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

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Foreword

This special issue of *Revista de Estatística — Statistical Review* contains the short summaries of the contributed papers accepted for presentation at the *23rd European Meeting of Statistics*, Funchal, 13-18 August 2001.

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 the INE — the Instituto Nacional de Estatística. The Programme Committee:

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

both built up the invited programme and screened the contributed papers before acceptance.

We are grateful to *FCT — Fundação Para a Ciência e Tecnologia* and *Calouste Gulbenkian Foundation* for their sponsorship of the meeting and help in preparation of this special issue of *Revista de Estatística — Statistical Review*.

We wish to express our gratitude to Mr. Nuno Barreto for his careful retyping and editing of the papers, and to Mrs. Liliana Martins for her skill in desktop publishing. But our largest debt is to all the authors who submitted their work, thus making the *EMS 2001* such a lively meeting.

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

Lisboa, 2001 March 30th

GLOBAL INDEX
ÍNDICE GLOBAL

REVISTA DE
ESTATÍSTICA
2º QUADRIMESTRE DE 2001

CONTRIBUTED PAPERS

Discriminating Between Time-Dependent and Non-Linear Effects of Continuous Predictors in Survival Analysis: Regression Spline Modeling and Inference

Michal Abrahamowicz
McGill University, Department of Epidemiology and Biostatistics
1650 Cedar Avenue, Montreal, Quebec H3G 1A4 CANADA
michal@michal.ri.mgh.mcgill.ca

Todd MacKenzie
Department of Biometry, University of Colorado
The Children's Hospital Association, 1056 East 19th Avenue, Box B321
Denver, Colorado 80218
Mackenzie.Todd@tchden.org

Thomas Perneger
University of Geneva, Department of Social and Preventive Medicine
CMU, CH-1211 Geneva 4, Switzerland
perneger@cmu.unige.ch

1. Background

Conventional version of the, extremely popular, Cox (JRSS 1972) Proportional Hazards (PH) regression model for censored survival data imposes two assumptions: (1) linearity of the effects of continuous predictors on log hazard; and (2) PH assumption that requires the Hazard Ratio (HR) expressing the predictor's impact to remain constant over entire follow-up period. In 1990's, several researchers proposed flexible non- or semi-parametric generalizations of the Cox model, in order to relax *either* the linearity *or* the PH assumption. Applications of these methods in various areas of epidemiological and biomedical research yielded new insights into the role of prognostic factors for colon cancer (Quantin et al, Am J Epi 1999), breast cancer (Gray, JASA 1992), stroke (Lewis et al, Ann Int Med 1997), and many other diseases. In contrast, little work has been done on simultaneous relaxation of both assumptions.

2. Objectives

1/ To propose a new method for simultaneous flexible modeling of both (i) non-linear dose-response functions; and (ii) changes over time in HR, within a generalized version of Cox model;

2/ to evaluate the usefulness of simultaneous estimation using both simulated and real-life data;

3/ to assess issues related to testing the linearity and/or PH hypotheses, and to model selection.

3. Method

We propose a generalization of our previous work on flexible modeling of time-dependent HR (Abrahamowicz, MacKenzie and Esdaile, JASA 1996), to estimate the "flexible product model":

$$h(t|x) = h_0(t) \exp[f(t)*g(x)]$$

where $h_0(t)$ and $h(t)$ denote, respectively, the baseline hazard function and hazard conditional on the predictor X. Low-dimension regression splines (Ramsay, Stat Science 1988) are employed to simultaneously model two "marginal" functions: f(t) and g(x) representing, respectively, the time-dependent and non-linear effects of predictor X. Several Likelihood Ratio Tests (LRT) are proposed to test various hypotheses of interest. To avoid the problems related to model identification and over-parametrization, the model is re-formulated so that g(x) represents only the relative changes in hazard associated with different predictor values, while the strength of their impact is represented by f(t). This implies a specific interpretation of the estimates, and of their pointwise confidence intervals, which will be illustrated using an empirical example. Finally, a simple test of the adequacy of the assumptions underlying the product model will be proposed.

4. Simulations:

One of the two main goals of the simulations is to assess the accuracy of the point estimates of the two marginal functions, and of the resulting estimates of the conditional hazards. To this end, we will simulate different, practically relevant, data structures using our previously developed permutational algorithm for generating censored survival data with abitrary patterns of time-dependent relative risks (MacKenzie and Abrahamowicz, J Stat & Computing, in press). The performance of the proposed LRT tests, in terms of both empirical test size and power will be also evaluated. The second major goal of the simulations is to demonstrate the impact of model mis-specification on the bias in the estimates, and on the inflation of type I error rates of relevant tests. Specifically, we will demonstrate that constraining the estimate by imposing an (incorrect) linearity assumption may result in type I error rate as high as 70% when testing the proportional hazards hypothesis, and - vice versa - an incorrect assumption of the constant hazard ratio may produce a dramatic inflation of type I error rate for testing the linearity of the dose-response relationship.

5. Illustration

The ability of the new method to yield new insights in the structure of real-life data will be illustrated by assessing the effects of popular markers (CD4 cell count and viral load) of disease progression in HIV-positive patients. This example is of practical interest as previous analyses, that focused on only one of the two aspects of the relationships between these HIV markers and clinical outcomes, suggested, respectively, that: (i) their effects are non-linear; and (b) they change during follow-up time. These conclusions will be revisited using our flexible product model.

References

Cox DR. Regression models and life tables (with discussion). *J R Stat Soc* 1972;Series B **34**:187-220.

Quantin C, Abrahamowicz M, Moreau T, et al. Variation over time of the effects of prognostic factors in a population based study of colon cancer: Comparison of statistical models. *Am J Epidemiol* 1999;**15011**:1188-200.

Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association* 1992;**87**:942-51.

Lewis RF, Abrahamowicz M, Côté R, et al. Predictive power of duplex ultrasonography in asymptomatic carotid disease. *Ann Intern Med* 1997;**127**:13-20.

Abrahamowicz M, MacKenzie T, Esdaile JM. Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis. *J Am Stat Assoc* 1996;**91436**:1432-9.

MacKenzie T, Abrahamowicz M. Marginal and hazard ratio specific random data generation: applications to semi-parametric bootstrapping. *J Stat Comput*. [In Press]

Assessing and Modelling the Behaviour of the Gross National Product of Argentina: 1875 - 1999

Juan Carlos Abril, María Beatriz Blanco
Faculty of Economics, National University of Tucumán
Casilla de Correo 209, 4000 San Miguel de Tucumán, Argentina
jabril@herrera.unt.edu.ar

1. The Statistical Analysis of the Data

A form of measuring the economic activity of a country is by its gross national product (GNP) and by its gross national product per capita (GNPpc). Since our objective is to study the evolution of the Argentinean economy from 1875 to 1999, we analyse the annual data of the GNP and the GNPpc for this period. We take the logarithm of the series and apply a structural time series model of the form

$$\begin{aligned} y_t &= \mu_t + \psi_t + \mathbf{z}_t'\delta + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, \sigma_\varepsilon^2), \quad t = 1, \ldots, n, \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t, \quad \eta_t \sim \text{NID}(0, \sigma_\eta^2), \\ \beta_t &= \beta_{t-1} + \zeta_t, \quad \zeta_t \sim \text{NID}(0, \sigma_\zeta^2), \end{aligned} \tag{1}$$

where y_t is the log of the series, μ_t is the trend, ψ_t is the cycle, ε_t is an irregular component serially independent, normally distributed with mean zero and constant variance, i. e. $\varepsilon_t \sim \text{NID}(0, \sigma_\varepsilon^2)$, $\mathbf{z}_t$ is a $p\times1$ vector of observed explanatory variables and δ is $p\times1$ vector of unknown parameters. A stochastic formulation of the trend allows the level μ_t and the slope β_t to evolve over time. The stochastic cycle is

$$\begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix} = \rho \begin{bmatrix} \cos\lambda_c & \text{sen}\,\lambda_c \\ -\,\text{sen}\,\lambda_c & \cos\lambda_c \end{bmatrix} \begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix}, \; t = 1, \ldots, n, \tag{2}$$

where λ_c is the frequency in radians within the interval $0 \le \lambda_c \le \pi$, κ_t and κ_t' are two white noise mutually uncorrelated with mean zero and common variance σ_κ^2, and ρ is a smoothing factor such that $0 < \rho \le 1$. The period is $2\pi / \lambda_c$.

In our case and for both series we identify and estimate a model with fixed level, i. e. $\sigma_\eta^2 = 0$, stochastic slope and cycle, with some structural changes in the slope and outliers. The outliers are identified as those large values in the irregular residuals and are captured using dummies variables as explanatory variables in the measurement equation [the first equation of (1)]. The changes in the slope are identified as those large values in the slope residuals and are captured using dummies variables in the corresponding transition equation [the third equation of (1)]. The estimation procedure was carried out by means of the Kalman filter and smoother and using the STAMP package [Koopman, Harvey, Doornik and Shephard (1995)]. Due to the lack of space, we are going to analyse only the GNPpc because it is usually considered the most important from the economic point of view. Anyway, the results for both series are highly similar.

2. More than a Century of Argentinean Economy

In general, the Argentinean economy has shown an important growth during the period under consideration. In fact, the trend was growing at an increasing rate with some intermediate periods of decrease, such as 1909-1916, 1928-1931 and 1979-1989 (for the GNP, they are 1913-1916, 1929-1931 and 1980-1989). On the other hand,

both series have a stochastic cycle with a period of 5 years and 7 months and amplitude of 2% of the trend, approximately. There are two outliers in both series; one in 1891 corresponding to an external debt crisis and another in 1899 which corresponds to an important monetary reform.

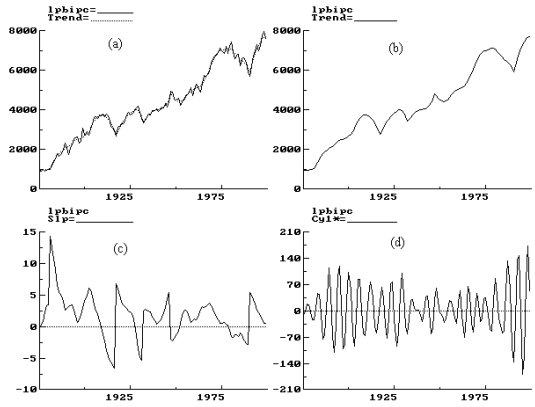


Figure 1. GNPpc of Argentina 1875-1999 with trend [(a) and (b)], its slope (c) and its cycle (d)

In the period under consideration there are five structural changes corresponding to changes in the slope. The first one occurred in 1881 as a result of new economic policies implemented by the national government from 1880 and the increasing use of better technologies mainly in transport. The second one occurred in 1917, and it could be attributed to the consequences of the First World War. The third happened in 1932 and it is the result of the world recession. The forth came about in 1947 and it is the result of the economic reforms introduced by the newly appointed government headed by J. D. Perón. Finally, the fifth occurred in 1990 and clearly we can attribute it to the passage from an hyperinflationary period which reached its maximum in 1989 to a stabilisation period which started in 1991. For the GNP the only changes in the slope that took effect where in 1881, 1917 and 1990.

The five structural changes occurring in the GNPpc determine six periods in the Argentinean economy associated with different political circumstances inside the country and with important international situations that affected Argentina.

References

Abril, J. C. (1999). Análisis de Series de Tiempo Basado en Modelos de Espacio de Estado. EUDEBA. Buenos Aires.

Cortés Conde, R. (1997). La Economía Argentina en el Largo Plazo. Sudamericana. Buenos Aires.

Cortés Conde, R. (1998). Progreso y Declinación de la Economía Argentina. *Fondo de Cultura Económica*. Buenos Aires.

Dornbusch, R. and de Pablo, J. C. (1988). Deuda Externa e Inestabilidad Macroeconómica en la Argentina. Sudamericana. Buenos Aires.

Gerchunoff, P. and Llach, J. J. (1975). Capitalización industrial, desarrollo asociado y distribución del ingreso entre los dos gobiernos peronistas: 1950-1972. *Desarrollo Económico*, 15, num. 57, April-June.

Koopman, S. J., Harvey, A. C., Doornik, J. A. and Shephard, N. (1995). *STAMP 5.0: Structural Time Series Analyser, Modeller and Predictor*. Chapman and Hall. London.

Unemployment Duration: Competing and Defective Risks

John T. Addison
University of South Carolina, Department of Economics
Columbia, SC, USA
ecceaddi@darla.badm.sc.edu

Pedro Portugal
Banco de Portugal, Departamento de Estudos Económicos
Av. Almirante Reis, 71, Lisboa, Portugal
Jppdias@bportugal.pt

This paper uses a competing risks model of unemployment duration in which exit from unemployment can result from finding a job or becoming inactive, which destinations are properly viewed as behaviorally distinct states (Flinn and Heckman, 1983). Use of a competing risks specification while familiar is not commonplace in empirical unemployment duration analysis Altogether less familiar, is the notion that risks may be defective (Yamaguchi, 1992).

Whereas in single-risk duration models one aggregates over a number of exit modes, for competing risks one has to confront the possibility that some exit routes are simply not viable. Furthermore, if indeed there exists a subpopulation with a zero hazard rate, we should observe a declining aggregate hazard function.

The empirical model used here reflects the sample information and the sample plan (observation over a fixed interval). We use a grouped duration model in which remaining duration is conditioned on elapsed duration (Cox and Oakes, 1985). A flexible semiparametric baseline hazard function is specified, namely, a piecewise-constant hazard function with 13-segments. Modes of failure are treated as independent competing risks. This competing risks framework is next extended to encompass defective risks, which are then allowed to depend on the characteristics of the individual while allowing for gamma heterogeneity of the "susceptible" subpopulation. The model is estimated via maximum likelihood methods

Our data are taken from the nationally representative Portuguese quarterly employment surveys (Inquérito ao Emprego), conducted by the National Institute of Statistics (INE) (Instituto Nacional de Estatistica). The sample period is 1992(2)-1997(4), the starting date being dictated by changes in survey design after the first quarter of 1992.

The quarterly employment survey has a quasi-longitudinal capacity. One sixth of the sample rotates out each quarter, allowing us to track transitions out of unemployment for up to five quarters, and hence pursue the conditional approach. Transition rates are obtained simply by identifying those unemployed individuals in the survey, and their elapsed duration in a given quarter, who move out of unemployment over the subsequent quarter. The destination states of previously unemployed workers can also be identified. For present purposes, we distinguish between the two destination states of employment and inactivity.

More technically the stock sampling basis of the employment survey provides backward recurrence times for the relevant labor market state. Information on forward recurrence times has thus to be inferred. Specifically, remaining duration of unemployment, conditional on elapsed duration, distributed as the entrant conditional density function (Lancaster, 1990).

The main lesson of this paper is that a substantial proportion of (Portuguese) long-term unemployment can be explained either by the failure of individuals to receive acceptable job offers or by their non-consideration of the inactivity option. We showed that some factors preempt options at the same time as they independently shape transition rates out of unemployment (i.e. for viable options). Defective risks are clearly manifested in cause-specific survival functions.

All of this is consistent with the conventional view of ossified European labor markets. The argument is that high firing costs not only decrease flows into unemployment but also strengthen the bargaining power of insiders. The result is a lower arrival rate of job offers, and higher unemployment duration to reestablish equilibrium (Blanchard and Portugal, 2001). In this setting, it is indeed likely that an important subset of the unemployed population (especially older individuals) will see their already slim chances of receiving acceptable job offers being reduced to zero.

We singled out for special attention the role of age and unemployment benefits. Each has statistically significant effects on hazard rates and defective risks. Age increases the proportion of those who will never receive acceptable job offers and symmetrically decreases the proportion of those active in the labor market. It also independently increases hazards into inactivity. Unemployment benefit effects are a little more tricky to the extent that even in Portugal one cannot be a recipient for ever. Subject to this caveat - although benefits can be received in one form or other for up to 5 years - benefits are associated with increases in the proportion of those who will never find work. Benefits also decrease the hazard rates into employment and inactivity among those for whom these options are not preempted.

In search of a more adequate specification of unemployment duration, modern duration analysis should not simply recognize alternative destination states (and here it would also seem profitable to consider a variety of employment options) but also explicitly incorporate defective risks. The relevance of the latter is most obvious in terms of understanding long-term unemployment and interpreting negative duration dependence. In general, if there are defective risks to begin with, and these are ignored in modeling unemployment duration, there is a bias toward a declining hazard function that results from the mixture of the two subpopulations. (In our case, there are in essence two hazard functions: one that declines before trending up, and another for which the hazard rate is zero. In conjunction, they produce a declining hazard.)

References

Blanchard, Olivier, and Pedro Portugal. (2001). What Hides Behind an Unemployment Rate: Comparing Portuguese and U.S. Unemployment *American Economic Review*, forthcoming.

Cox, D.R., and D. Oakes. (1985) *Analysis of Survival Data*. London: Chapman and Hall.

Flinn, Christopher J., and James J. Heckman. (1983). "Are Unemployment and Out of the Labor Force Behaviorally Distinct Labor Market States?" *Journal of Labor Economics* **1,** 28-42.

Lancaster, Tony. (1990) The Econometric Analysis of Transition Data. Cambridge: Cambridge University Press.

Yamaguchi, Kasuo. (1992) "Accelerated Failure-Time Regression Models with a Regression Model of Surviving Fraction: An Application to the Analysis of 'Permanent Employment' in Japan." *Journal of the American Statistical Association* **87,** 284-292.

Non- and Semiparametric Multiple Imputation

Marc Aerts
Limburgs Universitair Centrum, Universitaire Campus
B3590 Diepenbeek, Belgium
marc.aerts@luc.ac.be

Gerda Claeskens
Texas A&M University, College Station
Texas 77843, U.S.A.
gerda@stat.tamu.edu

Niel Hens, Geert Molenberghs
Limburgs Universitair Centrum, Universitaire Campus
B3590 Diepenbeek, Belgium
niel.hens@luc.ac.be, geert.molenberghs@luc.ac.be

1. Introduction

Datasets with missing values arise frequently in statistical practice. Population surveys inevitably face the problem of incomplete data, missing data create difficulties in quality of life studies, in cancer clinical trials, etc. There exist many ways to deal with missing data problems, ranging from the most naive one focusing on the complete cases only to well-defined parametric, semiparametric and nonparametric approaches.

A large part of the literature deals with missing covariate values. Here however we consider missing response data. We focus attention on smoothing methods to obtain multiple imputation estimators and this in a non-Bayesian framework.

The onset to the use of kernel methods for imputation of missing values was given by Titterington and Sedransk (1989), who used kernel density estimation in combination with a nonparametric bootstrap for imputing values. In their method, relationships between variables are not directly accounted for. For single imputation in a nonresponse setting, Cheng (1994) and Chu and Cheng (1995) used kernel estimators in a regression model. We make use of a nonparametric regression relationship between a partially observed response variable and a fully observed covariate to create multiple imputations for the missing data.

2. A local Imputation Method

Our approach is based on local imputation methods. Whereas multiple imputation is mainly regarded as a Bayesian technique (see Little and Rubin 1987, Rubin 1987), the proposed methods are essentially bootstrap based (see Efron 1994). The parameter of interest is a marginal parameter of an incompletely observed variable. The regression relationship with a completely observed variable is exploited to impute values for the missing items. Throughout, we assume an ignorable nonresponse mechanism.

We introduce two local bootstrap methods, a local resampling method which is fully nonparametric (not needing model specification), and a local semiparametric bootstrap method which still assumes that the conditional response distribution is correctly specified but which allows any smooth conditional mean response structure.

The method requires two smoothing parameters used in two different stages of the algorithm, a resampling step and an imputation step.

Asymptotic expressions for bias, variance and mean squared error are derived. These can be used to determine the optimal order of the sample size for both smoothing parameters. To get fully data-driven bandwidth selectors, a new adjusted jackknife procedure is presented. A simulation study and an example on quality of life data illustrate the performance and the applicability of the method.

References

Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random, *Journal of the American Statistical Association*, **89**, 81 - 87.

Chu, C.K. and Cheng, P. E. (1995). Nonparametric regression estimation with missing data, *Journal of Statistical Planning and Inference*, **48**, 85 - 99.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, **89**, 463 - 475.

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. J. Wiley & Sons, New York.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. J. Wiley & Sons, New York.

Titterington, D.M. and Sedransk, J. (1989). Imputation of missing values using density estimation, *Statistics & Probability Letters*, **8**, 411 - 418.

Repeated Measurement Designs for Self and Mixed Carryover Effects

Kasra Afsarinejad
AstraZeneca R&D Mölndal, Department of Biostatistics
S-43183 Mölndal, Sweden
Kasra.afsarinejad@astrazeneca.com

1. Two-Period Repeated Measurement Designs

In such diverse fields as agriculture, animal husbandry, food science, education, psychology, social engineering, marketing, medicine, pharmacology, and industry, researchers often perform experiments designed in such a way that each experimental unit is assigned more than once to a treatment, either different or identical. Such designs are known by different names in the literature: repeated measurement (crossover or changeover) designs.

In many situations due to lack of enough experimental units and for many other practical reasons, including substantial variability among units we need to use repeated measurement designs for assessments and the estimation of differences between the treatments under the study. But then for reasons beyond the control of the experimenter effects such as carryovers (residuals) or interactions will enter into the data. In some cases, it might be desirable to generate and measure such carryover effects. When repeated measurement designs are recommended because of their economic use of experimental material, then the estimation or elimination of these effects becomes a secondary and often important aspect of the design and its related analysis. Of course, we should make every effort to avoid these undesirable effects in the design stage. But this might not be possible in all cases. Therefore, we need to identify efficient designs, which allow the study and the elimination of these effects.

We assume t ($\geq$2) treatments are to be studied utilizing n experimental units. Each unit is to be used in two periods. These two periods are the same for all n units. In the absence of missing observations the design will yield 2n observations. Each unit can be given the same treatment or different treatments in two periods. Thus, we are allowed to select n sequences from t^2 possible sequences of two treatments each. There is no restriction that these n sequences must be distinct. Thus, our design problem is which n sequence give us the best design and how should we analyze the related data. Clearly, the choice will depend on the model of observations and the goal we expect to achieve from the study. Throughout the paper, we use d to denote the design and d (i, j) to denote the treatment being assigned to unit j in period i, i = 1, 2; j = 1, 2, …, n.

The Model If d (1, j) = k, then the model of response for the observation, y_{1jko} , collected on unit j in period 1 is postulated to be:

(1) $$y_{1jko} = \mu + \pi_1 + \beta_j + \tau_k + e_{1jko},$$

where, μ is the general mean, π_1 is the effect of period 1, β_j is the effect of unit j, and τ_k is the direct effect of treatment k. These are unknown constants. e_{1jko} is the only random (noise) component of the model which is assumed to be distributed as normal with mean zero and variance σ^2. All these n observations are independent. Note that we use y_{1jko} rather than y_{1jk} to signify that there is no carryover effect on this

observation. However, for the n observations collected in the second period two cases arise:

Case 1 If d (2, j) = d (1, j) = k, then the model of response for y_{2jkk} is:

(2) $$y_{2jkk} = \mu + \pi_2 + \beta_j + \tau_k + \rho_k + e_{2jkk,}$$

where, π_2 is the effect of period 2 and ρ_k is the carryover effect of treatment k from period 1 on itself in period 2. This carryover effect is called the *self-carryover effect* of treatment k.

Case 2 If d (2, j) = k, d (1, j) = *l*, k ≠ *l*, then the model of response for y_{2jkl} is:

(3) $$y_{2jkl} = \mu + \pi_2 + \beta_j + \tau_k + \rho_l^* + e_{2jkl} ,$$

here ρ_l^* is the carryover effect of treatment *l* on treatment k in period 2. Note that this carryover effect is the same for all k ≠ *l*. This carryover effect is called the *simple mixed carryover effect* of treatment *l*. Self and simple mixed carryover effects are assumed to be constants. The error e_{2jkl} is assumed to be normal with mean zero and variance σ^2. The observations on different units are independent but the two observations on each unit are allowed to be correlated with cov (y_{1jko} , y_{2jkk}) = cov (y_{1jko} , y_{2jkl}) = $\delta\ \sigma^2$. If possible, we should plan the study so that δ becomes positive.

Therefore, our model could have up to 5+3t+n unknown parameters. However, the t direct treatment effects and σ^2 are primary parameters of the study. Note that we could entertain a more general mixed carryover effect. For example, we could assume that the mixed carryover effect of a treatment in period 1 on the treatment in period 2 depends on what treatment was applied in period 2. However, we believe this might unnecessarily overparameterize and often saturate or even worse super saturate the model.

Study and the estimation of self and simple mixed carryover effects are very important in many fields including, but not limited to, medicine and life sciences. For example, in a single drug therapy for a chronic disease it will be very helpful to the physician to know the magnitude of the self-carryover effect of the drug the doctor is recommending for her patient. Or, for arranging the best crop rotation schedule it will be very useful to know the size and the impact of the simple mixed carryover effects.

To minimize the possible confusion about the carryover effects and for other practical reasons such as compliance in clinical trials two-period repeated measurement designs are very popular designs in practice. However, to avoid confounding and its related statistical problems for two-period designs we should not limit ourselves while choosing sequences of treatments. In this article we study, for the first time, two-period repeated measurement designs for comparing two or more treatments allowing the appearance of two types of carryover effects namely self-carryover and simple mixed carryover effects. It is shown that if we properly design the study then unbiased and efficient estimates of all contrasts in direct treatment effects can be obtained. If we carefully design the study we can eliminate simple mixed carryover effects in the analysis stage. We have shown that it is easy to design the study if we are also interested in optimally estimating contrasts in self-carryover effects. It is also pointed out that if we are not careful in our design we might partially or fully confound the design yielding damaged or unusable data. Finally, we have shown the structure of the smallest designs that allow unbiased estimation of independent contrasts in simple mixed carryover effects.

Bias Reducing Estimators in an Exponential Distribution

M. Masoom Ali
Ball State University, Department of Mathematical Sciences
Muncie, IN 47306 USA
mali@bsu.edu

Jungsoo Woo
Yeungnam University, Department of Statistics
Gyongsan, South Korea
jswoo@yu.ac.kr

1. Introduction

We consider the parametric estimation in an exponential distribution with *pdf* given by,

$$f(x;\theta)=\frac{1}{\theta}e^{-\frac{x-\theta}{\theta}}, \quad 0<\theta<x,$$

which has mean 2θ and variance θ^2. We consider the jackknife and other point and interval estimates of the location and scale parameters and also the estimates of the right tail probability of the above exponential distribution with equal location and scale parameters.

2. Estimation of θ and θ^2

We consider the following estimates of θ in the above exponential distribution where the location and scale parameters are both equal.

$\hat{\theta}_1 = X_{(1)}$, $\hat{\theta}_2 = \frac{n}{n+2}X_{(1)}$, $\hat{\theta}_3 = \frac{1}{2n}\sum_{i=1}^{n}X_i$, where the estimator $\hat{\theta}_2$ and the method of moments estimator $\hat{\theta}_3$ are both unbiased. The ordinary jackknife estimator of θ_1 (Gray and Schucany (1972)) can be written as $J(\hat{\theta}_1)=\frac{2n-1}{n}X_{(1)}-\frac{n-1}{n}X_{(2)}$. From the fact that $E(\hat{\theta}_1)=\theta+\frac{\theta}{n}$, a bias reducing estimator $\hat{\theta}_4$ can be defined by $\hat{\theta}_4=\hat{\theta}_1-\frac{\hat{\theta}_1}{n}=\frac{n-1}{n}X_{(1)}$. Considering the mean square errors of these five estimators, we observe the following fact:

Fact 1 MSE $(J(\hat{\theta}_1))$ > MSE $(\hat{\theta}_3)$ > MSE $(\hat{\theta}_1)$ > MSE $(\hat{\theta}_2)$ > MSE $(\hat{\theta}_4)$.

From the maximum likelihood estimate $\hat{\theta}_1$ and the bias reducing estimator $\hat{\theta}_4$ of θ, we define an estimator $\hat{\theta}_{1k}$ in the following manner. Let

$$\hat{\theta}_{10}=\hat{\theta}_1=X_{(1)}, \text{ and } E(\hat{\theta}_{10})=\theta+\frac{1}{n}\theta,$$

$$\hat{\theta}_{11}=\hat{\theta}_4=(1-\frac{1}{n})X_{(1)}, \text{ and } E(\hat{\theta}_{11})=\theta-\frac{1}{n^2}\theta,$$

$$\hat{\theta}_{12} = (1 - \frac{1}{n} + \frac{1}{n^2}) X_{(1)}, \text{ and } E(\hat{\theta}_{12}) = \theta + \frac{1}{n^3}\theta .$$

By induction we can define an estimator of θ as follows:

$$\hat{\theta}_{1k} = \sum_{i=0}^{k} \frac{(-1)^i}{n^i} X_{(1)}, \quad k = 0, 1, 2, 3, \ldots .$$

We observe the following Fact 2.

Fact 2 (a) $E(\hat{\theta}_{1k}) = \theta + \frac{(-1)^k}{n^{k+1}}\theta$, $k = 0, 1, 2, \ldots$

(b) $Var(\hat{\theta}_{1k}) = \sum_{i=0}^{k} \sum_{j=0}^{k} \frac{(-1)^{i+j}}{n^{i+j+2}}\theta^2$, $k = 0, 1, 2, \ldots$

(c) $\lim_{k\to\infty} \hat{\theta}_{1k} = \hat{\theta}_2$

(d) By numerical evaluations, we find that the bias reducing estimate $\hat{\theta}_{11} = \hat{\theta}_4$ is better than $\hat{\theta}_{1k}$ for $k = 0, 2, 3, \ldots$.

We also considered several confidence intervals based on these point estimators of θ and observe that the confidence interval based on $\hat{\theta}_4$ has the shortest expected length.

In a similar fashion, we observe that the bias reducing estimator $\hat{\theta}_{11}^2 = (1 - \frac{2(n-1)}{n^2}) X_{(1)}^2$ of θ^2 performs better than the other proposed estimators of θ^2 in the sense of the mean square error.

3. Estimation of the Right-Tail Probability

For the exponential distribution in Section 1, the right-tail probability *R(t)* is given by $R(t) = e^{-\frac{t-\theta}{\theta}}$, $0 < \theta < t$. We consider four estimators of *R(t)* based on the various point estimates of θ. The estimator based on the bias reducing point estimator $\hat{\theta}_4$, namely,

$$\hat{R}_4(t) = \exp[-\frac{t - \frac{n-1}{n} X_{(1)}}{\frac{n-1}{n} X_{(1)}}], \quad t > \frac{n-1}{n} X_{(1)},$$

is observed to have less mean square error that the other three estimators when *R(t)* = 0.01, θ = 0.5 and the sample size *n* = 10(5)30.

Reference

Gray, H. L. and Schucany, W. R. (1972). The Generalized Jackknife Statistics. Mercel Dekker, Inc. New York.

Modelling and Performance Evaluation of a Broadband Wireless Network*

Nelson Antunes
Universidade do Algarve, Área Dept. de Matemática e Centro de Matemática Aplicada
Campus de Gambelas, 8000 Faro, Portugal
nantunes@ualg.pt

António Pacheco, Rui Rocha
Instituto Superior Técnico, Dept. de Matemática e Centro de Matemática Aplicada
Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
apacheco@math.ist.utl.pt, rmr@digitais.ist.utl.pt

Broadband wireless networks are seen today as one of the key factors for the success of the global communication infrastructure in the near future. Their design, planning and control must be supported by suitable traffic models that are able to deal with new sets of constraints in which the quality of service (QoS) management and mobility play an important role.

Traffic models for wireless networks that have been proposed in the literature may be classified according to their dimension. One-dimensional models consider a street or a highway and are reviewed in Antunes *et al* (2001). Two-dimensional models are proposed for scenarios where mobility in multiple directions in the plane is considered. These models are mainly focused on mobility aspects as the circuit switching characteristics of today cellular systems impose special constraints on the teletraffic component Antunes.*et al* (1998) On the contrary, third and fourth generation networks, will support fixed and variable rate transmission.

In this paper we propose a new traffic model for broadband wireless networks that allows fairly general mobility behaviour and describes the changes in the bandwidth requirements over the duration of each connection. It can be applied to both sub-urban and urban scenarios, as well as linear scenarios.

The model considers a finite set of cells, called a cluster, where mobiles arrive according to a Poisson process. The mobile users move between positions, independently of each other, according to a Markov renewal process, which allows for non-exponential holding times in a cell. We note that simulation studies revealed that the holding time in a cell is not exponential and showed that the generalized gamma distribution can be a good approximation for the cell residence time with random movement Zonnozi and Dassanayake (1997) and D. To restrict the movement of mobiles to existence paths or roads in the cluster, each position is modelled as a pair of cells representing the former and the current cell location of the mobile. According to Bacceli and Zuyev (1997), Su and Gerla (2000), it is more realistic to consider that mobiles move according to the presence of paths (highways, streets, roads and walking paths) than with complete random mobility.

For the teletraffic process, we assume that a Markov modulated fluid process Pacheco and Prabhu (1996) (MMFP) describes the individual bandwidth of a mobile during its presence in the system. The state space of the MMFP is the union of $K+1$ disjoint sets, $A_0, A_1, ..., A_K$, where K is the total number of types of calls. The set A_0

* This research was partially supported by Fundação para a Ciência e a Tecnologia, the Project PRAXIS/P/MAT/10002/1998 ProbLog, the SAPIENS Project CPS/34826/99-00 SCALE and the SAPIENS Project 40004/2001 TOWN initiative.

represents inactive states (without call) and the set A_k ($1 \le k \le K$) represents the states associated to the traffic class k.

For each type of call state or traffic class, we investigate the transient and limit number of mobiles in different cells, which turn out to be independent Poisson random variables. Moreover, we conclude that the handoff processes between cells of the cluster are the sum of independent Poisson cluster processes, and we characterize their univariate distributions through generating functions. If mobiles make at most one visit to a cell coming from another cell, then the handoff process between the two cells is a non-homogeneous Poisson process. As a consequence, the handoff process from a border cell to the outside in the general model and all the handoff processes between cells in the highway model per traffic class are non-homogeneous Poisson processes. Our analysis uses intensively the theory on translations and thinnings of Poisson processes along with the theory of Markov renewal and semi-Markov processes.

The performance analysis for network planning involves a longer time scale. Hence for capacity planning we propose the use of limit results. Since our model does not assume capacity constraints, we use the sample-path analysis version of Palm probabilities to approximate the handoff and new call blocking probabilities.

Unlike network planning, connection admission control and congestion control involve a short-time scale requiring a transient analysis. A key to network control is to exploit the detailed knowledge of the network state in order to obtain good estimates of the mean and variance of the demand in the near future. We assume that we know the number of mobiles in the cluster and, for each mobile: the position, the elapsed time in that position and the teletraffic state. Using the Lindeberg-Feller central limit theorem for non-identically-distributed summands we predict the total required bandwidth at some future time in a cell.

Finally, results of simulation studies are used to validate the theoretical analysis, leading to the following conclusions: for blocking probabilities below 3% the accuracy of the analytical results is very good; and for a large number of mobiles in the system, the estimation of the required capacity in a near future can be accurately done through the use of a normal approximation.

References

Antunes, N., Pacheco, A. and Rocha, R. (2001). Traffic modelling for broadband wireless networks: the highway scenario. Advances in Performance Analysis (In Print).

Antunes, N., Rocha, R., Pinto, P. and Pacheco, A. (1998). Impact of next-generation wireless networks requirements on teletraffic modelling. *Interoperable Communication Networks*, **1** (2-4): 706-715, Baltzer Science Publishers, 1998.

Baccelli, F. and Zuyev, S. (1997). Stochastic geometry models of mobile communication networks. In J.H. Dshalalow, ed., Frontiers in Queueing: Models and Applications in Science and Engineering, p. 227-243. CRC Press, Boca Raton, Florida.

Pacheco, A. and Prabhu, N.U. (1996). A Markovian storage model. *The Annals of Applied Probability*, **6** (1):76-91.

Su, W. and Gerla, M. (2000). Bandwidth allocation strategies for wireless ATM networks using predictive reservation. *In Proc. INFOCOM-2000*, p. 1400-05, Telavive, Israel.

Zonnozi, M. and Dassanayake, P. (1997). User mobility modeling and characterization of mobility patterns. *IEEE Journal of Selected Areas in Communications*, **15** (7):1239-52.

Time Series Estimation through Optimal Filtering: Non-Gaussian Series

Xavier Font Aragonés
Escola Universitària Politècnica de Mataró
Av. Puig I Cadaflach 101-111, 08303 Mataró - Spain
font@eupmt.es

Pilar Muñoz Gràcia, Manuel Martí Recober
Universistat Politècnica de Catalunya - DEIO
Pau Gargallo, 5, 08071 Barcelona - Spain
pilar.munyoz@upc.es; manuel.marti-recober@upc.es

1. Introduction

The state space representation and its application on time series analysis provides us with new ways for analysing time series. From pioneering first research by Akaike (1974) to most recent work Maravall (1994), Kitagawa (1998) or Durbin (1998), researchers have worked based on state space representation and have tried to model situations where time series go from simple stationary situations to the most complex non-stationary, non-linear and non-gaussian situations. In our paper different time series have been modelled through state space representation. A similar approach in gaussian and stationary ARMA time series models proposed Muñoz(1988).

(1) $$x_t = F\, x_{t-1} + G\, v_t$$

(2) $$y_t = H\, x_t + w_t$$

Our work is based on the usual state space representation characterized by two equations: state equation (1) and observation equation (2). Both non-restricted in terms of the density function associated with the random variable (v_t) and (w_t), respectively. Both equations allow us to express our filtering problem in terms of density functions. We are interested in the following densities:

(3) $$P(x_t/Y_t) = K\, P(y_t/x_t)\, P(x_t/Y_{t-1})$$

(4) $$P(x_t/Y_{t-1})$$

The first one is known as the filtering density (3) and the second one prediction density (4). Note that Y_t expresses the observation vector up to t and that K is a normalized constant.

2. Methodology

On the one hand, in order to obtain the complex densities (3) and (4) related to each time instant, we introduce a simple numerical method. The filtering densities are obtained using an approximation to the real density. This approximation procedure defines some nodes over the state domain where each density will be evaluated.

In fact, any density is taken as discrete and is defined in the same set of points as the nodes mentioned above. With 25 or 50 nodes we achieved enough accuracy in simple-model estimation.

On the other hand, the parameters of our model will be reached by applying the usual likelihood approach. This means maximizing the log likelihood function defined as:

(5) $$l(\theta) = \log \; p(y_1,...,y_N) = \sum_{i=1}^{N} \log \; p(y_i / y_1,...,y_{i-1}) = \sum_{i=1}^{N} \log \; p(y_i / Y_{i-1})$$

If we have an analytical expression for the log likelihood (5), we just need to define the partial derivatives and apply a well-known quasi-newton optimization method. Sometimes, it is not easy to obtain the partial derivatives in non-gaussian and non-linear situations. In such situations we use non-derivative methods.

3. Simulations and Results

In situations where there is no doubt about the results we have proved our methodology. This simulation comes from stationary, linear and gaussian time series and is defined by its easier model AR(1). Then we apply the methodology to a more sophisticated and critical situation, typically known as unit root estimation, and also to a AR(1) model without gaussian error. Statistical software package has been used to compare results.

Our last simulation comes from non-linear, non-stationary and non-gaussian time series. We take the Kitagawa's (1987) example of Tokyo rain falls to illustrate the type of approach where time series is discrete and obviously non-gaussian.

4. Conclusions

Our approach reduces significantly the complexity related to evaluate the filtering densities in each time instant. This approach is non restrictive in three ways: Firstly, we do not need to impose a gaussian density or make any linear or stationary assumption to apply the methodology. Secondly, it is not necessary to obtain an analytical and close expression for the partial derivatives of our cost function (likelihood function). Thirdly, the computational cost and the calculation time are severely reduced.

References

Akaike, H (1974) "Markovian Representation of Stochastic Processes and Its Applications to the analysis of autoregressive moving average process" *Anals of the Institute of Statistical Mathematics*, **26**, 236-387.

Durbin, J, Koopman SJ (1998) "Time Series analysis of non-gaussian observations based on state space models from both classical and bayesian perspective" *Center for Economic Research - Discussion paper*

Gómez, V ; Maravall A (1994) "Estimation, Prediction, and interpolation for nonstationary series with the Kalman Filter" *Journal of the American Statistical Association V89 n426*.

Muñoz, MP (1988) "Estimació del pol i de la variància del soroll d'un model AR(1) mitjançant filtratge no lineal" Qüestió Vol12 no 1 pp 21-42

Kitagawa,G (1998)"A self-Organizing State-space model" *American Statistical Association*. Vol **93**, No 443 Sept. 1998.

Kitagawa,G. (1987) "Non-gaussian state-space modeling of nonstationary time series" *America Statistical Association* Vol **82** No400 .

A New Approach to Bayesian Sensitivity Analysis

J. Pablo Arias, Jacinto Martín, Carlos J. Pérez
Universidad de Extremadura, Department of Mathematics
Carretera de Trujillo, s/n, Cáceres, Spain
{jparias,jrmartin,carper}@unex.es

1. Introduction

Robust Bayesian analysis studies the sensitivity of the results of Bayesian analysis with respect to the inputs of the analysis, mainly the prior distribution and the loss function. Usually, the imprecision in beliefs and preferences are modelled by a class Γ of prior distributions and a class Z of loss functions. The most commonly used measures in global sensitivity analysis are the range of the bayes actions and the range of the posterior expected loss of the bayes action for a fixed pair (loss, prior), see Berger *et al* (2000).

The usual interpretation of the range is that the quantity of interest is robust when the range is small, and is not robust when the range is large. This interpretation is quite intuitive; however it is very much problem-dependent: it is not clear what is meant by "large" and "small", see Ríos *et al* (2000). The following example shows this shortcoming:

Example 1 Suppose that we have a decision problem with precision in the loss function $L(a,\theta)=(a-\theta)^2$.

1. If we considered a class of prior distributions so that its posterior means μ are in [0,5] and its posterior variances σ^2 are in [0.1,0.2], the range of the bayes alternatives could be considered quite large, (5). Moreover, the range of the posterior expected loss for a=2.5 is 6.35.

2. If we considered a class of prior distributions, so that its posterior means μ are in [0,5] and its posterior variances σ^2 are in [100.1,100.2], the range of the bayes alternatives could be considered, again, quite large, (5). Moreover, the range of the posterior expected loss for a=2.5 is 6.35, too.

Apparently the two classes could consider themselves equal, i.e., either both robust or both nonrobust. Moreover, since, for all prior distribution π the expected loss of a is $(a-\mu_\pi)^2+\sigma_\pi^2$ and the expected loss for π–bayes action is σ_π^2, in both cases the greater difference between posterior expected loss of a=2.5 and posterior expected loss of bayes action μ_π is 6.25. It is achieved for μ_π=0 and σ_π^2=0.1 (among other values) in the first class and for μ_π=0 and σ_π^2=100.1 (among other values) in the second. For the first class the posterior expected loss of 2.5 is 6.35 and for the bayes action is 0.1. For the second class the posterior expected loss of 2.5 is 106.35 and for the bayes action is 100.1. Then, we think that the second class is less sensitive than the first one.

This example motivates the use of other sensitivity measures. We propose one in the following section.

2. Sensitivity Measure

Definition We define the sensitivity measure of a with respect to $\Gamma\times Z$ by

$$S(a) = \sup_{\pi\in\Gamma, L\in Z}\left(\frac{T(a,L,\pi)}{T(a_{(L,\pi)},L,\pi)}-1\right)$$

where

$T(a,L,\pi)=\int L(a,\theta)\pi(\theta\mid x)d\theta=\dfrac{\int L(a,\theta)f(x\mid\theta)\pi(\theta)d\theta}{\int f(x\mid\theta)\pi(\theta)d\theta}$, and $a_{(L,\pi)}$ the bayes actions for the pair (L,π)

$S(a)$ measures the maximum relative error between the expected posterior loss of the alternative a and the expected posterior loss of the bayes actions for any pair prior-loss function.

The smaller is $S(a)$, the less sensitivity will be found in alternative a respect to changes in the loss function or prior distribution, i.e. the selection of a as optimal alternative is quite good although there are imprecision in beliefs and preferences. A particular case of this measure can be found in Ruggeri and Sivaganesan (2000)

In an opposite way, we define a robustness measure $R(a)$.

$$R(a)=\inf_{\pi\in\Gamma,L\in Z}\left(\frac{T(a_{(L,\pi)},L,\pi)}{T(a,L,\pi)}\right)=\frac{1}{1+S(a)}$$

Example 2 (Cont. of Example 1) After some simple computations, we get for the first class $S(2.5)=62.5$ and for the second class $S(2.5)= 0.0624$. This means that in the first case we have an error of 6250% and in the second case the error is only 6.24%, showing the difference between robustness of the two classes. Figure 1 shows some posterior distributions for classes 1 and 2. Note than for class 2 there is no visual difference between distributions.

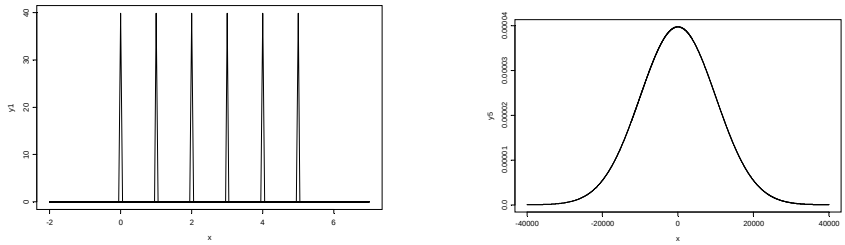


Figure 1. Posterior distributions for classes 1 and 2.

We apply our sensitivity measure to some classes from bayesian robustness literature, see e.g. Berger (1994).

This work has been supported by the grant number IPR00A075 from the Junta de Extremadura.

Reference

Berger, J. O. (1994). An overview of robust Bayesian analysis (with discussion), *Test.* **3**, 5-59.

Berger, J. O., Ríos Insua, D. and Ruggeri, F. (2000). Bayesian robustness. *In Robust Bayesian Analysis (eds D. Ríos Insua and F. Ruggeri)*, **1-32**. New York: Springer-Verlag.

Ríos Insua, D., Ruggeri, F. and Martín, J. (2000). Bayesian sensitivity analysis: a review. *In Sensitivity Analysis (eds A. Saltelli et al.)*, chapter. **10**. New York:Wiley.

Ruggeri, F. and Sivaganesan, S. (2000). Global sensitivity measure, *Sanky□: The Indian Journal of Statistics,* vol **62**, Serie A, Pt. 1, 110-127.

Successive Sampling for the Ratio of Population Parameters

Eva M. Artés Rodríguez, Amelia-V García Luengo
University of Almería, Departament of Statistics and Applied Mathematics
Spain
eartes@ual.es, amgarcia@ual.es

1. Introduction

Usually, in many national sample surveys, information collected regularly on the same population from one period to the next. In such repetitive surveys, three possible sampling procedures may be used:

1. Extracting a new sample on each occasion (repeated sampling).
2. Using the same sample every occasion (panel sampling).
3. Performing a partial replacement of units from one occasion to another (sampling on successive occasions, which is also called rotation sampling when the units are constructed in the number of stages in which they are to become part of the sample, as it happens with the EPA -Spanish Survey of Working Population- which are performed quarterly, and most of the family surveys carried out by the INE -Spanish Statistics Institute-).

The third possibility, has been discussed extensively by several authors in the case of estimating the population mean (total) (Rao and Mudholkar, 1967; Artés and García, 2000). However, in many of such repetitive surveys, the estimate of the population ratio and product of two characters for the most recent occasion may be of practical interest.

We build the optimum estimator of the ratio population means at the second occasion, $\hat{R}_{(i)2}$, $(i=1,2,3)$. For the matched portion an estimate improved may be obtained using a double sampling estimate for the ratio of two means $\hat{R}_{(i)2m}$, $(i=1,2,3)$, and a simple ratio estimate, $\hat{R}_{2u}$, based on the unmatched part of the sample on the second occasion, with weights Q and $1-Q$, that minimize $V\left(\hat{R}_{(i)2}\right)$, $(i=1,2,3)$. Thus $\hat{R}_{(i)2}=Q\hat{R}_{2u}+(1-Q)\hat{R}_{(i)2m}$, $i=1,2,3$ and

(1) $$V_{\min}\left(\hat{R}_{(i)2}\right)=\frac{R^2}{n}\left(\mathrm{C}_1^2+\mathrm{C}_2^2-2\rho_0\mathrm{C}_1\mathrm{C}_2\right)\frac{1+\mathrm{qZ_i}}{1+\mathrm{q}^2Z_i}\quad i=1,2,3$$

with $Z_1=\dfrac{C_0^2(1+A)}{\left(\mathrm{C}_1^2+\mathrm{C}_2^2-2\rho_0\mathrm{C}_1\mathrm{C}_2\right)}$. $Z_2=\dfrac{C_0^2(1-A)}{\left(\mathrm{C}_1^2+\mathrm{C}_2^2-2\rho_0\mathrm{C}_1\mathrm{C}_2\right)}$, $Z_3=\dfrac{\theta^2C_0^2\left(1-\frac{A}{\theta}\right)}{\left(\mathrm{C}_1^2+\mathrm{C}_2^2-2\rho_0\mathrm{C}_1\mathrm{C}_2\right)}$

where $A=2\left[\rho_1(C_1/C_0)-\rho_2(C_2/C_0)\right]$ and $\theta_{opt}=\rho_1(C_1/C_0)-\rho_2(C_2/C_0)$.

The optimum matching fraction is obtained minimizing in (1) with respect to u, and so, we have

$$p_{opt}=\frac{1+Z_i-\sqrt{1+Z_i}}{Z_i},\ i=1,2,3$$

We can compote the gain in precision, G, of the combined estimate $\hat{R}_{(i)2}$, obtained by using a double-sampling ratio estimate of means from the matched portion of the sample on the second occasion, over the direct estimate, $\hat{R}_2$

$$G = \frac{V\left(\hat{R}_2\right) - V\left(\hat{R}_{(i)2}\right)}{V\left(\hat{R}_{(i)2}\right)} = \frac{-Z_i p(1-p)}{1+(1-p)Z_i};\ i = 1,2,3.$$

Necessarily $p \le 1$. If $p = 1$ (perfect matching) or $p = 0$ (no matching), the gain is zero. For other $p(0 < p < 1)$, there will be positive gain for $\hat{R}_{(1)2}$ if $A < -1$, for $\hat{R}_{(2)2}$ if $A > 1$, and for $\hat{R}_{(3)2}$ if $A < \theta < 0$ or $0 < \theta < A$.

The theory has been applied to provide more accurate estimations of the analysed variables over a study on schoolchildren's health habits and fitness carried out in Almeria schools. From these data we can state that

COMPARISON OF ESTIMATORS

Estimators	% Gain in precision
Direct $\hat{R}_{(1)2}$	-30.06%
$\hat{R}_{(2)2}$	2.04%
$\hat{R}_{(3)2}$	26.51%

References

Artés E. M. and García, A. V. (2000), A note on successive sampling using auxiliary information. *Proceedings of the 15th International Workshop on Statistical Modelling*, 376-379.

Rao, P.S.R.S. and Mudholkar, G.S. (1967), Generalized multivariate estimators for the mean of finite populations, *Journal of the American Statistical Association*, **62**, 1008-1012.

Ray, S. K. and Singh, R. K. (1985), Some Estimators for the Ratio and Product of Population Parameters, *J. Ind. Soc. Agri. Stat*, **37** (1), 1-10.

The Fan Plot in the Forward Search for Transformations in Regression and the Distribution of the Test Statistic for Transformation

Anthony Atkinson
London School of Economics, Department of Statistics
Houghton Street, London WC2A 2AU, UK
a.c.atkinson@lse.ac.uk

Marco Riani
Università di Parma, Dipartimento di Economia
Via J. Kennedy 6, 43100 Parma , Italy
mriani@unipr.it

1. Introduction

This paper is concerned with the distribution of the approximate score statistic for power transformation of the response in regression that was introduced by Atkinson (1973). This statistic is the *t* test for regression on a constructed variable. Since the constructed variable is a function of the response, the null distribution of this statistic is not exactly Student's *t*. The simulation study of Atkinson and Lawrance (1989) shows that, although the distribution of the statistic is roughly symmetrical, the variance is too large. However the extent of this variance inflation is different in different examples. In the most extreme case it is roughly two rather than one, an effect which is large enough to give misleading interpretation of the observed value of the statistic. The results in this note characterise conditions under which good, or less good, agreement with the null *t* distribution can be expected.

2. The Forward Search

We are concerned with the Box-Cox parametric family of transformations for positive data, for which, for example, a value of one for the parameter λ indicates no transformation, whereas zero indicates the log transformation. Information on the need for a transformation often comes from the largest or smallest observations. As a result, the estimated transformation, or equivalently, the value of $T(\lambda)$, the approximate score statistic for λ, can be misleading if one or more outliers are present. The presence of a single outlier is revealed by the constructed variable plot. If $z(\lambda)$ is the transformed response for a particular λ and $w(\lambda)$ the constructed variable for the transformation, the constructed variable plot is that of residual $z(\lambda)$ against residual $w(\lambda)$, both residuals being from regression on the variables X in the regression model. However, Atkinson and Riani (2000, Chapter 4) show that the constructed variable plot will fail if more than one or two outliers are present. Instead they suggest a forward search through the data, from which they produce a "fan plot" of values of $T(\lambda)$ which leads to a clear indication of the presence and effect of multiple outliers.

The forward search starts from a robustly selected subset of the observations which is outlier free. The size *m* of the subset used in fitting the data is augmented by one and the model refitted, each growth in sample size selecting the $m + 1$ observations with the smallest residuals. The procedure continues until $m = n$, the number of observations. As a result of the structure of this search the outliers, if any, enter at the end, that is as *m* approaches *n*. The plot of the statistic $T(\lambda)$ during this

search indicates the effect of individual observations on the evidence for a transformation. Outliers often cause sudden jumps in the value of the statistic. Both the order of observations produced by the search and the values of $T(\lambda)$, depend on the value of λ used in the search. Usually the results for five values of λ are plotted on the same graph. The evidence of wrong values of λ increases as observations are added during the search, giving plots $T(\lambda)$ which move steadily away from zero, often some in positive and some in negative directions. Because of the shape of the resulting five curves, this picture is called a "fan plot".

3. Tests and the Fan Plot

Given the ordering of the observations in the forward searches on which the fan plot is based, it is not clear what is the null distribution of $T(\lambda)$ in the plot. The simulations of Atkinson and Riani (2001) show that, for much of the search, the *t* distribution provides a good guide to significance, the simulated distribution rapidly becoming indistinguishable from the normal as *m* increases. However, at the end of the search, two interesting effects are observed which account for the distributional properties noted by Atkinson and Lawrance when $m = n$. It is this last part of the search which is important for the detection of influential observations.

The first effect is a lengthening of the tails of the distribution in the last few steps, so that the graph of the estimated percentage points diverges with *m* like the bell of a trumpet. The simulations show that this effect increases as the value of the multiple correlation coefficient R^2 decreases, being negligible when strong regression is present and strongest for a simple random sample. The explanation lies in the changing structure of the constructed variable plots during the search, which are highly structured for simple samples, but seemingly random if strong regression is present. This reflects the dependence of the correct transformation on the structure of the linear model, if any, as well as on the distribution of errors.

The second effect which is also sometimes present is that of skewness in the simulated distribution. This is caused by the need to reject non-positive values of the simulated response. Although the distribution of the transformed observations cannot, in general, be exactly normal, it should involve a negligible truncation of negative values. That these occur sufficiently often to have an effect on the distribution of $T(\lambda)$ is an indication that a model has been fitted to the data for which the predictive distribution has an appreciable probability of generating negative observations. Such a model cannot be correct for observations that should be positive and is an indication that a satisfactory model has not been found. This difficulty does not arise with the lognormal model ($\lambda = 0$), where all predictions are positive when exponentiated back to the original scale.

References

Atkinson, A.C. (1973). Testing transformations to normality, *J. R. Statist. Soc*. **B 35**, 473-479.

Atkinson, A.C. and Lawrance, A.J. (1989). A comparison of asymptotically equivalent test statistics for regression transformation, *Biometrika* **76**, 223-229.

Atkinson, A.C. and Riani, M. (2000). Robust Diagnostic Regression Analysis. New York: Springer-Verlag.

Atkinson, A.C. and Riani, M. (2001?). Tests in the fan plot for robust, diagnostic transformations in regression. (Submitted).

Estimation of the Mean of a Wiener sheet

Sándor Baran
University of Debrecen, Institute of Mathematics and Informatics
Egyetem square 1., 4032 Debrecen, Hungary
barans@math.klte.hu

Gyula Pap
University of Debrecen, Institute of Mathematics and Informatics
Egyetem square 1., 4032 Debrecen, Hungary
papgy@math.klte.hu

Martien C. A. van Zuijlen
University of Nijmegen, Department of Mathematics
Toernooiveld 1., 6525 ED Nijmegen, The Netherlands
martien@sci.kun.nl

Let $\{W(s,t): s,t \geq 0\}$ be a standard Wiener sheet and consider the process $Z(s,t) = W(s,t) + m$ with an unknown parameter $m \in \mathbb{R}$. Let $[a,b] \subset (0,\infty)$ and let $\gamma : [a,b] \to \mathbb{R}$ be a continuous, strictly decreasing function with $\gamma(b) > 0$. Consider the curve $\Gamma = \{(s,\gamma(s)) : s \in (a,b)\}$ and the set

$$G = \{(s,t) \in \mathbb{R}^2 : a \leq s \leq b, t \geq \gamma(s) \text{ or } s > b, t \geq \gamma(b)\}.$$

Let $\tilde{G} \subset G$ be a subset containing an ε-strip of Γ, i.e.,

$$\{(s,t) \in \mathbb{R}^2 : s \in [a, a+\varepsilon], t \in [\gamma(s), \gamma(a)]\} \cup$$
$$\{(s,t) \in \mathbb{R}^2 : s \in [a+\varepsilon, b], t \in [\gamma(s), \gamma(s)+\varepsilon]\} \subset \tilde{G}$$

with some $\varepsilon > 0$. Arató, N. M. (1997) considered a twice continuously differentiable function γ on $[a,b]$ and determined the maximum likelihood estimator (MLE) of m based on the observation of $\{Z(s,t) : (s,t) \in \tilde{G}\}$ by the help of stochastic partial differential equations. We derive the MLE of m under less assumptions on the function γ applying direct discrete approach instead.

First we verify that for a partition $\Im : a = s_1 < s_2 < \cdots < s_{N-1} < s_N = b$ of $[a,b]$ the MLE of m based on any finite sample containing the observations

$$\{Z(s_i, \gamma(s_i)) : 1 \leq i \leq N\} \cup \{Z(s_i, \gamma(s_{i-1})) : 2 \leq i \leq N\}$$

and possibly also finitely many observations from

$$\bigcup_{i=1}^{N} \{Z(s,t) : s \geq s_i, t \geq \gamma(s_i)\}$$

has the form $\tilde{m}_N = \varsigma_N / A_N$, where

$$A_N = \sum_{i=1}^{N} \frac{1}{s_i \gamma(s_i)} - \sum_{i=2}^{N} \frac{1}{s_i \gamma(s_{i-1})},$$
$$\varsigma_N = \sum_{i=1}^{N} \frac{Z\left(s_i \gamma(s_i)\right)}{s_i \gamma(s_i)} - \sum_{i=2}^{N} \frac{Z\left(s_i \gamma(s_{i-1})\right)}{s_i \gamma(s_{i-1})}.$$

We prove that if we take a sequence of partitions $\Im_N : a = s_1^{(N)} < s_2^{(N)} < \cdots < s_{N-1}^{(N)} < s_N^{(N)} = b,\ N = 1,2,\ldots$ with $\max_{2 \le i \le N} \left(s_i^{(N)} - s_{i-1}^{(N)}\right) \to 0$ than the above MLE converges in L^2 norm to the MLE of m based on the observation of $\left\{Z(s,t) : (s,t) \in \tilde{G}\right\}$. This MLE is a weighted linear combination of the values at the endpoints $(a, \gamma(a))$ and $(b, \gamma(b))$ of the curve Γ and a weighted integral of the observed process and its normal derivative along the curve Γ.

References

Arató, N. M. (1997). Mean estimation of Brownian sheet, *Comput. Math. Appl.* **33**, 13-25.

Non Parametric Estimation of Multivariate Extreme Value Dependence

M. Isabel Barão*
University of Lisbon, Department of Statistics and Operational Research
Faculty of Sciences, Campo Grande, Edificio C2, Piso 2, 1749-016 Lisboa, Portugal
mibarao@fc.ul.pt

Jonathan Tawn
Lancaster University, Department of Mathematics and Statistics
Fylde College, LA1 4YF, Lancaster, U.K.
J.Tawn@lancaster.ac.uk

The theory and methods for multivariate extreme values are aimed at characterising and estimating features of the joint distribution of vector random variables in the tail region. Recently there has been much work developing parametric and non-parametric methods for estimating the characteristics of the dependence between extreme values. In this study we overview the two approaches, illustrating how advances in the parametric approach can be incorporated into the nonparametric approach. We illustrate how this leads to improved performance, especially in cases of weak asymptotic dependence.

An advantage of parametric methods for multivariate extreme value is their ability to adjust estimates of the marginal distributions to account for information in the other variables. This feature is particularly important when marginal outliers are present or when marginal observations are missing through a non-random mechanism. We will illustrate how a fully multivariate analysis helps in such problems.

References

Barão, M. I. and Tawn, J. A. (1999). Extremal analysis of short series with outliers: sea-levels and athletics records. *Appl. Statist*., **48**, 469--487.

Bruun, J. T. and Tawn, J. A. (2001). Comparison of parametric and non-parametric methods for bivariate extreme values. (In preparation).

Capéraà, P. and Fougères, A.-L. (2001). Estimation of a bivariate extreme value distribution. To appear in *Extremes*.

Coles, S. G. and Tawn, J. A. (1994). Statistical methods for multivariate extremes: an application to structural design (with discussion). *Appl. Statist.,* **43**, 1--48.

Einmahl, J., de Haan, L. and Sinha, A. K. (1997). Estimating the spectral measure of an extreme value distribution. *Stoch. Proc. Appl*., **70**, 143--171.

de Haan, L. and de Ronde, J. (1998). Sea and wind: multivariate extremes at work. *Extremes*, **1**, 7--45.

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

Smith, R. L. (1994). Multivariate threshold methods. In *Extreme Value Theory & Applications*, 225--248, eds. J. Galambos,

* Research partially supported by FCT/POCTI/FEDER

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Non Asymptotic Minimax Rates of Testing in Signal Detection

Yannick Baraud
Ecole Normale Supérieure, DMA
45 Rue D'Ulm, 75230 Paris Cedex 05
yannick.baraud@ens.fr

We consider the statistical model given by

(1) $$Y_i = f_i + \sigma\varepsilon_i,\ i \in \mathbb{N}^* = \mathbb{N}\setminus\{0\}$$

where $f = (f_i)_{i\geq 1}$ is an unknown sequence of real numbers (called the signal), or a σ positive number and the ε_i's a sequence of i.i.d. standard Gaussian random variables. One observes the sequence $Y = (Y_i)_{i\geq 1}$. Let $\varepsilon_{a,p}(R)$ denotes the l_p-body

$$\varepsilon_{a,p}(R) = \left\{ f \in l_2(\mathbb{N}^*), \sum_{k\in\mathbb{N}^*} \left|\frac{f_k}{a_k}\right|^p \leq R^p \right\},$$

where R is a positive integer, p ome real number in $]0,2]$ and $(a_k)_{k\geq 1}$ a non increasing sequence of positive numbers converging towards 0. We consider the problem of testing "$f = 0$" against the alternative "$f \in \varepsilon_{a,p}(R)\setminus\{0\}$" and establish nonasymptotic bounds for the separation rate. Those bounds hold under no condition on the decay of the a_k's. The upper bound derives from the performance of a test described in Baraud, Huet, Laurent BHL which uses model selection technics. To establish the lower bound, we reduce to the case of an alternative of the form

$$\left\{ f \in l_2(\mathbb{N}^*), \left|\{i, f_i \neq 0\}\right| \leq D, \forall i > N\ f_i = 0 \right\},$$

where the integers D and $N\ (D \leq N)$ are judiciously calibrated, and use combinatoric arguments. Our results allow to recover asymptotic bounds established in the pioneer papers by Ermakov (91), Ingster (93a), Lepski and Spokoiny (99).

References

Baraud, Y., Huet, S. and Laurent, B. (1999). Adaptive tests of linear hypotheses by model selection. *Technical report*. **99-13**. Ecole Normale Supérieure.

Ermakov, M.S. (1991). Minimax detection of a signal in a Gaussian white noise. *Theory Probab. Appl.* **35** 667 -679.

Ingster, Yu. 1. (1993a). Asymptotically minimax hypothesis testing for nonparametric alternatives I. *Math. Methods Statist.* **2**, 85 - 114.

Ingster, Yu. 1. (1993b). Asymptotically minimax hypothesis testing for nonpararnetric alternatives II. *Math. Methods Statist.* **3** 171 - 189.

Ingster, Yu. 1. (1993c). Asymptotically minimax hypothesis testing for nonparametric alternatives III. *Math. Methods Statist.* **4** 249 - 268.

Lepskjj, O. V., and Spokoiny, V. G. (1999). Minimax nonparametric hypothesis testing: the case of inhomogeneous alternative. *Bernoulli* **5**, 333 - 358.

Age Regularities of Mitotic Clock and Mortality Index

Alexander Begun
MPI for Demographic Research
Doberaner Strasse 114, D-18057 Rostock, Germany
begun@demogr.mpg.de

Anatoli Yashin
MPI for Demographic Research
Doberaner Strasse 114, D-18057 Rostock, Germany
yashin@demogr.mpg.de

1. Introduction

The phenomenon of the exponential increase of the mortality rate in human population in the age interval 30-85 years was firstly described by Gompertz in1825. Later studies showed this regularity to be hold at suitable ages for a number of multicellular organisms. In spite of deceleration of the hazard in human after 85 and leveling-off or leveling-off in *Drosophila* it seems that the phenomena of exponential growth in mortality is closely related to the essence of senescence processes in the living organism. A number of approaches aim to explain exponential increase of adult mortality with age. One class of such models is based on the theory of extreme values and regards a living organism as a system of interacting elements performing different physiological functions. Gradual decline of their amount and capacities lead to increasing risk of the failure for the organism as a whole.

In the model proposed by Abernethy(1998) the living organism is regarded as a system components or clones of replicating cells dedicated to perform a specific vital function. All the mature cells evolve from a unique stem cell and the waiting-time for completion of the fixed-length string of replications is a random variable with a distribution function of exponential type. The members of a mature cell-pair, that is the outcome of completed string, discharge immediately their vital function. Since the number of replications is bounded, any given stem cell can produce a large but finite number of cell-pairs. The clone is failed when all but a small critical number q out of its large population of $n=2^{h-1}$ cell-pairs have been generated. The death of organism is associated with the failure of at least one clone. This model assumes the mitotic cycle as a random duration and terminated by a counting mechanism resident in the DNA of somatic cells. The counter is duplicated during mitosis and passed on to the daughter cells with its new-updated state. The counter terminates mitosis after $h+k$ cycles, where k is the replication number at which the embrionic cell mass differentiates into 2^k clones. If the times-to-replication-completion are asymptotically independent, the extreme values theory can be applied to this model and an exponential hazard $h(t)=\exp(-Rexp(\alpha t))$ for the life span is derived from this model. By unlimited cellular replicability $h(t)$ vanishes and by default death would be restricted to catastrophic events, accidents, or overwhelming infection, which typically are constant risks, independent on age.

Another way to get different forms of hazard curves is changing of the age-scale. In the next section we will show how to choose the age-scale to get any kind of a hazard.

2. Age-Scale Transform of the Waiting Time for Mitotic Events

The survival function corresponding to the Gompertz hazard can be represented in a form *W(t)=exp(-Rexp(αt))*. In principle, for any given continuous survival function *S(t)* one can find such an age-scale transform *x(t)* that *W(x(t))=S(t)*. Particularly, *x(t)*=ln*(*-ln*(S(t))/R)/α*. But what in reality may this age-scale transform mean on the cellular level? We have only one possibility for changing the age-scale. This possibility is in changing the exponential increase with age of the waiting time for mitotic events. Now we define such a transform of the waiting time for mitotic events with age, that lead to a given hazard *μ(t)* and supply our general formulae with some important examples. It can be proven that the hazard $\tilde{\mu}(t)$ for the modified waiting-time-for-completion-of-replication, corresponding to survival function *S(t)* is given by the function

$$\tilde{\mu}(t) \approx \frac{\ln\ln(-\ln S(t))}{\ln(-\ln S(t))(-\ln S(t))}\mu(t) \text{ as } t\to\infty.$$

As a corollary of this result we receive three important forms of hazard rates as t→∞.

1. Exponential (Gompertzian) hazard, $\mu(t) \approx \exp(t)$

$$t\tilde{\mu}(t) \approx \ln(t)\ ;$$

2. The leveling-off of a hazard, $\mu(t) \approx const$

$$t\tilde{\mu}(t) \approx \frac{\ln\ln(t)}{\ln(t)}\ ;$$

3. Weibull hazard, $\mu(t) \approx t^{\lambda}$,λ>-1,

$$t\tilde{\mu}(t) \approx \frac{\ln\ln(t)}{(\lambda+1)\ln(t)}\ .$$

The constant hazard can be regarded as a special case for Weibull hazard when λ=0. An exponential hazard corresponds to the exponential increase with age of the waiting time for mitotic events, $t \approx e^{\tau}$. In the case of leveling-off or of Weibull hazard the waiting time for mitotic events increases as double exponent, i.e. $t \approx e^{e^{\tau}/(\lambda+1)}$.

References

Abernethy, J.D. (1998) Gompertzian Mortality Originates in the Winding-down of the Mitotic Clock, *J.Theor.Biol*. **192**, 419-435.

Estimation of the Extreme Value Index and Regression on Generalized Quantile Plots

Jan Beirlant, Goedele Dierckx, Gunther Matthys
Universitair Centrum voor Statistiek
W. de Croylaan 52b, 3001 Heverlee, Belgium
Jan.Beirlant@wis.kuleuven.ac.be, Goedele.Dierckx@ucs.kuleuven.ac.be,
Gunther.Matthys@ucs.kuleuven.ac.be

Armelle Guillou
Université Paris VI, Laboratoire de Statistique Théorique et Appliquée,
Tour 45-55, E3, Boîte 158, 4 place Jussieu, 75252 Paris cedex 05, France
guillou@ccr.jussieu.fr

Let $X_1, X_2, ..., X_n$, be a sequence of independent and identically distributed random variables with distribution function F and with tail quantile function U defined by

$$U(x) = \inf\{y \ ; \ F(y) \geq 1 - 1/x\}.$$

Denoting the order statistics by $X_{1,n} \leq ... \leq X_{n,n}$, the basic statistical model considered here is given by the maximal domain of attraction condition which governs extreme value theory : Suppose that there exist sequences of constants $(a_n; a_n > 0)$ and (b_n), and some $\gamma \in \Re$, such that

(1) $$\lim_{n\to\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = \exp\left(-(1+\gamma x)^{-\frac{1}{\gamma}}\right) \text{ for all } x.$$

The main aim of this paper is to discuss the estimation problem of the extreme value index γ under this model. Most research in this area concentrates on the heavy tailed distributions with $\gamma > 0$. In such a case, it follows from (1) that X is of Pareto-type. Therefore, the tail function $U(x) = Q(1 - x^{-1})$ with Q the quantile function of F, is regularly varying, that is $U(x) = x^{\gamma} L(x)$ where L is a slowly varying function, i.e. satisfying $L(tx)/L(x) \to 1$ as $x \to \infty$, for all $t > 0$. For the estimation of γ under this regular variation model, Hill (1975) proposed the following estimator:

$$H_{k_n,n} = \frac{1}{k_n} \sum_{j=1}^{k_n} \log X_{n-j+1,n} - \log X_{n-k_n,n}.$$

Here (k_n) is a sequence of positive integers $(1 \leq k_n < n)$ which, in theoretical asymptotic considerations, satisfies the conditions $k_n \to \infty$ and $k_n / n \to 0$ as $n \to \infty$. It has been mentioned in literature that this and many other estimators can be viewed as estimators of the slope in a Pareto quantile plot. A large group of estimators of γ evolves from different possible regression fits on Pareto quantile plots. An important subclass of Pareto-type distributions was defined by Hall (1982) with

(2) $$L(x) = C_1(1 + C_2 x^{\rho}\{1 + o(1)\})$$

where $\rho < 0, C_1 > 0$ and $C_2 \in \Re$. The asymptotic nature of the definition of the Pareto-type model implies that any estimator will contain quantities, the selection of which plays a crucial role for successful application of such a technique: especially the choice of the number of extremes k_n has received a lot of interest. Minimizing the mean squared error of the estimation technique has been a constant guideline throughout almost all publications on this topic. However, next to the choice of k_n, the appearance of a systematic and important bias is considered to be a serious problem. This typically happens when $|\rho|$ is small in the Hall model discussed above. Recently, in Beirlant et al. (1999) and Feuerverger and Hall (1999), the regression problem defined by the upper subsets

$$\left(\log(n+1) - \log j, \log X_{n-j+1,n}\right), \ (j = 1,...,k),$$

of a Pareto quantile plot was further specified taking into account a second order slow variation condition, essentially specified by the Hall model (2). In this way, the bias is reduced and the problem of volatility in the plots of the estimates as a function of k can be weakened, while mean squared error rates comparable with those of the more classical estimators are retained.

The estimation of $\gamma \in \Re$ has been studied less extensively. Beirlant et al. (1996) proposed in this case an estimator based on a generalized quantile plot, which takes over the role of the Pareto quantile plot in this more general setting. As in the $\gamma > 0$ case, one can now construct several regression based estimators for $\gamma \in \Re$. All estimators exhibit bias problems for different underlying distribution functions. The aim of this paper is to perform the same program as in Beirlant et al. (1999) starting from the generalized quantile plot, inspecting the induced regression problem in more detail. This leads to new estimators for $\gamma \in \Re$, which follow from working with a full, respectively a reduced, regression model. The estimators resulting from the full model possess a smaller bias. Then, we derive the regression model which allows to attain our goal through least squares estimates of the parameters of the model. The basic asymptotic results are also given. The behaviour of the resulting novel estimates of γ is discussed with a small sample simulation study. Finally, we provide a diagnostic for selecting the number of extremes k when using the generalized Hill estimator following from the reduced regression model. This method is inspired by a new approach developed in Guillou and Hall (2000) for the Hill estimator itself.

References

Beirlant, J., Dierckx, G., Goegebeur, Y. and Matthys, G. (1999). Tail index estimation and an exponential regression model, *Extremes* **2**, 177-200.

Beirlant, J., Vynckier, P. and Teugels, J.L. (1996). Excess functions and estimation of the extreme value index, *Bernoulli* **2**, 293-318.

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. (2000). A diagnostic for selecting the threshold in extreme-value analysis, to appear in *J. Roy. Statist. Society* Ser. B.

Hall, P. (1982). On some simple estimates of an exponent of regular variation, *J. Roy. Statist. Society* Ser. B **44**, 37-42.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Ann. Statist.* **3**, 1163-1174.

A Frailty Model for Interval-Censored Observations

Bo Martin Bibby
The Royal Veterinary and Agricultural University, Institute of Mathematics and Physics
Thorvaldsensvej 40, DK-1871 Frederiksberg C, Denmark
bibby@dina.kvl.dk

Ib Michael Skovgaard
The Royal Veterinary and Agricultural University, Institute of Mathematics and Physics
Thorvaldsensvej 40, DK-1871 Frederiksberg C, Denmark
Ib.M.Skovgaard@imf.kvl.dk

We consider the mortality in m populations and suppose that it depends randomly on some covariates X. For each individual this dependence is assumed to be on two levels, namely an individual and a population level. More precisely, for the j'th individual in the i'th population we consider hazard functions of the form

$$\lambda_{ij}(t) = a(t) + (Z_i + Z_{ij}) \cdot b(t) \cdot g(\beta X),$$

where a is the natural death intensity. Here Z_i and Z_{ij} represent frailty on the individual and population levels and have means that sum to one.

This model is well studied in Yashin et al. (1995) and Petersen (1998), but only in the case where complete observations are available. We are mainly interested in interval censored observations, a very common type of incomplete observations in dose-time mortality studies. In theory it is straight forward to calculate the joint survival function for each population. However, in practice it involves summation of a very large number of survival probabilities if the population sizes are not very small. We discuss possible solutions to this problem.

As an example we consider a dose-time mortality experiment in which groups of ticks were subjected to different doses of an insect pathogen fungus. At 7 days in a 11 day period after the start of the experiment the number of dead ticks was observed. For this kind of experiment Nowierski et al. (1996) proposed to use log-dose as a covariate and an exponential link function g.

References

Nowierski, R.M., Zeng, Z., Jaronski, S., Delgado, F. and Swearingen, W. (1996). Analysis and modelling of time-dose-mortality of Melanoplus sanguinipes, Locusta migratoria migratorioides, and Schistocerca gregaria (Orthoptera: Acridae) from Beauveria, Metarhizium, and Paecilomyces isolates from Madagascar. *J. Invertebrate Pathology* **67**, 236-252.

Petersen, J. H. (1998). An additive frailty model for correlated life times. *Biometrics* **54**, 646-661.

Yashin, A. I., Vaupel, J. W., and Iachine, I. A. (1995). Correlated individual frailty: An advantageous approach to survival analysis of bivariate data. *Mathematical Population Studies* **5**(2), 145-159.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

An Empirical Approach to Model Uncertainty

John Bithell
University of Oxford, Department of Statistics
1 South Parks Road, Oxford OX1 3TG, UK
bithell@stats.ox.ac.uk

Model uncertainty is increasingly being recognised as an important component of our assessment of the strength of evidence in inferences drawn from statistical observations. Scientific assessments that merely report the latest confidence intervals to be published may reflect only a part of the uncertainty in our scientific conclusions.

Most approaches to model uncertainty proceed on the assumption that we have access to original data and typically involve a full Bayesian analysis from first principles. In practice, published results of investigations are often restricted to point and interval estimates of a particular parameter.

Accordingly we propose a simple method of using summary data in this form to construct an overall interval for a single parameter that reflects the possibility that just one of the models under consideration may in fact be correct. The approach is Bayesian and requires a specification of the prior probabilities that each model is correct. It also assumes (1) that the intervals on which we operate are effectively Bayesian and (2) that the posterior distributions they yield are normal with variances that may be assumed to be estimated without significant error. Under these assumptions it is straightforward to obtain an implicit equation for the upper and lower limits of an overall (unconditional) Bayesian interval and the equation is easily solved numerically.

The way the algorithm operates under changes in the parameters is explored and the method is illustrated by applications in the field of radiation epidemiology. The underlying assumptions are discussed and argued to be less restrictive than at first appears.

WWW Cache Statistical Modelling

Barbara Bogacka
Queen Mary, University of London, School of Mathematical Sciences
Mile End Road, London E1 4NS, UK
B.Bogacka@qmw.ac.uk

Marcus Keogh-Brown
Queen Mary, University of London, School of Mathematical Sciences
Mile End Road, London E1 4NS, UK
Marcus@flexnet.co.uk

World Wide Web (WWW) traffic is a significant part of network communications. A net of servers (caches) provides delivery service of requested files. However, the clients demand high quality of service, which grows together with the business of the network. Thus accurate statistical models of cache behaviour might help in dealing with these contradictory demands.

In the vast internet traffic literature there are several measures proposed to characterise cache behaviour. The most common are hit rate and file popularity. These are well examined and discussed, yet there are no satisfactory conclusions about the traffic behaviour or about prediction and modelling of cache characteristics optimising the Quality of Service.

The servers are set in a hierarchy reflecting the size and so the number of stored and transferred files. The available data for analysis are cache log files giving values of several variables, such as Time-stamp, Elapsed Time, Size of file, and other (see *ftp://ftp.ircache.net/Traces/readme*). Analysis of these data for various caches shows that they are all very bursty but they have very different correlation structures for different caches.

Hence we have calculated a new variable, based on the Time-stamp and the Elapsed Time, which is the number of requests being processed at a given time point, for example at every second. We call it *Queue Length*, see Figure 1.

- The Queue Length time series shows typical characteristics of a long memory, self-similar process (cf. Beran, 1994):
- There are periods where the observations stay at high level and periods where they stay at low level.
- At short time periods, there seem to be cycles and local trends.
- At long time periods, there is no apparent trend or cycle; the series looks stationary.

These properties could only be noticed looking at a minimum of one week's data since there are daily differences. Hence, for the large caches, self-similarity is a very useful characteristic as it allows aggregating of the series without changing its properties, and so decreasing its size to a manageable one. Furthermore, one of the mathematical consequences of these properties is that the autocorrelation function $\rho(\tau)$ tends, with lag τ tending to infinity, to the following simple form

(1) $$\rho(\tau) \xrightarrow[\tau \to \infty]{} c\tau^{-\alpha},$$

where c is a constant, $\alpha = 2 - 2H$, $H \in (0.5,\ 1)$ being the so-called Hurst parameter, widely used as a measure of self-similarity; the closer H is to one, the stronger the self-similarity of the series.

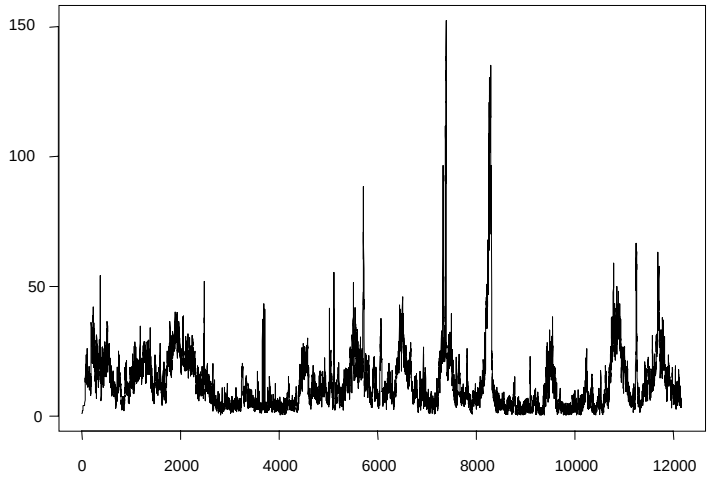


Figure 1. NLANR BO1 cache's Queue Length time series for a two-week period (aggregated at 100 seconds non-overlapping periods of time).

We have examined data from a wide spectrum of caches, both of different size and different hierarchy level. The non-linear least square method gives a good approximate for the parameter H, which seems to take values between 0.75 and 0.85 for all considered servers. Another important property of the Queue Length is its distribution. It is a very skewed, long tail kind of distribution and finding the one fitting best to the data is not an easy task. Also the question arises whether there is a connection between high values of H and the long tail distribution.

The purpose of our work is to examine further the Queue Length properties and to find out which cache parameters or variables influence the Queue Length performance. Currently, we are looking at the so-called *flight delay*, which, briefly, is the total time spent on communication between a client and a cache minus the time of the file transfer. We expect that this variable will prove significant to the Quality of Service, measured by the Queue Length properties. We will present the results of our investigation and discuss possible further work needed to optimise the Quality of Service.

References

Beran, J. (1994). Statistics for Long-Memory Processes. Chapman and Hall. New York.

Potential Function Approach to Time Series Modelling

Svetlana Borovkova
Delft University of Technology
Mekelweg 4, 2628 CD, Delft, The Netherlands
S.A.Borovkova@its.tudelft.nl

Herold Dehling
Ruhr-Universitat Bochum
Universitatsstrasse 150, 44780, Bochum, Germany
herold.dehling@ruhr-uni-bochum.de

John Renkema
Shell International Trading and Shipping Co Ltd
Shell-Mex House, Strand, London, WC2R 0ZA, UK
John.J.Renkema@STASCO.com

Herbert Tulleken
Shell Research and Technology Centre Amsterdam
P.O.Box 38000 1030 BN, Amsterdam, The Netherlands
Herbert.J.A.Tulleken@opc.shell.com

There are examples of real-life time series, which exhibit behavior, whereby their values seem to concentrate in a number of "attraction regions", preferring some values to others. A striking example is the series of daily crude oil prices: the oil price has a number of "preferred" regions (approximately at 18, 14 and 23 dollars per barrel), and most trading occurs here. The price lies outside these regions only relatively briefly and is then rather unstable. We can relate this feature of the time series to a certain property the invariant distribution of the underlying stochastic process. Namely, it can be expressed by the multimodality of the density of the invariant distribution. This phenomenon is in general not possible to capture by using traditional linear techniques of time series analysis.

We present a new approach to the problem of time series modeling in which we capture the multimodal invariant distribution of time series data within the (nonlinear) model. We propose to apply what we call a potential function approach whereby we let the underlying process be governed by a potential field that has its local minima at the attracting values with the addition of some random fluctuations.

We consider a continuous time process $(y_t)_{t\in R}$ in R, evolving according to

(1) $$dy_t = -U'(y_t)dt + \sqrt{\beta}dw_t,$$

where $U : R \to R$ is a potential function, (w_t) is the standard Brownian motion and $\beta \in R$ is the factor that measures the magnitude of random fluctuations. Under suitable conditions on U (Stroock and Varadhan (1979)), the distribution of (y_t) approaches weakly an equilibrium, which is a Gibbs distribution with density given (up to a constant) by

(2) $$\pi_\beta(y) = e^{-2U(y)/\beta}.$$

The discretization of the equation (1) (the Euler scheme) gives

$$y_{t+\Delta t} = y_t - U'(y_t)\Delta t + \sqrt{\beta}\Delta w_t.$$

Motivated by this scheme, the potential function model for a discrete time series $(y_i)_{i\in N}$ is given by

(3) $$y_{i+1} = y_i - U'(y_i)h + \varepsilon_i,$$

where h is an (unknown) time step, which we assume to be small, and (ε_i) are i.i.d. random variables having normal distribution with mean 0 and variance $h\beta$.

Using the relation (2) and a suitable estimate for the invariant density computed from the data, we obtain the estimate for the (scaled) potential. The combined parameter $h\beta$ that measures the effect of random fluctuation together with time discretization is subsequently estimated by the method of least squares.

This approach naturally extends to modeling multivariate time series. In that case the potential function U in R is replaced by the potential field in R^k and the derivative – by the gradient, so that the model (3) becomes $y_{i+1} = y_i - \nabla U(y_i)h + \varepsilon_i$, where (ε_i) are i.i.d. multivariate normal random vectors with k uncorrelated components each having the variance $h\beta$. In fact, we can show that the assumption of the components being uncorrelated with the same variance can be relaxed.

In dimensions higher than one, the main challenge in fitting the model is the estimation of the potential field and its gradient. We show how to estimate the potential field in higher dimensions by a combination of Gaussian kernels whereby parameters are estimated by a variant of the maximum likelihood method.

Testing the resulting model against historical data of oil prices (univariate as well as multivariate) shows that the essential price behavior is captured remarkably well. The model can generate copies of time series with the same distributional properties as the observed series, which is useful for applications such as parametric bootstrap, Monte Carlo simulations, scenario testing and other applications that require a large number of independent copies of the original time series. Furthermore, using the presented model for forecasting reduces the uncertainty about future time series behavior and allows to make better predictions.

To conclude, we note that our work was inspired by an optimisation technique called simulated annealing, where the global minimum of a function U is found by the application of diffusion in the form of (1) (Cherny (1985), Kirkpatrick (1083), Geman and Hwang (1986)). Introducing diffusion into the evolution, allows it to escape from local minima. This idea is in turn related to the physical procedure called annealing where a physical substance is melted and then slowly cooled to find all low energy configurations. The factor β then gets a special meaning and corresponds to a "temperature" of the substance.

References

Cherny, V. (1985). Thermodynamic approach to the travelling salesman problem: an efficient simulation algorithm. *J. Opt. Theory Appl.*, **45**, pp. 41-51.

Geman, S., Hwang, C.R. (1986). Diffusions for global optimisation. *SIAM J. Cont. Opt.*, **24**, pp. 1031-1043.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P. (1983). Optimisation by simulated annealing, *Science*, **220**, pp. 671-680.

Stroock, D.W., Varadhan, S.R.S. (1979). Multidimensional diffusion processes. Springer-Verlag, New York.

Regression with Partially Informative Censoring

Roel Braekers, Noël Veraverbeke
Universitaire Campus, Limburgs Universitair Centrum
B-3590 Diepenbeek, Belgium
roel.braekers@luc.ac.be, noel.veraverbeke@luc.ac.be

1. Introduction

We consider a fixed design regression model in which the non-negative response Y_x at covariate value $x \in]0,1[$ is subject to random right censoring by two independent and non-negative censoring variables C_x and D_x. The censoring variable D_x has a general distribution function G_{2x}, while C_x has a distribution function G_{1x} that satisfies the following proportional hazards assumption

(1) $$1 - G_{1x}(t) = (1 - F_x(t))^{\beta_x}, \qquad t \geq 0$$

for some $\beta_x > 0$ and where $F_x(t) = P(Y_x \leq t)$. Since (1) expresses a type of informative censoring, we call the model partially informative. In the absence of covariates, it has been introduced in Gather and Pawlitschko (1998). The observed variables at design point x are (Z_x, δ_x), where $Z_x = \min(Y_x, C_x, D_x)$ and $\delta_x = 1, 0, -1$ according as $Z_x = Y_x$, C_x, D_x. We write $H_x(t) = P(Z_x \leq t)$ and $T_{H_x} = \inf\{t : H_x(t) = 1\}$.

2. Estimator for the Conditional Distribution Function

For our inference, we consider the observed data (Z_{x_i}, δ_{x_i}) (i=1,…,n) at n fixed design points $0 < x_1 \leq x_2 \leq \ldots \leq x_n < 1$. We assume that F_x and G_x are continuous and also that the probabilities $p_{xk} = P(\delta_x = k)$ (k = -1, 0, 1) are strictly between 0 and 1. It can be seen that $\beta_x = p_{x0}/p_{x1}$ and that $1 - F_x(t) = (1 - K_x(t))^{\gamma_x}$, where $\gamma_x = 1/(1+\beta_x) = p_{x1}/(p_{x0} + p_{x1})$ and $K_x(t) = P(\min(Y_x, C_x) \leq t)$. An estimator for $F_x(t)$ is given by $F_{xh}(t)$ where $1 - F_{xh}(t) = (1 - K_{xh}(t))^{\gamma_{xh}}$, where γ_{xh} and $K_{xh}(t)$ are estimators for γ_x and $K_x(t)$ (see Braekers and Veraverbeke (2001)). They involve smoothing weights $\{w_{ni}(x;h_n)\}$ of Gasser-Müller type, involving a probability density kernel K and a bandwidth sequence $0 < h_n \to 0$.

3. Uniform Consistency and Weak Convergence

Under regularity conditions on the design points and on the kernel K and under smoothness conditions on the underlying quantities (existence of second order partial derivatives), we have the following results.

Theorem 1 Assume $T < T_{H_x}$ and $1-H_x(t) > \delta > 0$.

If $h_n \to 0$, $nh_n^5/\log n = O(1)$, then as $n \to \infty$,

$$\sup_{0 \leq t \leq T} |F_{xh}(t) - F_x(t)| = O((nh_n)^{-1/2}(\log n)^{1/2}) \quad \text{a.s.}$$

Theorem 2 Assume $T < T_{H_x}$.

(a) If $nh_n^5 \to 0$ and $\log^3 n/(nh_n) \to 0$, then as $n \to \infty$,

$$(nh_n)^{1/2}(F_{xh}(\cdot) - F_x(\cdot)) \to W_x(\cdot) \quad \text{in} \quad D[0,T]$$

(b) If $h_n = Cn^{-1/5}$ for some $C > 0$, then as $n \to \infty$,

$$(nh_n)^{1/2}(F_{xh}(\cdot) - F_x(\cdot)) \to \tilde{W}_x(\cdot) \quad \text{in} \quad D[0,T]$$

where $W_x(\cdot)$ and $\tilde{W}_x(\cdot)$ are Gaussian processes.

4. Testing for the Model

A formal test for this model is based on the following characterization: (1) holds for some $\beta_x > 0$ if and only if $H_x^1(t) = \tilde{a}_x H_x^{0,1}(t)$ $(t \geq 0)$, where $H_x^1(t) = P(Z_x \leq t, \delta_x = 1)$ and $H_x^{0,1}(t) = P(Z_x \leq t, \delta_x \neq -1)$. Following an idea of Csörg (1998) we prove weak convergence of a normalized version of the empirical process $\{H_{xh}^1(t) - \tilde{a}_{xh} H_{xh}^{0,1}(t); t \geq 0\}$ where $H_{xh}^1(t)$ and $H_{xh}^{0,1}(t)$ are empirical distribution functions of the kernel type:

$$H_{xh}^1(t) = \sum_{i=1}^{n} w_{ni}(x;h_n) I(Z_{x_i} \leq t, \ddot{a}_{x_i} = 1) \text{ and } H_{xh}^{0,1}(t) = \sum_{i=1}^{n} w_{ni}(x;h_n)\, I(Z_{x_i} \leq t, \ddot{a}_{x_i} \neq -1).$$

The limiting process is (under certain conditions on the bandwidth) equal to $B(H_x^{0,1}(t)/(p_{x0} + p_{x1}))$ where $\{B(t); 0 \leq t \leq 1\}$ is a Brownian bridge. This result opens the way to obtain limit distributions for goodness-of-fit statistics which are functionals of this empirical process.

References

Braekers, R. and Veraverbeke, N. (2001). The partial Koziol-Green model with covariates, *Journal of Statistical Planning and Inference* **91**, 55-71.

Csörg , S. (1998). Testing for the partial proportional hazards model of random censorship. In *Proceedings of Prague Stochastics '98*, vol.**1**, (eds M. Hušková, P. Lachout, A. Višek), 87-92. Union of Czech Mathematicians and Physicists.

Gather, U. and Pawlitschko, J. (1998). Estimating the survival function under a generalized Koziol-Green model with partially informative censoring, *Metrika* **48**, 189-209.

Criteria for the Choice of Tuning Constants In Robust Regression

João Branco
Universidade Técnica de Lisboa, Instituto Superior Técnico, Dept. de Matemática
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
joao.branco@math.ist.utl.pt

Maria Manuela Souto de Miranda
Universidade de Aveiro, Dept. de Matemática e UI&D Matemática e Aplicações
Campus de Santiago, 3810-193 Aveiro, Portugal
manuela.souto@mat.ua.pt

Robust regression is often carried out using generalised M-estimators. If these estimators are based on Huber type functions then a tuning constant has to be chosen before the estimator is fully defined. The choice of the tuning constant is critical since it determines the degree of efficiency and of robustness of the estimator.

Choosing the constant is almost a subjective decision. Different criteria of choice and several specific values have been proposed in the literature. One common approach is to fix a priori the cutting value without taking into account the probability distribution of the population. This has the advantage of simplifying the process of estimation but it can introduce variability on the efficiency of the estimator, since efficiency depends on the true underlying distribution and on the form of the estimator, as pointed out by Kelly (1992). The need of objective criteria for the choice of the constant has motivated suggestions directed to specific models.

For the estimation of the parameters of a structural linear relation using instrumental variables, Branco and Souto de Miranda (2000) suggest a method based on the empirical influence function of the classical least squares estimator of the relation parameters. This method can also be applied to robust estimation in linear regression models, since the regression model can be seen as a particular case of a structural relation model.

Kelly (1992) considers the simple linear regression model and suggests cutting values that result from setting up a high asymptotic efficiency. Staude and Seather (1990) consider the linear regression with p regressors and take $c = k((p+1)/n)^{1/2}$, observing that $k = 1$ seems to be a better compromise between efficiency and robustness. Wilcox (1997) chooses $k = 2$ to define the tuning constant, but he also recommends another bounded influence estimator, presented in Coakley and Heltmansperger (1993), which is based on a Huber function with $c = 1.345$.

The cutting values suggested for these bounded influence estimators and the cutting values obtained using the influence function criterion are compared and the corresponding estimators are studied.

References

Branco, J.A. and Souto de Miranda, M.M. (2000). The tuning constant problem in regression type models. In *Multivariate Statistics* (Eds. T. Kollo, E.-M. Tiit and M. Srivastava), New Trends on Probability and Statistics, 5, pp 45 – 50. TEV, Vilnius.

Coakley, C.W. and Heltmansperger, T.P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association* **88**, 872-880.

Kelly, G. (1992). Robust regression estimators – the choice of tuning constants. *The Statistician*, **41**, 303-314.

Staudte, R.G. and Seather, S.J. (1990). Robust estimation and testing. Wiley, New York.

Wilcox, R.R. (1997). Introduction to robust estimation and hypothesis testing. Academic Press, San Diego.

Inference on the Location Parameter of Exponential Populations — Externally Studentized Statistics

Maria de Fátima Brilhante*
Universidade dos Açores e
Centro de Estatística e Aplicações da Universidade de Lisboa
fbrilhante@notes.uac.pt

Let $X_1,\ldots,X_n$ be a random sample from a two parameter exponential population. From the indpendence between $X_{1:n}$ and $X_{n:n}$- $X_{1:n}$, and from

$$f_{X_{k:n}-X_{k-1:n}}(x)=\frac{n+1-k}{\delta}\ \exp\left(-\frac{n+1-k}{\delta}\ x\right)\ I_{(0,\infty)}$$

it is easily derived that the probability density function of the externally studentized expression

$$T_{n-1}=\frac{X_{1:n}-\lambda}{X_{n:n}-X_{1:n}}$$

is

$$f_{T_{n-1}}(t)=\int_0^\infty n\exp(-ns)\left\{(n-1)\exp\left(-\frac{s}{t}\right)\left[1-\exp\left(-\frac{s}{t}\right)\right]^{n-2}\frac{1}{t}\right\}ds.$$

After some elementary algebra,

$$f_{T_{n-1}}(t)=-n\left[\ B(n,nt)+nt\frac{\partial B(n,nt)}{\partial(nt)}\ \right]$$

and using the fact that

$$\frac{\partial B(n,nt)}{\partial(nt)}=B(n,nt)\left[\psi(nt)-\psi(n+nt)\right]$$

and the recurrence expression for the digamma function, we finally get

$$f_{T_{n-1}}(t)=nB(n,nt)\sum_{k=1}^{n-1}\frac{nt}{(n-k)+nt}\ I_{(0,\infty)}.$$

With this expression it is easy to compute explicit expressions for $f_{T_{n-1}}$, and quantile tables for effective use.

* Research partially supported by FCT/POCTI/FEDER

References

Brilhante, M. F. (1996) Inferência sobre o parâmetro de localização de uma população exponencial . I. Studentização externa. *A Estatística a Decifrar o Mundo*, 47-55, Salamandra, Lisboa.

Brilhante, M. F., Pestana, D. D. and Rocha, J. (1996) Inferência sobre o parâmetro de localização de uma população exponencial . II. Studentização interna. *A Estatística a Decifrar o Mundo*, 57-63, Salamandra, Lisboa.

Erdélyi, A., Magnus, W., Oberhettinger, F. and Tricomi, F. G. (1953). Higher Transcendental Functions. McGraw-Hill, New York.

MCMC Estimation of Multilevel Models in the MLwiN Software Package

William Browne
Institute of Education, Multilevel Models Project
20 Bedford Way,
London, UK
w.browne@ioe.ac.uk

1. History

The Multilevel models project (*http://multilevel.ioe.ac.uk/)* at the Institute of Education, London has produced a series of PC DOS based software packages (ML2, ML3, MLn) for fitting multilevel models for over 15 years. The original packages were based on a maximum likelihood based estimation engine using the iterative generalised least squares (IGLS) algorithm (Goldstein 1986) for Gaussian outcomes and quasi-likelihood methods (MQL and PQL) for dichotomous outcomes.

In 1998 the software package MLwiN (Rasbash et al. 2000A) was first released and contained many advances over its forerunners. It contained a user-friendly Windows interface allowing users to set up their models using an equation-based interface and to use extensive graphical displays to view their data. It also offered the user the option of using Monte Carlo Markov chain (MCMC) estimation methods as an alternative to the maximum likelihood methods. Gibbs sampling methods were used for Gaussian outcomes and hybrid Metropolis-Hastings Gibbs sampling algorithms were used for dichotomous outcomes (see Browne 1998 for details).

In 2000 a second release of the package was released containing a large number of improvements particularly in the data manipulations features of the package. This release also allowed the user greater flexibility over their choice of MCMC estimation methods, the ability to fit Poisson response models using MCMC and enhanced documentation (Rasbash et al 2000B). MLwiN is now the leading multilevel modelling package in Europe and the project team currently run introductory workshops on the use of the package several times a year.

Later this year a new development version of MLwiN with far greater MCMC capabilities will be released to the user community and this presentation will detail some of the new functionality that will be included.

2. MCMC Development Version

Currently the MLwiN package has a user community of over 2,000 people worldwide who are generally academics from the social and medical sciences. Many users do not have a strong statistical background and tend to use the default maximu likelihood based methods as these are conceptually simpler to understand and converge to point estimates. We do also have a more advanced userbase who are interested in fitting more complex models that cannot be easily fit using the IGLS algorithm and learning about other estimation methods.

For this user group we will be releasing a development version of MLwiN that has far greater MCMC estimation functionality and will also contain additional documentation on Bayesian statistics and MCMC algorithms in general. This

development version will contain an estimation engine that can fit all of the models that can be fit in the current version of MLwiN along with many more complex models.

The advances here will include MCMC estimation of the following models:

- Multivariate response multilevel models with missing responses.
- Complex multilevel data structures including cross-classified and multiple membership models
- Mixed response Normal and Binomial multilevel models
- Multilevel factor analysis models
- Models with measurement error in the predictor variables.

The development version will also include the ability to convert any MLwiN model into WinBUGS (Spiegelhalter et al. 1998) code. This will allow the user to compare for some models the MH methods used in MLwiN with the adaptive rejection algorithms used in WinBUGS.

References

Browne, W.J. (1998). *Applying MCMC Methods to Multilevel Models*. PhD dissertation, Department of Mathematical Sciences, University of Bath, UK.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, **73**, 43-56.

Rasbash, J, Browne, W.J., Healy, M., Cameron, B. and Charlton, C. (2000A) *MLwiN version 1.1* London: Institute of Education, University of London.

Rasbash, J, Browne, W.J., Goldstein, H. Yang, M. et al. (2000B). *A User's guide to MLwiN (version 2.1)* London: Institute of Education, University of London.

Spiegelhalter, D., Thomas, A. and Best N. (1998) *WinBUGS: Bayesian Inference using Gibbs Sampling, Manual v1.2*. Cambridge: Medical Research Council Biostatistics Unit.

Confidence Intervals for Logistic Regression in Sparse Data

Shelley B. Bull
University of Toronto, Department of Public Health Sciences,
Samuel Lunenfeld Research Institute of Mount Sinai Hospital, 600 University Avenue,
Toronto, Canada M5G 1X5
bull@mshri.on.ca

Juan Pablo Lewinger
University of Toronto, Department of Statistics,
Samuel Lunenfeld Research Institute of Mount Sinai Hospital, 600 University Avenue,
Toronto, Canada M5G 1X5
pablo@utstat.utoronto.ca

1. Introduction

Logistic regression modelling of mixed binary and continuous covariates is common in practice, but conventional estimation methods may not be appropriate for small samples. It is well known that the usual maximum likelihood estimates (MLEs) of the log odds ratio parameters are biased in finite samples, and there is a non-zero probability that an MLE is infinite. In exponential family models with canonical parameterization, Firth (1993) showed that modifying the score function to remove first order bias is equivalent to penalizing the likelihood by the Jeffreys' prior and removes the order n^{-1} bias of the MLEs. Rubin and Schenker (1987) and others have noted the equivalence between a Bayesian estimator based on Jeffreys prior and the correction that adds ½ to each cell in a 2 by 2 table (Haldane 1956). In small sample studies of multinomial logistic regression with general covariate types, these penalized estimates (PMLEs) were found to have smaller bias and MSE than the MLEs (Bull et al. 2001), but little is known about corresponding interval estimates.

Asymptotically, the MLEs are normally distributed around the true parameter with variance given by the inverse of the Fisher information matrix, but in finite samples, the quadratic approximation to the log likelihood may not apply. Wald test statistics and confidence intervals (CIs) based on large sample standard errors can have poor properties when the parameter is far from zero (Hauck and Donner 1977, Agresti 1999); CIs based on the profile likelihood may be preferred in small samples (Alho 1992). The first order asymptotic covariance matrix of the PMLEs is the same as that of the MLEs (Firth 1993), but construction of symmetric Wald-type CIs based on the PMLEs may be ill-advised because the small samples in which PMLEs are most useful will also be those in which the log-likelihood is not quadratic. We evaluated alternative methods of CI construction in small sample simulations and in a sparse data application.

2. Methods

Maximization of the unconditional and penalized likelihoods to obtain the overall supremum and the profile likelihood was implemented with the Qnewton function in the GAUSS software (Aptech 1990), which uses the BFGS descent algorithm. Finiteness of the MLEs was determined with algorithms developed in previous studies (Bull et al 2001). Symmetric Wald-type CIs were constructed with standard errors from the Fisher information evaluated at the MLEs/ PMLEs. Profile CIs for each regression parameter

were calculated by inversion of the likelihood ratio test based on the respective likelihoods, using a simple secant method to obtain the interval endpoints.

Simulations of small datasets were conducted using GAUSS, including correlated binary and continuous covariates in regressions with binary outcomes. The performance of the CIs was compared with respect to coverage probability and mean interval length.

3. Application to a Disease Prevention Trial

One of the disease outcome of interest, occurrence of hepatitis C, was rare, producing empty cells in some subgroups (Blajchman et al, 1995). As a result, in the model with an interaction between Treatment and Time Period, the usual unconditional logistic regression MLEs are infinite for two of the parameters. The Wald CIs, which are undefined, were set to be the entire real line (Agresti 1999). The PMLEs, however, could be obtained for all parameters. The profile likelihood CIs for the infinite MLEs have one finite and one infinite limit, but in contrast, those for the PMLEs have two finite limits.

	Treatment	Time Period	Treatment by Time
MLE	– 4	–1.57	+ 4
Wald CI	(– 4, + 4)	(–2.81, –0.32)	(– 4, + 4)
Profile CI	(– 4, –0.78)	(–2.85, –0.28)	(–0.15, + 4)
PMLE	–2.43	–1.57	1.96
Wald CI	(–5.33, 0.47)	(–2.75, –0.38)	(–1.24, 5.16)
Profile CI	(–7.30, –0.24)	(–2.79, –0.34)	(–0.70, 6.95)

Table 1. Unconditional and penalized maximum likelihood estimates with 95% CIs.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada and the Network for Centres of Excellence in Mathematics (Canada).

References

Agresti, A. (1999). On logit confidence intervals for the odds ratio with small samples, *Biometrics* **55**, 597-602.

Alho, J.M (1992). On the computation of likelihood and score test based confidence intervals in generalized linear models. *Statistics in Medicine,* **11**, 923-990.

Aptech Systems Incorporated (1990). *The GAUSS System, Version 2.0*, Kent, Washington.

Blajchman, M.A., Bull, S.B. and Feinman, S.V. for the Canadian Post-Transfusion Hepatitis Prevention Study Group (1995). Post-transfusion hepatitis: Impact of the non-A non-B hepatitis surrogate tests. *The Lancet* **345**, 21-25.

Bull, S.B., Mak, C. and Greenwood, C.M.T. (2001). A modified score function estimator for multinomial logistic regression in small samples. *submitted for publication*.

Firth, D (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27-38.

Haldane, J.B.S (1956). The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics* **20**, 309-311.

Hauck, W.W. and Donner, A (1977). Wald's test as applied to hypothesis testing in logit analysis. *Journal of the American Statistical Association* **81**, 471-476.

Rubin, D.B. and Schenker, N (1987). Logit-based interval estimation for binomial data using Jeffreys prior. In *Sociological Methodology 1987* (ed. C.C. Clogg), 131-144. American Sociological Association, Washington, D.C.

Minimax Bounds for Supersmooth Deconvolution Density Estimation

Cristina Butucea
ModalX, Université Paris X
200, Av. de la République; 92 001 Nanterre Cedex, France
and
Laboratoire de Probabilités et Modèles Aléatoires,
Université Paris VI; 6, rue Clisson; 75 013 Paris, France
cbutucea@u-parisl0. fr, butucea@ccr. jussieu. fr

We estimate in the minimax framework the common probability density f of i.i.d. random variables $X_1,\ldots,X_n$ which are not directly observed. Instead, we have at our disposal random variables $Y_1,\ldots,Y_n$ from the convolution model

$$Y_i = X_i + e_i,\ i = 1,\ldots,n$$

where the noise variables e_i are i.i.d. and independent of the X_i.

The deconvolution model was thoroughly studied in the litterature, for densities f in various smoothness classes and noise densities either ordinary smooth or supersmooth. We consider here the case where both the unknown density f and the errors' density f_e are supersmooth. They are described by their Fourier transforms F and F_e respectively, as follows. We suppose that for some $0 < a \le 2$ and $r > 0$, f belongs to the analytic class of densities $A_{a,r}(L)$ such that

$$\int |F(u)|^2 \exp\left(2a|u|^r\right) du \le L.$$

This condition describes an ellipsoid in the class of infinitely differentiable functions. Direct estimation of such functions and very detailed description of these objects is given by Lepski and Levit 1998.

As for the noise density, its Fourier transform has an exponential decay:

$$b_{\min} \exp\left(-b|u|^s\right) \le |F_e(u)| \le b_{\max} \exp\left(-b|u|^s\right),$$

for some strictly positive real numbers $b, b_{\min}, b_{\max}$ and $0 < s \le 2$.

In minimax theory we introduce a maximal risk measuring the quality of approximation of an arbitrary estimator for the worst function to estimate in our class. For fixed regularity parameters a and r, the minimax rate of convergence is the convenient normalizing sequence for the following criteria. First, we search for an estimation procedure whose normalized risk stays finite asymptotically and second, the normalized risk must stay strictly positive asymptotically for any possible estimation method.

Thus, for fixed a, r, b, s we introduce the maximal risk

$$\sup_{f \in A_{a,r}(L)} R_f[f_n, f],$$

and call it pointwise if the risk is the mean squared error (MSE) at some fixed real x, respectively, L_2 risk associated to the mean integrated squared error (MISE). In this context, we describe kernel estimators f_n of f attaining minimax rates in both cases.

The attained rates depend on whether $r < s$, case in which the bias effect is stronger than the variance of the estimator, or $r > s$, where the variance is dominating. In the case $r = s$ we obtain almost polynomial convergence rates and exactly polynomial if $r = s = 1$. Our rates improve in this context the L_2 rates of wavelet estimators in Pensky and Vidakovic 2000.

Moreover, in the case of noises having stable density we obtain efficient bounds, which means evaluating the normalized maximal risks up to constants.

In a discrete version of the deconvolution model, Tsybakov 2000 obtained adaptive rates of the L_2 risk and analytic classes of parameters. These rates are surprisingly slower with a logarithmic factor than the minimax rates and this phenomenon occurred previously only for the pointwise risk. We expect in our model the same effect and adaptive, rates slower than the minimax by a logarithm. of *n* in the case where $r \geq s$.

In the other case, if $r < s$ the estimators can be free of the smoothness parameter of the unknown density f and thus they are adaptive to the order of smoothness for the same minimax rates of convergence.

References

Lepski, O. V. And Levit, B. Y.(1998). Adaptive mihimax estimation of infinitely differentiable functions, *Mathem. Meth. Statist.* **7**, 123-156.

Pensky, M. And Vidakovic, B. (2000). Adaptive wavelet estimator for nonparametric density deconvolution, *Ann. Statist.* **27**, 2033-2053.

Tsybakov, A. B. (2000). On the best rate of adaptive estimation in some inverse problems, *C. R. Acad. Sci. Paris Ser.l Math.* **330**, 835-840.

Adaptive Estimation in Systems with Uncertain Observations and Unknown False Alarm Probability*

Raquel Caballero-Águila
Universidad de Jaén, Dpto. de Estadística e I.O.
Paraje Las Lagunillas s/n. 23071 Jaén, Spain
raguila@ujaen.es

Aurora Hermoso-Carazo, Josefa Linares-Pérez
Universidad de Granada, Dpto. de Estadística e I.O.
Campus Fuentenueva s/n. 18071 Granada, Spain
ahermoso@ugr.es, jlinares@ugr.es

1. Introduction

In many practical situations, such as communication systems, there may be a nonzero probability (false alarm probability) that any observation consists of noise alone; this may be caused by an intermittent failure in the observation mechanisms. These situations can be described by linear systems whose observation equation includes not only an additive noise, but also a multiplicative noise component, modelled by a sequence of Bernoulli random variables; such systems are called *Systems with Uncertain Observations*.

The linear state estimation problem in systems with uncertain observations, under the hypotheses of mutual independence of the noises and the initial state and independence of the Bernoulli random variables, was treated by Nahi (1969). Later on, García-Ligero et al. (1997) and Caballero et al. (2000) obtained the quadratic and polynomial filters, respectively. In all these works it is assumed that, at any time, the false alarm probability or, equivalently, the probability that the signal exists in the observations, is known.

In this paper we consider a linear discrete-time system with uncertain observations, whose observation equation is given by

$$z_k = u_k C_k x_k + v_k, \quad k \geq 0$$

where x_k is the state vector at time k, C_k is the observation matrix and $\{v_k;\ k \geq 0\}$ is a gaussian white noise. The uncertainties are governed by the sequence $\{u_k;\ k \geq 0\}$ of independent Bernoulli random variables, and we assume that $P[u_k = 1] = p$ for all $k \geq 0$, being p an unknown parameter. Also, we assume that the initial state and the noises of the system are mutually independent.

Our aim is to obtain estimators for the probability p, based on the successive observations, $z_0, z_1, ..., z_k$, of the system, that can be obtained recursively.

The proposed estimators of the unknown parameter p can be used for adapting the linear, quadratic and polynomial filtering algorithms established in Nahi (1969), Garcia-Ligero et al. (1997) and Caballero et al. (2000), respectively.

This work has been partially supported by the "Ministerio de Ciencia y Tecnología" under contract BFM2000-0602.

2. False Alarm Probability Estimation

We set the problem of obtaining $p_k = E\{p / Z^k\}$, the Bayes estimator of the probability p given the observations $Z^k = \{z_0,...,z_k\}$, assuming a quadratic loss function. Denoting by $f(p / Z^{-1})$ the prior density for p, the posterior density, $f(p / Z^k)$, can be obtained from Bayes' theorem, and it becomes

$$f(p / Z^k) = \frac{f(z_k / p, Z^{k-1}) f(p / Z^{k-1})}{\int f(z_k / p, Z^{k-1}) f(p / Z^{k-1}) dp}, \quad k \geq 0$$

where $f(z_k / p, Z^{k-1}) = pf(z_k / u_k = 1, Z^{k-1}) + (1-p) f(z_k / u_k = 0, Z^{k-1})$.

Accordingly, the computation of the posterior densities requires obtaining the densities $f(z_k / u_k = i, Z^{k-1})$, $i = 0,1$. By the independence hypotheses of the system, the density $f(z_k / u_k = 0, Z^{k-1})$ agrees with that of the observation noise vector v_k. However, as a result of the uncertainty in the observations, the determination of $f(z_k / u_k = 1, Z^{k-1})$ is not simple, since its computation grows in complexity as k increases. To avoid this difficulty, it seems natural to consider approximations for these densities. So, approximations for the posterior densities and, consequently, for the Bayes estimators of the parameter p are obtained.

We propose to approach this problem by approximating mixtures of gaussian distributions by gaussian distributions with their corresponding parameters.

However, even though this procedure provides a recursive method for obtaining approximations of the required densities, $f(z_k / u_k = 1, Z^{k-1})$, the computation of the posterior densities may not be simple since, for each k, the densities $f(z_k / p, Z^{k-1})$ have a mixture form; obviously, the difficulty will depend on the prior distribution selected.

We have treated this problem by assuming that the prior density $f(p / Z^{-1})$ is a Beta density. The reason for that supposition is that p is the parameter of a Bernoulli distribution and the family of Beta distributions is a conjugate family for the sampling from a Bernoulli distribution. Then, it is easily seen that the posterior densities build up as weighted averages of Beta densities. To avoid the computational complexity that this fact involves, we propose to approximate, in each stage, the mixture of Beta densities by a suitable Beta density. The main advantage of this approximation is that the proposed estimators of the probability p can be easily obtained by a recursive relation.

References

Caballero, R., Hermoso, A. and Linares, J. (2000). Least Mean-Squared Error Polynomial Estimation in Systems with Uncertain Observations. *Proceedings of the 2000 IEEE International Symposium on Information Theory*, **110**. Sorrento, Italy.

García-Ligero, M. J., Hermoso, A. and Linares, J. (1997). Second Order Polynomial Filtering for Discrete Systems with Uncertain Observation. *Proceedings of the VIII International Symposium on Applied Stochastic Models and Data Analysis*, **157-162**. Anacapri, Italy.

Nahi, N. E. (1969). Optimal Recursive Estimation with Uncertain Observation, *IEEE Transactions on Information Theory*. **IT-15**, 457-462.

Steady-State Analysis of the Polynomial Filter in Stationary Systems with Uncertain Observations

Raquel Caballero-Águila
Universidad de Jaén, Dpto. de Estadística e I.O.
Paraje Las Lagunillas s/n. 23071 Jaén, Spain
raguila@ujaen.es

Aurora Hermoso-Carazo, Josefa Linares-Pérez
Universidad de Granada, Dpto. de Estadística e I.O.
Campus Fuentenueva s/n. 18071 Granada, Spain
ahermoso@ugr.es, jlinares@ugr.es

1. Introduction

The estimation problem in linear systems in which some observations may not contain the signal has been considered as an important research field, because of its applications in many practical situations, such as communication systems, in which there can exist an intermittent failure in the observation mechanisms. So, there may be a nonzero probability (false alarm probability) that any observation consists of noise alone.

In these systems, called *Systems with Uncertain Observations*, the observation equation includes not only an additive noise, but also a multiplicative noise component, modelled by a Bernoulli random variables sequence. Because of this multiplicative noise, even if the additive noises are gaussian, the conditional expectation of the state given the observations, which provides the least mean-squared error (LMSE) estimator, is not a linear function of the observations and its computation requires an exponentially growing memory. Consequently, for this class of systems, attention has been directed to suboptimal estimators which are easier to achieve.

Nahi (1969) treated the LMSE linear filtering problem in systems with uncertain observations, when the multiplicative process is an independent Bernoulli random variables sequence, under the hypothesis of mutual independence of the noises and the initial state. Later on, García-Ligero et al. (1997) considered the LMSE quadratic filter and Caballero et al. (2000a) generalised this study considering the polynomial filter of an arbitrary order ν $(\nu \geq 1)$.

In systems with uncertain observations in which the additive noises of the state and observation equations are correlated at consecutive instants of time, the LMSE linear filtering problem was treated by Hermoso and Linares (1994). Recently, on the assumption that the additive noises are correlated at the same instant of time, Caballero et al. (2000b) have studied the LMSE ν th-order polynomial filtering problem. Our aim in this paper is to examine the asymptotic behaviour of this ν th-order polynomial filter for the case of stationary systems, more exactly, we propose to determine if this polynomial filter admits a steady-state form.

2. Steady-State Behaviour of the Polynomial Filter

In order to study the ν th-order polynomial filtering problem in systems with uncertain observation and correlated disturbances, Caballero et al. (2000b) defined a new system (augmented system), whose state and observation vectors were obtained as the

This work has been partially supported by the "Ministerio de Ciencia y Tecnología" under contract BFM2000-0602.

aggregate of the original vectors and their Kronecker powers up to the ν th-order. In this way, the LMSE linear filter of the augmented state based on the augmented observations provides the LMSE ν th-order polynomial filter for the original state.

The application of this polynomial filter to stationary systems motivates the study of conditions under which it admits a steady-state form. This form would be very advantageous from a computational point of view, since some of the calculations, such as the computation of the gain matrices and the error covariance matrices, would not have to be performed at each iteration of the algorithm. Thus, for practical applications it is important to see if the linear filtering algorithm for the augmented system admits a steady-state form. This would allow us to calculate the steady-state linear filter for the augmented state which provides the steady-state polynomial filter for the state of the original system.

The existence of a steady-state linear filter for the augmented system is not an immediate issue because, even though the original system is stationary, the augmented system may not be. Hence, our aim in this paper is, on the one hand, to establish conditions which guarantee that the augmented system is asymptotically stationary and, on the other, to determine under these conditions the steady-state form of the linear filtering algorithm proposed by Caballero et al. (2000b).

With regard to the first aim, we show that the augmented system is asymptotically stationary provided that the original stationary system is asymptotically stable or, equivalently, its transition matrix is strictly stable.

In order to obtain the steady-state linear filter in the asymptotically stationary augmented system, we define a new system, whose state equation coincides with that of the augmented system, but its observation matrix is deterministic (there is no uncertainty in the observations). This system satisfies the conditions to apply the Kalman filter for systems with correlated noises; moreover, this filter coincides with the linear filter for the state of the augmented system, provided that the observations are identical. On the other hand, it is easy to prove that, if the transition matrix of the original system is strictly stable, this new system is also asymptotically stationary. Then, we can apply the steady-state Kalman filtering algorithm which provides the steady-state algorithm for the linear filtering problem of the augmented state.

Finally, as we have indicated above, the steady-state polynomial filter for the state of the original system is obtained from the steady-state linear filter for the augmented state.

References

Caballero, R., Hermoso, A. and Linares, J. (2000a). Optimal Polynomial Filtering in Systems with Uncertain Observations. *Proc. of the XIX IASTED Int. Conf. on Mod., Ident. Control*, **326-332**. Innsbruck, Austria.

Caballero, R., Hermoso, A. and Linares, J. (2000b). Least Mean-Squared Error Polynomial Estimation in Systems with Uncertain Observations. *Proc. of the 2000 IEEE Int. Symp. on Inf. Th.*, **110**. Sorrento, Italy.

García-Ligero, M. J., Hermoso, A. and Linares, J. (1997). *Second Order Polynomial Filtering for Discrete Systems with Uncertain Observation. Proc. of the VIII Int. Symp. on Appl. Stoch. Models and Data An.*, **157-162**. Anacapri, Italy.

Hermoso, A. and Linares, J. (1994). Linear Estimation for Discrete-Time Systems in the Presence of Time-Correlated Disturbances and Uncertain Observations. *IEEE Transactions on Automatic Control*. **AC-39**, 1636-1638.

Nahi, N. E. (1969). Optimal Recursive Estimation with Uncertain Observation, *IEEE Transactions on Information Theory*. **IT-15**, 457-462.

A Class of Asymptotically Unbiased Semi-Parametric Estimators of the Tail Index*

Frederico Caeiro
F.C.T. – Universidade Nova de Lisboa, Departamento de Matemática
Quinta da Torre 2825 – 114 Caparica, Portugal
fac@mail.fct.unl.pt

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

We shall consider here a class of consistent semi-parametric estimators of a positive tail index γ, parametrized in a *tuning* parameter α, which enables us to have access, for any available sample, to an *Asymptotically Unbiased* estimator of γ, with a reasonably flat *Mean Squared Error* pattern, as a function of k, the number of top order statistics considered, and a high efficiency relatively to the classical Hill estimator (Hill, 1975), provided we may have access to a larger number of top order statistics than the number needed for optimal estimation through the Hill estimator. Such a class is given by

$$\gamma_n^{(\alpha)}(k) := \frac{\Gamma(\alpha)}{M_n^{(\alpha-1)}(k)} \left(\frac{M_n^{(2\alpha)}(k)}{\Gamma(2\alpha+1)} \right)^{1/2} , \quad \alpha \geq 1 ,$$

where $M_n^{(\alpha)}(k) := \frac{1}{k} \sum_{i=1}^{k} \left(\ln X_{n-i+1:n} - \ln X_{n-k:n} \right)^{\alpha}$, $\alpha > 0$, are consistent estimators of $\Gamma(\alpha+1)\,\gamma^{\alpha}$, whenever k is an intermediate sequence, i.e., $k = k_n \to \infty$, and $k = o(n)$, as $n \to \infty$. $M_n^{(1)}(k)$ is the above mentioned Hill estimator for γ. Under these restrictions on k, the statistics $\gamma_n^{(\alpha)}(k)$ are consistent for γ, and under some extra mild conditions on the second order behaviour of F they are asymptotically normal, with an asymptotically bias eventually non-null, and given by $b_{\alpha} \lim_{n\to\infty} \sqrt{k} A(n/k)$, $\alpha > 1$, where $A(t)$ measures the rate of convergence of $\ln U(tx) - \ln U(t)$ towards $\gamma \ln x$, $U(t) = F^{\leftarrow}(1-1/t)$, $F^{\leftarrow}$ denoting the generalized inverse function of F. The bias term b_{α} is given by:

$$b_{\alpha} = \frac{1}{2\rho} \left\{ (1-\rho)^{-2\alpha} - 2(1-\rho)^{1-\alpha} + 1 \right\} ,$$

and consequently, for every $\rho < 0$ there is always a value α_0 such that $b_{\alpha_0} = 0$. In table 1 we present the values $\alpha_0(\rho)$ for a few values of the second order parameter ρ. For the value $\rho = -1$ and for a sample of size n = 5000 from a Burr model, $F(x) = 1 - (1 + x^{-\rho/\gamma})^{1/\rho}$, $x \geq 0, \gamma > 0, \rho < 0$, with $\gamma = -\rho = 1$, we illustrate, in figure 1,

* Research partially supported by FCT / POCTI / FEDER.

the smoothness of the sample path of $\gamma_n^{(\alpha_0)}(k)$, $\alpha_0 = 1.9$, relatively to the sample path of $\gamma_n^{(1)}(k)$ and of $\gamma_n^{(5)}(k)$.

ρ	-0.10	-0.25	-0.50	-0.75	-1.00	-1.25	-1.50	-2.00	-2.50	-3.00	-
$\alpha_0(\rho)$	4.654	3.106	2.374	2.071	1.900	1.789	1.710	1.605	1.536	1.488	1

Table 1: $\alpha_0(\rho)$ as a function of the second order parameter ρ.

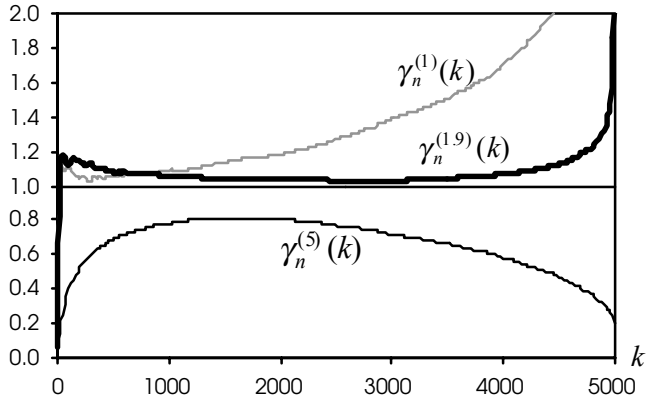


Figure 1. Sample paths of $\gamma_n^{(1)}(k)$, $\gamma_n^{(1,9)}(k)$ and of $\gamma_n^{(5)}(k)$.

The simulation results obtained on the basis of a multi-sample procedure of size 1000×10 enable us to make the following comments:

1. As expected, there is a great reduction in the *MSE* of the estimator, $\gamma_{n0}^{(\alpha_0)}(k)$ and consequently, provided we get to know ρ, we may easily work with an estimator highly efficient relatively to the Hill estimator.
2. Also the stability of the sample path of $\gamma_n^{(\alpha)}(k)$ for value α_0 may provide a selection of the optimal value (defined in an adequate way, like has been done, for instance, by Gomes and Martins (2001)), which will on its turn provide an estimator of ρ.
3. Even in the region $\rho < -1$, where it is difficult to find competitors to the Hill estimator, we are able to obtain, with $\gamma_n^{(\alpha_0)}(k)$, a better performance of this estimator relatively to the Hill estimator, provided we go deeper in the sample.

References

Gomes, M. I., Martins, M. J. and M. Neves (1998). Alternatives to a semi-parametric estimator of parameters of rare events – the Jackknife methodology. *Notas e Comunicações CEAUL* **18/98**. Accepted at Extremes.

Gomes, M. I. and M. João Martins (2001). Alternatives to Hill's estimator – asymtotic versus finite sample behaviour. *J. Statist. Planning and Inference* **93**, 161-180.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Ann. Statist.* **3**, 1163-1174.

Spatial Regression Models: Linking Community Air Pollution and Health

Sabit Cakmak
Health Canada, 203 Environmental Health Center
Tunney's Pasture, Ottawa, Canada K1A 0L2, PA: 0800B1
sabit_cakmak@hc-sc.gc.ca

Richard Burnett
Health Canada, 200 Environmental Health Center
Tunney's Pasture, Ottawa, Canada K1A 0L2, PA: 0800B1
rick_burnett@hc-sc.gc.ca

Daniel Krewski
University of Ottawa, Dep. of Epid. and Community Medicine, Faculty of Medicine
Ottawa, Canada
dkrewski@uottawa.ca

Cohort study designs are often used to assess the association between community based ambient air pollution concentrations and health outcomes, such as mortality, development and prevalence of disease, and pulmonary function. Typically, a large number of subjects are enrolled in the study in each of a small number of communities. Fixed site monitors are used to determine long-term exposure to ambient pollution. The association between community average pollution levels and health is determined after controlling for risk factors of the health outcome measured at the individual level (i.e., smoking). Health responses, however, often cluster by community, indicating that responses of subjects within the same community are more similar than responses of subjects in different communities. This implies that community itself poses some risk to health. Community-level variables, such as measures of socioeconomic status of the community, can be used to model this unexplained risk in addition to individual-level risk factors. Failure to account for all the variation between community health outcomes even after controlling for individual and community level risk factors can lead to downward biased estimates of the uncertainty in the community-level risk factors, including air pollution (Ware and Stram, 1988). Additional bias in the uncertainty of the risk estimates can occur if the community average health outcomes display spatial auto-correlation. That is, health responses for communities close together are more similar than responses for communities farther apart. Auto-correlation in the residuals of these models could be due to missing or systematically mis-measured risk factors that are spatially auto-correlated. Failure to account for spatial auto-correlation can yield downward biased estimates of uncertainty in the community-level risk factors and may suggest un-complete control for potentially confounding community-level factors with the variables of primary interest, such as air pollution (Miron, 1984). We present a new spatial regression model linking spatial variation in ambient air pollution to health. Health outcomes can be measured as continuous variables (pulmonary function), binary (prevalence of disease), or time to event data (survival or development of disease). The model incorporates risk factors measured at the individual level, such as smoking, and at the community level, such as air pollution. We demonstrate that the spatial auto-correlation in community health outcomes, an indication of not fully

characterizing potentially confounding risk factors to the air pollution-health association, can be accounted for through the inclusion of location in the deterministic component of the model assessing the effects of air pollution on health. We present a statistical approach that can be implemented for very large cohort studies. Our methods are illustrated with an analysis of the American Cancer Society cohort to determine whether the prevalence of heart disease is associated with concentrations of sulfate particles.

References

Miron, J. Spatial autocorrelation in regression analysis: a beginner's guide. In: *Spatial Statistics and Models*. Gaile GL, Willmott CJ eds. D. Reidel Publishing Company, Boston. 1984.

Ware, J.H., and Stram, D,O. 1988. Statistical issues in epidemiologic studies of the health effects of ambient air pollution. *Can. J. Statist.* **16**:5-13.

Homogeneity versus Inhomogeneity in Spatial Point Processes: Misfitting Issues

Mª Angeles Calduch, Jorge Mateu
Universitat Jaume I, Department of Mathematics
Campus Riu Sec, 8029-AP, Catellon, Spain
mcalduch@mat.uji.es, mateu@mat.uji.es

1. Introduction

Spatial point processes can be considered random geometric structures where the typical measure is based on the spatial locations of the individuals considered. Different models for point structures have been developed, starting from the Poisson process and ending up with Gibbs point processes. The latter class has been mostly used to model interaction between locations with several degrees of spatial structure. They are built having as reference the Poisson point process and usually the homogeneity assumption has been considered. However, in recent years, models for inhomogeneous point processes with interaction have been suggested by several authors (Stoyan and Stoyan, 1998; Baddeley et al., 2000; Jensen and Nielsen, 2000a,b). This appears to be a very natural step towards a more realistic modelling, where both first and second order properties of the point pattern are taken into account. Inhomogeneous point patterns may arise in many applied experimental sciences such as forestry, where both the spatial location and possible interactions between the trees might be subject to soil fertility, defined through a spatial random field.

The aim of this paper is to show fruitful comparisons between both point process structures with an emphasis on the misfitting issue. Questions like what is the error shown when an original inhomogeneous point process is being fitted by an homogeneous structure will be discussed. The results will be used to analyze real data sets coming from several applied fields.

2. Theoretical Set-Up

A point process is said to be a *Poisson point process* with intensity function λ if a) The number of points in any region B follows a Poisson distribution with mean $\int_B \lambda(x)dx$; and b) Given n points, their positions can be considered as an independent sample from the distribution with density $\lambda(x)/\int_B \lambda(x)dx$. If λ is constant the Poisson point process is said to be *homogeneous*, otherwise the process is *inhomogeneous*. A homogeneous Poisson process is often used as a reference model.

Markov or *Gibbs point processes* are useful to describe inhibition (van Lieshout, 2000 and references therein). These processes are defined upon a reflexive and symmetric relationship which in turn defines the neighbourhood condition between points in the pattern. Let X be a point process with density f with respect to the homogeneous Poisson point process with intensity 1. If X is Markov, its conditional intensity $f(x \cup u)/f(x)$ depends only on those points in x which are neighbours of u. These processes are also characterized by the Hammersley-Clifford theorem and the density is of the form

(1) $$f(x) = C \exp\left(-\sum\sum_{i<j} \phi\left(\|x_i - x_j\| \right) \right),$$

where C is a normalising constant and ϕ is a pair potential function modelling the interaction between points.

Following Jensen and Nielsen (2000) there are three basic ways of introducing inhomogeneity into a Markov model: a) Inhomogeneity induced by a non-constant first-order interaction (Stoyan and Stoyan, 1998); b) By thinning of a homogeneous Markov point process (Baddeley et al., 2000); c) By transformation of a homogeneous Markov point process (Jensen and Nielsen, 2000). For any of the three ways, the inhomogeneity may be described by a function λ defined on the same set as the points. In addition to the point pattern, explanatory variables may be observed at each point, for the purpose of explaining the inhomogeneity. The interaction specified in the models may or may not be location dependent.

For example, inhomogeneous Gibbs processes where the point density follows a deterministic trend are very useful (Baddeley et al., 2000; Jensen and Nielsen, 2000; Stoyan and Stoyan, 1998). In this case, equation (1) becomes

(2) $$f(x) = K \exp\left(-\sum\sum_{i<j} \phi\left(\|x_i - x_j\| \right) \right) \prod_i p(x_i) .$$

Now p(.) is a non-negative function modelling the trend in the density of the points. While the potential function models the short-range interaction, p determines the long-range variability.

3. Misfitting Issues

The aim in this section is to present a simulation study to quantify misfitting situations. We will focus on the three different types of introducing inhomogeneity in point patterns and on several degrees of trend modelling, varying from smooth trends to stronger ones. We will also use several point pattern models, such as the Strauss process and other more general Gibbs processes.

Finally we will analyze real point patterns and show how the misfitting affects the concluding results.

References

Baddeley, A., Moller, J. and Waagepetersen, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns, *Statistica Neerlandica.* To appear.

Jensen, E.B.V. and Nielsen, L. S. (2000a). Inhomogeneous spatial point processes. *Laboratory for Computational Stochastics*. Aarhus, Denmark.

Jensen, E.B.V. and Nielsen, L. S. (2000b). Inhomogeneous Markov point processes by transformation, *Bernoulli* **6**, 761-782.

Stoyan, D. and Stoyan, ,H. (1998). Non-homogeneous Gibbs process models for forestry: A case study, *Biometrical Journal* **40**, 521-531.

Van Lieshout, M.N.M. (2000). Markov point processes and their applications. *World Scientific*. To appear.

Smoothing Methods in Catchment Modification Detection

Gorana Capkun
Swiss Federal Institute of Technology, Department of Mathematics
1015-Lausanne, Switzerland
gorana.capkun@epfl.ch

Rapidly increasing population pressure in many rural areas of developing countries has led to changes in land use owing to deforestation, reclamation of wetlands, etc., or due to other catchment modifications like floodgates and roads. Such changes are intended to increase agricultural production, the use of waterpower, improve the quality of life and so forth. However, land mismanagement may have inadvertent negative effects on a hydrological regime, such as increasing the occurrence of floods and decreasing dry season flows. Therefore there is a need for improved knowledge and quantitative analysis of the impact of changes in land use and management practice on land and water resources.

The hydrological literature contains three basic approaches dealing with impact of land use change on catchment runoff: the experimental catchment approach, the land use modelling approach, and studies involving the use of hydrological models (sometimes combined with basic statistical methods such as linear regression and simple parametric and nonparametric tests). Lumped catchment models must be used carefully as they may fail to predict the impact of land-use change on catchment runoff due to limitations in the model conceptualization of the hydrological processes involved (Kuczera *et al*.,1993). Therefore, rigorous model validation procedures are required before the model capabilities can be assessed (Ewen and Parkin, 1996).

Our proposition below can be viewed as the extension of the model estimate and dynamic response variable comparison methods (before and after changes) proposed in Jakeman *et al*. (1993). In previous work we have developed a simple Markov generalized linear model of the mean and variance structures of runoff at time t, given previous rainfall and runoff (Capkun *et al*., 2000). Its mean is taken to be a linear autoregressive combination of present and previous rainfall and previous runoff, while its variance also depends on rainfall history. Inference for its parameters may be performed using classical likelihood methods, and also using the more robust technique of quasilikelihood, presupposing no particular distribution for runoff. Robust "sandwich" confidence intervals for the model parameters are constructed using both likelihood and quasilikelihood approaches. The model generally seems to fit rather well.

In this talk I shall discuss a smoothing method for correlated data that allows us to evaluate the impact of floodgate construction on runoff. Our data example uses daily observed rainfall and runoff on the Viege catchment in Switzerland, 1922–1985. The Mattmark floodgate was constructed in the 1960's and our goal is to evaluate the effect of this on the model parameters. When fitted using a local polynomial model and quasilikelihood (Fan and Gijbels, 1996), we obtain a correlated series of estimates such as that shown in Figure 1. The vertical lines show the approximate beginning and end of the floodgate construction and dashed lines are "sandwich" based confidence intervals for two of the model parameters. Parameter beta_0 explains the direct impact of the rainfall on simultaneous runoff, while gamma_1 is the autoregressive parameter of lag one.

This approach seems not to have been applied *per se* in the statistical literature, which apart from applications of smoothing methods to financial times series has largely concentrated on the situation where the data are independent though not identically distributed. Not many smoothing methods have been used in hydrology, and if so, almost entirely in the density estimation context. However they seem potentially very useful.

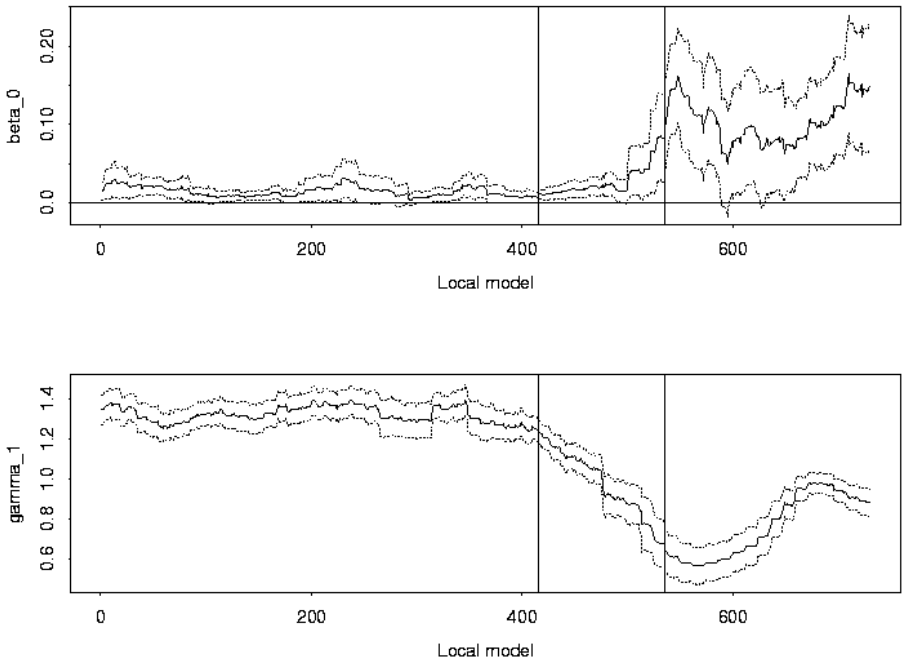


Figure 1. Local parameter estimates for fit of our autoregressive model to data from Viege. Vertical lines show the approximate beginning and end of the floodgate construction; dashed lines are "sandwich" confidence intervals.

References

Capkun, G., Davison, A. C. and Musy, A. (2001). A robust rainfall-runoff transfer model, *Submitted to Water Resources Research*.

Ewen, J. and Parkin, G. (1996).Validation of catchment models for predicting land-use and climate change impacts. 1. Method, *Journal of Hydrology*, **175**, 583-594.

Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and Its Applications. Chapman and Hall. London.

Jakeman, A. J., Littlewood, I. G. and Whitehead, P. G. (1993). An assessment of the dynamic response characteristics of streamflow in the Balquhidder catchments, *Journal of Hydrology*, **145**, 337-355.

Kuczera, G., Raper, G. P., Brah, N. S. and Jayasuriya, M. D. (1993). Modelling yield changes after strip thinning in a mountain ash catchment: an exercise in catchment model validation, *Journal of Hydrology*, **150**, 433-457.

Asymptotic Properties of a Simple TCP Model

Niclas Carlsson
Åbo Akademi University, Department of Mathematics
Fänriksgatan 3 B, FIN-20500 Åbo, Finland
nkarlsso@abo.fi

1. Introduction of the Model

Consider a data transmission channel using the transmission control protocol (TCP), in combination with a congestion control algorithm. Congestion control limits the traffic on a loaded channel by controlling the number of data packets sent at a time before waiting for an acknowledgement, that is, it controls the maximum size of bursts of packets. If one burst is sent without errors, the maximum size of the next one is increased by an additive constant. If, on the other hand, an error is detected, then the maximum size of the next burst is decreased by a constant multiplicative factor.

Our model mimics this behaviour under some simple assumptions. We assume that there is always data to be sent so that we always send the maximum amount of packets allowed. We also assume that the probability of error for each packet is independent of all other packets and equal to a constant p. The round trip time (RTT), the time taken for a signal to travel to the destination and back, is assumed to be constant and for simplicity we choose it to be equal to one. We also assume that we always can detect if an error has occurred before the next burst of packets is to be sent.

Consequently we study the family of Markov chains $\{X_n\}_{n=0}^{\infty}$ which have the transition probabilities:

(1)
$$\begin{cases} P(X_{n+1} = X_n + b) = (1-p)^{X_n} \\ P(X_{n+1} = aX_n) = 1-(1-p)^{X_n} \end{cases}$$

where $0 < a < 1$ and $b > 0$ are arbitrary but fixed. In real-world applications we usually have $a = ½$.

Similar models have been studied in the past, for instance in *Mathis et al.* and *Padhye et al.*, but our approach is different from these in that we are not making further mathematical assumptions, for instance, by assuming that the time between drops is independent of X. Also, we study not only the asymptotic mean of the process, but the stationary distribution and its convergence as well as the convergence of the process itself.

2. Summary of Results

A known feature of congestion control models is that the throughput is asymptotic to $c/\sqrt{p}$ for some c as $p \to 0$. We show that our process (1), when scaled by a factor $\sqrt{p}$ in both time and space, converges to an easily characterized process as $p \to 0$.

More exactly, we prove the following results: Consider the process (1). Let $Y_{p,n} = \sqrt{p}X_n(p)$. Then the following holds:

1 For any $p > 0$, the process $Y_{p,n}$ has a unique invariant measure μ_p.

2. Let $Y_{p,t}$ be the continuous time process received by replacing discrete time steps by exponential times with mean $1/\sqrt{p}$. Then the set of invariant measures for $Y_{p,t}$ and $Y_{p,n}$ coincide.
3. The process $Y_{p,t} \xrightarrow{w} Y_t$ in $D_{[0,\infty[}$ as $p \to 0$, where Y_t is a Poisson-intensity type jump process with linear drift.
4. The process Y_t has a unique invariant measure μ_0. This measure has a density, which can be represented as an alternating sum of the form $f(x) = \sum a_k \exp(-b_k x^2)$.
5. $\mu_p \xrightarrow{w} \mu_0$ as $p \to 0$, where μ_0 is the invariant measure of Y_t.

Thus, for *p* small, the unique stationary distribution of (1) can be approximated by a simple linear transformation of μ_0. The density of μ_0 is depicted in figure 1, for a particular choice of parameters.

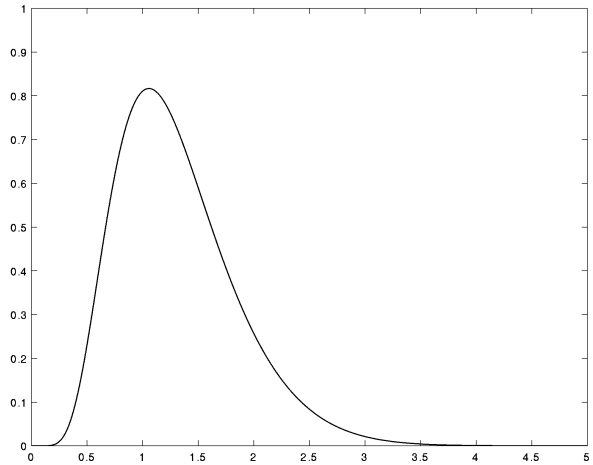


Figure 1. The invariant density *f(x)* for *a=1/2, b=1*.

References

Mathis, M., Semke, J. and Mahdavi, J. (1997). The Macroscopic Behaviour of the TCP Congestion Avoidance Algorithm, *Computer Communication Review* **27** (3)

Padhye, J., Firoiu, V., Towsley, D. and Kurose, J. (1998). Modeling TCP Throughput: A Simple Model and its Empirical Validation, *ACM SIGCOMM*

Fair Estimation: An Alternative to Maximum Likelihood in General Models

Eric Cator
Delft University of Technology
cator@edutiosa.twi.tudelft.nl

We consider the general set-up of a model of probability measures on a space X, parameterised by a space Θ. We want to estimate the parameter θ using a likelihood-method called fair estimation. This means that we consider an estimator T such that the distribution of T on Θ given the underlying parameter θ has a likelihood in θ that is independent of θ. In other words, each parameter has the same likelihood of being estimated correctly. An estimator with this property is called fair; we then take that fair estimator that maximizes the likelihood of estimating correctly. Of course, one has to make precise what is meant by the likelihood of T in θ, and this entails a choice that reflects how accurate each parameter has to be estimated with respect to the other parameters.

If we follow this procedure for classical multivariate models or normal linear regression, we find the classical (unbiased) estimators back, which are significantly better than the maximum likelihood estimator, especially when the number of parameters is close to the number of data points. Furthermore, for some relatively simple parametric models, we get more natural and more efficient estimators than the maximum likelihood method. For general parametric models we can get the same asymptotic behaviour as the (asymptotically efficient) maximum likelihood estimator.

In general it is hard to find the optimal fair estimator, but it does exists also in non-parametric models. We were able to find this estimator for interval censoring case 1, but we can still only calculate it for small sample sizes. However, especially for small sample sizes, fair estimation outperforms the maximum likelihood method in almost every sense. In order to calculate the fair estimator in this non-parametric model, we had to develop a completely new methodology with some very interesting theoretical results. Present work also entails trying to use this methodology to prove asymptotic efficiency (in a certain sense stronger than just the optimal rate) of the maximum likelihood estimator (and hence also of the right fair estimator) in this case.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Application of the Generalized Pareto Distribution to Flood Exceedances

V. Choulakian
Université de Moncton, Département de mathématiques et de statistique
Moncton, NB, Canada E1A 3E9
choulav@umoncton.ca

Michael A. Stephens
Simon Fraser University, Department of Statistics and Actuarial Science
Burnaby, BC, Canada V5A 1S6
stephens@stat.sfu.ca

The Generalized Pareto distribution (GPD) is a useful distribution for describing flood exceedances, that is, the distribution of heights of flood waters over a specific level called the threshold. In this paper we illustrate the above application. The paper is based on Choulakian and Stephens (2001), where much greater detail is given. The GPD has the following distribution function:

$$F(x) = 1-(1-kx/a)^{1/k}$$

where a is a positive scale parameter, and k is a shape parameter. The density function is

$$f(x) = (1/a)(1-kx/a)^{(1-k)/k};$$

the range of x is $0 \le x < \infty$ for $k \le 0$ and $0 \le x \le a/k$ for $k > 0$. For the special values $k = 0$ and 1, the GPD becomes the exponential and uniform distributions respectively. The distribution is sometimes called simply Pareto when $k \le 0$. Since the GPD has three parameters, it can be a versatile distribution for use with long-tailed data. This is demonstrated in Choulakian and Stephens (2001).

When the GPD is used in the modelling of extreme values in hydrology, as described above, the distribution is often called the "peaks over thresholds" (POT) model. Davison and Smith (1990) discuss this application in their section 9, using river flow exceedances for a particular river over a period of 35 years. Davison and Smith also suggest an interesting method of deciding the threshold: essentially the threshold is raised until the exceedances fit the GPD. This relies on an attractive feature of the GPD, similar to that for the exponential distribution: if a variable has a GPD distribution, its conditional distribution given that the value exceeds a value, say T, is also GPD. This method of choosing the threshold is investigated in some detail. For this purpose one needs tests of fit for the GPD, and these are given, based on the Cramér-von Mises and Anderson-Darling statistics. Examples are given to illustrate the estimation techniques and the goodness-of-fit procedures.

References

Choulakian, V. and Stephens, M.A. (2001). Goodness-of-Fit Tests for the Generalized Pareto Distribution. To appear, *Technometrics*, August, 2001.

Davison, A.C. and Smith, R.L. (1990). Models for exceedances over high thresholds (with comments). *Journal of the Royal Statistical Society*, series B, **52**, 393-442.

Nonparametric Goodness of Fit Tests: Data-Driven and Easy to Use, also in the Multidimensional Case

Gerda Claeskens
Texas A&M University, Department of Statistics
447 Blocker Building, College Station, TX 77843, USA
Gerda@stat.tamu.edu

Nils Lid Hjört
University of Oslo, Department of Statistics
P.O. Box 1053 Blindern, N-0316 Oslo, Norway
Nils@math.uio.no

1. Motivation

To investigate goodness of fit, we test whether a set of independent data has a certain density $f_0(.,\theta)$, where the vector θ may or may not be completely specified. Of course, there is a number of tests available for this situation. We discuss density-based omnibus goodness of fit tests based on estimated versions of likelihood ratio or score tests, incorporating nonparametric density estimation in a natural fashion. We focus on estimators constructed via log linear expansions, as they lead to a particular revealing structure regarding both construction of tests and limit distributions.

2. Data-Driven Test Statistics

Consider the following log linear expansion, where f_0 is the density under the null hypothesis,

$$f_S(x \mid a) = f_0(x) c_S(a)^{-1} \exp\{ \sum_{j \in S} a_j \psi_j(x) \}$$

for x in the interval of interest, where the ψ_j functions are orthogonal and normalized with respect to f_0, and also orthogonal to the function $\psi_0=1$. The set S is a subset of the natural integers, like $\{1, \dots, m\}$, and $c_S(a)$ is a normalizing constant. Employing this model, the likelihood ratio test statistic becomes

$$Z_n^* = Z_{n,S_n^*} = 2 \sum_{i=1}^{n} \log \frac{f_{S_n^*}(X_i \mid \hat{a})}{f_0(X_i)} = 2n\{ \sum_{j \in S_n^*} \hat{a}_j \bar{\psi}_j - \log c_{S_n^*}(\hat{a}) \},$$

where $\hat{a}$ is arrived at via maximum likelihood in the particular model indexed by the selected set S_n^*, and where $\bar{\psi}_j = n^{-1} \sum_{i=1}^{n} \psi_j(X_i)$. An alternative to the likelihood-ratio inspired test statistic is the score test, which here takes the particularly simple form $T_n^* = \sum_{j \in S_n^*} n \bar{\psi}_j^2$.

3. Behavior of Tests Using the AIC or BIC Regime

Denote S_n^* the maximizer w.r.t S, of the Akaike information criterion (AIC): $\mathrm{AIC}_{n,S} = Z_{nS} - C|S|$ where C is a constant bigger than 1.

Assume that at the outset all subsets S of $\{1, \dots, m_0\}$ are considered, where m_0 is fixed. The test proceeds by rejecting the null hypothesis when $Z_n^* > z_0$, with this positive constant appropriately adjusted. Another option is to reject the null hypothesis

if $Z_n^* > 0$, with the threshold parameter C adjusted to lead to a required significance level. The two types of test given here have certain parallels to ideas worked with earlier, but only in regression contexts, and with a nested sequence of models (Aerts, Claeskens and Hart, 1999, 2000), rather than as here where all models inside a certain range are allowed consideration.

If one wishes to allow subsets of $\{1, \ldots, m_0\}$ with a growing m_0, the traditional solution is to work with the sequence of nested subsets, say $\{1, \ldots, m\}$. We refer to Claeskens and Hjort (2001) for the asymptotic distribution of Z_n^* and T_n^* under a sequence of local alternatives.

The Bayesian information criterion in the present case takes the form $\mathrm{BIC}_{n,S} = Z_{nS} - (\log n)|S|$. We show that the BIC applied to nested models only, as is commonly done, has disadvantages and that, surprisingly, in the context of all subsets, there is a version of the BIC based tests, which is asymptotically equivalent with the different-looking corresponding AIC version.

Finite sample simulations, where we compare a number of different tests, show that none of the tests appears to be uniformly best.

4. Testing a Parametric Family

The null hypothesis to be tested is that the density belongs to a parametric family $f_0(.,\theta)$, where θ is p-dimensional, and traditional regularity conditions apply.

Let basis functions $\psi_j(x,\theta)$ be orthogonal with respect to $f_0(x,\theta)$. For a bounded set S, consider the extended parametric model

$$f_S(x,\theta\,|a) = f_0(x,\theta)c_S(a,\theta)^{-1}\exp\{\sum_{j\in S} a_j\psi_j(x,\theta)\},$$

where $c_S(a,\theta)$ is a normalizing constant. There are several ways to perform tests in parametric families. The apparatus developed is very general, and can be applied to test the adequacy of any parametric family, subject to the usual regularity conditions, also in higher dimensions. An important example is to test for multivariate normality, see Claeskens and Hjort (2001) for more details.

References

Aerts, M., Claeskens, G. and Hart, J.D. (1999). Testing the fit of a parametric function. *J. Amer. Statist. Assoc.,* **94**, 869-879.

Aerts, M., Claeskens, G. and Hart, J.D. (2000). Testing lack of fit in multiple regression. *Biometrika*, **87**, 405-424.

Claeskens, G. and Hjört, N.L. (2001). Goodness of fit via nonparametric likelihood ratios. Technical report.

Semiparametric Estimation in Single Index Poisson Regression: A Practical Approach

Daniela Climov, Léopold Simar
Institut de Statistique, Université Catholique de Louvain

Michel Delecroix
CREST-ENSAI, Rennes
delecroi@ensai.fr

We address the problem of estimating the direction parameter and the regression function in a Poisson single index model. The observed data $(X_i, Y_i) \in IR^K xIN$, for $i = 1,...,n$ are independent, the conditional distribution of Y_i given the vector of explicative variables X_i is Poisson, and we assume that we have a Single Index Model :

$$R(x) = E\left[Y_i \middle| X_i = x\right] = g_{\beta_0}(\beta_0 x),$$

where $\beta_0 x$ is the usual product of two vectors from IR^k, and g_β is defined by:

$$g_\beta(z) = E\left[Y_i \middle| \beta X_i = z\right].$$

The Single Index Models have been extensively used in the literature in actuarial sciences, in biometrics or in econometrics, but with a fixed link function in the framework of General Linear Models (GLM, see McCullagh and Nelder, 1989). Here we focus on the problem of estimating simultaneously the link and the parameters β in the case of a Poisson regression model. One of the most attractive approaches for estimating this kind of models is based on M-estimation methods: under the only above SIM condition a consistent estimator $\hat{\beta}_n$ of β_0 can be defined by maximizing with respect to β the empirical mean of some objective function Ψ :

$$\hat{\beta}_n = \arg\max_\beta \frac{1}{n}\sum_{i=1}^{n} \Psi\left(Y_i, \hat{g}_{\beta,h_n}(\beta X_i)\right),$$

where $\hat{g}_{\beta,h_n}$ is a nonparametric estimator of the function g_β ,defined below, and h_n is the serie of corresponding bandwidths, which tends to zero at some appropriate rate as $n \to \infty$. The Nadaraya-Watson leave-one-out estimator of $g_\beta(\beta X_i)$ will be used. It is defined as:

$$\hat{g}^{(-i)}_{\beta,h_n}(\beta X_i) = \frac{\sum_{j\neq i} Y_j K_{h_n}(\beta X_i - \beta X_j)}{\sum_{j\neq i} K_{h_n}(\beta X_i - \beta X_j)},$$

where $K_{h_n}(x) = h_n^{-1} K(x/h_n)$ and K is a fixed kernel function (typically a symmetric probability function).

On can find in Delecroix and Hristache (1999), asymptotic efficiency arguments justifying the maximum likelihood principle to choose here the function Ψ. Our estimator of β_0 is then the solution of:

$$\hat{\beta}_n = \arg\max_{\beta} \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i \log\left(\hat{g}^{(-i)}_{\beta,h_n}(\beta X_i)\right) - \hat{g}^{(-i)}_{\beta,h_n}(\beta X_i)\right\}.$$

Once β_0 has been consistently estimated, the regression function $R(x) = E(Y|X=x)$ can be estimated, in a second stage, from the nonparametric regression of Y_i on the estimated index $\hat{\beta}_n X_i$, using the Nadaraya-Watson estimator, which has the same form as in (1.5), except that here, the ith observation is included in the sum and that we use another serie of bandwidth $h_n^{'}$.To resolve the problem of choosing practically the two series of bandwidth h_n and $h_n^{'}$, Delecroix, Hristache, Patilea (1999), following Härdle, Hall and Ichimura (1993), suggest to define:

$$\left(\hat{\beta}_n, \hat{h}_n\right) = \arg\max_{h,\beta} \frac{1}{n}\sum_{i=1}^{n}\Psi\left[Y_i, \hat{g}^{(-i)}_{\beta,h}(\beta X_i)\right].$$

and

$$\hat{R}_n(x) = \hat{g}_{\hat{\beta}_n;\hat{h}_n}\left(\hat{\beta}_n x\right),$$

They prove the $\sqrt{n}$ asymptotic normality of $\hat{\beta}$ and $\hat{R}_n(x)$.

The aim of our study is twofold. First we investigate by Monte Carlo experiments the finite sample properties of that estimator. It shows that it is very difficult to use in practice, for reasonable sample sizes, the above limit laws. Then we suggest a bootstrap procedure to construct confidence intervals for $\hat{\beta}$, and show the practical efficiency of the method by simulation arguments. We conclude by the study of a real data set.

References

Delecroix, M. and M.Hristache (1999). M-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique. *Bull. Belg. Math. Soc*. **6**, 161-185.

Delecroix, M., Hristache, M. and V.Patilea (1999). Optimal smoothing in semiparametric index approximation of regression functions. *Cahiers du CREST* n°. **9952**, INSEE, Paris.

Härdle, W., Hall, P. and H. Ichimura (1993).Optimal smoothing in single-index models. *Ann. Statist*. **21**, 157-178.

McCullagh, P. and J.A. Nelder (1989). *Generalized Linear Models*. London: Chapman and Hall.

Combining Vasicek and Robust Estimators for Estimating Systematic Risk

G. S. Cloete
Market Risk, Standard Corporate and Merchant Bank
3 Simmonds Street, Johannesburg 2001, South Africa

P. J. de Jongh
Centre for Business Mathematics and Informatics
Potchefstroom University for CHE, Private Bag X6001, Potchefstroom 2531, South Africa
BWIPJD@puknet.puk.ac.za

T. de Wet
Department of Statistics and Actuarial Science, University of Stellenbosch
Stellenbosch 7600, South Africa
TDEWET@akad.sun.ac.za

The problem of estimating and forecasting systematic risk, or the so-called beta parameter in the market model, is well-known and has been studied by several authors (see e.g. Lam 1999, Lally 1998, Bowie and Bradfield 1998, Boabang 1996, Draper and Paudyal 1995, Murray 1995 and Bartholdy and Riding 1994). Some time ago, a paper by Fama and French (1992) sparked a debate about the relevance of beta. According to Fama and French (1992) "beta as the sole variable explaining returns on stocks is dead". Their findings have sparked renewed interest in the beta parameter and its applications in modern portfolio theory. Using a similar methodology, Davis (1994), and He and Ng (1994) came to a similar conclusion to that of Fama and French. However, Kothari et al.(1995) and Clare et al.(1998) used alternative methodologies and found that beta still has an important role to play. Earlier, this viewpoint was strongly supported by Black (1993) who claimed that Fama and French used "data mining" to reach their conclusions. In spite of these criticisms, many practitioners today still use estimates of beta in their decision-making processes and various services that provide beta estimates exist. Therefore research into finding more efficient estimators for beta remains relevant. The classical estimator for beta is the well-known ordinary least squares (OLS) estimator, but several authors have shown that this estimator suffers from several deficiencies, e.g. it has a mean reversion tendency, is inefficient when return distributions are non-normal, and has significant bias problems when shares are thinly traded. Several alternatives for OLS have been proposed in the literature. Amongst others, Vasicek (1973) and Blume (1973) proposed estimators to improve the mean reversion tendency of beta, Chan and Lakonishok (1992) proposed robust estimators to ensure more efficient estimation of beta, and Scholes and Williams (1977) proposed estimators to deal with the bias problem when shares are infrequently traded. A host of empirical studies have been carried out in order to evaluate the performance of the estimators under various conditions (see e.g. the recent studies by Draper and Paudyal 1995, Murray 1995, Boabang 1996, and Lally 1998). Of the above-mentioned estimators, the Vasicek-estimator and the robust estimators seem to perform well over a wide range of empirical studies.

In this paper we will base the so-called Vasicek-estimator (see Vasicek 1973) on the class of L-estimators and will evaluate its performance empirically, using data from the Johannesburg Stock Exchange (JSE). By doing this we combine the properties of the Vasicek-estimator with that of robust estimators and show that the new estimators perform much better than OLS and better or as good as some of the other popular estimators. In order to combine the Vasicek-estimator with the class of L-estimators, good scale estimators are needed for the various L-estimators considered. We define suitable scale estimators and show their consistency.

Incomplete Data: A Unifying Approah

Daniel Commenges
INSERM U330, Bordeaux, France
Daniel.Commenges@bordeaux.inserm.fr

The topic of incomplete data has focussed much attention in recent years. Two main directions can be distinguished: missing data and censored data. Although other types of incomplete data can be considered I will focus on the two concepts mentioned. Censored data have been known for a long time especially in survival data analysis. One of the first method dealing with such types of data is the Kaplan-Meier estmator (Kaplan and Meier, 1958). The work of Rubin (1976) has been a cornerstone in the classification and the study of missing data. This topic has gained a particular importance with the development of analyses for longitudinal studies. The two types of incomplete data seem to have something in common but the theories developed for them are completely different. One unifying attempt has been made by Heitjan (1994).

The aim of the present work is to unify the two concepts by a stochastic process approach. In a general framework we have a process of interest, say X, which is not completely observed. We define a continuous time response process R(t) which takes the value 1 if the process is oberved at time t, 0 otherwise. This framework covers for instance the case where X is a continuous process observed at discrete times and the case where X is a counting process observed in continous time during a certain period. In the former case there may be missing data (if R(t)=0 where it was planned that R(t)=1); in the latter case we may have the classical right censored data if the event was not observed during the observation period.

The general framework, a process X observed when a continuous response process is equal to one, certainly covers other interesting cases. The aim of the work is to give conditions under which the process leading to incomplete data is ignorable, thus generalizing results of Rubin (1976) and results known for censored data to this broader context. The challenge is here to write the likelihood of the observation of the processes. The correct definition of the likelihood which is necessary here in the context of stochastic processes relies on the Radon-Nikodyme derivative (Feigin, 1976).

The first step is to write the likelihood of the complete data (as if we observed the two processes X and R), then to take the expectation, conditional on what has been observed. Then conditions will be given for R(t) under which a simple likelihood, ignoring the process leading to incomplete data, can be obtained.

References

Feigin, Paul D. (1976) Maximum likelihood estimation for continuous-time stochastic processes. *AdvAppPr* **8** , 712- 736.

Heitjan, Daniel F. (1994) Ignorability in general incomplete-data models, *Biomtrka* **81**, 701- 708.

Kaplan, E. L. and Meier, Paul (1958), Nonparametric estimation from incomplete observations. *JASA* **53** , 457- 481.

Rubin, Donald B. (1976) Inference and missing data. *Biomtrika* **63** , 581- 590.

A Discrete Distribution Spanned by the Gaussian Hypergeometric Function with Complex Parameters

Antonio Conde Sánchez, María José Olmo Jiménez, José Rodríguez Avi, Antonio José Sáez Castillo
University of Jaén, Department of Statistics and Operations Research
Paraje Las Lagunillas s/n, D3, 23071 Jaén, Spain
aconde@ujaen.es, mjolmo@ujaen.es, jravi@ujaen.es, ajsaez@ujaen.es

1. Model Description

The aim of this work is to present a new family of Pearson's Discrete Distributions that may be obtained when the second polynomial coefficient in the difference equation does not have real solutions but complex. Therefore, we consider the following difference equation which the family of Pearson's Discrete Distributions verifies, that is

(1) $$G(r)f_{r+1} - L(r)f_r = 0 \quad r \in Z^+$$

where $L: Z^+ \to R$ and $G: Z^+ \to R - \{0\}$ are quadratic polynomials with real coefficients such as

(2) $$L(r) = (\alpha + r)(\bar{\alpha} + r)\lambda \qquad G(r) = (\gamma + r)(r+1)$$

with $\alpha = a + ib \in C, \bar{\alpha} = a - ib$ the conjugate of α and γ, λ real numbers.

The solution of (1) is given by

(3) $$f_r = f_0 \frac{(\alpha)_r (\bar{\alpha})_r \lambda^r}{(\gamma)_r r!} = f_0 \frac{\prod_{j=0}^{r-1}\left[(a+j)^2 + b^2\right] \lambda^r}{(\gamma)_r r!}$$

which is always real-defined and it coincides with the terms of the Gaussian Hypergeometric Function ${}_2F_1(\alpha, \bar{\alpha}; \gamma; \lambda)$ except a constant. So, imposing that f_r is a probability mass function, f_0 may be obtained applying the Gauss Summation Theorem for $\lambda = 1$.

The convergence of the function ${}_2F_1(\alpha, \bar{\alpha}; \gamma; 1)$ does not depend on the complex part of the roots of the polynomial L because the convergence condition is

(4) $$\Re(\gamma - a - bi - a + bi) = \gamma - 2a > 0$$

Moreover, the probability generating function, $g(t)$, is given by this expression

(5) $$g(t) = f_0 \sum_{r=0}^{\infty} \frac{(\alpha)_r (\bar{\alpha})_r t^r}{(\gamma)_r r!} = \frac{{}_2F_1(\alpha, \bar{\alpha}; \gamma; t)}{{}_2F_1(\alpha, \bar{\alpha}; \gamma; 1)}$$

so it is said that this discrete distribution is spanned by the Gaussian Hypergeometric Function with complex parameters.

2. Estimation

The following recurrence relation is verified by the moments about zero of the obtained distribution

(6) $$\omega_1 \mu'_{h+1} = \sum_{m=0}^{h} \binom{h}{m} \{\omega_2 \mu'_{m+1} + \omega_3 \mu'_m\}$$

where $\omega_1 = \gamma - 1, \omega_2 = 2a$ and $\omega_3 = a^2 + b^2$. As a consequence, the first three moments μ, μ'_2 and μ'_3 satisfy the equation system indicated below

(7) $$\begin{aligned} \mu(\omega_1 - \omega_2) - \omega_3 &= 0 \\ \mu'_2(\omega_1 - \omega_2 - 1) - \mu(\omega_2 - \omega_3) - \omega_3 &= 0 \\ \mu'_3(\omega_1 - \omega_2 - 2) - \mu'_2(2\omega_2 + \omega_3 - 1) - \mu(\omega_2 - 2\omega_3) - \omega_3 &= 0 \end{aligned}$$

from which we can solve the problem of estimation in two steps: first, solving the linear system with ω_1, ω_2 and ω_3 as unknown values and replacing μ'_i by the sample moments about zero m_i; secondly, obtaining the explicit values of the parameters by means of the relation with ω_i. But the drawback of this method is that the existence of the third moment about zero is not guaranteed and it is also very sensitive to changes in sample values. So, an alternative method is obtained from the quotient between consecutive probabilities

(8) $$\frac{f_r}{f_{r+1}} = \frac{G(r)}{L(r)} = \frac{(\gamma + r)(r+1)}{(a+r)^2 + b^2} \Rightarrow f_{r+1} = \frac{(a+r)^2 + b^2}{(\gamma + r)(r+1)} f_r, \; r = 0,1,\ldots$$

whose disadvantage is that the whole sample information is not been used. In consequence, we use mixed estimation methods that consider simultaneously the first relations between moments and frequencies until obtaining the necessary number of equations.

3. Applications

Finally, we conclude the description of this kind of distributions applying them to model and describe some discrete situations in fields as Sports and Economics, among others, and that may not be fitted adequately with the usual discrete distributions.

References

Bowman, K. O., Shenton, L. R., Kastenbaum, M. A. (1991). Discrete Pearson Distributions, Oak Ridge National Laboratory. Technicak Report TM-11899 Oak Ridge, Tennessee.

Dacey, M. F. (1972). A Family of Discrete Probability Distributions defined by the Generalized Hypergeometric Series. *Sankhya, serie* B, **34**, 243-250.

Gutiérrez-Jáimez, R. and Rodríguez Avi, J. (1997). Family of Pearson Discrete Distributions Generated by the Univariate Hypergeometric Function ${}_3F_2(\alpha_1, \alpha_2, \alpha_3; \gamma_1, \gamma_2; \lambda)$. *In Applied Stochastics Models and Data Analysis*, **13**, 115-125.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions*. Wiley, New York. Second edition.

Ord, J. K. (1972). *Family of frequency distributions*. Griffin, London.

Rodríguez-Avi, J., Conde-Sánchez, A. and Sáez-Castillo, A.J. (2001). A new class of Discrete Distributions with complex parameters. Submitted to *Statistical Papers.*

Rodríguez-Avi, J., Gutiérrez-Jáimez, R. and Conde-Sánchez, A. (2000). Study of a wide class of univariate discrete distributions generated by the hypergeometric function ${}_3F_2$. Submitted to *Theory of Probability and Its Applications.*

Summation of Hypergeometric Series of Matricial Argument and its Application in the Distribution of the Smallest Root of a Wishart

Antonio Conde Sánchez, María José Olmo Jiménez, José Rodríguez Avi, Antonio José Sáez Castillo
University of Jaén, Department of Statistics and Operations Research
Paraje Las Lagunillas s/n, D3, 23071 Jaén, Spain
aconde@ujaen.es, mjolmo@ujaen.es, jravi@ujaen.es, ajsaez@ujaen.es

1. Introduction

Hypergeometric series and functions of matricial argument, defined as sums of zonal polynomials, have been used in the expression of the distribution of quadratic forms in multivariate normal samples. Nevertheless, the great majority of these series of matricial argument do not have a explicit summation result, so many theoretic results that depends on them can not be applied to real data.

In this paper, we describe a new expression of every hypergeometric series expressed on terms of zonal polynomials, the latter calculated as a linear combination of symmetric monomials, which are a more usual and single base of symmetric and homogeneous polynomials. For this reason, it is necessary to calculate a great number of zonal polynomials of a high degree on terms of these symmetric monomials. This new expression permits the summation of hypergeometric series of matricial argument (that involves zonal polynomials) with a finite number of addends, and the approximate summation of series with a infinite number of addends, making possible the application of theoretic results that involves this type of series. In the paper we present, in fact, an application where is calculated the probability distribution of the smallest root of a Wishart .

2. Results

The family of hypergeometric functions of matricial argument is given by

(1) $$ {}_pF_q\left(a_1,...,a_p;b_1,...,b_q;Y_{m\times m}\right)=\sum_{k=0}^{\infty}\sum_{\kappa}\frac{(a_1)_\kappa...(a_p)_\kappa}{(b_1)_\kappa...(b_q)_\kappa}\frac{C_\kappa(Y)}{k!}, $$

where for each k it's necessary to sum over all partitions κ of k and $(a)_\kappa$ is the generalized hypergeometric coefficient.

In this series, the zonal polynomial associated with the partition κ, $C_\kappa(Y)$, is a symmetric and homogeneous polynomial of degree k in the eigenvalues, $y_1,...,y_m$, of the argument matrix and such that (a) the term of high weight is $y_1^{k_1}\cdot...y_m^{k_m}$, (b) it's an eigenfunction of the differential operator

(2) $$ \Delta_Y=\sum_{i=1}^{m}y_i^{\,2}\frac{\delta^2}{\delta y_i^{\,2}}+\sum_{i=1}^{m}\sum_{\substack{j=1\\ j\neq i}}^{m}\frac{y_i^{\,2}}{y_i-y_j}\frac{\delta}{\delta y_i} $$

and (c) the sum of all the zonal polynomials of degree k is $(trY)^k$.

Given that there is no a general explicit expression for zonal polynomials, it's necessary to evaluate them through another known base of the symmetric and homogeneous polynomials. In this sense, it's known that (James, 1968)

(3) $$C_\kappa(Y) = \sum_{\lambda \le \kappa} c_{\kappa,\lambda} M_\lambda(Y)$$

where $c_{\kappa,\lambda}$ are constants which are calculated iteratively through an algorithm (James, 1968) that has been programmed in MATLAB (Gutiérrez, Rodríguez and Sáez, 2000), and $M_\kappa(Y)$ is the symmetric monomial associated with κ. In this manner, it's possible to express any hypergeometric series that involves zonal polynomials as a series given by symmetric monomials. This methodology permits to apply, by example, the next theoretic result.

Theorem If l_m is the smallest root of S, where A=n·S is $W_m(n,\Sigma)$, and r=(n-m-1)/2 is a positive integer, then

(4) $$P_\Sigma[l_m > x] = etr\left(-\frac{1}{2} nx\Sigma^{-1}\right) \sum_{k=0}^{mr} \sum_{\kappa}^{*} \frac{C_\kappa\left(\frac{1}{2} nx\Sigma^{-1}\right)}{k!}$$

where Σ_κ^* denotes the sum over all the partitions $\kappa=(k_1,...,k_m)$ such that $k_1 \le r$.

As a real application of this theorem, we have considered samples of 10 data of the variables *Sepal length* for *Iris setosa, Iris versicolor* and *Iris virginica* in the well known example by Fisher (1936). If we denote these samples z_{10x3} , centred in mean, supposed $Z \rightarrow N_3(0,\Sigma)$, then $A=Z'\cdot Z$ is a $W_3(10,\Sigma)$ distribution, where Σ is approximated by its maximum likelihood estimation.

We have validated this result in a empirical sense in two ways: First, simulating sample data of *Z* and so, of *A*, checking that the empirical and the theoretical distribution functions are the same (Conde *et al*, 2001); secondly, we have generated a sample from the Fisher's original data, we have calculated the empirical distribution of this sample and we have tested the goodness of fit through a Kolmogorov-Smirnoff test.

References

Conde, A; Rodríguez, J; Sáez, A. (2001). Cálculo de polinomios zonales y algunas aplicaciones en Análisis Multivariante. Submmited to Qüestiio.

Fisher R.A. (1936). The use of Multiple Measurements in taxonomic problems, *Ann. Eugenics*, **7**, 179-188.

Gutiérrez R., Rodríguez J., and Sáez A. J. (2000). Approximation of hypergeometric functions of matricial argument through their development in series of zonal polynomials, *Electronic Transactions in Numerical Analysis* **11**, 121-130

James, A. T. (1968). Calculation of zonal polynomial coefficient by use of the Laplace-Beltrami operator, *Ann. Math. Statis.***39**, 1711-1718.

Muirhead, R. J. (1982). Aspects of multivariate statistical theory. Wiley. New York.

Principal Directions for the Normal Random Variable

Carles M. Cuadras, Daniel Cuadras
Universitat de Barcelona. Departament d'Estadística.
Diagonal 645, 08028 Barcelona, Spain.
carlesm@porthos.bio.ub.es

1. Introduction

Let X be a random variable with continuous cdf $F(x)$, probability density $f(x)$ and range $[a,b]$. Let us consider the symmetric kernel

$$K(s,t)=\min\{F(s),F(t)\}-F(s)F(t),$$

the normalized eigenvectors $\phi_n(x)$ of K with eigenvalues λ_n, i.e.,

$$K(s,t)=\sum_{n=1}^{\infty}\lambda_n\phi_n(s)\phi_n(t),$$

and the Bernoulli process $\boldsymbol{X}=\{X_t, t\in[a,b]\}$, where X_t is the indicator of $[X>t]$.

The process $\boldsymbol{X}$ describes the random variable X, namely, if a is finite, then

$$X=a+\int_a^b X_t dt\ .$$

Let us define

$$h_n(x)=\int_a^x \phi_n(t)dt.$$

Then $X_n=h_n(X),\ n\geq 1,$ are the principal components in the Karhumen-Loève expansion of $\boldsymbol{X}$. Thus (X_n) is a set of uncorrelated variables with $\text{var}(X_n)=\lambda_n$ such that $tr(K)=\sum_{n=1}^{\infty}\lambda_n$.

The r.v. X can be expanded as

$$X=x_0+\sum_{n=1}^{\infty}h_n(b)(X_n-h_n(x_0)) \text{ and } \quad |X-X'|=\sum_{n=1}^{\infty}(X_n-X_n)^2,$$

where X, X' are iid. These expansions were found for the uniform, exponential, logistic and Pareto distributions (Cuadras and Fortiana, 1995; Cuadras and Lahlou, 2000a,b).

2. Continuous Scaling Expansions

The functions h are solutions of the ordinary differential equation

$$\lambda\ h''(x)+(h(x)-\mu)f(x)=0,\ \ h(a)=h'(a)=0,$$

with $\mu=E(h(X)),\ \lambda=\text{var}(h(X))$.

Alternatively, these principal components can also be obtained by continuous scaling on the distance function $\delta(x,x')=\sqrt{|x-x'|}$. Let $G(x,x')$ be the centralized (inner product) function for $\delta(x,x')$, i.e.,

$$G(x,x')=-\frac{1}{2}[\delta(x,x')^2-E_X\delta(x,X)^2-E_{X'}\delta(x',X')^2+E_{XX'}\delta(X,X')^2],$$

where X, X' are iid. Let us consider the eigendecomposition

$$f(x)^{1/2} G(x,x') f(x')^{1/2} = \sum_{n=1}^{\infty} \lambda_n u_n(x) u_n(x'),$$

and define

$$c_n(x) = f(x)^{-1/2} \sqrt{\lambda_n}\ u_n(x).$$

Then

$$G(x,x') = \sum_{n=1}^{\infty} c_n(x) c_n(x'),$$

it can be proved that

$$E(c_n(X)) = 0,\ \ \mathrm{var}(c_n(X)) = \lambda_n,$$

$$\mathrm{cov}(c_m(X), c_n(X)) = 0,\ m \neq n,\ E_X(G(X,X)) = tr(K),$$

and

$$h_n(x) = c_n(x) - c_n(a).$$

Thus the functions h_n can be obtained from c_n.

3. Normal Distribution

Let $f(x), \Phi(x)$ be the density and cdf for the r.v. X with N(0,1) distribution, res-pectively. For this distribution, we could not find in closed form, the functions *h* solutions of the above ordinary differential equation.

The centralized (inner product) function is given by

$$G(x,x') = \min\{x,x'\} + f(x) + f(x') + x(\Phi(x) - 1) + x'(\Phi(x') - 1) - 1/\sqrt{\pi}.$$

The eigendecomposition of $G(x,x')$ is obtained by using numerical methods, allows us to find the functions *h* and therefore the principal components $h_n(X)$ of ***X*.**

These principal components can be used in distinguishing the normal from the logistic distribution, in goodness-of-fit tests, in studying the asymptotic distribution of Rao's quadratic entropy (Liu and Rao, 1995).

References

Cuadras, C.M. and Fortiana, J. (1995). A continuous metric scaling solution for a random variable. *J. of Multivariate Analysis,* **52**, 1-14.

Cuadras, C.M. and Lahlou, Y. (2000a). Some orthogonal expansions for the logistic distribution. *Comm. Stat.- Theory Methods,* **29**, 2643-2663.

Cuadras, C.M. and Lahlou, Y. (2000b). Principal components for the Pareto distribution. Distributions with Given Marginals and Statistical Modelling, Barcelona, report.

Liu, Z. and C. R. Rao (1995) Asymptotic distribution of statistics based on quadratic entropy and bootstrapping. *J. of Statistical Planning and Inference*, **43**, 1-18.

Unimodality of Copulas

Ioan Cuculescu
Universitatea Bucuresti, Facultatea de matematica
Str. Academiei 14, RO-70109 Bucuresti, România
icucul@pro.math.unibuc.ro

Radu Theodorescu
Université Laval, Département de mathématiques et de statistique
Sainte-Foy, Québec, Canada G1K 7P4
radutheo@mat.ulaval.ca

1. Introduction

It is Sklar (1959) who coined the term *copula* for a distribution whose margins are uniform on $I = [0,1]$. An important property of a distribution is unimodality. It is then natural to ask whether copulas are unimodal. Multivariate unimodality takes different forms so we choose here central convex, block, and star ones and examine copulas with respect to them.

We refer to Nelsen (1999) for copula concepts and to Dharmadhikari and Joag-dev (1988) and Bertin, Cuculescu, and Theodorescu (1997) for unimodality ones.

We examine the case of dimension $r = 2$ and comment about the case $r > 2$.

2. Unimodality

Let $W(u,v) = \max(u + v - 1,0)$ and $M(u,v) = \min(u,v)$ be the lower and the upper Fréchet-Hoeffding bounds; W and M are copulas. Further set $\Pi(u,v) = uv$ for the independence copula. Fréchet's family of copulas consists of all convex combinations of W, M, and Π.

We have the following result concerning central convex unimodality:

Proposition 2.1 A copula may be central convex unimodal only about (0.5,0.5). It is so if and only if it belongs to Fréchet's family.

The next result concerns block unimodality:

Proposition 2.2 A copula block unimodal about an interior point $(a,b) \in I^2$ has the probability density function

$$f = q1_{(0,a)\times(0,b)} + (1-aq)(1-a)^{-1}1_{(a,1)\times(0,b)} + (1-bq)(1-b)^{-1}1_{(0,a)\times(b,1)}$$

$$+ (1-a-b+abq)(1-b)^{-1}(1-a)^{-1}1_{(a,1)\times(b,1)},$$

where 1_A stands for the indicator function of A and $\max((1/a) + (1/b) - (1/ab),0) \leq q \leq \min(1/a,1/b)$. If (a,b) is not an interior point then the only block unimodal copula is Π.

Next we examine copulas in the class of star unimodal distributions, broader than that of block unimodal distributions. We also indicate examples of star unimodal copulas, absolutely continuous, with a nonull singular part, and even singular.

3. Diagonals

We now characterize diagonals of copulas star unimodal about (0,0):

Proposition 3.1 Let δ be a diagonal and c $\in$ [0,0.5]. There exists a copula C star unimodal about (0,0) such that $\delta = \delta_C$ and

(2) $$C_u'(1,v) = (1-c)v, \qquad C_v'(u,1) = (1-c)u, \qquad u, v < 1,$$

if and only if δ'(u)/u is absolutely continuous nonincreasing and

$$\delta'(1) = 2(1-c), \quad (\delta'(u)/u)' \geq -4c/u^2, \ \delta(u) - u\delta'(u)/2 \geq cu.$$

If c = 0 then C = Π and δ (u) = δ_C(u) = u^2.

We construct copulas *C* star unimodal about (0,0) satisfying (2) and indicate their diagonal sections δ_C.

4. Unimodality of Archimedean Copulas

Let us now examine Archimedean copulas.

Proposition 4.1 An Archimedean absolutely continuous star unimodal copula C (particularly block unimodal) coincides with Π.

Proposition 4.2 An Archimedean star unimodal copula C having a nonnull singular part coincides with W.

As a by-product we obtain

Corollary 4.3 With the exception of Π and W, Fréchet's copulas are not Archimedean.

5. About the Case of Higher Dimension

For higher dimension $r > 2$ Proposition 2.2 is valid with selfexplanatory modifications: I^r splits generally into 2^r parallelipipeds, the probability density function is constant on each of them, the constants depending on a parameter analogous to q. A similar remark holds for star. The extension to higher dimension of Proposition 3.1 has to start with a study of the corresponding diagonal sections. As far as Section 4 is concerned, it appears that the methods used may also work for $r > 2$.

References

Bertin, E., Cuculescu, I., and Theodorescu, R. (1997). *Unimodality of probability Measures*. Kluwer. Dordrecht.

Dharmadhikari, S. W. and Joag-dev, K. (1988*). Unimodality, convexity, and Applications*. Academic Press. New York.

Nelsen, R. B. (1999). *An introduction to copulas*. Springer. New York.

Sklar, A. (1959). Fonctions de répartition à *n* dimensions et leur marges, *Publ. Inst. Statist. Univ*. Paris **8**, 229-231.

Testing the Proportional Odds Model Under Random Censoring

Jean-Yves Dauxois
CREST-ENSAI
Campus de Ker-Lann, 35170, Bruz, France
jean-yves.dauxois@ensai.fr

Syed N.U.A. Kirmani
Department of Mathematics, University of Northern Iowa
Cedar Falls, Iowa 50614-0506 USA
kirmani@math.uni.edu

In practical applications, it is not uncommon for the hazard functions obtained for two groups to converge with time. One of the approaches to allow for converging hazard functions is the proportional odds model. We develop a procedure for testing the proportional odds assumption when the available data consists of two independent random samples of randomly right censored lifetimes. Asymptotic normality of the test statistic is proved and the testing procedure illustrated through application to two well-known survival data sets.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Goodness-of-Fit Tests for Location and Scale Families Based on a Weighted L_2-Wasserstein Distance Measure

Tertius de Wet
University of Stellenbosch, Department of Statistics and Actuarial Science
Victoria Street, Stellenbosch 7600, South Africa
TDEWET@akad.sun.ac.za

In two recent papers, del Barrio, Cuesta-Albertos, Matran and Rodriguez (1999) and del Barrio, Cuesta-Albertos and Matran (2000) considered a new class of goodness-of-fit statistics based on L_2-Wasserstein distance. They derived the limiting distribution of these statistics as quadratic functionals of the Brownian bridge process (which can be reduced to an infinite weighted sum of chi-one squared random variables) and showed that the normal distribution is the only location-scale family for which this limiting distribution has the "loss of degrees of freedom" property (ie the loss of terms in the infinite sum), due to the estimation of the unknown parameters.

In this talk we consider a weighted L_2-Wasserstein distance and show that these statistics retain the loss of degrees of freedom property for general classes of distributions, if applied separately to the location family and to the scale family and if the "right" weight function is used. These weight functions are such that the corresponding minimum distance estimators for the location parameter and for the scale parameter are asymptotically efficient. To get a loss of degrees of freedom, the weight function has to be chosen so that a function of it is an eigenfunction of a certain covariance kernel. Solving the integral equation defined in terms of this covariance kernel, gives an explicit expression for the weight function in terms of the underlying distribution.

For the location case, the normal and logistic cases are discussed as examples and for the scale case, the normal and exponential cases are discussed as examples. For these examples, the weight functions are obtained explicitly.

References

del Barrio, Eustasio, Cuesta-Albertos, Juan A, Matran Carlos and Rodriguez-Rodriguez, Jesus M (1999). Tests of goodness-of-fit based on the L_2-Wasserstein distance, *Ann. Statist.* **27,** 1230-1239.

del Barrio, Eustasio, Cuesta-Albertos, Juan A, and Matran, Carlos (2000). Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests, *TEST* **9,** 1-96.

Bandwidth Selection in Deconvolving Kernel Estimation

Aurore Delaigle
Université catholique de Louvain, Institut de Statistique
voie du Roman Pays, 20, 1348 Louvain-la-Neuve, Belgium
delaigle@stat.ucl.ac.be

We consider kernel estimation of a density from an iid sample that is contaminated by random noise. More precisely we observe an iid sample $Y_1,...,Y_n$, where for all i, $Y_i=X_i+Z_i$, with $X_i \sim f_X$ the density of interest, and where $Z_i \sim f_Z$ represents the random noise on the ith observation, independent of X_i. We also assume that the error distribution f_Z is known.

In this context Carroll and Hall (1988) and Stefanski and Carroll (1990) introduced a consistent estimator of the density f_X, the deconvolving kernel density estimator. See for example Stefanski and Carroll (1990), Fan (1991a,b,c,1992) or Wand and Jones (1995) for a theoretical study of the estimator. This estimation procedure requires the choice of a smoothing parameter depending on the sample size and called the bandwidth. The performance of the deconvolving kernel density estimator depends crucially on the choice of the bandwidth, but however very few papers focus attention on how to choose this parameter in practice.

Stefanski and Carroll (1990) and Hesse (1999) study a method of selection of the bandwidth, based on cross-validation techniques. Their procedure works in practice, but it suffers from a few drawbacks such as a large variability, multiplicity or non-existence of the solution. We propose two other practical methods of bandwidth selection that possess good theoretical properties and illustrate their performance in finite sample size via simulations.

The first method is based on an asymptotic approximation of the Mean Integrated Squared Error (MISE), derived by Stefanski and Carroll (1990) and Wand and Jones (1995). The bandwidth is then selected through minimization of the asymptotic MISE. This approximation still involves an unknown quantity that we estimate either by simply referring to a normal density or by a kernel method developed by Delaigle and Gijbels (2001c). The consistency of the method is provided and it is shown that this method can bring considerable improvement on the density estimation.

The second method uses a bootstrap approximation of the MISE as described in Delaigle and Gijbels (2001a). This estimation procedure requires the choice of a second bandwidth, and we show how to choose it properly. We establish the consistency of the bootstrap method. We also explain why we do not need to resample from the data in practice and thus unlike many other bootstrap procedures our method does not require extensive computations.

Finally we study the performance of our two practical methods and compare them with the cross-validation method, by presenting results of a simulation study conducted by Delaigle and Gijbels (2001b).

---

* This is joint work with Irène Gijbels, Université catholique de Louvain, Belgium

References

Carroll, R.J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, **83**, 1184 - 1186.

Delaigle, A. and Gijbels, I. (2001a). Bootstrap selection of the bandwidth in kernel estimation of a density from a contaminated sample. *Discussion paper #**0101***, Institut de Statistique, Université catholique de Louvain, Belgium.

Delaigle, A. and Gijbels, I. (2001b). Comparison of several data-driven bandwidth selection procedures in kernel estimation of a density when the data are contaminated by random noise. In preparation, Institut de Statistique, Université catholique de Louvain, Belgium.

Delaigle, A. and Gijbels, I. (2001c). Estimation of integrated squared density derivatives from a contaminated sample. In preparation, Institut de Statistique, Université catholique de Louvain, Belgium.

Fan, J. (1991a). Asymptotic normality for deconvolution kernel density estimators. *SankhyaA*, **53**, 97-110.

Fan, J. (1991b). Global behaviour of deconvolution kernel estimates. *Statistica Sinica*, **1**, 541-551.

Fan, J. (1991c). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, **19**, 1257-1272.

Fan, J. (1992). Deconvolution with supersmooth distributions. *The Canadian Journal of Statistics*, **20**, 155-169.

Hesse, C. (1999). Data-driven deconvolution. *Journal of Nonparametric Statistics*, **10**, 343-373.

Stefanski, L. and Carroll, R.J. (1990). Deconvoluting kernel density estimators. *Statistics*, **2**, 169 - 184.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

On Bivariate Beta Distributions

Fernanda Diamantino*
Faculty of Science, Department of Statistics and Operations Research
Bloco C2, Piso 2, Campo Grande-Cidade Universitária, 1749-016 Lisboa
fernanda@fc.ul.pt

The statistics $Q_{1n} = \frac{X_{n:n} - X_{n-1:n}}{X_{2:n} - X_{1:n}}$ and $Q_{2n} = \frac{X_{n:n} - X_{2:n}}{X_{n-1:n} - X_{1:n}}$ have been extensively studied by Mendonça (2000), for the investigation of tail weight, kurtosis and skewness of several models (see also Diamantino & Pestana, 1996).

We focus our interest in quotients of generalized spacings

$$Q_1^* = \frac{X_{n:n} - X_{k:n}}{X_{i:n} - X_{1:n}} \quad \text{and} \quad Q_2^* = \frac{X_{k:n} - X_{j:n}}{X_{i:n} - X_{l:n}}.$$

When the underlying model is a generalized Pareto, if i<k we get simplified formulas. Moments are, in this case, easy to compute, but $E\left[\frac{1}{X_{i:n} - X_{l:n}}\right]$ may be infinite.

The special case of Uniform parent distribution is dealt with in detail. And the results therein are used to present and investigate several multivariate Beta models, and to put forward simulation algorithms.

References

Arnold, B.C. (1983). Pareto Distributions. Statistical Distributions,Vol.5. ICPH.

David, H.A. (1981). Order Statistics , John Wiley & Sons.

Diamantino, F. and Pestana, M. (1996). Perturbações da Gaussiana – sua Influência em Estatísticas Studentizadas. *Notas e Comunicações do CEAUL* **15/96**, Universidade de Lisboa.

Johnson, N.L. and Kotz, S. (1972). Distributions in Statistics: Continuous Multivariate Distributions. John Wiley & Sons.

Mendonça, S. (2000). Tópicos sobre a Convergência Fraca de Sucessões de Variáveis Aleatórias. Tese de Doutoramento. Universidade da Madeira.

Rohatgi, V. K. (1984). Statistical Inference. John Wiley & Sons.

* Research partially supported by FCT/POCTI/FEDER

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Asymptotic Expansions of the Robbins-Monro Process

Jürgen Dippon
Universität Stuttgart, Mathematisches Institut A
Pfaffenwaldring 57, 70511 Stuttgart, Germany
dippon@mathematik.uni-stuttgart.de

To estimate the root x_0 of an unknown regression function $f : R \to R$, whose function values $f(x)$ at points x can be observed with some random error V only, Robbins and Monro (1951) suggested to run the recursion

$$X_{n+1} = X_n - \frac{a}{n} Y_n$$

with observation $Y_n = f(X_n) + V_n$ of $f(X_n)$ at step n. Under regularity assumptions, the normalized Robbins-Monro process (Z_n), given by $Z_n := (X_n - x_0)/\sqrt{Var(X_n)}$, is asymptotically normal.

In this talk we present Edgeworth expansions which provide approximations of the distribution function of Z_n up to an error of order $o(1/\sqrt{n})$ and $o(1/n)$, for instance

$$P\left(\frac{X_n - x_0}{\sqrt{Var(X_n)}} \le x \right) = \Phi(x) + \frac{1}{\sqrt{n}} p_1(x)\phi(x) + \frac{1}{n} p_2(x)\phi(x) + o\left(\frac{1}{n}\right)$$

where Φ and ϕ denote the distribution function and the density of the $N(0,1)$-distribution, respectively, and p_1 and p_2 are known polynomials.

As corollaries we obtain asymptotic confidence intervals for the unknown parameter x_0 whose coverage probability errors are of order $O(1/n)$. Further results concern Cornish-Fisher expansions of the quantile function of Z_n, an Edgeworth correction of the distribution function of Z_n, and a stochastic expansion of Z_n in terms of powers of both $1/\sqrt{n}$ and a standard normal random variable Z.

The proofs use ideas of Helmers, Callaert, Janssen, Veraverbeke, Bickel, Goetze and van Zwet who investigated Edgeworth expansions for L- and U-statistics.

Predictive Comparisons

Sofia Dias, Ian R. Dunsmore
University of Sheffield, Department of Probability and Statistics
The Hicks Building, Hounsfield Road, Sheffield S3 7RH, United Kingdom
S.Dias@sheffield.ac.uk, I.R.Dunsmore@sheffield.ac.uk

1. Introduction

In many areas of inference interest lies in comparisons between groups. Classical and Bayesian approaches developed to date usually involve comparison between parameters, such as mean values. However, the predictive paradigm suggests that comparisons should be made in terms of future observables, rather than parameters.

Our interest lies, within the Bayesian framework, in developing alternative approaches that will use the concept of predictive probability for a number of models that may arise in practice. Attention will be dedicated to the applications of this predictive probability to Medical Statistics, namely Clinical Trials, but the application of predictive probabilities is by no means restricted to this area.

If a clinical trial is conducted with the purpose of comparing the effects of two treatments T_1 and T_2, by considering the responses obtained Y_1 and Y_2, interest will be in determining which of these treatments has a better response for a future patient. Assuming samples from Y_1 and Y_2 are available, we can use the data and any prior information to estimate the effectiveness of the treatments in a future experiment, i.e. when given to a patient that has not entered the original trial, through a predictive probability.

Considering Y_1 and Y_2 as two random variables with distributions of known form but unknown parameters, and letting $\boldsymbol{\theta}$ represent the vector of all unknown parameters, we will write $\Pr(Y_1 < Y_2 \mid \boldsymbol{\theta})$ for the probability that Y_1 is less than Y_2 given these parameters. Classical parametric procedures attempt to estimate $\Pr(Y_1 < Y_2)$ by finding "best" estimators of the parameter $\boldsymbol{\theta}$ and 'plugging-in' the results in $\Pr(Y_1 < Y_2 \mid \boldsymbol{\theta})$. This relies on parameter (point or interval) estimates whose adequacy can never be checked in practice since parameters are not observable. A predictive approach however will give a measure that is expressed in terms of observable quantities and is therefore easier to interpret by non-statisticians. In particular, in the case of comparing two treatments using a Bayesian predictive methodology, we will be interested in estimating the probability that a future patient will have a better response on T_2 than on T_1. This can be written as $\Pr(Y_1 < Y_2 \mid \text{Data})$, assuming a large response is a good response. This predictive probability will be an *actual* probability: the probability of the event $\{Y_1<Y_2\}$, which patients and clinicians can easily relate to.

Assigning a prior distribution to $\boldsymbol{\theta}$, the informative experiment is used to obtain the posterior distribution of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta} \mid \text{Data})$. The predictive probability is then given by

$$\textbf{(1)} \qquad \Pr(Y_1 < Y_2 \mid \text{Data}) = \int_{\boldsymbol{\theta}} \Pr(Y_1 < Y_2 \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \text{Data})\, d\boldsymbol{\theta}.$$

The expression in (1) involves a multidimensional integral in $\boldsymbol{\theta}$, whose posterior distribution might be complicated, or even unattainable in closed form. Not only that, the form of $\Pr(Y_1 < Y_2 \mid \boldsymbol{\theta})$ may be an additional complicating factor. This integral will often be impossible to solve analytically and other methods such as numeric or Monte Carlo integration need to be considered.

2. Predicting $Pr(Y_1 < Y_2)$

Letting Y_1^f and Y_2^f represent the responses obtained by a future patient, we will present methods of predicting the probability that $Y_1^f < Y_2^f$ for the Normal, Normal Regression and the Ordinal Logistic Models.

Looking at (1) it is easily seen that it can be written as

$$\Pr(Y_1 < Y_2 \mid \text{Data}) = \mathrm{E}_{\text{post}}\left[\Pr(Y_1^f < Y_2^f \mid \boldsymbol{\theta})\right]$$

which is the mean of $\Pr(Y_1^f < Y_2^f \mid \boldsymbol{\theta})$ with respect to the posterior distribution of $\boldsymbol{\theta}$, termed the posterior mean (Enis and Geisser, 1971). If we can simulate a large enough sample of vectors $\{\boldsymbol{\theta}^{(k)}, k=1,...,\mathrm{M}\}$ from the posterior, we can approximate $\Pr(Y_1 < Y_2 \mid \text{Data})$ as

$$\Pr(Y_1 < Y_2 \mid \text{Data}) \approx \frac{1}{\mathrm{M}} \sum_{k=1}^{\mathrm{M}} \Pr(Y_1^f < Y_2^f \mid \boldsymbol{\theta}^{(k)}) .$$

In the Normal model, comparing future responses with equal variances based on (1) is straightforward but when the variances differ this is no longer the case. We will outline a method of approximating $\Pr(Y_1 < Y_2 \mid \text{Data})$ for this latter case, comment on the implications of comparing responses with different predictive variability and suggest possible alternatives to $\Pr(Y_1 < Y_2 \mid \text{Data})$. The case of correlated future observations when the informative experiments are independent is also of interest, since it may be more reasonable for predicting a future patient's response to different treatments.

If covariates are measured in the informative experiment, this information should be incorporated in predictions. When attempting to predict the outcome for a future patient, it is not unrealistic to assume that a preliminary observation of that patient has been carried out, providing the necessary information on the covariates. It will therefore be assumed that predictions will be obtained for cases where the values of the future covariates are known and that the covariates measured for patients on T_1 and patients on T_2 are the same. Letting $\boldsymbol{x}^f$ represent the vector of covariates for the future patient, we will present methodologies for predicting $\Pr(Y_1 < Y_2 \mid \boldsymbol{x}^f, \text{Data})$, for the Normal and Ordinal Logistic Regressions. Some comments on the predictive influence of covariates will also be made.

Acknowledgements

Sofia Dias gratefully acknowledges the sponsorship of *Fundação para a Ciência e Tecnologia* (grant no. PRAXIS XXI/BD/18301/98) to carry out and present this work as part of her research for a PhD.

References

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge University Press.

De Moraes, A. R. and Dunsmore, I. R. (1995). Predictive Comparisons in ordinal models, *Comm .in Statist – Theory and Methods.* **24**, 2145-2164.

Enis, P. and Geisser, S. (1971). Estimation of the probability that $Y < X$, *J. of the American Statist. Association.* **66**, 162-168.

Checking Extreme Value Conditions

Daniel Dietrich
University of Bern

Laurens de Haan
Erasmus University, Rotterdam
ldehaan@few.eur.nl

Jürg Hüsler
University of Bern
Juerg.huesler@stat.unibe.ch

Extreme value theory is useful mainly because it is the only realistic framework for extending the empirical distribution function or the empirical quantile function beyond the range of the available data. Since the extreme value context is unavoidable in this extension, not much attention has been paid to the problem of testing extreme value conditions, that is, checking whether the extreme value conditions are reasonable for a data set without specifying the shape parameter of the limit extreme value distribution. The present paper aims at providing such a check. Our approach is loosely related to the Cramér-von Mises test statistic

(1) $$n\int_{-\infty}^{\infty}\left(F_n(x)-F(x)\right)^2 dF(x)$$

in the following sense. Here F and F_n denote the distribution function and the empirical distribution function of a sample of size n, respectively. Two modifications will be made.

First, we consider the related statistic comparing inverse distribution functions rather than distribution functions:

(2) $$n\int_0^1\left(Q_n(y)-Q(y)\right)^2 \Big/ \left(Q'(y)\right)^2 dy$$

with Q the inverse distribution function and Q_n, the empirical quantile function. The criteria (1) and (2) have the same limiting distribution albeit under stronger conditions for the latter (see e.g. Csörgö and Revesz (1981), Cor. 5.4.1 and Cor. 5.5.2).

Secondly,a tail version is needed. For distribution tails criterion (2) seems slightly more natural. We present a tail version of it, with estimated parameters. An asymptotic expansion for the tail inverse empirical distribution function due to Drees (1998) will prove useful. As suggested by this expansion we consider the tail statistic,

$$E_{k,n} := \int_0^1\left(\frac{\log X_{n-kt,n}-\log X_{n-k,n}}{\hat{\gamma}_+}-\frac{t^{-\hat{\gamma}_-}-1}{\hat{\gamma}_-}\left(1-\hat{\gamma}_-\right)\right)^2 t^2 dt$$

for $k \square n$, where $\hat{\gamma}_+$ is an estimator for $\gamma_+ = \max(0,\gamma)$, $\hat{\gamma}_-$ is an estimator for $\gamma_- = \min(0,\gamma)$, and γ is the extreme value index. The factor t^2 has been introduced to ensure finiteness of all the integrals involved.

Another closely related interpretation of our test statistic is via the Pareto approximation: the observations exceeding a high threshold follow approximately one of the Pareto distributions

$$1-\left(1+\gamma x\right)^{-1/\gamma}, 1+\gamma x>0, \gamma \in \mathbb{R}$$

under proper normalisation (Balkema and de Haan (1974), Pickands (1975)), with the convention that for $\gamma = 0$ the distribution is equal to the exponential one: $1-\exp\left(-x\right)$. Now

$$\frac{\left(1-s\right)^{-\gamma}-1}{\gamma}$$

(similar to the function appearing to the right in the definiton of $E_{k,n}$) is the quantile function of the Pareto distribution which approximates high values of the log of the observations. For $\gamma = 0$, this is set to $-\log\left(1-s\right)$.

We prove that $E_{k,n} \to 0$ i.p. under the domain of attraction condition if $k = k\left(n\right) \to \infty, k/n \to 0$ as $n \to \infty$. Moreover, $kE_{k,n}$, has a specified limit distribution under an extra condition on the distribution function for sequences $k = k\left(n\right) \to \infty$ that do not increase too fast.

In a situation where only nonnegative values of γ play a role (for example for distributions which are unbounded in the right) a simplified version can be used:

$$T_{k,n} := \int_0^1 \left(\frac{\log X_{n-kt,n} - \log X_{n-k,n}}{\hat{\gamma}_+} + \log t \right)^2 t^2 dt.$$

Its properties are similar.

References

Balkema A.A. and L. de Haan (1974) Residual life time at great age. *Ann. Probab.* **2**, 792-808

Csörgö, M. and P. Revesz (1981) *Strong approximations in probability and statistics*. Akadémiai Kiadó, Budapest.

Drees, H. (1998). On smooth statistical tail functionals. *Scand. J. Statist.* **25**, No 1, 187-210.

Pickands, J. III (1975) Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119-131.

A New Testing Strategy for the Unit Root vs Dickey-Fuller Test: A Monte Carlo Comparison

Rafaela Dios-Palomares, Jose A. Roldan-Casas, Antonio Ramos-Millán
University of Cordoba, Department of Statistics
Avda. Menendez Pidal,14080, Cordoba, Spain
ma2rocaj@uco.es

1. Introduction

A problem arising in many time series applications is the question of whether a series should be differenced. This is equivalent to asking if the time series has a unit root.

Dickey and Fuller (1979, 1981) proposed some test statistics for the unit root hypothesis for a time series. They derive the finite and limiting distributions of test statistics for a unit root when the estimated model is a random walk, a random with shift in mean, and a random walk with shift in mean and a linear time trend. The distribution of Dickey and Fuller (DF) tests relied on the innovation process being white noise. In 1981 they extend the DF test to an AR(*p*) process which is called 'augmented' Dickey-Fuller (ADF) test.

On the other hand, Nankervis and Savin (1985, 1987) show that the statistics proposed by Dickey and Fuller yield non-similar tests of the unit root hypothesis. Non-similarity implies that the distribution of a test statistic is affected by the value, under the null, of a nuisance parameter. So, if the nuisance parameter is unknown, we can reject or not reject the null hypothesis wrongly and the consequences are low powers and size distortions. Due to low power and these size distortions an acceptance of the random walk hypothesis should be treated with caution.

A testing strategy which takes into account the non-similarity of Dickey-Fuller tests has been proposed by Roldan (2000) and Roldan and Dios (2000) to test the unit root hypothesis in the context of a first-order autoregressive process with unknown intercept and a linear trend.

In order to demonstrate the relevance of non-similarity and its consequences, we use Monte Carlo simulations to compare the performance of the strategy and the DF test in the context of the model mentioned above

2. Results and Conclusions

We compare the powers of the two-sided test of the random walk with drift hypothesis considered by Dickey and Fuller with the powers of the strategy in a Monte Carlo study using the model

$$Y_t = \mu + \beta\, t + \rho Y_{t-1} + e_t \qquad t = 1,2,\ldots,T$$

where μ, β and ρ are unknown real numbers and t is a linear trend. We assume that Y_0 is a known constant and equal to zero and the $\{e_t\}$ is a sequence of independent normal random variables with mean zero and variance σ_e^2.

Ten thousand samples of size $T = 50$, 100, 250 and 500 were generated for $\rho = 0.8$, 0.9, 1.00, 1.1, 1.2; $\mu = 0$, 1, 10; and $\beta = 0$, 0.1, 0.5, 1. All simulations were carried out using routines developed in Eviews 3.1 with the random number generator contained therein.

Table 1 reports Monte Carlo powers of 0.05 two-sided size tests (Dickey-Fuller test and strategy) for $\rho = 1$. These values are the highest estimated rejection probability for each T when the null hypothesis ($\rho = 1$) is true, that is, they represent the empircal size for each T considered in the experiment.

There are two conclusions to be drawn from results presented in Table 1. DF test presents important size distorsions since the empirical size is always greater than 0.71, increasing with T. However, the strategy only presents size distorsion at T = 50 and 100, since the empirical size tends to nominal size (0.05) as T increases.

We conclude the paper by comparing the powers of the strategy and the DF test at explosive and estable alterantives. For ρ >1 the powers of the strategy and the DF test are much the same and equal to 1 for all β, μ and T.

At stables alternatives (ρ <1) the strategy is uniformly more powerful than the DF test for all β, μ and T. The powers of both tests are strongly influenced by the values of T and β and are low when $T \leq 100$ and β is very close to zero. However, the powers converge to 1 as the sample size and the value of β increase. The results show much more rapid convergence in the strategy case.

T	Two-sided Dickey-Fuller test	Strategy
50	0.7145	0.6357
100	0.7314	0.6991
250	0.7428	0.0606
500	0.7457	0.0608

Table 1. Empirical size

Hence we recommend the strategy proposed by Roldán (2000) and Roldan and Dios (2000) since compared to the DF test the strategy has superior power at stable alternatives and its size distorsions dissapear when the sample size increases.

References

Dickey, D.A. and Fuller, W.A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root, *Journal of the American Statistical Association*, **74**, 427-431.

Dickey, D.A. and Fuller, W.A. (1981). Likelihood Ratio Statistics for Autoregressive Time Series With a Unit Root, *Econometrica*, **49**, 1057-1072.

Nankervis, J.C. and Savin, N.E. (1985). Testing the autoregressive parameter with the *t* statistic. *Journal of Econometrics*, **27**, 143-161.

Nankervis, J.C. and Savin, N.E. (1987). Finite sample distributions of *t* and *F* statistics in an AR(1) model with an exogenous variable. *Econometric Theory*, **3**, 387-408.

Roldán, J.A. (2000), Análisis sobre la detección de raíces unitarias desde la perspectiva de la no similaridad. Estudio de integración en el mercado del aceite de oliva, Ph.D. thesis, University of Cordoba, Cordoba, Spain.

Roldán Casas, J.A. y Dios Palomares, R. (2000). Análisis de detección de raíces unitarias en series de tiempo. Un enfoque metodológico con tests no similares, *Qüestiio*, Vol. **24**, 3, 415-440.

Inference for Observations of Integrated Diffusions

Susanne Ditlevsen
University of Copenhagen, Department of Biostatistics
Blegdamsvej 3, 2200 Copenhagen N, Denmark
s.ditlevsen@biostat.ku.dk

Michael Sørensen
University of Copenhagen, Department of Theoretical Statistics
Universitetsparken 5, 2200 Copenhagen N, Denmark
michael@math.ku.dk

1. Introduction

Estimation of parameters in diffusion models is usually based on observations of the process at discrete time points. Here we investigate estimation when a sample of discrete observations is not available but instead, observations of a running integral of the process with respect to some weight function. This type of observations could, for example, be obtained when a realization of the process is observed after passage through an electronic filter. The integrated process is no longer a Markov-process. A generalization of martingale estimating functions, namely prediction-based estimating functions (PBEF), is applied to estimate parameters in the underlying process. The estimators can be shown to be consistent and asymptotically normal.

2. Prediction-Based Estimating Functions Applied to Integrated Diffusions

Consider the one-dimensional diffusion $dX_t = b(X_t;\theta)dt + \sigma(X_t;\theta)dW_t, X_0 \sim \mu_0$, where θ is an unknown *p*-dimensional parameter belonging to the parameter space $\Theta \subseteq \mathrm{R}^p$ and *W* is a one-dimensional standard Wiener process. We assume that *X* is an ergodic, stationary diffusion with invariant measure μ_0, and that X_0 is independent of *W*. We assume that the stochastic differential equation has a unique weak solution. Suppose that a sample of observations at discrete time points is not available but instead, a running integral of the process with respect to some weight function. Suppose the interval of observation [0,*T*] is subdivided into *n* smaller intervals of length Δ=*T*/*n*, and let ν be a probability measure on the interval [0, Δ]. Our observations will then be

(1) $$Y_i = \int_0^\Delta X_{(i-1)\Delta+s} d\nu(s) \ ; \ i = 1,\ldots,n.$$

If our observations are obtained by integrating uniformly over the time axis, ν is simply the uniform distribution on [0,Δ] with density φ= 1/Δ, and we get the more simple observations

$$Y_i = \frac{1}{\Delta}\int_{(i-1)\Delta}^{i\Delta} X_s ds \ ; \ i = 1,\ldots,n.$$

Note that since X_t is stationary also Y_i is stationary. The problem is to estimate the parameter θ in the underlying process X. We solve it by applying the method of PBEF, introduced in Sørensen (2000). In the following we will briefly outline this method. Assume that $f_j, j = 1,\ldots,N$, are one-dimensional functions such that $E_\theta(f_j(Y_i)^2) < \infty$ for all $\theta \in \Theta$. We denote the expectation when θ is the true parameter value by $E_\theta(\cdot)$. Let

$P^{\theta}_{i-1,j}$, $j=1,\ldots,N$, be finite-dimensional, closed linear subspaces of the L^2-space of square integrable F_{i-1}-measurable one-dimensional variables when θ is the true parameter value. Here F_{i-1} is the σ-algebra generated by $Y_1,\ldots,Y_{i-1}$. The space $P^{\theta}_{i-1,j}$ can be interpreted as a set of predictors of $f_j(Y_i)$ based on $Y_1,\ldots,Y_{i-1}$. Let $P^{\theta}_{i-1,j}$ be spanned by $(1, Z_{j1}^{(i-1)},\ldots,Z_{jq_j}^{(i-1)})$. We assume that for $k=1,\ldots,q_j$, $Z_{jk}^{(i-1)} = h_{jk}(Y_{i-1},\ldots,Y_{i-r})$ are linearly independent. Here h_{jk} is a function, independent of *i* and θ, from $\mathbf{R}^r$ into **R**, and $r \in N$. Note that $Z_{jk}^{(i-1)}$ is well-defined only when $i \geq r+1$. We write the elements in $P^{\theta}_{i-1,j}$ as $a_0 + a^T Z_j^{(i-1)}$, where $a^T = (a_1,\ldots,a_{q_j})$ and $Z_j^{(i-1)} = (Z_{j1}^{(i-1)},\ldots,Z_{jq_j}^{(i-1)})^T$ are q_j-dimensional vectors. We denote transposition by ^{T}. We will study the following estimating function

$$G_n(\theta) = \sum_{i=r+1}^{n} \sum_{i=1}^{N} \Pi_j^{(i-1)}(\theta)(f_j(Y_i) - \pi_j^{(i-1)}(\theta))$$

where Y_i is of the form (1), $\Pi_j^{(i-1)}(\theta)$ is a *p*-dimensional stochastic vector, the coordinates of which belong to $P^{\theta}_{i-1,j}$, and $\pi_j^{(i-1)}(\theta)$ is the minimum mean square error predictor of $f_j(Y_i)$ in $P^{\theta}_{i-1,j}$. If, for instance, we take $f_j(y) = y$, $Z_{jk}^{(i-1)} = Y_{i-k}$ and φ= 1/Δ, we need to calculate the moments $E_\theta(Y_1)$ and $E_\theta(Y_1 Y_k)$ for $k=1,\ldots,r$. We have $E_\theta(Y_1) = E_\theta(X_0)$ and $E_\theta(Y_1 Y_k) = \frac{1}{\Delta^2}\int_0^{\Delta} \int_{(k-1)\Delta}^{k\Delta} E_\theta(X_s X_u)\, du\, ds$. Note that since X_t is stationary, $E_\theta(X_s X_u)$ is simply a function of the distance |s-u|.

3. The Optimal PBEF for Integrated Diffusions

Natural choices for $f_j(y)$ and $Z_{jk}^{(i-1)}$ would be $f_j(y) = y^{\alpha_j}$ and $Z_{jk}^{(i-1)} = Y_{i-l_k}^{\alpha_{jk}}$, where α_j and α_{jk} are such that $E_\theta(Y^{2\alpha_j\alpha_{jk}})$ exists. For simplicity we assume α_j and α_{jk} are integers. To derive the optimal PBEF, we need higher order moments of the form $E_\theta(Y_1^{k_1} Y_{t_1}^{k_2} Y_{t_2}^{k_3} Y_{t_3}^{k_4})$, where $1 \leq t_1 \leq t_2 \leq t_3$, as functions of the moments of X_t, moments we assume known or possible to simulate. Assume that $1 < t_1 < t_2 < t_3$, and that φ(t)=1/Δ, and let $(k_1+k_2+k_3+k_4) \leq 2\alpha_j\alpha_{jk}$. Arguments of symmetry yield that

$$E_\theta(Y_1^{k_1} Y_{t_1}^{k_2} Y_{t_2}^{k_3} Y_{t_3}^{k_4}) = \frac{k_1!k_2!k_3!k_4!}{\Delta^{(k_1+k_2+k_3+k_4)}} \int_0^{\Delta} dv_1 \int_{v_1}^{\Delta} dv_2 \cdots \int_{v_{(k_1-1)}}^{\Delta} dv_{k_1} \cdots\cdots \int_{(t_3-1)\Delta}^{t_3\Delta} dr_1 \int_{r_1}^{t_3\Delta} dr_2 \cdots \int_{r_{(k_4-1)}}^{t_3\Delta} E_\theta(X_{v_1}\cdots X_{r_{k_4}}) dr_{k_4}$$

where $r_{k_4} \geq \cdots \geq r_1 \geq s_{k_3} \geq \cdots \geq s_1 \geq u_{k_2} \geq \cdots \geq u_1 \geq v_{k_1} \geq \cdots \geq v_1$. Thus an explicit PBEF can be found if we know explicit expressions of the moments of X_t. It can be shown that the estimators are consistent and asymptotically normal, given conditions on X_t.

References

Sørensen, M. (1999): On asymptotics of estimating functions. *Braz.J.of Prob.Stat.*, **13**, 111-136.
Sørensen, M. (2000): Prediction-Based Estimating Functions. *Econ.J.*, **3**,123-147.

Estimation for Discrete Distributions

Louis G. Doray
Département de mathématiques et de statistique, Université de Montréal
C .P. 6128, Succursale Centre-ville, Montréal Canada
doray@DMS.UMontreal.CA

Andrew Luong
École d'actuariat, Université Laval
Québec, Canada
alnong@act.ulaval.ca

1. Recurrence Relationship

Many useful discrete distributions have a complicated pmf involving various forms of series expansion, as in the cases of the Hermite, or Poisson-generalized inverse Gaussian distribution. Classical methods for estimating the parameters of the distribution, such as the likelihood method, which are based on the pmf can be difficult to implement. For these pmf's, the recursive relationship between successive probabilities can be used for estimating the parameters.

Let us consider parametric families where the recursive relationship between successive probatilities of the pmf can be written as an homogeneous difference equation of order r,

(1) $$p_i = \phi_1(\theta,i)\,p_{i-1} + \ldots + \phi_r(\theta,i)\,p_{i-r} \quad i = a+r, a+r+1, \ldots, m$$

where

a) $\theta = \left[\theta_1, \ldots, \theta_p\right]$ is the vector of parameters, $\theta \in \Theta \subseteq \mathbb{R}^p$

b) $p_i = P\left[X = i\right], \quad i = a+r, \ldots, m$

c) $\phi_1, \ldots, \phi_r$ are fumtions determined by the parametric family, functions assumed to be differentiable.

With a relationship of order 2 or less, we will be able to represent most parametric families found in Johnson et al. (1992); the Poisson, negative binomial, binomial, logarithmic, zeta and ETNB distributions, the Generalized Yule, Good, Generalized Poisson and Exponential family can all be represented by (1) with $r = 1$. The Hermite, Polya-Aeppli and Sichel distribution satisfy the recurrence formula of order 2, as well as the Poisson-Pareto and Poisson-inverse gamma distributions. Certain mixed Poisson distributions, such as the Poisson-Weibull and the Poissontransformed gamma will yield a recurrence relationship of order *r* greater than 2 (see Willmot (1993)).

2. Asymptotic Properties of QDE

Let n_i represent the number of observations which take the value i in the sample $X_1, \ldots, X_n$, let $\hat{p}_i$ represent the relative frequency $\hat{p}_i = n_i/n$ and let us define

$$f_i = \phi_1(\theta,i)\,p_{i-1} + \ldots + \phi_r(\theta,i)\,p_{i-r}.$$

Using relation (1) and fixing a value for k with $k \le m$, we then have the representation $p_i = f_i(\theta) + \varepsilon_i$. In practice, the choice of k is made so that $n_a, \ldots, n_k$ are all positive.

Using matrix notation, let $\hat{p} = [\hat{p}_{a+r}, \ldots, \hat{p}_r]'$, $f(\theta) = [f_{a+r}(\theta), \ldots, f_r(\theta)]'$ and $\varepsilon = [\varepsilon_{a+r}, \ldots, \varepsilon_k]'$. We then have $\hat{p} = f(\theta) + \varepsilon$. The mean and the variance-covariance matrix of ε are given by $E(\varepsilon) = 0$ and $\Sigma(\theta)$. Let $\Sigma = \Sigma(\theta_0)$, where θ_0 is the true vector value of the parameter. $\Sigma(\theta)$ can be obtained using the variance-covariance matrix of a multinomial distribution. Also, let us define $\Sigma^*(\theta) = n\Sigma(\theta)$ and $\Sigma^* = \Sigma^*(\theta_0)$. Σ^* differs from Σ only by a known constant multiple n.

The quadratic distance estimator (QDE) θ is defined by the vector value which minimizes

(2) $$Q(\theta) = [\hat{p} - f(\theta)]' \Sigma^{*-1} [\hat{p} - f(\theta)] = u'(\theta) \Sigma^{*-1} u(\theta)$$

where we define $u_i(\theta) = \hat{p}_i - f_i(\theta)$ and $u(\theta)$ is the vector $[u_{a+r}(\theta), \ldots, u_k(\theta)]'$.

The estimator $\hat{\theta}$ obtained by replacing Σ^{*-1} by a consistent estimate $\hat{\Sigma}^{*-1}$ in (2) is a consistent estimator of θ, (i.e. $\hat{\theta} \xrightarrow{p} \theta_0$); it can be shown, with a Taylor series' expansion and the multivariate central limit theorem, that it has an aymptotically normal distribution

(3) $$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{L} N(o, \Sigma_{\hat{\theta}}), \text{ with } \Sigma_{\hat{\theta}} = (S' \Sigma^{*-1} S)^{-1},$$

where the matrix $S = (s_{ij})$ has elements

$$s_{ij} = E\left(\frac{\partial f_i}{\partial \theta_j}\right) = \sum_{l=1}^{i-r} \frac{\partial \phi_{i-l}(\theta, i)}{\partial \theta_j} \hat{p}_l \text{, evaluated at } \theta = \theta_0 .$$

Thus, the asymptotic variance-covariance matrix of $\hat{\theta}$ is $(1/n)\Sigma_{\hat{\theta}}$.

Admittedly, there is some arbitrariness for fixing a value for k. The QDE remains consistent for all choices of values for k. For efficiency sake, we should fix k at a large value or let $k \to \infty$, as the sample size $n \to \infty$. For robustness sake, we might fix $k = k_0$, discarding, possible outlier observations at the tail, or values exceeding k_0.

A variation of the above general QD method which can lead to simplifications in computations exists for parametric families which allow a recursive relationship of order 1 (see Doray and Luong (1997)).

References

Doray, L.G. and Luong, A. (1997). Efficient estimators for the good family. *Communication in statistics: Simulation and Computation*, **26**, 1075-1088.

Johnson, N.L., Kotz, S and Kemp, A.W. (1992). Univariate discrete distributions. Wiley, New York.

Willmot, G.E. (1993). Recursive evaluation of mixed Poisson probabilities and related quantities. *Scandinavian Actuarial Journal*, 114-133.

Tail Dependence in Bivariate EVT

Gerrit Draisma
Econometrisch Instituut, Erasmus Universiteit Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
draisma@few.eur.nl

Holger Drees
Institute for Applied Mathematics, Ruprecht-Karls-University
Im Neuenheimer Feld 294, 69120 Heidelberg, Germany
hdrees@statlab.uni-heidelberg.de

Ana Ferreira
Eurandom, Technical University Eindhoven
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
On leave from ISA, Universidade Tecnica de Lisboa, Portugal
ferreira@eurandom.tue.nl

Laurens de Haan
Econometrisch Instituut, Erasmus Universiteit Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
ldehaan@few.eur.nl

Suppose (X_i, Y_i) is a sequence of i.i.d. random vectors with d.f. F. In this paper we are interested in probabilities of the type

(1) $$P_r\{X > u \ \ and \ \ Y > v\},$$

where u and v are large threshold values. The probability that both thresholds are exceeded is of interest, e.g., if the levels of two different air pollutants, the losses suffered in two different investments or different variables relevant for the probability of a flooding are observed. Since only large values of X and Y are involved, one would expect the multivariate extreme value theory to provide the appropriate framework for systematic estimation of the probability (1). Namely, the assumption that there exist normalising constants $a_n, c_n > 0$ and b_n, d_n real values such that

(2) $$\lim_{n\to\infty} F^n(a_n x + b_n, c_n y + d_n) = \lim_{n\to\infty} P_r\left\{\frac{\max\{X_i,\cdots,X_n\} - b_n}{a_n} \le x \ , \frac{\max\{Y_i,\cdots,Y_n\} - d_n}{c_n} \le y\right\} = G(x,y)$$

for all but denumerable many vectors (x, y). Here G is a distribution function with non-degenerate marginals (Resnick, 1987 - cf. Chapter 5).

Unfortunately, if the marginals of the limit distribution are independent, in which case we say that the maxima of the X_i and those of the Y_i are asymptotically independent, the previous limit assumption is of little help. Indeed, this is a rather common situation; for instance, it holds for nondegenerate bivariate normal distributions.

In order to overcome this problem, Ledford and Tawn (1996,1997,1998) introduced a quite general sub-model, where the tail dependence is characterised by a coefficient $\eta \in]0,1]$. More precisely, in the setting with uniform marginals, they assumed that the function $t \mapsto P_r\{1-X<t \ \ and \ \ 1-Y<t\}$ is regularly varying at 0 with index $1/\eta$. Then $\eta = 1$ in case of asymptotic dependence, whereas $\eta < 1$ implies asymptotic independence. This sub-model can also be used to device a test for asymptotic independence in the basic relation (2).

Let (X,Y) have continuous marginal distribution functions F_1 and F_2. Our basic assumption is that

$$\textbf{(3)} \qquad \lim_{t \downarrow 0} \frac{\dfrac{P_r\{1-F_1(X)<tx \ \ and \ \ 1-F_2(Y)<ty\}}{q(t)} - c(x,y)}{q_1(t)} = c_1(x,y)$$

exists, for $x, y \geq 0$ (but $x+y>0$), with q positive, $q_1 \to 0$ as $t \to 0$ and c_1 non-constant and not a multiple of c. Moreover we assume that convergence is uniform on $\{(x,y) \in [0,\infty[^2 : x^2+y^2=1\}$.

We propose a new estimator of the parameter η, introduced by Ledford and Tawn (1996). We prove asymptotic normality of this estimator and two other estimators proposed in the quoted paper. Our estimator for η is inspired by the work of Peng (1998).

Also a procedure is set up to estimate the probability of a failure set that works under asymptotic dependence as well as under asymptotic independence. Under our model, we prove consistency of this estimator.

References

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

Ledford, A. and Tawn, J.A. (1997). Modelling dependence within joint tail regions. *J. Royal Statistical Society*, B, **59**, 475-499.

Ledford, A. and Tawn, J.A. (1998). Concomitant tail behaviour for extremes. *Adv. Appl. Prob.*, **30**, 197-215.

Peng, L. (1999). Estimation of the coefficient of tail dependence in bivariate extremes. *Statistics & Probability Letters*, **43**, 399-409.

Resnick, S.I. (1987). Extreme Values, Regular Variation, and Point Processes. Springer-Verlag.

Functional Data Analysis of Complex Computer Simulation Output: A Case Study on Nuclear Waste Disposal Risk Assessment

David Draper
University of California, Department of Applied Mathematics and Statistics
Baskin School of Engineering, 1156 High Street
Santa Cruz CA 95064 USA

Bruno Mendes
University of Bath
UK
mendes@cse.ucsc.edu

A key issue in the consolidation process of the nuclear fuel cycle is the safe disposal of radioactive waste. Deep geological disposal based on a multibarrier concept is at present the most actively investigated option (visualize a deep underground facility within which radioactive materials such as spent fuel rods or reprocessed waste, previously encapsulated, are placed, surrounded by other man-made barriers). While the safety of this concept ultimately relies on the safety of the mechanical, chemical and physical barriers offered by the geological formation itself, the physico-chemical behavior of such a disposal system over geological time scales (hundreds or thousands of years) is far from known with certainty.

From 1996 to 1999, with partners in Italy, Spain, and Sweden, we were involved in a project for the European Commission, GESAMAC, which aimed in part to capture all relevant sources of uncertainty in predicting what would happen if the disposal barriers were compromised in the future by processes such as geological faulting, human intrusion, and/or climatic change. One major goal of the project was the development of a methodology to predict the radiologic dose for people in the biosphere as a function of time, how far the disposal facility and the other components of the multibarrier system are underground, and other factors likely to be related to dose. For this purpose we developed a complex computer simulation environment called GTMCHEM which "deterministically" models the one-dimensional migration of radionuclides through the geosphere up to the biosphere. We describe the application of methods of functional data analysis (FDA) to explore the dependence of predicted radiologic dose curves as a function of time on inputs to the computer simulations.

FDA includes extensions of traditional statistical methods such as principal components analysis and the analysis of variance (ANOVA) to the case where the outcome, instead of a single real number, is a curve, in our case the logarithm of radiologic dose as a function of the logarithm of time. Previous work in this field was limited to methods such as ANOVA applied to the maximum of such curves; FDA thus permits a much more complete investigation of the relationship between dose and time, and how this relationship depends on the computer simulation inputs.

REVISTA DE
ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Propagating Model Uncertainty in Geochemical Calculations

David Draper
University of California, Department of Applied Mathematics and Statistics
Baskin School of Engineering, 1156 High Street
Santa Cruz CA 95064 USA

Bruno Mendes
University of Bath
UK
mendes@cse.ucsc.edu

The safe disposal of radioactive waste from nuclear power plants is an important problem in energy policy. The most actively investigated option at present is deep geological disposal based on a multibarrier concept (visualize a facility far underground within which radioactive materials such as spent fuel rods or reprocessed waste, previously encapsulated, are placed, surrounded by other man-made barriers). If such a facility is compromised in the future, for instance through geological faulting, the most likely means of transport for the escaping radionuclides is groundwater passing through the rock between the disposal facility and the surface. Thus geochemistry, which studies the behavior of such radionuclides in solution, is central to this energy policy problem.

The basic chemical problem faced in this work is how ions interact in a medium of high ionic concentration. Considerations from physical chemistry have led to a collection of competing mathematical models, without a clear consensus as to which model is best. In the work presented here we use Bayesian hierarchical modeling to capture and propagate the uncertainty across these models in predictions of radionuclide activity coefficients. We use two competing models for the calculation of the activity coefficients of ions: the Specific Interaction Theory (SIT) and the Pitzer equations. Working with experimental data in which groundwater samples are contaminated with radionuclides, we use Markov chain Monte Carlo methods to assess the posterior plausibility of the two competing theories and to produce well-calibrated predictive distributions, for the concentration of radionuclide contaminant in solution, that correctly account for model uncertainty.

On the Minimax Regret Estimation of a Restricted Normal Mean, and Implications

Bernd Droge
Humboldt-Universität zu Berlin, Sonderforschungsbereich 373
Unter den Linden 6, 10099 Berlin, Germany
droge@mathematik.hu-berlin.de

We consider the problem of estimating the mean of a normal distribution with known variance, when that mean is known to lie in a bounded interval. In a decision-theoretic framework we study finite sample properties of a class of nonlinear estimators. These estimators are based on thresholding techniques which have become very popular in the context of wavelet estimation. Under squared error loss we show that there exists a unique minimax regret solution for the problem of selecting the threshold.

For comparison we investigate the properties of a variety of competitors such as the maximum likelihood estimator, the minimax linear estimator and the minimax regret linear estimator. It turns out that, for example, the latter estimator may dominate even the optimal nonlinear threshold estimator in cases where the prior information is strong compared to the noise level. In most cases, however, the nonlinear estimation approach is preferable.

By examples we illustrate the implications of our results for the problem of estimating the regression function in a nonparametric situation. This is possible since, as usual, a coordinatewise application of the scalar results leads immediately to results for multivariate (sequence space) problems. Then it is well known that orthogonal transformations can be employed to turn statements about estimation over coefficient bodies in sequence space into statements about estimation over classes of smooth functions in noisy data.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Spectral Analysis of Fractional Brownian Motions

Kacha Dzhaparidze
CWI
Kruislaan 413, 1098 SJ Amsterdam, The Netherlands
Kacha@cwi.nl

A fractional Brownian motion with Hurst index different from ½ is neither a Markov process nor a semimartingale. It is, however, a self-similar Gaussian process of a simple covariance structure: the increments constitute a stationary process with a power structure function (see e.g. Yaglom (1986), Example 3 on p. 406). This allows one to carry out thorough time domain analysis by rather elementary means, see Norros et al. (1999) and references therein. Basic results are a linear representation of a fractional Brownian motion through a standard Brownian motion and vice versa, as well as explicit formulae for prediction and likelihood inference.

In parallel, one can cover these and related problems in a frequency domain. The spectral density of the stationary process of increments is again a simple power function. In this paper we associate with this function Toeplitz forms in the spirit of Grenander and Szego (1958) and derive their extremal properties. To this end we need to extend basic notions of the existing theory to the present continuous time situation and to define counterparts to Szego polynomials, their reciprocals and associated reproducing kernel polynomials. It is shown that new notions retain useful properties, for instance, we again have recurrence relationships and the Christoffel-Darboux formula. Similarly to the theory of Toeplitz forms in discrete time, the results obtained in the frequency domain can be translated in terms of random processes in question. In this way new aspects of the linear theory of fractional Brownian motions are brought forward and a new light is shed on known results.

References

Norros, I., Valkeila E. and Virtamo, J. (1999). An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motions, *Bernoulli* **5**, 571-587.

Grenander, U. and Szego, G. (1958). *Toeplitz Forms and Their Applications*. University of California Press. Berkeley.

Yaglom, A. M. (1987). *Correlation Theory of Stationary and Related Random Functions I.* Springer. New York.

Core Distances

Zden□k Fabián
Institute of Computer Sciences, Academy of Sciences of the Czech republic
Pod vodárenskou v□ží 2, Prague, Czech Republic
zdenek@uivt.cas.cz

1. Core Function

The core function of continuous probability distribution has been recently introduced by Fabián (2001). Density and core function represent equivalent description of distributions which are regular in the usual sense. However, description by means of the core function is usually simpler.

Definition The core function of a distribution Q with support $S_Q=R$ is

$$T_Q(x) = -\frac{q'(x)}{q(x)}$$

where q is the density of Q. The core function of $P=Q\varphi^{-1}$ with support $S_P \neq R$ is

(1) $$T_P(x) = T_Q(\varphi^{-1}(x))$$

where $\varphi: R \to S_P$ is a 'suitable' smooth one-to-one mapping.

It has been shown by Fabián (2001) that for most currently used model distributions the word 'suitable' means a generalization of Johnson´s transformations (Johnson, 1949)

(2) $$\varphi: R \to (0,\infty): \quad x = \varphi(y) = e^y$$
$$\varphi: R \to (0,1): \quad x = \varphi(y) = e^y/(1+e^y)$$

for arbitrary $S_P \neq R$. Moreover, it has been proved that (1) can be expressed indepen- dently of Q by means of density p of distribution P and of Jacobian $L(x) = \varphi'(\varphi^{-1}(x))$ of the transformation $\varphi: R \to S_P$ as

(3) $$T_P(x) = \frac{1}{p(x)} \frac{d}{dx} [-\sigma L(x)\, p(x)] .$$

(3) is also valid for parametric distributions P_θ, $\theta \in \Theta \subset R^m$ with a general structure $\Theta = S_P \times (0,\infty) \times \Delta$, $\Delta \in R^{m-2}$. Thus $\theta = (\tau, \sigma, \lambda)$ where

$$\tau = \varphi(\mu)$$

is the transformed location parameter μ of distribution Q, which we call *induced location,* σ is the scale and $\lambda \in \Delta$ other (shape) parameters.

The sense of the core function is stated by the following theorem (Fabián 2001).

Theorem If P_θ has the induced location □, its core function is the inner part of the maximum likelihood score for □, $s_□(x|\theta) = (\partial/\partial\tau) \log p_\theta(x)$.

Examples The core function of the standard normal distribution with density $q(x)=\exp(-x^2/2)/\sqrt{2\pi}$ is $T_Q(x)=x$, the core of the standard lognormal $P=Q□^{-1}$ is $T_P(x)=T_Q(\log x)=\log x$. □ is the induced location of the exponential distribution with density $p(x)=1/□\ \exp(-x/□)$ and core $T_P(x)=x/□-1$. The Weibull distribution with density $p(x)=□x^{-1}(x/□)^□\ \exp(-(x/□)^□)$ has likelihood score for □ in the form $s_□(x)=□□^{-1}((x/□)^□ -1)$ and the core $T_P(x) = (x/□)^□ -1$ is its inner part. The core of the beta distribution with density $p(x)=B(□,□)^{-1}x^{\alpha-1}(1-x)^{\beta-1}$ where B is the beta function is $T_P(x)=(□+□)x-□$, the function which was yet unknown.

2. Core Distances in the Sample and Probability Spaces

Denote by F the distribution function of P. Taking into consideration the existence (Fabián, 1997) of moments of core functions,

$$M_k = \int_{S_P} T_P(x)^k \, dF(x) ,$$

we examined the distance of points $x_1, x_2 \in S_P$ in the sample space,

(4) $$d_T(x_1,x_2) = \frac{1}{\sqrt{M_2}} \, | T_P(x_2) - T_P(x_1) | ,$$

and the mean square distance of core functions of distributions P_1, P_2,

(5) $$D_T(P_1|P_2) = \frac{1}{M_2} \int_{S_P} [T_{P_2}(x) - T_{P_1}(x)]^2 \, dF_1(x) .$$

Let us present some results.

(i) Consider a distribution P_θ with density $p_\theta(x)$. It holds that

$$d_T(x_1,x_2) = \frac{1}{\sqrt{I(\theta)}} \, | s_\square(x_2|\theta) - s_\square(x_1|\theta) |$$

where $I(\theta)$ is the Fisher information for τ. Thus, in cases of distributions with induced location d_T represents the distance introduced at S_P by the maximum likelihood estimator of this parameter.

(ii) The zero of the core function, x^*: $T_P(x^*)=0$, minimizes the mean square distan-ce (4) between x^* and any other possible sample from P. x^* (in a parametric case $x^*=\tau$) can thus be considered as an alternative 'centre of mass' of the distribution P.

(iii) The distance (5) has properties typical for 'good' distances of distributions, namely it is invariant to sufficient data transformations. It can be taken as a new distance of distributions P_1, P_2.

(iv) Distances (5) within the members of parametric families are often simpler than other distances. Using the mean square difference of core functions instead of functionals of the ratio of densities we avoid in distance formulas the terms originating due to the norming constants. As an example, the core distance between two normal distributions, $N(\mu_j,\sigma_j), j=1,2$, is

(6) $$D_T(P_{\mu_1,\sigma_1}|P_{\mu_2,\sigma_2}) = \frac{1}{2}\left[\frac{(\mu_1-\mu_2)^2}{\sigma_2^2} + \left(\frac{\sigma_1}{\sigma_2} - 1\right)^2\right] .$$

If $\sigma_1 = \sigma_2$, (6) reduces to the Kullback-Leibler and Rényi distance. The distance of two members of the exponential family is simply $D_T(P_1|P_2)=(\tau_1/\tau_2-1)^2$. Moreover, distances (5), not influenced by somewhat arbitrary functions (f-divergences), exhibit consistent behaviour: they are finite in cases when core functions (which are proportional to the influence functions for the induced location parameter) of the family are bounded and vice versa. We hope that they can be used in testing hypotheses about estimated parameters.

References

Fabián, Z. (1997). Information and entropy of continuous random variables. *IEEE Trans. on Information Theory*, **43**, 3, 1080-83.

Fabián, Z. (2001). Induced cores and their use in robust parametric estimation. *Comun. in Statistics, Theory and Methods*, **30**, 3.

Johnson, N.L. (1949). Systems of frequency curves generated by methods of translations, *Biometrika* **36**, 149-176.

Multi-Site Modelling of Rainfall Based on the Neyman-Scott Process

Anne-Catherine Favre
Université de Neuchâtel, chaire de statistique appliquée
Espace de l'Europe 4, case postale 1825, 2002 Neuchâtel, Suisse
anne-catherine.favre@unine.ch

This paper studies rainfall model that are based on point processes. First, it considers models applicable to a single site, in which storms arrive randomly according to a Poisson process, each storm consisting of a random number of cells that deposit a random amount of rain for a random period. The arrivals of cells form a stochastic series of points subject to clustering. Two models have been used in the literature to represent such a clustered point process : the Neyman-Scott process and the Bartlett-Lewis process.

The paper proposes a new method of parameter estimation for the Neyman-Scott Rectangular Pulses Model (NSRPM). It also derives the statistical properties of this new estimator, which permits the construction of confidence intervals.

To take into account the spatial variability of precipitation, a multi-site model (MS-NSRPM), reflecting the underlying spatial-temporal structure of rainfall directly through between-sites interactions, has been developed. The new model consists of a two-dimensional rainfall process, which marginally is a Neyman-Scott process in each dimension. The association between the two processes is handled through the generation of correlated random variables and the thinning of a Poisson process representing storms at a base (fictitious) location. Figure 1 shows a simplified representation of the multi-site model.

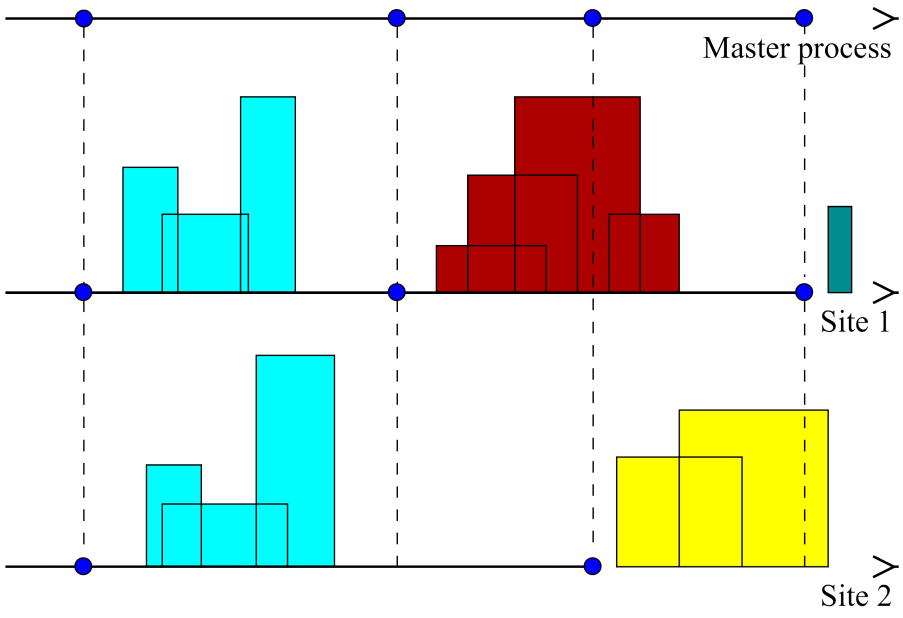


Figure 1. Schematic depiction of the multi-site Neyman-Scott rectangular pulses model: master process and bivariate correlated generation.

Two of the biggest advantages of this model are its simplicity and flexibility. An extensive analysis of 23 rainfall stations situated on the Swiss Plateau has provided guidance for the development of the model. The model has been validated by its capacity of reproducing the cross-correlation function at various time lags and the probability that two sites are simultaneously dry during an arbitrary time interval of given length. The theoretical values of the cross-correlation has also been derived.

References

Cowpertwait, P.S.P. (1995). A generalized spatial-temporal model of rainfall based on a clustered point process. *Proc. Roy. Soc. London Ser.* A., **450**, 163-175

Favre, A.C. (2000). Single and Multi-Site Modelling of Rainfall Based on the Neyman-Scott Process, Ph.D. thesis no 2320, Swiss Federal Institute of Technology, Lausanne.

Lawrance, A.J. and Lewis, P.A.W. (1983). A new autoregressive time series model in exponential variables (near(1)). *Adv. In Appl. Probab.*, **13**(4),826-845.

Bayesian Wavelet Regression on Curves with Application to a Spectroscopic Calibration Problem

Tom Fearn
University College London, Department of Statistical Science
Gower Street, London WC1E 6BT, UK
tom@stats.ucl.ac.uk

Philip J. Brown
University of Kent, Institute of Mathematics and Statistics
Canterbury, Kent CT2 7NF, UK
philip.j.brown@ukc.ac.uk

Marina Vannucci
Texas A&M University, Department of Statistics
College Station, TX 77843, USA
mvannucci@stat.tamu.edu

The talk will describe some recently published work (Brown, Fearn and Vannucci, 2001) in which a spectroscopic calibration problem is tackled by regression on selected wavelet coefficients. The problem is to predict the composition of biscuit doughs from measurements of their near-infrared spectra, a regression problem where the predictor variables correspond to a continuous curve measured at hundreds of discrete points. This is tackled by applying a wavelet transform to the curve (the spectrum) and then selecting, using Bayesian methodology, a modest number of wavelet coefficients as predictor variables.

Reference

Brown, P.J., Fearn, T. and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem, *J. Amer. Statist. Assoc.* 96, to appear.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Nonparametric Simulated Maximum Likelihood Estimator

Jean-David Fermanian
ENSAE, CREST
fermania@ensae.fr

Bernard Salanié
CREST, Univ. Chicago

Let $(x_t, y_t)_{t=1,\dots,T}$. be an i.i.d. sample of vector valued r.v. satisfying the regression model $y_t = g(x_t, \theta, \varepsilon_t)$ for each *t*. The disturbance ε follows a known distribution, *g* is known and θ belongs to a compact subset of $\mathbb{R}^q$. The associated loglikelihood is $L_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} \ln l_t(\theta)$, denoting $l_t(\theta)$ the law of y_t knowing (x_t, θ). Instead of maximizing the loglikelihood, it can be sometimes easier to maximize an approximation of this quantity. Nonetheless, the criterion functions have sometimes no analytical expressions. This is often due to the presence of integrals of large dimensions. This theoretical and numerical difficulty has been circumvented by simulation techniques. The main idea is to use a random function whose expectation provides the criterion. This function is called a simulator. This related expectation is approximated drawing independent realizations of some underlying random variables.

The application of this general principle has lead to the following methods of estimation: Simulated Nonlinear Feast Squares (SNLS), Simulated Maximum and Pseudo-Maximum Likelihood (SML and SPML), Method of Simulated Moments (MSM)...

In this paper, we propose to approximate each term $l_t(\theta)$ by a kernel estimator based on some i.i.d. simulated sample $(\varepsilon_t^s)_{s=1,\dots,S}$ drawn from the law of ε. Denoting $y_t^s = g(x_t, \theta, \varepsilon_t^s)$, the likelihood $l_t(\theta)$ is estimated by the kernel method viz

$$l^S(y_t \mid x_t, \theta) \equiv l_t^S(\theta) \equiv \frac{1}{Sh^m} \sum_{s=1}^{S} K\left(\frac{y_t - y_t^s}{h}\right)$$

Here, *h* it denotes a bandwidth sequence which tends to 0 when *S* tends to the infinity.

For technical reasons, it is necessary to trim the too small values of l_t^S. It can be done by considering the nonparametric simulated loglikeIihood

$$\tilde{L}_T^S(\theta) = \frac{1}{T}\sum_{t=1}^{T} \tau_S\left(l_t^S(\theta)\right) \ln l_t^S(\theta),$$

where τ_S is a sufficiently regular function such that $\tau_S(x) = 0$ if $|x| < h^\delta$ and $\tau_S(x) = 1$ if $|x| < 2h^\delta$, $\delta > 0$. Thus, our estimator is defined by

$$\hat{\theta}_T^S = \arg\max_{\theta \in \Theta} \tilde{L}_T^S(\theta).$$

Under some regularity conditions, it is shown that $\hat{\theta}_T^S$ is strongly consistent, asymptotically normal and asymptotically efficient, when *S* and *T* grow to the infinity simultaneously at some convenient rates. Moreover, the same methodology can be applied to nonlinear dynamic models.

Bayesian Inference for Exponential Populations under Double Failure-Censoring

Arturo J. Fernández, Francisco J. Salmerón
Universidad de La Laguna, D.E.I.O.C.
C/Astrofísico Fco. Sánchez, s/n
La Laguna, Tenerife. Spain
ajfernan@ull.es

Marta I. López
Universidad de La Laguna, Dpto Econ. Instituc., Estadística y Econometría
Camino de la Hornera, s/n
La Laguna, Tenerife. Spain
mlopez@ull.es

1. Introduction and Modelling

Statistical methods for survival data and other time-to-event data are widely used in many fields. When managing this kind of data, some lifetimes of individuals may be censored. By censored data we mean that, in a potential sample of size *n*, a known number of observations may be missing at either end (single censoring) or at both ends (double censoring). This type of censoring is often called Type II- or failure-censoring. In this paper the important exponential lifetime model is considered and studied from a Bayesian perspective based on a doubly failure-censored sample. Consider then a random sample of size *n* from an exponential lifetime distribution with unknown mean μ, and let $x_r,...,x_s$ be the ordered observations remaining when the $(r-1)$ smallest and the $(n-s)$ largest observations have been censored. The likelihood function for μ given $\mathbf{x}=(x_r,...,x_s)$ is proportional to:

(1) $$L(\mu\,|\,\mathbf{x}) \propto \mu^{-m}\exp\{-\xi(\mathbf{x})/\mu\}\{1-\exp(-x_r/\mu)\}^{r-1},$$

where $m=s-r+1$ and $\xi(\mathbf{x})=\sum_{i=r}^{s}x_i+(n-s)x_s$.

In the case of censored data, *ML* estimators may be of limited value, so it is important in our situation to asses a prior distribution for μ. In this paper we consider prior densities of the form

(2) $$g(\mu)\propto \exp(-a/\mu)\mu^{-(b+1)},\ \mu>0,$$

where to be a proper (inverted gamma) density, we must have $a>0$ and $b>0$. The moment- and percentile-matching methods may be used to fit (2). With prior ignorance about μ, Jeffreys' prior, $(a=0,b=0)$ can reasonably be accepted.

By combining (1) with (2), the posterior density of μ is obtained to be

(3) $$g(\mu\,|\,\mathbf{x})=\frac{\{a+\xi(\mathbf{x})\}^{b+m}\exp\left[-\{a+\xi(\mathbf{x})\}/\mu\right]\{1-\exp(-x_r/\mu)\}^{r-1}}{\mu^{b+m+1}\Gamma(b+m)F_r\left[a+\xi(\mathbf{x}),b+m\right]},\ \mu>0,$$

where $F_r[u,v]=\sum_{j=0}^{r-1}(-1)^j\binom{r-1}{j}(1+j\,x_r/u)^{-v},\ u,v>0\ \ (F_1\equiv 1)$.

2. Point and Interval Estimation

Under squared-error loss function, the Bayes estimator of μ is given by

$$\tilde{\mu} = E[\mu \mid \mathbf{x}] = \frac{F_r\left[a+\xi(\mathbf{x}), b+m+1\right]}{F_r\left[a+\xi(\mathbf{x}), b+m\right]}\left\{\frac{a+\xi(\mathbf{x})}{b+m-1}\right\}, \quad (b+m>1).$$

Similarly, the Bayes estimators of the hazard rate, $\lambda = 1/\mu$, and the survival function at $t>0$, $R_t = R(t \mid \mu) = \Pr(X > t \mid \mu)$, are $\tilde{\lambda} = E[\lambda \mid \mathbf{x}]$ and $\tilde{R}_t = E[R_t \mid \mathbf{x}]$. If the absolute-error loss function is deemed suitable, the posterior medians represent the appropriate Bayes estimators. If there is no compelling reason to accept some specific loss function, the highest posterior density (*HPD*) estimator will be used.

Another common Bayesian approach to inference is to present credible sets or intervals for μ. For unimodal posterior density (3), the $100(1-\alpha)\%$ *HPD* credible interval $[c_1, c_2]$ for μ must simultaneously satisfy $S(c_1 \mid \mathbf{x}) - S(c_2 \mid \mathbf{x}) = 1-\alpha$ and $g(c_1 \mid \mathbf{x}) = g(c_2 \mid \mathbf{x})$.

3. Hypothesis Testing

In Bayesian analysis, the problem of deciding whether μ lies in Ω_0 or in Ω_1, where Ω_0 and Ω_1 are two disjoint subsets of the parameter space $\Omega = (0,\infty)$ is solved by just calculating the posterior probabilities $p_0(\mathbf{x}) = \Pr(\mu \in \Omega_0 \mid \mathbf{x})$ and $p_1(\mathbf{x}) = \Pr(\mu \in \Omega_1 \mid \mathbf{x})$ and deciding between H_0 and H_1 accordingly. When considering a "$0-k_i$" loss (i.e., $L(\mu, a_i) = 0$ if $\mu \in \Omega_i$ and $L(\mu, a_i) = k_i$ if $\mu \in \Omega_{i-1}$, where a_i represents the action of accepting H_i, $i = 0,1$), action a_1 will be taken if and only if (*iff*) $B_{10}(\mathbf{x}) > (k_1 q_0)/(k_0 q_1)$, where $q_0 (q_1)$ denotes the positive prior probability of $\Omega_0 (\Omega_1)$, $B_{10}(\mathbf{x}) = \{p_1(\mathbf{x})/p_0(\mathbf{x})\}/(q_1/q_0)$ is the Bayes factor in favour of Ω_1.

An appropriate approach to conduct a Bayesian test of the form $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ is to give μ_0 a probability $q_0 > 0$, while giving $\mu \in \Omega_1 = \Omega - \{\mu_0\}$ the density $q_1 g_1(\mu)$, where $q_1 = 1 - q_0$ and g_1 is a proper density.

Another method states that the null hypothesis $H_0 : \mu = \mu_0$ can reasonably be accepted *iff* the $100(1-\alpha)\%$ *HPD* credible set for μ, $C_\alpha(\mathbf{x})$, contains μ_0. It is limited to cases where the prior information on μ is vague; in particular, where $q_0 = 0$.

In order to test $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$, a reasonable loss function is "$0 - K_i(\mu)$" loss, i. e., $L(\mu, a_i) = K_i(\mu) I(\mu \in \Omega_{1-i})$, where $K_0(\mu)$ and $K_1(\mu)$ are non-decreasing positive functions of $(\mu - \mu_0)$ and $(\mu_0 - \mu)$, respectively. The Bayes test rejects H_0 *iff* $\int_{\Omega_1} K_0(\mu) g(\mu \mid \mathbf{x}) d\mu > \int_{\Omega_0} K_1(\mu) g(\mu \mid \mathbf{x}) d\mu$, where $\Omega_0 = (0, \mu_0]$ and $\Omega_1 = (\mu_0, \infty)$. In particular, under "$0-k_i$" loss and prior (3), the Bayes test rejects $H_0 : \mu \leq \mu_0$ *iff* $S(\mu_0 \mid \mathbf{x}) > k_1/(k_0 + k_1)$.

Multiscale Reconstruction and Extrapolation of Fractional Random Fields

R. Fernandez Pascual
University of Jaen, Department of Statistics and Operations Research
Jaen, Spain
rpascual@ujaen.es

M.D. Ruiz-Medina, J.M. Angulo
University of Granada, Department of Statistics and Operations Research
Faculty of Sciences, Campus Fuente Nueva s/n
Granada, Spain
mruiz@ugr.es, jmangulo@ugr.es

Random field signal processing problems arise in different applied areas such as engineering (Ekstrom, 1982), medicine and biology (Christakos, 1998; Louis, 1992; Lubbig, 1995), geophysics (Campi, 1980; Santosa and Symes, 1988; Trampert, Leveque and Cara, 1992), hydrology (Dietrich and Newsam, 1989; Kitanidis and Vomvoris, 1983; Sun, 1994), metereology (Huang and Cressie, 1996; Wikle and Cressie, 1997), etc. In the case where the random model of interest represents the input of the system, and the random output model is known, with some prior information about the random input model being available, the problem can be formulated as an inverse reconstruction problem. When the available information comes from the observation of the output random field in a subregion of its domain, the problem can be formulated as an inverse extrapolation problem. Under suitable conditions, the Orthogonal Projection Theorem provides, in both cases, a least-squares linear pointwise approximation of the random input. In this paper, however, we are interested in a functional linear approximation of the input random field, which leads to considering a weak-sense formulation of these problems. That is, we study the problems of functional linear inverse reconstruction and extrapolation of an input random field in a system defined by an integral equation. These problems are solved using the theory of reproducing kernel Hilbert spaces and of distributions on fractional Sobolev spaces. More specifically, we consider a class of generalised ordinary random fields with associated reproducing kernel Hilbert space (RKHS) isomorphic to a fractional Sobolev space. For this class, the mentioned problems can be solved in the second-order weak-sense. Furthermore, a discretization of the information available in both cases is obtained by considering a wavelet-based orthogonal expansion of the type derived in Angulo and Ruiz-Medina (1998, 1999) for the random fields involved. Truncation of these orthogonal expansions leads to a finite-dimensional approximation of the inverse reconstruction and extrapolation problems considered.

An important example of random field linear systems within the class studied in this paper is given by fractional integration of fractional Brownian motion with different orders of regularity. Systems of this type often arise in the study of processes that display long-range dependence, for example, in turbulence theory (Anh *et al*., 1998, 1999). For these systems, both the input and output random fields admit second-order weak-sense linear representations, in terms of generalised random functions with associated RKHSs isomorphic to appropriate fractional Sobolev spaces. Hence, the wavelet-based approach presented in this paper can be applied to solving the corresponding inverse reconstruction and extrapolation problems. Simulation studies using different values of the parameters defining the system considered in this example have been developed to illustrate this approach.

References

Angulo, J.M. and Ruiz-Medina, M.D. (1998). A series expansion approach to the inverse problem. *J. Appl. Prob.* **35**, 371-82.

Angulo, J.M. and Ruiz-Medina, M.D. (1999). Multiresolution approximation to the stochastic inverse problem. *Adv. Appl. Prob.* **31**, 1039-57.

Anh, V.V., Angulo, J.M. and Ruiz-Medina, M.D. (1999). Possible long-range dependence in fractional random fields. *J. Statist. Plan. Infer.* **80**, 95-110.

Anh, V.V., Angulo, J.M., Ruiz-Medina, M.D. and Tieng, Q. (1998). Long-range dependence and second-order intermittency of two-dimensional turbulence. *Environm. Model. & Soft.* **13**, 233-238.

Campi, S. (1980). An inverse problem related to the travel time of seismic waves. *Bolletino della Unione Matematica Italliana* B **17**, 661-674.

Christakos, G. and Hristopulos, D.T. (1998). Spatio-temporal environmental health modelling. Kluwer Academic Publishers.

Dietrich, C.D. and Newsam, G.N. (1989). A stability analysis of geostatistical approach to aquifer transmissivity identification. *Stoch. Hydrol. Hydraul.* 293-316, Springer-Verlag,.

Ekstrom, M. (1982). Realizable Wiener filtering in two dimensions. IEEE Trans. on Acoustics, *Speech and Signal Proc.* **30**, 31-40.

Huang, H.C. and Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Comput. Stat. and Data Anal.* **22**, 159-175.

Kitanidis, P.K. and Vomvoris, E.G. (1983). A geostatistical approach to the inverse problem for generalized random variables. *Inverse Problems* **5**, 599-612.

Louis, A.K. (1992). Medical imaging: state of the art and future development. *Inverse Problems* **8**, 709-738.

Lubbig, H. (ed). (1995). The inverse problem. Akademie Verlag.

Santosa, F. and Symes, W.W. (1988). Computation of Hessian for least-squares solutions of inverse problems of reflection seismology. *Inverse Problems* **4**, 211-233.

Sun, N.Z. (1994). Inverse problems in groundwater modeling. Theory and applications of transport in porous media. Kluwer Academic Publishers.

Trampert, J., Leveque, J.J. and Cara, M. (1992). Inverse problems in seismology. *In Inverse Problems in Scattering and Imaging* (eds M. Bertero and E.R. Pike), 345-369. Adam Hilger.

Wikle, K. and Cressie, N. (1997). A dimension-reduction approach to space-time Kalman filtering. *Preprint* 97-24, Statistical Laboratory, ISU, Departament of Statistics, Iowa State University, Ames, Iowa (USA).

Statistical Analysis of Infectious Diseases

Juan Ferrandiz, Antonio López
Universitat de Valencia, Departament d'Estadística i I.O.
Dr. Moliner 50, 46100 Burjassot, Spain.
Juan.Ferrandiz@uv.es, Antonio.Lopez@uv.es

Pilar Sanmartín
Universidad Politécnica de Cartagena, Dep. de Matemática Aplicada y Estadística
Paseo Alfonso XIII s/n Cartagena (Spain)
pilar.sanmartin@upct.es

Ferran Martinez
Intituto de Salud Carlos III
Sinesio Delgado 6, 28029 Madrid, Spain
fmartinz@isciii.es

1. Introduction

Infectious diseases dynamics is usually expressed in terms of the Susceptible-Infective-Removed states of members in the community under study. Models based on this framework are denoted with the acronym SIR (see Becker and Britton(1999)).

For discrete time data the *chain binomial epidemic model* offers two parameters of easy interpretation: the number of susceptible individuals s_t at time t, and the probability of any of these susceptible individuals being infected π_t. The number of new infected individuals appearing at time t+1, y_{t+1} is considered a Binomial(s_t, π_t) variable and, accordingly, the number of susceptible individuals decreases to $s_{t+1} = s_t - y_t$. The probability π_t should depend on the number of infective individuals at time t, and it seems reasonable to take this probability increasing with this number of infective individuals.

Although this models has been conceived for household outbreaks, we can borrow its main structure to cope with aggregated data in a larger temporal and geographical scale. It offers a simple and natural explanation of the bell-shaped form of epidemics waves and can explain the non-linear structure of infectious diseases. However in this case, the number of susceptible and infective individuals are usually unknown. The extension of such models to the stochastic time series approach is consider by Finkestadt and Grenfell (2000) for the study of chilhood diseases as measles (TSIR models).

By the other hand, Geographical studies of epidemiological data over time require space-time regression models in order to capture the influence of covariates and spatial interactions along the studied period. When considering infectious diseases, a direct influence of mortality/morbidity values between neighbouring regions is expected. Spatial auto-regressive models, such as Besag's auto-models (see Besag (1974)) and Cressie(1993)), seem a natural and intuitive approach.

In this paper we consider a dynamic auto-Poisson model which embodies the SIR structure as in Finkestadt and Grenfell (2000) but taking into account the spatial dependences of infectious diseases. We apply this model to the study of meningitis mortality data in Spain from 1960-1990.

2. Dynamic Auto-Poisson Model

We consider data collected in a given geographical region (provinces in a country) in a large temporal scale (such as years). Dealing with meningitis data, we face an important infra-declaration in such reported cases due to the large proportion of asymptomatic infective individuals who are not detected as sick. They are active transmitters of the disease and become immune eventually decreasing the number of susceptible individuals.These limitations impede a detailed assignment of s_t as before and we will resort to an indirect estimate as in Finkestadt and Grenfell (2000). Let be y_{jt}, r_{jt}, s_{jt} and π_{jt} the number of cases, reported cases, number of susceptibles and probability of infection at time t and location j, for j=1,…,L and t=1,…,T. (with s_{jt} being the number of susceptible at the end of the aggregated period). If ρ_{jt} denotes the reported rate, we have $y_{jt} = \rho_{jt} r_{jt}$. Moreover, if the probability of infection is small with respect to the number of susceptible population, we can consider :

$$Y_{jt} \sim Po(\lambda_{jt}) \quad \log(\lambda_{jt}) = \log(\pi_{jt}) + \log(s_{jt-1})$$

For every region j, $\log(\pi_{jt})$ depends on infective individuals in past at this site and neighbouring regions ($\delta(j)$):

$$\log(\pi_{jt}) = \beta_1 y_{j,t-1} + \beta_2 \sum_{i \in \delta(j)} \omega_{ij} y_{i,t-1}$$

with ω_{ij} fixed weights. As in Finkestadt and Grenfell (2000) we approach $\log(s_{jt})$ in a first step, using the SIR mechanism and locally linear regressions (see Fan and Gijbels (1996)). ρ_{jt} are also estimated at this stage. Taking $y_{jt} = \hat{\rho}_{jt} r_{jt}$ we can apply GLM procedures in order to obtain parameter estimations for (β_1, β_2) (see Fahrmeir and Tutz (94)). The study of meningitis mortality data in Spain from 1960-1990 is performed with this model in order to undertake the mechanism of the disease and perform one-step-ahead predictions.

References

Becker, N.G. and Britton, T. (1999) Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society,* series B **37**, 297-328.

Besag. (1974) Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society ,* series B **36**, 192-225.

Cressie. (1991). *Statistics for Spatial Data*. Wiley. New York.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall. London.

Fahrmeir, L. and Tutz G., (1994*). Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag. London.

Finkenstadt, B. and Grenfell, B. (2000) Time series modelling of childhood diseases: A dynamical systems approach, *Applied Statistics.* **49**, 187-205.

Estimation and Testing of the Pareto Index

Alena Fialova
Charles University, Fac. of Math. and Physics
Dep. of Prob. and Math. Statistics
Sokolovská 83, 186 00, Praha 8
Czech Republic
fialova@karlin.mff.cuni.cz

Assuming that the distribution function F is heavy tailed satisfying

(1) $$\lim_{x\to\infty} \frac{-\log\ (1-F(x))}{m\ \log x} = 1$$

we propose a new type of inference on the parameter m, namely an estimate of m and a test of the hypothesis

$$H_{m_0} : F \text{ satisfies (1) and}$$

(2) $$\liminf_{x\to\infty} x^{m_0}(1-F(x)) \geq 1 .$$

The inference is based on the tail behavior of the sample mean $\overline{X}_n$ under (1), which is shown to distinguish sharply the type of tails. The estimate and test are based on the empirical distribution function of $\overline{X}_n$ taken from the k independent samples of fixed size n and the inference is asymptotic as k to infinity. The asymptotic (normal) distribution of the estimate and of the test criterion under H_{m_0} is derived. Numerical results demonstrate good properties of both estimate and the test of m.

References

Csörg□, S., Deheuvels, P. and Mason, D. (1985). Kernel estimates of the tail index of a distributions. *Ann. Statist*., **13**, 1050—1077

Davis, R.A. and Resnick, S.T. (1984). Tail estimates motivated by extreme value theory. *Ann. Statist.*, **12**, 1467—1487

Dekkers, A.L.M., Einmahl, J.H.J. and de Haan, L. (1989). A moment estimator for the index of an extreme value distribution. *Ann. Statist*., **17**, 1833—1855

Embrechts, P., Kl□ppelbelg, C. and Mikosch, T. (1997). *Modelling Extremal Events*. Springer-Verlag.

de Haan, L. (1975). On regular variation and its application to the weak convergence of sample extremes. *Math Centre Tracts* **32**. Centre for Mathematics and Computer Science, Amsterdam.

Jureckova, J. (1981). Tail-behavior of location estimators. *Ann. Statist*., **9**, 578—585

Jureckova, J. (2000). Test of tails based on extreme regression quantiles. *Statist & Probab. Letters*, **49**, 53—61

Mason, D.M. (1982). Laws of large numbers for sums of extreme value. *Ann. Probab*., **10**, 756—764

Pickands, J. III (1975). Statistical inference using extreme order statistics, *Ann. Statist*. **3**, 119-131

The Performance of a Test of Discordancy for the Bingham Distribution

Adelaide Figueiredo
FEP
Rua Dr. Roberto Frias, 4200 Porto – Portugal
Adelaide@fep.up.pt

Paulo Gomes
INE
Rua de Vilar, 235-9º, 4050 – 626 Porto – Portugal
Paulo.Gomes@ine.pt

1. Introduction

Suppose the data represent a sample of *p* normalized variables for a given set of *n* individuals. The selection of those variables from a population of variables had new contributions in the context of a Bingham population (see Gomes, 1987 and Figueiredo, 2000). These authors admitted that the sample of variables comes from a Bingham population and then it is important to verify if there are any outliers in the sample.

In this paper to test the null hypothesis that all variables of the sample come from a Bingham distribution, we generalize to the *n*-sphere a test of discordancy proposed by Best (1986) for the sphere. We analyze the performance of the test in the case of an alternative hypothesis that admits one variable from a Bingham distribution with another concentration parameter or with another directional parameter and the remain variables of the sample from the same distribution of the null hypothesis.

2. Test of Discordancy

Let $X = \left[\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, ... \,|\, \mathbf{x}_p\right]$ be a sample of p variables from a Bingham distribution. We consider a case particular of the Bingham distribution $B_n(\mathbf{u},\xi)$ with density probability function given by

(1) $$f(\mathbf{x}) = K \exp\left(\xi \left(< \mathbf{u}, \mathbf{x} >\right)^2\right) \quad \mathbf{x} \in S_n,\ \mathbf{u} \in S_n,\ \xi > 0$$

where K is a normalizing constant, S_n represents the surface of the *n*-sphere; **u** is a directional parameter and ξ a concentration parameter.

Let *w* be the largest eigenvalue of $X\ {}^tX$ and *w(j)* the largest eigenvalue of $X\ {}^tX$-$\mathbf{x}_j\ {}^t\mathbf{x}_j$, where $\mathbf{x}_j$ is the variable that we remove from the sample. The statistic of the test is defined by

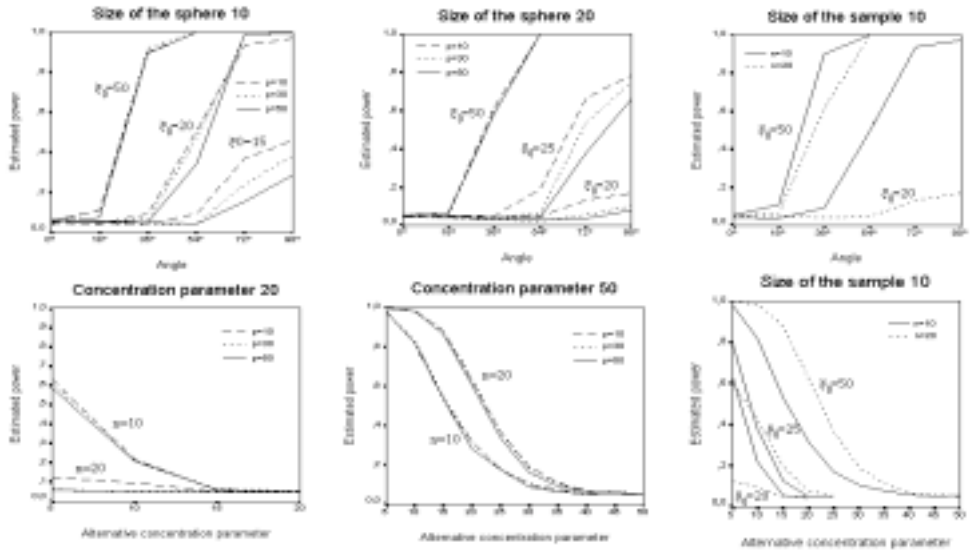
(2) $$H_p^{(1)} = \max_j \frac{2\xi_0 \left(1 + w(j) - w\right)/(n-1)}{2\xi_0 \left(p - 1 - w(j)\right)/(p-2)(n-1)}$$

3. Simulation Study

To evaluate the performance of the test we use the power that is the probability of the test reject the most discordant variable.

We determine the estimate of the power of the test for some sizes of the sphere, some sizes of the sample and some values of the parameters. We consider *n*=10,

p=10,30,50, ξ_0=15,20,50 and n=20, p=10,30,50, ξ_0=20,25,50. We use two models A and B. For both models the null hypothesis H_0 is that all variables of the sample come from the Bingham distribution $B_n(\mathbf{e}_n, \xi_0)$. In model A the alternative hypothesis H_1 is that one of the p variables comes from $B_n(\mathbf{u},\xi_0)$, where the angle between $\mathbf{u}$ and $\mathbf{e}_n$ is θ, and the others *p-1* variables come from $B_n(\mathbf{e}_n, \xi_0)$. In model B the alternative hypothesis H_1 is that one variable of the sample comes from $B_n(\mathbf{e}_n, \xi_1)$, where $\xi_1<\xi_0$ and the others *p-1* variables from $B_n(\mathbf{e}_n, \xi_0)$. We determine the 0.95-empiric percentiles of the statistic, using 10000 replicates of that statistic, under the null hypothesis. Then, we use 3000 replicates of the statistic, under the alternative hypothesis to estimate the power of test in both models.



4. Conclusion

Firstly, in model A the estimated powers of the test increase with the angle θ since if ξ_o is not very small. Additionally, for an angle θ fixed, the estimated power increases with ξ_o and for ξ_o not very small the estimated power decreases with n if θ is not very small.

Secondly, in model B the estimated powers decrease with ξ_1, since the parameter ξ_o is not very small. Additionally, for ξ_1 fixed, the estimated power of the test increases with ξ_o and increases with n if ξ_o sufficiently large for each n considered.

Finally, the estimated power of the test decreases with the sample size, which is valid for every discordancy test. This is due to the fact: as the size of the sample becomes larger, more variables are in alternative with the same distribution of the null hypothesis and consequently smaller is the effect of the contaminant variable.

References

Best, D. and Fisher, N.I. (1986). Goodness-of-fit and discordancy tests for samples from the Watson distribution on the sphere. *Australian Journal of Statistics*, vol.**28**, nº1, p.13-31.

Gomes, P. (1987). Distribution de Bingham sur la *n*-sphere: une nouvelle approche de l'Analyse Factorielle. Thèse d'État. Université des Sciences et Techniques du Languedoc – Montpellier.

Figueiredo, A. (2000). Classificação de variáveis no contexto de um modelo probabilístico definido na *n*-esfera. Tese de Doutoramento. Faculdade de Ciências da Universidade de Lisboa.

The Total Median in Statistical Quality Control*

Fernanda Otília Figueiredo
Oporto University, CEAUL, School of Economics
otilia@fep.up.pt

Maria Ivette Gomes
Lisbon University, CEAUL, DEIO, School of Sciences
ivette.gomes@fc.ul.pt

Control charts are tools widely used in industry to detect abnormal behaviour in manufacturing processes. In general we assume that the process observations are from a normal population with mean μ and standard deviation σ. However, even if in potential normal situations there is some possibility of having disturbances in the data, and we have then the need of finding an efficient and robust estimator to monitor the process parameters. Simulation studies for some symmetric and asymmetric distributions related to the normal allow us to suggest the total median as a robust location estimator, and we here analyse the robustness of the total median chart comparatively to the sample mean chart.

Let $(X_1,\ldots,X_n)$ be a random sample of size *n* taken from a population *F*(.). The total median is given by

$$MdT=\sum_{i=1}^{n}\sum_{j=i}^{n}\alpha_{ij}\,\frac{X_{(i)}+X_{(j)}}{2},\ \alpha_{ij}=P\left(MdB=\frac{x_{(i)}+x_{(j)}}{2}\right),\ 1\le i\le j\le n,$$

where *MdB* denotes the median of the bootstrap sample $(X_1^*,\ldots,X_n^*)$ associated to the *n* observations $(x_1,\ldots,x_n)$, through a resampling with replacement. The total median can also be expressed as a linear combination of the sample order statistics, $MdT=\sum_{i=1}^{n}a_i X_{(i)}$, where the coefficients a_i are related with the previous coefficients α_{ij}. The notation $X_{(i)}, 1\le i\le n$, is used for the ascending order statistics associated to the sample $X_i, 1\le i\le n$. Simulation studies have been used to evaluate the performance of the estimators $\overline{X}$ and *MdT* in terms of the resulting mean square error. We have concluded that the sample mean is an efficient estimator for the mean value of a symmetric distribution with moderate tails or with moderate skewness, although it is not robust. For heavy-tailed and/or high-skewed distributions we suggest the total median as a robust and efficient location estimator for small-to-moderate samples. Details may be found in Figueiredo and Gomes (2000). More details about robust estimators and their properties can be found in Hoaglin et al. (1983).

To investigate the robustness of the $\overline{X}$ and the *MdT* control charts to monitor the parameters of a process we have computed the false alarm rate of these charts, in order to evaluate deviations from normality. A similar study of robustness of the *EWMA* control chart to non-normality was done by Borror et al. (1999). In our study we have considered, without loss of generality, known μ and σ, and we have computed the 3-sigma control limits for samples of size *n* = 3, 4, 5, 10, 20. Tables 1 and 2 present those control limits and the rate of false alarms whenever we consider standardized data from some distributions with different skewness and tail-weight, γ and τ, respectively. Both charts $\overline{X}$ and *MdT* cannot be considered robust to

* Research partially supported by FCT / POCTI / FEDER.

deviations from the normality assumption because the false alarm rates are quite variable. This behaviour may be due to the skewness of the sampling distribution of the control statistics and to the consideration of 3-sigma control limits. Nevertheless, for small samples, the differences to the normal-case are smaller when considering the *MdT* chart.

Distribution	γ	τ	n =3	n=4	n=5	n=10	n=20
Normal	0.000	1.000	0.00266	0.00267	0.00280	0.00277	0.00275
$\chi^2_{(20)}$	0.632	1.001	0.00428	0.00385	0.00365	0.00318	0.00327
$\chi^2_{(10)}$	0.894	1.004	0.00550	0.00486	0.00442	0.00356	0.00330
Exponential	2.000	1.062	0.01175	0.01053	0.00931	0.00665	0.00489
T(20)	0.000	1.067	0.00343	0.00319	0.00315	0.00299	0.00280
Lognormal	1.750	1.143	0.01706	0.01679	0.01611	0.00635	0.00481
T(10)	0.000	1.145	0.00458	0.00428	0.00387	0.00335	0.00309
Logistic	0.000	1.212	0.00509	0.00459	0.00413	0.00356	0.00310
$\chi^2_{(1)}$	2.828	1.218	0.01557	0.01412	0.01280	0.00934	0.00669
T(3)	0.000	1.721	0.01167	0.01117	0.01070	0.00910	0.00791
Control Limits			±1.73205	±1.50000	±1.34164	±0.94868	±0.67082

Table 1. False alarm rate of 3-sigma control limits $\overline{X}$ chart

Distribution	γ	τ	n=3	n=4	n=5	n=10	n=20
Normal	0.000	1.000	0.00270	0.00262	0.00276	0.00283	0.00277
$\chi^2_{(20)}$	0.632	1.001	0.00405	0.00346	0.00295	0.00227	0.00237
$\chi^2_{(10)}$	0.894	1.004	0.00511	0.00415	0.00316	0.00193	0.00268
Exponential	2.000	1.062	0.00950	0.00704	0.00433	0.00101	0.00064
T(20)	0.000	1.067	0.00325	0.00288	0.00268	0.00236	0.00206
Lognormal	1.750	1.143	0.01201	0.00805	0.00302	0.00097	0.00087
T(10)	0.000	1.145	0.00397	0.00340	0.00264	0.00181	0.00141
Logistic	0.000	1.212	0.00449	0.00356	0.00273	0.00161	0.00108
$\chi^2_{(1)}$	2.828	1.218	0.01150	0.00797	0.00431	0.00057	0.00003
T(3)	0.000	1.721	0.00814	0.00550	0.00197	0.00020	0.00003
Control Limits			±1.74689	±1.52029	±1.39027	±1.02141	±0.75144

Table 2. False alarm rate of 3-sigma control limits *MdT* chart

Overall conclusion: although the *MdT* is an efficient and robust location estimator for small-to-moderate samples, the use of control charts based on this statistic must be carefully planned. In some situations of non-normality the false alarm rate of this chart can be much smaller than expected, particularly as *n* increases.

References

Borror, C.N., Montgomery, D.C. and Runger, G.C. (1999). Robustness of the EWMA Control Chart to Non-normality. *Journal of Quality Technology* **31**, 309-316.

Figueiredo, F. and Gomes, M.I. (2000). Estimadores robustos de localização e escala. *Proceedings VIII Annual Congress of S.P.E., 134-145.*

Hoaglin, D.C., Mosteller, F., and Tukey, L.G. (1983). Análise exploratória de dados. Técnicas robustas. John Wiley and Sons, New York.

Modelling Porcine Muscle Fibre Patterns

Sofia Fonseca
University of Aberdeen & Biomathematics and Statistics Scotland, Rowett Research
Meston Building, Dept. of Mathematical Sciences, AB24 3UE
Aberdeen, UK
sofia@bioss.sari.ac.uk

Graham Horgan
Biomathematics and Statistics Scotland, Rowett Research Institute
Greenburn Road, Bucksburn,Aberdeen AB219SB
Aberdeen UK
graham@bioss.sari.ac.uk

Ian Wilson
University of Aberdeen
Meston Building, Dept. of Mathematical Sciences, AB24 3UE
Aberdeen, UK
i.wilson@maths.abdn.ac.uk

Charlotte Matin
Rowett Research Institute
Greenburn Road, Bucksburn,Aberdeen AB219SB
Aberdeen, UK
C.Maltin@rri.sari.ac.uk

1. Background

Skeletal muscle constitutes the largest single organ of the body as it makes up 40-45% of the total body mass in humans and other mammals. Muscle fibres are elongated cells arranged in parallel bundles that make up around 75-90% of the total muscle volume. They can be classified into at least 3 groups according to their biochemical (glycolytic/oxidative) and contractile (fast/slow) properties: (i) slow twitch oxidative (SO); (ii) fast twitch oxidative glycolytic (FOG) and (iii) fast twitch glycolytic (FG).

Their arrangement varies between species and is particularly unusual in pigs, which present SO fibres organised in clusters. This pattern can be seen in cross-sections taken from mature animals. This unique fibre arrangement is the final result of the whole process of development, termed *myogenesis*. As the number of SO fibres is relevant to meat quality, there are important commercial implications. The improvement of meat quality therefore relies on the understanding of muscle development and in particular of cluster formation.

2. Spatial Tesselations

Voronoi Tessellations are being used to model characteristics of individual fibres and models are being developed to understand: (1) the distribution of the cluster positions, (2) interactions between clusters and (3) the number and arrangement of fibres within clusters. The aim of this project is to develop stochastic

models of the processes creating observed fibre distributions, and to enable interpretation of variations in these patterns in terms of variations in the model parameters.

3. The Model

A two stage Simple Sequential Inhibition Processes (SSI) is used to simulate the clustered patterns as follows:

Let $SS_1=\{a_1, \ldots, a_n\}$ be the set of points obtained in the first stage ($a_i = (x_i,y_i)$) and $\vartheta_1=\{V(a_1),\ldots, V(a_n)\}$ the Voronoi diagram generated by SSI_1. This "Voronoi grid" is the framework for the remainder of the simulation. In the second stage $SSI_2=\{b_1, \ldots, b_n\}$ is a SSI process generated from SSI_1 ($SS_2 \subset SS_1$) and it allocates cluster seeds. Each cluster is formed in another two "stages": i) the number of fibres per cluster is chosen randomly (U(a,b)) ii) the nearest neighbour of an existing cell is chosen to be a member of the cluster until the number in i) is achieved.

4. Results

Fibre clusters appear not to be distributed at random. Moreover, it was found that there was a minimum interpoint distance between cluster centers, which indicates some cluster repulsion. Muscle fibre samples show a generally good agreement with the suggested model.

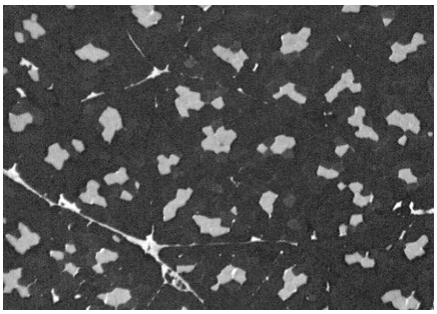


fig.1: muscle fibre sample

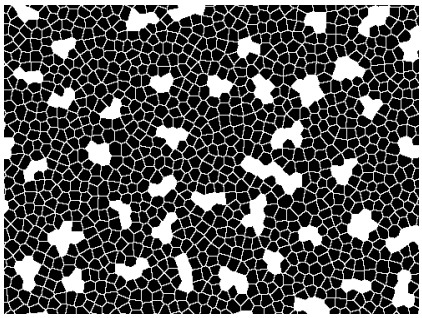


fig2: simulated pattern of clusters

References

Gatrell, A.C., Bailey, T.C., Diggle, P.J. and Rowlingson, B.S. (1996) Spatial point pattern analysis and its applications in geographical epidemology. *Trans Inst Geogr NS* **21** 256-274, Royal Geographical Society

van Lieshout, M.N.M. (2000) Markov Point Processes and their Applications. Imperial College Press

Okabe, A., Boots, B., Sugihara, K. and Chiu, S.N. (2000) Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, John Wiley & Sons, Inc.

Ripley, B.D. (1981) Spatial Statistics. John Wiley & Sons, Inc.

Schabenberger, O (1999,2000) Spatial Point Patterns (Chapter 7). STAT 5544 Spatial Statistics – Summer 2000

Stoyan, D., Kendall, W. S. and Mecke, J. (1996) Stochastic Geometry and Its Applications (2nd Edition). John Wiley & Son Ltd

Heavy Tails – How to Weigh Them?

Maria Isabel Fraga Alves*
CEAUL, DEIO, Faculty of Science– Lisbon University
isabel.alves@fc.ul.pt

1. Introduction

A non-negative random variable or its distribution is frequently designed as a *risk*. At a first approach any distribution in the positive axis can be used as the associated model of severity for individual claims. However, we instinctively distinguish between "well behaved" and "dangerous" distributions – ***heavy tails***. For these, the large claims have great influence for the total claim amount. In fact, when in practice actuaries try to estimate the mean (or variance!) of the claim amount, they use resampling techniques in order to obtain an estimate; however, sometimes the successive values do not stabilize for any limit value. One possible reason is that some moments of the underlying distribution do not exist, as is the case of heavy tail distributions.

One parameter that is used to have some insight of the weight of the tail is the well-known ***tail index***. Hill estimator (Hill, 1975) has been largely used in the extreme value theory in order to estimate the tail index associated to a distribution function with a positive index. One possible criticism to its use is the possible associated bias, much depending on the top portion of the original sample used, and also the fact that it is not location invariant which produces a big bias too.

Location invariance is specially important for insurance data. Here a new Hill-type estimator is studied, based on the original Hill estimator, but made location invariant by a random shift.

2. The Fire Insurance Data

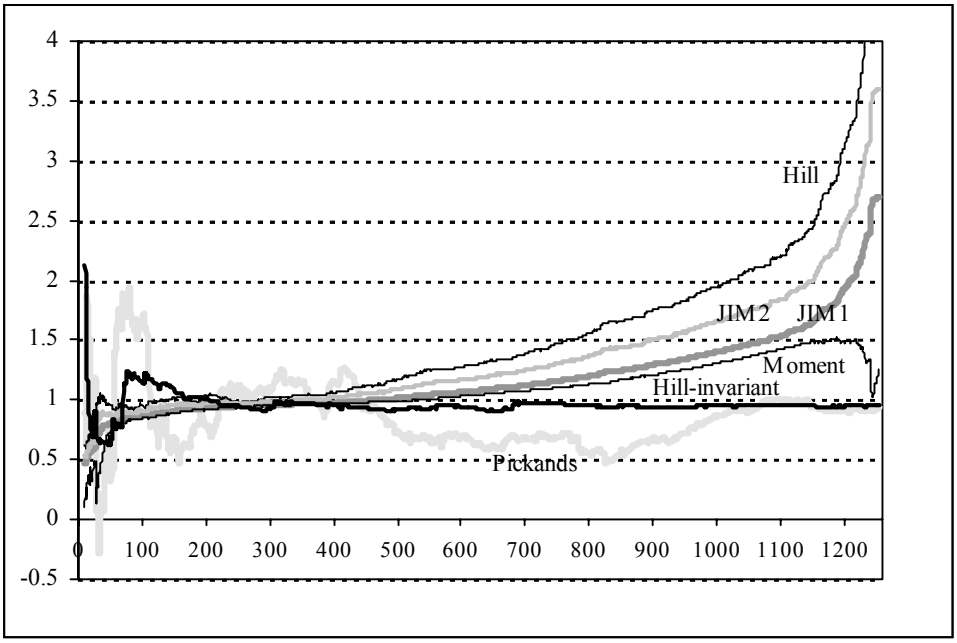


Figure 1. Sample paths of tail index for the 1256 portuguese data fire insurance

* This project was partially supported by FCT/ POCTI/FEDER.

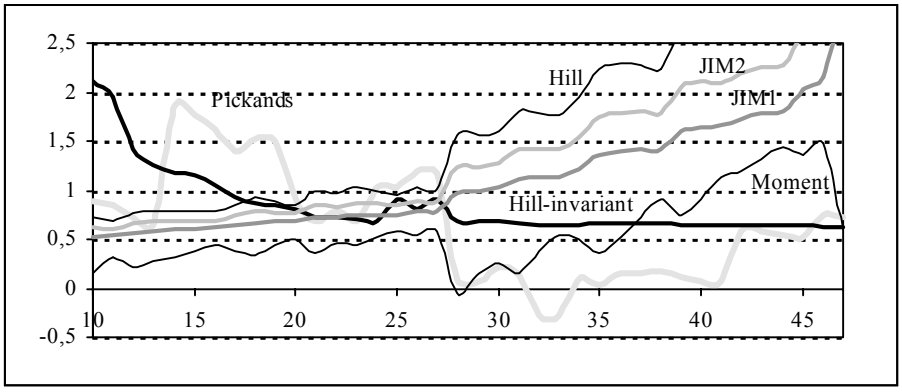


Figure 2. Sample paths of tail index only for the top 47 portuguese data fire insurance (excesses over 5 000 000 PTE)

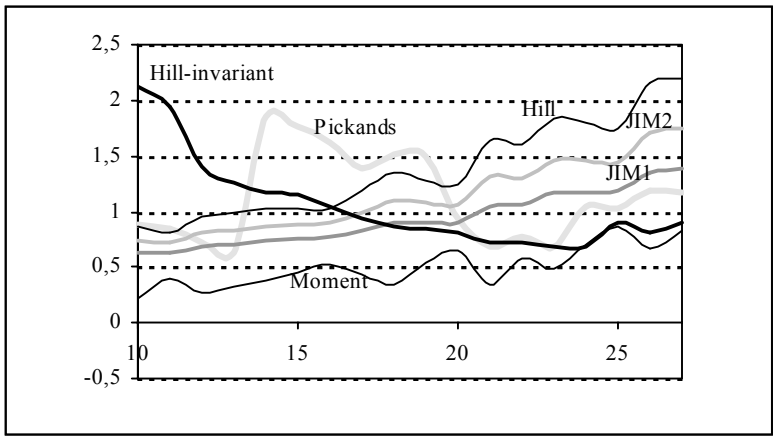


Figure 3. Sample paths of tail index only for the top 27 portuguese data fire insurance (excesses over 10 000 000 PTE)

In Figures 1-3 it is evident the different pattern behaviour for the tail index, much depending on the high level considered for the retained excess values, if the estimators are not location invariant.

The estimators here plotted are the new proposed Hill-invariant (Fraga Alves, 2000), the traditional Hill (Hill, 1975), the Moment (Dekkers et al., 1989), the Pickands (Pickands,1975), which is also location invariant, JIM1 and JIM2 denoting some recent estimators considered in Gomes et al. (2001).

Notice the flat pattern of the new estimator, adapted in such way that beforehand it is possible to estimate the weight tail index, independent of the high level considered, once its distribution is independent of whatever location considered.

References

Dekkers, A.L.M., Einmahl, J.H.J. and De Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. *Ann. Statist.* 17, 1833-1855.

Fraga Alves, M. I.(2000). A Location Invariant Hill-type estimator. *Technical Report 3/2000 (CEAUL)- FCUL*.

Gomes, M.I., Martins,M.J. (2001). Generalizations of the Hill estimator - asymptotic versus finite sample behaviour. *J. Statist. Plan. Inf.* 93, 161-180.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution*, Ann. Statist*. 3, 1163-1174.

Pickands, J. (1975). Statistical inference using extreme value orderstatistics. *Ann. Statist* 3, 119-131.

Estimation of the Parameter Controlling the Speed of Convergence in Extreme Value Theory

M.I. Fraga Alves*, M.I. Gomes*
Lisbon University, CEAUL, DEIO, Faculty of Science
isabel.alves@fc.ul.pt, ivette.gomes@fc.ul.pt

L.de Haan*, T.Lin
Erasmus University of Rotterdam
ldehaan@few.eur.nl

Let $X_1, X_2, \ldots$ be i.i.d. random variables with distribution function F. The distribution is in the domain of attraction of an extreme value distribution if and only if $U := \left(\frac{1}{1-F}\right)^{\leftarrow}$ satisfies an extended regular variation property:

$$\lim_{t\to\infty} \frac{U(tx)-U(t)}{a(t)} = \frac{x^{\gamma}-1}{\gamma},$$

for all $x>0$ with γ a real-valued parameter and a a suitable positive function. The limit function should be interpreted as $\log x$ for $\gamma = 0$. The speed of convergence of the partial maxima towards the limit distribution and also the asymptotic normality of estimators of the parameter γ (and other quantities) are controlled by a second order extended regular variation property:

$$\lim_{t\to\infty} \frac{\frac{U(tx)-U(t)}{a(t)} - \frac{x^{\gamma}-1}{\gamma}}{A(t)} = \frac{1}{\rho}\left(\frac{x^{\gamma+\rho}-1}{\gamma+\rho} - \frac{x^{\gamma}-1}{\gamma}\right),$$

for all $x>0$ with ρ the non-positive second order parameter and A a suitable positive or negative function. As before the limit function is defined by continuity for $\gamma = 0$ and/or $\rho = 0$. The function $|A|$, which is regularly varying of order ρ and tends to zero, represents the speed of convergence. Hence large values of $|\rho|$ corresponds to rapid convergence, whereas for example $\rho = 0$ means vary slow convergence (logarithmic or even worse).

Estimators of e.g. γ are generally functions of a number, say k, of top order statistics. In order to get asymptotic normality (or even consistency) for those estimators one needs as $k(n)/n \to 0$, as the sample size n tends to infinity, but this still leaves much freedom. Adaptive optimal choices for k are known. For those choices one needs to estimate ρ, the second order parameter. Up to now only estimators are known with rather bad behaviour.

We present an estimator for ρ, under the assumption that $\rho < 0$, which converges at a polynomial rate. The idea behind the estimator (or in fact a class of estimators) is as follows.

* This project was partially supported by FCT/ POCTI / FEDER.

Most estimators of γ have the following type of expansion:

$$\hat{\gamma} \equiv \hat{\gamma}(k) = \gamma + \frac{P_n}{\sqrt{k}} + B(\gamma, \rho) A(n/k) + o_p\left(A(n/k)\right)$$

($n \to \infty$) where P_n is asymptotically normal, the letter B indicates bias and the function A is as before.

Now take three estimators $\hat{\gamma}_1$, $\hat{\gamma}_2$, $\hat{\gamma}_3$, depending on the same number k of top order statistics. Then

$$\hat{\gamma}_1 - \hat{\gamma}_2 = \frac{P_n^{(1)} - P_n^{(2)}}{\sqrt{k}} + \left(B_{(\gamma,\rho)}^{(1)} - B_{(\gamma,\rho)}^{(2)}\right) A(n/k) + o_p(A(n/k))$$

Hence, if k is chosen such that $\sqrt{k} A(n/k) \to \infty$ (which one can achieve in practice), then

$$\frac{\hat{\gamma}_1 - \hat{\gamma}_2}{A(n/k)} \xrightarrow{p} B_{(\gamma,\rho)}^{(1)} - B_{(\gamma,\rho)}^{(2)}$$

and in fact

$$\frac{\hat{\gamma}_1 - \hat{\gamma}_2}{\hat{\gamma}_2 - \hat{\gamma}_3} \xrightarrow{p} \frac{B_{(\gamma,\rho)}^{(1)} - B_{(\gamma,\rho)}^{(2)}}{B_{(\gamma,\rho)}^{(2)} - B_{(\gamma,\rho)}^{(3)}}$$

($n \to \infty$), thus identifying ρ asymptotically.

We have carried out this program and have obtained asymptotic normality for the resulting estimator at a polynomial rate.

For this we need third order extended regular variation which has been developed for the purpose. Simulations will be presented.

Aging Classes Based on the ICV and ICX Orders

Manuel Franco, José-María Ruiz
Universidad de Murcia, Departamento de Estadística e I.O.
Campus de Espinardo 30100, Murcia, Spain
mfranco@um.es; jmruizgo@um.es

M. Carmen Ruiz
Univ. Politécnica de Cartagena, Dpto. de Mat. Apl. y Estadística
Paseo Alfonso XIII, 52. 30203, Cartagena (Murcia), Spain
maricarmen.ruiz@upct.es

1. Introduction and Main Results

It is well known that the concept of positive aging describes the adverse effects of age on the lifetimes of units and it has been found very useful to classify life distributions by stochastic orderings, as one can observe in the recent literature. For definitions of several classes of life distributions, e.g. *IFR*, *DMRL*, *NBU*, *NBUE* and their duals, see Bryson and Siddiqui (1969) and Barlow and Proschan (1975). Further, there are many results in the literature about the preservation of these classes under the formation of different types of coherent systems such as k-out-of-n, series and parallel systems; see Esary, Marshall and Proschan (1970), Sabnis and Nair (1997), Abouammoh and El-Neweihi (1986) and Sengupta and Nanda (1999) among others.

Recall that a k-out-of-n system functions if and only if at least k of its n components work. In particular, a series system is an n-out-of-n system and a parallel system is a 1-out-of-n system.

Deshpande, Kochar and Singh (1986) introduce some new aging criteria based on the stochastic dominance of first and higher orders. In fact, second order stochastic dominance (1) and (2) defined in their paper coincides with the increasing concave ($\leq_{ICV}$) and increasing convex ($\leq_{ICX}$) orderings. Using the increasing concave ordering, Deshpande, Kochar and Singh (1986) propose the following aging classes.

Definition *Let X be the lifetime of a unit with distribution function F. It is said that:*

(i) F is an increasing (decreasing) failure rate of second order distribution, $X \in IFR(2)$ *(DFR(2)), if* $X_{t_1} \geq_{ICV} (\leq_{ICV}) X_{t_2}$ *for all* $0 \leq t_1 \leq t_2$.

(ii) F is a new better (worse) than used of second order distribution, $X \in NBU(2)$ *(NWU(2), if* $X \geq_{ICV} (\leq_{ICV}) X_t$ *for all* $t \geq 0$,

where $X_t = (X - t / X > t)$ *is the residual life of the unit of age t.*

Analogously, Cao and Wang (1991) introduce a new class of life distributions: *F* is *new better than used in the convex ordering*, $X \in NBUC$ *(NWUC), if* $X \geq_{ICX} (\leq_{ICX}) X_t$ *for all* $t \geq 0$.

The closure of the *NBUC* class under the formation of parallel systems has been proved by Hendy, Mashhour and Montasser (1993) in the case of independent identically distributed (i.i.d.) components, and recently by Li, Li and Jing (2000) and Pellerey and Petakos (2000) for independent components.

The objective of this work is to study the preservation of these aging classes under the formation of different types of coherent systems. In particular, we obtain the closure of the *IFR*(2) and *NBU*(2) classes and their dual classes under formation of

series systems with independent and not necessarily identically distributed components. Likewise, we get a shorter proof of the closure of the *NBUC* class under parallel systems mentioned above.

On the other hand, note that *DFR*(2) and *NWU*(2) classes are not preserved under the formation of parallel systems with i.i.d. components, and so does for *k*-out-of-*n* and coherent systems. For example, if two i.i.d. components have exponential distribution, then they are *DFR*(2) and *NWU*(2), but its parallel system does not belong to these classes.

Likewise, a parallel system of two independent and exponentially distributed components with different means is not *IFR*(2), but its components have this property.

Thus, *IFR*(2) class is not closed under the formation of parallel systems with independent components. However, we prove its closure under parallel systems with i.i.d. components.

The following relationships among classes may be obtained using the well known properties of stochastic dominances:

$$\begin{array}{ccccc} \mathbf{IFR} & \Rightarrow & \text{IFR(2)} & \Rightarrow & \text{DMRL} \\ \Downarrow & & \Downarrow & & \Downarrow \\ \text{NBU} & \Rightarrow & \text{NBU(2)} & \Rightarrow & \text{NBUE} \end{array}$$

and in this context, we give some examples which check that some of the above relations are only one-way implications.

References

Abouammoh, A. and El-Neweihi, E. (1986) Closure of the *NBUE* and *DMRL* classes under formation of parallel systems. *Statist. Probab. Letters* **4**, 223-225.

Barlow, R.E. and Proschan, F. (1975) *Statistical Theory of Reliability and Life Testing*, *Probability Models*. To Begin With, Silver Spring, MD, USA.

Bryson, M.C. and Siddiqui, M.M. (1969) Some criteria for ageing. *J. Amer. Statist. Assoc.* **64**, 1472-1483.

Cao, J. and Wang, Y. (1991) The *NBUC* and *NWUC* classes of life distributions. *J. Appl. Prob.*, **28**, 473-479.

Deshpande, J.V., Kochar, S.C. and Singh, H. (1986) Aspects of positive ageing. *J. Appl. Prob.*, **23**, 748-758.

Esary, J.D., Marshall, A.W. and Proschan, F. (1970) Some reliability applications of the hazard transform. *SIAM J. Appl. Math.* **18**, 849-860.

Hendi, M.I., Mashhour, A.F. and Montasser, M.A. (1993) Closure of the *NBUC* class under formation of parallel systems. *J. Appl. Prob.* **30**, 975-978.

Li, X. and Kochar, S.C. (2000) Some new results involving *NBU*(2) class of life distributions. To appear in *J. Appl. Prob.*

Li, X., Li, Z. and Jing, B. (2000) Some results about the *NBUC* class of life distributions. *Statist. Probab. Letters* **46**, 229-237.

Pellerey, F. and Petakos, K. (2000) On closure property of the *NBUC* class under formation of parallel systems. Technical Report, Dipartimento di Matematica, Statistica ed Informatica, Universit´ a di Bergamo.

Sabnis, S.V. and Nair, M.R. (1997) Coherent structures and unimodality. *J. Appl. Prob.* **34**, 812-817.

Sengupta, D. and Nanda, A.K. (1999) Log-concave and concave distributions in reliability. *Naval Research Logistics* **46**, 419-433.

Shaked, M. and Shanthikumar, J.G. (1994) *Stochastic orders and their applications*. Academic Press, New York.

Preservation of Some Stochastic Orders

Manuel Franco
Universidad de Murcia
Departamento de Estadística e I.O.
Campus de Espinardo 30100
Murcia, Spain
mfranco@um.es

José M. Ruiz
Universidad de Murcia
Departamento de Estadística e I.O.
Campus de Espinardo 30100
Murcia, Spain
jmruizgo@um.es

M. Carmen Ruiz
Univ. Politécnica de Cartagena
Dpto. de Mat. Apl. y Estadística.
Paseo Alfonso XIII, 52. 30203
Cartagena (Murcia), Spain
maricarmen.ruiz@upct.es

1. Introduction and Main Results

In the literature, several papers have been devoted to compare two coherent systems in the usual stochastic, hazard rate, reversed hazard rate and likelihood ratio orders, which play an important role in reliability theory. In particular, Boland et al. (1994) and (1998), Block et al. (1998), Khaledi and Kochar (1999) and (2000) and Belzunce et al. (2001), among others, have studied stochastic comparisons, most of them based on the preservation under the formation of k-out-of-n systems.

A of k-out-of-n system is a system with n components which works if and only if at least k of the components function. In particular, the parallel and series systems are 1-out-of-n and n-out-of-n systems, respectively. In this context, the survival function of a k-out-of-n system coincides with the (n-k+1)th order statistic of the component lifetimes, so the study of order statistics is also of interest to compare the aging of these systems.

Likewise, some shifted orders have been introduced and studied by Shanthikumar and Yao (1986), Brown and Shanthikumar (1998), Lillo et al. (2000) and (2001) and Hu and Zhu (2001), and such orders are useful tools for establishing interesting stochastic inequalities. In general, the shifted and proportional versions are stronger orderings and easy to verify in many situations, so they are helpful to check which components are more reliable, and consequently systems formed from them.

In this paper, we first study the preservation of the shifted and proportional versions of the well known hazard rate, reversed hazard rate and likelihood ratio orderings under the formation of coherent systems with different structure and independent and identically distributed components.

Then, we consider a set of independent but not necessarily identically distributed components, and we establish comparisons in the above orderings between two coherent systems with different structures formed from this set of components.

Finally, we show sufficient conditions to preserve the shifted and proportional versions under the formation of two coherent systems with different structures and formed from two sets of independent but not necessarily identically distributed components.

All the results mentioned above allow us to establish new comparisons between k-out-of-n systems with different number of components and supported failures, and to get several known results on closure of the IFR, DRF and ILR classes given by Esary and Proschan (1963), Nanda et al. (1998) and Franco et al. (2001), respectively.

Moreover, as consequences of our results, we obtain some recent results of Lillo et al. (2001) on comparisons in the shifted likelihood ratio ordering of order statistics formed from one and two different samples.

References

Belzunce, F., Franco, M., Ruiz, J.M. and Ruiz, M.C. (2001). On partial orderings between coherent systems with different structure. *Probab. Engrg. Inform. Sci.*, **15**, 273-293.

Block, H.W., Savits, T.H. and Singh, H. (1998). The reversed hazard rate function. *Probab. Engrg. Inform. Sci.* **12**, 69--90.

Boland, P.J., El-Neweihi, E. and Proschan, F. (1994). Applications of the hazard rate ordering in reliability and order statistics. *J. Appl. Probab.* **31**, 180-192.

Boland, P.J., Shaked, M. and Shanthikumar, J.G. (1998). Stochastic ordering of order statistics. In: Balakrishnan, N., Rao, C.R. (eds.), *Handbook of Statistics: Order Statistics and Their Applications*, vol. **16**. Elsevier, Amsterdam, pp. 89-103.

Brown, M. and Shanthikumar, J.G. (1998). Comparing the variability of random variables and point processes. *Probab. Engrg. Inform. Sci*. **12**, 425-444.

Esary, J.D. and Proschan, F. (1963). Relationship between system failure rate and component failure rates. *Technometrics* **5**, 183-189.

Franco, M., Ruiz, M.C. and Ruiz, J.M. (2001). A note on closure of the ILR and DLR classes under formation of coherent systems. Technical Report, Dpto. Estadística e I.O., Universidad de Murcia, Spain.

Hu, T. and Zhu, Z. (2001). An analytic proof of the preservation of the up shifted likelihood ratio order under convolutions. Technical Report, Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui, China.

Khaledi, B.E. and Kochar, S. (1999). Stochastic orderings between distributions and their sample spacings-II. *Statist. Probab. Lett*. **44**, 161-166.

Khaledi, B.E. and Kochar, S. (2000). On dispersive ordering between order statistics in one-sample and two-sample problems. *Statist. Probab. Lett*. **46**, 257-261.

Lillo, R., Nanda, A.K. and Shaked, M. (2000). Some shifted stochastic orders. In: Limnios, N., Nikulin, M. (eds.), *Recent Advances in Reliability Theory*. Birkhäuser, Boston, pp. 85-103.

Lillo, R., Nanda, A.K. and Shaked, M. (2001). Preservation of some likelihood ratio stochastic orders by order statistics. *Statist. Probab. Lett*. **51**, 111-119.

Nanda, A.K., Jain, K. and Singh, H. (1998). Preservation of some partial orderings under the formation of coherent systems. *Statist. Probab. Lett*. **39**, 123-131.

Shanthikumar, J.G. and Yao, D.D. (1986). The preservation of the likelihood ratio ordering under convolutions. *Stochastic Process. Appl*. **23**, 259-267.

On the trends of Gini Coefficient in the Greek Economic Environment during the years 1960 to 1996

Chris Frangos
Technological Educational Institution of Athens
Aghiou Spyridonos Str.,P.C. 122 10, Aigaleo, Athens, Greece
cfrangos@teiath.gr

In this paper we compute the values of Gini Coefficient(Gini(1912), Gini and Galvani(1929), Kendall and Stuart(1969)), based on actual data of the annual income declaration for all the Greek taxpayers as from 1960 to 1996.

We calculate by exponential smoothing the trend of Gini Coefficient and we show that it has an upward direction from 1980 to 1996.

Specifically, we show that the gap between rich and poor people had a constant width between the years 1980 to 1992, whereas the same gap has been widened between the years 1993 to 1996.

Treating the data as a Time Series, we compute the Residual Standard Error (RSE) based on the methods of Quennouille (1949) Jackknife and Bootstrap. For an extensive bibliography and detailed presentation of the Bootstrap and Jackknife resampling methods of nonparametric estimation,the interested reader can consult Efron(1979a, 1979b, 1981a, 1987), Frangos(1980a, 1980b, 1983, 1984, 1987, 1991, 1994) and Miller(1974).

Moreover, we forecast values of Gini coefficient for 1997, 1998,1999.

The importance of Gini Coefficient as a characteristic measure of the distribution of income is obvious for the Economy of a country. High values of Gini Coefficient, (values approaching 1), show an asymmetric distribution of income and they are a signal to the ministry of Economic Affairs of the government concerned that it must provide the lower income groups of taxpayers with more jobs, relaxation of tax regulations and more social benefits.

The policy, also, of the concerned government must be to increase the investment leading to the construction of factories and the improvement of the infrastructure of the country in order to create more jobs and to bring to all the population groups the benefits of economic development.

Recently, Greece has experienced a good degree of economic development. The interest rates have been lowered, many public projects (highways, etc) are under construction, inflation is 2.5% and the Economy is booming. On the other hand, we have a high rate of unemployment (12%) ,particularly between the young people and the women, and some population groups, like the farmers and the pensioners are experiencing “the big stick” of the measures of economic austerity.

The Greek government is aspiring to bring the country into the EUROZONE by 1-1-2002 and Greece is already a full member of the European Monetary Union.

In this climate of high expectations and sings of poverty, it is instructive and it could be beneficial to examine the values and the trend of Gini Coefficient for the Greek Economy during the years 1960 to 1996.

In section 2 ,Gini coefficient is defined.

In section 3, treating the values of Gini Coefficient for Greece as a Time Series,we find the trend and we forecast its value for 1997.

In sections 4 and 5 we use exponential smoothing and 5 year moving averages in order to forecast the values of Gini coefficient for 1997.

In section 6 we propose some nonparametric methods for finding the Residual Standard Error of Gini Coefficient, like the Bootstrap and the Jackknife method for nonparametric interval estimation.

A possible factor for the recent high values of Gini Coefficient is the huge problem of unemployment in Greece . A possible solution could be the adoption of a more definite social policy for the creation of more jobs through investments and business incentives with money coming from the Third Package of Economic Assistance which is going to be provided to Greece by the European Union in 2000.

References

Efron, B. (1981). Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and other methods. *Biometrika*, **68**, 589-599.

Efron, B. (1987). Better Bootstrap Confidence Intervals (with discussion).*Journal of the American Statist. Assoc.*, **82**, 171-200.

Frangos, C. C. (1980a) . Modified Jackknife methods in Statistics with particular reference to second-order effects. Ph.D. Dissertation, London University, London School of Economics.

Frangos, C. C.(1980b). Variance Estimation for the second-order Jackknife. *Biometrika*, **67**, 3, 715-718.

Frangos, C. C. and Knott, M. (1983).Variance Estimation for the Jackknife using von Mises Expansions. *Biometrika*, **70**, 2, 501-504.

Frangos, C. C. and Stone, M. (1984).On Jackknife, Cross-Validatory and Classical Methods of Estimating a proportion with batches of different sizes. *Biometrika* , **71**, 2, 361-366.

Frangos, C. C. and Schucany, W.R. (1990). Jackknife Estimation of the Bootstrap acceleration constant.*Computational Statistics and Data Analysis*, **9**, 271-281.

Frangos, C. C. and Swanepoel, C.J. (1994). Bootstrap confidence intervals for the slope parameter of a logistic Model. *Communications in Statistics-Simulation*, **23**, 4, 1115-1126.

Kendall, M. G. and Stuart, A. (1969). *The Advanced Theory of Statistics,* vol. 1, 3rd Edit., C. Griffin and Co., London.

McConnell, C. R. and Brue, S. L.(1993). Microeconomics: Principles, Problems and Policies, 12th Edit., McGraw-Hill, Inc., New York, p.380-385.

Online Monitoring of High-dimensional Physiological Time Series

Roland Fried, Ursula Gather, Vivian Lanius
University of Dortmund, Department of Statistics
Vogelpothsweg 87, 44221 Dortmund, Germany
{fried, gather, lanius}@statistik.uni-dortmund.de

Michael Imhoff
Community Hospital Dortmund, Surgical Department
Beurhausstr. 37, 44137 Dortmund, Germany
mike@imhoff.de

1. Introduction

In intensive care, intelligent alarm systems are needed which detect critical situations and intervention effects quickly and reliably. Clinical information systems acquire the vital signs of the critically ill patients online at least every minute. Statistical methods have already shown to be useful for online detection of patterns of change in univariate physiological time series (Gather, Fried and Imhoff, 2000).

For modelling multivariate time series of vital signs we have to estimate a huge number of parameters. Moreover, patterns in high dimensions are difficult to interpret. Physicians typically select the most important variables according to their experience and base their decisions on the patterns found in these variables.

Instead of selecting a subjective subset we apply statistical methods for dimension reduction to compress the data into a few relevant variables. Dynamic factor analysis (Peña and Box, 1987) allows to find latent variables which capture the major part of the variability in the data. We use graphical models as a preliminary step to impose a structure on the loading matrices since we want to ensure that the latent factors can be interpreted by the physician. In the following, we apply this approach to a 10-dimensional time series of vital signs with about 3800 observation times.

2. Dynamical Dimension Reduction

Graphical correlation models for multivariate time series allow to identify linear, possibly time-lagged partial associations between the variables (Dahlhaus, 2000). An analysis of the data considered in this case-study identifies four groups of variables which are strongly associated. These are the arterial pressures, the pulmonary artery pressures including the central venous pressure, the pulse and the heart rate, and the blood temperature, which does not have strong associations to any of the other variables.

For the reason of interpretability we neglect the weak partial correlations between these groups and search common factors for each group individually. Using the approach suggested by Peña and Box (1987) we calculate the eigenvalues and the eigenvectors of the sample autocovariance matrices. We find one factor to be sufficient for every group and the correlations between the factors to be weak.

3. Online Monitoring of the Factor Series

An experienced senior physician judged the factors as well as the original variables and classified the patterns found into interesting and clinically relevant. Almost every pattern detected in the original variables is also visible in the factors with just two exceptions.

Bauer, Gather and Imhoff (2000) suggest a procedure for online detection of (patchy) outliers and level changes, which is based on the Mahalanobis distance between a vector containing the most recent observations and the centre of a multivariate control ellipsoid. In Fried (2001) an additional tool for online detection of slow monotone trends is suggested, which applies a weighted sum of the observations with weights chosen according to a minimax-criterion (Abelson and Tukey, 1953). We apply a combined chart consisting of these two tools to the factors and compare the results to the classifications of the physician.

Almost every pattern judged to be clinically relevant by the physician is also detected by the combined procedure, and most of the interesting patterns are detected, too. The percentage of false alarms is found to be lower for the factors than for the individual variables. This might be due to some smoothing effects since combining closely related variables helps to reduce the noise.

4. Conclusion

Statistical methods for dimension reduction may be applied successfully to compress the information contained in a multivariate time series into a few important variables. Graphical models can be used to derive a partitioning of the variables into strongly associated groups. Possibly there is common movement within such a group, so that a factor model is useful to identify a few latent variables which actually drive the multivariate time series. In our case-study, the factors obtained capture almost every important pattern detected in any of the variables in the corresponding group. This 'coverage' of patterns is better than for any of the observed variables. Hence the factors found in the analysis can be considered to be a suitable lower dimensional summary of the multivariate time series.

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.

Bauer, M., Gather, U. and Imhoff, M. (1999). The Identification of Multiple Outliers in Online Monitoring Data. Technical Report 29/1999, SFB 475, University of Dortmund, 44221 Dortmund, Germany (submitted).

Dahlhaus, R. (2000). Graphical Interaction Models for Multivariate Time Series. *Metrika* **51**, 157-172.

Fried, R. (2001). Online Detection of a Monotone Trend in a Time Series. Preprint, Department of Statistics, University of Dortmund, Germany.

Gather, U., Fried, R. and Imhoff, M. (2000). Online classification of states in intensive care. In Data Analysis. Scientific Modeling and Practical Application (eds. W. Gaul, O. Opitz and M. Schader), 413-428. Springer, Berlin.

Peña, D. and Box, G.E.P. (1987). Identifying a Simplifying Structure in Time Series. *J. Am. Statist. Assoc*. **82**, 836-843.

Plague in Kazakhstan: a Bayesian Hierarchical Model for the Temporal Dynamics of a Vector-Transmitted Infectious Disease

Arnoldo Frigessi
Norwegian Computing Centre
P.O. Box 114 Blinder, Oslo, Norway
frigessi@nr.no

We propose a discrete-time Bayesian hierarchical model for the plague-rodent-flea ecological system which captures many of its essential features. The model accounts for the sampling variability arising from multiple independent sources of data. The prior for the unknown population counts incorporates specific biological hypotheses regarding the interacting dynamics of the two species, and the transmission of Plague.

The population dynamics of the rodents are characterised by a discrete time stochastic model, with density-dependent effects mediating the survival rate. Posterior estimates of the lag-coefficients suggest the presence of a specialised non-migratory predator, together with a *huddling effect* which causes an increase in winter survival with increased autumn abundance. We also deduce a relationship between the summer growth rate of the fleas and the number of fleas-per-rodent in the previous spring.

We propose an infection process which acknowledges the different roles played by the fleas and rodents, and reflects a belief that epizootics arise when the populations of the two species reach certain (relative) levels. Although the data do provide some support for this hypothesis, weak prediction of the plague-periods suggest that external forces may also be important in determining the spread of the disease.

This is joint work with E. Clare Marshall, Department of Epidemiology and Public Health, Imperial College of Science and Medicine, London; Nils-Christian Stenseth, Department of Biology, University of Oslo; Marit Holden, Norwegian Computing Centre, Oslo; Vladimir Ageyev and Nikolay Klassovskiy, Anti-Plague Research Institute, Almaty, Republic of Kazakhstan.

References

C. Marshall, A. Frigessi, N. C. Stenseth, M. Holden, V. S. Ageyev and N. L. Klassovskiy, Plague in Kazakhstan: a Bayesian model for the temporal dynamics of a vector-transmitted infectious disease, *Norwegian Research Center Report* n. **959**, July 2000.

* This research was partially supported by EU TMR network ERB-FMRX-CT96-0095 on Computational and statistical methods for the analysis of spatial data.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Statistical Surveillance by Hidden Markov Models or Likelihood Ratios

Marianne Frisén, Eva Andersson, David Bock
Department of Statistics, Göteborg University
Box 660, SE-40530 Göteborg, Sweden
Marianne.Frisen@Statistics.gu.se, Eva.Andersson@Statistics.gu.se,
David.Bock@Statistics.gu.se

Timely detection of a change in a process from one state to another is of importance in many different areas. The optimality of likelihood ratio based methods is discussed by e.g., Shiryaev (1963) and Frisén and de Maré (1991). Likelihood ratio based methods are evaluated with respect to several measures of performance, such as expected delay and predictive value Frisén (1992) in a frequentistic framework by Frisén and Wessman (1999).

When more than two states are of interest, or when it is relevant to follow the process when it wanders between the states, hidden Markov models are useful. Those models are usually analysed in a Bayesian setting. Surveillance methods based on the posterior probability are suggested by e.g., Smith and West (1983) and Hamilton (1989).

It will be demonstrated that the decision problem for many applications can be expressed using either of the two approaches described above. Requirements for identical results are determined. Differences in performance when the two approaches give different results are described. The different ways to control false alarms in the two subcultures are of special concern and the consequences are demonstrated.

The two approaches are used for detection of turning points in business cycles Andersson et al. (2001) and for natural family planning Andersson (2000). Maximum likelihood estimators under different order restrictions Frisén (1986) are used in the alarm statistics.

References

Andersson, E. (2000) Monitoring Cyclical Processes Using Non-parametric Statistical Surveillance. *IBC 2000* Berkeley, San Francisco, US.

Andersson, E., Bock, D. and Frisén, M. (2001) Likelihood based methods for turning point detection in business cycles., Department of Statistics, Göteborg University.

Frisén, M. (1986) Unimodal regression. *The Statistician,* **35,** 479-485.

Frisén, M. (1992) Evaluations of Methods for Statistical Surveillance. *Statistics in Medicine,* **11,** 1489-1502.

Frisén, M. and de Maré, J. (1991) Optimal Surveillance. *Biometrika,* **78,** 271-80.

Frisén, M. and Wessman, P. (1999) Evaluations of likelihood ratio methods for surveillance. Differences and robustness. *Communications in Statistics. Simulations and Computations,* **28,** 597-622.

Hamilton, J. D. (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica,* **57,** 357-384.

Shiryaev, A. N. (1963) On optimum methods in quickest detection problems. *Theory of Probability and its Applications.,* **8,** 22-46.

Smith, A. F. and West, M. (1983) Monitoring Renal Transplants: An Application of the Multiprocess Kalman Filter. *Biometrics,* **39,** 867-878.

Remarks on Fuzzy Hypotheses Testing

Joanna Gerstenkorn
□ód□ *University*
Fac. of Sociology and Economy
Poland
tocha@krysia.uni.lodz.pl

In the traditional statistics it is assumed that the decisions taken up are based on precise (crisp) data. If the data are imprecise or when a hypothesis, even if based on crisp data, provides parameters that are not exactly determined then such a situation leads the procedure to ideas of the fuzzy sets theory and to application of its concepts and methods.

Fuzzy random variables were first introduced by Kwakernaak (1978) and analysed by Puri and Ralescu (1986), by Kruse and Meyer (1987) in their text-book, by T. Gerstenkorn and E. Rakus (1990) and others. Testing of hypotheses with fuzzy data was analysed by Casals et al. (1986), Kruse and Meyer (1987), and Saade and Schwarzlander (1990). Decision making in statistics based on fuzzy information was examined by Tanaka, Okuda and Asai (1979), and by Buckley (1985). As it is impossible to omit the notion of probability in fuzzy statistics, we find many papers with quite different ideas proposed. A survey of these concepts is given in T. Gerstenkorn and J. Ma□ko (1996).

Not long ago, in 1996, B.F.Arnold became engaged in fuzzy hypotheses with crisp data. His idea was taken up and evolved by Taheri and Behboodian (1999) who have proposed a new definition of the fuzzy hypothesis and the probability of the 1st and 2nd type. On this ground they presented their version of the Neyman-Pearson Lemma.

In the paper presented some interesting examples are given to compare numerically and graphically the classical and fuzzy methods in the sens of Taheri and Behboodian. In that manner the advantage of the proposed fuzzy testing method has been enhanced.

References

Arnold, B.F. (1990) An approach to fuzzy hypothesis testing, *Metrika* **44**, 119-126.

Buckley, J.J. (1985) Fuzzy decision making with data: application to statistics, *Fuzzy Sets and Systems* **16**, 139-147.

Casals, M.R., Gil, M.A., Gil, P. (1986) On the use of Zadeh's probability definition for testing statistical hypotheses from fuzzy information, *Fuzzy Sets and Systems* **20**, 175-190.

Gerstenkorn, T., Rakus, E. (1990) On the utility of the notions of a fuzzy variable and a linguistic variable in natural sciences, *Biometrical letters* **27** (1,2), 3-12.

Gerstenkorn, T., Ma□ko, J. (1996) Fuzziness and randomness: various conceptions of probability, *Proc. del III Congreso Internacional de la Sociedad Internacional de Gestión y Economia Fuzzy (SIGEF), 10-13 Nov. 1996*, Buenos Aires, Argentina, Facultad de Ciencias Economicas, Universidad de Buenos Aires, Vol. III, paper 2.45.

Kruse, R., Meyer, K.D. (1987) Statistics with vague data, *Reidel Publ. Comp.*, Dordrecht, Netherlands.

Kwakernaak, H. (1978), Fuzzy random variables: definition and theorem, *Inform. Sci.* **15**, 1-29.

Puri, M.L., Ralescu, D.A. (1986) Fuzzy random variables, *J. Math. Anal. Appl.* **114**, 409-422.

Saade, J.J., Schwarzlander, H. (1990) Fuzzy hypothesis testing with hybrid data, *Fuzzy Sets and Systems* **35**, 197-212.

Taheri, S.M., Behboodian, J. (1999) Neyman-Pearson Lemma for fuzzy hypotheses testing, *Metrika* **49**, 3-17.

Tanaka, H., Okuda, T., Asai, K. (1979) Fuzzy information and decision in statistical model. In: *Advances in Fuzzy Set Theory and Applications*, North-Holland, Amsterdam, pp. 303-320.

Gini's Mean Difference in the Theory and Application to Inflated Distributions

Tadeusz Gerstenkorn
□ód□ *University, Fac. of Mathem*
Poland
aticstadger@math.uni.lodz.pl

Joanna Gerstenkorn
□ód□ *University,Fac. of Sociology and Economy*
Poland
tadger@math.uni.lodz.pl

In 1911 Prof. Corrado Gini published a very vast statistical study initiating consideration on the mean called later in the literature Gini's mean difference (m.d.).The period of World War I undoubtedly disturbed the extension of Gini's ideas. We have failed to ascertain whether someone was dealing with the m.d. in the twenties. It did not gain popularity among theoreticians of statistics. However it is worthy of notice since, unlike other quantities designed for measuring the dispersion of a random variable, the m.d. is independent of any central measure of localization, which can be seen from its definition

$$\Delta = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |x-y|\,dF(x)dF(y).$$

The analytic investigation of the discussed characteristic is made difficult because of the absolute value occuring in the formula. However, it facilitates the computations on numerical data, which also concerns, as is well known, the mean deviation (m.dv.). Hence, we sometimes encounter the investigations concerning the m.d. connected with the m.dv. This is the case in Ramasubban (1958). For the normal distribution, the exact standard error of the m.d. was given by Nair (1936). In 1952 Lomnicki obtained the very result by using a simpler method. In 1953, Kamat calculated the third moment of the m.d. Following Kamat, Ramasubban (1956) obtained an approximation of values for the fourth moment. Interesting properties of the m.d. we can find in Yule and Kendall (1953) and also in Kendall and Stuart (1963). An extensive bibliography of papers based on Gini's ideas is presented in Giorgi (1990). In the paper we show an application of the m.d. to inflated distribution (composing of any discrete distribution with the degenerate, i.e. one-point distribution) introduced by S.N. Singh (1963) and M.P. Singh (1965/66 and 1966).

Definition We say that a discrete random variable Y is subject to the *generalized inflated distribution* (i.e. the one with a deformation at any point i=l) if its probability function is expressed by the formula

$$P(Y=i)=\begin{cases}\beta+\alpha h(l) & \text{if } i=l\\ \alpha h(i) & \text{if } i=0,1,2,...,l-1,l+1,...\end{cases}$$

where $\alpha \in (0,1]$ and $\beta = 1\text{-}\alpha$ and h(i) is probability function of uninflated distribution.

Making use of a result of Ramasubban (1958) we show the following

Theorem Gini's m.d. for the generalized inflated distribution is expressed by the formula

$$\Delta = 2\alpha\beta\{2lF(l+1)-1]-m_1+2m_1(l+1)\}$$

$$+2\alpha^2[\sum_{j=1}^{j-1}\sum_{i=0}^{j-1}(j-i)h(i)h(j)-\sum_{j=1}^{l-1}\sum_{i=0}^{j-1}(j-i)h(i)h(j)],$$

where:

m_1 - the expected value of an uninflated distribution $h(i)$,

$m_1(l+1)$ - the right-hand incomplete moment (i.e.the one with the truncation of the value of the variable to $x=l$ inclusive) of the uninflated distribution,

$F(l+1)$ - the distribution function of the uninflated distribution at a point $x=l+1$.

In a few Corollaries we show an application of this formula to some discrete inflated distributions.

References

Gini, C. (1911) Variabilità e mutabilità - contributo allo studio delle distribuzioni e delle relazioni statistiche, *Studi Economico-Giuridici delle R. Universita di Cagliari*, vol. III, Parte II, p. 3-159.

Giorgi, G.M. (1990) Bibliographic portrait of the Gini concentration ratio, *Metron* 48, 183-221.

Kamat, A.R. (1953) The third moment of Gini's mean difference, *Biometrika* 40 (3-4) 451-452.

Kendall, M.G., Stuart, A. (1963) *The Advanced Theory of Statistics*, Vol. 1, Distribution Theory, II ed. *Charles Griffin & Comp., London*, Sec. 2.20-2.23.

Lomnicki, Z.A. (1952) The standard error of Gini's mean difference, *Ann. Math. Stat*. 23, 635-637.

Nair, U.S. (1936) Standard error of Gini's mean difference, *Biometrika* 28, 428.

Ramasubban, T.A. (1956) A □-approximation to Gini's mean difference, *J. Indian Soc. Agric. Statist*. 8, 116.

Ramasubban, T.A. (1958) The mean difference and the mean deviation of some discontinuous distributions, *Biometrika* 45 (3-4), 549-556.

Singh, S.N. (1963) A note on inflated Poisson distribution*, J. Ind. Stat. Assoc*. 1 (3), 140-144.

Singh, M.P. (1965/66) Inflated binomial distribution, *J. Sci. Res. Banares Hindu University* 16 (1), 87-90.

Singh, M.P. (1966) A note on generalized inflated binomial distribution, *Sankhyã, The Indian J. of Statistics* 28 (1), 99.

Yule, G.Udny, Kendall, M.G. (1953)

Nonparametric Density Estimation Using the Sinc Kernel: Finite Sample Analysis

Ingrid K. Glad
University of Oslo, Department of Mathematics
P.B. 1053 Blindern, 0316 Oslo, Norway
glad@math.uio.no

Nils L. Hjört
University of Oslo, Department of Mathematics
P.B. 1053 Blindern, 0316 Oslo, Norway
nils@math.uio.no

Nikolai G. Ushakov
Norwegian University of Science and Technology, Department of Mathematical Sciences
7491 Trondheim, Norway
ushakov@math.ntnu.no

Based on n independent samples from a density f, the aim is to estimate f nonparametrically by means of a kernel density estimation method. Usually the kernel function K is taken to be a probability density with at least a couple of finite moments; this ensures that the estimator itself becomes a density function, and methods based on Taylor expansions make it possible to analyse its behaviour to a satisfactory degree.

The present paper deals however with a non-standard choice for K, the so-called sinc kernel

$$K(u) = \sin(u) / \pi u.$$

It is symmetric with (Riemann-)integral 1, but K and hence the estimator of f take also negative values. We argue that this kernel is interesting and should be considered despite this problem. In fact, it is known that the sinc kernel estimator has good asymptotic properties in terms of mean squared and integrated mean squared errors (MSE and MISE), see Davis (1975, 1977). We go further and find the exact finite sample MISE of the sinc estimator, which turns out to be an appealingly simple expression compared to most other kernels. The exact MISE formula allows to derive expressions for the bandwidth minimising the MISE, which for example in the normal density case $f = N(\mu, \sigma^2)$ simply reads $h^* = \sigma (\log(n+1))^{-1/2}$.

Comparing the minimum finite sample MISE of the sinc kernel estimator with the corresponding minimum MISE for the normal kernel (Marron and Wand (1992) has expressions for this case), we find for example that the sinc kernel performs better than the normal one for $n \geq 42$, when the density f to be estimated is itself standard normal. When n grows large, the minimum MISE goes to zero like $(\log(n))^{1/2}/n$, faster than the usual $n^{-4/5}$ rate achieved by the traditional kernels.

Similarly we compare the sinc estimator with the normal kernel estimator for f belonging to the classes of normal mixtures and skewed, normal densities (see Azzalini (1985)). We also look to the sinc kernel estimator in higher dimensions.

The problem of negativity of the estimator can easily be solved by a simple modification described in Glad, Hjort and Ushakov (1999). Any density estimator which is not a density, that is, is negative in regions or does not integrate to one, can be corrected in a way that the outcome is a density and still guarantees that the MISE is not increased. Hence, a corrected sinc estimator maintains the often superior precision properties, and does not exhibit negativity problems.

References

Azzalini, A. (1985). A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics* **12**, 171-178.

Davis, K.B. (1975). Mean squared error properties of density estimates, *Ann. Statist* **3**, 1025-1030.

Davis, K.B. (1977). Mean integrated square error properties of density estimates, *Ann. Statist.* **5**, 530-535.

Glad, I.K., Hjort, N.L., Ushakov, N.G. (1999). Correction of density estimators which are not densities, *Statistical Research Report*, Department of Mathematics, University of Oslo, No.17.

Marron, J.S. and Wand, M.P. (1992). Exact mean integrated squared error, *Ann. Statist.* 20, 712-736.

On Asymptotically Minimax Estimation of Linear Functionals for Some Classes of Infinitely Differentiable Functions

A. Goldenshluger
University of Haifa, Department of Statistics
Haifa 31905, Israel
goldensh@rstat.haifa.ac.il

B. Levit
Queen's University, Department of Mathematics & Statistics
Kingston ON, K7L 3N6, Canada
blevit@mast.queensu.ca

We consider estimating the value of a linear functional of infinitely differentiable functions from noisy observations. Some classes of infinitely differentiable functions with additional restrictions on smoothness in the frequency domain are introduced. Such classes describe functions with a good localization in both the time and frequency domains. We develop asymptotically minimax estimators for such families and compute corresponding asymptotics of the minimax risk.

These results have two interesting features. First, the resulting optimal rate of convergence depends, in general, on the smoothness of the signal in both the time and frequency domains. Depending on how these two relate to each other, the optimal rate can go all the way up to the parametric rate. Second, the corresponding asymptotically minimax estimators are not coordinatewise. That is, they are not described by the classical Wiener filter.

REVISTA DE
ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Truncated Sequential Change-Point Detection Based on Renewal Counting Processes

Allan Gut
Uppsala University, Department of Mathematics
Box 480, SE-751 06 Uppsala, Sweden
allan.gut@math.uu.se

Josef Steinebach
University of Marburg, FB Mathematik & Informatik
Hans-Meerwein-Straße, D-35032 Marburg, Germany
jost@mathematik.uni-marburg.de

The typical approach in change-point theory is to perform the statistical analysis based on a sample of fixed size. Alternatively, one observes some random phenomenon sequentially and takes action as soon as one observes some statistically significant deviation from the "normal" behaviour.

Based on the, perhaps, more realistic situation that not every observation is actually observed, we consider the counting process related to the original process, and assume that this process is observed at equidistant time points, after which action is taken or not depending on the number of observations between those time points. In order for the procedure to stop also when everything is in order, we introduce a fixed time horizon n at which we stop declaring "no change" if the observed data did not suggest any action until then.

We propose some stopping rules and consider their asymptotics under the null hypothesis as well as under alternatives. We also discuss possible (sequential) estimators for the case when parameters in the model are unknown. The main basis for the proofs are strong invariance principles for renewal processes and extreme value asymptotics for Gaussian processes.

Reference

Gut, A. and Steinebach, J. (2000). Truncated sequential change-point detection based on renewal counting processes. Uppsala University, *U.U.D.M. Report* **2000:18**, 34 pp.

Generalized Jackknife Estimators Revisited*

M. Ivette Gomes
Faculdade de Ciências, Departamento de Estatística e Investigação Operacional
Bloco C2, Piso 2, Campo Grande, 1749-016 Lisboa Codex, Portugal
ivette.gomes@fc.ul.pt

M. João Martins, Manuela Neves
Departamento de Matemática, Instituto Superior de Agronomia
Tapada da Ajuda, 1349-017 Lisboa Codex, Portugal
mjmartins@isa.utl.pt, manela@isa.utl.pt

Gomes et al. (1998, 1999) have worked with estimators of the tail index $\gamma > 0$, of an underlying heavy tail model F, based on the Generalized Jackknife methodology (Gray and Schucany, 1972). Those estimators had the peculiarity of reducing the dominant component of the asymptotic bias, through a linear combination of an adequate pair of semi-parametric estimators of γ, being competitors to the well-known Hill estimator for the tail index γ (Hill, 1975), given by

(1.1) $$\gamma_n^{(1)}(k) := \frac{1}{k} \sum_{i=1}^{k} \ln\left(X_{n-i+1:n} / X_{n-k:n}\right),$$

where, as usual, $X_{i:n}$ denotes the *i*-th ascending order statistic (o.s.) associated to the sample $(X_1, X_2, ..., X_n)$. The discrepancy we have got there between the behaviour of one of the Generalized Jackknife estimators investigated, here denoted by $\hat{\gamma}_n^G(k) := 2\hat{\gamma}_n^{(1)}(k/2) - \hat{\gamma}_n^{(1)}(k)$, and the theoretical developments, lead us to turn back to the Generalized Jackknife random variables based on an affine combination of the Hill estimator at two different levels, and to consider the Generalized Jackknife estimators

(1.2) $$\hat{\gamma}_n^{G_j}(k) := \frac{\hat{\gamma}_n^{(1)}(k) - p_j(k/n)\,\hat{\gamma}_n^{(1)}(k/2)}{1 - p_j(k/n)}, \quad j = 1, 2,$$

with $p_1(t) = \ln(1-t)/\ln(1-t/2)$, and $p_2(t) = 2 + t$. The first estimator is specially devised for Fréchet parents, the second one has a much wider scope, but they are both going to be considered in a general semi-parametric set-up. They are both related to different approximations of the quotient of asymptotic bias of $\hat{\gamma}_n^{(1)}(k)$ and of $\hat{\gamma}_n^{(1)}(k/2)$, and they are, together with $\hat{\gamma}_n^G(k)$, asymptotically undistinguishable. Despite of that, they have quite distinct exact properties. While $\hat{\gamma}_n^G(k)$ has always some bias, $\hat{\gamma}_n^{G_1}(k)$ is almost unbiased for an underlying Fréchet parent, and has a *MSE* at the optimal level $k_0^{G_1} := \arg\min_k MSE[\hat{\gamma}_n^{G_1}(k)]$ much lower than that of $\hat{\gamma}_{n0}^G \equiv \hat{\gamma}_n^G(k_0^G)$, which is on its turn lower than that of $\hat{\gamma}_{n0}^{(1)} \equiv \hat{\gamma}_n^{(1)}(k_0^{(1)})$, the original Hill estimator at its optimal level.

We shall here present a robustness study of the estimators in (1.2), on the basis of a multi-sample simulation of size 5000×10, through the computation of the Relative Efficiencies $(\mathrm{Re}.)$ of the two estimators, $\mathrm{Re}_{G_j} = \sqrt{MSE_s(\hat{\gamma}_{n0}^{(1)}) / MSE_s(\hat{\gamma}_{n0}^{G_j})}$, j = 1, 2,

* Research partially supported by FCT / POCTI / FEDER.

where MSE_s denotes the simulated *MSE* of the estimator at its simulated optimal level, for the following set of models in Hall's class (Hall and Welsh, 1985): the *Fréchet* model, $F(x) = \exp(-x^{-1/\gamma})$, $x \geq 0$, with $\gamma = 1$, for which the second order parameter is $\rho = -1$, the *Burr* model, $F(x) = 1 - \left(1 + x^{-\rho/\gamma}\right)^{1/\rho}$, $x \geq 0,\ \gamma > 0,\ \rho < 0$, with $\gamma = 1$ and for $\rho = -.25,\ -.5,\ -1,\ -2$, the *Student-t* model with $\upsilon = 8, 4, 2, 1$ degrees of freedom, for which $\gamma = .125, .25, .5, 1$ and $\rho = -.25,\ -.5,\ -1,\ -2$, respectively, and, as a curiosity, for a model outside Hall's class, the *Out-Hall* model, with a quantile function $F^{\leftarrow}(1-t) = t^{-1}e^{-2t(\ln t - 1)}$, $0 < t \leq 1$, for which $\rho = -1$, $\gamma = 1$.

n	100	500	1000	5000	10000	20000
$\mathrm{Re}_{G_1}/\mathrm{Re}_{G_2} \mid Burr(-2)$	0.77/ 0.71	0.73/ 0.72	0.71/ 0.72	0.68/ 0.72	0.67/ 0.71	0.66/ 0.70
$\mathrm{Re}_{G_1}/\mathrm{Re}_{G_2} \mid Stu(-2)$	1.09/ 0.75	1.74/ 0.86	1.36/ 0.79	0.68/ 0.71	0.66/ 0.70	0.66/ 0.69
$\mathrm{Re}_{G_1}/\mathrm{Re}_{G_2} \mid Fréchet$	1.19/ 0.92	1.48/ 1.09	1.62/ 1.18	2.02/ 1.45	2.25/ 1.59	2.51/ 1.75
$\mathrm{Re}_{G_1}/\mathrm{Re}_{G_2} \mid Burr(-1)$	1.02/ 0.94	1.15/ 1.13	1.22/ 1.22	1.38/ 1.49	1.45/ 1.63	1.53/ 1.79
$\mathrm{Re}_{G_1}/\mathrm{Re}_{G_2} \mid Stu(-1)$	1.17/ 0.86	1.01/ 1.02	1.18/ 1.10	1.20/ 1.32	1.26/ 1.43	1.33/ 1.55
$\mathrm{Re}_{G_1}/\mathrm{Re}_{G_2} \mid Out-Hall$	1.05/ 1.04	1.10/ 1.09	1.12/ 1.11	1.16/ 1.14	1.17/ 1.15	1.18/ 1.17
$\mathrm{Re}_{G_1}/\mathrm{Re}_{G_2} \mid Burr(-.5)$	1.26/ 1.18	1.70/ 1.60	1.96/ 1.84	2.78/ 2.63	3.25/ 3.08	3.81/ 3.60
$\mathrm{Re}_{G_1}/\mathrm{Re}_{G_2} \mid Stu(-.5)$	1.34/ 0.97	1.88/ 1.30	1.48/ 1.43	2.07/ 2.00	2.41/ 2.34	2.82/ 2.73
$\mathrm{Re}_{G_1}/\mathrm{Re}_{G_2} \mid Burr(-.25)$	1.22/ 1.25	1.48/ 1.82	1.61/ 2.18	1.90/ 3.42	2.04/ 4.18	2.20/ 5.11
$\mathrm{Re}_{G_1}/\mathrm{Re}_{G_2} \mid Stu(-.25)$	1.41/ 0.97	1.56/ 1.25	1.66/ 1.46	2.20/ 2.21	2.67/ 2.68	3.25/ 3.26

Two general comments:

1. For small values of the second order parameter ρ, here illustrated with $\rho = -2$, the Generalized Jackknife estimators cannot overpass the performance of the Hill estimator and have, for most of the models, simulated relative efficiencies close to 70%. Anyway their sample paths are quite stable and close to the target value.
2. For values $\rho \geq -1$ the Generalized Jackknife estimators perform quite well, with stable sample paths and *MSE*'s often much smaller than the *MSE* of the Hill estimator at its optimal level, even when we work with models outside Hall's class.

References

Gomes, M.I., Martins, M.J. and M. Neves (1998). *Alternatives to a semi-parametric estimator of parameters of rare events — the Jackknife methodology*. Preprint CEAUL 18/98. Accepted at *Extremes*.

Gomes, M.I., and M.J. Martins (1999). *Asymptotic efficiency of generalized Jackknife semi-parametric estimators of a heavy tail*. Preprint CEAUL 16/99.

Gray, H.L. and W.R. Schucany (1972). *The Generalized Jackknife Statistic*. Marcel Dekker.

Hall, P. and A.H. Welsh (1985). Adaptive estimates of parameters of regular variation. *Ann. Statist.* **13**, 331-341.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distributions. *Ann. Statist*. **3**, 1163-1174.

A Censoring Estimator of a Positive Tail Index*

M. Ivette Gomes, Orlando Oliveira
Faculdade de Ciências, Departamento de Estatística e Investigação Operacional
Bloco C2, Piso 2, Campo Grande, 1749-016 Lisboa Codex, Portugal
ivette.gomes@fc.ul.pt, orlando.oliveira@fc.ul.pt

Under a heavy tail framework, i.e., whenever we assume that the tail of the model $F(.)$, underlying the data, is a regularly varying function with index $-1/\gamma$, $\gamma > 0$, the Pareto behaviour of the top scaled order statistics (o.s.), $X_{n-i+1:n} / X_{n-k:n}$, $1 \le i \le k$, leads us to a maximum likelihood estimator of γ given by

(1.1) $$\gamma \quad k := \frac{1}{k} \sum_{=1}^{k} \left[\quad X_{\ -+1:} - \quad X_{\ -k:} \ \right],$$

which was introduced by Hill (1975), and is a consistent estimator of γ whenever k is intermediate, i.e., $k = k_n \to \infty$, and $k = o(n)$, as $n \to \infty$ (Mason, 1982). As usual, $X_{i:n}$ denotes the i-th ascending o.s., $1 \le i \le n$, associated to the sample $(X_1, X_2, ..., X_n)$ of independent random variables with common distribution function (d.f.) $F(.)$.

If we instead consider the Fréchet behaviour of X/a, and estimate jointly γ and a through maximum likelihood, under a type II censoring scheme, where we have access to the top k+1 o.s., $\underline{X}_k = \left(X_{n-k:n} \le X_{n-k+1:n} \le ... \le X_{n:n}\right)$, we get an estimator $\hat{\gamma}$ which may be implicity written as

(1.2) $$\hat{\gamma} = \frac{k}{k+1} \gamma_n \ (k) - \frac{\frac{1}{n} \sum_{i=1}^{k} \left(X_{n-i+1:n} / X_{n-k:n}\right)^{-1/\hat{\gamma}} \quad \left(X_{n-i+1:n} / X_{n-k:n}\right)}{\frac{1}{n} \sum_{i=1}^{k} \left(X_{n-i+1:n} / X_{n-k:n}\right)^{-1/\hat{\gamma}} + 1 - \frac{k}{n}}.$$

In this paper, we have not worked with the estimator in (1.2), which is easy to get iteratively for one sample, but leads to time-consuming simulations. We have worked instead with an explicit estimator, denoted by $\gamma_n^C(k)$, given by the expression in the second member of (1.2), but with $\hat{\gamma}$ replaced by the Hill estimator $\gamma_n^H(k)$ .Also, since for intermediate sequences, the denominator of the last term in (1.2) may be written as 1 / (/), we suggest the explicit estimator

(1.3) – /

or alternatively,

(1.4) —— – / ,

* Research partially supported by FCT / POCTI / FEDER.

which also seem to be able to reduce the asymptotic bias of the Hill estimator for intermediate, but reasonably large, values of k.

We shall consider here the finite sample properties of the above mentioned estimators of the tail index, for the *Fréche*t model, , with . The simulation results were based on a multi-sample simulation of size in order to guarantee small standard errors for the simulated characteristics, the mean value , the *M*ean *S*quared *E*rror , the optimal sample fraction, , with , and the *R*elative *EFF*iciency , defined as = = , with , and where denotes the simulated of the estimator at its simulated optimal level. The simulator of for instance , denoted by , is the average of 10 independent replicates of . In the following Table we show some finite sample properties of (reproduction of results in Gomes and Oliveira (1999)), , and , , for a *Fréchet* model.

n											
100	1.108	0.952	1.058	1.040	0.045	0.013	0.022	0.019	1.855	1.440	1.544
500	1.063	0.980	1.032	1.027	0.014	0.003	0.006	0.005	2.114	1.555	1.607
1000	1.048	0.986	1.024	1.023	0.008	0.002	0.003	0.003	2.269	1.610	1.649
5000	1.030	0.995	1.013	1.012	0.003	0.000	0.001	0.001	2.726	1.755	1.776
10000	1.023	0.997	1.010	1.010	0.002	0.000	0.001	0.001	2.975	1.820	1.836
20000	1.018	0.998	1.008	1.008	0.001	0.000	0.000	0.000	3.313	1.903	1.916

A few general remarks:

1. The estimator reduces excessively the bias of the Hill estimator, for large values of k, but, at the optimal level, which is attained deep into the tail, provides high efficiencies relatively to the Hill estimator.
2. The simplified estimators and , although with a smaller relative efficiency than , at their optimal levels, provide an interesting bath-tube pattern for the , , flat for a wide range of k-values, making thus less relevant the choice of the treshold.
3. The mean squared error of any of the censoring estimators is smaller than that of the Hill estimator at its optimal level, for a wide region of k-values.

References

Gomes, M.I. and O. Oliveira (1998). The bootstrap methodology in Statistical Extremes-the choice of the optimal sample fraction. *Notas CEAUL* **15**/98.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distributions. *Ann. Statist*. **3**, 1163-1174.

Mason. D.M. (1982). Laws of large numbers for sums of extreme values. *Ann. Probab*. **10**, 754-774.

Moment Method Estimation of the Offspring Variance for a Controlled Galton-Watson Branching Process*

M. Gonzalez, M. Molina, and I. Del Puerto
University of Extremadura, Department of Mathematics
06071 Badajoz. Spain
mmolina@unex.es

1. Introduction

In this paper we consider the controlled Galton-Watson process (CGWP) introduced by Sevast'yanov and Zubkov (1974) and defined in the recursive form:

where the empty sum is considered to be 0 and X_{ni} are integer-valued i.i.d. random variables with mean and variance denoted by m and respectively. The variable X_{ni} is interpreted as the number of offsprings produced by the i-th individual in the n-th generation and Z_n represents the population size in the n-th generation. Each individual generates new individuals, independently of all others, with indentical probability distribution. The population size in the -th generation is controlled by the function with range and domain , assumed integer-valued for integer-valued arguments and verifying that . It can be verified that the process is a Markov chain with stationary transition probabilities. In this work, we derive some estimators for the offspring variance and investigate their asymptotic properties.

2. Estimation of the Offspring Variance

Suppose the sample for a CGWP is available. Then, taking into account that:

a.s.

and considering the moment method estimators for m studied in González et al. (2000), we propose providing that , the following estimators for :

* Research supported by the Plan Nacional de Investigación Científica, Desarrollo a Innovación Tecnológica, grant BFM2000-0356

3. Asymptotic Properties

In order to investigate asymptotic properties for the above estimators we consider a CGWP with and we assume the conditions given in Bagley(1986) or Molina et al.(1998) which guarantee the almost sure convergence of , suitably normed, to a non-negative, finite and nondegenerate in 0 random variable *W*.

Proposition 1 On it is verited, for every , that:

i) converges in probability to 0 as

ii) converges in probability to 0 as

where

Proposition 2 On and are weakly consistent estimators for

Proposition 3 It is verified for every real number that:

i) converges to as

ii) converges to as

where denotes the distribution function of the standard normal probability distribution.

References

Bagley, J.H. (1986). On the almost sure convergence of controlled branching processes. *Journal of Applied Probability*, **23**, 827-831.

Gonzalez, M. , Martinez, R. , Mota, M. and Del Puerto, 1. (2000). Nonparametric estimation for controlled Galton-Watson processes: method of moments. *Proceeding of the International Seminar on Nonparametric Inference*. Santiago de Compostela. Spain.

Sevast'yanov, B.A. and Zubkov, A.M. (1974). Controlled branching processes. *Theory of Probability and its Applications*, **19** , 14-24.

Maximum Posterior Density Estimators for Branching Processes with Immigration*

M. Gonzalez, M. Molina, and M. Mota
University of Extremadura, Department of Mathematics
06071 Badajoz. Spain.
mvelasco@unex.es

1. Introduction

The Galton-Watson branching process with immigration (GWBPI) is a wellknown modification of the standard Galton-Watson branching model in which the immigration of individuals from an outer source is allowed.

We consider a GWBPI with the reproduction and immigration distributions belonging to the power series family of distributions, i. e. a sequence
defined recursively by

(1)

(with the empty sum defined to be 0) where and are two independent sequences of i.i.d. non-negative integer valued random variables with non-degenerate probability distributions and , respectively, being , where is a function of k or constant, , with and ; and , where is a function of k or constant, , with and . We also assume that Z_o is a random variable with distribution law .

Intuitively, Z_n denotes the number of individuals (particles) and Y_n the number of immigrants, both in the nth generation. It is easy to verify that is a Markov chain with stationary transition probabilities.

Let and , respectively, the means and variances of the offspring and immigration distributions. It easily follows that ,
and .

There are some previous works on estimation for this branching model, either using classical methods (e.g. see Wei and Winnicki (1990) or considering sequential or bootstrap estimation (e.g. see Datta and Sriram (1995), Sririam *et al* (19919). The purpose of this paper is to study the estimation problem from a Bayesian outlook,

* Research supported by the Plan National de Investigation Gientifica, Desarrollo a Innovacíon Tecnológica, grant BFM2000-0356

using the zero-one loss function. Thus, the maximum posterior density (MPD) estimators for the main parameters are obtained.

2. Results

Suppose we observe , namely the number of individuals and immigrants per generation until the nth generation. If we consider the following conjugate class of prior distributions for :

(2)

where , we obtain the posterior, distribution:

with , , , , and (i.e. the -field generated by

Theorem 1 For a GWBPI (1) and considering the conjugate class (2), the MPD estimator for the (i,j)-cumulant of the reproduction and immigration joint distribution is

being such that .

In Particular, the MPD estimators for the mean of tire reproduction and immigration distributions are and , respectively.

References

Datta, S. and Sriram, T. (1995). A modified bootstrap for branching processes with immigration. *Stochastic Processes and their Applications*, **56**, 275-294.

Sriram, T., Basawa, I. and Huggins, R. (1991). Sequential estimation for branching processes with immigration. *Annals of Statistics*, **19**, 4, 2232-2243.

Wei, C. and Winnicki, J. (1990). Estimation of the means in the branching process with immigration. *Annals of Statistics*, **18**, 4, 1757-1773.

A First Approach to the Efficient Linear Estimation in the Elliptical Bivariate Pearson Type VII Distribution

R. Gutiérrez-Jáimez
Universidad de Granada. Dpto. de Estadística e I.O.
Avda. Fuentenueva s/n. 18071 Granada, Spain
rgjaimez@ugr.es

J. D. Jiménez-López
Universidad de Jaén. Dpto. de Estadística e I.O.
Paraje Las Lagunillas s/n. 23071 Jaén, Spain
jdomingo@ujaen.es

1. Introduction

For a long time, the main point of interest of the *Classic Multivariate Analysis* was the study of the multivariate normal distribution. Nevertheless, from the middle of the 20th century, it was necessary to extend the multivariate analysis to non-normal populations, since there were many practical situations that might be better modelled by other alternative probability laws.

An important class of these alternative distributions is the family of the multivariate elliptically contoured distributions. This is a rich family which contains some of the better known distributions, such as the normal law, the uniform one, and t-Student, Cauchy and Laplace distributions, among others. Moreover, the distributions of this family verify many of the most important and useful properties that are certain under gaussian hypotheses. So, for these distributions, the generalization of the normal distribution properties drove to a new theory called *Generalized Multivariate Analysis*.

The problem of estimating the parameters of the multivariate elliptically contoured distributions has been studied by several authors, for example, Fang and Zhang (1990), Gupta and Varga (1992), etc. These authors investigated several properties of the point estimators as unbiasedness, sufficiency, completeness and consistency, assuming that the random vectors were identically distributed, but with a particular matrix-variate joint distribution. They concluded that, under this assumption, many of the known results about the normal multivariate law (collected by Anderson (1971) or Muirhead (1982)) still held true for the elliptically contoured distributions.

Our aim in this paper is to study the efficiency property of a particular elliptically contoured distribution: the bivariate Pearson type VII. More specifically, we consider the class of unbiased linear estimators of the parameters, concluding that, although there are not efficient linear estimators, the sample mean and the sample covariance matrix are, in this class, those that minimize the determinant of the covariance matrix.

2. Efficiency Property of the Linear Estimators of the Parameters

We consider a sample of random vectors identically distributed as a bivariate Pearson type VII distribution and with a joint distribution that belongs to the family of matrix-variate Pearson type VII distributions. This choice is justified by the fact that the bivariate marginal distributions of this family are also Pearson type VII.

This work has been partially supported by the "Ministerio de Ciencia y Tecnología" under contract BFM2000-0602 .

Under this dependence hypothesis, we prove the Cramer-Rao regularity conditions, by using the results established by Magnus and Neudecker (1988), and we also obtain the Fisher information matrix. A very important and useful property of this matrix is the similarity to that of the normal law under independence hypothesis: the submatrices situated in the diagonal are the Fisher information matrices of the individual parameters and the rest are zero matrices.

In order to study the efficiency property, it is firstly necessary to find unbiased estimators of the parameters. From the analogy with the normal distribution, it seems reasonable to think that the sample mean and the sample covariance matrix, weighted by an unbiasedness constant, might be efficient estimators of μ and Σ, respectively. However, we show, on the one hand, that the sample mean is unbiased but the determinant of its covariance matrix does not reach the Cramer-Rao lower bound; and, on the other, we obtain an efficiency equation (as a function of the parameter q and the sample size, n) which allows us to get particular conditions about q and n, under which the sample covariance matrix, weighted by a constant, is an efficient estimator. Nevertheless, this observation is not useful in practice because the parameter q must be fixed as a function of the sample size, n, in the bivariate Pearson type VII distribution.

In order to look for efficient estimators of μ and less restrictive conditions on the efficiency property of the sample covariance matrix, we have focused our study on wider classes of unbiased estimators: the families of unbiased linear estimators of μ and Σ, respectively. Firstly, we prove that, in general, there are not efficient estimators of the individual parameters inside these families, applying the results about moments collected in Díaz and Gutiérrez (1996). Secondly, we establish some conditions about the weights of the linear forms, under which the minimum of the distance between the determinant of the covariance matrix and the Cramer-Rao lower bound is reached. And finally, we demonstrate that the sample mean and the sample covariance matrix, weighted by an unbiased constant, verify these conditions. So, we conclude that these estimators are the linear unbiased estimators of μ and Σ, respectively, which minimize the determinants of their covariance matrices. Finally, bearing in mind the structure of the Fisher information matrix, we also show that the best unbiased linear estimator of the joint parameter, in the sense of minimizing the determinant of its covariance matrix, is precisely the joint estimator.

References

Anderson, T. W. (1971). *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, New York.

Díaz, J. A. and Gutiérrez, R. (1996). *Cálculo Diferencial Matricial y Momentos de Matrices Aleatorias Elípticas*. Servicio de Reprografía. Facultad de Ciencias. Universidad de Granada, Spain.

Fang, K. T. and Zhang, Y. T. (1990). *Generalized Multivariate Analysis*. Springer-Verlag, New York.

Gupta, A. K. and Varga, T. (1992*). Elliptically Contoured Models in Statistics*. Kluver Academic Publishers, Dordrecht/ Boston/ London.

Magnus, J. R. and Neudecker, H. (1988*). Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, New York

Muirhead, R. J. (1982). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, New York.

A Rank Test Based on the Dyadic Expansion of a Number for Comparing Two Populations

David Gutiérrez-Rubio, Fernando López-Blázquez, Begoña Salamanca-Miño, Teresa Gómez-Gómez
Facultad de Matemáticas, Dpto. Estadística e Investigación Operativa
C/Tarfia s/n, Sevilla (Spain)
davidi@cica.es

1. Preliminaries

Let $X_1, X_2, ..., X_{n_1}$, $Y_1, Y_2, ..., Y_{n_2}$ be i.i.d. random variables with continuous distributions F and G respectively. Without any additional assumption assumption about F and G, we consider the contrast

(1)
$$\begin{cases} H_0 : \ G(x) = F(x) \\ H_1 : \ G(x) = F(x - \Delta) \end{cases}$$

There is an extensive literature about these contrasts using nonparametric techniques, in particular, the linear rank statistics

(2)
$$L = \sum_{=1}^{n_1} c \ a(R \ )$$

where c_j are fixed constants, a is real valued function, R_j are the ranks of X_j among X, Y. The null hypothesis is rejected for small or large values of L.The Wilcoxon's test is a particular case of (2) taking $c_j = 1, \ \ a(x) = x$. Various results concerning asymptotic laws, efficiency, etc. have been discovered for a broad subclass of these tests. Here we present a new type of rank statistics with an election of a that does not fit in the cases studied before.

2. Definition and Main Results.

Given $x \in (0,1)$, we denote β_j as the j-th digit in the dyadic expansion of x, i.e., x expressed in base 2 is of the form $0.\beta_1\beta_2...$

In this paper we present a rank statistic of the form (1) given by the expression

$$T = \sum_{j=1}^{n_1} 2^{-R_j}$$

We will call dyadic test to test based on the statistic T.Note that T takes values on the subset of $(0,1)$

$$A_{n_1,n_2} = \{x \in (0,1) : \ \sum_{j=1}^{n_1+n_2} \beta_j = n_1 \quad \wedge \quad \beta_j = 0 \quad \forall j > n_1 + n_2\}$$

The null distribution of this statistic, unlike Wilcoxon's Statistic, is easily computed using combinatorial reasonings.

$$F_T(\ ) = \frac{1+\sum_{j=1}^{n_1}\binom{n_1+n_2-k_j}{n_1-j+1}}{\binom{n_1+n_2}{n_1}}, \qquad = \sum_{j=1}^{n_1} 2^{-k_j}, \qquad 1 \le k_{\ 1} < k_2 < \ldots < k_{n_1} \le n_1.$$

The asymptotic distribution of T for a wide class of distributions F, G is given in the following theorem. Here BMF(p) denotes the Binomial Multifractal Distribution with parameter p, which is a continous singular distribution on (0,1).

Theorem Suppose that $\lim_{n\to\infty} n_1/n = p$ and $\lim_{n\to\infty} n_2/n = q$, and the limit $\lambda = \lim_{x\to G^{-1}(0)} \frac{F(x)}{G(x)}$ exists, finite or not. Then $T \xrightarrow{L} F\left(\frac{\lambda p}{q+\lambda p}\right)$.

From the previous theorem, it follows that the asymptotic law of T only depends on the left tails of F, G, which indicates a significative loss of information about the distributions. However, this is not always the case. As established in the following theorem, the dyadic test can have infinite efficiency with regard to Wilcoxon's test. Let F', F'' denote in this case the first and second derivatives of F to the right.

Theorem Let $e_{T,W}$ be the assymptotic relative efficiency of the dyadic test with regard to Wilcoxon's test in the contrast (1), and suppose that exists the right derivative of F at $\vartheta = F^{-1}(0)$. Then

1. If $\vartheta = -\infty$ then $e_{T,W} = \infty$.
2. If $\vartheta > -\infty$ and
   a. $F'(\vartheta) > 0$ then , .
   b. then . In this case, the efficiency depends on the level . .
   c. then .

References

R. H. Randles, D. A. Wolfe (1979). Introduction to the theory of Nonparametric Statistics. Wiley Series in Probability and Mathematical Statistics.

J. Hájek, Z. Sidák (1967) Theory of Rank Tests. Academic Press.

Estimation des Paramètres du Modèle de Cox Généralisé: Etude par Simulation

Hafdi Mohammed Ali, El Himdi Khalid
Faculté des Sciences, Universitée Mohammed V, Dép. de Mathématiques et d'Informatique
B.P. 1014, Rabat, Maroc
elhimdi@fsr.ac.ma

Mikhail Nikulin
Université Victor Segalen, Bordeaux, UFR MI2S, Université de Bordeaux-2, BP 69
33076 Bordeaux Cedex, France
nikou@mi2s.u-bordeaux2.fr

Avant d'appliquer un modèle, quel qu'il soit, a des données réelles, une phase de simulations est indispensable dans le but d'étudier la qualité des méthodes d'estimations utilisées et partant, les propriétés asymptotiques des estimateurs.

Dans cet article, nous présentons une étude par simulation de la vraisemblance partielle modifiée utilisée pour estimer les paramètres du modèle de Cox généralisée tel que défini dans Bagdonaviçisus and Nikulin (1999).

Soit $T_{x(.)}$ une variable aléatoire qui exprime l'instant de décès d'un individu soumis à la covariable $x(.)=(x_1(.),...,x_m(.))$; notons sa fonction de survie $S_{x(.)}(t) = P(T_{x(.)} >t)$ et soit

---

le taux de décès sous x(.). On note $\Lambda_{x(.)}(t)$ le taux de hasard cumulée sous x(.) :

$\Lambda_{x(.)}(t)$ est aussi appelée la ressource exponentielle utilisée jusqu'à l'instant t.

Les modèles de survie interviennent lorsque l'on veut exprimer la liaison entre le taux de décès et le stress auquel sont soumis les individus. A cet effet, différents modèles ont été proposées par plusieurs auteurs. Citons en particulier le modèle suivant de hasard proportionnel, (PH) introduit par Cox (1972), et qui a connu une grande célébrité et un grand champ d'application:

$$\alpha_{x(.)}(t) = r(x(t))\alpha_0(t),$$

où α_0 est le taux de hasard de base (taux de hasard associée à $x \equiv 0$) et r est une fonction positive. Une généralisation du modèle de Cox qui tient compte de la dépendance, à un instant t, entre le taux de décès et la ressource utilisée ($\Lambda_{x(.)}(t)$) est proposée dans Bagdonaviçisus and Nikulin (1999). Le modèle proposée est désignée par Generalized Proportional Hazard model (GPH). Il suppose que le taux de décès à un instant t ne dépend pas seulement des stress mais aussi de leur passée exprimée par la ressource utilisée jusqu'à cet instant, i.e.:

$$\alpha_{x(.)}(t) = r(x(t))q(\Lambda_{x(.)}(t))\alpha_0(t)$$

où r et q sont deux fonctions positives.

Différentes paramétrisations de r et q permettent de déduire différents modèles. Dans cet article on se restreint à la paramétrisation suivante:

et .

Le modèle de Cox linéaire généralisée GLPH qui en résulte est alors défini par:

Dans ce travail, nous donnons une formulation du problème d'estimation du modèle GLPH, puis nous présentons les étapes des simulations réalisées et puis nous commentons les résultats obtenus. Une application sur des données réelles est aussi proposée.

References

Bagdonaviçisus, V.and Nikulin, M. (1999). Generalized Proportional Hasards Model Based on Modified Partial Likelihood. *Life time Data Analysis*, **5**, 329-350.

Cox, D. R. (1972). Regression models and life tables. *J. R. Statist. Soc.*, B, **34**, 187-220.

Sensitivity Analysis in the Presence of Discontinuities

Ernst Hansen
University of Copenhagen, Department of Statistics
Denmark
erhansen@math.ku.dk

The problem of computing magnitudes of the form

where is a stochastic variable and is a parameter (which we will think of as onedimensional), is frequently attacked by a simulation study: A natural estimator is

where are independent variables with the same distribution as the original . Sensitivity analysis in this problem amounts to the calculation of . If differentiation and integration can be interchanged, so that

(1)

a natural estimator based on simulation is

(2)

However, (1) may fail blatantly, for instance in PERTs where is not even continuous. If (1) fails, better results may be obtained by an estimate based on finite differences,

(3)

We will analyze the bias and variance of (3) in the presence of discontinuities of . The approach is differential geometric, and the main results are based on the coarea formula.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Stochastic Ricker Models

Göran Högnäs
Åbo Akademi University, Department of Mathematics
FIN-20500 Åbo, Finland
ghognas@abo.fi

We discuss some stochastic versions of the classical deterministic Ricker model

of the time evolution of the density of a population. Here models the intrinsic growth rate and is an inhibitive environmental factor. The introduction of demographic stochasticity leads us to a size-dependent branching process whose quasi-stationary distribution (for some values of and small values of) tends to concentrate around the attracting period cycle of the deterministic system. When we allow the environment to vary, modelled by an i.i.d. sequence of parameters , the branching process may exhibit growth-catastrophe behaviour. The effects of introducing a further independent source of randomness in the growth rate are also analyzed.

The results are connected to modern chaos theory as our stochastic models are non-local perturbations of a deterministic dynamical system.

References

Fagerholm, H. (2000). Stabilitetsundersökning av en stokastisk populationsmodell (in Swedish) [Stability analysis of a stochastic population model]. Unpublished M.Sc. thesis, Department of Mathematics, Åbo Akademi University, Åbo.

Högnäs, G. (2000). On some one-dimensional stochastic population models. In Probability on Algebraic Structures (eds G. Budzban, Ph. Feinsilver and A. Mukherjea), 209 – 220. *Contemporary Mathematics* **261**, American Mathematical Society, Providence.

Adaptive Tests in Additive Regression

Wolfgang Härdle
Humboldt-Universität zu Berlin
Institut für Statistik und Ökonometrie, Spandauer Str.1, D - 10178 Berlin, Germany

Stefan Sperlich
Universidad Carlos III de Madrid, Departamento de Estadística y Econometría
c/ Madrid 126, E - 28903 Getafe-Madrid, Spain
stefan@est-econ.es

Vladimir Spokoiny
Weierstrass-Institut
Mohrenstr. 39, D – 10117 Berlin, Germany

In multivariate regression problems we study the structural relationship between the response variable Y and the d-dimensional vector of covariates X via the regression curve $F(x) = E(Y|X=x)$ with x being the realizations. Purely nonparametric models do not make any assumption about the form of $F(x)$. The problem is then to fit a d-dimensional surface to the observed data. A serious problem arising here is the amount of data in higher dimensions needed in order to have enough data in a local neighborhood of each point. A promising approach for dimension reduction to deal with this so-called curse of dimensionality is additive modeling, an in economics favorite structure anyway, see e.g. Deaton and Müllbauer (1980). Such a nonparametric additive regression model has the form $F(x) = f1(x1) + ... + fd(xd)$, where the fm, $m=1,...d$ are a set of unknown component functions. In the statistical literature the nonparametric additive regression has been promoted by Buja, Hastie and Tibshirani (1989), introducing the so called backfitting algorithm. An further advantage in additive models is that they allow component-wise inferences. Important problems of component analysis in economics are the question of significance as well as of linearity, since nonlinearities can raise serious problems, e.g. of identification in equation systems. In our article we focus on the general problem of testing for $f1$ the null hypothesis of being of polynomial form. Here certainly, the cases of being linear or constant, i.e. having no impact at all, are the most interesting one. For the ease of presentation we give details only for the case of linearity, the constant case is the more trivial one.

Ingster (1982, 1993) has shown that a test could be uniformly consistent against a smooth alternative only if this alternative deviates from the null with the distance of order n to the $-2s/(4s+1)$ with s being the degree of smoothness. The structure of the proposed rate-optimal tests essentially relies on the smoothness properties of the underlying function though such kind of prior information is typically lacking in practice. Spokoiny (1996) offered an adaptive data-driven testing procedure which does not require this knowledge and allow for a near optimal testing rate up to a $\log \log(n)$ factor. The latter can be viewed as the price for adaptation. However, his procedure is essentially a theoretical device rather than a practically applicable method since it is developed for the idealized signal + white noise model, simple null, known noise variance, no explicit determination of the critical value, etc.

Practically relevant procedures should address numerous issues arising in applications. This is where we account for in detail in this article. In the context of multidimensional additive modeling, an additional challenge comes from the fact that the considered component *f1* even being completely specified, does not specify the whole model since nothing is assumed about the other components which can be viewed as an infinite-dimensional nuisance parameter. This particularly creates a serious problem with evaluating the critical value of the proposed test statistic. Therefore, the task is to develop a procedure which, independent of the functional form of the nuisance components leads to the given type I error if *f1* is (e.g.) linear, and is sensitive against a smooth alternative with unknown degree of smoothness. In view of practical applications we proceed with a deterministic non-regular design allowing for discrete components and unknown noise variance.

We apply a Haar decomposition which is a particular and non-regular case of the wavelet transform. For the hypothesis testing the application of this basis leads not only to the desired optimal testing rate but also provides a test which is more stable w.r.t. possible design non regularity. This again is important for practical applications, and we finally reach a reduction of the computational burden. Our approach is based on the simultaneous approximation of all components by Haar sums: we first estimate the Haar coefficients for all components and then analyze the coefficients corresponding to the first one. After this, the test is made independent of the chosen resolution level (the smoothing parameter) what is meant when speaking of adaptive. Unfortunately, this makes the distribution of the resulting statistic infeasible to calculate. Therefore we have to introduce additionally a Monte Carlo- (or bootstrap-like) method, as is oftens recommended in small samples anyway.

The results demonstrate that each component of the model can indeed be tested with the rate corresponding to the case when all the remaining components are known. The proposed procedure is feasible in practice and computationally straightforward. We discuss modifications and extensions e.g. to test additivity or to detect local deviations from the null hypothesis like jumps. An intensive simulation study and a real data example about female labor supply demonstrate the good performance of the proposed test procedure.

References

Buja, A., Hastie, T.J., and Tibshirani, R.J., (1989). Linear smoothers and additive models, *Ann. Statist.* **17**, 453-555.

Deaton, A. and Müllbauer, J. (1980). Economics and Consumer Behavior. Cambridge University Press, New York.

Ingster, Yu.I., (1982). Minimax nonparametric detection of signals in white Gaussian noise, Problems Inform. *Transmission*, **18**, 130--140.

Ingster, Yu.I., (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives," I--III. *Math. Methods of Statist.*, **2**, 85--114; 3, 171--189; 4, 249--268.

Spokoiny, V., (1996). Adaptive hypothesis testing using wavelets, *Ann. Statist.*, **24**, 2477--2498.

Sichel's Compound Poisson

William V. Harper
Otterbein College, Mathematical Sciences Department
Towers Hall 136, 1 Otterbein College, Westerville, OH 43085 USA
WHarper@otterbein.edu

Isobel Clark
Geostokos Limited
Alloa Business Centre, Whins Road, Alloa FK103SA, Central Scotland
drisobelclark@yahoo.co.uk

The following is the general definition of a compound or mixed Poisson.

and is the mixing

(compounding, weighting) distribution.

Different mixing distributions have been studied for various applications. One of the more flexible choices is the one recommended by Sichel (1973) below. The mathematics appear a little complex; however, the end result is straightforward to apply.

where are the three parameters. is a modified Bessel function of the second kind of order . Using this mixing distribution , the resulting family of discrete distributions is given below. Parameter characterizes the tail length of with a short tail as and a long tail as .

This encompasses many discrete distributions as special cases (e.g., Poisson, negative binomial). In particular, Sichel recommends the use of -½ which results in the mixing distribution below with the corresponding which we call Sichel's compound Poisson.

may be reparameterized for easier estimation by setting giving below.

There are two parameters to be estimated from the observed data. It can be and or and . For the distribution becomes a Poisson distribution. The parameter characterizes the frequencies at the start of the distribution. For the distribution is a reverse J-shaped curve whereas for a unimodal distribution results. Using the first of the two methods in Sichel (1973), the observed proportion for the barren observations (i.e., r = 0), , and the sample average (includes barren observations), , are used to solve for estimates of and and from these may also be estimated.

This leads to . One can calculate any desired probabilities by substituting the above estimates for the corresponding parameter into the following recursive formula and given the estimates for and below.

In the full paper examples are given comparing the above to more common discrete statistical distributions. Sichel's Compound Poisson is found to fill a gap not covered by standard distributions.

Reference

Sichel, H.S., Statistical valuation of diamondiferous deposits, *Journal of the South African Institute of Mining and Metallurgy*, February, 1973, pp. 235.243.

Small Sample Properties of Robust Fixed-Width Confidence Intervals

Zden□k Hlávka
Institut für Statistik, Humboldt-Universität zu Berlin
Spandauer Str. 1, Berlin, Germany
hlavka@wiwi.hu-berlin.de

1. Introduction

The most popular sequential methods are based on the sample mean. This implies that these methods are nonrobust and that they can produce misleading results if the real distribution is not Gaussian.

The behaviour of the "ordinary" fixed-width confidence intervals can be improved if we consider procedures based on some robust estimators. This approach was developed by Jure□ková and Sen (1981) who proposed robust version of the fully sequential Chow-Robbins procedure. However, in some situation, it might be more feasible to use three-stage procedures. Further improvement of the three-stage sequential procedure based on sample mean can be achieved by considering bootstrap approximation of the critical points (Aerts and Gijbels, 1993). The robust three-stage procedure based on M-estimators and bootstrap critical points was proposed and its basic asymptotic properties were established in Hlávka (2000).

In this paper, we will propose and discuss some modifications which improve the small sample behaviour of the sequential procedure based on bootstrap critical points.

2. Robust Three-Stage Procedure

Consider iid random variables with a common distribution function Denote by , , and the M-estimate, an estimate of its asymptotic variance, and the bootstrap approximation of the critical point of its (studentized or standardized) asymptotic distribution based on the observations, respectively.

Let denote the integer part of . The robust three-stage procedure goes as follows. In the first stage we fix the parameter and we draw

observations, where is the desired width of

the resulting confidence interval, is quantile of the standard Normal distribution, and is minimal starting sample size. In the intermediate stage of the

sequential procedure we draw

observations, where controls the sample size. The final sample size is then

given by .

The interval is an approximate confidence interval for the unknown location parameter . Under general conditions, we show that the square root of the final sample size has asymptotically Normal distribution with variance depending on the parameter and on the score function generating the M-estimator. Our results correspond to the earlier results for the distribution of the final sample size for the Chow-Robbins procedure (Jure□ková 1978).

The drawback of the above procedure for small sample sizes is that the calculation of the final sample size is based on the critical points , i.e. only on the first observations even though the (unknown) critical points would be more appropriate in this situation. The difference between the bootstrap critical points and is negligible for large number of observations (i.e. for small). For smaller sample size, we propose some modifications of the original procedure.

3. Simulation Study

In the simulation study, we compare various types of approximations of the critical points: approximation by critical points of standard Normal distribution, standardized bootstrap critical points, and studentized bootstrap critical points. The bootstrap critical points can be "adjusted" to provide better results for smaller sample sizes. The proposed methods are compared from the points of view of the coverage probability, the mean, the median, and the standard deviation of the final sample size. The simulations show that for small samples, the best results are obtained by the procedure based on the "adjusted" studentized bootstrap critical points.

References

Aerts, M. and Gijbels, I. (1993). A three stage procedure based on bootstrap critical points, *Sequential Analysis*, **12(2)**, 93-113.

Hlávka, Z. (2000). Asymptotic properties of robust three-stage procedure based on bootstrap critical points, Discussion Paper 94/2000, SFB 373, Humboldt-Universität zu Berlin.

Jure□ková, J. (1978). Bounded-length confidence intervals for regression and location parameters, The Second Prague Symposium on Asymptotic Statistics, Charles University Prague.

Jure□ková, J. and Sen, P.K. (1981). Sequential procedures based on M-estimators with discontinuous score functions, *Journal of Statistical Planning and Inference*, **5**, 253-266.

Publication, Presentation, and Teaching Statistics Using MD*Book

Zden□k Hlávka
Institut für Statistik, Humboldt-Universität zu Berlin
Spandauer Str. 1, Berlin, Germany
hlavka@wiwi.hu-berlin.de

Sigbert Klinke
Institut für Statistik, Humboldt-Universität zu Berlin
Spandauer Str. 1, Berlin, Germany
sigbert@wiwi.hu-berlin.de

Rodrigo Witzel
Institut für Statistik, Humboldt-Universität zu Berlin
Spandauer Str. 1, Berlin, Germany
witzel@informatik.hu-berlin.de

1. Introduction

Any information can be displayed in many different ways. This is reflected in the file formats which are mostly used to move the information around internet. For printing, it is desirable to have the postscript file. For displaying the information interactively, we prefer html or pdf formats. MD*Book is a set of tools which allows to create easily various output formats from a single LaTEX source. It is also very simple to create links from your files to other places in internet and thus to make the information even more alive and accessible.

Moreover, the usual text can be integrated with programs in Java and the reader has the chance to run the presented method directly from the browser without the necessity to install the software used by the author of the paper or textbook.

2. Various Output Formats

The technical description of the MD*Book package can be found on *www.md-book.com*. By running MD*Book, you can obtain postscript, pdf, html, or java output formats. The translation is driven by the so-called .sk file. This file determines what will be written to a temporary LaTEX files on which MD*Book runs latex, pdflatex, or latex2html. The big advantage of MD*Book is that a definition of a certain LaTEX command can depend on the desired output format.

The .sk file can be created automatically from any given LaTEX file. Thus, you can take your old LaTEX files, transform it to the .sk file and you immediately obtain all output formats supported by MD*Book.

The layout of the resulting html, pdf or ps files can be easily modified and the result depends only on your fantasy.

3. Teaching Statistics

MD*Book has been designed for professional presentation of statistical methods. The description of a method can be directly linked to an executable and editable computer code.

MM*Stat, the interactive tool for teaching statistics and a set of class transparencies prepared with MD*Book can be found at *www.md-stat.com*. The lecture notes, also prepared with MD*Book, can be found on the www page *www.xplore-stat.de*. These books and transparencies are integrated with the Java version of the XploRe computing environment (XQC). However, any other Java based computing environment could be used.

MD*Book is currently running only on Linux systems. Anyway, this doesn't limit its use since the files can be transferred to and translated on our Linux machine via the html forms on the internet page *www.md-book.com*.

Stereology of Extremes; Shape Factor

Daniel Hlubinka*
Charles University of Prague, Department of Probability and Statistics
Sokolovská 83, 186 75 Praha 8, Czech Republic
daniel.hlubinka@mff.cuni.cz

1. Spheroidal Parameter of the Planar Section

We shall consider oblate spheroidal particles in our presentation. The particle has two equal major semiaxes and one minor semiaxis and hence the particle is fully characterized by the size and shape considering its orientation being isotropic uniform random.

The size of the particle is the length of its major semiaxis . Denoting the length of the particle's minor semiaxis, the shape factor is . It is clear, that both and are nonnegative random variables, and we shall denote their upper bounds (possibly infinite) by and respectively.

In what follows we consider random spheroidal particles in given volume. The random planar section of the volume is the only available observation, hence we will observe size and shape factor of the section of particle. The section form ellipses and the ellipses are again fully characterized by the size and the shape factor defined by analogy to the three dimensional particles. Cruz-Orive (1976) gives joint density of

(1)

where is a population mean size of particles (half of the mean caliper diameter), and is the joint density of .

2. Extremes of the Shape Factors

In the paper we study the extremes both of and of . We will prove the stability of the domain of attraction, and in a particular case we shall provide the normalizing constants. Therefore we will need one-dimensional distributions of and . The problem can be divided into three subcases:

Problem with known particle's size. We will study the conditional distribution of given , which is

(2) ,

where is the conditional density of given .

* This research was supported by a postdoc grant GA□R 201/99/D059 (supporting project KONTAKT ME 335), and by the project MSM 113200008

Problem with unknown particle's size. We will study the conditional distribution of given , which can be derived from (1) easily.

Marginal distribution of the shape factor. We will study the marginal distribution rather than the conditional. However it seems to us that the extremes of shape factor should be related to the size in practical applications. This part is then complementary only.

3. Domain of Attraction

Recall that a distribution function belongs to the domain of attraction of c.d.f. , if the normalized sample extremes converge in distribution to . There are three limiting distributions only:

Frechet

Weibull

Gumbel

The main results of the paper are:

Theorem 1 *Suppose that for* any *fixed size the density is in the domain of attraction of . Then the distribution is in the domain of attraction of , where for and for .*

Theorem 2 *Assume that the* density *is in the domain of attraction of uniformly in size . Then the distribution is in the domain of attraction of , where for and for .*

Theorem 3 *Assume that the* density *is in the domain of attraction of uniformly in size . Then the marginal distribution is in the domain of attraction of , where for and for .*

References

Cruz-Orive, L.-M. (1976). Particle size-shape distributions; The general spheroid problem, *Jour. of Microscopy* **107.3**, 235–253.

Drees, H., Reiss, R.-D. (1992). Tail behavior in Wicksell's corpuscle problem, *Probability Theory and Applications*, J. Galambos, J. Kátai, eds., 205–220. Kluwer. Dordrecht.

de Haan, L. (1975). On regular variation and its application to the weak convergence. *Math. Centre Tracts* **32**. Mathematisch Centrum, Amsterdam.

Takahashi, R. (1987). Normalizing constants of a distribution which belongs to the domain of attraction of the Gumbel distribution, *Stat. Prob. Letters* **5**, 197–200.

Takahashi, R., Sibuya, M. (1996). The maximum size of the planar sections of random spheres and its application to metalurgy, *Ann. Inst. Statist. Math.* **48.1**, 361–253.

Takahashi, R., Sibuya, M. (1998). Prediction of the maximum size in Wicksell§s corpuscle problem, *Ann. Inst. Statist. Math.* **50.2**, 361–377.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the *k* largest observations. *Jour. American Stat. Assoc.* **73.364**, 812–815.

Statistical Measures for Compatibility and Relevance of Difference between Study Results
- An Approach Based on Confidence Intervals and Fuzzy Sets -

Josef Högel, Martina Kron
University of Ulm, Department of Biometry and Medical Documentation
Schwabstr. 13, D-89075 Ulm
josef.hoegel@medizin.uni-ulm.de

1. Goal

The paper proposes two sets of statistical measures: one to assess the degree of compatibility between study results, the other to assess the extent to which an observed difference in the results of trials or groups of a single trial is relevant. A pre-requisite is that outcome is measured by parameters for which confidence intervals are available.

2. Result in one Trial Arm Regarded as a Fuzzy Set

The information obtained about a parameter under investigation (e.g. the mean reduction of systolic blood pressure after the intake of a beta inhibitor within the framework of a clinical trial) can be condensed as exemplified in the following situation.

Example A clinical trial demonstrates that among the participants systolic blood pressure could be lowered by an average of . As a point estimate of the unknown μ, we may attribute to a "relative degree of acceptance" equal to 1. The smaller the observed standard deviation *s* of the outcome is, and the larger the number of participants *n* in the trial, the more we can be sure that further good guesses for μ are close to . The concept of (1-α)%-confidence intervals (CI) with in their center supports this idea. Having to assign to a certain value t other than a degree of acceptance less than 1 but larger or equal to 0, the error probabilities α of two-sided (1-α)%-intervals with the respective value as lower ( ) or upper boundary ( ), would serve this purpose.

This procedure yields a positive "function of acceptance" , $t \in \mathbb{R}$, which could also be called a normal fuzzy set (Zadeh, 1965). In the example given, asymptotic (1-α)%-CIs are

(1) ,

being the -quantile of the standard normal distribution Φ. This leads to

(2)

3. Compatibility of Two Trial Arms

Let us assume now that the results of certain treatments (not necessarily different) have to be compared between two groups of patients. Assume that outcome is quantified by the same statistical measure and that CIs can be derived. Then for each of the two treatments *A* and *B*, corresponding fuzzy sets and can be determined.

To bring these functions in relation to each other, the following measures can be considered, if the integrals exist, denoting by *S* a common support:

(3)

indicates the degree to which the two results are compatible or not contradictory.

(4)

assesses the degree to which the precision of the estimates observed in two groups is similar. Of course, both measures (3) and (4) have values between 0 and 1; the larger they are the better. To prevent the assignment of a high compatibility by measure (3) in case of trials with very different precision in the treatment arms, (4) could be used as a penalty for (3). Multiplication of (3) and (4) yields

(5) ,

and (5) is always less or equal to (3).

4. Relevance of Difference Between Estimates

Let us assume that the smallest difference between the outcome of two therapies which is clinically relevant, is given by . We can assess the relevance of the magnitude the outcome is different between two groups of patients using

(6) ,

the starred quantities above being the point estimates of the outcome under therapies *A* and *B*.
(6) ranges from 0 to ∞,. Weighting (6) by a coefficient of uncertainty to prevent the assignment of large values to (6) as a consequence of imprecise studies yields

(7)

as a measure of relevance adjusted for precision ranging from 0 to ∞.

5. Conclusion

Two sets of measures have been introduced: one assessing the degree of compatibility between two estimates, the other one to quantify the degree of relevance of an observed difference. Both sets have been constructed in an analogous way. They could prove to be useful, e.g. in comparing the results of different clinical trials in meta-analyses. No statistical testing is required, and the measures can be derived from the "ingredients" of confidence intervals the former being published in most articles.

Reference

Zadeh, L.A. (1965). Fuzzy Sets. *Information and control* **8**, 338-353.

Regional Trends in Rural Sulfur Dioxide Concentrations Over the Eastern U.S.

David M. Holland
U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC 27711, U.S.A.
holland.david@epa.gov

Petrutza Caragea, Richard L. Smith
Department of Statistics, University of North Carolina Chapel Hill, NC 27599, U.S.A.
piac@email.unc.edu, rls@email.unc.edu

1. Introduction

The implementation of the Clean Air Act (CAA), from its passage in 1970 to the 1990 amendments, has always required an assessment of the effects of atmospherically-transported pollutants on the environment. The 1990 amendments included new requirements that appreciably reduced sulfur dioxide (SO_2) emissions in two phases occurring around 1995 and 2000. The estimation of emission-related trends in airborne concentrations has been the subject of many investigations since the implementation of national monitoring networks in the late 1970's. Most of these studies focused on developing models either for site-specific trends or models for trend in a summary statistic that represents a network-typical value. In recent years, the focus of environmental policy has shifted toward regional-scale strategies that require regional estimates of trend for both their development and subsequent evaluation.

In an effort to provide meaningful regional trend information, this paper describes a two-stage modeling approach to estimate trends in rural airborne concentrations of SO_2 for 1990-1998 that have been adjusted for the effects of meteorology and season. After the large decrease in large electric utility SO_2 emissions in 1994-1995, SO_2 emission levels increased through 1998. This analysis is intended to provide accurate and precise trend information for most of the 1990's with particular interest given to the recent period of emission increases. The first stage uses a linear additive model to estimate site-specific trend, and the second stage uses an extension of classical Kriging methodology to estimate regional trends and standard errors. Finally, Bayesian techniques are used to estimate standard errors to quantify the effect of ignoring the uncertainty of the spatial covariance parameters.

2. Data

This analysis is applied to airborne SO_2 concentration ($\mu g/m^3$) data measured at 32 rural long-term monitoring sites in the eastern U.S. that are part of the Clean Air Status and Trends Monitoring Network (CASTNet) (U.S. Environmental Protection Agency, 1998a). Continuous measurements of temperature (degrees Celsius), wind speed ($m\ s^{-1}$), and wind direction (degrees clockwise from north) were summarized hourly at each site, and weekly measurements of SO_2 concentrations were obtained from filter pack measurements. The east-west wind component (u) is calculated as –windspeed × sine(wind direction) and the north-south wind component (v) is calculated as –windspeed × cosine(wind direction). It was necessary to summarize hourly meteorological data on the same scale as the SO_2 measurements, i.e., weekly. These meteorological summaries were

calculated by averaging all hourly meteorological variables between 10 a.m. and 5 p.m. across the week to characterize conditions during periods of atmospheric mixing. All sites in this analysis were required to have 80 percent of all weeks with concurrent SO_2 and meteorological data. This analysis was applied to data observed between 1990 and June, 1999.

3. Regional Trend Estimation

The first stage uses a linear additive model to relate the logarithm of weekly SO_2 concentrations to prevailing meteorological conditions, season, and time. The model is of the form,

(1)

where $SO_{2(ijkl)}$ refers to the measured pollutant concentration in the i^{th} week of the j^{th} month of the k^{th} year at the l^{th} site, $\varepsilon \sim N(0,\sigma^2_l)$, 1 is an indicator function for the year and month variables. B is a cubic spline function, and R is a thin-plate spline function. The variable *year* is defined in years starting in 1990. For each site location l, the estimated effect due to time is assumed to have a normal distribution with parameters and . Then site-specific trend expressed as total percent change between 1990 and 1998 can be defined as . For this analysis, the variances were assumed to be equal. The *delta method* based on retaining the first term of a Taylor series expansion of is used to approximate the variance of trend at locations l and the covariance of trend for different locations l and l'. Estimates of site-specific trend for SO_2 were all negative and the majority of trends were in the –20% to –40% range. For 21 of the 32 sites, R^2 values exceeded 0.6, providing good *de facto* evidence that the site-specific trend model in (1) is accounting for the seasonal and meteorological influences affecting SO_2 concentrations.

The second stage of the analysis uses an extension of standard Kriging methodology to predict smoothed surfaces of trend based on the site-specific estimates obtained in the first stage (see Holland *et al*., 2000). For this, trends are assumed to vary over the eastern U.S. as a realization of a Gaussian random field and maximum likelihood is used to fit the model. Kriging estimators are used to calculate averages of trend (defined as regional trend) and standard errors for three geographic areas: Midwest, Mid-Atlantic, and the South. A Bayesian analysis with Markov Chain Monte Carlo methods was used to account for the extra variability induced when parameters of the spatial covariance function are estimated. In all regions, the Kriging and Bayesian regional trend estimates were quite similar: -41% in the Midwest, -33% in the Mid-Atlantic, and –30% in the South. The Kriging standard errors were approximately 0.3% in each region. After accounting for the uncertainty of the spatial covariance parameters, the standard errors increased to approximately 2% in each region. Future work will apply a fully Bayesian approach to model the errors associated with estimating trend in the first stage.

References

Holland, D. M., De Oliveira, V., Cox, L. H., Smith, R. L. (2000). Estimation of regional trends in sulfur dioxide over the eastern United States. *Environmetrics* **11**, 373-393.

U.S. Environmental Protection Agency (1998). Clean Air Act Status and Trends Network (CASTNet) Deposition Summary Report (1989-1995). EPA/600/R-98/027. Office of Research and Development: Research Triangle Park, NC, 27711, USA.

Constrained Empirical Orthogonal Function Analysis with Application to Global Sea Surface Temperature Records

Jian Huang, Finbarr O'Sullivan
University College Cork, Department of Statistics
Ireland
jian@stat.ucc.ie

Climatological variables such as sea level pressure, sea surface temperature and precipitation are affected by a large variety of physical processes. Records of climatological variables, sampled in space and time, have multi-temporal and multi-spatial scale variations (Chelton, 1994). Therefore in climatological data analysis it is of interest to extract dominant patterns of various spatial and temporal scale variations that are supposed to represent the dynamics of the field under study. A key statistical tool used in this context has been empirical orthogonal function (EOF) or the principal component analysis (e.g., Kutzbacu1967, Von Storch and Navarra 1995). In standard EOF analysis one commonly compute the spatial patterns as the eigenvectors of the estimated marginal spatial covariance (kernel). Then the associated temporal variation patterns are obtained by the least squares regression. In climatology studies it is common that the spatial dimensionality is large, hence the statistical reliability of the eigenvectors of the estimated marginal covariance cannot be guaranteed. Here we describe a constrained EOF methodology that focus on the estimation of temporally defined features whose spatial intensity varies throughout the domain. By incorporating seasonal constraint into the definition of temporal features the method can be applied to data with large temporal and spatial dimensionality and interpretation of the temporal features becomes easier.

Suppose $y(x,t),\ x=1,...,N,\ t=1,...T$ is a space-time data set. The standard EOF analysis constructs a unique expression of form .

Now we consider introducing some constraints on the temporal variation pattern

$$m=1,...M,\ s=1,...S \text{ and } t=m+(s-1)*M,$$

where b represents annual or sub-annual variation pattern and c is the inter-annual modulation of b. (x), and are sequentially computed by minimizing

,

where is residual from removing the previous patterns and $w(x,t)$ represent the relative accuracy of the measurement recorded at spatial location x and time t-a weight of zero is assigned to the missing measurement. The optimization is carried out using the E-M algorithm.

The constrained EOF analysis has been applied to global monthly sea surface temperature records over a 47 years period. The first EOF (Figure 1), explained 5.38% variability in the anomaly, seems to show structures associated with the familiar El-Nino southern oscillation (for example, see Diazand Kiladis, 1992). The second EOF (Figure 2), explained 2.41% variability, seems to show structures associated with El-Nino like event in tropical Atlantic.

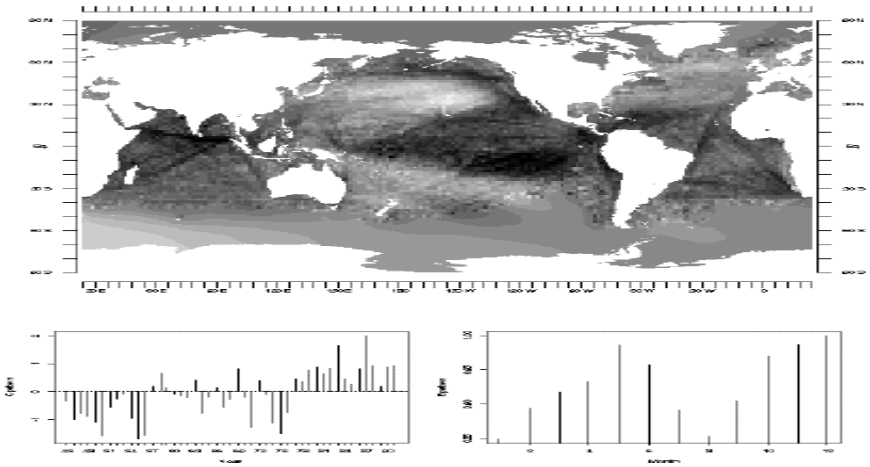


Figure 1. The First constrained EOF and the associated spatial pattern. Shades of gray represent the SST anomaly.

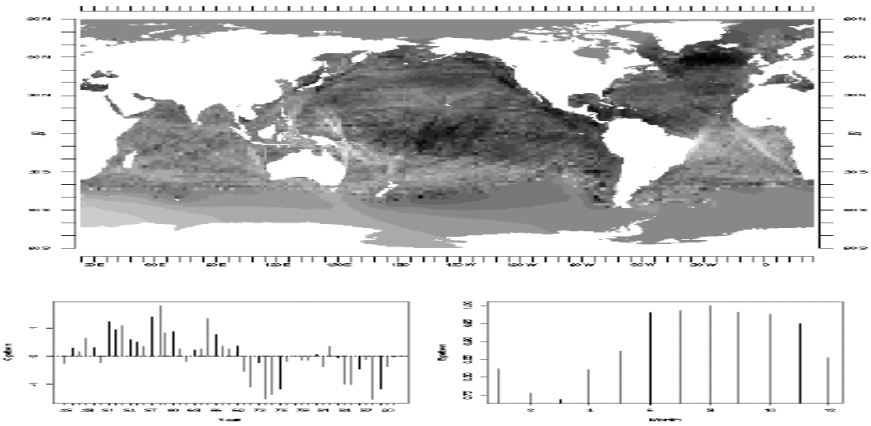


Figure 2. The First constrained EOF and the associated spatial pattern. Shades of gray represent the SST anomaly.

References

Chelton, D. B. (1994). Physical Oceanography: a Brief Overview for statisticians, *Statistical Science*, **9** 150-166.

Diaz, H. and Kiladis, G. (1992). Atmospheric Teleconnections Associated with the Extreme Phases of the Southern Oscillation. El Nino: Historical and Paleoclimatic Aspects of the Southern Oscillation (eds H. Diaz and V. Markgraf), Cambridge University Press.

Von Storch, H. and Navarra A. (1995). Analysis of Climate Variability: Application of Statistical Techniques (eds H. Von Storch and A. Navarra), Springer-Verlag, Berlin.

Ward, M. N. (1995). Analyzing the Boreal Summer Relationship between World-wide Sea-Surface Temperature and Atmospheric Variability. In Analysis of Climate Variability: Application of Statistical Techniques (eds H. Von Storch and A. Navarra), Springer-Verlag, Berlin.

Model Selection for Estimating the Nonzero Coefficients in a Gaussian Model

Sylvie Huet
INRA, Laboratoire de biométrie
78352 Jouy-en-Josas, France.
huet@banian.jouy.inra.fr

We propose a generalised Akaike criterion for Gaussian model selection when the variance is unknown. We consider the particular case of estimating the nonzero coefficients in a Gaussian vector when the number of these coefficients is unknown. We define the penalty term involved in our criterion in order to minimise the Kullback-Leibler risk of the resulting estimator.

1. The Estimating Problem

We consider the following model

where *m* is an unknown vector of and is a Gaussian vector with i.i.d. components . Some of the components of are nonzero, say . is unknown, and we aim at estimating *m* and .

We define a collection of subsets of *k* indices: where . By convention, . Let be some integer strictly smaller than *N*. We denote by J the collection of all subsets of with cardinality less or equal to .

For any vector *x* of , we set , and for all we denote by the vector of such that equals if and 0 if . is the complement of in . We say that *m belongs to J* if for all , , that is if .

If *m belongs to J* the maximum likelihood estimator of is . We propose to estimate *J* via a penalised maximum likelihood criterion.

2. The Modified Akaike Criterion

Let

(1)

be a penalised maximum likelihood criterion with penalty and let be the subset in that minimises , then the final estimator of is

.

The Akaike procedure consists in choosing . It is well known that for small samples or when the dimension of the parameter space is large, then the Akaike procedure leads to choose models of high dimension. Several authors have already proposed modified Akaike criteria, by considering a larger penalty term, see Hurvich and Tsai (1989), or McQuarrie and Tsai (1999). In the context of density estimation based on histograms, G. Castellan (1999) proved that the Akaike procedure does not work when the considered family of histograms is large.

Following the work of Birgé and Massart (2001), we calculate the penalty term such that the risk of the final estimator is minimised in some sense. Considering the Kullback-Leibler information as a loss function,

where is the likelihood function, we would like to estimate *J* by minimising the risk of , that is . Unfortunately, this is impossible because this function depends on . Nevertheless we show that we can build an estimator , such that

where is a set of probability close to 1, and some constant greater than 1.

The penalty term involved in (1) is written in the following way:

for any constant , and where the terms and are greater than 1. takes into account the complexity of the collection . This term is similar to the complexity term involved in the result of Birgé and Massart (2001). The term is the price to pay for estimating the unknown variance .

References

Birgé, L. and Massart, P. (2001) Gaussian model selection. To appear in J.E.M.S.

Castellan, G. (1999) Modified Akaike's criterion for histogram density estimation. *Technical report* **99**.61. Université de Paris-Sud.

McQuarrie, A.D. and Tsai, C-L. (1999) Regression and Time Series model Selection. World Scientific. Singapore.

Hurvich, C.M. and Tsai, C-L. (1989) Regression and time series model selection in small samples, *Biometrika* **76**, 297-307.

U. S. National & Regional Ozone Air Quality Trends, 1980-99

William F. Hunt, Jr.,
North Carolina State University, Visiting Senior Scientist
Raleigh, NC, USA
whunt@stat.ncsu.edu

How do we take millions of hourly ozone measurements and turn them into environmental information to help decision-makers properly direct and focus the National and State air pollution control programs in the United States? For the past 25 years, the United States Environmental Protection Agency (USEPA) has evaluated the trends and status of the Nation's air quality and has published the results in an annual report. Both ambient measurements, collected across the United States, and emission trends, based upon engineering estimates, will be presented focussing on the 20-year trend between 1980 and 1999 and the 10-year period between 1990 and 1999. Progress in measuring the effectiveness of the air pollution control program is based upon examining the trends in both ambient air quality measurements and emission inventory data.

1. Background

Ground level ozone has remained a pervasive pollution problem throughout the United States. Ozone is not emitted directly into the air but is formed by the reaction of volatile organic compounds (VOCs) and nitrogen oxides (NOx) in the presence of heat and sunlight. VOCs are emitted from motor vehicles, chemical plants, refineries, factories, consumer and commercial products, and other industrial sources. Nitrogen oxides are emitted from motor vehicles, power plants, and other sources of combustion. Ozone is effected by the weather – hotter summers produce more exceedances of the ozone standard. Ozone and the precursor pollutants that cause ozone can be transported into an area from pollution sources found hundreds of miles upwind. Short-term (1-3 hours) and prolonged (6-8 hours) exposures to ambient ozone have been linked to a number of health effects of concern. The daily maximum one-hour ozone standard requires that the expected number of days per calendar year with daily maximum hourly concentrations exceeding 0.12 parts per million (ppm) be less than or equal to one. The new daily maximum 8-hour average standard for ozone is defined as a 3-year average of the annual fourth highest daily maximum 8-hour average ozone values and must be less than or equal to 0.08 ppm. The health and welfare related NAAQS(s) apply to all of our 50 States and must be achieved. Each State is required to submit a State Implementation Plan (SIP), which specifies a plan as to how each State will reduce air pollution in order to achieve the NAAQS(s). The Federal government implements a series of programs to help achieve the NAAQS(s), including programs for mobile and stationary sources, etc. The question to ask is: "How well does the air pollution control program work, given all these control measures?"

2. Air Pollution Data

The ambient air quality concentrations are based upon actual measurements of pollutant concentrations. These measurements are made at monitoring sites across the United States. Emission estimates are calculated from the total tonnage of these

pollutants, or their precursors, released into the air annually. Air monitoring in the United States is largely conducted by state and local air pollution control agencies. In 1999, there were 705 ozone trend sites meeting 10-year trend criteria. Because only a few sites have monitored continuously for two decades (1980-1999), the ozone trend line is composed of two segments – 441 sites with complete data for the first ten years (1980-1989) and 705 sites meeting the criteria in the most recent 10 year period.

3. Major Findings

The National 20 year ozone trend, between 1980 and 1999, showed ambient ozone levels have decreased 20 percent based upon one-hour and 12 percent based upon the 8-hour data. Because only a few sites have monitored continuously for two decades, this trend line is composed of two segments – 441 sites with complete during the first ten years (1980-89) and 705 sites meeting the data completeness criteria in the most recent ten year period. Between 1980 and 1999, emissions of volatile organic compounds (VOCs) have decreased 33 percent. During that same period, emissions of nitrogen oxides (NOx) increased one percent. The gross domestic product increased 147 percent and the U. S. population increased 33 percent over the 1980-99 time period. Over this 20-year period, the ozone air quality improvements are a result of the effective implementation of the clean air laws and regulations, as well as improvements in the efficiency of industrial technologies.

Across the nation, however, there has been little progress over the past ten years – a 4 percent decrease in the 2nd highest daily maximum one-hour O3 value and "no change" in the 4th highest daily maximum 8-hour average. The emissions of VOC(s) decreased 15 percent, while the NOx emissions increased 2 percent. The U.S. population increased 10 percent and the gross domestic product increased 60 percent. The air quality improvement has not responded in the same way as the 20-year ozone trend.

While there has not been deterioration in the national ambient ozone trend between 1990 and 1999, there has been a lack of progress. For both the one- and 8-hour ozone measurements, ozone air quality trends in the Mid-Atlantic, Southeast, South Central, and Northwestern United States increased over the ten year period from 1989 to 1998. The highest increase of 17 percent occurred in the Southeast for the 4th highest daily maximum 8-hour measurements.

Because of this lack of progress over the past ten years, the USEPA proposed the NOx SIP Call Rule in September 1997, requiring a cap-and–trade program for large sources of NOx emissions in the eastern half of the United States. It would establish a NOx cap-and–trade program for sources in 22 eastern states. The NOx SIP Call, once implemented, should have a significant impact on reducing ozone levels and should reverse the "stalled" trend in ozone.

References

National Air Quality and Emissions Trends Report, 1998, EPA 454/R-00-003, U. S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, NC 27711, March 2000.

National Ambient Air Quality Standard for Ozone; Final Rule, 62FR 38856, Washington, DC, July 18, 1997.

Latest Findings on National Air Quality: 1999 Status and Trends. EPA-454/F-00-002, U. S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, NC 27711, August 2000.

Permutation Principle in Change Point Analysis

Marie Hušková
Charles University, Department of Statistics
Sokolovská 83, 186 00 Prague, Czech Republic
huskova@karlin.mff.cuni.cz

One of the simpliest models considered in the change point analysis is the model for abrupt change the mean in the local model. It can be described as follows:

where are independent observations, is the mean before the change and + (0) is the mean after the change, *m* is the change point, are iid random variables with zero mean and finite nonzero variance, denotes the indicator.

The primary interest is to test the null hypothesis that the observations are iid,, i.e. the mean does not change against the alternative that at some moment the mean changes. In case the null hypothesis is rejected to get an estimator of the location of the change point is of interest.

There are many variations of the above formulated problem. One can meet it practice, e.g. changes in hydrological and meteorological time series, structural changes in econometrics models, statistical quality control problems. There is a long list of books and papers devoted to this problem, e.g. Csorgo and Horvath (1997) that reflects development up to 1997.

Test procedures are usually constructed by maximum likelihood or Bayesian principle. The former one leads to the so called max-type statistics and the latter one to the weighted-type statistics. Having a test statistic one needs the corresponding critical value or at least their approximation. Usually the asymptotic distribution of the considered test statistic under the null hypothesis provides such approximation.

It is known that under the null hypothesis the limit distributions of the max-type test statistics belong to the extreme value type and the convergence to the limit distribution is very slow and therefore the approximation to the critical values through limit distribution is not satisfactory. Usually the resulting tests is are conservative.

In the talk an alternative approximation based on permutation principle will be proposed and discussed, some theoretical results will be presented for a number of models. Results of simulation study will be presented also.

Theoretical results say that the permutation principle provides asymptotically correct approximations for critical values. The proofs are based on asymptotic behavior various functionals of simple linear rank statistics.

The talk will be based on joint work with Antoch (2001), Slabý (2001) and Steinebach (2000).

The work was partially supported by grants GA□R 201/00/0769 and MSM 113200008.

References

Antoch, J. and Hušková, M. (2001). Permutation tests for change point analysis, to appear *Statistics and Probability Letters* .

Csorgo, M. and Horvath, L. (1993). *Limit Theorems in Change-Point Analysis*. J.Wiley, New York.

Good, P. (2000). *Permutation Tests*. Springer Verlag, New York.

Hušková, M. and Slabý A.(2001). Permutation tests for multiple changes, preprint.

Hušková, M. and Steinebach, J.A.(2000). Limit theoerems for a class of tests of gradual changes, *Journal of Statistical Planning and Inference* **89**, 57 – 77.

Romano, J.P.(1989) Bootstrap and randomization tests of some nonparametric hypotheses, *Annals of Statistics* **17**, 141-159.

Accelerating Diffusions

Chii-Ruey Hwang
Academia Sinica, Institute of Mathematics
Taipei, Taiwan
crhwang@sinica.edu.tw

Shu-Yin Hwang-Ma
Soochow University, Department of Business Mathematics
56, Sec. 1, Kwei-Yang St., Taipei, Taiwan
sym@bmath.scu.edu.tw

Shuenn-Jyi Sheu
Academia Sinica, Institute of Mathematics
Taipei, Taiwan
sheusj@math.sinica.edu.tw

An underlying distribution (x) in R is assumed to have a density proportional to with satisfying some regularity conditions. The following diffusion is used commonly to approximate (x).

(1)

To accelerate the convergence, an extra vector field is added to the gradient drift, where . Note that (x) is again the equilibrium distribution of the following family of diffusions.

(2)

Let be the corresponding infinitesimal generator and define

(3) = sup{ real part of : is in the spectrum of and is not zero}.

is used as a comparison criterion in our study. We proved that with a complete characterization of the equality. In other words the extra-added drift does help in improving the convergence rate.

References

Chen, M. F. (2000). *Ergodic Convergence Rates of Markov Processes: Eigenvalues Inequalities and Ergodic Theory*. Beijing.

Hislop, P. D. and Sigal, I. M. (1996). Introduction to Spectral Theory. *Appl. Math. Sci.* **113**, Springer.

Hwang, C. -R., Hwang-Ma, S. -Y. and Sheu, S. -J. (1993). Accelerating Gaussian diffusions, *Ann. Appl. Probab.*, **3**, 897-913.

Nagel, R. (1986). *One-parameter Semigroup of Positive Operators*. LN 1184, Springer.

Reed, M. and Simon, B. (1978). Methods of Modern Mathematical Physics 4. AP, NY

Stannat, W. (1999) (Nonsymmetric) Dirichlet operators in : existence, uniqueness and associated Markov process, *Ann. Scola Norm. Sup. Pisa Cl. Sci.*, **XXVIII**, 99-140.

Varadhan, S. R. S. (1980*). Lectures on Diffusion Problems and Partial Differential Equations*. Springer, NY.

Combinaison d'une Approche par Compétences et des Techniques de Scoring pour l'Élaboration d'un Plan de Formation

Berrada Ilham
Université Mohammed V-Souissi
ENSIAS-LMD, B.P. 713, Rabat, Maroc
iberrada@ensias.um5souissi.ac.ma

Les besoins en formation n'existent pas en soi. Ils constituent des écarts qu'il est nécessaire d'identifier et analyser par rapport à des référentiels existants tels que : un projet, les activités propres à un emploi, une avancée technologique, etc.

L'approche par compétences utilisée dans ce travail permet d'analyser, à l'aide d'outils statistiques, les besoins en formation du personnel d'une organisation afin d'établir leur portefeuille de compétences, en regard de la performance attendue et de planifier les séminaires permettant l'acquisition de ces compétences en tenant compte des priorités d'apprentissage accordées aux compétences par chacune des personnes.

Cette approche a été mise en œuvre selon les trois phases de diagnostic, de planification et de suivi qui seront succinctement présentés.

La première partie de ce travail présente un état de l'art dans le domaine de l'ingénierie de formation en informatique utilisant l'approche par compétence et les technologies de l'information, cf. Lesy-Leboyer (1993) et Flück (1992). Une présentation du référentiel des compétences reliées aux différents métiers de l'informatique au Maroc est aussi faite. Ce référentiel se trouve au cœur de l'approche par compétence aussi bien dans la première phase de diagnostic qui permet le recueil des objectifs organisationnels, que dans la phase d'auto positionnement des individus lors de la planification de leurs parcours de perfectionnement.

La deuxième partie de ce travail décrit le processus de réalisation qui offre des outils, des méthodes et des stratégies d'orientation requises lors d'une élaboration d'un plan de formation.

Les outils qui ont été élaborés permettent essentiellement :

1. d'automatiser les questionnaires conçus pour la collecte des données dans les phases de diagnostic et d'analyse des besoins;
2. de préparer les données à une analyse statistique avancée basée sur des techniques de Scoring et de segmentation, cf. Kass (1980) et Magidson (1993);
3. de produire l'arbre décisionnel des compétences qui segmente les compétences par rapport à trois critères à savoir, l'importance de la compétence pour le poste (mesuré sur une échelle de mesure allant de 1 à 5), le degré de maîtrise de la compétence (mesuré sur une échelle de 3 à 5) et la priorité d'apprentissage pour chacune des compétences (mesurée sur une échelle de 1 à 3). Ainsi, chaque nœud de l'arbre représente une population homogène par rapport aux compétences à acquérir selon un niveau de maîtrise et dans un ordre de priorité ;
4. de planifier la formation en tenant compte de l'arbre décisionnel obtenu et des éventuels conflits entre les modules de formations pouvant répondre à plusieurs compétences de niveaux différents.

L'ensemble de ces outil seront illustrés à l'aide d'une application réelle relative à l'élaboration du plan de formation des informaticiens du Ministère de l'économie et des Finances Marocain.

Références

Kass, G. (1980) An explanatory technique for investigating large quantities of categorical data. *Applied Statistics,* **29:2**, 119-127.

Magidson, J., and SPSS Inc. (1993) SPSS for Windows CHAID Release 6.0. Chicago: SPSS Inc.

Flück, C. and Le Brunchoquet, C. (1992) Développer les emplois et les compétences. Insep-Edition.

Lesy-Leboyer, C. (1993) Le bilan des compétences. Les éditions d'organisation.

A Comparison of Shewhart Control Charts Based on Normality, Nonparametrics, and Extreme-Value Theory

Roxana A. Ion
University of Amsterdam, Korteweg-de Vries Institute for Mathematics
Plantage Muidergracht 24, 1018 TV Amsterdam, Nederlands
roxana@science.uva.nl

Ronald J. M. M. Does
University of Amsterdam, Korteweg-de Vries Institute for Mathematics
Plantage Muidergracht 24, 1018 TV Amsterdam, Nederlands
rjmmdoes@science.uva.nl

Chris A. J. Klaassen
University of Amsterdam, Korteweg-de Vries Institute for Mathematics
Plantage Muidergracht 24, 1018 TV Amsterdam, Nederlands
chrisk@science.uva.nl

Since Shewhart in the early Twenties originated the concept of control chart, it has become a powerful tool in Statistical Process Control (Shewhart, 1931). In the present paper, several control charts for individual observations are compared. Traditional ones are the well-known moving range Shewhart control charts with control limits based on the average of the moving ranges of the individual measurements.

The availability of modern computing power in statistical process control enables one to consider nonparametric Shewhart control charts with control limits estimated via such diverse computationally intensive techniques from mathematical statistics as the bootstrap, empirical quantiles, kernel estimators, and extreme-value theory.

The alternative control charts we will compare, are four charts based on empirical quantiles, which are related to the bootstrap method , two control charts based on kernel estimators, and two based on extreme-value theory. For these nonparametric control charts the underlying distribution function, denoted by F, is assumed to be unimodal, but otherwise unknown. This means that we include distributions which have an increasing-decreasing density, such as the Normal, Logistic, Laplace, Cauchy, Student, Uniform and Exponential distribution.

The estimation of the control limits is based on the observations obtained in the so-called Phase 1, in which the data are collected from the production process and parameters are estimated (cf. Woodall and Montgomery (1999)). It is important to note that we consider the monitoring phase which is usually called Phase 2. Since the in-control parameters are unknown, we will study the statistical performance of the classical and newly proposed control charts by the average and standard deviation of the in-control run length.

The use of all control charts is demonstrated by a real-life example from a printers assembling company. Their performance is studied by Monte Carlo simulation.

It turns out that even under normality, the alternatives behave quite well especially when sufficiently many data are available. The performance of our Alternative Empirical Quantile control chart is excellent for all distributions considered.

References

Shewhart, W. A. (1931). Economic Control of Quality of Manufactured Product. Van Nostrand, Princeton, New York.

Woodall, W. H. and Montgomery, D. C. (1999). Research Issues and Ideas in Statistical Process Control, *Journal of Quality Technology* **31**, 376-386.