ISSN 1645-6726

INSTITUTO NACIONAL DE ESTATÍSTICA
PORTUGAL

REVSTAT
Statistical Journal

Volume 1
2003

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

- **PUBLISHER**
  - *Instituto Nacional de Estatística (INE)*
    Av. António José de Almeida, 2
    1000-043 LISBOA
    PORTUGAL
    Tel.: (0351) 21 842 61 00
    Fax: (0351) 21 842 63 64
    Web site: http://www.ine.pt
- **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**
  - *500 copies*
- **LEGAL DEPOSIT REGISTRATION**
  N.º 191915/03

PRICE

[VAT 5% included]

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

EDITORIAL

Some years ago INE — Instituto Nacional de Estatística — launched a statistical journal, *Revista de Estatística / Statistical Review*, intended to fill the gap of scientific publication in the area of statistics, in Portuguese. This was thought, at the time, to be an important step towards the normalization of Portuguese scientific terms in the area. The editorial board, inspired by D.R. Cox authoritative paper on "The Current Position of Statistics" (*ISI Review*, 1997), also tried to bring together the academic and official statistics interests, through a careful planning of invited papers, and we feel happy to acknowledge that those initial goals have been fully accomplished.

When the editorial board decided to publish a special issue containing the extended abstracts of invited and contributed papers presented at the 23rd European Meeting of Statisticians, the time seemed to be ripe for internationalization: aside from the printing know-how, some of us had gained, under the leadership of Anthony Davison, experience in the editorial and refereeing process, and could make the first contacts to build up a sound editorial board during the meeting.

Under the invitation of INE, I have accepted in 2002 to take charge of the first steps of this new journal, REVSTAT. Together with the direction of INE, a rich board of associate editors has been chosen — and we take the opportunity to thank the very prompt and friendly response of the statistical community. This editorial board, representing a broad field of interests in Probability, Statistics and their applications, reflects our purpose of accepting contributions in any of these areas, basing our decisions only on creativity and merit, upon the recommendation of referees. The launching of REVSTAT has been slower than we would hope for, and we renew our invitation to all researchers in Probability, Statistics, or their applications, to consider REVSTAT as a suitable scientific journal to publish their achievements. Our commitment is to deal with submissions fast, and to guarantee quality of the new journal through the appraisal of expert referees.

Once again, in my name and in the name of INE, thanks to all those who generously have donated their time and efforts to this enterprise.

M. Ivette Gomes

INDEX

A RANDOM-EFFECTS LOG-LINEAR MODEL WITH POISSON DISTRIBUTIONS

Authors: Maria Alexandra Seco
– Department of Mathematics of ESTG, Instituto Politécnico de Leiria, Portugal (aseco@estg.ipleiria.pt)

António St.Aubyn
– Department of Mathematics of ISA, Universidade Técnica de Lisboa, Portugal (staubyn@isa.utl.pt)

Received: October 2002 Revised: April 2003 Accepted: June 2003

Abstract:

- In several applications data are grouped and there are within-group correlations. With continuous data, there are several available models that are often used; with counting data, the Poisson distribution is the natural choice. In this paper a mixed log-linear model based on a Poisson–Poisson conditional distribution is presented. The initial model is a conditional model for the mean of the response variable, and the marginal model is formed thereafter. Random effects with Poisson distribution are introduced and a variance-covariance matrix for the response vector is formed embodying the covariance structure induced by the grouping of the data.

Key-Words:

- *log-linear models; grouped data; random effects; mixed models; overdispersion; iterative reweighted generalized least squares.*

AMS Subject Classification:

- 62J02, 62J12, 62J99, 62P12.

1. INTRODUCTION

In many applications in biology, agriculture, engineering and economics, for instance, grouped data reveal within-group correlation. For continuous data there are several available models which are used. These include Variance Component Models and Mixed Models (Laird and Ware [2], Pinheiro and Bates [6]) which embody fixed and random effects. Both models are based on the Multivariate Normal distribution, which has friendly properties, as the marginal and conditional distributions are still Normal.

Goldstein [1] gives several examples where ignoring the group structure can lead to imprecise estimates, confidence intervals and significant tests. He alerts that grouped data should be modelled respecting its particular structure.

A mixed log-linear model based on the Poisson–Poisson hierarchical distribution will be presented for grouped count data. The initial model is a conditional model for the mean of Y, and the marginal model is derived afterwards. It will be shown that building the model this way and introducing random Poisson effects, is a means of introducing overdispersion in a pseudo-Poisson model (overdispersion is said to exist when $\mathrm{var}(Y) = \phi\, E(Y),\ \phi > 1$). Moreover, the variance-covariance matrix is built for the response vector $\mathbf{Y}$, which embodies the covariance structure induced by the grouping of the data.

Several authors (McCulloch and Searle [5], Vonesh and Chinchilli [7]) have made references to some mixed models based on Poisson–Gamma or Bernoulli–Beta distributions as they are conjugate families. Starting from a model where $Y_{ij}|b_i$ follows a Poisson law and b_i a Gamma one, and as the $Y_{ij}|b_i$ are conditionally independent, the derived density function for $\mathbf{Y}_i$, a density product, is computationally unfriendly. In this paper a practical and simpler approach is proposed, that starts from a Poisson–Poisson model and uses the marginal moments of the response variable. The parameters are then estimated, with the iterative, non-linear, generalized least squares method.

In this presentation, attention is given to the simplest case of a single random effect. This is not as restrictive as it seems because, as was referred above, it portrays a situation of overdispersion with within-group correlation.

2. THE LOG-LINEAR CONDITIONAL MODEL

Consider M groups, with n_i observations per group (counts), where a within-group correlation structure is expected. Define the mixed log-linear model

$$\log\left[E(\mathbf{Y}_i|b_i)\right] \;=\; \mathbf{X}_i\boldsymbol{\beta} + \mathbf{1}_{n_i} b_i\;, \qquad i = 1,...,M, \quad j = 1,...,n_i\;. \tag{2.1}$$

Here $\mathbf{Y}_i = [Y_{i1} \; ... \; Y_{in_i}]^T$ is a random vector $n_i{\times}1$, b_i is a random variable $(1{\times}1)$, $\mathbf{X}_i$ is a known model matrix of order $n_i{\times}p$, $\boldsymbol{\beta}$ is a $p{\times}1$ vector of unknown fixed parameters and $\mathbf{1}_{n_i}$ is a vector $n_i{\times}1$ of ones. $\mathbf{Y}_i$ and b_i are independent for different i's.

Consider that each $Y_{ij}|b_i$ is a random variable conditionally independent for different j's following the Poisson law

$$Y_{ij}|b_i \sim P\Big(\exp\left\{\mathbf{x}_j^T\boldsymbol{\beta} + b_i\right\}\Big) \; , \qquad i = 1, ..., M, \;\; j = 1, ..., n_i \; ,$$

where $\mathbf{x}_j^T$ is row j of the model matrix $\mathbf{X}_i$ and $\boldsymbol{\beta}$ is the same as before. Let

$$\begin{aligned} &b_i \sim P(\theta_i) \; , \\ &\theta_i > 0 \; , \end{aligned}$$

independent for different i's.

Hence $E(b_i) = \mathrm{var}(b_i) = \theta_i, \;\; i = 1, ..., M$.

Note that $\mathbf{Y}$, the vector of all the random variables, is an $N{\times}1$ vector which is partioned as M components $\mathbf{Y}_i$, each of which is a random n_i-vector, $i = 1..., M$,

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_M \end{bmatrix} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{Mn_M} \end{bmatrix} .$$

N is the total number of observations, $N = \sum_{i=1}^{M} n_i$. Note that $\mathrm{cov}(Y_{ij}, Y_{ik}) \neq 0$, $j \neq k$, i.e., the Y_{ij} for $j = 1, .., n_i$, are not independent as they represent the same group, but they are independent for different i's (groups). Each b_i random variable is introduced to portray the situation of within group correlation for group i, $i = 1, ..., M$.

3. THE MARGINAL MODEL FOR Y

The parameter estimates are computed from a model based on the marginal moments of $\mathbf{Y}$. The mean value, variance and covariance of the $\mathbf{Y}$ marginals are then computed.

Let Y_{ij} be the variable that corresponds to the j-th observation in group i, $i = 1, ..., M, \; j = 1, ..., n_i$. As it is assumed that $Y_{ij}|b_i \sim P\left(\exp\left\{\mathbf{x}_j^T\boldsymbol{\beta} + b_i\right\}\right)$

and $b_i \sim P(\theta_i)$,

$$\begin{aligned} E(Y_{ij}) &= E_{b_i}\Big[E(Y_{ij}|b_i)\Big] \\ &= E\Big(\exp\{\mathbf{x}_j^T\boldsymbol{\beta} + b_i\}\Big) \\ &= \exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}M_{b_i}(1) \ , \end{aligned}$$

where $M_{b_i}(\cdot)$ is the b_i moment generating function. Then

$$\begin{aligned} E(Y_{ij}) &= \exp\{\mathbf{x}_j^T\boldsymbol{\beta}\} \, \exp\{(e-1)\,\theta_i\} \\ &= \exp\Big\{\mathbf{x}_j^T\boldsymbol{\beta} + (e-1)\,\theta_i\Big\} \ , \end{aligned}$$

where e is the Neper number, and

$$\log\left[E(Y_{ij})\right] = \mathbf{x}_j^T\boldsymbol{\beta} + (e-1)\,\theta_i \ .$$

Note the offset, $(e-1)\,\theta_i$, that comes out in the marginal expected value of Y_{ij}, derived from the introduction of the random effect b_i in the conditional model.

For the Y_{ij} variance,

$$\begin{aligned} \mathrm{var}(Y_{ij}) &= \mathrm{var}\Big[E(Y_{ij}|b_i)\Big] + E\Big[\mathrm{var}(Y_{ij}|b_i)\Big] \\ &= \mathrm{var}\Big(\exp\{\mathbf{x}_j^T\boldsymbol{\beta} + b_i\}\Big) + E\Big(\exp\{\mathbf{x}_j^T\boldsymbol{\beta} + b_i\}\Big) \\ &= E\Big(\exp\Big\{2(\mathbf{x}_j^T\boldsymbol{\beta} + b_i)\Big\}\Big) - \Big[E\Big(\exp\Big\{\mathbf{x}_j^T\boldsymbol{\beta} + b_i\Big\}\Big)\Big]^2 \\ &\quad + E\Big(\exp\Big\{\mathbf{x}_j^T\boldsymbol{\beta} + b_i\Big\}\Big) \\ &= \exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}\Big[\exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}M_{b_i}(2) - \exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}(M_{b_i}(1))^2 + M_{b_i}(1)\Big] \\ &= E(Y_{ij})\left[\exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}\frac{M_{b_i}(2)}{M_{b_i}(1)} - \exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}M_{b_i}(1) + 1\right] . \end{aligned}$$

It is known that the distribution of Y_{ij} is not Poisson, but it may be called pseudo-Poisson with overdispersion. Note that

$$\mathrm{var}(Y_{ij}) = \varphi\, E(Y_{ij}) \ ,$$

where the contribution of b_i for the "overdispersion component" is highlighted,

$$\varphi = \exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}\frac{M_{b_i}(2)}{M_{b_i}(1)} - \exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}M_{b_i}(1) + 1 \ .$$

Finally,

$$\begin{aligned}
\operatorname{var}(Y_{ij}) \;&=\; \exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}\exp\{(e-1)\,\theta_i\}\times\\
&\times\left[\exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}\,\frac{\exp\{(e^2-1)\,\theta_i\}}{\exp\{(e-1)\,\theta_i\}} \;-\; \exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}\exp\{(e-1)\,\theta_i\} \;+\; 1\right]\\
&=\; \exp\{2\,\mathbf{x}_j^T\boldsymbol{\beta}\}\,\Big[\exp\{(e^2-1)\,\theta_i\} - \exp\{2\,(e-1)\,\theta_i\}\Big]\\
&\quad+\exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}\exp\{(e-1)\,\theta_i\}\\
&=\; C(\theta_i)\exp\{2\,\mathbf{x}_j^T\boldsymbol{\beta}\} + K(\theta_i)\exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}\ ,
\end{aligned}$$

where

$$C(\theta_i) \;=\; \exp\{(e^2-1)\,\theta_i\} - \exp\{2(e-1)\,\theta_i\}\ ,$$

and

$$K(\theta_i) = \exp\{(e-1)\,\theta_i\}\ . \tag{3.1}$$

For the covariance, with $j \neq k$, and for the i group,

$$\begin{aligned}
\operatorname{cov}(Y_{ij},Y_{ik}) \;&=\; \operatorname{cov}\Big[E(Y_{ij}|b_i),E(Y_{ik}|b_i)\Big] + E\Big[\operatorname{cov}(Y_{ij},Y_{ik}|b_i)\Big]\\
&=\; \operatorname{cov}\Big[E(Y_{ij}|b_i),E(Y_{ik}|b_i)\Big] + E(0)\\
&=\; \exp\{\mathbf{x}_j^T\boldsymbol{\beta}+\mathbf{x}_k^T\boldsymbol{\beta}\}\operatorname{var}[\exp\{b_i\}]\\
&=\; \exp\{\mathbf{x}_j^T\boldsymbol{\beta}+\mathbf{x}_k^T\boldsymbol{\beta}\}\,\Big[M_{b_i}(2) - (M_{b_i}(1))^2\Big]\\
&=\; \exp\{\mathbf{x}_j^T\boldsymbol{\beta}+\mathbf{x}_k^T\boldsymbol{\beta}\}\,\Big[\exp\{(e^2-1)\,\theta_i\} - \exp\{2(e-1)\,\theta_i\}\Big]\\
&=\; C(\theta_i)\,\exp\{\mathbf{x}_j^T\boldsymbol{\beta}+\mathbf{x}_k^T\boldsymbol{\beta}\}\ .
\end{aligned}$$

3.1. Parameter estimation

The parameter estimates are obtain minimizing

$$\sum_{i=1}^{M}\Big(\mathbf{y}_i - K(\theta_i)\exp\{\mathbf{X}_i\,\boldsymbol{\beta}\}\Big)^T\,\mathbf{V}_i^{-1}\Big(\mathbf{y}_i - K(\theta_i)\exp\{\mathbf{X}_i\boldsymbol{\beta}\}\Big) \tag{3.2}$$

where $\mathbf{y}_i$ is a n_i-dimension vector of responses and $K(\theta_i) = \exp\{(e-1)\,\theta_i\}$, $i = 1,...,M$. Matrix $\mathbf{V}_i$, the variance-covariance matrix of $\mathbf{Y}_i$, is symmetric of order $n_i \times n_i$, with generic element υ_{jk}:

$$\begin{aligned}
&\mathbf{V}_i = [\upsilon_{jk}]_{j,k=1,\ldots,n_i}\ , \qquad i = 1,...,M\ ,\\
&\upsilon_{jj} \;=\; C(\theta_i)\exp\{2\,\mathbf{x}_j^T\boldsymbol{\beta}\} + K(\theta_i)\exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}\ ,\\
&\upsilon_{jk} \;=\; C(\theta_i)\exp\{\mathbf{x}_j^T\boldsymbol{\beta}+\mathbf{x}_k^T\boldsymbol{\beta}\}\ , \qquad j \neq k\ .
\end{aligned}$$

As $\mathbf{V}_i$ depends on $\boldsymbol{\beta}$ and θ_i it becomes necessary to apply an iterative method. It is possible to apply the IRGLS — Iterative Reweighted Generalized Least Squares method. This is an improvement of the Estimated Generalized Least Squares (EGLS) procedure which iterates using updated values of $\mathbf{V}_i(\hat{\boldsymbol{\beta}}, \hat{\theta}_i)$ to wash out any inefficiency associated with the initial estimates of $\boldsymbol{\beta}$ and θ_i. At each iteration $\mathbf{V}_i$ is updated using current estimates of the parameters. IRGLS may be applied to small or moderate samples (Vonesh and Chinchilli [7]).

Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_M)$ and $\boldsymbol{\tau} = (\boldsymbol{\beta}, \boldsymbol{\theta})$. IRGLS corresponds to solving a set of generalized estimating equations (Liang and Zeger [3]):

$$\mathbf{U}(\boldsymbol{\tau}) = \sum_{i=1}^{M} \mathbf{U}_i(\boldsymbol{\beta}, \theta_i) = \mathbf{0} \ ,$$

or

$$\sum_{i=1}^{M} \left\{ \mathbf{D}_i^T(\boldsymbol{\beta}, \theta_i)\, \mathbf{V}_i^{-1}(\boldsymbol{\beta}, \theta_i) \left(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}, \theta_i) \right) \right\} \;=\; \mathbf{0} \ , \tag{3.3}$$

where $\mathbf{D}_i(\boldsymbol{\beta}, \theta_i) = \dfrac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta}, \theta_i)}{\partial(\boldsymbol{\beta}, \theta_i)^T}$ and $\boldsymbol{\mu}_i = E(\mathbf{Y}_i)$. A solution to (3.3) can be obtained using the Gauss–Newton algorithm whereby estimates of $\boldsymbol{\tau}$ are updated as

$$\hat{\boldsymbol{\tau}}^{(t+1)} \;=\; \hat{\boldsymbol{\tau}}^{(t)} + \Omega(\hat{\boldsymbol{\tau}}^{(t)})\, \mathbf{U}(\hat{\boldsymbol{\tau}}^{(t)}) \ ,$$

with

$$\Omega(\hat{\boldsymbol{\tau}}^{(t)}) \;=\; \left[\sum_{i=1}^{M} \mathbf{D}_i^T(\hat{\boldsymbol{\beta}}^{(t)}, \hat{\theta}_i^{(t)})\ \mathbf{V}_i^{-1}(\hat{\boldsymbol{\beta}}^{(t)}, \hat{\theta}_i^{(t)})\ \mathbf{D}_i(\hat{\boldsymbol{\beta}}^{(t)}, \hat{\theta}_i^{(t)}) \right]^{-1} .$$

3.2. Inference and asymptotic properties

It is known (Vonesh and Chinchilli [7]) that the $\boldsymbol{\tau}$ IRGLS estimator, under regularity conditions that are usually satisfied, is asymptotically strongly consistent and has a Normal asymptotic distribution with mean zero and variance matrix given by:

$$\Omega(\hat{\boldsymbol{\tau}}) \;=\; \mathrm{var}(\hat{\boldsymbol{\tau}}) \;=\; \left[\sum_{i=1}^{M} \mathbf{D}_i^T(\boldsymbol{\beta}, \theta_i)\, \mathbf{V}_i^{-1}(\boldsymbol{\beta}, \theta_i)\, \mathbf{D}_i(\boldsymbol{\beta}, \theta_i) \right]^{-1} .$$

In terms of inference $\mathrm{var}(\hat{\boldsymbol{\tau}})$ is replaced by

$$\hat{\Omega}(\hat{\boldsymbol{\tau}}) \;=\; \left[\sum_{i=1}^{M} \mathbf{D}_i^T(\hat{\boldsymbol{\beta}}, \hat{\theta}_i)\, \mathbf{V}_i^{-1}(\hat{\boldsymbol{\beta}}, \hat{\theta}_i)\, \mathbf{D}_i(\hat{\boldsymbol{\beta}}, \hat{\theta}_i) \right]^{-1} .$$

To protect against possible misspecification of $\mathbf{V}_i(\boldsymbol{\beta},\theta_i)$ one can use, if necessary, robust inference based on the robust estimator suggested by Liang and Zeger [3],

$$\hat{\Omega}_R(\hat{\boldsymbol{\tau}}) \;=\; \hat{\Omega}(\hat{\boldsymbol{\tau}}) \left[\sum_{i=1}^{M} \mathbf{U}_i(\hat{\boldsymbol{\beta}},\hat{\theta}_i)\, \mathbf{U}_i^T(\hat{\boldsymbol{\beta}},\hat{\theta}_i)\right] \hat{\Omega}(\hat{\boldsymbol{\tau}}) \;,$$

where

$$\mathbf{U}_i(\hat{\boldsymbol{\beta}},\hat{\theta}_i) \;=\; \mathbf{D}_i^T(\hat{\boldsymbol{\beta}},\hat{\theta}_i)\, \mathbf{V}_i^{-1}(\hat{\boldsymbol{\beta}},\hat{\theta}_i) \Big(\mathbf{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}},\hat{\theta}_i)\Big) \;.$$

3.3. Computational issues and model linearization

To optimize the objective function (3.2), it is advisable, in practical and computational terms, to find a linearization of the model that transforms the expected value of the variable in a linear function of the parameters $\boldsymbol{\beta}$, as it simplifies the objective function and the variance-covariance matrix considered in it.

Let $\mu_{ij} = E(Y_{ij}) = K(\theta_i)\exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}$ and $\eta_{ij} = \log(\mu_{ij})$. Consider the new random variable

$$\zeta_{ij} \;=\; \eta_{ij} - \log\left[K(\theta_i)\right] + (Y_{ij} - \mu_{ij})\, \frac{d\eta_{ij}}{d\mu_{ij}} \;;$$

then

$$E(\zeta_{ij}) \;=\; \eta_{ij} - \log\left[K(\theta_i)\right] \;=\; \mathbf{x}_j^T\boldsymbol{\beta} \;,$$

which is linear in $\boldsymbol{\beta}$.

Or

$$\begin{aligned} \zeta_{ij} \;&=\; \mathbf{x}_j^T\boldsymbol{\beta} \;+\; (Y_{ij} - \mu_{ij}) \times \frac{1}{\mu_{ij}} \\ &=\; \mathbf{x}_j^T\boldsymbol{\beta} \;+\; \frac{Y_{ij}}{K(\theta_i)\exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}} - 1 \;. \end{aligned}$$

Let $\boldsymbol{\zeta}$ be the $N\times 1$ vector, $\boldsymbol{\zeta} = \left[\boldsymbol{\zeta}_1^T\ \boldsymbol{\zeta}_2^T\ ...\ \boldsymbol{\zeta}_M^T\right]^T$, $\boldsymbol{\zeta}_i = \left[\zeta_{i1}\ \zeta_{i2}\ ...\ \zeta_{in_i}\right]^T$, $i = 1,...,M$ and $\mathbf{W}$ the block diagonal variance-covariance matrix in $\boldsymbol{\zeta}$, $\mathbf{W} = \bigoplus_{i=1}^{M} \mathbf{W}_i$, where $\mathbf{W}_i$ is a matrix $n_i\times n_i$, symmetric, with generic element w_{jk}. For each group i, $i = 1,...,M$ and $j = 1,...,n_i$,

$$\begin{aligned} w_{jj} \;&=\; \mathrm{var}(\zeta_{ij}) \\ &=\; \left[\frac{1}{K(\theta_i)\exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}}\right]^2 \mathrm{var}(Y_{ij}) \\ &=\; \frac{C(\theta_i)}{\left[K(\theta_i)\right]^2} + \frac{1}{K(\theta_i)\exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}} \;. \end{aligned}$$

On the other hand, for $j \neq k$ in the i group,

$$\begin{aligned} w_{jk} &= \operatorname{cov}(\zeta_{ij}, \zeta_{ik}) \\ &= \frac{\operatorname{cov}(Y_{ij}, Y_{ik})}{\left[K(\theta_i)\exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}\right]\left[K(\theta_i)\exp\{\mathbf{x}_k^T\boldsymbol{\beta}\}\right]} \\ &= \frac{C(\theta_i)}{[K(\theta_i)]^2} \ . \end{aligned}$$

The minimization problem (3.2) becomes equivalent to,

$$\min(\boldsymbol{\zeta} - \mathbf{X}\boldsymbol{\beta})^T\,\mathbf{W}^{-1}(\boldsymbol{\zeta} - \mathbf{X}\boldsymbol{\beta}) \ , \tag{3.4}$$

where $\mathbf{X}$ is a model matrix of order $N{\times}p$, $\boldsymbol{\zeta}$ is a $N{\times}1$ vector, $\boldsymbol{\zeta} = [\boldsymbol{\zeta}_1^T\ \boldsymbol{\zeta}_2^T\ ...\ \boldsymbol{\zeta}_M^T]^T$, $\boldsymbol{\zeta}_i = [\zeta_{i1}\ \zeta_{i2}\ ...\ \zeta_{in_i}]^T$, $i = 1, ..., M$ and $\mathbf{W} = \bigoplus_{i=1}^{M}\mathbf{W}_i$, $\mathbf{W}_i = [w_{jk}]_{j,k=1,...,n_i}$, with

$$\begin{aligned} w_{jj} &= \frac{C(\theta_i)}{[K(\theta_i)]^2} + \frac{1}{K(\theta_i)\exp\{\mathbf{x}_j^T\boldsymbol{\beta}\}} \ , \\ w_{jk} &= \frac{C(\theta_i)}{[K(\theta_i)]^2} \ , \qquad j \neq k \ . \end{aligned}$$

The following algorithm is proposed.

Algorithm:

1. Let $t = 0$. Obtain initial estimates for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}^{(0)}$.

   A log-linear model considering all variables as independent can be used, so that,

   $$\log\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \ ,$$

   where $\boldsymbol{\mu} = E(\mathbf{Y})$, $\mathbf{Y}$ is the $N{\times}1$ vector of all variables, each obeying a Poisson law with mean μ_{ij}, $i{=}1, ..., M$, $j{=}1, ..., n_i$, $\mathbf{X}$ is a model matrix of order $N{\times}p$, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters to be estimated, considering in $\boldsymbol{\beta}$ all the main effects of the model. Thereby $\hat{\boldsymbol{\beta}}^{(0)}$ is found and it will be used in 4.

2. Obtain initial estimates for θ_i, $\hat{\theta}_i^{(0)}$, $i{=}1, ..., M$.

   The estimates can be initialized near zero, or can be obtained by finding the Ordinary Least Squares estimates $\hat{\theta}_i$, that minimizes the objective function

   $$\sum_{i=1}^{M}\Big(\mathbf{y}_i - K(\theta_i)\exp\{\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(0)}\}\Big)^T\Big(\mathbf{y}_i - K(\theta_i)\exp\{\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(0)}\}\Big) \ ,$$

   where $\hat{\boldsymbol{\beta}}^{(0)}$ was found in 1.

3. Compute $K_i^{(t)}\!=\!K(\hat{\theta}_i^{(t)})$, $C_i^{(t)}\!=\!C(\hat{\theta}_i^{(t)})$, following (3.1) and also $A_i^{(t)}\!=\!\frac{C_i^{(t)}}{(K_i^{(t)})^2}$, $i\!=\!1,...,M$.

4. Compute

$$\hat{\zeta}_{ij}^{(t)} = \mathbf{x}_j^T\hat{\boldsymbol{\beta}}^{(t)} + \frac{y_{ij}}{K_i^{(t)}\exp\{\mathbf{x}_j^T\hat{\boldsymbol{\beta}}^{(t)}\}} - 1 , \qquad i = 1,...,M, \quad j = 1,...,n_i ,$$

$$\hat{\mathbf{W}}_i^{(t)} = \mathbf{J}_{n_i}A_i^{(t)} + \mathrm{diag}\left\{\frac{1}{K_i^{(t)}\exp\{\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(t)}\}}\right\} , \qquad i = 1,...,M ,$$

(where $\mathbf{J}_{n_i}$ is a square n_i dimensional matrix of ones and $\mathbf{X}_i\boldsymbol{\beta}$ is a $n_i\!\times\!1$ vector with elements $\mathbf{x}_j^T\boldsymbol{\beta}$, $j\!=\!1,...,n_i$),

$$\hat{\mathbf{W}}^{(t)} = \mathrm{diag}\left\{\hat{\mathbf{W}}_1^{(t)},...,\hat{\mathbf{W}}_M^{(t)}\right\} ,$$

and

$$\hat{\Sigma}^{(t)} = \left[\hat{\mathbf{W}}^{(t)}\right]^{-1} .$$

5. Update $\hat{\boldsymbol{\beta}}^{(t+1)}$ and $\hat{\theta}_i^{(t+1)}$ that minimize

$$(\boldsymbol{\zeta} - \mathbf{X}\boldsymbol{\beta})^T\,\hat{\Sigma}^{(t)}(\boldsymbol{\zeta} - \mathbf{X}\boldsymbol{\beta}) ,$$

where $\mathbf{X}$ is a model matrix of order $N \times p$, $\boldsymbol{\zeta}$ is a $N \times 1$ vector, $\boldsymbol{\zeta} = \left[\boldsymbol{\zeta}_1^T\ \boldsymbol{\zeta}_2^T\ ...\ \boldsymbol{\zeta}_M^T\right]^T$, $\boldsymbol{\zeta}_i\!=\![\zeta_{i1}\ \zeta_{i2}\ ...\ \zeta_{in_i}]^T$, $i = 1,...,M$.

6. Let $t = t + 1$. Iterate steps 3 to 6 until the estimates have all stabilized.

Notice that the algorithm uses the IRGLS estimation.

In the final model the fitted values are given by

$$\hat{y}_{ij} = K(\hat{\theta}_i)\exp\{\mathbf{x}_j^T\hat{\boldsymbol{\beta}}\} , \qquad i = 1,...,M, \quad j = 1,...,n_i .$$

Note the i-group effect $K(\theta_i)$ present in the fitted values.

In summary, in this proposed modelling strategy, the starting point is a conditional model in $\mathbf{Y}_i|b_i$, considering $\log\left[E(\mathbf{Y}_i|b_i)\right] = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{1}_{n_i}b_i$. A distribution for the random variable b_i is introduced that allows correlation structure representation within the groups. The parameters are then estimated using the IRGLS method, based on $\mathbf{Y}$ moments.

4. A MODELLING EXAMPLE WITH WATER SAMPLES

The total number of coliforms (rod-shaped bacteria) in a water sample is measured in MPN/100ml, number of coliforms (in thousands) per 100 ml of water.

A set of grouped data is analyzed here. The number of coliforms in three collection spouts was registered in Lis river of the Leiria district, Portugal, in 54 occasions [source: INAG, Portugal].

The data is presented in the following graphics by *temperature* and *pH* which are the covariates of the modelling process.

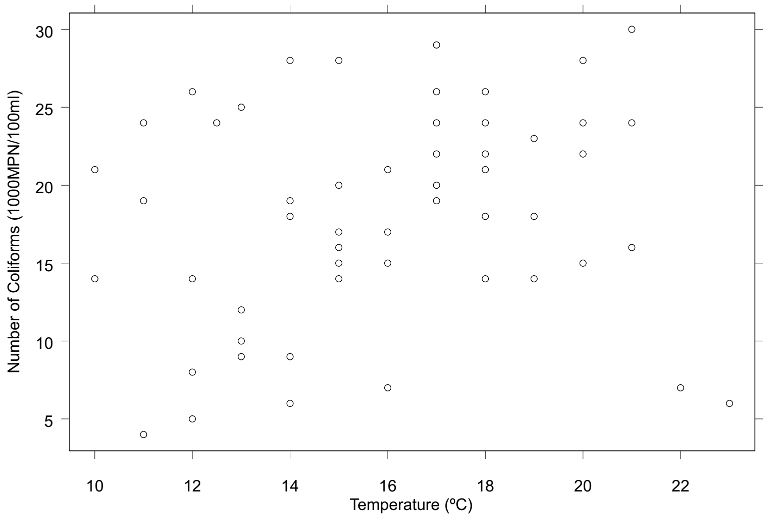


Figure 1: Number of coliforms by temperature.

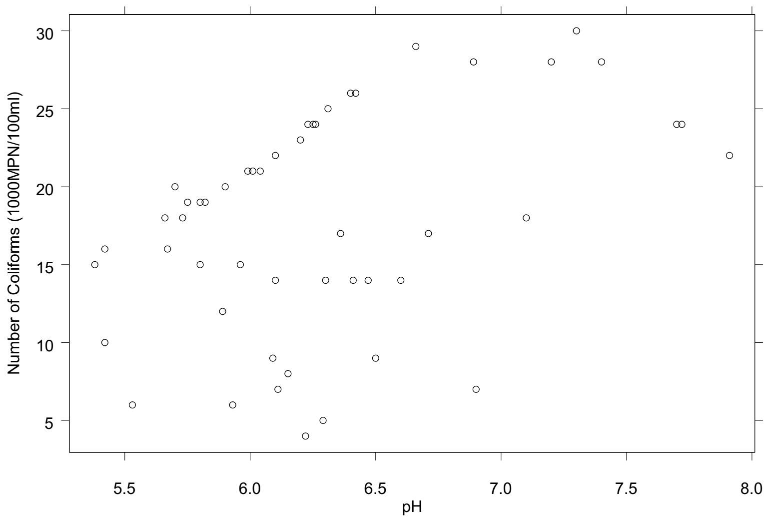


Figure 2: Number of coliforms by pH.

Observing the earlier graphics no systematic pattern is observed. However, looking at Figure 3, which represents the same observations per group — Amor, Milagres and Ponte das Mestras collection spouts, a dependence between the response variable and the covariates is highlighted. It may be also noticed that the response behaves differently for different groups.

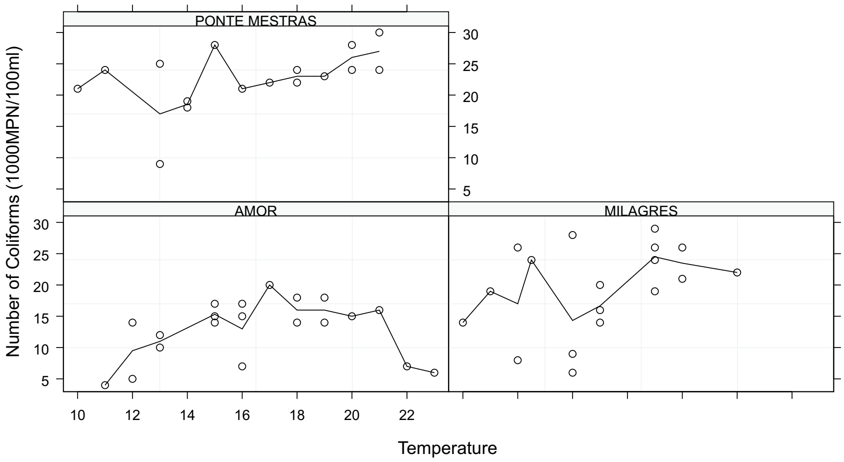


Figure 3: Number of coliforms by temperature and captation.

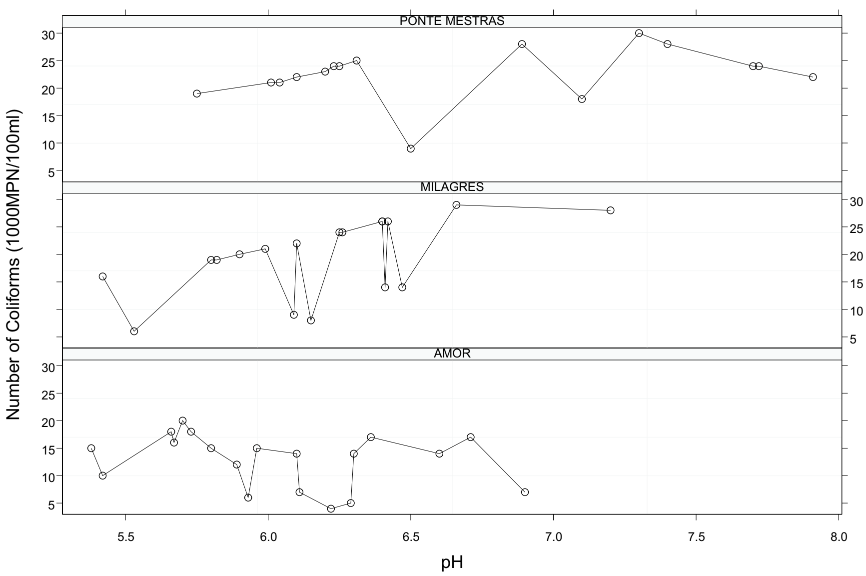


Figure 4: Number of coliforms by pH and captation.

In fact, at Ponte das Mestras and Milagres, the number of coliforms seems to follow the temperature and pH increase. However, at Amor, this is not observed.

The response variable, the number of coliforms, is a discrete variable (counting), suggesting a model based on some Poisson distribution and the method described earlier was implemented. This was done using S language and a small program that supports the method.

The modelling process will start with a log-linear Poisson model (point 1 of the proposed algorithm) considering all the variables as independent, and therefore, $\boldsymbol{\beta}$ initial values are obtained.

Considering the linear predictor

$$\beta_0 + \beta_1\, temp + \beta_2\, pH \ ,$$

$\hat{\beta}_0 = 1.22$, $\hat{\beta}_1 = 0.01$ and $\hat{\beta}_2 = 0.23$ are obtained, where *temp* is the *temperature* covariate. Overdispersion is observed in the model.

The $\theta_i, \ i = 1, 2, 3,$ parameters were initialized near zero.

It was observed that models with an intercept (β_0) have worst convergence, so all the models were considered without this parameter. Starting from $\hat{\beta}_1^{(0)} = 0.02$ and $\hat{\beta}_2^{(0)} = 0.39$, which were obtained from a log-linear Poisson model without intercept, the proposed methodology leads to the estimates

$$\hat{\beta}_1 = 0.03, \quad \hat{\beta}_2 = 0.14, \quad \hat{\theta}_1 = 0.77, \quad \hat{\theta}_2 = 0.98 \quad \text{and} \quad \hat{\theta}_3 = 1.00 \ ,$$

where θ_1 comes from Amor, θ_2 from Milagres and θ_3 from Ponte das Mestras.

However the $\hat{\beta}_1$ and $\hat{\beta}_2$ standard errors were estimated as 0.02 and 0.09, respectively, so they are not jointly significant. The θ_i standard errors were all significant.

So the models whose linear predictor has only one covariate, *temperature* or pH, will be compared.

Model with linear predictor	**Objective function (3.4) value**
$\beta_1\, temp$	78.10
$\beta_2\, pH$,	81.77

The model with the *temperature* covariate is chosen, as it has a lower value for function (3.4). The following estimates and standard errors were obtained in the selected model.

Parameter	**Referred to**	**Estimate**	**Standard Error**
β_1	temperature	0.04	0.01
θ_1	Amor	1.16	0.16
θ_2	Milagres	1.49	0.13
θ_3	Ponte das Mestras	1.48	0.14

The normalized residuals are concentrated in $[-2.04, 1.16]$.

It can be noticed that the water temperature influences the number of coliforms, because the coefficient of the *temperature* covariate is significant, although it has a low estimate ($\hat{\beta}_1 = 0.04$). The number of coliforms increases with water temperature, but not in the same way in all the spouts. In fact, in Amor this is not evident, thereby the correspondent θ_i estimate is the lower one. Probably, in this group, there are some other factors important to the coliform concentrations that were not considered here.

The select quasi-log-linear model, based on the quasi-likelihood function (as overdispersion is present), has linear predictor $\beta_0 + \beta_2\, pH$, considering pH the most significant covariate, but this model has no better fit than the mixed Poisson–Poisson considered in this paper.

As a result, clusters in data should not be ignored. It is possible to model grouped count data with the mixed Poisson–Poisson model and the algorithm proposed above. This methodology estimates the fixed and covariance parameters respecting the between groups correlations structure. Using the IRGLS method it becomes possible to obtain consistent estimates.

REFERENCES

[1] Goldstein, H. (1995). *Multilevel Statistical Models*, 2^a ed., Arnold and Oxford University Press.

[2] Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics*, **38**, 963–974.

[3] Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13–22.

[4] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2^a ed., Chapman & Hall, London.

[5] McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear and Mixed Models*, John Wiley & Sons.

[6] Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-Plus*, Springer-Verlag.

[7] Vonesh, E.F. and Chinchilli, V.M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*, Marcel Dekker.

THE EXTREMAL INDEX OF SUB-SAMPLED PERIODIC SEQUENCES WITH STRONG LOCAL DEPENDENCE

Authors: H. Ferreira
– Department of Mathematics, University of Beira Interior, Portugal
(ferreira@fenix2.ubi.pt)

A.P. Martins
– Department of Mathematics, University of Beira Interior, Portugal
(amartins@noe.ubi.pt)

Received: January 2003 Revised: July 2003 Accepted: July 2003

Abstract:

- Let $\mathbf{X} = \{X_n\}_{n\geq 1}$ be a T-periodic sequence. We define a family of local dependence conditions $D_T^{(k)}(\mathbf{u})$, $k \geq 1$, and calculate the extremal index $\theta_{\mathbf{X}}$ from the distributions of k consecutive variables of $\mathbf{X}$. For a periodic sub-sampled sequence $\mathbf{Y} = \{X_{g(n)}\}_{n\geq 1}$, where g generates blocks of I_1 observations separated by J observations, we present results on local and long range dependence conditions and compute the extremal index $\theta_{\mathbf{Y}}$.

Key-Words:

- *sub-sampling; periodic sequences; extremal index; extreme values.*

1. INTRODUCTION

In this paper we consider that $\mathbf{X} = \{X_n\}_{n\geq 1}$ is a T-periodic sequence of random variables, i.e., there exists an integer $T \geq 1$ such that, for each choice of integers $1 \leq i_1 < ... < i_n$, $(X_{i_1}, ..., X_{i_n})$ and $(X_{i_1+T}, ..., X_{i_n+T})$ have the same distribution. The period T will be considered the smallest integer satisfying the above definition.

We say that a T-periodic sequence $\mathbf{X}$ has extremal index $\theta_{\mathbf{X}}$ when, $\forall\, \tau > 0$, $\exists\, \mathbf{u}^{(\tau)} = \{u_n^{(\tau)}\}_{n\geq 1}$ such that

$$\lim_{n\to\infty} n\,\frac{1}{T}\sum_{i=1}^{T} P\Big(X_i > u_n^{(\tau)}\Big) \;=\; \tau$$

and

$$\lim_{n\to\infty} P\Big(\max\{X_1, ..., X_n) \leq u_n^{(\tau)}\Big) \;=\; e^{-\theta_{\mathbf{X}}\tau} \;.$$

The elements of $\mathbf{u}^{(\tau)}$ are called normalized levels for $\mathbf{X}$.

Such as happens for stationary sequences, the extremal index of a periodic sequence (Alpuim ([1]), Ferreira ([4])) enables us to infer the limiting behaviour of M_n from the limiting behaviour of $\hat{M}_n = \max\{\hat{X}_1, ..., \hat{X}_n\}$, $n \geq 1$, where $\hat{\mathbf{X}} = \{\hat{X}_n\}_{n\geq 1}$ is a periodic sequence of independent variables such that $F_{X_i} = F_{\hat{X}_i}$, $\forall\, i \geq 1$. Specifically,

$$\lim_{n\to\infty} P\Big(\max\{X_1, ..., X_n) \leq u_n^{(\tau)}\Big) \;=\; \bigg(\lim_{n\to\infty} P\Big(\max\{\hat{X}_1, ..., \hat{X}_n) \leq u_n^{(\tau)}\Big)\bigg)^{\theta_{\mathbf{X}}}$$

holds true.

By evaluating its extremal index $\theta_{\mathbf{X}}$, we describe in section 2 the asymptotic behaviour of the partial maximum $M_n = \max\{X_1, ..., X_n\}$, $n \geq 1$, under the condition $D(\mathbf{u})$ of Leadbetter ([5]) and a local dependence condition that generalizes the $D^{(k)}(\mathbf{u})$ of Chernick et al. ([2]).

In section 3 we give sufficient conditions for the analogous dependence conditions to hold for a sub-sampled sequence $\mathbf{Y} = \{X_{g(n)}\}_{n\geq 1}$ and we relate the extremal indexes $\theta_{\mathbf{X}}$ and $\theta_{\mathbf{Y}}$.

There are important situations in finance, for instance, where it seems reasonable to sub-sample the process by blocks matching them with bussiness periods (Dacorogna et al. ([3])). For a complete description of the extremal behavior of sub-sampled sequences $\mathbf{Y}$ from moving averages $\mathbf{X}$ with regularly varying tails see Scotto and Ferreira ([10]) and references therein.

Robinson and Tawn ([9]) pointed out the importance of the sampling frequency on the extremal properties and they have showed that if the sequence

$\mathbf{X}=\{X_n\}_{n\geq 1}$ and the sub-sampled sequence $\mathbf{Y}=\{X_{Tn}\}_{n\geq 1}$ have extremal indexes $\theta_{\mathbf{X}}$ and $\theta_{\mathbf{Y}}$, respectively, then

$$\theta_{\mathbf{X}} \;\leq\; \theta_{\mathbf{Y}} \;\leq\; T\,\theta_{\mathbf{X}}\left(1-\sum_{j=1}^{T-1}\left(1-\frac{j}{T}\right)\Pi(j)\right) ,$$

where $\Pi(j),\ j\geq 1$, are the asymptotic cluster size distributions for $\mathbf{X}$. Moreover, the upper bound is obtained under the condition $D''(u_n)$ from Leadbetter and Nandagopalan ([6]).

Our results in section 3 enable the computation of the extremal index of periodic sub-sampled sequences $\mathbf{Y}=\{X_{g(n)}\}_{n\geq 1}$ for g such that $\lim_{n\to\infty}\frac{g(n)}{n}=G$, under a family of local dependence conditions for T-periodic sequences. They generalize the main result in Martins and Ferreira ([7]) concerning stationary sequences satisfying the condition $D''(u_n)$ and g defined as $g(n)=(n-1) \bmod I + T\left[\frac{(n-1)}{I}\right],\ n\geq 1$.

2. COMPUTING THE EXTREMAL INDEX UNDER $D_T^{(k)}(\mathbf{u})$

We introduce a family of local dependence conditions for T-periodic sequences satisfying the long range dependence condition $D(\mathbf{u})$ from Leadbetter ([5]). The sequence of dependence coefficients in this condition will be referred as $\alpha^{(\mathbf{X},\mathbf{u})}=\{\alpha^{(\mathbf{X},\mathbf{u})}_{n,l}\}_{n\geq 1}$ and it is such that $\alpha^{(\mathbf{X},\mathbf{u})}_{n,l_n}=o(1)$ for some $l_n=o(n)$. For simplicity we omit the sequences $\mathbf{X}$ and $\mathbf{u}$ in these notations whenever no doubt is created.

Definition 2.1. Let $k\geq 1$ be a fixed integer and $\mathbf{X}$ a T-periodic sequence satisfying $D(\mathbf{u})$. The condition $D_T^{(k)}(\mathbf{u})$ holds for $\mathbf{X}$ when there exists a sequence of integers $\mathbf{k}=\{k_n\}_{n\geq 1}$ such that

$$\lim_{n\to\infty} k_n=+\infty\ , \qquad \lim_{n\to\infty} k_n\frac{l_n}{n}=0\ , \qquad \lim_{n\to\infty} k_n\alpha_{n,l_n}=0\ , \tag{2.1}$$

and

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

where

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

and, for $k\geq 2$,

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

The extremal behaviour of $\mathbf{X}$ has already been considered in Ferreira ([4]) under the conditions $D_T^{(k)}(\mathbf{u})$, for $k=1,2$.

If $\max\{X_i, X_{i+1}, ..., X_j\}$ is denoted by $M_{i,j}^{(\mathbf{X})}$ and we put $M_{i,j}^{(\mathbf{X})} = -\infty$ for $i>j$, then $\lim_{n\to\infty} S^{(k)}_{[\frac{n}{k_n T}]} = 0$ implies

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

which leads to

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

This last restriction, when $T=1$, is the one considered in $D^{(k)}(\mathbf{u})$ by Chernick et al. ([2]) for stationary sequences. Under $D^{(k)}(\mathbf{u})$ they compute $\theta_{\mathbf{X}}$ from the distribution of the first k variables of $\mathbf{X}$ and apply the result to several autoregressive sequences. In the following we will extend their results for periodic sequences.

Proposition 2.1. *If the T-periodic sequence $\mathbf{X}$ satisfies $D(\mathbf{u})$ and $D_T^{(k)}(\mathbf{u})$ then*

$$P\Big(M_n \leq u_n\Big) - \exp\left(\frac{n}{T} \sum_{i=1}^{T} P\Big(X_i > u_n \geq M_{i+1,i+k-1}\Big)\right) \;=\; o(1) \;.$$

Proof: Under $D(\mathbf{u})$ we have, for $\mathbf{k}$ as in (2.1),

$$P\Big(M_n \leq u_n\Big) - P^{k_n}\Big(M_{[\frac{n}{k_n T}]T} \leq u_n\Big) \;=\; o(1) \;,$$

and therefore it is enough to proof that

$$P\Big(M_{[\frac{n}{k_n T}]T} > u_n\Big) - \frac{\dfrac{n}{T} \displaystyle\sum_{i=1}^{T} P\Big(X_i > u_n \geq M_{i+1,i+k-1}\Big)}{k_n} \;=\; o(1) \;. \tag{2.2}$$

Since, by applying $D_T^{(k)}(\mathbf{u})$,

$$\begin{aligned} P\Big(M_{[\frac{n}{k_n T}]T} > u_n\Big) \;&=\; P\left(\bigcup_{i=1}^{[\frac{n}{k_n T}]T} \Big\{X_i > u_n \geq M_{i+1,[\frac{n}{k_n T}]T}\Big\}\right) \\ &=\; \left[\frac{n}{k_n T}\right] \sum_{i=1}^{T} P\Big(X_i > u_n \geq M_{i+1,i+k-1}\Big) - A_n \;, \end{aligned}$$

holds with $k_n A_n \leq S^{(k)}_{[\frac{n}{k_n T}]} = o(1)$, we conclude (2.2). $\square$

As a consequence of this result we compute the extremal index as follows.

Corollary 2.1. *If the T-periodic sequence $\mathbf{X}$ satisfies $D(\mathbf{u})$ for all $\mathbf{u}=\mathbf{u}^{(\tau)}$ and $D_T^{(k)}(\mathbf{v})$ for some $\mathbf{v}=\mathbf{v}^{(\tau_0)}$ then there exists $\theta_{\mathbf{X}}$ if and only if there exists*

$$\nu_{\mathbf{X}} = \lim_{n\to\infty} n\,\frac{1}{T}\sum_{i=1}^{T} P\Big(X_i > v_n \geq M_{i+1,i+k-1}\Big) ,$$

and in this case it holds

$$\theta_{\mathbf{X}} = \frac{\nu_{\mathbf{X}}}{\tau_0} . \qquad \square$$

We can apply this result to calculate the extremal index of a T-periodic moving average, following the approach of Chernick et al. ([2]) for the stationary case.

Let $\mathbf{Z} = \{Z_n\}_{n\geq 1}$ be a T-periodic sequence of independent variables with regularly varying equivalent tails with exponent $-\alpha$ satisfying

$$\lim_{x\to\infty}\frac{P(Z_i > x)}{P(Z_j > x)} = \gamma_{i,j}^{(+)} > 0, \quad \lim_{x\to\infty}\frac{P(Z_i < -x)}{P(Z_j < -x)} = \gamma_{i,j}^{(-)} > 0 , \qquad i,j = 1,...,T ,$$

and

$$\lim_{x\to\infty}\frac{P(Z_i > x)}{P(|Z_i| > x)} = p_i \in [0,1] , \qquad i = 1,...,T .$$

For $\tau_i > 0$, $i=1,...,T$, and $\tau = \frac{1}{T}\sum_{i=1}^{T}\tau_i$, let $\mathbf{u}^{(\tau)}$ be defined by

$$\lim_{n\to\infty} nP(|Z_i| > u_n) = \tau_i \bigg/ \left\{ p_i \sum_{s=0}^{T-1}\gamma_{i-s,i}^{(+)} \sum_{j=-\infty}^{\infty}[c_{jT+s}^{+}]^{\alpha} + q_i \sum_{s=0}^{T-1}\gamma_{i-s,i}^{(-)} \sum_{j=-\infty}^{\infty}[c_{jT+s}^{-}]^{\alpha} \right\},$$

where $q_i = 1 - p_i$, $c_j^{+} = \max\{c_j, 0\}$, $c_j^{-} = \max\{-c_j, 0\}$ and $\mathbf{c}=\{c_j\}$ is a sequence of constants such that $\sum_{j=-\infty}^{+\infty}|c_j|^{\delta} < +\infty$ for some $\delta < \min\{\alpha, 1\}$.

For the T-periodic moving average $X_n = \sum_{j=-\infty}^{+\infty} c_j Z_{n-j}$, $n \geq 1$, by applying our result to the $2m$-dependent T-periodic sequence $X_n^{(m)} = \sum_{j=-m}^{m} c_j Z_{n-j}$ and following in a straighforward way the reasoning of Chernick et al. ([2]), we find

$$\theta = \frac{\displaystyle\sum_{i=1}^{T}\gamma_{i,1}\left\{ p_i \sum_{s=0}^{T-1}\gamma_{i-s,i}^{(+)}\, c_s^{+}(\alpha) + q_i \sum_{s=0}^{T-1}\gamma_{i-s,i}^{(-)}\, c_s^{-}(\alpha)\right\}}{\displaystyle\sum_{i=1}^{T}\gamma_{i,1}\left\{ p_i \sum_{s=0}^{T-1}\gamma_{i-s,i}^{(+)} \sum_{j=-\infty}^{\infty}[c_{jT+s}^{+}]^{\alpha} + q_i \sum_{s=0}^{T-1}\gamma_{i-s,i}^{(-)} \sum_{j=-\infty}^{\infty}[c_{jT+s}^{-}]^{\alpha}\right\}} ,$$

where

$$c_s^{+}(\alpha) = \sum_{j=-\infty}^{\infty}\Big([c_{jT+s}^{+}]^{\alpha} - \max_{r > jT+s}\{c_r^{+}\}^{\alpha}\Big)^{+} , \quad c_s^{-}(\alpha) = \sum_{j=-\infty}^{\infty}\Big([c_{jT+s}^{-}]^{\alpha} - \max_{r > jT+s}\{c_r^{-}\}^{\alpha}\Big)^{+} .$$

For details on the proofs of this example see Martins and Ferreira ([8]).

3. PERIODIC SUB-SAMPLED SEQUENCE

We first set sufficient conditions for the previous results to hold for $\mathbf{Y} = \{X_{g(n)}\}_{n\geq 1}$. Let $g\colon \mathbb{N}\to\mathbb{N}$ be a strictly increasing function for which there exists positive integers I_1 and I_2 such that, $\forall\, n,k\in\mathbb{N}$, it holds $g(n+kI_1) = g(n)+kI_2$. We will refer such g as an I_1, I_2-periodic function and suppose that I_1 and I_2 are the smallest integers satisfying the definition.

Therefore $\mathbf{Y}=\{X_{g(n)}\}_{n\geq 1}$ is obtained from $\mathbf{X}$ by sub-sampling blocks of I_1 variables separated by $J = I_2 - (g(I_1)-g(1)) - 1 \geq 1$ variables.

In a particular case considered in Scotto and Ferreira ([10]), $\mathbf{X}$ is a stationary moving average with heavy-tailed innovations and g generates blocks of I_1 consecutive observations separated by $J\geq 1$ observations.

Proposition 3.1. *If $\mathbf{X}$ is a T-periodic sequence and g is an I_1, I_2-periodic function with I_2 a multiple of T, then $\mathbf{Y} = \{X_{g(n)}\}$ is an I_1-periodic sequence.*

Proof: For each choice of integers $1\leq i_1 < ... < i_n,\ p\geq 1$, we have

$$(Y_{i_1+I_1}, ..., Y_{i_n+I_1}) \;=\; (X_{g(i_1+I_1)}, ..., X_{g(i_n+I_1)}) \;=$$
$$= (X_{g(i_1)+I_2}, ..., X_{g(i_n)+I_2}) \overset{d}{=} (X_{g(i_1)}, ..., X_{g(i_n)}) \;=\; (Y_{i_1}, ..., Y_{i_n}) \;. \quad \square$$

In the next result, we denote a sequence $\mathbf{u}$ such that $\lim\limits_{n\to\infty} nP(X_i > u_n^{(\tau_i)}) = \tau_i$ by $\mathbf{u} = \mathbf{u}^{(\tau_i, X_i)}$. From the definition of normalized levels and $\mathbf{Y}\subset\mathbf{X}$ we give a simple procedure to get $\mathbf{v}=\mathbf{v}^{(\tau,\mathbf{Y})}$ with $\tau = \frac{1}{I_1}\sum_{i=1}^{I_1} G^{-1}\tau_{g(i)}$ and $G = \lim\limits_{n\to\infty}\dfrac{g(n)}{n}$.

Proposition 3.2. *Let $\mathbf{X}$ be a T-periodic sequence and g an I_1, I_2-periodic function with I_2 a multiple of T. If $\lim\limits_{n\to\infty}\dfrac{g(n)}{n} = G$ and $\mathbf{u} = \mathbf{u}^{(\tau_i,X_i)}$, $i=1,...,T$, then $\mathbf{v} = \{u_{g(n)}\}$ satisfies:*

(i) $\mathbf{v} = \mathbf{v}^{(G^{-1}\tau_i, X_i)},\ \ i = 1, ..., T.$

(ii) $\mathbf{v} = \mathbf{v}^{(G^{-1}\tau_{g(i)}, Y_i)},\ \ i = 1, ..., I_1,$ *and* $\{\tau_{g(1)}, ..., \tau_{g(I_1)}\}\subset\{\tau_1, ..., \tau_T\}$. $\square$

For $\mathbf{u} = \mathbf{u}^{(\tau_i', X_i)}$, with $\tau_i' = G\tau_i$, $i=1,...,T$, we have $\mathbf{v} = \{u_{g(n)}\} = \mathbf{v}^{(\tau_i, Y_i)}$ and we can easily get $\alpha^{(\mathbf{Y},\mathbf{v})}_{n,l^{(\mathbf{X})}_{g(n)}} \leq \alpha^{(\mathbf{X},\mathbf{u})}_{g(n),l^{(\mathbf{X})}_{g(n)}}$ with $l^{(\mathbf{X})}_{g(n)} = o(n)$.

Moreover, if $\mathbf{v} = \mathbf{v}^{(\tau_{0,i}, X_i)}$, $i=1,...,T$, then $\mathbf{w} = \{v_{[nI_2/I_1]}\}$ satisfies

$$\mathbf{w} = \mathbf{w}^{(\tau_{0,i} I_1/I_2, X_i)} \;, \qquad i = 1, ..., T \;,$$
$$\mathbf{w} = \mathbf{w}^{(\tau_{0,g(i)} I_1/I_2, Y_i)} \;, \qquad i = 1, ..., I_1$$

and
$$S^{(k,\mathbf{Y},\mathbf{w})}_{[\frac{n}{k_n I_1}]} \leq A\, S^{(k,\mathbf{X},\mathbf{w})}_{[\frac{n}{k'_n T}]}\ ,$$
where A is a constant and $k'_n = k_{[nI_1/I_2]}$.

These are the main arguments to obtain the following result.

Proposition 3.3. *Let* $\mathbf{X}$ *be a* T*-periodic sequence* $\mathbf{X}$ *satisfying* $D(\mathbf{u})$ *for all* $\mathbf{u}=\mathbf{u}^{(\tau_i,X_i)}$ *for some* $i\in\{1,...,T\}$ *and* $D^{(k)}_T(\mathbf{v})$ *for some* $\mathbf{v}=\mathbf{v}^{(\tau_{0,i},X_i)}$, $i=1,...,T$, *with* $\mathbf{k}'=\{k_{[nI_1/I_2]}\}$ *and* $\mathbf{k}=\{k_n\}$ *as in (2.1). Then, for* g *as in the above proposition,* $\mathbf{Y}=\{X_{g(n)}\}$ *satisfies:*

(i) $D(\mathbf{u})$ *for all* $\mathbf{u}=\mathbf{u}^{(\tau_i,Y_i)}$, $i=1,...,I_1$,

(ii) $D^{(k)}_{I_1}(\mathbf{w})$ *for* $\mathbf{w}=\{v_{[nI_2/I_1]}\}=\mathbf{w}^{(\tau_{0,g(i)I_1/I_2},Y_i)}$, $i=1,...,I_1$, *with* $\mathbf{k}=\{k_n\}$. □

We will assume that $\mathbf{X}$ is in the conditions of Proposition 3.3 and calculate the extremal index of the periodic sub-sampled sequence $\mathbf{Y}=\{X_{g(n)}\}$ as a consequence of this proposition and Corollary 2.1.

Proposition 3.4. *Let* $\mathbf{X}$ *be a* T*-periodic sequence* $\mathbf{X}$ *satisfying* $D(\mathbf{u})$ *for all* $\mathbf{u}=\mathbf{u}^{(\tau_i,X_i)}$ *for some* $i\in\{1,...,T\}$ *and* $D^{(k)}_T(\mathbf{v})$ *for some* $\mathbf{v}=\mathbf{v}^{(\tau_{0,i},X_i)}$, $i=1,...,T$, *with* $\mathbf{k}'=\{k_{[nI_1/I_2]}\}$ *and* $\mathbf{k}=\{k_n\}$ *as in (2.1). Then, for* g *as in the above proposition,* $\mathbf{Y}=\{X_{g(n)}\}$ *has extremal index* $\theta_{\mathbf{Y}}$ *if and only if there exists*

$$\nu_{\mathbf{Y}} = \lim_{n\to\infty} n\,\frac{1}{I_1}\sum_{i=1}^{I_1} P\Big(X_{g(i)} > v_{[nI_2/I_1]} \geq \max\Big\{X_{g(i+1)}, X_{g(i+2)}, ..., X_{g(i+k-1)}\Big\}\Big)\ .$$

In this case
$$\theta_{\mathbf{Y}} = \frac{I_1\,\nu_{\mathbf{Y}}}{\displaystyle\sum_{i=1}^{I_1}\tau_{0,g(i)}}\ . \qquad \square$$

Let
$$\nu_{\mathbf{X}} = \lim_{n\to\infty} n\,\frac{1}{T}\sum_{i=1}^{T} P\Big(X_i > v_n \geq M^{(\mathbf{X})}_{i+1,i+k-1}\Big)\ ,$$
and $\theta_{\mathbf{X}} = \frac{\nu_{\mathbf{X}}}{\tau_0}$, with $\tau_0 = \frac{1}{T}\sum_{i=1}^T \tau_{0,i}$.

For the particular case of $I_1=T$ and $g(i{+}1)=g(i)$, for $i=1,...,I_1$, we find $\theta_{\mathbf{Y}}=\theta_{\mathbf{X}}+\frac{\rho}{T\tau_0}$ where

$$\begin{aligned}\rho \;=\; &\lim_{n\to\infty} n\,P\Big(X_{g(I_1)} > v_{[nI_2/I_1]} \geq \max\Big\{X_{g(1)+I_2}, X_{g(2)+I_2}, ..., X_{g(k-1)+I_2}\Big\}\Big)\\ &- \lim_{n\to\infty} n\,P\Big(X_{g(I_1)} > v_{[nI_2/I_1]} \geq M^{(\mathbf{X})}_{g(I_1)+1,g(I_1)+k-1}\Big)\ .\end{aligned}$$

If $k=1$ then $\rho=0$, as expected, and for the particular cases where $1=T=I_1$ and $k=2$ we have very simple expressions for ρ (Martins and Ferreira ([7])). They can be applied, for instance, to calculate the extremal index of the sub-sampled ARMAX(α) process considered in Robinson and Tawn ([9]). For that example we find

$$\theta_{\mathbf{Y}} \;=\; \theta_{\mathbf{X}} + \frac{\rho}{\tau_0} \;=\; 1-\alpha+\frac{\alpha\,(1-\alpha^{I_2-1})\,\tau_0}{\tau_0} \;=\; 1-\alpha^{I_2}\;,$$

equal to the value of Robinson and Tawn ([9]) for the sampling case $\mathbf{Y}=\{X_{nI_2}\}$.

4. CONCLUDING REMARKS

Under the local dependence condition $D_T^{(k)}(\mathbf{u}^{(\tau)})$ we compute the extremal index of the T-periodic sequence $\mathbf{X}$ from the T distributions of k consecutive variables as well as the extremal index of some sub-sampled I_1-periodic sequences $\mathbf{Y}=\{X_{g(n)}\}$.

It would be interesting to apply these results to functions g used in applications and moving averages or Markov sequences $\mathbf{X}$ where $D''(u_n)$ fails. This remains as topic of future research.

ACKNOWLEDGMENTS

We are grateful to a referee's corrections and rigorous report.

REFERENCES

[1] Alpuim, M.T. (1988). Contribuições à teoria de extremos em sucessões dependentes. Ph.D. Thesis, DEIOC, Univ. of Lisbon.

[2] Chernick, M.R.; Hsing, T. and McCormick, W.P. (1991). Calculating the extremal index for a class of stationary sequences, *Adv. Appl. Prob.*, **23**, 835–850.

[3] Dacorogna, M.M.; Müller, U.A.; Nagler, R.J.; Olsen, R.B. and Pictet, O.V. (1993). A geopraphical model for the daily and weekly seasonal volatility in the foreign exchange market, *Journal of International Money and Finance*, **12**, 413–438.

[4] Ferreira, H. (1994). Multivariate extreme values in T-periodic random sequences under mild oscillation restrictions, *Stochastic Process. Appl*, **49**, 111–125.

[5] Leadbetter, M.R. (1983). Extremes and local dependence in stationary sequences, *Z. Wahrschtheor*, **65**, 291–306.

[6] Leadbetter, M.R. and Nandagopalan, L. (1989). On exceedance point processes for stationary sequences under mild oscillation restrictions, *Lect. Notes Statist.*, **51**, 69–80.

[7] Martins, A. and Ferreira, H. (2003a). The extremal index of sub-sampled processes. To appear in *J. Statist. Plann. and Inference.*

[8] Martins, A. and Ferreira, H. (2003b). Índice extremal de médias móveis periódicas com caudas de variação regular. Pre-print. Univ. of Beira Interior.

[9] Robinson, M.E. and Tawn, J.A. (2000). Extremal analysis of processes sampled at different frequencies, *J.R. Statist. Soc. B*, **62**, 117–135.

[10] Scotto, M.G. and Ferreira, H. (2002). Extremes of deterministic sub-sampled moving averages with heavy-tailed innovations. Preprint Univ. of Lisbon.

LIFETIME MODELS WITH NONCONSTANT SHAPE PARAMETERS

Authors: JOSMAR MAZUCHELI
– Departamento de Estatística, Universidade Estadual de Maringá, Maringá, P.R. — Brazil (jmazucheli@uem.br)

FRANCISCO LOUZADA-NETO
– Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, S.P. — Brazil (dfln@power.ufscar.br)

JORGE ALBERTO ACHCAR
– Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, S.P. — Brazil (jorge@icmc.sc.usp.br)

Received: October 2002 Revised: September 2003 Accepted: October 2003

Abstract:

- In its standard form, a lifetime regression model usually assumes that the time until an event occurs has a constant shape parameter and a scale parameter that is a function of covariates. In this paper we consider lifetime models with shape parameter dependent on a vector of covariates. Two special models are considered, the Weibull model and a mixture model incorporating long-term survivors, when we consider that the incidence probability is also dependent on covariates. Classical parameters estimation approach is considered on two real data sets.

Key-Words:

- *accelerated life tests; bootstrap; long-term survivors; nonconstant shape parameter; Weibull distribution.*

AMS Subject Classification:

- 49A05, 78B26.

1. INTRODUCTION

To express the distribution of a nonnegative random variable, T, which represents the lifetime of individuals (or components) in some population subjected to covariate effects, several mathematically equivalent functions that uniquely determine the distribution can be considered; namely, the cumulative distribution, the density, the survival and the hazard functions [16]. For lifetime data, the survival function at a particular time t is defined as

$$S_0(t \,|\, \mu(\mathbf{x}), \gamma) \;=\; P(T > t \,|\, \mu(\mathbf{x}), \gamma) \;, \tag{1.1}$$

where $\mu(\mathbf{x})$ is a scale parameter that is a function of covariate involving unknown parameters and γ is a constant unknown shape parameter. It is particularly useful to define the survival model in terms of (1.1), because of its interpretation as the probability of an individual (or component) surviving till time t [16].

Besides, in several applications, it is clear that a non-zero proportion of patients or components can be considered cured, or do not fail in their testing time [20]. In this context, we consider the model

$$S(t \,|\, \mathbf{x}) \;=\; p + (1-p)\, S_0(t \,|\, \mu(\mathbf{x}), \gamma) \;, \tag{1.2}$$

where S is the population survival function and $0 < p < 1$ represents the cured fraction, which is cured or never fails with respect to the specific cause of death (or failure). Observe that (1.2) is a mixture model with two components, where S_0 is the survival function of the individuals which are not cured. For the cured patients, the survival function is equal to one for all finite values t. Mixture survival models provide a way of modelling time to death when cure is possible, simultaneously estimating death hazard of fatal cases and the proportion of cured cases.

In many applications however the usual assumption of constant shape parameter γ cannot be appropriate. For instance, in some studies with fatigue of materials, usually, it is assumed that the shape parameter of the Weibull distribution depends on the stress levels, as we can see in Wang and Kececioglu ([30]), Meeker and Escobar ([22]), Pascual and Meeker ([26]), Meeter and Meeker ([23]), Meeker and Escobar ([21]), Hirose ([12]), Chan ([4]), Smith ([27]) and Nelson ([25]). Anderson ([1]) considers a Weibull accelerated regression model with the dispersion parameter depending on the location parameter. In the context of risk modelling, Hsieh ([13]) introduces heteroscedastic risk models, and Louzada-Neto ([19, 17]) introduces an extended risk model. Applications in the context of regression models with normal errors and nonconstant scale are considered by Zhou *et al.* ([31]) and Tanizaki and Zhang ([28]). Cepeda and Gamerman ([3]) consider Bayesian modelling of variance heterogeneity in normal regression models.

In this paper we consider a general survival model with shape and cured fraction parameters depending on covariates. The approach with constant shape parameter was first used by Farewell [8]. The advantage of such a formulation is to have several usual survival models as particular cases. Maximum likelihood estimation procedure is adopted for two special cases: the Weibull distribution with shape parameter depending on a vector of covariates and a long-term Weibull survival mixture model in the presence of covariates. In Section 2 we introduce a general survival model with shape and scale parameters depending on covariates. The Weibull case is introduced in Section 3. Two real data sets are presented in Section 4. Some concluding remarks in Section 5 finalize the paper.

2. A GENERAL SURVIVAL MODEL

Consider a survival model with shape parameter depending on covariates. The corresponding survival function is

$$S_0(t\,|\,\mu(\mathbf{x}),\gamma(\mathbf{y})) \;=\; P(T>t\,|\,\mathbf{x},\,\mathbf{y}) \;, \tag{2.1}$$

where $\mu(\mathbf{x})$ is a scale parameter depending on a covariate vector, $\mathbf{x}$, and $\gamma(\mathbf{y})$ is the shape parameter depending on a covariate vector, $\mathbf{y}$. Both $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ may involve unknown parameters. Of course, the vectors $\mathbf{x}$ and $\mathbf{y}$ can be equal.

For fitting long-term survival data, where a proportion of the individuals are cured [20], we consider the general survival model

$$S(t|\mathbf{x},\mathbf{y},\mathbf{z}) \;=\; p(\mathbf{z}) + (1-p(\mathbf{z}))\, S_0(t\,|\,\mu(\mathbf{x}),\gamma(\mathbf{y})) \;, \tag{2.2}$$

where $\mu(\mathbf{x})$ and $\gamma(\mathbf{y})$ are scale and shape parameters of the lifetime distribution of non-cured patients and $0<p(\mathbf{z})<1$ is the incidence probability depending on a covariate vector, $\mathbf{z}$, involving unknown parameters. For $p(\mathbf{z})=0$ we have the model (2.1).

A special case is given by the Weibull survival function for the non-cured patients, given by

$$S_0(t\,|\,\mu(\mathbf{x}),\gamma(\mathbf{y})) \;=\; \exp\left[-\left(\frac{t}{\mu(\mathbf{x})}\right)^{\gamma(\mathbf{y})}\right] . \tag{2.3}$$

Let us assume a random sample $T_1,...,T_n$, such that, associated to each T_i there are covariate vectors $\mathbf{x}_i^t=(1,x_{i1},...,x_{ik})$, $\mathbf{y}^t=(1,y_{i1},...,y_{ik})$ and $\mathbf{z}^t=(1,z_{i1},...,z_{ik})$, and an indicator variable δ_i, $\delta_i=1$ if t_i is an observed lifetime or $\delta_i=0$ if t_i is a censored observation (rigth-censored observations). Then, for an uninformative censoring mechanism, the likelihood function [16] can be written as

$$L \;=\; \prod_{i=1}^{n} f(t_i|\mathbf{x}_i,\mathbf{y}_i,\mathbf{z}_i)^{\delta_i}\, S(t_i|\mathbf{x}_i,\mathbf{y}_i,\mathbf{z}_i)^{1-\delta_i} \;, \tag{2.4}$$

where $f(t_i|\mathbf{x}_i,\mathbf{y}_i,\mathbf{z}_i)$ is the density function and $S(t_i|\mathbf{x}_i,\mathbf{y}_i,\mathbf{z}_i)$ is defined in (2.2).

Let $\boldsymbol{\theta}' = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ be the parameter vector indexing (2.2). The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ can be obtained by solving the system of nonlinear equations, $\partial \log L/\partial\boldsymbol{\theta} = \mathbf{0}$. However, it can be hard to solve the system of nonlinear equations above by pure Newton-type methods, since it is easy to overstep the true minimum. An alternative algorithm is proposed by [30] based on [15, 2, 9]. However, a straightforward procedure, which we prefer, is to maximize (2.4). This procedure can be implemented in a standard statistical software such as *R* [14] or a SAS via a routine that finds a local maximum of a nonlinear function using general-purpose optimization procedure. In the appendix, we present the SAS code of the NLP procedure [10, 11] used to find out the maximum likelihood estimates presented in our examples.

3. THE WEIBULL PARTICULAR CASE

Consider the general Weibull survival model obtained by considering (2.2) with (2.3). Assuming that the scale parameter, the shape parameter and the incidence probability are affected by covariate vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$, respectively, let us to consider $p(\mathbf{z}_i)$ as a logit link, such as, $\log\left(\frac{p(\mathbf{z}_i)}{1-p(\mathbf{z}_i)}\right) = \eta_0 + \sum_{j=1}^k \eta_j z_{ij}$, the log-linear models $\log(\mu(\mathbf{x}_i)) = \alpha_0 + \sum_{j=1}^k \alpha_j x_{ij}$ and $\log(\gamma(\mathbf{y}_i)) = \beta_0 + \sum_{j=1}^k \beta_j y_{ij}$. Thus, the log-likelihood function for $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is given by

$$\begin{aligned} l(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}|\mathbf{x},\mathbf{y},\mathbf{z}) \;\propto\; & \sum_{i=1}^n \delta_i \Big[\mathbf{y}_i^t\boldsymbol{\beta} + e^{\mathbf{y}_i^t\boldsymbol{\beta}}\mathbf{x}_i^t\boldsymbol{\alpha} + e^{\mathbf{y}_i^t\boldsymbol{\beta}}\log(t_i)\Big] \\ & + \sum_{i=1}^n \delta_i \log(p(\mathbf{z}_i)) - \sum_{i=1}^n \delta_i (t_i\, e^{\mathbf{x}_i^t\boldsymbol{\alpha}})^{e^{\mathbf{y}_i^t\boldsymbol{\beta}}} \\ & + \sum_{i=1}^n (1-\delta_i)\log\left[p(\mathbf{z}_i) + (1-p(\mathbf{z}_i))\, e^{(-t_i\, e^{\mathbf{x}_i^t\boldsymbol{\alpha}})^{e^{\mathbf{y}_i^t\boldsymbol{\beta}}}}\right], \end{aligned} \tag{3.1}$$

where $p(\mathbf{z}_i)^{-1} = e^{-(\gamma_0+\sum_{j=1}^k\gamma_j z_{ij})}(1+e^{\gamma_0+\sum_{j=1}^k\gamma_j z_{ij}})$, $\boldsymbol{\alpha}^t = (\alpha_0, ..., \alpha_k)$, $\boldsymbol{\beta}^t = (\beta_0, ..., \beta_k)$, $\boldsymbol{\gamma}^t = (\gamma_0, ..., \gamma_k)$, $\mathbf{x}_i^t = (1, x_{i1}, ..., x_{ik})$, $\mathbf{y}^t = (1, y_{i1}, ..., y_{ik})$ and $\mathbf{z}^t = (1, z_{i1}, ..., z_{ik})$.

4. SOME APPLICATIONS

4.1. A first application

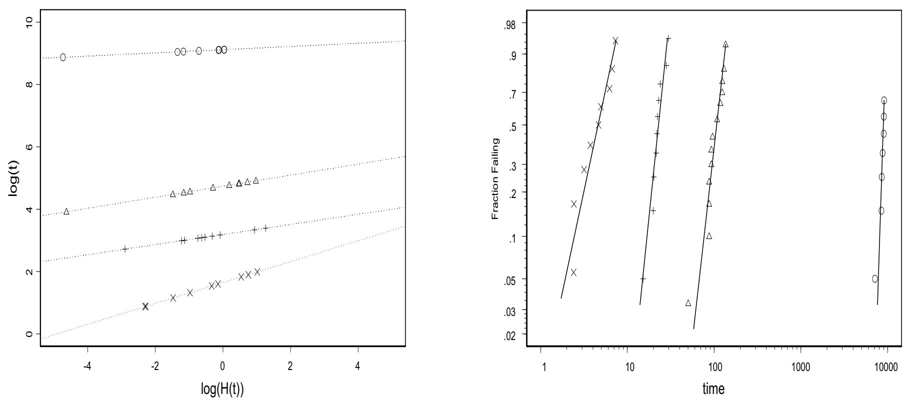
To check the assumption of shape parameter dependent on the covariates we can use graphical diagnostic methods. As a special case, consider the accelerated lifetime test (ALT) data on PET film, (see, Table 1), introduced by Hirose [12], see also Wang and Kececioglu ([30]). The ALT was performed at

Table 1: Failure times (hours) from an accelerated life test on PET film in SF_6 gas insulated transformers, [12].

Voltage	Failure times
5 kV	7131, 8482, 8559, 8762, 9026, 9034, 9104, 9104.25*, 9104.25*, 9104.25*
7 kV	50.25, 87.75, 87.76, 87.77, 92.90, 92.91, 95.96, 108.3, 108.3, 117.9, 123.9, 124.3, 129.7, 135.6, 135.6
10 kV	15.17, 19.87, 20.18, 21.50, 21.88, 22.23, 23.02, 28.17, 29.70
15 kV	2.40, 2.42, 3.17, 3.75, 4.65, 4.95, 6.23, 6.68, 7.30

Starred quantities denote censored observations.

four levels of the voltage; $v = 5, 7, 10$ and 15, with $10, 15, 10$ and 9 observations each, respectively. Three censored values were observed at $v = 5$. Denoting by $S(t) = P(T > t)$, the survival function, we should have parallel straight lines for the plots of $\log(-\log \hat{S}(t))$ versus $\log(t)$ for each stress level considering the Weibull distribution [16]. This is also true for the Weibull probability plot, Figure 1-b. In Figures 1-a and 1-b we observe straight lines which indicates that the Weibull distribution is appropriate, but we do not have parallel lines which indicates different shape parameters for each stress level. Interested readers can refer to Chapters 2, 7 and 8 of Meeker and Escobar ([22]), which present different methods to search for an appropriate lifetime distribution for fitting a set of data.



(a) (b)

Figure 1: Weibull fit for PET film data, Table 1.
(a): Hazard plot.
(b): Probability plot. 5 kV ($\circ$), 7 kV ($\triangle$), 10 kV ($+$) and 15 kV ($\times$).

Figure 1 indicate that the scale and shape parameter of the Weibull distribution should be affected by the stress levels. Moreover, following [30], plots show that $\log \hat{\mu}$ and $\log \hat{\gamma}$ have linear relationships with $x_1 = y_1 = -\log(v - 4.76)$, where $\hat{\mu}$ and $\hat{\gamma}$ are the MLEs of μ and γ, obtained by considering each individual covariate level, which are given in Table 2, the constant 4.76 is a fixed threshold level [12], below which a failure is unlikely to occur.

Table 2: Maximum likelihood and standard deviation estimates considering a Weibull model for each stress level.

Level	log-likelihood	MLE $\hat{\mu}$	MLE $\hat{\gamma}$
5 kV	−57.7394	9.1145 (0.0196)	2.9721 (0.3496)
7 kV	−67.5903	4.7367 (0.0480)	1.7315 (0.2100)
10 KV	−28.1308	3.1873 (0.0541)	1.8230 (0.2375)
15 kV	−17.4361	1.6474 (0.1179)	1.0938 (0.2676)

Table 3 shows the MLEs for the parameter of (2.3) and their standard deviations assuming $\log(\mu(x_1)) = \alpha_0 + \alpha_1 \log(v - 4.76)$ and $\gamma(y_1) = \text{constant}$ (hereafter called Model A) and $\log(\gamma(y_1)) = \beta_0 + \beta_1 \log(v - 4.76)$ (hereafter called Model B).

Table 3: Maximum likelihood estimates considering two Weibull models.

Model	Parameter	Estimates MLE	Estimates StDev
model A	α_0	6.3480	0.0399
	α_1	−1.9629	0.0265
	β	1.6080	0.1281
model B	α_0	6.3285	0.0213
	α_1	−1.9529	0.0156
	β_0	2.2311	0.1776
	β_1	−0.4636	0.1152

Locally at the MLEs, the values of the log-likelihood functions are −179.9849 (for Model A) and −173.2728 (for Model B). The values of the likelihood ratio statistics, Wald and score statistics to test model A against model B, that is, $H_0\colon \beta_1 = 0$ against $H_1\colon \beta_1 \neq 0$, are equal to 13.4240, 16.1896 and 17.0416, respectively. Their empirical p-values obtained from 10 000 bootstrap simulations are equal to 0.0007, 0.0007 and 0.0014, respectively, leading to a strong evidence in favour of the complete model (Model B). The empirical distributions of these statistics are given in Figure 2 together with their Q-Q plots. We do not observe a good approximation to the chi-square distribution with one degree of freedom.

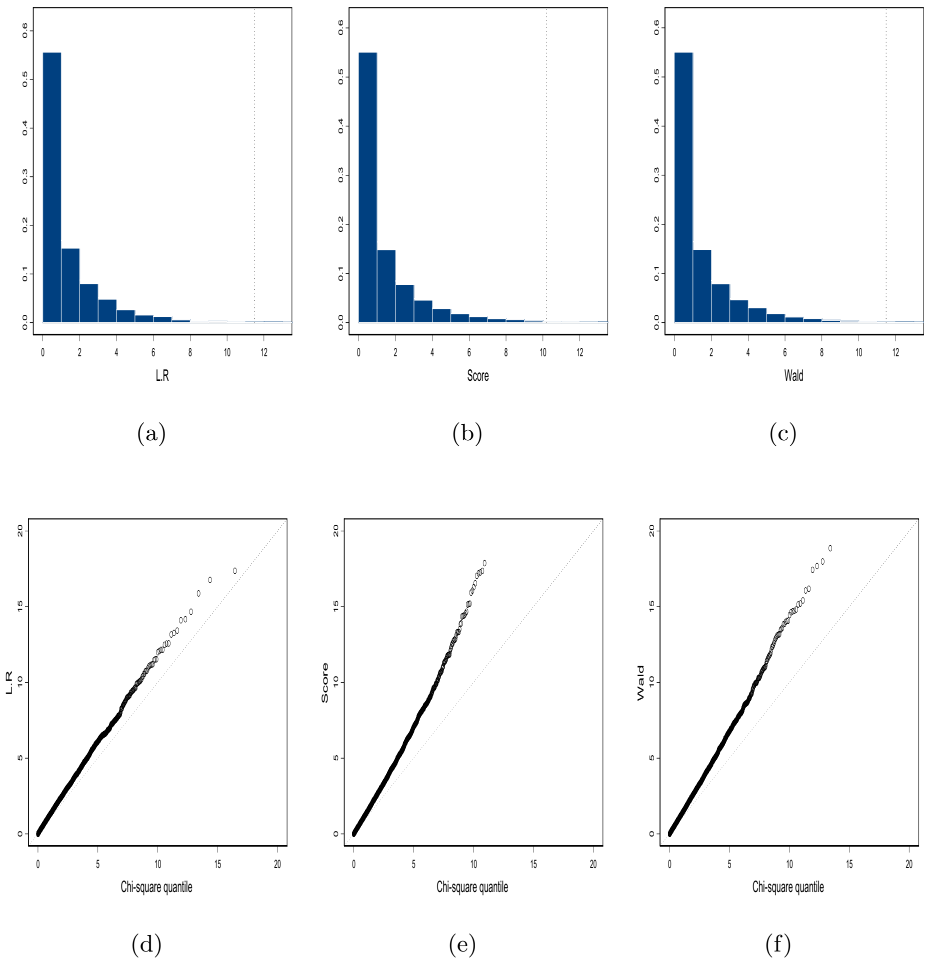


Figure 2: Empirical distributions. (a): Likelihood ratio statistic; (b): Score statistic; (c): Wald statistic; (d-f): Q-Q plots for (a), (b) and (c).

4.2. A second application

As an example where scale, shape and the proportion of immunes parameters may depend on covariates, consider the ovarian cancer data given by Edmunson *et al.* ([7]) and Therneau ([29]) (see also, [20] pp. 134 and [5] pp. 142).

The response variable (see, Table 4) was the survival time, in years, for 26 women following randomization to one or other of the two chemotherapy treatments. In Table 4 the censor indicator variable is 1 if t_i is an observed survival time and 0 if t_i is a right-censored observation. As pointed out in [20], we notice that large survival times tend to be censored, so there is some evidence of the existence of an immune component. To verify the possible difference between two treatments (treatment 1: standard chemotherapy — cyclophosphamide alone; treatment 2: combined chemotherapy — cyclophosphamide combined with adriamycin), [5] considered the usual Weibull regression model with covariates affecting only the scale parameter and concluded that there is a nonsignificant difference between the two treatments. In fact, the final model considered by Collett (1994) included age and treatment as covariates.

Table 4: Survival times (in years) of ovarian cancer patients.

Survival Time Group 1	Censor Indicator	Survival Time Group 2	Censor Indicator
0.1616	1	0.9671	1
0.3151	1	1.0000	1
0.4274	1	1.2712	1
0.7342	1	1.3014	1
0.9014	1	1.5425	1
1.1808	1	1.0329	0
1.7479	1	1.1534	0
1.2274	0	2.0384	0
1.3068	0	2.1068	0
2.2000	0	2.1096	0
2.3425	0	3.0932	0
2.8493	0	3.3041	0
3.0301	0	3.3616	0

From the survival curves (see, Figure 3), we observe that there are large censored observations, which could indicate the presence of immune individuals [20]. Therefore, we assume the model (2.2) with $S_0(t)$ given by (2.3) with $\log(\frac{p_i}{1-p_i}) = \eta_0 + \eta_1 x_i$, $\log(\mu_i) = \alpha_0 + \alpha_1 x_i$ and $\log(\gamma_i) = \beta_0 + \beta_1 x_i$, where x_i taking the value 1 if individual i is in the treatment group 1 or the value 2 if i is in the treatment group 2.

In this way, we can have the following hypothesis tests:

$H_0\colon \eta_1 = 0$ (no treatment effect in the proportion of cured patients),
$H_0\colon \alpha_1 = 0$ (no treatment effect in the ratio of susceptible patients) or
$H_0\colon \beta_1 = 0$ (no treatment effect in the shape of the lifetime distribution).

In Table 5 we have the MLE and their asymptotic standard-deviation estimates considering 4 models:

Model 1: $\log(\frac{p_i}{1-p_i}) = \eta_0$, $\log(\mu_i) = \alpha_0$ and $\log(\gamma_i) = \beta_0$;
Model 2: $\log(\frac{p_i}{1-p_i}) = \eta_0$, $\log(\mu_i) = \alpha_0 + \alpha_1 x_i$ and $\log(\gamma_i) = \beta_0$;
Model 3: $\log(\frac{p_i}{1-p_i}) = \eta_0$, $\log(\mu_i) = \alpha_0 + \alpha_1 x_i$ and $\log(\gamma_i) = \beta_0 + \beta_1 x_i$ and
Model 4: $\log(\frac{p_i}{1-p_i}) = \eta_0 + \eta_1 x_i$, $\log(\mu_i) = \alpha_0 + \alpha_1 x_i$ and $\log(\gamma_i) = \beta_0 + \beta_1 x_i$.

Locally at the MLE, the values of $-2\log(likelihood)$ are given by 49.3512 (Model 1), 48.1652 (Model 2), 40.6565 (Model 3) and 40.2318 (Model 4). We observe that Model 4 seems to give better fit for the data. This result is corroborated by Figure 3, where we have the plots of the fitted survival curves obtained from Models 2, 3 and 4 and the nonparametric Kaplan–Meier survival curve. We omitted the fitted survival curve from Model 1, which is very far from the Kaplan–Meier survival curve.

Table 5: Maximum likelihood estimates — long-term survivors models.

	Parameter					
Model	η_0	β_0	α_0	α_1	β_1	η_1
1	0.0284	0.7457	0.1423			
	(0.4300)	(0.2658)	(0.1572)			
2	0.0614	0.7222	-0.3759	0.3600		
	(0.4464)	(0.2663)	(0.5293)	(0.3764)		
3	0.0420	-1.0535	-0.4173	0.3482	1.4744	
	(0.4240)	(0.7314)	(0.5749)	(0.2936)	(0.4686)	
4	0.8870	-1.0782	-0.3627	0.3201	1.4833	-0.5614
	(1.3954)	(0.7615)	(0.6232)	(0.3175)	(0.4812)	(0.8725)

It is important to point out that in this application we have a small data set (26 patients) and should be careful to conclude that model 4 provides a better fit. In fact, model 3 and model 4 give similar fits for the survival curves (see Figure 3) and the difference $40.6565 - 40.2318 = 0.4247$ is nonsignificant. In this case the cured proportions and rates of failure do not seem to differ significantly between the treatment groups.

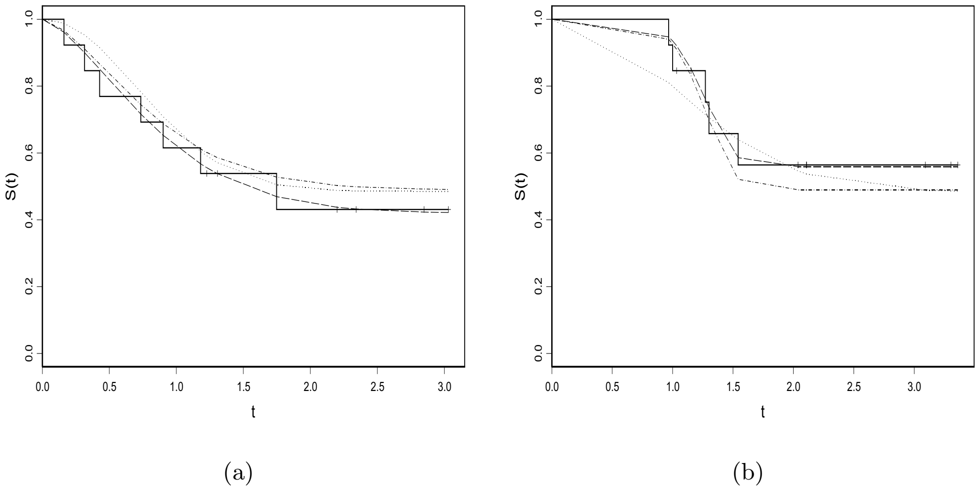


Figure 3: Survival curves.
(a): standard chemotherapy;
(b): combined chemotherapy;
(– – –) Kaplan–Meier;
(· · ·) model 2;
(—·—) model 3;
(——) model 4.

5. CONCLUDING REMARKS

In this paper, we consider a general class of survival models where the shape, the scale and the incidence probability parameters can be dependent on covariates. The major advantage of the general survival class of models lies on its ability to accommodate several usual survival models. From the practical viewpoint the methodology can be implemented straightforwardly and runs immediately using existing statistical packages.

ACKNOWLEDGMENTS

This work has been supported by the Brazilian Organizations CAPES and CNPq. The authors would like to thank the referees for their helpful comments.

Appendix A — Maximum Likelihood, First Application

In this appendix, we present the SAS code used to get the maximum likelihood estimates presented the both examples. The optimization of the log-likelihood was made by using the nonlinear programming SAS procedure considering the trust-region algorithm, [6, 24].

Listing 1: Single Weibull model.

```
proc nlp data=hirose tech=tr phes cov=2 vardef=n;
 max L;
 parms alpha0 = 6.0, beta0   = 1.0;
 mu      = exp(alpha0);
 beta    = exp(beta0);
 logH    = log(beta)-beta*log(mu)+beta*log(t);
 logS    = -(t/mu)**beta;
 L       = delta*logH+logS;
 by voltage;
run;
```

Listing 2: Weibull model with constant shape parameter.

```
proc nlp data=hirose tech=tr phes cov=2 vardef=n;
 max L;
 parms alpha0 = 6.0, alpha1 = 0.9, beta0 = 1.0;
 mu      = exp(alpha0+alpha1*voltage);
 beta    = exp(beta0);
 logH    = log(beta)-beta*log(mu)+beta*log(t);
 logS    = -(t/mu)**beta;
 L       = delta*logH+logS;
run;
```

Listing 3: Weibull model with nonconstant shape parameter.

```
proc nlp data=hirose tech=tr phes cov=2 vardef=n;
 max L;
 parms
 alpha0 = 6.0, alpha1 = -2.0, beta0 = 2.2, beta1 = -0.4;
 mu      = exp(alpha0+alpha1*voltage);
 beta    = exp(beta0+beta1*voltage);
 logH    = log(beta)-beta*log(mu)+beta*log(t);
 logS    = -(t/mu)**beta;
 L       = delta*logH+logS;
run;
```

Appendix B — Maximum Likelihood, Second Application

Listing 4: Long-term survivors model — model 4.

```
proc nlp data = dados tech=tr cov=2 vardef=n phes;
  max L;
  parms   alpha0 =-0.4, alpha1 = 0.3, beta0  = -1.0,
          beta1  = 1.4, g0     = 0.0, g1     = 0.0;
  mu      = exp(alpha0+alpha1*treatment);
  beta    = exp(beta0+beta1*treatment);
  p       = exp(g0+g1*x1)/(1+exp(g0+g1*treatment));
  h       = (beta/mu)*(t/mu)**(beta-1);
  S       = exp(-(t/mu)**beta);
  Lc      = log(p)+log(h)+log(S);
  Li      = log(1-p+p*S);
  L       = delta*Lc+(1-delta)*Li;
run;
```

REFERENCES

[1] Anderson, M.K. (1991). A nonproportional hazards Weibull accelerated failure time regression model, *Biometrics*, **47**, 281–288.

[2] Barbosa, E.P.; Colosimo, E.A., and Louzada-Neto, F. (1996). Accelerated life tests analysed by a piecewise exponential distribution via generalized linear models, *IEEE Transactions on Reliability*, **45**, 4, 619–623.

[3] Cepeda, E., and Gamerman, D. (2000). Bayesian modeling of variance heterogeneity in normal regression models, *Brazilian Journal of Probability and Statistics*, **14**, 207–221.

[4] Chan, C.K. (1991). Temperature-dependent standard deviation of log(failure time) distributions, *IEEE Transactions on Reliability*, **40**, 2, 157–160.

[5] Collett, D. (1994). *Modelling Survival Data in Medical Research*, Chapman and Hall, New York.

[6] Dennis, J.E.; Gay, D.M. and Welsch, R.E. (1981). An adaptive nonlinear least-squares algorithm, *ACM Transactions on Mathematical Software*, **7**, 348–368.

[7] Edmunson. J.H.; Fleming, T.R.; Decker, D.G.; Malkasian, G.D.; Jorgenson, E.O.; Jeffries, J.A.; Webb, M.J. and Kvols, L.K. (1979). Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease. *Cancer Treatment Reports*, **63**, 241–247.

[8] Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics*, **38**, 1041-1046.

[9] Green, P.J. (1984). Iteratively re-weighted least squares for maximum likelihood estimation and some robust alternatives, *Journal of the Royal Statistical Society B*, **46**, 149–192.

[10] Hartmann, W. (1992). *Applications of Nonlinear Optimization with PROC NLP and SAS/IML Software*, Technical Report, Cary, N.C.: SAS Institute Inc.

[11] Hartmann, W. (1992). *The NLP Procedure: Extended User's Guide*, Cary: SAS Institute Inc.

[12] Hirose, H. (1993). Estimation of threshold stress in accelerated life-testing, *IEEE Transactions on Reliability*, **42**, 650–657.

[13] Hsie, F. (2001). On heteroscedastic hazards regression models: theory and application, *Journal of the Royal Statistical Society B*, **63**, 1, 63–79.

[14] Ihaka, R. and Gentleman, R.R. (1996). A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

[15] Jensen, S.T.; Johansen, S. and Lauritzen, S.L. (1991). Globally convergent algorithm for maximizing a likelihood function, *Biometrika*, **78**, 4, 867–877.

[16] Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*, John Wiley, New York.

[17] Louzada-Neto, F. (1997). Extended hazard regression model for reliability and survival analysis, *Lifetime Data Analysis*, **3**, 367–381.

[18] Louzada-Neto, F. (1999). Modelling lifetime data by hazard models: a graphical approachs, *Applied Stochastic Models in Business and Industry*, **15**, 123–129.

[19] Louzada-Neto, F. (2001). Bayesian analysis for hazard models with nonconstant shape parameters, *Computational Statistics*, **16**, 243-254.

[20] Maller, R.A. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*, John Wiley, New York.

[21] Meeker, W.Q. and Escobar, L.A. (1993). A review of recent research and current issues in accelerated testing, *International Statistical Review*, **61**, 1, 147–168.

[22] Meeker, W.Q. and Escobar, L.A. (1998). *Statistical Methods for Reliability Data*, John Wiley, New York.

[23] Meeter, C.A. and Meeker, W.Q. (1994). Optimum accelerated life tests with a nonconstant scale parameter, *Technometrics*, **36**, 71–83.

[24] Moré, J.J. and Sorensen, D.C. (1983). Computing a Trust-Region Step, *SIAM Journal on Scientific and Statistical Computing*, **4**, 553–572.

[25] Nelson, W. (1984). Fitting of fatigue curves with nonconstant standard deviation to data with runouts, *Journal of Testing and Evaluation*, **12**, 69–77.

[26] Pascual, F.G. and Meeker, W.Q. (1997). Regression analysis of fatigue data with runouts based on a model with nonconstant standard deviation and a fatigue limit parameter, *Journal of Testing and Evaluation*, **25**, 292–301.

[27] Smith, R.L. (1991). *Weibull regression models for reliability data*. In "Reliability Engineering and System Safety", Elsevier Publishers, 55–77.

[28] Tanizaki, H. and Zhang, X. (2001). Posterior analysis of the multiplicative heterocedasticity model, *Communications in Statistics. Theory and Methods*, **30**, 5, 855–874.

[29] Therneau, T.M. (1986). *The COXREGR Procedure*, In "SAS SUGI Supplemental Library Users's Guide, Version 5", SAS Institute Inc., Gary, IN.

[30] Wang, W. and Kececioglu, D.B. Fitting the Weibull log-linear model to accelerated life-test data, *IEEE Transactions on Reliability*, **49**, 2, 217–223.

[31] Zhou, X., Stroupe, K.T. and Tierney, W.M. (2001). Regression analysis of health care charges with heteroscedasticity, *Applied Statistics*, **50**, 3, 303–312.

ON THE CONNECTION BETWEEN THE DISTRIBUTION OF EIGENVALUES IN MULTIPLE CORRESPONDENCE ANALYSIS AND LOG-LINEAR MODELS *

Authors: S. Ben Ammou
– Department of Quantitative Methods, Faculté de Droit et des Sciences Economiques et Politiques de Sousse, Tunisia, (saloua.benammou@fdseps.rnu.tn)

G. Saporta
– Chaire de Statistique Appliquée & CEDRIC, Conservatoire National des Arts et Métiers, Paris, France (saporta@cnam.fr)

Received: December 2002 Revised: September 2003 Accepted: October 2003

Abstract:

- Multiple Correspondence Analysis (MCA) and log-linear modeling are two techniques for multi-way contingency table analysis having different approaches and fields of applications. Log-linear models are interesting when applied to a small number of variables. Multiple Correspondence Analysis is useful in large tables. This efficiency is balanced by the fact that MCA is not able to explicit the relations between more than two variables, as can be done through log-linear modeling. The two approaches are complementary. We present in this paper the distribution of eigenvalues in MCA when the data fit a known log-linear model, then we construct this model by successive applications of MCA. We also propose an empirical procedure, fitting progressively the log-linear model where the fitting criterion is based on eigenvalue diagrams. The procedure is validated on several sets of data used in the literature.

Key-Words:

- *Multiple Correspondence Analysis; eigenvalues; log-linear models; graphical models; normal distribution.*

AMS Subject Classification:

- 49A05, 78B26.

*We thank Professor M. Bourdeau for his careful reading.

1. INTRODUCTION

Multiple Correspondence Analysis and log-linear modeling are two very different, but mutually beneficial approaches to analyzing multi-way contingency tables: log-linear models are profitably applied to a small number of variables. Multiple Correspondence Analysis is useful in large tables. This efficiency is balanced by the fact that MCA is not able to explicit relations between more than two variables, as can be done through log-linear modeling. The two approaches are complementary. After a short reminder on MCA and log-linear approaches, we study the distribution of eigenvalues in MCA under modeling hypotheses, especially in the case of independence. At the end we propose an algorithmic approach for fitting log-linear models where the fitting criterion is based on eigenvalues diagram.

2. A SHORT SURVEY OF MULTIPLE CORRESPONDENCE ANALYSIS AND LOG-LINEAR MODELS

We first introduce MCA and log-linear modelling, then we present some works using both methods.

2.1. Multiple Correspondence Analysis

Correspondence Analysis (CA) has quite a long history as a method for the analysis of categorical data. The starting point of this history is usually set in 1935 [28], and since then CA has been reinvented several times. We can distinguish simple CA (CA of contingency tables) and MCA or Multiple Correspondence Analysis (CA of so-called indicator matrices). MCA traces back to Guttman [23], Burt [8] or Hayashi [25]. In France, in the 1960s, Benzecri [6] proposes, other developments of this method. Outside France, MCA has been developed by J. de Leeuw since 1973 [22] under the name of Homogeneity Analysis, and the name of Dual Scaling by Nishisato [38].

Multiple Correspondence Analysis (MCA) is a multidimensional descriptive technique of categorical data. A theoretical version of Multiple Correspondence Analysis of p variables can be defined as the limit, when the number of statistical units increases, of the CA of a complete disjunctive table.

Let X be a complete disjunctive table of p categorical variables $X_1, X_2, ...,$ X_p, with respectively $m_1, m_2, ..., m_p$ modalities observed over a sample of n individuals. CA of this complete disjunctive table is equivalent to the analysis of B [8], where $B = X'X$ is the Burt table associated with X. The two analyses have the same factors, but the eigenvalues in MCA equal to the squared root of the eigenvalues in the CA of the associated Burt table. MCA of X corresponds to the diagonalization of the matrix $\frac{1}{p}(D^{-1}X'X) = \frac{1}{p}(D^{-1}B)$ where $D = \mathrm{Diag}(X'X) = \mathrm{Diag}(B)$.

The structure of the eigenvalue diagram depends on the variable interactions. It is well known that in the case of pairwise independent variables, the q non-trivial eigenvalues are theoretically equal to to $\frac{1}{p}$, where

$$q = \sum_{i=1}^{p} m_i - p \ . \tag{1}$$

2.2. Log-linear modeling

Log-linear modeling is a well-known method for studying structural relationships between categorical variables in a multiple contingency table when all the variables have no particular role. Relatively recent and not as well known in France as MCA, log-linear modeling has many classical references. After first works of Birch [7] in 1963 and Goodman [17], we must mention the basic books of Haberman [24], Bishop, Fienberg & Holland [8], Fienberg [15].

More Recently, Dobson [12], Agresti [1], Christensen [10] have written syntheses on the subject supplemented with personal contributions.

Whittaker [41] devotes a large part of his book to log-linear models before defining associated graphical models.

2.2.1. Log-linear modeling in the binomial case

Let $X = (X_1, X_2, ..., X_p)$ be a k-dimensional random vector, with values in $\{0,1\}^k$. The expression for the k-dimensional probability density of X is:

$$\begin{aligned} f_k(X) \;=\; & p(0,0,...,0)^{(1-x_1)(1-x_2)\cdots(1-x_k)} \cdot p(1,0,...,0)^{x_1(1-x_2)\cdots(1-x_k)} \\ & \cdot p(0,1,...,0)^{(1-x_1)x_2\cdots(1-x_k)} \cdots\ p(0,0,...,1)^{(1-x_1)(1-x_2)\cdots x_k} \\ & \cdots\ p(1,1,...,0)^{x_1x_2\cdots(1-x_k)} \cdots\ p(1,1,...,1)^{x_1x_2\cdots x_k} \ . \end{aligned}$$

We can write the density function as a log-linear expansion:

$$\begin{aligned} \log[f_k(X)] \;=\; & u_o + \sum_{i=1}^{k} u_i\, x_i + \sum_{\substack{i,j=1,\\ i\neq j}}^{k} u_{ij}\, x_i\, x_j + \sum_{\substack{i,j,\ l=1,\\ i\neq j\neq l}}^{k} u_{ijl}\, x_i\, x_j\, x_l \\ & + \cdots + u_{123\ldots k}\, x_1\, x_2 \cdots x_k \end{aligned}$$

where $u_o = \log[p(0,0,...,0)]$, $u_i = \log[\frac{p(0,0,...,0,1,0,...0)}{p(0,0,...,0)}]$ and the u-terms $u_{ij}, ..., u_{123\ldots k}$ are a log cross product ratio in the (k,k) probability table. The u-term u_{ij} is set to zero when X_i and X_j are independent variables.

2.2.2. Log-linear modeling in the multinomial case

Let $X = (X_1, X_2, ..., X_k)$ be a k-dimensional random vector, with values in $\{0, 1, ..., m_1-1\} \times \{0, 1, ..., m_2-1\} \times ... \times \{0, 1, ..., m_k-1\}$ instead of in $\{0, 1\}^k$ as in the preceding case.

The generalisation to the k-dimensional cross-classified multinomial distribution is the log-linear expansion:

$$\log[f_k(X)] \;=\; u_o + \sum_{i=1}^{k} u_i(x) + \sum_{\substack{i,j=1,\\ i\neq j}}^{k} u_{ij}(x) + \sum_{\substack{i,j,\ l=1,\\ i\neq j\neq l}}^{k} u_{ijl}(x) + \cdots + u_{123\ldots k}(x) \;.$$

Each u-term is a coordinate projection function with the coordinates indicated by its index; and each u-term is constrained to be zero whenever one of its indicated coordinates is zero.

The importance of log-linear expansions rests with the fact that many interesting hypotheses can be generated by setting some u-terms to zero.

We are interested particularly in this paper with graphical and hierarchical log-linear models.

2.2.2.1. Graphical log-linear models

Let $G = (K, E)$ be the independence graph of the k-dimensional random vector X, with k vertices in $K = \{1, 2, ..., k\}$ and edge set E. G is the set of pairs (i, j) such that whenever (i, j) is not in E the variables X_i and X_j are independent conditionally on the other variables.

Given an independence graph G, the cross classified multinomial distribution for the random vector X is a graphical model for X, if the distribution of X is different from constraints of the form that for all pair of coordinates not in the edge set E of G, the u-terms constraining the selected coordinates are identically zero.

2.2.2.2. Hierarchical log-linear models

A graphical model satisfies constraints of the form that all u-terms 'above' a fixed point have to be zero to get conditional independence. A larger class of models, the hierarchical models, is obtained by allowing more flexibility in setting the u-terms to zero.

A log-linear model is hierarchical if, whenever one particular u-term is constrained to zero then all higher u-terms containing the same set of subscripts are also set to zero.

We note here that every distribution with a log-linear expansion has an interaction (or independence) graph, and a hierarchical log-linear model is graphical if and only if its maximal u-terms correspond to cliques in the independence graph.

When all the u-terms are non-zero, we have the **saturated** model.

In the case when only the u_i are non-zero, the model is called the **mutual independence model**:

$$\log[f_k(X)] = u_o(x) + \sum_{i=1}^{k} u_i(x) .$$

When only u_i and some of u_{ij} are non-zero, the model is called a **conditional independence model**:

$$\log[f_k(X)] = u_o(x) + \sum_{i=1}^{k} u_i(x) + \sum_{i,j} u_{ij}(x) .$$

These conditional independence models refer to simple interactions between some variables.

2.2.3. Parameters estimation and related tests

Theoretical frequencies are generally estimated using the maximum-likelihood method. Weighted regression, or iterative methods can be also used as well since log-linear models are particular cases of the generalized linear model. Usually the classical χ^2 or the G^2 tests (the likelihood ratio) are used to assess log-linear models. The values of the two statistics increase with the number of variables, and decrease with the number of interactions. The closer the statistics are to zero, the better the models.

Model selection becomes difficult when the number of variables grow: e.g. with four variables there are 167 different hierarchical models. To avoid the "combinatory explosion" we can use criterions based on the Kullback information like the Akaike criterion:

$$AIC = -2\log(\widehat{L}) + 2\,k \qquad (\textit{An Information criterion}) ,$$

or the Schwartz criterion:

$$BIC = -2\log(\widehat{L}) + k\log(n) \qquad (\textit{Bayesian Information criterion}) ,$$

where $\widehat{L}$ is the maximum of the likelihood function (L), and k the number of parameters maximising L.

2.3. Multiple Correspondence Analysis and log-linear model as complementary tools of analysis

In this section, we present some works that show how CA (or MCA) and log-linear modeling can be related. This leads to a better understanding of CA, and to a combined use of both methods.

CA is often introduced without any reference to other methods of statistical treatment of categorical data with proven usefulness and flexibility.

A major difference between CA and most other techniques for categorical data analysis lies in the use of probability models. In log-linear analysis (LLA), for example, a distribution is assumed under which the data are collected, then a log-linear model for the data is hypothesized and estimations are made under the assumption that this probability model is true, and finally these estimates are compared with the observed frequencies to evaluate the log-linear model. In this way it is possible to make inferences about the population on the basis of the sample data.

In CA, it is claimed that no underlying distribution has to be assumed and no model has to be hypothesized, but a decomposition of the data is obtained to study the 'structure' in the data.

A vast literature has been devoted to the assessment of CA (or MCA) and LLA. We briefly report here some of that literature.

Several works compare CA or MCA and LLA. Daudin and Trecourt [11] demonstrate empirically that LLA is better adapted to study global relationships between the variables, in the sense that marginal liaisons are eliminated in the computation of profiles.

Goodman [17],[18],[19],[20],[21] defines two models belonging to the same family: the saturated row column correspondence analysis model as a generalization of MCA, and the row column association model as a generalization of LLA. He demonstrates, with illustrations by examples, that using these models is better than using the classical methods.

Baccini, Mathieu and Mondot [3] use an example to compare the two methods. Jmel [30], De Falguerolles, Jmel and Whittaker [13],[14] use graphical models compared to MCA.

Other works use CA or MCA and LLA as a combined approach to contingency table analysis: Van der Heijden and de Leeuw [26],[27],[28], Novak and Hoffman [39] and others, use CA as a tool for the exploration of the residuals from log-linear models, and give an example of the procedure.

Worsley [42] shows that in certain cases CA leads directly to the appropriate log-linear model.

Lauro and Decarli [31] used AC as a procedure for the identification of best log-linear models.

3. EIGENVALUES IN CORRESPONDENCE ANALYSIS

It is well known that MCA is an extension of CA, although we first present eigenvalues in CA, and some simple rules for the selection of the number of eigenvalues.

3.1. Asymptotic distribution of eigenvalues in Correspondence Analysis

Let N be a contingency table with m_1 rows and m_2 columns, and let us assume that N is the realization of a multinomial distribution $M(n, p_{ij})$ which is realistic. In this framework the observed eigenvalues μ_i are estimators of the eigenvalues λ_i of nP, where P is the table of unknown joint probabilities.

Lebart [32] and O'Neill [34],[35],[36] proved the following result:
if $\mu_i = 0$ then λ_i has the same distribution as the corresponding eigenvalues of a $(m_1-1)(m_2-1)$ degrees of freedom from the Wishart matrix: $W_{(m_1-1)(m_2-1)}(r, l)$ where $r = \min(m_1-1,\, m_2-1)$.

If $\mu_j = 0$ then $\sqrt{\lambda_j}$ is asymptotically normally distributed, but with parameters depending on the unknown p_{ij}. Since it is difficult to test this hypothesis, some authors have proposed a bootstrap approach, which unfortunately is not valid: since the empirical eigenvalues, on which the replication is based, are generally not null, we cannot observe the distribution based on the Wishart matrix.

3.2. Malinvaud's test

Based upon the reconstitution formula, which is a weighted singular value decomposition of N:

$$n_{ij} \;=\; \frac{(n_{i.}\, n_{.j})}{n} \left(1 + \frac{\sum_k (a_{ik}\, b_{ki})}{\sqrt{\lambda_k}} \right),$$

where a_{ik}, b_{ik} are the factorial components associated to the row and column profiles.

We may use a chi-square test comparing the observed n_{ij} from a sample of size n to the expected frequencies under the null-hypothesis H_k of only k non zeros. The μ_i weighted least squares estimates of these expectations are precisely the $\widetilde{n_{ij}}$ of the reconstitution formula with its first k terms. We then compute the classical chi-square goodness of fit statistic:

$$Q_k = \sum_i \sum_j \frac{(\widetilde{n}_{ij} - n_{ij})^2}{\widetilde{n}_{ij}} \; .$$

If $k = 0$ (independence), Q_0 is compared to a chi-square with $(m_1 - 1)\,(m_2 - 1)$ degrees of freedom. Under H_k, Q_k is asymptotically distributed like a chi-square with $(m_1 - k - 1)\,(m_2 - k - 1)$ degrees of freedom. However Q_k suffers from the following drawback: if n_{ij} is small, $\widetilde{n}_{ij}$ can be negative and the test statistic cannot be used. This is not the case for the modification proposed by E. Malinvaud [37] who proposed to use $\frac{(n_{i.}\, n_{.j})}{n}$ instead of $\widetilde{n}_{ij}$ for the denominator. Furthermore, this leads to a simple relation with the sum of the discarded eigenvalues:

$$Q'_k = \sum_i \sum_j \frac{(\widetilde{n}_{ij} - n_{ij})^2}{\frac{(n_{i.} n_{.j})}{n}} = n\,(\lambda_{k+1} + \lambda_{k+2} + \ldots + \lambda_r) \; .$$

Q'_k is also asymptotically distributed like a chi-square with $(p - k - 1)\,(q - k - 1)$ degrees of freedom.

4. BEHAVIOUR OF EIGENVALUES IN MCA UNDER MODELING HYPOTHESES

Let $X = (X_1|X_2|\ldots|X_p)$ be a disjunctive table of p categorical variables X_i (with respectively m_i modalities) observed on a sample of n individuals, and q the number of non trivial eigenvalues (as defined in § 2.1).

Multiple Correspondence Analysis is the CA of disjunctive table X.

The rank of X: $\operatorname{rank}(X) = \min(q{+}1; n)$, so equals $q{+}1$ if $n > q{+}1$.

We suppose, without loss of generality, that n is large enough, which is the usual case. CA factors are the eigenvectors of the matrix $\frac{1}{p}\,D^{-1}B$ (where B and D are defined in § 2.1). So $D^{-1}B$ is a diagonal unit matrix.

Its trace is: $\operatorname{Tr}(D^{-1}B) = \sum_{i=1}^{p} m_i$ and $\frac{1}{p}\operatorname{Tr}(D^{-1}B) = \frac{1}{p}\sum_{i=1}^{p} m_i$.

Since $\sum_{i=1}^{q} \mu_i = \frac{1}{p}\sum_{i=1}^{p} m_i - 1$, we can conclude that

(2) $$\frac{1}{q}\sum_{i=1}^{q} \mu_i = \frac{1}{p}$$

and

$$(3) \qquad \sum_{i=1}^{q} (\mu_i)^2 = \frac{1}{p^2} \sum_{i=1}^{p} (m_i - 1) + \frac{1}{p^2} \sum_{i \neq j} \sum \varphi_{ij}^2$$

where φ_{ij}^2 is the observed Pearson's φ^2 crossing of X_i with X_j, and

$$\varphi^2 = \frac{1}{n} \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i\cdot}\, n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}\, n_{\cdot j}}{n}} = \frac{\chi^2}{n} ,$$

($n_{i\cdot}$ and $n_{\cdot j}$ are margin effectives).

Although MCA is an extension of CA, the results of § 3 are not valid and one cannot use Malinvaud's test: elements of X being 0 or 1 and not frequencies, Q_k and Q'_k do not follow a chi-square distribution.

However it is possible to get information about the dispersion of the q eigenvalues in particular cases [5].

4.1. Distribution of eigenvalues in MCA under independence hypothesis

Under the hypothesis of pairwise independence of the variables X_i, all $\varphi_{ij}^2 = 0$ and equation (3), becomes

$$\sum_{i=1}^{q} (\mu_i)^2 = \frac{1}{p^2} \sum_{i=1}^{p} (m_i - 1) .$$

Using (2) we get

$$\sum_{i=1}^{q} (\mu_i)^2 = \frac{1}{p^2}\, q ,$$

and finally:

$$\sum_{i=1}^{q} (\mu_i)^2 = \frac{1}{p^2} = \left[\frac{1}{q} \sum_i (\mu_i)\right]^2 .$$

Since the mean of the squared μ_i equals their squared means only if all the terms are equal, we can conclude that all the eigenvalues have the same value, so that:

$$\mu_i = \frac{1}{p} , \qquad \forall\, i .$$

We conclude that the theoretical MCA (i.e. for the population), under the hypothesis of pairwise independence of the variables X_i leads to one q-multiple non-trivial non-zero eigenvalue $\lambda = \frac{1}{p}$. And the eigenvalue diagram has the particular shape shown in *Figure 1*:

λ_I	Eigenvalues diagram
λ_1	****************************
λ_2	****************************
λ_3	****************************
λ_4	****************************
λ_5	****************************
$\vdots$	****************************
λ_q	****************************

Figure 1: Theoretical eigenvalues diagram in the independence case.

This result is not true when we have a finite sample, since sampling fluctuations make the observed $\varphi_{ij}^2 \neq 0$. Although the trace of $\frac{1}{p}(D^{-1}B)$ and $\overline{\mu}$ the mean of the observed non-trivial eigenvalues, are constants, we observe q different non-trivial eigenvalues $\mu_i \neq \frac{1}{p}$, and the eigenvalue diagram takes the shape shown in *Figure 2*:

λ_I	Eigenvalues diagram
λ_1	****************************
λ_2	***************************
λ_3	**************************
λ_4	*************************
λ_5	************************
$\vdots$	***********************
λ_q	**********************

Figure 2: Observed eigenvalues diagram in the independence case.

4.1.1. Dispersion of eigenvalues

Let S_μ^2 be the measure of μ_i around $\frac{1}{p}$ given by:

$$S_\mu^2 = \frac{1}{q}\sum_{i=1}^{q}\left(\mu_i - \frac{1}{p}\right)^2 = \frac{1}{q}\sum_{i=1}^{q}(\mu_i)^2 - \frac{1}{p^2} ,$$

which implies

$$\sum_{i=1}^{q}(\mu_i)^2 = q\left(S_\mu^2 + \frac{1}{p^2}\right) .$$

Using equations (1)&(3), we have:

$$\sum_{i=1}^{q}(\mu_i)^2 = \frac{q}{p^2} + \frac{1}{p^2}\sum_{i\neq j}\sum \varphi_{ij}^2 = \frac{q}{p^2} + \frac{1}{n\,p^2}\sum_{i\neq j}\sum \chi_{ij}^2 .$$

Under the hypothesis of pairwise independence of the variables, the χ^2_{ij} are realizations of $\chi^2_{(m_i-1)(m_j-1)}$ variables, so their expected values are $(m_i-1)\,(m_j-1)$.

We can then easily compute $E(\sum_{i=1}^{q}(\mu_i)^2)$, and get:

$$E\left(\sum_{i=1}^{q}(\mu_i)^2\right) = \frac{q}{p^2} + \frac{1}{p^2}\,\frac{1}{n}\,\sum_{i\neq j}\sum (m_i-1)\,(m_j-1)\ .$$

Finally:

$$E(S^2_\mu) = \frac{1}{q}\,E\left(\sum_{i=1}^{q}(\mu_i)^2\right) - \frac{1}{p^2}$$

and we obtain:

$$E(S^2_\mu) = \frac{1}{p^2}\,\frac{1}{n}\,\frac{1}{q}\,\sum_{i\neq j}\sum (m_i-1)\,(m_j-1)\ .$$

Now, since $E(S^2_\mu)=\sigma^2$, we may assume that $\frac{1}{p}\pm 2\,\sigma$ contains roughly 95% of the eigenvalues. Moreover, since the kurtosis of the set of eigenvalues is lower than for a normal distribution, this proportion is actually probably larger then 95%.

4.1.2. Estimation of the Burt table

Let X be the disjunctive table associated to p categorical variables X_i, with m_i modalities respectively, observed on a sample of n individuals, where $X_i = (X_{i1}, X_{i2}, ..., X_{im_i})$, X is a matrix made (of p-block) of p blocks X_i

$$X = (X_1 \mid X_2 \mid ... \mid X_i \mid ... \mid X_p)\ .$$

Let $(X^j_{i1}, X^j_{i2}, ..., X^j_{ip})$ be the observed value of X_i on the j^{th} individual.

We can write

$$X = \begin{bmatrix} X^1_{11} & \cdots & X^1_{1m_1} & X^1_{21} & \cdots & X^1_{2m_2} & \cdots & X^1_{p1} & \cdots & X^1_{pm_p} \\ X^2_{11} & \cdots & X^2_{1m_1} & X^2_{21} & \cdots & X^2_{2m_2} & \cdots & X^2_{p1} & \cdots & X^2_{pm_p} \\ & \vdots & & & \vdots & & \vdots & & \vdots & \\ X^n_{11} & \cdots & X^n_{1m_1} & X^n_{21} & \cdots & X^n_{2m_2} & \cdots & X^n_{p1} & \cdots & X^n_{pm_p} \end{bmatrix}\ .$$

The Burt table of X is then

$$B = \begin{bmatrix} X'_1X_1 & X'_1X_2 & \cdots & X'_1X_p \\ X'_2X_1 & X'_2X_2 & \cdots & X'_2X_p \\ \vdots & \vdots & \ddots & \vdots \\ X'_pX_1 & X'_pX_2 & \cdots & X'_pX_p \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ B_{p1} & B_{p2} & \cdots & B_{pp} \end{bmatrix}\ ,$$

where

$$
\begin{aligned}
B_i &= B_{ii} \\
&= X_i' X_i = \begin{bmatrix}
\sum_{j=1}^{n} (X_{1i}^j)^2 & \sum_{j=1}^{n} (X_{1i}^j)\,(X_{2i}^j) & \cdots & \sum_{j=1}^{n} (X_{1i}^j)\,(X_{m_i i}^j) \\
\sum_{j=1}^{n} (X_{2i}^j)\,(X_{1i}^j) & \sum_{j=1}^{n} (X_{2i}^j)^2 & \cdots & \sum_{j=1}^{n} (X_{2i}^j)\,(X_{m_i i}^j) \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{j=1}^{n} (X_{m_i i}^j)\,(X_{1i}^j) & \sum_{j=1}^{n} (X_{m_i i}^j)\,(X_{2i}^j) & \cdots & \sum_{j=1}^{n} (X_{m_i i}^j)^2
\end{bmatrix}
\end{aligned}
$$

and

$$
X_{ki}^j = \begin{cases} 0 \\ 1 \end{cases}
$$

with $\sum_{k=1}^{m_i} X_{ki}^j = 1$. Since there is only one k in $\{1, ..., m_i\}$ such as $X_{ji}^k = 1$, all other being zero, we obtain:

$$
\sum_{k=1}^{n} (X_{ki}^j)^2 \; = \; \sum_{k=1}^{n} X_{ki}^j \qquad \text{in} \;\; \{1, ..., n\}, \quad \forall\, k \in \{1, ..., m_i\}
$$

and

$$
\sum_{k=1}^{n} (X_{ki}^j)\,(X_{k'i}{}^j) \; = \; 0 \qquad \forall\, k, \quad k \in \{1, ..., m_i\}\ .
$$

And so can conclude that $\forall\, i = 1, ..., p$ the diagonal sub-matrices of the Burt table are themselves diagonal matrices:

$$
X_i' X_i \; = \; B_i \; = \begin{bmatrix}
\sum_{j=1}^{n} (X_{1i}^j)^2 & & & & 0 \\
& \ddots & & & \\
& & \sum_{j=1}^{n} (X_{ki}^j)^2 & & \\
& & & \ddots & \\
0 & & & & \sum_{j=1}^{n} (X_{m_i i}^j)^2
\end{bmatrix} .
$$

Furthermore, we know that

$$
\sum_{k=1}^{m_i} \left(\sum_{j=1}^{n} X_{ki}^j \right) = \sum_{k=1}^{m_i} (n_{ki}) \; = \; n \ ,
$$

where

$$
n_{ki} \; = \; \sum_{j=1}^{n} X_{ki}^j \; = \; n_i^k
$$

is the number of individuals that have the k^{th} modality of the i^{th} variable (for $1 \le i \le p$ and $1 \le k \le m_i$).

So the diagonal sub-matrices of the Burt table are:

$$B_i = B_{ii} = \begin{bmatrix} n_i^1 & & & & 0 \\ & \ddots & & & \\ & & n_i^k & & \\ & & & \ddots & \\ 0 & & & & n_i^{m_i} \end{bmatrix} \qquad \text{where } \sum_{k=1}^{m_i} \frac{n_{ki}}{n} = 1 \quad \forall\, i{=}1,...,p\ .$$

Consider now two independent variables X_α and X_β amongst the p variables having respectively m_α and m_β modalities.

Let B_α be the (m_α, m_α) square matrix $B_\alpha = X'_\alpha X_\alpha$, and $B_{\alpha\beta}$ the (m_α, m_β) rectangular matrix $B_{\alpha\beta} = X'_\alpha X_\beta$.

We have

$$(B_\alpha)_{ii} = \sum_{k=1}^{n} X_{i\alpha}^k = X_{\cdot i}^\alpha \qquad \text{and} \qquad (B_\alpha)_{ij} = 0 \quad \text{if } i \neq j\ ,$$

and where $(B_{\alpha\beta})_{ij} = X_{i\alpha}^k\, X_{i\beta}^k \leq n$.

Under the hypothesis that X_α and X_β are independent

$$(B_{\alpha\beta})_{ij} = \frac{(B_\alpha)_{ij}\,(B_\beta)_{ij}}{n} = \frac{X_{\cdot i}^\alpha\, X_{\cdot i}^\beta}{n}\ .$$

Since $X_{\cdot i}^\alpha = n_i^\alpha$ and $X_{\cdot i}^\beta = n_i^\beta$, we can write

$$\left[(B_{\alpha\beta})_{ij} = \sum_{k=1}^{n} X_{ki}^\alpha\, X_{kj}^\beta = \frac{X_{\cdot i}^\alpha\, X_{\cdot i}^\beta}{n} = \frac{n_i^\alpha\, n_j^\beta}{n} \right]$$

and, more generally, we can conclude that

$$X'_i X_j = B_{ij} = \begin{bmatrix} \frac{n_1^i n_1^j}{n} & \frac{n_1^i n_2^j}{n} & \cdots & \frac{n_1^i n_{m_j}^j}{n} \\ \frac{n_2^i n_1^j}{n} & \frac{n_2^i n_2^j}{n} & \cdots & \frac{n_2^i n_{m_j}^j}{n} \\ \vdots & \vdots & & \vdots \\ \frac{n_{m_i}^i n_1^j}{n} & \frac{n_{m_i}^i n_2^j}{n} & \cdots & \frac{n_{m_i}^i n_{m_j}^j}{n} \end{bmatrix}$$

if the p variables are mutually independent.

Now consider a sample of p multinomial random variables X_i. Let $p_i^k = p_{ik}$ be the probability that an individual be in the k^{th} category of the i^{th} variable, and p_{ij}^k be the probably that the j^{th} individual be in the k^{th} category of the i^{th} variable.

The observed Burt table is:

$$B = X'X = \begin{bmatrix} X_1'X_1 & X_1'X_2 & \cdots & X_1'X_p \\ X_2'X_1 & X_2'X_2 & \cdots & X_2'X_p \\ \vdots & \vdots & \vdots & \vdots \\ X_p'X_1 & X_p'X_2 & \cdots & X_p'X_p \end{bmatrix},$$

with

$$X_i'X_i = N_i = \begin{bmatrix} \sum_{j=1}^{n}(X_{ij}^1)^2 & & & & 0 \\ & \ddots & & & \\ & & \sum_{j=1}^{n}(X_{ki}^j)^2 & & \\ & & & \ddots & \\ 0 & & & & \sum_{j=1}^{n}(X_{m_i i}^j)^2 \end{bmatrix} = \operatorname{diag}\{n_i^1, ..., n_i^{m_i}\} .$$

But $n_i^k = \sum_{j=1}^{n}(X_{ki}^i)^2 = np_i^k$ and $\sum_{k=1}^{m_i} p_i^k = 1$, so that $\sum_{k=1}^{m_i} n_i^k = n\sum_{k=1}^{m_i} p_i^k = n,\ \forall\, i=1, ..., p$

and $$X_i'X_j = \begin{bmatrix} n\,p_i^1 & & & & 0 \\ & \ddots & & & \\ & & n\,p_i^k & & \\ & & & \ddots & \\ 0 & & & & n\,p_i^{m_i} \end{bmatrix}.$$

Since X_i and X_j are independent variables, $X_i'X_j = N_{ij}$ and $(N_{ij})_{kk'} = (X_i'X_j)_{kk'} = n\,p_i^k p_j^{k'}$, which implies

$$X_i'X_j = N_{ij} = \begin{bmatrix} n\,p_1^i p_1^j & n\,p_1^i p_2^j & \cdots & n_1^i n_{m_j}^j \\ n\,p_2^i p_1^j & n\,p_2^i p_2^j & \cdots & n\,p_2^i p_{m_j}^j \\ \vdots & \vdots & & \vdots \\ n\,p_{m_i}^i p_1^j & n\,p_{m_i}^i p_2^j & \cdots & n\,p_{m_i}^i p_{m_j}^j \end{bmatrix}.$$

The maximum-likelihood estimator of p_i^k is $\hat{p}_i^k = \frac{n_i^k}{n}$, so

$$\hat{N}_i = \begin{bmatrix} n_i^1 & & & & 0 \\ & \ddots & & & \\ & & n_i^k & & \\ & & & \ddots & \\ 0 & & & & n_i^{m_i} \end{bmatrix} = B_{ii}$$

and

$$\hat{N}_{ij} = \begin{bmatrix} \frac{n_1^i n_1^j}{n} & \frac{n_1^i n_2^j}{n} & \dots & \frac{n_1^i n_{m_j}^j}{n} \\ \frac{n_2^i n_1^j}{n} & \frac{n_2^i n_2^j}{n} & \dots & \frac{n_2^i n_{m_j}^j}{n} \\ \vdots & \vdots & & \vdots \\ \frac{n_{m_i}^i n_1^j}{n} & \frac{n_{m_i}^i n_2^j}{n} & \dots & \frac{n_{m_i}^i n_{m_j}^j}{n} \end{bmatrix} = B_{ij} \ .$$

We can conclude that the the maximum-likelihood estimator $\hat{B}$ of the theoretical Burt table is $\tilde{B}$ the observed one. Using the invariance functional propriety we can affirm that the maximum-likelihood estimators of the eigenvalues of $D^{-1}B$ are the eigenvalues of $D^{-1}\tilde{B}$, so that each μ_i is the maximum-likelihood estimator of $\lambda_i = \lambda$.

Maximum-likelihood estimators are asymptotically normal, and so, asymptotically, each μ_i is normally distributed. But due to the fact that eigenvalues are ordered, the eigenvalues are not identically and independently distributed. However:

$$E(\mu_1) > \frac{1}{p}, \quad E(\mu_q) < \frac{1}{p} \qquad \text{but} \ \ E(\mu_1) \underset{n\to\infty}{\longrightarrow} \frac{1}{p} \quad \text{and} \quad E(\mu_q) \underset{n\to\infty}{\longrightarrow} \frac{1}{p} \ .$$

Furthermore the eigenvalue variances are not the same. And from simulations of large samples of n observations ($n = 100$, ..., $n = 10\,000$), we notice that the convergence of the eigenvalue distribution to a normal one is slow, especially for the extremes (μ_1 and μ_q), even for very large samples [4].

4.2. Distribution of eigenvalues in MCA under non-independence hypotheses

4.2.1. Distribution of the theoretical eigenvalues

Let μ be an eigenvalue of $D^{-1}X'X$. Since μ can be also obtained by diagonalization of $\frac{1}{p}XD^{-1}X'$, μ is a solution of $\frac{1}{p}XD^{-1}X'z = z$, where z is an eigenvector associated to μ.

So

$$\frac{1}{p}\left(\sum_{i=1}^{p} X_i\big(X_i'X_i\big)^{-1}X_i'\right)z = \mu\, z \quad\Longleftrightarrow\quad \frac{1}{p}\sum_{i=1}^{p} P_i\, z = \mu\, z\ ,$$

where $P_i = \sum_{i=1}^{p} X_i(X_i'X_i)^{-1}X_i'$ is the orthogonal projector on the space spanned by linear combinations of the indicators of variables categories X_i.

Let A_i the centered projector associated to P_i:

$$A_i = P_i - \frac{1_{m_i m_i}}{n} \qquad \text{where}\ \ 1_{m_i m_i} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}\ .$$

And so we get

$$\frac{1}{p}\sum_{i=1}^{p} A_i\, z = \mu\, z\ . \tag{4}$$

4.2.1.1. The Case of two-way interactions

Let us assume that among the p studied variables, there is a two-way interaction between X_j and X_k, and that the $(p-2)$ reminding variables are mutually independent. Multiplying equation (4) by A_j we get:

$$\frac{1}{p}\Big(\underbrace{A_jA_1}_{0} + \underbrace{A_jA_2}_{0} + \cdots + \underbrace{A_jA_j}_{Aj} + \cdots + A_jA_k + \cdots + \underbrace{A_jA_p}_{0}\Big)z = \mu\, A_j\, z\ ,$$

since all variables are pairwise independent except X_j, X_k, and the A_i are orthogonal projectors. Thus:

$$A_jA_k\, z = (p\,\mu - 1)\, A_j\, z\ . \tag{5}$$

Similarly, multiplying (4) by A_k, we get:

$$A_kA_j\, z = (p\,\mu - 1)\, A_k\, z\ . \tag{6}$$

Now let us multiply (5) by A_k to get:

$$A_kA_jA_k\, z = (p\,\mu - 1)\, A_kA_j\, z\ .$$

Using (6) we obtain

$$A_kA_j\, \underbrace{A_k\, z}_{z_1} = (p\,\mu - 1)^2\, \underbrace{A_k\, z}_{z_1}\ .$$

With the notation $\lambda = (p\,\mu - 1)^2$, we finally write:

$$A_k A_j\, z_1 \;=\; \lambda\, z_1\ . \tag{7}$$

Equation (7) implies that λ is an eigenvalue of the product of the centered projector $A_k A_j$ associated to the eigenvector z_1.

In general: $\forall\, j, k = 1, ..., p$, if there is an interaction between X_j and X_k, the orthogonal projector $A_j A_k$ admits a non zero eigenvalue $\lambda = (p\,\mu - 1)^2$.
If $\lambda \neq 0 \Leftrightarrow \mu \neq \frac{1}{p}$, the trace of Burt table being constant, there is, at least, another eigenvalue not equal to $\frac{1}{p}$.

Let n_0 be the number of eigenvalue non equal to $\frac{1}{p}$, so that $\sum_{i=1}^{n_0} \lambda_i = \frac{n_0}{p}$.

Theoretically, (except for the particular case, where $\lambda = 1$, for which we have $\mu = \frac{2}{p}$ and $\mu' = 0$), the number of non-trivial-eigenvalues greater than $\frac{1}{p}$ is equal to the number of non-trivial eigenvalues smaller than $\frac{1}{p}$.

The eigenvalue diagram shape is shown on *Figure 3*:

λ_I	Eigenvalues diagram
λ_1	****************************
λ_2	***************************
λ_3	************************
λ_4	************************
λ_5	************************
$\vdots$	******************
λ_q	*****************

Figure 3: Theoretical eigenvalues diagram in two-way interaction case.

The number n_0 depends on the number of categories of X_j and X_k, on the number of variables and on the number of dependent variables.

Let n_1 be the multiplicity of $\frac{1}{p}$, we will show that $n_1 = q - 2\min((m_j - 1); (m_k - 1))$, when $p > 2$, and when there is only one two-way interaction between the variables.

This result can be shown as follows:

Let us consider equation (4), and suppose, without loss of generality, that X_1 and X_2 are dependant. So, upon multiplication by A_3: $\frac{1}{p} \sum_{i=1}^{p} A_i z = \mu z$ becomes $\frac{1}{p}(A_3 A_1 + A_3 A_2 + A_3 A_3 + \cdots + A_3 A_P)\, z = \mu\, A_3\, z,$ and we get $\mu = \frac{1}{p}$.

Now multiply equation (4) by A_2 and A_1 in turn to get:

$$\begin{cases} \Big(A_1A_1 + A_1A_2 + A_1A_3 + \cdots + A_1A_P\Big)\, z \;=\; p\,\mu A_1\, z \\ \Big(A_2A_1 + A_2A_2 + A_2A_3 + \cdots + A_2A_P\Big)\, z \;=\; p\,\mu A_2\, z \end{cases} \iff$$

$$\iff \quad \begin{cases} (A_1 + A_1A_2)\, z \;=\; p\,\mu\, A_1\, z \\ (A_2A_1 + A_2)\, z \;=\; p\,\mu\, A_2\, z \end{cases}$$

$$\iff \quad \begin{cases} A_1A_2\, b \;=\; \lambda\, z \\ A_2A_1\, b \;=\; \lambda\, z \end{cases}$$

where $\lambda = (p\,\mu - 1)^2$, $a = A_1\, z$ and $b = A_2\, z$.

We recognize here the CA equations, so that the CA of Burt tables, when only two variables are dependent is equivalent to the CA of the contingency tables crossing the two dependent variables. It is well known that the number of eigenvalue in CA equals $q - 2\min((m_j-1);(m_k-1))$, and for all non trivial λ_i, there corresponds the values μ_i and μ_i' such that:

$$\mu_i = \frac{1+\sqrt{\lambda_i}}{p} \qquad \text{and} \qquad \mu_i' = \frac{1-\sqrt{\lambda_i}}{p}\;.$$

Finally, the CA of the Burt table may have $2\min((m_j-1);(m_k-1))$ eigenvalues non trivial and not equal to $\frac{1}{p}$, associated to the CA of the contingency table. So the number of supplementary eigenvalues equals $q - 2\min((m_j-1);(m_k-1))$.

There is, in addition, one n_1 multiple eigenvalue, where n_1 is at least $q - 2\min((m_j-1);(m_k-1))$.

4.2.1.2. The case of higher order interactions

Since the Burt table is constructed with pairwise cross products of variables, its observation cannot give us information about multiway interactions.

However the observation of the bi-dimensional theoretical Burt sub-tables, for all pairwise variable combinations, can provide us with all the two-way interactions.

The theoretical Burt table can reveal the existence of higher order interactions in the following case:

If $B_{ij} \neq B_{ii}\, 1_{m_j m_j} B_{jj}$ and $B_{ik} \neq B_{ii}\, 1_{m_k m_k} B_{kk}$: there may be a triple interaction between X_i, X_j and X_k.

In general, a Burt table doesn't give either the order of the interactions, or supplementary information on the eigenvalue behavior.

4.2.2. Distribution of observed eigenvalues

Exceptionally, with a small number of interactions, we observe the particular shape of the eigenvalue diagram exhibited in *Figure 4*, where we can distinguish eigenvalues near $\frac{1}{p}$ (theoretically equal to $\frac{1}{p}$), and so we are able to recognize the existence of the independent variables in the analysis.

λ_I	Eigenvalues diagram
λ_1	****************************
λ_2	***************************
λ_3	***********************
λ_4	**********************
λ_5	*********************
$\vdots$	**************
$\vdots$	*************
λ_q	************

Figure 4: Observed eigenvalues diagram in a two-way interaction case.

When the number of interaction grows, we cannot distinguish eigenvalues theoretically equal to $\frac{1}{p}$ from the eigenvalues non equal to $\frac{1}{p}$.

To detect the existence or interactions, we can first check if the observed variables are mutually independent. In that case, the eigenvalues distribution diagram should have a particular shape (see § 4.1.), with more than 95% of the eigenvalues within the confidence interval $\frac{1}{p} \pm 2\,\sigma$ (see § 4.1.1).

If there is one or more eigenvalues out of the confidence interval, we can then assume the existence of one or more two-way interaction between variables.

5. AN EMPIRICAL PROCEDURE FOR FITTING LOG-LINEAR MODELS BASED ON THE MCA EIGENVALUE DIAGRAM

We propose an empirical procedure for progressively fitting a log-linear model where the fitting test at each step is based on the MCA eigenvalues diagram.

Let X_i, X_j and X_k, three categorical variables, with respectively m_i, m_j and m_k modalities, and let a cross variable with $(m_i \times m_j)$ modalities. We suppose that X_{ij} and X_k, have the same behavior if $m_k = m_i \times m_j$.

Under the hypothesis that two dependant variables X_i and X_j have the same behaviour as the variable X_k with the same characteristics of the cross variable X_{ij}, we propose here an empirical procedure for fitting progressively, with p steps, the log-linear model where the fitting criterion at each step is based on the MCA eigenvalue diagram. Distribution of observed eigenvalues

5.1. Description of the procedure steps

The first step of the procedure consist to test the pairwise independence hypothesis of the variables. To detect existence of interactions, we must first check if all variables are mutually independent. For that matter, we calculate the eigenvalues of MCA on all the p variables, and construct the related confidence interval: the eigenvalue distribution diagram should have a particular shape (cf. § 4.1.). If all the eigenvalues belong to the confidence interval $\frac{1}{p} \pm 2\,\sigma$ (cf. § 4.1.1), we can conclude that the p variables are mutually independent. The log-linear model associated to the variables is a simple additive one:

$$\log[f_p(X)] \;=\; u_0(x) + \sum_{i=1}^{p} u_i(x) \;,$$

and the procedure is stopped.

If one or more eigenvalue are not in the confidence interval, we conclude that there is at least one double interaction between variables, and we go to the second step of the procedure.

In the second step, we look at all two-way interaction u-terms. We isolate one variable amongst the p variables that we note X_p, without loss of generality, and so we obtain a set of $(p-1)$ variables X_i, and apply the first step to test pairwise independence of the $(p-1)$ variables.

If the $(p-1)$ variables are independent, we can conclude that the doubles interactions are amongst X_p and at least one of the X_i, so we construct correspondent cross variables X_{ip} by using the first step to test independence between variables (X_i, X_p) where $i = 1, ..., p-1$. The log-linear model associated to the variables is:

$$\log[f_p(X)] \;=\; u_0(x) + \sum_{i=1}^{p} u_i(x) + \sum_{i=1}^{p-1} u_{ip}(x)\,\delta_{ip} \;,$$

and the procedure stopped, (with $\delta_{ip} = 1$ if the interaction between X_p and X_i exists, otherwise it is set to zero.)

If the $(p-1)$ variables are not independent, we can conclude that there is double interaction between X_i and X_j where $i, j = 1, ..., p$–1, and perhaps between X_i and X_p.

We can construct correspondent cross variables X_{ip} and X_{ij} by using the first step to test independence of variables (X_i, X_p) and variables (X_i, X_j) where $i, j = 1, ..., p-1$. The log-linear model associated to the variables is:

$$\log[f_p(X)] = u_0(x) + \sum_{i=1}^{p} u_i(x) + \sum_{i=1}^{p-1} u_{ip}(x)\, \delta_{ip} + \text{terms due to the interaction between three or more variables}$$

and we go to the third step of the procedure

In the third step, we look at three-way interaction u-terms, by testing the dependence of variables X_i and cross variables X_{jk}, where $i, j, k = 1, ..., p$ and i, j, k are different, and construct cross variables X_{ijk}. The independence test is based on the eigenvalue pattern of the related MCA as described in the first step.

Continuing this way, in the k^{th} step, we look at k-way interaction u-terms, ... and in the least step we look at the p-way interaction u-term.

This algorithm is summarized in *Figure 5.*

5.2. An example for a graphical model

For this example we use a data set given by Haberman [24] that was used in Falguerolles *et al.* [14] to fit a graphical model. The data reports attitudes toward non therapeutic abortions among white subjects crossed with three categorical variables describing the subjects.

The data set is a contingency table observed for 3181 individuals, crossing four three modality variables X_1, X_2, X_3 and X_4, defined in *Table 1.*

The first step of the procedure consists of testing the pairwise independence hypothesis of the variables. We first transform the contingency table in a complete disjunctive table, then calculate the parameters (defined in § 2.1 and § 4.1.1) needed for the test (*Table* 2).

MCA on the four variables gives the eigenvalues diagram of *Figure 6.*

The shape of eigenvalues diagram refers clearly to the existence of dependent variables.

Eigenvalues λ_1, λ_7 and λ_8 are not in the interval I_c, so there is at least two dependent variables: there is one or more two-way interactions between variables.

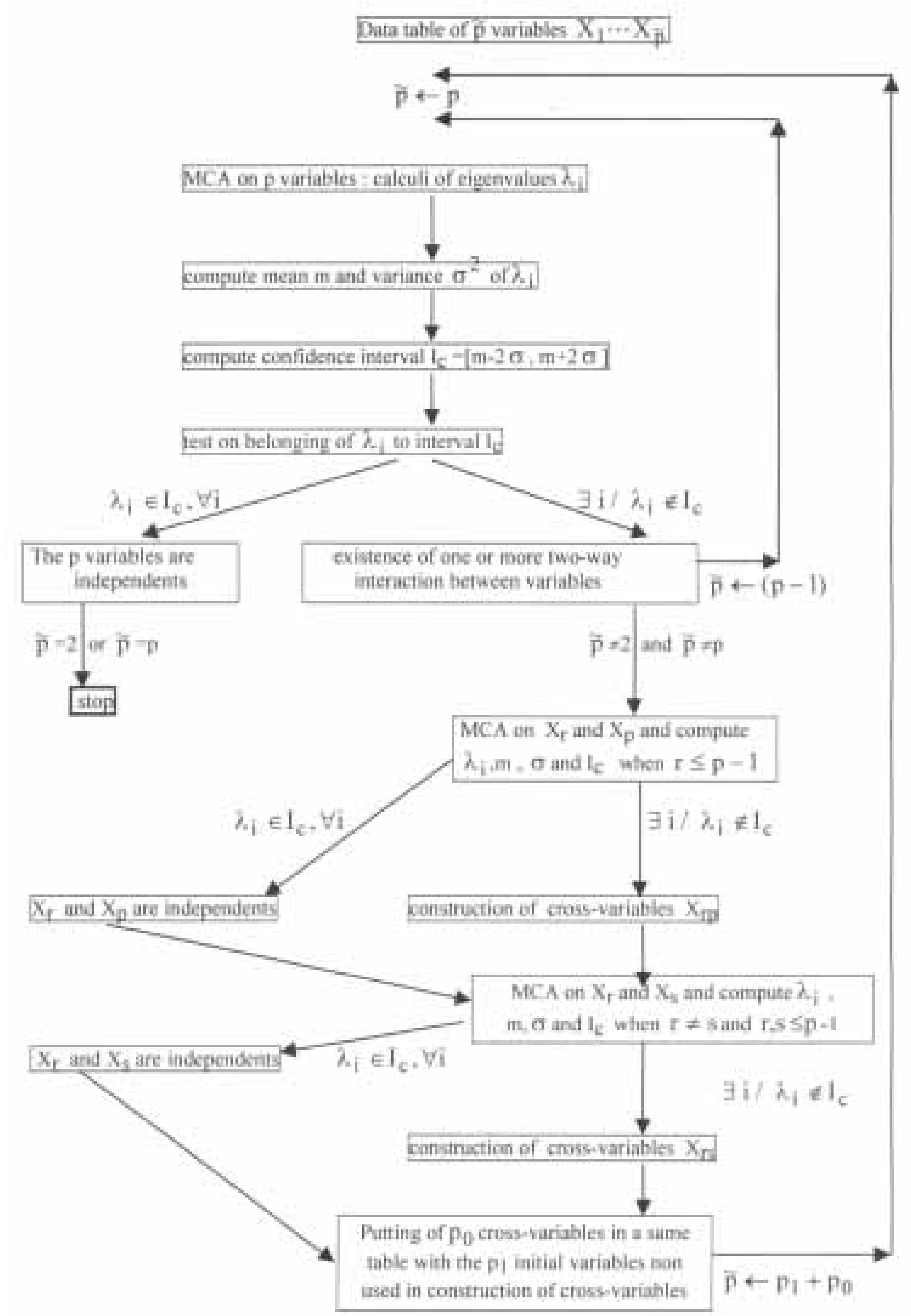


Figure 5: Block diagram for the Empirical procedure.

Table 1: Attitudes toward non therapeutic abortions among white.

Year X_1	Religion: X_2	Education in years: X_3	Attitude: X_4 positive	mixed	negative
1972	northern Protestant	≤ 8	09	16	41
		9–12	85	52	105
		≥ 13	77	30	38
	southern Protestant	≤ 8	08	08	46
		9–12	35	29	54
		≥ 13	37	15	22
	Catholic	≤ 8	11	14	38
		9–12	47	35	115
		≥ 13	25	12	42
1973	northern Protestant	≤ 8	17	17	42
		9–12	102	38	84
		≥ 13	88	15	31
	southern Protestant	≤ 8	14	11	34
		9–12	61	30	59
		≥ 13	49	11	19
	Catholic	≤ 8	06	16	26
		9–12	60	29	108
		≥ 13	31	18	50
1974	northern Protestant	≤ 8	23	13	32
		9–12	106	50	88
		≥ 13	79	21	31
	southern Protestant	≤ 8	05	15	37
		9–12	38	39	54
		≥ 13	52	12	32
	Catholic	≤ 8	08	10	24
		9–12	65	39	89
		≥ 13	37	18	43

Table 2: Parameters needed for the test
(first step of the example for a graphical model).

n	p	m_1	m_2	m_3	m_4	q	m	σ	I_c
3181	4	3	3	3	3	8	0.25	0.0109	[0.2283, 0.2717]

$\lambda_1 = 0.3221$	**************************
$\lambda_2 = 0.2704$	*********************
$\lambda_3 = 0.2599$	********************
$\lambda_4 = 0.2531$	*******************
$\lambda_5 = 0.2451$	******************
$\lambda_6 = 0.2393$	*****************
$\lambda_7 = 0.2277$	****************
$\lambda_8 = 0.1823$	***********

Figure 6: Eigenvalues diagram
(first step of the example for a graphical model).

The second step consists of the detection of two-way interactions. In a first time, we use our first step with only three variables X_1, X_2 and X_3.

With the values of n and m_i (for $i = 1, ..., 3$) still the same, the other parameters become (*Table 3*):

Table 3: Parameters for the test
(second step of the example for a graphical model).

q	m	σ	I_c
6	0.33333	0.0118	[0.3097, 0.3569]

We get the following eigenvalue diagram (*Figure 7*):

$\lambda_1 = 0.3606$	**************************
$\lambda_2 = 0.3448$	*************************
$\lambda_3 = 0.3385$	************************
$\lambda_4 = 0.3305$	**********************
$\lambda_5 = 0.3025$	*********************

Figure 7: Eigenvalues diagram
(second step of the example for a graphical model).

λ_1 and λ_5 are not in interval I_c, so there is one or more two-way interaction between X_1, X_2 and X_3, as also as interactions between X_4 and others.

In a second step we look at the interactions between X_4 and X_i ($i = 1, 2, 3$).

For $i = 1$ to $i = 3$ we look at the eigenvalues of the MCA of X_4 with X_i, and calculate their variances and intervals I_c.

Crossing X_1 with X_4 we get (*Table 4*):

Table 4: MCA on X_1 and X_4 (parameters and eigenvalues).

q	m	σ	I_c	λ_1	λ_2	λ_3	λ_4
4	0.5	0.0125	[0.4750, 0.5250]	0.5389	0.5156	0.4644	0.4611

Crossing X_2 with X_4 we get (*Table 5*):

Table 5: MCA on X_2 and X_4 (parameters and eigenvalues).

q	m	σ	I_c	λ_1	λ_2	λ_3	λ_4
4	0.5	0.0125	[0.4750, 0.5250]	0.5741	0.5076	0.4924	0.4259

Crossing X_3 with X_4 we get (*Table 6*):

Table 6: MCA on X_3 and X_4 (parameters and eigenvalues).

q	m	σ	I_c	λ_1	λ_2	λ_3	λ_4
4	0.5	0.0125	[0.4750, 0.5250]	0.6112	0.5041	0.4959	0.3979

In the three cases, λ_1 and λ_4 are not in the intervals I_c, so there is a two-way interaction between X_1 and X_4, X_2 and X_4 and between X_3 and X_4, so we can construct cross variables X_{4i} having 9 modalities $(i = 1, 2, 3)$.

Now, we use the first step with only two variables X_1 and X_2, after we look for interactions between X_3 and X_i $(i = 1, 2)$.

Crossing X_1 with X_2 we get (*Table 7*):

Table 7: MCA on X_1 and X_2 (parameters and eigenvalues).

q	m	σ	I_c	λ_1	λ_2	λ_3	λ_4
4	0.5	0.0125	[0.4750, 0.5250]	0.5153	0.5045	0.4955	0.4848

All the eigenvalues are in the confidence interval, so X_1 and X_2 are independent conditionally on the other, and there is no cross variable X_{12}. The corresponding u-term u_{12} equals to zero.

Let us now look, when $i = 1$ and $i = 2$, at the eigenvalues of the MCA of X_3 with X_i, with their variances and intervals I_c:

Crossing X_1 with X_3 we get (*Table 8*):

Table 8: MCA on X_1 and X_3 (parameters and eigenvalues).

q	m	σ	I_c	λ_1	λ_2	λ_3	λ_4
4	0.5	0.0125	[0.4750, 0.5250]	0.5134	0.5023	0.4978	0.4866

All the eigenvalues are in the confidence interval I_c, so X_1 and X_3 are independent conditionally on the other, and there is no cross variable X_{13}: the corresponding u-term u_{13} equals to zero.

Crossing now X_2 with X_3 we get (*Table 9*):

Table 9: MCA on X_2 and X_3 (parameters and eigenvalues).

q	m	σ	I_c	λ_1	λ_2	λ_3	λ_4
4	0.5	0.0125	[0.4750, 0.5250]	0.5401	0.5128	0.4872	0.4599

Here, λ_1 and λ_4 are not in the interval I_c, so there is a two-way interaction between X_2 and X_3, u_{23} is not set to zero, and we can add the cross variable X_{32} (as well as X_{23}) with 9 modalities to the model.

The third step consists of the detection of triple interactions between variables, that is to two-way interactions between the variables X_i and the cross variables X_{jk}.

We first put the cross variables (X_{41}, X_{42}, X_{43}, X_{32}) with the initial variables that were deemed non dependent in the second step of the procedure, i.e. X_1 and X_2, and then we use the first step of the procedure with the set of obtained variables.

So we get the following results (*Table 10* and *Figure 8*):

Table 10: MCA on X_1, X_2, X_{41}, X_{42}, X_{43} and X_{32}
(parameters third step of the example for a graphical model).

q	m	σ	I_c
36	0.1667	0.0168	[0.1331, 0.2003]

$\lambda_1 = 0.5201$	**************************
$\lambda_2 = 0.5006$	*************************
$\lambda_3 = 0.3447$	******************
$\lambda_4 = 0.3347$	******************
$\lambda_5 = 0.3303$	******************
$\lambda_6 = 0.3193$	*****************
$\lambda_7 = 0.1810$	************
$\lambda_8 = 0.1796$	***********
$\lambda_9 = 0.1732$	***********
$\lambda_{10} = 0.1710$	***********
$\lambda_{11} = 0.1664$	***********
$\lambda_{12} = 0.1627$	***********
$\lambda_{13} = 0.1626$	***********
$\lambda_{14} = 0.1578$	***********
$\lambda_{15} = 0.1538$	**********
$\lambda_{16} = 0.1423$	*********

Figure 8: MCA on X_1, X_2, X_{41}, X_{42}, X_{43} and X_{32}
(eigenvalues diagram, third step of the example for a graphical model).

The first six eigenvalues are not in I_c: there is one or more two-way interaction between the initial variables X_i, and the crossed ones X_{ik}, so there exists a triple interaction between simple variables.

We drop X_{32} and use the first step with the five other variables to get the following results (*Table 11* and *Figure 9*):

Table 11: MCA on X_1, X_2, X_{41}, X_{42} and X_{43} (parameters for the test).

q	m	σ	I_c
28	0.2	0.0162	[0.1671, 0.2324]

$\lambda_1 = 0.6105$	**************************
$\lambda_2 = 0.6006$	**************************
$\lambda_3 = 0.4143$	****************
$\lambda_4 = 0.4028$	****************
$\lambda_5 = 0.3982$	****************
$\lambda_6 = 0.3831$	***************
$\lambda_7 = 0.2262$	**********
$\lambda_8 = 0.2220$	**********
$\lambda_9 = 0.2162$	**********
$\lambda_{10} = 0.2083$	*********
$\lambda_{11} = 0.2054$	*********
$\lambda_{12} = 0.2017$	*********
$\lambda_{13} = 0.1952$	*********
$\lambda_{14} = 0.1986$	*********
$\lambda_{15} = 0.1952$	*********
$\lambda_{16} = 0.1928$	*********
$\lambda_{17} = 0.1878$	********
$\lambda_{18} = 0.1837$	********
$\lambda_{19} = 0.1815$	********
$\lambda_{20} = 0.1711$	********

Figure 9: MCA on X_1, X_2, X_{41}, X_{42} and X_{43} (eigenvalues diagram, third step of the example for a graphical model).

The first six eigenvalues are not in I_c, so there is at least one two-way interaction between the variables. We know that simple variables X_1, X_2 and the crossed variables X_{41}, X_{42}, X_{43} are dependent so we have to test dependence between X_1 and X_{32} only. Crossing X_1 and X_{32} we get the following results (*Table 12)*:

Table 12: MCA on X_1 and X_{32} (parameters and eigenvalues).

q	m	σ	I_c
10	0.5	0.0159	[0.4682, 0.5318]

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
0.5287	0.5194	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.4806	0.4713

All the eigenvalues are in the confidence interval I_c, so X_1 and X_{32} are independent conditionally on the other, and there is no cross variable X_{132}. The corresponding u-term u_{123} equals zero.

Now we can drop the cross variable X_{43}. The remaining variables X_1, X_2, X_{41}, X_{42} are dependent by construction. We have only to test for dependence between X_1 and X_{43}.

Crossing X_1 with X_{43}, we get the same parameter as the crossing of X_1 and X_{32}, and the following eigenvalues (*Table 13*):

Table 13: MCA on X_1 and X_{43} (eigenvalues).

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
0.5445	0.5232	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.4768	0.4555

We remark that λ_1 and λ_{10} are not in the interval I_c, so X_1 and X_{43} seem to be dependent. But we have to fit a graphical model, that is a particular case of hierarchical models (as defined in § 2.2.2.2, a log-linear models is hierarchical if, whenever one particular u-term is constrained to zero then all higher u-terms containing the same set of subscripts are also set to zero).

Here the u-term u_{13} is set to zero, so the u-term u_{134} is also set to zero.

Crossing X_2 with X_{43}, we get the same parameter as the crossing of X_1 and X_{32}, and the following eigenvalues (*Table 14*):

Table 14: MCA on X_2 and X_{43} (eigenvalues).

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
0.5871	0.5466	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.4534	0.4143

Eigenvalues λ_1, λ_2, λ_9 and λ_{10} are not in the interval I_c, the u-terms u_{23} and u_{24} are not set to zero, and since X_2 and X_{43} are not dependent the u-term u_{234} is not set to zero.

Crossing X_1 with X_{42} (or equivalently X_2 with X_{41}) we get the same parameter as the crossing of X_1 and X_{32}, and the following eigenvalues:

Table 15: MCA on X_1 and X_{42} (eigenvalues).

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
0.5434	0.5289	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.4711	0.4566

Eigenvalues λ_1 and λ_{10} are not in the interval I_c, the u-term u_{14} is equal to zero, X_1 and X_{42} are dependent, and the u-term u_{124} is set to zero.

Finally, variables X_1 and X_{41} are dependent by construction.

The procedure stops here because we can't have more than triple interactions in a hierarchical model when all the two-way interactions are not present. We obtain the following model (see *Figure 10* for the associated graph):

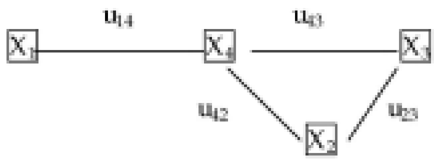


Figure 10: Lattice diagram (example for a graphical model).

$$\begin{aligned}\log[f_4(X)] \;=\;& u_0 + u_1 x_1 + u_2\, x_2 + u_3\, x_3 + u_4\, x_4 + u_{32}\, x_2 x_3 + u_{41} x_4 x_1 + u_{42}\, x_4 x_2 \\ & + u_{43}\, x_4 x_3 + u_{432}\, x_4 x_3 x_2\ .\end{aligned}$$

5.3. An example for a saturated model

Here we use a data set given by Israëls [29] that was also used by Van der Heijden et al. [28] about 'shop-lifting' habits.

Table 16 is a contingency table crossing three variables: sex (2 modalities), age (9 modalities) and type of goods (13 modalities: Clothing (C), Clothing accessories (Ca), Provision-Tobacco (PT), Writing materials (Wm), Books (B), Records (R), Household goods (Hg), Sweets (S), Toys (T), Jewellery (J), Perfume (P), Hobbies tools(Ht), and Others(O)) observed over 33 101 individuals.

In the first step, we test the pairwise independence of variables X_1, X_2 and X_3. We first transform the contingency table in a complete disjunctive table, then compute the parameters (defined in § 2.2 & § 4.1.1) needed for the test to get (*Table 17*).

A MCA on the three variables gives the eigenvalue diagram of *Figure 11*.

The eigenvalue diagram shows clearly that the variables are not independent: only 8 eigenvalues $(\lambda_7, ..., \lambda_{15})$ are in the confidence interval.

Using the second step of the procedure, we get the two-way interactions.

Table 16: Multicontingency table for the shop-lifting data.

Sex: X_1	Age: X_2	Goods: X_3 C	Ca	PT	Wm	B	R	Hg	S	T	J	P	Ht	O
Male	≤ 11	81	66	150	667	67	24	47	430	743	132	32	197	209
	12–14	138	204	340	1409	259	272	117	637	684	408	57	547	550
	15–17	304	193	229	527	258	368	98	246	116	298	61	402	454
	18–20	384	149	151	84	146	141	61	40	13	71	52	138	252
	21–29	942	297	313	92	251	167	193	30	16	130	111	280	624
	30–39	359	109	136	36	96	67	75	11	16	31	54	200	195
	40–49	178	53	121	36	48	29	50	5	6	14	41	152	88
	50–64	137	68	171	37	56	27	55	17	3	11	50	211	90
	≥ 65	45	28	145	17	41	7	29	28	8	10	28	111	34
Female	≤ 11	71	19	59	224	19	7	22	137	113	162	70	15	24
	12–14	241	98	111	463	60	32	29	240	98	138	178	29	58
	15–17	477	114	58	91	50	27	41	80	14	548	141	9	72
	18–20	436	108	76	18	32	12	32	12	10	303	70	14	67
	21–29	1180	207	132	30	61	21	65	16	12	74	104	30	157
	30–39	1009	165	121	27	43	9	74	14	31	100	81	36	107
	40–49	517	102	93	23	31	7	51	10	8	48	46	24	66
	50–64	488	127	214	27	57	13	79	23	17	22	69	35	64
	≥ 65	173	64	215	13	44	0	39	42	6	12	41	11	55

Table 17: Parameters needed for the test
(first step of the example for a satured model).

n	p	m_1	m_2	m_3	q	m	σ	I_c
33101	3	2	9	13	21	0.3333	0.0061	[0.3211, 0.3455]

$\lambda_1 = 0.5759$	***
$\lambda_2 = 0.4256$	***********************************
$\lambda_3 = 0.3966$	********************************
$\lambda_4 = 0.3899$	*******************************
$\lambda_5 = 0.3542$	****************************
$\lambda_6 = 0.3494$	****************************
$\lambda_7 = 0.3407$	***************************
$\lambda_8 = 0.3384$	**************************
$\lambda_9 = 0.3344$	**********************
$\lambda_{10} = 0.3333$	**********************
$\lambda_{11} = 0.3333$	**********************
$\lambda_{12} = 0.3333$	**********************
$\lambda_{13} = 0.3322$	*********************
$\lambda_{14} = 0.3271$	*********************
$\lambda_{15} = 0.3260$	*********************
$\lambda_{16} = 0.3177$	********************
$\lambda_{17} = 0.3103$	*******************
$\lambda_{18} = 0.2802$	******************
$\lambda_{19} = 0.2632$	****************
$\lambda_{20} = 0.1925$	************
$\lambda_{21} = 0.1423$	*******

Figure 11: MCA on X_1, X_2 and X_3
(eigenvalues diagram, third step of the example for a saturated model).

MCA of X_1 and X_3 gives the following results (*Table 18* and *Figure 12*):

Table 18: MCA on X_1 and X_3
(parameters).

n	p	q	m	σ	I_c
33101	2	13	0.5	0.00002	[0.5000, 0.5000]

$\lambda_1 = 0.7032$	**
$\lambda_2 = 0.5000$	*************************
$\lambda_3 = 0.5000$	*************************
$\lambda_4 = 0.5000$	*************************
$\lambda_5 = 0.5000$	*************************
$\lambda_6 = 0.5000$	*************************
$\lambda_7 = 0.5000$	*************************
$\lambda_8 = 0.5000$	*************************
$\lambda_9 = 0.5000$	**********************
$\lambda_{10} = 0.5000$	**********************
$\lambda_{11} = 0.5000$	**********************
$\lambda_{12} = 0.5000$	**********************
$\lambda_{13} = 0.2968$	**********

Figure 12: MCA on X_1 and X_3
(eigenvalues diagram, second step of the example for a saturated model).

The first and the last eigenvalues are not in the confidence interval so the u-term u_{13} is not set to zero.

We notice here the peculiar form of the eigenvalues diagram, due to the fact that multiple eigenvalue $\lambda = \frac{1}{2}$ that have a multiplicity $11 = m_3 - m_1$ is an artificial one (cf. § 4.2.1.1).

MCA of X_2 and X_3 gives the following results (*Table 19* and *Figure 13*):

Table 19: MCA on X_2 and X_3
(parameters).

n	p	q	m	σ	I_c
33101	2	20	0.5	0.0001	[0.4998, 0.5002]

The 8 first and the 8 last eigenvalues are not in the confidence interval so the u-term u_{23} is not set to zero.

$\lambda_1 = 0.7852$	**
$\lambda_2 = 0.6074$	*******************************
$\lambda_3 = 0.5903$	*****************************
$\lambda_4 = 0.5346$	**************************
$\lambda_5 = 0.5245$	*************************
$\lambda_6 = 0.5112$	************************
$\lambda_7 = 0.5109$	************************
$\lambda_8 = 0.5019$	***********************
$\lambda_9 = 0.5000$	********************
$\lambda_{10} = 0.5000$	********************
$\lambda_{11} = 0.5000$	********************
$\lambda_{12} = 0.5000$	********************
$\lambda_{13} = 0.4981$	*******************
$\lambda_{14} = 0.4891$	******************
$\lambda_{15} = 0.4888$	******************
$\lambda_{16} = 0.4755$	*****************
$\lambda_{17} = 0.4654$	****************
$\lambda_{18} = 0.4097$	************
$\lambda_{19} = 0.3926$	***********
$\lambda_{20} = 0.2148$	******

Figure 13: MCA on X_2 and X_3
(eigenvalues diagram, second step of the example for a saturated model).

MCA of X_1 and X_2 gives the following eigenvalue results (*Table 20, Figure 14*):

Table 20: MCA on X_1 and X_2
(parameters).

n	p	q	m	σ	I_c
33101	2	9	0.5	0.0037	[0.4926, 0.5074]

$\lambda_1 = 0.6241$	**
$\lambda_2 = 0.5000$	*************************
$\lambda_3 = 0.5000$	*************************
$\lambda_4 = 0.5000$	*************************
$\lambda_5 = 0.5000$	*************************
$\lambda_6 = 0.5000$	*************************
$\lambda_7 = 0.5000$	*************************
$\lambda_8 = 0.5000$	*************************
$\lambda_9 = 0.3759$	**********

Figure 14: MCA on X_1 and X_2
(eigenvalues diagram, second step of the example for a saturated model).

The first and the last eigenvalues are not in the confidence interval so the u-term u_{12} is not set to zero. At the end of the second step, we obtain all three two-way interactions. To know if the model is a saturated one we can built one of the crossed variables and test its dependence with the remaining simple variable.

MCA of X_{32} with X_1 gives the following eigenvalues:

$$\lambda_1 = 0.7285 \,, \quad \lambda_2 = \lambda_3 = \cdots = \lambda_{116} = 0.5 \,, \quad \lambda_{117} = 0.2715 \quad \text{and} \quad I_c = [0.4615,\, 0.5384] \,.$$

The first and the last eigenvalues are not in the confidence interval so the u-term u_{123} is not set to zero.

At the end we get the following saturated model:

$$\log[f_3(X)] = u_0 + u_1 x_1 + u_2\, x_2 + u_3\, x_3 + u_{12}\, x_1 x_2 + u_{23}\, x_2 x_3 + u_{13}\, x_1 x_3 + u_{123}\, x_1 x_2 x_3 \,.$$

5.4. An example for a mutual independence model

Here we use a data set given by Andersen [2] as a contingency table crossing four variables observed over 299 individuals corresponding to a retrospective study of ovary cancer, defined in Table 21:

Table 21: Retrospective study of ovary cancer.

X_1 stage	X_2 operation	X_3 survival	X_4 X-ray	
			No	Yes
Early	radical	no	10	17
	limited	yes	41	64
		no	1	3
		yes	13	9
Advanced	radical	no	38	64
	limited	yes	6	11
		no	3	13
		yes	1	5

In the first step of procedure, we test for the pairwise independence of variables X_1, X_2, X_3 and X_4. We first transform the contingency table in a complete disjunctive table, then compute the parameters (see § 4.1.1) needed for the test.

The MCA on the four variables gives the following results (*Table 22* and *Figure 15*):

Table 22: Parameters needed for the test
(first step of the example for a mutual independence model).

n	p	m_1	m_2	m_3	m_4	q	m	σ	I_c
299	4	2	2	2	2	4	0.25	0.0250	[0.2000, 0.3000]

$\lambda_1 = 0.4145$	**********************************
$\lambda_2 = 0.2512$	********************
$\lambda_3 = 0.2449$	*******************
$\lambda_4 = 0.0894$	*********

Figure 15: MCA on X_1, X_2, X_3 and X_4
(eigenvalues diagram, first step of the example for a mutual independence model).

The eigenvalue diagram shows clearly that variables are not independent, only λ_2 and λ_3 are in the confidence interval.

Let's drop X_4 and use the second step of the procedure. MCA on the three remaining variables gives the following results (*Table 23* and *Figure 16*):

Table 23: MCA on X_1, X_2 and X_3
(parameters).

n	p	q	m	σ	I_c
299	3	3	0.3333	0.0273	[0.2787, 0.3879]

$\lambda_1 = 0.3639$	**********************
$\lambda_2 = 0.3342$	********************
$\lambda_3 = 0.3019$	*******************

Figure 16: MCA on X_1, X_2 and X_3
(eigenvalues diagram).

The eigenvalue diagram shows clearly that variables are independent, since all the eigenvalues are in the confidence interval, so there is surely one or more interaction X_4 and X_i, $i=1,\dots,3$.

The MCA on X_4 and X_i gives the following results (*Table 24* and *Figure 17*):

Table 24: MCA on X_4, X_i
(parameters).

n	p	q	m	σ	I_c
299	2	2	0.5	0.0283	[0.4434, 0.5566]

X_4 and X_1	X_4 and X_2	X_4 and X_3
$\lambda_1 = 0.5365$ **********	$\lambda_1 = 0.8198$ **********	$\lambda_1 = 0.5058$ **********
$\lambda_2 = 0.4635$ *********	$\lambda_2 = 0.1802$ ***	$\lambda_2 = 0.4942$ *********

Figure 17: Eigenvalues diagram for MCA on X_4 and X_1,
MCA on X_4 and X_2 and MCA on X_4 and X_3.

It's clear that there exists only an interaction between X_4 and X_2, X_1 and X_3 are non dependent of X_4, then $u_{14} = u_{13} = 0$ and $u_{24} \neq 0$ and we build the crossed variable X_{24}.

The MCA of X_1, X_3 and X_{24} gives the following results (*Table 25* and *Figure 18*):

Table 25: MCA on X_1, X_3 and X_{24}
(parameters).

n	p	q	m	σ	I_c
299	3	5	0.3333	0.0273	[0.2787, 0.3879]

$\lambda_1 = 0.3647$	**********************
$\lambda_2 = 0.3624$	**********************
$\lambda_3 = 0.3333$	*********************
$\lambda_4 = 0.3047$	********************
λ_5=0.3016	********************

Figure 18: Eigenvalues diagram for MCA on X_1, X_3 and X_{24}.

The eigenvalue diagram shows that the variables are independent, all the eigenvalues being within the confidence interval, and there is no triple interaction between variables.

We finally obtain the same model as Andersen:

$$\log[f_4(X)] \;=\; u_0 + u_1 x_1 + u_2\, x_2 + u_3\, x_3 + u_4\, x_4 + x_{24} x_4 x_2 \;.$$

6. CONCLUSION

Log-linear modeling and MCA are two complementary techniques for the analysis of categorical data. In this framework, we propose a method for fitting progressively log-linear models, using the eigenvalue shape of MCA.

We show that, in MCA, under the independence hypothesis for the variables, each observed eigenvalue is asymptotically normally distributed. These distributions have the same mean, different variances and converge to normal distributions. In this case, the eigenvalue diagram takes a peculiar shape. This shape is different if there is one or more interactions between variables, and we can recognize the log-linear model fitted for the data in some special cases.

Then, based on these results, we propose a simple procedure for progressively fitting log-linear models, where the fitting criterion is based on MCA eigenvalue diagrams: the chosen model is constructed by successive utilizations of MCA (non constrained by the variables number). Finally, we validate this procedure on three sets of data drawn from the literature.

REFERENCES

[1] Agresti, A. (1990). *Categorical Data Analysis*, Wiley–Interscience.

[2] Andersen, E.B. (1991). *The Statistical Analysis of Categorical Data* (Second edition), Springer-Verlag.

[3] Baccini, A.; Mathieu, J.R. and Mondot, A.M. (1987). Comparaison sur un exemple, d'analyse des correspondances multiples et de modélisations, *Revue de Statistique Appliquée*, **XXXV**(3), 21–34.

[4] Ben Ammou, S. (1996). *Comportement des valeurs propres en analyse des correspondances multiples sous certaines hypothèses de modéles*, Doctoral Thesis of University Paris IX, Dauphine.

[5] Ben Ammou, S. and Saporta, G. (1998). Sur la normalité asymptotique des valeurs propres en ACM sous l'hypothèses d'indépendance des variables, *Revue de Statistique Appliquée*, **XLVI**(3), 21–35.

[6] Benzecri, J.P. (1973). *Analyse des Données [The Analysis of Data]* (2 vol), Paris: Dunod.

[7] Birch, M.W. (1963). Maximum likelihood in three-way contingency tables, *J. Royal Satist. Soc. (B)*, **25**, 220–233.

[8] Bishop, Y.M.M.; Fienberg, S.E. and Holland, P.W. with the collaboration of Light, R.J. and Mosteller, F. (1975). *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press.

[9] Burt, C. (1950). The factorial analysis of qualitative data, *British J. of Statist. Psychol.*, **3**(3), 166–185.

[10] Christensen, R. (1990). *Log-Linear Models*, Springer-Verlag, New York.

[11] Daudin, J.J. and Trecourt, P. (1980). Analyse factorielle des correspondances et modéle log-linéaire: comparaison des deux méthodes sur un exemple, *Revue de Statistique Appliquée*, **XXVIII**(1).

[12] Dobson, A. (1983). *An Introduction to Statistical Modelling*, Chapman and Hall, New York.

[13] de Falguerolles, A. and Jmel, S. (1993). Un modèle graphique pour la sélection de variables qualitatives, *Revue de Statistique Appliquée*, **XLI**(2), 23–41.

[14] de Falguerolles, A.; Jmel, S. and Whittaker, J. (1995). Correspondence analysis and association models constrained by a conditional independence graph, *Psychometrika*, **60**(2), 161–180.

[15] Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data*, MIT Press, Cambridge, Mass.

[16] Goodman, L.A. (1970). The multivariate analysis of qualitative data: interaction among multiple classifications, *J. of Amer. Statist. Assoc.*, **65**, 226–256.

[17] Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories, *Journal of the American Statistical Association*, **74**, 537–552.

[18] Goodman, L.A. (1981). Association models and the bivariate normal for contingency tables with ordered categories, *Biometrika*, **68**, 347–355.

[19] Goodman, L.A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories, *Journal of the American Statistical Association*, **76**, 320–334.

[20] Goodman, L.A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables (with comments), *International Statistical Review*, **54**(3), 243–309.

[21] Goodman, L.A. (1991). Measures, models and graphical display in the analysis of cross-classified data, *Journal of the American Statistical Association*, **86**, 1085–1110.

[22] Gifi, A. (1990). *Non Linear Multivariate Analysis*, J. Wiley.

[23] Guttman, L. (1941). *The quantification of a class of attributes: a theory and method of a scale construction.* In "The Prediction of Personal Adjustment" (P. Horst, Ed.), SSRC, New York, 251–264.

[24] Haberman, S.J. (1974). *The Analysis of Frequency Data*, University of Chicago, University Press, Chicago.

[25] Hayashi, C. (1956). Theory and examples of quantification (II), *Proc. of Institute of Statist. Math.*, **4**(2), 19–30.

[26] van der Heijden, P.G.M. and de Leeuw, J. (1985). Correspondence analysis used complementary to log-linear analysis, *Psychometrika*, **50**(4), 429–447.

[27] van der Heijden, P.G.M. and Worsley, K.J. (1986). Comment on "Correspondence analysis used complementary to log-linear Analysis", *Leiden Psychological Reports*, Psychometrics and Research Methodology, Department of Psychology, Leiden University – The Netherlands.

[28] VAN DER HEIJDEN, P.G.M.; DE FALGUEROLLES, A. and DE LEEUW, J. (1989). A combined approach to contingency table analysis using correspondence analysis and log-linear analysis (with discussion), *Applied Statitics*, **38**, 249–292.

[29] ISRAËLS, A. (1987). *Eigenvalue Techniques for Qualitative Data*, Leiden, The Netherlands, DSWO-Press.

[30] JMEL, S. (1991). Modeĺes graphiques, analyse en composantes principales et analyse des correspondances multiples: comparaisons sur des exemples. Poster présenté lors des *23èmes Journées de Statistique á Strasbourg*.

[31] LAURO, N.C. and DECARLI, A. (1982). Correspondence analysis and log-linear models in multi-way contingency tables. Some remarks on experimental data, *Rivista Internationale di Statistica Metron*, **XL**(1–2).

[32] LEBART, L. (1976). The significance of eigenvalues issued from correspondence analysis, *COMPSTAT, Physica Verlag, Vienne*, 38–45.

[33] DE LEEUW, J. (1984). *Canonical Analysis of Categorical Data* (Doctoral dissertation, University of Leiden, 1973), Leiden, DSWO-Press.

[34] O'NEILL, M.E. (1978). Asymptotic distributions of the canonical correlations from contingency tables, *Austral. J. Statist.*, **20**(1), 75–82.

[35] O'NEILL, M.E. (1978). Distributional expansion for canonical correlations from contingency tables, *J.R. Statist. Soc. B.*, **40**(3), 303–312.

[36] O'NEILL, M.E. (1980). A note on the canonical correlations from contingency tables, *Austral. J. Statist.*, **20**(1), 58–66.

[37] MALINVAUD, E. (1987). Data analysis in applied socio-economic statistics with special consideration of correspondence analysis, *Marketing Science Conference Proceedings, HEC-ISA, Joy en Josas.*

[38] NISHISATO, S. (1980). *Analysis of Categorical Data. Dual Scaling and Its Application*, Univ. of Toronto Press.

[39] NOVAK, T.P. and HOFFMAN, D.L. (1990). Residual scaling: an alternative to correspondence analysis for the graphical representation of residuals from log-linear models, *Multivariate Behavioral Research*, **25**, 351–370.

[40] SICILIANO, R. (1990). Asymptotic distribution of eigenvalues and statistical tests in non symmetric correspondence analysis, *Statistica Applicata*, **2**(3), 259–276.

[41] WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*, Wiley & Sons Ltd, England.

[42] WORLSLEY, K.J. (1987). Un exemple d'identification d'un modèle log-linéaire grâce à une analyse des correspondances (avec discussion), *Revue de Statistique Appliquée*, **XXXV**(3), 13–20.

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 23rd 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.
- For 2004 two volumes are scheduled for publication.
- 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 should be typed 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 to the e-mail: 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:

Adrião Ferreira da Cunha
Executive Editor, REVSTAT – STATISTICAL JOURNAL
Instituto Nacional de Estatística
Av. António José de Almeida 5
1000-043 LISBOA
PORTUGAL

Copyright and Reprints

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.