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

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

How do Bootstrap and Permutation Tests Work?

Arnold Janssen
University of Duesseldorf
janssena@uni-duesseldorf.de

Resampling methods are frequently used to adjust critical values of nonparametric tests. In practice these two-step testing procedures benefit from the strong computational efforts of the new computer generation. In this talk we will discuss permutation tests, the m(n) out of k(n) data points bootstrap tests, the weighted bootstrap and wild bootstrap tests.

A comprehensive and unified theory for the analysis of two-step linear resampling statistics is presented. Under fairly mild assumptions tightness of the underlying resampling statistics is proved. The weak accumulation points of their conditional weak limit distribution are derived. From this representation it becomes clear which part of the resampling statistic is responsible for asymptotic normality. Based on this discussion we find equivalent conditions for the asymptotic normality of the resampling statistics. In this case it is shown that resampling tests work which means that the resampling tests are asymptotically equivalent to unconditional tests. The results can also be applied to power functions of tests. In the normal case, we derive the asymptotic power under local alternatives and consistency under non contiguous sequences of alternatives. Special attention is devoted to two-sample testing problems. Here we show what conditional tests are really doing. In this case it turns out that permutation tests are better than bootstrap tests. In order to cover this sort of example we present results for arbitrary non-i.i.d. triangular arrays of random variables. Further examples are distributional convergent partial sums of infinitesimal rowwise independent random variables. In this case the limit variable can be compared with the unconditional limit law of the resampling variable. In this connection we prove the following result.

The unconditional limit variable coincides in distribution with the unconditional resampling variable iff the limit variable is normal or when it is symmetric and the resampling scheme is asymptotically given by random signs as resampling scheme.

The results about permutation tests continue earlier work of Janssen (1997). The research is joint work with T. Pauls from Duesseldorf University.

We are grateful for a DFG-grant.

Reference

Janssen, A. (1997). Studentized permutation tests for non i.i.d. hypotheses and the generalized Behrens-Fisher problem. *Stat. Probab. Letters* **85**, 255-277.

Score Functions, their Role and Applications

Jana Jureèková
Charles University, Department of Statistics
Sokolovská 83, CZ-186 75 Prague 8, Czech Republic
Jurecko@karlin.mff.cuni.cz

Consider a random sample from a distribution with density $f(x,\grave{e})$, differentiable with respect to the components of $\grave{e}$. Let $L(x,\grave{e})$ be its likelihood function and $l(x,\grave{e})$ the vector of its partial derivatives with respect to the components of $\grave{e}$, i.e. the (Fisher) score function corresponding to $f(x,\grave{e})$.

This function plays a basic role in the statistical inference. Some of its basic features are well known; e.g., because the score function of the normal distribution is linear in the location and quadratic in the scale, the score function plays a similar role as the mean and scale also for other distributions. We shall give some further illustratations of its basic role, as

(1) a characterization of the score function by a constant regression with respect to the maximal invariant;
(2) possible skrinkage of the score function in the multivariate location model;
(3) an expression of the score function of an arbitrary statistic by means of the score function of the observations, and various applications of this identity, e.g. in (finite sample) approximations of powers of tests and of moments of estimators.

References

Bondesson, L. (1974). Characterization of probability laws through constant regression. *Z. Wahrscheinlichkeitsth. Verw. Geb* . **30**, 93-115.

Jureèková, J. (1999). Equivariant estimators and their asymptotic representations. *Tatra Mountains Publications* **17**, 1-9.

Jureèková, J. and Milhaud, X. (1994). Shrinkage of maximum likelihood estimator of multivariate location. *Asymptotic Statistics* (P. Mandl and M. Hušková, eds.), pp. 303-318. Physica-Verlag, Heidelberg.

Jureèková, J. and Milhaud, X. (1999). Characterization of distributions in invariant models. *J. Statist. Planing Infer.* **75**, 353-361.

Jureèková, J. and Milhaud, X. (2001). Some finite-sample properties of statistics and tests. Preprint.

Klaassen, C. A. J. (1980). Statistical Performance of Location Estimators. PhD Disertation, Mathematisch Centrum, Amsterdam.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Testing Hypotheses for Gibbs Random Fields with an Application to the Ising Model

Martin Janž ura
Institute of Information Theory and Automation, Acad. of Sciences of the Czech Republic
Pod vodárenskou vìž í 4, 182 08 Prague, Czech Republic
janzura@utia.cas.cz

1. Gibbs Distributions

The statistical inference for Gibbs distributions has been recently widely studied because of its relevance for image processing and spatial statistics. The Gibbs distributions were originally used in the frame of statistical physics to describe the equilibrium states of large systems. They can be also understood as a infinite-dimensional generalization of the usual exponential families of distributions or the log-linear models for contingency tables.

The role of exponential statistics is played by systems of interactions. But, and that is the main difference to compare with finite systems or random sequences, the Gibbs random fields may not be given uniquely by the system of interactions (the phenomenon is called as "phase transitions"), and, moreover, there may be a non-translation-invariant Gibbs distribution with respect to a translation invariant system of interactions ("symmetry breakdown") - cf., e.g., Georgii (1988).

2. Maximum Pseudolikelihood Estimation

A natural parametrization, given by the system of interactions, turns the statistical inference problems to standard parametric procedures. Since the ML estimate meets both the theoretical and numerical limitations, the maximum pseudo-likelihood (MPL) method was introduced (Besag (1975)) as an alternative. It consists in replacing the likelihood

$$p_V^q(x_V)$$

where $q \in \Theta \subset R^k$ is the parameter, $V \subset Z^d$ is the observation region and $x_V \in c$ is the data configuration (observed image), by its pseudo likelihood counterpart

$$\prod_{t \in V} p^q_{\Lambda+t \mid \Lambda^c+t}\left(x_{\Lambda+t} \middle| x_{\Lambda^c+t}\right)$$

where $\Lambda \subset Z^d$ is „small" (usually $\Lambda = \{0\}$). The MPL estimate is consistent (cf., e.g., again Comets (1992)), in general it is not efficient (Janž ura (1997)), but, to the contrary to the ML estimate, it is asymptotically normal for every parameter regardless of phase transition or symmetry breakdown (Comets and Janž ura (1998)). The latter includes proving a proper version of the central limit theorem (CLT) for non-translation-invariant random fields.

3. Testing Hypotheses

Now, we would like to test the submodel, i.e. the composite hypothesis

$$H^0 : q^{\ell+1} = q^{\ell+2} = \cdots = q^k = 0$$

for some $\ell < k$. We shall construct the test statistics in the standard c^2 form, namely

$$\hat{c}^2 = \left(\hat{q} - \tilde{q}\right)^{\mathrm{T}} M \left(\hat{q} - \tilde{q}\right)$$

where $\hat{q}$ and $\tilde{q}$ are the MPL estimates in the full model and undepr the hypothesis, respectively, and M is a suitable random matrix of order $k - \ell$, making $\hat{c}^2$ asymptotically $c^2_{k-\ell}$ distributed. The proof is based on a further generalized version of

4. Ising Model

The (two dimensional) Ising model is the most elementary non-trivial case of the Gibbs random fields. It can be defined by specifying the local characteristics

$$p^{q}_{t|\{t\}^c}\left(x_t \middle| x_{\{t\}^c}\right) = \frac{\exp\left\{x_t\left(q_2 + q_1 \sum_{s \in \partial t} x_s\right)\right\}}{\cosh\left\{q_2 + q_1 \sum_{s \in \partial t} x_s\right\}}$$

where $(q_1, q_2) \in R^2$ is the parameter, $x_t \in \{-1,1\}$ for every $t \in Z^2$, and $\partial t \subset Z^2$ is the nearest neighborhood of the site $t \in Z^2$ (for a detailed treatment cf., e.g., Georgii (1988)).

Let us consider the hypothesis

$$H^0 : q_2 = 0$$

which, using the statistical physics terminology, means the absence of an external field.

The above method was applied and tested with simulated data. It was shown that for a sufficiently large observation region (200×200) the procedure works reliably even in the phase transitions area, i.e. for $q_2 = 0$ and $q_1 > q_c \doteq 0.44$.

Acknowledgement

This work is supported by the Grant Agency of the Czech Republic under Grant No. 201/00/1149.

References

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician* **24**, 179-195.

Comets, F. (1992). On consistency of a class of estimators for exponential families of Markov random fields on a lattice. *Ann. Statist.* **20**, 455-468.

Comets, F. and Janžura, M. (1998). A central limit theorem for conditionally centered random fields with an application to Markov fields. *J. Appl. Prob.* **35**, 608-621.

Georgii, H.O. (1988). *Gibbs Measures and Phase Transitions*. DeGruyter, Berlin.

fields.

Kybernetika **33**, No.2, 133-159.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Minimax Prediction Under Random Sample Size

Alicja Jokiel-Rokita
Wroclaw University of Technology, Institute of Mathematics
Wybrzeze Wyspianskiego 27, PL-50-370 Wroclaw, Poland
arokita@im.pwr.wroc.pl

The problem considered in the paper belongs to a class of problems for which the aim is to predict the value of a random variable Y on the basis of the observation of a random variable X, where X and Y have a distribution dependent on the same unknown parameter.

The paper deals with a special form of such problems - namely, with the problem of finding a minimax predictor of the random variable Y having the multinomial distribution $\mathrm{M}(m,p)$, where m is known and $p=(p_1,\ldots,p_r)\in P$, $P=\left\{p=(p_1,\ldots,p_r): p_i\geq 0, i=1,\ldots,r, \sum_{i=1}^{r} p_i=1\right\}$, is an unknown parameter, or multivariate hypergeometric distribution $\mathrm{H}(m,w)$, where m is known and $w=(w_1,\ldots,w_r)\in W=\left\{w=(w_1,\ldots,w_r): w_i\in N, i=1,\ldots,r, \sum_{i=1}^{r} w_i\geq m\right\}$ is an unknown parameter. The random variable X is such that $X\,|\,N=n$ has the multinomial distribution $\mathrm{M}(n,p)$, or multivariate hypergeometric distribution $\mathrm{H}(n,w)$, respectively.

We assume that the loss function connected with the predictor $d(X)$ is of the form

(1) $$L(Y,d(X))=(Y-d(X))^T C(Y-d(X)),$$

where C is a nonnegative matrix.

We have solved this problem in two cases, namely, when the sample size N is a random variable, whose distribution is known and in the case when it is unknown. In both cases we assume that this distribution does not depend on the unknown parameter p or w, i.e., that N is an ancillary statistic.

Contrary to the widely hold notion that the appearance of an ancillary statistic should not change the statistical inference, the following results are obtained: the minimax predictor of Y in the case when the sample size is fixed (Wilczynski (1985), Jokiel-Rokita (1998)) is seen to be neither minimax nor admissible if a random sample size is considered. The first example of such an ancillarity paradox was given by Brown (1990). He showed that in multiple linear regression the admissibility of the ordinary estimator of the constant term depends on the distribution of the design matrix, which is an ancillary statistic. Next example was presented by He (1990) who considered estimation of the multinomial probabilities with respect to the loss function given by (1), in which C was the identity matrix and the distribution of a random sample size is known. He proved that the estimator of p, obtained by Steinhaus (1957), which is minimax when the sample size is fixed, is neither minimax nor admissible when the sample size is random except for some trivial cases. Analogous

results were presented by Amrhein (1995) who studied minimax estimation of the multivariate hypergeometric proportions with respect to the same loss as He. He also proved that the minimax estimator obtained by Trybula (1958) in the fixed sample size case is neither minimax nor admissible when the sample size is random except for some trivial cases.

The following two situations explain the importance of considering of a random sample size and which lead to a random sample size in a natural way. First, suppose that a sample is drawn at random from a frame that contains the original population - also called the target or the domain of interest - as a subset. Then the size of the subsample belonging to the target has a hypergeometric distribution. This may happen in connection with telephone surveys, if households are selected by random-digit dialing with a preassigned number of calling attempts. Second, there are so-called nonrensponse models that take into account that for certain units in the sample, it may not be clear to which stratum they belong. Nonrensponse typically occurs in surveys concerning sensitive data, such drug abuse or tax evasion. But even in general the facility of nonrensponse can never be ruled out, because the selected units may just not be available during the investigation. To model this so-called phenomenon of non-at-homes, we assume that the selected units independent of each other and independent of their strata fail to answer with the same probability. The effective sample size is then an ancillary statistic with a binomial distribution. In practice we are often faced with those and similar problems simultaneously.

References

Amrhein, P. (1995). Minimax estimation of proportions under random sample size. *JASA* **90**, 1107-1111.

He, K. (1990). An ancillarity paradox in the estimation of multinomial probabilities. *JASA* **85**, 824-828.

Jokiel-Rokita, A. (1998). Minimax prediction for the multinomial and multivariate hypergeometric distribution. *Applicationes Mathematicae* **25**, 271-283.

Steinhaus, H. (1957). The problem of estimation. *Ann. Math. Statist.* **28**, 633-648.

Trybula, S. (1958). Some problems of simultaneous minimax estimation. *Ann. Math. Statist.* **29**, 245-253.

Wilczynski, M. (1985). Minimax estimation for the multinomial and multivariate distributions. *Sankhya* **47**, 128-132.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

MAREG and WinMAREG: A Tool for Analysing Longitudinal Data with Drop-Outs

Christian Kastner
IZB Soft
St.-Martin-Str. 47, 81541 München, Germany

Andreas Fieger
ServiceBarometer AG
Gottfried-Keller-Straße 12, 81245 München, Germany

Christian Heumann
Institut für Statistik, Ludwig-Maximilians-Universität
SFB 386, Teilprojekt C3: Fehlende Daten, Ludwigstr. 33, 80535 München, Germany

Sandro Scheid
Institut für Statistik, Ludwig-Maximilians-Universität
SFB 386, Teilprojekt C3: Fehlende Daten, Ludwigstr. 33, 80535 München, Germany
scheid@atheme.uni-muechen.de

Marginal regression models are an extension of the usual generalised linear model (GLM) to the case of longitudinal data. Beginning with the stimulating paper of Liang and Zeger (1986) a lot of methods for handling correlated data were proposed. The generalised estimating equations (GEE) approach of Liang and Zeger (1986) is a semiparametric quasi-likelihood approach for correlated data using the correlation as a measure of association. It was extended to several measures for the association and several methods for estimating the parameters. An overview of these methods is given e.g. by Ziegler, Kastner and Blettner (1998).

Many studies suffer from missing or incomplete data. In this situation, either all available cases or the complete cases are used by most computer programs. The GEE approach may yield biased estimates if data are not missing completely at random (Robins, Rotnitzky and Zhao, 1995).

In literature, two different approaches have been proposed to deal with the problem of missing data. Imputation methods impute missing data. By contrast, weighting methods discard the incomplete data but weight observations inversely proportional to their observation probability (Paik, 1997). Thus, the weighting estimating equations (WEE) generally follow the classical Horvitz-Thompson approach. They have been extensively discussed in recent years (Robins et al., 1995; Robins and Rotnitzky, 1995; Rotnitzky and Robins, 1995) but have rarely been applied in practice.

While the GEE was implemented in some software packages during the last years (Ziegler and Grömping, 1998), the WEE are not available in accessible form with computer software.

MAREG and WinMAREG implement these methods with our requirements for user friendly software, easy to handle for the user, run as stand alone program, support a standard database file format, coding of categorical variables, give online help and chance to handle big data sets.

MAREG is the programm for estimating the marginal regression models. It is available for DOS and UNIX platforms (currently Sun Solaris).

WinMAREG is the WINDOWS user-interface to specify the model you want to analyse.

The first step is to open the data file in WinMAREG. Currently supported file formats are dBase and Paradox database tables. The data will then be displayed in a grid .After this the estimation procedure can be chosen through the menu, specifying GEE or ML and the type of the link function depending on the kind of response. Available link functions are the identity link (continuous response), the logit link (binary response), the cumulative logit link (multi-categorical, especially ordinal response) and the multinomial logit link (for general multicategorical response). WinMAREG then opens the dialog for model specification.

Here the mean model and the association structure can be specified. For the GEE approach the IEE and the method of Prentice (correlation method) are implemented, the ML procedure supplies the IEE and the conditional odds ratio method. As a next step a design matrix has to be created from the selected variables.

WinMAREG provides automatic construction of design matrices including an intercept, that allow the user to specify models with fixed effects, varying intercepts, varying covariate effects, varying intercepts and covariate effects and a user defined design. For the association structure exchangeable, stationary, unspecified or a user-defined association structure is available. Categorical variables are usually coded as 1,2,... . For marginal regression models these variables have to be coded. WinMAREG gives you the opportunity to do this automatically. Dummy and effect coding are available. Another point in the analysis of real data sets is the problem of missing data. As in standard software packages any numeric value can be specified to represent a missing value. MAREG then performs a complete case analysis. For the GEE approach the WEE method is also available.

If the model is chosen, WinMAREG produces a plain text file of the data, where the categorical variables chosen in the model, are already coded as specified. It also writes a so-called CAI file (plain text), containing the information that is needed by MAREG.

Finally it runs MAREG in a DOS-window. As MAREG is also available for UNIX platforms, WinMAREG gives you the opportunity to write the data and the CAI file without starting MAREG. By transferring the data and CAI files to a UNIX machine, they can be used with the UNIX version of MAREG. This is very helpful when the data sets are large.

As there is no software tool that is able to handle marginal regression models in an easy way we hope MAREG and WinMAREG will come up to the requirements for an easy to use software. WinMAREG requires MS Windows 95 or above. There are no restrictions concerning the number of variables or cases. Clearly there are a lot of features not available in this version of MAREG, but we hope that new features as diagnostic methods and new approaches can be implemented soon.

The latest version of MAREG and WinMAREG is available from the authors or at *http://www.stat.uni-muenchen.de/~andreas/mareg/winmareg.html*.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

An Upper Bound on the Binomial Process Approximation to the Exceedance Process

E. Kaufmann, R.-D. Reiss*
University of Siegen, FB 6 – Mathematik
Walter-Flex-Str. 3, D- 57068 Siegen, Germany
edgar@stat.math.uni-siegen.de, reiss@xtremes.stat.math.uni-siegen.de

1. Introduction

The decisive, innovative step forward in extreme value statistics during the last decades was the parametric modeling of distribution functions (dfs) of exceedances over a threshold u, which are left-truncated dfs at u, by means of generalized Pareto (GP) dfs. The special modeling of such truncated dfs by means of GP dfs is motivated by the fact that the possible, continuous limiting dfs - as the threshold u goes to the upper endpoint of the support of the original df F - are GP dfs.

The exceedances over the threshold u can be distributionally represented by a binomial point process. In this context, we are interested in the overall error which occurs when the truncated df is replaced by an appropriate GP df. More precisely, given n iid rvs with common df F, the exceedances over a threshold u can be distributionally described by a binomial point process N(0) (cf. Reiss (1993)). The original binomial process N(0) will be replaced by a binomial process N(1), where the truncation of F is replaced by a GP df W.

By computing remainder terms one is also able to discuss the concept of penultimate distributions. Within the GP model one may find a df which provides a more accurate approximation to the left truncation of F than the limiting one. Such a GP df is called penultimate df.

2. The Ultimate Approximation

The aim is to establish an upper bound on the variational distance

$$\Delta(n,u) = \sup |P\{N(0) \in M\} - P\{N(1) \in M\}|$$

between the processes N(0) and N(1), where M ranges over all measurable sets in the space of point measures. Thus, one gets a bound on the overall error which occurs, when F is replaced by W.

To obtain a sharp upper bound on the variational distance between the distributions of the point processes one must deduce a bound on the Hellinger distance between the left truncation of F and W (cf. Corollary 1.2.4 in Falk et al. (1994)). If a corresponding inequality is applied, which is formulated in terms of the variational distance between the left truncation of F and W, one gets an inaccurate rate (cf. Reiss (1993)).

An upper bound on the Hellinger distance will be established in terms of an auxiliary function which is based on the hazard function. Thus, our bound is related to the well-known von Mises condition which is sufficient for F to belong to the pot-domain of attraction of W.

---

* representing the paper

3. The Penultimate Approximation

The starting points are the conditions which determine an upper bound for the ultimate rate. Under these conditions a penultimate approximation exists if, and only if, the auxiliary function, based on the hazard function, is slowly varying (cf. Kaufmann (2000)).

Under some additional condition we compute a bound on the remainder term in the penultimate GP approximation which corresponds to that in the ultimate approximation.

4. Concluding Remarks

An extended outline of these results and applications in real world problems may be found in the recent 2nd edition of the book by Reiss and Thomas (2001).

References

Falk, M., Hüsler, J. and Reiss, R.-D. (1994*). Laws of Small Numbers: Extremes and Rare Events.* DMV-Seminar Bd **23**. Birkhäuser, Basel.

Kaufmann, E. (2000). Penultimate approximations in extreme value theory. *Extremes* **3**, 39-55.

Reiss, R.-D. (1993). A Course on Point Processes. Springer, New York.

Reiss, R.-D. and Thomas, M. (2001). *Statistical Analysis of Extreme Values.* Birkhäuser, Basel (1st ed., 1997).

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

On Robust Forecasting under Distorted Regression Models and Systems of Simultaneous Equations

Yurij Kharin
Belarussian State University, Dept. of Mathematical Modeling and Data Analysis
4 Fr. Skoriny av., 220050 Minsk, BELARUS
kharin@fpm.bsu.unibel.by

Sviatlana Staleuskaya
Belarussian State University, Statistical Analysis and Modeling Research Laboratory
4 Fr. Skoriny av., 220050 Minsk, BELARUS
stalev@fpm.bsu.unibel.by

Modeling and forecasting of dynamics of stochastic systems are always based on a set of prior assumptions (hypothetical models). The validity of these assumptions can have important consequences for the validity of the final forecasting results. Forecasting procedures are called robust if "they are not affected much by small changes in the assumptions". Robustness is a topical subject from both a theoretical and a practical point of view (Hampel et al., 1986), (Huber, 1981), (Kharin, 1996), (Kharin and Staleuskaya, 1997), (Krishnakumar J. and Ronchetti E., 1997) (Lucas, 1996). From a theoretical perspective, robustness stimulates researchers to determine the crucial assumptions underlying their results. The practical relevance of robustness is easily illustrated by considering, for example, the development of economic policy recommendations based on statistical forecasts. If a forecast alters dramatically when the model assumptions are changed only slightly, the policy maker might be just off as without any forecast.

This paper is devoted to the problems of robustness analysis, robust estimation and forecasting by regression models and systems of simultaneous equations.

We consider a complete linear system of simultaneous equations in N jointly dependent variables and K predetermined variables, observed for T successive time periods. Let y_t be the tth observation on the N-vector of jointly dependent variables; x_t be the tth observation on the K-vector of predetermined variables; $x_t \in \mathbf{R}^N$ represents the disturbance term at the tth observation. Then the system of simultaneous equations is written as (Greene, 1983)

(1) $$A' y_t + B' x_t = x_t,$$

where A and B are $N \times N$ and $K \times N$ matrices of unknown coefficients respectively under prior restrictions:

(2) $$R\,\mathrm{vec}(\Gamma) = b, \qquad \Gamma = \begin{bmatrix} A \\ B \end{bmatrix},$$

where R is a fixed matrix and b is a fixed vector.

Assuming that A is nonsingular matrix we can solve (1) to obtain its reduced form, which is the well-known multivariate regression model:

(3) $$y_t = q' x_t + u_t,$$

where $q = -BA^{-1}$, $u_t = \left(A^{-1}\right)' x_t$.

We consider the family of "plug-in" algorithms of forecasting of y_{T+t} $(t \geq 1)$:

(4) $$\hat{y}_{T+t} = \hat{q}' x_{T+t} = \left(-\hat{B}\hat{A}^{-1}\right)' x_{T+t},$$

where $\hat{A}, \hat{B}$ are some statistical estimators of the coefficients.

The classical estimators are sensitive to deviations from the model distribution, to outlying observations, to model misspecifications, etc. We concentrate our attention mainly on the most typical cases of these distortions: errors-in-variables, Tukey-Huber outliers and additive distortions (Kharin Yu., Staleuskaya S., 1998).

We use the following quantitative characteristics of robustness. Let $\hat{\Psi}$ denotes a statistical estimators of the $(m_1 \times m_2)$-matrix Ψ (e.g., $\Psi = \Gamma$ or $\Psi = q$). Then as measures of robustness we consider the following functionals (Hampel et al., 1986), (Kharin, 1996):

$(m_1 \times m_2)$-matrix of bias: $b\left\{\hat{\Psi}\right\} = \mathrm{E}\left\{\hat{\Psi} - \Psi\right\}$;

$(m_1 \times m_1)$-mutual covariance matrix of the i th and j th column vectors $\hat{\Psi}^i, \hat{\Psi}^j \in \mathbf{R}^{m_1}$ of the matrix $\hat{\Psi}$: $V_{ij}\left\{\hat{\Psi}\right\} = \mathrm{E}\left\{\left(\hat{\Psi}^i - \Psi^i\right)\left(\hat{\Psi}^j - \Psi^j\right)'\right\}, i, j \in \{1, 2, \ldots, m_2\}$;

mean square risk of forecasting: $r = \mathrm{E}\left\{ \| \hat{y}_{T+t} - y_{T+t} \|^2 \right\} \geq 0$.

We construct exact formulas and asymptotic expansions of these functionals under distortions. To robustify the forecasting algorithm we consider the family of *M*-estimators for coefficients of simultaneous equations (1), (2):

$$\sum_{t=1}^{T} r\left(A' y_t + B' x_t\right) \to \min_{A,B}, \quad R\,\mathrm{vec}(\Gamma) = b,$$

where $r(\cdot)$ is a specially chosen function.

The theoretical analysis is illustrated by the results of computer modeling.

References

Greene W.M., (1993). *Econometric Analysis.* Macmillan Publishing Company, N.Y.

Hampel F.P., Ronchetti E.M., Rousseeuw P.J., and Stahel W.A., (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, N.Y.

Huber P., (1981). *Robust Statistics.* Wiley, N.Y.

Kharin Yu., (1996). *Robustness in Statistical Pattern Recognition.* Kluwer Academic Publishers, Dordrecht/Boston/London.

Kharin Yu., Staleuskaya S., (1997). On stability of multivariate linear regression forecasting. *Proceedings of the Academy of Sciences of Belarus (Series of Phys.-Math. Sciences)*, **4**, pp. 9–13.

Kharin Yu., Staleuskaya S., (1998). Robustness in statistical analysis of regression and simultaneous equations models. *Prague Stochastics'98*, UCMP, pp. 289–293.

Krishnakumar J. and Ronchetti E. (1997). Robust estimators for simultaneous-equations models. *J. of Econometrics.* Vol. **78**, p. 295–314.

Lucas A., (1996.) *Outliers Robust Unit Root Analysis.* Thesis Publishers, Amsterdam.

Staleuskaya S., Kharin Yu., (2000). Robustness of approximating approach in simultaneous equations models. *New Trends in Probability and Statistics*, vol. **5**. TEV, Vilnius, pp. 143–150.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Extremal Behaviour of Stochastic Processes in Finance

Claudia Klüppelberg
Munich University
Center for Mathematical Sciences of Technology
D-81290 München, Germany
cklu@matematik.tu-muenchen.de

One of the most prominent problems of the financial industry is the measurement of portfolio risk. Two standard methods use the empirical or normal method for estimation; both methods have been heavily criticised for not capturing risk sufciently. Alternative risk measures are based on quantiles, as for instance the Value-at-Risk (VaR) or the shortfall. The VaR, for instance, is based on the 1%-quantile of the profit-loss distribution and has become a benchmark risk measure, also accepted by regulators.

Estimation methods for quantiles have been developed under the acronym *let the tails speak for themselves*; it only uses such data which are responsible for the extremal behaviour. Standard procedures exist for iid observations anal are applicable to estimate VaR, shortfall and other quantile risk measures [see e.g. Embrechts, Klüppelberg and Mikosch (1997) for mathematical and statistical background and Emmer, Klüppelberg and Trüstedt (1998) for an explicit example].

Financial data, however, are not iid but exhibit a rather complex dependence structure, which can be modelled by diffusion models or (G)ARCH models. We describe the extremal behaviour of such volatility models and explain the estimation procedure of risk measures in this context [Borkovec (2000), Borkovec, M. and Klüppelberg, C. (1999), Borkovec, M. anal Kluppelberg, C. (1997)].

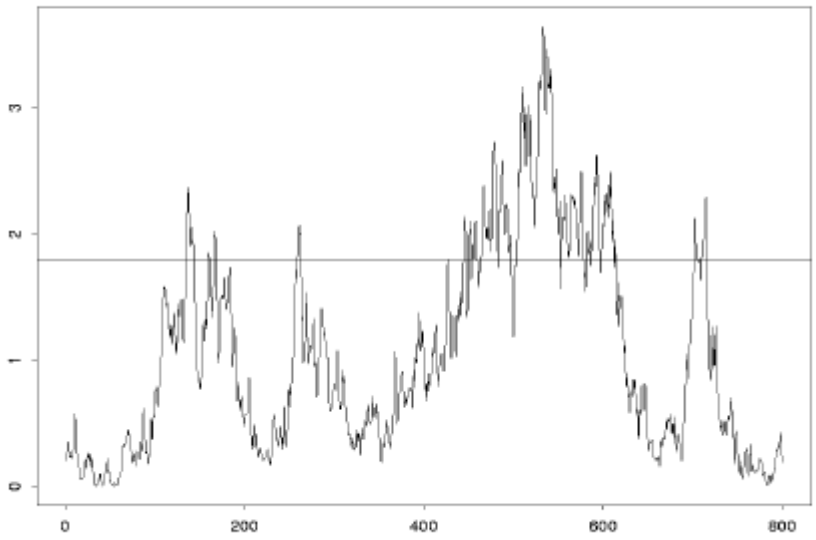


Figure 1. Simulated sample path of the Cox Ingesoll-Ross term structure model, given as solution to the SDE $dX_t = (c - dX_t)dt + s\sqrt{X_t}dB_t, t > 0,$

with (B_t) standard Browmsche motion and parameters $c = d = s = 1$.

References

Borkovec, M. (2000) Extremal behavior of the autoregressive process with ARCH(1) errors. *Stoch. Proc. Appl.* **85**, 289-207.

Borkovec, M. and Klüppelberg, C. (1999) The tail of the stationary distribution of an autoregressive process with ARCH(1) errors. *Ann. Appl. Probab.* To appear.

Borkovec, M. and Klüppelberg, C. (1997) Extremal behaviour of diffusion models in finance. *Extremes* 1, 47-80.

Embrechts, P. Klüppelberg, C. and Mikosch, T. (1997) Modelling Extremal Events for Insurance and Finance. Springer, Berlin.

Emmer, S., Klüppelberg, C. and Trüstedt, M. (1998) VaR - ein Mass fr das extreme Risiko. *Solutions* **2**, 53-63. English version available at *http://www.ma.tum.de/stat/*

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

On the Multivariate Skew Normal Distribution

Tõnu Kollo
University of Tartu, Department of Mathematical Statistics
J.Liivi Street 2, Tartu, Estonia
kollo@ut.ee

Imbi Traat
University of Tartu, Department of Mathematical Statistics
J.Liivi Street 2
Tartu, Estonia
imbi@ut.ee

Multivariate skew normal distribution has been defined by Azzalini, Dalla Valle (1996) as a two-parameter distribution. We observe it in a slightly different way defined in Gupta, Kollo (2000).

A random p-vector Z is distributed skew-normally, $Z \sim SN_p(\Sigma, a)$, with parameters a : p-vector and Σ: $p \times p$ positive definite matrix if its density has the form

$$f_Z(x) = 2 \quad f_{N(0,\Sigma)}(x) \quad \Phi(a^T x),$$

where $f_{N(0,\Sigma)}(x)$ is the density of $N(0,\Sigma)$ and $\Phi(\cdot)$ denotes the distribution function of $N(0,1)$. The distribution has several favourable properties. Its moment generating function

$$M(t) = 2 \ \exp \ (\frac{1}{2} t^T \Sigma \ t) \ \Phi(\ \frac{a^T \Sigma \ t}{(1 + a^T \Sigma \ a)^{1/2}} \ ),$$

makes it easy to find moments and cumulants. The expectation, dispersion matrix and the third central matrix moment $m_3(Z)$ are:

$$EZ = m = \frac{1}{c(a)} \ \sqrt{\frac{2}{p}} \ \Sigma \ a \ ,$$

$$DZ = \Sigma - m \quad m^T ,$$

$$m_3(Z) = \frac{1}{c^3(a)} \sqrt{\frac{2}{p}} \ (\frac{4}{p} \quad -1) \ (\Sigma \ \ a) \ ^{\otimes 2} \ a^T \Sigma ,$$

where $c(a) = (\ 1 + a^T \quad \Sigma \ a \ )^{1/2}$ In the presentation we will give also expressions of higher order moments and cumulants. The simulation of Z is also very convenient

$$Z = \begin{cases} X, \ if \quad X_0 > 0, \\ -X, \ otherwise, \end{cases}$$

where

$$\begin{pmatrix} X_0 \\ X \end{pmatrix} \sim N_{p+1}(0,\ \Sigma^*),$$

with the dispersion matrix

$$\Sigma^* = \begin{pmatrix} 1 & s^T \\ s & \Sigma \end{pmatrix}$$

and $s = Cov\ (X_0,\ X)$ being a function of a and Σ.

In the talk estimators of the parameters a and Σ will be examined with the emphasis on their asymptotic properties. In Azzalini, Dalla Valle (1996) and Azzalini, Capitanio (1999) the correlation matrix is considered as a parameter of a skew normal distribution instead of the dispersion matrix in our presentation. Our parameterization makes it possible to describe the distributions of estimators of the parameters in a simpler way. Results of an simulation experiment will be also presented and the possibilities of approximation by the skew normal distribution examined.

References

Azzalini, A., Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution, *J. R. Statist. Soc.* B **61**, 579-602.

Azzalini, A., Dalla Valle, A. (1996). The multivariate skew-normal distribution, *Biometrika* **83**, 715-726.

Gupta, A. K., Kollo, T. (2000) Multivariate skew normal distribution: Some properties and density expansions. Technical Report No. 00-05, Department of Mathematics and Statistics, Bowling Green State University, Ohio.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

An Approach to Stochastic Inverse Problems Using the Kalman Smoother and EM Algorithm

Franz Konecny
University of Agricultural Sciences, Vienna
konecny(@mail.boku.ac.at

In this paper we are concerned with parameter estimation of distributed systems, specified by a stochastic partial differential equation (SPDE). As an illustrating example, we consider a groundwater flow through a porous medium which is modeled by the continuity equation

(1) $$S\frac{\partial h}{\partial t}+\nabla\cdot \mathrm{q}=s(t,\mathrm{x})+x(t,\mathrm{x})$$

and Darcy's law

$$\mathrm{q}=-K\nabla h$$

where h = hydraulic head, q = fluid velocity vector, S = specific storage coefficient, K = hydraulic conductivity, s = deterministic source term and x = stochastic forcing term. Equations (1) and (2) can be combined to give a stochastic partial differential equation (SPDE) for the hydraulic head

(2) $$S\frac{\partial h}{\partial t}=\nabla\cdot(K\nabla h)+s(t,x)+x(t,x).$$

The equation may be adapted to 1D, 2D or 3D-flow, subject to some initial and boundary conditions.

The hydraulic conductivity K is a measure how freely the fluid is flowing in the medium. Since the medium is heterogenous, the hydraulic conductivity depends on the position x. The lack of information about $K(\mathrm{x})$ makes it natural to represent it as a random field. In the simplest case the log-conductivity is a homogeneous and isotropic Gaussian random field, which can be written as

(3) $$Y(\mathrm{x})=m_Y+\tilde{\mathrm{Y}}(\mathrm{x}),$$

where $m_Y=E[Y(\mathrm{x})]$ and $\tilde{Y}(\mathrm{x})$ is the fluctuation part with $E[\tilde{Y}(\mathrm{x})]=0$. In the literature, the following covariance function is often used to represent the spatial variability of hydraulic conductivity:

(4) $$R(\mathrm{x},\mathrm{x}')=s_Y^2\exp\left[-\frac{|\mathrm{x}-\mathrm{x}'|}{l_Y}\right].$$

s_Y^2 is the variance of Y and l_Y the correlation length.

We are concerned with the identification of the mean and covariance parameters of the log conductivity field. Usually only few measured values of the conductivity are available, which can not provide reliable estimates of the unknown parameters. Therefore estimation has be based on the noisy measurements of the hydraulic heads. The problem under consideration is an inverse problem of the SPDE (2).

For this task we propose an EM algorithm of the type, investigated by Dembo and Zeitouni (1986). The algorithm involves iterations of fixed interval smoothing and the maximization of some pseudolikelihood function. Since we have a linear state equation and Gaussian system-and measurement noise, we are in the realm of the Kalman smoother, which is applied to the augmented state equation. The implementation of the smoother requires the solution of a Riccati differential equation for the smoother covariance and a stochastic linear differential equation to obtain the optimal estimate of the augmented state. We shall discuss finite dimensional approximations of the smoothing problem and the implementation of the EM algorithm. The filter is tested by application to a hypothetical aquifer.

References

Dembo, A. and Zeitouni, 0.(1986) Parameter estimation of partially observed continuous time stochastic processes via the the EM algorithm. *Stochastic Proc. and Appl.*, **23**, 91-113.

Omatu, S. and Seinfeld, J.H.(1989) *Distributed Pammeter Systems, Theory and Applications*. Clarendon Press, Oxford.

Sun, N.-Z.: *Inverse Problems in Groundwater Modeling*. Dordrecht: Kluwer Acad. Publ., 1994.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Goodness of Fit of the Mathematical Model of the Process Grain Threshing and Separating in Multi-Drum Threshing Device

Andrzej Kornacki
Agricultural University,Institute of Applied Mathematics
Akademicka 13, 20-950 Lublin
AKORNAC@URSUS.AR.LUBLIN.PL

1. Introduction

Combine-harvesters are generally used for harvesting grains and other plants. To increase their effectiveness, multi-drum threshing and separating complexies are applied. Threshing and subsequent separation of grain by the concave grate depend on many factors . Reliable determination of the effect of these factors is carried out experimentally. In two previous papers of the author ([Kornacki , A (2000)],[Kornacki , A (2000)]) the mathematical model of the process of threshing and separating of grain in multi-drum threshing device was constructed. In the present paper we examine the goodness of fit of this model.

2. Results Symbols

$k = 1,2,\ldots,n$	the number of yielding drum
l	coordinate of the point on the concave grate
$X_k = X_k(l)$	unthreshed grain on kth-drum
$Y_k = Y_k(l)$	free grain on the kth-drum
$Z_k = Z_k(l)$	separating grain on the kth-drum
L_k	the length of the concave grate of the kth-drum
$X_k^* = X_k(L_k)$	unthreshed grain at the output of the kth-drum
$Y_k^* = Y_k(L_k)$	free grain at the output of the kth-drum
$Z_k^* = Z_k(L_k)$	separated grain at the output of the kth-drum
A_k	the coefficient of intensity of threshing in the I stage on kth-drum
A_k B_k	the coefficient of intensity of threshing in the II stage on kth-drum
m_k	the coefficient of intensity of separation on kth-drum
N_k	the mass of cereal introduced into the kth-drum

The mathematical model of threshing and separating of grain in multi-drum threshing device can be described (cf 3,23,25 in [Kornacki , A (2000)]):

(1) $$X_k = X_{k-1}^* e^{-A_k - B_k l},$$

(2) $$Z_k = \left[X_{k-1}^* \left(1 - e^{-A_k}\right) + Y_{k-1}^* \right]\left(1 - e^{-m_k L_k}\right) + X^*_{k-1} \left[e^{-A_k} + \frac{m_k e^{-A_k - B_k L_k} - B_k e^{-A_k - m_k L_k}}{B_k - m_k} \right]$$

(3) $$Y_k = N_k e^{-m_k l} + X_{k-1}^* \left[\frac{m_k e^{-A_k - B_k l} - B_k e^{-A_k - m_k l}}{B_k - m_k} \right].$$

Moreover, the coefficients of intensity we can get from equations (cf 3.5, 3.7, 3.11 in [2]):

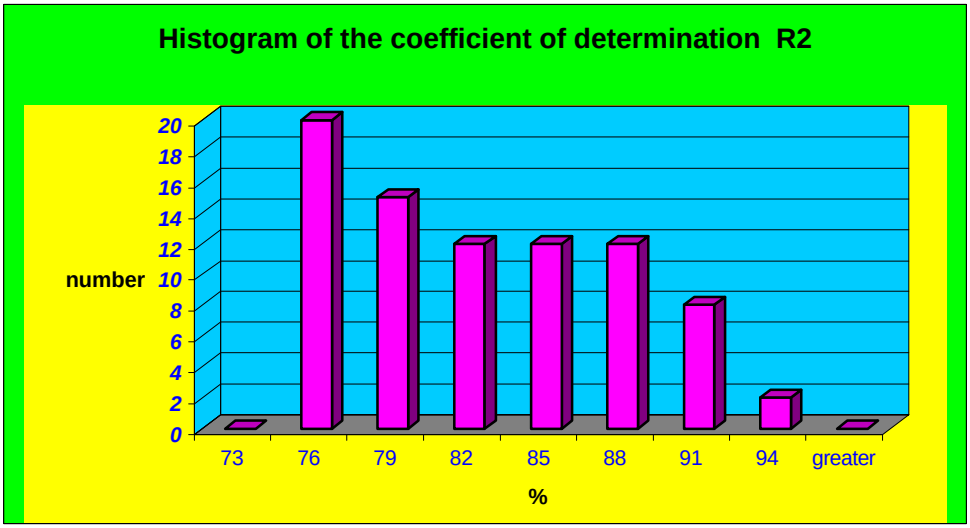
$$A_1 = -\ln(\frac{s}{X_0^*}) = \ln(\frac{X_0^*}{s}); \qquad B_1 = \frac{\ln(\frac{X_1^*}{s})}{-L_1}; \qquad e^{-m_1 L_1} = \frac{Y_1^* m_1 - B_1(X_1^* + Y_1^*)}{Y_0^* m_1 - N_1 B_1}$$

Now, we test a mathematical model for a real experimental data. Experiment was carried out in Department of Agricultural Engineering in Academy of Agriculture in Lublin in 1989-1990. Was examined on a bay eight-drum threshing device on threshing wheat Gran's. Using results from proceeding paper of the author (Kornacki,[2]) we can estimate coefficients A_1, B_1, m_1. Because a course of the process of threshing and separating is identical on a next drums so we assume $A_1 = A_2 = \ldots = A_n$, $B_1 = B_2 = \ldots = B_n$, $m_1 = m_2 = \ldots = m_n$

On the basis of model we can forecast the mass of separating grain in next drums. Then we can compare obtained results with experimental data using coefficients of determination:

$$R^2 = \frac{\sum_{i=1}^{8}(z_i - \bar{z})^2 - \sum_{i=1}^{8}(z_i - Z_i)^2}{\sum_{i=1}^{8}(z_i - \bar{z})^2}$$

In our case we have 81 observations. The values of determination coefficient varies from 73,36% to 93,01%. So, meaningful part of the changeability is explained by the mathematical model given by formulas (1)-(3). The histogram of values of the coefficient of determination is shown in figure below:



References

Kornacki , A (2000): Mathematical model of the threshing and separating process in multi-drum threshing device. *Annual Journal of Agricultural Engineering*, Vol **2**/1 pp 165-173.

Kornacki , A (2000): Verification of the mathematical model of the process of threshing and separating of grain in the two-drum threshing device. In¿ynieria rolnicza, In press

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Two Sample Test in the Situation of the Interval Censoring

Vladimir Koulikov, Piet Groeneboom, Hendrik Lopuhaä
Delft University of Technology, ITS faculty, CROSS department
Mekelweg 4, 2628CD Delft, The Netherlands
V.N.Koulikov@ITS.TUDelft.NL, P.Groeneboom@ITS.TUDelft.NL,
H.P.Lopuhaa@TWI.TUDelft.NL

Suppose the sample $\{(X_1,Z_1),\ldots,(X_n,Z_n)\}$ consists of realizations X_i of two random variables ξ_1 and ξ_2 with possibly different distributions and values Z_i indicating whether X_i is a realization from ξ_1 or ξ_2. We consider the situation where X_i becomes a realization of ξ_1 or ξ_2 randomly with probability p or (1-p). In this classical situation there are a lot of advanced methods to test the hypothesis H_0 "ξ_1 and ξ_2 are identically distributed" against any kind of alternative, but we want to test this hypothesis in the situation of interval censoring.

Suppose that instead of the random variable X we observe another random variable T (that is independent of X) and the indicator of the event {X<T}. The sample then consists of the vectors (T_i, d_i, Z_i), where T_i is the censoring variable, $d_i = 1_{X_i<T_i}$ and Z_i are either 0 or 1. We will consider testing procedures for the hypothesis H_0 in this setup, which are needed for some practical applications.

The procedures we consider most are intuitively obtained as a result of "non-formal" differentiating. Suppose we restrict ourselves to the case of the Lehmann alternative (this means: $\xi_1 \sim F_0$, $\xi_2 \sim F_o^{1+q}$, $q > -1$). Also suppose that the loglikelihood

$$l(F,q) = \sum_{i=1}^{n} \left(d_i (1+Z_i q) \log F(T_i) + (1-d_i) \log \left(1 - F(T_i)^{1+Z_i q} \right) \right)$$

is maximized by $(\hat{F}_n, \hat{q}_n)$. Since the maximum likelihood estimator is consistent we can suppose that under the null hypothesis θ=0 should more or less maximize $l(\hat{F}_n, q)$ as a function of θ (here and later $\hat{F}_n$ is the maximum likelihood estimator of the distribution function under the null hypothesis) or

$$\sum_{i=1}^{n} Z_i \hat{F}_n(T_i) \log \hat{F}_n(T_i) \left(\frac{d_i}{\hat{F}_n(T_i)} - \frac{1-d_i}{1-\hat{F}_n(T_i)} \right)$$

should be small. Notice that $\hat{F}_n(T_i) \log \hat{F}_n(T_i)$ is a "good" function of $\hat{F}_n$. One can suppose that something of that kind should hold in more general case. Our first theorem states:

Theorem 1 Suppose distributions of X (F_0) and of T (G) have a finite interval of support [0,M] and finite positive densities. Let the function $w(t)=t|log(t)|^m$ for m>0 or let $w \geq 0$ be a Lipschitz function with $w(0)=w(1)=0$. Then under the null hypothesis

$$\sqrt{n}\sum_{i=1}^{n} Z_i w\left(\hat{F}_n(T_i)\right)\left(\frac{d_i}{\hat{F}_n(T_i)}-\frac{1-d_i}{1-\hat{F}_n(T_i)}\right)\xrightarrow{D} N(0,s^2)\,.$$

The analogous theorem holds for the interval censoring, case II as well.

This theorem allows to construct the testing procedure. To determine the asymptotic power of the test we can make use of the following:

Theorem 2 In the conditions of Theorem 1 under the contiguous Lehmann alternative with $q_n = \frac{1}{\sqrt{n}} q_0$ (so for Z=0 X~F_0 and for Z=1 X~$F_0^{1+q_n}$)

$$\sqrt{n}\sum_{i=1}^{n} Z_i w\left(\hat{F}_n(T_i)\right)\left(\frac{d_i}{\hat{F}_n(T_i)}-\frac{1-d_i}{1-\hat{F}_n(T_i)}\right)\xrightarrow{D} N(m,s^2)$$

Nothing like the last theorem holds for the case of the location shift alternative; computer simulations show that immediately.

Computer simulations made for the testing procedures proposed by the Theorems 1 and 2 showed correctness of our results and really fast convergence to the limiting distributions. A big advantage of the method is that we only have to calculate the MLE under the null hypothesis. This makes application of it simpler and the computer program much faster.

The proof of Theorems 1 and 2 is mostly based on the chaining lemma, Hoeffding's inequality and partly relies on the ideas of Geskus and Groeneboom (1999). Some representations make it only necessary to prove the stochastic equicontinuity of the loglikelihood and finish the proof by the application of the central limit theorem.

If time allows some aspects of the likelihood ratio test will be considered.

References

Ronald Geskus and Piet Groeneboom (1999) "Asymptotically optimal estimation of the smooth functionals for the interval censoring, case II", *Annals of Statistics*, **27**, 627-674.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Inferences on Arima Model Selection and Suitability in Synodic Time Scale

T. Krishnan
Indian Statistical Institute
Chennai, India

Sujata Mukherjee
B14/168,Kalyani, West Bengal, India
sujmuk@hotmail.com

A *Stochastic model* is usually formulated from a given data, after several important *IDA and EDA*, normally seen as the stepping stone to a model based analysis. Many models may be tried involving appropriate *inference* procedures till the winner model, passing through the theory of inferences is selected from the rest.

Inference is based on a probability model having specific *Statistical equilibrium* of the form,

DATA = FIT + NOISE

Where **general interest is concerned on fit or forecasts**, the *Systematic component,* in the presence of residuals or noise, the *Random component* (Chatfield, 1995).

To fit a suitable model to the mathematical series $[y_{ti}]$ in Synodic time scale, part l (*STAT'2000, Poland*) modified to $[y_{tpi}]$ in Synodic time scale, part ll (*Stability Problem in Stochastic Model, 2001, Hungary*) with **exponential smoothing,** forecasts for linear and non-linear models, were not found to fit the series well.

Fitting a **quadratic polynomial** model, followed by a **cubic polynomial** model, with the *pre whitened series*, failed the tests to fit the data well.

Attempting to the class of **ARIMA models** (*Box & Jenkins, 1976*), following the principle of parsimony, with (p,d,q) lying between 0 and 2, allowing for total 27 possible models (*Chatfield, 1996*), a 2 parameter model, fitted the series well, with periodicity ***S***, the number of phases in one **S**ynodic month.

The 2parameter seasonal moving average ARIMA model successfully passed through the Box-Pierce test Statistics with smaller RMSE.

The **ARIMA model selection** was checked by over fitting with a 3parameter model, which did not show significant improvement over the 2parameter model.

1. The ARIMA (0,1,2) $(0,1,1)_S$

The 2parameter model, for the discrete series, $[y_{ti}]$ changed to Synodic time scale, $[y_{tpi}]$, with periodicity S, in the difference equation form is given as,

(1) $$(1-B^S)\, W_{tp} = (1-\theta_1 B- \theta_2 B^S)\,(1- \theta_S B^S)\, \hat{a}_{tp}$$

$W_{tp} = (1-B)\ln y_{tp}$, the non seasonal difference of the $\log_e$ transformed data in removing trend. The *seasonality* is taken as one Synodic month of S phases in its modified form when similarities in the series occur after S time interval, θ_1, θ_2 are the *non seasonal moving average parameters*, $(1-B^s)W_{tp}$, *the seasonal difference in removing seasonality*, θ_S *seasonal moving average parameter* and $\hat{\mathbf{a}}_{\mathbf{tp}}$, the *normally distributed uncorrelated random shock* with zero mean and constant variance, which generates the process.

The model is multiplicative in the sense that the observed series results from the successive filtering of the random noise series a_{tp}, through the non seasonal filter (between adjacent phases), then the seasonal filter (between Synodic months).

2. The Difference Equations of ARIMA (0,1,2) (0,1,1)$_S$

From equation (1), the forecast for origin = tp, in the difference equation form is,

(2) $$W_{tp}–W_{tp-s} = \hat{a}_{tp} - \theta_1 \hat{a}_{tp-1} - \theta_2 \hat{a}_{tp-2} - \theta_S \hat{a}_{tp-s} + \theta_{s+1} \hat{a}_{tp-s-1} + \theta_{s+2} \hat{a}_{tp-s-2}$$

The forecast for lead time T, is the mathematical expectation of y(tp+T), given as,

$$Y_{tp}(T) = E [y_{(tp+T)} \mid I_{tp}], \text{ where } I_{tp} = y_{tp}, y_{tp-1}, y_{tp-3},$$

$$W_{tp}(T) = [W_{(tp+T)}]$$

or,

(3) $$W_{(Tp)} = W_{(Tp-s)} - \theta_1 \hat{a}_{Tp-1} - \theta_2 \hat{a}_{Tp-2} - \theta_S \hat{a}_{Tp-s} + \theta_{s+1} \hat{a}_{Tp-s-1} + \theta_{s+2} \hat{a}_{Tp-s-2} + \hat{a}_{Tp}$$

The forecasts are affected by the moving average terms **q,** through (*S+2*) phases and the successive forecasts are the changes forecasted (-*S*) phases earlier. This way, the forecasts for the first phase of the ns Synodic month is estimated from the observed value of the first phase of the (n-1)s Synodic month.

(4) $$[W_{(ns+1)}]=$$

$$W_{((n-1)s+1)} - \theta_1 \hat{a}_{(ns)} - \theta_2 \hat{a}_{(ns-1)} - \theta_s \hat{a}_{((n-1)s+1} + \theta_{(s+1)} \hat{a}_{(n-1)s} + \theta_{(s+2)} \hat{a}_{((n-1)s-1)} + \hat{a}_{(ns+1)}$$

3. The Forecasting Methodology & Confidence Interval

To forecast the phase k data from a discrete mathematical series [z_{ti}] changed to the modified Synodic time scale [z_{tpi}] upto phase (k-1), is added to [y_{tpi}]. This now increases the data counts to (ns+(k-1)) phases, and then the ARIMA model is once again fitted, once again the parameters estimated and after checking the residual acfs, the forecast for the phase k is computed.

(5) $$[W_{(ns+k-1)}] = W_{(n-1)s} - \theta_1 \hat{a}_{(ns+k-2)} - \theta_2 \hat{a}_{(ns+k-3)} - \theta_s \hat{a}_{((n-1)s+k-1)} + \theta_{(s+1)} \hat{a}_{((n-1)s+k-2)} + \theta_{(s+2)} \hat{a}_{((n-1)s+k-3)} + \hat{a}_{(ns+k-1)}$$

The accuracy of forecast are checked by calculating 95% convenient set of probabilities. The confidence intervals are constructed with the variance of the forecast error, to find how reliable the forecasts are. The forecast error for origin tp and leadtime T is given by,

(6) $$e_{tp}(T) = W_{(tp+T)} - W_{tp}(T) = \hat{a}_{(tp+T)}$$

Where $W_{(tp+T)}$ is the observed value for period (tp+T) and $W_{tp}(T)$ is the forecast value for that period. The 95% confidence interval is calculated by the expression,

(7) $$W_{tp}(T) \pm 1.96\acute{o} [e_{tp}(T)]$$

The **ARIMA model Suitability,** will be presented in the paper, with the Statistical Inferences on **NOISE,** as the one step ahead forecast error, which generated the process and the possible nature of model inadequacy, if any.

Acknowledgement

SM thanks all at Computer Science Unit & Computer Science Servicing Center for their help during 1987–1993 at I S I, Calcutta, India.

References

Box,GEP & Jenkins,GM (1976) *Time series analysis,forecasting & control* Holden-Day

Chatfield C (1995) Editorial. *International Journal of Forecasting*.**11**

Chatfield C (1995) *Problem Solving. A Statisticians Guide*. Chapman & Hall.

Chatfield C (1996) *Model Uncertainty & Forecast Accuracy*.*J of Forecasting*.**15**

Krishnan T (1986) Editorial. *Data analysis in life sciences*.Proc.I S I, Calcutta, India

Nelson,CR (1973) *Applied Time Series Analysis for Managerial Forecasting*.Hol.-Day.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Comparisons in Location Based on a Quotient of Independent Measurements

Martina Kron, Wilhelm Gaus, Josef Högel
University of Ulm, Department of Biometry and Medical Documentation
Schwabstr. 13, D-89075 Ulm
martina.kron@medizin.uni-ulm.de

1. Introduction

The methodology was developed for analysis of data from a biological experiment since standard statistical methodology was inappropriate or not applicable. The peculiarities of the experiment were: 1. The outcome W:=X/(Y·Z) could not be measured in the same experimental unit. Each of its components had to be measured in different subjects because for measuring the subject was destroyed. 2. The distribution of X, Y, and Z was unknown and some outliers might have occurred.

A test statistic is proposed to compare two treatments in W for differences in location. The statistic is based on robust measures of location and dispersion in order to account for outliers in the data.

2. Robust Measures of Location and Dispersion

Assume that the outcome W:=X/(Y·Z) is unmeasurable and its components X, Y, and Z cannot be derived from the same experimental unit since it will have been destroyed after determination of either X, Y, or Z. Therefore, the distribution of W must be described by location and scale parameters of X, Y, and Z which have to be derived from independent experimental units.

Approximate formulas for E(X/Y) and Var(X/Y) are given in Mood et al. (1974). This approach can be extended to three stochastically independent variables X, Y, and Z. Thus, mean and variance of W can be approximated by

(1) $$E(W) \approx \frac{m_X}{m_Y m_Z}\left(1+\frac{s_Y^2}{m_Y^2}+\frac{s_Z^2}{m_Z^2}\right) \text{ and } Var(W) \approx \left(\frac{m_X}{m_Y m_Z}\right)^2 \cdot \left(\frac{s_X^2}{m_X^2}+\frac{s_Y^2}{m_Y^2}+\frac{s_Z^2}{m_Z^2}\right),$$

where μ_X, μ_Y, μ_Z and σ_X, σ_Y, σ_Z are mean and standard deviation of X, Y, and Z.

The existence of E(W) and Var(W) is not generally guaranteed by the finiteness of E(X) and E(Y·Z). However, E(W) and Var(W) exist if measurements are positive, have finite upper bounds, and are bounded away from 0.

Mean and variance of W can be estimated by using measures of location and dispersion for realisations of X, Y, and Z. If outliers might be observed, robust measures should be preferred. Appropriate robust and strongly consistent measures of location and dispersion are the (α,β)-trimmed mean

(2) $$\bar{x}_{a,b} := \frac{1}{n-\lceil na\rceil-\lceil nb\rceil}\sum_{i=\lceil na\rceil+1}^{n-\lceil nb\rceil} x_{(i)}$$

and the adjusted (α,β)-trimmed standard deviation

(3) $$\hat{s}_{a,b,x} := \sqrt{\frac{1}{n-SSQ_{a,b}}\cdot\sum_{i=\lceil na\rceil+1}^{n-\lceil nb\rceil}\left(x_{(i)}-\bar{x}_{a,b}\right)^2}$$

where $SSQ_{a,b} := \sum_{i=1}^{\lceil na\rceil} u_i^2 + \sum_{i=n-\lceil nb\rceil+1}^{n} u_i^2$ and u_i is the $\frac{i}{n+1}\cdot 100\%$-quantile of the standard normal distribution (Hampel et al. 1986, Högel et al. 1994).

3. Comparisons in Location

In experiments, a primary aim is to compare two experimental groups in location. Thus, e.g. the one-sided hypothesis

(4) $$H_0 : m_{W_1} - m_{W_2} \leq 0$$

shall be tested and the asymptotically normally distributed test statistic

(5) $$T := (\hat{m}_{W_1} - \hat{m}_{W_2}) / \hat{t}_{W_1,W_2}$$

is proposed, where $\hat{m}_{W_i}$, i=1,2, is the estimator for the mean of W in group i and

(6) $$\hat{t}_{W_1,W_2} := \sqrt{\hat{t}_{W_1}^2 + \hat{t}_{W_2}^2}\,, \text{ with } \hat{t}_{W_i}^2 := \left(\frac{\bar{x}_{a,b}}{\bar{y}_{a,b}\bar{z}_{a,b}}\right)^2 \cdot \left(\frac{s_{a,b,x}^2}{n_x \bar{x}_{a,b}^2} + \frac{s_{a,b,y}^2}{n_y \bar{y}_{a,b}^2} + \frac{s_{a,b,z}^2}{n_z \bar{z}_{a,b}^2}\right),$$

is the estimator for the standard deviation of W in group i, and n_X, n_Y, n_Z are the number of realisations of the random variables X, Y and Z.

4. Simulation Study

A simulation study with the following design was conducted: Normally distributed random variables X, Y, and Z were generated according to the parameter sets of the biological experiment. Realisations less than 1.0 or greater than mean plus 4 standard deviations were dropped. The test statistic T was calculated and simulations were repeated 50,000 times. The results show an approximate standard normal distribution.

5. Conclusion

A statistical test was proposed applicable if the outcome is the quotient of three independently measured random variables. A robust statistic was constructed to take care for outliers in the data. Simulations show that the test statistic is approximately normally distributed.

References

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P.J. and Stahel, W. A.. (1986). *Robust Statistics – The Method based on Influence Functions.* Wiley. New York.

Högel, J., Schmid, W., and Gaus, W. (1994). Robustness of the standard deviation and other measures of dispersion. *Biom. J.* **36**, 411-427.

Mood, A.. M., Graybill, F. A.. and Boes, D. C. (1974). *Introduction to the theory of statistics.* McGraw-Hill. Singapore.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

t -Products and $\sum$-Products in Probabilistic Normed Spaces

Bernardo Lafuerza-Guillén
blafuerz@pop.ual.es

In this work we give first a generalization of the results by Alsina and Schweizer. In addition we study $\sum$-products and finally, the product topologies in PN spaces which are products of countable families of PN spaces.

1. Finite t -Products of PN spaces

Definition 1 Let (V_1,ν_1,t,t^*) and (V_2,ν_2,t,t^*) be two PN spaces under the same triangle functions t and t^*. Let t_1 be a triangle function. The t_1-product of both PN spaces is the pair

$$(V_1\times V_2,\nu_1 t_1\nu_2)$$

where

$$\nu_1 t_1\nu_2 : V_1\times V_2 \to \Delta^+$$

is a probabilistic seminorm given by

$$(\nu_1 t_1\nu_2)(p,q) := t_1(\nu_1(p),\nu_2(q))$$

for any $(p,q)\in V_1\times V_2$.

Definition 2 Let $\{(V_i,\nu^i,t_i,t_i^*)\,|\,i\in\square\}$ be a countable family of PN spaces. The $\sum$-product of this family is the space $\left(\prod_{i=1}^{\infty}V_i,\nu^{\Sigma}\right)$ where $\nu^{\Sigma}:\prod_{i=1}^{\infty}V_i\to\Delta^+$ is a map given by

$$\nu_{\bar{p}}^{\Sigma} := \sum_{i=1}^{\infty}2^{-i}\nu_{p_i}^i$$

for every sequence $(p_i)=\bar{p}\in\prod_{i=1}^{\infty}V_i$.

In order to simplify the writing of this paper we replace $\nu_{\bar{p}}^{\Sigma}$, by $\nu_{\bar{p}}$.

Theorem 1 Let (V_1,ν_1,t,t^*), (V_2,ν_2,t,t^*) and t_1 be two PN spaces under the same triangle functions and a triangle function t_1 respectively. Assume that $t^*\square\ t_1$ and $t_1\square\ t$, then the t_1-product $(V_1\times V_2,\nu_1 t_1\nu_2)$ is a PN space under t and t^*.

Theorem 2 Let $\left(V_1,\|\cdot\|_1,G,t_M,M\right)$ and $\left(V_2,\|\cdot\|_2,G,t_M,M\right)$ and $\|\cdot\|_3$ be the two above mentioned PN spaces and the classical norm defined on $V_1\times V_2$ by

$$\|\bar{p}\|_3 := \|p_1\|_1\vee\|p_2\|_2,$$

with $\overline{p} = (p_1, p_2) \in V_1 \times V_2$. Then $\left(V_1 \times V_2, \| \cdot \|_3, G, t_M, M\right)$ is a PN space, which coincide with the M-product of both simple spaces.

Theorem 3 The t_M-product of two simple PN spaces $\left(V_1, \| \cdot \|_1, G, M\right)$ and $\left(V_2, \| \cdot \|_2, G, M\right)$ is the simple space under M generated by $\left(V_1 \times V_2, \| \cdot \|_s\right)$ and the same d.d.f G, namely, $\left(V_1 \times V_2, \| \cdot \|_s, G, M\right)$, where $\| \cdot \|_s$ is the classic norm defined via

$$\| \cdot \|_s := \| \cdot \|_1 + \| \cdot \|_2,$$

whatever the norms $\| \cdot \|_1$ and $\| \cdot \|_2$ may be.

2. Countable $\sum$-Products of PN Spaces

Theorem 4 Let $\left\{\left(V_i, v^i, t_i, t_i^*\right) | i \in \square\right\}$ be a countable family of PN spaces and let $t_i \geq t_W$ and $t_i^* \leq t_{W^*}$ for all $i \in \square$, then the $\sum$-product of this family denoted by

$$\left(\prod_{i=1}^{\infty} V_i, v^{\Sigma}, t_W, t_{W^*}\right)$$

is a Menger space under *W*.

3. Product Topology for Countable t-Products

Theorem 5 Let each of the PN spaces $\left(V_i, v^i, t_i, t_i^*\right)$ endowed with the strong topology corresponding to $v^i, i \in \square$, and Δ^+ with the topology of weak convergence. Then the product topology is weaker than the strong topology in (V, G).

Theorem 6 Let $\left\{\left(V_i, v^i, t_i, t_{W^*}\right) | i \in \square\right\}$ and V, v^{Σ} be as in Theorem 8. Let each V_i be endowed with the strong topology induced by v^i. Then the strong topology on V induced by v^{Σ} is the product topology.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Adaptive Estimation of a Quadratic Functional of a Density by Model Selection

Béatrice Laurent
Université Paris 11
Laboratoire de Mathématiques. Bât. 425, 91405 Orsay Cedex. France.

Let $X_1,\ldots,X_n$ be i.i.d. random variables with common density f belonging to $L^2(R)$. We propose an adaptive estimator of the quantity $\int_R f^2(x)dx$ which is based on model selection via some penalized criterion. Bickel and Ritov (1988) and Laurent (1996) have built estimators of $\int_R f^2(x)dx$ in a density model but these estimators depend on some prior information on f. Bickel and Ritov assumed that f belongs to some class of Hölderian functions of order a. They built an estimator $\hat{q}_a$ of $\int_R f^2(x)dx$ that is efficient if $a > 1/4$ and achieves the rate $n^{-4a/(1+4a)}$ if $a \leq 1/4$. They also proved that this rate of convergence is optimal. Similar results are obtained by Laurent (1996) with a simpler method of estimation based on projection estimators.

Following the ideas given in Laurent and Massart (2000) to estimate quadratic functionals in a Gaussian framework, we propose here an adaptive estimator of $\int_R f^2(x)dx$ in a density model. To define this estimator, we introduce some notations. Let

$$f(x)=\mathbf{1}_{[0,1]}(x), \quad y(x) = \mathbf{1}_{[0,1/2[}(x)- \mathbf{1}_{[1/2,1[}(x),$$

for any $k \in Z$ and $j \in N$, we define

$$f_{j,k}(x)=2^{j/2} f(2^j x-k), \quad y_{j,k}(x)=2^{j/2} y(2^j x-k).$$

The functions $(f_{0,k}, y_{j,k}, j \in N, k \in Z)$ form the Haar basis of $L^2(R)$. For any $J \in N$, the decomposition of f onto this basis can be written as

$$\sum_{k\in Z} a_{J,k} f_{J,k} + \sum_{j\geq J}\sum_{k\in Z} b_{j,k} y_{j,k}$$

where $a_{j,k} = \int f f_{j,k}$ and $b_{j,k} = \int f y_{j,k}$ and thus

$$\int_R f^2(x)dx = \sum_{k\in Z} a_{j,k}^2 + \sum_{j\geq J}\sum_{k\in Z} b_{j,k}^2 .$$

We consider an unbiased estimator of $\sum_{k\in Z} a_{J,k}^2$ namely

$$\hat{q}_J = \frac{1}{n(n-1)} \sum_{k\in Z} \sum_{l\neq l'=1}^{n} f_{J,k}(X_l)\ f_{J,k}(X_{l'}) .$$

Our adaptive estimator of $\int_R f^2(x)dx$ is defined as

$$\hat{q} = \sup_{J \in \Im} \left[\hat{q}_J - pen(J) \right],$$

where $\Im$ is a subset of N and $pen(J)$ is a penalty term that has to be conveniently chosen.

We give a non asymptotic risk bound for this estimator. We derive from this bound adaptive properties in the minimax sense over classes of functions including Hölderian classes and functions such that the sequence of coefficients $b = (b_{j,k})_{j \geq 0, k \in Z}$ belongs to some Besov body $B_{a,2,\infty}(R)$. Up to a logarithmic factor, our procedure is rate optimal simultaneously over all these classes. A crucial point in the proof of our results is an exponential inequality for U-statistics of order 2 due to Bretagnolle (1999).

References

Bickel, P. and Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence, *Sankhya Ser.* A, Vol. **50**, 381-393.

Bretagnolle, J. (1999). A new large deviation inequality for U-statistics of order 2. ESAIM : *Probability and Statistics*, **3**, 151-162.

Laurent, B. (1996). Efficient estimation of integral functionals of a density. *Ann. Statist.* **24**, No 2, 659-681.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28**, No 5.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection

Olivier Ledoit
Anderson Graduate School of Management, Finance Department
110 Westwood Plaza, Los Angeles, U.S.A.
oledoit@anderson.ucla.edu

Michael Wolf
Universidad Carlos III de Madrid, Departamento de Estadística y Econometría
Calle Madrid 126, Getafe, Spain
mwolf@est-econ.uc3m.es

The objective of this paper is to estimate the covariance matrix of stock returns. This is a fundamental question in empirical Finance with implications for portfolio selection and for tests of asset pricing models such as the CAPM.

The traditional estimator, the sample covariance matrix, is seldom used because it imposes too little structure. When the number of stocks N is larger than the sample size T, the sample covariance matrix is always singular (that is, not invertible) and in typical applications, there can be over a thousand stocks to choose from, but rarely more than ten years of monthly data, that is, $N = 1000$ and $T = 120$. Since the standard methods of portfolio selection as well as of testing asset pricing models require an estimate of the inverse of the covariance matrix of stock returns, this situation is clearly problematic.

The cure is to impose some structure on the estimator. Ideally, the particular form of the structure should be dictated by the problem at hand. In the case of stock returns, a low-dimensional factor structure seems natural. But this leaves two very important questions: How much structure should we impose? And what factors should we use? The first factor model to be suggested dates back to Sharpe (1963) who proposed the market (that is, a portfolio of all stocks) as the single factor. The market model implies a certain covariance matrix of stock returns that can be easily estimated from the data. Unfortunately, the market model is rejected by the data and therefore the ensuing covariance matrix estimator is not reliable. Intuitively, the market model imposes too much structure to be compatible with real data. The common solution, so far, has been to impose less structure buy building models with several factors. One approach is to use factors with economic interpretation (such as industry factors, P/E ratio, book-to-market, etc.) as is done by the financial services firm BARRA; e.g., see Kahn (1994). Another approach is to use statistical factors (such as principal components) without economic interpretation as is done by the firm APT; e.g., see Connor and Korajczyk (1992). Still, the exact nature and number of the factors to be included in a multi-factor model remains an open question.

This is why we propose a different approach to impose structure on the estimation of the covariance matrix of stock returns. Our idea is to take a weighted average of the sample covariance matrix and Sharpe's market-model covariance matrix, that is, to *shrink* the sample covariance matrix towards the market-model matrix. The idea of shrinking an unbiased but very variable estimator towards a biased estimator with little variation has a long and successful history in statistics, dating back to the seminal work of James and

Stein (1961). The intuition is that by properly combining an estimator with no bias but high variance and an estimator with high bias but small variance, one can obtain a new, improved estimator (in the mean squared error sense). The only remaining problem is to determine the *shrinkage intensity*, that this, the amount by which we should shrink the unbiased estimator towards the biased estimator. This is a problem which needs to be solved in a case-by-case analysis. In the paper, we develop a method to determine the optimal shrinkage intensity by a fully automatic procedure from the data. It should be noted that our methodology is very flexible, meaning it can be easily adapted to shrinkage targets different from the market-model covariance matrix as seen fit by the portfolio manager.

To see how well our method works in practice, we compare it to a number of existing covariance matrix estimators using real data from 1962 to 1995; the data consist of monthly returns of all the stocks contained in the NYSE and the AMEX stock exchanges. Using the well-known Markowitz (1952) portfolio selection algorithm, we consider the problems of constructing the minimum global variance portfolio as well as the minimum variance portfolio with an expected return of 20%. Starting in 1972, we use the last ten years of data to estimate the covariance matrix (by the different estimators included in the study) and the mean vector (by a unique method) of stock returns; then we construct the optimal portfolio and hold it for one year, keeping track of the resulting monthly returns; we repeat this procedure the next year until year 1994. This gives us a total of 23 years of monthly data, which allows us to estimate with high precision the portfolio variance corresponding to the various covariance matrix estimators. It turns out that our shrinkage estimator, for both portfolio selection problems, yields the portfolio with the smallest variance. The improvement over the other estimators, among them the market-model estimator, the industry-factor estimator, and the principal-components estimator, is significant both in statistical and economical terms.

In summary, we have proposed a new estimator for the covariance matrix of stock returns that is rooted in a statistical technique with a proven track record (namely, the shrinkage methodology), can be implemented easily, and appears to be superior to the commonly used estimators from a study using historical data. For details, the reader is referred to Ledoit and Wolf (2000).

References

Connor, G. and Korajczyk, R.A. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics* 15, 373-394.

Kahn, R. (1994). The E3 project. In Barra Newsletter, Summer 1994, page 11. (Available at http://www.barra.com/Research_Library/BarraPub/te3p-n.asp.)

Ledoit O. and Wolf M. (2000). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Working paper 00-77, Universidad Carlos III de Madrid, *Statistics and Econometrics Series.*

Markowitz, H. (1952). Portfolio selection. *Journal of Finance* **7**, 77-91.

Sharpe, W.F. (1963). A simplified model for portfolio analysis. *Management Science* **9**, 277-293.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (ed. J. Neyman), Vol. **1**, pages 361-379. UC Berkeley, Berkeley.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Problems in Inference after Model Selection

Hannes Leeb, Benedikt M. Pötscher
Department of Statistics, University of Vienna
Universitätsstr. 5, 1010 Vienna, Austria.
Hannes.Leeb@univie.ac.at, Benedikt.Poetscher@univie.ac.at

The traditional theory of parametric statistical inference is primarily concerned with the statistical properties of estimators or inference procedures, like tests or confidence sets, under the central assumption of an a priori given model. That is, it is assumed that the model is known to the researcher prior to the statistical analysis, except for the value of the true parameter vector. In practice, however, the specification of the model (choice of functional form, choice of regressors, number of lags, etc.) is often also determined only after the data have been observed, violating the central assumption of an a priori given parametric model. As a consequence, the actual statistical properties of estimators or inference procedures following such a data-driven model selection step are not described by the traditional theory which relies on an a priori given model; in fact, they may differ substantially from the properties predicted by traditional theory; cf. Pötscher (1991). Ignoring the additional uncertainty originating from the data-based model selection step and (inappropriately) applying traditional theory can hence result in very misleading conclusions.

Only recently, the distribution of parameter estimators computed after a data-driven model selection step, i.e., the distribution of what we call post-model-selection estimators, has been investigated. Sen (1979) obtained the large-sample distribution of a post-model-selection estimator in an iid maximum likelihood framework, when there are two competing models. In Pötscher (1991), the asymptotic properties of post-model-selection estimators (based on a sequence of tests) were studied in a rather general setting covering non-linear models, dependent processes, and more than two competing models. In particular, the asymptotic distribution of the post-model-selection estimator, both unconditional as well as conditional on having chosen a correct model (minimal or not), was derived. While constituting an important step towards understanding the statistical properties of post-model-selection estimators, these results appear to be limited in at least two ways:

- They do not provide information on the distribution of the post-model-selection estimator conditional on selecting an incorrect model. (Although, asymptotically, such models are never selected by any reasonable model selection procedure, the finite-sample probability of selecting an incorrect model can be substantial; see, e.g., Table III in Pötscher and Novak (1998).
- The convergence of the finite-sample distributions to the asymptotic distributions is not uniform over the parameter space (cf. Pötscher (1991)), which indicates potential problems with the accuracy of these approximations. (This is confirmed by a simulation study in Pötscher and Novak (1998); see also Kabaila (1995) and Pötscher (1995) for further discussions.)

The first part of this talk addresses these two issues. For simplicity, we restrict the discussion to linear regression problems. We begin by deriving the unconditional as well as the conditional finite-sample distribution of the post-model-selection estimator, which turns out to be quite complicated and difficult to interpret. Then we present approximations to the finite-sample distributions that are as simple and easy to

interpret as the asymptotic distributions obtained in Pötscher (1991), but at the same time are close to the finite-sample distributions uniformly with respect to the underlying parameters. As a by-product, we also obtain the asymptotic distribution conditional on choosing an incorrect model.

Having thus obtained satisfactory theoretical approximations to the distribution of the post-model-selection estimator, we next turn to the issue of obtaining `computable' approximations. Indeed, the large-sample distribution of the post-model-selection estimator as well as our uniform approximation still depend on unknown parameters and therefore have to be estimated. As described in Pötscher (1991), one can construct consistent estimators for, say, the finite-sample (unconditional) distribution of the post-model-selection estimator, even in the very general setting considered in that paper. This is, if the finite-sample cdf of the post-model-selection estimator is denoted by $G_{n,\theta}(t)$, where n is the sample size and θ is the unknown parameter, one can construct a consistent estimator $\bar{G}_n(t)$, i.e., one which satisfies

(1) $$P_{n,\theta}(\mid G_{n,\theta}(t) - \bar{G}_n(t) \mid > \varepsilon) \to 0$$

as $n \to \infty$ for each $\varepsilon>00$, where $P_{n,\theta}$ denotes the distribution of a sample of size n with the true parameter being θ. In the second part of this talk we describe a fundamental flaw of such consistent estimators: We show that the finite-sample quality of any such consistent estimator $\bar{G}_n(t)$ varies heavily with the unknown parameter θ. In fact, under very mild conditions, we show that

(2) $$\sup_\theta P_{n,\theta}(\mid G_{n,\theta}(t) - \bar{G}_n(t) \mid > \varepsilon) \to 1$$

as $n\to\infty$ for some $\varepsilon>00$. As the true parameter is unknown, this implies that (1) does not guarantee a small estimation error at any given sample size (however large). In particular, (2) implies that no uniformly consistent estimator for the cdf of the post-model-selection estimator can exist. Results similar to (1) and (2) also hold for estimators for the conditional distribution of the post-model-selection estimator, conditional on having chosen a fixed model. Extensions to other model selection procedures are also discussed.

In light of Hajek's famous quote: ``Especially misinformative can be those limit results that are not uniform. Then the limit may exhibit features that are not even approximately true for any finite n" (cf. Hajek (1970)), the lower bound result (2) raises fundamental questions concerning inference after model selection, most of which are still open.

References

Hajek, J. (1970): ``Limiting properties of likelihoods and inference," in Foundations of Statistical Inference: by V.P. Godambe, and D.A. Sprott, pp. 142-161, Toronto, Canada. Holt, Rinehard and Winston.

Kabaila, P. (1995): ``The effect of model selection on confidence regions and prediction regions," *Econometric Theory*, **11**, 537-549.

Pötscher, B.M. (1991): ``Effects of model selection on inference," *Econometric Theory*, **7**, 163-185.

Pötscher, B.M. (1995): ``Comment on `The effect of model selection on confidence regions and prediction regions'," *Econometric Theory*, **11**, 550-559.

Pötscher, B.M., and A.J. Novak (1998): ``The distribution of estimators after model selection: large and small sample results," *J. Statist. Comput. Simul.*, **60**, 19-56.

Sen, P.K. (1979): ``Asymptotic properties of maximum likelihood estimators based on conditional specification," *Ann. Stat.*, **7**, 1019-1033.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Statistical Analysis of Ozone Variability: a Case Study in Oporto

Solange Mendonça Leite
University of Lisbon, Geophysical Center
Rua da Escola Politécnica, nº 58, 1250-102 Lisbon, Portugal
solange@utad.pt

Fernando de Pablo Dávila
University of Salamanca, Department of Atmospheric Physics
Plaza de la Merced s/n, 37008 Salamanca, Spain
fpd123@gugu.usal.es

Clemente Tomás Sánchez
University of Salamanca, Department of Atmospheric Physics
Plaza de la Merced s/n, 37008 Salamanca, Spain
cts50@gugu.usal.es

Ozone has been known as a constituent of the atmosphere since the middle of the nineteenth century and its presence was then a cause of much interest to professional people with a basic knowledge of science (Bojkov, 1986). Using simple inorganic chemical techniques for analysis, many studies of its behavior were made but few deductions were ever drawn from the large data base that was assembled.

Much of the increase in knowledge of atmospheric ozone in the early part of the twentieth century had to do with the ozone layer. Its preponderance in the stratosphere was clearly recognized, leading to the suggestion that ozone would make an excellent tracer of stratospheric air.

Against this background it was perhaps surprising to find that ozone could be produced in the troposphere, leading *in extremis* to the infamous Los Angeles smog. Since that observation was made over 40 years ago, ozone production in the lower atmosphere has been observed world-wide as a predominant type of regional air pollution.

There is good experimental evidence that the tropospheric ozone concentrations in the Northern Hemisphere is increasing (Bojkov, 1986; Volz and Kley, 1987). Plausible reasons can be advanced that this is associated with increased emissions of precursor molecules, particularly nitrogen oxides, from anthropogenic sources.

But the field of atmospheric science is at least as large as the field of statistics. Nowadays, in the literature of atmospheric sciences are many examples of the application of statistical methods. The involvement of statisticians in atmospheric science, and of atmospheric physicists and chemistries in statistical science, which has certainly not been negligible, is growing. This can only improve the standard of statistics in the field of atmospheric sciences.

Several reviews of ozone trends have been recently formulated (Logan, 1985; Bojkov, 1986). Probably the longest recent record of data on ozone, which is still actively being collected, is that from the network of Germany meteorological service. This was commissioned in 1952 and has been maintained until the present (Feister and Warmbt, 1987).

Another example of the application of statistical analysis to characterize ground level ozone is reported to Castilla-León (Spain), where a network of 26 urban, suburban and rural stations have been providing data to analyze atmospheric contamination in a regional scale and its temporal evolution (Alvarez *et al*., 2000).

Ozone concentrations at ground level has been collected in Portugal only from the first half of last decade, in Lisbon and Oporto (Environmental Governmental Services), both situated in the Atlantic coast. We have been analyzing this ozone data by applying the adequate statistical methodology.

In this presentation some results of the statistical analysis of ozone data collected at six measurement stations in Oporto (Figure 1) are shown. The common period of data record is used in this presentation, that is from January 1, 1999 to December 31, 1999. Annual evolution of maximum and mean monthly values are analyzed, as well as the seasonal variation of the daily evolution. Particularly, winter (December, January and February) and summer (June, July and August) ozone concentration data are compared. Finally, seasonal variation of weekly evolution of ozone concentration is analyzed.

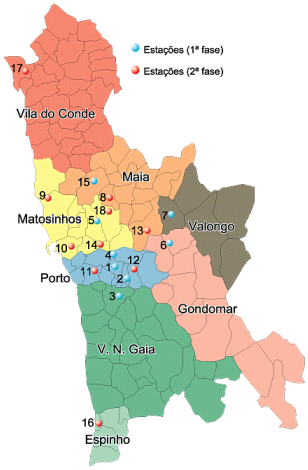


Figure 1. Geographical location of the six measurement stations in Oporto (source: Direcção Regional do Ambiente-Norte).

References

Alvarez, E., Pablo, F., Tomás, C. and Leite, S.M. (2000). Statistical predictive models of hourly mean and daily maximum concentrations of ozone in Castilla-León (Spain), TIES/SPRUCE 2000, University of Sheffield, UK.

Bojkov, R. D. (1986). Surface ozone during the second half of the nineteenth century, *J. Clim. Appl. Meteor.* **25**, 343-352.

Feister, U. and Warmbt, W. (1987). Long-term measurements of surface ozone in the German Democratic Republic, *J. Atmos. Chem.* **5**, 1-21.

Logan, J. A. (1985). Tropospheric ozone: seasonal behavior, trends and anthropogenic influence, *J. Geophys. Res.* **90**, 463-482.

Volz, A. and Kley, D. (1987). Ozone measurements in the 19th century: an evaluation of the Montsouris series, Nature.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Bayesian Analysis of the Skew-Normal Distribution

Brunero Liseo
Università di Roma "La Sapienza"
Via del Castro Laurenziano, 9 I-00161, Roma, Italy
Brunero.Liseo@uniroma1.it

Nicola Loperfido
Università di Urbino
nicola@econ.uniurb.it

1. Introduction

The Skew Normal (SN, hereafter) class of densities has been introduced by Azzalini (1985) and recently generalised to the multivariate case (Azzalini and Dalla Valle, 1996; Azzalini and Capitanio, 1999). This class of densities extends the Normal model by allowing a shape parameter to account for skewness. The density function of the generic element of the class is

(1) $$f(x;\lambda,\mu,\sigma) = \frac{2}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\lambda\frac{x-\mu}{\sigma}\right)$$

where φ and Φ represents the pdf and the cdf of the standard Normal density, respectively, and λ is a real parameter. Positive (negative) values of λ indicate positive (negative) skewness; when $\lambda=0$, one gets back to the Normal density.

The SN class enjoys remarkable properties in terms of mathematical tractability and it proved itself quite useful in modelling real data sets (see Azzalini and Capitanio, 1999). The SN class of densities plays a role in a Bayesian context too: it has been considered as a sampling model in Liseo (1990) and as a class of prior densities in O'Hagan and Leonard (1976).

Despite its nice properties, problems arises in the estimation of the parameters. For simplicity consider the standard case ($\sigma=1$, $\mu=0$). From (1) one sees that the likelihood function associated to a *n*-dimensional sample, is the product of *n* cdf's of the standard normal density: if all the observed x_i's are positive (negative) then the likelihood function will be monotonically increasing (decreasing) and the maximum likelihood estimate for λ will be (minus) infinite! This is maybe the worst case: even with positive and negative observations, the behaviour of the MLE is not satisfactory. In the general three-parameters case things can be even more difficult because the Fisher information matrix is singular as $\lambda \to 0$. Then "an alternative estimation method is called for…" (Azzalini and Capitanio, 1999).

2. A Bayesian Approach

In this paper we propose a fully Bayesian approach to the estimation of the parameters of the SN class of densities, when it is used as a sampling model.

Here we list the main results of the paper

Standard case ($\sigma=1$, $\mu=0$).

We calculate the Jeffreys' prior for λ and we prove that it is a proper density with tails of order $O(n^{-3/2})$. This is a quite unusual fact because noninformative priors for real parameters are usually improper. In the SN model, the propriety of the Jeffreys' prior seems to compensate the unusual behaviour of the likelihood function.

This enables us to use a simple Metropolis-Hastings algorithm to obtain a sample from the posterior distribution of λ. Simulation results indicates that the behaviour of the Bayes estimate (i.e. the posterior mean of λ) is better than that of MLE, even from a frequentist viewpoint.

The propriety of the Jeffreys' prior also enables us to obtain a simulation based approximation of the Bayes factor for testing the null hypothesis $H_0: l = l_0$ against the alternative $H_1: l \neq l_0$.

General Case

We derive the reference prior (Berger and Bernardo, 1992) for the three-parameter case. This is given by the product of the usual noninformative prior for the location-scale parameters ($\pi_R(\mu,\sigma)=1/\sigma$) times a complicated (but marginally proper!) density for λ.

We obtain a nice form of the integrated likelihood for λ, after that the location and scale parameters are eliminated with respect a Normal-Gamma type prior (which includes, as a special case, $\pi_R(\mu,\sigma)$).

We compare the frequentist behaviour of the Bayes estimates with the ML estimates, as obtained in Azzalini and Capitanio (1999)

References

Berger, J.O. and Bernardo, J.M. (1992). On the development of the reference prior method. *Bayesian Statistics IV,* (J.M. Bernardo et al. Eds.) 35-60, OUP, Oxford.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.* **12**, 171-178

Azzalini, A. and Dalla Valle, A.(1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715-726.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew-normal distribution. *J. Roy. Statist. Soc., B* **61**, 579-602.

Liseo, B. (1990). La classe delle densità normali sghembe: aspetti inferenziali da un punto di vista bayesiano. *Statistica* **50**, 59-70.

O'Hagan, A. and Leonard, T. (1976). Bayes estimation subject to uncertainty about parameter constraints. *Biometrika* **63**, 201-202.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Generalizations of the Thurstone-Mosteller Model

Igor Litvine
University of Port Elizabeth, Department of Mathematical Statistics
University Way, Summerstrand, Port Elizabeth, Republic of South Africa
msainl@upe.ac.za

David Friskin
University of Port Elizabeth, Department of Mathematical Statistics
University Way, Summerstrand, Port Elizabeth, Republic of South Africa
msadgf@upe.ac.za

1. Thurstone-Mosteller Model

In 1927 L. Thurstone published the following description of a model for paired comparisons and suggested a corresponding estimation procedure:

1. There is a set of stimuli that can be located on a subjective continuum (a sensation scale, usually not having a measurable physical characteristic).
2. Each stimulus when presented to an individual gives rise to a sensation in the individual.
3. The distribution of sensations from a particular stimulus for a population of individuals is normal.
4. Stimuli are presented in pairs to an individual, thus giving rise to a sensation for each stimulus. The individual compares these sensations and reports which is greater.
5. The standard deviations of the stimuli are equal and correlations are zero.
6. Our task is to space the stimuli (the sensation means), except for a linear transformation

The model also assumes that the data is balanced, i.e. each pair is compared the same number of times. In 1951 F. Mosteller suggested a generalization of the above model for correlated stimuli (equal correlations) and the model is now known now as the Thurstone-Mosteller model. Since then various other models for paired comparisons were developed (e.g. Bradley-Terry model, etc., see David (1988) and references).

However the Thurstone-Mosteller model remains one of the most popular models for various applications. Numerous extensions of this model include cases of order effects, possibility of ties, within pair effects, unequal correlations, unequal number of comparisons, etc.

This paper suggests two new generalizations of the Thurstone-Mosteller model.

2. Method of Payoff Functions for Incomplete Data

This method was suggested by Chebotarev (1989) for the row sum method. The idea of the method is to replace the actual and missing scores with so-called payoff functions $f(x_i, x_j, r_{ij})$ of unknown weights x_i and x_j and known scores r_{ij}.

Eventually the weights are found from a system of linear equations

(1) $$x_i = \sum_j (r_{ij} + e(x_j - x_i + r_{ij}mn))$$

Where n is the number of objects (stimuli) compared, m is the number of observations per pair (if the two were compared), e is a positive constant and summation is taken over such j that r_{ij} is defined. While the row sum method uses actual observed scores r_{ij}, in case of the Thurstone-Mosteller model we find them from the equation:

(2) $$p_{ij} = \Phi(r_{ij})$$

where Ö is the CDF of Standard Normal Distribution and p_{ij} is the observed relative frequency of event that stimulus i was preferred to stimulus j.

3. Bayesian Solution to the Thurstone-Mosteller Model

This approach allows us easily to accommodate for cases of unbalanced data, missing comparisons, etc. As an example we suggest the following theorem.

Theorem. Let prior distributions of the weights be Standard Normal Distributions for all stimuli. Then the Bayesian estimates of the unknown weights x_i may be found as solutions to the following system of the linear equations:

(3) $$x_i - x_j = \frac{r_{ij}^k}{s^2 + 1}$$

where ó is the common standard deviation of the stimuli.

It should be noted that the knowledge of the actual ó is not necessary to find the estimates. We can use instead any positive number, because "our task is to space the stimuli (the sensation means), except for a linear transformation" (see item 6 of the model description.

References

Thurstone, L.L. (1927). Phychophysical Analysis, *Ame. Journal of Phychology*. **38**, 363.

Mosteller, F. (1951). Remarks on the Method of Paired Comparisons: I. The Least Squares Solution Assuming Equal Standard Deviations and Equal Correlations, *Phychometrica*. **16**, 3-9.

David, H. A. (1988). The Method of Paired Comparisons. *Oxford University Press. New York.*

Chebotarev, P. Yu. (1989). Generalization of the Row Sum Method for Incomplete Paired Comparisons, *Automatika i Telemekhanika.* **8(50)**, 125-137.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Bootstrapping the Chambers-Dunstan Estimate of a Finite Population Distribution Function

M.J. Lombardía, W. González-Manteiga, J.M. Prada Sánchez
Santiago de Compostela University, Dpt. of Statistics and Operation Research
Faculty of Mathematics. Campus South, Santiago de Compostela, Spain
majose@zmat.usc.es, wences@zmat.usc.es, prada@zmat.usc.es

1. The Chambers-Dunstan Estimator and Bootstrap Schemes

It is of interest to estimate the distribution of a random variable, Y, defined for a finite population. Let P be the set of integers $\{1,\ldots, N\}$; S an n-element subset of P and $P\backslash S$ the complement of S in P. We consider a finite population $\boldsymbol{P} = \{(Y_k, x_k)\}_{k\in P,}$ where the values x_k of an auxiliary variable X are known for all population elements and the random variable Y is related to X by the model

(1) $$x: \quad Y_k = a + bx_k + e_k$$

where a and b are unknown parameters and e_k ($k \in P$) are independent and identically distributed random variables with zero mean. The Y_k are only known for a sample $\boldsymbol{S} = \{(Y_i,x_i)\}_{i\in S}$, which is taken without replacement from $\boldsymbol{P}$.

The objective is to estimate the finite population distribution function of Y,

(2) $$F_N(t) = N^{-1}\sum_{k\in P} I\left(Y_k \le t\right) = N^{-1}\left[\sum_{i\in S} I\left(Y_i \le t\right) + \sum_{j\in P-S} I\left(Y_j \le t\right)\right] \quad t\in \boldsymbol{R},$$

where I(ζ) is the indicator function of the event ζ. The x-based estimator proposed by Chambers and Dunstan (1986) is

(3) $$\hat{F}(t) = N^{-1}\left\{\sum_{i\in S} I\left(Y_i \le t\right) + \sum_{j\in P\backslash S} \hat{G}\left(t - \hat{a} - \hat{b}x_j\right)\right\}$$

where $\hat{a}$, $\hat{b}$ are the $\boldsymbol{S}$-based least-squares estimates of a and b respectively, and $\hat{G}$ is the empirical distribution of the residuals.

Given an estimate $\hat{G}_\bullet(u)$ of the distribution G(u) of e (see below), a bootstrap finite population $\boldsymbol{P}^* = \{(Y_k^*, x_k)\}_{k\in P}$ conforming to the model

(4) $$x^*: \quad Y_k^* = \hat{a} + \hat{b}x_k + e_k^*$$

can be generated by sampling $\hat{G}_\bullet(u)$ to obtain the e_k^* ($k\in P$). The distribution

(5) $$F_{N,\bullet}^*(t) = N^{-1}\sum_{k\in P} I\left(Y_k^* \le t\right) = N^{-1}\left[\sum_{i\in S^*} I\left(Y_k^* \le t\right) + \sum_{j\in P-S^*} I\left(Y_j^* \le t\right)\right]$$

of the variable Y^* can be estimated from an n-member sample $\boldsymbol{S}^* = \left\{\left(Y_i^*,\ x_i\right)\right\}_{i\in S^*}$ of $\boldsymbol{P}^*$ using the corresponding Chambers-Dunstan estimate:

(6) $$\hat{F}_\bullet^*(t) = N^{-1}\left\{\sum_{i\in S^*} I\left(Y_i^* \le t\right) + \sum_{j\in P\backslash S^*} \hat{G}^*\left(t - \hat{a}^* - \hat{b}^* x_j\right)\right\}$$

where $\hat{a}^*$, $\hat{b}^*$ are the $\boldsymbol{S}^*$-based least squares estimates of $\hat{a}$ and $\hat{b}$ respectively, and where $\hat{G}^*(u)$ is the empirical distribution of the residuals $\hat{e}_i^* = Y_i^* - \hat{a}^* - \hat{b}^* x_i$.

In this research we used two different estimators $\hat{G}_{\bullet}$ of G. One was obtained by recentring the empirical distribution of the errors $\hat{e}_i$ on their mean $\bar{\hat{e}}$, thus:

(7) $$\hat{G}_e(u) = n^{-1}\sum_{i\in S} I\left(\hat{e}_i - \bar{\hat{e}} \le u\right).$$

The other was a smoothed version:

(8) $$\hat{G}_h(u) = n^{-1}\sum_{i\in S} K\left(\frac{u-\left(\hat{e}_i - \bar{\hat{e}}\right)}{h}\right).$$

The next theorems show that the smoothed bootstrap estimator $\hat{F}_h^*$ is consistent. Some regularity conditions are necessary:

Theorem (Chambers, Dorfman and Hall, 1992)

$$MSE\left\{\hat{F}(t) - F_N(t)\right\} = n^{-1}(1-p)^2\left\{ t^{-2} s^2 \left(\int (x-m) g(t-a-bx) d(x) dx\right)^2 \right. +$$

$$+\int\int G\left\{ (t-a-bx)\wedge(t-a-by)\right\} d(x) d(y) dx dy \; - \left(\int G(t-a-bx) d(x) dx\right)^2 \Big\} +$$

$$+ \; N^{-1}(1-p) \int \left\{ G(t-a-bx) - G(t-a-bx)^2 \right\} d(x)dx + o(n^{-1}).$$

Theorem

$$MSE_*\left\{\hat{F}_h^*(t) - F_{N,h}^*(t)\right\} = n^{-1}(1-p)^2\left\{ t^{-2} \hat{s}^2 \left(\int (x-m) \hat{g}_h\left(t-\hat{a}-\hat{b}x\right) d(x) dx\right)^2 \right. +$$

$$+\int\int \hat{G}_h\left\{ \left(t-\hat{a}-\hat{b}x\right)\wedge\left(t-\hat{a}-\hat{b}y\right)\right\} d(x) d(y) dx dy - \left(\int \hat{G}_h\left(t-\hat{a}-\hat{b}x\right) d(x) dx\right)^2 \Big\} +$$

$$+ \; N^{-1}(1-p) \int \left\{ \hat{G}_h\left(t-\hat{a}-\hat{b}x\right) - \hat{G}_h\left(t-\hat{a}-\hat{b}x\right)^2 \right\} d(x)dx + o_p(n^{-1}).$$

2. Simulation Study

A big simulation study was made. We have not distinguished between both bootstraps since they both gave identical results within the level of precision used. In general, the behaviour of the $\hat{MSE}_*$ imitates $\tilde{MSE}$ at all points regardless of error distribution, population size or sampling fraction (n/N). As N and n increase, the discrepancy between $\hat{MSE}_*$ and $\tilde{MSE}$ decreases, reaching zero at several points within this level of precision, and remaining as $\hat{MSE}_*$ in the majority of cases - slightly lower than $\tilde{MSE}$.

References

Chambers, R.L., Dorfman, A.H., Hall P. (1992). Properties of estimators of the finite population distribution function. *Biometrika*, **79**, 3, pp577-582.

Chambers, R.L., Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, **73**, 3, pp597-604.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Formal Relationships Between Distribution Functions

Fernando López-Blázquez, Teresa Gómez-Gómez, Begoña Salamanca-Miño, David

Universidad de Sevilla, Departmento de Estadística e Investigación Operativa
Tarfia, s/n, 41012, Sevilla, Spain
lopez@cica.es

1. Introduction

We propose a formal method for obtaining a relationship between two arbitrary distribution functions. The method can be extended also to density functions an to probability mass functions in the discrete case.

Given two arbitrary distribution functions F and G we propose the following formal approximation

$$\textbf{(1)} \qquad F(x) \approx \sum_{k\geq 0} \left\{ \int_{(-\infty,x]} y_k(s)dG(s) \right\} E_F\left[y_k(X)\right],$$

where X is a random variable having the distribution function F and $\{y_k\}$ is an orthonormal system with respect to the measure dG(s). The symbol '$\approx$' means here formal approximation. The convergence and the sense in which the formal series in the right hand side of (1) converges to the left hand side of (1) should be stated in each particular case.

2. Applications

As an application of (1), we give the following exact formulas relating:

An arbitrary binomial distribution and an arbitrary poisson distribution:

$$\textbf{(2)} \qquad b(x;N,p) = p(x;l)\sum_{k=0}^{\infty} \frac{C_k^{(l)}(x)}{k!(l/p)^k} C_k^{(l/p)}(N)$$

An arbitrary hypergeometric distribution and an arbitrary binomial distribution:

$$\textbf{(3)} \qquad h(x,N,N_1,n) = b(x,N,p)\sum_{m=0}^{N_1} \binom{N}{m} (pq)^{-m} k_m(x,N,p)k_m(n,N,p)$$

where b(x; N, p), p(x; l) and h(x; N, N_1 ,n) denote the pmf of binomial, poisson and hypergeometric distributions respectively; C and k denote, respectively, the Charlier and Meixner polynomials of the first kind, see Chihara (1978)

It is possible to obtain some of the classical approximations and some new ones from (2) and (3). For instance, if in (2) we put l =Np and we use only one or two terms of the series in the right-hand side we obtain the classical approximation of the binomial to the poisson distribution. If we use three terms, we obtain the Kolmogorov approximation. With a similar argument, in (3) if we put and truncate adequately the series in the right hand side of (3), we obtain the classical binomial approximation to the hypergeometric distribution, (when only one term is used). Improvements in the approximations can be obtained if more terms are used, see López-Blázquez and

Some other applications of our results include an exact relationship between the negative binomial and the negative hypergeometric distributions, Gram-Charlier type A and B expansions, etc.

References

Chihara, T. (1978). *An introduction to orthogonal polynomials* Gordon & Breach, New York

Johnson, N. L., Kotz, S. And Kemp, A. (1992). *Univariate discrete distributions*. John Wiley and Sons, New York.

López-Blázquez, F. and Salamanca-Miño, B. (2001), Binomial approximation to hypergeometric probabilities, *J. Stat. Plann, Inference*, **87**, 21-29

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Distribution of the Sum of Weighted Central Chi-Square Variables

Fernando López-Blázquez
Universidad de Sevilla, Departmento de Estadística e Investigación Operativa
Tarfia, s/n, 41012, Sevilla, Spain
lopez@cica.es

Antonia Castaño-Martínez
Universidad de Cádiz, Departmento de Estadística e Investigación Operativa
Porvera,54, Jerez de la Frontera (Cádiz), Spain
antonia.castano@uca.es

1. Introduction

We consider $Q_n = \sum_{i=1}^{n} a_i X_i$ where a_i are known positive constants and X_i are independent chi-square variables with n_i degrees of freedom respectively. Our aim is to obtain the density and distribution functions of Q_n. We also derive bounds on the truncation error in the given expansions.

The method that we present is based on the inverse Laplace transform. The method of inversion that we propose in section 2 is based on the property of uniqueness of minimum variance unbiased estimators (MVUE) in the gamma distribution. Then, in section 3, we apply this method for the obtention of the density and distribution functions of Q_n.

2. The Inversion of Laplace Transforms

Let $h(l)$ be a parametric function MVU-estimable in a distribution following a $Ga(p,l)$ distribution with $p>0$ known and $l>0$ the unknown parameter, i.e. there exits a function T such that: $E_l\left[T^2(Y)\right]<\infty$ and $E_l\left[(T(Y)\right]=h(l)$, for all $l>0$.

In such case, T is the minimum variance unbiased estimator, MVUE, of $h(l)$. The set of all the MVU-estimable functions will be denoted by U.

From the uniqueness of the UMVU estimators, we can obtain the following result which gives an expression for the inversion of Laplace transforms:

Theorem Let $G(l)$, $l>0$, be a function such that for certain $p>0$, $h(l)=l^{p}G(l)$ is MVU-estimable function, then:

(1)
$$L^{-1}\left(G(l)\right)(x)=\frac{x^{p-1}}{\Gamma(p)}\sum_{j=0}^{\infty}\frac{(-m_0)^p\, g^{j)}(m_0)}{(p)_j}L_j^{(p-1)}\left(\frac{px}{m_0}\right)$$

for any $m_0>0$, with $g(m)=h(p/m)$, and with $L_j^{(p-1)}$ denoting the j-th generalized Laguerre polynomial.

Note that the choice of m_0 is arbitrary, so adecuate choices of the parameters in (1) may yield formulas computationally efficient.

3. The Distribution of Q_n

As an application of the previous theorem, we will obtain the density and distribution functions of Q_n.

Let *f* and *F* be the density and distribution function of Q_n respectively, then:

(2) $$f(x) = \frac{e^{-\frac{x}{2b}}}{(2b)^{n/2}} \frac{x^{n/2-1}}{\Gamma(n/2)} \sum_{k\geq 0} \frac{k! c_k}{(n/2)_k} L_k^{(n/2-1)}\left(\frac{nx}{4bm_0}\right), \ \forall m_0 > 0$$

and

(3) $$F(x) = \frac{e^{-\frac{x}{2b}}}{(2b)^{n/2+1}} \frac{x^{n/2}}{\Gamma(n/2+1)} \sum_{k\geq 0} \frac{k! d_k}{(n/2+1)_k} L_k^{(n/2)}\left(\frac{(n+2)x}{4bm_0}\right), \ \forall m_0 > 0$$

where c_k and d_k are some constants that can be easily obtained by recurrence formulas.

As our objective is to implement these formulas in a computer, we study the errors produced when the infinite series given in (2) and (3) are truncanted. We also compare our results with those given by Kotz et al. (1967).

References

Chihara, T. (1978). *An introduction to orthogonal polinomials* Gordon & Breach, New York

Davies, R. B. (1980). The distribution of a linear combination of c^2 random variables. *Applied Statistics*, **29**, 323-333.

Davies, A.W. (1977). A differential equation approach to linear combinations of independent chi-squares. *J. A. S. A.*, **72**, 212-214.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, **48**, 419-426.

Johnson, N. L., Kotz, S. (1968). Tables of distributions of positive definite quadratic forms in central normal variables. *Sankhya Ser B,* **30**, 303-314.

Johnson, N. L., Kotz, S., Balakrishnan, N. (1994*). Continuous Univariate distributions, 2nd ed.*, **1**, John Wiley & Sons, New York.

Kotz, S., Johnson, N. L., Boyd, D. W. (1967). Series representations of distributions of quadratic forms in normal variables I. Central case. *Ann. Math. Statist.* **38**, 823-837.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist*., **10**, 65-80.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Kernel-Type Estimators for the Extreme Value Index

Hendrik P. Lopuhaä
Delft University of Technolgy
Faculty of Information Technology and Systems
Department of Mathematics
Mekelweg 4, 2628 CD
Delft, The Netherlands
h.p.lopuhaa@its.tudelft.nl

P.Groeneboom
Delft University of Technolgy
Faculty of Information Technology and Systems
Department of Mathematics
Mekelweg 4, 2628 CD
Delft, The Netherlands
p.groeneboom@its.tudelft.nl

P.P. de Wolf
Statistics Netherlands(CBS)
Methods and Informatics Department
Division of Technology and Facilities
PO Box 4000, 2270 JM
Voorburg, The Netherlands
PWOF@cbs.nl

A large part of the theory of extreme value index estimation is developed for positive extreme value indices. The best known estimator for that case is the Hill estimator (see [3]).This estimator can be considered to be either a moment estimator or a (quasi) maximum likelihood estimator and was generalized to a kernel-type estimator, still only valid for positive extreme value indices.

The Hill estimator has been extended to a moment-type estimator valid for all extreme value indices (see [2]). Also the quasi-maximum likelihood estimators (see [4]) based on the generalized Pareto distribution have been given for a restricted region of negative extreme value indices. Both the moment-type approach and the likelihood approach lead to estimators that are based on the k largest observations. A major drawback of both approaches is the discrete character of the behavior of these estimators: adding a single large order statistic in the calculation of the estimator, i.e. increasing k by one, can change the actual value of the estimate considerably. Plotting these estimators as a function of the order statistics used, therefore often results in a zigzag figure. In [1], the Hill estimator is smoothed by a kernel in order to obtain a more stable figure. Unfortunately, this kernel type estimator is still valid only for positive extreme value indices.

In this talk we present kernel-type estimators valid for *all* real extreme value indices and compare their performance with the (generalized) moment estimator and (quasi) maximum likelihood estimator. It should be emphasized that our estimator is not a smoothed version of the moment estimator introduced in [2], but is based on the so-called von Mises conditions. The resulting estimator is shown to be consistent

under the single condition that the underlying distribution function is in the domain of attraction of an extreme value distribution. Under additional assumptions on the underlying distribution, asymptotic normality will be derived and sufficient conditions are provided under which the asymptotic bias vanishes.

References

[1] Csörgó, S., Deheuvels, P. and Mason, D.(1985) Kernel estimates of the tail index of a distribution. *Ann. Statist*. **13** 1050-1077.

[2] 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.

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

[4] Smith, R.L.(1987) Estimating tails of probability distributions. *Ann. Statist*. **15** 1174-1207.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Quantile Regression Analysis of Transition Data

José António Machado
Universidade Nova de Lisboa, Faculdade de Economia
TravessEstevão Pinto-Campolide, Lisboa, Portugal
jafm@fe.unl.pt

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

Quantile regression constitutes a natural and flexible framework for the analysis of duration data in general and unemployment duration in particular (Koenker and Basset, 1978). For instance, comparison of the quantile regressions for lower and upper tails of the duration distribution may shed important insights on the different determinants of short or long-term unemployment. Using quantile regression techniques, we estimate conditional quantile functions of US unemployment duration and the implied hazard functions. One of the most interesting conclusions pertains the role of ``advanced notice of firing'', which was found to impact short durations---low quantiles---but not relatively long durations. Overall, the results provide clear indications of the interest of quantile regression to the analysis of duration data The number of applications of quantile regression techniques has greatly increased in recent years. Labor economics has been one of the most popular fields for applications, but attention has been almost exclusively devoted to the study of wage equations (see, for example and with no claim to being exhaustive, Buchinsky (1994), Chamberlain (1994), Fitzenberger (1997) and Machado and Mata (2001)). Yet, quantiles seem quite appropriate to analyze unemployment duration for, at least, two main reasons. First, they provide a natural way of characterizing important concepts as short or long-term unemployment by focusing on the relevant tails of the duration distribution. Consequently, comparison of the quantile regressions for the 20th and for the 80th percentiles (say) may shed important insights on the different determinants of short or long-term unemployment. From a methodological vintage point, it is worth noticing that quantile regression enable the performance of the aforementioned comparisons within a unified and flexible framework. Moreover, quantile regression, as the seminal work of Powell (1984) reveals, is particularly well equipped to perform consistent inferences with censored data, a typical situation in duration studies.

The present paper has two chief aims. The first one is to explore the potential of models for conditional quantile functions as a tool to analyze duration data. Second, we wish to illustrated the approach with a well known and important data set--the U.S. ``Displaced Workers Survey''--in order to highlight the potential information gains from using quantile regression in duration analysis (Addison and Portugal, 1987). The works by Horowitz and Neuman (1987, 1989) constitute early attempts of using quantile estimates for unemployment duration. However, somehow, they do not appear to have made their way into the mainstream econometric analysis of duration. Be as it may, the emphasis there was the consistent estimation of a parameter vector in presence of censoring rather than exploitiong the full potential of quantile regression as a tool to the statistical analysis of conditional distributions.

In this paper it is argued that quantile regression analysis offers a fruitful semi-parametric alternative to study transition data. On one hand, the censored quantile

regression estimator enables the accommodation of incomplete duration data. And on the other hand, quantile regression renders itself naturally to estimation of Accelerated Failure Time models without imposing any distributional assumptions. Given the decreasingly costs of computer intensive statistical methods such as these, it is puzzling to realize that just a few empirical studies have applied quantile regression models to duration data.

Apart from being a distribution-free model, there are other advantages accruing from using quantile regression models. First, it is flexible approach in the sense that it allows for the covariates to have different impacts at diferent points of the distribution. Second, the estimators of the regression coefficients are robust to the presence of unobserved individual heterogeneity. Third, the estimators are resilient to misspecification of the functional form. And fourth, in comparison with conventional models, the quantile regression approach provides a much more complete characterization of the duration distribution.

It is arguable that a reason why researchers shy away from using the quantile regression estimator is its uneasiness in dealing with standard survival analysis concepts. It is shown, however, that it is straightforward to obtain typical survival outputs from quantile regression estimates (e.g., hazard and survival functions, mean residual life, conditional mean duration, etc.).

Finally, in some instances, the quantile regression approach offers a natural and intuitive way to deal with some economic concepts. This is clearly the case of earnings inequality. It is, in our view, also the case of unemployment duration. In particular, the notions of short and long-term unemployment can be given an unambiguous empirical content. In the empirical illustration with US unemployment duration, it was shown that some covariates impact differently at distinct regions of the unemployment duration distribution. The usefulness of the quantile regression approach is suggested by the conclusion that some variables impact solely at short durations (e.g., advance notice, schooling, and previous wage), other variables fade significantly over the course of the spell of unemployment (plant closing), while the effect of other variables remain constant across the board (gender and race). Those varying effects would be ignored if conventional duration models were to be employed.

References

Addison, J. and Portugal, P. (1987). On the distributional shape of unemployment duration. *Review of Economics and Statistics*, **68**, 520-526.

Buchinsky, M. (1994). Changes in the U.S. wage structure 1963-1987: Application of quantile regression. *Econometrica*, **62**, 405-458.

Chamberlain, G. (1994). Quantile regression, censoring and the structure of wages. in C. Sims, eds. *Advances in Econometrics 6th World Congress* vol. **1**, Cambridge Univ. Press.

Fitzenberger, B. (1997). A guide to censored quantile regression. in G. Maddala and C. Rao, eds. *Handbook of Statistics*, vol. **15**, Elsevier Science B. V.

Horowitz, J. and G. Neumann (1987). Semiparametric estimation of employment duration models. *Econometric Reviews*, **6**,5-40.

Horowitz, J. and G. Neumann (1989). Specification testing in censored regression models: parametric and semiparametric methods.*J. of Applied Econometrics*, **4**, S61-S86.

Koenker, R. and G.S. Bassett Jr. (1978). Regression Quantiles. *Econometrica*, **46**, 33-50.

Koenker, R. and J. Machado (1999). Goodness of fit and related inference processes for quantile regresssion. *J. of the American Statistical Association*, **94**.

Machado, J. and J. Mata (2001). Earning functions in Portugal 1982-1994: evidence from quantile regressions. *Empirical Economics*, forthcoming.

Powell, J. (1984). Least absolute deviations for the censored regression model. *J. of Econometrics*, **25,** 303-325.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

G-Minimax Sequential Estimation for Markov-Additive Processes

Ryszard Magiera
Wroclaw University of Technology, Institute of Mathematics
Wybrzeze Wyspianskiego 27, PL-50-370 Wroclaw, Poland
magiera@im.pwr.wroc.pl

The problem of estimating unknown parameters of a Markov-additive process from data observed up to a random stopping time is considered in the case when the set of prior distributions of the parameters is restricted.

Let $(A(t), X(t)), t \geq 0,$ (the time parameter t is continuous) be a Markov-additive process (in accordance with the definition of Pacheco and Prabhu (1995)) with the state space $R \times I$, where $I = \{1,2,...,m\}$. It is assumed that the conditional distribution of $A(t) - A(s)$, given $X(u) = i$ for all $u \in [s,t]$, is given by the density

$$\exp[v_i x - f_i(v_i)(t-s)]$$

with respect to a s-finite measure which may depend on the state i in general, and v_i is a real parameter, $v_i \in V_i \subset R$. This means that the sojourn time distributions belong to one-dimensional exponential families. Let $(l_{i,j})_{i,j=1}^{m}$ be the transition intensity matrix of the embedded m-state Markov chain $X(t)$.

The model of processes considered is a class of the Markov-additive processes which have important applications to queueing and data communication models. They are used to model queueing-reliability systems, arrival processes in telecommunication networks, environmental data, neural impulses etc. A particularly important class of the Markov-additive processes is the class of Markov-additive processes $(A(t), X(t))$ of arrivals, i.e., those ones with the additive component $A(t)$ taking values in the set of nonnegative integers. A typical example is that of arrivals at a queueing system.

Sequential estimation procedures of the form $d = (t, d(t))$ will be considered, where t is a stopping time and $d(t)$ is an estimator based on the observation of the process up to t. The parameter $J = (l_{i,j}, i, j = 1,...,m; v_1,...,v_m)$ of the Markov-additive process considered is unknown and the problem is to find optimal sequential procedures, i.e., optimal stopping times t and the corresponding sequential estimators $d(t)$ for J. It is supposed that if the observation is stopped at time t and the estimate $d(t)$ is reported, then the loss incurred is

$$L_t(J, d(t)) = L(J, d(t)) + c(t),$$

where $L(J, d(t))$ denotes the loss function (representing the error of estimation) and $c(t)$ is the cost function. The loss function is defined by a weighted squared error loss. The cost for a given procedure is determined by a function of one of the components of the Markov-additive process; for example, it is the cost depending on arrivals at a queueing system up to the moment of stopping.

Let p be a prior distribution on Θ. Then the Bayes risk of the sequential procedure $d = (t, d(t))$ is $r(p, d) = \int_{\Theta} E_J[L_t(J, d(t))]dp(J)$.

If there is precise prior information on the distribution of the unknown parameter J which can be described by a prior p, then usually the Bayes principle is used. If on the other hand no prior information is available, then the minimax principle can be applied. In the paper, to find optimal sequential estimation procedures, the intermediate approach between the Bayes and the minimax principle is chosen. The use of the Γ-minimax principle is appropriate if vague prior information is available which can be described by a subset Γ of a set Π of all priors. The problem is to find stopping times t and the corresponding sequential estimators $d(t)$ subject to the minimax criterion: a sequential procedure $d^0 = (t^0, d^0)$ is said to be Γ-minimax if i.e., if it minimizes the maximum of the total Bayes sequential risk when the set of prior

$$\sup_{p \in \Gamma} r(p, d^0) = \inf_{d \in D} \sup_{p \in \Gamma} r(p, d),$$

distributions of the unknown parameter is restricted to a subset Γ of all priors. D is a class of all sequential procedures d having finite Bayes risk for each $p \in \Gamma$.

The set Γ is determined by certain conditions imposed on the moments of the prior distributions. The idea and tools are exhibited to obtain Γ-minimax sequential procedures for estimating important quantities of the unknown parameters of the Markov-additive process. As one of the tools for solving the problem, a minimax theorem, which is a considerable generalization of a theorem of Dvoretzky, Kiefer and Wolfowitz (1953), is given for a general class of stochastic processes and a wide class of stopping times.

Several classes of Γ-minimax sequential procedures for estimating the unknown parameters of the Markov-additive process are presented. For example, a class of Γ-minimax sequential procedures is derived explicitly in the case when for a fixed state i the ratios of $l_{i,j}, j = 1, ..., m;\ j \neq i,$ to $f_i^{'}(v_i)$ are of interest. In particular, the results presented are applicable for the Markov-additive processes of arrivals most frequently involved in the literature, i.e., for the Markov-modulated Poisson processes. The results obtained constitute a generalization of the results given by Magiera (1999).

References

Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1953). Sequential decision problems for processes with continuous time parameter. Problems of estimation, *Ann. Math. Statist.*, **24**, 403-415.

Magiera, R. (1999). Minimax sequential procedures for Markov-additive processes, *Stochastic Models*, **15**, 871-888.

Pacheco, A. and Prabhu, N. (1995). Markov-additive processes of arrivals. In *Advances in Queueing: Theory, Methods and Open Problems* (ed. J.H. Dshalalow), 167-194, Boca Raton: CRC Press.

Stefanov, V. (1995). Explicit limit results for minimal sufficient statistics and maximum likelihood estimators in some Markov processes: exponential families approach, *Ann. Statist.*, **23**, 1073-1101.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Authorship Investigation Using Statistical Tools

Madalena Malva*
Dept. de Mat. do Ins. Politécnico de Viseu e
Centro de Est. e Apli. da Univ. de Lisboa
malva@mat.estv.ipv.pt

Statistical tools may be an effective help in deciding disputed authorship (Mosteller and Wallace, 1983) and in many other linguistic and stylistic studies. Zipf's law, for instance, may be used to characterize the vocabular wealth of a text or of an author (Kruskal and Tanur, 1978), and multidimensional scaling is an effective tool to investigate the stylistic evolution of an author (Malva and Pestana, 1999).

In 1974, Barreno, Horta and Costa (1974) published a collection of texts; the three authors never wanted to reveal who wrote what. We used word and phrase lengths, paragraph structure and length, and tagged noncontextual words and a random selection of texts in that and in other books by the authors, published in the same decade, to try to give a tentative answer to the problem of authorship identification, using some new statistics whose distribution we studied.

In particular, we put forward an objective rule to select discriminant noncontextual words: to any noncontextual word we associate (x,y,z), the number of occurences in the reference texts by Barreno, Horta and Costa, respectively, and we retained only those whose score $\max(x,y,z) - \min(x,y,z) > 2$. For the retained words, we have calculated

$$R = \frac{(x-y)^2}{x+y} + \frac{(x-z)^2}{x+z} + \frac{(y-z)^2}{y+z}$$

Under the hypothesis that this does not discriminate among authors, it has a chi-square distribution with 3 d.f. Using this, only 3o words have been retained as potential indicators. These have been used, together with paragraph, sentence and word lengths, to make author identifications.

With our approach we could compute the probability of each of the writers being the author of each of the selected texts. One of the authors confirmed that our solution was the right one in a vast majority of cases.

References

Barreno, M. I., Horta, M. T. e Costa, M. V. (1974) *Novas Cartas Portuguesas*, 2ª ed. Futura, Lisboa.

Kruskal, W. H. and Tanur, J. M. (1978) *International Encyclopedia of Statistics.* The Free Press, London.

Machado de Sousa, M. L. (1979) *Portugal e o Mundo nos Primeiros Decénios do Século XIX*. Francisco Solano Constâncio, 1777- 1846 (compilação). Arcádia, Lisboa

Malva, M. (1998). *Quem Foi Que? — Um Desafio à Estatística: Questões de Autoria em "Novas Cartas Portuguesas"*. Dissertação de Mestrado, DEIO, FCUL.

* This project was partially supported by FCT/ POCTI/FEDER.

Malva, M. and Pestana, D. (1999). Fronteiras da Estatística — as Palavras e a Estatística. *Bol. Soc. Port. Mat.* **39**

Mosteller, F. and Wallace, D. L. (1983) *Applied Bayesian and Classical Inference* — The Case of the Federalist Papers, Springer, New York.

Rosenberg, S., Nelson, C. and Vivekananthan, P. S. (1968) "A multidimensional approach to the structure of personality impressions". *J. Personality and Social Psychology*, **9**, 283-294.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

On Weak Converngence to Tweedie Laws and Regular Variation of Natural Exponential Families

José Raúl Martínez
FaMAF – Universidad Nacional de Córdoba
Ciudad Universitaria – 5000 – Córdoba – ARGENTINA
jmartine@mate.uncor.edu

1. Introduction

Tweedie Laws (or Tweedie exponential dispersion models) generalize an important subclass of infinitely invisible distributions, induding positive and extreme stable Laws. It was can jetured in Jorgensen (1997) and proved in Winogradov (1999) that under some assemptions, the results on weak convergence to certain Tweedie Laws can be derived from those on weak convergence to stable laws.

In this paper, we prove a result that relates regular variation of the variance function for a natural exponential family to regular variation of a miltably defined generating measure for the family. In particular this enables us to interpret the Tweedie convergence condition directly in terms of the asymptatic behavion of the under lying probability measures, and we have this characterized an important part of the domain of attraction to the Tweedie models.

2. Exponential Dispersion Models and Regular Variation.

Consider a particular univariate nonnegative natural exponential family given by the following probability density function:

(2.1) $$p(y;è)dy = 1/b_0(è)e^{èy}\, v(dy)$$

for y 0, where v is a ó-finite measure on [0,), and for a fixed ä 0,

$$b_ä(è) = \int_{[0,\ )} e^{èy} y^ä v(dy),$$

Here, canonical parameter , the canonical parameter domain , in the present case, is an interval finite or infinite.

Clearly, the exponential family (2.1) has mean

$$\mu = ô(è) = b_1(è)/b_0(è),$$

with mean domain Ù = (0, μ_0), where μ_0 is either be infinite (the steep case) or finite (the non-steep case). The variance function for (2.1) is

$$V(\mu) = ô´\{ ô^{-1}(\mu)\} \text{ with the same domain } Ù.$$

The **reproductive exponential dispersion model ED (μ,$ó^2$)** generated by the natural exponential family (2.1) is given by densities of the following form:

$$P(y;\mu,ó^2)dy = 1/bë(è).e^{ëèy}\, õ_ë(dy),$$

for suitable measures $õ_ë$. Here, $\mu = ô(è)$ is the mean mapping and $ó^2 = 1/ë$ is the dispersion parameter. Also, the variance is $ó^2\, V(\mu)$. Here V(.) is understood as the unit variance function.

The Tweedie exponential dispersion model with power parameter p å R\ (0,1), denoted by $Tw_p(\mu,\acute{o}^2)$, is defined in term of the following unit variance function:

$$V(\mu) = \mu^p$$

Here, the domain for μ is R_+, with one exception p = 0, for which it is R.

A ó – finite measure v en [0,) is said to be (bounded) regularly varying with exponent ñ > 0 if function x {(0,x)}is (bounded) regularly varying with exponent ñ at either zero or in finity, as the case may be. Here in after we refer to regular (bounded) variation with exponent ñ as ñ -variation.

Given a natural exponential family represented by (2.1), the measure $m_i(dy) = y^i v(dy)$, for i = 1,2 are called the **first–and second–moment measures** of the family, respectivy.

3. Main Results

Theorem 3.1 Consider the natural exponential family (2.1) generated by $_0$). Then

1 -
varying at infinity if and only if the unit variance function V(.) is bounded p(á)-varying at $ì_0$. For á ª {1,2}, the result remains valid with bounded regular variation replaced by regular variation.

2 - A particular variance function V(.) is bounded 2-varying at infinity if and only if corresponding generating measure í is (-á)-varying at infinity for

Theorem 3.2 The second-moment measure m^2 is (bounded) regularly varying at zero with exponent 2 – á > 2 if and only if the corresponding unit variance function V(.) is (bounded) regularly varying at zero either with exponent p(á) ª (1,2) if í(0) > 0 or exponent 2 if í(0) = 0, where paramet á on p are related as follows: p = (á – 2) / (á –1), with the convertion p = when á = 1

Theorem 3.3 (Jorgensen, Martínez, Tsao). Suppose the unit variance function V is regular of order p at either zero or infinity. Then p å R\ (0,1) and for any ì > 0 $^2 > 0$,

$$c^{-1}ED(cì,ó^2c^{2-p}) \quad Tw_p(ì,c_o ó^2)$$

as c 0 or c , respectively, where the convergence is through values of c such that ^{2} å Ë. The latter condition requires the model to be infinitely divisible if c^{2-p}

This is a Joint work with Ben Jorgensen (Odense University, Denmark) and Wladimir Vinogradov (Athens University, USA).

References

Jorgensen, B. (1997). The Theory of Dispersion Models. London: Chopman and Hall.

Winagradov, V., (1999) On a conjeture of B.Jorgensen and A.O. Wentrell: from Extreme Stable Laws to Tweedie Exponential Dispersion Models. *Technical Report Series of Department of Mathematics and Computer Science* N° **9**.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

The Central Limit Theorem for Trace Class Operators

André Mas
Crest-Université Paris VI
3, Av. Pierre Larousse, 92245 Malakoff Cedex, France
mas@ensae.fr

1. Introduction

The statistical study of random functions is a quite new but promising area in modern statistics. The monograph by Ramsay and Silverman (1991) underlines its practical aspects whereas, for instance, Dauxois Pousse and Romain (1982) provided a solid abstract framework to this topic.

We focus on the covariance operators of random (hilbertian) functions, which are basic tools for the statistical inference (principal component analysis, canonical covariance analysis, decomposition of gaussian curves…).

These empirical covariance operators are usually studied as Hilbert-Schmidt operators for several reasons related to the particular geometry of Hilbert spaces. Yet, the « natural » space should be the space of trace class operators since any Hilbert-valued random variable has second order strong moment (in other words the expectation of its square norm is finite) if and only if its covariance operator is of trace class (see for instance Ledoux and Talagrand, (1991)).

This separable Banach space, denoted C_1 in the following, is of cotype 2 as was proved by Tomczack-Jaegerman (1974). We give a sufficient condition, which may be easily checked, for a random operator to satisfy the CLT in this space. These results are applied to covariance operators of independent and time dependent random functions.

2. Main Results

All the random elements considered are centered.

The first result deals with independent random operators, the second one is more general and connected with m-dependance. We denote (A) the following assumption :

$$S_p\,(Ä|T(e_p)|^{\,2})^{1/2} \text{ is a convergent series,}$$

where e_p denotes any basis in H, T a random trace class operator and $\otimes$ stands for the expectation.

Theorem 1 Let T be a random operator in C_1, satisfying (A), then T also satisfies the CLT.

Theorem 2 Let $T_1,\ldots,T_n$ be a stationnary sequence of m-dependent random operators in C_1. If (A) holds for T_1, the CLT is still satisfied.

The first theorem was proved without any type or cotype argument. Assumption (A) suffices to ensure that the corresponding sequence of measures is tight and the conclusion follows by Prohorov's Theorem. Note the (A) is really quite close to the condition for a random variable with values in the sequence space l_1 to be pre-gaussian.

3. Application to Covariance Operators of Random Functions

The previous results may be applied to a sample $X_1,\ldots,X_n$ of random variables on H. Set $T_i=<X_i,.>X_i-C$ where $<.,.>$ denotes the inner product on H and C the covariance operator of X_i. or $T'_i=<X_i,.>Y_i$ where X_i and Y_i are independent (T'_i is the cross-covariance operator of the couple(X_i,Y_i)). We refer to Bosq (2000) (functional ARMA processes) and Cardot, Ferraty, Sarda (2000) (functional linear regression) for interesting examples of models where such operators are studied. A crucial point is to express assumption (A) by means of the data (namely the X_i's) only.

Proposition Let λ_i be the eigenvalues of C. If the two following conditions hold:

$$S_i(l_i)^{1/2} \text{ is finite and } Ä< X_1,e_p>^4=O(l_p^2),$$

T_i satisfies the CLT.

A similar result may be obtained with T'_i instead of T_i and for many linearly dependent processes.

References

Bosq, D. (2000). Linear Processes in Function Spaces. Lecture Notes in Statistics. Springer Verlag.

Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Stat. and Proba. Letters* **b45**, 11-22.

Dauxois, J. Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function : some applications to statistical inference. *J.Multivar. Anal.* **12**, 136-154.

Dunford, N. and Schwartz, J.T. (1988). Linear Operators Vol I, II. 2nd ed. Wiley Classics Library.

Ledoux M, Talagrand M (1991). Probability in Banach Spaces. Springer.

Ramsay, J.O., Silverman, B.W. (1997). Functional Data Analysis. Springer Series in Statistics. Springer.

Tomczak-Jeagerman, N. (1974). The modulii of convexity and smoothness and the Rademacher averages of trace classes. *Studia Math.* **50**, 163-182.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

The Time to a Given Drawdown in Brownian Motion: Connection to the Pricing of Look-Back American Options

Isaac Meilijson
School of Mathematical Sciences, Tel Aviv University
Israel
MEILIJSON@MATH.TAU.AC.IL

This talk deals with two different but inter-connected topics on stopping times t in Brownian Motion $B(t),\ t \geq 0$ that are related to its cumulative maximum

$$M(t) = \sup_{s \leq t} B(s):$$

- The distribution of the first time Brownian Motion differs from its cumulative maximum by some fixed amount, the *drawdown*, *gap* (Dubins and Schwarz, 1988) or *extent* (Goldhirsch and Noskovicz, 1990).
- Optimal stopping when the sampling cost is linear in time and the reward upon stopping is a non-decreasing function f of the cumulative maximum. This can be viewed as pricing and management of a type of look-back American option $f(M(t)) - ct$. The case of linear reward function was studied by Dubins and Schwarz.

It is possible to see on general principles that these optimal stopping times must be Azéma-Yor-type stopping times (Azéma and Yor, 1978a, 1978b), to embed à la Skorokhod (1965) distributions in Brownian Motion, a notion to be briefly introduced in the talk.

The solution will be explicitely presented for step functions f by a Dynamic Programming construction, to be taken more generally to the limit under finer and finer discretizations. This limit identifies the differential equation

$$H(x) - \frac{1}{4c}\left(H'(x)\right)^2 = f(x)$$

for the (always absolutely continuous) optimal reward function

$$H(x) = \sup_t E\left[f\left(x - M(t)\right) - ct\right]$$

and allows the representation of the solution as: stop as soon as the gap $M(t) - B(t)$ reaches the value $\frac{H'\left(M(t)\right)}{c}$.

References

Azéma, J. and Yor, M. (1978a). Une solution simple au problème de Skorokhod. *Sem. Prob. Strasb. XIII, Springer Lecture Notes in Math.* **721**.

Azéma, J. and Yor, M. (1978b). Le problème de Skorokhod: complements. *Sem. Prob. Strasb. XIII, Lecture Notes in Math.* **721**.

Dubins, L.E. and Schwarz, G. (1988). A sharp inequality for sub-martingales and stopping-times. Soc. Math. de France, *Astérisque*, **157/8**, 129-145.

Goldhirsch, I. and Noskovicz, S.H. (1990). The first passage time distribution in random random walk. *Physics Review*, **A42**, 2047-2064.

Meilijson, I. (1981/82). On the Azéma-Yor stopping time. *Sem. Prob. Strasb. XVII, Springer LN in Math*. **986**, 225-226.

Skorokhod, A. (1965). *Studies in the Theory of Random Processes.* Addison Wesley: Reading.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Extension of Kolmogorov's Strong Law to Multiple Regression

João Mexia
Faculdade de Ciências e Tecnologia, Departamento de Matemática
Quinta da Torre, 2825-114, Monte de Caparica, Portugal

Pedro Corte Real
Faculdade de Ciências e Tecnologia, Departamento de Matemática
Quinta da Torre, 2825-114, Monte de Caparica, Portugal
parcr@mail.fct.unl.pt

1. Generalization of Kolmogorov's Strong Law

We start by establishing the following Lemma;

Lemma 1 Consider the random variables $V_1,\ldots,V_n$, ..., i.i.d., and the real constants, $c_1,\ldots,c_n$, ..., whose absolute values are less than c. If $E[V_i]=0,\ i=1,\ldots,n$, and $\lim n^{-2}\sum_{i=1}^{n} V[V_i]<\infty$, then $\lim n^{-1}\sum_{i=1}^{n} c_i V_i \xrightarrow{a.s.} 0$.

Proof With $W_i=c_iV_i$, we have, $E[W_i]=0$, as well as, $\lim n^{-2}\sum_{i=1}^{n} V[W_i]\le c^2\lim n^{-2}\sum_{i=1}^{n} V[V_i]<\infty$, so, by [Williams, D. (1991), page 118], $\lim n^{-1}\sum_{i=1}^{n} c_iV_i=\lim n^{-1}\sum_{i=1}^{n} W_i \xrightarrow{a.s.} 0$.

Let us now consider the (classic) linear model, $\vec{y}^{\,n}=X_{n,s}\vec{b}^{\,s}+\vec{e}^{\,n}$. We may now state

Theorem 2 If $\lim n^{-1}X'X=W$, a positive definite matrix, the lines of the matrix X belong to a compact (contained in $\mathbb{R}^s$) and the components $e_i,\ i=1,\ldots,n$, of the error's vector are i.i.d. with $E[e_i]=0$. Then $\tilde{\vec{b}}_n^{\,s}=(X'X)^{-1}X'\vec{y}^{\,n}\xrightarrow{a.s.} b$.

Proof With $m=Xb$ we have,

(1) $$\tilde{b}-b=\left(n^{-1}X'X\right)^{-1}\left(n^{-1}X'(y-m)\right)=\left(n^{-1}X'X\right)^{-1}\left(n^{-1}X'e\right).$$

If the elements of matrix X are the $\left[x_{i,j}\right],\ i=1,\ldots,n;\ j=1,\ldots,s$, the generic component of the vector $n^{-1}X'e$, will be $n^{-1}\sum_{i=1}^{n} x_{i,j}e_i,\ j=1,\ldots,s$. Let us now consider the random vector $\vec{u}^{\,n}$ originated by truncating the components of $\vec{e}^{\,n}$, in the form

$$u_i=e_iI_{[-i,i]}(e_i),\ i=1,\ldots,n.$$

By the Kolmogorov's Truncation Lemma (see Williams, D. (1991), page 118]), we have, for n large enough

(2) $$P[u_n=e_n\text{ "}eventually\text{"}]=1 \text{ as well as } n^{-2}\sum_{i=1}^{n} V[u_i]<\infty.$$

Thus Lemma (1) implies $n^{-1}\sum_{i=1}^{n} x_{i,j}u_i \xrightarrow{a.s.} 0$, j=1,…,s. Hence, by (2) we know that, $n^{-1}\sum_{i=1}^{n} x_{i,j}e_i \xrightarrow{a.s.} 0$, j=1,…,s, as well as

(3) $$n^{-1}\sum_{i=1}^{n} x_{i,j}\left(y_i - m_i\right) = n^{-1}\sum_{i=1}^{n} x_{i,j}e_i \xrightarrow{a.s.} 0,\ \text{j=1,…,s.}$$

Now, with $r(W)$ the spectral radius of matrix W, we have $r\left(\left(n^{-1}X'X\right)^{-1}\right) \to r\left(W^{-1}\right) = r(W)^{-1} < \infty$. So, an $m \in \square$ exists, such that, for all $n \geq m$ we have $r\left(\left(n^{-1}X'X\right)^{-1}\right) < 2\,r\left(W^{-1}\right)$.

Then, using (1), we get,

$$\left\|\tilde{b} - b\right\| < 2\,r\left(W^{-1}\right)\left\|n^{-1}X'(y-m)\right\|$$

and the thesis follows immediately from (3).

If we make $X_{n,s} = X_{n,1} = \vec{1}^n$, we have exactly the case considered by Kolmogorov, so we can look at Theorem (2) as a generalization of the classical Kolmogorov's Law of Large Numbers.

References

Williams, D. (1991). Probability with Martingales. Cambridge Mathematical Textbooks.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Traffic Control at a Bottleneck

O. Moeschlin
FernUniversität Hagen, Department of Mathematics
L=tzowstr. 125, D-58084 Hagen
otto.moeschlin@FernUni-Hagen.de

The control of traffic lights at a bottleneck will be understood as a (non-standard) queueing problem assuming Poisson arrival processes. One lane of a two-lane road is blocked, so that the traffic from both sides has to share the one remaining lane. The organisation of the traffic flow is realized by traffic lights, which give mutually free course to at most one side. The control parameters in the hand of the operator of the traffic lights are the green times for the both sides.

The study of the bottleneck case is insofar important, as it can be generalized to more complicated forms of traffic organisation as roundabouts and junctions.

The technical part of a bottleneck is described for the imsymmetric case by Δ_i, t_{R_i}, $i=1,2$. Δ_i in [veh/s] denotes the passage capacities, t_{R_i} in [s] the clearance time for side *i*.

The arrival process for the both sides are assumed as Poisson processes over $(\Omega, \mathrm{A}, \mathrm{P})$ with intensities I_i, $i=1,2$. $t_{F_i} > 0$, $i=1,2$, are the times of free passage (signalized by 'Green' or 'Yellow'), acting as control variables.

Define

$$a_i(t_F) := \left[t_{F_i} \cdot \Delta_i \right],$$

[.] denoting the Gaussbrackets and

$$l_i(t_{F_1}, t_{F_2}) := I_i(t_{F_1} + t_{F_2} + t_{R_1} + t_{R_2}).$$

Basing on a detailed modeling of the control of the bottleneck the sequence $\left(q_j^{(i)}\right)$ of the distributions of the queue lengths $L_j^{(i)}$ (in the j-th period), $j \in \mathbb{N}$, follow a recursion with the operator

$$T_i : \mathrm{M}^{\infty}(\mathbb{N}^+) \to \mathrm{M}^{\infty}(\mathbb{N}^+),$$

($\mathrm{M}^{\infty}(\mathbb{N}^+)$ denoting the set of probability measures over $\mathbb{N}^+$ with finite support), defined by

$$T_i q(l) := \begin{cases} \sum_{k=0}^{a(t_{F_i})} \left(q * p_{l}(t_{F_1}, t_{F_2}) \right)(k) & \text{for } l = 0 \\ \left(q * p_{l_i}(t_{F_1}, t_{F_2}) \right)\left(l + a_i(t_{F_i}) \right) & \text{for } l \geq 1,\ i = 1,2. \end{cases}$$

For $a_i(t_{F_i}) > l_i(t_{F_1}, t_{F_2})$, $i=1,2$ it follows from the properties of T_i (together with E. Grycko), that the operators T_i $i=1,2$ enjoy a fixed point, i.e. the queueing process on both sides are ergodic; moreover it holds

$$\lim \int L^{(i)}(t)\,\mathrm{dP} < \infty,\ i = 1,2.$$

Calling a pair $\left(t_{F_1}, t_{F_2}\right)$ of times of free passage ergodic, the obtained results give rise to the following

Definition An ergodic pair $\left(t_{F_1}, t_{F_2}\right)$ is called optimal, if

$$\max_{i=1}^{2} \lim_{j\to\infty} \lim_{t+jt_u+t_{C_i}} \left(I^{-1} \int L^{(i)}(t)\,\mathrm{dP}\right),$$

where $t_C := t_{R_1} + t_{R_2} + t_{F(3-i)}$, $i = 1,2$ and $t_u := 2\left(t_{F_i} + t_{R_i}\right)$, i.e. t_u is the duration of a full period while t_{C_i} is the closed time signalized by 'Red', $i = 1,2$. It can be proved on one hand that the (asymptotic) efficiencies (throughput per time unit) equals the intensities I_i, $i = 1,2$; on the other hand by computer experimentation it yields, that

$$\left(I^{-1} \int L^{(i)}(t)\,\mathrm{dP}\right)$$

is a good estimate of the expected time a newly arriving vehicle from side i has to wait, which justifies the definition of an optimal pair of times of free passage.

Under stronger assumptions the results hold true even for renewal processes as arrival processes; the proof (together with C. Poppinga) is based on a dominance principle for stochastic processes.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Goodness-of-Fit Test For Linear Process

Zaher Mohdeb
University Mentouri, Constantine, Algeria
mohdeb@wissal.dz

Let $X_1, X_2, ..., X_n$ be *n* consecutive observations generated by a stationary time series $(X_t)_{t\in\square}$ with $E(X_t^2) < \infty$. We discuss the problem of testing the null hypothesis H_0 that the data is generated by a specified linear process linear process of the form $X_t = \sum_{j=-\infty}^{\infty} y_i e_{t-j}$, where the (e_t) are *i.i.d* random variables with mean zero and variance s^2; we assume that $E(e_t^4) < \infty$ and $\sum_{j=-\infty}^{\infty} |y_j||j|^b < \infty$. More precisely, let $I_{n,X}(w)$ denote the periodogram of $\{X_1, ..., X_n\}$ and let $f_X(w)$ be the spectral density of (X_t). Suppose that $0 < l_1 < ... < l_m < p$, we give an asymptotic distribution of the random variables $U_i = R_{i,n}/R_{m,n}$, where $R_{i,n} = \sum_{k=1}^{i} \{I_{n,X}(l_k)/f_X(l_k)\}, i = 1, ..., m-1$, and two corollaries wich can be used to construct two tests of the null hypothesis H_0. The first idea is to reject the null hypothesis H_0 if the "mormalized periodogram" $\{I_{n,X}(l_k)/f_X(l_k)\}$ contains a value substantially larger than the average value. The second corollary suggests another test of the null hypothesis H_0 applying the theory of well known Kolmogorov-Smirnov test. Note that *AR*, *MA* and *ARMA* models may be regarded as special cases of the null hypothesis. We give a numerical illustration of our test applied to simulated series.

Squared Skewness Minus Kurtosis Bounded by 186/125 for Unimodal Distributions

Philip J. Mokveld
University of Amsterdam, Korteweg-de Vries Institute for Mathematics
Plantage Muidergracht 24, 1018 TV Amsterdam, Nederlands
pmokveld@science.uva.nl

Chris A.J. Klaassen
University of Amsterdam, Korteweg-de Vries Institute for Mathematics
chrisk@science.uva.nl

Bert van Es
University of Amsterdam, Korteweg-de Vries Institute for Mathematics
vanes@science.uva.nl

Let F be a nondegenerate distribution with finite fourth moment. The coefficient of kurtosis of F is defined as

$$E_F(X - E_F X)^4 s_F^{-4} - 3$$

and denoted by k_F with s_F the standard deviation. We will denote the coefficient of skewness

$$E_F(X - E_F X)^3 s_F^{-3}$$

by t_F.

For different classes of distributions sharp inequalities have been derived for the squared skewness minus the kurtosis. For instance, Pearson (1916) derived an upper bound of 2, with equality iff X takes on two distinct values a.s. under F.

If F belongs to the class of unimodal distributions for which the mean and mode coincide, the inequality has an upper bound of 6/5, which is attained iff F is uniform. Consequently, the kurtosis has a lower bound of -6/5 for symmetric unimodal distributions.

We have derived a sharp inequality for all nondegenerate unimodal distributions with finite fourth moment. This is given by

$$t_F^2 \le k_F + \tfrac{186}{125},$$

which holds with equality iff F is a one-sided boundary-inflated uniform distribution with mass 1/2 at the atom.

References

Pearson, K. (1916). Mathematical contributions to the theory of evolution, XIX; Second supplement to a memoir on skew variation, Philos. *Trans. Roy. Soc. London* Ser. A 216, 429-457.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Two-Step Sequential Sampling

J.J.A. Moors, L.W.G. Strijbosch
Tilburg University, Department of Econometrics and Operations Research
PO Box 90153, 5000 LE Tilburg, the Netherlands
J.J.A.Moors@kub.nl, L.W.G.Strijbosch@kub.nl

1. Introduction

Deciding upon the optimal sample size in advance is a difficult problem in general. Often, the investigator regrets not having drawn a larger sample; in many cases additional observations are done. This implies that the actual sample size is no longer deterministic; hence, even if all sample elements are drawn at random, the final sample is *not* a simple random sample. Although this fact is widely recognized, its consequences are often grossly underrated in our view. Too often, these consequences are ignored: the usual statistical procedures are still applied.

This paper shows in detail the dangers of applying standard techniques to extended samples. To allow theoretical derivations only some elementary situations are considered. More precisely, the following features hold throughout the paper:

- The population variable of interest is normally distributed,
- Estimation concerns population mean and variance,
- All sample elements are drawn at random, with replacement,
- Only standard estimators, like sample mean and sample variance, will be considered.

2. Fixed Sample Extension

An initial sample of size n_1 is drawn from $N(m, s^2)$, with m and s unknown; sample mean and variance are denoted by $\bar{y}_1$ and s_1^2, respectively. An additional sample of again size n_1 is drawn if some criterion C_i is satisfied; here we consider:

$$C_2 = \{\bar{y}_1 > c\}, \quad C_3 = \{s_1^2 > d\}$$

with c and d given constants. (Note that C_1 may be applied when $H_0 : m \geq m_0$ is 'nearly' rejected, while C_2 is used when the attained accuracy of a sample proves to be unsatisfactory.)

Let $\bar{y}_3$ and s_3^2 denote mean and variance of the extended sample (of size $2n_1$); then for C_2 as well as C_3, the natural estimator $\bar{y}$ for m and the natural estimator $\text{var}(\bar{y})$ of its variance are given by

$$\bar{y} = \begin{cases} \bar{y}_1 & \text{if} \quad C_i' \\ \bar{y}_3 & \text{if} \quad C_i \end{cases} \qquad \text{var}(\bar{y}) = \begin{cases} s_1^2 / n_1 & \text{if} \quad C_i' \\ s_3^2 / (2n_1) & \text{if} \quad C_i \end{cases}$$

Table 1 presents some worst-case results: the extreme relative bias that these two estimators may have.

Criterion	$ERB(\bar{y})$	$ERB[\mathrm{var}(\bar{y})]$
C_2	$\frac{-0.20s/m}{\sqrt{n_1}}$	0.18
C_3	0	$-\frac{0.42}{\sqrt{n_1}}$

Table 1. Extreme relative bias (ERB) **of** $\bar{y}$ **and** $\mathrm{var}(\bar{y})$; C_2 **and** C_3**.**

3. Stochastic Sample Extension

Now two independent samples of size n_1 are drawn independently from $N(m_1, s^2)$ and $N(m_2, ts^2)$, leading to sample variances s_1^2 and t_1^2, respectively. Extension criterion $C_4 = \{s_1^2 > t_1^2\}$ is used; if C_4 occurs, the second step sample size n_2 for population 1 is

$$n_2 = n_1 \ \mathrm{ent}(s_1^2 / t_1^2 - 1)$$

(Criterion C_4 is useful when the two populations means m_1 and m_2 have to be estimated with about equal accuracy.) Table 2 presents the (simulated) relative bias of the variance estimator of the final sample mean $\bar{y}$.

t \ n_1	5	9	13	17	25	50
1/3	-7.5	-2.3	1.2	-1.0	0.0	-1.5
1/2	-12.3	-5.0	-3.7	-1.4	-3.9	-1.0
1	-21.5	-13.9	-10.5	-9.5	-7.3	-6.4
2	-30.5	-19.8	-17.2	-12.2	-9.3	-2.9
3	-32.3	-19.9	-14.6	-11.2	-5.0	-2.9

Table 2. Relative bias (in %) of $\mathrm{var}(\bar{y})$; C_4**.**

References

MOORS, J.J.A. & L.W.G. STRIJBOSCH, Two-step sequential sampling, CentER Discussion Paper 2000-39; see grey *www.kub.nl:2080/greyfiles/center/ctr_py_2000.html*

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Evaluating the Impact of Misleading Signals in Joint Schemes for m and s *

Manuel Cabral Morais, António Pacheco
Instituto Superior Técnico, Dept. de Matemática e Centro de Matemática Aplicada
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
{maj,apacheco}@math.ist.utl.pt

One of the potential risks of using joint schemes for the process mean (m) and standard deviation (s) is the emission of what is called misleading signal (MS) by St. John and Bragg (1991) and Morais and Pacheco (2000a). Two possibilities of triggering a MS are as follows: (1) A signal is only triggered by the scheme for m, although this parameter is on-target and s is out-of-control (MS of Type III); (2) m is off-target and s is in-control; however, a signal is exclusively given by the chart for s (MS of Type IV).

This sort of signal can send the user of the joint scheme to try to diagnose and to correct a non-existent assignable cause - and, thus, increase inspection costs -, because the diagnostic procedures that follow an out-of-control signal can differ depending on whether the signal was given by the scheme for m or the scheme for s.

In this extended abstract we use the probability of a misleading signal (PMS) to assess the impact of MS's in the joint monitoring of m and s. Another possibility that also springs to mind is the run length to a misleading signal (RLMS); for a further discussion on this alternative performance measure please refer to Morais and Pacheco (2001). We provide convincing examples that alert the user - namely of five joint schemes for the mean and standard deviation of a quality characteristic with normal distribution - to the phenomenon of MS's of Types III and IV.

The individual schemes for m which constitute the joint schemes with acronyms SS^+, CC^+, EE^+, CCS^+ and CES^+ are the upper one-sided X-bar, CUSUM, EWMA, combined CUSUM-Shewhart and combined EWMA-Shewhart schemes, respectively. As for the individual schemes for s, the joint schemes SS^+, CC^+, EE^+, CCS^+ and CES^+ are associated to the upper one-sided S^2, CUSUM, EWMA, combined CUSUM-Shewhart and combined EWMA-Shewhart schemes, respectively.

Let the increase in m and in s be measured in terms of $d = \sqrt{n}(m - m_0)/s_0$ and $q = s/s_0$, respectively. Then the PMS's of Types III and IV are equal to: PMS(III;q)=P[$RL_s(q) > RL_m(1,q)$], $q > 1$, and PMS(IV;d)= P[$RL_m(d,1) > RL_s(1)$], $d > 0$ (respectively), where $RL_m(d,q)$ and $RL_s(q)$ repre-sent the run lengths of the individual schemes for m and s.

To illustrate the occurrence of MS's of Types III and IV in the joint scheme EE^+, we consider an example with two simulated data sets; their corresponding sample means, variances and statistics of the individual schemes for m and s can be found in Table 1. This table also has PMS's of Types III and IV for the five joint schemes considered here. (For further considerations on the parameters of these joints schemes see Morais and Pacheco (2000b).) The results in Table 1 suggest that the scheme SS^+ compares

* 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 was done under the SAPIENS Project 40004/2001 TOWN initiative.

unfavorably to the joint schemes CC^+, EE^+, CCS^+ and CES^+, in terms of MS's of both types, in most cases. They also give the distinct impression that the values of PMS's are far from negligible, especially for small and moderate shifts in m and s. All these schemes but SS^+ tend to give more MS's due to changes in d (i.e. MS's of Type IV) than due to changes in q (MS's of Type III).

N	(m,s)=(100,1.3)				(m,s)=(100.075,1)			
	mean	var	$stat_m$	$stat_s$	mean	var	$stat_m$	$stat_s$
1	100.597	0.668	0.067	0.000	100.02	0.804	0.002	0.000
2	99.992	3.268	0.063	0.059	100.337	2.604	0.040	0.048
3	100.386	0.644	0.103	0.034	99.945	1.759	0.032	0.074
4	100.400	0.471	0.142	0.000	100.666	1.777	0.105	0.099
5	98.796	0.517	0.001	0.000	100.036	2.229	0.103	0.134
6	100.144	3.345	0.017	0.060	99.975	0.985	0.095	0.126
7	101.531	1.039	0.187	0.059	99.918	1.896	0.082	0.152
8	100.507	2.143	**0.234***	0.094	99.931	2.030	0.070	**0.180****
9	100.259	2.711	**0.252***	0.140	100.138	1.300	0.082	**0.184****
10	99.893	4.136	0.227	**0.204****	99.705	0.555	0.045	0.145

*MS of Type III; ** Non MS; *** MS of Type IV;

m_0 =100; s_0 =1; n=5; $[LCL_m, UCL_m]$=[0,0.195597]; $[LCL_s, UCL_s]$=[0,0.157079]

Table 1. Example, and PMS's of Types III and IV for five joint schemes.

PMS III						PMS IV					
q	SS^+	CC^+	EE^+	CCS^+	CES^+	d	SS^+	CC^+	EE^+	CCS^+	CES^+
1.05	.4319	.3486	.2981	.3650	.3278	0.05	.4611	.4381	.3911	.4502	.4154
1.10	.3771	.2603	.1821	.2770	.2031	0.10	.4238	.3828	.2970	.4021	.3304
1.20	.2976	.1874	.0810	.1969	.0888	0.25	.3214	.2406	.1319	.2581	.1551
1.30	.2452	.1660	.0449	.1735	.0488	0.50	.1919	.0980	.0463	.1075	.0569
1.40	.2091	.1607	.0294	.1682	.0318	0.75	.1115	.0375	.0213	.0436	.0286
1.50	.1827	.1610	.0216	.1693	.0232	1.00	.0650	.0139	.0108	.0186	.0169
2.00	.1106	.1762	.0118	.1892	.0120	1.50	.0237	.0017	.0031	.0047	.0077

We strongly believe that the probability of MS's of Types III and IV should be taken in consideration not only as an additional performance measure in the design of joint schemes for m and s as suggested by Morais and Pacheco (2000a), but also as a guideline in diagnosing which parameter(s) has (have) changed after the emission of a signal (see Reynolds Jr. and Stoumbos (2000)). An analytical justification - from a stochastic ordering point of view - for the monotonic behaviour (and other monotonicity properties) of the PMS's and other monotonic properties of these five joint schemes is given in Morais and Pacheco (2000b).

References

Morais, M.C. and Pacheco, A. (2000a). On the performance of combined EWMA schemes for m and s : A Markovian approach, *Communications in Statistics - Simulation and Computation* **29**, 153-174.

Morais, M.C. and Pacheco, A. (2000b). Misleading signals in joint schemes for μ and σ. Technical Report 17/2000, DM-IST, Lisboa, Portugal.

Morais, M.C. and Pacheco, A. (2001). Misleading signals em esquemas combinados EWMA para m e s . *VIII Annual Meeting of the Portuguese Statistical Society*.

Reynolds Jr., M.R. and Stoumbos, Z.G. (2000). Monitoring the process mean and variance using individual observations and variable sampling intervals. To appear in *Journal of Quality Technology*.

St. John, R.C. and Bragg, D.J. (1991). Joint X-bar & R charts under shift in mu or sigma, *ASQC Quality Congress Transactions - Milwaukee*, 547-550.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

A Conditional Quantile Regression Approach to Returns to Education

Daniel Mota
INE – Instituto Nacional de Estatística
daniel.mota@ine.pt

There is wide evidence that wage inequality has grown in western countries for the period covering the eighties and the first half of the nineties. A very common belief rests on the assumption that a more educated society would strongly reduce this inequality. Using data from the Portuguese Employment Survey covering the years 1998 to 2000, I intend to show that there is no empirical evidence supporting this view.

There are basically two alternative economic theories to explain the phenomenon of growing inequality. The first rests on labour demand and supply framework. A very fast technological evolution has induced a growing demand for high skilled workers but not accompanied with a larger supply. The second theory emphasises the importance of world trade. In a world of growing competition, exchange of technology has become essential to many countries and also easier. Thus, both theories assert that the increase in wage inequality was caused mainly by the pressure for high skilled workers' wages to rise.

Usually, mean regression results (usually estimated by ordinary least squares) are used to study the effects of a predefined set of variables on the dependent variable. However, this technique only gives the effects on the mean of the conditional distribution. This approach seems insufficient when the aim is to uncover the effects of a set of characteristics on the workers' wages. In fact, it is not reasonable to expect the impact of each variable to be the same on the wages of two people standing on opposite extremes of the distribution. Therefore, it is important to use a technique that gives more information concerning the conditional distribution of the wages instead of just its mean. Quantile regression techniques are used inasmuch as they will provide information on the effects of any variable on the dependent variable at any desired quantile of the conditional distribution. This technique has another advantage over the traditional mean regression in that it allows the implicit control for workers' heterogeneity.

In this framework, log hourly wages will depend on the degree of scholarity attained, experience, tenure and workplace area. This specification involved separate estimation for working males and females. Nonetheless, the sample selection bias problem is also addressed in a joint estimation setting.

From the list of empirical results obtained, some deserve a small comment. Working in 'Lisboa e Vale do Tejo' (hereafter, LVT) is clearly beneficial in wage terms. As an example, considering only the year 2000, paid male employees of the region 'Norte' receive, in average, less 17% than their counterparts of LVT. However, the wage gap along the conditional distribution is not uniform. In the first decile, the gap is about 13% whereas at last decil the gap widens to 22,5%.

Experience and tenure, in spite of being well defined, are shown to have little influence on wages formation. Still, they have positive impact which fades away in time. Concerning these two variables, there is not large variation between deciles.

Finally, special attention is paid to returns to education. The effect of education on workers' wages is not constant along the whole conditional wage distribution, as there are clear benefits for the workers standing on the top deciles. As suspected, the returns vary greatly from decile to decile and even for different number of years of experience. Results, based on the functional form described earlier or on a simpler form, both reveal that an increase in wage inequality should be expected if the level of education of the population significantly increases. Hence, for people with up to 6 years of education, an eventual additional schooling year would lead to a decrease in inequality, while for people with more than 9 years of education, it would contribute to strengthen the inequality. In fact, if one has studied no more than 6 years, education,contributes to a contraction of the wage inequality as its impact is smaller in upper deciles. On the other hand, if one has at least 9 years of education, its impact is bigger on the upper deciles. These results confirm that the conditional quantile regression approach is appropriate whereas sticking to the usual mean regression would have been clearly insufficient and perhaps misleading.

The findings herein are not as surprising as they may seem at first sight. Raising the education level of a population leads implicitly to the widening the wage range as dispersion is larger within higher education levels.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Temporal Characterization of Coastal Upwelling Index off Portugal*

Helena Mouriño
University of Lisbon, Department of Statistics and Operational Research
Bloco C2, Piso2, Campo Grande, 1749-016 Lisbon, Portugal
mhnunes@fc.ul.pt

Isabel Barão
University of Lisbon, Department of Statistics and Operational Research
Bloco C2, Piso2, Campo Grande, 1749-016 Lisbon, Portugal
mibarao@fc.ul.pt

Evidence from several different regions suggests that the major coastal upwelling systems of the world have been growing in upwelling intensity as greenhouse gases have accumulated in the earth's atmosphere. Effects of enhanced upwelling on the marine ecosystem are uncertain but potentially dramatic.

The upwelling indexes, measured through some of the most reliable Portuguese coastal stations, are very important variables to establish the climatic framework of coastal upwelling off Portugal. From 1 January 1985 to 31 December 1999 the wind regime at the meteorological stations of Cabo Carvoeiro, Sagres and Sines was analysed. The variables correspond to climatological averages of the mean daily values of speed and frequency. Unfortunately, there are some missing values that are related to damages in the equipment.

In order to estimate the seasonal and long term trend of each time series in the presence of missing values, we had applied two different approaches: the first one was related with the concept of local and seasonal moving averages; secondly we had developed an sine-cosine wave at some significant frequencies. To compare the results obtained from those meteorological stations, we had applied the Bootstrap Methodology. Afterwards, we had established a linear model to describe the relationship between Sagres / Cabo Carvoeiro and Sines / Cabo Carvoeiro.

* Research partially supported by FCT/POCTI/FEDER

Tests in Sparse Multinomial Data Sets

Hannelore Liero
University of Potsdam, Institute of Mathematics
Germany
liero@rz.uni-potsdam.de

We consider the following problems:

1. testing cell probabilities in sparse multinomial data sets, that is to test H: $p_{nj} = p_{nj}$ for all $j=1,\ldots,k_n$ versus K: $p_{nj'} \neq p_{nj'}$ for some j', where the p_{nj}'s are unknown and the p_{nj}'s are given cell probabilities. The number of cells k_n is assumed to tend to infinity as the sample size n tends to infinity.
2. testing independence in contingency tables, i.e. we check whether H: $p_{n\,ij} = q_{n\,i} r_{n\,j}$ for all $i=1,\ldots,k_{1n}$, $j=1,\ldots,k_{2n}$ versus K: $p_{n\,i'j'} \neq q_{n\,i'} r_{n\,j'}$ for some (i',j'), where the $p_{n\,ij}$'s are cell probabilities in a two-dimensional contingency table. Here $q_{n\,i}$ and $r_{n\,j}$ are the corresponding marginal probabilities. Sparseness is described by the condition $k_{1n} \to \infty$ and $k_{2n} \to \infty$ as $n \to \infty$.

The unknown cell probabilities are estimated by kernel smoothing, and the application of central limit theorems for the sum of squares of errors of local polynomial estimators for cell probabilities in sparse data sets leads to asymptotic α-tests for both problems.

The behavior of the power under local alternatives of the proposed test procedures is investigated, the influence of sparseness and comparisons with other tests are discussed. An analogy to L_2-type test procedures for densities is considered.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

A New Approach to the Linear Mean-Square Estimation Problem*

Jesús Navarro-Moreno
Universidad de Jaén, Departamento de Estadística e I.O.
Paraje "Las Lagunillas", s/n, 23071 Jaén (Spain)
jnavarro@ujaen.es

Juan Carlos Ruiz-Molina
Universidad de Jaén, Departamento de Estadística e I.O.
Paraje "Las Lagunillas", s/n, 23071 Jaén (Spain)
jcruiz@ujaen.es

Rosa María Fernández Alcalá
Universidad de Jaén, Departamento de Estadística e I.O.
Paraje "Las Lagunillas", s/n, 23071 Jaén (Spain)
rmfernan@ujaen.es

This poster is concerned with the problem of the linear mean-square estimation of a signal or system process based upon noisy observations. Series representations of stochastic processes by means of deterministic functions and uncorrelated random coefficients are one of the most used techniques to solve this problem. Several types of series representations with uncorrelated coefficients can be found in the literature: Karhunen-Loève expansion, simultaneous orthogonal expansions (Kadota, 1967), Cambanis representation (Cambanis, 1973), etc. The most general representation has been given by Cambanis, since the only assumptions are that the process is of second-order, measurable and defined on any interval of the real line. If a double orthogonality is imposed in this representation, then it becomes a series expansion depending on the eigenvalues and eigenfunctions of the integral operator whose kernel is the autocorrelation of the process, generalizing the Karhunen-Loève expansion. Moreover, it is the optimal representation in the sense that the mean-square error resulting from a finite representation of the process is minimized.

From the practical standpoint, this optimal expansion is very limited because there is no standard method to find the eigenvalues and eigenfunctions of the autocorrelation function. Navarro et al. (2000) have recently developed a methodology to clear up this difficulty. This approach provides a class of finite expansions, called approximate expansions. Such finite expansions are based on the approximate eigenvalues and eigenfunctions calculated from the Rayleigh-Ritz method in order to solve numerically the associated Fredholm integral equation and they have similar properties to the optimal expansion. Moreover, this solution includes, as a particular case, the approximate Karhunen-Loève expansions given by Gutiérrez et al. (1992) and Ruiz-Molina et al. (1999).

Our aim is to apply these approximate expansions to the linear mean-square estimation problem. Thus, a new solution will be provided overcoming the difficulty of computing the true eigenvalues and eigenfunctions. The advantage of this solution is that it is easily implementable on a computer.

* This work was supported in part by Project BFM2000-1103 of the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I), Ministerio de Ciencia y Tecnología, Spain.

References

Cambanis, S. (1973). Representation of Stochastic Processes of Second Order and Linear Operations, *J. Math. Anal. Appl.*, **41**, 603-620.

Gutiérrez, R., Ruiz-Molina, J.C. and Valderrama M.J. (1992). On the Numerical Expansion of a Second Order Stochastic Process, *Appl. Stochastic Models Data Anal*., **8**(2), 67-77.

Kadota, T.T. (1967). Simultaneous Diagonalization of Two Covariance Kernels and Application to Second Order Stochastic Processes, *SIAM J. Appl. Math*., **15**(6), 1470-1480.

Navarro-Moreno, J., Ruiz-Molina, J.C. and Fernández, R.M. (2000). Approximate Series Representations of Second-Order Stochastic Processes. Applications to Signal Detection and Estimation, *I.E.E.E. Signal Process*, submitted for publication.

Ruiz-Molina, J.C., Navarro-Moreno, J. and Valderrama, M.J. (1999). Differentiation of the Modified Approximative Karhunen-Loève Expansion of a Stochastic Process, *Statist. Prob. Lett.*, **42**, 91-98.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Reliability Measures in Weighted Distributions

Jorge Navarro, José M. Ruiz
Universidad de Murcia, Departamento de Estadística e Investigación Operativa
30100 Murcia, Spain
jorgenav@um.es, jmruizgo@um.es

Yolanda del Aguila
Universidad de Almeria, Departamento de Estadística y Matemática Aplicada

yaguila@ual.es

1. Introduction

Let X be a positive random variable having absolutely continuous distribution function $F_X(t)$, survival function $\bar{F}_X(t)=1-F_X(t)$, probability density function $f_X(t)$, failure rate function $r_X(t)=f_X(t)/\bar{F}_X(t)$ and mean residual life function $e_X(t)=E(X-t|X>t)$. It is well known that $\bar{F}_X(t)$, $r_X(x)$ and $m_X(t)$ are equivalent, in the sense that given one of them, the other two can be determined.

Let $w(t)$ be a positive real function with $0<E(w(X))<\infty$. The random variable Y with probability density function

$$f_Y(t)=\frac{w(t)\,f_X(t)}{E(w(X))}$$

is called the weighted random variable corresponding to X, and its distribution in relation to that of X is called the weighted distribution with weight function w. The concept of weighted distribution was formulated by Rao (1965) to model various situations in which the recorded observations cannot be considered as a random sample form the original distribution. In particular, the weighted distribution with weighted function $w(t)=t$ is called the length-biased distribution. A similar definition can be given when X is a discrete random variable, replacing $f_X(t)$ by $p_X(t)=\Pr[X=t]$.

In this paper we obtain general characterizations of probability distributions from relationships between failure rate and mean residual life from the original random variable X and the associated weighted random variable Y.

2. Characterization Results

Nair and Sankaran (1991) characterized the Pearson system of distributions (P.s.d.) defined by

$$f'(t)=\frac{-(t+a)}{b_0+b_1t+b_2t^2}\,f(t)$$

through the relation $m(t)=m+(a_0+a_1t+a_2t^2)r(t)$.This result was generalized by Ruiz and Navarro (1994), giving a general way to obtain $f(x)$ from the relation $m(t)=k+q(t)r(t)$, where k is a real number and $q(t)$ is a real function.

Asadi (1998) (using the result of Nair and Sanckaran (1991)) has characterized the distributions of the P.s.d. with $b_0 = 0$ through the relation

$$\frac{r_Y(t)}{r_X(t)} = d\,\frac{m_Y(t) - E(Y)}{m_X(t) - E(X)},$$

where d is a constant and Y is the weighted distribution associated to X and $w(t) = ta$ $(a > 0)$.

In this paper, using the results given in Ruiz and Navarro (1994), we extend the result of Asadi (1998) to obtain a general result, characterizing any density function from a similar relationship by using the weighted distribution.

Theorem Let X be a random variable, let $w(t)$ be a positive, differentiable and non constant real function and let Y be the weighted r.v. associated to X and $w(t)$. If X has a differentiable density function $f(t)$, k_1 and k_2 are two real numbers and

$$\frac{r_Y(t)}{r_X(t)} = d(t)\frac{m_Y(t) - k_2}{m_X(t) - k_1}$$

then $d(t)$ uniquely determines $f(t)$. Moreover,

$$\frac{f'(t)}{f(t)} = \frac{k_1 - t - q'_1(t)}{q_1(t)}$$

where

$$q_1(t) = \frac{k_2 - t - (k_1 - t)/d(t)}{(w(t)/d(t))'}\,w(t)$$

Using this theorical result we obtain characterizations of some usual distributions (see Navarro, del Aguila and Ruiz (2001)).

References

Asadi, M. (1998). Characterization of the Pearson system of distributions based on reliability measures. *Statist. Papers,* **39**, 347-360.

Nair, N. U. and Sankaran, P.G. (1991). Characterization of the Pearson system of distributions. *IEEE Trans. Reliability*, **40**, 75-77.

Navarro, J., del Aguila, Y. and Ruiz, J. M. (2001). Characterizations through reliability measures from weighted distributions. *To appear in Statistical Papers.*

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhya A*, **27**, 311-324.

Ruiz, J. M. and Navarro, J. (1994). Characterization of distributions by relationships between failure rate and mean residual life. *IEEE Trans. Reliability*, **43**, 640-644.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

The Classical Linear Regression Model with one Incomplete Binary Variable

Thomas Nittner
Ludwigs-Maximilians-Universität München, Department for Statistics
Akademisstraße 1, D – 80799 München
nittner@stat.uni-muenchen.de

1. Introduction

The linear regression as a main tool often is affected by missing values within statistical analyses. Binary variables however prevent using standard methods, e.g. conditional mean imputation. Some of these standard procedures are adapted to this problem, seven procedures are compared when missing data are confined to one independent binary variable: complete case analysis, zero order regression, categorical zero order regression, pi imputation, single imputation, multiple imputation and a single imputation based on a modified first order regression. Assuming a continuous and completely observed response vector y the partially incomplete regressor X_{K-1} could without restriction to generality be reorganized according to $X_{K-1} = (x_c, x_{mis})'$, where c and mis indicate complete and missing.

2. Standard Methods

The unconditional mean imputation, also known as **zero order regression (ZOR)** is a simple alternative to the **complete case analysis (CCA)** which discards all cases containing at least one missing value. The ZOR (Wilks [1932]) imputes the empirical mean of the observed values of X_{K-1} for all missing values which leads to an underestimated variance, however. It should be adapted to the non-continuous scaling, i.e. for a dummy variable the mode should be used. The so-called **first order regression (FOR)**, also known as conditional mean imputation, incorporates the structure of the design matrix X (Buck [1960]). An auxiliary regression based on the complete cases where the incompletely observed variable is the new response vector and all complete variables are the independent part of this regression realizes the consideration of the X - structure. Including the complete response vector y leads to the **modified first order regression (mFOR).** In general, the auxiliary regression is formulated according to

$$x_{ij} = q_{0j} + \sum_{m=1, m\neq j}^{K} x_{im}\ q_{mj} + u_{ij}\ ,\ i \notin \Phi_j$$

where x_{ij} is the missing value and Φ_j the index set of missing values. Thus, x_{ij} can be replaced by

$$\hat{x}_{ij} = \hat{q}_{0j} + \sum_{m=1, m\neq j}^{K} x_{im}\ \hat{q}_{mj}\ ,\ i \in \Phi_j$$

Having a binary x_{ij} however prevents the modelling of the auxiliary regression using the classical linear model. Logistic regression, solves this problem. Before estimating, it could be written as

$$P(x_{i,K-1} = 1 \mid x_{ij}) = \frac{\exp(h)}{1 + \exp(h)} = p_i \quad \text{where} \quad h = b_0 + \ldots + b_{K-2}\, x_{i,K-2},$$

$i \notin \Phi$ and $j = 0,\ldots, K-2$. Based on p_i, several procedures can be built.

3. The Methods

The idea of a probability imputation is realized in the hence called **pi imputation**. As within the classical prediction the estimate b - here based on the complete cases - is used to get substitutes for the missing values ($i \in \Phi$).

For the **Single imputation** conditional distribution of the incomplete variable given the complete variables has to be estimated line-by-line. From these (binomial) distributions a random draw has to be made for every missing case. The **multiple imputation** consists of M imputations leading to M completed data sets. The estimate of the multiple imputation is the average of the different parameters.

4. A Simulation Experiment

Considering the MSE-Ratios (ratio between the scalar MSE of the CCA and the MSE of the alternative) showed that the pi imputation was the only method which had a minor MSE than the CCA for all settings, i.e. for 10, 30 and 50% missingness, each correlation structure and for both models (two binary covariates, two binary plus one continuous). MFOR shows reverse trends concerning its behavior dependent on the missing percentage (see Figure 1), multiple imputation decreases its **MSE** within severe correlation (Figure 2), ZOR decreases ist MSE with increasing percentage of missingness (Figure 3). Analyzing **variance** and **bias** confirmed that mean imputation underestimates variance, mFOR shows maximal variances within severe correlation and pi imputation has the smallest biases within imputation methods. Multiple imputation didn't have higher variances than single imputation. Despite more or less individual deviations it can be stated that variances will increase with increasing percentages of missingness. The pi imputation has the smallest biases within the imputation methods. Multiple and single imputation underestimate the true parameter, mFOR overestimates.

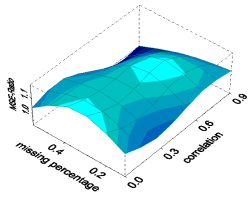


Figure 1

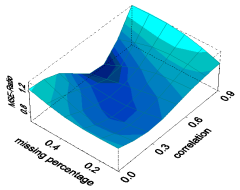


Figure 2

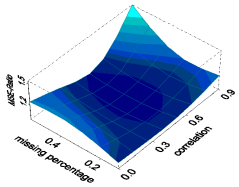


Figure 3

References

Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer, *Journal of the Royal Statistical Society*, Series B **22**, 302-307.

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

Wilks, S.S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples, *Annals of Mathematical Statistics* **3**, 163-195.

Rubin, D.B. (1996). Multiple imputation after 18+ years, *Journal of the American Statistical Association* **91**, 473-489.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

An Approach to Liouville Equation Concerning Predicting Forecast

Alvaro M. D. Nunes
University of Macau, Faculty of Science and Technology
P. O. Box 3001, Macau
fstadn@umac.mo

1. Introduction

We can say that a weather forecast is incomplete unless it is accompanied by a predictive statement about the skill of the forecast. Stochastic-dynamic prediction is considered for predicting the skill of dynamical forecasts made through numerical models. The Liouville equation (LE) is the basis of the temporal evolution of the pdf of the model state vector. Although, it is a linear partial differential equation, its explicit solution may pose a substantial problem from the analytical and numerical viewpoints. The approach currently under investigation concentrate on providing only partial information about the temporal development of pdf without explicitly solving the Liouville equation. This equation provides the theoretical framework and summarizes the information obtained. In this paper we consider the properties of Liouville equation for a simple, one-dimensional autonomous dynamical system.

2. The Liouville Equation

Let N be a finite number of variables X_i at a particular instant of the state of a dynamical system which pdf of the state vector X(t) is phase space in denoted byρ (X,t). Let:

2.1 $$\dot{X}_i = \dot{X}_i (X(t), t)$$

The LE is represented by the pdf ñ, and it is valid as long as no realizations are created or destroyed:

2.2 $$\frac{\partial \rho(X,t)}{\partial t} + \sum_{k=1}^{N} \frac{\partial}{\partial X_k} (\rho(X,t)\dot{X}_k(X,t)) = 0$$

The LE describing the temporal development of the pdf ⟩(X, t) in phase space is a homogenous quasi-linear first-order partial differential equation. Further, knowledge of ρ permits the solution of (2.2) which depends crucially on the dynamical system (2.1) under consideration through the direct insertion of the model dynamics into the LE. Thus, its solution must be expected to reflect qualitatively the characteristic behavior of the system.

3. Analytical and Numerical Results

Several features of the LE, and as well its usefulness in the context of forecasting skill are investigated. Let a simple, though nonlinear, one-dimensional autonomous dynamical system in the form of a Riccati equation:

3.1 $$\dot{X} = aX^2 + bX + c$$

with constant coefficients a, b, c such that:

3.2 $$\Delta = \frac{b^2}{4} - ac > 0$$

For the system (3.1) under consideration of LE, written in its general form in eq. (2.2), takes on the following specific form:

3.3 $$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial X}(\rho(aX^2 + bX + c)) = 0$$

This equation describes the temporal evolution of the role dependent variable, namely the pdf $\rho(X,t)$ as a function of the two independent variables X and t. The analytical solution of the LE (3.3) is explicitly found to be employing the method of characteristic (Zwillinger, 1989).

To study the analytical results in the context of predicting forecast skill, consider in the Ricatti equation (3.1) the case when the parameters take on the following values:

$$a = -1;\ b = 1\ ;\ c = 2$$

In this situation the two steady-state solutions to eq. (3.1) are $X_{s,1} = 2$ (stable) and $X_{s,2} = -1$ (unstable). For $Xo \geq 2$ all trajectories approach the stable steady-state solution $X_{s,1}$ without undergoing any singularity.

4. Conclusions

The Liouville equation represents the theoretical basis for dealing with imperfect initial conditions and model errors in the context of forecasting kill. Both kinds of uncertainty can be accounted for within the Liouville equation that describes the temporal evolution of the probability density function of the model state vector in phase space. The explicit analytical solution of the Liouville equation and its behavior have been derived and illustrated for a simple nonlinear model. It is found that the pdf reflects the basic characteristics of the dynamical model under consideration. The explicit solution is extreme valuable to check the validity of other approaches to assessing the statistics of the pdf. So, to investigate in greater detail the analytical approaches to solve the Liouville equation may be of considerable interest in more general situations than that considered in this paper, much results could be especially useful in identifying successful procedures for the purpose of forecasting forecast skill.

References

Palmer, T., and Tibaldi, S. (1988). On the prediction of forecast skill. *Mon. Wea. Rev.*, **116**, 2453-2480.

Nunes, A. M. D. (1992). Identification of nonlinear stochastic models from seismic records. In *Proceedings of the 5th International Meeting on statistical climatology*, 569-574, Toronto, Canada.

Pitcher, E. (1977). Application of stochastic dynamic prediction to real data. *J. Atmos. Sc.*, **34**, 3-21.

Walshaw, D. (2000). Modelling extreme wind speeds. *Appl. Statistics*, **49**, part 1, 51-

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Some General Remarks on the Analysis of Aggregated Environmental and Health Data

Markku Nurminen, Tuula Nurminen
Finnish Institute of Occupational health
Topeliuksenkatu 41 a A, 00250 Helsinki, Finland
Markku.Nurminen@occuphealth.fi, Tuula.Nurminen@occuphealth.fi

Epidemiologic studies of environmental exposures and their impacts on disease risk are an important and increasingly applied approach in health effects assessment (Nurminen, et al., 2000). However, environmental epidemiology often uses data that have been collected as temporal-spatial and demographic statistics, and thus are only available for analysis at the level of aggregate information (e.g. average exposure intensities and disease rates). The need to conduct aggregate-level studies springs primarily from the difficulty of obtaining high-quality, individual-level data on environmental exposures and extraneous covariates. Because of the special characteristics of grouped data, methodologic care must be exercised when links between environment and heath are analyzed.

The foremost requirement for valid inferences when one is linking and analyzing of environmental and health data is to ensure that the collected aggregate data are of good quality (Nurminen and Nurminen, 2000). Differences among datasets that have been constructed from publicly available databases may have arisen, for example, from variation in procedures for the selection of data items and in methods for handling missing data. However, in environmental epidemiology, considerable natural variation in the data may be expected, even for data relating to short time periods and distances (for example, in exposure measurements). Such variation is an intrinsic part of the environment-health system, and must be retained. Conversely, sampling and measurement errors must be identified, and either eliminated or assessed and controlled for. This may present major problems since the genealogy and quality of the data used in environment and health analysis are often unknown.

Although one of the premises in health and hazard surveillance is that existing data should be used when possible, this should not be seen as a justification for using inappropriate or invalid data. Typically, the use of erroneous data leads to further propagation of error, and even sophisticated and innovative statistical analysis cannot compensate for intrinsically poor data. At best, the results are uncertain, and allow no conclusions to be drawn. Moreover, even if the investigators discuss the limitations of the data when presenting their results to the scientific community, they may not always take these into account when making their policy recommendations. Data known to be faulty should therefore be rejected; data of uncertain quality should be evaluated carefully and then rejected if they cannot be validated adequately.

Suitable methods for linking aggregated environmental and health data must meet two criteria. First, they must be simple, inexpensive to implement and operable with the available data, thus allowing rapid assessment. Second, they must produce statistically valid and scientifically credible results if they are to be used as a basis for interventive action. This means that they should be unbiased and sensitive to the variations in the data at hand. Ideally, they should yield results that agree with those that would be obtained from individual-level studies, for which the statistical precision can be quantified in more detail. Applicable statistical methods include ecologic

analysis, time series analysis, and multilevel modeling (For a discussion of the application of these methods, see Nurminen, 2000). Recently, health risk assessment has accepted the use of epidemiologic (cohort and case-referent analysis) methods to quantify the impact of environmental and occupational exposures on public health (Nurminen, et al., 1999).

Unlike in traditional epidemiology, the aim of linking grouped environmental and health data aim is not to seek new environmental-health relations or confirm hypotheses; rather it is to use existing knowledge on such relations to help inform management and policy decisions, and raise awareness about the associations between environment and health. The linkage methods are thus used essentially as a means of describing and monitoring the relations between environment and health, and to help assess and demonstrate the existing risks to the population concerned (Corvalán, et al, 1997).

Any such data analysis must nevertheless be undertaken with care, for the relations between environment and health are often complex and fraught by uncertainties. On the one hand, this may lead to complacency and lack of action, if risks are not correctly identified. Assessing the risk of lung cancer due to exposure to diesel exhaust provides a current example of a situation in which the regulatory agencies appear to be in a state of 'paralysis by analysis' (Stayner, 1999). On the other hand it may cause unnecessary anxiety and fear, if non-existent risks are inferred. Data linkage thus needs to be recognized as a powerful but treacherous tool. Applied carefully and correctly, it can greatly strengthen decision-making; used carelessly, it will mislead. It is incumbent on the analyst, therefore, to ensure not only that environment and health linkage is conducted rigorously, but also that the results are presented and explained clearly and unambiguously.

Despite its limitations, aggregate data analysis may be the only feasible approach to estimating health outcomes of environmental exposures, for example, in regions where health monitoring is not undertaken, or for obtaining crude estimates of health impacts among very large populations. Many researchers are also faced with the problem of wishing to investigate individual-level relations, but having to use aggregate-level data, because of confidentiality or other restrictions on the availability of individual data. The application of aggregate data analysis as well as the development of new study designs and methods for data analysis are therefore important research needs in environmental epidemiology.

References

Corvalán, C., Nurminen, M. and Pastides, H. (eds). 1997. *Linkage Methods for Environment and Health Analysis.* Technical guidelines. WHO, Geneva.

Nurminen, M. (2000). Linking environment and health data: statistical and epidemiological issues. In *Decision-Making in Environmental Health.* From Evidence to Action (eds C. Corvalán, D. Briggs D and G. Zielhuis), 103-131, E & FN Spon.London.

Nurminen, M. and Nurminen, T. (2000). Methodologic issues in in linking aggregated environmental and health data, *Environmetrics*, **11**, 63-73.

Nurminen, M. Nurminen, T. and Corvalán, C.F. (1999). Methodologic issues in epidemiologic risk assessment. *Epidemiology*, **10**, 585-593.

Nurminen, T., Nurminen, M., Corvalán, C. and Briggs, D. (2000). Assessment of exposure and health effects. In *Decision-Making in Environmental Health. From Evidence to Action* (eds C. Corvalán, D. Briggs, D. and G. Zielhuis), 77-102, E & FN Spon. London.

Stayner, L. (1999). Protecting public health in the face of uncertain risks: The example of diesel exhaust. *Am J Public Health.* **89**, 991-993.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

How Long do Firms Spend in Bankruptcy?

Jesus Orbe, Eva Ferreira, Vicente Núñez-Antón
Universidad del País Vasco, Departamento de Estadística y Econometría
Euskal Herriko Unibertsitatea, Avenida Lehendakari Agirre 83, 48015 Bilbao, Spain
etporlij@bs.ehu.es

1. A Censored Partial Regression Model:

In this work, we propose a new method to consider situations where we have: (i) a response variable T with unknown probability distribution, (ii) censored observations, and (iii) several covariates where the effect of one of these, on the response variable is introduced in a nonparametric way. Let us assume that $T_1,\ldots,T_n$ are independent observations from some unknown distribution function F and, because of the censoring, not all of the T's are available. That is, rather than observing T_i, we observe $Y_i = \min(T_i, C_i)$ and the indicator d_i, which takes value 1 if the observation is not censored (i.e., if $T_i \le C_i$), and takes value 0 if the observation is censored. We are assuming that $C_1,\ldots,C_n$ are the values of the censoring variable C, which is independent of the variable T.

We consider a model where the effect of the covariates can be separated in two components: a parametric one and a nonparametric one,

$$\ln T_i = X_i b + h(r_i) + e_i$$

This proposal allows us to model situations where we do not know the functional form of the effect of one covariate on the response variable, or situations where the assumption of a lineal dependence, or any other different one is a restrictive assumption, or, maybe, it does not make any sense.

In order to estimate the model, we need to consider the goodness of the fit and the smoothness of the h function. This can be handled by minimizing the following penalized weighted least squares expression

$$\textbf{(1)} \qquad \sum_{i=1}^{n} W_{in}\left[\ln Y_{(i)} - X_i^T b - h(r_i)\right]^2 + a\int h''(r)^2\,dr$$

As for the goodness of the fit, this is controlled through the sum of the weighted squared residuals using the Kaplan-Meier weights W_{in}, where

$$W_{in} = \hat{F}_n(\ln Y_{(i)}) - \hat{F}_n(\ln Y_{(i-1)}) = \frac{d_i}{n-i+1}\prod_{j=1}^{i-1}\left[\frac{n-j}{n-j+1}\right]^{d_j},$$

being $\hat{F}_n$ a Kaplan-Meier (Kaplan and Meier (1958)) estimator of the distribution function F and $\ln Y_{(i)}$ is the i-th ordered value of the observed response variable. Thus, using these weights we take into account the existence of censored observations in the sample (Stute (1993)). As for the smoothness of h, we measure it using the integral of the square of the second derivatives. It can be shown that a smoothing cubic spline function is chosen to minimize (1) and, in this way, we can rewrite (1) as

$$(\ln Y - X\beta - Nh)^T W(\ln Y - X\beta - Nh) + a\, h^T Kh,$$

where h is a vector with the different values for each value of the covariate R. The estimation is carried out by taking derivatives with respect to both β and h.

The inference on the estimates can be carried out using bootstrap techniques. In order to do this, we propose a new procedure to generate the bootstrap resamples for the case of random censorship and heterogeneous model. The proposed bootstrap procedure is a very general one because it does not assume any model for the relation between the censoring mechanism and the covariates.

2. Aplication and Conclusions:

We illustrate the proposed methodology by analyzing the effect of several factors on the duration or time that firms spend under Chapter 11 of the U.S. Bankruptcy Code. After estimating the model and, using the posterior bootstrap study of the estimates, we can summarize the most relevant results. In relation with the estimation of the effect of the covariables introduced in a parametric way, we obtain, that the firms which have filed for prepackaged Chapter 11 spend less time in Chapter 11, the length of time negotiating before filing for Chapter 11 reduces the duration in Chapter 11, the more profitable firms emerge sooner from this situation, if the firm is involved in different disputes, it has more difficulties to leave Chapter 11, the firms that have realized highly leveraged transactions in the past leave bankruptcy before others. As for the estimation of the nonparametric component (the period of default), we obtain an increasing function up to 1985, then a decreasing function but with a deceleration on this decrease in the final part. Thus, this decreasing tendency indicates that the length of time spent in Chapter 11 is going to be shorter when we move the default date from the beginnings of the period under study, early eighties, to the end of the study in the early nineties. Thus, it seems that the reasons leading towards an effect of reduction of the duration in Chapter 11 are stronger than the reasons to increase the time spent under this situation. Therefore, this result may suggest that the courts and bankruptcies professionals have been acquiring more experience resolving different conflicts and this derives in faster negotiations. Other possible positive factor is the growing participation of vulture funds in reorganizations procedures. The final deceleration in the decrease could be reflecting the increment effect (larger durations) provoked by the sentence of the LTV firm and the change of tax treatments in 1990.

As a conclusion, we have to indicate that in the literature this problem has been studied using different approaches (Bandopadhyaya (1994), Li (1999), Helwege (1999), Orbe et al. (2001)). However, if we compare our work with the previous ones, here we propose a more flexible approach because: (i) we have not assumed any distribution for the duration, (ii) we have not assumed proportional hazard functions, (iii) the nonparametric component allows for more flexible study than if we had used indicator variables, and (iv) our method allows to have censored observations in the sample.

References

Helwege, J. (1999). How Long Do Junk Bonds Spend in Default?, *The Journal of Finance*, **44**, 341-357.

Kaplan, E.L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations, *Journal of the American Statistical Association*, **53**, 457-481.

Stute, W. (1993). Consistent Estimation under Random Censorship when Covariables are Present, *Journal of Multivariate Analysis.* **45**, 89-103.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

D-Optimal Designs for a Regression Curve

Isabel Mª Ortiz, Ignacio Martínez, Carmelo Rodríguez
Universidad de Almería, Dpto. Estadística y Matemática Aplicada
Campus Universitario de La Cañada, 04120 Almería, Spain
iortiz@ual.es, ijmartin@ual.es, crt@ual.es

1. Introduction

The regression curve models have received much attention in the literature of optimal designs for nonlinear models. We consider a specific case of these models, included in the growth curves model and it is expressed by:

(1) $$Y = b_1 + b_2 x + x^a + e\ .$$

The response variable is Y and e is the experimental error, with mean zero and variance s^2, which is assumed to be 1 without loss of generality. The controlled variable x is chosen from the interval $[a,b]$, with $0 < a < b$ and $q^T = (b_1, b_2, a)$ a vector of unknown parameters, a a integer nonnegative value.

The information matrix for nonlinear models depends on some unknown parameters and additional information about these parameters is necessary to calculate the optimal design. If this additional information is an initial value of the parameters, the optimal design is said to be Locally optimal design. If a prior distribution of the parameters is used then Bayesian optimal designs are obtained.

In this paper Locally and Bayesian three-point D-optimal designs are characterized and their support points are calculated for model (1). These designs have the minimum number of support points, equal to the number of unknown parameters. So these D-optimal designs are equally supported (Fedorov, 1972).

2. D-optimal Designs

Let be $Y = f(x,q) + e$, and $\nabla f(x,q)$ the gradient vector, then for an approximate design x with n support points $x_1 < x_2 < ... < x_n$ and weight $x(x_i)$ for each, for the best guess for the parameter $q = q^0$, the information matrix is written

$$M(x, q^0) = \sum_{i=1}^{n} x(x_i) \nabla f(x_i, q^0) \nabla f(x_i, q^0)^T$$

and the variance function is defined as

$$d(x, x, q^0) = \nabla f(x, q^0)^T M^{-1}(x, q^0) \nabla f(x, q^0)\ .$$

D-optimality criterion searches the design that maximizes the determinant of the information matrix. For independent observations, $M^{-1}(x,q)$ is proportional to the asymptotic covariance matrix for the maximum likelihood estimate for q. So, this criterion minimizes the generalized variance of the parameter estimates.

For the model (1), to obtain D-optimal designs, only additional information about the parameter a is necessary, because the information matrix is only function of this parameter.

Proposition The Locally D-optimal design supported at three different points for the guess $a = a^0$ in the design space $[a,b]$, $0 < a < b$, is equally supported on the points $\{x_1, x_2, x_3\}$, two of which are the extremes of the interval, i.e. $x_1 = a$ and $x_3 = b$, and x_2 is the solution of the following equation:

$$x_2^{a^0-1}\left(1+a^0 \ln(x_2)\right) = \frac{b^{a^0} \ln(b) - a^{a^0} \ln(a)}{b-a}.$$

The value x_2 depends on the extremes of the design space and the parameter a. Locally D-optimal designs have been calculated for different values of parameter a. In Figure 1, the different values of x_2 are shown for several integer nonnegative values of a, with design space $[0.1\ ,\ 5]$.

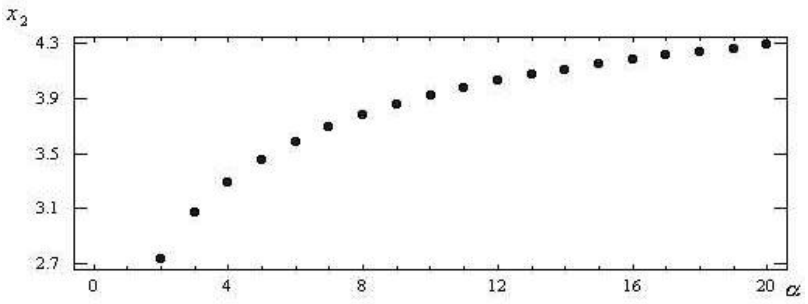


Figure 1. Support point x_2 as function of a.

Bayesian D-optimal designs maximize the expectation of the determinant of the information matrix over a prior distribution on the parameters. If the prior distribution used is discrete consists of p points, then the Bayesian D-optimal design for a model with k unknown parameters is supported at most at $pk(k+1)/2$ different points (Dette and Neugebauer, 1996). There is not a similar property for a continuous prior distribution.

Several numerical examples have been considered for different discrete prior distributions for the parameter a. In all these cases, Bayesian D-optimal designs obtained are supported at three different points with equal weights.

References

Dette, H. and Neugebauer, H.M. (1996). Bayesian optimal one point designs for one parameter nonlinear models, *Journal of Statistical Planning and Inference*, **52**, 17-31.

Fedorov, V.V. (1972). *Theory of Optimal Experiments.* Academic Press, New York.

Martínez, I., Ortiz, I. and Rodríguez, C. (2001). Optimum Experimental Designs for a Modified Inverse Linear Model. In *Proceedings of the 6th International Workshop on Model-Oriented Data Analysis* (eds P. Hackl and W.G. Müller), 171-181. Physica, Heidelberg.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Non-Uniform Cauchy Approximations for Windings of the Planar Brownian Motion

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

Vidmantas Bentkus
Institute of Mathematics and Informatics
Akademijos 4., 2600 Vilnius, Lithuania

Marc Yor
Laboratoire de Probabilites, CNRS URA 224, Universite Paris VI
4, Place Jussieu, Tour 56, 3 etage, F-75252 Paris Cedex 05, France

Consider a 2-dimensional Brownian motion $(Z_t)_{t\geq 0}$ starting from $Z_0 = (1,0)$ (for simplicity), and let $(q_t)_{t\geq 0}$ denote the continuous determination of the argument of $(Z_u)_{u\leq t}$ around $(0,0)$ as t evolves in $[0,\infty)$ such that $q_0 = 0$.

Spitzer (1958) proved the celebrated result:

$$\lim_{t\to\infty} P\left(\frac{q_t}{\log\sqrt{t}} < x\right) = G_0(x), \qquad x \in \mathbf{R},$$

where G_0 denotes the standard Cauchy distribution function:

$$G_0(x) = \frac{1}{2} + \frac{1}{p}\arctan x, \qquad G_0'(x) = \frac{1}{p\left(1+x^2\right)}, \qquad x \in \mathbf{R}.$$

Pap and Yor (2000) showed that for all $k \in \mathrm{N}$ the following estimate is valid:

$$\sup_{x\in\infty}\left|P\left(\frac{q_t}{\log\sqrt{t}} < x\right) - G_0(x) - \sum_{j=1}^{k-1}\frac{a_j}{\left(\log\sqrt{t}\right)^j}G_j(x)\right| \leq c_k\left(\log t\right)^{-k}$$

with an absolute constant c_k, with

$$G_j(x) = \frac{d^j}{dy^j}\bigg|_{y=1} G_0\left(\frac{x}{y}\right)$$

and with some coefficients a_k which can be computed explicitly. We remark that the Fourier-Stieltjes transform of the function G_j is

$$\int_{-\infty}^{\infty} e^{i l x} dG_j(x) = \left(-|l|\right)^j e^{-|l|}.$$

The aim of the present paper is to prove the following nonuniform bound for the remainder terms in Spitzer's theorem.

Theorem For every $k \in \mathrm{N}$ we have

$$\sup_{x\in\mathbf{R}}\left(1+|x|^k\right)\left|P\left(\frac{q_t}{\log\sqrt{t}}<x\right)-G_0(t)-\sum_{j=1}^{k-1}\frac{\tilde{a}_j(t)}{\left(\log\sqrt{t}\right)^j}G_j(x)\right|\le c_k\left(\log t\right)^{-k}$$

with an absolute constant c_k and with

$$\tilde{a}_j(t)=\frac{e^{-1/(2t)}}{j!}\sum_{l=1}^{\infty}\frac{h_l^{(j)}(0)}{l!t^l},\qquad h_l(u)=\frac{2^{u/2-l}\Gamma(1+l-u/2)}{\Gamma(1+l-u)}.$$

We note that the bound in our theorem is optimal in the sense that it is impossible to replace $1+|x|^k$ by $1+|x|^{k+e}$ (respectively $(\log t)^{-k}$ by $(\log t)^{-k-e}$) with an $e>0$. We also note that the functions $\tilde{a}_j$, $j\ge 0$ are bounded and $\lim_{t\to\infty}\tilde{a}_j(t)=a_j$, $j\in \mathrm{N}$.

The spirit of our expansion is close to the Edgeworth type expansions in the central limit theorems, that is, to the asymptotic expansions of the distribution function F_n of $(X_1+\cdots+X_n-a_n)/b_n$, where (X_n) is a sequence of independent identically distributed random variables, and (a_n), (b_n) with $b_n>0$ are sequences of constants such that F_n tends to some stable law. The question of asymptotic expansion in case of a non-normal stable law is less studied; see, e.g., Christoph (1981), Christoph and Wolf (1993). For example, under some condition on the distribution of X_1, F_n tends to the standard Cauchy law G_0 as $n\to\infty$ where $a_n=0$ and $b_n=n$, and

$$\sup_{x\in\mathbf{R}}\left(1+|x|^{k+1}\right)\left|F_n(x)-G_0(x)-\sum_{j=1}^{k-1}\tilde{Q}_j(x)n^{-j}\right|\le c_k n^{-k},$$

where the functions $\tilde{Q}_j$ are linear combinations of G_l, $l=j+1,\ldots,2j$. Particularly,

$$\sup_{x\in\mathbf{R}}(1+x^2)\left|F_n(x)-G_0(x)\right|\le c_1 n^{-1},$$

but in our Theorem we have slightly slower speed:

$$\sup_{x\in\mathbf{R}}\left(1+|x|\right)\left|P\left(\frac{q_t}{\log\sqrt{t}}<x\right)-G_0(x)\right|\le c_1\left(\log t\right)^{-1}.$$

References

Christoph, G. (1981). Asymptotic expansion in the case of stable law I., *Lithuanian Math J.*. **21**, 137-145.

Christoph, G. and Wolf, W. (1993). *Convergence Theorems with a Stable Limit Law.* Akademie Verlag, Berlin.

Pap, G. and Yor, M. (2000). The accuracy of Cauchy approximation for the windings of planar Brownian motion, *Periodica Math. Hungar.* **41** (1-2), 213-226.

Spitzer, F. (1958).Some theorems concerning 2-dimensional Brownian motion, *Trans. Amer. Math.Soc.* **87**, 187-197.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Asymptotic Error Rates in the Discriminant Analysis using Feature Selection

Tatjana Pavlenko
Department of Physics and Mathematics
Mid Sweden University, 851 70 Sundsvall, Sweden
tatjana@fmi.mh.se

Considering feature selection in discriminant analysis can be useful in situations where measuring all features relevant to the classifcation is "expensive". It is likely that the omission of certain features or sets of features, while naturally destroying the possible optimality of standard discriminant analysis, will not seriously effect the error probability (or any other criterion of interest). The effect of such selection on the discrimination error is evaluated in a *growing dimension asymptotics*, i.e., in the case when the relationship between dimensionality of observations p and population sample size n satisfy the condition $\lim_{n\to\infty} p/n = p$, where p is a certain constant. We consider the case in which the populations are represented by densities of the form $L\left(\mathrm{x},q^n\right)=\prod_{i=1}^{K} L\left(\mathrm{x}_i,q_i^n\right), n=1,2$, which means that both the observed vector x and vector of parameters q^n are decomposable into a family of K mutually disjoint, independent subsets (blocks) of size m. In this case the optimal decision rule minimizing the overall error probability can be obtained by applying Bayes theorem, which gives the discriminant rule

$$\mathrm{D}\left(\mathrm{x},q^1,q^2\right)=\ln\frac{L\left(\mathrm{x},q^1\right)}{L\left(\mathrm{x},q^2\right)}=\sum_{i=1}^{K}\ln\frac{L\left(\mathrm{x}_i,q_i^1\right)}{L\left(\mathrm{x}_i,q_i^2\right)}=\sum_{i=1}^{\mathrm{K}}\mathrm{D}\left(\mathrm{x},q_i^1,q_i^2\right)\quad d_0,$$

where d_0 is a specifed point independent of x.

The common need for subset selection procedures, is an evaluation function by which a discriminating power of a feature, or a subset of features, is assessed. We consider an evaluation function based on a distance measure and focus on the Jeffreys distance, which under the preceding assumptions turns out to be

$$J(n)=\int\ln\frac{L\left(\mathrm{x},q^1\right)}{L\left(\mathrm{x},q^2\right)}\left(L\left(\mathrm{x},q^1\right)-L\left(\mathrm{x},q^2\right)\right)m(d\mathrm{x})=\sum_{i=1}^{K}J_i(n)==\sum_{i=1}^{K}\left(J_i^1(n)-J_i^2(n)\right),$$

where $J_i^n(n)=\int\ln\frac{L_i\left(\mathrm{x}_i,q_i^1\right)}{L_i\left(\mathrm{x}_i,q_i^2\right)}L\left(\mathrm{x},q^n\right)m(d\mathrm{x}), n=1,2, i=1,\ldots,K$. The discriminating power the ith set of features (or ith block) is assessed by its *informativeness*, $\frac{nJ_i(n)}{2}$. Then the threshold based feature selection procedure is represented by means of an *inclusion-exclusion* factor defined by $1_{\left\{y_n^2,\infty\right\}}\left(\frac{n\hat{J}_i(n)}{2}\right)$ where $\hat{J}_i(n)$ is the plug-in estimate of $J_i(n)$ and y_n^2 is a given *threshold* such that $\lim_{n\to\infty} y_n^2 = y^2$. This procedure is embedded into the discriminant function by the following modification

$$\mathrm{D}_{y_n^2}\left(\mathrm{x},\hat{q}^1,\hat{q}^2\right)=\sum_{i=1}^{K}1_{\{y_n^2,\infty\}}\left(\frac{n\hat{J}_i(n)}{2}\right)\mathrm{D}_i\left(\mathrm{x}_i,\hat{q}_i^2,\hat{q}_i^2\right).$$

It is clear that using factor $1_{\{y_n^2,\infty\}}$ in the discrimination procedure we retain only the sets of features whose informativeness exceed the given threshold y_n^2. For this type of discrimination problem, the following limiting relationship between the *fraction of selected features*, $h\left(y^2\right):=\lim_{n\to\infty}\frac{k\left(y_n^2\right)}{K}$ and the given threshold $y^2=\lim_{n\to\infty}y_n^2$ is established in the growing dimension asymptotics:

$$h\left(y^2\right)=\int\left(1-\mathrm{F}\ \left(y^2;m,g^2\right)\right)dH\left(g^2\right),$$

where $\mathrm{F}\ \left(y^2;m,g^2\right)$ is the non-central c^2 distribution with *m* degrees of freedom and non-centrality parameter g^2 and *H* is the distribution of feature informativeness.

Furthermore, the asymptotic normality of the modified discriminant function is proved, which makes it possible to establish the limiting error rate,

$$e_1=P\left\{\mathrm{D}_{y_n^2}\le d_0\mid x\in\prod_1\right\}\to\Phi\left(-\frac{E\left(y^2\right)-d_0}{\sqrt{D\left(y^2\right)}}\right),$$

as $n\to\infty$, where Φ is the standard normal distribution function and

$$E\left(y^2\right)=r\int g^2\left[1-\mathrm{F}\ \left(y^2;m+2,g^2\right)\right]dH\left(g^2\right),$$
$$D\left(y^2\right)=2r\int\left[g^2\left(1-\mathrm{F}\ \left(y^2;m,g^2\right)\right)+m\left(1-\mathrm{F}\ \left(y^2;m,g^2\right)\right)\right]dH\left(g^2\right)$$

are the moments of the asymptotic distribution of $\mathrm{D}_{y_n^2}$. Note that $e_2=1-e_1$. When selecting features by means of function $1_{\{y_n^2,\infty\}}$, the discrimination errors are affected by two factors: one is the selection itself and the other reflects the estimation error induced by high dimensionality. This *combined effect* is studied using obtained asymptotic expressions for e_1 and e_2.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Analysis of Spatial Point Processes Based on The Output of Clustering Algorithms

Sandra M.C. Pereira
The University of Western Australia
Department of Mathematics and Statistics
35 Stirling Hwy, Crawley, WA, 6009, Australia
spereira@maths.uwa.edu.au

We propose a new strategy for analysing spatial point patterns. Algorithms of classical multivariate cluster analysis are applied to the point pattern data, and the output of the algorithms is used as a spatial summary statistic. In this paper we apply hierarchical clustering algorithms to the Euclidean distances between the points of a pattern, and take the resulting list of `fusion distances' as a summary statistic. It is demonstrated that this statistic is very good at discriminating between different types of spatial patterns. Several graphical techniques are proposed for exploratory data analysis, and formal inference is made using a modification of the standard Monte Carlo test.

1. Statistics of Spatial Point Patterns

A spatial point pattern is a set of points irregularly distributed within a region of interest. Statistical methods for such datasets (e.g. Diggle, 1983) are usually based on `summary statistics' of the point pattern. Most of the popular summary statistics are functions, for example, the empirical distribution function of the Euclidean distances between all pairs of points of the pattern. They are usually motivated by intuition, but have a rigorous interpretation as unbiased estimators of characteristics of the point process, under the assumption of spatial homogeneity. They may also be sufficient statistics under a parametric model. However, very little is known about the statistical behaviour of these summary functions, and in practice they often perform poorly at discriminating between different types of point patterns. Hence there is an interest in alternative methods.

2. Multivariate Clustering

Clustering algorithms (e.g. Everitt, 1993) are designed to partition a multivariate dataset into groups or clusters. Hierarchical clustering algorithms make a series of successively coarser partitions, starting with the lowest level in which each cluster consists of a single data point, and ending with a single cluster containing the entire dataset. At each intermediate stage the algorithm fuses the two clusters that are most similar according to some criterion. Important examples are the Single Linkage, Average Linkage and Complete Linkage criteria. These classical multivariate clustering techniques usually do not have a rigorous statistical interpretation in terms of a stochastic model. However, they have a strong pragmatic advantage in that the most popular clustering techniques have been found to work well on a large variety of real datasets.

3. New Strategy

In the hope of developing effective practical techniques for discriminating between different types of point pattern, we propose the following `hybrid' strategy.

1. A classical multivariate clustering algorithm is applied to the point pattern dataset, using the Euclidean distances between the points as the dissimilarity measures.
2. The output of the clustering algorithm is used as a spatial summary statistic.
3. For exploratory data analysis or formal inference, this summary statistic is compared with the results of the same statistic applied to simulations from the Poisson point process.

In the present paper, step 2 is implemented by taking the summary statistic to be the list of `fusion distances' chosen by the hierarchical algorithm. The k-th fusion distance is the dissimilarity between the two clusters that are fused at level k+1, for k = 1,…, n-1 where n is the number of points in the dataset. Although the fusion distances are stochastically dependent, the empirical cumulative distribution function (edf) of the fusion distances, $H(t)$, may be regarded as a summary statistic in the same sense as the popular spatial statistics $F(t)$, $G(t)$ and $K(t)$.

Initial experiments show that $H(t)$ is very good at discriminating between different types of spatial pattern in standard, real, datasets.

We argue that $H(t)$ should be compared with $\overline{H}(t)$, the average of the edf's of fusion distances from m simulations of the binomial point process (n independent uniform points in the same observation region). For graphical comparisons one may use a P-P or Q-Q style plot, angular transformations of the P-P plot, or relative distribution plots (Handcock and Morris, 1999). Formal inference can be performed using a version of the Monte Carlo hypothesis test. We describe graphical methods for performing this test on the P-P and Q-Q style plots.

In conclusion it appears that this approach is a very effective alternative to currently accepted methods in spatial statistics, with the disadvantage that we lack theoretical insight into its performance.

References

Diggle, P. J. (1983). Statistical Analysis of Spatial Point Patterns. Academic Press, London.

Everitt, B. S. (1993). Cluster Analysis. Edward Arnold, London, Melbourne, Auckland.

Handcock, M. S. and Morris, M. (1999). Relative Distribution Methods in the Social Sciences. Springer-Verlag, New York.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Reward Study of a Repairable Model With Three Types of Failures

Rafael Pérez Ocón
Universidad de Granada, Departamento de Estadística e I. O.
Avenida de Severo Ochoa s/n, 18.071 Granada. España
rperezo@ugr.es

Inmaculada Torres Castro
Universidad de Extremadura¸Departamento de Matemáticas
Avenida de Elvas s/n, 06.071 Badajoz. España
inmatorres@unex.es

Delia Montoro Cazorla
Universidad de Jaen, Departamento de Estadística e I. O.
Alfonso X el Sabio, 23600 Linares. Jaen. España
dmontoro@ujaen.es

A repairable unit-system with repair not "so good as new", replacement policy N (when the system has just repaired N times, in the next failure is replaced for a new one), submitted to three types of failures, is considered. Successive operational times define a geometric process, repair times define a renewal process, and the operational and repair times follow phase-type distributions. When the unit is operational a reward of A_0 m.u per unit time is obtained, when it is in repair there is a loss of A_R m.u per unit time, and when it is replaced for a new one a loss of A* m.u. is produced. The optimal replacement policy N is determined in the long-run average cost per unit time.

1. The model

Assumptions 1 Unit undergoes three types of failures. External failures occur according to a Poisson process of rate λ. Bernoulli trials determine whether these failures are repairable (with probability p) or non-repairable (with probability q = 1−p). The unit can also fail by ageing. That is always a non-repairable failure. A repairman attends to the repairable failures. Non-repairable failures require replacing the unit by a new one. We assume that all failures occur independently of each other.

Assumptions 2 We represent by X_n operating times after (n-1)-repair. Following repair the unit returns to service but it is not as good as new and suffers deterioration in its lifetime modelled by a Geometric Process. We assume X_n follows a PH distribution with representation $(\alpha, a^{n-1}T)$, where $T \in M_m$ a >1.

Assumptions 3 $\{Y_n\}$ is repair time after n-failure. We assume $\{Y_n\}$ is a renewal process and Y_n follows a PH distribution with representation (β, S) where $S \in M_n$ and $E(Y_n) = \mu_r$. We assume $\{X_n, n=1,2,..\}$ and $\{Y_n, n=1,2,\ldots\}$ are independent.

Assumptions 4 When unit is operating is obtained a reward (A_o) u.m./u.t , when unit is repairing is produced a cost of (A_r) u.m/u.t. When the unit is replaced by a new one, it is produced a cost of (A^*) u.m.

Let $\{X(t), t \geq 0\}$ be the stochastic process that represents the state of the system in time t.$\{X(t), t \geq 0\}$ is a Markov Process (Neuts et. al 2000). Its space state is given by $S=\{0, 1_R, 2, 2_R, \ldots N\}$. We said the process is in state *i* when unit is operative and it

has completed *i* repairs, and the process is in state i_R when unit is being repaired and it has completed (i-1) repairs before. Steady state distribution is given by

$$\delta_i e = Kp^i \lambda^{-1} \prod_{k=o}^{i} \left(1-\varphi\left(\frac{\lambda}{a^k}\right)\right) \quad i=0, \ldots, N,$$

$$\delta_{i_R} e = Kp^i \pi_r \prod_{k=0}^{i-1} \left(1-\varphi\left(\frac{\lambda}{a^k}\right)\right) \quad i=1, \ldots N$$

where $\varphi(\lambda;a)$ is the Laplace Transform of the density of a PH distribution with representation $(\alpha,\mathbf{T})$. We note by $D_i = p^i \prod_{k=0}^{i} \left(1- \left(\frac{\lambda}{a^k}\right)\right)$

2. Reward Problem

We call R(t) the reward function per unit time at the time t. It is known (Ross, 1983) that

$$\frac{R(t)}{t} \to \frac{E(R)}{E(T)},$$

where E(R) is the reward expected in a cycle and E(T) is the expected time of the cycle. We want to find N that maximizes $R(N) = \frac{E(R)}{E(T)}$ where R(N) is given by

$$R(0) = \frac{-A^* + A_o \lambda^{-1} D_0}{\lambda^{-1} D_0} \qquad R(N) = \frac{-A^* + A_o \lambda^{-1} \sum_{i=0}^{N} D_i - A_r \pi_r \sum_{i=0}^{N-1} D_i}{\lambda^{-1} \sum_{i=0}^{N} D_i + \pi_r \sum_{i=0}^{N-1} D_i} \qquad \forall N \geq 1$$

$$R(1)\text{-}R(0) \geq 0 \Leftrightarrow \frac{A^* \lambda D_0 \pi_r D_0}{(A_r + A_o)\lambda D_0 + \pi_r D_0} \geq B(0)$$

$$R(N+1) - R(N) \geq 0 \Leftrightarrow \frac{A^*}{A_r + A_o} \geq \frac{\lambda^{-1} \pi_r (D_N \sum_{i=0}^{N} D_i - D_{N+1} \sum_{i=0}^{N-1} D_i)}{\lambda D_0 + \lambda^{-1} \pi_{N+1} D_N + \pi_r D_N} = B(N)$$

It is showed that $B(N)$ are positives and strictly increasing for all N.
Therefore

$$N_{opt} = \min_{N \geq 0} \left\{ \frac{A^*}{A_0 + A_r} \leq B(N) \right\}.$$

References

Neuts M.F., Pérez-Ocón R., Torres-Castro, I. (2000). Repairing model with operating and repairing times governed by phase type distributions. *Adv. Appl. Prob.* .**32,**2.

Ross, S. (1983). Stochastic process. Wiley.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Approximating Posterior Distributions Using Quasi-Monte Carlo Methods

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

J. Carlos Rojano
Universidad de Málaga, Department of Statistics
rojano@uma.es

1. Introduction

One of the most important problems in Bayesian analysis is computing the posterior distributions. As Berger (2000) points out most Bayesian computations are focused on the calculation of posterior expectations, which are mainly integrals. Let p be a prior distribution, l the likelihood function, and x the observation from an experiment. Then the posterior distribution is:

(1) $$p(q \mid x) = \frac{l(q \mid x)p(q)}{\int_{\Theta} l(q \mid x)p(q)dq} = \frac{l(q \mid x)p(q)}{m(x)}$$

and there are several methods to compute $\int_{\Theta} g(q)p(q|x)dq$ for some functions $g(q)$.

In this work we compare some stochastic simulation methods for one-dimensional posterior distributions. Furthermore, we improve some of the methods. First, we adapt quasi-Monte Carlo (QMC) aproximations. Second, we apply the ratio of uniform deviates method in order to approximate posterior distributions. This technique allows us to compute posterior quantities generating values directly from the posterior distribution.

2. Quasi-Monte Carlo methods

Monte Carlo (MC) methods are statistical sampling techniques that have been extensively applied to approximate integrals and other purposes. QMC methods can be considered as deterministic versions of those. In the integration problem, an advantage of QMC methods is that they provide deterministic error bounds, as given by the Koksma-Hlawka inequality, with a better order than the Monte Carlo´s, which is probalistic, see Niederreiter (1992). QMC methods emphasize "uniformity" of the points, instead of "randomness" of them. The concept of discrepancy, which measures the uniformity of a set of points, is crucial to these methods. The more general concept of F-discrepancy, a measure of the representation of a set of points with respect to a distribution F, was suggested by Wang and Fang (1990). The so-called low-discrepancy point sets are used.

We compare the results obtained by using pseudo-random and quasi-random numbers applying methods like inverse transformation, ratio of uniforms and weighted bootstrap. The low-discrepancy point sets are used to compute quasi-random numbers, which yield better approximations to the posterior distribution. These improved approximations are reflected by the attainment of better estimations of posterior expectations.

3. Ratio of Uniform

The method of ratio of uniform deviates is based on the following result:

Theorem Let $C_h=\{(u,v): 0\leq u\leq (h(v/u))^{1/2}\}$ for any nonnegative function h such that $0<\int h<\infty$. Then C_h has finite volume and if we can generate (U,V) uniformly over C_h, then $X=V/U$ has density function $f=h/\int h$.

Notice that when applying this method in order to generate from posterior distributions we don't need to compute $m(x)$ in (1). This is a very important fact because in many problems it is impossible to compute analytically $m(x)$. In this work, we apply the method to unidimensional posterior distributions, but it is easy to extend, at least theoretically, to multidimensional settings. Note that actually there are too many researchers working on Monte Carlo methods based on Markov chains to compute posterior distributions. But these methods only approximate posterior distributions. We provide several methods to generate uniformly from C_h. One of them is based on adaptative sampling, see Thompson (1992). We also use quasi-random numbers to generate uniformly on C_h. Although the theoretical results are not encouraging, the applications to some examples produce better results than using pseudo-random numbers.

4. An example

Consider a prior *Beta*(1,10). Assume that the likelihood is *Binomial*(20,q). Then we know that the posterior distribution is *Beta*($x+1$, 20+10-x). We apply the proposed methods. For example, Figure 1 shows posterior distribution using ratio of uniform variates with pseudo-random numbers and quasi-random numbers for x=2.

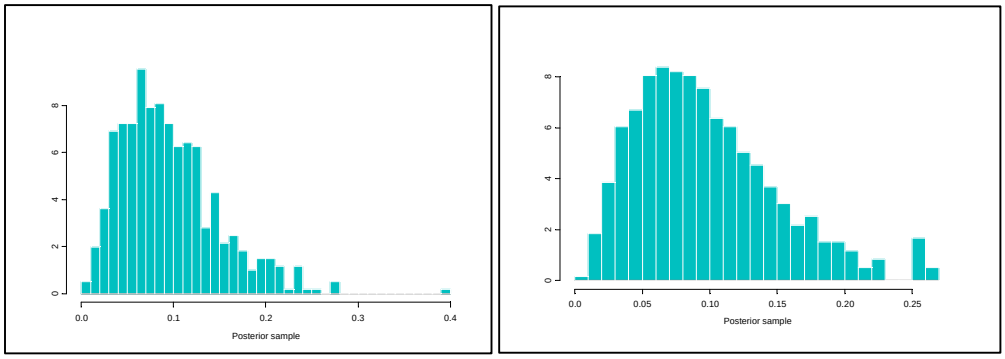


Figure 1. Histogram of the posterior distribution using ratio of uniforms.

We provide comparative results from different approximation methods in this and other examples. The empirical work shows that the QMC approximations improve the results obtained using pseudo-random numbers in terms of the F-discrepancy and the error of the estimation.

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

References

Berger, J. O. (2000). Bayesian analysis: A look at today and thoughts of tomorrow. *J. Amer. Stat. Assoc.* **95**, 452, 1269-1276.

Wang, Y and Fang, K. T. (1990). Number theoretic methods in applied statistics. *Chinese Ann. Math.* Ser.B, **11**, 41-55.

Niederreiter, H. (1992). Random number generation and quasi-Monte Carlo methods. *Society for Industrial and Applied Mathematics.* Philadelphia, Pennsylvania.

Thompson, S. K. (1992). *Sampling.* Wiley. New York.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Parseval's Relation and Self-Reciprocal Characteristic Functions

Dinis Duarte Pestana, Fernando Sequeira
Dept. de Est. da Univ. de Lisboa e Centro de Est. e Apl. da Universidade de Lisboa
dinis.pestana@fc.ul.pt

Sílvio Filipe Velosa
Universidade da Madeira e Centro de Estatística e Aplicações da Universidade de Lisboa
svelosa@math.uma.pt

Reciprocal pairs of characteristic functions such as $\left[e^{-|x|}, \frac{1}{1+t^2}\right]$ and self-reciprocal characteristic functions such as the Gaussian characteristic function $e^{-\frac{t^2}{2}}$ are well-known, and have been studied by Lévy (1967), Feller (1971), Teugels (1975), Lewis (1975), Shanbhag (1977) and Pestana (1978).

The topological properties of the convex pointed cone of positive and integrable characteristic functions could be the basis for an Krein-Milman-Choquet integral representation, but unfortunatelly extremal rays are not easily identified. An interesting characterization of random variables Y having self-reciprocal characteristic function uses the characteristic function of ZY, where Z is standard Gaussian independent of Y: j satisfies the functional equation $j(t) = \frac{1}{|t|} j\left(\frac{1}{t}\right)$ for some non-empty set (t_0,t_1)-{0}.

If a reciprocal pair [f,g] is known, we can derive a self-reciprocal characteristic function,

$$j(t) = \frac{\sqrt{g(0)}\, f(t) + \sqrt{f(0)}\, g(t)}{\sqrt{f(0)} + \sqrt{g(0)}},$$

and this accounts for the fact that self-reciprocal characteristic functions are in general the sum of two very different analytic expressions. The only exceptions we have found are the Gaussian, $e^{-\frac{t^2}{2}}$, and the hyperbolic cosine characteristic function $\frac{1}{\cosh\left(\frac{p}{2}t\right)}$.

Besides their intrinsic interest, self-reciprocal characteristic functions may be used to establish important results in Probability Theory. For instance, using Parseval's relation $\int_{-\infty}^{\infty} e^{-ity} j_X(t) dF_Y(t) = \int_{-\infty}^{\infty} j_Y(x-y) dF_X(x)$ with Y=Z/σ, Z the standard Gaussian random variable, in view of the fact that $e^{-\frac{t^2}{2}}$ is the characteristic function corresponding to the probability density function $\frac{e^{-\frac{x^2}{2}}}{\sqrt{2p}}$, we obtain

* Research partially supported by FCT/POCTI/FEDER.

$$\frac{1}{2p}\int_{-\infty}^{\infty} e^{-ity} j_X(t) e^{-\frac{s^2 t^2}{2}} dt = f_{X+sZ}(y),$$

one of the most elegant ways of establishing the unicity, inversion formula and continuity theorems for the characteristic functions.

We use Parseval's relation together with the self-reciprocal characteristic function $\frac{1}{\cosh\left(\frac{p}{2}t\right)}$ an alternative inversion theorem, and correlated results in Probability Theory.

References

Choquet, G. (1969) *Lectures in Analysis*, vol II. Benjamin, New York.

Feller, W. (1971) *An Introduction to Probability Theory and Some of its Applications*, vol. II, Wiley, New York.

Krein, M. and Milman, D. (1940) On extreme points of regular convex sets. *Studia Mathematica* **9**, 133-138.

Lévy, P. (1967) Fonctions caractéristiques positives. *C. R. Acad. Sci. Paris* **265A**, 249-252.

Lewis, T. (1975) Probability functions which are proportional to characteristic functions and the infinite divisibility of the von Mises distribution. *Perspectives in Probability and Statistics* (J. Gani, ed.), Academic Press, New York.

Pestana, D. D. (1978) *Some Contributions to Unimodality, Infinite Divisibility, and Related Topics*. Sheffield.

Phelps, R. R. (1966) *Lectures on Choquet's Theorem.* van Nostrand, New Jersey.

Shanbhag, D. N. (1977) On a conjecture of T. Lewis. *Bull. Austral. Math. Soc*. **2**, 253-255.

Teugels, J. (1975) On self-reciprocal characteristic functions. Bull. Soc. Belge Math.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Properties of Polynomial-Gaussian Processes

Agnieszka Pluciñska
Warsaw University of Technology
pawlicka@alpha.mini.pw.edu.pl

Let $X_d = (X_1,...,X_d)$ denote a d-dimensional random variable with a nondegenerate distribution. We suppose that X_d has a polynomial-Gaussian distribution (PGD_d), i.e. the density of X_d is the product of a non-negative polynomial in x_1 and a d-dimensional Gaussian density. It is known that every polynomial can be represented as a linear combination of Hermite polynomials H_r. Thus throughout the paper the density of X_d will be

(1) $$f(x_d) = \frac{\sqrt{\det A}}{(2p)^{\frac{d}{2}}} \sum_{r=0}^{2l} \frac{c_r}{k_{11}^{\frac{r}{2}}} H_r\left(\frac{x_1}{\sqrt{k_{11}}}\right) \exp\left[-\frac{1}{2}(Ax_d, x_d)\right] = p_{2l}(x_1)\, \tilde{f}(x_d)$$

where $A = K^{-1}$, $K = [k_{rs}]_{r,s=1}^{d}$ is a symmetric, positive definite $d \times d$ matrix, $x_d = (x_1,...,x_d) \in \mathbb{R}^d$, p_{2l} is a non-negative polynomial in x_1 of degree $2l$, $\tilde{f}(x_d)$ denotes a d-dimensional Gaussian density, H_r is the Hermite polynomial of degree r.

The characteristic function of X_d has the following form:

$$j(x_d) = E\exp[i(x_d, X_d)] =$$

(2) $$= \sum_{r=0}^{2l} c_r (ih)^r \exp\left[-\frac{1}{2}(Kx_d, x_d)\right] = \Psi_{2l}(h)\, \tilde{j}(x_d),$$

where $h = \frac{1}{k_{11}}[x_1 k_{11} + ... + x_d k_{1d}]$, $\Psi_{2l}(h)$ is a polynomial of degree $2l$ and $\tilde{j}(x_d)$ denotes the characteristic function of a Gaussian distribution.

We say that $c = (X_t, t \geq 0)$ is a polynomial Gaussian process (PGP) if for every $d \geq 1$ and every $t_1 < ... < t_d$ the vector $X_d = (X_{t_1},...,X_{t_d})$ has PGD_d given by (1) where $k_{rs} = \mathrm{K}(t_r, t_s)$. We denote by F_t the natural filtration of c.

Every one-dimensional distribution of PGP is PGD_1. The ch.f. of X_{t_s} has the folllowing from

(3) $$j(x_s) = E\exp(ix_s X_{t_s}) = \sum_{r=0}^{2l} c_r \left(ix_s \frac{k_{1s}}{k_{11}}\right)^r \exp\left(-\frac{1}{2}k_{ss}x_s^2\right)$$

Now we are going to construct quite general examples of PGP.

Let $\tilde{c} = (\tilde{X}_t, t \geq 0)$ be a centred Gaussian process with covariance function $k_{rs} = \mathrm{K}(t_r, t_s)$. Let X_{t_s} be a r.v. with ch.f. given by (3) where $k_{11} > 0$. For $k_{11} = 0$ we

put $P\left(X_{t_1}=0\right)=1$. We suppose that $\tilde{c}, X_{t_1}$ are idependent. We define a stochastic process c in the following way:

(4) $$c=\left(X_t, t\geq t_1\geq 0\right)=\left(\tilde{X}_{t_1}-\frac{\mathrm{K}\left(t_1,t\right)}{k_{11}}\left(\tilde{X}_{t_1}-X_{t_1}\right), t\geq t_1\geq 0\right)$$

for $k_{11}>0$. If $k_{11}=0$ we put $\frac{\mathrm{K}(t_1,t)}{k_{11}}=1$.

Proposition 1 The stochastic process defined by (4) is a PGP.

Now let us consider a special case of (4), namely $k_{rs}=\mathrm{K}\left(t_r,t_s\right)=\min\left(t_r+a, t_s+a\right)$, $a\geq 0$, $\mathrm{t}_1=0$. For $a=0$ we put $P\left(X_0=0\right)=1$. Moreover let $\mathrm{W}=\left(W_t, t\geq 0\right)$ be a Wiener process. Then c defined by (4) has the following form: $c=\left(W_t+X_0, t\geq 0\right)$. Now we are going to give a slightly modified version of P. Lévy's characterization theorem for martingales.

Proposition 2 If a stochastic process $c=\left(X_t, t\geq 0\right)$ has continuous trajectories, is square integrable, X_0 is a r.v, with ch.f. given by (3) where $k_{00}=a$, $a\geq 0$

$$E\left(X_t\middle|F_s\right)=X_s,\ E\left(\left(X_t-X_s\right)^2\middle|F_s\right)=t-s \text{ for all } s\leq t,$$

then c is PGP with $k_{rs}=\mathrm{K}\left(t_r,t_s\right)=\min\left(t_r+a, t_s+a\right)$.

For $a=0$ we put $P\left(X_0=0\right)=1$.

Proposition 3 If a stochastic process $c=\left(X_t, t\geq 0\right)$ has continuous trajectories, is square intergrable, there exist a function g such that $E\left(X_t\middle|F_s\right)=\frac{g(t)}{g(s)}X_s$, $Var\left(X_t\middle|F_s\right)$ is non random for all $s\leq$, X_0 is a r.v. with ch.f. given by (3) then c is a PGP.

Proposition 4 Let $c=\left(X_t, t\geq t_1\right)$ be a PGP with densities given by (1). Let $q_{rs}=E\left(X_{t_r}X_{t_s}\right)$. Then $q_{rs}=k_{rs}+2c_2k_{1s}k_{1r}k_{11}^{-2}$. Moreover c is a Markov process iff for $t_1\leq t_s\leq t_r\leq t_u$ one of the following equivalent conditions holds:

$$k_{sr}k_{ru}=k_{su}k_{rr},$$
$$q_{sr}q_{ru}=q_{su}q_{rr}.$$

We say that c is a process with independent linear forms if there exist functions $a_{rs}=a_{rs}\left(t_1,...,t_s\right)$, $r\leq s$ such that $X_{t_1}+X_{t_1}+a_{12}X_{t_1},...,X_{t_d}+a_{1d}X_{t_1}+...+a_{d-1,d}X_{t_{d-1}}$ are independent for $d=2,3,...$.

Proposition 5 If c is a process with independent linear forms and for every $t_r\geq t_1$ the ch.f. of X_{t_r}, is given by (3) then c is a PGP.

Proposition 6 Let $c=\left(X_t, t\in\left[t_1,T\right]\right)$ be a separable PGP with ch.f. given by (2) where $\mathrm{K}\in\square^1$. Then c has a continuous modification.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Testing the Goodness-of-Fit in a Gaussian Regression

Poilleux Hélène
INRA, Laboratoire de Biométrie
78352 Jouy-en-Josas, France
helene@banian.jouy.inra.fr

We propose a test for testing that the regression function of a Gaussian regression model belongs to a parametric family against a nonparametric alternative. The testing procedure is free from any knowledge on f and on the variance of the observations. The test is asymptotically of level a and we characterise a class of vectors over which the test is asymptotically powerful.

1. Statistical Model

We observe $Y_1,\cdots,Y_n$ given by the regression model

$$Y_i = f(x_i)+e_i$$

where f is an unknown real valued function defined on $\mathbf{R}^d$, i.e. $f \in \mathrm{F}\left(\mathbf{R}^d,\mathbf{R}\right)$, $x_1,\cdots,x_n$ are deterministic points belonging to $\mathbf{R}^d$ and $e_1,\cdots,e_n$ is an i.i.d. sample of centred Gaussian variables with an unknown variance s^2.

We propose a test for testing that the regression function f belongs to the parametric family $\mathrm{F}_\Theta = \{F(.,q)\ ,q \in \Theta\}$, where F is a known function and Θ is a subset of $\mathbf{R}^p$. We denote by $\tilde{f}$ the vector of $\left(f(x_i)\right)_{1\le i\le n}$ and by $\tilde{F}_q$ the vector of $\left(F(x_i,q)\right)_{1\le i\le n}$. We test the null hypothesis

$$H_0 : \exists q_0 \in \Theta, \text{such that } \forall x \in R^d,\ f(x) = F(x,q_0)$$

against the alternative

$$H_n : d_n(f,\mathrm{F}_\Theta) = \inf_{q\in\Theta} \left\| \tilde{f} - \tilde{F}_q \right\|_n \ge r_n(f)$$

where $\sqrt{n}\,\|\ .\ \|_n$ denotes the Euclidean norm in $\mathbf{R}^n$ and $r_n(f)$ depends on f and on the errors of first and second kind that we choose for the test.

2. The Testing Procedure

In the particular case of testing that f is a linear in the parameters, the testing procedure is similar to that proposed by Y.Baraud, B.Laurent and S. Huet.

We estimate f by $F_{\hat{q}} = F\left(.,\hat{q}\right) \in \mathrm{F}_\Theta$ where $\hat{q}$ is the least square estimator of q. We assume that the usual assumptions for $\hat{q}$ to converge to q^* that minimises $\left\| \tilde{f} - \tilde{F}_q \right\|_n^2$ are true. The idea of the proposed test is to construct several Fisher tests of the hypothesis "$Y - \tilde{F}_{\hat{q}} = 0$" against alternatives of the form "$Y - \tilde{F}_{\hat{q}} \in \tilde{S}_m$" where $\tilde{S}_m$ is a linear

subspace of $\mathbf{R}^n$. Noting that $\left\| P_{q^*}\left(\tilde{f} - \tilde{F}_{\hat{q}}\right) \right\|_n^2$ is close to 0 where P_{q^*} is the projection matrix on the columns of $\frac{\partial \tilde{F}_{q^*}}{\partial q}$, we consider spaces $\tilde{S}_m$ such that $P_{q^*}\tilde{S}_m = 0$.

For example, assuming for the sake of simplicity that $f \in \mathrm{F}\left([0\ ;\ 1]\ ,\ \mathbf{R}\right)$ and that $x_i = i/n$, we can take $\tilde{S}_m$ as

$$\tilde{S}_m(q) = \left(I_n - P_q\right)\tilde{E}_m,$$

where

1. I_n is the identity matrix,
2. m belongs to some set $M_n = \left\{ m \in \{1,\cdots,n\}\ ,\ m \in \left\{2^j, 0 \le j \le J_n\right\}\right\}$ such that $n^{-1}2^{J_n}$ tends to 0 when n tends to $+\infty$.
3. $\tilde{E}_m$ is the $n \times p$ matrix defined by : $\left(\tilde{E}_m\right)_{i,j} = d_{x_i \in \left]\frac{j-1}{m};\frac{j}{m}\right]}$ with $1 \le i \le n$, $1 \le j \le p$ and d denotes the Dirac function.

We denote by $\hat{\Pi}_m$ the orthogonal projection onto $\tilde{S}_m(\hat{q})$ and by D_m the dimension of $\tilde{S}_m(\hat{q})$. Let us fix $a \in [0;1]$. We associate to each $m \in M_n$ a level $a_m \in [0;1]$ such that

$$\sum_{m \in M_n} a_m = a.$$

Denoting by $\bar{F}_{D,N}(u)$ the probability for a Fisher statistic with D and N degrees of freedom to be larger than u, we define the test statistic by

$$\hat{T}_a = \sup_{m \in M_n}\left(\frac{\left\| \hat{\Pi}_m\left(Y - \tilde{F}_{\hat{q}}\right)\right\|_n^2 (n - D_m)}{\left\| \hat{\Pi}_m^{\perp}\left(Y - \tilde{F}_{\hat{q}}\right)\right\|_n^2 D_m} - \bar{F}^{-1}_{D_m, n-D_m}(a_m)\right).$$

We reject the null hypothesis H_0 if $\hat{T}_a > 0$.

3. The Results

The test is asymptotically of level a and we characterise a class of vectors over which the test is asymptotically powerful. More precisely, for any $b \in [0;1]$ we compute $r_n(f)$ such that the asymptotic power of the test is greater than $1-b$ for all f provided that $d_n(f, \mathrm{F}_\Theta) \ge r_n(f)$. If we consider the local alternative "$f = F(.,q_1) + \frac{g}{\sqrt{n}}$" for some bounded function g and some $q_1 \in \Theta$, then we show that for an adequate choice of $\left\{ \tilde{E}_m, m \in M_n\right\}$ the separation rate of testing is the parametric rate. A simulation study shows that the test is powerful even for small samples.

Reference

Baraud, Y., Huet, S., Laurent, B. (2000). Adaptative test of linear hypothesis by model selection. Technical report available at : *www.dma.ens.fr/~baraud/*

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Some Remarks on the Bayesian Analysis of Non-Dominated Statistical Models

Silvia Polettini
ISTAT, Serv. Metodologie Base Produzione Statistica
Via C. Balbo, 16, Roma, Italy
polettin@istat.it

Claudio Macci
Università di Torino, Dipartimento di Matematica
Via Carlo Alberto 10, Torino, Italy
macci@dm.unito.it

Brunero Liseo
Università "La Sapienza", Dipart. Studi Geoeconomici
Via del Castro Laurenziano 7, Roma, Italy
Brunero.Liseo@uniroma1.it

1. Notation and Definitions

Let (Θ,B) (the parameter space) and (X,A) (the sample space) be two measurable spaces. Denote by m on B the prior distribution and by $\{P_q: q\in\Theta\}$ the family of sampling distributions. A Bayesian experiment is the (unique) probability space $\mathsf{E}_m=(\Theta\times\mathsf{X}, \mathsf{B}\times\mathsf{A}, \Pi_m)$ such that $\forall\ B\in\mathsf{B}$ and $\forall\ X\in\mathsf{A}$

(1) $$\Pi_m(B\times X)=\int_B P^q(X)\ dm(q).$$

The *predictive distribution* is the probability measure P_m on A

$$X\in\mathsf{A}\mapsto P_m(X)=\Pi_m(\Theta\times X)=\int_\Theta P^q(X)\ dm(q);$$

the experiment E_m is *regular* if a family $\{m^x: x\in\mathsf{X}\}$ of probability measures on B exists such that $\forall\ B\in\mathsf{B}$ and $\forall\ X\in\mathsf{A}$

(2) $$\Pi_m(B\times X)=\int_X m^x(B)\ dP_m(x).$$

The *statistical experiment* $\{P_q: q\in\Theta\}$ is *dominated* (by a s-finite measure l) if the sampling distributions are all absolutely continuous w.r.t. a s-finite measure l.

In such case a function f_l (the *likelihood*) exists such that

$$\forall q\in\Theta\quad P^q(dx)=f_l(q;x)\ l(dx),$$

and, for any prior m, the marginal distribution P_m can be written as $[\int_\Theta f_l(q;x)\,dm(q)]l(dx)$; moreover, the *Bayes' theorem* holds, in that

$$P_m\left[\ x\in\mathsf{X}: B\in\mathsf{B}\mapsto m^x(B)=\frac{\int_B f_l(q;x)\ dm(q)}{\int_\Theta f_l(q;x)\ dm(q)}\ \right]=1.$$

(On the other hand, the Bayesian experiment E_m is said to be *dominated* (e.g. Florens *et al.*, 1990, page 30, Definition 1.2.4) if $P_m\approx m\times P_{m_\bullet}$)

As in non-dominated models one cannot refer to the likelihood and the Bayes' formula, in order to derive the posterior, expressions (1) and (2) need be used directly. The

posterior m^x is hence derived via the relation:

(3) $$\Pi_m(B\times X)=\int_B P^q(X)\,dm(q)=\int_X m^x(E)\,dP_m(x).$$

Note also that neither of the previous formulae for the marginal and the Bayes theorem are valid under a non-dominated statistical model, and that their misuse can result in paradoxes, e.g. Christensen and Utts (1992) and Blachman *et al.* (1996).

2. The Bayes Factor

In Bayesian hypothesis testing, the evidence provided by the data to a null hypothesis H_0 versus an alternative H_1 is usually carried by the Bayes factor of H_0 vs H_1. For an extensive review, see Kass and Raftery (1995). In Macci and Polettini (2001) we investigate use of the Bayes factor in non-dominated statistical models and illustrate its behaviour by a very simple example (e.g. Berger and Wolpert, 1988).

For a given statistical experiment, consider the hypothesis testing problem:

$$H_0:q\in\Theta_0 \text{ vs } H_1:q\in\Theta_1,\quad \Theta_0,\Theta_1\in \mathsf{B}:\ \Theta_0\cap\Theta_1=\emptyset.$$

Set $m_k=m(\Theta_k)$, assuming $m_k>0$, $k=0,1$. The *Bayes factor* (BF) in favour of H_0 vs H_1 is defined (e.g. Kass and Raftery, 1995) as the posterior to prior odds ratio:

(4) $$BF_m(x)=\frac{m^x(\Theta_0)/m_0}{m^x(\Theta_1)/m_1};$$

when the statistical model is dominated, denoting by $\tilde{m}_k(\cdot)$ the prior conditional to Θ_k (k=0,1), the BF may be written in terms of the likelihood as follows:

(5) $$BF_m(x)=\frac{\int_{\Theta_0} f_l(q,x)\,d\tilde{m}_0(q)}{\int_{\Theta_1} f_l(q,x)\,d\tilde{m}_1(q)}$$

When both hypotheses are simple, formula (5) returns the likelihood ratio.

We express BF_m as a ratio between suitable densities of probability measures on the sample space:

(6) $$BF_m(x)=\frac{dP_m(\cdot\,|\,\Theta_0)/dP_m(x)}{dP_m(\cdot\,|\,\Theta_1)/dP_m(x)}$$

Formula (6) shows that the BF only depends on the conditional distributions; in this sense, the definition enjoys the same property as the dominated case. Moreover, when the model is dominated, formula (5) can be recovered from the definition in (6).

Formula (6) also gave us the start for investigating a non dominated analogous of the likelihood function. We discuss results of work in progress on this topic, based on the Lebesgue decomposition of a measure with respect to a reference.

References

Berger, J.O. and Wolpert, R.L. (1988): The Likelihood Principle (2nd ed.). *IMS Lecture Notes*, Hayward, California.

Blachman, N.M., Christensen, R. and Utts, J.M.(1996): Comment on "Bayesian resolution of the exchange paradox" [Letter]. *Amer. Statist.* **50**, 98-98

Christensen, R. and Utts, J (1992): Bayesian resolution of the "exchange paradox". *Amer. Statist.* **46**, no. 4, 274--276.

Florens, J., Mouchart, M. and Rolin, J. (1990): *Elements of Bayesian Statistics.* Marcel Dekker Inc., New York.

Kass, R. E. (1993): Bayes Factors in Practice. *Statistician*, **42**, 551-560.

Kass, R. E. and Raftery, A. E. (1995): Bayes Factors. *Journ. Amer. Stat. Assoc.*, **90**, 773-795.

Macci, C. and Polettini, S. (2001): Bayes Factor for non-dominated statistical models, *Stat. Prob. Lett.* (to appear).

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Multidimensional Appell Polynomials

Denys Pommeret
CREST-ENSAI
Campus de Ker Lann, Rue Blaise Pascal, 35170 Bruz – France
pommeret@ensai.fr

More than one hundred years ago, Appell introduced a polynomial class characterized by the exponential form of its generator function. This class is well known on R and the aim of this work is to offer a multi-dimensional analysis in two directions: we indicate a classical characterization of the d-dimensional Appell polynomials via their generator function and we investigate their orthogonality. Two types of orthogonality are characterized with respect to Gaussian distributions. We relate this study to the theory of Lie algebras and martingales.

References

Appell, M.P. (1880). Sur une classe de polynomes, *Ann. Sci.Ecole Norm. Sup.* **9**, 119-144.

Sheffer, I.M. (1935). A differential equation for Appell polynomials. *Bull. Am. Math. Soc.* **40**, 914-923.

Pommeret, D. (2001). Orthogonal and pseudo-orthogonal multi-dimensional Appell polynomials. *Appl. Math. Comp.* **117**, 285-299.

Schoutens, W. and Teugels, J.L. (1998). Levy processes polynomials and martingales. *Stoch. Models* **14**, 335-349.

Numerical Taxonomy Methods for Statistical Data Processing

Tiberiu Postelnicu
Centre for Mathematical Statistics, Romanian Academy
Casa Academiei, Calea 13 Septembrie nr. 13, 76100 Bucharest 5, ROMANIA
tposteln@k.ro

The purpose of numerical taxonomy can be briefly defined as the construction of objective clusters of units by means of a quantitative measure of their affinity. Its name comes from the fact that the first methods were proposed for, and essentially applied to, biological classification.

Numerical taxonomy methods present a very powerful multiple comparison instrument. More generally, cluster analysis is the name given to various procedures whereby a set of individuals or units, termed as "Operational Taxonomic Units" (OTU). Techniques of cluster analysis can be applied in different fields of medicine: the recognition of various clinical forms of a disease, separation of distinctive racial groups, treatment of quantitative biogeographical data, etc.

An important case for statistical data processing deals with OTUs described by binary attributes. Homogeneities for binary and for ordered multistates data are presented. Methods of automatic classification are described and two types of homogeneities for the classification in biology and the genetics of the human populations are given.

The new extension concerns the inference in contingency table and it is applicable in any field. The connection between numerical taxonomy, one side, and the cluster analysis, as well as the discriminant analysis, on the other side, is useful to be considered

References

Dragomirescu L., Postelnicu T., (1994), Specific numerical taxonomy methods in biological classification. In "Statistical Tools in Human Biology"., *World Scientific*, 31-46.
Sneath P.H.A., Sokal R.R., (1973), Numerical taxonomy. San Francisco Freeman.
Buser M.W., Baroni-Urbani C., (1982), A direct nondimensional clustering method for binary data. *Biometrics*, **38**, 351-360.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Nonstationary Autoregression: Bootstrap and Subsampling

Zuzana Prásková
Charles University
Department of Probability and Mathematical Statistics
Sokolovska 83, 186 75 Prague
Czech Republic
praskova@karlin.mff.cuni.cz

It is known that the classical Efron bootstrap when applied to linear regression models with heterogeneous errors does not estimate consistently the variances of the least squares estimators of regression parameters. It is also known that the wild bootstrap is robust procedure against heteroscedasticity, i.e. it provides consistent estimators, but it is less efficient than the classical procedure in the case of homogeneous errors (see Liu and Singh (1992)).

In this contribution we will study the problem of consistency and efficiency of bootstrap in case of a causal autoregression process with independent but heterogeneous innovations. It will be shown that the (Efron) bootstrap based on estimated residuals does not yield consistent estimators of the variance of the least-squares estimators of the autoregression parameters, while the wild bootstrap is robust against heteroscedasticity. On the other hand, the efficiency of the wild bootstrap with respect to the Efron bootstrap varies with the values of the parameters of the model.The obtained results follow from asymptotic representation of the estimators and their mean square errors and from central limit theorems for martingales.

Alternatively, parameters of autoregression can be estimated by using subsampling, dividing the observed series into overlapping blocks. General theory for subsampling stationary and nonstationary strong mixing sequences is presented in Politis, Romano and Wolf (1999). We formulate conditions on the error process under which the sampling distribution of least-squares estimators of autoregressive heteroscedastic sequence is consistently estimated by subsampling (Praskova (2001)).

Subsampling variance estimators will be also considered and their efficiency will be studied in dependence on the block size.

References

Liu, R. Y. and Singh, K. (1992). Efficiency and robustness in resampling. *Ann. Statist.* **20**, 370-384.

Politis, D. N., Romano, J. P. and Wolf, M. (1999). Subsampling. Springer, New York, Berlin.

Praskova, Z. (2001). Subsampling and its application to time series. (Czech). In *Proceedings of Robust 2000 Summer School* (J. Antoch, ed.), Charles University, Prague.

Sherman, M. (1998). Efficiency and robustness in subsampling for dependent data. *J. Statist. Plan. Inference.* **75**, 133-146

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Biasedness and Unbiasedness of Seat Apportionments in Three Party Proportional Representation Systems

Friedrich Pukelsheim, Norman R. Draper, Mathias Drton, Karsten Schuster
Universität Augsburg, Institut für Mathematik
86135 Augsburg, GERMANY
Pukelsheim@Math.Uni-Augsburg.De

From the inception of the proportional representation movement, it has been an issue whether a given apportionment method favors larger parties at the expense of smaller parties. For three parties that are ordered by their vote counts form largest to smallest, we calculate the expected difference between the seat numbers and the ideal share of seats, as a function of the district magnitude, for three traditional methods. These are (i) the divisor method with standard rounding, also called the Webster method or the method of Sainte-Laguë; (ii) the quota method with fit by largest remainders (Hamilton method, method of Hare); and (iii) the divisor method with rounding down (Jefferson method, method d'Hondt). The first two methods are seen to be practically unbiased, whereas the last is clearly biased in favoring larger parties at the expense of smaller parties. The theoretical findings are confirmed by empirical election results from the Swiss Canton of Solothurn, and from the German State of Bavaria. Some historical remarks draw attention to three Europeans who contributed to the subject about eighty years ago, André Sainte-Laguë, Ladislaus von Bortkiewicz,

1. Overview

In proportional representation electoral systems, one important problem is how to compare the various methods of translating votes into specific seat apportionments. The seat numbers are, of course, integer numbers while, by comparison, the votes are almost continuous quantities. The translation of votes into seats nearly always involves adjusting, in some manner, the fractional seats that would arise if a naive calculation were made to obtain the actual seat apportionment. See *Balinski/Young* (1982) for an excellent exposition of apportionment methods and their structural properties. For an application of these principles to the electoral systems in Germany, both on the federal level and state levels, see *Pukelsheim* (2000 a-c).

A particular issue is whether an allocation method is biased in favor of larger parties at the expense of smaller parties, or vice versa. Here we discuss a statistical approach to the problem, assuming three parties are involved, and concentrating on three traditional apportionment methods. For repeated applications of each method, we evaluate the biases of the seat numbers for the various parties. The biases here considered are the averages, over all possible electoral outcomes, of the differences between the (integer) seats actually apportioned, and the (fractional) ideal share of seats that would have been awarded had fractional seats been possible.

Section 2 specifies this bias concept. Section 3 to 5 give, respectively for the three methods chosen, a description of the method and formulas for the theoretical biases of the seat numbers for the largest party, the second-largest party, and the third-largest party, the ranking being determined by vote counts.

Section 6 applies these formulas to empirical election data, confirming that the theoretical findings are sensible. Combination of the theoretical and empirical results leads to our main conclusion, that the divisor method with standard rounding (Webster, Sainte-Laguë), and the quota method with fit by largest remainders (Hamilton, Hare) have "practically unbiased" seat numbers for all three parties. In contrast, the divisor method with rounding down (Jefferson, Hondt) is visibly biased, in favoring larger parties at the expense of smaller parties.

Section 7 lists the bias formulas for a larger family of divisor methods that includes the two divisor methods of Section 3 and 5 as special cases. Section 8 provides an alternative approach to the comparison of apportionment methods, in terms of the majorization ordering. Section 9 reviews some often neglected contributions made about eighty years ago, by the three European scientists André Sainte-Laguë, Ladislaus von Bortkiewicz, and Georg Pólya.

References

Balinski, M.L./Young, H.P. (1982). Fair Representation - Meeting the Ideal of One Man, One Vote. New Haven CT.

Pukelsheim, F. (2000). Mandatszuteilungen bei Verhältniswahlen: Erfolgswertgleichheit der *Allgemeines Statistisches Archiv–Journal of the German Statistical Society* **84**, Heft 4.

Pukelsheim F. (2000). Mandatszuteilungen bei Verhältniswahlen: Vertretungsgewichte der Mandate. *Kritische Vierteljahresschrift für Gesetzgebung und Rechtswissenschaft* **83**, 76-103.

Pukelsheim, F. (2000). Mandatszuteilungen bei Verhältniswahlen: Idealansprüche der Parteien. *Zeitschrift für Politik–Organ der Hochschule für Politik München* **47**, 239-273.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

An Additive Intensity Model in a Multivariate Process Counting

Jose Manuel Quesada-Rubio, Ana M. Lara-Porras, Julia Garcia-Leal, Esteban Navarrete-Alvarez
Univ de Granada, Department de Estadística e I.O.Facultad de Ciencias
Avda. Fuente nueva s/n, Granada. Spain
quesada@ugr.es

1. Background and Estimation

Let us consider a multivariate process counting $\mathbf{N} = (N_1,...,N_k)$ with intensity process $\mathbf{l} = (\lambda_1,...,\lambda_k)$. We suppose that the intensity process $\lambda_h(t) = \lambda_h(t;\mathbf{q})$ is a predictable process with respect to some filtration $(\mathcal{F}_t : t \in T = [0,\tau))$ (for a given terminal time τ, $0 < \tau \leq \infty$) on a probability space $(\Omega,\mathcal{F},P)$, which may be specified by a parameter $\mathbf{q}$ belonging to an open subset of the q-dimensional Euclidean space $\mathbf{q} = (\theta_1,...,\theta_q) \in \Theta$.

We suppose that there are n individuals and a vector of covariates $\mathbf{Z}_i(t)$, i=1,...,n, observed for individual i, where the process $\mathbf{Z}_i(.)$ is $\mathcal{F}_t$-predictable and we consider the additive intensity model.

$$l_{hi}^{q} = Y_{hi}(t)a_{hi}^{q}(t;Z_i(t)) \quad ; \quad h = 1,...,k \quad ; \quad i = 1,...,n \quad ,$$

where $\alpha_{hi}^{\theta}(t; \mathbf{Z}_i(t))$ is of the form

$$a_{hi}^{q}(t;Z_i(t)) = a_{h0}(t,g) + r(b^{T}Z_{hi}(t)) \quad ; \quad h = 1,...,k \quad ; \quad i = 1,...,n \quad ,$$

and

- $r(\mathbf{b},\mathbf{Z})$ is a non-negative function. This function may accept different parameterizations with the condition $r(\mathbf{b},\mathbf{0}) = 0$.
- The $Y_{hi}(.)$ are predictable and do not depend on $\mathbf{q}$. Usually $Y_{hi}(t)$ contains information on whether individual i is observed to be at risk for experiencing a type h event just before time t.

Once we set the model, our aim is to estimate the parameters using the log-partial likelihood whose expression is

$$C_t(q) = \sum_{h,i}\left(\int_0^t \log\left\{ l_{hi}^{q}(t)\right\}dN_{hi}(t) - \int_0^t l_{hi}^{q}(t)dt\right),$$

in this way we get the vector of derivates with respect to $\mathbf{b}$ given by

$$U_t(b) = \int_0^t \sum_{h,i} \frac{1}{\left\{a_{h0}(t,g) + r\left(b^{T}Z_{hi}(t)\right)\right\}} r^{(1)}(b^{T}Z_{hi}(t))Z_{hi}(t)dN_{hi}(t) - \int_0^t \sum_h nS_h^{(1)}(b,t)dt$$

where $r^{(1)}(x) = dr(x)/dx$ and

$$S_h^{(1)}(b,t) = \frac{\sum_i Y_{hi}(t)Z_{hi}(t)r^{(1)}(b^{T}Z_{hi}(t))}{n}.$$

Below we have calculated the information matrix and analyzed some aspects of the additive intensity model.

The information matrix for β is given by

$$I(b) = -E\left[\frac{\partial U_t(b)}{\partial b}\right]$$

being

$$\frac{\partial U_t(b)}{\partial b} = \int_0^t \sum_{h,i} \frac{r^{(2)}(b^T Z_{hi}(t))\left\{a_{h0}(t,g) + r\left(b^T Z_{hi}(t)\right)\right\} - \left(r^{(1)}(b^T Z_{hi}(t))\right)^2}{\left\{a_{h0}(t,g) + r\left(b^T Z_{hi}(t)\right)\right\}^2} \{Z_{hi}(t)\}^{\otimes 2}\, dN_{hi}(t) -$$

$$-\int_0^t \sum_h n S_h^{(2)}(b,t)dt.$$

with $r^{(2)}(x) = dr^{(1)}(x)/dx$

$$S_h^{(2)}(b,t) = \frac{\sum_i Y_{hi}(t)\{Z_{hi}(t)\}^{\otimes 2}\, r^{(2)}(b^T Z_{hi}(t))}{n}$$

and $X^{\otimes 2}$ for an X vector representing XX^T.

References

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10**, 1100-1120.

Andersen, P.K., Borgan, Gill, R.D., and Keiding, N. (1993). *Statistical Models Based* on *Counting Processes.* Springer-Verlag. New York.

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall, London.

Fleming, T.R. and Harrington, D.P. (1991). *Counting Process and Survival Analysis.* Wiley, New York..

Prentice, R.L. and Self, S.G. (1983). Asymptotic distribution theory for Cox-type regressions models with general relative risk form. *Ann. Statist.* **11**, 804-813.

Self, S.G., and Prentice, R.L. (1982). Commentary on Andersen an Gill's Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10**, 1121-1124.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Sequential Spectral Test

María del Pino Quintana Montesdeoca
Campus Universitario de Tafira, Department of Mathematics
Las Palmas de Gran Canaria, 35017
mpino@dma.ulpgc.es

Pedro Saavedra Santana
Campus Universitario de Tafira, Department of Mathematics
Las Palmas de Gran Canaria, 35017
saavedra@dma.ulpgc.es

1. Introduction

Two populations are considered and a signal, which can be modeled as a stationary process, is evaluated over all its elements. The aim of this paper is to examine if the considered signal has predictive value for both populations. We analyse this problem through of the frequency domain. A sequential test is proposed for the spectral comparison of both poputlations. For prefixed errors alpha and beta, the regions of acceptance, rejection and continuation are determinated by means of an equations system. Since this equations can not be resolved analytically, we give a computational method which is valid for gaussian linear processes. A such method is based on the known asymptotic results for the periodogram of these processes. However, we also examine the robustness of this method by means of the application to non gausiann linear processes.

2. The Sequential Spectral Test

Let $\{X_{li}(t); l=1,2; i=1,\cdots,r; t=1,\cdots,N\}$ be a set of time series evaluated at the same times on random samples of *r* objects choosen from populations C_1 and C_2. The periodogram of each time serie for *j*th Fourier frequency is defined as:

(1) $$I_{li}(w_j)=\frac{1}{2p\,N}\left|\sum_{t=1}^{N}X_{li}(t)e^{-\mathbf{i}w_j t}\right|^2$$

Suppose that each individual periodogram verifies the model:

(2) $$I_{li}(w_j)= f_l(w_j)\cdot U_{lij}^N$$

being $f_l(w)$ the spectral density function corresponding to *l*th population and $\{U_{lij}^N\}$, for each *l*=1,2 e $i=1,\cdots,r$, independent random variables for $j=1,\cdots,n=[N/2]$ and exponentially distributed of parameter one. This model is based on the asymptotic representation for the periodogram of gausiann linear processes. (Priestley, 1981).
This model can be transformed as:

(3) $$Y_{lij}=m_l(w_j)+x_{lij}-C$$

being $Y_{lij}=\log I_{li}(w_j)-C$, $m_l(w_j)=\log f_l(w_j)$, $C=E[\log U_{lij}]\approx -0.57721$ (Euler constant) and $x_{lij}=\log U_{lij}$. It is easy to see that $\text{var}(x_{lij})=s^2=1.648124$.

For testing the null hypothesis $H_0: m_1(w)\equiv m_2(w)$, we consider the statistic test:

(4) $$J(r)=\frac{1}{n}\sum_{j=1}^{n}\left(\overline{Y}_{1.j}-\overline{Y}_{2.j}\right)^2 \text{ where } \overline{Y}_{l.j}=\sum_{i=1}^{r}Y_{lij}\,.$$

In order to define a sequential test we consider $m(r)=E\left[J(r)\middle|H_0\right]$, $t^2(r)=\text{var}\left(J(r)\middle|H_0\right)$ and for adecuate constants *a, b* and *c* the actions:

- Accept H_0 if $J(r)<c+m(r)-b\cdot t(r)$
- Reject H_0 if $J(r)>m(r)+a\cdot t(r)$
- Take a new observation in each group if

$$c+m(r)-b\cdot t(r)\le J(r)\le m(r)+a\cdot t(r)$$

Let *R* be the stopping time. The acceptance and reject regions are respectively:

(5) $$A=\bigcup_{r=r_0}^{r_{\max}}\{R=r;J(r)<c+m(r)-b\cdot t(r)\} \text{ and}$$

(6) $$B=\bigcup_{r=r_0}^{r_{\max}}\{R=r;J(r)>m(r)+a\cdot t(r)\},$$

being r_0 the number of initial observations. Thus, for errors and defined as:

(7) $$a=P(B\mid H_0)=\sum_{r=r_0}^{r_{\max}}P_{H_0}(R=r;J(r)>m(r)+a\cdot t(r))$$

(8) $$b=P(A\mid H_1)=\sum_{r=r_0}^{r_{\max}}P_{H_1}(R=r;J(r)<c+m(r)-b\cdot t(r))$$

the values of *a*, *b* and *c* are determinated.

The method for approaching this constants is as follows:

Step one: Random variables $\left\{U_{lij}^N\right\}$, exponentially distributed of parameter one are simullated for *l*=1,2; *i*=1,…, *r* and $j=1,\cdots,n=\left[N/2\right]$

Step two: Let $J(r)=\frac{1}{n}\sum_{j=1}^{n}\left(\overline{e}_{1.j}^{r}-\overline{e}_{2.j}^{r}\right)^2$ be the statistical test, where

$$\overline{e}_{lj}^{r}=\frac{1}{r}\sum_{i=1}^{r}\log(U_{lij}^N)$$

Step three: For fixed *a* and *b*, we have considered an specificed difference $H_1: d(w)=m_1(w)-m_2(w)$, being $D^2=\frac{1}{n}\sum_{j=1}^{n}d\left(w_j\right)^2$. By means of doing several repetitions of the test, on each simulation, we look for the values *a*, *b* and *c* that let us accept H_0 if $J(r)<c+m(r)-b\cdot t(r)$ or reject H_0 if $J(r)>m(r)+a\cdot t(r)$.

References

Diggle, P. J. And Al-Wasel, I.(1993). On Periodogram-Based Spectral Estimation for Replicated Time Series, in: Subba Rao (Ed), *Developments in Time Series Analysis.* (Chapman and Hall, Great Britain) 341-354.

Hernández-Flores, C.N., Artiles-Romero, J. And Saavedra-Santana, P. (1999). Estimation of the Population Spectrum with Replicated Time Series. *Comp. Stat. And Data Anal.* **30** 271-280.

Priestley, M. B**.** (1981). *Spectral Analysis and Time Series.* Wiley, New York.

Saavedra, P., Hernández, C.N. and Artiles, J. (1999). Spectral Analysis with Replicated Time Series. *Comm. Stat. Theory and Methods.* **29,** 2343-2362.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Double Checking for two Error Types

V.M. Raats
Tilburg University, Department of Econometrics and Operations Research
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
V.M.Raats@kub.nl

J.J.A. Moors
Tilburg University, Department of Econometrics and Operations Research
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
J.J.A.Moors@kub.nl

1. Introduction

Auditing a large population of recorded values is usually done by means of sampling. Based on the number of incorrect records that is detected in the sample, a point estimate and a confidence limit for the population fraction of incorrect values can be determined. In general it is (implicitly) assumed that the auditor does not make mistakes while judging the correctness of the values. However, in practice this assumption does not necessarily hold: auditors are human and can make errors. To take this possibility into account, a subsample of the audited records is checked once more by a second auditor who is assumed never to make mistakes. The information obtained from these two samples should be combined to derive an estimate for the error rate in the population.

The starting point for this type of double checking was Moors et al. (2000). Only one posssible error type was considered: auditors could only miss (fail to detect) existing errors. For the case of random sampling, the maximum likelihood estimator as well as an upper confidence limit for the error rate were derived, treating the probability of an unnoticed error as a nuisance parameter. It was shown that the introduction of the possible error type causes a considerably increase of the upper limits, even if the second auditor finds not a single additional error. Based on data from one of the Dutch institutions for social security payments the estimate for the error rate in social securtiy payments proved to be 5% with an upper 95% confidence limit of 12%.

2. The Second Error Type

In our paper, we first introduce a second error type: the auditor may consider a correct value as an error. Again, the sample information of both auditor and infallible expert is combined to give point and interval estimates for the fraction of errors in the population, while treating the two probabilities of errors of judgement by the first auditor as nuisance parameters. The impact of the second error type on the upper confidence limit of the error rate in the population turns out to be considerably smaller than the impact of the first error type. Moreover, the introduction of the second possible error type does not lead to a decrease of the upper limit. This may be explained from the constrution of the confidence interval: for all possible values of the nuisance parameters the upper confidence limits (given these values) are determined; subsequently the ‘all over’ confidence limit is defined as the maximum of these

In both the model with one error type and the extended model, the upper limit is realized for a very high value of a nuisance parameter. In reality, such high values will not often occur, so the Bayesian approach for both models is desirable.

3. The Bayesian Approach

In our Bayesian approach independent beta distributions were used as priors for the three parameters involved, leading to beta posteriors as well. Integration over the nuisance parameters then gives the (marginal) posterior for the population error rate. A Bayes estimate and Bayes upper limit for this main parameter follow.

In a sense, a weighted average over all possible values of the nuisance parameters is taken. Hence, the Bayes upper confidence limit is in general lower than according to the classical approach.

Extending the model with the second error type causes a reduction of the Bayesian upper limits; this in constrast to the classical approach where the extension did not have much influence at all.

References

Cox, D.R. and D.V. Hinkley (1974). *Theoretical statistics.* Chapman and Hall. London.

Lehmann, E.L. (1959). *Testing statistical hypotheses.* Wiley. New York.

Moors, J.J.A. (1999). Double checking for two error types. *CentER Discussion Paper 9923*, Tilburg University.

Moors, J.J.A., B.B. van der Genugten and L.W.G. Strijbosch (2000). Repeated Audit controls, S*tatistica Neerlandica* **54**, 3-13.

Raats, V.M. (1999). *Herhaalde steekproefcontrole*. Masters Thesis. Tilburg University

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

A Multivariate CLT for Decomposable Random Vectors

Martin Raič
Institute of Mathematics, Physics and Mechanics
Jadranska 19
SI-1000 Ljubljana
SLOVENIA
martin.raic@fmf.uni-lj.si

The aim of this paper is to derive bounds on the error in the multivariate CLT for sums of random vectors in R^m with a certain type of dependence structure. Namely, for $W = X_1 + \ldots + X_n$ with $\mathrm{var}(W) = \mathbf{I}$ and $\mathbf{E}X_i = 0$, we have:

$$\left|\mathbf{E}f(W) - \mathbf{E}f(Z_m)\right| \le e(X_1, \ldots X_n)$$

uniformly in f belonging to a 'good' class of non-smooth test functions (e. g. the indicators of measurable convex sets), where Z_m is the standard normal vector in R^m and $e(X_1, \ldots X_n)$ is usually of order $O(n^{-1/2})$, which in general cannot be improved.

Decomposable random variables were studied in Barbour, Karoński and Ruciński (1989), where sharp bounds for Lipschitz test functions were derived. The idea is to find decompositions $W = U_i + W_i$, where W_i is independent of X_i and the U_i's are sufficiently small. The next step is to split the U_i's to $U_i = \sum_j X_{ij}$ and to find further decompositions $W = U_{ij} + W_{ij}$ with W_{ij} independent of X_{ij}. The error in the CLT is of order $O\left(\sum_{i,j} \mathbf{E}\left|X_i X_{ij} U_{ij}\right|\right)$. As shown by Barbour, Karoński and Ruciński (1989), this concept can be applied in numerous problems related to random graphs. Many other problems, such as random permutations and U-statistics, can also be treated this way.

In order to derive bounds in the CLT, we use Stein's method, which is a powerful tool in treating dependent random variables. Unfortunately, Stein's method only works well for sufficiently smooth test functions. There has been a lot of effort to extend it to non-smooth functions. For independent random vectors and random permutation statistics, this was achieved by Bolthausen (1984), Götze (1991) Bolthausen (1984). In more general case, sharp bounds have only been obtained for bounded random vectors. Moreover, an additional factor of order $O(\log n)$ appears in the bounds in the multivariate case (see Rinott and Rotar (1996)).

In this paper, we refine the argument of Götze (1991) to derive bounds which are, under some uniformity conditions, of the same order as those appearing in Barbour, Karoński and Ruciński (1989). In particular, the random vectors need not be bounded – only third moments are required. We apply our result to local dependence (in particular U-statistics) and random graph degree statistics.

References

Barbour, Karoñski and Ruciñski (1989). A central limit theorem for decomposable random variables with applications to random graphs. *J. Comb. Theory B* **47** 125-145.

Bolthausen, E. (1984). An estimate of the remainder in a combinatorial central limit theorem. *Z. Wahrsch. Verw. Gebiete* **66** 379-386.

Bolthausen, E. and Götze, F. (1993). The rate of convergence for multivariate sampling statistics. *Ann. Statist.* **21**, No. 4 1692-1710.

Götze, F. (1991). On the rate of convergence in the multivariate CLT. *Ann. Probab.* **19** 724-739.

Rinott, Y. and Rotar, V. (1996). A multivariate CLT for local dependence with $n^{-1/2}\log n$ rate and applications to multivariate graph related statistics. *J. Multivariate Anal.* **56** 333-350.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

On the Comparison of the Cumulative Distribution Functions of the Pólya Distribution and the Binomial Distribution

Héctor M. Ramos
Universidad de Cádiz¸Departamento de Estadística e I. O.
Duque de Nájera, 8., 11002 Cádiz. Spain.
hector.ramos@uca.es

David Almorza
Universidad de Cádiz, Departamento de Estadística e I. O.
david.almorza@uca.es

Alfonso Suárez
Universidad de Cádiz, Departamento de Estadística e I. O.
alfonso.suarez@uca.es

In a recent work, Ollero and Ramos (1995) using the description of the Pólya distribution as a generalised binomial distribution, compared the cumulative distribution functions (cdf) of the Pólya distribution *P(N,p,n,c)* when $c<0$ and the binomial distribution *B(p,n)*. This comparison is done everywhere but in an interval with amplitude equal to 1.

In this work, using the expression of the cdf of the Pólya distribution *P(N,p,n,c)* when $c<0$ and two auxiliary functions, we present this comparison everywhere but in an interval with amplitude less than 1, that is included in the one obtained by Ollero and Ramos (1995).

1. Introduction

The Pólya distribution (Eggenberger and Pólya, 1923), is generally presented in terms of random drawings of balls from an urn. Initially, it is assumed that there are N balls in the urn, M white balls ($p=M/N$) and $N-M$ black balls. One ball is drawn at random and then replaced with c additional balls of the same colour. This procedure is repeated n times. The total number X of white balls in the sample will have the Pólya distribution *P(N,p,n,c)*. The constant c is interpreted as a parameter of contagion. To obtain the following results, we are going to use two auxiliary functions in the way

$F_{N,p,n,c}(x)$ is,

$$F_{N,p,n,c}(x)=\begin{cases} 0 \;;\; x<0 \\ \displaystyle\sum_{m=0}^{k}\binom{n}{m}\frac{(pN)^{(m,c)}(qN)^{(n-m,c)}}{N^{(n,c)}} \;;\; k\le x<k+1 \;\; (k:0,1,\cdots,n) \\ 1 \;;\; x\ge n \end{cases}$$

If $c<0$, then we will assume $(-c)(n-1)<\min\{M;N-M\}$. The expression $A^{(u,v)}$ is a factorial polynomial of the u-th degree with respect to A, which is given by: $A^{(u,v)}=A(A+c)(A+2c)\ldots(A+(u-1)c)$; $A^{(0,v)}=1$. If $c=0$ the outcome is the cdf of the binomial distribution *B(p,n)*.

2. The Comparison

Lemma 2.1 is a previous result that is necessary to obtain the main result, Theorem 2.1. The interval

$$\left(\frac{p(n-1)(N-c)+c(n-1)}{N}; \frac{p(n-1)(N-c)}{N} \right)$$

is strictly included in the one obtained by Ollero and Ramos (1995).

Lemma 2.1 Let the Pólya distribution $P(N,p,n,c)$ $(c<0)$ and the binomial distribution $B(n,p)$, with cdfs $F_{N,p,n,c}(x)$ and $F_{p,n}(x)$ respectively, then:

(i) $F_{N,p,n,c}(0) < F_{p,n}(0)$

(ii) $F_{N,p,n,c}(n\text{-}1) > F_{p,n}(n\text{-}1)$

Theorem 2.1 Let the $_{Pólya}$ distribution $P(N,p,n,c)$ $(c<0)$ and the binomial distribution $B(n,p)$, it is verified for all $x = 0, 1, 2, \ldots, n\text{-}1$ $(n > 1)$, that:

$$F_{N,p,n,c}(x) - F_{p,n}(x) = \begin{cases} < 0 \; ; \; x \leq \dfrac{p(n-1)(N-c)+c(n-1)}{N} \\ > 0 \; ; \; x \geq \dfrac{p(n-1)(N-c)}{N} \end{cases}$$

References

Eggenberger and Pólya (1923). Uber die Statistik verketteter Vorgänge. *Zeitschrift fur Angewande Mathematik und Mechanik*. Vol. **3**, 279-289.

Ollero and Ramos (1995). Description of a Subfamily of the Discrete Pearson System as Generalised-Binomial distribution. *Journal of the Italian Statistical Society*, **2**, 235-249.

Uhlmann (1966). Vergleich der hypergeometrischen mit der Binomial-Verteilung. *Metrika*, **10**, 145-158.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Characterizations of Aging Properties of the Logarithmic Transformations by Means of Star Ordering

Héctor M. Ramos
Universidad de Cádiz, Departamento de Estadística e I. O.
Duque de Nájera, 8., 11002 Cádiz. Spain.
hector.ramos@uca.es

Miguel A. Sordo
Universidad de Cádiz, Departamento de Estadística e I. O.
mangel.sordo@uca.es

Alfonso Suárez
Universidad de Cádiz, Departamento de Estadística e I. O.
alfonso.suarez@uca.es

1. Introduction

Let X be a non-negative random variable with distribution function F and survival function $\overline{F} = 1 - F$. We say that X is an increasing failure rate (IFR) random variable if $\overline{F}$ is log-concave and we say that it is a decreasing failure rate (DFR) random variable if $\overline{F}$ is log-convex. The IFR and DFR random variables are of interest in reliability theory (see, v.g., Barlow and Proschan (1975) and Ross (1983)).

Several authors have studied characterizations of random variables that have IFR and DFR properties in terms of stochastic orders (see, for example, Ross (1983) or Belzunce et al. (1996)). From this note, characterizations of random variables that their logarithmic transformation have the IFR (DFR) properties are obtained by means of the star ordering. We apply these results to comparisons of truncated income random variables in terms of the variance of the logarithm of income.

For non-negative random variables X and Y with distribution functions F and G, respectively, we say that X is smaller than Y in star order (denoted as $X \leq_* Y$) if

$$\frac{F^{-1}(b)}{F^{-1}(a)} \leq \frac{G^{-1}(b)}{G^{-1}(a)}, \text{ whenever } 0 < a \leq b < 1.$$

Basic references describing star ordering are Barlow and Proschan (1975) and Arnold (1987).

Let X be a non-negative random variable. Denote by $X_{(a,\infty)}$ and $X_{(0,a)}$, respectively, the left and the right truncated random variables of X in a.

2. Characterizations

Theorem 1 Let X be a non-negative random variable with strictly increasing distribution function F. Then

$$\log X \text{ is IFR (DFR)} \Leftrightarrow X_{(a,\infty)} \geq_* X_{(b,\infty)} (\leq_*) \text{ for all } a < b,\ a, b \in \text{supp}(X)$$

Theorem 2 Let X be a non-negative random variable with strictly increasing distribution function F. The following conditions are equivalent:

1. $\log X$ has a log-concave (log-convex) distribution function.
2. $X_{(0,a)} \leq_* X_{(0,b)} \left(\geq_*\right)$ for all $a < b$, $a, b \in \text{supp}\, X$

Theorems 1 and 2 generalize the characterizations given by Belzunce et al. (1995), which were restricted to absolutely continuous random variables.

3. Application

In economics, a number of hypotheses concerning the shape of the income distributions may be interpreted in terms of the logarithm of income. For instance, a conventional measure of income inequality is the variance of the logarithm of income. If a non-negative random variable X represents the income of a community, the variance of logarithms is defined as

$$J_X = s^2_{\log X} = E\left[\log X - m_X\right]^2$$

where m_X is the mean of the logarithm of income. The properties of J_X have been discussed by Atkinson (1970) and Creedy (1977) among others. On the other hand, the effect of truncation upon certain inequality measures has been studied in the literature (see Ord et al (1983) and Belzunce et al. (1995)). In the next theorems, we obtain sufficient conditions for truncated income distributions to be ordered in terms of the variance of the logarithms of incomes.

Theorem 3 Let X be a non-negative random variable with strictly increasing distribution function F. Then,

$$\log X \text{ is IFR (DFR)} \Rightarrow J_{X_{(a,\infty)}} \geq J_{X_{(b,\infty)}} \left(\leq\right) \text{ for all } a < b.$$

Theorem 4 Let X be a non-negative random variable with strictly increasing distribution function F. If $\log X$ has a log-concave (log-convex) distribution function then $J_{X_{(0,a)}} \leq J_{X_{(0,b)}} \left(\geq\right)$ for all $a < b$.

Conditions of theorems 3 and 4 are satisfied for many income distributions. Some of the most widely used models of income distributions, the logarithms of which have the IFR property, are the well-known Lognormal and Gamma distributions and the models proposed by Singh and Maddala (1976) and Dagum (1979).

References

Arnold, B. C. (1987). Majorization and the Lorenz order: A brief Introduction. Springer, New York.

Atkinson, A.B. (1970). On the measurement of inequality. *J. Econ. Theory* **2**, 244-263.

Barlow, R. E and Proschan, F. (1975). Statistical Theory of Reliability and Life Testing. Holt, Rinehart and Winston, New York.

Belzunce, F., Candel, J. and Ruiz, J. M. (1995). Ordering of truncated distributions through concentration curves. *Sankhya* **57**, Series A, 375-383.

Belzunce, F., Candel, J. and Ruiz, J. M. (1996). Dispersive orderings and characterizations of ageing classes. *Statist. Probab. Lett.* **28**, 321-327.

Creedy, J. (1977). The principle of transfers and the variance of logarithms. *Bull. Econ. Statist.* **39**, 153-158.

Dagum, C. (1980). The generation and distribution of income, the Lorenz curve and the Gini ratio. *Economie Appliquee* **33**, 327-367.

Ord, J. K., Patill, G. P. and Taillie (1983). Truncated distributions and measures of income inequality. *Sankhya* **45**, Series B, 413-430.

Ross, S. M. (1983) Stochastic Processes. Wiley, New York.

Singh, S. K. and Maddala, G. S. (1976). A function for size distribution of incomes. *Econometrica* **44**, 963-970.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Generalised Inverse versus Factor Analysis

Raquel Redondo
Universidad Complutense de Madrid, Dpto. Estadística e I.O. II
Madrid, Spain
eciop27@sis.ucm.es

The use of analysis methods in different areas to the one they were created for, has meant, in several occasions, new solutions to old problems. This paper is devoted to show how has occurred so with Generalised Inverse (GI) and Factor Analysis (FA), both multivariate methods, but with different applications. This paper shows how these two answers to different problems have a common methodology.

1. Generalised Inverse

Definition Let $A: X_n \to Y_m$ a linear application. We define the GI matrix of A, and denote A^+, as the matrix given by $A^+ = \sum_{i=1}^{r} \frac{1}{\sqrt{l_i}} u_i w_i^*$, where l_i is a non-zero eigenvalue of A^*A, u_i is an unitary eigenvector of A^*A associated to l_i, w_i is an unitary eigenvector of AA^* associated to l_i and A^* is A transposed matrix .

Other expressions are: $A^+ = \left[\sum_{i=1}^{r} \frac{1}{l_i} u_i u_i^*\right] A^* = A^* \left[\sum_{i=1}^{r} \frac{1}{l_i} w_i w_i^*\right]$

By generalised inverse matrix, a point $x \in X_n$ is transformed in other point $A^+Ax \in X_n$, who is expressed by its coordinates in the original base $B=\{e_1, \ldots, e_n\}$ of the space: $x \to A^+Ax$. Developing the expression A^+Ax, we have:

$$A^+Ax = \left[\sum_{i=1}^{r} \frac{1}{l_i} u_i u_i^*\right] A^*Ax = \left[\sum_{i=1}^{r} \frac{1}{l_i} u_i \left(u_i^* A^* A\right)\right] x$$

Since $u_i^*A^*A = (A^*A\, u_i)^* = (l_i\, u_i)^* = l_i\, u_i^*$, then $A^+Ax = \left[\sum_{i=1}^{r} \frac{1}{l_i} l_i u_i u_i^*\right] x$

(1)
$$x \xrightarrow{GI} A^+Ax = \left(\sum_{i=1}^{r} u_i u_i^*\right) x$$

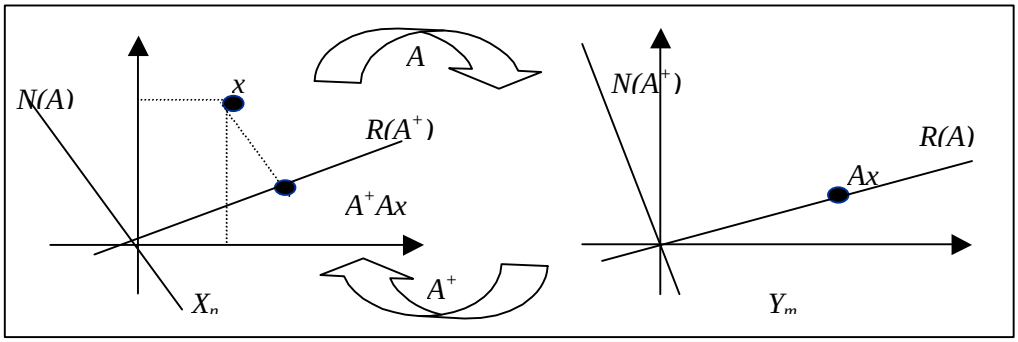


Figure 1. Generalised inverse transformation. Graphical description.

2. Factor Analysis

Let m points x_i, $i=1,2,\ldots,m$ and n characteristics y_j, $j=1,2,\ldots,n$, where x_{ij} is the value of characteristic y_j on point x_i. We build matrix $A=(x_{ij})$ and $A^* A = \sum_{i=1}^{n} x_i x_i^*$.

By FA, a point $x \hat{I} X_n$ is transformed in other point belonging to the same space, whose coordinates give the initial point projection on the new axes, called Factorial Axes: $x \mathbb{R} (x^* u_i)_{i=1,\ldots,k}$, referred to base $B'=\{u_1,\ldots,u_k\}$, composed by unitary eigenvectors of A^*A (this implies that, if we consider the whole space and the base $B'=\{u_1,\ldots,u_k,\ldots,u_n\}$, the corresponding coordinates for vectors $u_{k+1},\ldots,u_n$ are zero).

Taking matrix $C = (u_1,\ldots,u_k,\ldots,u_n)$ for base change, so that $Cx_{B'} = x_B$, as $x^* u_i \hat{I} R$ and coincides with its transposed, we have $x^* u_i = (x^* u_i)^* = u_i^* x$ and then $Cx_{B'} = \left(\sum_{i=1}^{k} u_i u_i^* \right) x$. That is:

(2) $$x \xrightarrow{FA} \left(\sum_{i=1}^{k} u_i u_i^* \right) x$$

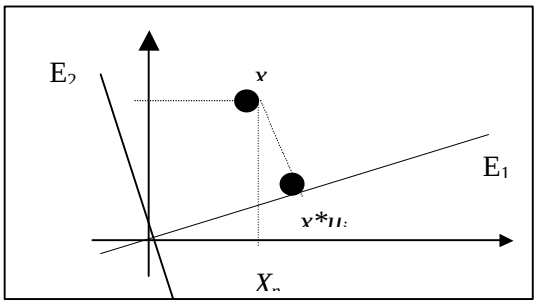


Figure 2. Factor analysis transformation. Graphical description.

With the notation used, if $r = k$ the resulting point of applying a FA projecting on the largest r factorial axes is coincident with the resulting point of applying transformation A^+A, but that transformed point is given in different bases of the space.

3. Conclusions

In the case $r<k$, some factorial axes associated to zero-eigenvalues have been considered. So we can get the identity of both expressions eliminating those axes.
In the case $r=k$, factorial axes considered are associated to non-zero eigenvalues, and exactly those. In this situation the identity of transformed point and expressions is guaranteed by expressions (1) and (2).

In the case $r>k$, the expressions and the resulting points are not coincident. So, next papers will get the coincidence, in a result already got.

References

Redondo, R. (1999). *Aproximación multidimensional al análisis de objetivos: la programación por objetivos y el análisis factorial*, Doctoral Thesis. UCM. Madrid.

Volle, M. (1997). Analyse des données. Economica. Paris.

Wiberg, D. M. (1985). *Espacio de estado y sistemas lineales.* McGraw-Hill. New York.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Mount Sigma: The Secret of Statistics

Raquel Redondo, Cristina del Campo, Roque Piñole, Enrique García-Pérez,

Dept. de Estadística e Investigación Operativa II, Universidad Complutense de Madrid
Madrid, Spain
eciop27@sis.ucm.es, campocc@ccee.ucm.es, eciop14@sis.ucm.es, eciop04@sis.ucm.es, jjrienda@ccee.ucm.es, eciop23@sis.ucm.es

New technologies are being increasingly applied in many areas, including learning. Due to that, the authors intend to contribute in computer aided learning with a computer software that will allow pupils learn or confirm knowledge in a more didactical and amusing way. This software is a game in which the hero has to climb Mount Sigma in order to achieve the secrets of Statistics.

1. Introduction

The introduction of new technologies nowadays has opened many possibilities for learning since these new technologies are able to aid the teaching-learning process. That is the reason why the authors work is devoted to search for new ways of applying these new technologies to foment self-learning and pupil personal work.

In this paper the authors intend to introduce the result of their work that consists on a interactive instrument for pupils that keeps away failure and gives support for their studies, particularly in Statistics.

2. Objective

The new instrument designed consists on a software program presented as a game, that allows pupils to fix and apply the knowledge acquired during lessons in a more interactive, didactical and attractive way for them. This software does not intend to substitute traditional and practical lessons, but to complement them.

The contents of the prototype are presented in a sequential way similarly as done in traditional lessons. But it is also possible to "walk through" the different difficulty stages in order to test the level of knowledge on the subject, being able to pass all the barriers that take place in the development of the game.

As mentioned before, the version presented is a prototype and therefore only includes a part of the basic Statistics contents. Specifically, the prototype is only devoted to the teaching of the probability calculus, including events algebra, probability axiomatization and properties and Total Probability and Bayes theorems. Anyway the possibilities are immense. The authors are already preparing the other parts, but it has to be pointed out that it can be also used in any other subject.

The prototype presented is designed to be used in every PC with Windows 95 or superior and it has very limited software and hardware requirements.

One of the advantages of the software the authors want to highlight is the possibility of easy translation into every language. In fact, the original software has been done in Spanish, but the translation into English is in process in order to make the software demonstration in this meeting understandable for every one attending the conference.

3. Final Aspect of the Prototype

The objective of the game is shown at the beginning and consists on getting to the top of Mount Sigma and achieving the secret of Statistics. The student plays the role of the hero, a pupil who is being taught by his master. The pupil wants to get the secret of knowledge and martial arts so, along with his lessons, he tries to scale Mount Sigma.

The game is divided into steps, each one representing a Statistical set of concepts as Probability Calculus, Ramdom Variable, Probability Distribution Models, etc. Each step has the following structure: the theoretical concepts are given at the beginning as if they were instructions to climb that step of the mount, then the hero gets into the lane. In order to get closer to the top the hero must pass some obstacles, represented as test questions that the hero must solve. If he is able to solve it he will go on, but if the hero fails the program will take him right to the beginning in order to let him firmly fix the knowledge.

Each time he passes a level of knowledge his master gives him a colored belt, simulating his martial arts abilities. The belt colors go from white to black.

As the authors teach in a Business School the questions are full of economical meaning and describe real examples.

In case the pupil fails, as it has being already said, the hero is taken to the base of Mount Sigma and he has to start again. In order to avoid repetitions in the questions proposed, the program is provided with a large base of questions aleatorily chosen each time a question is required.

Meanwhile the game is amusing, full of color, sound and animation. The pupil may read the instructions, but he can also listen to them. Besides the rhythm of contents and questions presentations is adaptable to each pupil because they are who establish that rhythm clicking the mouse when ready to continue.

Acknowledgements

This paper is partially supported by Universidad Complutense de Madrid under project PIE 20/2001.

References

P. Peralta, A. Rúa, R. Redondo y C. del Campo, Estadística: Problemas resueltos, Pirámide, Madrid, (in Spanish).

V.K. Rohatgi, An Introduction to Probability Theory and Mathematical Statistics, John Willey, New York, 1977.

Díaz Godino, M.C. Batanero y M.J. Cañizares, Azar y Probabilidad, Síntesis, Madrid, 1987 (in Spanish).

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Cluster Setting of Socioeconomical Patterns in Local Economies. An Application

Raquel Redondo
Universidad Complutense de Madrid, Dept. de Estadística e Investigación Operativa II
Madrid, Spain
eciop27@sis.ucm.es

Antonio Rúa
Universidad Pontificia Comillas de Madrid, Departamento de Métodos Cuantitativos
Madrid, Spain
rvieites@cee.upco.es

Cristina del Campo
Universidad Complutense de Madrid, Dept. de Estadística e Investigación Operativa II
Madrid, Spain
campocc@ccee.ucm.es

Once the appropriate variables for socioeconomical knowledge of different regions have been established and reduced into factors, it is proposed to group regions with similar characteristics into clusters so that different clusters are formed by regions with dissimilar behavior patterns attending to socioeconomical characteristics.

1. Introduction

The future development of European Union (EU) regions could be greatly influenced by the economical and political decisions involving local economies made at Brussels. In that sense, to make fair, balanced and homogeneous decisions, it is necessary to acquire adequate quantitative mechanisms that allow to get precise knowledge of the socioeconomical reality.

Once the search for suitable socioeconomical variables, the reduction of their dimension, the conversion of them into factors and the interpretation of those factors have been done, it is pretended to classify different regions so areas with similar characteristics will be integrated in a cluster and different clusters will then show different patterns of behavior in socioeconomical terms.

The fact that a region belongs to a concrete cluster with particular characteristics would be very useful for future decision making.

2. Methodology

Cluster Analisys (CA) is a technique that allows to classify the different elements in a sample (villages of the Spanish province of Segovia in this particular case) into groups called clusters, so, on one hand, each cluster is as homogeneous as possible and, on the other hand, the clusters are as heterogeneous as possible among them.

In our particular case and due to the large number of elements considered (207) the hierarchical method of the *k*-averages have been applied. This method splits the whole set of elements into *k* groups, where *k* is a value previously fixed. The metric used to evaluate the distance among the elements has been the Euclidean one.

3. An Application

The previous methodology has been applied to determine patterns of socioeconomical behavior in the Spanish province of Segovia. With the 17 factors achieved from the previously done Factor Analysis, six clusters have been obtained.

Cluster 1: 1 element (Segovia, the province capital). Focuses, with great difference from the rest, on economic activity due to tourism, basically.

Cluster 2: 1 element (La Granja). Focuses on forest, young populations and important economic activity.

Cluster 3: 69 elements. Focuses on non agriculture living style and old population.

Cluster 4: 5 elements. Focuses on economic activity with similar levels to cluster 2. These 5 elements are the biggest villages, apart from the capital, in the province.

Cluster 5: 19 elements. Focuses on new housing percentage because there is a lot of weekend and holidays visitors.

Cluster 6: 112 elements. Focuses on agriculture living style and pig feeding with low economic activity. This pattern represents the most common description of Segovian villages.

In order to compose the clusters the factor that makes the biggest differences has been "Economic Activity", that has divided the province into three very different groups:

group 1: cluster 1

group 2: clusters 2 and 4

group 3:clusters 3, 5 and 6.

The factor that makes smallest differences is "Medical Attention", where no significant differences have been found.

Acknowledgements

This paper is partially supported by Caja Segovia under project "Búsqueda de Patrones Socioeconómicos en la Provincia de Segovia".

References

Peralta, M.J., Rúa, A., Fernández, L., Borrás, F. (2000). Tipología Socioeconómica de las regiones . Edit.: Consejería de Hacienda de la Comunidad de Madrid. Madrid. In Spanish.

Rúa, A., Borrás, F., Fernández, L., Peralta, M.J. (2000a): Búsqueda de patrones socioeconómicos en la Unión Europea I: Análisis de componentes principales y factorial. *XXV Congreso Nacional de Estadística e Investigación Operativa*, pp. 513-514. Vigo. In Spanish.

Rúa, A., Borrás, F., Fernández, L., Peralta, M.J. (2000 b): Búsqueda de patrones socioeconómicos en la Unión Europea II: Análisis de conglomerados. *XXV Congreso Nacional de Estadística e Investigación Operativa*, pp. 515-516. Vigo. In Spanish.

Rúa, A., Redondo R., del Campo, C., Peralta, M.J. (2001). Búsqueda de tipologías socioeconómicas en los municipios de la provincia de Segovia. Edit.: Caja Segovia. To appear. In Spanish.

REVISTA DE
ESTATÍSTICA
2º QUADRIMESTRE DE 2001

On the Rank-Size (Zipf) Law and the Size Distribution of Human Settlements

William J. Reed
University of Victoria, Department of Mathematics & Statistics,
P.O. Box 3045, Victoria, B.C., Canada, V8W 3P4
reed@math.uvic.ca

The distribution of the sizes (population) of towns and cities exhibits a remarkable degree of regularity both over time and across different regions and jurisdictions. When cities are ranked in descending order by size and the rank plotted against size, the resulting points lie close to a straight line. This empirical property is known as the rank-size law, or in the special case when the slope of the line is negative one, as Zipf's law. It is easy to show that the rank-size property will hold when the distribution of sizes follows a power-law or Pareto distribution in the upper tail. Not surprisingly there have been many attempts to explain the the rank-size property. These fall mainly into two classes : (a) the so-called hierarchical models, (based upon micro-economic assumptions concerning production, consumption, congestion etc.) ; and (b) statistical models based on simple assumptions concerning the underlying stochastic mechanisms generating the size distribution.

While there have been many models from both classes which can replicate the rank-size property, they all apear to be limited in that they only explain the upper tail of the size distribution. This paper offers an explanation for the size distribution of human settlements over the full range of sizes.

The explanation is based on simple stocahstic models for the growth in time of settlements (geometric Brownian motion) and for the foundation of settlements (Yule process). When combined these components lead to a distribution of settlement sizes following a distribution known as the *double Pareto-lognormal* distribution. This distribution has Paretian (power-law) behaviour in *both* tails. Whether or not observed size distributions exhibit power-law behaviour in the lower tail as well as in the upper tail, can be checked empirically. Furthermore the double-Pareto-lognormal distribution can be fitted to the observed size distribution by maximum likelihood and the adequacy of the fit assessed. This provides two separate tests for the adequacy of the model as an explanation for the rank-size property and the overall size distribution.

In the paper a brief description of the components of the model is followed by an outline of the derivation of the double Pareto-lognormal distribution. The model predictions are discussed along with an examination of four datasets for settleement sizes in two provinces in Spain and two states in the U.S.A. Also a brief discussion of the use of statistical models for explaining distributional phenomena is given. It is argued that to explain the size of any individual settlement, economic and geographical factors must be considered, but that this may not be necessary to explain the *distribution* of sizes, since variations in these factors lead to variations in growth rates and times since foundation, which in turn can be regarded as random components in a stochastic model.

A Bayesian Approach to Optimal Alarm

Marília Reis,
Fac. de Ciências da Univ. de Lisboa, Dep. de Estatística e Inv. Oper.l, CEAUL
Cidade Universitária, Bloco C2, piso 2, 1749-016 Lisboa, Portugal
marilia.reis@fc.ul.pt

M.A. Amaral Turkman
Fac. de Ciências da Univ. de Lisboa, Dep. de Estatística e Inv. Oper.l, CEAUL
Cidade Universitária, Bloco C2, piso 2, 1749-016 Lisboa, Portugal
antonia.turkman@fc.ul.pt

K.F. Turkman
Fac. de Ciências da Univ. de Lisboa, Dep. de Estatística e Inv. Oper.l, CEAUL
Cidade Universitária, Bloco C2, piso 2, 1749-016 Lisboa, Portugal
kamil.turkman@fc.ul.pt

Let $\{X_t\}$ be a stationary sequence. The upcrossing of a level u at time $n+j$ is an event $C^{u}_{n+j} = \{X_{n+j-1} < u \leq X_{n+j}\}$, which we will refer to as a catastrophe. In these situations there is much interest in getting an accurate prediction of the time at which the catastrophe will occur so that an alarm can be given in advance and action in order to prevent major damage can be taken.

The most straightforward procedure is to consider the naive alarm system, which is based on linear prediction of the stochastic process. In this case the alarm is given when the predictor exceeds a certain alarm level. This alarm system will not necessarily perform well since it is optimised to produce good predictions of the level of the process and not to predict the occurrence of an upcrossing.

An alarm system is evaluated by it's Operating Characteristics (size of the alarm region, probability of correct alarm, probability of detecting the catastrophe, probability of false alarm and probability of not detecting the event) and is said to be optimal if, for a given probability of detecting catastrophes, it gives the minimum number of false alarms.

In this paper the optimal alarm policy for detecting future upcrossings of the sequence is studied in a Bayesian predictive context for a general AR(p) process and particular calculations are carried for an AR(2) process.

References

M.A. Amaral Turkman and K.F. Turkman (1990). Optimal alarm systems for autoregressive processes, *Computational statistics & Data Analysis* **10**, 307-314.

M.A. Amaral Turkman, Marília Reis and K.F. Turkman (2000). A Bayesian Approach to Event Prediction, *Notas e Comunicações* nº **9**, CEAUL.

Broemeling, L. D. (1985). Bayesian Analysis of Linear Models. Marcel Dekker Inc. New York and Basel.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Xtremes - Frontiers of Computational Extremes

R. D. Reiss, M. Thomas*
University of Siegen, FB 6 - Mathematik
Walter-Flex-Str. 3, D-57068 Siegen, Germany
reiss@stat.math.uni-siegen.de, michael@stat.math.uni-siegen.de

1. Overview of Xtremes

Xtremes is a statistical software package specially tailored for extreme value analysis. It provides an arsenal of visualization tools and parametric procedures, ranging from classical Gaussian models to multivariate extreme value and generalized Pareto distributions.

It is embedded in the statistical environment Risktec which includes a textual and a graphical programming language. A CORBA-based client/server architecture, supported by a DLL-wrapper, facilitates access to the statistical components of Risktec from clients like R or MS Excel.

2. New Statistical Methods in Xtremes

Recently, we were particularly interested in the following questions:

- the statistical modeling of tails in conjunction with the global modeling of distributions with special emphasis laid on heavy-tailed distributions such as sum-stable and Student distributions;
- the Bayesian methodology with applications to regional flood frequency analysis and credibility estimation in reinsurance business;
- conditional extremes;
- multivariate extreme value and peaks-over-threshold models;
- risk assessment of financial assets and portfolios in the presence of fat and heavy-tailed distributions by means of the Value-at-Risk (VaR); also VaR under the Black-Scholes pricing and for general derivative contracts.

3. Features of the Risktec Environment

The Risktec environment suggests innovative ways for the implementation of a user-friendly and extensible statistical software system. It can be used at different levels:

- The menu system of Xtremes is accomponied by a context-sensitve help system and provides an easy access even for the unexperienced user.
- An integrated formula interpreter allows simple enhancements of the menu system.
- The Pascal-based programming language StatPascal can be employed to implement extensions to the Risktec environment. StatPascal provides vector and matrix operations like other statistical languages, while retaining

* representing the paper in poster session

compatibility to standard Pascal for an easy adoption of existing statistical procedures and allowing compilation for an efficient execution.

- A CORBA-based component architecture enables external clients to utilize the statistical components of the Risktec environment. Further components can be added to the environment easily.
- The graphical programming environment XGPL is a visual tool for the combination of statistical components of the Risktec environment.

In the poster session, we demonstrate the application of the Risktec environment and discuss details of its component architecture.

References

Reiss, R.-D. and Thomas, M. (2001). *Statistical Analysis of Extreme Values*. Birkhäuser, Basel (1st ed., 1997).

Thomas, M. and Reiss, R.-D. (2000). Graphical Programming in Statistics: The XGPL Prototype. In: Decker, R. and Gaul, W. *Classification and Information Processing at the Turn of the Millenium.* Springer, Berlin.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

L^1 Density Estimation for Dependent Random Vectors

Noureddine Rhomari
CREST-INSEE, France and University of Oujda, Morocco[*]
CREST-Labo. Stat, Timbre J340 3, Av. Pierre Larousse, 92 245 Malakoff Cedex, France
rhomari@ensae.fr

The nonparametric probability density estimation plays a significant role in the statistical inference. In this paper we are interested in L^1 convergence of the kernel estimator f_n of a common probability density f of weakly dependent random vectors, not necessarily stationary. As pointed by Devroye and Györfi (1985), in addition to L^1 is the natural space of the densities, the study of the L^1 error $J_n = \int_{IR^d} |f_n(x) - f(x)| \, dx$ is justified by its characteristic properties such invariance by some one-one onto transformations (e. g. scaling), its aspect visual and it is 2 times the total variation of associated probability measures; see also Devroye (1987). This quantity, J_n, was largely studied for independent observations; Devroye (1983,87,91), Györfi and Devroye (1985) and Pinelis (1990) and the references therein. But, within the dependent framework it was treated little, one can quote for example Györfi (1987), Györfi et al. (1989), Tran (1989) and Danga (1992,94).

By considering at the same time strongly mixing and absolutely regular processes (a and b -mixing), we prove the almost sure convergence or in probability of J_n to 0 under weak conditions and we also specify their rates. We show for example that for regular processes (*i.e.* $b(n) \to 0$) the density kernel estimate is L^1 consistent in probability for a suitable smoothing parameter, without any condition on the density. Moreover the conditions and the rates are optimal as soon as the coefficients $b(\cdot)$ are summable; they are the same that the independent case. But in the a -mixing case the conditions and the rate are almost optimal when the coefficients decrease geometrically. Some of our results improve and extend those of Györfi (1987), Györfi et al. (1989), Tran (1989) and Danga (1992,94) obtained for weakly dependent and strictly stationary processes.

More precisely, let $X_1, \ldots, X_n$ be n observations from the unknown density f on IR^d, the kernel estimator f_n is defined for $x \in IR^d$ by: $f_n(x) = \frac{1}{nh_n^d} \sum_{t=1}^{n} K\left(\frac{x - X_t}{h_n}\right)$, where K is a real function on IR^d, integrable of integral 1 and (h_n) a sequence of positive numbers.

Under some conditions on the mixing coefficients related to the smoothing parameter h_n we show that $J_n \to 0$ as $n \to 0$. These conditions are very weak that $a(n) = o(n^{-1})$ or $b(n) = o(1)$ suffices to have probability consistence for a suitable parameter smoothing. When the b -mixing coefficients are summable we find the optimal conditions of independent case; this class of processes contains among others, the m -dependent, ARMA, some linear, nonlinear and functional AR processes.

---

[*] Permanent Adress: Université Mohamed 1er, Faculté des Sciences, Département de Maths. et Info., Route Sidi Maafa, 60 000 Oujda, Morocco, rhomari@sciences.univ-oujda.ac.ma

* For example if $a(n) = O(e^{-t\ n}), t > 0$, then $h \to 0$ and $nh^d / \log n \to \infty$ imply $J_n \to 0$ as $n \to 0$ and if in addition f has a regularity s (as Holder regularity) then $h_n \approx (n/\log n)^{-1/(2s+d)}$ yields $J_n \overset{a.s.}{=} O\left((n/\log n)^{-s/(2s+d)}\right)$. So the loss compared to the optimal rate of the independent case is only logarithmic.

* But in b -mixing case the conditions are simpler than the previous and close to those under independence. We prove that for any density f we have (with $p = p_n \le n/2$), $h \to 0$ and $nh^d \left(\sum_{i=1}^{p} b(i)\right)^{-1} \to \infty \Rightarrow J_n \to 0$, *i)* in probability if $p/n \to 0$ and $(n/p)b(p) \to 0$ and *ii)* almost surely if $p \log n / n \to 0$ and $\sum_n (n/p) b(p) < \infty$. Thus

1) If $\sum_n b(n) < \infty$ then $h \to 0$ and $nh^d \to \infty \Leftrightarrow J_n \to 0$, a.s. and $J_n \overset{a.s.}{=} O\left(n^{-s/(2s+d)}\right)$ with s the regularity of f and $h_n \approx n^{-1/(2s+d)}$.

2) If $b(n) = O(n^{-1})$ then $h \to 0$ and $nh^d / \log n \to \infty \Rightarrow J_n \to 0$, and $J_n = O_P\left((n/\log n)^{-s/(2s+d)}\right)$ for $h_n \approx (n/\log n)^{-1/(2s+d)}$.

3) If $b(n) = O(n^{-t}),\ 0 < t < 1,$ then for all $e_n \to 0, J_n = O_P\left(\left\{(1-t)e_n\ \ n^{2t/(1+t)}\right\}^{-s/(2s+d)}\right)$.

4) If $b(n) = o(1)$ there exists $h \to 0$ such that $J_n \to 0$.

We note that all the constants in the above $O(\cdot)$'s are explicit and that similar results are valid for L^p, $p \ge 1$ and in mean for L^p. The principal tools to prove these results are Bernstein type inequalities for dependent processes with values in a separable Banach space due to Rhomari (2000).

References

Danga, A. (1992). Moment de sommes partielles et estimation de la densité dans L^1 pour les processus mélangeant, *C. R. Acad. Sc. Paris*, Série A, **315**, 459-463.

Danga, A. (1994). Estimation de la densité dans L^1 pour des variables dépendantes ou censurées, *C. R. Acad. Sc. Paris*, Série A, **319**, 739-744.

Devroye, L. (1983). The equivalence of weak, strong and complete convergence in L^1 for kernel density estimates, *Ann. Statist.*, **11**, 896-904.

Devroye, L. (1987). A course in Density Estimation, Birkhäuser, Boston.

Devroye, L. (1991). Exponential inequalities in nonparametric estimation. In *Nonparametric Functional Estimation and Related Topics*, (G.G. Roussas, editor), 33-44. Kluwer.

Devroye, L. and Györfi, L. (1985). Nonparametric density estimation: The L^1 View. Wiley.

Györfi, L. (1987). Density estimation from dependent sample. In *Statistical Data Analysis Based on the L^1-Norm and Related Topics*, (Y. Dodge, editor), 393-402. Elsevier, North-Holland.

Györfi, L., Härdle, W., Sarda, P. and Vieu, Ph. (1989). Nonparametric curve estimation from time series. *Lecture Notes in Statistics*, **60**.

Pinelis, I.F. (1990). Inequalities for distribution of sums of independent random vectors and their application to estimating density, *Theory Probab. Appl.* **35**, 605-607.

Rhomari, N. (2000). Approximation et inégalités exponentielles pour les sommes de vecteurs aléatoires dépendants avec applications. *Actes 32e Journées Stat. de la SFdS.*

Tran, L.T. (1989). The L^1 convergence of kernel density estimates under dependence, *Canadian J. Statist.* **17**(2), 197-208.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Toward an Accurate Model of Random Events

Paolo Rocchi
IBM
via Shangai 53, 00144 Roma, Italy
paolorocchi@it.ibm.com

1. Foreword

Down the centuries several theories have been proposed but the debate on the roots of probability and statistics is still open. Specialists are used to present and discuss their theories in the whole; the scientific analysis is lacking. The debate between the schools becomes a philosophical confrontation and we do not get results in this dialog of the deaf.

All the probability theories assume the random event as the argument of probability therefore a logical order requires that first we must discuss the argument and then its measure. This procedure traces an analytical way out and the random event modeling is a right start for clarifying the foundations of probability. A contribution in this direction comes from Quantum Physics that since years put to light several experiments critical to mathematical modeling.

As first we remember Kolmogorov who affirms the random event X is a set of the particular events Ex

(1) $$X = \{Ex\}$$

when X is a subset of the sample space and the probability is the measure of X

(2) $$P = P(X)$$

1. Kolmogorov interprets the set X as the "result" of the event. However the result is a part and the event is the whole. The properties of the event are quite different from the properties of the result and we cannot merge the *set of events* and the *set of results* without a logical justification.

2. The set model cannot refer to all the results. Quantum Physics brings up the “two slits experiment” as an exception of significant importance.

Several subjectivists and bayesians appreciate the linguistic model for the random event. However

a. Several terms of the natural language are generic and ambiguous thus the sentence X appears inadequate to represent the random event in general.

b. From the sentence X we cannot formally derive the result, in other words X is a qualitative model inadequate to a mathematical theory.

We find conclusive confirmations of the inadequate theoretical models of events in applied calculations. E.g. The probability that the variable y is greater the constant k is written

(3) $$P(y > k)$$

In such a way we refuse the linguistic and set models and we prove their failing.

2. Structural Model

We searched for a solution of the above written difficulties and we designed a theoretical framework based on a new model for the random event.

We was convinced that interacting and connecting is the inner nature of events and we make the following assumption

2.1. *The idea of relating, of connecting, of linking is a primitive.*

This primitive suggests two elements specialized in relating and in being related that we define as such

2.2. *The relationship R connects the entities and we say R has the property of connecting.*

2.3. *The entity E is connected by R and we say E has the property of being connected.*

They are symmetric and complete since they exhaust the Primitive 2.1). From Definitions 2.2) and 2.3) follows that the relationship *R* links the entity *E* and they give the ensemble

(4) $$S = (E; R)$$

which is an algebraic structure.

The expression (4) provides an accurate model for events since *E* and *R* describe the parts of an event. As an example an entity is a dice, a spade, heads, tails, a product. The relationship that connects two or more entities is, for instance, a mechanism producing the output from the input, a force, a physical interaction.
The introductory presentation of (4) is to be completed. In (4) the event is given as a whole that is *S*; then it is defined in terms of the details *E* and *R*. This analysis can be insufficient and we reveal the entities *(E1,E2..,Em*) and the relations *(R1, R2..,Rp)*; these are exploded at a greater level, and so forth

$$\begin{aligned} S &= \\ &= (E;R) = \\ &= (E1,E2...,Em;R1,R2...,Rp) = \\ &= (E11,E12.....,Em1,Em2..,Emk;R11,R12.....,Rp1,Rp2..,Rph) \end{aligned} \quad \textbf{(5)}$$

In conclusion the *structure of levels* (5) is the complete and rigorous model of any event. The levels can also be written as

$$\begin{array}{ll} \textit{level 0} & S \\ \textit{level 1} & E;R \\ \textit{level 2} & E1,E2...,Em;R1,R2...,Rp \\ \textit{level 3} & E11,E12.....,Em1,Em2..,Emk;R11,R12.....,Rp1,Rp2..,Rph \end{array} \quad \textbf{(6)}$$

The multiple level decomposition is already used in software methodologies, in modern ontology and in various other sectors. The progressive explosion is also known in Probability Calculus E.g. We use the tree in detailing a decisional event.

The structure of levels meets the Kolmogorov theory when the result *Ej* is a set and *Ej1…,Ejs* are the subsets of *Ej*. E.g. The assumption is largely valid in gambles. The structure is compatible with the linguistic representation and in some cases is exactly symmetrical. E.g.

$$\begin{array}{ccccc} & \textit{“The coin} & | & \textit{comes down} & | \ \textit{heads”} \\ & \textit{Ein} & & \textit{R} & \textit{Eout} \end{array} \quad \textbf{(7)}$$

References

P. Rocchi (1998). La Probabilità è oggettiva o soggettiva ? - Pitagora, Bologna.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Inference on the Location Parameters — Internally Studentized Statistics

José Rocha
Universidade dos Açores e Centro de Estatística e Aplicações da Universidade de Lisboa
jrocha@notes.uac.pt

With the usual notations for the mean, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and for the sum of squares $SS_n = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2$, and from the recurrence relations $\overline{X}_{n+1} = \frac{n}{n+1}\overline{X}_n + \frac{1}{n+1}X_{n+1}$ and $SS_{n+1} = SS_n + \frac{n}{n+1}\left(\overline{X}_n - X_{n+1}\right)^2$ it is easy to get that the joint probability density function f_{n+1} of $\left(\overline{X}_{n+1}, SS_{n+1}\right)$ may be expressed in terms of f_n and of the probability density function f of the parent population:

$$f_{n+1}(w,s) = \sqrt{\frac{n+1}{n}}\, s \int_{-1}^{+1} f_n\left[w + \sqrt{\frac{s}{n(n+1)}}v, s\left(1-v^2\right)\right] f\left(w - \sqrt{\frac{ns}{n+1}}v\right) dv.$$

From this expression, and from the easily established formula

$$f_2(w,s) = \sqrt{\frac{2}{s}}\; f\left(w + \sqrt{\frac{s}{2}}\right) f\left(w - \sqrt{\frac{s}{2}}\right), w \in IR,\ s > 0$$

we derive

$$f_3(w,s) = \sqrt{3} \int_{1}^{+1} \sqrt{\frac{1}{1-v^2}}\, f\left(w + \frac{v + \sqrt{3\left(1-v^2\right)}}{\sqrt{6}}\sqrt{s}\right) f\left(w + \frac{v - \sqrt{3\left(1-v^2\right)}}{\sqrt{6}}\sqrt{s}\right) f\left(w - \frac{2v}{\sqrt{6}}\sqrt{s}\right) dv$$

that can be rewritten in the form $f_3(w,s) = \sqrt{3} \int_{-1}^{+1} \sqrt{\frac{1}{1-v^2}} \prod_{i=1}^{3} f\left(w + a_{i3}(v)\sqrt{s}\right) dv$ with

$$a_{13}(v) = \frac{v + \sqrt{3(1-v^2)}}{\sqrt{6}},\ a_{23}(v) = \frac{v - \sqrt{3(1-v^2)}}{\sqrt{6}},\ a_{33}(v) = -\frac{2v}{\sqrt{6}},$$ i. e.,

$$\sum_{i=1}^{3} a_{i3}(v) = 0 \text{ and } \sum_{i=1}^{3} a_{i3}{}^2(v) = 1.$$

From there we may obtain the general expression,

* Research partially supported by FCT/POCTI/FEDER

$$f_{n+1}(w,s)=\sqrt{\frac{(n+1)s}{n}}\;2^{n-2}\;\sqrt{n}\left[\prod_{i=3}^{n}\left(1-x_i^2\right)^{\frac{i-4}{2}}\right]s^{\frac{n-3}{2}}$$

$$\int_{-1}^{+1}\left\{\left(1-v^2\right)^{\frac{n-3}{2}}\prod_{i=1}^{n}f\left[w+\left(\frac{v}{\sqrt{n(n+1)}}+a_{in}\sqrt{1-v^2}\right)\sqrt{s}\right]f\left(w-\sqrt{\frac{ns}{n+1}}\right)\right\}dv\,.$$

From that, under mild regularity conditions, we may obtain an approximate expression for internally studentized statistics $T_{(n-1)}=\sqrt{n(n-1)}\,\frac{\overline{X}_n}{\sqrt{SS_n}}$, that in the case of gaussian populations (a special case, where the studentization is external) is exact:

$$f_{T_{(n-1)}}(t)\propto\int_0^{\infty}u^{n-1}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{t^2u^2}{n(n-1)}+a_{in}^2u^2+2\frac{ta_{in}u^2}{\sqrt{n(n-1)}}\right\}du \quad \text{with} \quad \sum_{i=1}^{n}a_{in}=0 \quad \text{and}$$

$\sum_{i=1}^{n}a_{in}^2=1$, and hence $f_{T_{(n-1)}}(t)\;\propto\;\int_0^{\infty}u^{n-1}\exp\left\{-\frac{u^2}{2}\left(\frac{t^2}{n-1}+1\right)\right\}du\propto\;\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}$.

Hotteling's (1961) expressions for Cauchy and Laplace parent distributions, and many others, may be obtained via the general expression.

References

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

David, H. A. (1971) *Order Statistics*. Wiley, New York.

Hotelling, H. (1961) The behavior of some standard statistical tests under nonstandard conditions. *Proc. 4th Berkeley Symp. Math. Statist. Probab*. **I**, 319-359.

REVISTA DE
ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Measures of Performance for Discordancy Tests in Normal Populations

Fernando Rosado
University of Lisbon, Faculty of Science and Center of Statistics
Edifício C2 - Campo Grande, Lisboa, Portugal
fernando.rosado@fc.ul.pt

José Palma
Superior Scholl of Technology – EST/IPS, Mathematics Department
Rua Vale de Chaves, Estefanilha, Setúbal, Portugal
jpalma@est.ips.pt

It is obvious the interest for the detection of outliers in samples from normal populations, since they can be contaminated by "surprising" observations. The treatment to give to outlying observations were, for a long time, subjected to study. Traditionally the approach to its study was in the sense of discovering them through subjective discordancy tests.

The tests of outliers as any other tests of hypotheses, should have a null and an alternative hypotheses. The null hypotheses it should express some probabilistic basic model for the generation of the whole sample without outliers contemplation and the alternative expresses one way in which the model should be modified to explain or incorporate them.

The construction of the tests depends therefore, in first analysis, of the alternative hypothesis formulated in the discordancy model. The study of the power of the test, the construction of tests with certain desirable properties, always demands that the outliers model is specified.

The most studied tests consider alternative hypotheses like: the inherent, contamination, slippage, and natural generative alternative (GAN). This last model is more general in the sense of the non restriction from initial alternative.

To each alternative model, it correspond a very special situation to the tests that have been formulated, in most of the cases it has not been possible to present discordancy tests that use in full such hypotheses without any restriction.

Many tests have been proposed in the literature, more than forty only for populations with normal distribution. In most of the statistics of test proposal its construction resulted of the application of an obvious beginning, use of extreme order statistics with the unknown parameters substituted by extreme order statistics sufficient for them. Later on, for many of those statistics they were discovered optimal properties, generally long after they have been proposed.

The outliers detection in most of the cases has not been done by rigorous and objective methods, not only in terms of construction of the test statistics as in the selection of the observations to test, having been used above all intuitive processes (for example candidates to outliers are chosen empirically firstly). Only with the method GAN the problem is been treated on an objective form, being the observation rejected as outlier chosen a posteriori, once rejected the homogeneity of the observations.

In the great diversity of discordancy tests it is fundamental to gauge its performance. To chose between rival tests we need to have some useful measures of

their relative performance, for example their power. The comparison of tests with the same power should depend on the alternative hypothesis that we have in mind to explain the outliers and demand other performance measures. It then requires knowledge of the distributional behaviour of the test statistic under this alternative hypothesis. This often presents difficult and complicated problems, and in the past many people have either ignored it or have confined themselves to simulated results.

In view of the central position which the normal distribution occupies in statistical theory, it is not surprising to find that the question of outliers from normal samples has received both the earliest and the most concentrated study in the outlier theory. However, in spite of the great variety of discordancy tests proposed in the literature, they are quite limited the performance studies that allow to compare them and the ones that exist are restricted to simulated results of the power of the tests.

With this communication we intend to identify significant performance measures in the field of the normal distribution, we also discusses some of the problems placed to outliers detection and to the construction of test statistics that drove to its great diversity.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Factor Socioeconomical Description of Local Economies: an Application

Antonio Rúa
Universidad Pontificia Comillas de Madrid, Departamento de Métodos Cuantitativos
Madrid, Spain
rvieites@cee.upco.es

Raquel Redondo, Cristina del Campo
Universidad Complutense de Madrid, Dept. de Estadística e Investigación Operativa II
Madrid, Spain
eciop27@sis.ucm.es, campocc@ccee.ucm.es

A socioeconomical characterization of local economies is looked for in this paper. The first thing to be done consists on selecting, defining and describing the appropriated socioeconomical variables. Afterwards, in order to allow an easier interpretation, a factor analysis is applied to reduce data dimension and detect variable relations. Finally an exam the different resulting factors is required.

1. Introduction

We are involved in a complicated political, social and economical structure as a result of the integration of villages in counties, counties in regions, regions in countries inside European Union. This complexity interferes with decision making what has effects on local economies. When making those decisions it is always possible to chose wrongly due to the absence of objective information about the real socioeconomical situation of that local economy. Then disposing of objective quantitative methods to get that knowledge becomes necessary so that decisions will be just, balanced and homogeneous. The main objective of this paper is to set an appropriated methodology in order to characterize socioeconomicaly regions inside European Union. Therefore it is necessary to:

1. Search for suitable socioeconomical variables, reduce their dimension, converting them into factors and interpret those factors.
2. Classify different regions so that areas with similar characteristics would be integrated in a cluster and different clusters would suppose different patterns of behaviour in socioeconomical terms.

First point will be developed in the present communication, while the second one will be the main subject of contribution entitled *"Cluster setting of socioeconomical patterns in local economies. An application"*, also presented to this meeting.

2. Data

The starting point of the analysis must be an appropriated database, official statistics if possible. In order to detect possible missing data, a revision of that base is recommended. If collect information present missing data for any variable, this lack of information can be completed using multiple regression with other related variables with "bondad de ajuste" greater than 75%. Afterwards, to avoid scale problems, ratios should be defined. So it is useful to group the variables into different sets, for example, demography, activity and unemployment, R&D, agriculture, energy, transport, living conditions and so on. To define ratios, it is necessary to have control variables, commonly

those control variables are population and extension. Then the ratios are made as: ratio=variable/control variable and all the ratios are classified in the same groups as the original variables.

3. Methodology

Principal Component Analysis (PCA) and Factor Analysis (FA) are both multidimensional methods to examine the interdependence among variables. PCA pretends to reduce the sets of original ratios into a new smaller set of variables called Principal Components so that the minimum possible number of components explains the maximum ratio variability. That is the reason why it must be used with exploratory intention. Meanwhile FA is used with confirmatory intention and defines factors that show interaction among ratios, and the meaning of factors is very close to components. Then PCA must be applied on data first and, if the analysis shows that is appropriated a dimension reduction FA will be applied to get the factors that summarize the variables information.

Factors must be interpreted related to the variables with largest scores in the factor.

4. An Application

The previous methodology has been applied to determine patterns of socioeconomical behavior in the Spanish province of Segovia. A number of 58 variables has been defined and classified into seven groups: control, demography, activity, economy, living conditions, agriculture and tourism. PCA had shown it was possible to reduce data dimension and FA had confirmed it. A number of 17 factors had been obtained explaining more than 75% of the total variance. The most important factors have been named: ageing, living style, economic activity, public expenses, medical attention and unemployment.

Acknowledgements

This paper is partially supported by Caja Segovia under project “Búsqueda de Patrones Socioeconómicos en la Provincia de Segovia”.

References

Peralta, M.J., Rúa, A., Fernández, L., Borrás, F. (2000). Tipología Socioeconómica de las regiones . Edit.: Consejería de Hacienda de la Comunidad de Madrid. Madrid. In Spanish.

Rúa, A., Borrás, F., Fernández, L., Peralta, M.J. (2000a): Búsqueda de patrones socioeconómicos en la Unión Europea I: Análisis de componentes principales y factorial. *XXV Congreso Nacional de Estadística e Investigación Operativa*, pp. 513-514. Vigo. In Spanish.

Rúa, A., Borrás, F., Fernández, L., Peralta, M.J. (2000 b): Búsqueda de patrones socioeconómicos en la Unión Europea II: Análisis de conglomerados. *XXV Congreso Nacional de Estadística e Investigación Operativa*, pp. 515-516. Vigo. In Spanish.

Rúa, A., Redondo R., del Campo, C., Peralta, M.J. (2001). Búsqueda de tipologías socioeconómicas en los municipios de la provincia de Segovia. Edit.: Caja Segovia. To apear. In Spanish.

Department of Mathematics
Las Palmas de Gran Canaria. 35017
Saavedra@dma.ulpgc.es, jartiles@dma.ulpgc.es, cflores@dma.ulpgc.es

Inmaculada Luengo-Merino
Campus Universitario de Tafira, Department of Informática y Sistemas
Las Palmas de Gran Canaria. 35017
mluengo@dis.ulpgc.es

1. Introduction

The consistency of several bootstrap procedures related to longitudinal data analysis require to estimate any probability covergence orders corresponding to Mallows distance. This distance, in our paper, is considered between the probability distribution of a random vector with increasing dimension *n* and its empirical probability distribution for a random sample with also increasing size *r*. This paper gives the conditions under which the above mentioned metric converges to zero in probability. Bickel and Freedman (1981) prove that the Mallows distance between a probability distribution defined on a Banach Space and its corresponding empirical distribution converges to zero almost sure, but they do not give the convergence order. In section 2 a condition is given under which the convergence in probability order for distributions on R is $r^{-1/2}$, being *r* the sample size. A such condition is satisfied by the uniform distribution over $[0,1]$. Using this result, in section 3 a theorem gives the convergence order of Mallows distance between the probability distribution of a random vector with increasing dimension *n* and its empirical distribution based on random sample with increasing size *r*.

2. Convergence of Mallows Metric for Distributions on R

First, we gives some definitions and notations. For $p \geq 1$, let $\Gamma_p(\mathrm{R}^n)$ be the set of probability distributions F on R^n such that $\int \|x\|^p \, dF(x) < \infty$. For the probability distribution functions *F* and *G* in $\Gamma_p(\mathrm{R}^n)$, the Mallows distance $d_p^{(n)}(F,G)$ is defined as the inferior of $E\left[|X-Y|^p \right]^{1/p}$ over the pairs of random vectors *X* and *Y*, such that *X* has law *F* and *Y* has law *G*. If *X* and *Y* are *n*-dimensional random vectors with the probability distributions *F* and *G* respectively, we can understand $d_p^{(n)}(X,Y)$ for $d_p^{(n)}(F,G)$.

According to lemma (4) of Bickel and Freedman (1981), $d_2^1(F_r,F) \to 0$ almost sure. We now give a condition under which this order is $r^{-1/2}$.

Theorem 1 Let $X_1,\ldots,X_r$ be independent and R-valued random variables, with law *F* and density function *f*. Let F_r be the empirical distribution function corresponding

to the random sample $X_1,\cdots,X_r$ and $X_{(r,1)},\cdots,X_{(r,r)}$ the order statistics. If $\sum_{i=1}^{r}\int_{F^{-1}((i-1)/r)}^{F^{-1}(i/r)} E\left[\left(X_{(r,i)}-x\right)^2\right] f(x)\cdot dx = O\left(r^{-1}\right)$ is satisfied then $d_2^1\left(F_r,F\right)=O_p\left(r^{-1/2}\right)$.

3. Convergence of the Mallows Metric for Probability Distributions on R^n

Finally, the follow result gives the form of Mallows metric between the probability distribution of a random vector, with incorrelated and identically distributed components and increasing dimension ν and its empirical probability distribution.

Theorem 2 Let $\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_r$ be independent *n*-dimensional random vectors $\mathbf{X}_i=\left(X_{i1},\cdots,X_{in}\right)$ with law $F^n\in\Gamma\left(\mathbf{R}^n\right)$ and being F the law of the components X_{ij}. Let F_r^n be the empirical distribution of $\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_r$ and F_r the empirical distribution of X_{1j}, X_{2j}, $\cdots$, X_{rj}. Then $d_2^n\left(F^n,F_r^{\nu}\right)\le n^{1/2}d_2^1\left(F,F_r\right)$.

References

Artiles-Romero, J., Hernández-Flores, C.N., Luengo-Merino, I. and Saavedra-Santana, P. (2001) A Comparison of two Population Spectrums. University of Las Palmas de Gran Canaria. *Preprint.*

Bickel, P. and Freedman, D. (1981) Some Asymptotic Theory for the Bootstrap. *Ann. Statist*., **9**, 1196-1217.

Hernández-Flores, C.N., Artiles-Romero, J. and Saavedra-Santana, P. (1999) Estimation of the Population Spectrum with Replicated Time Series. *Comp. Stat. and Data Anal.*, **30**, 271-280.

Saavedra, P., Hernández, C.N. and Artiles, J. (2000) Spectral Analysis with Replicated Time Series. *Communications in Statistics Theory and Methods*. **29**, 2343-2362.

Saavedra-Santana, P., Luengo-Merino, I., Hernández-Flores, C.N., and Artiles-Romero, J., (2001). Homogeneity Test for a Set of Time Series. University of Las Palmas de Gran Canaria. *Preprint.*

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Temperature Extremes in the North-Western Iberian Peninsula

Clemente Tomás Sánchez
University of Salamanca, Department of Atmospheric Physics
Plaza de la Merced s/n, 37008 Salamanca, Spain
cts50@gugu.usal.es

Fernando de Pablo Dávila
University of Salamanca, Department of Atmospheric Physics
Plaza de la Merced s/n, 37008 Salamanca, Spain
fpd123@gugu.usal.es

Solange Mendonça Leite
University of Lisbon, Geophysical Center
Rua da Escola Politécnica, nº 58, 1250-102 Lisbon, Portugal
solange@utad.pt

Much of environmental research has dwelt on average behavior, as well as possible changes in these averages. Nevertheless, some of the principal impacts of environment on society arise throughout its variability and, in particular, the occurrence of extreme events (Wigley, 1985). Moreover, temperature plays an important role, especially when assessing the impacts of extreme values variability in agriculture, safety, infrastructure, health, energy, economy, etc.

For instance, crop yield is dependent partly on extreme weather events. The relationships are difficult to unravel, however, because other factors are important. Sometimes, the occurrence of a single rare event such as a hail storm is the limiting stress. The investigator must recognize the existence of weather-sensitive factors, which produce significant correlations between yield and particular meteorological elements but which do not participate in a direct cause-effect relation. Accident rates are also weather dependent. For instance, traffic accidents are most frequent during fog and freezing rain. Many common diseases also show seasonal cycles but there is no obvious explanation in most cases. As a particular example, a number of investigators have suggested that ulcers are aggravated by temperature changes (Rao and Chakraborty, 1991).

Extreme temperature events elsewhere in the world will have global implications through geophysical, socio-economic and political mechanisms. Improved understanding of these occurrences is essential to assess the likely range of future climate extreme events and the extent to which these extreme events are predictable (Mearns *et al*., 1984; Katz and Brown, 1992; Sánchez *et al*., 1997).

Keeping this issues in mind, a statistical analysis of extreme temperature events will be performed in this presentation.

This presentation will focuse on the extremes of temperature of four weather stations in North-Western Iberian Peninsula. Monthly and annual absolute temperature maxima and minima observed at four weather stations for the period 1941-1999 were used. The selection of these four weather stations intended to bring prominence to the contrast in climatic types, which is a basic goal of extreme value analysis in Climatology. Salamanca and Zamora, the two Spanish weather stations selected, are

plainly continental, while Braga and Coimbra, the two Portuguese weather stations selected, clearly present oceanic characteristics. This is easily confirmed by the difference between the mean maximum temperature of the hottest and the mean minimum temperature of the coldest month, for all the period analyzed. Generally speaking, higher values of the thermal amplitude link with continental places, while maritime regions present lower values of the thermal amplitude.

Geographical parameters of the four weather stations are presented in Table 1.

Weather Station	Latitude (ºN)	Longitude (ºW)	Altitude (m)	Year
Braga	41º 33'	08º 24'	190	1941-88
Coimbra	40º 13'	08º 27'	35	1941-88
Salamanca	40º 56'	05º 29'	789	1945-97
Zamora	41º 31'	05º 44'	667	1933-99

Table 1. Geographical parameters of the four selected weather stations in North-Western Iberian Peninsula.

Monthly and annual extreme temperatures are analyzed using the non-parametric Spearman test in search for possible significant trends. In spite of the impossibility to succeed in deterministic forecasting for temporal scales longer than some days, it is necessary to simulate climatic variability on a wide range of time scales, by means of probabilistic techniques.

For now, we present the results obtained by applying the classical Gumbel distribution function to annual maximum and minimum temperature series, the risk of occurrence for several return periods and the best-fitting distribution.

References

Katz, R. W. and Brown, B. G. (1992). Extreme events in a changing climate: variability is more important than averages, *Climatic Change*, **21**, 289-302.

Mearns, L. O., Katz, R. W. and Schneider, S. H. (1984). Extreme high-temperature events: changes in their probabilities with changes in mean temperature, *J. Clim. Appl. Meteorol.*, **23**, 1601-1613.

Rao, C. R. and Chakraborty, R (1991). Statistical methods in biological and medical sciences. North-Holland. Amsterdam

Sánchez, J. M., Tomás, C., de Pablo, F. (1997). Consideraciones sobre el clima de Matácan (Salamanca). Ed. Caja Duero, Spain.

Wigley, T. M. L. (1985). Impact of extreme events, *Nature*, **316**, 106-107.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Pairwise Multiple Comparisons for Repeated Measurements

Ernst Schuster
University of Leipzig, Institute for Medical Informatics, Statistics and Epidemiology
Liebigstr. 27, 04103 Leipzig, Germany
Schuster@imise.uni-leipzig.de

Siegfried Kropf
University of Leipzig, Coordination Centre for Clinical Trials
Prager Straße 34, 04317 Leipzig, Germany
Siegfried.Kropf@kksl.uni-leipzig.de

1. Introduction

Frequently a response variable is measured to several fixed time points. We deal with the following multivariate Gauss model:

(1) $$\mathbf{x}_j = \begin{pmatrix} x_{j1} \\ \vdots \\ x_{jk} \end{pmatrix} \sim N_k \left(\begin{pmatrix} m_1 \\ \vdots \\ m_k \end{pmatrix}, \mathbf{S} \right)$$

for $j = 1,...,n$ independent sample vectors and $i = 1,...,k$ dependent time points. The variances s_{ii} $(i = 1,...,k)$ will often be equal in applications.

For small *k*, usually a multivariate analysis of variance is applied, where all elements of the covariance matrix are estimated. Under the compound symmetry assumption, one can alternatively use a univariate ANOVA test (Timm,1980).

In addition to the global test of H_0: $m_1 = m_2 = ... = m_k$, usually orthogonal contrasts, such as Helmert contrasts or polynomial contrasts, are calculated from statistical packages as SPSS or SAS.

The so-called experimentwise error rate of a multiple comparison method is the supremum of the probability of making at least one incorrect assertion (Hsu, 1996) in all decisions of the procedure. The simplest way to ensure this experimentwise error rate is the Bonferroni adjustment, where each single test of the procedure uses the local level á/*m* for *m* simultaneous tests. Frequently, especially for nonlinear curves, the user wishes to *m* may be small.

Subsequently an alternative procedure for multiple comparisons is developed. We start with the principle of a-priori ordered hypotheses (Maurer, Hothorn and Lehmacher, 1995). Testing the hypotheses in the given a-priori order, we can use the full level á in each comparison, however, we have the stop the procedure when for the first time a hypothesis cannot be rejected. The remaining hypotheses are considered as not significant at the experimentwise level regardless of their results in the local tests.

The problem of defining a useful a-priori order of the hypothesis can be avoided by a theorem of Kropf (2000) which considers tests for the univariate hypotheses $H_i : m_i = 0$ (i=1,…,k) in the above Gauss model (1) (i.e., the time points are considered separately, not the differences among different points in time):

- The *k* time points are ordered for decreasing values of $\sum_{j=1}^{n} x_{ji}^2$ for $i = 1,\ldots,k$.

- In this order, the usual two-tailed *t* tests for the hypothesis $m_i = 0$ are applied at the full level á as long as all tests are significant. The procedure stops with the first non-significant result.

variances at different time points. However, the assumption of equal variances is necessary in order to have an indication for a convenient order of hypotheses, i.e. for the power of the multiple procedure.

The theorem is now applied to the all-pair comparisons between the different time points.

2. A New Procedure for All-Pairwise Comparisons of Dependent Samples

We consider the $p = k(k-1)/2$ pairs (1,1), (1,2),…,$(k-1,k)$ of different time points and calculate the corresponding differences $d_{j1},\ldots, d_{jp}$ for each sample vector $\mathbf{x}_j$ ($j = 1,\ldots, n$): $d_{j1} = x_{j1} - x_{j2}, \ldots, d_{jp} = x_{j,k-1} - x_{jk}$.

For $j = 1,\ldots, n$ the vectors $(d_{j1},\ldots,d_{jp})'$ are independent from each other and have a multivariate normal distribution with expectation $(q_1,\ldots,q_p)'$. Under the additional compound symmetry assumption for the vectors $\mathbf{x_j}$, the p components of the vector of differences have also equal variances. Therefore the above theorem is applicable for the hypotheses $q_l = 0$, $l = 1,\ldots, p$, resulting in the following procedure:

- Order the $p = k(k-1)/2$ differences of time points for decreasing values of
$$\sum_{j=1}^{n} d_{jl}^2,\ l = 1,\ldots, p\ .$$
- In this order, carry out the usual two-tailed *t* tests (corresponding to the usual *t* test for pair differences) for the hypotheses $q_l = 0$, $l = 1,\ldots, p$ at the full significant result.

hypotheses may be useful because $\sum_{j=1}^{n} d_{jl}^2 / n = \frac{n-1}{n} s_l^2 + \bar{d}_l^2$ for each $l = 1,\ldots, p$. Therefore with equal variances for all differences, the order of hypotheses is mainly determined by the mean differences. Pairs of time points with large mean differences and hence large *t* values should be in the front part of the ordered sequence of pairs.

Again, if the variances or correlations in model (1) are unequal, then the procedure keeps the experimentwise error rate á nonetheless but the power of the tests may be insufficient.

References

Hsu, J.C. (1996). Multiple comparisons – theory and methods. Chapman & Hall. London.

Kropf, S. (2000). Hochdimensionale multivariate Verfahren in der medizinischen Statistik. Shaker Verlag. Aachen.

Maurer, W.; Hothorn, L.A. and Lehmacher, A.E. (1995). Multiple comparisons in drug clinical trials and preclinical assays. In *Biometrie in der chemisch-pharmazeutischen Industrie*. **6**. (ed J. Volmar), 3-18. Gustav Fischer Verlag. Stuttgart Jena New York.

Schuster, E. (2000). Markov Chain Monte Carlo Methods for Handling Missing Covariates in Longitudinal Mixed Models. In *Proceedings in Computational Statistics 2000* (eds J. Bethlehem and P. van der Heijden), 439-444. Physica Verlag. Heidelberg.

Timm, N.H.(1980). Multivariate Analysis of Variance of Repeated Measurements. In: *Krishnaiah, P.R.* (ed.). Handbook of Statistics. Volume 1 - Analysis of Variance. Amsterdam: North-Holland, 41-87.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

A Linear Model for Bridge Scores

Paul Seeger, John Öhrvik
Swedish University of Agricultural Sciences, Department of Biometry and Informatics
P.O. Box 7013, SE-750 07 Uppsala, Sweden
john.ohrvik@bi.slu.se

1. Introduction

In bridge competitions for pairs two pairs meet at a table and play a number of deals. These deals are kept in a board and played also by other pairs at other tables. This is repeated for a number of different boards. For each deal and table a score, measuring the difference in skill for the two pairs, NS and EW, is calculated.

Statistically this set up can be seen as a design with the pairs as 'treatments' and with a 'block structure' containing the factor boards, deals within boards and tables within boards. In a round the tables may share the *n* deals from the same board or different boards may be played at some or all of the tables. If there are $p=2t$ or $p=2t+1$ pairs, *t* tables will be used for each round and usually also for each of *b* boards. The scores y may be modeled as a linear mixed model

(1) $$\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{u},$$

where **y** is a columnvector with *btn* rows. The design matrix **X** has *btn* rows and *p* columns and specifies how each pair contributes to the score and the columnvector **a** contains the *p* unknown skill-scores that we wish to estimate. There is also a vector **u** with *btn* residuals which has mean $E(\mathbf{u}) = \mathbf{0}$ and variance $\mathrm{Var}(\mathbf{u}) = \mathbf{V}$. Thus
$E(\mathbf{y}) = \mathbf{X}\mathbf{a}$ and $\mathrm{Var}(\mathbf{y}) = \mathbf{V}$.

2. Model

The vector of residuals, **u**, includes possible effects of boards, deals within boards and tables within boards. Thus **V** contains variances and covariances referring to boards, deals and tables. The model (1) can in principle be written as

(2) $$\mathbf{y}=\mathbf{X}\mathbf{a}+\mathbf{Z}_b\mathbf{Bo}+\mathbf{Z}_t\mathbf{Ta}+\mathbf{Z}_d\mathbf{D}+\mathbf{e},$$

where **Bo** stands for board effects, **Ta** for table effects within boards, **D** for deal effects within boards and **e** for uncorrelated residuals (including an interaction between deals and tables). In order to allow negative covariances we introduce the following variance matrix:

(3) $$\mathbf{V} = \sigma^2\rho_b\mathbf{Z}_b\mathbf{Z}'_b + \sigma^2\rho_t\mathbf{Z}_t\mathbf{Z}'_t + \sigma^2\rho_d\mathbf{Z}_d\mathbf{Z}'_d + \sigma^2\rho\mathbf{I},$$

where $\sigma^2\rho_d$, $\sigma^2\rho_t$ and $\sigma^2\rho_b$ denote the covariances between scores within deal, within table within board and within board and $\rho = 1-\rho_b- \rho_t -\rho_d$. If we have a proper game, i.e. if there really are *b* boards and *n* deals per board without any accidental rotations $\mathbf{Z}_d\mathbf{Z}'_d=\mathbf{I}_b*\mathbf{J}_t*\mathbf{I}_n$, $\mathbf{Z}_t\mathbf{Z}'_t=\mathbf{I}_b*\mathbf{I}_t*\mathbf{J}_n$ and $\mathbf{Z}_b\mathbf{Z}'_b = \mathbf{I}_b*\mathbf{J}_t*\mathbf{J}_n$. $\mathbf{I}_b$ is a *b* by *b* identity matrix, $\mathbf{J}_t$ a *t* by *t* matrix with 1 in every position and * denotes the Kronecker product.

If we use raw scores or some direct transformation of them the fixed part of the score, **Xa**, can be said to depend on the difference between skills for the NS-pair and the EW-pair at the table corresponding to the row in question. In a row of **X** there is a 1 for the NS-pair at the table, -1 for the EW-pair and 0 for the other *p*-2 pairs.

3. Analysis

Using generalized least squares, gLS, we get the normal equations $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{a}^0 = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, where $\mathbf{a}^0$ is a solution to the equations. In a proper game, i.e. with *n* deals and *t* tables for each of *b* boards $\mathbf{V}^{-1}$ can be expressed simply as

(4) $$\sigma^2\rho\mathbf{V}^{-1} = \mathbf{I} - k_2\mathbf{Z}_b\mathbf{Z}'_b - k_3\mathbf{Z}_t\mathbf{Z}'_t - k_4\mathbf{Z}_d\mathbf{Z}'_d$$

where k_i are functions of the correlations that can be solved from the equations $\rho_b = k_2(tn\rho_b + n\rho_t + t\rho_d + 1) + k_3(n\rho_b + \rho_d) + k_4(t\rho_b + \rho_t)$, $\rho_t = k_3(\rho + n\rho_t)$ and $\rho_d = k_4(\rho + t\rho_d)$. Then we can write the normal equations as $\mathbf{A}\mathbf{a}^0 = \mathbf{B}\mathbf{y}$ with $\mathbf{B} = \mathbf{X}' - k_2\mathbf{X}'\mathbf{Z}_b\mathbf{Z}'_b - k_3\mathbf{X}'\mathbf{Z}_t\mathbf{Z}'_t - k_4\mathbf{X}'\mathbf{Z}_d\mathbf{Z}'_d$ and $\mathbf{A} = \mathbf{B}\mathbf{X}$.

With *n* deals per table and board we can write $\mathbf{X} = \mathbf{X}_1 * \mathbf{1}_n$ where $\mathbf{X}_1$ is the design matrix for *n*=1 and $\mathbf{1}_n$ a column vector of 1s. Then $\mathbf{X}'\mathbf{Z}_b\mathbf{Z}'_b = n\mathbf{X}'\mathbf{Z}_d\mathbf{Z}'_d$ and $\mathbf{X}'\mathbf{Z}_t\mathbf{Z}'_t = n\mathbf{X}'$. This means that the normal equations can be written with $\mathbf{B} = \mathbf{X}' - K\mathbf{X}'\mathbf{Z}_d\mathbf{Z}'_d$, where $K = (nk_2+k_4)/(1-nk_3)$. Replacing the k_i by their respective functions of the correlations it takes some straightforward algebra to show that $K = \gamma/(1+\gamma t)$, where $\gamma = (\rho_d+n\rho_b)/(\rho+n\rho_t)$. If γ were known a solution could be obtained as $\mathbf{a}^0 = \mathbf{A}^-\mathbf{B}\mathbf{y}$ where $\mathbf{A}^-$ is a generalized inverse of $\mathbf{A}$. The variance matrix would then be $\mathrm{Var}(\mathbf{a}^0) = \sigma^2(\rho+n\rho_t)\mathbf{A}^-$.

The constant γ is unknown and we have to estimate or guess it. With a positive ρ_d and with ρ_b and ρ_t not far from zero, γ could be large or even unlimited. This case corresponds to K=1/t and thus we obtain the oLS-estimator as a limit of the gLS-estimator as we are used to.

Although the scores are hardly normally distributed, especially not for small *p*, we have used a normal likelihood as an approximation. To get REML estimates of σ^2, ρ_b, ρ_d and ρ_t, we have written a program in Matlab. In fact, for proper games we get the same results for contrasts $\mathbf{c}'\mathbf{a}^0$ and its variance by using least squares and the anova method which does not use the normality assumption.

The vector $\mathbf{a}^0 = \mathbf{A}^-\mathbf{B}\mathbf{y}$ is not really an estimator of $\mathbf{a}$, which is not estimable. However certain contrasts in $\mathbf{a}$ are, and if all pairwise contrasts $\alpha_i - \alpha_j$ are estimable we can use $\mathbf{a}^0$ to rank the pairs. To see the estimability, we can check if $\mathbf{C} = \mathbf{C}\mathbf{A}^-\mathbf{A}$, where $\mathbf{C}$ is the matrix of contrasts.

The matrix $\mathbf{A}$ in the normal equations is called information matrix and is especially easy to use when it is on the form $c\mathbf{I}_p - d\mathbf{J}_p$. Then the generalized inverse $\mathbf{A}^-$ and $\mathrm{Var}(\mathbf{a}^0)$ are on the same form and are easily expressed with help of *c* and *d*. Such designs are said to be balanced. In this case it is easy to derive formulas for pairwise variances, PV, in gLS (true γ known) and oLS (K=1/t). If a guessed $K_g = \gamma_g/(1+t\gamma_g)$ is used ,ggLS, some algebra is needed to find PV. The results are for $PV/(\sigma^2(\rho+n\rho_t))$:

gLS $(2(p-1)(1+\gamma t))/(pr(1+\gamma(t-1)))$
ggLS $(2(p-1)(1+\gamma+\gamma_g(t-1)(2+\gamma_g t)))/(pr(1+\gamma_g(t-1))^2)$
oLS $(2(p-1)t)/(pr(t-1))$,

where *r* is the number of deals played by each pair. To illustrate the possible efficiencies of oLS and ggLS we consider the case *t*=3 and the true γ=1.5 see Table 1.

γ_g	-1	0	.5	1	1.5	2	2.5	3	5	10	∞
	4.5	2.5	1.5	1.39	1.375	1.38	1.39	1.40	1.43	1.46	1.50
			(=oLS)		(=gLS)						

Table 1. $prPV/(2(p-1)\sigma^2(r+nr_t))$ for different guessed values of γ

Apparently ggLS performs better than oLS with a large guess of γ.

4. Discussion

Applying a statistical linear model, generally used for block experiments, is useful for understanding the computations in bridge competitions. The method of least squares can be used to estimate differences in skill between the pairs.

If we are interested in the statistical error margins in the estimates and to reduce these errors by using gLS the detailed specification of the linear model becomes more important.The model used here contains a random part and a fixed part. Both positive and negative covariances are allowed in the random part. The pairs-design-matrix is defined in such a way that the factor pair is orthogonal to boards and deals. The generalized least squares estimates of pair differences are then independent of the random part of the model. Thus there is no difference between gLS- and oLS-estimators, the information matrix is simply $\mathbf{X}'\mathbf{X}$.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Quasi Copulæ

Carlo Sempi
Università di Lecce, Dipartimento di Matematica "Ennio De Giorgi"
Lecce, Italy
sempi@ilenic.unile.it

1. Introduction

Quasi-copulæ were introduced in Alsina *et al.* (1993) in order to characterize, in a class of operations on distribution functions, those ones that derive from corresponding operations on random variables defined on the same probability space. The concpet of quasi-copula turned out to be useful, but the original definition was impractical.

By a *track* B it is meant a subset of the unit square $[0,1]\times[0,1]$ that can be written in the form $B=\{(F(t),G(t)): t\in[0,1]\}$, for some continuous distribution functions F and G such that $F(0)=G(0)=0$ and $F(1)=G(1)=1$. A quasi-copula is a function Q from $[0,1]\times[0,1]$ to $[0,1]$ such that for every track B there exists a copula C_B that coincides with Q on B, namely

$$Q(x,y)= C_B(x,y), \quad (x,y)\in B.$$

This definition makes it very hard to recognize wether a given function Q from $[0,1]\times[0,1]$ to $[0,1]$ is in fact a quasi-copula.

2. Characterization of Bivariate Quasi-Copulas

The following characterization was proved in Genest *et al*. (1999)

Theorem 1 A function Q from $[0,1]\times[0,1]$ to $[0,1]$ is a quasi-copula if, and only if, it satisfies the following conditions:

(1) $Q(0,x)=Q(x,0)=0$ and $Q(x,1)=Q(1,x)=x$ for all $x\in[0,1]$;

(2) $Q(x,y)$ is nondecreasing in each of its arguments;

(3) Q satisfies a Lipschitz condition, that is, for all x,x',y,y' in $[0,1]$,

$$\le |x' - x| + |y' - y|.$$

The proof is based on an existence argument and relies on the following crucial

Lemma 1 Let $(x_1,y_1),\ldots,(x_n,y_n)$ be distinct points in $[0,1]\times[0,1]$ with $0\le x_1 \le\ldots\le x_n \le 1$ and $0\le y_1 \le\ldots\le y_n\le 1$. Let also $q_1,\ldots,q_n$ be reals with $0\le q_1 \le\ldots\le q_n\le 1$ and suppose that

(a) $0 \le q_{i+1} - q_i \le (x_{i+1} - x_i) + (y_{i+1} - y_i)$ $\quad 1 \le i \le n-1$;
(b) $\max\{0, x_i + y_i - 1\} \le q \le \min\{x_i, y_i\}$, $\quad 1 \le i \le n$.

Then there exists a copula C such that $C(x_i,y_i)= q_i$ for every $i=1,\ldots,n$.

It will be recalled that a copula satisfies, along with the boundary conditions (1), also the condition, which is stronger than (2),

(4) $$Q(x',y') - Q(x',y) - Q(x,y') + Q(x,y) \ge 0,$$

for all x,x',y,y' in [0,1] such that $x \le x'$ and $y \le y'$. Then a second characterization of a quasi copula is provided by the following

Theorem 2 A function Q:[0,1] × [0,1] → [0,1] is a quasi-copula if, and only if, it satisfies condition (1) and inequality (4) holds true whenever at least one of x, x', y,

By means of the above characterizations it is possible to construct absolutely continuous quasi-copulas.

The results of Theorems 1 and 3 for two-dimensional quasi-copulas can be extended to the n-dimensional case. However the proofs turn out to be more delicate than in the two-dimensional case (see Genest *et al*. (2001)). For n 3, an n-copula C satisfies the analogue of inequality (4). For an n-box

$$[\mathbf{x},\mathbf{y}]:=[x_1, y_1] \times [x_2,y_2] \times \dots] \times [x_n,y_n],$$

let $\mathbf{v} = (v_1,v_2,\dots,v_n)$ with $v_i \in \{x_i,y_i\}$ (i=1,2,…,n) denote a vertex of [**x**,**y**] and let s(**v**) be equal to 1 or to –1 if $v_i = x_i$ for an even (respectively, odd) number of i's. Then the C-volume of the box [**x**,**y**] is defined by

$$V_C([\mathbf{x},\mathbf{y}]) := \Sigma\, s(\mathbf{v})\, C(\mathbf{v}),$$

where the summation is over all the vertices **v** of the box. By definition, an n-copula C satisfies the inequality

(5) $$V_C([\mathbf{x},\mathbf{y}]) \ge 0,$$

for every box [**x**,**y**].

Theorem 3 A function Q: $[0,1]^n \to [0,1]$ is an n-quasi-copula if, and only if, it satisfies the following conditions:

(6) $Q(x_1,\dots,x_{i-1},0,x_{i+1},\dots,x_n) = 0$ and $Q(1,\dots,1,x_i,1,\dots,1) = x_i$ for all $x_i \in [0,1]$ (i=1,2,…,n).

(7) Q is nondecreasing in each variable;

(8) Q satisifes the Lipschitz condition

$$|Q(x_1,x_2,\dots,x_n) - Q(y_1,y_2,\dots,y_n)| \le \Sigma_{i=1,2\dots,n} |x_i - y_i|$$

for all$(x_1,x_2,\dots x_n)$ and $(y_1,y_2,\dots,y_n)$ in $[0,1]^n$.

A characterization analogous to that provided in Theorem 2 holds.

Theorem 4 A function Q: $[0,1]^n \to [0,1]$ is an n-quasi-copula if, and only if, it satisfies condition (6), condition (5) for every n-box [**x**,**y**] such that all the components x_i of **x,** but for at most one of them, are equal to zero, and moreover the inequality

$$Q(x_1,\dots, x_{i-1}, x_i, x_{i+1},\dots,x_n) - Q(y_1,\dots,y_{i-1},y_i,y_{i+1},\dots,y_n) \le y_i - x_i$$

for all x_i and y_i in [0,1] with $x_i \le y_i$ (i=1,2,…,n).

References

Alsina, C., Nelsen, R.B. and Schweizer, B. (1993), On the characterization of a class of binary operations on distribution functions, *Statist. Probab Lett.* **17**, 85-89.

Genest, C., Quesada Molina, J.J., Rodríguez Lallena, J.A. and Sempi, C.(1999), A characteriaztion of quasi-copulas, *J. Multivariate Anal.* **69**, 193-205.

Genest, C., Quesada Molina, J.J., Rodríguez Lallena, J.A. and Sempi, C. (2001), Multivariate quasi-copulas, to appear.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Entropy and the Portfolio Management: An Application to the Portuguese Stock Market

Amílcar Serrão, Andreia Dionísio
Evora University, Management Department
Largo dos Colegiais, P7000-554 Évora, Portugal
aserrao@uevora.pt

1. Introduction

This research work studies the entropy and the mutual information as uncertainty measures applied to the Portuguese stock market. The utilization of variance as measure of uncertainty is well entrenched by tradition in statistics. The mean-variance approach is appropriate only for distributions that are unimodal and symmetrical, since the first two moments of a population can be estimated from the respective moments of the sample distributions. When there are distributions that are non-symmetric, a different measure of uncertainty is required. This measure of uncertainty must be more dynamic and general than variance. This study proposes another measure of uncertainty called entropy or expected information to make a portfolio management. It is intended with this study to verify the entropy as uncertainty measure is adapted to the financial theory, more properly to the portfolio theory.

2. Mean-Variance Model versus Mean-Entropy Model

The efficiency frontiers obtained by the mean-variance model and the mean-entropy model is compared, where entropy is calculated from the stock portfolio selected by the mean-variance model. As it is unknown the true distribution of probability followed by the earnings yields of the stocks, the calculation of the conditional probabilities becomes an uncertainty source and to overcome such difficulty, the portfolio return is calculated from their stocks. If the mean-variance model for a specific coefficient risk aversion (K) is run a couple of times, it is possible to generate a return rate for the different selected portfolios and then to compute a relative frequency curve. The relative frequency is used here as a close approximation of the probability of occurrence of the return rate of each one of the portfolios. The entropy of the portfolio is calculated as follows:

$$H_p = -\sum_{p=1}^{n} p_p \, \mathrm{Log}_2 \, p_p \qquad (1)$$

Where: H_p - Entropy of the portfolio; and, p_p - Probability of occurrence of the return rate of each one of the portfolio selected by Mean-Variance model.

K	**Mean (%)**	Variance	**Entropy (bits)**
0.000	0.2700	5.8515	3.0291
0.005	0.2647	4.2798	2.8469
0.008	0.2627	3.9523	2.8533
0.010	0.2620	3.8767	2.8482
0.050	0.1866	1.7263	2.3343
0.100	0.1003	0.4315	1.5639
0.200	0.0572	0.1078	1.0996
1.000	0.0226	0.0043	0.9226

Table 1. Mean, variance and entropy for the selection portfolios

For the analysis of the Table 1 an attendance of the entropy is verified relatively to the variance, that is, as the variance decreases due to the diversification effect, the entropy also tends to decrease, in spite of form no so significant, being revealed sensitive to the diversification effect. Such tendency is verified in all of the portfolios except in the third where the variance decreases and the entropy increases.

3. Conclusions

The Portuguese stock market has a small dimension and a weak liquidity that move away it of the applicability of the traditional models of portfolio selection and management, where risk is simply measured by the variance. Through statistical analysis realized to all stocks that constitute the sample and to the BVL 30 index, it was verified that the normal probability distribution doesn't represent the empirical data faithfully, what at once puts in cause the application of the mean-variance model as an efficient model of portfolio selection.

In order to verify if the selection made by the empirical model provides all information to the investor about the true uncertainty, it was calculated the mean-entropy for each one of the optimal portfolios. The entropy is independent of the type of probability distribution, measuring the global dispersion unlike the variance that is limited to measure the dispersion around the mean. So the entropy provides correct information about dispersion of each one of the selected portfolios. Model results allowed to verify that the variance and the entropy goes in the same way, however the entropy doesn't represent the traditional curve of the efficiency frontier, being denoted the existence of a bias in the measure of the risk and the uncertainty. Such bias will only be able to be due to the fact of the empirical distribution for the data not to be normal, what at once demonstrates the inadequacy of the mean and of the variance as the only measures used to characterize a distribution. The entropy appears then as a new form of measuring the uncertainty for any type of probability distribution, constituting a source of privileged information for the investor. As the entropy is calculated through a logarithm, the events with smaller occurrence probability are more valued that in the calculation of the variance, what constitutes an advantage for the entropy, because more importance is given to the possibility of occurrence of rare events, namely “crash's”. So the “mean-entropy frontier” is more robust of the statistical point of view and it promotes more trustworthy information for the investor. It was also verified, for the analysis of model results, that the entropy is sensitive to the diversification effect, what at once facilitates the acceptance as a global dispersion measure in the portfolio theory.

References

AUSLOOS, M. (1998). The Money Games Physicists Play, *Euro physics News*, March/April.

DACOROGNA, M. (1999). Econophysics Find to Forum; *Physics World*, Sep.

EBRAHIMI, N.; MAASOUMI, E. AND SOOFI, E. S. (1999). Ordering Univariate Distributions by Entropy and Variance, *Journal of Econometrics*, **90**, 2, 317-336.

MARKOWITZ, H. M. (1959). Portfolio Selection: Efficient Diversification of Investments, John Wiley & Sounds, New York.

PHILIPPATOS, G. C. AND WILSON, C. (1972). Entropy, Market Risk, and the Selection of Efficient Portfolios, *Applied Economics*, **4**, 209-220.

PHILIPPATOS, G. C. AND NAWROCKI, D. N. (1973). The Information Inaccuracy of Stock Market Forecasts: it Adds New evidence of Dependence on the N.Y.S.E., *The Journal of Financial and Quantitative Analysis.*

PHILIPPATOS, G. C. AND WILSON, C. (1974). Entropy, Market Risk, and the Selection of Efficient Portfolios: Reply, *Applied Economics*, **6**, 77-81.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

New Method of Linear Discriminant Function Using Integer Programming (IP-OLDF)

Shuichi Shinmura
Seikei Univ., Dept. of Economics
Kichijoji Kitamachi 3-3-1, Musashino, Tokyo 180-8633 Japan
shinmura@econ.seikei.ac.jp

1. Introduction

In this paper, I introduce two optimal linear discriminant functions (OLDF) using integer programming (IP) and linear programming (LP). Those are called as IP-OLDF and LP-OLDF. In order to evaluate these new methods with the Fisher's linear discriminant function (Fisher's method) and the quadratic discriminant function (Quadratic's methods), I applied these ones to three data sets such as - the iris data, a medical data, and 115 data sets of random number data.

2. Algorithm of OLDF

Miyake & Shinmura (1976) proposed a new criterion of the linear discriminant function, which minimises the sample miss-classify rate (error rate). And the heuristic algorithm was proposed. This one was applied for the above medical data, but only 6-variables model could be solved because of the restriction of CPU time. In this paper, I propose a new algorithm using IP and LP for OLDF.

LP-OLDF minimises the summation of distances of miss-classify samples from critical point. On the other hand, IP-OLDF minimises the sample error rate directly.

3. Data and Results

These new methods were applied to three kinds of data in order to show its usefulness by comparing these methods with Fisher's and Quadratic methods.
The analysis was done by the following procedure. First, all possible models were computed as a frame of analysis. And the forward and the backward basic sequences were derived from these models. Next, the error rates were computed on these basic sequences for the four discriminant functions. And lastly, these values denoted by (IP, LP, FP, QP) were evaluated by various statistical methods such as the t-tests for the differences in the averages and the regression analyses.

3.1 Iris Data

1. Error Rate

The error rates by IP-OLDF (IP), LP method (LP), Fisher's method (FP) and Quadratic method (QP) are obtained on 15 models that are composed of all combinations of four independent variables.

2. T-test for the differences in the means

The order of the means of 4 error rates is IP (8.933) < LP (9.733) < FP (10.667) < QP (11.267). T-tests of IP with (LP, FP, QP) are rejected. T-test of LP with QP is rejected too. In any case, IP-OLDF is superior to other methods.

3. Examination of regression lines and discriminant coefficients

In this paper, we propose a new idea to evaluate 4 discriminant methods by IP-OLDF. The error rates of FP, QP and LP are predicted by that of IP-OLDF. The

regression results are as follows; FP = 1.582 + 1.003 * IP (r=0.989), QP = 1.617 + 1.064 * IP (r=0.987), LP = 0.934 + 0.971 * IP (r=0.996).

The order of predicted values is IP < LP < FP < QP. So, this result is as same as t-test although the differences of the four methods are very small for small values of IP and large for larger values of IP.

3.2 Medical Data

Important results were obtained in the analysis of medical data, and they will be reported at the conference.

3.3 Random Number Data

The following results are obtained about the random number data. This data is designed to evaluate the influence about rotation and transportation, and the relation between the internal and external samples. The error rates for 115 data sets were computed. The range of those is [0,71]. Effects by the rotation are observed in the cases such as 'D10Ak' in which two groups are very close or sample error rates are high.

The order of the averages for the internal sample is IP < QP < FP <LP. All t-tests are rejected. The order of the averages for the external sample is QP < IP < FP <LP. T-tests are rejected except for IP and FP. QP is better than IP in the external sample whereas IP is better than QP in the internal sample.

(IP, LP, FP, QP) were regressed by IP. The order of predicted values of (IP, LP, FP, QP) are as same as the average.

4. Conclusion

To summarize above result, IP-OLDF is the best method for three kinds of data, but Quadratic method is the best for the external check of the random number data.

In the future, we intend to design a random number data with more than three variables and evaluate IP-OLDF on various points.

References

Miyake A. & Shinmura S. (1976). Miss-classify rate of linear discriminant function, (eds F.T. de Dombal & F.Gremy) 435-445, North- Holland Publishing Company.

Shinmura, S. (1998). Optimal Linear Discriminant Functions using Mathematical Programming, *Bulletin of The Computational Statistics of Japan* **11(2)**, 93-105.

Shinmura S. (1999). Optimal Linear Discriminant Function (OLDF) using Mathematical Programming, *Bulletin of ISI99 52nd Session Contributed Papers Book3*, 247-248.

REVISTA DE
ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Optimising Monte Carlo Algorithms for Option Pricing

Dmitrii Silvestrov, Anatoliy Malyarenko, Evelina Silvestrova
Mälardalen University, Department of Mathematics and Physics

dmitrii.silvestrov@mdh.se

Alexander Kukush
Kiev University, Department of Mathematical Analysis
Kiev, Ukraine
kuog@mechmat.univ.kiev.ua

Viktor Galochkin
Center of Practical Informatics, Academy of Sciences
Kiev, Ukraine
gal@nas.gov.ua

Monte Carlo algorithms for optimal pricing of American type options are described. These algorithms are based on new theoretical results, which show that optimal stopping strategies for American options have a threshold structure for general dynamical models of pricing processes and convex pay-off functions. The results of theoretical and experimental studies show that the direct simulation approach has advantages with respect to traditional numerical methods. It is much more flexible and less sensitive to the modifications of models of underlying pricing processes, pay-off functions and other characteristics of the models. The optimising Monte Carlo pricing algorithms give also opportunity to estimate some important statistical characteristics that are hardly available in the case of the use of traditional numerical methods. For example, profit histograms and dynamical quintile diagrams for optimal stopping strategies, confidence intervals, etc. can be effectively estimated.

The PC based programs have been elaborating at present time. They show promising results by accuracy, computing time and other characteristics. Results of computer experiments are displayed and discussed as well as prospective of application of Monte Carlo methods to problems of option pricing.

$_{å}$ n}·
$A_{å}^{n}$ Ð$_{å}$ as n . It is further assumed that matrix $A_{å}$ is a nonlinear perturbation of matrix A_0 in the sence that $A_{å} = A_0 + A_{[1]}å + \ldots + A_{[k]}å^k + o(å^k)$. The explicit algorithm $_{å}$ $_0$+ ñ$_{[1]}$ å +… +ñ$_{[k]}$ $å^k$ + o($å^k$) and $Ð_{å} = Ð_0 + Ð_{[1]}å + \ldots + Ð_{[k]}å^k + o(å^k)$. Applications to Markov chains with absorption and asymptotical expansions in mixed large deviation and ergodic theorems for lifetime functionals and quasi-stationary distributions for models of population dynamics, queueing systems and risk processes are discussed.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Space-Time Modelling of Visual Field Data

Amelia Simó-Vidal, Mª Victoria Ibañez-Gual
Universitat Jaume I, Department of Mathematics
Campus del Riu Sec, Castellon, Spain
simo@mat.uji.es

1. Introduction

The Glaucoma is a very serious and extended illness. It may in time result in damage to the optic nerve, loss of peripheral vision and finally in blindness. Patients are usually unaware of peripheral vision loss and they may remain undiagnosed until central vision is severely affected. The most extended way to detect Glaucoma and to assess the extent of visual field loss is to perform a visual field test, a set of fixed locations in the visual field is chosen and they are randomly exposed to light stimulus with different intensities. When the patient perceives a stimulus, he pushes a button and his response is saved. There are different modalities of perimetric test. Our data set is the output from an Automated Static Perimetry Test. The output consists in a map with *n* numerical values. Each value represents the brightness intensity perceived in each point. The device used has been a Humphrey Field Analyzer.

Up to now the study of Glaucoma has been restricted to the study and modelling of only one visual field. In this paper a model is proposed for the spatio-temporal distribution of visual fields, the ML parameter estimators are obtained and the goodness of fit is checked.

2. Notation and Data Description

Let *P* be the number of patients, N_p the number of visual field test from patient number *p* and *n* the grid size. We define two variables on each visual field position. The first variable $S \in \{1, -1\}$ is called defect status. This variable allows each visual field point to be classified as "normal" or "disease". The second variable Z gives the threshold intensity. We denote $S_{tp} = (s_{1tp}, s_{2tp}, \ldots, s_{ntp})$ and $Z_{tp} = (z_{1tp}, z_{2tp}, \ldots, z_{ntp})$, with s_{itp} and z_{itp} the observed value of the defect status and threshold intensity respectively at site *i*, time *t* and patient *p*.

In the literature we can find a set of assumptions about visual field distribution. A previous descriptive analysis of data has been done in order to check them. They are:

- The evolution of the disease is highly correlated with ocular nerve fibre directions.
- The threshold values decrease with age linearly.
- The rate of loss of sensitivity with age is higher in the periphery than in the middle of the visual field.
- The variability of the thresholds increases with distance from the fixation point.
- After a Box-Cox transformation the threshold distribution is Gaussian.

These assumptions will be incorporated in the models in sections 3 and 4.

3. A Constrained Autologistic Model for Modelling Defect Status

Olsson and Rootzen (Olsson et al., 1994) proposed an autologistic model for the defect status vector at time t. In order to introduce the temporal dependence we take into account that if a point is defective at time *t* this status should remain unchanged for *s>t*.

We define $C(S)=\sum_{p}\sum_{i}\sum_{t=2}^{N_p}(1-s_{itp})(1+s_{it-1p})$ and $\Omega_c=\{S:C(S)=0\}$. We propose the following model for the joint defect status distribution:

$$f_c(S)=1_{\Omega_c}(S)\frac{\exp\{\sum_{p=1}^{P}\sum_{t=1}^{N_p}\sum_{i=1}^{n}\sum_{j=1}^{n}b_{ij}s_{jtp}s_{itp}\}}{\sum_{U\in\Omega_c}\exp\{\sum_{p=1}^{P}\sum_{t=1}^{N_p}\sum_{i=1}^{n}\sum_{j=1}^{n}b_{ij}u_{jtp}u_{itp}\}}$$

with $b_{ij}=\frac{b}{r_{ij}}$ if $i\in\partial i$, 0 otherwise; di is the first order neighbourhood of i, $r_{ij}=\sqrt{d_{ij}^{\parallel}+kd_{ij}^{\perp}}$; being $d_{ij}^{\parallel},d_{ij}^{\perp}$ the decomposition of the Euclidean distance into a part parallel to the retina nerve fibre trajectory and a part perpendicular to it.

4. Threshold Modelling

A VARI (1,1) model is assumed, $\Delta Z_{tp}=C_0+\Delta S_{tp}c_1+\Psi\Delta Z_{t-1p}+e_{tp}$, being Δ the first difference operator; $e_{tp}/\Delta Z_{t-1p},S_p\sim N_n(0,M(I-\Gamma)^{-1})$ with e_{tp} independent of e_{tq} if $p\neq q$, $M=diag(s(1),\ldots,s(n))$, $s(i)$ denotes a function of the distance between *i* and the fixation point, $(I-\Gamma)^{-1}$ is a symmetric and positive defined matrix, $\Gamma=(g_{ij})_{i,j=1,\ldots,n}$ with $g_{ij}=g$ if $j\in d_2i$, 0 otherwise and d_2i is the second order neighbourhood of site *i*; $\Psi=(y_{ij})_{i,j=1,\ldots,n}$ with $y_{ij}=y$ if $j\in d_2i$, $\tilde{y}$ if $i=j$, 0 otherwise; $C_0=(c_{01},\ldots,c_{0n})$; finally, c_1 is the parameter that incorporates the change in the mean due to a change in the defect status variable.

5. Parameter Estimation: Stochastic EM Algorithm

The defect status data are missing and the normalisation constant of the probability distribution of defect status vector is unknown. So we must use the stochastic EM algorithm (Celeux and Diebolt, 1985) in order to obtain the Maximum Conditional Likelihood estimators of the parameters. The simulation of the defect status vector given the thresholds will be obtained using the Gibbs Sampler.

6. Model Checking

Models in section 3 and 4 were chosen on the basis of medical knowledge, some statistical tools have been used for investigating the goodness of the fit. Two hypotheses have been checked: the order of the VAR model by means of the Portmanteau test and the whiteness of the residuals through the quantile-quantile and autocorrelations plots of the residuals.

References

Olsson J, and Rootzen H. (1994). An Image Model for Quantal Response Analysis in Perimetry, *Scand J Stat* **21**:357-387

Lütkepohl H. (1993). Introduction to Multiple Time Series Analysis. Ed. Springer-Verlag.

Celeux G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comp. Statist. Quart.*, **2**:73-82

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Different Optimality Criteria of Surveillance and their Implications

Christian Sonesson
Göteborg University, Department of Statistics
Box 660, SE 405 30 Göteborg, Sweden
Christian.Sonesson@statistics.gu.se

Marianne Frisén
Göteborg University, Department of Statistics
Box 660, SE 405 30 Göteborg, Sweden
Marianne.Frisen@statistics.gu.se

When constructing a method of surveillance, optimality criteria are of interest. The most commonly used is the minimal ARL^1 (the minimal average run length when the process is out-of-control when the surveillance starts) for a fixed ARL^0 (average in-control run length). However, as a formal optimality criterion, the ARL-criterion has severe drawbacks. Degenerated methods, which cannot be recommended in practice are ARL-optimal. The other criterion studied is the minimal expected delay (from a change to the detection) for a fixed probability of a false alarm. This criterion is appropriate for most applications, since it takes into account also the possibilities of later changes. This is important since the ability of detection depends on the time-point of the change. Special attention is given to the EWMA method and different suggested variants. In this case important differences between one- and two-sided versions exists.

References

Frisén, M. (2000). Characterization of methods for surveillance by optimality. *Research Report 2000*:**13**, Department of Statistics, Göteborg University.

Frisén, M. and de Maré, J. (1991). Optimal surveillance, *Biometrika.***78**, 271-80.

Frisén, M. and Sonesson, C. (2000). Optimal surveillance with EWMA. *Research Report 2000*:**7**. Department of Statistics, Göteborg University.

Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of Probability and its Applications*. **8**, 22-46.

Sonesson, C. (2001) Evaluations of different exponentially moving average methods, Manuscript.

Small Noise Asymptotics and Option Pricing for Stochastic Volatility Models

Michael Sørensen
University of Copenhagen, Department of Statistics and Operations Analysis
Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark
michael@math.ku.dk

Nakahiro Yoshida
University of Tokyo, Graduate School of Mathematical Sciences
3-8-1 Komaba, Meguro-ku, Tokyo 153, Japan
nakahiro@ms.u-tokyo.ac.jp

An asymptotic expansion of the expectation of an irregular functional of a diffusion process that is perturbed around a random limit is derived by means of Malliavin calculus. We consider in detail the particular example given by a stochastic volatility model where the noise of the volatility process is small. In this case a third order expansion is presented. Also models with a random discounting factor are considered. A typical application is to the pricing of options, where we study European options in particular. A result of some financial interest is that the classical Black-Scholes formula for the price of a European option turns out to be correct to second order also for stochastic volatility models provided that there is no leverage effect, i.e. no correlation between the noise of the volatility process and the noise driving the price process.

References

Ikeda, N. and Watanabe, S. (1989). Stochastic Differential Equations and Diffusion Processes. North Holland, Amsterdam.

Kim, Y. and Kunitomo, N. (1999). Pricing options under stochastic interest rates: a new approach. *Asia-Pacific Financial Markets* **6**, 49-70.

Sørensen, M. and Yoshida, N. (2001). Small diffusion expansions for stochastic volatility modles. *Preprint*, Dept. of Theoretical Statistics, Univ. of Copenhagen.

Yoshida, N. (1992). Asymptotic expansions for statistics related to small diffusions. *J. Japan Statist. Soc.* **22**, 139 – 159.

Yoshida, N. (2000). Perturbation methods and option pricing. *Proceedings of an International Conference at the University of Tokyo*.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Robust Estimation for Polynomial Structural Relationship

Maria Manuela Souto de Miranda
Universidade de Aveiro, Departamento de Matemática
Campus de Santiago, 3810-193 Aveiro, Portugal
manuela.souto@mat.ua.pt

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

The structural relation model is a measurement error model that describes a relationship of dependence of the form $h = g(x; \boldsymbol{b})$ where both x and h are unobservable random variables. The model assumes that these variables are observed with additional additive errors, which makes it adequate for the modelling of many real situations in different scientific areas.

Assuming that g is a linear function of the variable x and of the vectorial parameter $\boldsymbol{b}$, there are many known methods to estimate $\boldsymbol{b}$ provided it is identifiable. But when g is a nonlinear function in at least one of the arguments, the estimation procedure has yet to be properly studied.

Herein it is admitted that the function g has a polynomial form. Following the conditional approach to the estimation problem suggested by Gleser (1990), the values of the dependent observable variable are expressed in terms of its regression on the other observable variable. When the parameter is identifiable this way of treating the problem can lead to a question of estimation in an heteroscedastic nonlinear regression model, as noted in Cheng and Van Ness (1999). The usual methods for the estimation of the parameters of the nonlinear regression model can then be applied.

However, the resulting estimators are very sensitive to violations of the assumptions of the model and it is convenient to look for robust estimation procedures. Bounded influence estimators for the polynomial structural relation model are developed from robust regression methods and their properties are studied.

References

Cheng, C.- L. and Van Ness, J.W. (1999). *Statistical Regression with Measurement Error*, [*Kendall's Library of Statistics*] **6**. Arnold, London.

Gleser, L.J. (1990). Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. In Statistical Analysis of Measurement Error Models and Applications. Ed. P. Brown and W. Fuller. *Contemporary Mathematics*, **112**, 99-114.

On the Application of Vandermonde Matrices to Time Series Analysis

Peter Spreij
Korteweg-de Vries Institute for Mathematics
Universiteit van Amsterdam, Plantage Muidergracht 24
1018 TV Amsterdam
spreij@science.uva.nl

In this paper we present a way to compute the Fisher information matrix of an ARMA process. The computation is based on the fact that this matrix satisfies a Stein equation.

The coefficients of the Stein equation under consideration turn out to be matrices in companion form and a basis of eigenvectors of companion matrices can be represented as the columns of a (confluent) Vandermonde matrix. Therefore, also from a statistical perspective there is an interest in analyzing (confluent) Vandermonde matrices. Solutions of this Stein equation are relatively easy to compute as soon as one knows how to invert a Vandermonde matrix (in the generic case where all zeros and poles of the transfer function have multiplicity one) or a confluent Vandermonde matrix (in the general case).

Therefore we present some general technical results on companion matrices and confluent Vandermonde matrices, the main results concerning inversion of Vandermonde matrices. Then we apply these results to describe solutions to Stein equations and investigate the special case where the solutions are given by blocks of the asymptotic Fisher matrix of an ARMA process.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

A New Local Influence Measure in Generalized Linear Models

M. Mercedes Suárez Rancel, Yenis Marisel González Mora
Univ. of La Laguna, Fac. of Mathematics, Dept. of Stat., Research Op. and Comp. Tenerife 38271., Spain.
MSUAREZ@ULL.ES, YMGMMAT@HOTMAIL.COM

This paper investigates the local influence assessment in generalized linear models. The concept of local influence was introduced by Cook (1986) and modified by Billor and Loynes (1993). Cook's local influence was motivated by the Cook measure (1977); so they only study the local influence on the regression coefficient and they are not resistant to masking and swamping effects. In this article we propose a new locally influential measure to mitigate these difficulties. We demonstrate the need to make it by giving examples where existing measures may fail to detect extrem points.

1. Introduction

We assume the observations consist of a vector y of n independent responses from the exponential family

$$f_y(y_i;q) = \exp\left\{\left[y_i q_i - b(q_i)\right]/a(f) + c(y_i,f)\right\}$$

with $q_i = g(h_i)$, $h_i = x_i b$, where x is an n × p matrix covariates, b is a p-dimensional column vector of unknown parameters, and $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ are known functions. The dispersion parameter f is usually regarded as a nuisance parameter. Let $\hat{b}$ be the maximum likelihood estimate (MLE) of b.

The idea of influence assessment is to monitor the sensitivity of statistical analysis when the subjected to minor changes in the model. For a review of GLM diagnostics, see e.g. Davison and Tsai (1992). Most attention in this area has in practice been focused on global influence mainly the case deletion method. The generalized Cook's distance (McCullagh and Nelder, 1989,p. 407) is defined by

$$D_i = \left(\hat{b}_{(i)} - \hat{b}\right)' (x'Wx)\left(\hat{b}_{(i)} - \hat{b}\right)/\hat{f}$$

where $W = diag\{W_i\}$, $W_i = b^{(2)}\left(\hat{q}_i\right)\left[g^{(1)}(\hat{h}_i)\right]^2$, $\hat{b}_i$ denotes the estimate of b without case i and the superscript (k) denotes the kth derivative of the function.

The purpose of this study is to gain additional insight regarding the local influence analysis and its implications on global influence.

2. A New Local Influence Measure in Generalized Linear Models

Cook (1986) developed a general technique for the assessment of local influence. Billor and Lyones (1993) show some practical and theoretical difficulties which arise in Cook's approach. To avoid these difficulties Billor and Loynes (1993) suggest, an alternative likelihood displacement:

$$LD^{*}(w) = -2\left[L(\hat{b}) - L(\hat{b}_{w} \mid w) \right] \ ,$$

where $\hat{b}_{w}$ is the MLE from the perturbed model and $L(\hat{b}_{w} \mid w)$ is the log-likelihood of the perturbed model, while Cook (1986), uses only the perturbation in the estimation of the parameters. Both proposes are not resistant to masking and swamping effects. So, to try to mitigate these effects we propose a new measure based on the following likelihood displacement:

$$LD_{(i)}(w_i) = -2\left[L(\hat{b}) - L_{(i)}(\hat{b}_{w_i} \mid w_i) \right] + \left[\text{var}\left(\hat{m}_i \right) - \text{var}\left(\hat{m}_{w_i} \right) \right]$$

where $\hat{m}_i = b^{(1)}\left(g\left(x_i \hat{b} \right) \right)$ and $Var\left(\hat{m}_i \right) = b^{(2)}\left(g\left(x_i \hat{b} \right) \right)$.

References

Billor, N. and Loynes, R. M. (1993). Local Influence: A New Approach .*Comm. Statist.-Theory Meth.*, **22**, 1595-1611.

Cook, R.D. (1977). "Detection of Influential Observations in Linear Regression". *Technometrics*, **19**, 15-18.

Cook, R.D. (1986). Assessment of Local Influence (with discussion) . *Journal of the Royal Statistical Society*, Ser. B., **48**, 133-169.

Davison, A.C. and Tsai, C.-L. (1992). "Regression Model Diagnostics". *International Statistical Review*, **60**, 3, 337-353.

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

Suárez Rancel, M. Mercedes, González Sierra, M. Angel. (1999) "Measures and Procedures for the Identification of Locally Influential Observations in Linear Regression". Communications in Statistics: *Theory and Methods*, Volume **28**, Issue 3.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

A Characterization of the Majorisation Ordering Applied to Aging Process

Alfonso Suarez
Universidad de Cádiz¸Departamento de Estadística e I. O.
Duque de Nájera, 8., 11002 Cádiz. Spain.
alfonso.suarez@uca.es

Jose María Fernández
Universidad de Sevilla, Departamento de Estadística e I. O.
Avda. Reina Mercedes., 41013 Sevilla. Spain.
ferpon@cica.es

Miguel A. Sordo
Universidad de Cádiz¸Departamento de Estadística e I. O.
mangel.sordo@uca.es

David Almorza
Universidad de Cádiz¸Departamento de Estadística e I. O.
david.almorza@uca.es

1. Introduction

Several approaches have taken in the literature to the ordering of probability distributions in terms of stochastic or variability properties. Lewis and Thompson (LT) (1981) studied the dispersion ordering. We say that X is less dispersed than Y in the LT sense, if any pair of quantiles of Y are at least more widely separated as corresponding quantiles of X. Shaked (1982) gave some characterizations of this partial ordering which can be used to a new interpretation of the dispersion ordering by the number of sign changes of the distribution functions. Muñoz-Pérez (1990) characterized the dispersion ordering by the concept of Q-addition of random variables and by the spread function under certain restriction on the respective quantile functions. Finally, Pellerey and Shaked (1997) characterized the IFR and DFR aging notions by means of the dispersive ordering.

Hickey (1986) studied a partial ordering in the majorisation sense, this ordering is called in randomness. Shaked and Shantikumar (1994) showed applications of the stochastic and variability orders in reability theory.

Suarez et al. (1998) studied the notion of comparing the probability in "Equally Lebesgue Measurable Intervals" to define a new dispersion ordering weaker than the dispersion ordering in the LT sense, they also obtained, under certain restrictions on the distribution functions, the equivalence between this weak dispersion ordering and the concept of majorisation.

2. Results

A lot of different stochastic orderings have been studied in the literature to characterize aging concepts using the stochastic process $\{X_t, t > 0\}$, where $X_t = \{ X-t \mid X>t \}$ is the additional residual life. We introduce a new aging concept for a lifetime distribution.

Definition 1 The stochastic process $\{X_t, t > 0\}$ is said to be increasing in randomness (decreasing in randomness) if X_{t2} is less (more) in randomness than X_{t1} for all $0< t1 < t2$.

Note that the concept of increasing in randomness means that the randomness of lifetime is increasing when lifetime is increasing too.

Since the equivalence, under certain restrictions, between the majorisation concept and the weak ordering we characterize this new aging concept by the IFR distributions and we propose several examples of classic stochastic processes as shock models and continuous wear.

References

Hickey, R. J. (1986), Concepts of Dispersion in Distributions: a Comparative Note, *J. Appl. Prob.* **23**, 914-921.

Lewis, T. and Thompson, J. W. (1981), Dispersive Distributions and the Connection Between Dispersivity and Strong Unimodality, *J. Appl. Prob.* **18**, 76-90.

Muñoz-Pérez, J. (1990), Dispersive Ordering by The Spread Function, *Stat. and Prob. Letters.* **10**, 407-410.

Pellerey, F. and Shaked, M. (1997) Charaterizations of the IFR and DFR Aging Notions by means of the Dispersive Order, *Statistics and probability letters.* **33**, 389-393

Shaked, M. (1982), Dispersive Ordering of Distributions, *J. Appl. Prob.* **19**, 310-320.

Shaked, M. and Shanthikumar, J. G. (1994), Sthocastic Orders and Their Applications, Academic Press. New York.

Suárez et al. (1998). Orden en Dispersión por Acumulación de Probabilidades en Intervalos. *Actas XXIV Congreso Nacional de Estadística e Investigación Operativa*.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Smoothing a Semi-Variogram

Yücel Tandoğdu
Eastern Mediterranean University, Department of Mathematics
Salamis Road, Famagusta, North Cyprus, Via Mersin 10, TURKEY
yucel.tandogdu@emu.edu.tr

1. What is a Semi-Variogram?

The semi-variogram is a measure of average dissimilarities between all possible points within a field *F*, for a certain variable of concern. It is defined as

(0.1) $$g(h) = \tfrac{1}{2} E\{[Q(x) - Q(x+h)]^2\}$$

where *Q* is a random function representing the distribution of a value of concern, *x* is a point, and *h* is a distance vector in a field of study. Equation (1.1) can be expressed as a spatial integral defined over the field *F*.

(0.2) $$g_F(h) = \tfrac{1}{F(h)} \int_{F \cap F_{-h}} [Q(x) - Q(x+h)]^2 \, dx$$

Here $F \bigcap F_{-h}$ is the intersection of *F* and its translate F_{-h} by vector *–h*. If $x \in F \bigcap F_{-h}$, then both *x* and *x+h* are in *F*, in which case *F(h)* becomes the measure of that intersection.

In nature it is anticipated that phenomena occurring in close proximity will tend to be very similar, while increasing distance will reduce the similarity. Over a field *F* the true but unknown dissimilarity defined in equation (1.2), can be estimated by the experimental semi-variogram $\hat{g}(h)$, that is given by (Rendu)

(0.3) $$\hat{g}(h) = \tfrac{1}{n(h)} \sum_{i=1}^{n} d_i^2 = \tfrac{1}{n(h)} \sum_{i=1}^{n(h)} [q(x_i) - q(x_i + h)]^2$$

2. Computation of $\hat{g}(h)$

Data obtained from a field of study representing a certain variable (mineral content, porosity, contaminants, etc.) is used in the computation of $\hat{g}(h)$ given in equation (1.3). Graphs, where distance versus $\hat{g}(h)$ values are plotted, gives an estimate of the true spatial function $g_F(h)$.

In application problems such as the irregular location of data values, and the extreme values in the data set are encountered in computing $\hat{g}(h)$. Locations of the data points are determined by the technicalities of the operation and can not be changed. On the other hand extreme values in a data set cause wild fluctuations in the computed $\hat{g}(h)$ values. This is a handicap in producing a spatial model for the variable under study. Effect of the extreme values can partially be alleviated by the use of the smoothing technique given below.

3. Smoothing a Semi-Variogram

Model fitting to an experimental semi-variogram (obtained from data) is a very demanding job, as the model parameters will have great influence on the estimation process. In application transitional models are widely used. For these models the nugget variance (*S*), the structural variance (*C*) and the range of influence (*a*) parameters are estimated using $\hat{g}(h_i)$, where *i* is an integer between 5 and 20 depending on the size of data. It can be shown that *S+C* takes a value around the true variance s_F^2 of the field *F*, (Goovaerts, Journel). The data variance s^2 can be used as an estimator of s_F^2. Confidence limits for s_F^2 determined using well known statistical methods give a good idea about the lower and upper boundaries for the sill. During modeling it is preferred that the sill value is closer to the lower confidence limit. This can be achieved either by excluding data pairs that result in very large d_i^2 values (Equation 1.3) or preferably reducing these d_i^2 values down to the upper confidence limit of s_F^2. This is the smoothing process of the experimental semi-variogram.

3.1. Case Study

A data set consisting of 359 data points containing the x, y coordinates and an attribute of concern (*A*) was used. Histogram of *A* has indicated that approximate normality can be assumed. Directional semi-variograms indicated that isotropy can be considered for attribute *A*, with variance $s_A^2 = 65.901$. Figure 1 shows the omnidirectional $\hat{g}(h)$, the smooth $\hat{g}(h)$ obtained by reducing the extreme d_i^2 values down to the upper confidence limit of $s_{F_A}^2$, and the upper and lower confidence limits of $s_{F_A}^2$ at 99% confidence level.

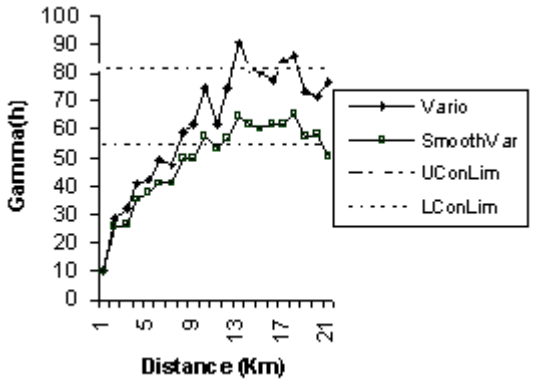


Figure 1. Experimental and smooth semi-variograms

References

Goovaerts, P. (1997). Geostatistics For Natural Resources Evaluation, 31-32, 101-103.
Journel, A. G., Huijbregts, J. Ch. (1978). Mining Geostatistics, 61-68.
Rendu, J. M. (1981). An Introduction to Geostatistical Methods of Mineral Evaluation, 15-17.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

A Data-Driven Non-Parametric Test for Component-Wise Independence Based on Sample Space Partitions

Olivier Thas, Jean-Pierre Ottoy
Ghent University, Dep. of Applied Mathematics, Biometrics and Process Control
Coupure Links 653, Ghent, Belgium
olivier.thas@rug.ac.be, jeanpierre.ottoy@rug.ac.be

1. The Test Statistic

Let S_n denote a sample of n i.i.d. p-variate observations $\boldsymbol{X}_i=(X_{i1}, \ldots, X_{ip})$ $(i=1,\ldots,n)$. The CDF of the joint distribution is denoted by F_x, which is assumed to be continuous throughout this paper. The sample space is denoted by S. We are interested in testing the hypothesis of independence between all components of $\boldsymbol{X}$ against any alternative (an omnibus test). Thus,

$$H_0 : F_x(x_1,...,x_p) = F_1(x_1)...F_p(x_p)$$

for all $\boldsymbol{x}$ in the sample space S. Further, let $[A]$ denote a partition of the p-dimensional sample space S, i.e. $[A]$ is a sample space partition (SSP). The, a partition construction rule may be defined such that each subsample P of q observations determines a $r_1 \times r_2 \times \ldots \times r_p$ SSP $[A]_P$ ($q=\max(r_1,\ldots,r_p)-1$). The set of all such subsamples that are induced by the sample, is denoted by $\boldsymbol{P}$, and $N=\#\boldsymbol{P}$. Conditional on any SSP $[A]_P$ the null hypothesis of component-wise independence may be tested by means of a simple Pearson c^2-test, which is denoted by $f^2(P)$.

Before the data-driven test is discussed, a family of test statistics, indexed by $r_1 \ldots r_p$, is introduced:

$$T_{r_1 \ldots r_p, n} = \frac{1}{N} \sum_{P \in \Pi} f^2(P) .$$

Note that this statistic is a rank statistic. Since, furthermore, the null hypothesis implies an invariance property, the exact null distribution of the test statistic is the permutational distribution. It may be interesting to note that this statistic is a generalisation of the test statistic introduced by Hoeffding (1948) and later formulated in terms of sample space partitions of size $2 \times 2 \times \ldots \times 2$ by Blum et al. (1961). They used however another measure for dependence in each table induced by the SSPs. Our test statistic may be considered as a generalisation of an Anderson-Darling statistic towards arbitrary SSP sizes as well.

We have shown that, under mild conditions, under H_0

$$T_{r_1 \ldots r_p, n} \xrightarrow{d} \sum_{j_1=1}^{\infty} \ldots \sum_{j_p=1}^{\infty} \frac{1}{\prod_{k=1}^{p} j_k (j_k+1)} Z^2_{j_1 \ldots j_p},$$

where the Z^2 are i.i.d. chi-squared distributed random variables with $(r_1-1)\ldots(r_p-1)$ degrees of freedom. Simulation experiments have shown, however, that the convergence is too slow to be useful for moderate sample sizes. Therefore, we suggest to use the (approximate) exact permutational null distributions. We will refer to this test as the SSP-test.

For $p=2$ this test is discussed in detail in Thas and Ottoy (2001). It may be shown that in a limiting case the test statistic becomes equivalent to a test statistic that was studied by Eubank et al. (1987) and Kallenberg and Ledwina (1999).

2. The Data-Driven Test

In the previous section a family of statistics was introduced, indexed by the size of the SSP. The arbitrariness of the SSP size may be considered as a flexible feature of the corresponding statistical test. This is clearly illustrated in a simulation in which the power of the SSP-test is estimated: by changing the SSP size the power may be considerably increased or decreased, depending on the alternative under study. Though, omnibus tests are often used in situations where the user does not have any idea on how the true dependence may look like, or, more specifically, he may be interested in all alternatives to independence equally well. In such cases, choosing a "wrong" SSP size may results in a power which is lower than optimal. In order to overcome this problem, we have made the test data-driven in the sense that the SSP size is estimated from the data. The estimated SSP size is given by

$$(R_1,...,R_p) = \text{Argmax}_{(r_1,...,r_p)\in\Gamma}\left\{T_{r_1...r_p,n} - 2(r_1-1)...(r_p-1)\ln(a_n)\right\},$$

where G is a set of permissible SSP sizes, and a_n, which determines the penalty in the selection rule, is such that $a_n \rightarrow$ as $n \rightarrow$. We have considered $a_n=n^{1/2}$ (cf. BIC), $a_n=(\ln(n))^{1/2}$ and $a_n=0.725\ \ln(n)$.

The data-driven test statistic now becomes

$$T_{R_1...R_p,n}.$$

3. Power Characteristics

We have proven that both the family of SSP-tests and the data-driven test are consistent against essentially any alternative. For moderate sample sizes ($n=20$ to 50), we have performed many simulation experiments in which the powers of the SSP-tests, as well as the powers of many other tests for independence are estimated and compared. The results suggest that the power of the SSP-tests with arbitrary SSP size are rather sensitive to the exact choice of the SSP size. With a good choice of the size the power can be made larger then almost any other test, but with a bad choice low powers may be obtained. The data-driven SSP-test solved this problem to a large extend, especially when the penalty based on $a_n=0.725\ln(n)$ is used.

References

Blum, J., Kiefer, J. and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function, *Ann. Math. Statist.* **32**, 485-498.

Eubank, R., LaRiccia, V. and Rosenstein, R. (1987). Test statistics derived as components of Pearson's phi-squared distance measure, *JASA* **82**, 816-825.

Hoeffding, W. (1948). A non-parametric test of independence, *Ann. Math. Statist.* **19**, 546-557.

Kallenberg, W. and Ledwina, T. (1999). Data-driven rank tests for independence, *JASA* **94**, 285-301.

Thas, O. and Ottoy, J.P. (2001). A nonparametric test for independence: a sample space partition approach, *J. of Nonparam. Statist.*, submitted.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Optimal Mean-Variance Robust Hedging under Asset Price Model Misspecification

T. Toronjadze
A. Razmadze Mathematical Institute, Georgian Academy of Sciences,
M. Aleksidze str.1, 380093, Tbilisi, Georgia
toro@rmi.acnet.ge

We consider the family X^{l}, $l \in \Lambda_e$, $\Lambda_e := \{ l : l = l^0 + eh,\ h \in K \}$, $e > 0$ (small parameter) of diffusions describing the misspecified discounted prices of a risky assets in a frictionless financial market, adapted to the filtration $F = (F_t)_{0 \le t \le T}$. A contingent claim is an F_T-measurable square integrable random variable, H, and a trading strategy q is a F-predictable process such that the stochastic integral $G(l,q) := \int q\, dX^{l}$, $l \in \Lambda_e$ is a well-defined real-valued square integrable semimartingale.

For each $l \in \Lambda_e$ the total loss of a hedger who starts with initial capital x, uses the strategy q, belives that the stock price process follows X^{l} and has to pay a random amount H at the date T, is thus $H - x - G_T(l;q)$. Denote $J(l,q) = E(H - x - G_T(l,q))^2$.

The robust mean-variance hedging means solving the optimization problem:

$$\text{minimize } \sup_{l \in \Lambda_e} J(l,q) \text{ over all strategies } q .$$

We solve this optimization problem approximating it (in the leading order e) by the problem:

$$\text{minimize } J(l^0,q) \text{ over all strategies subject to constraint}$$

$$\sup_{h \in K} \frac{DJ(l^0,h;q)}{J(l^0,q)} \le c ,$$

where $DJ(l^0,h;q)$ is the Gateaux differential of the functional J at the point l^0 in the direction h, c is some general constant.

Asymptotic Normality of the Deconvolution Kernel Estimator of the Distribution Function

Hae-Won Uh
University of Amsterdam, Korteweg-de Vries Institute for Mathematics
Plantage Muidergracht 24, 1018 TV Amsterdam, Nederlands
uh@science.uva.nl

Bert van Es
University of Amsterdam, Korteweg-de Vries Institute for Mathematics
vanes@science.uva.nl

Since the introduction of the deconvolution kernel density estimator there has been a great deal of results, such as asymptotic normality of the kernel density estimator and optimal rates for the density estimator and the distribution function estimator. However, the results concerning asymptotic normality of the deconvolution estimator of the distribution function are none too clear. Our results are related to those of Zhang (1990).

For simplicity we consider the estimation of $F(b)-F(a)$, $-\infty<a<b<\infty$, in the normal deconvolution problem, where we have observations from the convolution of the normal density and an unknown density f with distribution F. The estimator $F_{nh}(a,b)$ is defined as

$$F_{nh}(a,b) = \int_a^b f_{nh}(x)dx,$$

where

$$f_{nh}(x) = \frac{1}{2p}\int_{-1/h}^{1/h} e^{-itx} f_w(ht) e^{t^2/2} f_{emp}(t)dt,$$

and f_{emp} is the empirical characteristic function of the observations. Note that $f_{nh}(x)$ is the usual deconvolution kernel density estimator based on Fourier transform.

Under certain conditions we show that $F_{nh}(a,b)$ is asymptotically normally distributed with a standard deviation that is, apart from a constant, asymptotically equivalent to

$$\frac{1}{\sqrt{n}} h^{2(1+a)} e^{1/(2h^2)} \max(\sin(\tfrac{b-a}{2h}), h),$$

as $n \to \infty$ and $h \to 0$. Note the special role played by the bandwidths h for which the value $\sin((b-a)/2h)$ vanishes.

It turns out that the estimator can asymptotically be written as the sum of two means. Neither of them dominates for all bandwidths.

References

Zhang, C. H. (1990). Fourier methods for estimating mixing densities and distribution, *Ann. Statist.* Vol.18, 806-831.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Unemployment Duration Analysis for Married Women in Spain: Dealing with Length-Biased and Right-Censored Information

Jacobo de Uña-Álvarez
University of Vigo, Department of Statistics and O. R.
Campus Universitario Lagoas-Marcosende, 36200 Vigo, Spain
jacobo@correo.uvigo.es

María S. Otero-Giráldez
University of Vigo, Department of Applied Economics
Campus Universitario Lagoas-Marcosende, 36200 Vigo, Spain
sotero@correo.uvigo.es

Gema Álvarez-Llorente
University of Vigo, Department of Enterprises Organization and M.
Campus Universitario Lagoas-Marcosende, 36200 Vigo, Spain
galvarez@correo.uvigo.es

1. Introduction

It is very well-known that long unemployment duration greatly contributes to explain the high unemployment rates in Spain. On the other hand, unemployment is a particularly serious problem for Spanish women, mainly for those who are married. In this work we analyze the duration of unemployment spells for married women in Spain by using I. N. E. (the Spanish Institute for Statistics) data. The endpoint of the spell is defined by the transition to the "employed" and "out of the labour force" states.

Our data concern unemployment spells of 9950 Spanish women. This information was collected by means of repeated inquiries at the individuals' homes from 1987 to 1997. We included in the sample just those women being unemployed at the first inquiry time. As a result, each spell is sampled with a probability that is proportional to its length. This is typically referred as the *length-bias* problem. See, for example, Vardi (1982).

Moreover, because of the design of the inquiries, each individual was followed during no more than 18 months, so there was a risk of right-censoring for the unemployment duration time. Actually, 3774 spells were censored at the end of the period of observation, giving a censoring percentage of 38%. The statistical analysis of right-censored data is an old topic in theoretical and applied research. The product-limit Kaplan-Meier estimator has become the standard tool for inference from censored information. However, under length-bias, the product-limit is no longer consistent, and a suitable correction is needed. This problem, quite related to that of left-truncation, was considered by de Uña-Álvarez (2000), who introduced and analyzed a product-limit type estimate for length-biased censored data. His "corrected" version of the Kaplan-Meier curve allows for nonparametric estimation along with the fitting of parametric models. Parametric fits are useful, *e. g.*, for displaying in a smooth way curves such as the hazard rate and the density functions.

2. Main Results

Table 1 below summarizes our main results. This Table includes values for nonparametric estimates of the survival function and mean residual time functions. It also provides density and hazard rate values fitted under a loglogistic specification. The loglogistic model was found quite suitable for our data, at least during the first eleven years of unemployment.

Time (months)	Survival	Residual	Density	Hazard
3	0.9531	20	0.5534	0.6109
6	0.7321	23	0.5961	0.7868
9	0.5779	25	0.5108	0.8264
12	0.4673	28	0.4057	0.8055
15	0.3916	30	0.3150	0.7609
18	0.3354	32	0.2443	0.7093
21	0.2930	33	0.1911	0.6581
24	0.2653	33	0.1513	0.6104
27	0.2185	37	0.1213	0.5670
30	0.1903	39	0.0985	0.5281
33	0.1709	40	0.0810	0.4933
36	0.1616	39	0.0674	0.4622

Table 1. Unemployment duration of married women in Spain (1987-1997): survival, mean residual time (in months), density and hazard rate functions.

We see that the probability of staying unemployed for married women decreases quite rapidly at the beginning, this decreasement being much smoother after the first 2 years. The estimated mean unemployment duration time was 22 months. Compare this to the usual Kaplan-Meier mean (61 months!) and conclude about the importance of taking the length-bias problem into account. The increasing shape of the mean residual time function reveals how the possibilities of leaving unemployment disappear as time passes. Also, we outstand that the modal value of the density function is located at 6 months. Finally, note that the intensity of the transition to the "employed" and "out of the labour force" states (given by the hazard rate values) increases on the interval 3-9 months, then decreasing in a monotone way.

Acknowledgements

The first author acknowledges finantial support by the DGES grant PB98-0182-C02-02 and the Xunta de Galicia grant PGIDT00PXI20704PN.

References

de Uña-Álvarez, J. (2000). Product-limit estimation for length-biased censored data. Preprint.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias, *The Annals of Statistics* **10**, 616-620.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Optimization of Barron Density Estimates under the Chi-Square Criterion

Igor Vajda
Institute of Information Theory and Automation, Acad. of Sciences of the Czech Republic
18208 Prague, Czech Republic
vajda@utia.cas.cz

Edward van der Meulen
Department of Mathematics
Katholieke Universiteit Leuven, B-3001 Leuven, Belgium
ecvdm@gauss.wis.kuleuven.ac.be

1. Introduction

We consider independent observations $X_1,\ldots,X_n$, each distributed according to an unknown density f on the real line **R**. It is assumed that f is dominated by a known reference density f_0. Barron (1988) proposed a density estimate $f_n(\cdot) = f_n(\cdot;X_1,\cdots,X_n)$, which uses the reference density f_0 as a prior estimate of f. Let $\{A_{n1},\cdots,A_{nm}\}$ be a partition of **R** into $m = m_n < n$ a priori equiprobable intervals A_{nj}. Let $N(A_{nj})$ be the datacount in the bin A_{nj}. Then g_n, defined by $ng_n(x) = mN(A_{nj})f_0(x)$ if $x \in A_{nj}$, represents a histogram density estimate of f, which is shaped on each bin by the prior information f_0. The Barron density estimate f_n is defined as the convex mixture $f_n = \alpha g_n + (1-\alpha)f_0$, with $(n+m)\alpha = n$. Barron introduced the estimate f_n in his search for a density estimate which is consistent in information divergence, also called Kullback-Leibler distance and denoted by $I(f,f_n)$. Barron (1988) proved that if $I(f,f_0)$ is finite, and both $m \to \infty$ and $m/n \to 0$ as $n \to \infty$, then $EI(f,f_n)$ tends to zero as $n \to \infty$. Barron, Györfi and van der Meulen (1992) generalized and extended this result. These authors considered general convex mixtures of g_n and f_0 as possible density estimates, focused on distribution estimation (rather than density estimation), and proved a.s. consistency in information divergence of the Barron-type estimates. They also proved the consistency of these density estimates in reversed order information divergence under suitable conditions. Györfi et al. (1998) proposed to investigate the behaviour of the Barron density estimate f_n with the chi-square divergence $\chi^2(f,f_n)$ between f and f_n as error criterion. The chi-square divergence is topologically stronger than the information divergence. Györfi et al. (1998) proved the consistency in expected chi-square divergence of the Barron density estimate f_n, i.e. that $E\chi^2(f,f_n)$ tends to zero as $n \to \infty$, if $\chi^2(f,f_0)$ is finite, m satisfies the same condition with respect to n as above, and some additional conditions are fulfilled. They also showed that f_n is a.s. consistent in chi-square divergence if some stronger conditions are made. Berlinet, Györfi and van der Meulen (1997) showed that $I(f,f_n) - EI(f,f_n)$ is asymptotically normal, and Vajda and van der Meulen (1998a) proved the corresponding asymptotic normality of the chi-square divergence error. For consistency results of Barron-type density estimates with general divergence measures as error criterion see Berlinet, Vajda and van der Meulen (1998).

This work was supported by FWO Grant 1.5593.98N, GACR Grant 102/99/1137, and NATO Grant CRG 931030.

2. Main Results

In this paper we consider the problem of the optimization of partition sizes m_n, and of the selection of prior densities f_0 for Barron estimators f_n. First, new conditions are derived for the consistency of the Barron estimator in expected chi-square error, which are in particular satisfied for exponential f and f_0. This allows us later to apply our results to the case where both the unknown and the reference density are exponential. By showing that the maximum likelihood estimator of an exponential f is not consistent in expected chi-square error, we demonstrate the need for the nonparametric Barron estimator in this case. In solving the first problem we make use of a tight upper bound $B(n,m_n)$ on the expected chi-square error derived by Györfi et al. (1998), from which it can be concluded that the best rate of convergence of $B(n,m_n)$ to zero is achieved by $m_n = cn^{1/3}$ for some $c > 0$. In Vajda and van der Meulen (1998b) a begin was made with the optimisation of the constant c. Let m^0_n denote the value of m_n for which $B(n,m_n)$ achieves its minimum. We are interested in approximating the sequence m^0_n. Under the assumption that f and f_0 are more regular than required in the previous two papers we prove that, for the optimal sequence m^0_n, $n^{2/3}B(n,m^0_n)$ tends to a value d_0 as $n \to \infty$, where $16d^3_0 = 9J(f,f_0)$ and $J(f,f_0)$ denotes a Fisher information type distance between f and f_0. Moreover we prove that, with $c_0 = (2/3)d_0$,

$$B(n,m_n) = d_0 n^{-2/3} + o(n^{-2/3}) \qquad \text{iff} \qquad m_n = c_0 n^{1/3} + o(n^{1/3}).$$

Our results are illustrated by numerical studies carried out for several examples. For exponential f and f_0 satisfying our regularity assumptions and sample sizes $100 \le n \le 10.000$, the approximations $m^*_n = c_0 n^{1/3}$ are shown to coincide or almost coincide with exactly numerically calculated optima m^0_n. These conclusions remain valid for Rayleigh and Weibull densities f and f_0. The second problem, the optimization of f_0, is solved under the assumption that there is available an auxiliary estimate f* of f. In that case we either choose $f_0 = f^*$, or f_0 as the regular density different from f* which minimizes $J(f^*,f_0)$. This leads to two different versions f_n^I and f_n^{II} of the Barron estimator. The error bounds $B(n,m_n)$ achieved by these two versions are compared according to different preference criteria. The method and the results are illustrated by considering again the family of exponential densities. The complete version of this paper will appear as Vajda and van der Meulen (2001).

References

Barron, A.R. (1988). The convergence in information of probability density estimators. Presented at *IEEE Int. Symp. Inform. Theory*, Kobe, Japan.

Barron, A.R., Györfi, L. and van der Meulen, E.C. (1992). Distribution estimation consistent in total variation and in two types of information divergence, *IEEE Trans. Inform. Theory*, **38**, 1437-1454.

Berlinet, A., Györfi, L. and van der Meulen, E.C. (1997). Asymptotic normality of relative entropy in multivariate density estimation, *Publ. Inst. Stat. Paris*, **41**, 3-27.

Berlinet, A., Vajda, I. and van der Meulen, E.C. (1998). About the asymptotic accuracy of Barron density estimates, *IEEE Trans. Inform. Theory*, **44**, 999-1009.

Györfi, L., Liese, F., Vajda, I. and van der Meulen, E.C. (1998). Distribution estimates consistent in χ^2-divergence, *Statistics*, **32**, 31-57.

Vajda, I. and van der Meulen, E.C. (1998a). The chi-square error of Barron estimator of regular density is asymptotically normal, *Publ. Inst. Stat. Univ. Paris*, **42**, 93-110.

Vajda, I. and van der Meulen, E.C. (1998b). About the chi-square error of Barron density estimate, Prague Stochastics ′98 (eds M. Hušková, P. Lachout and J.A. Víšek), 2, 557-562. Prague : Union of Czech Mathematicians and Physicists.

Vajda, I. and van der Meulen, E.C. (2001). Optimization of Barron density estimates, *IEEE Trans. Inform. Theory*. To appear.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Estimating the Structural Distribution Function from a Large Number of Rare Events

Bert van Es
University of Amsterdam
Korteweg–de Vries Institute for Mathematics
Plantage Muidergracht 24, Amsterdam
The Netherlands
vanes@science.uva.nl

Chris A.J. Klaassen
University of Amsterdam
Korteweg–de Vries Institute for Mathematics
chrisk@science.uva.nl

Robert M. Mnatzakanov
Razmazde Mathematical Institut
Georgian Academy of Sciences
Tbilisi
Republic of Georgia
rob@rmi.acnet.ge

The concept of a structural distribution function originates from linguistics. A certain author has a vocabulary of words at his disposal and each text he writes is considered as a random collection of words from the vocabulary. Each word is assigned a probability of being used in the text. These probabilities are considered to be typical for the author and they are summarised in the so-called structural distribution function, i.e. an empirical distribution function of the word probabilities. For reasons of identification or comparison of texts it is desirable to be able to estimate this structural distribution function.

The basic probabilistic model we assume for the word counts consists of a Multinomial($n,p_{1M},...,p_{MM}$) random vector $n=(n_{1M},..., n_{MM})$ with typically large n, the number of words in the text, and a large M, the size of the authors vocabulary. We assume $n\to\infty$, $M\to\infty$ and $n/M\to l$, for some $0<l<\infty$. The structural distribution function F_M is given by

$$F_M(x)=\frac{1}{M}\sum_{j=1}^{M} I[Mp_{jM}\le x],\ \ x>0.$$

Additionally we assume that F_M converges weakly to a fixed distribution function F that we want to estimate.

The empirical distribution function of the word frequencies multiplied by M, the *natural estimator*, given by

$$\hat{F}_M(x)=\frac{1}{M}\sum_{j=1}^{M} I[\frac{M}{n}n_{jM}\le x],\ \ x>0,$$

is a straightforward estimator of *F*. However, this estimator turns out to be inconsistent, cf. Klaassen and Mnatsakanov (2000). We use the method of Poissonization to give a simple proof of the inconsistency.

Next we review several alternative weakly consistent methods, such as grouping of cells, an estimator based on Laplace inversion, and a kernel type estimator. The extra conditions, unfortunately needed to establish weak consistency, will be discussed.

References

Klaassen, C. A. J. and Mnatzakanov, R. M. (2000). Consistent estimation of the structural distribution function, *Scand. J. Statist.* 27, 733-746.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Semiparametric Estimation of Fractionally Cointegrated Time Series

Carlos Velasco
Universidad Carlos III de Madrid, Department of Statistics and Econometrics
Avenida de la Universidad 30, 28911 Leganés, Spain
cavelas@est-econ.uc3m.es

1. Nonstationary Time Series and Fractional Cointegration

Since the introduction of the concept of cointegration by Granger (1981), a vast literature has developed for the analysis of dynamic relationships among nonstationary time series. For nonstationary trending series with covariance stationary increments, cointegration implies that a (linear) combination of the observed series is stationary, at least less nonstationary, describing a long run equilibrium relationship. This idea of cointegration fits naturally in the broad field of fractionally integrated processes, generalising earlier analyses of integer integrated time series. Thus a series z_t is integrated of order $d_z \in (-0.5,0.5)$, i.e. $I(d_z)$, if its spectral density satisfies

(1) $$f_z(l) \sim G_z l^{-2d_z} \text{ as } |l| \to 0.$$

If z_t is nonstationary but has zero mean $I(d_z - 1)$ stationary increments Δz_t, then z_t is $I(d_z)$, $d_z \in [0.5,1.5)$. This characterisation of memory properties only attends to the relevant properties of the power spectrum, avoiding restrictions at high frequencies and covering the traditional $I(1)$ and $I(0)$ paradigms.

In this paper we consider that the observable $I(d)$ series y_t and x_t satisfy

(2) $$y_t = b\, x_t + u_t,$$

where u_t is $I(d)$, $d < d$, and propose semiparametric methods for the estimation of the degree of memory d of the residuals, the degree of cointegration $d - d$, and the regression coefficient b. Only the simple bivariate case is analysed. We first consider the situation where a preliminary consistent estimate $\tilde{b}$ is available to obtain residuals $\tilde{u}_t$. Semiparametric estimation avoids specification of short run system dynamics, being simpler to implement in practice.

2. Gaussian Semiparametric Residual Memory Estimation

Robinson and Marinucci (1998) and de Jong and Davidson (2000) show that least squares estimates of b in (2) given T observations of y_t and x_t are consistent under different regularity conditions. We take on their results and assume that

(3) $$\tilde{b} - b = O_p(T^{-r}), \quad r > 0,$$

r depending on d and d. For example, $r = d - d$ if $d + d > 1$ or $d = 1, d = 0$. We propose the estimate $\tilde{d}(\tilde{u})$ that minimises the following local Gaussian Whittle likelihood in the frequency domain

(4) $$Q_m(d, G_u, \tilde{u}) = \frac{1}{m}\sum_{j=1}^{m}\left\{\log G_u l_j^{-2d} + I_{\tilde{u}\tilde{u}}(l_j) \Big/ G_u l_j^{-2d}\right\},$$

where $I_{\tilde{u}\tilde{u}}(l_j)$ is the periodogram of the residuals $\tilde{u} = y_t - \tilde{b}x_t$. Alternatively, when nonstationary of u_t is suspected, it is possible to define $\tilde{d}(\Delta\tilde{u})$ based on the increments of residuals through $Q_m(d-1, G_u, \Delta\tilde{u})$. This Gaussian estimate was analysed in Robinson (1995), whose set up we follow here, including linear assumptions for u_t. The bandwidth m, growing with T, determines the number of Fourier frequencies l_j where the semiparametric model (1) is regarded as appropriate. An alternative semiparametric procedure is studied in Hassler *et al.* (2000), and despite its widespread use, it has less neat properties than ours.

We obtain that both versions of $\tilde{d}$ are consistent and maintain the same asymptotic distribution as if the u_t where observed, $\sqrt{m}(\tilde{d} - d) \to_d N(0, 1/4)$, as long as the order of cointegration $d - d$ is large enough, which implies a sufficiently fast rate of convergence in (3). This readily facilitates approximate inference rules for d. We remark that $\tilde{d}(\tilde{u})$ is also consistent for moderate nonstationary u_t, $d < 1$, as was shown in Velasco (1999) for observed series.

3. Joint Model Estimation

We also consider joint estimation of the vector parameter $q = (d, d, b)'$ through a multivariate version of the likelihood (4) for $(u_t(b), \Delta x_t)'$. This involves a generalisation of the semiparametric model (1) as in Lobato (1999). Following this reference, the joint estimate is based on a Newton-Raphson step,

$$\hat{q} = \tilde{q} - (\mathrm{L}_{qq}(\tilde{q}))^{-1}\mathrm{L}_q(\tilde{q}),$$

where $\mathrm{L}(\tilde{q})$ is the local likelihood evaluated at a preliminary estimate $\tilde{q}$. To develop asymptotic theory for $\hat{q}$, the initial estimates in $\tilde{q}$ have to converge fast enough, and these can be based indeed on the residual estimates of the previous section.

The memory estimates in $\hat{q}$ have the same asymptotic distribution as (residual) multivariate estimates, while $\hat{b}$ automatically corrects for the endogeneity bias that shows up in ordinary least squares estimates. This benefit is clearly reflected in a simulation study, where our estimates outperform other local proposals, e.g. in Robinson and Marinucci (1998), designed also to avoid this bias. The procedures of this paper and other issues regarding model specification are illustrated in an empirical
analysis of US monetary aggregates.

References

De Jong, R. M. and Davidson, J. (2000). The functional central limit theorem and convergence to stochastic integrals II: the fractional integrated case, *Economet. Th.* **16**, 643-666.

Granger, C. W. J. (1981). Some properties of time series data and their use in econometric model specification, *J. Economet.* **16**, 121-130.

Hassler, U., Marmol, F. and Velasco, C. (2000). Residual log-periodogram inference for long-run relationships, preprint.

Lobato, I. N. (1999). A semiparametric two step estimator in a multivariate long memory model, *J. Economet.* **90**, 129-153.

Robinson, P. M. (1995). Gaussian semiparametric estimation of long range dependence, *Ann. Statist.* **23**, 1630-1661.

Robinson, P. M. and Marinucci, D. (1998). Semiparametric frequency domain analysis of fractional cointegration, STICERD working paper, EM/98/350, LSE.

Velasco, C. (1999). Gaussian semiparametric estimation of non-stationary time series. *J. Time Ser. Anal.* **20**, 87-127.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Comparing Location Parameters of Exponential Populations

Sílvio Filipe Velosa
Universidade da Madeira e Centro de Estatística e Aplicações da Universidade de Lisboa
svelosa@math.uma.pt

An important feature of the exponential distribution is that the spacings $X_{1:n}$, $X_{2:n}- X_{1:n},\ldots, X_{n:n}- X_{n-1:n}$, where $X_{k:n}$ denotes the k-th ascending order statistic of the random sample $X_1,\ldots,X_c$, are independent, and that $X_{n:n}-X_{n-1:n} \stackrel{d}{=} X$; on the other hand, $nX_{1:n} \stackrel{d}{=} X$. The simple way of expressing a standard Pareto random variable Y as e^X, X an exponential random variable with appropriate shape parameter, may be used to establish similar results about quotients of consecutive order statistics of Pareto populations.

Brilhante (1996), and Brilhante *et al.* (1996) studied studentization in the two-parameter family of exponential random variables, establishing results about

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

where l is the location parameter, and

$$t_{n-1;i,k} = \frac{\overline{X} - l}{X_{k:n} - X_{i:n}};$$

thus inference on the location parameter may be performed using her results.

We establish results in the situation we have two independent exponential samples with location parameters l and l_2 respectively.

Our aim is to compare location parameters of two exponential populations. Assuming that both populations have the same scale parameter s, we first study the distribution of

$$t_1 = \frac{(X_{k:n} - l_1) - (Y_{j:n} - l_2)}{Z_{m+n:m+n} - Z_{m+n-1:m+n}}$$

and of

$$t_2 = \frac{(X_{k:n} - l_1) - (Y_{j:n} - l_2)}{(m+n) Z_{1:m+n}}$$

and their moments, where $Z_{I:m+n}$ denotes the i-th order statistic of the combined sample, and the implications we may get for Pareto populations.

* Research partially supported by FCT/POCTI/FEDER.

We discuss several possible ways of estimating the common scale parameter s; in case of different scale parameters, we use Welsh (1938) and Satterthwaite (1946) as guidelines.

References

Brilhante, M. F. (1996) Inferência sobre o parâmetro de localização de uma população exponencial . I. Studentização externa. , 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. *Decifrar o Mundo*, 57-63, Salamandra, Lisboa.

Satterthwaite, F. E. (1946) An approximate distribution of estimates of variance components. *Biometrics Bull*. (actual *Biometrics*) **2,** 110-114.

Welsch, B. L. (1938) The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350-362.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

On Influence of Block Effects on Growth Curve Fitting in Potthoff-Roy's Model

Miros³awa Weso³owska-Janczarek
University of Agriculture, Department of Applied Mathematics
Akademicka 13, 20-950, Lublin, Poland
janczar@ursus.ar.lublin.pl

The known growth curve model given by Potthoff and Roy (1964) as follows

$$\underset{np}{Y} = \underset{nm}{A}\,\underset{mq}{B}\,\underset{qp}{T} + \underset{np}{E}$$

under the assumptions $e(Y) = ABT$ and $\Sigma_{vec(Y)} = \underset{pp}{\Sigma} \otimes I_n$ where Σ is a covariance matrix for each of row-vectors of the observations matrix *Y*. The estimator of the matrix *B* of unknown coefficients in growth curves is

$$\hat{B} = (A'A)^{-} A'Y\hat{\Sigma}^{-1}T'(T\hat{\Sigma}^{-1}T')^{-1}$$

where

$$\hat{\Sigma} = \frac{1}{n}Y'[I_n - A(A'A)^{-}A']Y \ .$$

If the experimental units on which the explored feature was measured are divided into *a* homogeneous groups $(m = a)$, the matrix *A* is of full column rank and then $(A'A)^{-}$ is the usual inverse matrix.

But for a certain group of problems the measurements are obtained from experiments that are conducted in the complete block designs. Then the matrix *A* consists of two submatrices A_1 and A_2 defining the membership of each of the row-vectors of observations from the matrix *Y* in one of the *a* treatment groups and one of the *b* blocks and then $m = a + b,\ A = [A_1 \vdots A_2]$ and $B = \begin{bmatrix} B_1 \\ \cdots \\ B_2 \end{bmatrix}$. B_1 and B_2 are submatrices of coefficients of polynomials of $q-1$ degree for the treatments and blocks respectively.

The polynomials for the blocks are not interesting for the experimenter. What is interesting, the question whether it is possible to omit block effects in the growth curve model and what are the consequence of it.

Kala (1983). In this paper an attempt at the explanation of this problem is undertaken.

The study is based on the real-life data obtained from experiments on fruit-bearing of raspberries. The experiments were conducted in the complete block designs.

Two different cases are considered in this paper. First, when in the analysis of total yields from the entire fruit-bearing period the block effects are not significant and second, when these effects are significant.

The determination coefficients adapted into growth curves presented by Weso³owska-Janczarek (2000) are used as a measure of fitting of the curves assessed by Potthoff-Roy's method.

Generally, in both cases the calculated values of determination coefficients and mean determination coefficients when the block effects are ignored are a little higher. In the example, when block effects were unsignificant, the mean determination coefficient for the curves assessed, ignoring blocks is amounts to 22,6% whereas while taking block effects into consideration it is equal to 22,3%. In the latter case the same mean determination coefficients are equal to 41,6% and 36,2% respectively. The correlation coefficients calculated for the determination coefficients obtained without block effects and with them are equal to 0,965 in the former case and 0,734 in the latter one.

The conclusions are the following:

1°. Before the application of Potthoff-Roy's method to the estimation of the growth curves, when the experiment was conducted in the complete block design it is necessary to verify whether the block effects are significant.

2°. If the block effects are not significant, it is possible to ignore the block effects in the estimation of the growth curves.

References

Baksalary, J., Caliñski, T., Kala, R. (1983). Estymacja krzywych wzrostu w uk³adzie bloków kompletnych. Biuletyn Oceny Odmian, t. X. z. 1(15), 105-117.

Potthoff, R.F., Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* **51**, 313-326.

Weso³owska-Janczarek, M. (2000): On the use of determination coefficient to describe goodness of fit of assessed growth curves. *Biometrical Letters*, Vol. **37**, No 1, 13-20.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

Bayesian Approach to Parameter Estimation of the Generalized Pareto Distribution

P. de Zea Bermudez*
University of Lisbon, Center of Statistics and Applications
Campo Grande, Bloco C2, Piso 2, 1749-016 Lisbon, Portugal
patzea@fc.ul.pt

M. A. Amaral Turkman*
University of Lisbon, Center of Statistics and Applications
Campo Grande, Bloco C2, Piso 2, 1749-016 Lisbon, Portugal
antonia.turkman@fc.ul.pt

The distribution function of the Generalized Pareto Distribution (GPD) is given by:

$$F(x \mid k, s) = \begin{cases} 1-(1-kx/s)^{1/k} & \text{, if } k \neq 0 \\ 1-\exp(-x/s) & \text{, if } k = 0 \end{cases}$$

where k and $\sigma>0$ are respectively the shape and scale parameters of the distribution. For k 0, then $x > 0$ and when $k > 0$, then $0 < x < \sigma/k$.

The GPD was introduced by Pickands in 1975, who showed that the excesses above a sufficiently high threshold are distributed according to a GPD, provided the underlying distribution belongs to the domain of attraction of the generalized extreme value distribution.

Several methodologies have been used for estimating the parameters of the GPD, namely maximum likelihood (ML), the method of moments (MOM) and the probability-weighted moments (PWM). It is known that for these estimators to exist, certain constraints should be imposed on the range of the shape parameter k. Moreover, particularly for small sample sizes, the most efficient method to apply in any practical situation highly depends on a previous knowledge of the most likely values of k. This situation clearly suggests the use of Bayesian techniques as a way of introducing prior information on k.

Bayesian techniques have seldom been used for estimating the parameters of the GDP, probably due to the computational burden that generally comes associated with the implementation of Bayesian models. However, the development of powerful omputational tools during the last years has definitely enlarged the applicability of Bayesian procedures.

In the present work, we propose a simple Bayesian approach for estimating the parameters of the GPD. The Bayesian approach is compared, through a simulation study, with ML, PWM and with the elemental percentile method (EPM) developed by Castillo and Hadi (1997). The simulation study follows the same general lines as the ones performed by Hoskings and Wallis (1987) and Castillo and Hadi (1997). We

* The authors were partially supported by FCT, PRAXIS XXI and FEDER

consider sample sizes ranging from very small to very large (n = 15, 50, 100 and 550) and values of k (-2.0, -1.0, -0.4, -0.2, 0.2, 0.4, 1.0, 2.0) reflecting distributions going from very heavy-tailed to distributions with positive finite endpoints. As in the simulation studies performed by Hoskings and Wallis (1987) and Castillo and Hadi (1997), we set σ equal to one.

The quality of the estimates of k and σ is assessed by using bias and root mean squared error (RMSE).

Our main conclusion is that the Bayesian approach works extremely well to estimate the shape parameter of the GPD for k<0, both in terms of bias and RMSE. The results are particularly good for very large negative values of k (k = -2.0 and k = -1.0) for all sample sizes and when k = -0.4 and k = -0.2 for small sample sizes. Although the estimates of σ tend to have larger bias than the estimates produced by the other methods, the variances are smaller and hence they compare better in terms of RMSE. For k > 0 the performance of the Bayesian procedure is not as good as when k < 0, although it produces estimates of k and σ which are very reasonable in terms of RMSE.

The proposed estimation method is applied to two real data sets. The first set consists on a Norwegian fire insurance portfolio in 1981. These data are listed in Beirlant et al. (1996). The second data set, listed in Castillo and Hadi (1997), corresponds to zero-crossing hourly mean periods of the sea waves measured in the Bilbao buoy, in Spain. For both data sets several thresholds are considered and the excesses over the thresholds are modeled according to the GPD.

References

Beirlant, J., Teugels, J. L. and Vynkier, P. (1996), *Practical analysis of extreme values*, Leuven University, University Press.

Castillo, E. and Hadi, A. (1997) "Fitting the Generalized Pareto distribution to data," *Journal of the American Statistical Association*, **92**, 1609-1620.

Hoskings, J. R. M. and Wallis, J. R. (1987), "Parameter and quantile estimation for the generalized Pareto distribution", *Technometrics*, **29**, 339-349.

REVISTA DE ESTATÍSTICA
2º QUADRIMESTRE DE 2001

AUTHORS INDEX
ÍNDICE DE AUTORES

REVISTA DE
ESTATÍSTICA
2º QUADRIMESTRE DE 2001

General informations

CALENDÁRIO DE REUNIÕES

CALENDAR OF EVENTS

2001

- 13-19 August
  23rd European Meeting of Statisticians, Funchal, Island of Madeira, Portugal.
  Informações: Dinis Pestana, University of Lisbon and Rita Vasconcelos, University of Madeira,
  E-mail: dinis.pestana@fc.ul.pt
  rita@dragoeiro.uma.pt
  URL: http://www.fc.ul.pt/cea/ems2001

- 15-20 August
  SRTL-2: The Second International Research Forum on Statistical Reasoning, Thinking, and Literacy, to be hosted by the Centre for Cognition Research in Learning and Teaching and the School of Curriculum Studies in the University of New England, Armidale, Australia.
  Informações: Dr Chris Reading, Department of Curriculum Studies, University of New England, Armidale NSW 2351, Australia,
  Tel: +02-67735060,
  Fax: +02-67735078,
  Email: creading@metz.une.edu.au,
  URL: http://www.beeri.org.il/SRTL/

- 17-19 August
  5th ICSA International Conference, Co-sponsored by IMS, to be held in Hong Kong.
  Informações: IMS Program Chair: Howell Tong, Univ. of Hong Kong, Local Arrangements Chair: Wai Keung Li, University of Hong Kong
  Email: htong@hku.hk
  hrntlwk@hkucc.hku.hk
  URL: http://icsa.vlp.com/HK2001/

- 19-23 August
  22nd Annual Conference of ISCB (The International Society for Clinical Biostatistics) will be held in Stockholm, Sweden, August 19 - 23, 2001.
  Information: Scientific Secretariat.
  E-mail: Theresa.Westerstrom@iscb.stockholm2001.org
  URL: http://www.iscb.stockholm2001.org/.

- 21-25 August
  ICANN 2001, International Conference on Artificial Neural Networks of the European Network Society, to be held at the Vienna University of Technology, Austria.
  Informações: Conference Secretariat: ICANN 2001, Austrian Research Institute for Artificial Intelligence, Schottengasse 3, A-1010 Vienna, Austria.
  Email: icann@ai.univie.ac.at

- 22-29 August
  International Statistical Institute, 53rd Biennial Session (includes meetings of the Bernoulli Society, The International Association for Statistical Computing, The International Association of Survey Statisticians, The International Association for Official Statistics, The International Association for Statistical Education), Seoul, Korea.
  Informações: ISI Permanent Office, Prinses Beatrixlaan 428,
  P.O. Box 950, 2270 AZ Voorburg, The Netherlands.
  Tel.: +31–70–337–5737;
  Fax: +31–70–386–0025;
  E-mail: isi@cbs.nl
  or visit the Session website at http://www.nso.go.kr/isi2001

- 30-31 August
  IAOS Satellite Meeting on Statistics for Information Society, to be held in Tokyo, Japan.
  Informações: Akihito ITO, Japan Statistical Association, 2-4-6 Hyakunin-cho, Shinjuku-ku, Tokyo 169-0073, Japan.
  Tel: +81-3-5332-3151;
  Fax: +81-3-5389-0691;
  Email: jsa@jstat.or.jp or Ito@jstat.or.jp

- 30 August-1 September
  International Conference on Statistical Challenges in Environmental Health Problems, to be held at the Soft Research Park, Fukuoka City, Japan.
  Information: The Chairman, Organizing Committee, Takashi Yanagawa, Graduate School of Mathemathics, Kyushu University, Fukuoka 812-8581, Japan.
  E-mail: yanagawa@math.kyushu-u.ac.jp

- 30 August-1 September
  ICNCB - International Conference on New Trends in Computational Statistics with Biomedical Applications (ISI 2001 Satellite Meeting, co-sponsored by IASC), to be held at the Osaka University Convention Center, Osaka, Japan.
  Informações: ICNCB Office, Division of Mathematical Science, Graduate School of Engineering Science, Osaka University. 1-3 Machikaneyama-cho, Toyonaka, Osaka 560-8531, Japan; Fax +81(6)6850-6496.
  Email: ICNCB@jscs.or.jp
  URL: http://www.jscs.or.jp/ICNCB/

- 1-4 September
  The annual meeting of Japan Statistical Society will be held at Seinan Gakuin University.
  Informações: URL: http://sunyht2.ism.ac.jp

- 6-12 September
  International Association for MATHEMATICAL GEOLOGY 6th Int'l Conference Cancún, Mexico.
  Informações: Gina Ross, Kansas Geological Survey.
  Email: aspiazu@kgs.ukans.edu
  URL: http://www.kgs.ukans.edu/Conferences/IAMG

- 17-19 September
  Methodology and Statistics, to be held in Ljubljana, Slovenia at the Faculty of Social Science, University of Ljubljana, Kardeljeva pl. 5, Ljubljana.
  Information: Anuska Ferligoj.
  E-mail: anuska.ferligoj@uni-lj.si
  URL: http://vlado.fmf.uni-lj.si/trubar/preddvor/2001/.

- 20-22 September
  Rasch Symposium, in honour of Professor Georg Rasch 100 years birthday, to be held at the Copenhagen Business School, Copenhagen, Denmark.
  Informações: Marianne Andersen
  Email: ma.mes@cbs.dk
  URL: http://www.cbs.dk/news/200701.shtml

- 24-25 September
  Statistical methods in biopharmacy, 4th international meeting: "Integrating issues of efficacy, safety and cost-effectiveness", to be held in Paris, France.
  Informações: Jean Auclair, IRI Servier, 6 place des pléiades, 92415 Courbevoie cedex, France. Fax: 33 1 55 72 68 27.
  Email: sfds.2001@curie.net
  URL: http://www.sfds.asso.fr/groupes/congresbiophar/congres2001.htm

- 24-27 September
  Statistical Week 2001, to be held in Dortmund, Germany.
  Informações: URL: http://g2.www.dortmund.de/inhalt/statistik/statwoch/intro.htm

- 25-29 September
  32nd European Mathematical Psychology Group Meeting, to be held in Lisbon, Portugal. Includes a workshop on Teaching and Training Mathematical Psychology in an Interdisciplinary and International Context. An Introductory Course on "Mathematical Psychology and Data Analysis" will be held on September 25th.
  Information: Prof. Dr. Helena Bacelar-Nicolau, Tel: +351 21 793 45 54; Fax: +351 21 793 34 08.
  E-mail: hbacelar@fc.ul.pt
  or
  empg2001@fpce.ul.pt
  URL: http://correio.cc.fc.ul.pt/~cladlead/EMPG01.html.

- 1-3 October
  2nd International Symposium on PLS and Related Methods (PLS'01) to be held at Capri Palace, Island of Capri (Naples, Italy).
  Information: Dr. Vincenzo Esposito, Dipartimento di Matematica e Statistica, Facoltà di Economia, Università "Federico II" di Napoli, via Cintia, Monte Sant'Angelo. Tel. +39 081 675112, fax: + 39 081 675113;
  E-mail: binci@unina.it
  URL: www.dms.unina.it/PLS2001.html

- 29-31 October
  Statistics as bases of creation the economic policy and the economic development in the South-East Europe, to be held in Skopje, Republic of Macedonia.
  Information: Mr. Sasho Kjosev - Faculty of Economics, University " Sts. Cyril and Methodius", Skopje, Republic of Macedonia or Mrs. Biljana Apostolovska - State Statistical Office of the Republic of Macedonia.
  E-mail: skosev@eccf.ukim.edu.mk
  or
  biljanaa@stat.gov.mk

- 1-4 November
  Euroworkshop on Statistical Modelling - Nonparametric Models, to be held in Schloss Hoehenried, Bernried, near Munich, Germany.
  Information: Göran Kauermann (coordinator of the project) University of Glasgow, Dep of Statistics & Robertson Centre, Boyd Orr Building, Glasgow G12 8QQ.
  E-mail: goeran@stats.gla.ac.uk
  URL: http://www.stat.uni-muenchen.de/euroworkshop.

- 4-7 November
  IX Annual Congress of the Portuguese Statistical Society to be held at the Universidade dos Açores, Ponta Delgada, Portugal.
  Information: Comissão Organizadora Local do IX Congresso da SPE, Dep. Matemática, Universidade dos Açores, Apartado 1422 9501-801 Ponta Delgada, Portugal.
  E-mail: ix_congresso_spe@alf.uac.pt
  URL: http://www.ixcongressospe.uac.pt

- 12-16 November
  VIII Latin-American Congress in Probability and Mathematical Statistics, to be held at the University of Havana, Cuba.
  Information: Gonzalo Perera (Chairman Program Committee), Pablo Olivares (Chairman Local Organizing Committee).
  E-mail: gperera@fing.edu.uy
  or
  clapem@matcom.uh.cu
  URL: http://www.uh.cu/eventos/clapem/ehome.htm.

- 14-16 November

  The Federal Committee on Statistical Methodology, which is composed of the senior statisticians from several U.S. federal statistical agencies and is sponsored by the U.S. Office of Management and Budget is planning a research conference in Arlington, Virginia.

  Information: The conference will feature papers and software demonstrations on topics related to a broad range of government statistical research interests.
  URL: http://www.fcsm.gov/

- 21-22 November

  9th Conference on National Accounting: the measurement of the new economy; Paris, France. Simultaneous translation French-English.

  Information: Michel Boëda (INSEE) - Simultaneous translation French-English
  E-mail: michel.boeda@insee.fr
  URL: http://www.insee.fr/fr/av_service/colloques/cnat_accueil.html
  or
  http://www.insee.fr/en/av_service/colloques/cnat_accueil.html

- 7-9 December

  International Conference on "Characterization Problems and Applications", tentative Venue: Antalya, Turkey.

  Information: Omer L. Gebizlioglu, Ankara, Turkey; N. Balakrishnan, McMaster University, Canada; Ismihan Bayramov, Ankara, Turkey.
  E-mail: Omer.L.Gebizlioglu@science.ankara.edu.tr
  bala@mcmail.cis.mcmaster.ca
  Ismihan.Bayramov@science.ankara.edu.tr

- 19-22 December

  International Conference on Statistics, Combinatorics and Related Areas and The Eighth International Conference of the Forum for Interdisciplinary Mathematics, to be held at the University of Wollongong, Australia.

  Information: Chandra M. Gulati, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW 2522, Australia. Telephone:+61-2-4221-3836, fax:+61-2-4221-4845.
  E-mail: chandra_gulati@uow.edu.au
  or
  cmg@uow.edu.au
  URL: http://www.uow.edu.au/informatics/maths/statconference.

- 20-22 December

  Statistical Analysis for Global Environment, to be held at the Siam Intercontinental Hotel, Bangkok, Thailand.

  Information: Dr.Supol Durongwatana.
  E-mail: fcomsdu@phoenix.acc.chula.ac.th

- 20-23 December
  International Conference on History of Mathematical Sciences, to be held in Delhi, India.
  Information: Dr. Y. P. SABHARWAL, Department of Mathematics & Statistics, Ramjas College, University of Delhi, Delhi 110 007, India; Tel : (011) 294 1119.
  E-mail: ypsabharwal@yahoo.com
  or
  ichm2001rjc@yahoo.com

2002

- 15-18 January
  First International ICSC Congress on Neuro-Fuzzy NF'2002 to be held at The Capitolio de la Habana, Cuba.
  Informações: INTERNATIONAL COMPUTER SCIENCE CONVENTIONS Head Office: 5101C-50 Street, Wetaskiwin AB, T9A 1K1, Canada (Phone: +1-780-352-1912 / Fax: +1-780-352-1913)
  Email: operating@icsc.ab.ca
  or
  planning@icsc.ab.ca
  URL: http://www.icsc.ab.ca/NF2002.htm
  or
  http://www.icsc.ab.ca/

- 16-18 January
  Food-Industry and Statistics, to be held in Villeneuve d'Ascq (LILLE), France. Bât. EUDIL IAAL - Cité Scientifique, F 59655.
  Information: E-mail: agrostat2002@eudil.fr
  URL: http://www.eudil.fr/~agrostat.

- 4-8 February
  ProbaStat 2002, the 4th International Conference on Mathematical Statistics, to be held at Smolenice Castle, Smolenice, Slovak Republic.
  Information: E-mail: probastat@savba.sk
  URL: http://www.um.savba.sk/lab_15/probastat.html.

- 12-15 February
  First International ICSC-NAISO Congress on Autonomous Intelligent Systems ICAIS 2002 to be held at Deakin University, Geelong, Australia.
  Information: E-mail: icais02@itstransnational.com
  URL: http://www.icsc-naiso.org/conferences/icais2002/index.html

- 15-21 March
  ENAR/IMS Eastern Regional to be held in Washington, DC, USA.
  Informações: Program Chair: Jiayang Sun, Case Western Reserve University
  Local Arrangements Chair: Colin Wu, John Hopkins University
  Contributed Papers Chair: Nidhan Choudhuri;
  E-mail: jiayang@sun.STAT.cwru.edu
  colin@mts.jhu.edu
  nidhan@nidhan.cwru.edu
  URL: http://sun.cwru.edu/ims

- 2-5 June
  Annual Meeting of the Statistical Society of Canada, Hamilton, Ontario, Canada.
  Informações: Peter Macdonald, Department of Mathematics and Statistics, McMaster University, 1280 Main Street West, Hamilton, Ontario, L8S 4K1, Canada.
  E-mail: pdmmac@mcmail.cis.mcmaster.ca

- 17-20 June
  MMR 2002, Third International Conference on Mathematical Methods in Reliability, to be held at the Norwegian University of Science and Technology, Trondheim, Norway.
  Informações: Professor Bo Lindqvist, Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway. Tel.: +47-73 59 35 20 - Fax: +47-73 59 35 24.
  E-mail: mmr2002@math.ntnu.no
  URL: http://www.math.ntnu.no/mmr2002/

- 23-29 June
  The 8th International Vilnius Conference on Probability Theory and Mathematical Statistics, Vilnius, Lithuania.
  Informações: Professor Vytautas Statulevicius, Institute of Mathematics and Informatics, Akademijos str. 4, 2600 Vilnius, Lithuania.
  E-mail: conf@ktl.mii.lt

- 2-5 July
  MCQT'02 - First Madrid Conference on Queueing Theory, to be held at the Department of Statistics and OR, Faculty of Mathematics, University Complutense of Madrid, Spain.
  Information: Jesus R. Artalejo.
  E-mail: mc_qt@mat.ucm.es
  URL: http://www.mat.ucm.es/deptos/es/mcqt/conf.html.

- 7-12 July
  The Sixth International Conference on Teaching Statistics (ICOTS6), to be held in Durban, South Africa.
  Information: Maria-Gabriella Ottaviani - IPC Chair; Brian Phillips - International Organizer; , Dani Ben-Zvi - IPC Scientific Secretary.
  E-mail: mariagabriella.ottaviani@uniroma1.it; bphillips@swin.edu.au; dani.ben-zvi@weizmann.ac.il.
  URL: http://icots.itikzn.co.za/.

- 15-19 July
  Current Advances and Trends in Nonparametric Statistics, to be held on Crete, Greece.
  Informações: Michael G. Akritas and Dimitris N. Politis IMS Representative: Michael G. Akritas,
  E-mail: mga@stat.psu.edu
  URL: http://www.stat.psu.edu/~npconf/

- 21-26 July
  IBC 2002 - International Biometric Conference 2002, to be held at the University of Freiburg, Germany.
  Information: Chair: Robert Curnow; Chair Local Organizing Committee: Martin Schumacher.
  E-mail: r.n.curnow@reading.ac.uk ms@imbi.uni-freiburg.de
  URL: http://www.ibc2002.uni-freiburg.de/.

- 22-24 July
  26th Annual Conference of the Gesellschaft für Klassifikation (GfKl), to be held at the University of Mannheim, Germany.
  Informações: local organizer Prof. Dr. Martin Schader.
  URL: http://www.gfkl.de/gfkl2002

- 27 July – 1 August
  IMS Annual Meeting/Fourth International Probability Symposium, to be held in Banff, Canada.
  Informações: IMS Program Chair Tom DiCiccio, Cornell, Symposium Chair: Tom Kurtz, U. Wisconsin, IMS Local Chair: Subhash Lele, U. Alberta.
  E-mail: tjd9@cornell.edu Kurtz@math.wisc.edu slele@ualberta.ca

- 4-9 August
  Fourth International Conference on Statistical Data Analysis based on the L_1-Norm and Related Methods - to be held at the University of Neuchâtel, Switzerland.
  Information: Prof. Yadolah Dodge, Conference Organizer Statistics Group, Case Postale 1825, CH-2002 Neuchatel. Phone +41 32 718 13 80 Fax +41 32 718 13 81.
  E-mail: Yadolah.Dodge@unine.ch

- 11-15 August
  Joint Statistical Meetings, New York, Hilton and Sheraton New York.
  Sponsored by ASA, ENAR, WNAR, IMS, and SCC.
  Informações: ASA, 1429 Duke St., Alexandria, VA 22314-3415;
  Tel. (703) 684-1221;
  Email meetings@amstat.org

- 16-18 August
  Symposium on Stochastics and Applications (SSA) to be held at the National University of Singapore.
  Informações: E-mail: ssa@math.nus.edu.sg
  URL: http://www.math.nus.edu.sg/ssa

- 19-23 August
  24th European Meeting of Statisticians, Prague, Czech Republic.
  Informações: Martin Janzura, Institute of Information Theory and Automation, POB 18, 182 08 Praha 8, Czech Republic.
  Tel: 420 2 6605 2572.
  Fax: 420 2 688 4903.
  Email: janzura@utia.cas.cz

- 24-28 August
  Compstat2002 to be held in Berlin, Germany.
  E-mail: info@compstat2002.de, website http://www.compstat2002.de
  Informações: E-mail: info@compstat2002.de
  URL: http://www.compstat2002.de

- 25-28 August
  International Conference on Improving Surveys (ICIS-2002), to be held at the University of Copenhagen.
  Information: International Conference Services, P.O. box 41, Strandvejen 171, DK-2900 Hellerup, Copenhagen, Denmark. Telephone: +45 3946 0500, Fax +45 3946 0515.
  E-mail: ICIS2002@ics.dk

- 2-6 September
  RSS 2002 Conference to be held at the University of Plymouth, Plymouth, England.
  Information: The 2002 Conference of the Royal Statistitical Society (4-6 September) will preceded by short courses (2-3 September).
  E-mail: J.Stander@plymouth.ac.uk

- 13-17 November
  International Conference on Questionnaire Development, Evaluation, and Testing, probably to be held in the southeastern United States.
  Information: URL: http://www.jpsm.umd.edu/

- 28-30 December
  International Conference on "Ranking and Selection, Multiple Comparisons, Reliability, and Their Applications". Tentative Venue: Hotel Savera, Chennai, Tamilnadu, India.
  Organizers: bala@mcmail.cis.mcmaster.ca;, NKannan@utsa.edu; H. N. Nagaraja, Ohio State University, mailto:hnn@stat.ohio-state.edu
  Information: N. Balakrishnan, McMaster University; N. Kannan, University of Texas at San Antonio; H. N. Nagaraja, Ohio State University.
  E-mail: bala@mcmail.cis.mcmaster.ca
  NKannan@utsa.edu
  mailto:hnn@stat.ohio-state.edu

2003

- 10-20 August
  International Statistical Institute, 54th Biennial Session (includes meetings of the Bernoulli Society, The Intern. Assoc. for Statistical Computing, The Intern. Assoc. of Survey Statisticians, The Intern. Assoc. for Official Statistics and The Interna. Assoc. for Statistical Education), to be held in Berlin, Germany.
  Informações: ISI Permanent Office, Prinses Beatrixlaan 428,
  P.O. Box 950, 2270 AZ Voorburg, The Netherlands.
  Tel.: +31–70–337–5737;
  Fax: +31–70–386–0025;
  E-mail: isi@cbs.nl
  or visit the Session website at http://www.isi–2003.de

d) Informações sobre congressos, seminários, colóquios e conferências de

Para tal, são adoptadas as seguintes formas de contribuição para publicação na Revista:

- Quanto aos artigos referidos em a), contribuições da *iniciativa* dos próprios autores e por *convite* do Conselho Editorial, pertencentes ou não ao INE;
- Quanto às informações referidas em *b)*, *c)* e *d)*, contribuições dos departamentos do INE.

As contribuições de artigos por iniciativa dos próprios autores serão objecto de avaliação de mérito científico pelo Conselho Editorial, que decidirá ou não pela sua

Para a elaboração e envio das contribuições de artigos para publicação na *Normas de Apresentação de Originais*

Os autores dos artigos publicados, a que se refere a alínea a), receberão uma contribuição financeira paga pelo INE, de montante a fixar por despacho da Direcção mediante proposta do Director da Revista.

FOUNDATION, SUBJECT MATTER AND SCOPE OF THE REVIEW

INE is conscious of how statistical awareness is essential to the understanding of the majority of phenomena in the present world and is aware of its responsibility to disseminate statistical knowledge, making it available to the widest possible range of readers. INE has recognised the need to take a step in that direction and will begin publication of this *Statistical Review* three times yearly, designed to provide the following:

a) Within a scientific perspective, original articles on specialised areas of statistics, both pure and applied, as well as studies and analyses within the sphere of economics, social issues and demographics;

b) Information on activities and projects of the National Statistical System;

c) Information on activities developed by INE within the scope of co-operation;

d) Information on congresses, seminars and conferences of a statistical or related nature;

The following approaches for contributing material for publication in the review have been adopted:

- In relation to the articles referred to in section a), contributions are made by the authors themselves and by invitation of the Editorial Committee, whether they are employees of INE or not;
- In relation to the information referred to in section b), c) and d); contributions are from departments of INE.

The Editorial Committee who has sole discretion in deciding whether or not the material will be published will assess the scientific merit of contributions made on the initiative of the authors themselves.

The preparation and delivery of material for publication in the Review are subject to the *Rules for Submitting Originals* presented on the last page.

The authors of the published articles referred to in section a) will receive pecuniary compensation from INE in an amount to be determined by resolution of the Board on the recommendation of the Director of the Review.

resumo do artigo, com um máximo de 100 palavras,

b) Serão da responsabilidade dos respectivos autores as consequências de eventuais modificações da versão inicial aceite, bem como de atrasos na revisão das provas, que impossibilitem a publicação no número da Revista previsto, reservando-se o Director o direito de decidir a data da sua publicação futura;

c) Uma vez publicado o artigo, o autor receberá vinte exemplares da sua versão impressa e um exemplar do respectivo número da *Revista*.

7. Para *informações adicionais* contactar o Secretariado de Redacção:

Eduarda Liliana Martins
Instituto Nacional de Estatística
Avª. António José de Almeida, n.º 5 – 9º.
1000-043 Lisboa - Portugal

- Tel.: +351 21 842 62 05
- Fax.: +351 21 842 63 66
- e-mail: liliana.martins@ine.pt

RULES FOR SUBMITTING ORIGINALS

Within the terms of the *Regulation of the Statistical Review*, the Editorial Committee has approved the following **Rules for Submitting Originals**:

1. The original articles will be sent to the Review Director by the respective authors. They should be written in *Portuguese*, they should not have already been published in their entirety nor should they be in the process of being published in any other publication.

2. Articles may also be submitted in *English* to the Review's Director who will decide whether to accept them.

3. In relation to the *evaluation of the scientific merit* of the articles:

   a) The Editorial Committee will assess the articles submitted on the initiative of the authors on the basis of their scientific merit. The identity of both the author and the Committee members will be strictly confidential;

   b) The authors will receive information regarding the results of the evaluation of scientific merit within a maximum period of 30 days. If the article is accepted, the Committee will indicate the issue number of the *Review* in which the article will be published. If the article is not accepted, the original will be returned to the author.

4. The articles accepted for publication in the *Statistical Review* will also be made public on the Internet site of the INE.

5. The original articles having no more than thirty pages must be processed in *Word for Windows*, completely at black and white, with the information on the addicional(s) software(s) eventually used in the production of the original document, and they will be delivered in hard copy as well as on diskette, or sent by E-mail to: liliana.martins@ine.pt

6. With the presentation of the original articles, the authors must also respect the following rules:

   6.1 In relation to the *structure*:

   a) The text shall be printed on A4 format paper utilising the font *Times New Roman* size 11, spacing at least 12, and with the margins: *top* 4cm, *bottom* 3cm, *left* 2,5cm, *right* 5cm, *header* 1,25cm, *footer* 1,25cm;

   b) The first page shall contain only the title of the article as well as the name, address and telephone, fax and E-mail number of the author, indicating the position held and the institution that he/she belongs to. In the case of various authors, it is necessary to indicate the person to whom all correspondence received should be forwarded;

   c) The second page shall contain in *Portuguese* and *English* only the *title* and an *abstract* of the article with the maximum of 100

words followed by a paragraph indicating *key words* up to the limit of 15;

d) The third page will begin the text of the article with its respective sections or chapters sequentially numbered;

6.2 Regarding *Bibliographical References*:

a) Authors who are cited in the text of the article shall be indicated in parentheses with their name followed by the date of the respective publication and, if necessary, the page number (ex.: Malinvaud, 1989, 23);

b) All bibliographical references will be listed in alphabetical order by the surnames of the respective authors, immediately following the end of the text, as in the following example:

GREENE, W. H., "*Econometric Analysis*", Prentice-Hall, New Jersey, 1993.

6.3 Regarding *proof-reading and publication*:

a) Once the article is accepted and prior to its publication, the author will receive a copy for review. These copy will be returned to the Director of the Review within a maximum period of one week from the date of its reception;

b) The consequences of subsequent changes to the accepted first version are the responsibility of the respective authors as well as any delays in proof-reading that make its publication in the planned issue of the Review impossible. The Director reserves the right to decide upon the date for future publication;

c) Once the article is published, the author will receive twenty copies of his/her printed version and a copy of the respective issue of the *Review.*

7. For *further information* kindly contact the Editorial Secretary:

Eduarda Liliana Martins
Instituto Nacional de Estatística
Avª. António José de Almeida, nº. 5 – 9º.
1000-043 Lisbon - Portugal

- Tel.: +351 1 21 842 62 05
- Fax.: +351 1 21 842 63 66
- e-mail: liliana.martins@ine.pt