$

Suppose that there are K credit classes or states, the K^{th} state denoting the state of default. Thus, the risky bond can be in any of the states $i \in \mathcal{K} = \{1,\ldots,K\}$ which represents its credit quality. For any $i < K$, we write $g_i(t,T)$ the *conditional instantaneous* forward

rate for the risky bond that is in class i at time t. We assume the HJM-type dynamics for the instantaneous rates $f(t,T)$ and $g_i(t,T), i = 1,\ldots,K-1$, under the real-world probability $\mathbf{P}$, namely,

(2) $$df(t,T) = \alpha(t,T)dt + \sigma(t,T)dW_t,$$

and

(3) $$dg_i(t,T) = \alpha_i(t,T)dt + \sigma_i(t,T)dW_t,$$

where W is the Wiener process under $\mathbf{P}$. It should be stressed the process

(4) $$D_i(t,T) = exp\left(-\int_t^T g_i(T,u)du\right)$$

does not represent the price process of a tradable security. In other words, "the risky bond that is in state i at time t" is not a tradable asset. In the present framework, a particular defaultable bond is formally defined by its face value (by convention, equal to 1), the maturity date T, the bond's recovery covenants, and the bond's initial rating, that we denote by C_0^1. As a consequence of (2)-(3), we get the following dynamics of $B(t,T)$ and $D_i(t,T)$ under the equivalent risk-neutral probability $\mathbf{P}^*$

(5) $$dB(t,T) = B(t,T)\left(r_t dt + b(t,T)dW_t^*\right),$$

and

(6) $$dD_i(t,T) = D_i(t,T)\left(\left(r_t + \mu_i(t)\right)dt + b_i(t,T)dW_t^*\right),$$

where $\mu_i(t)$ is an $\mathbb{F}$-adapted stochastic process related to a Girsanov's transformation, and W^* denotes the Wiener process under $\mathbf{P}^*$.

2. Credit Migrations

Let $C_t = \left(C_t^1, C_t^2\right)$ denote a two-dimensional conditionally Markov process taking values in $\mathcal{K}\times\mathcal{K}$. In financial interpretation, process C models migrations between credit grades. More specifically, C_t^1 is the current rating of a bond, and C_t^2 represent its previous rating grade. It is thus natural to assume that the states $(K,i), i \in \mathcal{K}$ are absorbing. We wish to model the price process of a defaultable bond, for a given initial credit state C_0 at time 0. We need to take into account not only the fluctuations of the price due to the presence of the Wiener noise (interest rate risk); but also the sudden jumps which are due to rating upgrades or downgrades (credit risk).

Let $\delta_i \in [0,1), i = 1,\ldots,K-1$, denote the recovery rates. This means that if T-maturity unit bond defaults before or at time T, its owner is entitled to the payoff δ_i at maturity date T, provided that the bond belonged to class i just before default occurred.

In order to construct the arbitrage-free defaultable term structure, it will be important to appropriately specify the infinitesimal generator of C^1 at time t, given the σ-field $\mathcal{F}_t$, that is, the K-dimensional matrix

(7) $$\Lambda_t = \begin{pmatrix} \lambda_{1,1}(t) & \cdots & \lambda_{1,K}(t) \\ \cdot & \cdots & \cdot \\ \lambda_{K-1,1}(t) & \cdots & \lambda_{K-1,K}(t) \\ 0 & \cdots & 0 \end{pmatrix},$$

where $\lambda_{i,i}(t) = -\sum_{j=i} \lambda_{i,j}(t)$ for $i = 1,\ldots,K-1$, and where $\lambda_{i,j}$ are $\mathbb{F}$-adapted processes. To this end, we need to postulate that the processes $\lambda_{i,j}$ satisfy the following consistency condition: for $i = 1,\ldots,K-1$ and $t \in [0,T]$

(8) $$\sum_{j=1,j=i}^{K-1} \lambda_{i,j}(t)\big(D_j(t,T) - D_i(t,T)\big) + \lambda_{i,K}(t)\big(\delta_i Z(t,T) - D_i(t,T)\big) + \mu_i(t) D_i(t,T) = 0$$

where we set $Z(t,T) = B(t,T)/B_t$, so that $dZ(t,T) = Z(t,T)b(t,T)dW_t^*$. Let us stress that the entries of the matrix A should be chosen in such a way that $\lambda_{i,j}, i \neq j$, follow non-negative processes. In a very special case of zero recovery (i.e., when $\delta_i = 0$ for $i = 1,\ldots,K-1$) we may take, for instance $\lambda_{iK}(t) = \mu_i(t)$ for $i = 1,\ldots,K-1$ and $\lambda_{i,j} = 0$ for each i when $j \leq K-1$.

To produce a process C with desired properties we need to enlarge the underlying probability space $(\tilde{\Omega}, \tilde{\mathbb{F}}, \mathbf{Q}^*)$, where $\mathbf{Q}^*$ is the extended risk-neutral probability. The filtration $\tilde{\mathbb{F}} = (\mathcal{F}_t)_{t \in [0,T^*]}$ is an enlargement of Wiener filtration, and is also accounting for random shocks leading to credit migrations. Let us set

$$M_{i,j}(t) := H_{i,j}(t) - \int_0^t \lambda_{i,j}(s) H_i(s) ds, \quad \forall t \in [0,T],$$

where $H_i(t) = \mathrm{I}_{\{C_t^1 = i\}}$, and $H_{i,j}(t)$ represents the number of transitions from i to j by C^1 over the time interval $(0,t]$. It can be shown that the processes $M_{i,j}$ are $\tilde{\mathbb{F}}$-martingales under the extended risk-neutral probability $\mathbf{Q}^*$. To explain the conditional Markov feature of C^1, let us denote by $\mathcal{F}_t^C$ the σ-field generated by the observation of credit migration process C up to time t. Then for arbitrary $s > t$ and $i, j \in \mathcal{K}$ we have

$$\mathbf{Q}^*\{C_s = (i,j) \mid \mathcal{F}_t \vee \mathcal{F}_t^C\} = \mathbf{Q}^*\{C_s = (i,j) \mid \mathcal{F}_t \vee \{C_t = (C_t^1, C_t^2)\}\}.$$

The formula above provides the risk-neutral probability that the bond is in the credit grade i at time $s > t$, and the immediately preceding bond's class was j, given the bond was in the credit class C_t^1 at time t which was immediately premed by class C_t^2. (Note that the event $\{C_t = (i,i)\}$ indicates that bond has never left the credit class i prior to time t.)

3. Defaultable Bond Price

We specify the dynamics under the risk-neutral probability $\mathbf{Q}^*$ of the price process $D_C(t,T)$ of a defaultable bond by setting

$$dD_c(t,T) = \sum_{i,j=1,i=j}^{K-1} \big(D_j(t,T) - D_i(t,T)\big) dM_{i,j}(t) + \sum_{i=1}^{K-1} \big(\delta_j B(t,T) - D_i(t,T)\big) dM_{i,K}(t)$$
$$+ \sum_{i=1}^{K-1} H_i(t) D_i(t,T) b_i(t,T) dW_t^* + \sum_{i=1}^{K-1} \delta_i H_{i,K}(t) B(t,T) b(t,T) dW_t^*.$$

Notice that the process $D_C(t,T)$ follows a (local) martingale under $\mathbf{Q}^*$. It can be shown that the price process of a defaultable bond, for any initial condition C_0, is given by the following intuitive expression

(9) $$D_C(t,T) = \mathrm{I}_{\{C_t^1=K\}} \exp\left(-\int_t^T g_{C_t^1}(t,T)du\right) + \delta_{C_t^2}\mathrm{I}_{\{C_t^1=K\}} \exp\left(-\int_t^T (t,T)du\right)$$

for every $t \in [0,T]$. Put another way,

(10) $$D_C(t,T) = \mathrm{I}_{\{C_t^1=K\}} D_{C_t^1}(t,T) + \delta_{C_t^2}\mathrm{I}_{\{C_t^1=K\}} B(t,T).$$

Therefore, for any initial condition C_0, at any time t we have $D_C(t,T) = D_i(t,T)$ on the set $\{C_t^1 = i\}$ for every $i < K$. Furthermore, $D_C(t,T) = \delta_i B(t,T)$ on the set $\{(C_t^1, C_t^2) = (K,i)\}$. We thus see that $D_i(t,T)$ does indeed represent the price at time t of a T-maturity defaultable bond, provided that the bond is currently in the i^{th} credit class. Due to the conditionally Markovian structure of the model, the value $D_i(t,T)$ does not depend on the history of a particular defaultable bond, so that we have a unique price for all defaultable bonds which are currently in a given credit class. For each $i \in \mathcal{K}$, we define the i^{th} credit spread $\gamma_i(t,u)$ by setting $\gamma_i(t,u) = g_i(t,u) - f(t,u)$.

Combining (1) with (4), we get $D_i(t,T) = B(t,T)\exp\left(-\int_t^T \gamma_i(t,u)du\right)$.

Also

(11) $$D_C(t,T) = B(t,T)\left\{\mathrm{I}_{\{C_t^1=K\}} \exp\left(-\int_t^T \gamma_{C_t^1}(t,u)du\right) + \delta_{C_t^1}\mathrm{I}_{\{C_t^1=K\}}\right\}.$$

To simplify formulae (9) and (11), it is convenient to denote $f(t,T) = g_K(t,T)$, so that $\gamma_K(t,T) = 0$. Then (9) and (11) become $D_C(t,T) = X_t \exp\left(-\int_t^T g_{C_t^1}(t,u)du\right)$, and $D_C(t,T) = B(t,T)X_t \exp\left(-\int_t^T \gamma_{C_t^1}(t,u)du\right)$, respectively, where X_t is the promised payoff from the defaultable bond, as at time t; specifically, $X_t = \mathrm{I}_{\{C_t^1=K\}} + \delta_{C_t^2}\mathrm{I}_{\{C_t^1=K\}}$.

Finally let us introduce the default time by setting $\tau = \inf\{t \in \mathbf{R}_+ : C_t^1 = K\}$. Thus X_t can also be represented as follows $X_t = \mathrm{I}_{\{t<\tau\}} + \delta_{C_t^2}\mathrm{I}_{\{t>\tau\}}$..

References

Arvanitis A., J. Gregory and J.-P. Laurent (1999) Building models for credit spreads. *Journal of Derivatives* **6**(3), pages 27-43.

Bielecki T.R. and M. Rutkowski (2000) Multiple ratings model of defaultable term structure. *Mathematical Finance* **10**, 125-139.

Duffie D. and K. Singleton (1998) Ratings-based term structures of credit spreads. Working paper, Stanford University.

Heath D., R. Jarrow and A. Morton (1992) Bond pacing and the term structure of interest rates: a new methodology for contingent claim valuation. *Econometrica* **60**, pages 77-105.

Jarrow R.A., D. Lando and S.M. Turnbull (1997) A Markov model for the term structure of credit risk spreads. *Review of Financial Studies* **10**(2), pages 481-523.

Jarrow R.A. and S.M. Turnbull (1995) Pacing derivatives on financial, securities subject to credit risk. *Journal of Finance* **50**, pages 53-85.

Maksymiuk R. and D. Gatarek (1999) Applying HJM to credit ask. Risk, May, pages 67-68.

Pugachevsky D. (1999) Generalizing with HJM. Risk, August, pages 103-105.

Schönbucher P.J. (1998) Tern stricture modelling of defcaultable bonds. *Review of Derivatives Research* **2**, pages 161-192.

REVISTA DE ESTATÍSTICA
171ª PÁGINA
2º QUADRIMESTRE DE 2001

STOCHASTIC MODELS IN TELECOMMUNICATIONS

Organizer: Walter Willinger

Invited Speakers: Anja Feldmann
Henrik Nyberg
Darryl Veicht

REVISTA DE ESTATÍSTICA
173ª PÁGINA
2º QUADRIMESTRE DE 2001

Stochastic Models in Telecommunications

Walter Willinger
AT&T Labs-Research
180 Park Avenue, Room C284
Florham Park, NJ 07932
walter@research.att.com

1. Overview

The global Internet has experienced a fascinating evolution, especially since the early days of the Web. Unprecedented in its growth, unparalleled in its heterogeneity, and unpredictable or even chaotic in the behavior of its traffic, the Internet has become a gold mine for new, exciting and challenging scientific problems. Resolving some of these problems will be essential for the efficient design and effective engineering and management of the next generation communication networks. Furthermore, solving these problems will rely more and more on an interdisciplinary approach to Internet research that looks to other areas in the natural and social sciences where experimenting with and analyzing complex interacting dynamical systems has a long tradition, e.g., physical sciences, mathematical sciences, and economics.

Recent empirical discoveries concerning various scaling properties of the temporal dynamics of Internet traffic (e.g., self-similarity) or of some of the topological features associated with the physical structure of the Internet (e.g., power-law distributions) have resulted in a number of novel models or "explanations" of these "emergent" phenomena. In fact, unprecedented opportunities for providing mathematically rigorous phenomenological explanations for the observed scaling properties make the networking area distinctly different from other fields in science and engineering which have a rich history in dealing with scaling phenomena (e.g., hydrology, atmospheric sciences, freeway traffic, finance, biology), but where physical explanations are generally given only in an ad-hoc manner, often without serious attempts for validating them empirically. Realizing this difference and the ensuing opportunities for new scientific discoveries, researchers from many different scientific disciplines with a common interest in scaling phenomena have started to view the global Internet as a vast experimental playground, accessible to anyone with an Internet connection.

The observed scaling phenomena in Internet-related measurements have also lead to the emergence of wavelets as an important set of tools for analyzing and mining the vast amount of collected data. Internet measurements tend to be unique and outstanding, not only with respect to quantity and quality, but, more importantly, with respect to the amount of information that is typically contained in every single observation (e.g., TCP header, BGP table). While it is hard to think of many other areas in the sciences where the available data provide such detailed information about so many different facets of system behavior, surprisingly little of the collected data is, in general, diligently "mined" and most of the available data is simply ignored. By providing a "mathematical microscope" for analyzing the scaling behavior of network measurements over many time scales, wavelet-based methods have recently been proven to be a viable alternative that is able to exploit the many facets of the measured data and in the process can give rise to new insights into the complex dynamics of

large-scale networks such as today's Internet. in a system as immense as the Internet where scale is a major concern.

2. Session Structure

This session focuses on some recently considered wavelet-based approaches to mining Internet-related measurements and offers a glimpse at the increasingly experimentation- and measurement-driven nature of Internet research. To this end, the session consists of the following three talks:

1. "*Infinite Divisibility and Traffic Data*"
   Darryl Veitch (EMUlab, University of Melbourne, Australia)
   Email: *d.veitch@ee.mu.oz.au*
   URL: *http://www.emulab.ee.mu.oz.au/~darryl*
2. "*Limit Approximations of the Infinite Source Poisson Traffic Model and Comparisons with Measured Traffic*"
   Henrik Nyberg (Ericsson Radio Systems AB, Sweden)
   Email: *henrik.nyberg@era.ericsson.se*
   URL: see *http://www.md.chalmers.se/Stat/Research/researchgroups/telecom.html*
3. "*Dynamics of Internet Traffic*"
   Anja Feldmann (Computer Science Department, University of Saarbruecken, Germany)
   Email: *anja@cs.uni-sb.de*
   URL: *http://www.cs.uni-sb.de/~anja/*

3. References

In addition to the speakers home pages, the following web sites and references provide useful pointers to additional papers on topics related to the session theme.

1. P. Barford and S. Floyd, *http://www.cs.bu.edu/pub/barford/ss-Ird.html*
2. V. Paxson arid S. Floyd, "Why we don't know how to simulate the Internet" *http://www.aciri.org/floyd/papers/wsc.ps*
3. W. Willinger and V. Paxson, "Where Mathematics meets the Internet" *Notices of the American Mathematical Society*, Vol. 45, pp. 961-970, 1998. *ftp://ftp.ee.lbi.gov/papers/internet-math-AMS98.ps.gz*
4. W. Willinger, "The discovery of self-similar traffic" In: *Performance Evaluation: Origins and Directions*, Lecture Notes in Computer Science, Vol. 1769, pp. 493-505, Springer-Verlag, 2000.

Dynamics of Internet Traffic

Anja Feldmann
Universität des Saarlandes
Saarbrücken, Germany
anja@cs.uni-sb.de

Polly Huang
ETH Zürich
Zürich, Switzerland
huang@tik.ee.ethz.ch

Walter Willinger
AT&T Labs-Research
Florham Park, NJ
walter@research.att.com

In the past decade the Internet has expanded explosively in terms of size, heterogeneity, traffic volume; and diversity of protocols and applications, and will continue to grow and undergo changes in the foreseeable future. Even though humans design the networks, the protocols, the application we do not understand the resulting dynamics of the overall Internet. Yet, understanding the dynamics of Internet traffic is one of the fundamental task of understanding, locating, and fixing performance bottleneck in the Internet.

Within the last few years, the availability of traffic measurements from various different places in the network, at different times, and under various networking conditions has enabled us to study some components that contribute to the traffic dynamics, such as packet arrival process [l, 2, 3, 4, 5, 6, 7], TCP connection process [8], flow characteristics [9, 10, 11, 12, 13], and traffic matrices [l4, 15]. In this process we found traditional statistics inference and model fitting techniques often inadequate. Instead scientific inference and physical-based model building based on the available measurement data and the knowledge about the networking context proved extremely useful.

In this talk I will show what kind of information is available in network data by presenting the results of one study. The goal of this study is to detect performance problems, such as excessive packet delays, packet losses, load changes, or route changes, by relying solely on passive packet-level traces of existing traffic collected from a single tap point in the network. The study takes advantage of a number of structural properties of aggregate TCP/IP packet traces that can be compared across different time periods and across parts of the traffic destined to different subnets. To expose these properties we exploit the built-in scale-localization ability of wavelets.

References

[l] Will F:. Leland, Murad S. Taqqu, Willinger Walter, and Daniel V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEEIACM Trans. Networking*, vol. **2**, no. 1, pp. 1-15, 1994.

[2] Willinger Walter, Murad S. Taqqu, Robert Sherman, and Daniel V. Wilson, "Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Tans. Networking*, vol. **5**, pp. 71-86, 1997.

[3] Vern Paxson and Sally Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. Networking*, vol. **3**, no. 1, pp. 226-244, 1995.

[4] Mark E. Crovella and Azer Bestavros, "Self-similarity in world wide web traffic - evidence and possible causes.," *IEEE/ACM Trans. Networking*, vol. **5**, no. 6, pp. 835-846, 1997.

[5] K. Park, G. Kim, and M.E. Crovella, "On the relationship between file sizes, transport protocols, and self-similar network traffic," in Proc. *IEEE International Conference on Network Protocols*, 1996, pp. 171-180.

[6] Anja Feldmann, Anna Gilbert, and Walter Willinger, "Data networks as cascades: Explaining the multifractal nature of internet WAN traffic," in Proc. *ACM SIGCOMM, September* 1998, pp. 42-55.

[7] Anja Feldmann, Anna Gilbert, Polly Huang, and Walter Willinger, "Dynamics of IP traffic: A study of the role of variability and the impact of control," in Proc. *ACM SIGCOMM*, September 1999.

[8] Anja Feldmann, Anna Gilbert, Walter Willinger, and Tom G. Kurtz, "The changing nature of network traffic: Scaling phenomena.," 1998.

[9] Kimberly C. Claffy, Hans-Werner Braun, and George C. Polyzos, "A parameterizable methodology for internet traffic flow profiling," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 8, pp. 1481-1494, October 1995.

[10] Anja Feldmann, Jennifer Rexford, and Ramon Caceres, "Efficient policies for carrying Web traffic over flow-switched networks," December 1998.

[11] Peter Newman, Greg Minshall, and Tom Lyon, "IP switching: ATM under IP," *IEEE/ACM Trans. Networking*, vol. **6**, no. 2, pp. 117-129, April 1998.

[12] Steven Lin and Nick McKeown, "A simulation study of IP switching," in Proc. *ACM SIGCOMM*, September 1997, pp. 15-24.

[13] Kevin Thompson, Gregory J. Miller, and Rick Wilder, "Wide-area Internet traffic patterns and characteristics," *IEEE Network Magazine*, vol. **11**, no. 6, pp. 10-23, November/December 1997.

[14] Anja Feldmann, Albert Greenberg, Carsten Lund, Nick Reingold, Jennifer Rexford, and Fred True, "Deriving traffic demands for operational IP networks: Methodology and experience," in Proc. *ACM SIGCOMM*, August/September 2000.

[15] Anja Feldmann, Albert Greenberg, Carsten Lund, Nick Reingold, and Jennifer Rexford, "NetScope: Traffic engineering for IP networks," *IEEE Network Magazine*, March 2000.

REVISTA DE ESTATÍSTICA
177ª PÁGINA
2º QUADRIMESTRE DE 2001

Limit Approximations of the Infinite Source Poisson Traffic Model and Comparison with Measured Traffic

Henrik Nyberg
Ericsson Radio AB
Torshamnsgatan 23, Kista
Stockholm, Sweden
Henrik.Nyberg@era.ericsson.se

1. Introduction

The infinite source Poisson model (or M/G/∞ input model) is a fluid queue approximation of network data transmission that assumes that sources begin constant rate transmissions of data at Poisson time points for random lengths of time. This model has been a popular one as analysts attempt to provide explanations for observed features in telecommunications data such as self-similarity, long range dependence and heavy tails. Some features of this model is surveyed, in particular the asymptotic self-similar approximations when transmission lengths are governed by a heavy-tailed distribution. It turns out that different self-similar limits are possible depending on the scaling. This observation raises the issue of selecting the limit that best describes real aggregate traffic. Four traffic data sets are investigated with techniques to estimate tail parameters, Hurst exponents and other parameters of importance for modelling with self-similar models. The agreement with the infinite Poisson model is evaluated. This paper summarises a part of the investigation presented in [1].

2. Preliminaries

A common assumption is that transmission times are given by i.i.d file sizes with a common distribution F which may be heavy tailed with a regularly varying tail so that

(1) $$\bar{F}(x) := 1 - F(x) \propto x^{-\alpha} L(x),\ x \to \infty,$$

where $L(x)$ is a slowly varying function, i.e. $\lim_{t\to\infty} \frac{\bar{F}(tx)}{\bar{F}(t)} = x^{-\alpha},\ x > 0$.

Three important cases can be distinguished: 1) F has infinite expectation and $0 < \alpha < 1$; 2) F has a heavy tail with $1 < \alpha < 2$ so that the variance is infinite whereas the expectation is finite; 3) F has finite variance, which includes classical models with thin (e.g. exponential) tails.

To estimate α, exceedences of a given level was modelled with a generalised Pareto distribution of the form

(2) $$G_\gamma(x) = 1 - \left(1 + \frac{\gamma x}{\sigma}\right)^{-1/\gamma},\ 1 + \frac{\gamma x}{\sigma} > 0,\ \sigma > 0,\ \gamma \in R.$$

If $\gamma > 0$, this is a heavy-tailed distribution with $\alpha = 1/\gamma$. If $\gamma = 0$, (2) is the exponential distribution and if $\gamma < 0$, the distribution has a finite upper endpoint.

A stochastic process $\{X(t); 0 \le t < \infty\}$ is said to be self-similar if the finite dimensional distributions of the time changed and rescaled process $\theta^{-H} X(\theta t)$ are the same as for the original process, i.e.

(3) $$\theta^{-H} X(\theta \cdot) \overset{fidi}{=} X(\cdot) \text{ for } 0 < \theta.$$

A second order process is said to be second order self-similar if its covariance function C satisfies

(4) $$\theta^{-2H} C(\theta s, \theta t) = C(s,t) \text{ for } 0 < \theta.$$

A fractional Brownian motion (FBM) B_H is a centered continuous Gaussian process with a covariance function that satisfies (4). Thus, B_H is self-similar with Hurst parameter H. (H=0.5 gives the ordinary Brownian motion.) Another self-similar process is stable Lévy motion which has i.i.d. increments with a non-normal stable distribution with index α, $0<\alpha<2$. The Hurst parameter is in this case $H=1/\alpha$.

Let Γ_k and L_k be the arrival time and the life time respectively of the kth active source. The number of active sources, $N(t)$, at time t is a Poisson random variable with expectation $m(t):=\lambda\int_0^t \bar{F}(s)\,ds$ where λ is the Poisson arrival rate of active sources.

It is assumed that data is sent at unit rate from the source during transmission. The *cumulative traffic* in [0,t] is

$$A(t):=\int_0^t N(s)\,ds=\sum_{k:\Gamma_k\le t}\min(L_k,t-\Gamma_k). \tag{5}$$

The *traffic rate* process, taken as the cumulative traffic over intervals of length $\Delta>0$ is

$$Y_\Delta(k)=A\big((k+1)\Delta\big)-A(k\Delta),\quad k=0,1,2,\ldots \tag{6}$$

3. Limits and Approximations

For limits in the infinite source Poisson model, we have three principally different cases depending on the tail parameter α as mentioned in the previous section. The scope is here limited to the case $1<\alpha<2$. For this case, two different limits can be obtained. The FBM limit is obtained by first letting the Poisson arrival rate $\lambda\to\infty$ and then letting the time scaling $T\to\infty$. The Lévy stable noise limit is obtained by letting $T\to\infty$ while λ is fixed.

Let μ be the (finite) expectation with respect to F. The asymptotic variance of the cumulative traffic rate, as $T\to\infty$, is given by

$$V\big[A(T)\big]:=\sigma^2(T)=\lambda\int_0^T E\Big[\min(L_k,s)^2\Big]ds\propto\frac{2\lambda}{(2-\alpha)(3-\alpha)}T^3\bar{F}(T). \tag{7}$$

For the case that λ is constant and $T\to\infty$, the standardised process

$$\hat{G}_T(t):=\frac{A(Tt)-\lambda\mu Tt}{\sigma(T)} \tag{8}$$

does not converge (not even in the sense of finite dimensional distributions) to a self-similar Gaussian process. However, it is possible to get stable limits. Let $f^{\leftarrow}$ be the left continuous inverse of a monotone non-decreasing right continuous function $f:R\mapsto R$. Let

$$b(T)=\left(\frac{1}{1-F}\right)^{\leftarrow}(T)=T^{1/\alpha}L_1(T), \tag{9}$$

for some slowly varying function L_1. Observe that $b(T)/\sigma(T)\to 0$ as $T\to\infty$.

Then, for fixed λ,

$$\hat{G}_T(t)\frac{\sigma(T)}{b(T)}=\frac{A(Tt)-\lambda\mu Tt}{b(T)}\Rightarrow X_\alpha(t)\ \text{as}\ T\to\infty, \tag{10}$$

where $X_\alpha(\cdot)$ is α-stable Lévy motion with totally right-skewed marginal distribution. On the other hand, if $T\to\infty$ after letting $\lambda\to\infty$, $A(\cdot)$ approaches a fractional Brownian motion. The interpretation might be that on large time scales with moderate input rate, stable Lèvy motion is a reasonable approximation, whereas on small or

moderate time scales and for high input rate, fractional Brownian motion is a better approximation. The following result gives conditions for an input rate $\lambda(T)$ that depends on the time scaling T to give either the stable limit or the FBM limit. The complete theorem and a proof are given in [2].

Proposition 1 Assume a family of infinite source Poisson models indexed by T with Poisson rate $\lambda = \lambda(T)$ and with cumulative input process $A_T(\cdot)$. Assume that $\lambda = \lambda(T)$ depends on T so that the following slow growth conditions holds:

$$\lim_{T\to\infty} \lambda T\bar{F}(T) = 0$$

Then the cumulative process $(A_T(Tt), t \geq 0)$ satisfies the limit relation

$$\frac{A_T(T\cdot) - T\lambda\mu(\cdot)}{b(\lambda T)} \xrightarrow{fidi} X_\alpha(\cdot),$$

where $X_\alpha(\cdot)$ is a Lévy stable motion.

Assume that the following fast growth condition holds:

$$\lim_{T\to\infty} \lambda T\bar{F}(T) = \infty$$

Then the cumulative process $(A_T(Tt), t \geq 0)$ properly normalised as

$$\frac{A_T(T\cdot) - T\lambda\mu(\cdot)}{\sqrt{\lambda T^3 \bar{F}(T)}},$$

converges in $D[0,\infty)$ to a fractional Brownian motion with self-similarity parameter $H = (3-\alpha)/2$.

4. Data Analysis and Results

The four data sets investigated are 1) the Boston University 1995 study of WWW sessions; 2) the UC Berkeley home IP HTTP data collected in November 1996; 3) traces collected in 1997 at a Customer Service ATM Switch in Munich and 4) detailed data from a corporate Ericsson WWW server from October 1998. Only data sets 1, 2, and 4 are covered here due to space limitations. The form of the data analysed is as aggregate flow rate (bytes per 1 or 10 seconds) generated by the users during the measurement period. The BU data was modified ('BUburst') by lumping together file requests by a user that were at most 0.5 seconds apart. As a reference, simulated M/G/∞ traffic is added ('simM/G/∞'). For comparison, synthetic traffic traces, 'UCB CBR' and 'Eri CBR', were made from the UCB and Ericsson data using the original file sizes but transfer times corresponding to a constant bitrate (CBR).

Estimation results for the traces are summarised in Table 1. $\hat{H}$ is the estimated Hurst parameter. H^* is the Hurst parameter estimated from the file size tails as $H^* = (3-\alpha)/2$. The fourth column shows the estimated values of $\lambda T\bar{F}(T)$ used in Proposition 1. Columns 5-6 indicate the goodness of fit to normal and stable marginal distributions. The last column indicates the type of dependence. The precise definition of the dependence classification is given in [1].

Data set	$\hat{H}$	H^*	$\lambda T\bar{F}(T)$	Gaussian	Stable	dependence
simM/G/∞	.90 ±.01	.90	8	good	bad	strong/long
BUburst 10s	.89 ±.02	.67 ±.01	.09	med.	good	strong/long
BUburst 1s	.81 ±.01	.67 ±.01	.07	med.	bad	nonstationary
UCB 10s	.58 ±.03	.62 ±.01	16	bad	good	independent
UCB CBR 10s	.95 ±.07	.62 ±.01	16	good	bad	strong/long
Ericsson	.88 ±.02	.86 ±.08	.5	bad	bad	nonstationary
Eri CBR 1s	1.48 ±.02	.86 ±.08	.5	med.	med.	strong/long

Table 1. Results of traffic rate characterisation.

Table 2 shows the estimated tail parameter $\gamma = 1/\alpha$ associated with file transfer times, file sizes and the mean transfer rate per file.

Data set	transfer time: γ	file size: γ	mean rate: γ
BUburst	.60 ±.02	.69 ±.13	1.01 ±.14
UCB	.57 ±.02	.52 ±.02	.79 ±.04
Ericsson	.78 ±.16	1.15 ±.18	-

Table 2. Tail parameters related to file transfers

The Hurst parameter was estimated with wavelet methods (regression on time scales of squared wavelet coefficients). The tails of the distributions of file sizes, transfer times and traffic rates were investigated with maximum likelihood (ML) estimation of the generalised Pareto distribution given by equation (2) applied to the top 5% of the observations. The goodness of fit was checked with QQ-plots. The parameters of stable Lévy motion were estimated with ML methods. Dependence in the heavy-tailed data was investigated using the heavy-tailed acf. Details and references for the estimation methods are given in [1].

5. Conclusions

The infinite source Poisson model is a fairly flexible model to predict the limiting behaviour of aggregate traffic. Global statistical properties such as heavy tails and long range dependence appeared in the data as predicted by the model. File sizes were consistently heavy tailed, usually with $1 < \alpha < 2$. The scaling behaviour summarised in the estimated Hurst parameters was compatible both with FBM and the stable noise model. However, all measured marginal distributions are far from normal except for the synthetic traces (in particular the 'UCB CBR 10s' in Table 1).

The overall impression is that the infinite source Poisson model is not sufficient to adequately capture the properties of the data. A discrepancy from the model assumptions is that file transfer rates appear to have considerable variation, see Table 2. This is probably the center of the problem and has led to subsequent work on the impact of random transfer rates in the infinite source Poisson model [2]. See e.g. Taqqu etal. [4] for related research.

Acknowledgements

This paper is based on joint research with major contributions from Charles-Antoine Guerin, Olivier Perrin, Sidney Resnick, Holger Rootzén and Catalin Starica.

References

[1] Guerin, C.A., Nyberg, H., Perrin, O., Rootzén, H. And Starica, C. (2000), "Empirical Testing of the Infinite Source Poisson Data Traffic Model", Report 2000:4, *Mathematical Statistics*, Chalmers University of Technology.

[2] Maulik, K., Resnick, S. and Rootzén, H., "A Network Traffic model with Random Transmission Rate", Report 2000:85, Mathematical Statistics, Chalmers Univ. of Tech.

[3] Mikosch, T., Resnick, S., Rootzén, H. And Stegeman, A., "Is network traffic approximated by stable Lévy motion or fractional Brownian motion?", Report 1999:32, Mathematical Statistics, Chalmers University of Technology.

[4] Taqqu, M., Pipiras, V. (2000), "Convergence of weighted sums of random variables with long-range dependence", *Stochastic Processes and their Applications*. **90** (157-174).

REVISTA DE ESTATÍSTICA
181ª PÁGINA
2º QUADRIMESTRE DE 2001

Infinite Divisibility and Traffic Data

D. Veitch
The University of Melbourne
EMUlab, Department of Electrical and Electronic engineering
Victoria, 3010
Australia
d.veitch@ee.mu.oz.au

P. Abry
CNRS UMR 5672, Physics Lab.
Ecole Normale Sup\'erieure de Lyon
46, all\'ee d'Italie 69364 LYON Cedex 07
France
Patrice.Abry@ens-lyon.fr

Infinite divisibility is well known as a defining property of important families of probability distributions, and a key concept underlying the semi-group formulation of Markov processes. Its connection to scaling phenomena is far less appreciated, with the exception of the statistical theory of turbulence where it has been exploited for two decades, albeit to a limited extent. In this talk we clarify and explain the role of infinite divisibility in the description of scaling behaviour, through the unifying viewpoint of Infinitely Divisible Cascades. We introduce a wavelet formulation of these models as a compact description of the diverse forms of scaling behaviour found in packet data in telecommunications networks, and illustrate their advantages over other scaling models.

The Infinitely Divisible Cascade models contain as special cases important classes of scaling processes, used to date to model scaling behaviour in teletraffic as well as in diverse other fields. These include long range dependent, exactly self-similar, and multifractal models. More than just being a larger class however, the semi-group structure of IDCs allows one to put into context the underlying structural assumptions of the simpler alternatives. It constitutes a {\it natural} generalisation which does away with the notion of scaling as being synonymous with power-laws. As stable distributions form an infinitely divisible class, infinite moment models with scaling features are also included in a natural way.

An IDC framework could be used in the time domain, however there are many advantages in basing the study of time series with scaling properties on wavelet coefficients.

We will describe estimators, based on the wavelet coefficients of underlying time series, which can be used to first detect the presence of scaling and the corresponding range of scales over which it exists, and second, to estimate the corresponding parameters of the cascades. Particular attention is paid to correctly discriminating between non stationarity and scaling phenomena.

Through the IDC lens, we present an analysis of exceptionally precise TCP/IP data made available by the WAND group at the University of Waikato at *http://wand.cs.waikato.ac.nz/wand/wits/index.html*. This set of data, the `Auckland II' traces, are taken from both directions of the access link of the University of Auckland to the external Internet.

The capture hardware developped at WAND (measuring ATM technology at $155 $\,Mbits/s for Auckland II) is capable of loss-less measurement with synchronised timestamps accurate to below $1\,\mu$s. >From the raw data many time series have been extracted and analysed, for example at the IP level the byte and packet flow rates are examined. At a higher protocol level, TCP connections (and UDP flows) are examined through such series as the arrival rate, durations, and interarrival times of TCP connections (TCP is the protocol used for reliable data transfer over the Internet, including Web based data retrieval).

The traces we analyse offer a representative vision of TCP/IP traffic, the behaviour of which is a key problem in the current Internet. We will discuss how the broader framework that IDC models offer gives insight into the statistical origins of the different scaling behaviours found over different scaling regimes, and the relationships between the very different scaling properties found in different time series. Such insights are valuable in the search for physical, that is network based, origins of scaling in traffic, which are in turn essential in order to answer key questions such as the robustness of the fractal traffic phenomena in the face of the rapid, and interlinked, evolution of networks and the teletraffic they carry.

PART II
BERARDO PRIZE

Entries:

M. Fátima Brilhante
Sung Nok Chiu
Gerda Claeskens and Liang Peng
Konstantinos Fokianos
Roland Fried
Sangita Kulathinal and Dario Gasbarra
Domenico Marinucci
Sandra Mendonça
Jacobo de Uña-Álvarez

The prize has been awarded to Dr. D. Marinucci, for his paper
Gaussian semiparametric estimation for random fields with long range dependence

"*Marinucci is an innovator [...]. He works on a very difficult problem and finds a non-trivial reduction of the problem to a question that can be technically handled. [...] The structure of his work is great. His Summary should be studied at school.*"

[from the jury report]

On the Infinite Divisibility of the Spacings of Exponential Mixtures

M. Fátima Brilhante
University of Azores
Department of Mathematics and C.E.A.U.L.
Rua da Mãe de Deus, Apartado 1422
9501-801 Ponta Delgada (Codex), Portugal
fbrilhante@notes.uac.pt

The arithmetic properties of mixtures of distributions have been object of study of many statisticians. Goldie (1967) by proving that the product of two independent non-negative random variables is infinitely divisible if one of them is exponentially distributed, proved that an exponential mixture is infinitely divisible. Steutel (1967), on the other hand, established a sufficient condition for a generalized exponential mixture (i.e. an exponential mixture with some negative mixing coefficients) to be infinitely divisible.

In this paper we will be interested in proving the infinite divisibility of the spacings of an exponential mixture. We will do so by showing that the spacings are themselves exponential mixtures and, therefore, infinitely divisible.

Let

$$Z=\begin{Bmatrix} X_1 & X_2 & \cdots & X_m \\ p_1 & p_2 & \cdots & p_m \end{Bmatrix}$$

where $\sum_{k=1}^{m} p_k = 1$, $p_k > 0$, and X_k is exponentially distributed with scale parameter δ_k, i.e. with distribution function (df) $F_{X_k}(x) = 1 - e^{-x/\delta_k}$, $x > 0$.

Since the random variable Z is a mixture of exponential random variables, it has df

(1) $$F_Z(z) = \sum_{k=1}^{m} p_k \left(1 - e^{-z/\delta_k}\right), \quad z > 0.$$

Let $Z_{i:n}$ denote the ith ascending order statistic of a random sample of size n from a population with df (1). The probability density function (pdf) of the spacing $U_{i,n;m} = Z_{i:n} - Z_{i-1:n}$, $i = 1,\ldots,n$, with the usual convention $Z_{0:n} = 0$, is given by

$$f_{U_{i,n;m}}(u) = \frac{n!}{(i-2)!} \sum_{k_1+\cdots+k_m=n-i+1} \frac{1}{k_1!\ldots k_m!} \sum_{t=0}^{i-2} (-1)^t \sum_{j_1+\cdots+j_m=t} \frac{t!}{j_1!\ldots j_m!} \times$$

$$\times \left\{ \frac{\delta_2\delta_3\ldots\delta_m\, p_1^{k_1+j_1+1} p_2^{k_2+j_2} \ldots p_m^{k_m+j_m}}{\delta_2\delta_3\ldots\delta_m(k_1+j_1+1) + \delta_1\delta_3\ldots\delta_m(k_2+j_2) + \cdots + \delta_1\delta_2\ldots\delta_{m-1}(k_m+j_m)} \right.$$

$$+\cdots+$$

$$\left. + \frac{\delta_1\delta_2\ldots\delta_{m-1}\, p_1^{k_1+j_1} p_2^{k_2+j_2} \ldots p_m^{k_m+j_m+1}}{\delta_2\delta_3\ldots\delta_m(k_1+j_1) + \delta_1\delta_3\ldots\delta_m(k_2+j_2) + \cdots + \delta_1\delta_2\ldots\delta_{m-1}(k_m+j_m+1)} \right\}$$

$$\times \left(\frac{k_1}{\delta_1} + \frac{k_2}{\delta_2} + \cdots + \frac{k_m}{\delta_m} \right) \exp\left[-\left(\frac{k_1}{\delta_1} + \frac{k_2}{\delta_2} + \cdots + \frac{k_m}{\delta_m} \right) u \right], \quad u > 0$$

when $i = 2,\ldots,n$; and

$$f_{U_{1,n:m}}(u) = \sum_{k_1+\cdots+k_m=n} \frac{n!}{k_1!\ldots k_m!} p_1^{k_1}\ldots p_m^{k_m}\left(\frac{k_1}{\delta_1}+\cdots+\frac{k_m}{\delta_m}\right)\exp\left[-\left(\frac{k_1}{\delta_1}+\cdots+\frac{k_m}{\delta_m}\right)u\right]$$

for $u > 0$.

The expressions above remain valid if we consider a generalized exponential mixture instead. However, to ensure in this case that we will have a proper pdf, the parameters p_k and δ_k must satisfy some conditions (cf. Steutel (1967), Bartholomew (1969) and Harris et al. (1992)).

The "popularity" of the exponential model often relies in the fact that it allows, most of the time, a simple analytical treatment. Therefore, when investigating alternative models to the exponential, the Laplace model emerges as a natural candidate given its simple expression obtained as the difference of two independent and identically distributed (iid) exponential random variables. And just like the model itself, the laplacian spacings will prove to be infinitely divisible.

Let $(X_{1:n},\ldots,X_{n:n})$ be the vector of ascending order statistics associated with a random sample of size n from a Laplace population with location parameter λ and scale parameter δ, i.e. with pdf

$$f(x) = \frac{1}{2\delta}\exp\left(-\frac{|x-\lambda|}{\delta}\right), \quad -\infty < x < \infty,$$

and let

$$V_{i,n} = X_{i+1:n} - X_{i:n}, \quad i = 1,\ldots,n-1$$

denote the ith spacing.

Considering, without loss of generality, $\lambda = 0$ and $\delta = 1$, we obtain the following expressions for the pdf of $V_{i,n}$

$$\text{(2)} \qquad f_{V_{i,n}}(v) = i e^{-iv}\left[P(R \geq i+1) + \frac{n-i}{n-2i}P(R=i)\right] + \\ + (n-i)e^{-(n-i)v}\left[P(R \leq i-1) - \frac{i}{n-2i}P(R=i)\right], \; v > 0$$

if $n - 2i \neq 0$, and

$$\text{(3)} \qquad f_{V_{n/2,n}}(v) = \frac{n}{2}e^{-nv/2}\left[1 - P\left(R=\frac{n}{2}\right)\right] + \left(\frac{n}{2}\right)^2 v e^{-nv/2} P\left(R=\frac{n}{2}\right), \; v > 0$$

if $i = n/2$, where R is a binomial random variable with parameters n and $p = 1/2$.

A closer observation of the exponential coefficients of (2) reveals that these coefficients have opposite signs. In fact, if

$$p_1 = P(R \geq i+1) + \frac{n-i}{n-2i}P(R=i) \quad \text{and} \quad p_2 = P(R \leq i-1) - \frac{i}{n-2i}P(R=i)$$

then $p_1 > 0$ and $p_2 < 0$ if $i < n/2$, while $p_1 < 0$ and $p_2 > 0$ if $i > n/2$. Hence, the laplacian spacings are generalized exponential mixtures of two components, and

therefore are infinitely divisible (cf. Steutel (1967)). On the other hand, if n is even, the central spacing which is a mixture of an exponential and a gamma distributions also prove to be infinitely divisible, since we will see that the characteristic function has a Lévy-Khinchine's representation of an infinitely divisible characteristic function.

Although the symmetry of the Laplace model is not reproduced in the distributions of the spacings, other forms of symmetry will be pointed out. On one hand, the coefficients of the mixtures come from the symmetrical binomial model. This property is obviously the consequence of the generation of the Laplace distribution as the difference of iid exponential variables in a symmetrization process. On the other hand, the spacings having a symmetrical standing in what regards the central spacing, or central spacings if n is odd, are identically distributed.

Given that the exponential is a special case of the Generalized Pareto model, we will also see that it is possible to generalize the results obtained for the spacings of an exponential mixture to the ratios of consecutive order statistics of a Generalized Pareto mixture.

References

Bartholomew, D. J. (1969). Sufficient conditions for mixture of exponentials to be a probability density function. *Annals of Mathematical Statistics*, **40**, 2183-2188.

Brilhante, M. F. (1999). Inferência Estatística em Modelos Não Gaussianos com Recurso a Spacings e Outras Funções de Estatísticas Ordinais. Dissertação de Doutoramento, Universidade dos Açores, Ponta Delgada.

Feller, W. (1971). An Introduction to Probability Theory and Its Applications. vol. II. John Wiley & Sons, New York.

Goldie, C. M. (1967). A class of infinitely divisible distributions. *Proc. Cambridge Philos.* Soc., **63**, 1141-1143.

Harris, C. M., Marchal, W. G. and Botta, R. F. (1992). A note on generalized hyperexponential distributions. *Communications in Statistics-Stochastic Models*, **8**, 179-191.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). Continuous Univariate Distributions. 2nd ed. John Wiley & Sons, New York.

Reiss, R-D. (1989). Approximate Distributions of Order Statistics With Applications to Nonparametric Statistics. Springer-Verlag, New York.

Steutel, F. W. (1967). Note on the infinite divisibility of exponential mixtures. *Annals of Mathematical Statistics*, **38**, 1303-1305.

Steutel, F. W. (1968). A class of infinitely divisible mixtures. *Annals of Mathematical Statistics,* **39**, 1153-1157.

Steutel, F. W. (1969). Note on completely monotone densities. *Annals of Mathematical Statistics*, **40**, 1130-1131.

VOLUME II

2° QUADRIMESTRE DE 2001

REVISTA DE ESTATÍSTICA
189ª PÁGINA
2º QUADRIMESTRE DE 2001

Spatial Point Pattern Analysis by Using Voronoi Diagrams and Delaunay Tessellations — A Comparative Study

S. N. Chiu
Hong Kong Baptist University
Department of Mathematics,
Kowloon Tong, Hong Kong.
snchiu@hkbu.edu.hk

1. Introduction

An important task of the analysis of spatial point patterns is to check how well a model fits the observed data. The benchmark that point patterns are often compared with is the point process that is completely spatially random (CSR), which means that the patterns are realisations of a Poisson or binomial point process. The CSR or any other model is usually tested by comparing some summary characteristics of the observed pattern with those of the hypothesised model. Often used approaches for analysing mapped data are based on nearest-neighbours, quadrat counts and inter-point distances.

Another approach, which has not been widely developed, is based on the Voronoi polygons and Delaunay triangles. This approach was originally proposed by Evans (1967) and advocated by various researchers (see Okabe et al., 2000). For a given point pattern, we associate all locations on the plane with the closest member(s) of the point pattern. The result is a tessellation of the plane into a set of the regions associated with members of the point pattern. This tessellation is called the Voronoi diagram generated by the point pattern and the regions constituting the Voronoi diagram are called Voronoi polygons. If we join all generator points whose Voronoi polygons share a common Voronoi edge, we obtain a second tessellation called the Delaunay tessellation generated by the point pattern. Both the Voronoi diagram and the Delaunay tessellation are uniquely determined by the point pattern, and vice versa. Thus, characteristics of these tessellations are also characteristics of the point pattern.

Hutchings and Discombe (1986) compared the power of the Monte-Carlo simulation tests based on the distributions of area, the perimeter and the number of sides, and suggested that the best single Voronoi polygon characteristic for distinguishing both regular and clustered empirical patterns from a hypothesized pattern of CSR is area, while perimeter is sensitive to clustering only. They also found that the Monte-Carlo simulation test based on the length of an edge in the Delaunay tessellation is effective in distinguishing both regular and clustered patterns from those of CSR. Boots (1975) and Vincent et al. (1977) analysed patterns of urban settlements in various parts of the United States and showed that, in this context, the Pearson goodness-of-fit test for the distribution of an angle of a Delaunay triangle is more effective than the nearest neighbour approach in correctly rejecting a null hypothesis of CSR. Mardia et al. (1977) and Boots (1986) used the distributions of the minimum angle and the maximum angle, respectively, to construct tests for CSR and the latter found that the minimum angle has a higher overall power.

Recently, numerically tractable expressions of the distributions of various characteristics of Voronoi polygons and Delaunay triangles have been derived. These

expressions enable us to compare the empirical distributions obtained from a point pattern and the theoretical distributions, under the CSR hypothesis, by using the Kolmogorov—Smirnov test. The advantage of using the Kolmogorov—Smirnov test is that critical regions can be approximated without simulations.

2. Results

The distributions of the length of an edge in the Voronoi diagram, and the area and the perimeter of a Delaunay triangle have been derived by Mecke and Muche (1995), Rathie (1992) and Muche (1996), respectively. These expressions have not yet been used in analysing empirical point patterns.

The powers of the Kolmogorov—Smirnov tests based on the above three distributions, as well as the distributions of the length of an edge in the Delaunay tessellation, a randomly chosen angle, the minimum angle, the middle angle and the maximum angle of a Delaunay triangle, and the distance between a point in the point pattern and a vertex of the Voronoi polygon containing the point for testing CSR are compared.

Preliminary investigations have been done. For clustered patterns, the tests using angles of a Delaunay triangle are less powerful than the others, among which the distance between a point in the point pattern and a vertex of the Voronoi polygon containing the point is more effective than the others. For regular patterns, again the angles are less powerful but the differences are smaller than in the clustered cases.

References

Boots, B.N. (1975). Patterns of Urban settlements revisited, *The Professional Geographer* **27**, 426-431.

Boots, B.N. (1986). Using angular properties of Delaunay triangles to evaluate point patterns, *Geographical Analysis* **18(3)**, 250-260.

Evans, I.S. (1967). The properties of patterns of points, measured by space filling and angular relationships, *Geographical Articles (Cambridge)* **8**, 63-77.

Hutchings, M.J. and Discombe, R.J. (1986). The detection of spatial pattern in plant populations, *Journal of Biogeography* **13**, 225-236.

Mardia, K.V., Edwards, R. and Puri, M.L. (1977). Analysis of central place theory, *Bulletin of the International Statistical Institute* **47**, 93-110.

Mecke, J. and Muche, L. (1995). The Poisson Voronoi tessellation I. A basic identity, *Mathematische Nachrichten* **176**, 199-208.

Muche, L. (1996). Distributional properties of the three-dimensional Poisson Delaunay cell, *Journal of Statistical Physics* **84**, 147-167.

Okabe, A., Boots, B., Sugihara, K. and Chiu, S.N. (2000). Spatial Tessellations. Concepts and Applications of Voronoi Diagrams. Second Edition. Wiley, Chichester.

Rathie, P.N. (1992) On the volume distribution of the typical Poisson—Delaunay cell, *Journal of Applied Probability* **29**, 740-744.

Vincent, P.J., Howarth, J., Griffiths, J. and Collins, B. (1977). Urban settlement patterns and the properties of the simplical graph, *The Professional Geographer* **29**, 21-25.

Inference for a Receiver Operating Characteristic Curve via Smoothed Empirical Likelihood

Gerda Claeskens
Texas A&M University
Department of Statistics
447 Blocker Building
College Station, TX 77843, USA
Gerda@stat.tamu.edu

Liang Peng
Georgia Institute of Technology
School of Mathematics
Atlanta, GA 30332, USA
Peng@math.gatech.edu

1. The Receiver Operating Characteristic Curve

The receiver operating characteristic (ROC) curve is an important tool to summarize the performance of a medical diagnostic test for detecting whether or not a patient has a disease. In a medical test resulting in a continuous measurement T, the disease is diagnosed if $T > c$, for a given a threshold c. Suppose the distribution function of T is F_1 conditional on disease and F_2 conditional on non-disease. The ROC curve is defined as the graph $(1\text{-}F_1(c), 1\text{-}F_2(c))$ for various values of the threshold c, or in other words, sensitivity versus 1 - specificity, power versus size for a test with critical region $\{T > c\}$. This enables one to summarize a test's performance or to compare two diagnostic tests. For more information about ROC curves and their use, we refer to Swets and Pickett (1982) and references therein.

An alternative definition is $R(p)=1\text{-}F_1\{F_2^{-1}(1\text{-}p)\}$ for $0 \le p \le 1$, where F_2^{-1} denotes the inverse function of F_2. For estimation of this ROC curve, there are several approaches, such as fully parametrically (Goddard and Hinberg, 1990), fully nonparametrically, either via empirical distribution functions (Hsieh and Turnbull, 1996), or using kernel estimators (Lloyd, 1998). For a semiparametric approach we refer to Li, Tiwari and Wells (1999). We focus on smoothed empirical likelihood estimation.

2. Empirical likelihood estimation

Denote by θ_0 the parameter to be estimated, more specifically, $\theta_0 = R(t_0) = 1\text{-}F_1\{F_2^{-1}(1\text{-}t_0)\}$, where $0 \le t_0 \le 1$. Clearly, there exists a quantity η_0 such that $\eta_0 = F_1^{-1}(1\text{-}\theta_0) = F_2^{-1}(1\text{-}t_0)$. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be random samples from populations with (unknown) distribution functions F_1 and F_2, respectively; denote by $p = (p_1, \ldots, p_n)$ and $q=(q_1, \ldots, q_m)$ two probability vectors (that is, $\sum p_i = \sum q_j = 1$ and $p_i, q_j \ge 0$). For this setting, the smoothed empirical likelihood for θ, evaluated at θ_0, is defined as

$$L(\theta_0) = \sup_{(p,q,\eta)} \left(\prod_{i=1}^{n} p_i \right) \left(\prod_{j=1}^{m} q_j \right)$$

subject to the constraints

$$\sum_{j=1}^{m} q_j G_2\left(\frac{\eta - Y_j}{h_2}\right) = 1 - t_0, \qquad \text{and} \qquad \sum_{i=1}^{n} p_j G_1\left(\frac{\eta - X_i}{h_1}\right) = 1 - \theta_0,$$ where, for j=1, 2,

$G_j(t)= \int_{-\infty}^{t} K_j$ (u) du, for a kernel function K_j and where h_j is a bandwidth sequence.

To better appreciate this definition, consider the first constraint which is used in the non-disease population only, to estimate the quantile $\eta_0 = F_2^{-1}(1\text{-}t_0)$ via empirical likelihood. For ROC estimation, however, we add the second constraint, which determines the estimation of the $1\text{-}\theta_0$ quantile in the disease population. A challenging aspect of this approach is that the parameter of interest is not this quantile, but instead $R(t_0)=\theta_0$. The link between both populations, disease and non-disease is through the receiver operator characteristic relation.

Using the method of Lagrange multipliers, the empirical log likelihood ratio (multiplied by minus two) is defined as

$$l(\theta_0) = 2\sum_{i=1}^{n} \ln\{1 + \tilde{\lambda}_1 w_1(\tilde{\eta}, X_i)\} + 2\sum_{j=1}^{m} \ln\{1 + \tilde{\lambda}_2 w_2(\tilde{\eta}, Y_j)\},$$

where $(\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{\eta})$ are solutions to the following set of equations

$$\sum_{i=1}^{n} \frac{w_1(\eta, X_i)}{1 + \lambda_1 w_1(\eta, X_i)} = 0, \quad \sum_{j=1}^{m} \frac{w_2(\eta, Y_j)}{1 + \lambda_2 w_2(\eta, Y_j)} = 0,$$

$$\lambda_1 \sum_{i=1}^{n} \frac{w_1'(\eta, X_i)}{1 + \lambda_1 w_1(\eta, X_i)} + \lambda_2 \sum_{j=1}^{m} \frac{w_2'(\eta, Y_j)}{1 + \lambda_2 w_2(\eta, Y_j)} = 0$$

where $w_1(\eta, X_i) = G_1\left(\frac{\eta - X_i}{h_1}\right) - (1 - \theta_0)$ and $w_2(\eta, Y_j) = G_2\left(\frac{\eta - Y_j}{h_2}\right) - (1 - t_0)$.

The maximum smoothed empirical likelihood estimator for the ROC value at t_0 is that value θ for which the smoothed empirical likelihood function is maximized. Under some regularity conditions on the population distribution functions, the kernels, bandwidth sequences, and the rate at which the two sample sizes, n and m, are allowed to grow, we obtain existence, strong consistency and asymptotic normality of the estimator.

As an illustration, this estimation method will be applied to a medical dataset.

3. Pointwise confidence interval

One natural way to construct a confidence interval for $R(t_0)$ is via the normal approximation to the distribution of its estimator. This, however, can be improved upon in a number of ways. There are several advantages in employing empirical likelihood methods (Owen, 1988) for the construction of confidence intervals: the shape of the confidence region is determined automatically by the sample, as opposed to the symmetry imposed by a normal method, and the empirical likelihood regions are Bartlett correctable; see for example Hall and La Scala (1990), or DiCiccio, Hall and Romano (1991). Another advantage, especially for application to the ROC setting, is that the method of empirical likelihood avoids searching for transformations which

would result in less skewness of the estimator's distribution, in order for the normal approximation to perform better. See Zou, Hall and Shapiro (1997), who apply a logit transformation to estimators of both $1\text{-}F_1$ and $1\text{-}F_2$.

The definition of the ROC curve by means of $R(t)$ instead of as a graph of sensitivity versus 1- specificity, combines information about both populations, and hence avoids the philosophical issue of constructing a confidence interval vertically (for $1\text{-}F_1(p)$), horizontally (for $1\text{-}F_2(p)$), or perhaps a two-dimensional confidence region for $(1\text{-}F_1(p), 1\text{-}F_2(p))$.

Chen and Hall (1993) introduce the method of smoothed empirical likelihood for the construction of a confidence interval for quantiles. We adapt this method for application to the ROC curve, which is a combination of a distribution function and a quantile function.

Under some typical regularity conditions we show that the log empirical likelihood ratio statistic $l(\theta_0)$ converges in distribution to a chi-squared distribution with one degree of freedom. This is a nonparametric version of Wilks's theorem in the ROC context.

Since the limit distribution of $l(\theta_0)$ is asymptotically pivotal, a simple approach to construct a 1-α confidence interval for $\theta_0 = R(t_0)$, is via

$$I_{1-\alpha}=\{\theta: l(\theta)\leq c_{1-\alpha}\},$$

where $c_{1-\alpha}$ is the 1-α quantile of the chi-squared distribution with one degree of freedom. This approach will give a confidence interval for the ROC value $R(t_0)$, with asymptotically correct coverage probability. That is, $P(\theta_0 \in I_{1-\alpha}) = 1\text{-}\alpha + o(1)$.

A topic of current research is to derive the coverage error of this pointwise confidence interval, as well as for a confidence interval based on the normal approximation. Our conjecture is that for the empirical likelihood method, this coverage error is $o(n^{-1/2}+m^{-1/2})$, while for the normal approximation this would be $O(n^{-1/2}+m^{-1/2})$. The superiority of the smoothed empirical likelihood method over the normal approximation is clearly demonstrated by the results of a simulation study, see also Section 5.

4. Bootstrap confidence regions

In order to compare two diagnostic tests via their ROC curve, pointwise confidence intervals are not optimal. Instead of developing asymptotic properties for global curve estimation, which would be one way to construct a confidence region for the ROC curve, a bootstrap algorithm will provide a confidence region for $\{R(t): 0 \leq t \leq 1\}$. In the context of empirical likelihood estimation for density functions, Hall and Owen (1993) explain that very large sample sizes are needed to get accurate confidence regions based on an infinite-parameter verion of Wilks's theorem. In the ROC context, if anything, this is not expected to be better. Therefore, a bootstrap confidence region is constructed in such a way that there is equal pointwise coverage, without loosing advantages of automated shape-determination by the empirical likelihood method.

5. Results of a simulation study

A Monte Carlo study has been performed to compare the coverage accuracy of confidence intervals obtained by the smoothed empirical likelihood method and of those based on a normal approximation. We generated 10 000 pseudorandom samples of various sizes from $F_1=N(1,1)$ and $F_2=N(0,1)$. The two distribution functions G_1 and

G_2 have been chosen so that the corresponding kernel functions are $15/16(1-t^2)^2 1\{|t| \le 1\}$. Currently, bandwidths are chosen fixed throughout the simulations.

Table 1 shows simulated coverage probabilities and demonstrates that the smoothed empirical likelihood method outperforms the normal approximation method in all settings. For all the settings (except for the case t_0=0.9), the simulated coverage probabilities of the empirical likelihood method are all very close to the nominal level, while there is a sizable discrepancy between the simulated coverage probability and the nominal level in the normal approximation method, even for sample sizes as large as 200. The relatively poor performance for the case $t_0 = 0.9$ might be explained by the fact that the receiver operator characteristic value $R(0.9)$ is very close to 99%. Therefore, neither method can be expected to work very well in this case, although the empirical likelihood still outperforms the normal approximation method.

	Method	t_0=0.1	t_0=0.3	t_0=0.5	t_0=0.7	t_0=0.9
n=m=50	Emp.Lik	0.903	0.900	0.896	0.899	0.717
	N.Approx	0.844	0.858	0.868	0.849	0.569
n=m=100	Emp.Lik	0.898	0.897	0.893	0.896	0.811
	N.Approx	0.855	0.868	0.859	0.866	0.787
n=m=200	Emp.Lik	0.895	0.889	0.889	0.892	0.871
	N.Approx	0.854	0.856	0.863	0.863	0.795
n=50, m=100	Emp.Lik	0.880	0.878	0.875	0.878	0.670
	N.Approx	0.848	0.850	0.855	0.853	0.569
n=100, m=50	Emp.Lik	0.880	0.876	0.880	0.881	0.822
	N.Approx	0.832	0.846	0.856	0.823	0.762

Table 1. Simulated coverage accuracy, nominal level = 0.90.

References

Chen, S. and Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *Ann. Statist.*, **21**, 1166 - 1181.

DiCiccio, T.J., Hall, P. and Romano, J.P. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.*, **19**, 1053 - 1061.

Goddard, M. J. and Hinberg, I. (1990). Receiver operating characteristics (ROC) curves and non normal data: an empirical study. *Statist. Med.*, **9**, 325 - 337.

Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *Internat. Statist. Rev.*, **58**, 109 - 127.

Hall, P. and Owen, A.B. (1993). Empirical likelihood confidence bands in density estimation. *J. Comput. Graph. Statist.*, **2**, 273-289.

Hsieh, F. and Turnbull, B.W. (1996). Non-parametric and semi-parametric estimation of the receiver operating characteristic curve. *Ann. Statist.*, **24**, 25 - 40.

Li, G., Tiwari, R.C. and Wells, M.T. (1999). Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curves. *Biometrika*, **86**, 487 - 502.

Lloyd, C.J. (1998). The use of smoothed ROC curves to summarize and compare diagnostic systems. *J. Amer. Statist. Assoc.*, **93**, 1356 - 1364.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237 - 249.

Swets, I.A. and Pickett, R.M. (1982). Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. Academic Press. New York.

Zou, K.H., Hall, W.J. and Shapiro, D.E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statist. Med.*, **16**, 2143 - 2156.

REVISTA DE ESTATÍSTICA
195ª PÁGINA
2º QUADRIMESTRE DE 2001

Combining Information for Semiparametric Density Estimation

Konstantinos Fokianos
University of Cyprus
Department of Mathematics & Statistics
P.O. Box 20537
Nicosia 1678, CYPRUS
fokianos@ucy.ac.cy

1. On the Density Ratio Model

The density ratio model is introduced by the multinomial logits model. Suppose that y denotes a categorical random variable with m categories, and x' is a $p-$ dimensional vector of covariates. Then the multinomial logits model is given by (see Fahrmeir & Tutz (2000))

$$\mathbf{(1)} \qquad \mathrm{P}[y=i\,|\,x]=\frac{\exp\left(\alpha_i^*+x'\beta_i\right)}{\sum_{k=1}^{m}\exp\left(\alpha_k^*+x'\beta_k\right)},\quad i=1,\ldots,m.$$

Let $\pi_i=\mathrm{P}[y=i]$ for $i=1,\ldots,m$ and suppose that there are m independent retrospective samples of sizes $n_1,\ldots,n_m$ acquired from the population with $y=i$, $i=1,\ldots,m$, respectively. Denote the observed data by x_{ij}, $j=1,\ldots,n_i$, i=1,...,m and assume that the conditional distribution of x given y has density $g_i(x)\equiv g(x\,|\,y=i)=dG_i(x)$, $i=1,\ldots,m$. A straightforward application of Bayes theorem reveals that

$$g(x\,|\,y=i)=\frac{\mathrm{P}[y=i\,|\,x]f(x)}{\pi_i}.$$

Then, the density ratio model is given by (Qin (1998), Fokianos et al. (2001))

$$\mathbf{(2)} \qquad \frac{g(x\,|\,y=i)}{g(x\,|\,y=m)}=\frac{\pi_m}{\pi_i}\exp\left(\alpha_i^*+x'\beta_i\right)=\exp\left(\alpha_i+x'\beta_i\right),\ i=1,\ldots,m-1$$

where $\alpha_i=\alpha_i^*+\log(\pi_m/\pi_i)$ for $i=1,\ldots,m-1$. Clearly when $\beta_i=0$ then $\alpha_i=0$. In other words, all the m density functions are assumed unknown related however through an exponential tilt–or distortion–which determines the difference between them. Notice that model (2) is quite general and includes examples such as the exponential and partial exponential families of distributions.

We study a generalization of (2) given by the following model

$$\mathbf{(3)} \qquad g_i(x)=w(x,\theta_i)g_n(x)\ i=1,\ldots,q$$

where the densities $g_i(x), i=1,\ldots,q$ are not specified and θ_i is a finite dimensional parameter with dimension equal to $d_i, i=1,\ldots,q$. We assume throughout that w is a known positive function.

2. Inference

Consider m samples with corresponding densities satisfying equations (3). That is, consider $q=m-1$ weight functions $w(x,\theta_i)$, known up to a parameter, let $n=\sum_{i=1}^{m} n_i$ and consider the non–parametric likelihood based on the pooled data $\{x_{ij}, j=1,\ldots,n_i,\ i=1,\ldots,m\}$

$$\text{(4)}\qquad \begin{aligned} L(\theta,G_m) &= \left[\prod_{j=1}^{n_1} p_{1j} w(x_{ij},\theta_1)\right]\left[\prod_{j=1}^{n_2} p_{2j} w(x_{ij},\theta_2)\right]\cdots\left[\prod_{j=1}^{n_m} p_{mj}\right] \\ &= \left[\prod_{i=1}^{m}\prod_{j=1}^{n_i} p_{ij}\right]\left[\prod_{i=1}^{q}\prod_{j=1}^{n_i} w(x_{ij},\theta_i)\right] \end{aligned}$$

with $p_{ij}=dG_m(x_{ij})$ and $\theta=(\theta_1',\ldots,\theta_q')'$, a vector of dimension $d=\sum_{i=1}^{q} d_i$. The log–likelihood is

$$\text{(5)}\qquad l=\log L=\sum_{i=1}^{m}\sum_{j=1}^{n_i} p_{ij}+\sum_{i=1}^{q}\sum_{j=1}^{n_i}\log w(x_{ij},\theta_i).$$

Maximization of (5) is carried out by following a profiling procedure whereby first we express each p_{ij} in terms of some finite dimensional parameters and then we substitute them back into the likelihood to produce a parametric function. It turns out that (5) becomes

$$\text{(6)}\qquad l(\theta,\mu)=-\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log\left[1+\sum_{k=1}^{q}\mu_k\left(w(x_{ij},\theta_k)-1\right)\right]+\sum_{i=1}^{q}\sum_{k=1}^{n_i}\log w(x_{ij},\theta_i)-n\log n,$$

where $\mu=(\mu_1,\ldots,\mu_k)'$ a vector of Lagrange multipliers. Put $\hat{\theta}=(\hat{\theta}_1',\ldots,\hat{\theta}_q')'$ and $\hat{\mu}=(\hat{\mu}_1,\ldots,\hat{\mu}_q)'$ for the maximum likelihood estimates of θ and μ respectively. In addition let

$$\hat{p}_{ij}=\frac{1}{n}\frac{1}{1+\sum_{k=1}^{q}\hat{\mu}_k\left[w(x_{ij};\hat{\theta}_k)-1\right]}$$

to obtain the maximum likelihood estimator of G_l

$$\text{(7)}\qquad \hat{G}_l(x)=\sum_{i=1}^{m}\sum_{k=1}^{n_i}\hat{p}_{ij} w(x_{ij},\hat{\theta}_l) I(x_{ij}\le x)$$

(8)
$$= \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{w\left(x_{ij},\hat{\theta}_l\right)}{1+\sum_{k=1}^{q}\hat{\mu}_k\left[w\left(x_{ij},\hat{\theta}_k\right)-1\right]} I\left(x_{ij}\le x\right),$$

for $l=1,\ldots,m$ and I the indicator function. It can be shown that

$$\sqrt{n}\begin{pmatrix}\hat{\theta}-\theta\\ \hat{\mu}-\zeta\end{pmatrix}\to N(\mathbf{0},\mathbf{W})$$

in distribution, as $n\to\infty$, with $\zeta=\left(\zeta_1,\ldots,\zeta_q\right)'=\left(n_1/\tilde{n},\ldots,n_q/n\right)'$, under certain conditions.

We turn now to the question of density estimation based by smoothing the increments of $\hat{G}_i$, $i=1,\ldots,m$.

3. Main Results

Set $\rho_i=n_i/n_m$, $i=1,\ldots,m$ and $w(x,\theta_i)=w_i(x)$ for $i=1,\ldots,m$. In particular, $\rho_m\equiv 1$ and $w_m(x)\equiv 1$. In what follows, we consider only univariate measurements. Smoothing the increments of $\hat{G}_l$, for all l, amounts to the following estimators

(9)
$$\hat{g}_l(x)=\frac{1}{h_n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{p}_{ij}w_l\left(x_{ij}\right)K\left(\frac{x-x_{ij}}{h_n}\right),\ l=1,\ldots,m,$$

where h_n is a sequence of window widths such that $h_n\to 0$ as $n\to\infty$ and K is a kernel function. It turns out that under some regularity conditions

1. $\mathrm{AMISE}\left[\hat{g}_l(x)\right]=\frac{1}{4}h_n^4k_2^2\int\left(g_l''(x)\right)^2dx+\frac{1}{n_lh_n}\int\frac{\rho_lw_l(x)g_l(x)}{\sum_{k=1}^{m}\rho_kw_k(x)}dx\int K^2(t)dt.$
2. The asymptotically optimal bandwidth–which is found by minimizing $\mathrm{AMISE}\left[\hat{g}_l(x)\right]$–is equal to
$$h_n^*=\left(\int\frac{\rho_lw_l(x)g_l(x)}{\sum_{k=1}^{m}\rho_kw_k(x)}dx\right)^{1/5}\left(\int K^2(t)dt\right)^{1/5}\left(\int\left(g_l''(x)\right)^2dx\right)^{-1/5}k_2^{2/5}\zeta_l^{-1/5}n^{-1/5}.$$
3. Assigning h_n^* from the above expression, we obtain that the asymptotic mean integrated square error is equal to
$$\frac{5}{4}\left(\int\frac{\rho_lw_l(x)g_l(x)}{\sum_{k=1}^{m}\rho_lw_l(x)}dx\right)^{4}\left(\int K^2(t)dt\right)^{4/5}\left(\int\left(g_l''(x)\right)^2dx\right)^{1/5}k_2^{2/5}\zeta_l^{-4/5}n^{-4/5}.$$

Notice that the new density estimator reduces the asymptotic mean integrated square error (AMISE) when it is compared with that of of traditional density estimator (see Wand & Jones (1995)). Indeed, recall that

(10)
$$\mathrm{AMISE}\left[\overline{g}_l(x)\right]=\frac{1}{4}h_n^4k_2^2\int\left(g_l''(x)\right)^2dx+\frac{1}{n_lh_n}\int K^2(t)dt,$$

with $\bar{g}_l(x) = 1/n_l h_n \sum_{j=1}^{n_l} K\left(\frac{x-x_j}{h_n}\right)$. It follows that $\text{AMISE}\left[\hat{g}_l(x)\right] \leq \text{AMISE}\left[\bar{g}_l(x)\right]$. Thus, the proposed density estimator has less asymptotic mean integrated square error. Equality is obtained if $n_i = 0$ for $i \neq l$, that is only the l'th sample is available.

To choose the smoothing parameter we take an empirical approach by setting

$$\textbf{(11)} \quad \hat{h}_n = \left(\int \frac{\rho_l \hat{w}_l(x) \hat{g}_l(x)}{\sum_{k=1}^{m} \rho_k w_k(x)} dx \right)^{1/5} \left(\int K^2(t) dt \right)^{1/5} \left(\int \left(\hat{g}_l''(x)\right)^2 dx \right)^{-1/5} k_2^{2/5} \zeta_l^{-1/5} n^{-1/5}.$$

It is well known that we can either use an initial guess to get an estimate of $\left(\int \left(\hat{g}_l''(x)\right)^2 dx \right)^{-1/5}$ and then substitute back in (11) to obtain h_n or to iterate this scheme further until convergence.

4 Applications

We illustrate the new methodology to real data consisting of results from an experiment in visual perception using random dot sterograms. The subject observes two images which appear to be composed entirely of random dots. However, they are constructed so that a 3D image will be seen, if the images are viewed with a stereo viewer, causing the separate images to fuse.

An experiment was performed to determine whether knowledge of the form of the embedded image affected the time required for subjects to fuse the images. One group of subjects (group NV) received either no information or just verbal information about the shape of the embedded object. A second group (group VV) received both verbal information and visual information (e.g., a drawing of the object). The scientific question is whether there are differences between the mean time required to fuse the images for the two groups. Previous analysis do indicate that there are significant differences after log–transforming the data. We show that our approach identifies differences between the mean times without transforming the data. In addition we see show that the new semiparametric density estimator for the VV group reduces the AMISE by almost a factor of two when compared with the traditional density estimator while yhe corresponding semiparametric density estimator for the NV group reduces the AMISE by a factor of 1.5.

References

Fahrmeir, L. and Tutz, G. (2000). *Multivariate Statistical Modeling Based on Generalized Linear Models*, Second edition. Springer-Verlag. New York.

Fokianos, K., Kedem, B. Qin, J. and Short, D. (2001). A Semiparametric approach to the One–Way Layout. *Technometrics*, **43**, 56–64.

Qin, J.(1998). Inference for case–control and semiparametric two–sample density ratio model. *Biometrika*, **85**, 619–630.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall. London.

REVISTA DE ESTATÍSTICA
199ª PÁGINA
2º QUADRIMESTRE DE 2001

Online Detection of Trends in Time Series

Roland Fried
University of Dortmund
Department of Statistics
Vogelpothsweg 87
44221 Dortmund, Germany
fried@statistik.uni-dortmund.de

1. Introduction

Modern technology allows for high-frequency sampling of statistical data. In intensive care, e.g., the vital signs of critically ill patients are measured in short time intervals. A crucial task is the automatic, fast and proper identification of patterns of change in the observed time series.

Control charts suggested in statistical process control typically rely on the stationarity of the model parameters and on the existence of a target value for the observations. However, in intensive care the data generating processes are neither stationary nor can a single target value be specified in advance (Högel, 2000). Moreover, monitoring schemes are usually designed to detect a single sudden change, although in many cases the process deteriorates gradually or in several little steps (Chang and Fricker, 1999). Such slow monotone trends do not only influence the process mean, but they also have a large impact on the estimates of the model parameters. Simulation studies show that the properties of the monitoring schemes strongly depend on these estimates. Thus, there is also a loss in power versus sudden changes if there are undetected trends in the time series.

In Section 2 we modify Brillinger's (1989) approach for trend detection so that it is suitable for online monitoring. In Section 3 we present the results from a simulation study and conclude with a discussion of some possible improvements.

2. Online Detection of Trends

Monitoring schemes for autocorrelated data are often based on autoregressive models since these form a quite flexible model class and are rather easy to handle. We assume local stationarity of the time series and move a time window of length n through the series. The basic model for the observations belonging to the current time window reads

$$Y_t = S_t + E_t \ , \ t = 1,\ldots,n$$
$$E_t = \phi_1 E_{t-1} + \ldots + \phi_p E_{t-p} + U_t$$

That is, the observations at time t consists of a deterministic signal we are interested in and additive random noise from an AR(p)-process with unknown innovation variance and autoregressive coefficients. We assume that the parameters of the AR(p)-model change slowly over time in comparison to the window width n (compare Belitser, 2000).

For trend detection often a parametric form like

$$S_t = a \cdot t + b$$

is assumed and it is tested whether the slope is significantly different from zero. However, this means to specify a fixed form of the signal and in consequence non-linear trends might not be detected this way. On the other hand, control charts such as EWMA and CUSUM charts are often based on a weighted sum of the observations.

However, the power of these charts is usually best for sudden shifts of the mean. Another approach to trend detection is to mirror the start and the end of the series (i.e., of the current time window in our context) by comparing time delayed means calculated from the first and last m observations, m<n/2 . This is equivalent to using a weighted sum

$$\sum_{t=1}^{n} c_t Y_t = -\frac{1}{m} Y_1 - \cdots - \frac{1}{m} Y_m + \frac{1}{m} Y_{n-m+1} + \cdots + \frac{1}{m} Y_n$$

with weights summing up to zero. Values of such a weighted sum which are far away from zero indicate a monotone change during the time interval considered.

We follow Abelson and Tukey (1963) and Brillinger (1989) who apply a mini-max approach to get weights for which the worst-case discriminatory power for an extremely unfavourable monotone trend is as high as possible. This results in

$$c_t = \sqrt{(t-1)\left(1-\frac{t-1}{n}\right)} - \sqrt{t\left(1-\frac{t}{n}\right)}$$

and the corresponding worst case is a single step change. A monotone trend is detected if the absolute value of the weighted sum is too large in comparison to the variance

$$Var\left(\sum_{t=1}^{n} c_t Y_t\right) = \sum_{s,t=1}^{n} c_s c_t \gamma(t-s)$$

Hence, we need reliable estimates of the autocovariances

$$\gamma(0),\ldots,\gamma(n-1)$$

or, equivalently, of the parameters of the AR(p)-model. Since a trend has a serious impact on the usual sample autocovariances, we should try to reduce the influence of trend patterns in a simple way. For this reason, Brillinger (1989) replaces the overall mean by a running mean when analysing a long time series retrospectively. However, this is not a good solution if only a small or moderate number of observations can be used. Instead, we suggest to fit a parametric model as mentioned above using simple least squares estimates in a first step. Then we estimate the autocovariances up to the time lag p from the residuals using

$$\hat{\gamma}(h) = \frac{1}{n-2} \sum_{t=1}^{t-h} \left(Y_t - \hat{a}t - \hat{b}\right)\left(Y_{t+h} - \hat{a}(t+h) - \hat{b}\right)$$

and estimate the parameters of the AR(p)-model thereafter via the Yule-Walker equations. The denominator (n-2) helps to reduce the negative bias of the sample autocovariances (the means are estimated using two parameters).

3. Simulations

Using large sample asymptotics, Brillinger (1989) suggests to compare the standardized weighted sum to the standard Gaussian distribution. Simulations show

that this rule is too sensitive even in a retrospective application to a long time series (Woodward et al., 1997). Moreover, in online monitoring we apply multiple testing since we check at each time point whether a trend has occurred during the last n observations. Moreover, n should not be chosen too large since stationarity is judged to be only a local approximation. Thus, we want to find suitable critical values c and check the power of the procedure via some simulations.

Most monitoring schemes for autocorrelated data apply simple AR(1)-models. We simulate 200 time series of length N=300 and with innovation variance set to one for several autoregressive coefficients and several deterministic signals. Linear trends of length 50 or 100 starting at time point t=100 with slopes a=0, 0.05, 0.10, 0.15, 0.20, 0.25, as well as sinus-shaped trends causing the same total change in level as the linear trends with the same duration are considered. For the time window we choose n=60 observations corresponding to one hour of measurements observed in one minute intervals.

A crucial point is the estimation of the model parameters. In our simulations, we find the bias and the mean squared error MSE of all estimators to be small in case of small or moderate autocorrelations. For very strong positive autocorrelations the estimates become negatively biased and the MSE increases.

With respect to critical values, we consider c=5 to be a good choice if we want to restrict the probability of a false alarm within 300 subsequent observations to less than 5 %. We find this to be a conservative bound in case of small autocorrelations, while in case of very strong positive autocorrelations we might need a somewhat larger value. Application of normal theory for the asymptotic distribution would result in a smaller value.

Trends which are not very slow can be detected reliably using c=5 if the autocorrelations are small or moderate. For instance, the simulated power is larger than 80 % for all autocorrelations considered if a linear trend with a slope of 10 % occurs during 50 minutes. The detection of sinus-shaped trends is somewhat more difficult. This might be caused by the smooth beginning of this trend form, which has zero derivative at its endings.

The time needed for trend detection does not depend strongly on the slope, but increases with the autocorrelations. On the average, between 30 and 45 observations are needed to detect a trend with a slope of at least 10%. Strong positive correlations result in monotone sequences in the time series just like deterministic trends. Therefore, these mechanisms are hard to distinguish within short time series anyway (Woodward and Gray, 1993).

4. Conclusion

We have proposed a procedure for online detection of monotone deterministic trends in time series. Since reliable estimates of the autocovariances of the noise process during trend periods are needed, we have used a simple variation of the sample autocovariances which helps to improve the estimation of the variance of the test statistic during trend periods. So far we can say that the results of our simulations are encouraging.

The procedure was also applied to time series with several thousand observations of the vital signs of a critically ill patient. The procedure identified almost every trend pattern detected by an experienced senior physician. Approximately 40 % additional trend patterns were detected by the procedure which were not found to be important by the physician. These `false alarms' were caused by

a soft monotone change during the time window which came along with very small random fluctuations resulting in a small variance of the test statistic. Therefore the procedure was more sensitive to small systematic changes in the mean than the human expert. Such application-dependent problems could be overcome by using a lower limit for the variance based on existing knowledge or on an analysis of past data.

There are several possibilities for further improvements. Robust estimates of the autocovariances can reduce the impact of outliers on the classifications, and small sample corrections might help to overcome the negative bias of the estimators found in case of large correlations.

Acknowledgements

The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

References

Abelson, R. P. and Tukey, J. W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order, *Ann. Math. Statist.* **34**, 1347-1369

Belitser, E. (2000). Recursive estimation of a drifted autoregressive parameter, *Annals of Statistics* **28**, 860-870.

Brillinger, D. R. (1989). Consistent detection of a monotonic trend superposed by a stationary time series, *Biometrika* **76**, 23-30.

Chang, J.T. and Fricker, R.D. (1999). Detecting When a Monotonically Increasing Mean Has Crossed a Threshold, *J. Quality Technology* **31**, 217-234.

Fried, R. (2001). Online detection of a monotone trend in a time series, Preprint, Department of Statistics, University of Dortmund, Germany.

Högel, J. (2000). Applications of statistical process control techniques in medical field, *Allg. Stat. Archiv* **84**, 337-359.

Woodward, W.A. and Gray, H.L. (1993). Global warming and the problem of testing for trend in time series data, *J. of Climate* **6**, 953-962.

Woodward, W.A., Bottone, S. and Gray, H.L. (1997). Improved tests for trends in time series data, J. Agricultural, *Biological and Environmental Statistics* **2**, 403-416.

Testing Equality of Cause-Specific Hazard Rates Corresponding to M Competing Risks Among K Groups

Sangita Kulathinal
National Public Health Institute
Helsinki
sangita.kulathinal@ktl.fi

Dario Gasbarra
Rolf Nevanlinna Institute
University of Helsinki
dog@rni.helsinki.fi

1. Introduction

Consider a situation where m competing risks are acting simultaneously in the same environment. Suppose there are K independent groups of unit, and each unit is exposed to m competing risks. Let T_{ki} be the failure time and $\delta_{ki} \in \{0,1,2,\ldots,m\}$ be the cause of failure, $i = 1,2,\ldots,n_k$ and $k = 0,1,2,\ldots,K$ (0 corresponds to censoring). For each k, the pairs (T_{ki},δ_{ki}), $i = 1,2,\ldots,n_k$, are independent and identically distributed. We assume noninformative censoring and we do not make any regarding the dependence among the m risks.

Define the cumulative incidence function for the risk j in the group k by

$$F_{jk}(t) = P[T_{ki} \le t, \delta_{ki} = j], \tag{1}$$

which are assumed to be continuous with subdensities $f_{kj}(t)$. Also define the causespecific hazard rate by $h_{kj}(t) = \frac{f_{kj}(t)}{S_k(t)}$, where $S_k(t) = P[T_{kj} > t] = 1 - \sum_{j=0}^{m} F_{kj}(t)$ is an overall survival function of the group k.

The main purpose of this pager is to develop a test procedure for the hypothesis $H_0 : h_{1j}(t) = h_{2j}(t) = \ldots = h_{Kj}(t) = h_j(t)$, $\forall j = 1,2,\ldots,m$, where $h_j(\cdot), j = 1,2,\ldots,m$ are unspecified common cause-specific hazard rates.

The problem of testing equality of cause-specific hazard rates corresponding to m dependent risks has been discussed in the literature (see Lam, 1338, Aly *et al.*, 1994 and references therein). Gray (1988) gives a class of K-sample tests for comparing the cumulative incidence functions of a particular type of failure out of several competing risks among different groups. Lindkvist *et al.* (1998) propose a class of tests based on a two-dimesional vector statistic for testing equality of cumulative cause-specific hazard rates corresponding to two risks between two samples. Their approach can be easily applied to m risks.

The test is applied to the analysis of contraceptive failure data in intrauterine device (IUD) collected in several countries.

2. Test Statistic and Asymptotic Distribution

Let $n = \sum_{k=1}^{K} n_k$. Define the counting process $N_{kj}(t) = \sum_{i=1}^{n_k} I[T_{ki} \le t, \delta_{ki} = j]$, $Y_k(t) = \sum_{i=1}^{n_k} I[T_{ki} \ge t]$ and $M_{kj}(t) = N_{kj}(t) - \int_0^t Y_k(s)\, d\Lambda_{kj}(s)$, $j = 0,1,2,\ldots,m$, $k = 1,2,\ldots,K$. Then for $t \in [0,\tau]$, $M_{kj}(t)$'s are orthogonal square integrable

martingales with respect to the filtration $\{\mathcal{F}_t^{N,Y}\}$ which is generated by N_{kj} and Y_k. Define the overall counting process $N_{\cdot j}(t)=\sum_{k=1}^{K}N_{kj}(t)$, $Y_{\cdot}(t)=\sum_{k=1}^{K}Y_k(t)$, and the martingale $M_{\cdot j}(t)=\sum_{k=1}^{K}M_{kj}(t)$. Note that $\langle M_{kj},M_{k'j'}\rangle=0$, for $(k,j)\neq(k',j')$, and $\langle M_{\cdot j},M_{k'j'}\rangle=\delta_{jj'}\int_0^{\cdot}Y_{k'}(s)d\Lambda_{j'}(s)$.

A K-sample test can be based on the scores for $j=1,2,\ldots,m$, $k=1,2,\ldots,K$

$$Z_{ki}=\int_0^{\tau_k}\sum_{j=1}^{m}K_{kij}^n(t)\{d\hat{\Lambda}_{kj}(t)-d\hat{\Lambda}_{\cdot j}(t)\}, \tag{2}$$

where $K_{kij}^n(t)$ are suitably chosen locally bounded $\{\mathcal{F}_t^{N,Y}\}$-predictable processes. $d\hat{\Lambda}_{kj}(t)=dN_{kj}(t)/Y_k(t)$, and $d\hat{\Lambda}_{\cdot j}(t)=dN_{\cdot j}(t)/Y_{\cdot}(t)$. When $K_{kij}^n(t)=\delta_{ij}K_{ki}^n(t)$ then (2) simplifies to

$$Z_{ki}=\int_0^{\tau_k}K_{ki}^n(t)\{d\hat{\Lambda}_{ki}(t)-d\hat{\Lambda}_{\cdot i}(t)\}, \tag{3}$$

which can be shown as a generalisation of the test proposed by Lindkvist *et al.* (1998). The martingale central limit theorem is applied to derive the asymptotic distribution of Z (Andersen *et al.* 1993).

Theorem 1

Assume $n^{-1}Y_k^n(s)\to y_k(s)$ uniformly in probability, where $y_k(s)$ are deterministic functions. Denote $y_{\cdot}(s)=\sum_{k=1}^{K}y_k(s)$. Then, under the assumption that $n^{-1}K_{kij}^n(t)\to K_{kij}^0(t)$ uniformly in probability, with each $K_{kij}^0(\cdot)$ bounded on $[0,\tau]$ and under the null hypothesis, as $n\to\infty$, $n^{-1/2}Z$ converges in distribution to $N_{m\times K}(\underline{0},\Sigma)$, where

$$\Sigma_{(i,k),(i',k')}=\int_0^{\min(\tau_k,\tau_{k'})}\left(\frac{\delta_{kk'}}{y_k(t-)}-\frac{1}{y_{\cdot}(t-)}\right)\sum_{j=1}^{m}K_{kij}^0(t)K_{k'i'j'}^0(t)d\Lambda_j(t). \tag{4}$$

A consistent estimator of $\Sigma_{(i,k),(i',k')}$ is given by

$$n^{-1}\int_0^{\min(\tau_k,\tau_{k'})}\left(\frac{\delta_{kk'}}{Y_k^n(t-)Y_{\cdot}^n(t-)}-\frac{1}{Y_{\cdot}^n(t-)^2}\right)\sum_{j=1}^{m}K_{kij}^n(t)K_{k'i'j'}^n(t)dN_{\cdot j}(t),$$

where $\tau=\min(\tau_k,\tau_l)$. When $K_{kij}^n(t)=\delta_{ij}K_{ki}^n(t)$, the asymptotic covariance matrix becomes block-diagonal, $\Sigma=\text{diagonal}(D_1,D_2,\ldots,D_m)$ where $D_i(k,k')=\Sigma_{(i,k),(i,k')}$, $i=1,\ldots,m$, k, $k'=1,\ldots,K$.

Under the alternative, as $n\to\infty$, $n^{-1/2}Z$ converges in distribution to $N_{m\times K}(\underline{\mu},\Sigma)$, where

$$\mu_{ki}=\int_0^{\tau_k}\sum_{j=1}^{m}K_{kij}^0(t)\left[d\Lambda_{kj}(t)-\frac{\sum_{k=1}^{K}y_k(t)d\Lambda_{kj}(t)}{y_{\cdot}(t)}\right],$$

To generate a class of tests of H_0, we take the weight process of the form

(5) $$K_{kij}^{n}(t) = L_{ij}^{n}(t)Y_{k}^{n}(t),\ i,j = 1,2,\ldots,m,\ k = 1,2,\ldots,K.$$

For the weight process of this type $\sum_{k=1}^{K} Z_{ki} = 0$ for each i. Hence in this case rank of Σ is $(K-1)\text{rank}(A)$, where A is the $m \times m$ matrix

$$A(i,j) = \int_0^{\infty} L_{ij}^{0}(t)h_j(t)dt,$$

where $n^{-1}K_{kij}^{n}(t)$ tends to $L_{ij}^{0}(t)y_k(t)$ as $n \to \infty$. In the follow-up we assume that A has full rank m. Under the null hypothesis, the test statistic $n^{-1}Z'\Sigma^{-1}Z$ then has asymptotically chi-squared distribution with $m(K-1)$ degrees of freedom. Under the alternative, the test statistic has asymptotically noncentral chi-squared distribution with $m(K-1)$ degrees of freedom with the noncentrality parameter $\underline{\mu}'\Sigma^{-1}\underline{\mu}$. It is shown that the locally asymptotic efficient nonparametric test belongs to this class for the alternatives of the type

$$h_{kj}^{n}(t) = h_j(t) + a_n^{-1}h_j(t)\sum_{i=1}^{m}\gamma_{ij}(t)\phi_{ki} + o(a_n^{-1}),$$

for $k = 1,2,\ldots,K$ and $j = 1,2,\ldots,m$, and $a_n = n^{1/2}$ throughout. The motivation for this alternative comes from the Gumbel's distribution which is illustrated in the next section. By applying the technique given on pages 615-624 of Andersen *et al.* (1993) it is shown that the nonparametric test is asymptotically equivalent to an efficient parametric test.

When the matrices $(\gamma_{ij}(t))_{t\geq 0}$ can be diagonalized by the same linear trasformation simultaneously for all t, the sequence of local alternatives can be expressed as

$$h_{kj}^{n}(t) = h_j(t) + a_n^{-1}\phi_{kj}(t)\gamma_j(t)h_j(t) + o(a_n^{-1}),\ j = 1,2,\ldots,m;\ k = 1,2,\ldots,K,$$

where ϕ_{kj} are constants and $\gamma_j(t)$ are fixed functions, $j = 1,2,\ldots,m;\ k = 1,2,\ldots,K$.

An attractive process in our case is the type of process suggested by Harrington and Fleming (1982). We will consider the weight process

(6) $$L_j^{n}(t) = \left[1 - \hat{F}_j(t)\right]^{\rho}$$

where ρ is a fixed constant between 0 and 1 and $\hat{F}_j(t)$ is an estimate of the common incidence, function for risk j, $P[U \leq t, \varepsilon = j]$ and is given by

(7) $$\hat{F}_j(t) = \int_0^{t} \hat{S}(u-)\frac{dN_j(u)}{Y(u)},$$

where $\hat{S}(t-)$ is the left-hand limit of the Kaplan-Meier (1958) estimate of the survival function of U.

Example 1

Simulation Study

We consider a m-variate Gumbel exponential distribution, with parameters α and $\lambda = (\lambda_1,\ldots,\lambda_m)$, with the density, $f(x_1,x_2,\ldots,x_m) = \prod_{i=1}^{m}\lambda_i \exp\left(-\sum_{i=1}^{m}\lambda_i x_i\right) \times$

$$\times\left[1 + \alpha\prod_{i=1}^{m}\left(2\exp(-\lambda_i x_i) - 1\right)\right].$$

We will consider $K = 5$ groups and $m = 3$ risks, and the true alight function corresponding to the optimal test and also the weight function given in (6) with

$\rho = 0.5$. The level of significance used throughout is 0.05. The null hypothesis is rejected if the test statistic is greater than 21.026. The parameters used in the simulation are $\alpha = 0.5$ throughout and $\lambda = (0.8, 0.2, 0.6)$ for the null hypothesis. The censoring distribution is taken as exponential for each group with intensities (1, 0.6, 0.7, 0.8, 0.9), respectively. Samples of sizes (50, 60, 70, 80, 90) were generated for five groups with 1000 repetitions. The empirical distributions were observed to be quite close to the true distribution. The empirical level of significance using the optimal weight function is 0.056 while it is 0.06 when the weight function (6) is used.

To check the performance of the test as well as to compare the two weight functions, noncentrality parameters and empirical powers are computed when the parameters are (0.8, 0.2, 0.6), (0.79, 0.19, 0.59), (0.789, 0.199, 0.599), (0.7, 0.1, 0.5), and (0.69, 0.09, 0.49) for the five groups.

The empirical study showed that the Harrington and Fleming (1982) type of weight function can be used in practice since it gives power which is reasonably close to the power of the test when the optimal weight function is used.

Example 2

Application

The data are taken from a five year follow-up study of 1547 women from Finland, Sweden and Hungary, on termination of IUD conducted by a pharmaceutical company based in Finland. The summary of the data is given below:

Termination due to	Finland	Sweden	Hungary
pregnancy	2	0	0
expulsion	31	28	9
amenorrhea	11	7	50
bleeding and pain	60	96	64
hormonal disturbances	46	80	12
censoring	398	430	223
total	548	641	358

The main interest was in testing the equality of the five cause-specific hazards for the three countries. The weight functions used were the same as in (6) with $\rho = 0, 0.5, 1$. The values of the test statistic are 149.38, 149.52, and 149.63 respectively. These values are higher than the cut-off point, 18.3, of the chi-squared distribution with 10 degrees of freedom. Hence the hypothesis of equality of cause-specific hazards is rejected.

References

Aly, E. A. A., Kochar, S. C. and McKeague, I. W. (1994). Some tests for comparing cumulative incidence functions and cause-specific hazard rates, *J. Am. Statis. Assoc.* **89**, 994-999.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N.(1982). Linear nonparametric tests for comparison of counting processes, with application to censored survival data (with discussion). *Int. Statist. Rev.* **50**, 219-258.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical models based on counting processes*, Springer: New York.

Gray, R. L.(,1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *Ann. Statist.* **16**, 1141-1154.

Kalbfleisch, J. D. and Prentice, R. L.(1980). *The statistical analysis of failure time data*, Wiley: New York.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Statis. Assoc.* **53**, 457-481.

Karia, S. R., Toivonen, J. and Arjas, E.(1998). Analysis of contraceptive failure data in interauterine device studies. *Contraception* **58**, 361-374.

Lam, K. F. (1998). A class of tests for the equality of k cause-specific hazard rates in a competing risks model. *Biometrika* **85**, 179-188.

Lindkvist, H. and Belyaev, Y.(1998). A class of non paramteric tests in the competing risks model for comparing two samples. *Scand. J. Statis.* **25**, 143-150.

Ramlau-Haunsen, H.(1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11**, 453-466.

Gaussian Semiparametric Estimation for Random Fields with Long Range Dependence

Domenico Marinucci
University of Rome "La Sapienza", Dipartimento di studi Geoeconomici e Statistici
Via del Castro Laurenziano, 9, 00161 Roma, Italy
marinucc@scec.eco.uniroma1.it

We say a (wide sense homogenous and isotropic) random field is long range dependent if its spectral density has a singularity at the origin, i.e. it has a zero or it diverges to infinity at zero frequency. Random fields with long range dependence are now known to arise in many cases of interest, for instance, in the astrophysical literature, see Peebles (1990). Albeverio et al. (1994) and many subsequent authors have studied the asymptotic behaviour of solutions for nonlinear differential equation in three dimensions, as motivated by physical considerations, and found that long range dependent behaviour typically arises under a great variety of circumstances. Many other stochastic models can produce long range dependent behaviour in random fields; for instance, fractional and non-fractional diffusion-wave equations (Anh et al. (1999)), with applications including wave diffusions in porous media, nonlinear acoustic shock waves and other types of irrotational flow. In all these cases the exact form of the spectral density can be quite complicated or even not yet known, however typically spectral singularities do arise around the origin.

In the time series case, statistical inference and its mathematical foundations in the presence of long range dependence have now been investigated in great detail, under both parametric and semiparametric conditions, see Giraitis and Surgaili (1990), Beran(1994), Robinson(1995a,b), Giraitis and Taqqu(1999), and many others. On the other hand, for the random field case statistical inference procedure have not been developed to the same extent as for time series, the only references being provided by Heyde and Gay(1993), and by Leonenko and Woyczynski(1999), who consider a fully parametric specification over the whole frequency band. No effort has been devoted so far to allow for semiparametric procedures, making assumptions only on an arbitrary small neighbourhood around zero frequency. Nevertheless, data sets which are candidates for long range dependent behaviour (like catalogs of galaxy redshifts in the astronomical context) are often characterized by an extremely high number of observations; hence it can be computationally very hard to implement fully parametric estimates, which are typically not available in closed form and require lengthy iterations. More important, a full-band model entails by necessity a number of assumptions and approximations whose validity can be questioned, while many of them need not be necessary for the analysis of the behaviour of the system at the largest scale; the presence of observational error, moreover, can add a white noise additive component in the spectral density of the observables, so that a full band model may be misspecified, whereas our narrow band specification may still be valid.

It is also important to remark how most physical models are developed for continuous parameter fields, whereas observations are usually available on a discretized lattice; discretization procedures have a complicated nonlinear effect on the spectral density, which is often difficult to pin down exactly, especially as data collection is in many cases beyond the control of the statistician. The most common discretization procedures, however, such as neighborhood smoothing or grid sampling, do not have effects at zero frequency, except at most some rescaling in the constants, and this provides

in our opinion one further reason to favour local-to-zero specifications. The behaviour of the spectral density around the origin is often of considerable interest by itself; for instance, for the asymptotic analysis of geometric functionals of random fields, see Ivanov and Leonenko (1989); the slope of the spectral density at the origin can be used as the benchmark to discriminate between alternative models, such as different inflationary scenarios for the very early Universe.

The purpose of this paper is to develop a semiparametric procedure for statistical inference on the long range dependence parameters, imposing only local-to-zero conditions. Our basic idea is to extend to the random field case the Whittle semiparametric procedure considered for long range dependent time series by Robinson (1995b). As many semiparametric methods, Whittle estimates rely only on the information at the smallest frequencies, and therefore have asymptotic efficiency zero with respect to procedures based on a correctly specified parametric model. In the presence of misspecification of the high-frequency component, however, a parametric model will generally lead to inconsistent estimates, whereas semiparametric procedures have robustness properties that seem desirable. Moreover the loss of asymptotic efficiency seems acceptable in many random fields contexts, where data sets candidate for long range dependent behaviour are often characterized by an extremely high number of observations (e.g. the ongoing Sloan Digital Survey on stellar distribution aims at mapping the position of more than one million galaxies). Finally, from the computational point of view the procedure we advocate seems extremely convenient, requiring minimization of a globally concave univariate function, a task which can be easily accomplished by several well-known optimization routines.

In this paper, we establish first some results of independent interest on the asymptotic behaviour of the discrete Fourier transform of a homogenous and isotropic vector field with long range dependence; this material extends to the random field case analogous results by Robinson (1995a) and Velasco (2000), and we believe it may find applications for other semiparametric inference procedures in the presence of spectral singularities. We then focus more directly on statistical inference; in particular, we analyze Whittle semiparametric estimates, for which we prove consistency and asymptotic Gaussianity. The proofs are rather lengthy and collected separately.

References

Albeverio, S., Molchanov, S.A. and D. Surgailis (1994). Stratified Structure of the Universe and Burgers' Equation–a Probabilistic Approach, *Prob.Th.Rel.Fields*, **100**, 457-484

Anh, V.V., J.M. Angulo and M.D. Ruiz-Medina (1999). Possible Long-Range Dependence in Fractional Random Fields", *J. Statist. Plann. and Inf.*, **80**, 95-110

Beran, J., (1994). Statistics for Long Memory Processes, Chapman and Hall, London

Giraitis, L. and D. Surgailis (1990). A Central Limit Theorem for Quadratic Forms in Strongly Dependent Linear Variables and its Application to Asymptotic Normality of Whittle's Estimates", *Prob.Th.Rel.Fields*, **86**, 87-104

Giraitis, L. and M.S. Taqqu (1999). Whittle Estimator for Finite-Variance Non-Gaussian Time Series with Long Memory", *Ann. Statist.*, **27**, n.1, 178-203

Heyde, C.C. and R. Gay (1993). Smoothed Periodogram Asymptotics and Estimation for Processes and Fields with Possible Long-Range Dependence", *Stoch.Proc.Applic.*, **45**, 169-182

Ivanov, A.V. and Leonenko, N.N. (1989). Statistical Analysis of Random Fields, Kluwer Academic Publishers, Dordrecht

Leonenko, N.N. and W.A. Woyczynski (1999). Parameter Identification for Singular Random Fields Arising in Burgers' Turbulence", *J.Statist.Plann.Inf.*, **80**, 1-14

Peebles, J. (1990). Principles of Physical Cosmology, Princeton University Press

Robinson, P.M. (1995a). Log-Periodogram Regression of Time Series with Long Range Dependence, *Ann.Statist.*, **23**, 1048-1072

Robinson, P.M. (1995b). Gaussian Semiparametric Estimation of Long Range Dependence, *Ann.Statist.*, **23**, 1630-1661

Velasco,C. (2000). Non-Gaussian Log-Periodogram Regression, *Ec.Th.*, **16**, 44-79

On Sums and Extremes of Random Variables

Sandra Mendonça
University of Madeira
Department of Mathematics
Campus Universitário da Penteada
Funchal, Portugal
sandra@uma.pt

1. Introduction

The epistemological value of the theory of probability is revealed only by limit theorems.
Gnedenko and Kolmogorov [1954]

In the history of the Theory of Probability we can find names of great mathematicians who were concerned with the limit behaviour of sums and extremes, like Lévy, Khinchine, Gnedenko, Doeblin, Feller, Fréchet, Fisher, Tippet, Gumbel, Weibull and Mejzler. In fact, it was the study of limit theorems which changed the "Calculus of Probabilities" into "Probability Theory". In this work we present some partial results related to random normalizations.

2. The Quotient Between the Sum and the Maximum

The development of the theory of maxima of independent random variables (r.v.'s) was, in some way, parallel to the theory of sums of independent r.v.'s. Apart this possible parallel, some authors have been tempted to compare them in a more direct way using the quotient between $M_n = \max_{1\le i\le n} X_i$ and $S_n = \sum_{1\le i\le n} X_i$, revealing, in this way, the susceptibility of the sum being influenced by the maximum term. Among these authors we can find, for example, Darling, O'Brien, Maller, Resnick, Teugels and Bingham.

Let us consider $\{X_i\}_{i\in\mathbf{N}}$ a sequence of independent random variables, identically distributed (i.i.d.) as a random variable X, F its distribution function and let $\omega(F)$ be the supremum of the support of X, *i.e.*, $\omega(F) = \sup\{x: F(x) < 1\}$. Suppose that $X \ge 0$. Bingham *et al.*, [1987] have put together some results concerning this subject:

- $M_n / S_n \to 0$ in probability $\Leftrightarrow$ $\int_0^{\infty} y dF(y)$ is of slow variation*;
- $M_n/S_n \to 1$ in probability $\Leftrightarrow 1 - F$ is of slow variation;
- M_n/S_n has a non-degenerate limit distribution $\Leftrightarrow F$ is attracted to a stable law of index $\alpha\in(0,1) \Leftrightarrow \mathrm{E}(S_n/M_n)$ has a finite, grater than one limit;
- If X has a finite mean μ then: $(S_n - n\mu)/M_n$ has a non-degenerate limit distribution $\Leftrightarrow$ F is attracted to a stable law of index $\alpha\in(1, 2) \Leftrightarrow E((S_n - n\mu)/M_n) \to c \in (1,+\infty)$ (and then $\alpha = (1 + c)/c$).

---

* We say that a function g, defined for positive numbers, is of slow variation if $\lim_{x\to+\infty} \frac{g(ax)}{g(x)} = 1$, $\forall a>0$.

3. Random-Normalizations: Partial Results

I keep the subject before me and wait till the first dawnings open slowly, by little and little, into a full and clear light.

Isaac Newton

The partial results here presented *(the first dawnings)* show a new way of looking to classical normalizations (which means, linear normalizations); quoting Resnick, [1986]:

Limit theorems are useless in statistical contexts if they depend on parameters that must be computed from unknown distributions. A way of solving these problems is to replace these parameters by functions of the observations.

Building also an extra parallel between sums and extremes, and inspired by Logan *et al.* [1973] who have studied, for i.i.d. r.v.'s, the statistic

$$\sum_{i=1}^{n} X_i \Big/ \left(\sum_{i=1}^{n} |X_i|^p \right)^{1/p}, \tag{10}$$

and by Darling [1952] who has studied the limit case

$$\sum_{i=1}^{n} X_i \Big/ \max_{1 \le i \le n} |X_i|, \tag{11}$$

we have been studying the statistics

$$W_n(p) = \max_{1 \le i \le n} X_i \Big/ \left(\sum_{i=1}^{n} |X_i|^p \right)^{1/p} \quad \text{and} \quad W_n(+\infty) = \max_{1 \le i \le n} X_i \Big/ \max_{1 \le i \le n} |X_i|,$$

Both variables have support contained in the interval $[-1,1]$. When $X \geq 0$, $W_n(+\infty)=1$, and the statistic is no longer of interest. Also if the support (*supp*) of the variable is bounded, the weak limit of $W_n(+\infty)$ is

$$w - \lim_{n \to +\infty} W_n(+\infty) = \frac{\sup suppX}{\max\left[\left|\inf suppX\right|, \left|\sup suppX\right|\right]}.$$

If the support of X is inferiorly unbounded and superiorly bounded then the weak limit of $W_n(+\infty)$ is also easily calculated.

Some exact and asymptotic properties of $W_n(+\infty)$ and $W_n(p)$ will be now presented.

Proposition 1 If X is absolutely continuous then

- $F_{W_n(+\infty)}(x) = n \int_{-\infty}^{0} \left[F(xt) - F(t) \right]^{n-1} f(t)dt \; I_{(-1,1)}(x) + I_{[1,+\infty)}(x)$.
- If $\omega(F) > 0$ then P [Wn $(+\infty) = 1] > 0$.
- If X *is* simmetric then P [Wn $(+\infty) = 1] = ½$.

Proof The proof can be found in Mendonça [1999].

Proposition 2 Let us suppose the support of X is contained in $[0, +\infty)$. If 1–F is of *slow* variation then $\lim_{n \to +\infty} F_{W_n}(x) = F_1(x)$, where F_1 is the d.f. of the degenerate r.v. concentrated in the value 1.

Proof Since the variables are non-negative we can write

$$F_{W_n(p)}(x)=P\left[\frac{\max_{1\le i\le n} X_i}{\left(\sum_{i=1}^{n}|X_i|^p\right)^{1/p}}\le x\right]=1-P\left[\frac{\sum_{i=1}^{n} Y_i}{\max_{1\le i\le n} Y_i}\le x\right]$$

(where $Y_i = X_i^p$) and use Darling's [1952] results.

Once more, if we use Darling's [1952] results, we can state that

Proposition 3 If $X_i \ge 0$ and if $\sum_{i=0}^{n} X_i^p$, properly normed, converges to a stable distribution, with characteristic exponent $0 < \alpha < 1$, then

$$\lim_{n\to+\infty} F_{W_n(p)}(x)=1-G\left(\frac{1}{x^p}\right)$$

where G is such that

$$\int_0^{+\infty}\exp(ity)\,dG(y)=\frac{\exp(it)}{1-\alpha\int_0^{1}\frac{\exp(itu-1)}{u^{\alpha+1}}du}.$$

Still about $W_n(p)$, observe that, when $\max_{1\le i\le n} X_i = X_{i_0} < 0$,

$$\frac{\max_{1\le i\le n} X_i}{\left(\sum_{i=1}^{n}|X_i|^p\right)^{1/p}}<-\frac{1}{n^{1/p}}\Leftrightarrow -n^{1/p}\max_{1\le i\le n} X_i>\left(\sum_{i=1}^{n}|X_i|^p\right)^{1/p}\Leftrightarrow$$

$$n\left(-\max_{1\le i\le n} X_i\right)^p>\sum_{i=1}^{n}|X_i|^p\Leftrightarrow(n-1)\left|X_{i_o}\right|^p>\sum_{i=1,i\ne i_0}^{n}|X_i|^p\ge(n-1)\left|X_{i_o}\right|^p$$

which is an impossible event. We can then conclude that

Proposition 4 If $x\in[-1, -1/n^{1/p})$, then $P[W_n(p) \le x] = 0$.

Further: having in mind Resnick's tail equivalence concept, and that Paretian tails span all possible regular variation index, we have computed $W_n(+\infty)$ for the variable X with density function given by

$f(x) = \frac{1}{2}\beta_2\alpha^{\beta_2}(-x)^{-\beta_2-1}\mathbf{I}_{(-\infty,-\alpha]}(x) + \frac{1}{2}\beta_1\alpha^{\beta_1}(-x)^{-\beta_1-1}\mathbf{I}_{[\alpha,+\infty)}(x)$, $\alpha, \beta_1, \beta_2>0$.

X has Paretian tails, with regular variation index $-\beta_1$ (right tail) and $-\beta_2$ (left tail). In the interval $(-\alpha,\alpha)$ the integral $\int_{-\infty}^{0}\left[F(xt)-F(t)\right]^{n-1}f(t)\,dt$ is null. For $-1< x \le 0$ we obtain

$$F_{W_n(+\infty)}(x)=\frac{\left[1-(-x)^{\beta_2}\right]^{n-1}}{2^n}$$

which converges to zero when n goes to infinity; when $0 < x < 1$

$$F_{W_n(+\infty)}(x)=\frac{n\beta_2}{2}\int_0^{x}\left[1-\frac{1}{2}\left(\frac{y}{x}\right)^{\beta_1}-\frac{1}{2}y^{\beta_2}\right]^{n-1}y^{\beta_2-1}dy+\frac{\left[1-x^{\beta_2}\right]^n}{2^n}.$$

If $\beta_1 = \beta_2 = \beta$ then

$$F_{W_n(+\infty)}(x) = \frac{n\beta}{2}\int_0^x \left[1 - \frac{1}{2}\left(\frac{y}{x}\right)^{\beta} - \frac{1}{2}y^{\beta}\right]^{n-1} y^{\beta-1}dy + \frac{\left[1-x^{\beta}\right]^n}{2^n}$$

$$= \frac{1}{1+x^{\beta}}\left[x^{\beta} + \left(\frac{1-x^{\beta}}{2}\right)^n\right]$$

which converges, when n goes to infinity, to $\frac{x^{\beta}}{1+x^{\beta}}$.

References

Bingham, N., Goldie, C. and Teugels, J. (1987). Regular Variation. Cambridge University Press.

Darling, D. A. (1952). The Influence of the Maximum Term in the Addition of Independent Random Variables, *Transactions of the American Mathematical Society* **73**, 95-107.

Feller, W. (1971). An Introduction to Probability Theory and Its Applications, II. John Wiley & Sons.

Fisher, R. A and Tippett, L., Limiting forms of the frequency distributions of the largest or smallest member of a sample, *Proc. Cambridge Philos. Soc.* **24**, 180-190.

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum, *Ann. de la Soc. Polonaise de Math.* (Cracow) **6**, 93-116.

Galambos, J. (1987). The Asymptotic Theory of Extreme Order Statistics, 2nd Ed.. Robert E Krieger.

Gnedenko, B., and Kolmogorov, A. (1954). Limit Distributions for Sums of Independent Random Variables, Addison-Wesley.

Graça Martins, M. E. and Pestana, D. D. (1987). Nonstable limit laws in extreme value theory. (eds Puri *et al.*), *New Perspectives in. Theoretical and Applied Statistics,* 449-458. Wiley.

LeCam, L. (1986). The Central Limit around 1935, *Statistical* Science **1**, No. 1, 78-96.

Linnik, Ju. V. and Ostrovskii (1977). Decomposition of Random Variables and Vectors, Translations of Mathematical Monographs. American Mathematical Society.

Loève, M. (1973). Paul Lévy 1886-1971, *The Annals of Probability* **1**, 1-18.

Logan, B. F., Mallows, C. L., Rice, S. 0. and Shepp L. A. (1973). Limit Distributions of Self-Normalized Sums, *The Annals of Probability* **1,** No. 5, 788-809.

Mejzler, D. (1956). On the problem. of the limit, distribution for the maximal term of a variational series, *Lvov Pol. Inst. Nauen Zp.* **38**, 90-109.

Mendonça, S. (1999). A Note on Random Normalizations, *Proceedings of the 11th European Young Statistitians Meeting.* Marly-le-Roi.

Pestana, D. (1978). Some Contributions to Unimodality, Infinite Divisibility, and Related Topics, PhD Thesis. University of Sheffield.

Pestana, D. (1984). Urbanik's (L_r) Classes, Mejzler's (M_r) Classes, and Higher Order Monotone Functions, *Actas III Colóquio Estatísticas Investigação Operacional.* Lagos.

Resnick, S. (1971). Tail Equivalence and its Applications, *J. of Appl. Prob.* **8**. Israel, 135-136.

Resnick, S. (1972). Products of Distribution Functions Attracted to Extreme Value Laws, *J. Appl. Prob.* **8**, 781-793.

Resnick, S. (1987). Extreme Values, Regular Variation, and Point Processes. Springer-Verlag.

Seneta, E. (1976). Regularly Varying Functions. *Lecture Notes in Mathematics* **508**. Springer-Verlag.

Urbanik, K. (1973). Limit laws for sequences of normed sums satisfying some stability conditions, In *Multivariate Analysis*, **III** (ed Krishnaiah), 225-237. Academic Press.

Product-Limit Estimation for Length-Biased Censored Data

Jacobo de Uña-Álvarez
University of Vigo
Department of Statistics and O. R.
Campus Universitario Lagoas-Marcosende
36200 Vigo, Spain
jacobo@correo.uvigo.es

1. Introduction

When analyzing times of duration (*e. g.* survival times in clinical trials, failure times in reliability studies or the length of spells of economic interest), a number of problems related to "loss of information" (in a wide sense) typically emerge. The presence of *censored* data is one of the most common features in this field. Right-censoring is caused by the occurrence of a risk that precedes the end of the spell being analyzed. Another phenomenon (very important in renewal processes) is that provoked by *length-biased* sampling. Roughly speaking, the length-bias appears when each duration time is sampled with a probability that is proportional to its length.

Let T be the duration of interest, and let F be the distribution function of T. The length-biased distribution of F is

$$F^*(t) = \mu^{-1} \int_0^t u F(du)$$

where μ is the mean duration time (assumed to exist). Under length-bias, one observes a random sample (not from F but) from F^*. Put Y for a random variable with distribution F^*. The observable information under both random censoring and length-bias is represented by a pair (Z,δ) where $Z = \min(Y,C)$ and $\delta = I(Y \leq C)$, C being a censoring variable independent of Y. The role of δ is indicating if the corresponding datum is censored ($\delta = 0$) or not ($\delta = 1$). The problem considered here is that of the estimation of F (and related quantities) given an initial sample of independent pairs (Z_i,δ_i), $i=1,...,n$, with the same distribution as (Z,δ).

Consistent estimation of F from censored data is an old goal in statistical research. Kaplan and Meier (1958) introduced the nonparametric maximum likelihood estimator of this curve. The so-called product-limit Kaplan-Meier estimate $\hat{F}$ satisfies $\hat{F}(t) \to F(t)$ with probability one (provided that the censoring distribution is not too heavy). Under length-bias what we get is $\hat{F}(t) \to F^*(t)$, and we have no longer consistency. Nonparametric estimation under length-biased sampling mechanisms was considered, among others, by Vardi (1982, 1985), Horváth (1985) and Jones (1991). For extending their ideas under random censoring, consider the key relation

$$F(t) = \int_0^t u^{-1} F^*(du) \Big/ \int_0^\infty u^{-1} F^*(du).$$

Then, substitute the product-limit $\hat{F}$ for F^*. We come up with the *length-biased-corrected product-limit*

$$\tilde{F}(t) = \int_0^t u^{-1}\hat{F}(du) / \int_0^\infty u^{-1}\hat{F}(du).$$

This is the natural extension of Vardi (1982)'s estimate under both length-bias and rigth-censoring. Note that $\tilde{F}$ shares some properties with $\hat{F}$. For example, it gives positive mass to the uncensored observations but not to the censored ones. However, unlike the usual product-limit, $\tilde{F}$ takes the value one at the maximum uncensored time, irrespective of the status (censored-uncensored) of $Z_{(n)} = \max_{1\le i\le n} Z_i$.

From a formal viewpoint, $\tilde{F}(t)$ is nothing but a quotient between two Kaplan-Meier integrals as those analyzed in Stute and Wang (1993) and Stute (1995, 1996). In Section 2 we present large sample results for $\tilde{F}$ and related estimates. Section 3 gives an illustration of the proposed techniques with real data. See de Uña-Álvarez (2000) for proofs and further details.

2. Main Result

We state our main result for the empirical parameter

$$\tilde{\gamma} = g(\int \varphi_1 d\tilde{F}, ..., \int \varphi_r d\tilde{F})$$

where the functions $g, \varphi_1, ..., \varphi_r$ are chosen by the researcher. Special cases of $\tilde{\gamma}$ give estimates for the distribution function $F(t)$, the survival function $S(t) = 1 - F(t)$, the cumulative hazard function $\Lambda(t) = -\ln(1 - F(t))$, the mean residual time function

$$r(t) = \frac{1}{1 - F(t)} \int_t^\infty (u - t) F(du),$$

and parameters such as the mean duration time and the variance of T. Put

$$\gamma = g(\int \varphi_1 dF, ..., \int \varphi_r dF)$$

for the limit of $\tilde{\gamma}$.

In regular situations, we obtain $\tilde{\gamma} \to \gamma$ with probability one and $\sqrt{n}(\tilde{\gamma} - \gamma) \to N(0, \sigma^2)$ in law, where the limit variance σ^2 is a complicated function of g, the φ's, and the joint distribution of the pair (Z, δ). In practice, one will be interested in consistently estimating the σ^2 parameter. The *jackknife* has been analyzed to this purpose by Stute (1996) under random right-censoring. A similar approach can be followed in our context, but details are not given here.

In proofs, the key fact about $\tilde{\gamma}$ is that it can be expressed as a function of $(r+1)$ integrals with respect to the usual product-limit measure. Once this is noted, we proceed by applying the strong law in Stute and Wang (1993) and by using the iid representation in Stute (1995).

Length-biased sampling is closely related to the left-truncation phenomenon. Indeed, length-biased data can be seen as arising from a left-truncated situation in

which the truncation variable is uniformly distributed on an interval containing the support of T. In general, the sampled distribution is of the form

$$F^w(t) = W^{-1} \int_0^t w(u)F(du)$$

where w is a nonnegative weighting function (known to the researcher). See Vardi (1985) for further motivation in the context of selection bias models. The corresponding correction for estimating F becomes

$$\tilde{F}^w(t) = \int_0^t w(u)^{-1} \hat{F}(du) / \int_0^\infty w(u)^{-1} \hat{F}(du).$$

of course, one can establish for this empirical similar results to those obtained for $\tilde{F}$.

3. An example

As an illustration, we consider data concerning unemployment spells of 700 Galician women. These data, obtained from the I. N. E. (the Spanish Institute for Statistics), were collected by means of repeated inquiries at the individuals' homes from 1987 to 1997. We included in the sample just those women being unemployed at the first inquiry time, the resulting spells being thus length-biased. Moreover, because of the design of the inquiries, each individual was followed during no more than 18 months, so there was a risk of right-censoring for the unemployment duration time. Actually, 378 spells were censored at the end of the period of observation, giving a censoring percentage of 64%.

Time (months)	PLE	CPLE
3	.9914	.9121
6	.9700	.7707
9	.9386	.6481
12	.9029	.5487
15	.8729	.4843
18	.8269	.4017
21	.7895	.3446
24	.7732	.3225
27	.7475	.2929
30	.7151	.2590
33	.6912	.2364
36	.6799	.2266

Table 1. Ordinary product-limit estimate (PLE) and length-bias-corrected version (CPLE) for the survival function of female unemployment duration in Galicia.

For comparison purposes, we show in Table 1 the survival function $1-F(t)$ estimated by means of both the ordinary and the length-bias-corrected product-limit estimates. Note that the ordinary product-limit decreases uniformly, overestimating the probability of being unemployed. On the other hand, the corrected version *reveals* that the probability of leaving unemployment is relatively high at the beginning, then decreasing quite rapidly. The usual Kaplan-Meier mean and median for these data are 6.97 and 6.5 years, respectively; the corresponding quantities for $\tilde{F}$ are 2.56 and 1.25

years. So the conclusions may be very misleading if the length-bias problem is not taken into account.

Acknowledgements

I am very grateful to María Soledad Otero Giráldez and Gema Álvarez Llorente for providing me with the economic data used in Section 3. I also acknowledge finantial support by the DGES grant PB98-0182-C02-02 and the Xunta de Galicia grant PGIDT00PXI20704PN.

References

Horváth, L. (1985). Estimation from a length-biased distribution, *Statistics & Decisions* **3**, 91-113.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457-481.

Jones, M. C. (1991). Kernel density estimation for length biased data, *Biometrika* **78**, 511-519.

Stute, W. (1995). The central limit theorem under random censorship, *The Annals of Statistics* **23**, 422-439.

Stute, W. (1996). The jackknife estimate of variance of a Kaplan-Meier integral, *The Annals of Statistics* **24**, 2679-2704.

Stute, W. and Wang, J. L. (1993). The strong law under random censorship, *The Annals of Statistics* **21**, 1591-1607.

de Uña-Álvarez, J. (2000). Product-limit estimation for length-biased censored data. Preprint.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias, *The Annals of Statistics* **10**, 616-620.

Vardi, Y. (1985). Empirical distributions in selection bias models (with discussion), *The Annals of Statistics* **13**, 178-205.

AUTHORS INDEX
ÍNDICE DE AUTORES

GENERAL INFORMATIONS

CALENDÁRIO DE REUNIÕES

CALENDAR OF EVENTS

2001

- 13-19 August
  23rd European Meeting of Statisticians, Funchal, Island of Madeira, Portugal.
  Informações: Dinis Pestana, University of Lisbon and Rita Vasconcelos, University of Madeira,
  E-mail: dinis.pestana@fc.ul.pt
  rita@dragoeiro.uma.pt
  URL: http://www.fc.ul.pt/cea/ems2001

- 15-20 August
  SRTL-2: The Second International Research Forum on Statistical Reasoning, Thinking, and Literacy, to be hosted by the Centre for Cognition Research in Learning and Teaching and the School of Curriculum Studies in the University of New England, Armidale, Australia.
  Informações: Dr Chris Reading, Department of Curriculum Studies, University of New England, Armidale NSW 2351, Australia,
  Tel: +02-67735060,
  Fax: +02-67735078,
  Email: creading@metz.une.edu.au,
  URL: http://www.beeri.org.il/SRTL/

- 17-19 August
  5th ICSA International Conference, Co-sponsored by IMS, to be held in Hong Kong.
  Informações: IMS Program Chair: Howell Tong, Univ. of Hong Kong, Local Arrangements Chair: Wai Keung Li, University of Hong Kong
  Email: htong@hku.hk
  hrntlwk@hkucc.hku.hk
  URL: http://icsa.vlp.com/HK2001/

- 19-23 August
  22nd Annual Conference of ISCB (The International Society for Clinical Biostatistics) will be held in Stockholm, Sweden, August 19 - 23, 2001.
  Information: Scientific Secretariat.
  E-mail: Theresa.Westerstrom@iscb.stockholm2001.org
  URL: http://www.iscb.stockholm2001.org/.

- 21-25 August
  ICANN 2001, International Conference on Artificial Neural Networks of the European Network Society, to be held at the Vienna University of Technology, Austria.
  Informações: Conference Secretariat: ICANN 2001, Austrian Research Institute for Artificial Intelligence, Schottengasse 3, A-1010 Vienna, Austria.
  Email: icann@ai.univie.ac.at

- 22-29 August
  International Statistical Institute, 53rd Biennial Session (includes meetings of the Bernoulli Society, The International Association for Statistical Computing, The International Association of Survey Statisticians, The International Association for Official Statistics, The International Association for Statistical Education), Seoul, Korea.
  Informações: ISI Permanent Office, Prinses Beatrixlaan 428,
  P.O. Box 950, 2270 AZ Voorburg, The Netherlands.
  Tel.: +31–70–337–5737;
  Fax: +31–70–386–0025;
  E-mail: isi@cbs.nl
  or visit the Session website at http://www.nso.go.kr/isi2001

- 30-31 August
  IAOS Satellite Meeting on Statistics for Information Society, to be held in Tokyo, Japan.
  Informações: Akihito ITO, Japan Statistical Association, 2-4-6 Hyakunin-cho, Shinjuku-ku, Tokyo 169-0073, Japan.
  Tel: +81-3-5332-3151;
  Fax: +81-3-5389-0691;
  Email: jsa@jstat.or.jp or Ito@jstat.or.jp

- 30 August-1 September
  International Conference on Statistical Challenges in Environmental Health Problems, to be held at the Soft Research Park, Fukuoka City, Japan.
  Information: The Chairman, Organizing Committee, Takashi Yanagawa, Graduate School of Mathemathics, Kyushu University, Fukuoka 812-8581, Japan.
  E-mail: yanagawa@math.kyushu-u.ac.jp

- 30 August-1 September
  ICNCB – International Conference on New Trends in Computational Statistics with Biomedical Applications (ISI 2001 Satellite Meeting, co-sponsored by IASC), to be held at the Osaka University Convention Center, Osaka, Japan.
  Informações: ICNCB Office, Division of Mathematical Science, Graduate School of Engineering Science, Osaka University. 1-3 Machikaneyama-cho, Toyonaka, Osaka 560-8531, Japan; Fax +81(6)6850-6496.
  Email: ICNCB@jscs.or.jp
  URL: http://www.jscs.or.jp/ICNCB/

- 1-4 September
  The annual meeting of Japan Statistical Society will be held at Seinan Gakuin University.
  Informações: URL: http://sunyht2.ism.ac.jp

- 6-12 September
  International Association for MATHEMATICAL GEOLOGY 6th Int'l Conference Cancún, Mexico.
  Informações: Gina Ross, Kansas Geological Survey.
  Email: aspiazu@kgs.ukans.edu
  URL: http://www.kgs.ukans.edu/Conferences/IAMG

- 17-19 September
  Methodology and Statistics, to be held in Ljubljana, Slovenia at the Faculty of Social Science, University of Ljubljana, Kardeljeva pl. 5, Ljubljana.
  Information: Anuska Ferligoj.
  E-mail: anuska.ferligoj@uni-lj.si
  URL: http://vlado.fmf.uni-lj.si/trubar/preddvor/2001/.

- 20-22 September
  Rasch Symposium, in honour of Professor Georg Rasch 100 years birthday, to be held at the Copenhagen Business School, Copenhagen, Denmark.
  Informações: Marianne Andersen
  Email: ma.mes@cbs.dk
  URL: http://www.cbs.dk/news/200701.shtml

- 24-25 September
  Statistical methods in biopharmacy, 4th international meeting: "Integrating issues of efficacy, safety and cost-effectiveness", to be held in Paris, France.
  Informações: Jean Auclair, IRI Servier, 6 place des pléiades, 92415 Courbevoie cedex, France. Fax: 33 1 55 72 68 27.
  Email: sfds.2001@curie.net
  URL: http://www.sfds.asso.fr/groupes/congresbiophar/congres2001.htm

- 24-27 September
  Statistical Week 2001, to be held in Dortmund, Germany.
  Informações: URL: http://g2.www.dortmund.de/inhalt/statistik/statwoch/intro.htm

- 25-29 September
  32nd European Mathematical Psychology Group Meeting, to be held in Lisbon, Portugal. Includes a workshop on Teaching and Training Mathematical Psychology in an Interdisciplinary and International Context. An Introductory Course on "Mathematical Psychology and Data Analysis" will be held on September 25th.
  Information: Prof. Dr. Helena Bacelar-Nicolau, Tel: +351 21 793 45 54; Fax: +351 21 793 34 08.
  E-mail: hbacelar@fc.ul.pt
  or
  empg2001@fpce.ul.pt
  URL: http://correio.cc.fc.ul.pt/~cladlead/EMPG01.html.

- 1-3 October
  2nd International Symposium on PLS and Related Methods (PLS'01) to be held at Capri Palace, Island of Capri (Naples, Italy).
  Information: Dr. Vincenzo Esposito, Dipartimento di Matematica e Statistica, Facoltà di Economia, Università "Federico II" di Napoli, via Cintia, Monte Sant'Angelo. Tel. +39 081 675112, fax: + 39 081 675113;
  E-mail: binci@unina.it
  URL: www.dms.unina.it/PLS2001.html

- 29-31 October
  Statistics as bases of creation the economic policy and the economic development in the South-East Europe, to be held in Skopje, Republic of Macedonia.
  Information: Mr. Sasho Kjosev - Faculty of Economics, University " Sts. Cyril and Methodius", Skopje, Republic of Macedonia or Mrs. Biljana Apostolovska - State Statistical Office of the Republic of Macedonia.
  E-mail: skosev@eccf.ukim.edu.mk
  or
  biljanaa@stat.gov.mk

- 1-4 November
  Euroworkshop on Statistical Modelling - Nonparametric Models, to be held in Schloss Hoehenried, Bernried, near Munich, Germany.
  Information: Göran Kauermann (coordinator of the project) University of Glasgow, Dep of Statistics & Robertson Centre, Boyd Orr Building, Glasgow G12 8QQ.
  E-mail: goeran@stats.gla.ac.uk
  URL: http://www.stat.uni-muenchen.de/euroworkshop.

- 4-7 November
  IX Annual Congress of the Portuguese Statistical Society to be held at the Universidade dos Açores, Ponta Delgada, Portugal.
  Information: Comissão Organizadora Local do IX Congresso da SPE, Dep. Matemática, Universidade dos Açores, Apartado 1422 9501-801 Ponta Delgada, Portugal.
  E-mail: ix_congresso_spe@alf.uac.pt
  URL: http://www.ixcongressospe.uac.pt

- 12-16 November
  VIII Latin-American Congress in Probability and Mathematical Statistics, to be held at the University of Havana, Cuba.
  Information: Gonzalo Perera (Chairman Program Committee), Pablo Olivares (Chairman Local Organizing Committee).
  E-mail: gperera@fing.edu.uy
  or
  clapem@matcom.uh.cu
  URL: http://www.uh.cu/eventos/clapem/ehome.htm.

- 14-16 November
  The Federal Committee on Statistical Methodology, which is composed of the senior statisticians from several U.S. federal statistical agencies and is sponsored by the U.S. Office of Management and Budget is planning a research conference in Arlington, Virginia.
  Information: The conference will feature papers and software demonstrations on topics related to a broad range of government statistical research interests.
  URL: http://www.fcsm.gov/

- 21-22 November
  9th Conference on National Accounting: the measurement of the new economy; Paris, France. Simultaneous translation French-English.
  Information: Michel Boëda (INSEE) - Simultaneous translation French-English
  E-mail: michel.boeda@insee.fr
  URL: http://www.insee.fr/fr/av_service/colloques/cnat_accueil.html
  or
  http://www.insee.fr/en/av_service/colloques/cnat_accueil.html

- 7-9 December
  International Conference on "Characterization Problems and Applications", tentative Venue: Antalya, Turkey.
  Information: Omer L. Gebizlioglu, Ankara, Turkey; N. Balakrishnan, McMaster University, Canada; Ismihan Bayramov, Ankara, Turkey.
  E-mail: Omer.L.Gebizlioglu@science.ankara.edu.tr
  bala@mcmail.cis.mcmaster.ca
  Ismihan.Bayramov@science.ankara.edu.tr

- 19-22 December
  International Conference on Statistics, Combinatorics and Related Areas and The Eighth International Conference of the Forum for Interdisciplinary Mathematics, to be held at the University of Wollongong, Australia.
  Information: Chandra M. Gulati, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW 2522, Australia. Telephone:+61-2-4221-3836, fax:+61-2-4221-4845.
  E-mail: chandra_gulati@uow.edu.au
  or
  cmg@uow.edu.au
  URL: http://www.uow.edu.au/informatics/maths/statconference.

- 20-22 December
  Statistical Analysis for Global Environment, to be held at the Siam Intercontinental Hotel, Bangkok, Thailand.
  Information: Dr.Supol Durongwatana.
  E-mail: fcomsdu@phoenix.acc.chula.ac.th

- 20-23 December
  International Conference on History of Mathematical Sciences, to be held in Delhi, India.
  Information: Dr. Y. P. SABHARWAL, Department of Mathematics & Statistics, Ramjas College, University of Delhi, Delhi 110 007, India; Tel : (011) 294 1119.
  E-mail: ypsabharwal@yahoo.com
  or
  ichm2001rjc@yahoo.com

2002

- 15-18 January
  First International ICSC Congress on Neuro-Fuzzy NF'2002 to be held at The Capitolio de la Habana, Cuba.
  Informações: INTERNATIONAL COMPUTER SCIENCE CONVENTIONS Head Office: 5101C-50 Street, Wetaskiwin AB, T9A 1K1, Canada (Phone: +1-780-352-1912 / Fax: +1-780-352-1913)
  Email: operating@icsc.ab.ca
  or
  planning@icsc.ab.ca
  URL: http://www.icsc.ab.ca/NF2002.htm
  or
  http://www.icsc.ab.ca/

- 16-18 January
  Food-Industry and Statistics, to be held in Villeneuve d'Ascq (LILLE), France. Bât. EUDIL IAAL - Cité Scientifique, F 59655.
  Information: E-mail: agrostat2002@eudil.fr
  URL: http://www.eudil.fr/~agrostat.

- 4-8 February
  ProbaStat 2002, the 4th International Conference on Mathematical Statistics, to be held at Smolenice Castle, Smolenice, Slovak Republic.
  Information: E-mail: probastat@savba.sk
  URL: http://www.um.savba.sk/lab_15/probastat.html.

- 12-15 February
  First International ICSC-NAISO Congress on Autonomous Intelligent Systems ICAIS 2002 to be held at Deakin University, Geelong, Australia.
  Information: E-mail: icais02@itstransnational.com
  URL: http://www.icsc-naiso.org/conferences/icais2002/index.html

- 15-21 March
  ENAR/IMS Eastern Regional to be held in Washington, DC, USA.
  Informações: Program Chair: Jiayang Sun, Case Western Reserve University
  Local Arrangements Chair: Colin Wu, John Hopkins University
  Contributed Papers Chair: Nidhan Choudhuri;
  E-mail: jiayang@sun.STAT.cwru.edu
  colin@mts.jhu.edu
  nidhan@nidhan.cwru.edu
  URL: http://sun.cwru.edu/ims

- 2-5 June
  Annual Meeting of the Statistical Society of Canada, Hamilton, Ontario, Canada.
  Informações: Peter Macdonald, Department of Mathematics and Statistics, McMaster University, 1280 Main Street West, Hamilton, Ontario, L8S 4K1, Canada.
  E-mail: pdmmac@mcmail.cis.mcmaster.ca

- 17-20 June
  MMR 2002, Third International Conference on Mathematical Methods in Reliability, to be held at the Norwegian University of Science and Technology, Trondheim, Norway.
  Informações: Professor Bo Lindqvist, Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway. Tel.: +47-73 59 35 20 - Fax: +47-73 59 35 24.
  E-mail: mmr2002@math.ntnu.no
  URL: http://www.math.ntnu.no/mmr2002/

- 23-29 June
  The 8th International Vilnius Conference on Probability Theory and Mathematical Statistics, Vilnius, Lithuania.
  Informações: Professor Vytautas Statulevicius, Institute of Mathematics and Informatics, Akademijos str. 4, 2600 Vilnius, Lithuania.
  E-mail: conf@ktl.mii.lt

- 2-5 July
  MCQT'02 - First Madrid Conference on Queueing Theory, to be held at the Department of Statistics and OR, Faculty of Mathematics, University Complutense of Madrid, Spain.
  Information: Jesus R. Artalejo.
  E-mail: mc_qt@mat.ucm.es
  URL: http://www.mat.ucm.es/deptos/es/mcqt/conf.html.

- 7-12 July
  The Sixth International Conference on Teaching Statistics (ICOTS6), to be held in Durban, South Africa.
  Information: Maria-Gabriella Ottaviani - IPC Chair; Brian Phillips - International Organizer; , Dani Ben-Zvi - IPC Scientific Secretary.
  E-mail: mariagabriella.ottaviani@uniroma1.it;
  bphillips@swin.edu.au;
  dani.ben-zvi@weizmann.ac.il.
  URL: http://icots.itikzn.co.za/.

- 15-19 July
  Current Advances and Trends in Nonparametric Statistics, to be held on Crete, Greece.
  Informações: Michael G. Akritas and Dimitris N. Politis IMS Representative: Michael G. Akritas,
  E-mail: mga@stat.psu.edu
  URL: http://www.stat.psu.edu/~npconf/

- 21-26 July
  IBC 2002 - International Biometric Conference 2002, to be held at the University of Freiburg, Germany.
  Information: Chair: Robert Curnow; Chair Local Organizing Committee: Martin Schumacher.
  E-mail: r.n.curnow@reading.ac.uk
  ms@imbi.uni-freiburg.de
  URL: http://www.ibc2002.uni-freiburg.de/.

- 22-24 July
  26th Annual Conference of the Gesellschaft für Klassifikation (GfKl), to be held at the University of Mannheim, Germany.
  Informações: local organizer Prof. Dr. Martin Schader.
  URL: http://www.gfkl.de/gfkl2002

- 27 July – 1 August
  IMS Annual Meeting/Fourth International Probability Symposium, to be held in Banff, Canada.
  Informações: IMS Program Chair Tom DiCiccio, Cornell, Symposium Chair: Tom Kurtz, U. Wisconsin, IMS Local Chair: Subhash Lele, U. Alberta.
  E-mail: tjd9@cornell.edu
  Kurtz@math.wisc.edu
  slele@ualberta.ca

- 4-9 August
  Fourth International Conference on Statistical Data Analysis based on the L_1-Norm and Related Methods - to be held at the University of Neuchâtel, Switzerland.
  Information: Prof. Yadolah Dodge, Conference Organizer Statistics Group, Case Postale 1825, CH-2002 Neuchatel. Phone +41 32 718 13 80 Fax +41 32 718 13 81.
  E-mail: Yadolah.Dodge@unine.ch

- 11-15 August
  Joint Statistical Meetings, New York, Hilton and Sheraton New York.
  Sponsored by ASA, ENAR, WNAR, IMS, and SCC.
  Informações: ASA, 1429 Duke St., Alexandria, VA 22314-3415;
  Tel. (703) 684-1221;
  Email meetings@amstat.org

- 16-18 August
  Symposium on Stochastics and Applications (SSA) to be held at the National University of Singapore.
  Informações: E-mail: ssa@math.nus.edu.sg
  URL: http://www.math.nus.edu.sg/ssa

- 19-23 August
  24th European Meeting of Statisticians, Prague, Czech Republic.
  Informações: Martin Janzura, Institute of Information Theory and Automation, POB 18, 182 08 Praha 8, Czech Republic.
  Tel: 420 2 6605 2572.
  Fax: 420 2 688 4903.
  Email: janzura@utia.cas.cz

- 24-28 August
  Compstat2002 to be held in Berlin, Germany.
  E-mail: info@compstat2002.de, website http://www.compstat2002.de
  Informações: E-mail: info@compstat2002.de
  URL: http://www.compstat2002.de

- 25-28 August
  International Conference on Improving Surveys (ICIS-2002), to be held at the University of Copenhagen.
  Information: International Conference Services, P.O. box 41, Strandvejen 171, DK-2900 Hellerup, Copenhagen, Denmark. Telephone: +45 3946 0500, Fax +45 3946 0515.
  E-mail: ICIS2002@ics.dk

- 2-6 September
  RSS 2002 Conference to be held at the University of Plymouth, Plymouth, England.
  Information: The 2002 Conference of the Royal Statistitical Society (4-6 September) will preceded by short courses (2-3 September).
  E-mail: J.Stander@plymouth.ac.uk

- 13-17 November
  International Conference on Questionnaire Development, Evaluation, and Testing, probably to be held in the southeastern United States.
  Information: URL: http://www.jpsm.umd.edu/

- 28-30 December
  International Conference on "Ranking and Selection, Multiple Comparisons, Reliability, and Their Applications". Tentative Venue: Hotel Savera, Chennai, Tamilnadu, India.
  Organizers: bala@mcmail.cis.mcmaster.ca;, NKannan@utsa.edu; H. N. Nagaraja, Ohio State University, mailto:hnn@stat.ohio-state.edu
  Information: N. Balakrishnan, McMaster University; N. Kannan, University of Texas at San Antonio; H. N. Nagaraja, Ohio State University.
  E-mail: bala@mcmail.cis.mcmaster.ca
  NKannan@utsa.edu
  mailto:hnn@stat.ohio-state.edu

2003

- 10-20 August
  International Statistical Institute, 54th Biennial Session (includes meetings of the Bernoulli Society, The Intern. Assoc. for Statistical Computing, The Intern. Assoc. of Survey Statisticians, The Intern. Assoc. for Official Statistics and The Interna. Assoc. for Statistical Education), to be held in Berlin, Germany.
  Informações: ISI Permanent Office, Prinses Beatrixlaan 428,
  P.O. Box 950, 2270 AZ Voorburg, The Netherlands.
  Tel.: +31–70–337–5737;
  Fax: +31–70–386–0025;
  E-mail: isi@cbs.nl
  or visit the Session website at http://www.isi–2003.de

FUNDAMENTO, OBJECTO E ÂMBITO DA REVISTA

O INE, consciente de como uma cultura estatística é essencial para a compreensão da maioria dos fenómenos do mundo actual, e da sua responsabilidade na divulgação do conhecimento estatístico, fazendo-o chegar ao maior número possível de leitores, tendo reconhecido a necessidade de dar um passo nesse sentido, passou a editar quadrimestralmente a presente *Revista de Estatística* destinada a divulgar:

a) Numa perspectiva científica, artigos originais sobre temas especializados da estatística, tanto pura como aplicada, bem como sobre estudos e análises nos domínios económico, social e demográfico;

b) Informações sobre actividades e projectos importantes do Sistema Estatístico Nacional;

c) Informações sobre acções desenvolvidas pelo INE no âmbito da cooperação.

d) Informações sobre congressos, seminários, colóquios e conferências de interesse estatístico ou afim;

Para tal, são adoptadas as seguintes formas de contribuição para publicação na Revista:

- Quanto aos artigos referidos em a), contribuições da *iniciativa* dos próprios autores e por *convite* do Conselho Editorial, pertencentes ou não ao INE;
- Quanto às informações referidas em *b)*, *c)* e *d)*, contribuições dos departamentos do INE.

As contribuições de artigos por iniciativa dos próprios autores serão objecto de avaliação de mérito científico pelo Conselho Editorial, que decidirá ou não pela sua publicação.

Para a elaboração e envio das contribuições de artigos para publicação na Revista são adoptadas as *Normas de Apresentação de Originais* que figuram na última página.

Os autores dos artigos publicados, a que se refere a alínea a), receberão uma contribuição financeira paga pelo INE, de montante a fixar por despacho da Direcção mediante proposta do Director da Revista.

FOUNDATION, SUBJECT MATTER AND SCOPE OF THE REVIEW

INE is conscious of how statistical awareness is essential to the understanding of the majority of phenomena in the present world and is aware of its responsibility to disseminate statistical knowledge, making it available to the widest possible range of readers. INE has recognised the need to take a step in that direction and will begin publication of this *Statistical Review* three times yearly, designed to provide the following:

a) Within a scientific perspective, original articles on specialised areas of statistics, both pure and applied, as well as studies and analyses within the sphere of economics, social issues and demographics;

b) Information on activities and projects of the National Statistical System;

c) Information on activities developed by INE within the scope of co-operation;

d) Information on congresses, seminars and conferences of a statistical or related nature;

The following approaches for contributing material for publication in the review have been adopted:

- In relation to the articles referred to in section a), contributions are made by the authors themselves and by invitation of the Editorial Committee, whether they are employees of INE or not;
- In relation to the information referred to in section b), c) and d); contributions are from departments of INE.

The Editorial Committee who has sole discretion in deciding whether or not the material will be published will assess the scientific merit of contributions made on the initiative of the authors themselves.

The preparation and delivery of material for publication in the Review are subject to the *Rules for Submitting Originals* presented on the last page.

The authors of the published articles referred to in section a) will receive pecuniary compensation from INE in an amount to be determined by resolution of the Board on the recommendation of the Director of the Review.

NORMAS DE APRESENTAÇÃO DE ORIGINAIS

Nos termos do *Regulamento da Revista de Estatística*, o Conselho Editorial aprovou as seguintes **Normas de Apresentação de Originais**:

1. Os originais dos artigos serão enviados ao Director da Revista pelos respectivos autores, devendo ser escritos em *português* e não terem sido ainda totalmente publicados, ou estar em processo de edição em outra publicação.

2. Poderão também ser apresentados artigos escritos em *inglês*, cabendo ao Director da Revista a decisão sobre a sua aceitação.

3. Quanto à *avaliação do mérito científico* dos artigos:

   a) Os artigos apresentados por *iniciativa* dos respectivos autores serão submetidos à avaliação do mérito científico pelo Conselho Editorial, com garantia do anonimato tanto do autor como dos avaliadores;

   b) Os autores receberão a informação sobre o resultado da avaliação num prazo máximo de trinta dias, com indicação, nos casos de avaliação positiva, do número da *Revista* em que serão publicados, e nos casos de avaliação negativa com a devolução do original apresentado.

4. Os artigos aceites para publicação na *Revista de Estatística* serão igualmente divulgados no *site* do INE na *Internet*.

5. Os originais, com uma extensão não superior a trinta páginas, serão processados em *Word for Windows*, integralmente a preto e branco, com indicação do(s) software(s) adicional(ais) eventualmente utilizado(s) na produção do documento original, e entregues em suporte papel acompanhado da respectiva *disquette*, ou enviados por E-mail para o seguinte endereço: liliana.martins@ine.pt

6. Na apresentação dos originais, os autores respeitarão ainda as seguintes normas:

   6.1. Quanto à *estrutura*:

      a) O texto deve ser processado em formato A_4, com utilização do tipo de letra *Times New Roman 11*, espacejamento *at least 12*, e com as seguintes margens: *top*: 4 cm, *bottom*: 3 cm, *left*: 2,5 cm, *right*: 5 cm, *header*: 1,25cm, *footer*: 1,25cm;

      b) A primeira página conterá exclusivamente o título do artigo, bem como o nome, morada e telefone, fax e E-mail do autor, com indicação das funções exercidas e da instituição a que pertence, devendo, no caso de vários autores, ser indicado a quem deverá ser dirigida a correspondência da Revista;

      c) A segunda página conterá, em português e inglês, unicamente o título e um *resumo* do artigo, com um máximo de 100 palavras,

seguido de um parágrafo com indicação de *palavras-chave* até ao limite de 15;

d) Na terceira página começará o texto do artigo, sendo as suas eventuais secções ou capítulos numeradas sequencialmente;

6.2. Quanto a *referências bibliográficas*:

a) Os autores eventualmente citados no texto do artigo serão indicados entre parênteses curvos pelo seu nome seguido da data da respectiva publicação e, se for caso disso, do número de página (p. ex.: Malinvaud, 1989, 23);

b) As Referências Bibliográficas serão listadas, por ordem alfabética dos apelidos dos respectivos autores, imediatamente a seguir ao final do texto, de acordo com a fórmula seguinte:

GREENE, W. H., "*Econometric Analysis*", Prentice-Hall, New Jersey, 1993.

6.3. Quanto à *revisão de provas e publicação*:

a) Uma vez aceite o artigo e antes da sua publicação, receberá o autor provas para revisão, as quais serão devolvidas ao Director da Revista no prazo máximo de uma semana contado da data da sua recepção;

b) Serão da responsabilidade dos respectivos autores as consequências de eventuais modificações da versão inicial aceite, bem como de atrasos na revisão das provas, que impossibilitem a publicação no número da Revista previsto, reservando-se o Director o direito de decidir a data da sua publicação futura;

c) Uma vez publicado o artigo, o autor receberá vinte exemplares da sua versão impressa e um exemplar do respectivo número da *Revista*.

7. Para *informações adicionais* contactar o Secretariado de Redacção:

Eduarda Liliana Martins
Instituto Nacional de Estatística
Av^{a}. António José de Almeida, n.º 5 – 9º.
1000-043 Lisboa - Portugal

- Tel.: +351 21 842 62 05
- Fax.: +351 21 842 63 66
- e-mail: liliana.martins@ine.pt

RULES FOR SUBMITTING ORIGINALS

Within the terms of the *Regulation of the Statistical Review*, the Editorial Committee has approved the following **Rules for Submitting Originals**:

1. The original articles will be sent to the Review Director by the respective authors. They should be written in *Portuguese*, they should not have already been published in their entirety nor should they be in the process of being published in any other publication.

2. Articles may also be submitted in *English* to the Review's Director who will decide whether to accept them.

3. In relation to the *evaluation of the scientific merit* of the articles:

   a) The Editorial Committee will assess the articles submitted on the initiative of the authors on the basis of their scientific merit. The identity of both the author and the Committee members will be strictly confidential;

   b) The authors will receive information regarding the results of the evaluation of scientific merit within a maximum period of 30 days. If the article is accepted, the Committee will indicate the issue number of the *Review* in which the article will be published. If the article is not accepted, the original will be returned to the author.

4. The articles accepted for publication in the *Statistical Review* will also be made public on the Internet site of the INE.

5. The original articles having no more than thirty pages must be processed in *Word for Windows*, completely at black and white, with the information on the addicional(s) software(s) eventually used in the production of the original document, and they will be delivered in hard copy as well as on diskette, or sent by E-mail to: liliana.martins@ine.pt

6. With the presentation of the original articles, the authors must also respect the following rules:

   6.1 In relation to the *structure*:

      a) The text shall be printed on A4 format paper utilising the font *Times New Roman* size 11, spacing at least 12, and with the margins: *top* 4cm, *bottom* 3cm, *left* 2,5cm, *right* 5cm, *header* 1,25cm, *footer* 1,25cm;

      b) The first page shall contain only the title of the article as well as the name, address and telephone, fax and E-mail number of the author, indicating the position held and the institution that he/she belongs to. In the case of various authors, it is necessary to indicate the person to whom all correspondence received should be forwarded;

c) The second page shall contain in *Portuguese* and *English* only the *title* and an *abstract* of the article with the maximum of 100 words followed by a paragraph indicating *key words* up to the limit of 15;

d) The third page will begin the text of the article with its respective sections or chapters sequentially numbered;

6.2 Regarding *Bibliographical References*:

a) Authors who are cited in the text of the article shall be indicated in parentheses with their name followed by the date of the respective publication and, if necessary, the page number (ex.: Malinvaud, 1989, 23);

b) All bibliographical references will be listed in alphabetical order by the surnames of the respective authors, immediately following the end of the text, as in the following example:

GREENE, W. H., "*Econometric Analysis*", Prentice-Hall, New Jersey, 1993.

6.3 Regarding *proof-reading and publication*:

a) Once the article is accepted and prior to its publication, the author will receive a copy for review. These copy will be returned to the Director of the Review within a maximum period of one week from the date of its reception;

b) The consequences of subsequent changes to the accepted first version are the responsibility of the respective authors as well as any delays in proof-reading that make its publication in the planned issue of the Review impossible. The Director reserves the right to decide upon the date for future publication;

c) Once the article is published, the author will receive twenty copies of his/her printed version and a copy of the respective issue of the *Review*.

7. For *further information* kindly contact the Editorial Secretary:

Eduarda Liliana Martins
Instituto Nacional de Estatística
Avª. António José de Almeida, nº. 5 – 9º.
1000-043 Lisbon - Portugal

- Tel.: +351 1 21 842 62 05
- Fax.: +351 1 21 842 63 66
- e-mail: liliana.martins@ine.pt