ISSN 1645-6726

INSTITUTO NACIONAL DE ESTATÍSTICA
PORTUGAL

REVSTAT

Statistical Journal

Volume 4, No.3
November 2006

Catalogação Recomendada

REVSTAT. Lisboa, 2003-
Revstat : statistical journal / ed. Instituto Nacional
de Estatística. - Vol. 1, 2003- . - Lisboa I.N.E.,
2003- . - 30 cm
Semestral. - Continuação de : Revista de Estatística =
ISSN 0873-4275. - edição exclusivamente em inglês
ISSN 1645-6726

CREDITS

- **EDITOR-IN-CHIEF**
  - *M. Ivette Gomes*
- **CO-EDITOR**
  - *M. Antónia Amaral Turkman*
- **ASSOCIATE EDITORS**
  - *António Pacheco*
  - *Barry Arnold*
  - *Dani Gamerman*
  - *David Cox*
  - *Dinis Pestana*
  - *Edwin Diday*
  - *Gilbert Saporta*
  - *Helena Bacelar Nicolau*
  - *Isaac Meilijson*
  - *Jef Teugels*
  - *João Branco*
  - *Ludger Rüschendorf*
  - *M. Lucília Carvalho*
  - *Marie Husková*
  - *Nazaré Mendes-Lopes*
  - *Radu Theodorescu*
  - *Susie Bayarri*
- **EXECUTIVE EDITOR**
  - *Maria José Carrilho*
- **SECRETARY**
  - *Liliana Martins*

- **PUBLISHER**
  - *Instituto Nacional de Estatística (INE)*
    Av. António José de Almeida, 2
    1000-043 LISBOA
    PORTUGAL
    Tel.: (0351) 21 842 61 00
    Fax: (0351) 21 844 04 01
    Web site: http://www.ine.pt
    Customer Support Service: 808 201 808
- **COVER DESIGN**
  - *Mário Bouçadas, designed on the stain glass window at INE by the painter Abel Manta*
- **LAYOUT AND GRAPHIC DESIGN**
  - *Carlos Perpétuo*
- **PRINTING**
  - *Instituto Nacional de Estatística*
- **EDITION**
  - *350 copies*
- **LEGAL DEPOSIT REGISTRATION**
  - *N.º 191915/03*

PRICE

[VAT 5% included]

- Single issue .... **€ 10**
- Annual subscription ... **€ 16**

INDEX

REVSTAT – Statistical Journal
Volume 4, Number 3, November 2006, 189–200

LIMIT DISTRIBUTION FOR THE WEIGHTED RANK CORRELATION COEFFICIENT, r_W

Authors: Joaquim F. Pinto da Costa
– Dep. de Matemática Aplicada, Universidade do Porto, Portugal
jpcosta@fc.up.pt

Luís A.C. Roque
– Dep. de Matemática, ISEP, Instituto Politécnico do Porto, Portugal
lar@isep.ipp.pt

Received: May 2005 Accepted: July 2006

Abstract:

- A weighted rank correlation coefficient, inspired by Spearman's rank correlation coefficient, has been proposed recently by Pinto da Costa & Soares [5]. Unlike Spearman's coefficient, which treats all ranks equally, r_W weights the distance between two ranks using a linear function of those ranks, giving more importance to top ranks than lower ones. In this work we prove that r_W has a gaussian limit distribution, using the methodology employed in [7].

Key-Words:

- *ranking; correlation; limit distribution.*

AMS Subject Classification:

- 62H20, 62E20, 62F07, 62F12.

1. INTRODUCTION

The objective of rank correlation methods is to assess the degree of monotonicity between two or more series of paired data. By monotonicity we mean a tendency for the values in the series to increase or decrease together (positive correlation) or for one to increase as the other decreases (negative correlation). They are applicable to paired data, that is to data where there is some connection between corresponding members of the samples. To use these methods, we must first rank the observations in each sample, $\boldsymbol{X}$ and $\boldsymbol{Y}$, from 1 (highest rank) to n (lowest rank), where n is the number of pairs of observations. We, thus obtain, $r(X_i)$ and $r(Y_i)$ where X_i and Y_i are the pair of values corresponding to observation i in each sample and $r(X_i)$ returns the rank of value i in the first series. For sake of simplicity, let us use the ranks directly rather than the values in the series. That is, $R_i = r(X_i)$ and $Q_i = r(Y_i)$.

There has been a growing interest about weighted measures of rank correlation [5, 1, 10, 6]; that is, measures that unlike Spearman's [11] coefficient which treat all ranks equally, weight ranks proportionally to how high they are, although other types of weight functions could be considered.

In 2005 Pinto da Costa & Soares [5] have introduced a weighted rank correlation coefficient, r_W, that weights the distance between two ranks using a linear function of those ranks, giving more importance to higher ranks than lower ones. These authors have also analysed the distribution of r_W in the case of independence between the two vectors of ranks. A table of critical values has been provided in order to test whether a given value of the coefficient is significantly different from zero, and a number of applications for this new measure has also been given.

In this work we start by defining this new measure of correlation in section 2. Then, in section 3 we analyse the asymptotic distribution of r_W for the general case; that is, we make no assumption of independence between the two vectors of ranks. To do so, we use the same notation and analogous arguments of those used by Ruymgaart, Shorack and Van Zwet (1972) in the proof of their Theorem 2.1 (see [7]). We prove that r_W has a normal limit distribution.

2. WEIGHTED RANK CORRELATION COEFFICIENT, r_W

In this section we describe a weighted measure of correlation that has been introduced in [5]. r_S is the value obtained by calculating Pearson's linear correlation coefficient of the paired ranks (R_1, Q_1), (R_2, Q_2), ..., (R_n, Q_n). It is easy to see that in the case of no ties,

$$r_S = 1 - \frac{6 \sum\limits_{i=1}^{n} (R_i - Q_i)^2}{n^3 - n} = 1 - \frac{6 \sum\limits_{i=1}^{n} D_i^2}{n^3 - n} \ ,$$

where $D_i^2 = (R_i - Q_i)^2$. As it is obvious from this expression, r_S only takes into account the differences between paired ranks and not the values of the ranks themselves. For instance, if $D_1 = 2$, doesn't matter whether the values for (R_1, Q_1) are $(1,3)$ or $(n-2, n)$. Nevertheless, there are applications where top ranks are much more important than lower ones, and Spearman's rank correlation does not take this into account. For instance, when humans state their preferences, it is obvious that top preferences are more important and accurate than lower ones. Another example might be the evaluation of stock trading support systems. A potential invester would like to have a system which gives a grading of the stocks in question so that he/she can make a decision. In order to evaluate the output of the system, one can for instance calculate Spearman's correlation between the ranking predicted by the system and the true ranking of the stocks at that time. However, the top ranked alternatives are obviously more important than the lower ones, which makes weighted measures of correlation more suitable for this application also.

In [5, 8], Pinto da Costa & Soares propose a measure of correlation — adapted from Spearman's rank correlation coefficient — that weighs ranks proportionally to how high they are. Specifically, they propose the following alternative distance measure:

$$W_i^2 = (R_i - Q_i)^2 \Big((n - R_i + 1) + (n - Q_i + 1) \Big) = D_i^2 (2n + 2 - R_i - Q_i) \ .$$

The first factor, D_i^2, represents the distance bewteen R_i and Q_i, exactly as in Spearman's; the second factor represents the importance of R_i and Q_i.

The authors then prove that in order to have a coefficient of the form $A + B \sum_{i=1}^{n} W_i^2$ that yields values in the range $[-1, 1]$, A must be 1 and $B = \frac{-6}{n^4+n^3-n^2-n}$. Their weighted measure of correlation is therefore,

$$r_W = 1 - \frac{6 \sum\limits_{i=1}^{n} (R_i - Q_i)^2 \Big((n - R_i + 1) + (n - Q_i + 1) \Big)}{n^4 + n^3 - n^2 - n} \ .$$

In [5] it is proved that under the hypothesis of independence between the two vectors of ranks, the expected value of r_W is 0, which is a desirable property for a correlation coefficient. Under the same hypothesis, $\operatorname{var}(r_W) = \frac{31n^2+60n+26}{30(n^3+n^2-n-1)}$. In addition, the authors have also conducted an experimental evaluation of the differences between the values obtained by r_W and r_S in various situations, showing that large differences can occur.

3. THE ASYMPTOTIC DISTRIBUTION OF r_W

Let (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) represent n i.i.d. random vectors from a continuous distribution. In this section, we show that r_W is asymptotically normal distributed. We start by showing the results of some simulations that indicate that this new statistic convergs to the gaussian curve in a particular case; namely, that the two vectors of ranks are independent. Then, we study formally the asymptotic distribution of r_W for the general case.

We have calculated the exact distribution of r_W for n up to 14. Due to computational limitations, for larger values of n we estimated the distribution based on a random sample of one million permutations. In Figure 1 we plot the distribution for $n = 14$ and $n = 15$, respectively the last exact and the first estimated distributions. In the same figure we also plot the estimated distributions for $n = 20$ and 40, respectively. In all graphs, the values of r_W have been standardized and we plot the Normal curve for comparison. From these graphs it seems clear that at least in this special case, the statistic r_W converges to the gaussian as n increases.

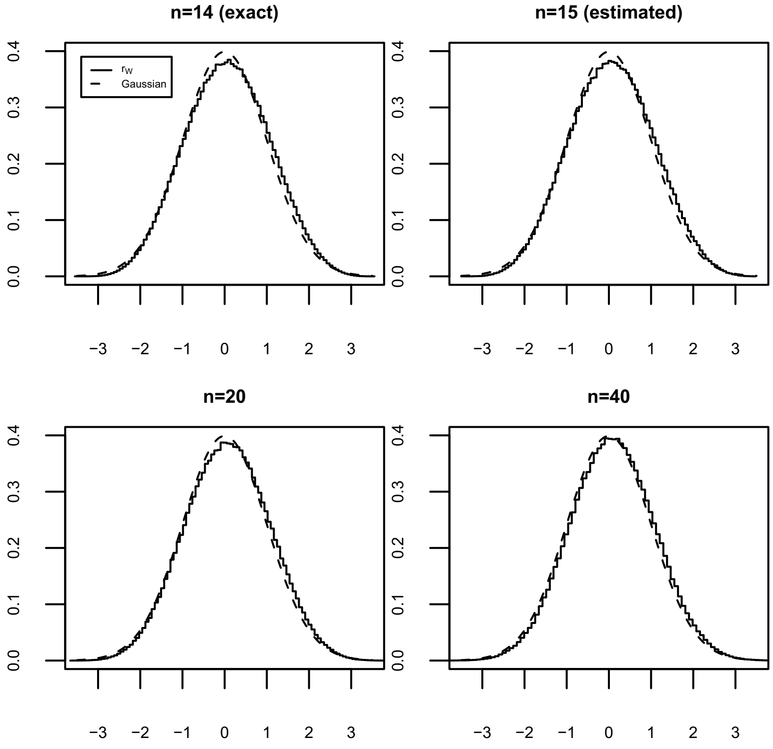


Figure 1: Exact distribution for $n = 14$ and estimated distribution for $n = 15$, 20 and 40, together with the Standard Normal curve.

Now we make no independence assumptions; that is, we study the asymptotic distribution of r_W for the general case. First,

$$\begin{aligned} r_W &= 1 - \frac{6 \sum\limits_{i=1}^{n} (R_i - Q_i)^2 \, (2n+2-R_i-Q_i)}{n^4+n^3-n^2-n} \\ &= 1 - \frac{6}{n} \sum_{i=1}^{n} \left(\frac{R_i}{n+1} - \frac{Q_i}{n+1}\right)^2 \left(\frac{2n+2-R_i-Q_i}{n-1}\right) . \end{aligned}$$

Therefore, the asymptotic behaviour of r_W is the same as the one of $1-6W_n$, where

$$W_n = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{R_i}{n+1} - \frac{Q_i}{n+1}\right)^2 \left(2 - \frac{R_i}{n+1} - \frac{Q_i}{n+1}\right) .$$

W_n is a statistic of the type $\frac{1}{n}\sum_{i=1}^{n} a_n(R_i, Q_i)$, where $a_n(i,j)$ is a real number for $i,j = 1,2,...,n$.

If we define $J(s,t) = (s-t)^2\,(2-s-t)$, $0 \le s,t \le 1$, then $J(s,t)$ is a limit of the score function

$$J_n(s,t) = a_n(i,j) = J\left(\frac{i}{n+1}, \frac{j}{n+1}\right) , \tag{3.1}$$

for i and j such that $\frac{i-1}{n} < s \le \frac{i}{n}$ and $\frac{j-1}{n} < t \le \frac{j}{n}$. Hence, W_n can be written as (see [2]),

$$W_n = \iint J_n(F_n, G_n)\; dH_n , \tag{3.2}$$

where F_n and G_n are the empirical marginal distribution functions of F and G, respectively; H_n is the bivariate empirical distribution function of H. Now, let us define the population moment $\mu = \iint J(F,G)\,dH$. By analogy to r_W, we define the population weighted rank correlation coefficient to be

$$\begin{aligned} \rho_W(X,Y) &= 1 - 6\mu \\ &= 1 - 6 \iint \Big(F(x) - G(y)\Big)^2 \Big(2 - F(x) - G(y)\Big)\; dH(x,y) , \end{aligned}$$

or, by using copulas [4]

$$\rho_W(X,Y) = 1 - 6 \int_{[0,1]^2} (u-v)^2\,(2-u-v)\; dc(u,v) ,$$

where the copula $c(u,v) = P\big(F(X) \le u,\, G(y) \le v\big)$, $0 \le u,v \le 1$.

Next we present the conclusion that r_W is assymptotically gaussian distributed.

Theorem 3.1. *r_W is an asymptotic normal and consistent (ANC) estimator of ρ_W.*

Proof: We want to prove that r_W is an asymptotic normal and consistent (ANC) estimator of ρ_W; first,

$$\sqrt{n}\,(r_W - \rho_W) \;=\; -6\,\sqrt{n}\,(W_n - \mu) \;=\; -6\,\sqrt{n}\left[\iint J_n(F_n, G_n)\; dH_n \;-\; \mu\right].$$

We start by considering the empirical processes $U_n(F) = \sqrt{n}\,(F_n - F)$, $V_n(G) = \sqrt{n}\,(G_n - G)$, $U_n^*(F) = \sqrt{n}\,(F_n^* - F)$, $V_n^*(G) = \sqrt{n}\,(G_n^* - G)$, where $F_n^* = \left[\frac{n}{n+1}\,F_n\right]$ and $G_n^* = \left[\frac{n}{n+1}\,G_n\right]$. Let now $\bar{\Delta}_n = [X_{1n}, X_{nn}] \times [Y_{1n}, Y_{nn}]$ where X_{in} and Y_{in} denote the i^{th} order statistics and $B_{0n}^* = \sqrt{n}\,\iint \left[J_n(F_n, G_n) - J(F_n^*, G_n^*)\right] dH_n$.

We will now prove that $J_n(F_n, G_n) = J(F_n^*, G_n^*)$ and so $B_{0n}^* = 0$ for all n. In fact the function F_n, for instance, is a step function and so there is always an $i \in \{0, 1, ..., n\}$ such that $F_n = \frac{i}{n}$; similarly for G_n. This means that by (3.1) $J_n(F_n, G_n) = J\left(\frac{i}{n+1}, \frac{j}{n+1}\right)$ for some i and j. Now, by the definition above, $\frac{i}{n+1} = F_n^*$ and $\frac{j}{n+1} = G_n^*$. So, $B_{0n}^* = 0$ for all n.

Because $B_{0n}^* = 0$ for all n, then an assumption similar to 2.3 b) in [7] (see Appendix A) is satisfied, that is, $B_{0n}^* \to_p 0$. We will now use the same argument of these authors, adapting it to our situation because our score function $a_n(i, j)$ is bivariate and the score functions used in [7], $a_n(i)$ and $b_n(i)$ have just one variable (see Appendix A). Nevertheless, the adaption follows from the same steps of their proof. The asymptotic convergence of r_W to the Normal distribution may be uniform over a class of distribution functions. However in this work we are not interested in proving uniform convergence, but only convergence for a single distribution.

Now we can write,

$$\sqrt{n}\,(W_n - \mu) \;=\; \sum_{i=1}^{3} A_{in} + B_{0n}^* + B_{1n}^* \;,$$

where

$$A_{1n} \;=\; \sqrt{n}\,\iint J(F, G)\; d(H_n - H)\;,$$

$$A_{2n} \;=\; \iint U_n(F)\,\frac{\partial J}{\partial s}(F, G)\; dH\;,$$

$$A_{3n} \;=\; \iint V_n(G)\,\frac{\partial J}{\partial t}(F, G)\; dH\;,$$

B_{0n}^* is defined above ,

$$B_{1n}^* \;=\; \sqrt{n}\,\iint \Big[J(F_n^*, G_n^*) - J(F, G)\Big]\, dH_n \;-\; A_{2n} \;-\; A_{3n}\;.$$

3.1. $\sum\limits_{i=1}^{3} A_{in}$ is asymptotically normal distributed

As in [7] we can prove the asymptotic normality of A_{1n}, A_{2n} and A_{3n} based on the fact that J is a continuous function and its partial derivatives are continuous and bounded on $(0,1)^2$.

Let us start by noting that $A_{1n}=\frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n} A_{1in}$ where $A_{1in}=J\big(F(X_i),G(Y_i)\big)-\mu$. In fact,

$$\begin{aligned} A_{1n} &= \sqrt{n}\iint J(F,G)\ d(H_n-H) \\ &= \sqrt{n}\left(\iint J(F,G)\ dH_n - \iint J(F,G)\ dH\right) . \end{aligned}$$

Now, as in equation 3.2 we get,

$$\begin{aligned} A_{1n} &= \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\Big(J\big(F(X_i),G(Y_i)\big)-\mu\Big) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(J\big(F(X_i),G(Y_i)\big)-\mu\Big) . \end{aligned}$$

The random variables A_{1in} are i.i.d. with mean zero. If we choose $\delta=\frac{1}{4}$, $D=p_0=q_0=2$, $r(u)=\frac{1}{u(1-u)}$ then we have an assumption similar to assumption 2.1 in the statement of Theorem 2.1 in [7] (See Appendix A), that is,

$$\begin{aligned} J(F,G) &\leq D\big(r(F)\big)^{a}\big(r(G)\big)^{b} && \text{with } a=\tfrac{\delta-\frac{1}{2}}{po}=-\tfrac{1}{8} \text{ and } b=\tfrac{\delta-\frac{1}{2}}{qo}=-\tfrac{1}{8} , \\ \frac{\partial J}{\partial s}(F,G) &\leq D\big(r(F)\big)^{a+1}\big(r(G)\big)^{b} && \text{with } a=\tfrac{\delta-\frac{1}{2}}{p1}=-\tfrac{1}{8} \text{ and } b=\tfrac{\delta-\frac{1}{2}}{q1}=-\tfrac{1}{8} , \\ \frac{\partial J}{\partial t}(F,G) &\leq D\big(r(F)\big)^{b}\big(r(G)\big)^{a+1} && \text{with } a=\tfrac{\delta-\frac{1}{2}}{p2}=-\tfrac{1}{8} \text{ and } b=\tfrac{\delta-\frac{1}{2}}{q2}=-\tfrac{1}{8} . \end{aligned}$$

Taking this assumption into account and by application of Holder's inequality,

$$\iint\big|\phi(F)\,\psi(G)\big|\,dH \leq \left[\int|\phi|^{p_0}\,dI\right]^{\frac{1}{p_0}}\left[\int|\psi|^{q_0}\,dI\right]^{\frac{1}{q_0}},\ \ \forall\, p_0>0,\ q_o>0\colon \frac{1}{p_0}+\frac{1}{q_0}=1 ,$$

where ϕ and ψ are functions on $(0,1)$, dI denotes Lebesgue measure restricted to the unit interval, we note that A_{1in} has a finite absolute moment of order $2+\delta_0$ for some $\delta_0>0$ (see appendix B).

Let us consider now A_{2n}. As $U_n(F)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(I(X_i\leq x)-F\big)$ we can write $A_{2n}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_{2in}$, where $A_{2in}=\iint\big(I(X_i\leq x)-F\big)\frac{\partial J}{\partial s}(F,G)\ dH$ are i.i.d. with mean zero. If we choose $\delta=\frac{1}{4}$, $D=p_1=q_1=2$, $r(u)=\frac{1}{u(1-u)}$ then an assumption similar to 2.1 in [7] is satisfied. Again, by applying Holder's inequality and similarly to A_{1in}, it follows that A_{2in} has a finite absolute moment of order $2+\delta_1$ for some $\delta_1 > 0$.

Let us consider now A_{3n}. As $V_n(G) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \big(I(Y_i \le y) - G\big)$ we can write $A_{3n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n A_{3in}$ where $A_{3in} = \iint \big(I(Y_i \le y) - G\big) \frac{\partial J}{\partial t}(F,G)\, dH$ are i.i.d. with mean zero. If we choose $\delta = \frac{1}{4}$, $D = p_2 = q_2 = 2$, $r(u) = \frac{1}{u(1-u)}$ then an assumption similar to assumption 2.1 in [7], is satisfied. By application of Holder's inequality and similarly to A_{1in}, it follows that A_{3in} has a finite absolute moment of order $2+\delta_2$ for some $\delta_2 > 0$.

From the above conclusions: $A_{1n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n A_{1in}$ where A_{1in} are i.i.d. with mean zero; $A_{2n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n A_{2in}$ where A_{2in} are i.i.d. with mean zero; $A_{3n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n A_{3in}$ where A_{3in} are i.i.d. with mean zero and because A_{1in}, A_{2in}, A_{3in} have a finite absolute moment of order larger than 2, we get $\sum_{i=1}^3 A_{in} \to_d N(0,\sigma^2)$ as $n \to \infty$. The expression for the variance corresponds to equation 3.10 in [7] and is given by

$$\sigma^2 \;=\; \mathrm{Var}\Big[J\big(F(X),G(Y)\big) + \iint \big(I(X \le x) - F\big) \frac{\partial J}{\partial s}\big(F(x),G(y)\big)\, dH(x,y) + \iint \big(I(Y \le y) - G\big) \frac{\partial J}{\partial t}\big(F(x),G(y)\big)\, dH(x,y)\Big]\,.$$

3.2. B_{1n}^* is asymptotically negligible

We have already seen that an assumption similar to 2.3 b) in [7] is satisfied. If we consider the mean value theorem (see [9]),

$$\sqrt{n}\, J(F_n^*, G_n^*) \;=\; \sqrt{n}\, J(F,G) + U_n^*(F)\, \frac{\partial J}{\partial s}(\phi_n^*, \psi_n^*) + V_n^*(G)\, \frac{\partial J}{\partial t}(\phi_n^*, \psi_n^*)$$

for all (x,y) in $\bar{\Delta}_n$ with $\phi_n^* = F + \alpha_3(F_n^* - F)$ and $\psi_n^* = G + \alpha_4(G_n^* - G)$, where α_3 and α_4 are numbers between 0 and 1, then B_{1n}^* can be decomposed as a sum of seven terms which are all asymptotically negligible by the same arguments used in section 5 of Ruymgaart et al. (1972) [7].

3.3. r_W is asymptotically normal distributed

We have thus that $\sqrt{n}(W_n - \mu) \to N(0,\sigma^2)$ in distribution and it is immediate that r_W is an asymptotic normal and consistent (ANC) estimator of $\rho_W :\ \sqrt{n}(r_W - \rho_W) \to N(0, 36\,\sigma^2)$. □

APPENDIX

A. Asymptotic Normality of Nonparametric Statistics

We present in this appendix Theorem 2.1 of Ruymgaart, Shorack and Van Zwet, 1972 (see [7]) as it is the fundamental tool used in the proof of our Theorem 3.1. We start by introducing some notation. Let $(X_1,Y_1),...,(X_n,Y_n)$ be a random sample from a continuous bivariate distribution function $H(x,y)$ (bivariate empirical df is denoted by H_n) having marginal dfs $F(x)$ and $G(y)$ and empirical df F_n and G_n, respectively. The rank of X_i is denoted by R_i and the rank of Y_i by Q_i. Let $T_n = \frac{1}{n}\sum_{i=1}^n a_n(R_i)\, b_n(Q_i)$, where $a_n(i)$, $b_n(i)$ are real numbers for $i = 1,...,n$. The standardization of T_n can be written as

$$\sqrt{n}(T_n - \mu) \;=\; \sqrt{n}\left[\iint J_n(F_n)\, K_n(G_n)\; dH_n \;-\; \mu\right] ,$$

where $J_n(s) = a_n(i)$, $K_n(s) = b_n(i)$, for $i = 1,...,n$ such that $\frac{(i-1)}{n} < s \le \frac{i}{n}$; $\mu = \iint J(F)\, K(G)\, dH$. The functions J and K can be thought of as limits of the score functions J_n and K_n. $\mathcal{H}$ denote the class of all continuous bivariate dfs H.

Assumption 2.1 (Ruymgaart, Shorack and Van Zwet, 1972)**.** The functions J and K are continuous on $(0,1)$; each is differentiable except at most at a finite number of points, and in the open intervals between these points the derivatives are continuous. The function J_n, K_n, J, K satisfy $|J_n| \le Dr^a$, $|K_n| \le Dr^a$ and $|J^{(i)}| \le Dr^{a+i}$ and $|K^{(i)}| \le Dr^{b+i}$ for $i = 0,1$. Here D is a positive constant, $a = \frac{(\frac{1}{2}-\delta)}{p}$, $b = \frac{(\frac{1}{2}-\delta)}{q}$ for some $0 < \delta < \frac{1}{2}$ and some $p,q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$.

Assumption 2.3 b (Ruymgaart, Shorack and Van Zwet, 1972)**.**

$$B^*_{0n} \;=\; \sqrt{n}\iint \Big[J_n(F_n)\, K_n(G_n) - J(F^*_n)\, K(G^*_n)\Big]\, dH_n \underbrace{\;\rightarrow\;}_{p} 0 \qquad \text{as} \quad n \to \infty$$

where $F^*_n = \big[\frac{n}{n+1}\big] F_n$ and $G^*_n = \big[\frac{n}{n+1}\big] G_n$.

Theorem 2.1 of Ruymgaart, Shorack and Van Zwet, 1972 (see [7])**.** *If H is in $\mathcal{H}$ and if assumptions 2.1 and 2.3 b) are satisfied, then*

$$\sqrt{n}(T_n - \mu) \underbrace{\;\rightarrow\;}_{d} N(0,\sigma^2) \qquad \text{as} \quad n \to \infty \;,$$

where μ *and* σ^2 *are finite and are given by*

$$\mu \;=\; \iint J(F)\,K(G)\;dH \qquad \textit{(expression 1.3 in [7])}$$

and

$$\begin{aligned}\sigma^2 \;&=\; \operatorname{Var}\Big[J\big(F(X)\big)\,K\big(G(Y)\big) \;+\; \iint(\phi_X-F)\,J'(F)\,K(G)\;dH\\ &\quad+\iint(\phi_Y-G)\,J(F)\,K'(G)\;dH\Big] \qquad \textit{(expression 3.10 in [7])}\end{aligned}$$

with $\phi_{X_i}(x)=0$ *if* $x<X_i$ *and* $\phi_{X_i}(x)=1$ *if* $x\ge X_i$.

B. A_{1in} has a finite absolute moment of order greater than 2

We show here that there exist $\delta_0>0$ and $\delta_0<\delta=\frac{1}{4}$ such that $E\,|A_{1in}|^{2+\delta_0}$ is bounded. Using Assumption 2.1 in the appendix above we can prove that

$$\iint \big|J\big(F(X_i),G(Y_i)\big)\big|^{2+\delta_0}\;dH \;\le\; D\iint \big|r(F)\big|^{a(2+\delta_0)}\,\big|r(G)\big|^{b(2+\delta_0)}\;dH\;.$$

By using now Holder's Inequality this quantity is

$$\begin{aligned}&\le\; D\,\frac{1}{n}\sum_{i=1}^{n}\Big\{r^{(2+\delta_0)(\delta-\frac12)}\Big(\frac{i}{n+1}\Big)\Big\}^{\frac{1}{p_0}}\,\Big\{\frac{1}{n}\sum_{i=1}^{n}r^{(2+\delta_0)(\delta-\frac12)}\Big(\frac{i}{n+1}\Big)\Big\}^{\frac{1}{q0}}\\ &=\;\frac{D}{n}\sum r^{(2+\delta_0)(\delta-\frac12)}\Big(\frac{i}{n+1}\Big)\\ &\le\; D\int_0^1\frac{1}{\big(u(1-u)\big)^{(2+\delta_0)(\frac12-\delta)}}\;du\end{aligned}$$

that is finite for $0<\delta_0<\delta=\frac{1}{4}$.

REFERENCES

[1] Blest, D. (2000). Rank correlation — an alternative measure, *Australian & New Zealand Journal of Statistics*, **42**(1), 101–111.

[2] Bhuchongkul, S. (1964). A class of nonparametric tests for independence in bivariate populations, *Ann. Math. Statist.*, **35**, 138–149.

[3] Chernoff, H. and Savage, I.R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics, *Ann. Math. Statist.*, **29**, 972–994.

[4] Nelsen, R.B. (1999). *An Introduction to Copulas*, Lecture Notes in Statistics No. 139, Springer, New York.

[5] Pinto da Costa, J.F. and Soares, C. (2005). A weighted rank measure of correlation, *Australian & New Zealand Journal of Statistics*, **47**(4), 515–529.

[6] Roque, L. (2003). *Métodos Inferenciais para o Coeficiente de Correlação ρ_W*, Tese de Mestrado em Estatística, Faculdade de Ciências, Universidade do Porto, Portugal.

[7] Ruymgaart, F.H.; Shorack, G.R. and Van Zwet, W.R. (1972). Asymptotic normality of nonparametric tests for independence, *The Annals of Mathematical Statistcs*, **43**, 1122–1135.

[8] Soares, C.; Costa, J. and Brazdil, P. (2001). *Improved statistical support for matchmaking: rank correlation taking rank importance into account.* In "Actas da JOCLAD 2001: VIII Jornadas de Classificação e Análise de Dados", p. 72–75.

[9] Salas, S.L.; Hille, E. and Etgen, G.J. (2003). *Calculus: One and several Variables*, 9th edition, Hardcover, Wiley.

[10] Soares, C.; Brazdil, P. and Costa, J. (2000). *Measures to compare rankings of classification algorithms.* In "Data Analysis, Classification and Related Methods", Proceedings of the Seventh Conference of the International Federation of Classification Societies IFCS (H. Kiers, J.-P. Rasson, P. Groenen and M. Schader, Eds.), pp. 119–124, Springer.

[11] Spearman, C. (1904). The proof and measurement of association between two things, *American Journal of Psychology*, **15**, 72–101.

REVSTAT – Statistical Journal
Volume 4, Number 3, November 2006, 201–225

COMBINING METHODS IN SUPERVISED CLASSIFICATION: A COMPARATIVE STUDY ON DISCRETE AND CONTINUOUS PROBLEMS

Authors: Isabel Brito
– Institut Curie, Service Bioinformatique, France
isabel.brito@curie.fr

Gilles Celeux
– INRIA Futurs, France
Gilles.Celeux@inria.fr

Ana Sousa Ferreira
– LEAD, FPCE, Universidade de Lisboa, Portugal
asferreira@fpce.ul.pt

Received: December 2005 Accepted: August 2006

Abstract:

- Often in discriminant analysis several models are estimated but based on some validation criterion, a single model is selected. In the purpose of taking profit from several potential models, *classification rules combining models* are considered in this article. More precisely two ways of combining models are considered: a serial combining method and a hierarchical combining method. Serial combining is a convex linear combination of a finite number of models. Hierarchical combining method leads to nested models structured in a binary tree. In this paper, several combining methods resorting from both points of view are presented and their performances are assessed on discrete and continuous classification problems.

Key-Words:

- *Gaussian classification; eigenvalue decomposition; multinomial classification; conditional independence model; convex combining; hierarchical combining.*

AMS Subject Classification:

- 62H30.

1. INTRODUCTION

In multivariate discriminant analysis, each object is assumed to arise from one of K exclusive groups $G_1, ..., G_K$ with prior probabilities $\pi_1, ..., \pi_K$, $\pi_k \geq 0$, $k=1, ..., K$, $\sum_k \pi_k = 1$. Each object is characterised by a multivariate vector $\mathbf{x}$ of d variables. In this article, all d variables are assumed to be either continuous or discrete. The conditional density that $\mathbf{x}$ belongs to group G_k is denoted by $f_k(\mathbf{x})$. Accordingly to the discrete or continuous case, $f_k(\mathbf{x})$ is a probability or a density probability function which has to be estimated from a n-dimensional training sample $\mathbf{t}$ ($\mathbf{t}_i = (\mathbf{x}_i, z_i)$, $i=1, ..., n$), where $\mathbf{x}_i$ is the d-dimensional vector measurement for unit i and $z_i \in \{1, ..., K\}$, denotes its group origin. Often, it is convenient to replace z_i with $\mathbf{y}_i$, a K-dimensional binary indicator vector of group membership for unit i: The k-th coordinate of $\mathbf{y}_i$ is 1 if i arises from group G_k and 0 otherwise.

The Bayes classifier assigns an individual vector $\mathbf{x}$ to G_g if

$$\pi_g\, f_g(\mathbf{x}) = \arg\max_k \pi_k\, f_k(\mathbf{x}) \;, \qquad k = 1, ..., K \,.$$

Usually, the group conditional probability function $f_k(\mathbf{x})$ is unknown and has to be estimated on the basis of the training sample $\mathbf{t}$. For continuous problems, the parametric paradigm is adopted and these functions are assumed to belong to a family of densities, in particular $f_k(\mathbf{x})$ are assumed to be d-normal with mean vector μ_k and covariance matrix Σ_d.

For discrete problems the most natural model is to assume that the group conditional probabilities $f_k(\mathbf{x})$ where $\mathbf{x} \in \{0,1\}^d$ are multinomial probabilities. (For simplicity, the discrete variables are supposed to be binary variables.) In this case, the group conditional probabilities are estimated by the observed frequencies in the training set $\mathbf{t}$. Goldstein and Dillon [14] call this model the full multinomial model (FMM). One way to deal with the curse of dimensionality consists of reducing the number of parameters to be estimated. The first-order independence model (FOIM) assumes that the d binary variables are independent in each group G_k ([14]).

In many situations M different classifiers are in competition for the same problem and one of those classifiers is selected, based on some validation criterion. Acting in such a way, leads to reject several classifiers for which the parameters have been estimated. Besides, misclassified objects can be different for the different classifiers. Thus, those classifiers may contain useful information about the supervised classification problem, and this information is lost by selecting a unique classifier. The idea of combining models is present in a growing number of papers, hoping to obtain a more robust and more stable model than any of the competing models ([27], [35], [36], [4], [7], [20], [29] and [25] are examples of such papers).

The aim of this paper is to gather and extend combining methods previously presented ([9], [10], [32] and [34]) and to assess their performances from numerical comparisons on real data set.

In this paper, two ways of combining classifiers, called serial combination method and hierarchical combination method, are considered on the basis of numerical experiments on real data sets. For serial combination, a convex linear combination of M models is considered

$$\sum_m \mathbf{c}^m(\mathbf{x})\,\beta_m\;, \qquad \beta_m \geq 0,\;\; \sum_m \beta_m = 1,\;\; m=1,...,M\;, \tag{1.1}$$

where $\mathbf{c}_m(\mathbf{x})$ indicates the output of model m. Usually, this output is the group conditional probabilities functions $f_k^m(\mathbf{x})$, $k=1,...,K$, or the posterior probabilities $p_k^m(\mathbf{x})$

$$p_k^m(\mathbf{x}) = \frac{\pi_k\, f_k^m(\mathbf{x})}{\sum\limits_g \pi_g f_g^m(\mathbf{x})}\;, \qquad g,k=1,...,K, \quad m=1,...,M\;, \tag{1.2}$$

or sometimes the membership estimation $z^m(\mathbf{x})$. To define the combining coefficients β_m, two strategies are possible: a single coefficient is associated to each model m (β_m is then a scalar) or K coefficients are associated to each model (β_m is then K-dimensional). The latter strategy can be thought of as attractive because it allows to choose a coefficient by model and by group. It means that it would be possible to weight differently the groups in the same combination of models. In fact, many numerical experiments on both real and simulated data ([33] and [10]) showed that this strategy produce awkward combining vectors. Moreover, in discrete problems, the training data sets are most often small in regard to the number of parameters to be estimated, and it is difficult to estimate several combining coefficients per model in a reliable way ([33]). A better strategy is to consider a single coefficient for each model. This strategy produces more stable and more interpretable combined models.

The methods that estimate a single coefficient per model are grouped according two different approaches based on least squares minimisation or on likelihood maximisation. In this work several methods have been considered according both approaches. Those methods are the committee of methods, which is a least squares minimisation technique and the other ones are based on likelihood ratios.

Hierarchical combining is different in spirit. It applies on polychotomous classification problems with $K>2$ groups and leads to nested models. Attention is focused on a method of combining models by a hierarchical coupling method related to an approach of Friedman [13]. This method is reducing the multigroup problem into several two-group problems. The hierarchical combined model is structured into a binary tree where each branch is associated to a model or a combination of models and a dichotomy between groups to be classified ([32], [34] and [9]).

The paper is organized as follows. In Section 2, the models in competition for both continuous and discrete classification problems are presented. In Section 3, the different convex combining strategies are described. Committee of methods and Likelihood ratios combining methods are presented in this section. Section 4 is devoted to the presentation of Hierarchical combining. Section 5 is concerned with the presentation of numerical experiments. The performances of combining models are compared on both discrete and continuous problems. For continuous data problems, serial and hierarchical combining methods are evaluated separately. Thus, when using hierarchical coupling, at each tree level only one model is chosen. For qualitative data problems, when using hierarchical combination at each node of the tree, a serial combination of models can be considered. Two sections, one about computer programs (Section 6) and another with a short discussion (Section 7) ends the paper.

2. CONTINUOUS AND DISCRETE CLASSIFIERS

In continuous supervised classification problems for assessing combining classification methods, the fourteen Gaussian models of EDDA ([3]) have been considered. Defined in the Gaussian setting, each group conditional probability function is supposed to be a d-dimensional Gaussian distribution with vector mean μ_k and covariance matrix Σ_k.

EDDA makes use of the variance matrix eigenvalue decomposition $\Sigma_k = \lambda_k\,\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ where $\lambda_k = |\Sigma_k|^{1/d}$, $\mathbf{D}_k$ is the eigenvector matrix of Σ_k and $\mathbf{A}_k$ is a diagonal matrix such that $|\mathbf{A}_k| = 1$, with the normalised eigenvalues of Σ_k on the diagonal in a decreasing order. This decomposition can lead to parsimonious and versatile models. Parameter λ_k denotes the volume of the k-th group, $\mathbf{A}_k$ its shape and $\mathbf{D}_k$ its orientation. Different assumptions on those parameters lead to fourteen models pooled into three families: eight elliptical models, four diagonal models and two spherical models. The eight elliptical models are

$$\big[\lambda\,\mathbf{D}\mathbf{A}\mathbf{D}^T\big]\,,\quad \big[\lambda_k\,\mathbf{D}\mathbf{A}\mathbf{D}^T\big]\,,\quad \big[\lambda\,\mathbf{D}\mathbf{A}_k\mathbf{D}^T\big]\,,\quad \big[\lambda_k\,\mathbf{D}\mathbf{A}_k\mathbf{D}^T\big]\,,$$
$$\big[\lambda\,\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T\big]\,,\quad \big[\lambda_k\,\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T\big]\,,\quad \big[\lambda\,\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T\big]\,,\quad \big[\lambda_k\,\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T\big]\,.$$

The absence of subscript k means that the parameter at hand has a fixed value over the groups and its presence that the parameter is free over the groups. For instance, models $[\lambda\,\mathbf{D}\mathbf{A}\mathbf{D}^T]$ and $[\lambda_k\,\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T]$ are respectively, the linear discriminant analysis model and the quadratic discriminant analysis model. Assuming that Σ_k are diagonal lead to the simplification $\Sigma_k = \lambda_k\mathbf{B}_k$, where $\mathbf{B}_k$ is a diagonal matrix where $|\mathbf{B}_k| = 1$. The four diagonal models are: $[\lambda\mathbf{B}]$, $[\lambda_k\mathbf{B}]$, $[\lambda\mathbf{B}_k]$, $[\lambda_k\mathbf{B}_k]$. The spherical models are $[\lambda\mathbf{I}]$, $[\lambda_k\mathbf{I}]$, $\mathbf{I}$ denoting the identity matrix. For each model, parameters $\mathbf{D}_k$, $\mathbf{A}_k$ or $\mathbf{B}_k$ and λ_k are estimated by maximizing the likelihood ([3]).

The output that has been considered for model m, in continuous combining context, is the posterior group probabilities $p_k^m(\mathbf{x})$ ([10]). In the hereunder considered examples, those probabilities have been derived by (1.2), where the prior probabilities π_k have been replaced with n_k/n (n_k is the number of units from G_k in the training set $\mathbf{t}$).

In discrete problems, only two reference models have been considered. They are the full multinomial model (FMM) and the first order independence model (FOIM). Those two models are expected to provide different classifiers in many circumstances. In the full multinomial model (FMM) the conditional probabilities are estimated with the observed frequencies

$$f_k(\mathbf{x}) = \frac{N(\mathbf{x}\,|\,k)}{n_k}\,, \qquad k=1,...,K\,, \tag{2.1}$$

where $N(\mathbf{x}\,|\,k)$ is the number of observations of the training sample, belonging to G_k, for which state $\mathbf{x}$ occurs. This model involves 2^d-1 parameters in each group. Hence, even for moderate d, not all of the parameters are identifiable.

Since data sets are small or very small in regard to the number of probabilities to be estimated, a problem of sparseness is encountered and some of the multinomial cells may have no data in the training sets. Thus smoothing the observed frequencies is desirable. Hand [16] has noticed that the choice of the smoothing method is not very important so that computationally less demanding methods may be used. Thus the observed frequencies are smoothed using a single smoothing parameter λ $(0<\lambda\leq 1)$ and the conditional densities takes the form (we omit the index k for simplicity)

$$f(\mathbf{x}\,|\,\lambda) \;=\; \frac{1}{n}\sum_i \lambda^{d-\|\mathbf{x}-\mathbf{x}_i\|}\,(1-\lambda)^{\|\mathbf{x}-\mathbf{x}_i\|}\;, \qquad i=1,...,n\;. \tag{2.2}$$

When $\lambda = 1.00$ no smoothing is proceeded and the amount of smoothing is increasing as λ decreases to 0. This method will be called KERNEL in the sequel.

The first-order independence model (FOIM) assumes that the d binary variables are independent in each group G_k, $k=1,...,K$. Then, the group probability function is of the form $\prod_j f(x_j\,|\,G_k)$, $j=1,...,d$, and is estimated by

$$f_k^I(\mathbf{x}) \;=\; \prod_j \frac{N(x_j\,|\,k)}{n_k}\;, \tag{2.3}$$

where $n_k = \sharp G_k$ and $N(x_j\,|\,k) = \sharp\{y\in G_k\colon\, y_j = x_j\}$. In this model the number of parameters to be estimated for each group is reduced from 2^d-1 to d. This method is simple but may be unrealistic in some situations.

The resulting serial combining classifier is using a single coefficient, producing an intermediate model between the full multinomial model and the first order independence model. Combining methods differ in the way this coefficient is derived.

3. CONVEX COMBINING STRATEGIES

3.1. Committee of methods

A natural way of deriving the coefficients β_m in serial combining is minimising the fitting error using a least squares criterion. The committee of methods introduced by Bishop [4] in the neural computing literature is such an approach. In the committee of methods that will be considered here to get a relevant convex combining of classifiers, the fit of a classifier m is measured with the group classification probabilities, $\mathbf{c}^m(\mathbf{x})$. The committee of models is of the form

$$\mathbf{c}_{\text{COM}}(\mathbf{x}) \;=\; \sum_m \mathbf{c}^m(\mathbf{x})\,\beta_m \;, \tag{3.1}$$

with $\beta_m > 0,\; m = 1, ..., M,$ and $\sum_m \beta_m = 1$. Writing $\mathbf{c}^m(\mathbf{x})$ as

$$\mathbf{c}^m(\mathbf{x}) \;=\; \mathbf{c}(\mathbf{x}) + \mathbf{e}^m(\mathbf{x}) \;, \tag{3.2}$$

where $\mathbf{c}(\mathbf{x})$ is the true group probabilities vector and $\mathbf{e}^m(\mathbf{x})$ represents the vector error of model *m,* leads to

$$\mathbf{c}_{\text{COM}}(\mathbf{x}) \;=\; \mathbf{c}(\mathbf{x}) + \sum_m \mathbf{e}^m(\mathbf{x})\,\beta_m \;. \tag{3.3}$$

Defining $\mathbf{C}$ the error correlation matrix of the models whose general term is

$$\mathbf{C}_{ml} = E\big[e^m(\mathbf{X})\,e^l(\mathbf{X})\big] \;, \qquad m, l = 1, ...M \;, \tag{3.4}$$

E denoting the expectation under the true distribution of the training dataset, the committee of methods consists of minimizing the error $Er = \sum_m \sum_l \beta_m \beta_l \mathbf{C}_{ml}$ under the constraint that the positive coefficients β are summing to one. Using standard Lagrangian manipulation leads to

$$\beta_m \;=\; \frac{\sum\limits_l (\mathbf{C}^{-1})_{ml}}{\sum\limits_m \sum\limits_l (\mathbf{C}^{-1})_{ml}} \;. \tag{3.5}$$

The correlation error matrix can be estimated by plug-in empirical values

$$\hat{\mathbf{C}}_{ml} \;=\; \frac{1}{n} \sum_i \big(\mathbf{y}_i - \mathbf{c}^m(\mathbf{x}_i)\big)\,\big(\mathbf{y}_i - \mathbf{c}^l(\mathbf{x}_i)\big)^T \;. \tag{3.6}$$

This formula means that in a natural way, the error vector $\mathbf{e}_m(\mathbf{x}_i)$ is estimated with

$$\hat{\mathbf{e}}_m(\mathbf{x}_i) \;=\; \Big(\hat{e}^k_m(\mathbf{x}_i) \;=\; y^k_i - c^k_m(\mathbf{x}_i)\Big) \,. \tag{3.7}$$

3.2. Likelihood ratios

LeBlanc and Tibshirani [20] presented an interesting combination method by likelihood ratios although they did not experiment it. It consists of choosing the combining coefficients as the ratio of the likelihood for model m over the sum of all models likelihoods,

$$\beta_m = \frac{L_m(\theta, \mathbf{x})}{\sum\limits_l L_l(\theta, \mathbf{x})} , \tag{3.8}$$

where, recalling that y_{ik} is the k-th coordinate of the indicator vector giving the label of unit i,

$$L_m(\theta, \mathbf{x}) = \prod_i \prod_k \left[f_k^m(\mathbf{x}_i)\, \pi_k\right]^{y_{ik}} .$$

In the discrete case the single coefficient β is

$$\beta_m = \frac{L_I}{L_I + L_M} , \tag{3.9}$$

where L_I, L_M represents the likelihood for the FOIM and the FMM models, respectively.

Since the likelihood increases with the model complexity, this weighting strategy will favour more complex models. Thus, it could be preferable to propose penalized versions of likelihood ratios.

A natural penalisation is inspired from Akaike Information Criterion (AIC) ([1]). Denoting ν_m the number of independent parameters of model m, the AIC criterion is AIC $= -2\ln(L_m(\theta, \mathbf{x})) + 2\,\nu_m$ and it leads to the combining coefficients

$$\beta_m = \frac{L_m(\theta, \mathbf{x}) \exp\{-\nu_m\}}{\sum\limits_l L_l(\theta, \mathbf{x}) \exp\{-\nu_l\}} . \tag{3.10}$$

In the discrete case, it takes the form

$$\beta_m = \frac{L_I \exp\{-Kd\}}{L_I \exp\{-Kd\} + L_M \exp\{-K(2^d - 1)\}} , \tag{3.11}$$

because Kd and $K(2^d - 1)$ are respectively the number of independent parameters for the FOIM and the FMM models.

Remark that in the discrete case, it appears that the likelihood ratio strategy derived from AIC leads always to a single coefficient with value one or zero and so this strategy is useless because it leads to a single model, FOIM or FMM (see [34]).

Another possibility, in the Bayesian model averaging spirit ([23] and [29]), is to base the combining weights on integrated likelihood ratios. The integrated or marginal likelihood for model m is

$$L(\mathbf{x}\,|\,m) \;=\; \int L_m(\theta,\mathbf{x})\, p(\theta_m)\, \mathrm{d}\theta_m \;, \tag{3.12}$$

where $p(\theta_m)$ is a prior probability distribution on θ_m.

Unfortunately, in most continuous cases, integral (3.12) is difficult to calculate. Kass and Wasserman [18] and Raftery [29] showed that integrated likelihood can be approximated using BIC criterion of Schwarz ([31]). This approximation leads to the combining coefficient for model m

$$\beta_m \;=\; \frac{L_m(\theta,\mathbf{x})\, n^{-0.5\,\nu_m}}{\sum L_l(\theta,\mathbf{x})\, n^{-0.5\,\nu_l}} \;. \tag{3.13}$$

In the discrete context, it is possible to get exact calculation of integral (3.12). In the non informative Bayesian setting, the prior distribution of FOIM parameters $p(a_j^k)$, $k\!=\!1,...,K$, $j\!=\!1,...,d$, are non informative Jeffreys distribution $\mathrm{B}(1/2,1/2)$ and prior distribution of FMM parameters $p(b_h^k)$, $k\!=\!1,...,K$, $h\!=\!1,...,s$, where s is the number of states, is a non informative distribution of Jeffreys $\mathrm{D}(1/2,1/2,...,1/2)$. From which, it follows directly that integrated likelihood for FOIM and FMM are

$$L_I(\mathbf{x}) \;=\; \frac{\prod\limits_k \prod\limits_j \mathrm{B}\big(x_k^j + 0.5\, n_k - x_k^j + 0.5\big)}{\mathrm{B}(0.5, 0.5)^{kd}} \;, \tag{3.14}$$

and

$$L_M(\mathbf{x}) \;=\; \frac{\Gamma(s/2)^k \prod\limits_k \prod\limits_h \Gamma\big(0.5 + c_k^h\big)}{\Gamma(1/2)^{ks} \prod\limits_k \Gamma\big(s/2 + n_k\big)} \;, \tag{3.15}$$

where c_k^h is the number of objects of group G_k with state h. And, the resulting combining coefficient β is estimated by

$$\beta \;=\; \frac{L_I(\mathbf{x})}{L_M(\mathbf{x}) + L_I(\mathbf{x})} \;. \tag{3.16}$$

4. HIERARCHICAL COMBINING

When the number of groups K to be discriminated is greater than two, as noted in Friedman [13], it can be advantageous to consider the polychotomous classification problem as a sequence of two group classification problem to get classifiers easier to be estimated and to be interpreted. Friedman proposed to decompose the K groups in all possible combinations of pairs of groups. For each pair of groups, a classifier is designed. The overall classifier is derived from all the pairwise classifiers by a majority vote.

The strategy we now present is different. A polychotomous problem is decomposed into several dichotomous problems but the dichotomous problems are nested in a hierarchical binary tree. It is the reason why this strategy is called hierarchical coupling. Let $\mathcal{G} = \{G_1, ..., G_K\}$ be the set of groups. Consider a partition of $\mathcal{G}$ in two elements. At this level the best two class partition of groups is designed according to some criterion and the model or combination of models leading to the two class classifier minimizing the cross validated error rate between the two classes is designed. According to the sample size of the learning sample, leave one out or v-fold cross validation is considered. If available, it is also possible to assess the error rate with a test sample.

The procedure is repeated until all the elements in the actual partition are single groups. The combining classifier obtained from this hierarchical coupling procedure can be represented in a hierarchical tree as exemplified in Figure 1.

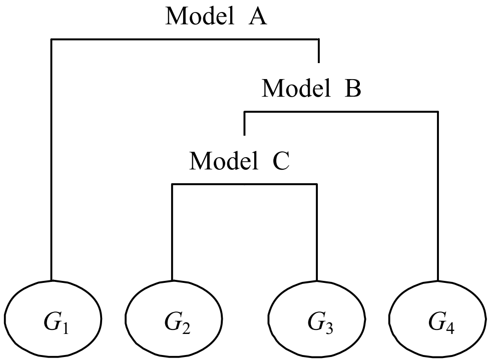


Figure 1: Example of hierarchical combined model for a four group problem.

The classifier depicted in Figure 1 is as follows. When a new observation is presented to the hierarchical classifier it passes through model A that classifies it in G_1 or $G_2 \cup G_3 \cup G_4$. If model A classifies the observation in G_1 the analysis is stopped. Otherwise, the observation passes through model B and the decision is G_4 or $G_2 \cup G3$. If model B does not classify the observation in G4 it passes finally through model C that assigns the observation to G_2 or G_3.

In order to choose, at each level, the best model or combination of models and the best partition, different strategies for continuous and discrete problems are employed.

In continuous data context, it was proceeded as follows:

1. For each possible binary partition all M models are estimated (at the beginning level there are $M(2^{K-1}-1)$ couples (model, partition)).
2. From those couples, the one providing the lowest misclassification error rate (ME) is chosen. In all the experiments, ME is evaluated by leave one out cross validation.

In the discrete case, the hierarchical coupling procedure is somewhat different.

1. At each level of the binary tree, the choice of the two-class decomposition of groups among the $2^{K-1}-1$ possible decomposition is done by minimizing the basic affinity coefficient ([24] and [2]) between the two classes of groups: Denoting $\mathrm{F}_1=\{p_j\}$ and $\mathrm{F}_2=\{q_j\}$, $j=1,...,d$, two discrete distributions defined on the same space, the affinity coefficient between F_1 and F_2 is given by $\rho(\mathrm{F}_1\mathrm{F}_2)=\sum_j\sqrt{p_j}\sqrt{q_j}$. Then the two classes of groups minimizing the affinity coefficient are selected.
2. After the two classes of groups have been chosen, the combining model is chosen by minimizing the error rate evaluated on a test sample or by v-fold cross validation.

Consider the example for a four group problem:

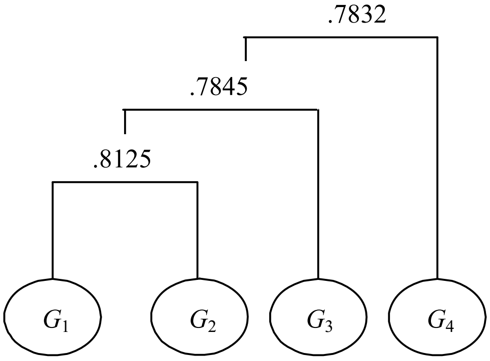


Figure 2: Example of hierarchical combined model for a four group discrete problem with the basic affinity coefficient values displayed.

It can be noticed that hierarchical combining method leads often to simple models at each step. From this point of view, it can lead to easily interpretable and stable decision rules, avoiding unnecessary complicated models.

5. RESULTS ON REAL DATA

In continuous context, combining methods have been applied on benchmark real datasets. Four of them were taken from the Machine Learning Repository of California University [5] (MLR), one from the Oxford University Repository [26] (OR) and another one from [15] (Hab). Table 1 provides a brief description of each dataset and their source.

Table 1: Continuous datasets description.

Dataset	Source	Description	Nb of units	Nb of features	Nb of groups
Bupa	MLR	Presence/absence of liver disorders that might arise from excessive alcohol consumption, measured by blood tests	345	6	2
Crabs	OR	Morphology of two species, blue and orange, by sex, of Australian crabs	200	5	4
Haberman	MLR	Survival of patients who had undergone surgery for breast cancer	306	3	2
Haemo	Hab	Presence of haemophilia on women	75	2	2
Iris	MLR	Measurements on the sepal and petal iris to determine iris specie (the famous Fisher dataset)	150	4	3
Thyroid	MLR	Medical records to predict the type of patients thyroidism	215	5	3

In discrete context, several real and simulated binary datasets were used to evaluate the performance of the considered strategies. Table 2 gives a brief description of each real dataset.

Table 2: Discrete datasets description.

<table>
<tr><th>Dataset</th><th>Source</th><th colspan="2">Description</th><th>Nb of units</th><th>Nb of features</th><th>Nb of groups</th></tr>
<tr><td>Medical Data</td><td>[30]</td><td colspan="2">Presence/absence of four symptoms liver disorders to predict the type of icterus</td><td>20</td><td>4</td><td>2</td></tr>
<tr><td>Psychological Data in older people</td><td>[11]</td><td colspan="2">Scores obtained for each older adult in the six dimensions of the Psychological Well-Being Scale taken as binary data into two groups</td><td>80</td><td>6</td><td>2</td></tr>
<tr><td>Psychological Data</td><td>[28]</td><td colspan="2">Six binary variables of a psychological test — Rorschach test — in 3 groups with different degrees of alexithymia</td><td>34</td><td>6</td><td>3</td></tr>
<tr><td rowspan="3">Psychological Counselling Career Data</td><td rowspan="3">[21]</td><td rowspan="3">Students of four licenciature's: Biology (B), Psychology (P), Language and Literature (LL), Engineering (E), described by the Psychological Questionnaire — My Vocational Situation — that is organised in three scales</td><td>Vocational Identity (VI) with 6 items</td><td>600</td><td>6</td><td>4</td></tr>
<tr><td>Occupational Information (OI) with 4 items</td><td>600</td><td>4</td><td>4</td></tr>
<tr><td>Barriers (B) with 4 items</td><td>600</td><td>4</td><td>4</td></tr>
</table>

5.1. Performance of serial combining techniques

The continuous case

Because several of the fourteen EDDA models lead to similar classifiers, combining all of them is useless. The more different models have been determined from the Correspondence Analysis of the fourteen models involved in EDDA described with their posterior densities $p_k^m(\mathbf{x})$ (see Brito [9]). For each dataset, the chosen models are given in Table 3.

Table 3: EDDA models chosen for each dataset by a Correspondence Analysis.

Dataset	Chosen models
Bupa	$[\lambda\mathbf{B}]$, $[\lambda_k\mathbf{B}]$, $[\lambda\mathbf{I}]$, $[\lambda_k\mathbf{I}]$
Crabs	$[\lambda\mathbf{DAD}^T]$, $[\lambda\mathbf{I}]$
Haberman	$[\lambda\mathbf{D}_k\mathbf{AD}_k^T]$, $[\lambda\mathbf{B}_k]$, $[\lambda\mathbf{I}]$
Haemo	$[\lambda\mathbf{DAD}^T]$, $[\lambda_k\mathbf{DAD}^T]$, $[\lambda\mathbf{I}]$
Iris	$[\lambda\mathbf{B}]$, $[\lambda\mathbf{I}]$
Thyroid	$[\lambda\mathbf{B}]$, $[\lambda\mathbf{I}]$

Serial combining methods were evaluated by leave-one-out cross validated misclassification error rate (ME). The purpose is to compare combining techniques opposite to single model techniques. In Tables 4 to 5, ME on each database are presented and compared with ME of model chosen with the standard EDDA strategy.

Table 4: Model and ME for each dataset using the committee of methods technique and EDDA.

Dataset	Committee of methods		EDDA	
	Model	ME	Model	ME
Bupa	$.79\,[\lambda\mathbf{B}] + .21\,[\lambda\mathbf{I}]$	.3971	$[\lambda\mathbf{B}]$	.4000
Crabs	$[\lambda\mathbf{DAD}^T]$	.5000	$[\lambda\mathbf{I}]$	.0500
Haberman	$.4\,[\lambda\mathbf{D}_k\mathbf{AD}_k^T] + .6\,[\lambda\mathbf{I}]$	.2549	$[\lambda\mathbf{B}_k]$	.2516
Haemo	$[\lambda\mathbf{DAD}^T]$	.1600	$[\lambda\mathbf{DAD}^T]$	.1467
Iris	$.82\,[\lambda\mathbf{B}] + .18\,[\lambda\mathbf{I}]$	.0400	$[\lambda\mathbf{B}]$	.0400
Thyroid	$.73\,[\lambda\mathbf{B}] + .27\,[\lambda\mathbf{I}]$	.0930	$[\lambda\mathbf{B}]$	.0977

For **Bupa** and **Thyroid** datasets, misclassification error rate is slightly better using the committee of methods technique. **Bupa** dataset contains information on the presence or absence of liver disorders caused by excessive alcohol consumption. **Thyroid** dataset resumes medical records in order to predict patient type of thyroidism. In both cases, the diagonal model $[\lambda\mathbf{B}]$ is the model chosen with EDDA method. And, in both cases, the committee of methods technique proposes combining that model to the spherical model $[\lambda\mathbf{I}]$. The resulting shrunk model gives somewhat better predictions than the diagonal model alone.

Haberman dataset describes survival of women who had undergone surgery to remove breast cancer. **Haemo** illustrates the presence or absence of haemophilia on women. For those two datasets, EDDA strategy is slightly better than the application of committee of methods. In the other hand, for **Crabs** dataset which describes the morphology of males and females of two species of Australian crabs and for the famous Fisher dataset **Iris**, both EDDA and committee of methods lead to the same misclassification error. For two of the six examples, the **Crabs** and **Haemo** datasets, the committee of methods technique, lead to a single model, the linear discriminant analysis model, and for all other datasets a combination of models was selected.

Table 5: Model and ME for each dataset using the penalised likelihood ratios technique and EDDA.

Dataset	Penalised likelihood		EDDA	
	Model	ME	Model	ME
Bupa	$[\lambda \mathbf{B}_k]$	.4000	$[\lambda \mathbf{B}]$	.4000
Crabs	$[\lambda \mathbf{DAD}^T]$	.5000	$[\lambda \mathbf{I}]$	.0500
Haberman	$[\lambda \mathbf{B}_k]$	.2516	$[\lambda \mathbf{B}_k]$	.2516
Haemo	$.32\,[\lambda \mathbf{DAD}^T] + .68\,[\lambda_k \mathbf{DAD}^T]$	.1600	$[\lambda \mathbf{DAD}^T]$	.1467
Iris	$[\lambda \mathbf{B}]$	.0400	$[\lambda \mathbf{B}]$	.0400
Thyroid	$[\lambda \mathbf{B}]$	.0977	$[\lambda \mathbf{B}]$	.0977

Using the penalised likelihood ratios technique did not produce improved performances on those datasets. The only case where it did not select a single model, for dataset **Haemo**, it provided a slightly higher misclassification error rate.

The discrete case

Since our samples are small the performance of the serial combining methods were evaluated by v-fold cross validation(ME). In Table 6, ME obtained on dataset **Psychological Data in older people** using the committee of methods technique and single models are compared. The performances of the classifiers have been assessed with half-sampling (two-fold cross validation error rate). Group prior probabilities were assumed to be equal, $\pi_k = .5 \;\; (k = 1, 2)$.

Table 6: Estimated error rate (half-sampling) and parameters values for the **Psychological Data in older people**.

	FOIM	FMM	KERNEL	C. MET.	C. MET.
Half-sampling	.30	.41	.32	.25	.25
λ		1.00	.95	1.00	.95
β				.555	.493

The goal of the present study is to explore the impact of playing with pets on psychological well-being among older people [11]. So, the two groups are constituted by 40 aged persons who have pets (group G_1) and 40 aged persons who don't have pets (group G_2).

Remark that this dataset is not very sparse ($2^6 = 64$ states and 80 observations) but, even so, the lowest error rate has been obtained with the committee of methods. The estimation obtained for β, through this strategy, is quite stable, producing a really intermediate model between the full multinomial model and the first order independence model. Also note that this approach seems to be no sensitive to the sparseness problem and so there is no need to smooth of the observed frequencies ($\lambda = 1$). On the basis of this study we can conclude that the involvement of playing with pets among older people can contribute for psychological well-being and thus, perhaps, for a successful ageing.

The numerical experiments performed for the model CMET on simulated binary data showed that good performances can be expected in a setting for which sample sizes are small or very small and population structures are identical in the two classes.

In Table 7, ME using the integrated likelihood ratio techniques and the single models have been compared on dataset **Medical Data**. In that case ME is the five-fold cross validation error rate of compared classifiers. Group prior probabilities were assumed to be equal, $\pi_k = .5 \ (k = 1, 2)$.

Table 7: Estimated error rate with five-fold cross-validation and parameters values for the **Medical Data**.

	FOIM	FMM	KERNEL	INT. LIK.	INT. LIK.
Five-fold cross-vali.	.45	.55	.55	.45	.45
λ		1.00	.95	1.00	.95
β				.832	.985

In this study, the goal is to predict the type of icterus, since it's not easy to make a diagnosis on the basis of liver disorders. Integrated likelihood ratio technique and FOIM provide the same performance for this dataset. The numerical experiments performed for this strategy on simulated binary data have shown that good performances can be expected with this technique in a moderate or large sample setting ([34]). In this small dataset setting (20 patients) it is no surprising that this method does not improve the performance since it involves the evaluation of an additional parameter β.

5.2. Assessing the performance of hierarchical combining

The continuous case

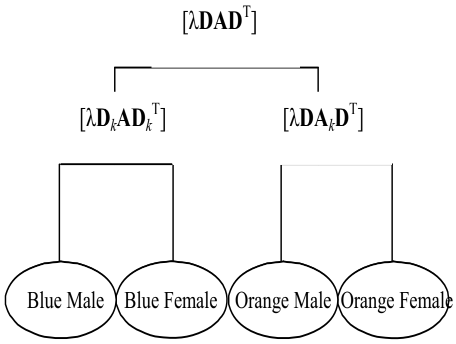
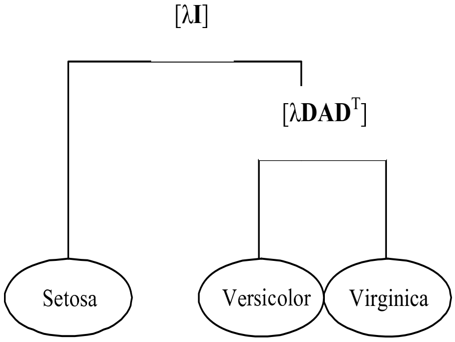
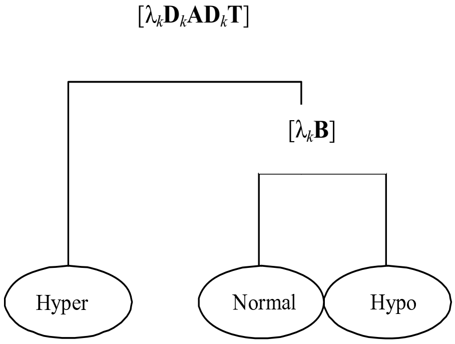
Hierarchical combining concerns only datasets with more than two groups. It has been assessed on **Crabs**, **Iris** and **Thyroid** datasets. All the fourteen models of EDDA were employed to get the hierarchical model. Hierarchical combining and EDDA methods are compared in Table 8.

Hierarchical combining concerns only datasets with more than two groups. It has been assessed on **Crabs**, **Iris** and **Thyroid** datasets. All the fourteen models of EDDA were employed to get the hierarchical model. Hierarchical combining and EDDA methods are compared in Table 8. As it can be seen from Table 8, the classification error rates of hierarchical methods and EDDA are quite similar. Here the interest of hierarchical coupling lies essentially in its ability to choose different models at each step of the classification procedure. Thus it can provide more subtle and interpretable results. For instance, for **Iris** dataset, it shows at a glance that the Setosa group can be easily separated from the two other groups with the simplest model $\left[\lambda \mathbf{I}\right]$. On the contrary, for **Thyroid** dataset, it appears that separating the "hyper" group from the other groups needs a more complex model than separating the normal group from the "hypo" group.

Hierarchical coupling model for **Crabs** dataset is also appealing. At the first level, the linear model $\left[\lambda \mathbf{DAD}^T\right]$ splits the Blue and Orange species. At the second level, males and females are separated inside each species. For Blue crabs, hierarchical coupling selects an elliptical model allowing for class of males and class of females to have different orientations $\left[\lambda \mathbf{D}_k \mathbf{AD}_k^T\right]$. For Orange Crabs, an elliptical model $\left[\lambda \mathbf{DA}_k \mathbf{D}^T\right]$ is preferred which differentiates the shape of males and females classes.

In contrast with EDDA strategy which selects $\left[\lambda \mathbf{D}_k \mathbf{AD}_k^T\right]$ for separate the four groups, hierarchical coupling is less strict, proposing more adequate models at the different levels. Only Blue males and females need the $\left[\lambda\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T\right]$ model to be separated, less complex models being proposed to distinguish other groups.

Table 8: Model and ME for each dataset using the hierarchical coupling technique and EDDA.

Dataset	Hierarchical coupling		EDDA	
	Model	ME	Model	ME
Crabs	 $[\lambda\mathbf{DAD}^T]$ $[\lambda\mathbf{D}_k\mathbf{AD}_k^T]$ $[\lambda\mathbf{DA}_k\mathbf{D}^T]$ Blue Male Blue Female Orange Male Orange Female	.045	$\left[\lambda\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T\right]$	.045
Iris	 $[\lambda\mathbf{I}]$ $[\lambda\mathbf{DAD}^T]$ Setosa Versicolor Virginica	.02	$\left[\lambda\mathbf{D}\mathbf{A}\mathbf{D}^T\right]$	.02
Thyroid	 $[\lambda_k\mathbf{D}_k\mathbf{AD}_k\mathbf{T}]$ $[\lambda_k\mathbf{B}]$ Hyper Normal Hypo	.0372	$\left[\lambda_k\mathbf{B}\right]$	.0326

The discrete case

For the **Psychological Data** the misclassification error is assessed by half-sampling. For the **Psychological Counselling Career Data** it is assessed from a test sample. A training sample of 200 students was drawn at random and the rest of the dataset constituted the test sample. Table 9 summarises the results of the four methods for these datasets and the coefficients of the combination obtained in each level of the tree.

The **Psychological Counselling Career Data** set consists of 600 students of the 1st and 2nd forms of four licenciature's degree: Biology (B), Psychology (P), Language and Literature (LL) and Engineering (E). The aim of the study is to know if those four groups of student are different regarding their Career Information.

For the **Psychological Counselling Career Data** the first decomposition chosen by hierarchical coupling for the several scales, suggest that Biology students are different from the other students in what concerns the definition of a clear and stable picture of their goals and interests, Engineering students revealing a distinct need for vocational information from the other students; and the students of odd groups show individually perceived external obstacles or limitations in pursuing occupational goals different from the students of even groups.

Remark that this dataset is not very sparse ($2^6 = 64$ or $2^4 = 16$ states and 200 observations), but again the hierarchical combining method using the integrated likelihood (HIER/IL) or committee of methods (HIER/CM) provides markedly the lowest misclassification error rate. The results of the hierarchical coupling provide markedly the lowest test estimates of the misclassification risk for all scales. However, HIER performs poorly for the Barriers scale.

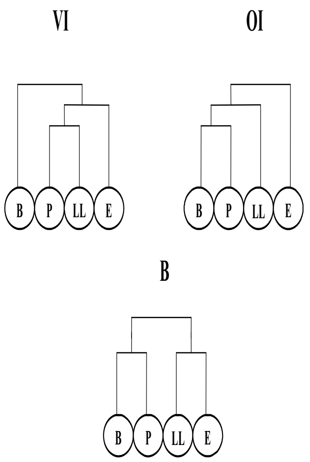
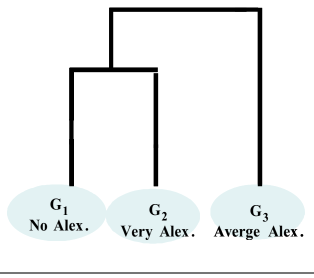
We noted that in some situations, particularly when the groups have very different sizes, usual methods and even the HIER method perform poorly. Moreover, the choice of the decomposition at each level of the tree may be unrealistic. Therefore, new developments on the hierarchical coupling approach are required in such a situation and this is a perspective for future research on this method.

The **Psychological Data** set consists of 34 dermatology's patients divided into three groups — Nonalexithymics Group (G_1), Alexithymics Group (G_2), Intermediate Group (G_3) — according to the value obtained in a psychological test (TAS-20: Twenty Item Toronto Alexithymia Scale) conceived to evaluate the presence of alexithymia[1]. The goal of the study is to evaluate how alexithymia influences personality characteristics (evaluated by another psychological test — Rorschach test).

[1] Alexithymia means "no words to express emotions".

For the **Psychological Data** the first decomposition chosen by hierarchical coupling, suggests that the union of the extremes groups forms a class well-separated from the intermediate group, since these subjects obtained balanced scores. Since the dataset is very sparse ($2^6 = 64$ states and only 17 observations) the hierarchical combining method using committee of methods (HIER/CM) provides the lowest estimated error rate.

Table 9: Model and ME for two datasets using the hierarchical coupling technique.

Dataset	Hierarchical coupling	Model	ME	λ	β		
					1st	2nd	3rd
Psycho-logical Coun-selling Career Data		VI Scale					
		FOIM	.69	1			
		FMM	.75	1			
		KERNEL	.73	.99			
		HIER/CM	.49	1	.51	.52	.53
		HIER/CM	.49	.99	.47	.47	.48
		HIER/IL	.38	1	.98	.99	1
		HIER/IL	.38	.99	1	1	1
		OI Scale					
		FOIM	.66	1			
		FMM	.66	1			
		KERNEL	.65	.99			
		HIER/CM	.45	1	.50	.51	.52
		HIER/CM	.46	.99	.48	.49	.49
		HIER/IL	.41	1	0	≈ 0	≈ 0
		HIER/IL	.38	.99	0	.02	1
		B Scale					
		FOIM	.66	1			
		FMM	.66	1			
		KERNEL	.65	.99			
		HIER/CM	.50	1	.50	.52	.50
		HIER/CM	.50	.99	.49	.49	.49
		HIER/IL	.52	1	.99	.99	1
		HIER/IL	.52	.99	1	1	1
Psycho-logical Data					1st	2nd	
		FOIM	.53	1			
		FMM	.71	1			
		KERNEL	.65	.99			
		HIER/CM	.29	1	.52	.55	
		HIER/CM	.29	.99	.47	.50	
		HIER/IL	.35	1	.18	.44	
		HIER/IL	.35	.99	.53	.78	

These results are in accordance with the numerical experiments performed for CM and IL strategies on simulated binary data that have shown that good performances can be expected with CM technique in a small or very small sample setting and with IL technique in moderate or large sample setting.

6. COMPUTER PROGRAMS

The efficiency of the combining approaches presented in this paper has been investigated on both real and simulated data. The computer programs realizing these combining approaches were implemented by the authors and are available from them.

The continuous case

All computer programs for the continuous case are written in Matlab® code. The different routines are structured as follows:

- **EDDA** — estimates all EDDA models and the leave-one-out cross validated misclassification error of each model;
- **COMMITTEE** — estimates the serial combined model by a committee of methods strategy;
- **SERIAL** — estimates the serial combined model by a penalized likelihood strategy;
- **HIERARCHICAL** — evaluates the combination of the models for all possible two class of groups. It calculates the leave-one-out cross validated misclassification error of each solution and builds the tree representation.

Run time execution is about five time more important for hierarchical coupling method than for serial combining method. It means that, for most applications, it remains a reasonable method.

The discrete case

The computer programs implemented for the discrete case use FORTRAN® 77 Language according to Microsoft FORTRAN Optimizing Compiler Version 5.0 and they use a structure in three main routines:

- **GESTAO** — determines the group conditional probabilities associated to the full multinomial model (FMM) and to the first-order independence model (FOIM) and their estimative by cross validation;
- **CALFA** — determines the combining coefficient according to the chosen combining strategy;
- **CRULE** — builds the new combining model and determines the error rate evaluated on a test sample.

For the **hierarchical combining**, an additional routine is implemented:

- **HIERQ** — builds the hierarchical binary tree, using the basic affinity coefficient.

After the selection of the two classes of groups have been chosen at each level of the binary tree, the combining model is chosen by minimizing the error rate evaluated on a test sample, using routines **GESTAO**, **CALFA** and **CRULE**.

Finally, it can be noticed that the run time execution for the hierarchical combining is quite similar to that of the serial combining in the $K=3$ group case. Otherwise, when $K>3$, the run time execution for the hierarchical combining triplicate or even more, due to the necessary reorganization of the groups for the evaluation of the basic affinity coefficient for all possible combination of couples of groups. However, the computational time for hierarchical combining remains quite reasonable and cannot be regarded as a drawback of this approach.

7. DISCUSSION

It is worth noticing that the combining methods that were considered in this paper are of different nature than other combining or ensemble methods. For example, Bagging and Boosting methods which are very efficient to improve unstable classifiers are committee-based approaches in which a single classification algorithm is applied to repeatedly modified versions of the data ([7], [8], [12], [17]-chapter 10). On the contrary the combing methods we considered are combining several methods but do not modified the weights of the data. On an other hand, the CRUISE ([19]) and QUEST ([22]) methods are classification tree algorithms different of the hierarchical combining methods we considered because the tree we designed is not a classification tree.

Many combining methods of classification have been considered in different contexts from a practical point of view. The main conclusions of this comparative experimental study are the following. Convex combining appears to be disappointing in the continuous case. In that case, at best, they lead to the same error rate obtained with the better single model. Moreover, they often prefer a single model to a combination of several models. Convex combining appears to be more efficient to propose a good compromise between FMM and FOIM models in discrete data context. Maybe the reason for this more satisfactory behaviour is that FMM and FOIM are quite different models.

On the contrary hierarchical coupling seems to be a promising technique of combining classification methods when more than two groups are to be classified. In different contexts, hierarchical coupling leads to a substantial improvement of the misclassification error rate and its easily interpretable representation is appealing. It provides original and parsimonious classification rules. An interesting perspective would be to explore all possible hierarchical coupling solutions. This is feasible when the number of groups is less than five. Otherwise, a *branch and bound* algorithm could be considered in order to search for the optimal tree solution in a reasonable time.

Finally, it can be noticed that there is a huge literature on combining models. For instance Bayesian Model Averaging (BMA) (see [23] or [29], among many others) has received a lot of attention. However, the practical implementation of Bayesian Model Averaging is far from being simple especially in the continuous case. Finally, we want to cite the interesting theoretical study of Yang ([37]) which proves that combining models cannot be expected to outperform an optimal single method for large samples.

REFERENCES

[1] Akaike, H. (1974). A new look at statistical model identification, *IEEE Transactions on Automatic Control*, **AU-19**, 716–722.

[2] Bacelar-Nicolau, H. (1985). The affinity coefficient in cluster analysis, *Methods on Operations Research*, **53**, 507–512.

[3] Bensmail, H. and Celeux, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition, *Journal of The American Statistical Association*, **91**, 716–722.

[4] Bishop, C. (1995). *Neural Networks for Pattern Recognition*, Oxford University Press London.

[5] Blake, C. and Merz, C. (1998). *UCI repository of machine learning databases*, http://www.ics.uci.edu/~mlearn/MLRepository.html, Dept. of Information and Computer Sciences Irvine University of California.

[6] Breiman, L. (1995). Stacked Regression, *Machine Learning*, **24**, 49–64.

[7] Breiman, L. (1996). Bagging predictors, *Machine Learning*, **26**(2), 123–140.

[8] Breiman, L. (1998). Arcing classifiers, *The Annals of Statistics*, **26**, 801–849.

[9] Brito, I. and Celeux, G. (2000). *Discriminant analysis by hierarchical coupling in EDDA context.* In "Proceedings of the 7$^{\text{th}}$ Conference of the International Federation of Classification Societies, IFCS-2000" (J. Jansen, Ed.), Springer-Verlag.

[10] Brito, I. (2002). *Combinaison de modles en analyse discriminante dans un contexte gaussien*, PhD Thesis, Université Joseph Fourier, Grenoble.

[11] Dourado, S.; Mohan, R.; Vieira, A.; Sousa Ferreira, A. and Duarte Silva, M.E. (2003). *Animais de estimação e bem-estar em idosos.* In "Resumos do V Simpósio Nacional de Investigação em Psicologia", Edições Associação Portuguesa de Psicologia.

[12] Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, **55**(1), 119–139.

[13] Friedman, J.H. (1996). *Another Approach to Polychotomous Classification*, Technical Report, Stanford University.

[14] Goldstein, M. and Dillon, W. (1978). *Discrete Discriminant Analysis*, Wiley and Sons.

[15] Habbema, J.; Hermans, J. and van den Broek, K. (1974). *A stepwise discriminant analysis program using density estimation.* In "Proceedings of Computational Statistics, Compstat 1974", Physica-Verlag, 101–110.

[16] Hand, D. (1982). *Kernel Discriminant Analysis*, Research Studies Press, Wiley.

[17] Hastie, T.; Tibshirani, R. and Friedman, J. (2000). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer-Verlag.

[18] Kass, R. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses with large samples, *Journal of The American Statistical Association*, **90**, 928–934.

[19] Kim, H. and Loh, W.-Y. (2001). Classification trees with unbiased multiway splits, *Journal of The American Statistical Association*, **96**, 589–604.

[20] LeBlanc, M. and Tibshirani, R. (1996). Combining Estimates in Regression and Classification, *Journal of The American Statistical Association*, **91**, 1641–1650.

[21] Lima, M.R. (1998). *Orientação e Desenvolvimento da Carreira em Estudantes Universitários*, PhD Thesis (in Portuguese), Univ. de Lisboa.

[22] Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees, *Statistica Sinica*, **7**, 814–840.

[23] Madigan, D.; Raftery, A. and Hoeting, J. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window, *Journal of The American Statistical Association*, **89**, 1535–1549.

[24] Matusita, K. (1955). Decision rules based on distances for problems of fit, two samples and estimation, *Annals of the Institute of Statistical Mathematics*, **26**(4), 631–640.

[25] Merz, C. and Pazzani, M. (1999). A principal component approach to combining regressions estimates, *Machine Learning*, **36**, 9–32.

[26] Oxford University Machine Learning Repository Statlib (1996). http://lib.stat.cmu.edu, Depart. of Statistics at Carnegie Mellon University.

[27] Perrone, M. and Cooper, L. (1973). *When networks disagree: ensemble methods for hybrid neural networks.* In "Artificial neural networks for speech and vision" (R. Mammone, Ed.), Chapman & Hall, 126–142.

[28] Prazeres, N.L. (1996). *Ensaio de um Estudo sobre Alexitimia com o Rorschach e a Escala de Alexitimia de Toronto (TAS-20)*, Master Thesis (in Portuguese), Univ. de Lisboa.

[29] Raftery, A. (1996). Approximate Bayes Factors and Accounting for Model Uncertainty in Generalised Linear Models, *Biometrika*, **83**, 251–266.

[30] Romeder, J. (1973). *Méthodes et Programmes d'Analyse Discriminante*, Dunod, Paris.

[31] Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461–464.

[32] Sousa Ferreira, A.; Celeux, G. and Bacelar-Nicolau, H. (1999). *Combining Models in Discrete Discriminant Analysis by a Hierarchical Coupling Approach.* In "Applied Stochastic Models and Data Analysis, ASMDA 99" (H. Bacelar-Nicolau, F. Costa Nicolau, J. Janssen, Eds.), INE, 159–164.

[33] Sousa Ferreira, A. (2000). *Combining Models in Discrete Discriminant Analysis*, PhD Thesis (in Portuguese), Univ. Nova de Lisboa.

[34] Sousa Ferreira, A.; Celeux, G. and Bacelar-Nicolau, H. (2000). *Discrete Discriminant Analysis: the Performance of Combining Models by a Hierarchical Coupling Approach.* In "Data Analysis, Classification and Related Methods" (Kiers, Rasson, Groenen, Schader, Eds.), Springer, 181–186.

[35] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society*, **B 36**, 111–147.

[36] Wolpert, D. (1992). Stacked generalization, *Neural Networks*, **5**, 241–259.

[37] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation, *Biometrika*, **92**, 937–950.

REVSTAT – Statistical Journal
Volume 4, Number 3, November 2006, 227–247

PEAKS OVER RANDOM THRESHOLD METHODOLOGY FOR TAIL INDEX AND HIGH QUANTILE ESTIMATION

Authors: Paulo Araújo Santos
– ESGS, Instituto Politécnico de Santarém, Portugal
paulo.santos@esg.ipsantarem.pt

M. Isabel Fraga Alves
– CEAUL and DEIO (FCUL), Universidade de Lisboa, Portugal
isabel.alves@fc.ul.pt

M. Ivette Gomes
– CEAUL and DEIO (FCUL), Universidade de Lisboa, Portugal
ivette.gomes@fc.ul.pt

Received: May 2006 Accepted: September 2006

Abstract:

- In this paper we present a class of semi-parametric *high quantile* estimators which enjoy a desirable property in the presence of linear transformations of the data. Such a feature is in accordance with the empirical counterpart of the theoretical linearity of a quantile χ_p: $\chi_p(\delta X + \lambda) = \delta\chi_p(X) + \lambda$, for any real λ and positive δ. This class of estimators is based on the sample of excesses over a random threshold, originating what we denominate *PORT* (*Peaks Over Random Threshold*) methodology. We prove consistency and asymptotic normality of two high quantile estimators in this class, associated with the *PORT*-estimators for the tail index. The exact performance of the new tail index and quantile *PORT*-estimators is compared with the original semi-parametric estimators, through a simulation study.

Key-Words:

- *heavy tails; high quantiles; semi-parametric estimation; linear property; sample of excesses.*

AMS Subject Classification:

- 62G32, 62E20.

1. INTRODUCTION

In this paper we deal with semi-parametric estimators of the *tail index* γ and *high quantiles* χ_p, which enjoy desirable properties in the presence of linear transformations of the available data. We recall that a *high quantile* is a value exceeded with a small probability. Formally, we denote by F the heavy-tailed distribution function (d.f.) of a random variable (r.v.) X, the common d.f. of the i.i.d. sample $\underline{X} := \{X_i\}_{i=1}^n$, for which the high quantile

$$\chi_p(X) := F^{\leftarrow}(1-p)\,, \qquad p = p_n \to 0, \quad \text{as } n\to\infty\,, \qquad n\,p_n \to c \geq 0\;, \tag{1.1}$$

has to be estimated. Here $F^{\leftarrow}(t) := \inf\{x\colon\ F(x) \geq t\}$ denotes the generalized inverse function of F.

We consider estimators based on the $k+1$ top order statistics (o.s.), $X_{n:n} \geq \cdots \geq X_{n-k:n}$, where $X_{n-k:n}$ is an intermediate o.s., i.e., k is an intermediate sequence of integers such that

$$k = k_n \to \infty\,, \quad k_n/n \to 0, \qquad \text{as } n \to \infty\;. \tag{1.2}$$

We assume that we are working in a context of heavy tails, i.e., $\gamma > 0$ in the extreme value distribution

$$G_\gamma(x) = \begin{cases} \exp\bigl\{-(1+\gamma\,x)^{-1/\gamma}\bigr\}, & 1+\gamma\,x > 0, \quad \gamma \neq 0 \\ \exp\bigl(-e^{-x}\bigr), & x \in \mathbb{R}, \quad \gamma = 0\;, \end{cases} \tag{1.3}$$

the non-degenerate d.f. to which the maximum $X_{n:n}$ is attracted, after a suitable linear normalization. When this happens we say that the d.f. F is in the Fréchet domain of attraction and we write $F \in D(G_\gamma)_{\gamma>0}$.

The paper is developed under the first order regular variation condition, which allows the extension of the empirical d.f. beyond the range of the available data, assuming a polynomial decay of the tail. This condition can be expressed by

$$F \in D(G_\gamma)_{\gamma>0} \qquad \text{iff} \qquad \overline{F} := 1-F \in RV_{-1/\gamma} \qquad \text{iff} \qquad U \in RV_\gamma\;, \tag{1.4}$$

where U is the quantile function defined as $U(t) := F^{\leftarrow}(1-1/t)$, $t \geq 1$; the notation RV_α stands for the class of regularly functions at infinity with index of regular variation α, i.e., positive measurable functions h such that $\lim\limits_{t\to\infty} h(tx)/h(t) = x^\alpha$, for all $x > 0$.

It is interesting to note that the p-quantile can be expressed as $\chi_{p_n} = U(1/p_n)$.

To get asymptotic normality of estimators of parameters of extreme events, it is usual to assume the following extra second regular variation condition, that involves a non-positive parameter ρ:

$$\lim_{t\to\infty} \frac{U(tx)/U(t) - x^\gamma}{A(t)} = x^\gamma\, \frac{x^\rho - 1}{\rho}\ , \tag{1.5}$$

for all $x > 0$, where A is a suitably chosen function of constant sign near infinity. Then, $|A| \in RV_\rho$ and ρ is called the second order parameter (Geluk and de Haan, 1987). For the strict Pareto model, with tail function $\overline{F}(x) = (x/C)^{-1/\gamma}$ and quantile function $U(t) = Ct^\gamma$, $U(tx)/U(t) - x^\gamma \equiv 0$. We then consider that (1.5) holds with $A(t) \equiv 0$.

More restrictively, we might consider that F belonged to the wide class of Hall [11], that is, the associated quantile function U satisfies

$$U(t) = Ct^\gamma\big(1 + Dt^\rho + o(t^\rho)\big)\ , \qquad \rho<0,\quad C>0,\quad D\in\mathbb{R},\quad \text{as } t\to\infty\ , \tag{1.6}$$

or equivalently, (1.5) holds, with $A(t) = D\rho t^\rho$. The strict Pareto model appears when both D and the remainder term $o(t^\rho)$ are null.

Returning to the problem of high quantile estimation, we recall the classical semi-parametric Weissman-type estimator of χ_{p_n} (Weissman, 1978),

$$\widehat{\chi}_{p_n} = \widehat{\chi}_{p_n}(\underline{X}) = X_{n-k_n:n}\left(\frac{k_n}{np_n}\right)^{\hat\gamma_n}\ , \tag{1.7}$$

with $\hat\gamma_n = \hat\gamma_n(\underline{X})$ some consistent estimator of the tail parameter γ.

In the classical approach one considers for $\hat\gamma_n$ the well known Hill estimator (Hill, 1975),

$$\hat\gamma_n^H = \hat\gamma_n^H(\underline{X}) = \frac{1}{k_n}\sum_{j=1}^{k_n} \log\frac{X_{n-j+1:n}}{X_{n-k_n:n}}\ , \tag{1.8}$$

or the Moment estimator (Dekkers et al., 1989),

$$\hat\gamma_n^M = \hat\gamma_n^M(\underline{X}) = M_n^{(1)} + 1 - \frac{1}{2}\left\{1 - \frac{\big(M_n^{(1)}\big)^2}{M_n^{(2)}}\right\}^{-1}\ , \tag{1.9}$$

with $M_n^{(r)}$, the r-Moment of the log-excesses, defined by

$$M_n^{(r)} = M_n^{(r)}(\underline{X}) = \frac{1}{k_n}\sum_{j=1}^{k_n}\left(\log\frac{X_{n-j+1:n}}{X_{n-k_n:n}}\right)^r\ , \qquad r = 1, 2\ . \tag{1.10}$$

We use the following notation:

$$\widehat{\chi}_{p_n}^H = X_{n-k_n:n}\left(\frac{k_n}{np_n}\right)^{\hat\gamma_n^H}\ , \qquad \widehat{\chi}_{p_n}^M = X_{n-k_n:n}\left(\frac{k_n}{np_n}\right)^{\hat\gamma_n^M}\ . \tag{1.11}$$

Finally, we explain the question that motivated this paper. It is well known that scale transformations to the data do not interfere with the stochastic behaviour of the tail index estimators (1.8) and (1.9), i.e., we can say that they enjoy scale invariance. The incorporation of (1.8) or (1.9) in the Weissman-type estimator in (1.7), allows us to obtain the following desirable exact property for quantile estimators: for any real positive δ,

$$\widehat{\chi}_{p_n}(\delta\underline{X}) \;=\; \delta X_{n-k_n:n}\left(\frac{k_n}{np_n}\right)^{\hat{\gamma}_n} \;=\; \delta\,\widehat{\chi}_{p_n}(\underline{X}) \;. \tag{1.12}$$

But we want a similar linear property in the case of location transformations to the data, $Z_j := X_j + \lambda$, $j = 1, ..., n$, for any real λ. That is, our main goal is that, for the transformed data $\underline{Z} := \{Z_j\}_{j=1}^n$, the quantile estimator satisfies

$$\widehat{\chi}_{p_n}(\underline{Z}) \;=\; \widehat{\chi}_{p_n}(\underline{X}) + \lambda \;. \tag{1.13}$$

Altogether, this represents the empirical counterpart of the following theoretical linear property for quantiles,

$$\chi_p(\delta X + \lambda) \;=\; \delta\chi_p(X) + \lambda\,, \qquad \text{for any real } \lambda \text{ and real positive } \delta\,. \tag{1.14}$$

Here we present a class of high quantile-estimators for which (1.12) and (1.13) hold exactly, pursuing the empirical counterpart of the theoretical linear property (1.14). For a simple modification of (1.7) that enjoys (1.13) approximately, see Fraga Alves and Araújo Santos (2004). For the use of reduced bias tail index estimation in high quantile estimation for heavy tails, see Gomes and Figueiredo (2003), Matthys and Beirlant (2003) and Gomes and Pestana (2005), where the second order reduced bias tail index estimator in Caeiro *et al.* (2005) is used for the estimation of the *Value at Risk*.

1.1. The class of high quantile estimators under study

The class of estimators suggested here is function of a sample of excesses over a random threshold $X_{n_q:n}$,

$$\underline{X}^{(q)} := \big(X_{n:n} - X_{n_q:n},\; X_{n-1:n} - X_{n_q:n},\; ...,\; X_{n_q+1:n} - X_{n_q:n}\big) \;, \tag{1.15}$$

where $n_q := [nq] + 1$, with:

- $0 < q < 1$, for d.f.'s with finite or infinite left endpoint $x_F := \inf\{x\colon F(x) > 0\}$ (*the random threshold is an empirical quantile*);

- $q = 0$, for d.f.'s with finite left endpoint x_F (*the random threshold is the minimum*).

A statistical inference method based on the sample of excesses $\underline{X}^{(q)}$ defined in (1.15) will be called a *PORT*-methodology, with *PORT* standing for*Peaks Over Random Threshold*. We propose the following *PORT-Weissman* estimators:

$$\hat{\chi}^{(q)}_{p_n} = (X_{n-k_n:n} - X_{n_q:n}) \left(\frac{k_n}{n p_n}\right)^{\hat{\gamma}^{(q)}_n} + X_{n_q:n} \ , \tag{1.16}$$

where $\hat{\gamma}^{(q)}_n$ is any consistent estimator of the tail parameter γ, made location/scale invariant by using the transformed sample $\underline{X}^{(q)}$. Indeed, the incorporation in the Adapted-Weissman estimator in (1.16), of tail index estimators, as function of the sample of excesses, allows us to obtain exactly the linear property (1.13).

1.2. Shifts in a Pareto model

To illustrate the behaviour of the new quantile estimators in (1.16), we shall first consider a parent X from a *Pareto*$(\gamma, \lambda, \delta)$,

$$F_{\gamma,\lambda,\delta}(z) = 1 - \left(\frac{z-\lambda}{\delta}\right)^{-1/\gamma}, \qquad z > \lambda + \delta, \quad \delta > 0 \ , \tag{1.17}$$

with $\lambda = 0$ and $\gamma = \delta = 1$. Let us assume that we want to estimate an upper $p = p_n = \frac{1}{n}$-quantile in a sample of size $n = 500$. Then, we want to estimate the parameter $\chi_p(X) = 500$. If we induce a shift $\lambda = 100$ to our data, we would obviously like our estimates to approach $\chi_p(X + 100) = 600$.

In Figure 1 we plot, for the *Pareto*$(\lambda, 1, 1)$ parents, with $\lambda = 0$ and $\lambda = 100$ and for $q = 0$ in (1.15), the simulated mean values of the *Weissman* and *PORT-Weissman* quantile estimators based on the Hill, denoted $\hat{\chi}^H_p$ and $\hat{\chi}^{H(q)}_p$, respectively. These mean values are based on $N = 500$ replications, for each value k, $5 \le k \le 500$, from the above mentioned models.

Similarly to the Hill horror plots (Resnick, 1997), associated to slowly varying functions $L_U(t) = t^{-\gamma}\, U(t)$, we also obtain here Weissman–Hill horror plots whenever we induce a shift in the simple standard Pareto model. Indeed, for a standard Pareto model ($\lambda = 0$ in (1.17)), Weissman type estimators in (1.7) perform reasonably well, with $\hat{\gamma}_n = \hat{\gamma}^H_n$. However, a small shift in the data may lead to disastrous results, even in this simple and specific case. For the *PORT-Weissman* estimates, the shift in the quantile estimates is equal to the shift induced in the data, a sensible property of quantile estimates. Figure 1 also illustrates how serious can be the consequences to the sample paths of the classical high quantile estimators, when we induce a shift in the data, as suggested in Drees (2003). We may indeed be led to dangerous misleading conclusions, like a systematic underestimation, for instance, mainly due to "stable zones" far away of the target quantile to be estimated.

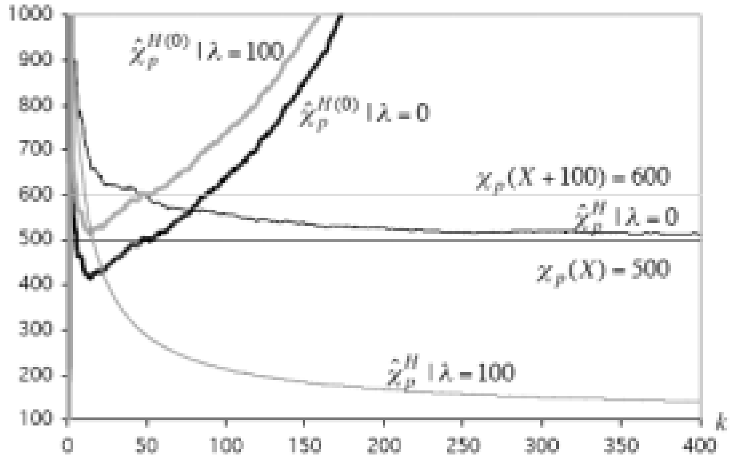


Figure 1: Mean values of $\widehat{\chi}^H_{p_n}$ and $\widehat{\chi}^{H(0)}_{p_n}$, $p_n = 0.002$ for samples of size $n = 500$ from a *Pareto*$(1, 0, 1)$ parent (target quantile $\chi_{p_n} = 500$) and from the *Pareto*$(1, 100, 1)$ (target quantile $\chi_{p_n} = 600$).

1.3. Scope of the paper

As far as we know, no systematic study has been done concerning asymptotic and exact properties of semi-parametric methodologies for tail index and high quantile estimation, using the transformed sample in (1.15). Somehow related with this subject, Gomes and Oliveira (2003), in a context of regularly varying tails, suggested a simple generalization of the classical Hill estimator associated to artificially shifted data. The shift imposed to the data is deterministic, with the aim of reducing the main component of the bias of Hill's estimator, getting thus estimates with stable sample paths around the target value. A preliminary study has also been carried out, by the same authors, replacing the artificial deterministic shift by a random shift, which in practice represents a transformation of the original data through the subtraction of the smallest observation, added by one, whenever we are aware that the underlying heavy-tailed model has a finite left endpoint.

With the purpose of tail index and high quantile estimation there is, in our opinion, a gap in the literature regarding classical semi-parametric estimation methodologies adapted for shifted data, the main topic of this paper.

In Section 2, we derive asymptotic properties for the adapted Hill and Moment estimators, as functions of the sample of excesses (1.15). In Section 3, we propose two estimators for χ_p that belong to the class (1.16) and prove their asymptotic normality. In Section 4, and through simulation experiments, we compare the performance of the new estimators with the classical ones. Finally, in Section 5, we draw some concluding remarks.

2. TAIL INDEX PORT-ESTIMATORS

For the classical Hill and Moment estimators, we know that for any intermediate sequence k as in (1.2) and under the validity of the second order condition in (1.5),

$$\hat{\gamma}_n^H \overset{d}{=} \gamma + \frac{\gamma}{\sqrt{k}}\, P_k^H + \frac{A(n/k)}{1-\rho}\,\bigl(1+o_p(1)\bigr) \tag{2.1}$$

and

$$\hat{\gamma}_n^M \overset{d}{=} \gamma + \frac{\sqrt{\gamma^2+1}}{\sqrt{k}}\, P_k^M + \frac{\bigl(\gamma(1-\rho)+\rho\bigr)\,A(n/k)}{\gamma(1-\rho)^2}\,\bigl(1+o_p(1)\bigr)\ , \tag{2.2}$$

where P_k^H and P_k^M are asymptotically standard normal r.v.'s.

In this section we present asymptotic results for the classical Hill estimator in (1.8) and the Moment estimator in (1.9), both based on the sample of excesses $\underline{X}^{(q)}$ in (1.15), which will be denoted respectively, by

$$\hat{\gamma}_n^{H(q)} := \hat{\gamma}_n^H\bigl(\underline{X}^{(q)}\bigr) \quad \text{and} \quad \hat{\gamma}_n^{M(q)} := \hat{\gamma}_n^M\bigl(\underline{X}^{(q)}\bigr)\,, \qquad 0 \le q < 1\ . \tag{2.3}$$

In the following, χ_q^* denotes the q-quantile of F: $F(\chi_q^*) = q$ (by convention $\chi_0^* := x_F$), so that

$$X_{n_q:n} \xrightarrow{p} \chi_q^*\,, \quad \text{as } n\to\infty\,, \qquad \text{for } 0 \le q < 1\ .$$

For the estimators in (2.3) we have the asymptotic distributional representations expressed in Theorem 2.1.

Theorem 2.1 (PORT-Hill and PORT-Moment). *For any intermediate sequence k as in* (1.2), *under the validity of the second order condition in* (1.5), *for any real q, $0 \le q < 1$, and with T generally denoting either H or M, the asymptotic distributional representation*

$$\hat{\gamma}_n^{T(q)} \overset{d}{=} \gamma + \frac{\sigma_T}{\sqrt{k}}\, P_k^T + \left(c_T\, A(n/k) + d_T\, \frac{\chi_q^*}{U(n/k)}\right)\bigl(1+o_p(1)\bigr) \tag{2.4}$$

holds, where P_k^T is an asymptotically standard normal r.v.,

$$\sigma_H^2 := \gamma^2\ , \qquad c_H := \frac{1}{1-\rho}\ , \qquad d_H := \frac{\gamma}{\gamma+1}\ , \tag{2.5}$$

$$\sigma_M^2 := \gamma^2+1\,, \quad c_M := \frac{\gamma(1-\rho)+\rho}{\gamma(1-\rho)^2} \quad \textit{and} \quad d_M := \left(\frac{\gamma}{\gamma+1}\right)^2. \tag{2.6}$$

Remark 2.1. Notice that $\sigma^2_{\scriptscriptstyle M} = \sigma^2_{\scriptscriptstyle H} + 1$, $c_{\scriptscriptstyle M} = c_{\scriptscriptstyle H} + \frac{\rho}{\gamma(1-\rho)^2}$ and $d_{\scriptscriptstyle M} = (d_{\scriptscriptstyle H})^2$. Consequently, $\sigma_{\scriptscriptstyle M} > \sigma_{\scriptscriptstyle H}$, $c_{\scriptscriptstyle M} \le c_{\scriptscriptstyle H}$ and $d_{\scriptscriptstyle M} < d_{\scriptscriptstyle H}$.

The proof of Theorem 2.1 relies on the the following Lemmas 2.1 and 2.2.

Lemma 2.1. *Let F be the d.f. of X, and assume that the associated U-quantile function satisfies the second order condition* (1.5). *Consider a deterministic shift transformation to X, defining the r.v. $X_q := X - \chi^*_q$ with d.f. $F_q(x) = F(x) + \chi^*_q$ and associated U_q-quantile function given by $U_q(t) := F^{\leftarrow}_q(1-1/t)$ $= U(t) - \chi^*_q$.*

Then U_q satisfies a second order condition similar to (1.5), *that is*

$$\lim_{t\to\infty} \frac{U_q(tx)/U_q(t) - x^\gamma}{A_q(t)} = x^\gamma \left(\frac{x^{\rho_q} - 1}{\rho_q}\right), \qquad \text{for} \;\; x > 0, \;\; \rho_{\scriptscriptstyle q} \le 0 \;, \tag{2.7}$$

with

$$\bigl(A_q(t), \rho_{\scriptscriptstyle q}\bigr) := \begin{cases} \bigl(A(t), \rho\bigr) & \text{if} \;\; \rho > -\gamma\,; \\[2ex] \left(A(t) + \dfrac{\gamma\chi^*_q}{U(t)}, -\gamma\right) & \text{if} \;\; \rho = -\gamma\,; \\[2ex] \left(\dfrac{\gamma\chi^*_q}{U(t)}, -\gamma\right) & \text{if} \;\; \rho < -\gamma\;. \end{cases} \tag{2.8}$$

Proof: Under (1.5), for $x > 0$,

$$\begin{aligned}
\frac{U_q(tx)}{U_q(t)} &= \frac{U(tx) - \chi^*_q}{U(t) - \chi^*_q} \\
&= \frac{U(tx)}{U(t)} \left\{\frac{1 - \chi^*_q/U(tx)}{1 - \chi^*_q/U(t)}\right\} \\
&= \frac{U(tx)}{U(t)} \left\{1 + \chi^*_q\, \frac{1/U(t) - 1/U(tx)}{1 - \chi^*_q/U(t)}\right\} \\
&= \frac{U(tx)}{U(t)} \left\{1 + \frac{\chi^*_q}{U(t)} \left[1 - \frac{U(t)}{U(tx)}\right] \bigl(1 + o(1)\bigr)\right\} \\
&= x^\gamma \left\{1 + \frac{x^\rho - 1}{\rho}\, A(t) \bigl(1 + o(1)\bigr)\right\} \left\{1 + \frac{\gamma\chi^*_q}{U(t)}\, \frac{x^{-\gamma} - 1}{-\gamma}\, \bigl(1 + o(1)\bigr)\right\} \\
&= x^\gamma \left\{1 + \frac{x^\rho - 1}{\rho}\, A(t) + \frac{\gamma\chi^*_q}{U(t)}\, \frac{x^{-\gamma} - 1}{-\gamma} + o\bigl(A(t)\bigr) + o\bigl(1/U(t)\bigr)\right\}.
\end{aligned}$$

Then U_q satisfies (2.7), for A_q and ρ_q defined in (2.8) and the result follows. □

Lemma 2.2. *Denote by* $M_n^{(r,q)}$ *the* $M_n^{(r)}$ *statistics in* (1.10), *as functions of the transformed sample* $\underline{X}^{(q)}$, $0 \le q < 1$ *in* (1.15)*; that is,*

$$M_n^{(r,q)} := M_n^{(r)}\big(\underline{X}^{(q)}\big) = \frac{1}{k}\sum_{j=1}^{k}\left(\log\frac{X_{n-j+1:n}-X_{n_q:n}}{X_{n-k:n}-X_{n_q:n}}\right)^r, \qquad r=1,2 \ .$$

Then, for any intermediate sequence k as in (1.2), *under the validity of the second order condition in* (1.5) *and for any real q,* $0 \le q < 1$,

$$M_n^{(r,q)} - \frac{1}{k}\sum_{j=1}^{k}\left(\log\frac{X_{n-j+1:n}-\chi_q^*}{X_{n-k:n}-\chi_q^*}\right)^r = o_p\left(\frac{1}{U(n/k)}\right), \qquad r=1,2 \ .$$

Proof: We will consider $r = 1$. Using the first order approximation $\ln(1+x) \sim x$, as $x \to 0$, together with the fact that $X_{n_q:n} = \chi_q^*\big(1+o_p(1)\big)$, we will have successively

$$\begin{aligned}
M_n^{(1,q)} - \frac{1}{k}\sum_{j=1}^{k}\log\frac{X_{n-j+1:n}-\chi_q^*}{X_{n-k:n}-\chi_q^*} &= \\
&= \frac{1}{k}\sum_{j=1}^{k}\log\frac{X_{n-j+1:n}-X_{n_q:n}}{X_{n-k:n}-X_{n_q:n}} - \log\frac{X_{n-j+1:n}-\chi_q^*}{X_{n-k:n}-\chi_q^*} \\
&= \frac{1}{k}\sum_{j=1}^{k}\log\frac{1-X_{n_q:n}/X_{n-j+1:n}}{1-X_{n_q:n}/X_{n-k:n}} - \log\frac{1-\chi_q^*/X_{n-j+1:n}}{1-\chi_q^*/X_{n-k:n}} \\
&= \frac{1}{k}\sum_{j=1}^{k}\left(\frac{X_{n_q:n}}{X_{n-k:n}} - \frac{X_{n_q:n}}{X_{n-j+1:n}} + \frac{\chi_q^*}{X_{n-j+1:n}} - \frac{\chi_q^*}{X_{n-k:n}}\right)\big(1+o_p(1)\big) \\
&= \frac{X_{n_q:n}-\chi_q^*}{X_{n-k:n}}\,\frac{1}{k}\sum_{j=1}^{k}\left(1-\frac{X_{n-k:n}}{X_{n-j+1:n}}\right)\big(1+o_p(1)\big) \\
&= \frac{o_p(1)}{X_{n-k:n}}\,\frac{1}{k}\sum_{j=1}^{k}\left(1-\frac{X_{n-k:n}}{X_{n-j+1:n}}\right)\big(1+o_p(1)\big) \, .
\end{aligned}$$

Denote by $\{Y_j\}_{j=1}^k$ i.i.d. Y standard Pareto r.v.'s, with d.f. $F_Y(y) = 1-y^{-1}$, for $y > 1$ and $\{Y_{j:k}\}_{j=1}^k$ the associated o.s.'s.

Since $X_{n-k:n} \overset{d}{=} U(Y_{n-k:n})$, with $Y_{n-k:n}$ the $(n-k)$-th o.s. associated to an i.i.d. standard Pareto sample of size n and $\big(\frac{k}{n}\big)Y_{n-k:n} \xrightarrow{p} 1$, for any intermediate sequence k, then $\frac{X_{n-k:n}}{U(n/k)} \xrightarrow{p} 1$; this together with the fact that

$$\left\{\frac{Y_{n-j+1:n}}{Y_{n-k:n}}\right\}_{j=1}^{k} \overset{d}{=} \Big\{Y_{k-j+1:k}\Big\}_{j=1}^{k}$$

allow us to write

$$
\begin{aligned}
M_n^{(1,q)} - \frac{1}{k}\sum_{j=1}^{k} \log\frac{X_{n-j+1:n}-\chi_q^*}{X_{n-k:n}-\chi_q^*} \; &= \\
&= \frac{o_p(1)}{U(Y_{n-k:n})}\,\frac{1}{k}\sum_{j=1}^{k}\left(1-\frac{U(Y_{n-k:n})}{U\Big(\frac{Y_{n-j+1:n}}{Y_{n-k:n}}\,Y_{n-k:n}\Big)}\right)\big(1+o_p(1)\big)\\
&= \frac{1}{k}\sum_{j=1}^{k}\Big(1-Y_{k-j+1:k}^{-\gamma}\Big)\,o_p\bigg(\frac{1}{U(n/k)}\bigg)\,\big(1+o_p(1)\big)\\
&= \frac{1}{k}\sum_{j=1}^{k}\Big(1-Y_j^{-\gamma}\Big)\,o_p\bigg(\frac{1}{U(n/k)}\bigg)\,\big(1+o_p(1)\big)\,.
\end{aligned}
$$

Now $E\big[Y^{-\gamma}\big]=\frac{1}{\gamma+1}$ and by the weak law of large numbers we obtain

$$
\begin{aligned}
M_n^{(1,q)} - \frac{1}{k}\sum_{j=1}^{k} \log\frac{X_{n-j+1:n}-\chi_q^*}{X_{n-k:n}-\chi_q^*} \; &= \\
&= \frac{\gamma}{\gamma+1}\,\Big(1+o_p(1/\sqrt{k})\Big)\,o_p\bigg(\frac{1}{U(n/k)}\bigg)\\
&= o_p\bigg(\frac{1}{U(n/k)}\bigg)\,.
\end{aligned}
$$

For $r=2$ steps similar to the previous ones lead us to the result. □

Remark 2.2. Note that if $q\in(0,1)$, $X_{n_q:n}-\chi_q^*=O_p(1/\sqrt{n})$ and for $r=1,2$, $\sqrt{k}\,\Big[M_n^{(r,q)}-\frac{1}{k}\sum_{j=1}^{k}\Big\{\log\frac{X_{n-j+1:n}-\chi_q^*}{X_{n-k:n}-\chi_q^*}\Big\}^r\,\Big]=O_p\Big(\sqrt{k/n}\;\frac{1}{U(n/k)}\Big)=o_p(1)$ holds.

Proof of Theorem 2.1: Taking into account Lemma 2.2

$$
\hat\gamma_n^{H(q)} = \frac{1}{k}\sum_{j=1}^{k}\log\frac{X_{n-j+1:n}-\chi_q^*}{X_{n-k:n}-\chi_q^*} + o_p\bigg(\frac{1}{U(n/k)}\bigg)\,.
$$

Now, considering the result in Lemma 2.1 and representation (2.1) adapted for the deterministic shift data from $X_q := X-\chi_q^*$ model, we obtain the following representation for *PORT*-Hill estimator

$$
\hat\gamma_n^{H(q)} \overset{d}{=} \gamma + \frac{\gamma}{\sqrt{k}}\,P_k^H + \frac{A_q(n/k)}{1-\rho_q}\,\big(1+o_p(1)\big) + o_p\bigg(\frac{1}{U(n/k)}\bigg)\,,
$$

with $A_q(t)$ provided in (2.8), and the result (2.4) follows with $T=H$.

Similarly, considering Lemmas 2.1 and 2.2 and the representation (2.2) adapted for the deterministic shift data from $X_q := X - \chi_q^*$ model, we obtain for the *PORT*-Moment estimator the representation

$$\hat{\gamma}_n^{M(q)} \overset{d}{=} \gamma + \frac{\sqrt{\gamma^2+1}}{\sqrt{k}}\, P_k^M + \frac{\Big(\gamma(1-\rho_q) + \rho_q\Big)\, A_q(n/k)}{\gamma(1-\rho_q)^2}\, \big(1 + o_p(1)\big) + o_p\left(\frac{1}{U(n/k)}\right) ,$$

and result (2.4) follows with $T = M$. □

Remark 2.3. Note that if we induce a deterministic shift λ to data X from a model $F =: F_0$, i.e., if we work with the new model $F_\lambda(x) := F_0(x-\lambda)$, the associated U-quantile function changes to $U_\lambda(t) = \lambda + \delta\, U_0(t)$. Then, as expected, (2.4) holds whenever we replace $\hat{\gamma}_n^{H(q)}$ by $\hat{\gamma}_n^H|\lambda$ (the Hill estimator associated with the shifted population with shift λ) provided that we replace χ_q^* by $-\lambda$. This topic has been handled in Gomes and Oliveira (2003), where the shift λ is regarded as a *tuning parameter* of the statistical procedure that leads to the tail index estimates. The same comments apply to the classical Moment estimator.

Corollary 2.1. *For the strict Pareto model, i.e., the model in* (1.17) *with* $\lambda = 0$ *and* $\gamma = \delta = 1$, *the distributional representations* (2.4) *holds with* $A(t)$ *replaced by 0.*

Under the conditions of Theorems 2.1 and with the notations defined in (2.5) and (2.6), the following results hold:

Corollary 2.2. *Let* μ_1 *and* μ_2 *be finite constants and let* T *generically denote either* H *or* M.

i) *For* $\gamma > -\rho$,

$$\hat{\gamma}_n^{T(q)} \overset{d}{=} \gamma + \frac{\sigma_T}{\sqrt{k}}\, P_k^T + c_T\, A(n/k)\, \big(1 + o_p(1)\big) .$$

If $\sqrt{k}\, A(n/k) \to \mu_1$, *then*

$$\sqrt{k}\Big(\hat{\gamma}_n^{T(q)} - \gamma\Big) \underset{n\to\infty}{\overset{d}{\longrightarrow}} Normal\big(\mu_1\, c_T,\, \sigma_T^2\big) .$$

ii) *For* $\gamma < -\rho$,

$$\hat{\gamma}_n^{T(q)} \overset{d}{=} \gamma + \frac{\sigma_T}{\sqrt{k}}\, P_k^T + d_T\, \frac{\chi_q^*}{U(n/k)}\, \big(1 + o_p(1)\big) .$$

If $\sqrt{k}/U(n/k) \to \mu_2$, *then*

$$\sqrt{k}\Big(\hat{\gamma}_n^{T(q)} - \gamma\Big) \underset{n\to\infty}{\overset{d}{\longrightarrow}} Normal\big(\mu_2\, d_T \chi_q^*,\, \sigma_T^2\big) .$$

iii) *For $\gamma = -\rho$,*

$$\hat\gamma_n^{T(q)} \stackrel{d}{=} \gamma + \frac{\sigma_T}{\sqrt{k}}\, P_k^T + \left[c_T\, A(n/k) + d_T\,\frac{\chi_q^*}{U(n/k)}\right] \bigl(1+o_p(1)\bigr) \ .$$

If $\sqrt{k}\, A(n/k) \to \mu_1$ *and* $\sqrt{k}/U(n/k) \to \mu_2$, *then*

$$\sqrt{k}\left(\hat\gamma_n^{T(q)} - \gamma\right) \xrightarrow[n\to\infty]{d} \mathit{Normal}\bigl(\mu_1\, c_T + \mu_2\, d_T \chi_q^*,\, \sigma_T^2\bigr) \ .$$

3. HIGH QUANTILE PORT-ESTIMATORS

On the basis of (1.16), we shall now consider the following estimators of χ_{p_n}, functions of the sample of excesses over $X_{n_q:n}$, i.e., of the sample $\underline{X}^{(q)}$ in (1.15):

$$\text{(3.1)}\qquad \widehat\chi_{p_n}^{H(q)} := \bigl(X_{n-k_n:n} - X_{n_q:n}\bigr)\left(\frac{k_n}{n p_n}\right)^{\hat\gamma_n^{H(q)}} + X_{n_q:n} \ , \qquad 0 \le q < 1 \ ,$$

$$\text{(3.2)}\qquad \widehat\chi_{p_n}^{M(q)} := \bigl(X_{n-k_n:n} - X_{n_q:n}\bigr)\left(\frac{k_n}{n p_n}\right)^{\hat\gamma_n^{M(q)}} + X_{n_q:n} \ , \qquad 0 \le q < 1 \ .$$

For these estimators we have the asymptotic distributional representations presented in Theorem 3.1.

Theorem 3.1. *In Hall's class* (1.6), *for intermediate sequences k_n that satisfy*

$$\text{(3.3)}\qquad \log(n p_n)/\sqrt{k_n} \to 0\,, \qquad \text{as} \quad n\to\infty \ ,$$

with p_n such that (1.1) *holds, then, with T denoting either H or M, (c_H, d_H, σ_H) and (c_M, d_M, σ_M) defined in* (2.5) *and* (2.6), *respectively, and for any real q, $0 \le q < 1$,*

$$\frac{\sqrt{k_n}}{\sigma_T \log\bigl(k_n/(n p_n)\bigr)} \left(\frac{\widehat\chi_{p_n}^{T(q)}}{\chi_{p_n}} - 1\right) = P_k^T + \sqrt{k_n}\left(c_T\, A(n/k) + d_T\,\frac{\chi_q^*}{U(n/k)}\right)\bigl(1+o_p(1)\bigr)\,,$$

where P_k^T is an asymptotically standard normal r.v.

Proof: From now on, we denote $a_n := \frac{k_n}{n p_n}$. With the underlying conditions in (1.1), a_n tends to infinity, as $n\to\infty$, and the quantile to be estimated can be expressed as

$$\chi_{p_n} = U\!\left(\frac{1}{p_n}\right) = U\!\left(\frac{n\, a_n}{k_n}\right) \ .$$

We will present the proof for $T = H$, since for $T = M$ the proof follows the same steps.

First notice that

$$\begin{aligned}\widehat{\chi}^{H(q)}_{p_n} &= \big(X_{n-k_n:n} - X_{n_q:n}\big)\, a_n^{\hat{\gamma}_n^{H(q)}} + X_{n_q:n} \\ &= X_{n-k_n:n}\left[\left(1 - \frac{X_{n_q:n}}{X_{n-k_n:n}}\right) a_n^{\hat{\gamma}_n^{H(q)}} + \frac{X_{n_q:n}}{X_{n-k_n:n}}\right] .\end{aligned}$$

Now, since $X_{n_q:n} \xrightarrow{p} \chi_q^*$, we have $\frac{X_{n_q:n}}{X_{n-k_n:n}} = o_p(1)$. Then

$$\widehat{\chi}^{H(q)}_{p_n} = X_{n-k_n:n}\left[a_n^{\hat{\gamma}_n^{H(q)}}\big(1+o_p(1)\big)\right] ,$$

which means that the proposed estimator $\widehat{\chi}^{H(q)}_{p_n}$ is asymptotically equivalent to the Weissman type estimator (1.7), whenever we use the consistent estimator $\hat{\gamma}_n \equiv \hat{\gamma}_n^{H(q)}$.

Consider now a convenient representation for the difference,

$$\widehat{\chi}^{H(q)}_{p_n} - \chi_{p_n} = X_{n-k_n:n}\left\{a_n^{\hat{\gamma}_n^{H(q)}} - a_n^{\hat{\gamma}_n^{H(q)}}\left(\frac{X_{n_q:n}}{X_{n-k_n:n}}\right) + \frac{X_{n_q:n}}{X_{n-k_n:n}} - \frac{\chi_{p_n}}{X_{n-k_n:n}}\right\} ,$$

and recall that we may write

$$\frac{\chi_{p_n}}{X_{n-k_n:n}} = \frac{U\big(\frac{n}{k_n}\, a_n\big)}{U\big(\frac{n}{k_n}\big)}\, \frac{U\big(\frac{n}{k_n}\big)}{U(Y_{n-k_n:n})} .$$

According to (1.5), for $\rho<0$, $U\big(\frac{n}{k_n}\, a_n\big)/U\big(\frac{n}{k_n}\big) = a_n^{\gamma}\big(1-A(n/k_n)/\rho\big)\,\big(1+o_p(1)\big)$.

Considering that for the estimator $\hat{\gamma}_n^{H(q)}$, the representation (2.4) holds, we get successively, for sequences k_n that verify (3.3),

$$a_n^{\hat{\gamma}_n^{H(q)}} = a_n^{\gamma}\left(1+\log a_n\big(\hat{\gamma}_n^{H(q)} - \gamma\big)\right)\big(1+o_p(1)\big)$$

and

$$\begin{aligned}&\widehat{\chi}^{H(q)}_{p_n} - \chi_{p_n} = \\ &= a_n^{\gamma}\, X_{n-k_n:n}\left\{1+\log a_n\big(\hat{\gamma}_n^{H(q)} - \gamma\big)\big(1+o_p(1)\big) - \big(1-A(n/k_n)/\rho\big)\big(1+o_p(1)\big)\right\} \\ &= a_n^{\gamma}\, X_{n-k_n:n}\left\{\log a_n\big(\hat{\gamma}_n^{H(q)} - \gamma\big) + A(n/k_n)/\rho\right\}\big(1+o_p(1)\big) .\end{aligned}$$

Now, we consider the following representation for intermediate statistics, proved in Ferreira et al. (2003),

$$X_{n-k_n:n} \stackrel{d}{=} U\!\left(\frac{n}{k_n}\right)\left(1+\frac{\gamma B_k}{\sqrt{k_n}} + o_p\!\left(\frac{1}{\sqrt{k_n}}\right) + o_p\big(A(n/k_n)\big)\right) , \tag{3.4}$$

with B_k an asymptotically standard normal r.v.

Using (2.4) and (3.4), we may write

$$\widehat{\chi}_{p_n}^{H(q)} - \chi_{p_n} \;=\; U\Big(\frac{n}{k_n}\Big)\, a_n^\gamma \Big(1 + O_p\big(1/\sqrt{k_n}\,\big)\Big) \,\Big\{W_n + A\Big(\frac{n}{k_n}\Big)/\rho\Big\}\,\big(1 + o_p(1)\big)\;,$$

where

$$\begin{aligned} W_n \;&=\; \log a_n \big(\hat{\gamma}_n^{H(q)} - \gamma\big) \\ &=\; \log a_n \left(\frac{\sigma_H}{\sqrt{k_n}}\, P_k^H + \left(c_{\scriptscriptstyle H}\, A(n/k) + d_{\scriptscriptstyle H} \frac{\chi_q^*}{U(n/k)}\right) \big(1 + o_p(1)\big)\right)\;, \end{aligned}$$

with P_k^H independent of the random sequence B_k in (3.4).

Consequently,

$$\frac{\widehat{\chi}_{p_n}^{H(q)} - \chi_{p_n}}{a_n^\gamma\, U\big(\frac{n}{k_n}\big)} \;=\; \big\{W_n + A(n/k)/\rho\big\}\,\big(1 + o_p(1)\big)$$

and

$$\frac{\sqrt{k_n}}{\sigma_H \log a_n} \left(\frac{\widehat{\chi}_{p_n}^{H(q)}}{\chi_{p_n}} - 1\right) \;=\; P_k^H + \sqrt{k_n}\left(c_{\scriptscriptstyle H}\, A(n/k) + d_{\scriptscriptstyle H} \frac{\chi_q^*}{U(n/k)}\right) \big(1 + o_p(1)\big)\,. \quad \square$$

The following result is a direct consequence of Corollary 2.2 and Theorem 3.1.

Corollary 3.1. *Under the same conditions of Theorem 3.1, then, with T replaced by H or M, and $(c_{\scriptscriptstyle H}, d_{\scriptscriptstyle H}, \sigma_{\scriptscriptstyle H})$ and $(c_{\scriptscriptstyle M}, d_{\scriptscriptstyle M}, \sigma_{\scriptscriptstyle M})$ defined in (2.5) and (2.6), respectively, the following results hold.*

i) *For $\gamma > -\rho$,*

$$\frac{\sqrt{k_n}}{\sigma_{\scriptscriptstyle T} \log\big(k_n/(n p_n)\big)} \left(\frac{\widehat{\chi}_{p_n}^{T(q)}}{\chi_{p_n}} - 1\right) \;=\; P_k^T + \sqrt{k_n}\,\big(c_{\scriptscriptstyle T}\, A(n/k)\big)\,\big(1 + o_p(1)\big)\;,$$

If $\sqrt{k_n}\, A(n/k_n) \to \mu_1$, finite, as $n \to \infty$, then the mean value is $\mu_1\, c_{\scriptscriptstyle T}$.

ii) *For $\gamma < -\rho$,*

$$\frac{\sqrt{k_n}}{\sigma_{\scriptscriptstyle T} \log\big(k_n/(n p_n)\big)} \left(\frac{\widehat{\chi}_{p_n}^{T(q)}}{\chi_{p_n}} - 1\right) \;=\; P_k^T + \sqrt{k_n}\left(d_{\scriptscriptstyle T} \frac{\chi_q^*}{U(n/k_n)}\right) \big(1 + o_p(1)\big)\;,$$

If $\sqrt{k_n}/U(n/k_n) \to \mu_2$, finite, as $n \to \infty$, then the mean values is $\mu_2\, d_{\scriptscriptstyle T}\, \chi_q^*$.

iii) *For* $\rho=-\gamma$,

$$\frac{\sqrt{k_n}}{\sigma_T \log\bigl(k_n/(n p_n)\bigr)} \left(\frac{\widehat{\chi}_{p_n}^{T(q)}}{\chi_{p_n}} - 1\right) =$$
$$= \; P_k^T + \sqrt{k_n}\left(c_T\, A(n/k) + d_T \frac{\chi_q^*}{U(n/k_n)}\right)\bigl(1+o_p(1)\bigr) \;,$$

If $\sqrt{k_n}\,A(n/k_n) \to \mu_1$, *finite, and* $\sqrt{k_n}/U(n/k_n) \to \mu_2$, *finite, as* $n\to\infty$, *then the mean value is* $\mu_1 c_T + \mu_2\, d_T \chi_q^*$.

4. SIMULATIONS

Here, we compare the finite sample behavior of the proposed high quantile estimators $\widehat{\chi}_{p_n}^{H(q)}$ in (3.1) and $\widehat{\chi}_{p_n}^{M(q)}$ in (3.2) with the classical semi-parametric estimators $\widehat{\chi}_{p_n}^{H}$ and $\widehat{\chi}_{p_n}^{M}$ in (1.11). We have generated $N=200$ independent replicates of sample size $n=1000$ from the following models:

- Burr Model: $X \frown \mathit{Burr}(\gamma,\rho)$, $\gamma=1$, $\rho=-2,-0.5$, with d.f.

$$F(x) \;=\; 1-\bigl(1+x^{-\rho/\gamma}\bigr)^{1/\rho}\,, \qquad x\geq 0\;.$$

- Cauchy Model: $X \frown \mathit{Cauchy}$, $\gamma=1$, $\rho=-2$, with d.f.

$$F(x) \;=\; \frac{1}{2}+\frac{1}{\pi}\,\mathrm{arctang}\; x\;, \qquad x\in\mathbb{R}\;.$$

At a first stage, we generate samples from the standard models $F_0 := F$. At a second stage, we introduce a positive shift $\lambda=\chi_{0.01}$, i.e., a new location chosen in a comparable basis as the percentile 99% of the starting point distribution F_0. This defines a new model $F_\lambda(x) := F_0(x-\lambda)$ from the same family.

We estimate the high quantile $\chi_{0.001}$, for each model F_0 or F_λ from the referred Burr and Cauchy families, and we present patterns of Mean Values and Root of Mean Squared Errors, plotted against $k=6, ..., 800$.

The simulations illustrate the dramatic disturbance on the behavior of the classical quantile estimators in (1.11), when a shift is introduced. We, again, enhance that the flat stable zones achieved with these estimators, in the presence of shifts, could lead us to dangerous misleading conclusions, unless we are aware of the suitable threshold k or of specific properties of the underlying model.

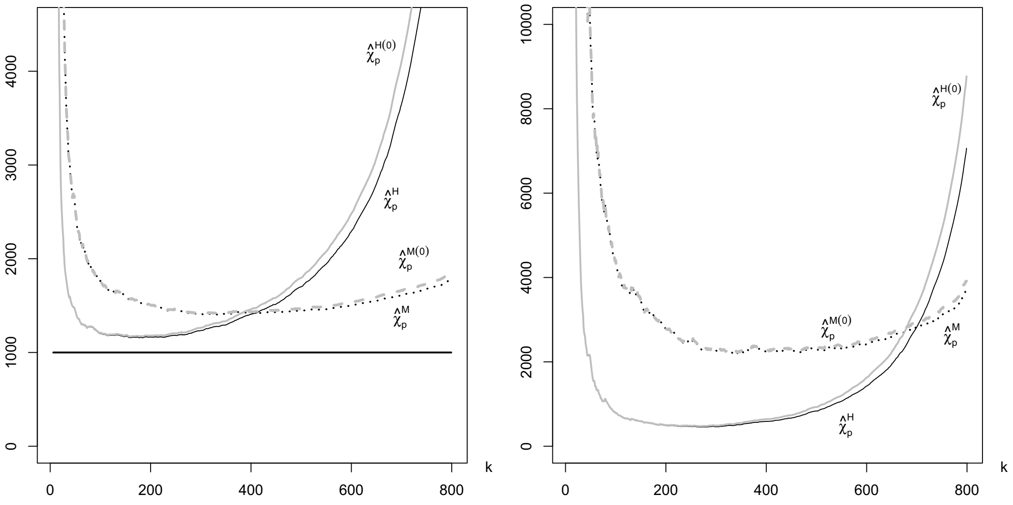


Figure 2: Mean values (left) and root mean squared errors (right), of $\widehat{\chi}_{p_n}^{H(0)}$, $\widehat{\chi}_{p_n}^{M(0)}$, $\widehat{\chi}_{p_n}^{H}$ and $\widehat{\chi}_{p_n}^{M}$, for a sample size $n = 1000$, from a Burr model with $\gamma = 1$, $\rho = -2$ and $\lambda = 0$ (target quantile $\chi_{0.001} = 1000$).

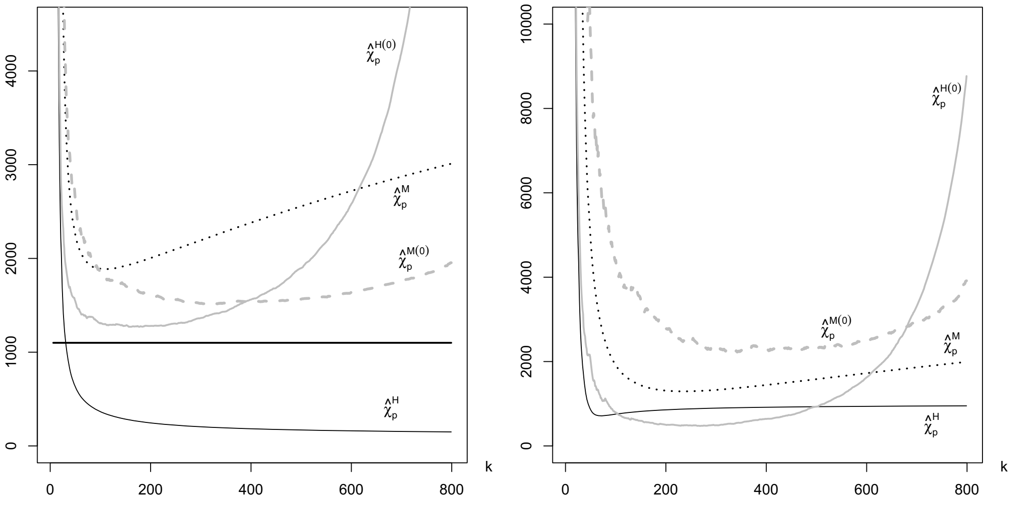


Figure 3: Mean values (left) and root mean squared errors (right), of $\widehat{\chi}_{p_n}^{H(0)}$, $\widehat{\chi}_{p_n}^{M(0)}$, $\widehat{\chi}_{p_n}^{H}$ and $\widehat{\chi}_{p_n}^{M}$, for a sample size $n = 1000$, from a Burr model with $\gamma = 1$, $\rho = -2$ and $\lambda = 99.99$ (target quantile $\chi_{0.001} = 1099.99$).

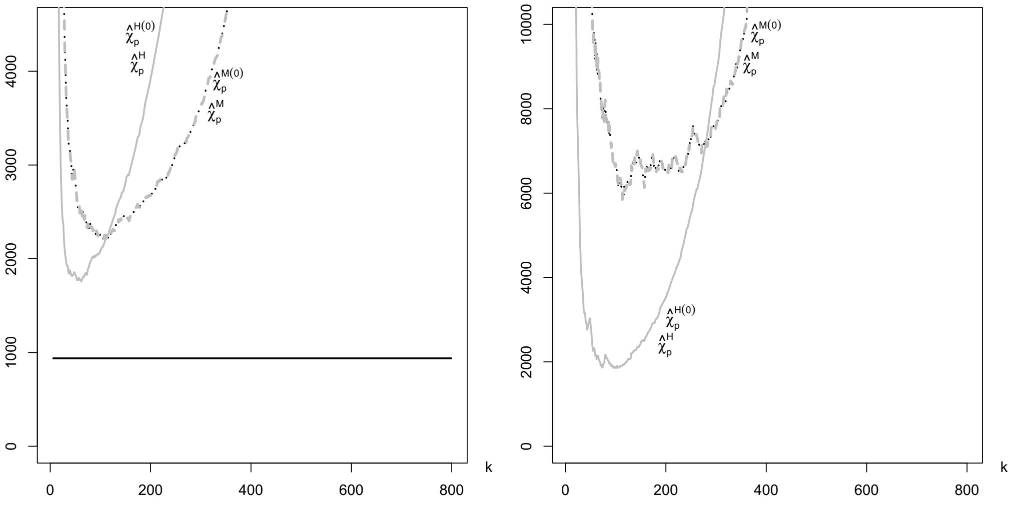


Figure 4: Mean values (left) and root mean squared errors (right), of $\widehat{\chi}^{H(0)}_{p_n}$, $\widehat{\chi}^{M(0)}_{p_n}$, $\widehat{\chi}^{H}_{p_n}$ and $\widehat{\chi}^{M}_{p_n}$, for a sample size $n=1000$, from a Burr model with $\gamma=1$, $\rho=-0.5$ and $\lambda=0$ (target quantile $\chi_{0.001}=937.731$).

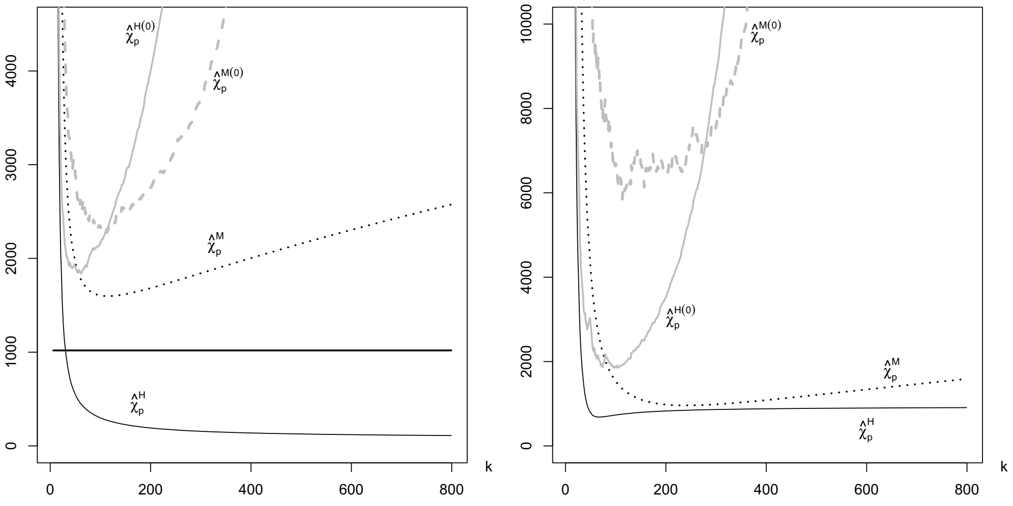


Figure 5: Mean values (left) and root mean squared errors (right), of $\widehat{\chi}^{H(0)}_{p_n}$, $\widehat{\chi}^{M(0)}_{p_n}$, $\widehat{\chi}^{H}_{p_n}$ and $\widehat{\chi}^{M}_{p_n}$, for a sample size $n=1000$, from a Burr model with $\gamma=1$, $\rho=-0.5$ and $\lambda=81.023$ (target quantile $\chi_{0.001}=1018.754$).

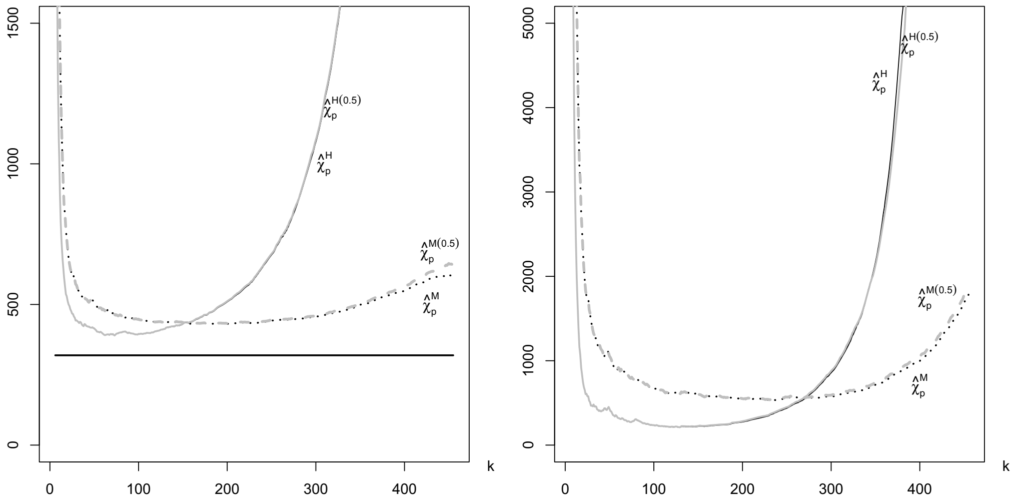


Figure 6: Mean values (left) and root mean squared errors (right), of $\widehat{\chi}_{p_n}^{H(0.5)}$, $\widehat{\chi}_{p_n}^{M(0.5)}$, $\widehat{\chi}_{p_n}^{H}$ and $\widehat{\chi}_{p_n}^{M}$, for a sample size $n=1000$, from a Cauchy model with $\gamma=1$, $\rho=-2$ and $\lambda=0$ (target quantile $\chi_{0.001}=319.309$).

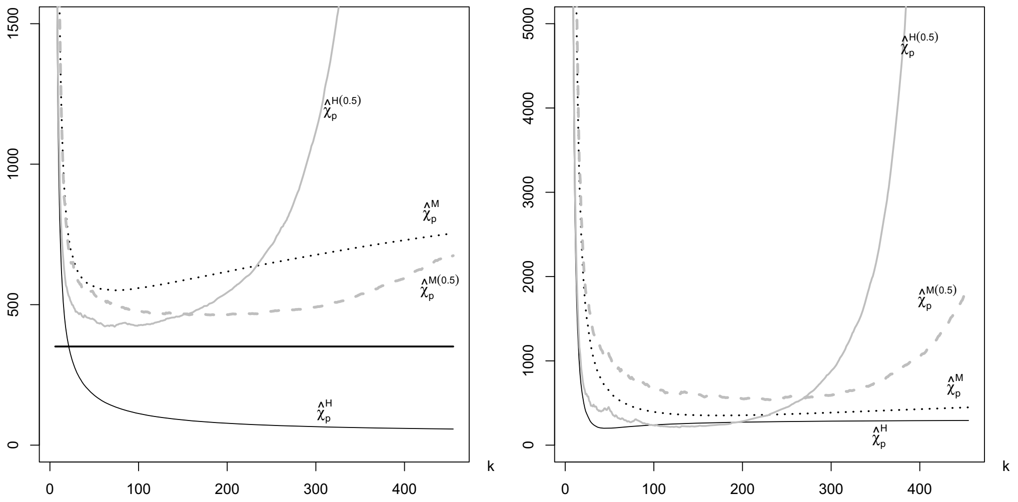


Figure 7: Mean values (left) and root mean squared errors (right), of $\widehat{\chi}_{p_n}^{H(0.5)}$, $\widehat{\chi}_{p_n}^{M(0.5)}$, $\widehat{\chi}_{p_n}^{H}$ and $\widehat{\chi}_{p_n}^{M}$, for a sample size $n=1000$, from a Cauchy model with $\gamma=1$, $\rho=-2$ and $\lambda=31.821$ (target quantile $\chi_{0.001}=351.13$).

From the figures, in this section, we observe that the classical quantile estimators diverge a lot from the important linear property (1.13). On the other hand, the estimators we propose, (3.1) and (3.2), enjoy exactly this property.

5. CONCLUDING REMARKS

- The *PORT* tail index and quantile estimators, based on the sample of excesses, $\underline{X}^{(q)}$, in (1.15), provide us with interesting classes of estimators, invariant for changes in location, as well as scale, a property also common to the classical estimators.
- In practice, whenever we use a *tuning parameter* q in $(0,1)$, we are always safe. Indeed, in such a case, the new estimators may or may not behave better than the classical ones, but they are consistent and asymptotically normal for the same type of k-values.
- A *tuning parameter* $q=0$ is appealing but should be used carefully. Indeed, if the underlying parent has not a finite left endpoint, we are led to non-consistent estimators, with sample paths that may be erroneously flat around a value quite far away from the real target.

ACKNOWLEDGMENTS

This work has been supported by FCT/POCTI and POCI/FEDER.

REFERENCES

[1] Caeiro, F.; Gomes, M.I. and Pestana, D. (2005). Direct reduction of bias of the classical Hill estimator, *Revstat*, **3**(2), 111–136.

[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] Drees, H. (2003). Extreme quantile estimation for dependent data, with applications to finance, *Bernoulli*, **9**(4), 617–657.

[4] Ferreira, A.; de Haan, L. and Peng, L. (2003). On optimizing the estimation of high quantiles of a probability distribution, *Statistics*, **37**, 401–434.

[5] Fraga Alves, M.I. and Araújo Santos, P. (2004). *Extreme quantiles estimation with shifted data from heavy tails*, Notas e Comunicações CEAUL 11/2004.

[6] Geluk, J. and de Haan, L. (1987). *Regular Variation, Extensions and Tauberian Theorems*, CWI Tract 40, Center of Mathematics and Computer Science, Amsterdam, Netherlands.

[7] Gomes, M.I. and Figueiredo, F. (2002). Bias Reduction in risk modelling: semi-parametric quantile estimation, *Test* (to appear).

[8] Gomes, M.I.; Martins, M.J. and Neves, M. (2005). *Revisiting the second order reduced bias "maximum likelihood" tail index estimators*, Notas e Comunicações CEAUL 10/2005 (submitted).

[9] Gomes, M.I. and Oliveira, O. (2003). How can non-invariant statistics work in our benefit in the semi-parametric estimation of parameters of rare events, *Commun. Stat., Simulation Comput.*, **32**(4), 1005–1028.

[10] Gomes, M.I. and Pestana, D. (2005). A sturdy second order reduced bias' Value at Risk estimator, *J. Amer. Statist. Assoc.* (to appear).

[11] Hall, P. (1982). On some simple estimates of an exponent of regular variation, *J. R. Statist. Soc.*, **44**(1) 37–42.

[12] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Ann. Statist.*, **3**(5), 1163–1174.

[13] Holton, G.A. (2003). *Value-at-Risk Theory and Practice*, Academic Press.

[14] Martins, M.J. (2000). *Estimação de Caudas Pesadas – Variantes ao Estimador de Hill*, Tese de Doutoramento, D.E.I.O., Faculdade de Ciências da Universidade de Lisboa.

[15] Resnick, S.I. (1997). Heavy tail modeling and teletraffic data, *Ann. Statist.*, **25**(5), 1805–1869.

[16] Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations, *J. Amer. Statist. Assoc.*, **73**, 812–815.

REVSTAT – Statistical Journal
Volume 4, Number 3, November 2006, 249–273

EXTREMES OF PERIODIC INTEGER-VALUED SEQUENCES WITH EXPONENTIAL TYPE TAILS

Authors: Andreia Hall
– Departamento de Matemática, UI&D Matemática e Aplicações,
Universidade de Aveiro, Portugal
andreia@mat.ua.pt

Manuel G. Scotto
– Departamento de Matemática, UI&D Matemática e Aplicações,
Universidade de Aveiro, Portugal
mscotto@mat.ua.pt

Received: July 2006 Accepted: October 2006

Abstract:

- This paper aims to analyze the extremal properties of periodic integer-valued sequences with marginal distribution belonging to a particular class defined by Anderson [1970. J. Appl. Probab. 7, 99–113] where the tail decays exponentially. An expression for calculating the extremal index of sequences satisfying certain local conditions, similar to those introduced by Chernick *et al.* [1991. Adv. Appl. Prob. 6, 711–731] is obtained. An application to infinite moving averages and max-autoregressive sequences is included. These results generalize the ones obtained for the stationary case.

Key-Words:

- *extreme value theory; binomial thinning; periodic sequences.*

AMS Subject Classification:

- 60G70, 60G10.

1. INTRODUCTION

The analysis of integer-valued time series has become an important area of research in the last two decades partially because its wide applicability to experimental biology (Zhou and Basawa [34]), social science (McCabe and Martin [24]), international tourism demand (Nordström [29], Garcia-Ferrer and Queralt [16], Brännäs *et al.* [12]), queueing systems (Ahn *et al.* [7]) and economy (Quoreshi [30]). We refer to McKenzie [28] for an overview of the early work in this area. Among the most successful integer-valued time series models proposed in the literature we mention the INAR(p) model and the INMA(q) model. The former was first introduced by McKenzie (e.g., [26]) and Al-Osh and Alzaid [1] for the case $p=1$. Empirical relevant extensions have been suggested by Brännäs ([9], explanatory variables), Blundell *et al.* ([8], panel data), Brännäs and Hellström ([11], extended dependence structure), and more recently by Silva *et al.* ([32], replicated data). Extensions and generalizations were introduced by Du and Li [14] and Latour [22]. The INMA(q) model was proposed by Al-Osh and Alzaid [2] and subsequently studied by Brännäs and Hall [10]. Related models were introduced by Aly and Bouzar ([4], [5]) and Zhu and Joe [35].

Within the reasonably large spectrum of integer-valued models proposed in the literature, little is known about its extremal properties. Anderson [6] gave a noticeable contribution to the study of the extremal properties of integer-valued independent and identically distributed (i.i.d.) sequences and as an example of application, the author analyzed the behavior of the maximum queue length for $M/M/1$ queues. Extensions of Anderson's results were proposed by Hooghiemstra *et al.* [21] who provide bounds and approximations for the distribution of the maximum queue length for $M/M/s$ queues, based on an asymptotic analysis involving the extremal index. McCormick and Park [25] were the first to study the extremal properties of some models obtained as discrete analogues of continuous models, replacing scalar multiplication by *random thinning*. Hall [17] analyzed the asymptotic behavior of the maximum term of a particular Markovian model. [18] provided results regarding the limiting distribution of the maximum of sequences within a generalized class of integer-valued moving averages driven by i.i.d. heavy-tailed innovations. Extensions for exponential type-tailed innovations have been studied by Hall [19]. More recently, Hall and Moreira [20] derived the extremal properties of a particular moving average count data model introduced by McKenzie [27].

It is worth to mention that all the references given in the previous paragraph deal with the case of stationary sequences. In contrast, however, the study of the extremal properties of integer-valued non-stationary sequences has been overlooked in the literature. This paper aims at giving a contribution towards this direction. In particular we consider periodic sequences with marginal distributions within a particular class of discrete distributions first considered by Anderson [6]. Potential applications can be found in the analysis of the number of hotel guest nights where the series exhibit strong seasonal pattern with a peak in July–August and a trough in December–February, and in the study of the number of claims of short-term disability benefits made by injured workers since it is expected to see fewer claims in the winter months and more in the summer months.

The term periodic is used in this paper in a different sense than in the literature of periodic stochastic processes in which a sequence $(X_n)_{n\in\mathbb{N}}$ is said to be periodically stationary (in the wide sense) if its mean and covariance structure are periodic functions of time with the same period. This class of processes, however, does not appear to be sufficiently flexible to deal with data which exhibit non-standard features like nonlinearity and/or heavy tails. In this paper by periodic sequence, with period say T, we mean that for a sequence of random variables (rv's) $(X_n)_{n\in\mathbb{N}}$ there exist an integer $T \geq 1$ such that, for each choice of integers $1 \leq i_1 < i_2 < \cdots < i_n$, $(X_{i_1}, ..., X_{i_n})$ and $(X_{i_1+T}, ..., X_{i_n+T})$ are identically distributed. The period T will be considered the smallest integer satisfying the above definition.

The rest of the paper is organized as follows: Section 2 provides the necessary theoretical background; Section 3 includes the main result that leads to the calculation of the limiting distribution of the maximum term; in Section 4 the previous results are applied to a particular class of max-autoregressive sequences generalizing the results of Hall [17]; finally, in Section 5 we look at the distribution of the maximum term of periodic moving average sequences obtained as discrete analogues of classical moving averages with periodic (but independent) innovations, generalizing the results given in Hall [19].

In this paper we want to highlight the following issues:

a) Under fairly general dependence conditions, integer-valued T-periodic sequences with marginal distribution in Anderson's class exhibit a quasi-stable non-degenerate limiting distribution of the maximum term which is obtained as a generalization of the stationary case.

b) The expression of the extremal index may be obtained from the joint distribution of a finite number of observations, calculated at T distinct sets of variables.

c) The results obtained for the integer-valued max-autoregressive and moving average models generalize the ones obtained for the stationary case: whereas for the max-autoregressive model the extremal index is less than unity (reflecting the influence of the dependence structure on the extremes), for the moving averages the extremal index is equal to one.

2. PRELIMINARY RESULTS

The study of the extremal properties of stationary sequences is frequently based on the verification of appropriate dependence conditions which assure that the limiting distribution of the maximum term is of the same type as the limiting distribution of the maximum of i.i.d. rv's with the same marginal distribution F. For stationary sequences, usual conditions used in the literature are Leadbetter's $D(u_n)$ condition (Leadbetter *et al.* [23]) and condition $D^{(k)}(u_n)$, $k \in \mathbb{N}$, (Chernick *et al.* [13]). For completeness and reader's convenience the definition of condition $D(u_n)$ is given below.

Definition 2.1. The condition $D(u_n)$ is said to hold for a stationary sequence $(X_n)_{n\in\mathbb{N}}$ with marginal distribution F, if for any integers $i_1 < \cdots < i_p < j_1 < \cdots < j_q < n$ such that $j_1 - i_p \geq l_n$ we have

$$\Big|F_{i_1,\ldots,i_p,j_1,\ldots,j_q}(u_n,\ldots,u_n) - F_{i_1,\ldots,i_p}(u_n,\ldots,u_n)\,F_{j_1,\ldots,j_q}(u_n,\ldots,u_n)\Big| \;\leq\; \alpha_{n,l_n}$$

with $\alpha_{n,l_n} \underset{n\to\infty}{\longrightarrow} 0$ for some sequence (l_n), $l_n = o(n)$.

For periodic sequences the following adaptation of condition $D^{(k)}(u_n)$ may be used:

Definition 2.2 (Ferreira and Martins [15])**.** Let $k \geq 1$ be a fixed integer and $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ a T-periodic sequence verifying $D(u_n)$ with mixing coefficient α_{n,l_n}. The condition $D_T^{(k)}(u_n)$ holds for $\mathbf{X}$ if there exists a sequence of integers $(k_n)_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} k_n = +\infty\,, \qquad \lim_{n\to\infty} k_n \frac{l_n}{n} = 0\,, \qquad \lim_{n\to\infty} k_n\,\alpha_{n,l_n} = 0\ ,$$

$$\lim_{n\to\infty} S^{(k)}_{[\frac{n}{k_n T}]} = 0\ ,$$

where

$$S^{(1)}_{[\frac{n}{k_n T}]} \;=\; \frac{n}{T}\sum_{i=1}^{T}\sum_{j=i+k}^{[\frac{n}{k_n T}]T} P\big(X_i > u_n,\; X_j > u_n\big)\,,$$

and for $k \geq 2$

$$S^{(k)}_{[\frac{n}{k_n T}]} \;=\; \frac{n}{T}\sum_{i=1}^{T}\sum_{j=i+k}^{[\frac{n}{k_n T}]T} P\big(X_i > u_n,\; X_{j-1} \leq u_n < X_j\big)\,.$$

Remark 2.1. If $\lim_{n\to\infty} S^{(k)}_{[\frac{n}{k_n T}]} = 0$, then

$$\lim_{n\to\infty} \frac{n}{T} \sum_{i=1}^{T} P\big(X_i > u_n \geq M_{i+1,i+k-1},\, M_{i+k,[\frac{n}{k_n T}]T} > u_n\big) \;=\; 0 \;,$$

with $M_{i,j} = \max_{i\leq r\leq j}(X_r)$ and $M_{i,j} = -\infty$ if $i > j$.

When $D(u_n)$ and $D^{(k)}_T(u_n)$ hold for a particular sequence the limiting distribution of the maximum term and its corresponding extremal index may be derived. Following Ferreira and Martins [15] the extremal index is given by

$$\theta \;=\; \lim_{n\to\infty} \frac{n\,\frac{1}{T}\sum\limits_{i=1}^{T} P\big(X_i > u_n \geq M_{i+1,i+k-1}\big)}{n\,\frac{1}{T}\sum\limits_{i=1}^{T} P\big(X_i > u_n\big)} \;.$$

Integer-valued sequences require extra care when the analysis of the extremal properties is in demand since in many cases, there is no non-degenerate limiting distribution for the maximum term. Anderson [6] defined a particular class of discrete distributions for which the maximum term (under an i.i.d. setting) possesses an almost stable behavior in the sense of the following theorem:

Theorem 2.1 (Anderson [6]). *Let F be a distribution function whose support consists of all sufficiently large integers. Then, there exists a sequence of constants (b_n) such that*

$$\begin{cases} \limsup\limits_{n\to\infty} F^n(x+b_n) \;\leq\; e^{-e^{-\alpha x}} \\[2ex] \liminf\limits_{n\to\infty} F^n(x+b_n) \;\geq\; e^{-e^{-\alpha(x-1)}} \end{cases},$$

for some $\alpha > 0$ and for every $x \in \mathbb{R}$, if and only if

$$\lim_{n\to\infty} \frac{1-F(n)}{1-F(n-1)} \;=\; \exp\{-\alpha\} \;.$$

In fact b_n may be obtained by $b_n = F_c^{-1}(1-\frac{1}{n})$ where F_c is any continuous distribution in the domain of attraction of the Gumbel distribution with $F_c([x]) = F_x$.

Whenever a distribution F satisfies the conditions of the theorem above we shall denote it by $F \in D_\alpha(Anderson)$. The study of stationary sequences with marginal distribution in the class of Anderson [6] was considered by Hall [17], who obtained the following result:

Theorem 2.2 (Hall [17]). *Suppose that for some $k \geq 1$, conditions $D(u_n)$ and $D^{(k)}(u_n)$ hold for the stationary sequence $\boldsymbol{X}$ with marginal $F \in D_\alpha$(Anderson), where $\boldsymbol{u_n}$ is a sequence of the form $u_n = x + b_n$. If $M_n = \max_{1 \leq k \leq n}(X_k)$, then there exists a value $0 \leq \theta \leq 1$ such that*

$$\begin{cases} \limsup\limits_{n\to\infty} P\big(M_n \leq x + b_n\big) \leq e^{-\theta e^{-\alpha x}} \\ \liminf\limits_{n\to\infty} P\big(M_n \leq x + b_n\big) \geq e^{-\theta e^{-\alpha(x-1)}} \end{cases},$$

if and only if

$$P\big(M_{2,k} \leq u_n | X_1 > u_n\big) \xrightarrow[n\to\infty]{} \theta \ .$$

Hall refers to the parameter θ as the extremal index due to its similarity with the conventional extremal index.

3. LIMITING DISTRIBUTION FOR THE MAXIMUM TERM

In this section attention is focused in the extremal behavior of periodic sequences with marginal distributions in Anderson's class. The first result extends Theorem 3 in Hall [17] for T-periodic integer-valued sequences.

Theorem 3.1. *Suppose that for $k \geq 1$ the conditions $D(u_n)$ and $D_T^k(u_n)$ hold for the T-periodic integer-valued sequence $\mathbf{X}$, with $F_r \in D_{\alpha_r}$(Anderson), for $r = 1, ..., T$ where $(u_n)_{n\in\mathbb{N}}$ is a sequence of the form $u_n = x + b_n$. If there exists $\underline{\theta}$ and $\overline{\theta}$, $0 \leq \underline{\theta} \leq \overline{\theta} \leq 1$, such that*

$$\begin{aligned} \underline{\theta} \;&=\; \liminf_{n\to\infty} \frac{\frac{n}{T} \sum\limits_{r=1}^{T} P\big(X_r > u_n > M_{r+1,r+k-1}\big)}{\frac{n}{T} \sum\limits_{r=1}^{T} P\big(X_r > u_n\big)} \\ &\leq\; \limsup_{n\to\infty} \frac{\frac{n}{T} \sum\limits_{r=1}^{T} P\big(X_r > u_n > M_{r+1,r+k-1}\big)}{\frac{n}{T} \sum\limits_{r=1}^{T} P\big(X_r > u_n\big)} \;=\; \overline{\theta} \ , \end{aligned}$$

then

$$\begin{cases} \limsup\limits_{n\to\infty} P\big(M_n \leq x + b_n\big) \leq e^{-\underline{\theta} \frac{1}{T} \sum_{r=1}^{T} e^{-\alpha_r x}} \\ \liminf\limits_{n\to\infty} P\big(M_n \leq x + b_n\big) \geq e^{-\overline{\theta} \frac{1}{T} \sum_{r=1}^{T} e^{-\alpha_r (x-1)}} \end{cases}.$$

Proof: First let us suppose that $\liminf_{n\to\infty} P(M_n \le x+b_n) > 0$, $\forall x$. By Proposition 2.1 in Ferreira and Martins [15] we have that

$$P\big(M_n \le u_n\big) - e^{-\frac{n}{T}\sum_{r=1}^{T} P(X_r > u_n > M_{r+1,r+k-1})} \to 0 , \qquad n\to\infty ,$$

which is equivalent to

$$P\big(M_n \le u_n\big) - \Big(e^{-\frac{n}{T}\sum_{r=1}^{T} P(X_r>u_n)}\Big)^{\frac{\frac{n}{T}\sum_{r=1}^{T} P(X_r>u_n>M_{r+1,r+k-1})}{\frac{n}{T}\sum_{r=1}^{T} P(X_r>u_n)}} \to 0 , \tag{3.1}$$

as $n\to\infty$. From Theorem 2.1 it follows that

$$\begin{aligned} 0 \;<\; e^{-\frac{1}{T}\sum_{r=1}^{T} e^{-\alpha_r(x-1)}} &\le \liminf_{n\to\infty} e^{-\frac{n}{T}\sum_{r=1}^{T} P(X_r>u_n)} \\ &\le \limsup_{n\to\infty} e^{-\frac{n}{T}\sum_{r=1}^{T} P(X_r>u_n)} \;\le\; e^{-\frac{1}{T}\sum_{r=1}^{T} e^{-\alpha_r x}} \;<\; 1 , \end{aligned}$$

and if we assume

$$\begin{aligned} \underline{\theta} \;&=\; \liminf_{n\to\infty} \frac{\frac{n}{T}\sum\limits_{r=1}^{T} P\big(X_r > u_n > M_{r+1,r+k-1}\big)}{\frac{n}{T}\sum\limits_{r=1}^{T} P(X_r > u_n)} \\ &\le\; \limsup_{n\to\infty} \frac{\frac{n}{T}\sum\limits_{r=1}^{T} P\big(X_r > u_n > M_{r+1,r+k-1}\big)}{\frac{n}{T}\sum\limits_{r=1}^{T} P\big(X_r > u_n\big)} \;=\; \overline{\theta} \end{aligned}$$

then, (3.1) leads to the stated result.

The case $P(M_n \le x+b_n) \to 0$ as $n\to\infty$ is easily handled by the results above and the arguments in Hall ([17], p. 725). We skip the details. □

As a consequence of Theorem 3.1 the extremal index can be computed as follows:

Corollary 3.1. *Suppose that for some $k\ge 1$ the conditions $D(u_n)$ and $D_T^k(u_n)$ hold for the T-periodic integer-valued sequence $\mathbf{X}$, with $F_r \in D_{\alpha_r}(Anderson)$, for $r=1,...,T$ where $\{u_n\}_{n\in\mathbb{N}}$ is a sequence of the form $u_n = x+b_n$. Then, there exists a value $0\le\theta\le 1$ such that*

$$\begin{cases} \limsup\limits_{n\to\infty} P\big(M_n \le x+b_n\big) \;\le\; e^{-\theta\frac{1}{T}\sum_{r=1}^{T} e^{-\alpha_r x}} \\ \liminf\limits_{n\to\infty} P\big(M_n \le x+b_n\big) \;\ge\; e^{-\theta\frac{1}{T}\sum_{r=1}^{T} e^{-\alpha_r(x-1)}} \end{cases} ,$$

if and only if

$$\frac{\frac{n}{T}\sum\limits_{r=1}^{T} P\big(X_r > u_n > M_{r+1,r+k-1}\big)}{\frac{n}{T}\sum\limits_{r=1}^{T} P\big(X_r > u_n\big)} \to \theta , \qquad n\to\infty .$$

4. MAX-AUTOREGRESSIVE PERIODIC SEQUENCES

Let $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ be a T-periodic non-negative integer-valued max-autoregressive sequence defined as

$$X_n \;=\; \max\{X_{n-1}, Z_n\} - c_n \;, \tag{4.1}$$

where $(c_1, ..., c_T) \in \mathbb{N}^T$, $c_{n+T} = c_n$ for all $n \in \mathbb{N}$ and $\mathbf{Z} = (Z_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. integer-valued rv's with common distribution F. Let H_n denote the distribution of X_n. The max-autoregressive sequence defined in (4.1) is an extension of the max-autoregressive model considered by Alpuim [3]. Her ideas will be extensively used throughout this section. First note that the following relations hold

$$\begin{aligned} H_n(x) \;&=\; P\big(X_n \le x\big) \;=\; P\big(X_{n-1} \le x + c_n,\, Z_n \le x + c_n\big) \\ &= \prod_{i=0}^{\infty} F\Big(x + \sum_{l=0}^{i} c_{n-l}\Big) \;=\; \prod_{s=0}^{T-1}\prod_{j=0}^{\infty} F\big(x + jS + S_{s,n}\big)\,, \end{aligned}$$

with $S = \sum_{i=1}^{T} c_i$ and $S_{s,n} = \sum_{l=0}^{s} c_{n-l}$. Moreover, it is also true that

$$F(x) = \frac{H_n(x - c_n)}{H_{n-1}(x)}\,, \qquad \text{for all} \;\; n \;. \tag{4.2}$$

Next result shows that if F belongs to Anderson's class then H_n will also belong to Anderson's class for all n.

Lemma 4.1. *Let* $\mathbf{X}$ *be a max-autoregressive integer-valued T-periodic sequence defined by* (4.1). *If* $F \in D_\alpha$*(Anderson) then* $H_n \in D_\alpha$*(Anderson),* $\forall\, n \in \mathbb{N}$. *Let* $u_n = x + b_n$ *be such that*

$$\begin{cases} \limsup\limits_{n\to\infty}\, n\big(1 - F(u_n)\big) \,\le\, e^{-\alpha x} \\ \liminf\limits_{n\to\infty} n\big(1 - F(u_n)\big) \,\ge\, e^{-\alpha(x-1)} \end{cases} .$$

Choosing $u'_n = x + b_n + \frac{\ln C_1}{\alpha}$ *with* $C_1 = \frac{\sum_{s=0}^{T-1} e^{-S_{s,1}\alpha}}{1 - e^{-S\alpha}}$ *it follows that*

$$\begin{cases} \limsup\limits_{n\to\infty}\, n\big(1 - H_1(u'_n)\big) \,\le\, e^{-\alpha x} \\ \liminf\limits_{n\to\infty}\; n\big(1 - H_1(u'_n)\big) \,\ge\, e^{-\alpha(x-1)} \end{cases}$$

and

$$\begin{cases} \limsup\limits_{n\to\infty}\, n\big(1 - H_r(u'_n)\big) \,\le\, \gamma_{r,1}\, e^{-\alpha x} \\ \liminf\limits_{n\to\infty}\; n\big(1 - H_r(u'_n)\big) \,\ge\, \gamma_{r,1}\, e^{-\alpha(x-1)} \end{cases} ,$$

where

$$\gamma_{i,r} = \lim_{x\to\infty} \frac{1 - H_i(x)}{1 - H_r(x)}\,, \qquad r = 1, ..., T, \;\; i = 0, ..., T-1 \;. \tag{4.3}$$

Furthermore for $i=0,...,T-1$

$$\begin{cases} \limsup\limits_{n\to\infty} \dfrac{n}{T}\sum\limits_{i=1}^{T} P\big(X_i > u_n'\big) \le \dfrac{1}{T}\sum\limits_{i=1}^{T} \gamma_{i,1}\, e^{-\alpha x} \\ \liminf\limits_{n\to\infty} \dfrac{n}{T}\sum\limits_{i=1}^{T} P\big(X_i > u_n'\big) \ge \dfrac{1}{T}\sum\limits_{i=1}^{T} \gamma_{i,1}\, e^{-\alpha(x-1)} \end{cases} .$$

Proof: First note that for any two integer-valued distribution functions, say F_1 and F_2, the following relation hold: If $F_1 \in D_\alpha(Anderson)$ and $\lim\limits_{n\to\infty} \frac{1-F_2(n)}{1-F_1(n)} = c > 0$ then $F_2 \in D_\alpha(Anderson)$. Furthermore, if b_n is such that

$$\begin{cases} \limsup\limits_{n\to\infty} n\big(1-F_1(x+b_n)\big) \le e^{-\alpha x} \\ \liminf\limits_{n\to\infty} n\big(1-F_1(x+b_n)\big) \ge e^{-\alpha(x-1)} \end{cases} ,$$

then for $b_n' = b_n + \frac{\ln c}{\alpha}$

$$\begin{cases} \limsup\limits_{n\to\infty} n\big(1-F_2(x+b_n')\big) \le e^{-\alpha x} \\ \liminf\limits_{n\to\infty} n\big(1-F_2(x+b_n')\big) \ge e^{-\alpha(x-1)} \end{cases} .$$

Now suppose that $F \in D_\alpha(Anderson)$.

$$\begin{aligned} \lim_{x\to\infty} \frac{1-H_n(x)}{1-F(x)} &= \lim_{x\to\infty} \frac{\prod\limits_{s=0}^{T-1}\prod\limits_{j=0}^{\infty} F\big(x+jS+S_{s,n}\big)}{1-F(x)} \\ &= \lim_{x\to\infty} \frac{\sum\limits_{s=0}^{T-1}\sum\limits_{j=0}^{\infty} 1-F\big(x+jS+S_{s,n}\big)}{1-F(x)} \\ &= \sum_{s=0}^{T-1} e^{-S_{s,n}\alpha} \lim_{x\to\infty} \sum_{j=0}^{\infty} \frac{1-F(x+jS)}{1-F(x)} . \end{aligned}$$

Since $\lim_{x\to\infty} \frac{1-F(x)}{1-F(x-1)} = e^{-\alpha}$ we may choose $\alpha' < \alpha$ so that there exists x_0 such that for all $x > x_0$ then $\frac{1-F(x+jS)}{1-F(x)} < e^{-jS\alpha'}$ for all j. By the dominated convergence theorem, limit and sum can be interchanged providing

$$\lim_{x\to\infty} \frac{1-H_n(x)}{1-F(x)} = \sum_{s=0}^{T-1} e^{-S_{s,n}\alpha} \sum_{j=0}^{\infty} e^{-jS\alpha} = \frac{\sum\limits_{s=0}^{T-1} e^{-S_{s,n}\alpha}}{1-e^{-S\alpha}} \equiv C_n .$$

Applying the relations stated in the beginning of the proof we conclude that $H_n \in D_\alpha(Anderson)$. □

We shall now obtain the asymptotic behaviour of the maximum term of the T-periodic non-negative integer-valued max-autoregressive sequence in (4.1).

Theorem 4.1. *Let* $\mathbf{X}$ *be the* T*-periodic non-negative integer-valued moving average sequence defined in* (4.1) *with* $F \in D_\alpha(Anderson)$. *If* $M_n = \max\limits_{1\le k\le n}(X_k)$ *and* $u_n = x + b_n$ *with*

$$b_n \;=\; b'_n + \frac{\ln\Bigl(\frac{1}{T}\sum\limits_{i=1}^{T} C_i\Bigr)}{\alpha}$$

where $C_i = \frac{\sum_{s=0}^{T-1} e^{-S_{s,i}\alpha}}{1-e^{-S\alpha}}$ *and* b'_n *is the sequence of normalizing constants of* F, *then*

$$\begin{cases} \limsup\limits_{n\to\infty} P\bigl(M_n \le u_n\bigr) \;\le\; e^{-\theta e^{-\alpha x}} \\ \liminf\limits_{n\to\infty} P\bigl(M_n \le u_n\bigr) \;\ge\; e^{-\theta e^{-\alpha(x-1)}} \end{cases}$$

and the extremal index θ *is given by*

$$\theta \;=\; \frac{\sum\limits_{i=1}^{T} \gamma_{i,1}\bigl(1 - \exp\{-\alpha\, c_{i+1}\}\bigr)}{\sum\limits_{i=1}^{T} \gamma_{i,1}}\;, \tag{4.4}$$

with $\gamma_{i,1} = C_i/C_1$.

Proof: First we prove that condition $D(u_n)$ holds for $\mathbf{X}$. Note that for any two indexes i_1, i_2 we obtain the following relations by (4.2):

$$\begin{aligned} P\bigl(X_{i_1} \le x,\, X_{i_2} \le x\bigr) \;&=\; P\bigl(X_{i_1} \le x\bigr) \prod_{l=0}^{i_2-i_1-1} F(x + S_{l,i_2}) \\ &=\; H_{i_1}(x)\, \frac{H_{i_2}(x)}{H_{i_1}(x + S_{i_2-i_1-1,i_2})}\;. \end{aligned} \tag{4.5}$$

Using (4.5) we obtain

$$\begin{aligned} &\Bigl|H_{i_1,\dots,i_p,j_1,\dots,j_q}(u_n,\dots,u_n) - H_{i_1,\dots,i_p}(u_n,\dots,u_n)\, H_{j_1,\dots,j_q}(u_n,\dots,u_n)\Bigr| \;= \\ &\qquad =\; \left| H_{i_1}(u_n) \prod_{m=2}^{p} \prod_{l=0}^{i_m-i_{m-1}-1} F(u_n + S_{l,i_m}) \prod_{m=2}^{q} \prod_{l=0}^{j_m-j_{m-1}-1} F(u_n + S_{l,j_m}) \right. \\ &\qquad\qquad \left. \times \left(\prod_{l=0}^{l_n-1} F(u_n + S_{l,j_1}) - H_{j_1}(u_n) \right) \right| \\ &\qquad \le\; \left| \frac{H_{j_1}(u_n)}{H_{i_p}(u_n + S_{j_1-i_p-1})} - H_{j_1}(u_n) \right| \\ &\qquad \le\; 1 - H_{i_p}(u_n + S_{j_1-i_p-1}) \;\le\; 1 - H_{i_p}(u_n)\;. \end{aligned}$$

Since $1 - H_i(u_n) \sim O(\frac{1}{n})$ for all i, the desired result is obtained.

Next we show that condition $D''_T(u_n)$ also holds for $\mathbf{X}$.

$$\begin{aligned} P\big(X_i > u_n \geq X_{i+1},\, X_{i+j} > u_n\big) \;=&\\ &= P\big(X_i > u_n,\, X_{i+j} > u_n | X_{i+1} \leq u_n\big)\, H_{i+1}(u_n)\\ &= P\big(X_i > u_n | X_{i+1} \leq u_n\big)\, P\big(X_{i+j} > u_n | X_{i+1} \leq u_n\big)\, H_{i+1}(u_n)\ , \end{aligned}$$

since the events $\{X_i > u_n | X_{i+1} \leq u_n\}$ and $\{X_{i+j} > u_n | X_{i+1} \leq u_n\}$ are independent for this type of sequences. Moreover

$$\begin{aligned} P\big(X_i > u_n | X_{i+1} \leq u_n\big) \;&=\; \frac{H_{i+1}(u_n) - H_i(u_n)\, F(u_n + c_{i+1})}{H_{i+1}(u_n)}\\ &=\; 1 - \frac{H_i(u_n)}{H_{i+1}(u_n)}\, F(u_n + c_{i+1})\ . \end{aligned}$$

Since $\frac{H_i(u_n)}{H_{i+1}(u_n)} \geq H_{i+1}(u_n)$ we have

$$P\big(X_i > u_n | X_{i+1} \leq u_n\big) \;\leq\; 1 - H_i(u_n) \;=\; O\left(\frac{1}{n}\right) .$$

For the second term we have

$$\begin{aligned} P\big(X_{i+j} > u_n | X_{i+1} \leq u_n\big) \;&=\; 1 - \frac{H_{i+j}(u_n)}{H_{i+1}\Big(u_n + \sum\limits_{m=0}^{j-2} c_{i+j-m}\Big)}\\ &\leq\; 1 - H_{i+j}(u_n)\\ &=\; O\left(\frac{1}{n}\right) . \end{aligned}$$

Hence

$$\begin{aligned} \lim_{n\to\infty} \frac{n}{T} \sum_{i=1}^{T} \sum_{j=i+2}^{[\frac{n}{k_n T}]T} P\big(X_i > u_n \geq X_{i+1},\, X_{i+j} > u_n\big) \;&\leq\; \lim_{n\to\infty} nT \left[\frac{n}{k_n T}\right] O\left(\frac{1}{n}\right) O\left(\frac{1}{n}\right)\\ &=\; 0\ . \end{aligned}$$

Note that by Corollary 3.1

$$\begin{aligned} \theta \;&=\; \lim_{n\to\infty} \frac{\frac{n}{T} \sum\limits_{i=1}^{T} P\big(X_i > u_n \geq X_{i+1}\big)}{\frac{n}{T} \sum\limits_{i=1}^{T} P\big(X_i > u_n\big)}\\ &=\; \lim_{n\to\infty} \frac{\sum\limits_{i=1}^{T} P\big(X_i > u_n \geq X_{i+1}\big) \,/\, P\big(X_1 > u_n\big)}{\sum\limits_{i=1}^{T} P\big(X_i > u_n\big) \,/\, P\big(X_1 > u_n\big)}\ . \end{aligned}$$

Since

$$\lim_{n\to\infty} \sum_{i=1}^{T} P\big(X_i > u_n\big) \,/\, P\big(X_1 > u_n\big) \;=\; \sum_{i=1}^{T} \gamma_{i,1}$$

and

$$\begin{aligned}
\lim_{n\to\infty} \sum_{i=1}^{T} & P\big(X_i > u_n \ge X_{i+1}\big) \,/\, P\big(X_1 > u_n\big) \;= \\
&= \lim_{n\to\infty} \sum_{i=1}^{T} \Big(P\big(X_i \le u_n\big) - P\big(X_i \le u_n,\, X_{i+1} \le u_n\big)\Big) \,/\, P\big(X_1 > u_n\big) \\
&= \lim_{n\to\infty} \sum_{i=1}^{T} \Big(H_{i+1}(u_n) - H_i(u_n)\, F(u_n + c_{i+i})\Big) \,/\, \big(1 - H_1(u_n)\big) \\
&= \lim_{n\to\infty} \sum_{i=1}^{T} \left(H_{i+1}(u_n) - H_i(u_n)\, \frac{H_{i+1}(u_n)}{H_i(u_n + c_{i+i})}\right) \,/\, \big(1 - H_1(u_n)\big) \\
&= \lim_{n\to\infty} \sum_{i=1}^{T} \frac{H_{i+1}(u_n)}{H_i(u_n + c_{i+1})} \Big(H_i(u_n + c_{i+1}) - H_i(u_n)\Big) \,/\, \big(1 - H_1(u_n)\big) \\
&= \lim_{n\to\infty} \sum_{i=1}^{T} \frac{1 - H_i(u_n)}{1 - H_1(u_n)} \left(1 - \frac{1 - H_i(u_n + c_{i+1})}{1 - H_i(u_n)}\right) \\
&= \sum_{i=1}^{T} \gamma_{i,1} \big(1 - \exp(-\,\alpha\, c_{i+1})\big) \;,
\end{aligned}$$

then

$$\theta \;=\; \frac{\sum\limits_{i=1}^{T} \gamma_{i,1} \big(1 - \exp(-\,\alpha\, c_{i+1})\big)}{\sum\limits_{i=1}^{T} \gamma_{i,1}} \;,$$

concluding the proof. □

5. MOVING AVERAGE MODELS WITH EXPONENTIAL TYPE-TAILS

Let $\mathbf{Z} = (Z_n)_{n\in\mathbb{Z}}$ be a sequence of T-periodic integer-valued random variables. Throughout this section we will assume that

$$(5.1) \qquad 1 - F_{Z_r}(x) \;\sim\; K_r\, x^{\xi_r} (1 + \lambda_r)^{-x} \;, \qquad x \to \infty, \;\; \xi \in \mathbb{R}, \;\; K_r, \lambda_r > 0 \;,$$

for $r = 1, ..., T$. Furthermore, we assume that X_n admits the representation

$$(5.2) \qquad X_n \;=\; \sum_{j=-\infty}^{\infty} \beta_j \circ Z_{n-j} \;, \qquad \beta_j \in [0,1] \;,$$

where the discrete operator $\circ$ denotes binomial thinning defined as $\beta \circ Z = \sum_{s=1}^{Z} U_s(\beta)$, where $(U_s(\beta))$ is a i.i.d. sequence of Bernoulli random variables verifying $P\big(U_s(\beta)=1\big) = \beta$. Moreover, the sequence of coefficients $(\beta_j)_{j\in\mathbb{Z}}$ will be taken to satisfy

$$\sum_{j=-\infty}^{\infty} \beta_j < \infty \ ,$$

in order to ensure the almost sure convergence of (5.2). All thinning operations involved in (5.2) are independent, for each n. Nevertheless, dependence is allowed to occur between the thinning operators $\beta_j \circ Z_n$ and $\beta_i \circ Z_n$, $j \neq i$ (which belong to X_{n+j} and X_{n+i} respectively).

Lemma 5.1. *Under the conditions set above, the sum*

$$\sum_{j=-\infty}^{\infty} \beta_j \circ Z_{n-j} \ ,$$

with

$$\beta_j = O\big(|j|^{-\delta}\big) \ , \tag{5.3}$$

as $j \to \pm\infty$, for some $\delta > 2$, converges almost surely to X_n.

Proof: Note that

$$E\left[\sum_{j=-\infty}^{\infty} \beta_j \circ Z_{n-j}\right] \;=\; \sum_{s=0}^{T-1} E\big[Z_{n-s}\big] \sum_{j=-\infty}^{\infty} \beta_{jT+s} \;<\; \infty \ .$$

Likewise,

$$\begin{aligned}
\mathrm{Var}&\left[\sum_{j=-\infty}^{\infty} \beta_j \circ Z_{n-j}\right] \;= \\
&= \sum_{s=0}^{T-1} \Big(\mathrm{Var}\big[Z_{n-s}\big] - E\big[Z_{n-s}\big]\Big) \sum_{j=-\infty}^{\infty} \beta_{jT+s}^2 + E[Z_{n-s}] \sum_{j=-\infty}^{\infty} \beta_{jT+s} \\
&< \infty \ .
\end{aligned}$$

Thus $\sum_{j=-\infty}^{\infty} \beta_j \circ Z_{n-j} \to X_n$ almost surely by the Corollary of page 112 in Tucker [33]. □

We now begin with a series of results designed to understand the tail behavior of $X_r^{(s)} = \sum_{j=-\infty}^{\infty} \beta_{jT+s} \circ Z_{r-jT-s}$ as well as sums of these variables. The first result we present is a simple modification of Theorem 8 in Hall [19] for the stationary case, but crucial for the characterization of the tail behavior of X_r.

Lemma 5.2. *Let* $\mathbf{Z}$ *be a* T*-periodic sequence verifying* (5.1)*. For fixed values of* $s = 0, ..., T-1$ *and* $r = 1, ..., T$*, it holds that, as* $x \to \infty$

$$P\big(X_r^{(s)} > x\big) \;\sim\; \breve{K}_{r-s}\, x^{\breve{\xi}_{r-s}} (1 + \breve{\lambda}_{r-s})^{-x} \;,$$

for $\xi_{r-s} \neq 1$*, with* $\breve{\lambda}_{r-s} = \frac{\lambda_{r-s}}{\beta^{(s)}}$*,* $\beta^{(s)} = \max\limits_{-\infty \le j \le \infty} \{\beta_{jT+s}\}$*,* $k_s = \#\big\{j\colon\; \frac{\beta_{jT+s}}{\beta^{(s)}} = 1\big\}$*,*

$$\breve{\xi}_{r-s} \;=\; \begin{cases} k_s\, \xi_{r-s} + k_s - 1 & \xi_{r-s} > -1 \\ \xi_{r-s} & \xi_{r-s} < -1 \end{cases} ,$$

$$K^*_{r-s} \;=\; \beta^{(s)}\, K_{r-s} \bigg(\frac{1 + \lambda_{r-s}}{\lambda_{r-s} + \beta^{(s)}} \bigg)^{\xi_{r-s}+1} ,$$

$$\breve{K}_{r-s} \;=\; \begin{cases} \breve{\lambda}_{r-s}^{k_s-1}\, {K^*_{r-s}}^{k_s} \dfrac{\big(\Gamma(\xi_{r-s}+1)\big)^{k_s}}{\Gamma\big(k_s(\xi_{r-s}+1)\big)}\; E\Big[(1+\breve{\lambda}_{r-s})^{\sum_{j' \notin \gamma_s} \beta_{j'} \circ Z_{r-s}}\Big] & \xi_{r-s} > -1 \\ k_s\, K^*_{r-s} \Big(E\big[(1+\breve{\lambda}_{r-s})\big] \Big)^{k_s-1} E\Big[(1+\breve{\lambda}_{r-s})^{\sum_{j' \notin \gamma_s} \beta_{j'} \circ Z_{r-s}}\Big] & \xi_{r-s} < -1 \end{cases} ,$$

with $j' = jT + s$ *and* $\gamma_s = \big\{i_1, ..., i_{k_s}\colon\; \beta_{i_h} = \beta^{(s)},\; h = 1, ..., k_s\big\}$.

We shall now obtain the tail behavior of F_{X_r}. For simplicity in notation we define $i_1, ..., i_T = 0, 1, ..., T-1$, being $i_1 \neq i_2 \neq ... \neq i_T$.

Lemma 5.3. *For the process defined in* (5.2) *it holds that, for* $r = 1, ..., T$*, as* $x \to \infty$*,*

$$P\big(X_r > x\big) \;\sim\; A^*_r\, x^{\xi^*_r} (1 + \lambda^*_r)^{-x} , \tag{5.4}$$

with $\lambda^*_r = \min(\breve{\lambda}_r, ..., \breve{\lambda}_{r-T+1})$*. Moreover, the constant* A^*_r *can be calculated as follows:*

1. *if* $\breve{\lambda}_{r-i_1} = \cdots = \breve{\lambda}_{r-i_T}$ *and*

(**a**) $\breve{\xi}_{r-i_1} = \cdots = \breve{\xi}_{r-i_T} < -1$ *then* $\xi^*_r = \breve{\xi}_{r-i_1}$ *and*

$$A^*_r \;=\; \begin{cases} C_{2,r} & T = 2 \\ C_{T,r} & T \ge 3 \end{cases} ,$$

with

$$C_{2,r} \;=\; \breve{K}_{r-i_1} E\Big[(1+\lambda^*_r)^{X_r^{(i_2)}}\Big] + \breve{K}_{r-i_2}\, E\Big[(1+\lambda^*_r)^{X_r^{(i_1)}}\Big] ,$$

$$C_{T,r} \;=\; C_{T-1,r}\, E\Big[(1+\lambda^*_r)^{X_r^{(i_T)}}\Big] + \breve{K}_{r-i_T}\, E\Big[(1+\lambda^*_r)^{\sum_{s=1}^{T-1} X_r^{(i_s)}}\Big] ;$$

(**b**) $\breve{\xi}_{r-i_1} > -1$, ..., $\breve{\xi}_{r-i_T} > -1$ *then* $\xi^*_r = \sum_{s=1}^T \breve{\xi}_{r-i_s} + T - 1$, *and*

$$A^*_r = \begin{cases} C^*_{2,r} & T = 2 \\ C^*_{T,r} & T \geq 3 \end{cases},$$

with

$$C^*_{2,r} = \lambda^*_r \breve{K}_{r-i_1} \breve{K}_{r-i_2} \frac{\Gamma\big(\breve{\xi}_{r-i_1} + 1\big)\,\Gamma\big(\breve{\xi}_{r-i_2} + 1\big)}{\Gamma\big(\breve{\xi}_{r-i_1} + \breve{\xi}_{r-i_2} + 2\big)},$$

$$C^*_{T,r} = C^*_{T-1,r}\,\lambda^*_r \breve{K}_{r-i_T} \frac{\Gamma\bigg(\sum\limits_{s=1}^{T-1} \breve{\xi}_{r-i_s} + T - 1\bigg)\,\Gamma\big(\breve{\xi}_{r-i_T} + 1\big)}{\Gamma\bigg(\sum\limits_{s=1}^{T} \breve{\xi}_{r-i_s} + T\bigg)};$$

2. *if* $\breve{\lambda}_{r-i_1} < \cdots < \breve{\lambda}_{r-i_T}$, *then* $\xi^*_r = \breve{\xi}_{r-i_1}$ *and*

$$A^*_r = \breve{K}_{r-i_1} \prod_{h=2}^T E\Big[(1+\lambda^*_r)^{X_r^{(i_h)}}\Big];$$

3. *if* $\breve{\lambda}_{r-i_1} < \cdots < \breve{\lambda}_{r-i_{l+1}} = \cdots = \breve{\lambda}_{r-i_{l+k}} < \breve{\lambda}_{r-i_{l+k+1}} < \cdots < \breve{\lambda}_{r-i_T}$, *then* $\xi^*_r = \breve{\xi}_{r-i_1}$ *and*

$$A^*_r = \breve{Q}^{(k)}_{r-i_1} \Bigg(\prod_{h=1}^{T-k-l} E\Big[(1+\lambda^*_r)^{X_r^{(i_{l+k+h})}}\Big]\Bigg)$$

with

$$\breve{Q}^{(k)}_{r-i_1} = \breve{K}_{r-i_1} \Bigg(\prod_{h=2}^{l} E\Big[(1+\lambda^*_r)^{X_r^{(i_h)}}\Big]\Bigg) E\Big[(1+\lambda^*_r)^{\sum_{h=1}^k X_r^{(i_{l+h})}}\Big]; \tag{5.5}$$

4. *if* $\breve{\lambda}_{r-i_1} < \cdots < \breve{\lambda}_{r-i_l} < \breve{\lambda}_{r-i_{l+1}} = \cdots = \breve{\lambda}_{r-i_T}$ *then* $\xi^*_r = \breve{\xi}_{r-i_1}$ *and*

$$A^*_r = \breve{Q}^{(T-l)}_{r-i_1},$$

with $\breve{Q}^{(\cdot)}_{r-i_1}$ *defined as in* (5.5);

5. *if* $\breve{\lambda}_{r-i_1} = \cdots = \breve{\lambda}_{r-i_l} < \breve{\lambda}_{r-i_{l+1}} < \cdots < \breve{\lambda}_{r-i_T}$ *and*

(**a**) $\breve{\xi}_{r-i_1} = \cdots = \breve{\xi}_{r-i_l} < -1$ *then* $\xi^*_r = \breve{\xi}_{r-i_1}$

$$A^*_r = \begin{cases} C_{2,r} & T = 2 \\ C_{T,r} \prod\limits_{h=l+1}^T E\Big[(1+\lambda^*_r)^{X_r^{(i_h)}}\Big] & 3 \leq l < T \end{cases};$$

(**b**) $\breve{\xi}_{r-i_1} > -1$, ..., $\breve{\xi}_{r-i_l} > -1$ *then* $\xi^*_r = \sum_{s=1}^l \breve{\xi}_{r-i_s} + l - 1$

$$A^*_r = \begin{cases} C^*_{2,r} & T = 2 \\ C^*_{T,r} \prod\limits_{h=l+1}^T E\Big[(1+\lambda^*_r)^{X_r^{(i_h)}}\Big] & 3 \leq l < T \end{cases}.$$

Proof: The result follows by applying repeatedly Lemma 7 in Hall [19] which is the discrete version of Theorem 7.1 in Rootzén [31], after some tedious calculations. □

We are now in conditions to obtain the limiting distribution of the maximum term of $\mathbf{X}$. An explicit expression for the sequence of norming constants (b_n) can be obtained though the following result. For clarification in notation we define $\check{\lambda} = \min_{1\le r\le T}\{\lambda_r^*\}$ and $(q_1, ..., q_k)$ the set of indices such that $\check{\lambda}/\lambda_{q_l}^* = 1$, for $l = 1, ..., k$ $(k \le T)$. In addition, we define $\check{\xi} = \max_{1\le l\le k}\{\xi_{q_l}^*\}$ and the set of indices $(p_1, ..., p_s)$ such that $\check{\xi}/\xi_{p_l}^* = 1$, with $l = 1, ..., s$, $(s \le k)$. Furthermore, let $A = \frac{1}{T}\sum_{j=1}^{s} A_{p_j}^*$.

Lemma 5.4. *For the T-periodic integer-valued sequence $\mathbf{X}$ given in* (5.2) *the normalizing constants b_n of Theorem 3.1 are given by*

$$b_n = \big(\ln(1+\check{\lambda})\big)^{-1} (\ln n + \check{\xi} \ln\ln n + \ln A) \; . \tag{5.6}$$

The demonstration of this lemma is based on the following result.

Lemma 5.5. *If a distribution function F belongs to the domain of attraction of an extreme value distribution, $(F \in D(G_\gamma(x)))$ and $F_* = F(x)(1+\epsilon(x))$ with $\lim_{x\to x_F} \epsilon(x) = 0$, then $F_* \in D(G_\gamma(x))$.*

Proof of Lemma 5.4: By Lemma 5.3, as $x \to \infty$

$$\begin{aligned} \frac{1}{T}\sum_{r=1}^{T} P\big(X_r > x\big) &\sim \frac{1}{T}\sum_{r=1}^{T} A_r^* \, x^{\xi_r^*} (1+\lambda_r^*)^{-x} \\ &= A\, x^{\check{\xi}} (1+\check{\lambda})^{-x} \left[1 + \sum_{l=s+1}^{T} \frac{A_{p_j}^*}{A} \left(\frac{1+\lambda_{q_l}^*}{1+\check{\lambda}}\right)^{-x} x^{\xi_{q_l}^* - \check{\xi}}\right] \\ &\sim A\, x^{\check{\xi}} (1+\check{\lambda})^{-x} \;, \end{aligned}$$

where the last step is justified by Lemma 5.5. □

Let $\hat{\mathbf{X}}$ be the associated independent T-periodic sequence of $\mathbf{X}$, i.e. $\hat{X}_1$, $\hat{X}_2$, ..., are independent random variables being the tail distribution of $\hat{X}_r$ as in (5.4) for $r = 1, ..., T$, and define $\hat{M}_n = \max(\hat{X}_n)$. Next result ensures that condition $D(u_n)$ holds for $\mathbf{X}$ with F_{Z_r} given as in (5.1).

Lemma 5.6. *Suppose that the T-periodic integer-valued sequence $\mathbf{X}$ given in* (5.2) *is defined by a.s. convergent sums and satisfies*

$$\begin{cases} \limsup\limits_{n\to\infty} P\big(\hat{M}_n \le x + b_n\big) \le e^{-\frac{1}{T}\sum_{r=1}^{T}(1+\lambda_r)^{-x}} \\ \liminf\limits_{n\to\infty} P\big(\hat{M}_n \le x + b_n\big) \ge e^{-\frac{1}{T}\sum_{r=1}^{T}(1+\lambda_r)^{-(x-1)}} \end{cases},$$

for all $x \in \mathbb{R}$ and some set of constants $\lambda_1, ..., \lambda_T > 0$, $b_n \in \mathbb{R}$. Then condition $D(x + b_n)$ holds for $\mathbf{X}$.

Proof: For any $\epsilon_n > 0$,

$$\sup_{i,j} \Big| F_{i_1,...,i_p,j_1,...,j_q}(u_n,...,u_n) - F_{i_1,...,i_p}(u_n,...,u_n)\, F_{j_1,...,i_q}(u_n,...,u_n) \Big| \;\le$$

$$\le\; \frac{n}{T} \sum_{r=1}^{T} P\Big(x + b_n - 2\,\epsilon_n < X_r \le x + b_n + 2\,\epsilon_n\Big)$$
$$+\; \frac{n}{T} \sum_{r=1}^{T} P\left(\left|\sum_{s=0}^{T-1} \sum_{j=[n^{\gamma_T}]+1}^{\infty} \beta_{jT+s} \circ Z_{r-jT-s}\right| > \epsilon_n\right)$$
$$+\; \frac{n}{T} \sum_{r=1}^{T} P\left(\left|\sum_{s=0}^{T-1} \sum_{j=-\infty}^{-[n^{\gamma_T}]-1} \beta_{jT+s} \circ Z_{r-jT-s}\right| > \epsilon_n\right)$$

where $j_1 - i_p \ge 2\, n^{\gamma_T}$, $\gamma \in (0,1)$. Note that

$$\frac{n}{T} \sum_{r=1}^{T} P\Big(x + b_n - 2\,\epsilon_n < X_r \le x + b_n + 2\,\epsilon_n\Big) \;=$$
$$=\; \frac{n}{T} \sum_{r=1}^{T} P\Big(X_r > x + b_n - 2\,\epsilon_n\Big) \;-\; \frac{n}{T} \sum_{r=1}^{T} P\Big(X_r > x + b_n + 2\,\epsilon_n\Big)\,.$$

Since $b_n \to \infty$ and $\epsilon \to 0$, if b_n is a normalizing constant for the maximum term, then $b_n^{\pm} = b_n \pm 2\,\epsilon_n$ are also constants for the maximum term. For each n and a fixed value of $r = 1,...,T$ $\;n\,P(X_r > x + b_n^-)$ and $n\,P(X_r > x + b_n^-)$ are step functions of x, with the same step width, and different location parameters, but whose difference converges to zero. Then

$$\frac{n}{T} \sum_{r=1}^{T} P\Big(x + b_n - 2\,\epsilon_n < X_r \le x + b_n + 2\,\epsilon_n\Big) \;\to\; 0\;, \qquad n \to \infty\;,$$

for all $x \in \mathbb{R}$, providing that

$$\text{(5.7)} \qquad \frac{n}{T} \sum_{r=1}^{T} P\left(\left|\sum_{s=0}^{T-1} \sum_{j=[n^{\gamma_T}]+1}^{\infty} \beta_{jT+s} \circ Z_{r-jT-s}\right| > \epsilon_n\right) \;\to\; 0$$

$$\text{(5.8)} \qquad \frac{n}{T} \sum_{r=1}^{T} P\left(\left|\sum_{s=0}^{T-1} \sum_{j=-\infty}^{-[n^{\gamma_T}]-1} \beta_{jT+s} \circ Z_{r-jT-s}\right| > \epsilon_n\right) \;\to\; 0$$

as $n \to \infty$, for some $\gamma_T \in (0,1)$ and $\epsilon = o(1)$ as $n \to \infty$, are sufficient conditions for $D(u_n)$. In proving (5.7) and (5.8) note that by Markov's inequality

$$\frac{n}{T} \sum_{r=1}^{T} P\left(\left|\sum_{s=0}^{T-1} \sum_{j=[n^{\gamma_T}]+1}^{\infty} \beta_{jT+s} \circ Z_{r-s}\right| > \epsilon_n\right) \;\le$$
$$\le\; \frac{n}{T} \sum_{r=1}^{T} \frac{E\left[\left(\sum_{s=0}^{T-1} \sum_{j=[n^{\gamma_T}]+1}^{\infty} \beta_{jT+s} \circ Z_{r-s}\right)^2\right]}{\epsilon_n^2}\;.$$

Using the properties of the thinning operation, for a fixed value $r=1,...,T$

$$\begin{aligned}
E\left[\left(\sum_{s=0}^{T-1}\sum_{j=[n^{\gamma_T}]+1}^{\infty}\beta_{jT+s}\circ Z_{r-s}\right)^2\right] &= \\
&= \sum_{s=0}^{T-1}\Bigl(\mathrm{Var}\bigl[Z_{r-s}\bigr]-E\bigl[Z_{r-s}\bigr]\Bigr)\sum_{j=[n^{\gamma_T}]+1}^{\infty}\beta^2_{jT+s} \\
&\quad+\left(\sum_{s=0}^{T-1}E\bigl[Z_{r-s}\bigr]\sum_{j=[n^{\gamma_T}]+1}^{\infty}\beta_{jT+s}\right)^2+\sum_{s=0}^{T-1}E\bigl[Z_{r-s}\bigr]\sum_{j=[n^{\gamma_T}]+1}^{\infty}\beta_{jT+s} \\
&= O\bigl(n^{-\gamma_T(\delta-1)}\bigr)\ ,
\end{aligned}$$

by (5.3). Hence by taking for instance $\epsilon_n=O\bigl((\ln n)^{-\zeta}\bigr)$, $\zeta>0$ and choosing $\gamma_T\in(0,1)$ such that $\gamma_T(\delta-1)>1$, we have that condition (5.7) is satisfied. For the expression in (5.8) the procedure is analogous. □

Next result provides sufficient conditions for $D'_T(u_n)$.

Lemma 5.7. *Denote* $n'_T=[n^{\gamma_T}]$ *and suppose that for some constants* $\gamma_T\in(0,1)$ *and* $\zeta>0$ *the following conditions hold, for* $u_n=x+b_n$, $\forall x\in\mathbb{R}$,

$$\frac{n}{T}\sum_{r=1}^{T}\sum_{t=r+1}^{2n'_T}P\Bigl(X_r+X_t>2u_n\Bigr)\to 0\ ,\qquad n\to\infty\ ; \tag{5.9}$$

$$\frac{n^2}{T}\sum_{r=1}^{T}P\left(\sum_{s=0}^{T-1}\sum_{j=n'_T+1}^{\infty}\beta_{jT+s}\circ Z_{r-jT-s}>\zeta\right)\to 0\ ,\qquad n\to\infty\ ; \tag{5.10}$$

$$\frac{n^2}{T}\sum_{r=1}^{T}P\left(\sum_{s=0}^{T-1}\sum_{j=-\infty}^{-n'_T-1}\beta_{jT+s}\circ Z_{r-jT-s}>\zeta\right)\to 0\ ,\qquad n\to\infty\ ; \tag{5.11}$$

$$\sum_{s=0}^{T-1}\sum_{j=n'_T+1}^{\infty}\beta_{jT+s}\circ Z_{r-jT-s}\xrightarrow{P}0\,,\qquad \sum_{s=0}^{T-1}\sum_{j=-\infty}^{-n'_T-1}\beta_{jT+s}\circ Z_{r-jT-s}\xrightarrow{P}0\ . \tag{5.12}$$

Then, condition $D'_T(u_n)$ *holds for the* T*-periodic integer-valued sequence* $\mathbf{X}$ *defined in* (5.2).

Proof: First note that, for a fixed value of $r=1,...,T$, $P(X_r>u_n,X_t>u_n)$ $\le P(X_r+X_t>2\,u_n)$ following from (5.9) that

$$\frac{n}{T}\sum_{r=1}^{T}\sum_{t=r+1}^{2n'_T}P\Bigl(X_r>u_n,\,X_t>u_n\Bigr)\to 0\ ,\qquad n\to\infty\ .$$

Next write $X'_r=\sum_{s=0}^{T-1}\sum_{j=-\infty}^{n'_T}\beta_{jT+s}\circ Z_{r-jT-s}$ and $X''_t=\sum_{s=0}^{T-1}\sum_{j=n'_T}^{\infty}\beta_{jT+s}\circ Z_{t-jT-s}$ so that X'_r and X''_t are independent for $t>2\,n'_T$. Following Rootzén [31], for a fixed value of $r=1,...,T$ it follows that

$$\begin{aligned}
P\Big(X_r>u_n,\,X_t>u_n\Big) \;\le\; & P\Big(X'_r>u_n-\zeta\Big)\,P\Big(X''_r>u_n-\zeta\Big)\\
&+P\left(\sum_{s=0}^{T-1}\sum_{j=n'_T+1}^{\infty}\beta_{jT+s}\circ Z_{r-jT-s}\;>\;\zeta\right)\\
&+P\left(\sum_{s=0}^{T-1}\sum_{j=-\infty}^{-n'_T-1}\beta_{jT+s}\circ Z_{t-jT-s}\;>\;\zeta\right),
\end{aligned}$$

and hence, writing $u^*_n=x-\zeta+b_n$ we have that

$$\begin{aligned}
&\frac{n}{T}\sum_{r=1}^{T}\sum_{t=2n'_T+1}^{[n/kT]T}P\Big(X_r>u_n,\,X_t>u_n\Big)\;\le\\
&\qquad\le\;\sum_{r=1}^{T}\frac{n^2}{k\,T}\;P\left(\sum_{s=0}^{T-1}\sum_{j=-\infty}^{n'_T}\beta_{jT+s}\circ Z_{r-jT-s}\;>\;u^*_n\right)\\
&\qquad\quad\times P\left(\sum_{s=0}^{T-1}\sum_{j=-n'_T}^{\infty}\beta_{jT+s}\circ Z_{r-jT-s}\;>\;u^*_n\right)\\
&\qquad\quad+\frac{n^2}{T}\sum_{r=1}^{T}P\left(\sum_{s=0}^{T-1}\sum_{j=n'_T+1}^{\infty}\beta_{jT+s}\circ Z_{r-jT-s}\;>\;\zeta\right)\\
&\qquad\quad+\frac{n^2}{T}\sum_{r=1}^{T}P\left(\sum_{s=0}^{T-1}\sum_{j=-\infty}^{-n'_T-1}\beta_{jT+s}\circ Z_{r-jT-s}\;>\;\zeta\right).
\end{aligned}$$

The last two terms tend to zero by (5.10) and (5.11). By the same line of reasoning as in Rootzén ([31], p. 622) it is easy to check that

$$\limsup_{n\to\infty}\frac{n}{T}\sum_{r=1}^{T}\sum_{t=2n'_T+1}^{[n/kT]T}P\Big(X_r>u_n,X_t>u_n\Big)\;\le\;\frac{1}{k}\times(\text{constant})\;\to\;0\,,\qquad k\to\infty\;.\;\square$$

The final result is formalized through the following theorem.

Theorem 5.1. *For the T-periodic integer-valued sequence* $\mathbf{X}$ *defined in* (5.2), *with* $k_s=1$, $s=0,...,T-1$ *and* $\xi_{r-s}\neq 1$ *for* $r=1...,T$, *it holds that*

$$\begin{cases}\limsup\limits_{n\to\infty}\,P\big(M_n\le x+b_n\big)\;\le\;e^{-\frac{1}{T}\sum_{r=1}^{T}(1+\lambda^*_r)^{-x}}\\[2ex]\liminf\limits_{n\to\infty}\,P\big(M_n\le x+b_n\big)\;\ge\;e^{-\frac{1}{T}\sum_{r=1}^{T}(1+\lambda^*_r)^{-(x-1)}}\end{cases},$$

with b_n *defined as in* (5.6).

Proof: First note that for $r = 1, ..., T$

$$X_r + X_t \;=\; \sum_{s=0}^{T-1} \sum_{j=-\infty}^{\infty} \Big(\beta_{jT+s} \circ Z_{r-jT-s} \;+\; \beta_{jT+s+t} \circ Z_{r-jT-s}\Big)\,.$$

For simplicity in notation we define $\lambda_{\min} = \min(\lambda_r, ..., \lambda_{r-T+1})$ and for $s = 0, ..., T-1$, $\beta_1^{(s)} = \max_j \big\{\beta_{jT+s} \colon\; j \notin \gamma_s\big\} < \beta^{(s)}$, $\beta_2^{(s)} = \max_t \big\{\max_j \{\beta_{jT+s} + \beta_{jT+s+t}\}\big\} < 2\,\beta^{(s)}$, and $\breve\beta_{\max} = \max_{0 \le s \le T-1} \big\{\breve\beta^{(s)}\big\}$ with $\breve\beta^{(s)} = \max\big\{\beta_1^{(s)},\, \beta_2^{(s)}/2\big\}$.

$$\begin{aligned} E\Big[(1{+}h)^{\beta_{jT+s}\circ Z_{r-s} + \beta_{jT+s+t}\circ Z_{r-s}}\Big] \;&=\; E\Big[E\big[(1{+}h)^{\beta_{jT+s}\circ Z_{r-s} + \beta_{jT+s+t}\circ Z_{r-s}} \big| Z_{r-s}\big]\Big] \\ &=\; \tilde P_{Z_{r-s}}\Big(\beta(jT{+}s,\, t)\, h^2 + (\beta_{jT+s} + \beta_{jT+s+t})\, h\Big) \end{aligned}$$

with $0 \le h < \lambda_{\min}$. Since, for $h \ge 0$ and $s = 0, ..., T-1$, $\beta(jT{+}s, t)\, h^2 + (\beta_{jT+s} + \beta_{jT+s+t})\, h \le \breve\beta_{\max} h^2 + 2\,\breve\beta_{\max} h$, the existence of $E\big[(1{+}h)^{\beta_{jT+s}\circ Z_{r-s} + \beta_{jT+s+t}\circ Z_{r-s}}\big]$ will be granted if it is possible to find an $h > 0$ such that $\breve\beta_{\max} h^2 + 2\,\breve\beta_{\max} h < \lambda_{\min}$.

$$\breve\beta_{\max} h^2 + 2\,\breve\beta_{\max} h - \lambda_{\min} = 0 \quad\iff\quad h = -1 \pm \sqrt{1 + \frac{\lambda_{\min}}{\breve\beta_{\max}}}\;.$$

Let $h_1 < 0 < h_2$ be the two solutions of this equation.

$$\begin{aligned} E\Big[(1{+}h)^{X_r + X_t}\Big] \;&=\; E\Big[(1{+}h)^{\sum_{s=0}^{T-1} \sum_{j=-\infty}^{\infty} (\beta_{jT+s}\circ Z_{r-jT-s} + \beta_{jT+s+t}\circ Z_{r-jT-s})}\Big] \\ &=\; \prod_{s=0}^{T-1} \left(\prod_{j=-\infty}^{[t/2]} \tilde P_{Z_{r-s}}\Big(\beta(jT{+}s,\, t)\, h^2 + (\beta_{jT+s} + \beta_{jT+s+t})\, h\Big) \right. \\ &\qquad \left. \times \prod_{j=[t/2]+1}^{\infty} \tilde P_{Z_{r-s}}\Big(\beta(jT{+}s, t)\, h^2 + (\beta_{jT+s} + \beta_{jT+s+t})\, h\Big) \right). \end{aligned} \tag{5.13}$$

Moreover, $\tilde P'_{Z_{r-s}}(\nu) = E\big[(1{+}\nu)^{Z_{r-s}}\big] < \infty$, if $0 < \nu < \lambda_{\min}$, and $\tilde P'_{Z_{r-s}}(\nu) \ge 1$ for $0 \le \nu \le \breve\beta_{\max} h^2 + 2\,\breve\beta_{\max} h$. By the mean value Theorem, $\tilde P_{Z_{r-s}}(\nu_1 + \nu_2) \le \tilde P_{Z_{r-s}}(\nu_1)\,(1 + C\nu_2)$, $\nu_1, \nu_2 > 0$, $\nu_1 + \nu_2 \le \breve\beta_{\max} h^2 + 2\,\breve\beta_{\max} h$, with

$$\begin{aligned} C \;=\; \sup &\left\{ \frac{\tilde P'_{Z_{r-s}}(\nu + x)}{\tilde P_{Z_{r-s}}(\nu)} \colon\; s = 0, ..., T-1,\; 0 < \nu + x < \breve\beta_{\max} h^2 + 2\breve\beta_{\max} h,\; \nu > 0,\; x > 0 \right\} \\ &< \infty\;. \end{aligned}$$

On the basis of this result we have for $\nu_1 = h(\beta_{jT+s} + \beta_{jT+s+t})$ and $\nu_2 = \beta(jT{+}s, t) h^2$

$$\begin{aligned} &\prod_{j=-\infty}^{[t/2]} \tilde P_{Z_{r-s}}\Big(\beta(jT{+}\,s,\, t)\, h^2 + (\beta_{jT+s} + \beta_{jT+s+t})\, h\Big) \;\le \\ &\quad \le\; \prod_{j=-\infty}^{[t/2]} \tilde P_{Z_{r-s}}\Big((\beta_{jT+s} + \beta_{jT+s+t})\, h\Big) \prod_{j=-\infty}^{[t/2]} \Big(1 + C\beta_{jT+s}\, \beta(jT{+}\,s,\, t)\, h^2\Big) \\ &\quad \le\; \prod_{j=-\infty}^{[t/2]} \tilde P_{Z_{r-s}}\big(\beta_{jT+s}\, h\big) \prod_{j=-\infty}^{-[t/2]} \Big(1 + C\beta_{jT+s}\, h\Big) \prod_{j=-\infty}^{[t/2]} \Big(1 + C\beta(jT{+}\,s,\, t)\, h^2\Big)\;. \end{aligned}$$

Noticing that $\tilde{P}_{Z_{r-s}}(\beta_{jT+s}\,h) = 1 + \beta_{jT+s}\,h\,E[Z_{r-s}]\,(1+o(1))$ and using (5.3), we may conclude that the last expression is bounded, uniformly in t. Using a similar argument for the second product in (5.13) we are lead to conclude that $E\big[(1+h)^{X_r+X_t}\big] < \infty$ for $r=1,...T$. By Lemma 5.4, $u_n \sim \frac{\ln n}{\ln(1+\check{\lambda})}$ as $n \to \infty$. By Bernstein's inequality

$$\begin{aligned} P\Big(X_r+X_t > 2\,u_n\Big) \;&\le\; E\Big[(1+h)^{X_r+X_t}\Big]\,(1+h)^{-2u_n} \\ &=\; O\Big((1+h)^{-2u_n}\Big) \\ &=\; O\Big(n^{\frac{\ln(1+2h+h^2)}{\ln(1+\check{\lambda})}}\Big) \\ &=\; o\Big(n^{-(1+\gamma_T)}\Big)\;, \end{aligned}$$

where the last equality follows by the arguments given in Hall ([19], p. 373). Moreover, in proving (5.10) and (5.11), it suffices to show by Bernstein's inequality that

$$E\Bigg[(1+h)^{\sum_{s=0}^{T-1}\sum_{j=Tn'_T+1}^{\infty}\beta_{jT+s}\circ Z_{r-jT-s}}\Bigg]$$

and

$$E\Bigg[(1+h)^{\sum_{s=0}^{T-1}\sum_{j=-\infty}^{-Tn'_T-1}\beta_{jT+s}\circ Z_{r-jT-s}}\Bigg]\;,$$

are bounded as $n \to \infty$, for some $h = n^{\eta} - 1$, $\eta > 0$. We can choose ζ and η such that $2 < \zeta\,\eta < \zeta\,\gamma_T(\delta-1)$. By (5.3), we have that

$$\begin{aligned} E\Bigg[(1+h)^{\sum_{s=0}^{T-1}\sum_{j=n'_T+1}^{\infty}\beta_{jT+s}\circ Z_{r-jT-s}}\Bigg] &= \prod_{s=0}^{T-1}\;\prod_{j=Tn'_T+1}^{\infty} E\Big[(1+h)^{\beta_{jT+s}\circ Z_{r-jT-s}}\Big] \\ &= \prod_{s=0}^{T-1}\;\prod_{j=n'_T+1}^{\infty} \tilde{P}_{Z_{r-s}}(\beta_{jT+s}\,h) \\ &= \prod_{s=0}^{T-1}\;\prod_{j=n'_T+1}^{\infty} \Big(1+\beta_{jT+s}\,h\,E\big[Z_{r-s}\big]\big(1+o(1)\big)\Big) \\ &< \infty\;, \end{aligned}$$

as $n \to \infty$ providing

$$\begin{aligned} &\frac{n^2}{T}\sum_{r=1}^{T} P\Bigg(\sum_{s=0}^{T-1}\;\sum_{j=n'_T+1}^{\infty} \beta_{jT+s}\circ Z_{r-jT-s} > \zeta\Bigg) \;\le \\ &\qquad\le\; \frac{n^2}{T}\,E\Bigg[(1+h)^{\sum_{s=0}^{T-1}\sum_{j=n'_T+1}^{\infty}\beta_{jT+s}\circ Z_{r-jT-s}}\Bigg]\,n^{-\zeta\eta} \;\to\; 0\;, \qquad n\to\infty\;. \end{aligned}$$

A similar procedure can be carried out to prove (5.11). We skip the details.

Finally, the proof is completed upon showing (5.12). Note that

$$\begin{aligned} E\left[\sum_{s=0}^{T-1}\sum_{j=n'_T+1}^{\infty}\beta_{jT+s}\circ Z_{r-jT-s}\right] &= \sum_{s=0}^{T-1}E\big[Z_{r-s}\big]\sum_{j=n'_T+1}^{\infty}\beta_{jT+s} \\ &< \sum_{s=0}^{T-1}E\big[Z_{r-s}\big]\sum_{j=n'_T+1}^{\infty}O(j^{-\delta}) \\ &= O\big(n^{\gamma_T(-\delta+1)}\big) \;\rightarrow\; 0\,, \qquad n\rightarrow\infty\,. \end{aligned}$$

Moreover

$$\begin{aligned} &\mathrm{Var}\left[\sum_{s=0}^{T-1}\sum_{j=n'_T+1}^{\infty}\beta_{jT+s}\circ Z_{r-jT-s}\right] = \\ &= \sum_{s=0}^{T-1}\sum_{j=n'_T+1}^{\infty}\beta^2_{jT+s}\Big(\mathrm{Var}\big[Z_{r-s}\big]-E\big[Z_{r-s}\big]\Big) + \sum_{s=0}^{T-1}\sum_{j=n'_T+1}^{\infty}\beta_{jT+s}\,E\big[Z_{r-s}\big] \\ &< \sum_{s=0}^{T-1}\sum_{j=n'_T+1}^{\infty}\Big(O(j^{-2\delta})+O(j^{-\delta})\Big) \\ &= O\big(n^{\gamma_T(-\delta+1)}\big) \;\rightarrow\; 0\,, \qquad n\rightarrow\infty\,. \end{aligned}$$

Hence, (5.12) holds concluding the proof. □

REFERENCES

[1] Al-Osh, M. and Alzaid, A. (1987). First order integer-valued autoregressive INAR(1) process, *J. Time Series Anal.*, **8**, 261–275.

[2] Al-Osh, M. and Alzaid, A. (1988). Integer-valued moving average (INMA) process, *Statistical Papers*, **29**, 281–300.

[3] Alpuim, M.T. (1988). An extremal Markovian sequence, *J. Appl. Probab.*, **26**, 219–232.

[4] Aly, E.-E. and Bouzar, N. (1994). Explicit stationary distributions for some Galton-Watson processes with immigration, *Commun. Statist.-Stochastic Models*, **10**, 499–517.

[5] Aly, E.-E. and Bouzar, N. (2005). Stationary solutions for integer-valued autoregressive processes, *International Journal of Mathematics and Mathematical Science*, **1**, 1–18.

[6] Anderson, C.W. (1970). Extreme value theory for a class of discrete distributions with applications to some stochastic processes, *J. Appl. Probab.*, **7**, 99–113.

[7] Ahn, S.; Gyemin, L. and Jongwoo, J. (2000). Analysis of the M/D/1-type queue based on an integer-valued autoregressive process, *Operational Research Letters*, **27**, 235–241.

[8] Blundell, R.; Griffith, R. and Windmeijer, F. (2002). Individual effects and dynamics in count data models, *Journal of Econometrics*, **108**, 113–131.

[9] Brännäs, K. (1995). Explanatory variables in the AR(1) count data model, *Umeå Economic Studies*, **381**.

[10] Brännäs, K. and Hall, A. (2001). Estimation in integer-valued moving average models, *Appl. Stochastic Models Bus. Ind.*, **17**, 277–291.

[11] Brännäs, K. and Hellström, J. (2001). Generalized integer-valued autoregression, *Econometric Reviews*, **20**, 425–443.

[12] Brännäs, K.; Hellström, J. and Nordström, J. (2002). A new approach to modelling and forecasting monthly guest nights in hotels, *Int. J. Forecast.*, **18**, 9–30.

[13] Chernick, M.; Hsing, T. and McCormick, W.J. (1991). Calculating the extremal index for a class of stationary sequences, *Adv. Appl. Prob.*, **6**, 711–731.

[14] Du, J.-G. and Li, Y. (1991). The integer valued autoregressive (INAR(p)) model, *J. Time Series Anal.*, **12**, 129–142.

[15] Ferreira, H. and Martins, A.P. (2003). The extremal index of sub-sampled sequences with strong local dependence, *Revstat*, **1**, 16–24.

[16] Garcia-Ferrer, A. and Queralt, R.A. (1997). A note on forecasting international tourism deman in Spain, *Int. J. Forecast.*, **13**, 539–549.

[17] Hall, A. (1996). Maximum term of a particular autoregressive sequence with discrete margins, *Commun. Statist.-Theory Meth.*, **25**, 721–736.

[18] Hall, A. (2001). Extremes of integer-valued moving averages models with regularly varying tails, *Extremes*, **4**, 219–239.

[19] Hall, A. (2003). Extremes of integer-valued moving averages models with exponential type-tails, *Extremes*, **6**, 361–379.

[20] Hall, A. and Moreira, O. (2006). A note on the extremes of a particular moving average count data model, *Statist. Probab. Lett.*, **76**, 135–141.

[21] Hooghiemstra, G.; Meester, L.E. and Hüsler, J. (1998). On the extremal index for the $M/M/s$ queue, *Commun. Statist.-Stochastic Models*, **14**, 611–621.

[22] Latour, A. (1988). Existence and stochastic structure of a non-negative integer-valued autoregressive processes, *J. Time Ser. Anal.*, **4**, 439–455.

[23] Leadbetter, M.R.; Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, New York.

[24] McCabe, B.P.M. and Martin, G.M. (2005). Bayesian Prediction of low count time series, *Int. J. Forecast.*, **21**, 315–330.

[25] McCormick, W.P. and Park, Y.S. (1992). Asymptotic analysis of extremes from autoregressive negative binomial processes, *J. Appl. Probab.*, **29**, 904–920.

[26] McKenzie, E. (1985). Some simple models for discrete variate time series, *Water Res. Bull.*, **21**, 645–650.

[27] McKenzie, E. (1986). Autoregressive analysis of extremes from autoregressive negative binomial processes, *J. Appl. Probab.*, **29**, 904–920.

[28] McKenzie, E. (2003). *Discrete variate time series*, In "Handbook of Statistics" (D.N. Shanbhag and C.R. Rao, Eds.), Elsevier Science, 573–606.

[29] Nordström, J. (1996). Tourism satellite account for Sweden 1992-93, *Tourism Economics*, **2**, 13–42.

[30] Quoreshi, A.M.M.S. (2006). Bivariate time series modelling of financial count data, *Commun. Statist.-Theory Meth.*, **35**, 1343–1358.

[31] Rootzén, H. (1986). Extreme value theory for moving average processes, *Ann. Probab.*, **14**, 612–652.

[32] Silva, I.; Silva, M.E.; Pereira, I. and Silva, N. (2005). Replicated INAR(1) processes, *Methodology and Computing in Applied Probability*, **7**, 517–542.

[33] Tucker, H. (1967). *A Graduate Course in Probability*, Academic Press, New York.

[34] Zhou, J. and Basawa, I.V. (2005). Least-squared estimation for bifurcation autoregressive processes, *Statist. Probab. Lett.*, **74**, 77–88.

[35] Zhu, R. and Joe, H. (2003). A new type of discrete self-decomposability and its applications to continuous-time Markov processes for modeling count data time series, *Stoch. Models*, **19**, 235–254.

REVSTAT – STATISTICAL JOURNAL

Background

Statistical Institute of Portugal (INE), well aware of how vital a statistical culture is in understanding most phenomena in the present-day world, and of its responsibility in disseminating statistical knowledge, started the publication of the scientific statistical journal *Revista de Estatística*, in Portuguese, publishing three times a year papers containing original research results, and application studies, namely in the economic, social and demographic fields.

In 1998 it was decided to publish papers also in English. This step has been taken to achieve a larger diffusion, and to encourage foreign contributors to submit their work.

At the time, the Editorial Board was mainly composed by Portuguese university professors, being now composed by national and international university professors, and this has been the first step aimed at changing the character of *Revista de Estatística* from a national to an international scientific journal.

In 2001, the *Revista de Estatística* published three volumes special issue containing extended abstracts of the invited contributed papers presented at the 23$^{\text{rd}}$ European Meeting of Statisticians.

The name of the Journal has been changed to REVSTAT – STATISTICAL JOURNAL, published in English, with a prestigious international editorial board, hoping to become one more place where scientists may feel proud of publishing their research results.

— The editorial policy will focus on publishing research articles at the highest level in the domains of Probability and Statistics with emphasis on the originality and importance of the research.

— All research articles will be refereed by at least two persons, one from the Editorial Board and another, external.

— The only working language allowed will be English.

— Two volumes are scheduled for publication, one in June and the other in November.

— On average, four articles will be published per issue.

Aims and Scope

The aim of REVSTAT is to publish articles of high scientific content, in English, developing innovative statistical scientific methods and introducing original research, grounded in substantive problems.

REVSTAT covers all branches of Probability and Statistics. Surveys of important areas of research in the field are also welcome.

Abstract/indexed in

REVSTAT is expected to be abstracted/indexed at least in *Current Index to Statistics*, *Mathematical Reviews*, *Statistical Theory and Method Abstracts*, and *Zentralblatt für Mathematic.*

Instructions to Authors, special-issue editors and publishers

Papers may be submitted in two different ways:

- By sending a paper copy to the Executive Editor and one copy to one of the two Editors or Associate Editors whose opinion the author(s) would like to be taken into account, together with a postscript or a PDF file of the paper to the e-mail: revstat@fc.ul.pt.
- By sending a paper copy to the Executive Editor, together with a postscript or a PDF file of the paper to the e-mail: revstat@fc.ul.pt.

Submission of a paper means that it contains original work that has not been nor is about to be published elsewhere in any form.

Submitted manuscripts (text, tables and figures) should be typed only in black, on one side, in double spacing, with a left margin of at least 3 cm and not have more than 30 pages.

The first page should include the name, affiliation and address of the author(s) and a short abstract with the maximum of 100 words, followed by the key words up to the limit of 6, and the AMS 2000 subject classification.

Authors are obliged to write the final version of accepted papers using LaTeX, in the REVSTAT style.

This style (REVSTAT.sty), and examples file (REVSTAT.tex), which may be download to *PC Windows System* (Zip format), *Mackintosh*, *Linux* and *Solaris Systems* (StuffIt format), and *Mackintosh System* (BinHex Format), are available in the REVSTAT link of the National Statistical Institute's Website: http://www.ine.pt/revstat.html

Additional information for the authors may be obtained in the above link.

Accepted papers

Authors of accepted papers are requested to provide the LaTeX files and also a postscript (PS) or an acrobat (PDF) file of the paper to the Secretary of REVSTAT: liliana.martins@ine.pt.

Such e-mail message should include the author(s)'s name, mentioning that it has been accepted by REVSTAT.

The authors should also mention if encapsulated postscript figure files were included, and submit electronics figures separately in .tiff, .gif, .eps or .ps format. Figures must be a minimum of 300 dpi.

Also send always the final paper version to:

Maria José Carrilho
Executive Editor, REVSTAT – STATISTICAL JOURNAL
Instituto Nacional de Estatística
Av. António José de Almeida
1000-043 LISBOA
PORTUGAL

Copyright and Reprints

Upon acceptance of an article, the author(s) will be asked to transfer copyright of the article to the publisher, the INE, in order to ensure the widest possible dissemination of information, namely through the National Statistical Institute's Website (http://www.ine.pt).

After assigning the transfer copyright form, authors may use their own material in other publications provided that the REVSTAT is acknowledged as the original place of publication. The Executive Editor of the Journal must be notified in writing in advance.

Authors of articles published in the REVSTAT will be entitled to one free copy of the respective issue of the Journal and twenty-five reprints of the paper are provided free. Additional reprints may be ordered at expenses of the author(s), and prior to publication.