ISSN 1645-6726

INSTITUTO NACIONAL DE ESTATÍSTICA
PORTUGAL

REVSTAT
Statistical Journal

Volume 2, No.2
November 2004

Catalogação Recomendada

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

CREDITS

- **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**
  - *400 copies*
- **LEGAL DEPOSIT REGISTRATION**
  - *N.º 191915/03*

PRICE

[VAT 5% included]

- Single issue ..... **€ 12**
- Annual subscription ... **€ 19,20**

INDEX

REVSTAT – Statistical Journal
Volume 2, Number 2, November 2004

LOCAL FITTING WITH A POWER BASIS

Authors: Jochen Einbeck
– Department of Statistics, Ludwig Maximilians University Munich, Germany (einbeck@stat.uni-muenchen.de)

Received: April 2003 Revised: September 2004 Accepted: October 2004

Abstract:

- Local polynomial modelling can be seen as a local fit of the data against a polynomial basis. In this paper we extend this method to the power basis, i.e. a basis which consists of the powers of an arbitrary function. Using an extended Taylor theorem, we derive asymptotic expressions for bias and variance of this estimator. We apply this method to a simulated data set for various basis functions and discuss situations where the fit can be improved by using a suitable basis. Finally, some remarks about bandwidth selection are given and the method is applied to real data.

Key-Words:

- *local polynomial fitting; Taylor expansion; power basis; bias reduction.*

AMS Subject Classification:

- 62G08, 62G20.

1. INTRODUCTION

The roots of local polynomial modelling as understood today reach back to articles from Stone [19] and Cleveland [1]. A nice overview of the current state of the art is given in Fan & Gijbels [7]. The basic idea of this nonparametric smoothing technique is simply described. Consider bivariate data $(X_1, Y_1), ..., (X_n, Y_n)$, forming an i.i.d. sample from a population (X, Y). Assume the data to be generated from a model

$$Y = m(X) + \sigma(X)\,\varepsilon\ , \tag{1.1}$$

where $E(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = 1$, and X and ε are independent. Of interest is to estimate the regression function $m(x) = E(Y \mid X = x)$ and its derivatives $m'(x), m''(x), ..., m^{(p)}(x)$. A Taylor expansion yields

$$m(z) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}\,(z-x)^j \equiv \sum_{j=0}^{p} \beta_j(x)\,(z-x)^j\ , \tag{1.2}$$

given that the $(p+1)^{\text{th}}$ derivative of $m(\cdot)$ in a neighbourhood of x exists. We define $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$, where K is a kernel function which is usually taken to be a non-negative density symmetric about zero, and h denotes the bandwidth. The task of finding the appropriate bandwidth is the crucial point of local polynomial fitting; see Section 6 for more details. Minimizing

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j(x)\,(X_i - x)^j \right\}^2 K_h(X_i - x)$$

leads to the locally weighted least squares regression estimator $\hat\beta(x) = (\hat\beta_0(x), ..., \hat\beta_p(x))^T$ and the corresponding estimators

$$\hat m^{(j)}(x) = j!\,\hat\beta_j(x) \tag{1.3}$$

for $m^{(j)}(x)$, $j = 0, ..., p$. Alternative approaches focussed on estimating the conditional quantiles instead of the mean function (Yu & Jones [21], [22]), where a special case is nonparametric robust regression by local linear medians, applying an L_1 norm instead of an L_2 norm (Fan, Hu & Truong [8]).

Local polynomial modelling can be interpreted as fitting the data locally against the basis functions 1, $X - x$, $(X - x)^2$, ..., $(X - x)^p$. An obviously arising question is now: why should just these basis functions be the best possible ones? In a general framework one may use the basis functions $\phi_0(X)$, $\phi_1(X)$, ..., $\phi_p(X)$, with arbitrary functions $\phi_j : \mathbb{R} \mapsto \mathbb{R}$, $j = 0, ..., p$. However, theoretical results are only available under some restrictions on the basis functions. Regarding (1.2) and (1.3), it is seen that estimation and interpretation of parameters is based on Taylor's expansion. Furthermore, nearly all asymptotic results, e.g. the bias of the estimator, are based on Taylor's theorem. Asymptotics provide a very

important tool to find bandwidth selection rules etc., so they play an important role for the use of the estimator in practice.

Thus, if some theoretical background is desired, one needs to develop a new Taylor expansion for every basis one wants to use. Of course this will not be possible for all choices of basis functions. In the following section we focus on a special case, namely the power basis, where this is in fact possible and describe the estimation methodology. In Section 3 we provide some asymptotics for estimating the conditional bias and variance of this estimator, analyze the results, and show that the asymptotic bias may be reduced with a suitable choice of the basis. In Section 4 we apply this method to a simulated data set and compare the results for various basis functions. In Section 5 we give some remarks on bandwidth selection. We apply the method on a real data set in Section 6, and finish with a short discussion in Section 7.

2. THE POWER BASIS

The family of basis functions that we will treat in this paper is motivated by the following theorem:

Theorem 2.1 (Taylor expansion for a power basis)**.** *Let I be a non-trivial interval, $m, \phi : I \to \mathbb{R}$, $p+1$ times differentiable in I, ϕ invertible in I, and $x \in I$. Then for all $z \in I$ with $z \neq x$, a value $\zeta \in (x, z)$ resp. (z, x) exists such that*

$$m(z) = \sum_{j=0}^{p} \frac{\psi_{(j)}(x)}{j!}\big(\phi(z) - \phi(x)\big)^j + \frac{\psi_{(p+1)}(\zeta)}{(p+1)!}\big(\phi(z) - \phi(x)\big)^{p+1}$$

with

$$\psi_{(j+1)}(\cdot) = \frac{\psi'_{(j)}(\cdot)}{\phi'(\cdot)}\ , \qquad \psi_{(0)}(\cdot) = m(\cdot)\ ,$$

holds.

The proof is omitted, since this theorem is simply obtained by applying Taylor's theorem, as found for example in Lay ([12], p. 211), on the function $g(\cdot) = (m \circ \phi^{-1})(\cdot)$ at point $\phi(x)$. Assuming the underlying model (1.1), Theorem 2.1 suggests to fit the data locally in a neighborhood of x against the basis functions $1, \phi(X) - \phi(x), ..., (\phi(X) - \phi(x))^p$. We call a basis of this type a *power basis* of order p. For $\phi = id$, the power basis reduces to the polynomial basis. For the rest of this paper, we assume that $\phi : \mathbb{R} \to \mathbb{R}$ is $p+1$ times differentiable and invertible in a neighborhood of x, though the estimation procedure itself, as outlined from (2.5) to (2.7), does not necessarily require this assumption.

Since the parameters

$$\gamma_j(x) := \frac{\psi_{(j)}(x)}{j!}$$

are constructed in a more complex way than the parameters $\beta_j(x)$ for local polynomial fitting, the simple relationship $m^{(j)}(x) = j!\,\beta_j(x)$ cannot be retained. However, by using the simple recursive formula

$$\gamma_j(x) = \frac{1}{j\phi'(x)}\,\gamma'_{j-1}(x)\ , \qquad \gamma_0(x) = m(x)\ ,$$

the parameters $\gamma_j(x)$ $(j \le p)$, which we abbreviate by γ_j from now on, can be calculated. In this manner the following relations between parameters and the underlying function and their derivatives are derived for the power basis:

$$m(x) = 0!\,\gamma_0 \tag{2.1}$$

$$m'(x) = 1!\,\phi'(x)\,\gamma_1 \tag{2.2}$$

$$m''(x) = 2!\,[\phi'(x)]^2\,\gamma_2 + \phi''(x)\,\gamma_1 \tag{2.3}$$

$$m'''(x) = 3!\,[\phi'(x)]^3\,\gamma_3 + 3!\,\phi''(x)\,\phi'(x)\,\gamma_2 + \phi'''(x)\,\gamma_1 \tag{2.4}$$

$$\vdots$$

Let $w_i(x) = K_h(X_i - x)$. Minimizing

$$\sum_{i=1}^{n}\left\{Y_i - \sum_{j=0}^{p}\gamma_j\big(\phi(X_i) - \phi(x)\big)^j\right\}^2 w_i(x) \tag{2.5}$$

in terms of $(\gamma_0, ..., \gamma_p)$, one obtains the local least squares estimator $\hat\gamma = (\hat\gamma_0, ..., \hat\gamma_p)^T$. The design matrix and the necessary vectors are given by

$$\mathbf{X} = \begin{pmatrix} 1 & \phi(X_1)-\phi(x) & \cdots & \big(\phi(X_1)-\phi(x)\big)^p \\ \vdots & \vdots & & \vdots \\ 1 & \phi(X_n)-\phi(x) & \cdots & \big(\phi(X_n)-\phi(x)\big)^p \end{pmatrix},$$

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}\ , \qquad \gamma = \begin{pmatrix} \gamma_0 \\ \vdots \\ \gamma_p \end{pmatrix} \qquad \text{and} \qquad \mathbf{W} = \begin{pmatrix} w_1(x) & & \\ & \ddots & \\ & & w_n(x) \end{pmatrix}.$$

The minimization problem (2.5) can be written as

$$\min_\gamma (\mathbf{y} - \mathbf{X}\gamma)^T\,\mathbf{W}(\mathbf{y} - \mathbf{X}\gamma)\ ,$$

yielding $\hat\gamma = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\,\mathbf{W}\mathbf{y}$, just as in the case of local polynomial fitting ([7]). Then $\hat m(x) = e_1^T\hat\gamma$, where $e_1 = (1, 0, ..., 0)^T$, is an estimator for the underlying function $m(\cdot)$ at point x. Using (2.2) to (2.4), estimators for the derivatives can be obtained in a similar way. Note that, to ensure that the matrix $\mathbf{X}^T\mathbf{W}\mathbf{X}$ is invertible, at least $p+1$ design points are required to satisfy $K_h(X_i - x) > 0$. Furthermore it can be shown that

$$\mathit{Bias}(\hat\gamma|\mathbb{X}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{W}\mathbf{r}\ , \tag{2.6}$$

where $\mathbf{r} = (m(X_1), ..., m(X_n))^T - \mathbf{X}\gamma$, and $\mathbb{X}$ denotes the vector of covariates $(X_1, ..., X_n)$. Finally the conditional covariance matrix is given by

$$\mathrm{Var}(\hat{\gamma}|\mathbb{X}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\,(\mathbf{X}^T\mathbf{\Sigma}\,\mathbf{X})\,(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\ , \tag{2.7}$$

where $\mathbf{\Sigma} = \mathrm{diag}(w_i^2(x)\,\sigma^2(X_i))$.

3. ASYMPTOTICS

Usually formulas (2.6) and (2.7) cannot be used in practice, since they depend on the unknown quantities $\mathbf{r}$ and $\mathbf{\Sigma}$. Consequently an asymptotic derivation is required. We denote

$$\mu_j\ = \int_{-\infty}^{\infty} u^j K(u)\,du \qquad \text{and} \qquad \nu_j\ = \int_{-\infty}^{\infty} u^j K^2(u)\,du$$

for the j^{th} moments of K and K^2. For technical ease we assume that the kernel K is a (not necessarily symmetric) bounded probability density function (i.e. $\mu_0 = 1$) with bounded support, though the latter assumption still can be relaxed significantly (Fan [4], Fan & Gijbels [6]). Further we define the kernel moment matrices

$$\begin{array}{ll}
\mathbf{S} = (\mu_{j+l})_{0\le j,l\le p} & \mathbf{c_p} = (\mu_{p+1}, ..., \mu_{2p+1})^T \\
\tilde{\mathbf{S}} = (\mu_{j+l+1})_{0\le j,l\le p} & \tilde{\mathbf{c}}_\mathbf{p} = (\mu_{p+2}, ..., \mu_{2p+2})^T \\
\bar{\mathbf{S}} = \big((j+l)\mu_{j+l+1}\big)_{0\le j,l\le p} & \bar{\mathbf{c}}_\mathbf{p} = \big((p+1)\mu_{p+2}, ..., (2p+1)\mu_{2p+2}\big)^T \\
\mathbf{S}^* = (\nu_{j+l})_{0\le j,l\le p}\ . &
\end{array}$$

Note that the matrix $\mathbf{S}$ is positive definite and thus invertible (Tsybakov [20], Lemma 1). Furthermore we introduce the denotation $\varphi(x) = \phi'(x)$, the matrices $\mathbf{H} = \mathrm{diag}(h^j)_{0\le j\le p}$ and $\mathbf{P} = \mathrm{diag}(\varphi^j(x))_{0\le j\le p}$ and recall that $e_{j+1} = (0, ..., 0, 1, 0, ..., 0)^T$ with 1 at $(j+1)^{\text{th}}$ position. $o_P(1)$ denotes a sequence of random variables which tends to zero in probability, and $O_P(1)$ stands for a sequence of random variables which is bounded in probability. Let $f(\cdot)$ be the design density of X. Firstly, we consider interior points, i.e. we assume x to be a fixed point in the support of the design density f.

Theorem 3.1. *Assume that $f(x) > 0$, $\sigma^2(x) > 0$, $\varphi(x) \neq 0$ and that $f(\cdot)$, $m^{(p+1)}(\cdot)$, $\phi^{(p+1)}(\cdot)$ and $\sigma^2(\cdot)$ are continuous in a neighbourhood of x. Further assume that $h \to 0$ and $nh \to \infty$. Then the asymptotic conditional covariance matrix of $\hat{\gamma}$ is given by*

$$\mathrm{Var}(\hat{\gamma}|\mathbb{X})\ =\ \frac{\sigma^2(x)}{nhf(x)}\,\mathbf{P}^{-1}\,\mathbf{H}^{-1}\,\mathbf{S}^{-1}\,\mathbf{S}^*\,\mathbf{S}^{-1}\,\mathbf{H}^{-1}\,\mathbf{P}^{-1}\big(1 + o_P(1)\big)\ . \tag{3.1}$$

The asymptotic conditional bias is given by

$$\mathit{Bias}(\hat{\gamma}|\mathbb{X})\ =\ h^{p+1}\,\varphi^{p+1}(x)\,\mathbf{P}^{-1}\,\mathbf{H}^{-1}\,\big(\gamma_{p+1}\mathbf{S}^{-1}\mathbf{c_p} + h\mathbf{b}_n\big)\ , \tag{3.2}$$

where $\mathbf{b}_n = O_P(1)$. *If in addition* $f'(\cdot)$, $m^{(p+2)}(\cdot)$ *and* $\phi^{(p+2)}(\cdot)$ *are continuous in a neighbourhood of* x *and* $nh^3 \to \infty$, *the sequence* $\mathbf{b}_n$ *can be written as*

$$\begin{aligned}\mathbf{b}_n \;=\; & \left(\gamma_{p+1}\,\frac{f'(x)}{f(x)} + \gamma_{p+2}\,\varphi(x)\right)\mathbf{S}^{-1}\,\tilde{\mathbf{c}}_{\mathbf{p}} \;+\; \gamma_{p+1}\,\frac{\varphi'(x)}{2\,\varphi(x)}\,\mathbf{S}^{-1}\,\bar{\mathbf{c}}_{\mathbf{p}} \\ & -\;\gamma_{p+1}\,\mathbf{S}^{-1}\left(\frac{f'(x)}{f(x)}\,\tilde{\mathbf{S}} - \frac{\varphi'(x)}{2\,\varphi(x)}\,\bar{\mathbf{S}}\right)\mathbf{S}^{-1}\,\mathbf{c}_{\mathbf{p}} \;+\; o_P(1)\;.\end{aligned} \tag{3.3}$$

This theorem was obtained in the case $\phi(x) = x$ and $p = 1$ by Fan [4], for general p by Ruppert & Wand [16], and for general $\phi(\cdot)$, $p = 1$, and symmetric kernels by Einbeck [3]. Based on Theorem 3.1 and formula (2.1) asymptotic expressions for bias and variance of the estimator of the conditional mean function can be derived. In particular we obtain

$$\begin{aligned}\mathrm{Var}(\hat{m}(x)|\mathbb{X}) \;&=\; \mathrm{Var}(e_1^T\hat{\gamma}|\mathbb{X}) \\ &=\; \frac{\sigma^2(x)}{nhf(x)}\,e_1^T\mathbf{S}^{-1}\,\mathbf{S}^*\,\mathbf{S}^{-1}\,e_1\big(1+o_P(1)\big)\;,\end{aligned} \tag{3.4}$$

which reduces for $p = 1$ to

$$\mathrm{Var}(\hat{m}(x)|\mathbb{X}) \;=\; \frac{\sigma^2(x)}{nhf(x)}\,\frac{\int(\mu_2 - u\mu_1)^2\,K^2(u)\,du}{(\mu_0\mu_2-\mu_1^2)^2}\,\big(1+o_P(1)\big)\;. \tag{3.5}$$

It is an important observation that the asymptotic variance of $\hat{m}(\cdot)$ does not depend on the basis function. Next, we take a look at the bias. Using (3.2) and (2.1) one gets

$$\begin{aligned}\mathit{Bias}(\hat{m}(x)|\mathbb{X}) \;&=\; \mathit{Bias}(e_1^T\hat{\gamma}|\mathbb{X}) \\ &=\; h^{p+1}\varphi^{p+1}(x)\,e_1^T\left(\frac{\psi_{(p+1)}(x)}{(p+1)!}\,\mathbf{S}^{-1}\,\mathbf{c}_{\mathbf{p}} + h\mathbf{b}_n\right),\end{aligned} \tag{3.6}$$

reducing for $p = 1$ to

$$\mathit{Bias}(\hat{m}(x)|\mathbb{X}) \;=\; \frac{h^2}{2}\left(m''(x) - \frac{\varphi'(x)}{\varphi(x)}\,m'(x)\right)\frac{\mu_2^2-\mu_1\mu_3}{\mu_0\mu_2-\mu_1^2}\,\big(1+o_P(1)\big)\;. \tag{3.7}$$

3.1. Derivatives

Similarly, one might take a look at the formulas for the derivatives. Using (2.2), one gets for the derivative estimator for $p = 1$

$$\begin{aligned}\mathrm{Var}\big(\hat{m}'(x)|\mathbb{X}\big) \;&=\; \mathrm{Var}\big(\varphi(x)\,e_2^T\hat{\gamma}|\,\mathbb{X}\big) \\ &=\; \frac{\sigma^2(x)}{nh^3f(x)}\,e_2^T\,\mathbf{S}^{-1}\,\mathbf{S}^*\,\mathbf{S}^{-1}e_2\big(1+o_P(1)\big) \\ &=\; \frac{\sigma^2(x)}{nh^3f(x)}\,\frac{\int(\mu_1-u\mu_0)^2K^2(u)\,du}{(\mu_0\mu_2-\mu_1^2)^2}\,\big(1+o_P(1)\big)\;,\end{aligned} \tag{3.8}$$

and

$$\begin{aligned} Bias\big(\hat{m}'(x)|\mathbb{X}\big) &= Bias\big(\varphi(x)\, e_2^T \hat{\gamma}|\, \mathbb{X}\big) \\ &= h^p \varphi^{p+1}(x)\, e_2^T \left(\frac{\psi_{(p+1)}(x)}{(p+1)!}\, \mathbf{S}^{-1}\mathbf{c_p} + h\mathbf{b}_n \right) \\ &= \frac{h}{2}\left(m''(x) - \frac{\varphi'(x)}{\varphi(x)}\, m'(x) \right) \frac{\mu_0\mu_3 - \mu_1\mu_2}{\mu_0\mu_2 - \mu_1^2} \big(1 + o_P(1)\big)\ , \end{aligned} \tag{3.9}$$

where (3.8) and (3.9) still hold for general p. Looking at (2.3) and (2.4), one might have the impression that the asymptotic formulas for higher derivatives will be extraordinarily complicated. However, first order expansions are easy to derive, since only the leading term $j!\,\varphi^j(x)\,\gamma_j$ determines the asymptotic behaviour. In particular, one gets for arbitrary $j \leq p$

$$\mathrm{Var}\big(\hat{m}^{(j)}(x)|\mathbb{X}\big) = \frac{(j!)^2\,\sigma^2(x)}{nh^{2j+1}f(x)}\, e_{j+1}^T\, \mathbf{S}^{-1}\, \mathbf{S}^*\, \mathbf{S}^{-1}\, e_{j+1}\big(1 + o_P(1)\big)$$

and

$$Bias\big(\hat{m}^{(j)}(x)|\mathbb{X}\big) \;=\; h^{p+1-j}\, j!\, \varphi^{p+1}(x)\, e_{j+1}^T \left(\frac{\psi_{(p+1)}(x)}{(p+1)!}\, \mathbf{S}^{-1}\mathbf{c_p} + o_P(1) \right).$$

Note that the formula for the variance is identical to the corresponding formula for local polynomial modelling ([7], p. 62), and that the variance is independent of the basis function for any choice of j and p.

3.2. Design adaption and automatic boundary carpentry

One might wonder why we provided a deeper derivation of $\mathbf{b}_n$ in Theorem 3.1. This is necessary due to a special property of symmetric kernels. Let us consider symmetric kernels throughout the rest of this section. Then, we have $\mu_{2k+1} = \nu_{2k+1} = 0$ for all $k \in \mathbb{N}_0$. The crucial point is that, when estimating the j^{th} derivative $\hat{m}^{(j)}(\cdot)$, the product $e_{j+1}^T \mathbf{S}^{-1}\mathbf{c_p}$ is zero iff $p - j$ is even. In the case $j = 0$, p even, one gets from (3.6)

$$Bias(\hat{m}(x)|\mathbb{X}) = \; h^{p+2}\varphi^{p+1}(x)\, e_1^T \mathbf{b}_n\ . \tag{3.10}$$

Suppose one increases the order of a power basis from an even order p to an odd order $p+1$. Obviously, the order $O\left(\frac{1}{nh}\right)$ of the variance (3.4) is unaffected, and Fan & Gijbels ([7], p. 77 f) show that the quantity $e_1^T\mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1}e_1$ remains constant when moving from an even p to $p+1$. Thus, there is not any change in variance. What about the bias? As can be seen from (3.10) and (3.6), the order of the bias remains to be $O(h^{p+2})$. However, for even p the bias involves the design density f and its derivative f', i.e. the estimator is not "design-adaptive" in the sense of Fan [4]. Regarding the case $j = 1$, the situation is similar: the matrix product $e_2^T\mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1}e_2$ remains constant when moving from an odd p to $p+1$, while the leading term of the bias simplifies. Summarizing, an odd choice

of $p-j$ should be preferred to an even choice, and local estimators based on a power basis show exactly the same behavior as local polynomial estimators in terms of design-adaptivity.

Beside this, "odd" local polynomial estimators have another strong advantage compared to "even" ones: they do not suffer from boundary effects and hence do not require boundary corrections. Does this property carry over to estimators based on a power basis as well? We answer this question by considering the case $p=1$ and $j=0$, though the findings remain valid for any odd choice of $p-j$. For a symmetric kernel and an interior point, (3.5) and (3.7) reduce to

$$\mathrm{Var}(\hat{m}(x)|\mathbb{X}) \;=\; \frac{\sigma^2(x)\,\nu_0}{nhf(x)}\,\big(1+o_P(1)\big) \tag{3.11}$$

and

$$\mathit{Bias}(\hat{m}(x)|\mathbb{X}) \;=\; \frac{h^2\mu_2}{2}\left(m''(x)-\frac{\varphi'(x)}{\varphi(x)}\,m'(x)\right)+o_P(h^2)\;, \tag{3.12}$$

respectively. The variance is exactly the same as for a local linear fit, while the bias expression includes an additional term expressing the interplay between the basis function and the underlying function. Let us consider boundary points now. Without loss of generality we assume that the density f has a bounded support [0;1]. We write a left boundary point as $x=ch$ ($c\geq 0$), and accordingly a right boundary point as $x=1-ch$. Calculation of the asymptotic bias and variance is straightforward as in Theorem 3.1; the only difference is that kernel moments μ_j and ν_j have to be replaced by

$$\mu_{j,c}=\int_{-c}^{\infty}u^jK(u)\,du \qquad \text{and} \qquad \nu_{j,c}=\int_{-c}^{\infty}u^jK^2(u)\,du$$

in case of a left boundary point, and analogously in case of a right boundary point. Thus, the kernel moments never vanish and the problem corresponds to finding bias and variance for asymmetric kernel functions. Indeed, one obtains at $x=ch$

$$\mathrm{Var}(\hat{m}(x)|\mathbb{X}) \;=\; \frac{\sigma^2(0+)}{nhf(0+)}\,\frac{\int(\mu_{2,c}-u\mu_{1,c})^2K^2(u)\,du}{(\mu_{0,c}\,\mu_{2,c}-\mu_{1,c}^2)^2}\,\big(1+o_P(1)\big) \tag{3.13}$$

and

$$\mathit{Bias}(\hat{m}(x)|\mathbb{X})=\frac{h^2}{2}\left(m''(0+)-\frac{\varphi'(0+)}{\varphi(0+)}m'(0+)\right)\frac{\mu_{2,c}^2-\mu_{1,c}\,\mu_{3,c}}{\mu_{0,c}\,\mu_{2,c}-\mu_{1,c}^2}\,\big(1+o_P(1)\big)\;. \tag{3.14}$$

Comparing (3.11) and (3.12) with (3.13) and (3.14) unveils that the rate of the estimator does not depend on the location of the target point x. For a nice demonstration of the dependence of the constant factors on c see Fan & Gijbels [5]. For even values of $p-j$, the rate of convergence at boundary points is slower than in the interior.

3.3. Bias reduction

According to equation (3.12), the bias of a first-order-fit depends on the basis $\phi(\cdot)$. This effect may be useful for bias reduction. To investigate this, firstly note that (3.12) reduces to the well-known formula

$$Bias(\hat{m}(x)|\mathbb{X}) = \frac{h^2\mu_2}{2}\, m''(x) + o_P(h^2)$$

in the special case of local linear fitting. Thus the subtraction of $\frac{\varphi'(x)}{\varphi(x)}m'(x)$ in (3.12) provides the chance for bias reduction. In the optimal case, the content of the bracket in (3.12) is zero, hence the differential equation

$$m''(x)\,\varphi(x) - m'(x)\,\varphi'(x) \;=\; 0$$

has to be solved, what leads to the solutions

$$\varphi(x) = c_1 m'(x) \qquad (c_1 \in \mathbb{R})$$

and hence

$$\phi(x) = c_1 m(x) + c_2 \qquad (c_1, c_2 \in \mathbb{R})\ . \tag{3.15}$$

Note that for symmetric kernels and $p-j$ odd one has $e_{j+1}^T \mathbf{b}_n = o_P(1)$. Thus, the remaining asymptotic bias is even of order $o_P(h^3)$. Having an optimal basis function in the form of (3.15), one may ask if there is any gain in increasing the order p? One finds immediately $\psi_{(1)}(x) = 1/c_1$ and thus

$$\gamma_p(x) = \psi_{(p)}(x)/p! = 0 \qquad \text{for} \;\; p \geq 2\ . \tag{3.16}$$

Thus any additional terms are superfluous, since their parameters should take optimally the value zero. The strategy should consequently be the following: work with $p = 1$, and try to find a basis which is as near as possible to the underlying function.

In particular, for $c_1 = 1$, $c_2 = 0$ we get $\phi_{opt}(x) = m(x)$, thus the underlying function $m(\cdot)$ is a member of the family of optimal basis functions. Certainly, the function $m(\cdot)$ is always unknown. However, there are still at least two ways to use this result. We want to approach them from a philosophical point of view. What does a basis function actually effect? For a given basis, the smoothing step in fact balances between the information given by the basis and the data. A similar concept is well-known from Bayesian statistics (see e.g. Smith & Kohn [18]). Though the Bayesian prior does not contain a basis function but an assumption about the distribution of unknown parameters, the principle, boldly compared, is the same, since the posterior distribution can be interpreted as a trade-off between information in the data and prior knowledge. Thus, having some ("prior") knowledge about m, the fitted ("posteriori") curve can be steered in the correct direction when incorporating this knowledge in the basis. If there does not exist any knowledge about m, one can calculate a pilot estimate via a local linear fit (or any other smooth fit, e.g. with splines) and use the estimated function as an improved basis. In the following section we will provide examples for the application of these strategies.

4. A SIMULATED EXAMPLE

Throughout this section, we consider the underlying function

$$m(x) = x + \frac{1}{1.2\sqrt{2\pi}}\, e^{-(x-0.2)^2/0.02} - \frac{1}{0.9\sqrt{2\pi}}\, e^{-(x-0.7)^2/0.0018}\ , \tag{4.1}$$

which we contaminate with Gaussian noise with $\sigma = 0.05$. The 50 predictors are uniformly distributed on $[0;1]$. We repeated this simulation 50 times, obtaining 50 data sets. See Fig. 3 for getting an impression of the data set. As a measure of performance, we use the relative squared error

$$\mathrm{RSE}(\hat{m}) = \frac{\|\hat{m} - m\|}{\|m\|} = \frac{\sqrt{\sum_{i=1}^{n} \left(m(X_i) - \hat{m}(X_i)\right)^2}}{\sqrt{\sum_{i=1}^{n} m(X_i)^2}}\ .$$

For each simulated data set and for each estimation $\hat{m}$ of m with different basis functions and polynomial orders we select the empirically optimal bandwidth h_{emp} by

$$h_{emp} = \min_h \mathrm{RSE}(\hat{m})\ .$$

This bandwidth h_{emp} is used for the corresponding fit, and the medians of the 50 RSE values obtained in this manner are shown in Table 1. (Of course, h_{emp} only may be calculated for simulated data. Bandwidth selection for real data is treated in Section 5.) The function $\mathrm{dnorm}(x)$ denotes the density of the standard normal distribution. We put a star $(*)$ behind the RSE if the value is better than that for local linear fitting $(\phi(x) = x)$ and two stars for the winner of the column.

Table 1: Medians of RSEs for various polynomial orders and basis functions.

$\phi(x)$	$p=1$	$p=2$	$p=3$	$p=8$
x	0.04819	0.05005	0.04915	0.04973
$\sin x$	0.04810 **	0.05003 *	0.04904 *	0.05008
$\arctan x$	0.04812 *	0.04997 *	0.04911 *	0.05011
$\cosh x$	0.04898	0.04919 *	0.04916	0.04634 **
dnorm x	0.04893	0.04888 **	0.04844 **	0.04844 *
$\exp x$	0.04829	0.05005	0.04917	0.04886 *
$\log(x+1)$	0.04811 *	0.04988 *	0.04917	0.05000

The corresponding boxplots of the RSE values are presented in Fig. 1. Taking a look at the table and the figure, one notes immediately that the differences between different basis functions are mostly negligible, and the performance does not improve when rising the polynomial order. Looking at the table in more depth, one observes that the group of odd basis functions behaves slightly different

than the group of even basis functions. In particular, for $p = 1$ the odd basis functions outperform the even ones. Recalling equation (3.15), this might be interpreted as that the underlying function $m(\cdot)$ possesses rather odd than even characteristics. Finally, one observes that the Gaussian basis yields the best RSE for $p = 2$ and $p = 3$. This is quite intuitive, since the underlying function contains a sum of Gaussians itself.

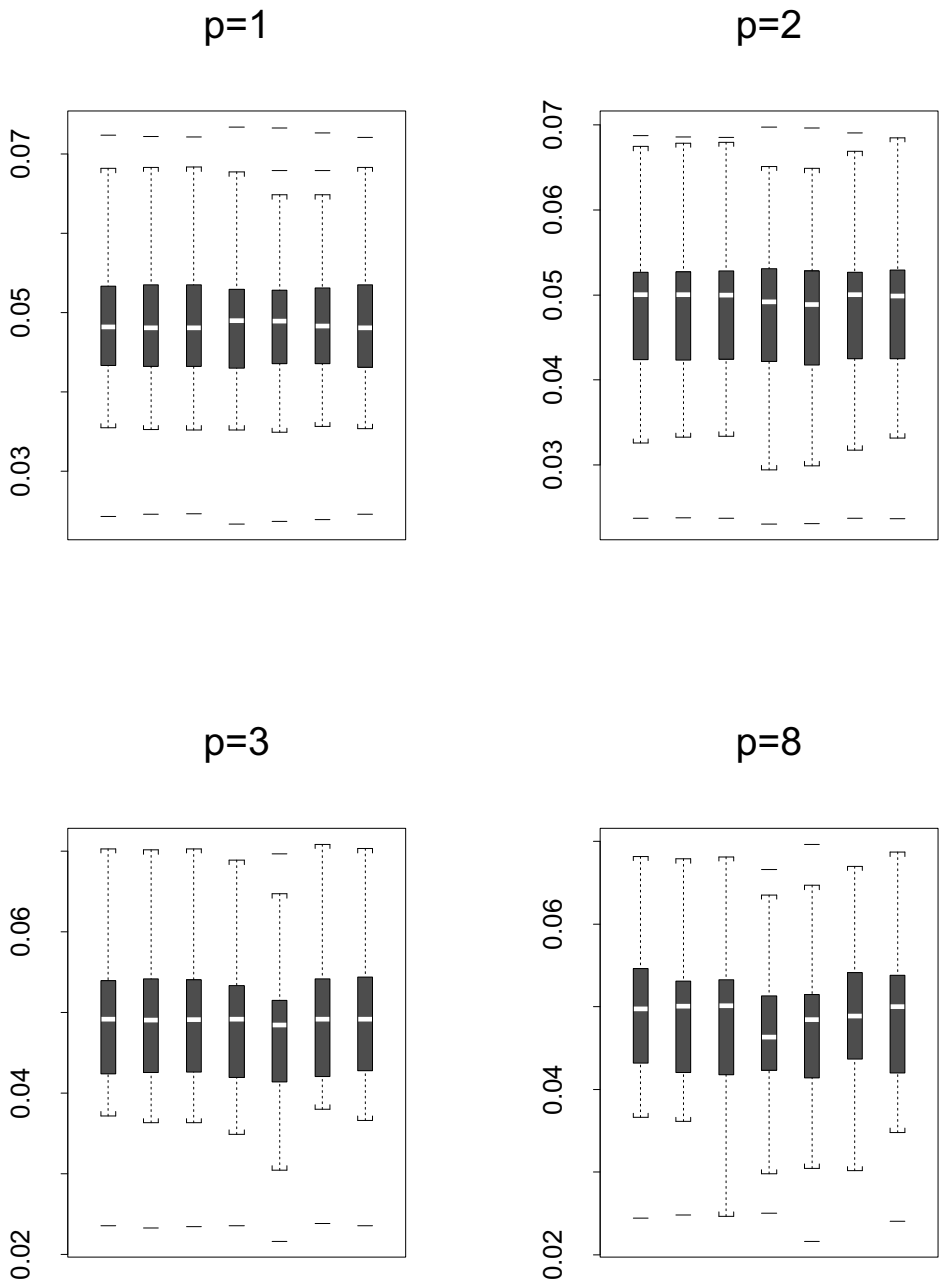

Figure 1: Boxplots of the relative errors using the basis functions $\phi(x) = x$, $\sin x$, $\arctan x$, $\cosh x$, $\mathrm{dnorm}(x)$, $\exp x$, $\log(x+1)$ and orders $p = 1, 2, 3, 8$.

Next, we will investigate if these results may be improved by the use of basis functions which contain information about the true function, as suggested by (3.15). We distinguish two situations:

a) *Some information about m is available.* We consider exemplarily two cases:

- Assume that the information about the true function is incomplete, e.g. due to a transmission problem, and the true function is only known on the interval $[0.25; 0.75]$ (i.e., only half of the true function is known!). A basis function $m_1(\cdot)$ is constructed by extrapolating the known part of the function by straight lines in a way that the first derivative is continuous.
- Assume somebody gave us a (partly wrong) information about the underlying function (4.1), namely

$$m_2(x) \;=\; x - \frac{1}{1.2\,\sqrt{2\pi}}\, e^{-(x-0.2)^2/0.02} - \frac{1}{0.9\,\sqrt{2\pi}}\, e^{-(x-0.7)^2/0.0018}\;,$$

  i.e. the first hump shows down instead of up.

We call basis functions like that "guessed" basis functions.

b) *No information about m is available.* In this case, we employ the pre-fit basis functions $\bar{m}(\cdot)$ and $\check{m}(\cdot)$ calculated with a local constant or linear fit, respectively. Let g_{emp} be the empirically optimal bandwidth of the pre-fit, i.e. $g_{emp} = h^{NW}_{emp}$ for a local constant (*N*adaraya-*W*atson) pre-fit and $g_{emp} = h^{LL}_{emp}$ for a local linear (LL) pre-fit. The bandwidth of the pre-fit is then selected as $g = \theta \cdot g_{emp}$, and the second bandwidth as $h = \lambda \cdot h^{LL}_{emp}$, where θ and λ are optimized in terms of RSE on $[1; 2] \times [1; 2]$.

Keeping in mind observation (3.16) and the conclusions drawn from Table 1, we only consider the case $p = 1$ from now on. The medians of 50 RSE values for each basis function are listed in Table 2. For comparison we added the results for the linear basis $\phi(x) = x$ and the (in practice unavailable) optimal basis $\phi(x) = m(x)$. The corresponding boxplots of RSE values are depicted in Fig. 2. In Fig. 3 the basis functions from Table 2 and the corresponding fitted curves are depicted. One notices again: the more similar basis and true function are, the better is the fitted curve. Further, one observes that there is not much gain in using a local linear instead of a local constant pre-fit. The benefit of applying a pre-fit basis is not overwhelming in this example, and is not as impressive as for multivariate predictors ([3]). Taking into account the difficulty of having to select two bandwidths, it is at least questionable if this additional work is worth the effort for univariate predictors. Nevertheless, in the next section we will give some insight in the nature of this two-dimensional bandwidth selection problem.

The "guessed" basis functions lead to a significant improvement, which does not require any extra work compared to a simple local linear fit. This is finally the principal message of this paper: if you have some information, use it in your basis, and your fit will improve. If this basis is wrong, but at least smooth, normally nothing serious should happen, since the commonly applied linear basis is a *wrong* basis as well in the most situations. Replacing one wrong and smooth basis by another wrong and smooth basis function will not make much difference, as demonstrated in Table 1.

Table 2: Medians of relative squared errors for improved basis functions.

ϕ	$p=1$	
x	0.04819	
$\bar{m}(x)$	0.04606	$*$
$\check{m}(x)$	0.04538	$*$
$m_1(x)$	0.04488	$*$
$m_2(x)$	0.03758	$**$
$m(x)$	0.01302	

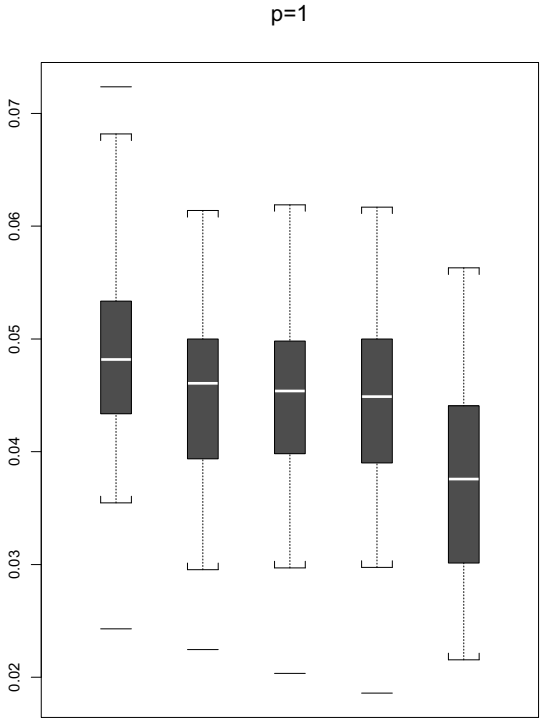

Figure 2: Boxplots of the relative errors using the basis functions $\phi(x) = x$, $\bar{m}(x)$, $\check{m}(x)$, $m_1(x)$ and $m_2(x)$ (from left to right) with $p = 1$.

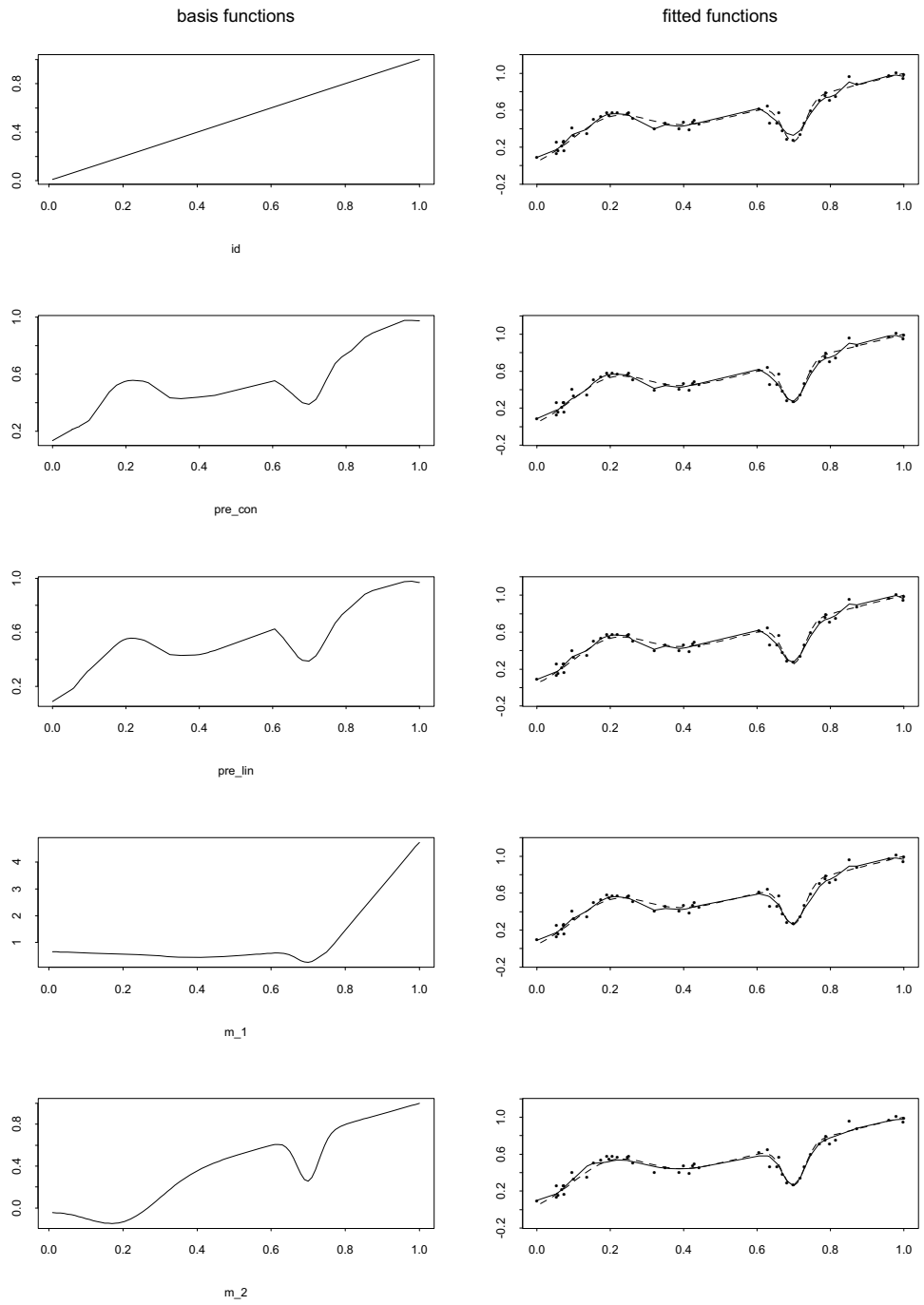

Figure 3: Left: basis functions; right: One particular of the 50 simulated data sets $(\cdot)$, true function (dashed line) and fitted functions (solid line) for $p = 1$. The denotations "pre_con" and "pre_lin" refer to the basis functions $\bar{m}$ and $\check{m}$, respectively.

5. NOTES ABOUT BANDWIDTH SELECTION

For bandwidth selection, one has the general choice between classical methods and plug-in methods. For an overview of bandwidth selection routines, we refer to Fan & Gijbels ([7], p. 110 ff). Classical methods as cross-validation or the AIC criterion can be applied directly on fitting with general basis functions. Promising extensions of the classical methods have been given by Hart & Yi [9] and Hurvich et al. [11]. In the last decades, classical approaches got the reputation to perform inferior in comparison to plug-in approaches, as treated by Fan & Gijbels [6], Ruppert et al. [15], and Doksum et al. [2], among others. However, this seems not to be justified, as Loader [13] explains, since plug-in-approaches require more theoretical assumptions about the underlying function than classical approaches. Plug-in estimators perform a pilot estimate in order to estimate the asymptotic mean square error, which is then minimized in terms of the bandwidth. Each plug-in-estimator is designed exclusively for a special smoothing method, so that application of these estimators for general basis functions requires some extra work.

Using Theorem 3.1, plug-in formulas for bandwidth selection can be derived straightforwardly by extending the corresponding methods for local polynomial fitting. We will not provide a complete treatment of this topic now, but only give some impressions of the results. Let us therefore consider the derivation of the asymptotically optimal variable bandwidth $h_{opt}(x)$, which varies with the target value x. Minimizing the asymptotic mean square error $MSE(\hat{m}(x)|\mathbb{X}) = Bias^2(\hat{m}(x)|\mathbb{X}) + \mathrm{Var}(\hat{m}(x)|\mathbb{X})$ for odd $p-j$, whereby (3.6) and (3.4) are employed for the bias resp. variance, we arrive at an asymptotically optimal bandwidth

$$h_{opt}^{(\phi)}(x) \;=\; C_{0,p}(K) \left[\frac{\sigma^2(x)}{\psi_{(p+1)}^2(x)\, f(x)\, \varphi^{2p+2}(x)} \right]^{\frac{1}{2p+3}} \cdot\, n^{-\frac{1}{2p+3}} \;, \tag{5.1}$$

where the constant $C_{0,p}(K)$, which only depends on p and the kernel K, is the same as in [7], p. 67. Recall from the end of Section 3 that $\psi_{(p+1)}(x)$ $(p \geq 1)$ approximates zero when $\phi(x)$ approximates $m(x)$. Consequently, the optimal bandwidth tends to infinity when the basis approximates the true function, what is in conformity to the observations which can be drawn from Fig. 4.

Bandwidth selection is especially difficult for data-adaptive basis functions as in the previous section: then we need the two bandwidths g and h for the first and second fit, respectively. We want to give some insight in this bandwidth selection problem, assuming for simplicity that the pre-fit $\bar{m}_g(x)$ is a local constant estimator with constant bandwidth g. Intuitively, one would firstly select an (in some sense, e.g. asymptotically) optimal bandwidth $\bar{g}$ of the pre-fit. Afterwards, one would use the resulting fit $\bar{m}_{\bar{g}}$ as a basis for the second fit, applying an optimized bandwidth $h_{opt}^{(\bar{m}_{\bar{g}})}$ for this pre-fit basis. However, this step-wise strategy in practice does not prove to be suitable: when the first fit is too wiggly, the wiggliness carries over to the second fit. Moreover, when the optimal bandwidth

is met in the first fit, then the optimal second bandwidth is very high and the minimum of the RSE curve is very flat. In other words: in this case the second fit is superfluous, and the improvement compared to a usual local fit is negligible.

Therefore, it is sensible to use somewhat higher bandwidths in the initial fit. To illustrate this, we return to the example from the previous sections, and examine exemplarily the particular data set depicted in Fig. 3. Following the step-wise strategy outlined above, we select $g = 0.015$ and $h = 0.048$. However, minimizing the RSE simultaneously over g and h, one obtains the empirically optimal bandwidth combination (0.030, 0.021). The dependence of the RSE on the bandwidth for different basis functions is demonstrated in Fig. 4. The RSE curve for the initial fit is the solid line, having a minimum at $g = 0.015$ and yielding an estimate $\bar{m}_{15}(x)$. Applying this estimate as a basis function, one gets the dotted line. However, applying the estimate $\bar{m}_{30}(x)$, obtained by a local constant fit with bandwidth $g = 0.030$, one gets the dashed curve. One sees that its minimum is deeper and more localized than that of $\bar{m}_{15}(x)$.

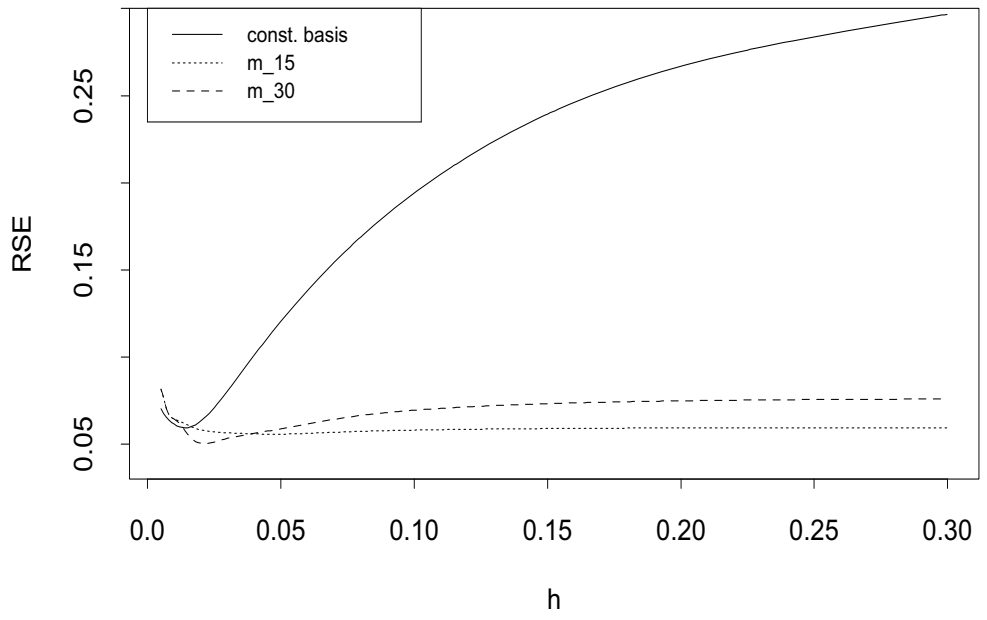

Figure 4: RSE as function of the bandwidth for a local constant fit, and for the basis functions $\bar{m}_{15}(x)$, and $\bar{m}_{30}(x)$.

In Section 4 we have already suggested that suitable bandwidths g and h for the pre-fitting algorithm are $1 - 2$ times bigger than the optimal bandwidths of a local constant or a local linear fit, respectively. We want to provide some heuristics to motivate this. Assume that the best bandwidth combination minimizing the RSE simultaneously over (g, h) is given by $(\theta \cdot \bar{g}, \lambda \cdot h_{opt}^{LL})$, where $\bar{g} = h_{opt}^{NW}$. Since we cannot access λ directly, we have to apply a sort of trick and to work with a *variable* second bandwidth. Setting (5.1) for $p = 1$ in relation to

the optimal variable bandwidth $h_{opt}^{LL}(x)$ for a local linear fit, one obtains

$$\frac{h_{opt}^{(\phi)}(x)}{h_{opt}^{LL}(x)} = \left[1 - \frac{\phi''(x)}{\phi'(x)} \cdot \frac{m'(x)}{m''(x)}\right]^{-2/5} .$$

We define the quantity

$$m^{\circ}(x) = \frac{m''(x)}{m'(x)} ,$$

and substitute for ϕ the pre-fit basis $\bar{m}_{\theta\cdot\bar{g}}$. Then one obtains

$$\lambda_x := \frac{h_{opt}^{(\bar{m}_{\theta\cdot\bar{g}})}(x)}{h_{opt}^{LL}(x)} = \left[1 - \frac{\bar{m}^{\circ}_{\theta\cdot\bar{g}}(x)}{m^{\circ}(x)}\right]^{-2/5} \approx \left[1 - \frac{\bar{m}^{\circ}_{\theta\cdot\bar{g}}(x)}{\bar{m}^{\circ}_{\bar{g}}(x)}\right]^{-2/5} . \tag{5.2}$$

What can be said about the relation between λ_x and θ? Writing $\bar{m}_g(x) = \sum_{i=1}^n w_i(x) Y_i / \sum_{i=1}^n w_i(x)$, where $w_i(x) = \frac{1}{g} K\big(\frac{X_i - x}{g}\big)$, one calculates

$$\begin{aligned} \bar{m}^{\circ}_g(x) &= \frac{\bar{m}''_g(x)}{\bar{m}'_g(x)} = \frac{\sum_{i=1}^n w_i\prime\prime(x)\big(Y_i - \bar{m}_g(x)\big)}{\sum_{i=1}^n w_i\prime(x)\big(Y_i - \bar{m}_g(x)\big)} - 2\,\frac{\sum_{i=1}^n w_i\prime(x)}{\sum_{i=1}^n w_i(x)} \\ &= -\frac{1}{g}\,\frac{\sum_{i=1}^n K''\big[(X_i - x)/g\big]\,\big(Y_i - \bar{m}_g(x)\big)}{\sum_{i=1}^n K'\big[(X_i - x)/g\big]\,\big(Y_i - \bar{m}_g(x)\big)} + \frac{2}{g}\,\frac{\sum_{i=1}^n K'\big[(X_i - x)/g\big]}{\sum_{i=1}^n K\big[(X_i - x)/g\big]} . \end{aligned} \tag{5.3}$$

One observes from (5.3) that, roughly approximated,

$$\frac{\bar{m}^{\circ}_{\theta\bar{g}}(x)}{\bar{m}^{\circ}_{\bar{g}}(x)} \approx \frac{1}{\theta} .$$

We substitute this quotient in (5.2) and get

$$\lambda_x \approx \left(1 - \frac{1}{\theta}\right)^{-2/5} \tag{5.4}$$

In order to get a notion about this relation, we assume for a moment equality in (5.4). The function

$$\lambda(\theta) = \big(1 - \theta^{-1}\big)^{-2/5} \tag{5.5}$$

is depicted in Fig. 5 (left). The hyperbolic shape of this function can be observed in reality as well. Let us consider the same data set as utilized in Fig. 4. Performing the pre-fit algorithm for g, h varying on a two-dimensional grid, the resulting RSE values are shown in Fig. 5 (right). The same hyperbola appears again. Thus, the minima of the RSE in terms of the pairs (g, h) are situated along a hyperbola-formed valley. We want to emphasis three special positions in this valley:

- $\theta \to \infty$. Then the first fit is a constant, and the resulting fit is the Nadaraya–Watson-estimator.

- $\lambda \to \infty$. Then the second fit is a parametric regression with a Nadaraya–Watson estimate as basis (which is approximately the same as the previous case).
- $\lambda = \theta$. Then one has $1 = \lambda - \lambda^{-3/2}$, which is solved at about $\lambda = 1.53$. This number corresponds to the magnitude recommended beforehand.

Yet, a generally optimal choice of λ and θ cannot be given. At least we can motivate that the main problem of bandwidth selection for the pre-fitting algorithm can be reduced to the problem of selecting the bandwidth of a local constant or a local linear fit, for the solution of which exist a variety of well established methods. The remaining problem is a problem of fine tuning of the parameters θ and λ. Though all considerations in this sections were outlined within the framework of a local constant pre-fit, they remain qualitatively the same for a local linear pre-fit. Indeed, there seems to be no observable advantage of a local linear compared to a local constant pre-fit. Since local constant fitting is more simple than local linear fitting, one might prefer local constant; however, it might be simpler to base both bandwidth selection problems on a local linear fit.

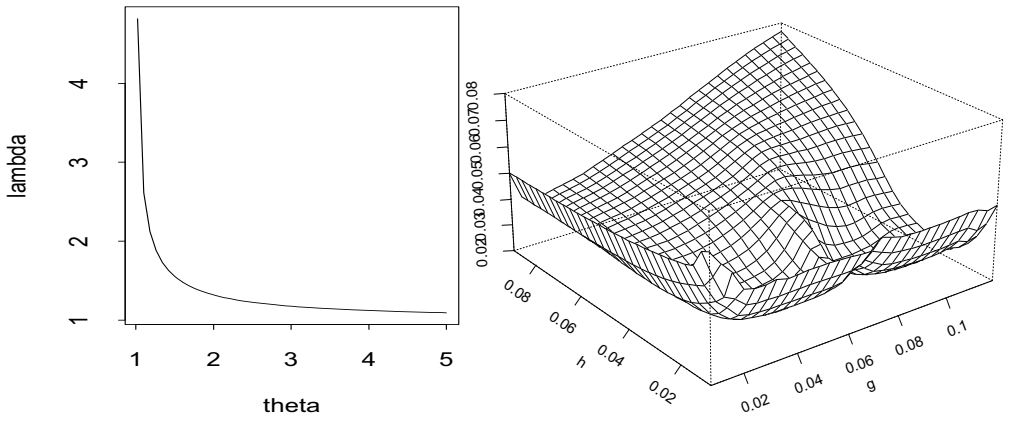

Figure 5: Left: function $\lambda(\theta)$; right: RSE for varying (g, h).

6. A REAL DATA EXAMPLE

In this section we consider the motorcycle data, firstly provided by Schmidt et al. [17], which have been widely used in the smoothing literature to demonstrate the performance of nonparametric smoothing methods (e.g. [7], p. 2). The data were collected performing crash tests with dummies sitting on motorcycles. The head acceleration of the dummies (in g) was recorded a certain time (measured in milliseconds) after they had hit a wall. (Note however that, strictly considered, these data are not fitting the models on which they are usually applied, since there were taken several measurements from every dummy at different

time points — thus the data possess an inherent dependence structure. As done in the other literature, we will ignore this problem in the following).

Fig. 6 shows the motorcycle data with a local linear fit (solid line). The bandwidth value 1.48 is obtained by cross-validation. According to the previous sections, the bandwidths g and h should be selected from the interval $[1.48; 2.96]$. Visually, the setting $g = h = 2.6$ was convincing for this data set. The dotted line shows the local linear pre-fit, and the dashed line is the local fit obtained using the pre-fit as basis function. For comparison, we also provide the result of a fit with smoothing splines.

For real data it is hard to judge which fit might be the best one — but at least it seems that the fit applying a local pre-fit basis is less biased at the first bend and the first hump, achieving at the same time a higher smoothness in the outer right area than a local linear fit. The performance seems now comparable to a spline fit.

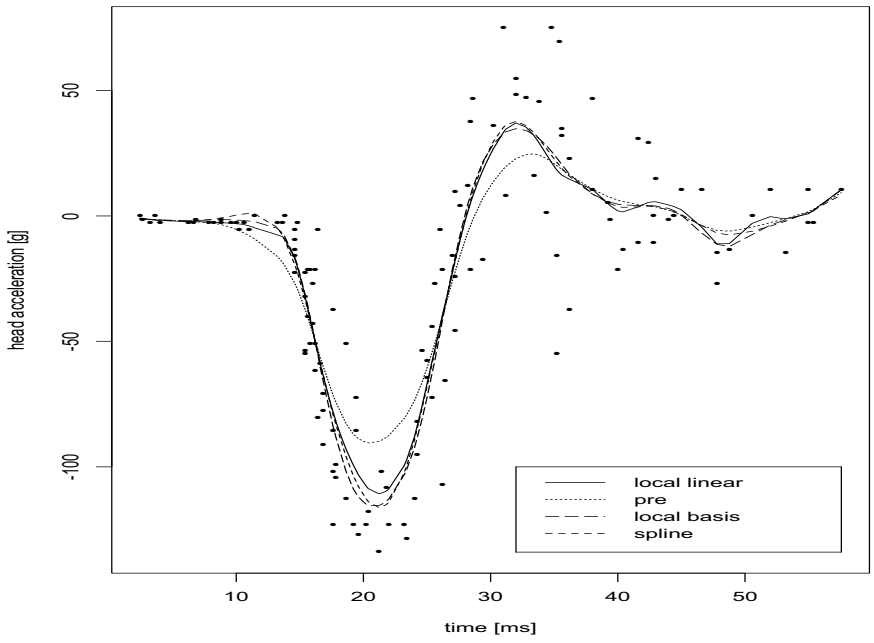

Figure 6: Motorcycle data with a local linear fit, a local pre-fit, a local fit using the latter fit as basis function, and a spline fit.

7. DISCUSSION

In a certain sense, the main findings of this paper are quite naive. Certainly, everyone has the notion that, when instilling more information about the true function in the basis, the resulting fit should improve. However, it seems that this notion never has been concretized neither from a theoretical nor from a practical point of view, though related ideas have already been mentioned in Ramsay & Silverman ([14], Section 3.3.3) and Hastie & Loader [10]. The main purpose of this work was to fill this gap, and we could confirm the intuitive notion by theoretical as well as practical results. Summarizing, bias reduction is definitely possible when using suitable basis functions, and the possible gain is much bigger than the possible loss by using wrong, but smooth, basis functions. However, application of the pre-fit algorithm can not be unrestrictedly recommended in general, since the possible gain compared to the effort is not overwhelming, at least in the univariate case.

In the framework of this paper it was not possible to solve all open questions completely. There remain some open problems especially concerning bandwidth selection in the case of pre-fitting. Furthermore, it would be useful to know when pre-fitting yields to a significant improvement and when not.

A. APPENDIX

Proof of Theorem 3.1

I. Asymptotic conditional variance

Whenever there appears in integral in this proof, the borders $-\infty$ and ∞ are omitted. We denote $S_{n,j} = \sum_{i=1}^n w_i(x)\,(\phi(X_i) - \phi(x))^j$ and $S^*_{n,j} = \sum_{i=1}^n w_i^2(x)\,\sigma^2(X_i)\,(\phi(X_i) - \phi(x))^j$. Then $\mathbf{S_n} := (S_{n,j+l})_{0\le j,l\le p} = \mathbf{X}^T\mathbf{W}\mathbf{X}$ and $\mathbf{S^*_n} := (S^*_{n,j+l})_{0\le j,l\le p} = \mathbf{X}^T\mathbf{\Sigma}\,\mathbf{X}$ hold, and the conditional variance (2.7) can be written as

$$\mathrm{Var}(\hat\gamma|\mathbb{X}) = \mathbf{S_n}^{-1}\,\mathbf{S^*_n}\,\mathbf{S_n}^{-1} \tag{A.1}$$

and thus approximation of the matrices $\mathbf{S_n}$ and $\mathbf{S^*_n}$ is required. Using that

$$\int K(u)\,u^j\,g(x+hu)\,du \;=\; \mu_j\,g(x) + o(1)$$

for any function $g : \mathbb{R} \mapsto \mathbb{R}$ which is continuous in x, we obtain

$$\begin{aligned} ES_{n,j} \;&=\; n\int K(u)\,\Big(\phi(x+hu) - \phi(x)\Big)^j\,f(x+hu)\,du \\ &=\; nh^j\int K(u)\,u^j\,\varphi^j(\zeta_u)\,f(x+hu)\,du \\ &=\; nh^j\Big(f(x)\,\varphi^j(x)\,\mu_j + o(1)\Big) \end{aligned}$$

where $\zeta_u \in (x, x+hu)$ exists according to Taylor's theorem. Similar we derive

$$\begin{aligned}
\operatorname{Var} S_{n,j} &= nE\Big(w_1^2\big(\phi(X_1)-\phi(x)\big)^{2j}\Big) - nE^2\Big(w_1\big(\phi(X_1)-\phi(x)\big)^{j}\Big) \\
&= nh^{2j-1}\Big(f(x)\,\varphi^{2j}(x)\,\nu_{2j} + o(1)\Big) \\
&= n^2h^{2j}O\left(\frac{1}{nh}\right) \qquad \text{(A.2)} \\
&= o\big(n^2h^{2j}\big)\ .
\end{aligned}$$

Since for every sequence $(Y_n)_{n\in\mathbb{N}}$ of random variables

$$Y_n = EY_n + O_P\left(\sqrt{\operatorname{Var} Y_n}\right) \qquad \text{(A.3)}$$

holds (what can be proven with Chebychev's inequality), we can proceed with calculating

$$\begin{aligned}
S_{n,j} &= ES_{n_j} + O_P\left(\sqrt{\operatorname{Var} S_{n,j}}\right) \\
&= nh^j f(x)\,\varphi^j(x)\,\mu_j(1+o_P(1)) \qquad \text{(A.4)}
\end{aligned}$$

which leads to

$$\mathbf{S_n} = nf(x)\,\mathbf{P}\,\mathbf{H}\,\mathbf{S}\,\mathbf{H}\,\mathbf{P}\,(1+o_P(1))\ . \qquad \text{(A.5)}$$

In the same manner, we find that

$$\begin{aligned}
S^*_{n,j} &= ES^*_{n_j} + O_P\left(\sqrt{\operatorname{Var} S^*_{n,j}}\right) \\
&= nh^{j-1}\Big(\varphi^j(x)\,\sigma^2(x)f(x)\,\nu_j + o(1)\Big) + O_P\left(\sqrt{o(n^2h^{2j-2})}\right) \\
&= nh^{j-1}\,\varphi^j(x)\,\sigma^2(x)f(x)\,\nu_j(1+o_P(1))
\end{aligned}$$

and thus

$$\mathbf{S^*_n} = \frac{n}{h}\,f(x)\,\sigma^2(x)\,\mathbf{P}\,\mathbf{H}\,\mathbf{S}^*\mathbf{H}\,\mathbf{P}\,(1+o_P(1)) \qquad \text{(A.6)}$$

and finally assertion (3.1) by plugging (A.5) and (A.6) into (A.1).

II. Asymptotic conditional bias

Finding an asymptotic expression for

$$\mathit{Bias}(\hat\gamma|\mathbb{X}) = \mathbf{S_n}^{-1}\,\mathbf{X}^T\,\mathbf{W}\mathbf{r} \qquad \text{(A.7)}$$

still requires to approximate $\mathbf{r} \equiv (r_i)_{1\le i\le n}$. Let $D_K(x)$ be the set of all data points within the kernel support. For all $i \in D_K(x)$ we obtain

$$\begin{aligned}
r_i &= m(X_i) - \sum_{j=0}^{p}\gamma_j\big(\phi(X_i)-\phi(x)\big)^j \\
&= \frac{\psi_{(p+1)}(\zeta_i)}{(p+1)!}\,\big(\phi(X_i)-\phi(x)\big)^{p+1} \\
&= \gamma_{p+1}(x)\,\big(\phi(X_i)-\phi(x)\big)^{p+1} + o_P(1)\,\frac{\big(\phi(X_i)-\phi(x)\big)^{p+1}}{(p+1)!}
\end{aligned}$$

where $\zeta_i \in (X_i, x)$ resp. (x, X_i) exists according to Theorem 2.1, and the term $o_P(1)$ is uniform over $D_K(x)$. Note that the invertibility demanded for $\phi(\cdot)$ in Theorem 2.1 is already guaranteed locally around x by the condition $\varphi(x) \neq 0$. Finally we calculate

$$\begin{aligned}
Bias(\hat{\gamma}|\mathbb{X}) &= \mathbf{S_n}^{-1}\,\mathbf{X}^T\,\mathbf{W}\Big[\big(\phi(X_i)-\phi(x)\big)^{p+1}\big(\gamma_{p+1}+o_P(1)\big)\Big]_{1\le i\le n} \\
&= \mathbf{S_n}^{-1}\,\mathbf{c_n}\big(\gamma_{p+1}+o_P(1)\big) \\
&= \mathbf{P}^{-1}\mathbf{H}^{-1}\mathbf{S}^{-1}\mathbf{H}^{-1}\mathbf{P}^{-1}\frac{1}{nf(x)}\left\{\gamma_{p+1}\mathbf{c_n}+\begin{pmatrix} o(nh^{p+1}) \\ \vdots \\ o(nh^{2p+1})\end{pmatrix}\right\}\big(1+o_P(1)\big) \\
&= \mathbf{P}^{-1}\mathbf{H}^{-1}\mathbf{S}^{-1}h^{p+1}\,\varphi^{p+1}(x)\,\gamma_{p+1}\,\mathbf{c_p}\,\big(1+o_P(1)\big)\ ,
\end{aligned}$$

by substituting the asymptotic expressions for $S_{n,j}$ (A.4) in $\mathbf{c_n} := (S_{n,p+1},...,S_{n,2p+1})^T$, and thus (3.2) is proven.

Now we proceed to the derivation of $\mathbf{b}_n$ which requires to take along some extra terms resulting from higher order expansions. With $(a+hb)^j = a^j + h(ja^{j-1}b + o(1))$ we find that

$$\begin{aligned}
ES_{n,j} &= nh^j\int K(u)\,u^j\left(\varphi(x)+\frac{hu}{2}\,\varphi'(\zeta_u)\right)^j\Big(f(x)+huf'(\xi_u)\Big)\,du \\
&= nh^j\int K(u)\,u^j\left[\varphi^j(x)+h\left(\frac{j}{2}\,\varphi^{j-1}(x)\,u\,\varphi'(\zeta_u)+o(1)\right)\right]\Big(f(x)+huf'(\xi_u)\Big)\,du \\
&= nh^j\left[f(x)\,\varphi^j(x)\,\mu_j + h\left(f'(x)\,\varphi^j(x)+\frac{f(x)}{2}\,j\,\varphi^{j-1}(x)\,\varphi'(x)\right)\mu_{j+1}+o(h)\right]
\end{aligned}$$

(A.8)

with ζ_u and ξ_u according to Taylor's theorem. Plugging (A.8) and (A.2) into (A.3) yields

$$S_{n,j} \;=\; nh^j\varphi^j(x)\left[f(x)\mu_j + h\left(f'(x)+\frac{f(x)}{2}\,\frac{\varphi'(x)}{\varphi(x)}\,j\right)\mu_{j+1}+o_n\right], \tag{A.9}$$

where $o_n = o_P(h) + O_P\left(\frac{1}{\sqrt{nh}}\right) = o_P(h)$ from the hypothesis $nh^3 \to \infty$, and further

$$\mathbf{S_n} \;=\; n\mathbf{PH}\left(f(x)\mathbf{S}+hf'(x)\tilde{\mathbf{S}}+h\frac{f(x)}{2}\,\frac{\varphi'(x)}{\varphi(x)}\,\bar{\mathbf{S}}+o_P(h)\right)\mathbf{HP}\ . \tag{A.10}$$

The next task is to derive a higher order expansion for $\mathbf{r}$. With Theorem 2.1 we obtain

$$\begin{aligned}
r_i &= \frac{\psi_{(p+1)}(x)}{(p+1)!}\big(\phi(X_i)-\phi(x)\big)^{p+1}+\frac{\psi_{(p+2)}(\zeta_i)}{(p+2)!}\big(\phi(X_i)-\phi(x)\big)^{p+2} \\
&= \gamma_{p+1}\big(\phi(X_i)-\phi(x)\big)^{p+1}+\gamma_{p+2}\big(\phi(X_i)-\phi(x)\big)^{p+2} \\
&\quad + \big(\psi_{(p+2)}(\zeta_i)-\psi_{(p+2)}(x)\big)\,\frac{(\phi(X_i)-\phi(x))^{p+2}}{(p+2)!} \\
&= \big(\phi(X_i)-\phi(x)\big)^{p+1}\,\gamma_{p+1}+\big(\phi(X_i)-\phi(x)\big)^{p+2}\big(\gamma_{p+2}+o_P(1)\big)
\end{aligned}$$

with $\zeta_i \in (X_i, x)$ resp. (x, X_i). Plugging this and (A.10) into (A.7) and denoting

$$\mathbf{T_n} := f(x)\mathbf{S} + h\left(f'(x)\tilde{\mathbf{S}} + \frac{f(x)}{2}\frac{\varphi'(x)}{\varphi(x)}\bar{\mathbf{S}}\right) + o_P(h)$$

leads to

$$\begin{aligned} Bias(\hat{\gamma}|\mathbb{X}) &= [n\,\mathbf{PHT_nHP}]^{-1}\Big[\mathbf{c_n}\gamma_{p+1} + \tilde{\mathbf{c}}_\mathbf{n}(\gamma_{p+2} + o_P(1))\Big] \\ &= \mathbf{P}^{-1}\mathbf{H}^{-1}{\mathbf{T_n}}^{-1}h^{p+1}\varphi^{p+1}(x)\cdot \\ &\quad \cdot \Big[\gamma_{p+1}f(x)\,\mathbf{c_p} + h\Big(\gamma_{p+1}f'(x) + \gamma_{p+2}\,\varphi(x)f(x)\Big)\,\tilde{\mathbf{c}}_\mathbf{p} \\ &\quad + h\,\gamma_{p+1}f(x)\,\frac{\varphi'(x)}{2\,\varphi(x)}\,\bar{\mathbf{c}}_\mathbf{p} + o_P(h)\Big]\,, \end{aligned}$$

where the asymptotic expressions (A.9) are substituted in $\mathbf{c_n}$ and $\tilde{\mathbf{c}}_\mathbf{n} = (S_{n,p+2}, ..., S_{n,2p+2})^T$. The matrix $\mathbf{T_n}$ still has to be inverted. Applying the formula

$$(\mathbf{A} + h\mathbf{B})^{-1} = \mathbf{A}^{-1} - h\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} + O(h^2)$$

yields

$${\mathbf{T_n}}^{-1} = \frac{1}{f(x)}\mathbf{S}^{-1} - h\frac{1}{f(x)}\mathbf{S}^{-1}\left(\frac{f'(x)}{f(x)}\tilde{\mathbf{S}} - \frac{\varphi'(x)}{2\varphi(x)}\bar{\mathbf{S}}\right)\mathbf{S}^{-1} + o_P(h)\,,$$

and we obtain finally

$$\begin{aligned} Bias(\hat{\gamma}|\mathbb{X}) = h^{p+1}\varphi^{p+1}(x)\,\mathbf{P}^{-1}\mathbf{H}^{-1}\cdot \\ \cdot \Bigg\{\gamma_{p+1}\mathbf{S}^{-1}\mathbf{c_p} + h\Bigg[\left(\gamma_{p+1}\frac{f'(x)}{f(x)} + \gamma_{p+2}\,\varphi(x)\right)\mathbf{S}^{-1}\tilde{\mathbf{c}}_\mathbf{p} \\ + \gamma_{p+1}\,\frac{\varphi'(x)}{2\varphi(x)}\mathbf{S}^{-1}\bar{\mathbf{c}}_\mathbf{p} + \gamma_{p+1}\mathbf{S}^{-1}\left(\frac{f'(x)}{f(x)}\tilde{\mathbf{S}} - \frac{\varphi'(x)}{2\varphi(x)}\bar{\mathbf{S}}\right)\mathbf{S}^{-1}\mathbf{c_p}\Bigg] + o_P(h)\Bigg\}\,. \end{aligned}$$

ACKNOWLEDGMENTS

The author is grateful to Gerhard Tutz, Torsten Scholz and Klaus Hechenbichler, University of Munich, for the fruitful inspirations and discussions during this work, which led to essential impacts on the development of this paper. Moreover, the author thanks two anonymous referees for a lot of helpful comments, which improved the paper substantially. The support by Sonderforschungsbereich 386 (DFG) in various aspects is gratefully acknowledged.

REFERENCES

[1] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association*, **74**, 829–836.

[2] Doksum, K.; Petersen, D. and Samarov, A. (2000). On variable bandwidth selection in local polynomial regression, *Journal of the Royal Statistical Society, Series B*, **62**, 431–448.

[3] Einbeck, J. (2003). Multivariate local fitting with general basis functions, *Computational Statistics*, **18**, 185–203.

[4] Fan, J. (1992). Design-adaptive nonparametric regression, *Journal of the American Statistical Association*, **87**, 998–1004.

[5] Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers, *Annals of Statistics*, **20**, 2008-2036.

[6] Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaption, *Journal of the Royal Statistical Society, Series B*, **57**, 371–395.

[7] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*, Chapman and Hall, London.

[8] Fan, J.; Hu, T.-C. and Truong, Y.K. (1994). Robust nonparametric function estimation, *Scandinavian Journal of Statistics* **21**, 433–446.

[9] Hart, J.D. and Yi, S. (1998). One-sided cross-validation, *Journal of the American Statistical Association*, **93**, 620–631.

[10] Hastie, T. and Loader, C. (1993). Rejoinder to: "Local regression: Automatic kernel carpentry", *Statistical Science*, **8**, 139–143.

[11] Hurvich, C.M.; Simonoff, J.S. and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society, Series B*, **60**, 271–293.

[12] Lay, S.R., (1990). *Analysis with an Introduction to Proof*, Prentice Hall, New Jersey.

[13] Loader, C.R. (1999). Bandwidth selection: Classical or plug-in?, *Annals of Statistics*, **27**, 415–438.

[14] Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis*, Springer, New York.

[15] Ruppert, D.; Sheather, S.J. and Wand, M.P. (1995). An effective bandwidth selector for local least squares regression, *Journal of the American Statistical Association*, **90**, 1257–1270.

[16] Ruppert, D. and Wand, M.P. (1994). Multivariate locally weighted least squares regression, *Annals of Statistics*, **22**, 1346-1370.

[17] Schmidt, G.; Mattern, R. and Schüler, F. (1981). Biochemical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column and without protective helmet under effects of impact, Technical Report Project 65, EEC Research Program on Biomechanics of Impacts, University of Heidelberg, Germany.

[18] Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection, *Journal of Econometrics*, **75**, 317–434.

[19] Stone, C.J. (1977). Consistent nonparametric regression, *Annals of Statistics*, **5**, 595–645.

[20] Tsybakov, A.B. (1986). Robust reconstruction of functions by the local approximation approach, *Problems of Information Transmission*, **22**, 133–146.

[21] Yu, K. and Jones, M.C. (1997). A comparison of local constant and local linear regression quantile estimators, *Computational Statistics and Data Analysis*, **25**, 159–166.

[22] Yu, K. and Jones, M.C. (1998). Local linear quantile regression, *Journal of the American Statistical Association*, **93**, 228–237.

REVSTAT – Statistical Journal
Volume 2, Number 2, November 2004

CENTRAL PARTITION FOR A PARTITION-DISTANCE AND STRONG PATTERN GRAPH

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

P.R. Rao
– Department of Computer Science and Technology, Goa University, India (pralhaad@rediffmail.com)

Received: January 2003 Revised: October 2004 Accepted: October 2004

Abstract:

- When several clustering algorithms are applied to a dataset E or the same algorithm with different parameters, we get several different partitions of the dataset. In this paper we consider the problem of finding a consensus partition between the set of these partitions. This consensus partition, called *central partition*, minimises the average number of disagreements between all of the partitions and has been considered for instance in [14, 5] in a different context from ours. We consider it in the context of partition-distance defined in [7]. We focus our attention in two particular distance functions between partitions and then do an experimental comparison between the two corresponding central partitions. In addition, by using the concept of strong patterns (maximal subset of elements that are always clustered together in all partitions), we define a new graph where the nodes are the strong patterns. This graph contains essentially the same information as the partition graph corresponding to the set E defined in [7], but is much simpler as the number of strong patterns is expected to be much smaller than the cardinal of E. Then, some properties of this new graph are proved.

Key-Words:

- *clustering; graph algorithms; node cover; combinatorial problems; strong pattern; central partition.*

AMS Subject Classification:

- 62-07, 62H30, 94C15.

1. INTRODUCTION

The concept of similarity between two partitions arises in several applications, such as molecular expression data in computational biology. When several different clustering methods are applied to the same data, or the same algorithm with different parameters, different partitions of the same data are produced. Also, if we have K qualitative variables describing our population, we might want to find a "central variable" which sumarizes these variables. These two problems are the same, because there is a one-to-one correspondence between qualitative variables and partitions. The problem of determining a central partition arises also in the case where the given partitions (qualitative variables) result from measurements at times $t, t+1, ..., t+K-1$ and we want to consider the notion of a moving consensus smoothing the partitions (or qualitative variables) at those times.

According to Barthelemy and Leclerc [2], there are three overlapping approaches that have been used to tackle the consensus problem:

(**i**) the *axiomatic approach*, where a central partition must satisfy some conditions that arise, for instance, from experimental evidence;

(**ii**) the *constructive approach*, where a way to construct the consensus is explicitly given, like the Pareto rule which states that two objects are linked in a consensus partition if and only if they are linked in all the K given partitions;

(**iii**) the *combinatorial optimization problem*, where we have some criterion measuring the *remoteness* (see equation (2.1)) of any partition to the given K partitions and we search for a partition that minimises this remoteness function.

This last approach, which goes back to Régnier [14], is the one we use in this work.

In order to find the best consensus, it becomes necessary to evaluate the closeness of the partitions produced. There are many distances that can be defined between two partitions of a dataset. The partition-distance is one such distance measure. This concept has been defined in [1], although Régnier [14] and Lerman (see p. 51 of [9]) had considered it before. This distance is further studied in [7], in which it is shown that the partition-distance between two partitions on a given set can be computed in polynomial time.

Further in [7], a new class of graphs called partition graphs has been defined. It is proved that the partition-distance between two partitions is equal to the size of the smallest node cover of the corresponding partition graph. By establishing the arrayed layout structure of the partition graph, it is shown in [7], that the partition graph is perfect.

Suppose $K \geq 2$ partitions of a nonempty set E consisting of n elements are given. In this paper, we define the notion of central partition with respect to the partition-distance used in [7]. The concept of central partition has been used in [5] in another context and with respect to a different measure of distance between partitions. The central partition is a partition that represents a consensus between all the initial K partitions obtained by different clustering algorithms or by the same algorithm with different parameters.

The computation of the central partition is hard. Hence, we have used an approximate algorithm (heuristic), described in [5], to compute an approximation to the central partition. In order to do this, we use the concept of strong patterns. A strong pattern is a maximal subset of elements of E that have been always clustered together in all of the K partitions. The heuristic consists in assuming that these elements should also be together in the central partition. In addition, by using this concept of strong patterns, we can define a graph where the nodes are the strong patterns, which contains essentially the same information as the partition graph corresponding to the K partitions, but is much simpler as the number of strong patterns is expected to be much smaller than n. The complexity is therefore dominated by determining the strong patterns.

The main goal of our work is first to make a summary of the works that have been done in the problem of consensus partitions. Then, the distance used in [5] and the partition-distance are compared using graph terminology. An experimental evaluation of the central partitions corresponding to these two distances is also presented. Next, a special graph, the strong pattern graph, is defined and some of its properties are given.

2. RELATED WORK

Suppose that we have K qualitative variables describing our set of objects E. Each such variable defines a partition of the set E. We can associate an equivalence relation on E with each variable: x and y are in the same equivalence class if the values of this variable are the same for x and y. Thus we obtain K equivalence relations on $E: R_1, R_2, ..., R_K$. In 1965 Régnier [14] proposes as a good clustering of E, a partition whose associated equivalence E_p minimises the quantity

$$\sum_{i=1}^{K} \delta(E_p, R_i) \ , \tag{2.1}$$

which is called a remoteness function. $\delta(R, E_p) = |R \cup E_p| - |R \cap E_p| = |R - E_p| + |E_p - R|$ is the number of non ordered pairs of points that are in the same cluster in one partition but not in the other. The partition which minimises equation (2.1) is called central partition by Régnier.

In 1981, Barthelemy and Monjardet [3] use the notion of median in order to unify the treatment of some problems which are based on the minimization of a remoteness function, like for instance aggregation problems in cluster analysis, social choice theory and paired comparisons methods. We will restrict and adapt the presentation of their median procedure to the case of clustering. These authors start by defining the partitions π_α (resp. π_β) to be such that two elements x and y are in the same cluster for this partition iff they are together in the same cluster for at least $K/2+1$ (resp. $K/2+0.5$) of the initial partitions. One can easily see that $\pi_\alpha \leq \pi_\beta$, which means that any cluster of π_α is included in a cluster of π_β. The authors define then the median interval of the K initial partitions to be $[\pi_\alpha, \pi_\beta]$. If K is odd, then $\pi_\alpha = \pi_\beta$ and so there is only one median partition; otherwise, every partition contained in that interval is a median partition. Barthelemy and Monjardet [3] then present some properties of this median procedure and survey some interesting mathematical problems related to the notion of median. In a later paper, Barthelemy and Leclerc [2], concentrate on the problem of finding a consensus partition that summarizes a K-tuple of partitions by using the median procedure. A detailed survey of the median procedure for partitions is given, from the axiomatic and the algorithmic points of view.

William Day [6] describes two models for the enumeration of metrics between partitions, focusing on the complexity of computing these metric distances. By doing so he rediscovers some metrics that already existed in the literature, but discovers some new metrics also. For some of them, there exist efficient algorithms with time complexities ranging from $\mathbf{O}(n)$ to $\mathbf{O}(n^3)$.

Strehl and Ghosh [15] propose three techniques for obtaining high-quality consensus partitions. The first one uses a similarity measure which is based on the given K initial partitions and then reclusters the objects using this new similarity measure. The second technique is based on hypergraph partitioning and the third technique collapses groups of clusters into meta-clusters which then compete for each object to determine the central partition. These authors claim that their techniques have low computational costs and so suggest further to use the three approaches for a given situation and then choose the best solution.

Monti *et al.* [10] use a resampling-based method to find the central (consensus) partition in the context of gene-expression microarray data. This type of data has the particularity of presenting many more variables (genes) than observations, which is a challenge for classical data analysis methods (see for instance [11]). Monti *et al.* [10] call their methodology consensus clustering which provides for a method to represent the consensus across multiple runs of a clustering algorithm and to assess the stability of the discovered clusters. They also provide a visualization tool to inspect and validate the number of clusters, membership and boundaries.

3. TWO DISTANCES BETWEEN PARTITIONS BASED ON THE PARTITION-GRAPH

Let E be a nonempty set consisting of n elements. A cluster of E is a nonempty subset of E. A partition of E is a collection of mutually exclusive clusters of E, whose union is E. Two partitions π and π' of E are identical if and only if every cluster in π is also a cluster in π'.

Given two partitions π and π', the *partition-distance*, $D_p(\pi, \pi')$, between π and π' is the minimum number of elements that must be removed from E such that the two induced partitions (π and π' restricted to the remaining elements) are identical.

In [7] this definition is extended to the case of $K > 2$ partitions. Also in [7], it is written that the partition-distance is equal to the minimum number of elements that must be moved between clusters in π, so that the resulting partition equals π'. This definition had already appeared before in the work of Régnier [14].

Example 3.1. Let $E = \{1, 2, 3, 4, 5, 6\}$. Consider the following partitions, π and π' of E:

$$\pi = \Big\{\{1, 2, 4, 6\},\, \{3, 5\}\Big\}\ , \qquad \pi' = \Big\{\{1, 2, 6\},\, \{3\},\, \{4, 5\}\Big\}\ ;$$

then the partition-distance between π and π' equals two, as the removal of two elements, namely 3 and 4, will make π and π' identical and no single element of E has this property.

Proposition 3.1. *The partition-distance, $D_p(\pi, \pi')$, between π and π' verifies the properties of a distance function.*

Proof: The first three properties are obvious. In fact, (**i**) $D_p(\pi, \pi) = 0$; (**ii**) $D_p(\pi, \pi') = D_p(\pi', \pi)$; (**iii**) $D_p(\pi, \pi') = 0 \ \Rightarrow\ \pi = \pi'$.

As for the triangular inequality, (**iv**) $D_p(\pi, \pi') \leq D_p(\pi, \pi'') + D_p(\pi'', \pi')$, let us start by denoting $D_p(\pi, \pi') = n_1$, $D_p(\pi, \pi'') = n_2$ and $D_p(\pi'', \pi') = n_3$. Suppose that $n_1 > n_2 + n_3$. If we remove n_2 elements from E, the two induced partitions of π and π'' become identical; the same happens between π'' and π' if we remove a certain set of n_3 elements. This means that if we remove at most $n_2 + n_3$ (corresponding to the union of the two previous sets to be removed) elements from E, the three induced partitions of π, π'' and π' become identical. This is absurd since by hypothesis, we need to remove at least n_1 elements, which is more than $n_2 + n_3$, in order to make the two induced partitions of π and π' identical. Therefore, we can not have $n_1 > n_2 + n_3$.

Given two partitions π and π' of the same set E, consider the graph $G(\pi, \pi')$ with one node for each element of the set E; two nodes are adjacent iff they are together in the same cluster of either π or π', but not in both. $G(\pi, \pi')$ is called a *partition graph* (see [7]). A *node-cover* of a graph is a subset of nodes Q such that every edge in the graph is incident with at least one node in Q.

As it is shown in [7], the partition-distance between two partitions π and π' is equal to the size of the smallest node-cover of the graph $G(\pi, \pi')$ (it has not been proved that the smallest node cover is unique). This means that the set of elements that must be removed so that the two induced partitions become identical is one of the smallest node covers. The distance used in [5] has also an interpretation in terms of this graph. For each partition π_l let v_l represent its associated equivalence relation: $v_l(i, i') = 1$ iff the two elements are in the same cluster. Then, the distance used in [5] is

$$D_C(\pi, \pi') = \frac{1}{2} \sum_{i,i' \in E} \left| v(i, i') - w(i, i') \right|$$

where the equivalence relations v and w correspond to the partitions π and π' respectively. It is easy to see that this distance is equal to the number of edges of the partition graph $G(\pi, \pi')$. □

4. THE CENTRAL PARTITION FOR A PARTITION-DISTANCE

In this section we start by defining the concept of strong pattern. Given K partitions of a dataset E, a strong pattern is a maximal subset of elements of E that have been always clustered together in all of the K partitions.

Now, in order to determine the strong patterns, we start by building a matrix R with n rows and K columns, where each column represents a partition. So, for instance, if the first partition has 5 clusters, the first column of R is composed of a sequence of numbers belonging to the set $\{1, 2, 3, 4, 5\}$. Thus, the element R_{ij} of this matrix is the cluster number attributed by partition π_j to the i^{th} observation.

From R we construct a square matrix A, of size n, such that $A_{ii'}$ is equal to the number of times that the objects i and i' are clustered together in the K partitions. The complexity of building the matrix A is therefore $n(n-1) \times K/2$.

Consider now the equivalence relation

$$\forall\, (i, i') \in E \times E\ , \qquad w^K(i, i') = \begin{cases} 1 & \text{if } A_{ii'} = K \\ 0 & \text{otherwise .} \end{cases}$$

The partition of strong patterns corresponds to this equivalence relation. To find this partition we look at the elements of matrix A row by row, starting with the first row. First, to the first element is attributed the first cluster, which we can call cluster 1. Then, in the first row, everytime we find that $A_{1i'} = K$, we put the element i' in cluster 1 also, and we delete the row corresponding to i' from consideration. Then we go to the next row to be considered, and we do the same, this time attributing its elements to cluster 2. We proceed in the same manner until there are no more rows to be considered. The complexity of this step is at most $n(n-1)/2$. Therefore the complexity for determining the strong patterns is $O(n^2K)$.

Suppose we have K partitions of E, $(\pi^1, \pi^2, ..., \pi^K)$. We are going to consider now how to obtain from these K partitions a new partition which best represents a consensus between all of the initial K partitions. We call it Central Partition. First of all, the partition corresponding to the strong patterns represents an unanimous consensus between all the K partitions; nevertheless, it usually cannot be considered as a central partition because it has got too many clusters (strong patterns) and is therefore too refined.

Let us denote by π^* the central partition that we are looking for. We define the central partition as the one that minimises the following criterium:

$$C(\pi^*) = \sum_{k=1}^{K} D_p(\pi^*, \pi^k)$$

where $D_p(\pi^*, \pi^k)$ is the partition-distance between the partitions π^* and π^k, that is, the number of elements that have to be removed so that the two induced partitions become identical. Intuitively, the central partition minimises the average number of disagreements between the K partitions. The problem of finding π^* is NP-hard and so we are going to use an heuristic to find an approximation of it. This heuristic has already been used and justified in [5]; we will adapt it to our context. In [5], the distance between two partitions, $D_C(\pi, \pi')$, is equal to the number of edges of the partition graph $G(\pi, \pi')$ that has been defined in [7]. In our case we use the partition-distance, $D_p(\pi, \pi')$.

Let us denote by S the set of strong patterns and q ($q \ll n$) its cardinality. We define now a square matrix B of size q such that $B_{pp'}$ is the number of times that the strong patterns p and p' are together in all of the K partitions.

Theoretically, the partition corresponding to the strong patterns is associated with an equivalence relation u^K:

$$\forall\,(p,p') \in S \times S\ , \qquad u^K(p,p') = \begin{cases} 1 & \text{if } B_{pp'} = K \\ 0 & \text{otherwise .} \end{cases}$$

In a similar way, other relations u^j, $j = 0, 1, ..., K-1$, can be defined:

$$\forall\,(p,p') \in S \times S\ , \qquad u^j(p,p') = \begin{cases} 1 & \text{if } B_{pp'} \geq j \\ 0 & \text{otherwise .} \end{cases}$$

These relations u^j, are in general not transitive and so cannot represent a partition. Only u^0 and u^K represent partitions. To u^0 is associated the elementary partition, where there is only one cluster; to u^K is associated the partition of strong patterns. For $j = 1, ..., K-1$, u^j does not represent a partition, because it is generally not transitive, and the authors in [5] associate with each u^j an equivalence relation $\overline{u}^j$, which is the transitive closure of u^j. Let Γ^j represent the partition associated with $\overline{u}^j$. Let Γ^0 represent the partition with only one cluster and Γ^K the partition of strong patterns. It is then shown that the partitions $\Gamma^0, \Gamma^1, \Gamma^2, ..., \Gamma^K$ are nested, that is, Γ^j is obtained from Γ^{j+1}, by merging two of its clusters.

The heuristic that is then used in order to find the approximate central partition consists in restraining the search to the partitions Γ^j. Each such partition is composed of clusters of strong patterns. In 1984 Celeux [4] has shown that in practice the approximate central partitions obtained by this heuristic are the same or very close to the exact central partition. That is, the clusters corresponding to both partitions, the exact and the one found using the heuristic, are similar.

Let S be the set of strong patterns and define the distance index

$$d(p, p') = K - B_{pp'} \ , \qquad \forall\, (p, p') \in S \ .$$

Let us now prove that this measure is really a distance index. In fact, (**i**) $d(p,p) = 0$ because $B_{pp} = K$. Next, (**ii**) $d(p,p') = d(p',p)$ because the matrix B is symetric. Now, if (**iii**) $d(p,p') = 0$, we have $B_{pp'} = K$; this only happens if the two strong patterns p and p' are in fact one, that is, $p = p'$.

Using this distance index, we build a matrix of distance indices between the strong patterns. The partitions Γ^j can be obtained in the following manner [5]. Start by building a minimal spanning tree (MST) containing q nodes (the strong patterns) and using the distance index $d(p, p') = K - B_{pp'}$ defined above. The edge joining two adjacent nodes p and p' has weight $d(p, p')$. Now, in order to determine the candidate central partitions, $\Gamma^0, \Gamma^1, ..., \Gamma^K$, we do the following: Γ^0 has just one cluster. Γ^1 is obtained from the MST by removing the edge of maximum weight and writing down the two obtained clusters. We continue by successively removing the edges of maximum weight, obtaining the other candidate central partitions $\Gamma^2, \Gamma^3, ..., \Gamma^K$. Everytime that we find two or more edges with maximum weight, we remove all of these at once. Celeux *et al.* [5] show that the candidate central partitions obtained by this methodology are the same defined above associated with $\overline{u}^j$.

For each candidate central partition, Γ^j, we compute the criterium defined above, that is,

$$C(\Gamma^j) \ = \ \sum_{k=1}^{K} D_p(\Gamma^j, \pi^k) \ ,$$

and we choose the partition which minimises this criterium. So, the central partition obtained is the one which minimises the sum of all the partition-distances

between the central partition and the initial K partitions.

Example 4.1. Let $E = \{1, 2, 3, 4, 5, 6\}$ and consider the following four partitions:

$$\pi^1 = \Big\{\{1,2\},\, \{3,4\},\, \{5\},\, \{6\}\Big\}\,, \qquad \pi^2 = \Big\{\{1,2,4\},\, \{3,5\},\, \{6\}\Big\}\,,$$
$$\pi^3 = \Big\{\{1,2,6\},\, \{3,4\},\, \{5\}\Big\} \quad \text{and} \quad \pi^4 = \Big\{\{1,2,5\},\, \{4,6\},\, \{3\}\Big\}\,.$$

This is a very small example with quite different partitions, but it serves to illustrate the determination of central partition. The strong patterns are therefore the subsets $\{1,2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$.

The symetric matrix B is:

	$\{1,2\}$	$\{3\}$	$\{4\}$	$\{5\}$	$\{6\}$
$\{1,2\}$	4	0	1	1	1
$\{3\}$		4	2	1	0
$\{4\}$			4	0	1
$\{5\}$				4	0
$\{6\}$					4

From B we construct the matrix of distance indices $d(p,p') = K - B_{pp'}$:

	$\{1,2\}$	$\{3\}$	$\{4\}$	$\{5\}$	$\{6\}$
$\{1,2\}$	0	4	3	3	3
$\{3\}$		0	2	3	4
$\{4\}$			0	4	3
$\{5\}$				0	4
$\{6\}$					0

Now, we build the minimal spanning tree (MST) between the strong patterns using for instance Prim's algorithm (see Figure 1).

Then, by starting to remove the edges of maximal weight, we get three candidate central partitions. Whenever two or more edges have maximum weight, we remove all of them at once.

The candidate central partitions are therefore:

$$\Gamma^0 = \Big\{1,\, 2,\, 3,\, 4,\, 5,\, 6\Big\}\,,$$
$$\Gamma^1 = \Big\{\{1,2\},\, \{3,4\},\, \{5\},\, \{6\}\Big\}\,,$$
$$\Gamma^2 = \Big\{\{1,2\},\, \{3\},\, \{4\},\, \{5\},\, \{6\}\Big\}\,.$$

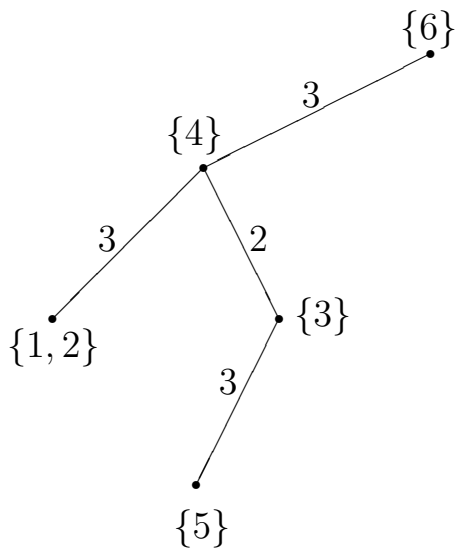

Figure 1: One possible MST between the strong patterns.

Now, in order to choose one of these three candidate central partitions as a central partition, we need to compute the value of $C(\Gamma^j),\ j = 0, 1, 2$; we will do this using the partition-distance defined above:

$$\begin{aligned} C(\Gamma^0) &= 4+3+3+3 = 13\ , \\ C(\Gamma^1) &= 0+2+1+2 = 5\ , \\ C(\Gamma^2) &= 1+2+2+2 = 7\ . \end{aligned}$$

The final partition chosen, that is the one which minimises the criterium $C(\Gamma^j)$, is the partition $\big\{\{1,2\},\{3,4\},\{5\},\{6\}\big\}$, which, in this case, coincides with one of the initial partitions.

5. EXPERIMENTAL COMPARISON BETWEEN THE TWO CENTRAL PARTITIONS

In this section we will show the results of some experiences in order to compare the two central partitions corresponding to the partition-distance used in [7], $D_p(\pi, \pi')$, and the distance used in [5], $D_C(\pi, \pi')$. As was shown above, the partition-distance between two partitions π and π' is equal to the size of the smallest node-cover of the graph $G(\pi, \pi')$ and the distance used in [5] corresponds to the number of edges of $G(\pi, \pi')$. Since the first of these two distances is more complicated to compute, it is of interest to know if the corresponding central partition represents a better consensus between the initial K partitions; otherwise it would be better to use the other distance. To see which of the two central partitions represents a better consensus, we use the Rand index [13], which was

latter corrected for chance in [8]. We start by computing the value of this index between the central partition and each of the initial K partitions and then find the average. The formula for the corrected Rand index between two partitions, one with L clusters and the other with C clusters, is

$$CRI = \frac{\sum_{i=1}^{L}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{L}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{L}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}\right] - \binom{n}{2}^{-1}\sum_{i=1}^{L}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}} \tag{5.1}$$

where n is the total number of objects, n_{ij} denotes the number of objects that are common to clusters u_i and v_j, $n_{i.}$ and $n_{.j}$ referring respectively to the number of objects in clusters u_i and v_j. This index takes values in the interval $[-1, 1]$ where the value 1 indicates a perfect agreement between the partitions, whereas values close to 0 correspond to cluster agreement found by chance.

We start by generating 19 random partitions of a dataset with 600 elements, with different numbers of clusters in each partition. We do not take into account the structure of the dataset underlying those partitions. In fact, the partitions were obtained by simulating an integer vector of size 600, where each component of this vector contains the cluster number attributed to the i^{th} element, $i = 1, 2, ..., 600$. This is because the two central partitions considered in this work only take into account the labels associated to each element of the dataset; that is, its cluster number, regardless of the structure of the dataset. So, the aim of this experiment is just to see which central partition best agrees with the initial partitions. Our aim is not to see if the initial partitions are a good clustering of any dataset. We suppose we are given K initial partitions and we want just to find the best possible consensus between them.

To generate the random partitions, we have used the code in [12], where it is also explained how the random partitions are generated. Then, we have written a program to compute the two central partitions. Let π^{*1} denote the central partition using the partition-distance $D_p(\pi, \pi')$ and π^{*2} the central partition using the distance $D_C(\pi, \pi')$. Now, we compute the corrected Rand index between each central partition and the initial 19 partitions and find the average. This procedure was repeated six times and the results are given in Table 1.

Table 1: *CRI* values for the two central partitions.

Dataset	Values relating to π^{*1}	Values relating to π^{*2}
1	.450616	.349814
2	.370913	.220207
3	.434782	.353463
4	.401694	.222835
5	.355193	.239976
6	.360283	.278106

As can be seen from these results, the central partition corresponding to the partition-distance presents higher *CRI* values, indicating therefore greater average similarity with the initial 19 partitions.

We have performed another controlled experiment that allows us to compare the two central partitions in the presence of noise. First, we partition a set with 500 elements into 10 clusters at random, as we did above, to obtain the original clustering. We duplicate this clustering 10 times, but, in each of these new 10 labelings, a fraction of the labels is replaced with random labels from a uniform distribution from 1 to 10 (number of clusters). Then, we find the two central partitions, π^{*1} and π^{*2}, for these 10 noisy partitions, and we compare each central partition with the initial partition which has no noise. The results, which are given in Table 2, contain the *CRI* values between π^{*1} and the initial partition, the average *CRI* values between π^{*1} and the given 10 partitions; and the same for π^{*2}.

Table 2: *CRI* values for the two central partitions in the presence of noise.

Fraction of noise	Average *CRI* values for π^{*1}	*CRI* between π^{*1} and initial part.	Average *CRI* values for π^{*2}	*CRI* between π^{*2} and initial part.
10%	.818964	.819189	.818964	.819189
20%	.672516	.667547	.666279	.651882
30%	.556944	.560007	.535590	.546835
40%	.454627	.487534	.398607	.414782
50%	.355307	.387085	.272208	.290658
60%	.274852	.298431	.167627	.174901
70%	.194300	.236023	..060229	.061390
80%	.119703	.149001	.024150	.028592

From these last results, we can see that the central partition corresponding to the partition-distance has higher *CRI* values with the initial partition than the other central partition; except for the case of 10% noise, where the results are the same. It seems also clear that the higher the presence of noise the larger the difference between the *CRI* values for the two central partitions. We can conclude therefore that in the presence of noise, the central partition using the partition-distance $D_p(\pi, \pi')$ is superior to the central partition using the distance $D_C(\pi, \pi')$. On the other hand, we can again see that the average *CRI* values are higher for π^{*1} than for π^{*2}, which confirms the results obtained above.

From this experimental study, we find that the partition-distance is more adequate to find a consensus partition.

6. STRONG PATTERN GRAPH

Having shown experimentally that the partition-distance is more adequate to find a consensus partition, we now present some independent results that were developped during the course of our investigation on the central partition. We start by defining a new graph based on the notion of strong pattern. This new graph contains essentially the same information as the partition-graph, but is much simpler. Then, some properties of this new graph are proved.

Let $U_1, U_2, ..., U_m$ be the strong patterns of K partitions on a set E of size n. *The strong pattern* graph $sp(G)$ consists of m nodes, $U_1, U_2, ..., U_m$ and any two nodes U_q, U_j are adjacent if the strong patterns U_q and U_j are together in the same cluster in at least one partition.

We will now prove that the smallest node-cover of $G(\pi, \pi')$, which is a subset of E, is the union of a set of strong patterns; that is, if an element of E belongs to the smallest node-cover, all of the elements belonging to the same strong pattern belong also to the smallest node-cover.

Proposition 6.1. *Any smallest node-cover of $G(\pi, \pi')$ is composed of a subset of strong patterns.*

Proof: In order to prove this proposition, consider two elements x and y belonging to the same strong pattern. Suppose now that x belongs to a smallest node-cover of $G(\pi, \pi')$. From the results above, x belongs also to a smallest set of elements that have to be removed so that the two induced partitions become identical. We want to prove that y belongs also to the same smallest node-cover; that is, that y has also to be removed. Suppose not; that is, after removing all the elements that have to be removed so that the two induced partitions become identical, y stays. This means that the cluster of the induced partition of π containing y and the cluster of the induced partition of π' containing y are the same. Hence, if we add x to these two clusters, these two clusters remain also the same, because x and y belong to the same strong pattern, that is, are always clustered together; and so x would not have to be removed, which is absurd by hypothesis. Therefore y has also to be removed. □

A *clique* in a graph is a subset of nodes which are pairwise adjacent; let $K(G)$ be the size of the largest clique in graph G. An *independent set* of nodes is a subset of nodes where no two nodes are adjacent; let $I(G)$ be the size of the largest independent set in graph G. If U is a non empty subset of the node set of graph G, then the subgraph H of G induced by U is the graph having the node set U and whose edge set consists of those edges of G incident with two distinct elements of U. The subgraph H is called a *node-induced* subgraph. A graph G is called *perfect* if $K(H) = I(H)$ for every *node-induced* subgraph H of G.

Proposition 6.2. *The strong pattern graph for two partitions of the same set is a perfect graph.*

Proof: The strong pattern graph corresponding to two partitions π^1 and π^2 is itself a partition graph. In fact we can form two partitions of the set of strong patterns: π_S^1 is composed of clusters of strong patterns whose individual elements were clustered together in π^1; similarly for π_S^2. The strong pattern graph defined above corresponds to the partition graph for π_S^1 and π_S^2. It is proved in [7] that any partition graph is a perfect graph. Therefore, the strong pattern graph, being a partition graph, is a perfect graph. □

7. CONCLUSIONS AND FUTURE WORK

We have considered in this paper the problem of finding a consensus partition (central partition) between a set of partitions corresponding for instance to the results of different clustering algorithms. The distance between partitions is the one defined in [7]. As the determination of the central partition is NP-hard, we have adapted an heuristic [5] which consists in assuming that if two elements are always clustered together in all of the initial partitions, they should also be together in the central partition. We have then shown experimentally that the central partition corresponding to the partition-distance represents a better consensus than the usual central partition, which uses the distance defined in [5]. By defining a strong pattern to be a maximal subset of elements which are always together, we have then defined a strong pattern graph where the nodes correspond to the strong patterns and two nodes are adjacent if the corresponding strong patterns are together in at least one partition. We have then proved that any smallest node-cover of a partition graph is composed of a subset of strong patterns and also that the strong pattern graph is a perfect graph.

As for the future work, we plan to implement a computer program to do some experiments in order to analyse the results of some clustering algorithms. This will serve as a way of summarising the results of several clustering algorithms, specially when we do not know which one is best suited to the particular problem at hand. Even if we do know which clustering algorithm to use, its results usually depend on a set of parameters which are not known. By trying different parameters, we will get different partitions and once again, it makes sense to find the central partition (corresponding to the partition-distance) as the one which minimises the average number of disagreements between the various outputs. We plan also to study more deeply the strong pattern graph which we introduce in this article.

ACKNOWLEDGMENTS

The second author would like to thank the Board of Directors of Fundação Oriente for awarding a Scholarship to undertake studies at LIACC, University of Porto. The encouragment given by Professor Pavel Brazdil is also gratefully acknowledged. The authors also acknowledge the suggestions from the referees.

REFERENCES

[1] Almudevar, A. and Field, C. (1999). Estimation of single generation sibling relationships based on DNA markers, *J. Agricultural, biological and environmental statistics*, **4**, 136–165.

[2] Barthelemy, J.P. and Leclerc, B. (1995). The Median Procedure for Partitions. In "Partitioning Data Sets" (J.J. Cox, P. Hansen and B. Julesz, Eds.), *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, **19**, Amer. Math. Soc., Providence, RI, pp. 3–33.

[3] Barthelemy, J.P. and Monjardet, B. (1981). The median procedure in cluster analysis and social choice theory, *Mathematical Social Sciences*, **1**, 235–267.

[4] Celeux, G. (1984). *Approximation rapide et interprétation d'une partition centrale pour les algorithmes de partitionnement*, Rapport de recherche INRIA n. 352.

[5] Celeux, G.; Diday, E.; Govaert, G.; Lechevalier, Y. and Ralambondrainy, H. (1989). *Classification Automatique Des Données*, Dunod, Paris.

[6] Day, W.H.E. (1981). The complexity of computing metric distances between partitions, *Mathematical Social Sciences*, **1**, 269–287.

[7] Gusfield, D. (2002). Partition-distance: A problem and class of perfect graphs arising in clustering, *Information Processing Letters*, **82**(3), 159–164.

[8] Hubert, L. and Arabie, P. (1985). Comparing Partitions, *Journal of Classification*, **2**, 193–218.

[9] Lerman, I.C. (1981). *Classification et Analyse Ordinale des Données*, Dunod, Paris.

[10] Monti, S.; Tamayo, P.; Mesirov, J. and Golub, T. (2003). Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data, *Machine Learning*, **52**(1–2), 91–118.

[11] Pinto da Costa, Joaquim F. and Silva, Luis M.A. (2003). *Feature Selection in DNA Microarrays*. In "Actes du Xème Congrès de la Société Francophone de Classification", Neuchatel, Switzerland, 10–12 September.

[12] Rafill, T. http://www.soe.ucsc.edu/∼raff/ac-match/

[13] Rand, W. (1971). Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association*, **66**, 846–850.

[14] Régnier, S. (1965). Sur quelques aspects mathématiques des problèmes de classification automatique, *ICC Bull.*, **4**, 175–191, repr. (1983) *Mathématiques et Sciences Humaines*, **82**, 13–29.

[15] Strehl, A. and Ghosh, J. (2002). Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions, *Journal of Machine Learning Research*, **3**, 583–617.

REVSTAT – Statistical Journal
Volume 2, Number 2, November 2004

EXTENSIONS OF KATZ–PANJER FAMILIES OF DISCRETE DISTRIBUTIONS *

Authors: Dinis D. Pestana
– Departamento de Estatística e Investigação Operacional,
Faculdade de Ciências da Universidade de Lisboa, Bloco C6, Piso 4,
Campo Grande, 1749-016 Lisboa, Portugal, e
CEAUL – Centro de Estatística e Aplicações da Universidade de Lisboa
(dinis.pestana@fc.ul.pt)

Sílvio F. Velosa
– Departamento de Matemática e Engenharias, Universidade da Madeira,
Campus Universitário da Penteada, 9000-390 Funchal, Portugal, e
CEAUL – Centro de Estatística e Aplicações da Universidade de Lisboa
(sfilipe@uma.pt)

Received: July 2004 Revised: September 2004 Accepted: September 2004

Abstract:

- Let $N_{\alpha,\beta,\gamma}$ be a discrete random variable whose probability atoms $\{p_n\}_{n\in\mathbb{N}}$ satisfy $\frac{f(n+1)}{f(n)} = \alpha + \beta \frac{\mathbb{E}(U_0^n)}{\mathbb{E}(U_\gamma^n)}$, $n=0,1,...,$ for some $\alpha,\beta \in \mathbb{R}$, where $U_\gamma \frown \mathit{Uniform}(\gamma,1)$, $\gamma \in (-1,1]$. When $\gamma \to 1$, $U_\gamma \to U_1$, the degenerate random variable with unit mass at 1, and the above iterative expression is $\frac{p_{n+1}}{p_n} = \alpha + \frac{\beta}{n+1}$ for $n = k, k+1, ...,$ used by Katz and by Panjer ($k=0$), by Sundt and Jewell and by Willmot ($k=1$) and, for general $k \in \mathbb{N}$, by Hess, Lewald and Schmidt.
We investigate the case $U_\gamma \frown \mathit{Uniform}(\gamma,1)$ with $\gamma \in (-1,1)$ in detail for $\alpha = 0$. We then construct classes $\mathcal{C}_\gamma$ of discrete infinitely divisible randomly stopped sums such that $N_{0,\beta,\gamma} \in \mathcal{C}_\gamma$. $\mathcal{C}_0$ is the class of compound geometric random variables, $\mathcal{C}_1$ is the class of compound Poissons, and $|\gamma_1| < \gamma_2 \leq 1$ implies $\mathcal{C}_{\gamma_1} \subset \mathcal{C}_{\gamma_2} \subseteq \mathcal{C}_1$.

Key-Words:

- *Poisson stopped sums (compound Poisson); geometric stopped sums (compound geometric); Panjer's algorithm.*

AMS Subject Classification:

- 60G50, 60E10, 91B30.

*Research partially suported by FCT/POCTI/FEDER.

1. INTRODUCTION

Let us consider the discrete random variables $N_{\alpha,\beta}$ whose probability mass functions (p.m.f.) $\{f_{N_{\alpha,\beta}}(n)\}_{n\in\mathbb{N}}$ satisfy

$$f_{N_{\alpha,\beta}}(n+1) = \left(\alpha + \frac{\beta}{n+1}\right) f_{N_{\alpha,\beta}}(n) \ , \qquad \alpha,\beta \in \mathbb{R}, \quad n = 0,1,... \tag{1.1}$$

From (1.1) it follows that $f_{N_{\alpha,\beta}}(n) = f_{N_{\alpha,\beta}}(0) \prod_{k=1}^{n} \left(\alpha + \frac{\beta}{k}\right)$. In particular,

$$f_{N_{\alpha,0}}(n) = f_{N_{\alpha,0}}(0)\,\alpha^{n} = (1-\alpha)\,\alpha^{n} \implies N_{\alpha,0} \frown \mathit{Geometric}(1-\alpha) \ ,$$

and we may write

$$f_{N_{\alpha,0}}(n+1) \ = \ \alpha\, f_{N_{\alpha,0}}(n) \ = \ \sum_{k=0}^{n} f_{N_{\alpha,0}}(n)\, r_{n-j} \ , \tag{1.2}$$

where $r_0 = \alpha$ is the ratio of a geometric series and $r_1 = \cdots = r_n = 0$.

On the other hand,

$$f_{N_{0,\beta}}(n) = f_{N_{0,\beta}}(0) \prod_{k=1}^{n} \frac{\beta}{n} = f_{N_{0,\beta}}(0)\, \frac{\beta^n}{n!} = \ \mathrm{e}^{-\beta}\, \frac{\beta^n}{n!} \implies N_{\alpha,0} \frown \mathit{Poisson}(\beta) \ ,$$

and we may write

$$(n+1)\, f_{N_{0,\beta}}(n+1) \ = \ \beta\, f_{N_{0,\beta}}(n) \ = \ \sum_{k=0}^{n} f_{N_{0,\beta}}(n)\, r_{n-j} \ , \tag{1.3}$$

where $r_0 = \beta$ and $r_1 = \cdots = r_n = 0$. Note that similar expressions do not hold for randomly stopped sums $S_{N_{\alpha,\beta}} = S_{N_{\alpha,\beta}}(Y) = \sum_{k=1}^{N_{\alpha,\beta}} Y_k$, where the summands Y_k are i.i.d. and independent of the subordinator $N_{\alpha,\beta}$, with p.m.f. satisfying (1.1), whenever both $\alpha \neq 0$ and $\beta \neq 0$. However, for geometric stopped sums $\sum_{k=1}^{N_{\alpha,0}} Y_k$ and for Poisson stopped sums, $\sum_{k=1}^{N_{0,\beta}} Y_k$ (i.e., when either $\beta = 0$ or $\alpha = 0$) we get nice similar expressions, with the $r_k \geq 0$ and convergence of $\sum_{k=0}^{\infty} r_k$, in the case of geometric stopped sums, and convergence of $\sum_{k=0}^{\infty} \frac{r_k}{k+1}$, for Poisson stopped sums. In the definition of randomly stopped sums, $\mathbb{P}\big[S_{N_{\alpha,\beta}}=0 \,\big|\, N_{\alpha,\beta}=0\big] = 1$, and therefore $\mathbb{P}[S_{N_{\alpha,\beta}}=0] = \mathbb{P}[N_{\alpha,\beta}=0] = f_{N_{\alpha,\beta}}(0)$ whenever $\mathbb{P}[Y_k>0] = 1$).

Panjer (1981) has remarked that the discrete (nondegenerate) random variables whose p.m.f.'s satisfy equation (1.1) are

- $N_{0,\beta} \frown \mathit{Poisson}(\beta),\ \beta > 0,$
- $N_{\alpha,\beta} \frown \mathit{Binomial}\Big(-1-\frac{\beta}{\alpha},\ \frac{\alpha}{\alpha-1}\Big),$ in case $\alpha < 0$ and $-\frac{\beta}{\alpha} \in \mathbb{N}^+,$ and
- $N_{\alpha,\beta} \frown \mathit{NegativeBinomial}\Big(\frac{\alpha+\beta}{\alpha},\ 1-\alpha\Big)$ if $\alpha \in (0,\ 1)$ and $\alpha + \beta > 0.$

The dispersion index $\frac{\mathrm{var}(N_{\alpha,\beta})}{\mathbb{E}(N_{\alpha,\beta})} = \frac{1}{1-\alpha}$ is less than 1 (underdispersion) for the binomial and greater than 1 (overdispersion) for the negative binomial. On the other hand, $N_{0,\beta} \frown \mathit{Poisson}(\beta)$ is a yardstick, with dispersion index 1. We denote $\mathbf{\Pi}$ the class of random variables $N_{\alpha,\beta}$ described above.

These random variables play an important role as subordinators in randomly stopped sums. Compound or generalized random variables (other names traditionally given to $S_{N_{\alpha,\beta}}$, cf. the discussion on terminology in Johnson, Kotz and Kemp, 1992) are at the core of branching processes and many other subjects where the aim is to obtain the distribution of randomly stopped sums, namely in the study of aggregate claims in the risk process, see Klugman, Panjer and Willmot (1998) and Rólski, Schmidli, Schmidt and Teugels (1999).

Katz (1965) had used an iterative expression equivalent to (1.1) to organize a coordinated presentation of count distributions. Panjer's (1981) pathbreaking result has been to use the iterative expression satisfied by the p.m.f. of the subordinator $N_{\alpha,\beta}$ to get an iterative algorithm to compute the density function (probability mass function or probability density function) of $S_{N_{\alpha,\beta}}$. This is used in section 2 to establish characterization theorems for infinitely divisible and for geometric infinitely divisible generating functions.

In section 3, we investigate discrete random variables $N_{\alpha,\beta,\gamma}$ whose probability mass function (p.m.f.) $\{p_n\}_{n\in\mathbb{N}}$ satisfies the more general relation

$$\frac{f_{N_{\alpha,\beta,\gamma}}(n+1)}{f_{N_{\alpha,\beta,\gamma}}(n)} = \alpha + \beta\,\frac{\mathbb{E}(U_0^n)}{\mathbb{E}(U_\gamma^n)} = \alpha + \frac{\beta}{\sum_{k=0}^{n}\gamma^k}\ , \qquad \alpha,\beta \in \mathbb{R},\quad n = 0,1,... \tag{1.4}$$

where $U_\gamma \frown \mathit{Uniform}(\gamma,\ 1),\ \gamma \in (-1,\ 1)$. As

$$\mathbb{E}(U_\gamma^n) = \frac{1}{n+1}\,\frac{1-\gamma^{n+1}}{1-\gamma} \ \xrightarrow[\gamma\to 1]{}\ 1\ , \tag{1.5}$$

Panjer's class corresponds to the degenerate limit case, letting $\gamma \to 1$ so that $U_\gamma \to U_1$, the degenerate random variable with unit mass at 1.

When $\alpha = 0$, the iterative expression for the p.m.f. of $N_{0,\beta,\gamma}$ verifies

$$\frac{1-\gamma^{n+1}}{1-\gamma}\, f_{N_{\alpha,\beta,\gamma}}(n+1) \;=\; \sum_{k=0}^{n} f_{N_{\alpha,\beta,\gamma}}(k)\, r_{n-k} \tag{1.6}$$

with $r_0 = \beta$ and $r_1 = \cdots = r_n = 0$, of which (1.2) and (1.3) aren't but the cases $\gamma = 0$ and $\gamma = 1$, respectively. We shall investigate the classes $\mathcal{C}_\gamma$ of randomly stopped sums $\sum_{k=0}^{N_{0,\beta,\gamma}} Y_k$, whose members satisfy (1.6) for nonnegative r_k, with of $\sum_{k=0}^{\infty} r_k < \infty$.

In section 4 we show that when $|\gamma_1| < \gamma_2 \leq 1$, $\mathcal{C}_{\gamma_1} \subset \mathcal{C}_{\gamma_2}$. Also, for $\gamma \in [0,1]$, the classes $\mathcal{C}_\gamma$ form an increasing chain of classes of infinitely divisible random variables, spanning from $\mathcal{C}_0$, the class of discrete geometric stopped sums, to $\mathcal{C}_1$, the class of discrete Poisson stopped sums.

Many of these results rely on properties of absolutely monotone functions scattered in the literature, that we shall discuss in section 2 below in conjunction with Panjer theory. Ospina and Gerber (1987) remarked that the representation theorem for the generating functions of discrete stopped Poisson sums (discrete infinitely divisible laws) follows from Panjer's theory, and the same is true for the representation of geometric infinitely divisible generating functions, see section 2, and for wider classes of generating functions whose bearing on general p-infinite divisibility is worth noting. This will be further discussed in the concluding section.

2. BASIC RESULTS

Let $\mathcal{G}(s) = \sum_{n=0}^{\infty} f(n) s^n,\ s \in [0, r),$ be the generating function of the sequence $\{f(n)\}_{n\in\mathbb{N}}$; in other words, $f(n) = \frac{\mathcal{G}^{(n)}(0)}{n!},\ n \in \mathbb{N}$.

If $p_n \geq 0$, $n \in \mathbb{N}$, then $\mathcal{G}^{(n)}(s) \geq 0$, $s \in [0, r)$, and we say that $\mathcal{G}$ is absolutely monotone (abs. mon.) in $[0, r)$. If there exists $r > 0$ such that $\mathcal{G}$ is abs. mon. in $[0, r)$, we say that the function $\mathcal{G}$ is abs. mon. (Bernstein, 1928).

We refer to Widder (1946, chapt. IV) and to Feller (1968, chap. XI) for basic information on absolutely monotone functions and generating functions; Skellam and Shelton (1957) or Srivastava and Manocha (1984) provide a thorough discussion. It is obvious that the sum or the product of abs. mon. functions is abs. mon.; we shall need the following results:

1. $\mathcal{G}$ is abs. mon. $\iff \mathcal{G}(0) \geq 0$ and $\frac{\mathrm{d}\mathcal{G}}{\mathrm{d}s}$ is abs. mon. $\iff \frac{\mathrm{d}}{\mathrm{d}s}\,[s\,\mathcal{G}(s)]$ is abs. mon. (since $p_n \geq 0$ iff $(1+n)\,p_n \geq 0$).

2. Let $\gamma \in (-1,\,1)$; then, $\mathcal{G}$ abs. mon. $\iff \mathcal{G}(s) - \gamma\,\mathcal{G}(\gamma s)$ abs. mon. (it is sufficient to note that $p_n \geq 0 \iff p_n(1-\gamma^{n+1}) \geq 0$).

Let $|\gamma| \leq \eta < 1$; then, $\mathcal{G}$ abs. mon. $\implies \eta\,\mathcal{G}(\eta s) - \gamma\,\mathcal{G}(\gamma s)$ abs. mon. ($p_n \geq 0$ implies $p_n(\eta^{n+1} - \gamma^{n+1}) \geq 0$).

Note that $\eta\,\mathcal{G}(\eta s) - \gamma\,\mathcal{G}(\gamma s)$ is no longer abs. mon. if $-1 < \eta < \gamma \leq 0$.

3. If $\mathcal{G}_1$ is abs. mon. in $[0, r_1)$, $\mathcal{G}_2$ is abs. mon. in $[0, r_2)$, and $\mathcal{G}_2(s) < r_1$ for all $s \in [0, r_2)$, the compound function $\mathcal{G}_1 \circ \mathcal{G}_2 = \mathcal{G}_1(\mathcal{G}_2)$ is abs. mon. in $[0, r_2)$. In particular:

(**a**) As $\mathcal{G}_1(s) = \mathrm{e}^{s}$ is the generating function of $p_n = \frac{1}{n!}$, $\mathcal{G}_2$ abs. mon. implies that $(\mathcal{G}_1 \circ \mathcal{G}_2)(s) = \mathrm{e}^{\mathcal{G}_2(s)}$ is abs. mon.

(**b**) As $\mathcal{G}_1(s) = \frac{1}{1-s}$ is the generating function of $p_n = 1$, $\mathcal{G}_2$ abs. mon. in $[0, r_2)$ with $\mathcal{G}_2(s) < 1$ for $s \in [0, r_2)$ implies that $(\mathcal{G}_1 \circ \mathcal{G}_2)(s) = \frac{1}{1-\mathcal{G}_2(s)}$ is abs. mon.

Let us now consider the randomly stopped sum $S_{N_{\alpha,\beta}} = \sum_{k=1}^{N_{\alpha,\beta}} Y_k$, where $Y_k \stackrel{\mathrm{d}}{=} Y$, $k = 1, 2, \ldots$, are i.i.d. counting random variables, with p.m.f. $\{f_Y(n)\}_{n\in\mathbb{N}}$, independent of the Panjer subordinator $N_{\alpha,\beta}$.

As

$$\mathbb{E}\left[\frac{k}{n+1}\,Y_1 \,\middle|\, \sum_{i=1}^{k} Y_i = n+1\right] = 1$$

and

$$\mathbb{P}\left[Y_1 = j \,\middle|\, \sum_{i=1}^{k} Y_i = n+1\right] = \frac{f_Y(j)\, f_Y^{\star(k-1)}(n+1-j)}{f_Y^{\star k}(n+1)}\,, \qquad j = 0, \ldots, n+1$$

(Rólski *et al.*, 1999, p. 119), where as usual $f^{\star k}$ denotes the k-fold convolution ($f^{1*} = f$, $f^{\star k} = f * f^{\star(k-1)}$), it follows that the probability mass function of a Poisson stopped sums ($N_{0,\beta}$, $\beta > 0$) verifies

$$\begin{aligned}(n+1)\, f_{S_{N_{0,\alpha}}}(n+1) &= \sum_{k=0}^{n} f_{S_{N_{0,\alpha}}}(k)\, \beta\,(n+1-k)\, f_Y(n+1-k) \\ &= \sum_{k=0}^{n} f_{S_{N_{0,\alpha}}}(k)\, r_{n-k}\,, \qquad (2.1)\end{aligned}$$

with $r_k = \beta\,(k+1)\, f_Y(k+1) \geq 0$, $k = 0, 1, \ldots$, and it therefore follows that the generating function $\mathcal{H}_{N_{0,\beta}}(s) = \sum_{k=0}^{\infty} r_k\, s^k$ of the $\{r_k\}_{k\in\mathbb{N}}$ is absolutely monotone, with $\sum_{k=0}^{\infty} \frac{r_k}{k+1} = \sum_{k=0}^{\infty} \beta\, f_Y(k+1) = \beta\,(1 - f_Y(0))$. Assuming that $f_Y(0) = 0$ (i.e., enforcing

a unique representation by fixing this free parameter), multiplying both sides of (1.6) by s^n and summing for $n = 0, 1, ...$, we get,

$$\mathcal{G}_{S_{N_{0,\beta}}}(s) = \exp\left[\beta\left(\frac{1}{\beta}\sum_{k=0}^{\infty}\frac{r_k}{k+1}\,s^{k+1} - 1\right)\right] = \mathrm{e}^{\beta[\mathcal{P}(s)-1]}\,, \tag{2.2}$$

where $\mathcal{P}(s) = \frac{1}{\beta}\sum_{k=0}^{\infty}\frac{r_k}{k+1}\,s^{k+1}$ is a (unique) p.g.f., such that $\mathcal{P}(0) = 0$.

On the other hand, for geometric stopped sums ($N_{\alpha,0}$, $0<\alpha<1$) we get

$$f_{S_{N_{\alpha,0}}}(n+1) = \sum_{k=0}^{n} f_{S_{N_{\alpha,0}}}(k)\, r_{n-k} \tag{2.3}$$

where $r_k = \frac{\alpha\, f_Y(k+1)}{1-\alpha\, f_Y(0)}$. As in the treatment of Poisson stopped sums, we may get a unique representation theorem by letting the free parameter $f_Y(0) = 0$, which implies $\sum_{k=0}^{\infty} r_k = \alpha$, multiplying both sides of (2.3) by s^n and summing for $n = 0, 1, \ldots$. In terms of generating functions,

$$\frac{\mathcal{G}_{S_N}(s) - f_{N_{\alpha,0}}(0)}{s} = \mathcal{G}_{S_N}(s)\,\mathcal{H}_{N_{\alpha,0}}(s)$$

where $\mathcal{H}_{N_{\alpha,0}}$ is the generating function of $\{r_n\}_{n\in\mathbb{N}}$, which are all nonnegative, with $\sum_{n=0}^{\infty} r_n = \alpha \in (0,1)$, i.e. $\mathcal{H}_{N_{\alpha,0}}$ is abs. mon. From $\mathcal{G}_{S_N}(1) = \frac{p_0}{1-\mathcal{H}_{N_{\alpha,0}}(1)} = \frac{p_0}{1-\alpha} = 1$, it follows that

$$\mathcal{G}_{S_N}(s) = \frac{p_0}{1 - s\,\mathcal{H}_{N_{\alpha,0}}(s)} = \frac{1-\alpha}{1 - s\sum_{k=0}^{\infty} r_k\, s^k} = \frac{1-\alpha}{1-\alpha\,\mathcal{P}(s)}$$

where $\mathcal{P}(s) = \sum_{k=0}^{\infty}\frac{r_k}{\alpha}\,s^{k+1}$, such that $\mathcal{P}(0) = 0$, is a p.g.f., because it is abs. mon. and $\mathcal{P}(1) = 1$. In other words, Panjer's iteration also provides a straightforward proof of the representation theorem for geometric infinitely divisible lattice distributions.

We record these representation theorems for the sake of the corollaries that we then establish, which will be instrumental in the proof of the extensions in sections 3 and 4.

Theorem 2.1. *The p.g.f. $\mathcal{G}_{S_{N_{0,\beta}}}$ of a discrete Poisson stopped sum such that $\mathbb{P}[S_{N_{0,\beta}}=0] = f_{S_{N_{0,\beta}}}(0) > 0$ has a unique representation $\mathcal{G}_{S_{N_{0,\beta}}}(s) = \mathrm{e}^{\beta[\mathcal{P}(s)-1]}$, where $\mathcal{P}$ is a p.g.f. such that $\mathcal{P}(0) = 0$, and $\beta = -\ln \mathcal{G}_{S_{N_{0,\beta}}}(0)$.*

The p.g.f. $\mathcal{G}_{S_{N_{\alpha,0}}}$ of a discrete geometric stopped sum such that $\mathbb{P}[S_N=0] = f_{S_{N_{\alpha,0}}}(0) > 0$ has a unique representation $\mathcal{G}_{S_{N_{\alpha,0}}}(s) = \frac{1-\alpha}{1-\alpha\,\mathcal{P}(s)}$, where $\mathcal{P}$ is a p.g.f. such that $\mathcal{P}(0) = 0$, and $\alpha = 1 - \mathcal{G}_{S_{N_{\alpha,0}}}(0)$.

Observe also that $\exp\Big(1-\frac{1}{\mathcal{G}_{S_{N_{\alpha,0}}}}\Big)=\mathrm{e}^{\frac{\alpha}{1-\alpha}[\mathcal{P}(s)-1]}=\mathcal{G}_{S_{N_{0,\frac{\alpha}{1-\alpha}}}}(s)$. On the other hand, $\dfrac{1}{1-\ln\Big(\mathcal{G}_{S_{N_{0,\beta}}}(s)\Big)}=\dfrac{1-\frac{\beta}{\beta+1}}{1-\frac{\beta}{\beta+1}\mathcal{P}(s)}=\mathcal{G}_{S_{N_{\frac{\beta}{\beta+1},0}}}(s)$.

Corolary 2.1.1.

(**1**) Let $\mathcal{G}$ be a probability generating function such that $\mathcal{G}(0)>0$; then, $\mathcal{G}$ is the p.g.f. of a discrete Poisson stopped sum iff $\frac{\mathcal{G}'(s)}{\mathcal{G}(s)}$ is abs. mon.

(**2**) Let $\mathcal{G}$ be a p.g.f. such that $\mathcal{G}(0)>0$, and $\gamma\in(-1,1)$. If $\mathcal{G}$ is the p.g.f. of a discrete Poisson stopped sum, then $\frac{\mathcal{G}(s)}{\mathcal{G}(\gamma s)}$ is abs. mon., and $\mathcal{G}_{\gamma}(s)=\frac{\mathcal{G}(\gamma)\,\mathcal{G}(s)}{\mathcal{G}(\gamma s)}$ is also the p.g.f. of a Poisson stopped sum.

(**3**) Let $\mathcal{G}$ be a p.g.f. such that $\mathcal{G}(0)>0$, and $|\gamma_1|\le\gamma_2<1$. If $\mathcal{G}$ is the p.g.f. of a discrete Poisson stopped sum, then $\frac{\mathcal{G}(\gamma_2 s)}{\mathcal{G}(\gamma_1 s)}$ is abs. mon. and $\mathcal{G}_{\gamma_1,\gamma_2}(s)=\frac{\mathcal{G}(\gamma_1)}{\mathcal{G}(\gamma_2)}\frac{\mathcal{G}(\gamma_2 s)}{\mathcal{G}(\gamma_1 s)}$ is also the p.g.f. of a Poisson stopped sum.

(**4**) Any discrete geometric stopped sum such that $\mathbb{P}[S_N=0]=\tilde{p}_0>0$ is a Poisson stopped sum, i.e. infinitely divisible.

Proof: (**1**) From Theorem 2.1 we know that $\mathcal{G}$, with $\mathcal{G}(0)>0$, is the p.g.f. of a Poisson stopped sum iff $\frac{\mathcal{G}'(s)}{\mathcal{G}(s)}=\mathcal{H}_{N_{0,\beta}}(s)=\sum_{k=0}^{\infty}r_k s^k$, where $r_k=\beta\,(k+1)\,f_Y(k+1)\ge 0,\ k=0,1,...,$ and therefore its generating function $\mathcal{H}_{N_{0,\beta}}(s)=\sum_{k=0}^{\infty}r_k\,s^k$ is absolutely monotone.

(**2**) From formula (2.2), we see that $\mathcal{G}(s)>0$ for all s, therefore $\frac{\mathcal{G}(s)}{\mathcal{G}(\gamma s)}\ge 1$ if $0\le s\le 1$. On the other hand, $\frac{\mathrm{d}}{\mathrm{d}s}\Big[\ln\frac{\mathcal{G}(s)}{\mathcal{G}(\gamma s)}\Big]=\frac{\mathcal{G}'(s)}{\mathcal{G}(s)}-\gamma\,\frac{\mathcal{G}'(\gamma s)}{\mathcal{G}(\gamma s)}$ is abs. mon., by 2.1.1.(1) and property 2 of abs. mon. functions. As $\ln\frac{\mathcal{G}(s)}{\mathcal{G}(\gamma s)}$ is nonnegative for $s=0$, it is also abs. mon., by property 1 of abs. mon. functions. From property 3(a) of abs. mon. functions, it follows that $\frac{\mathcal{G}(s)}{\mathcal{G}(\gamma s)}$ is abs. mon. Since $\mathcal{G}_{\gamma}(0)=\mathcal{G}(\gamma)>0$, $\mathcal{G}_{\gamma}(1)=1$, and $\frac{\mathcal{G}'_{\gamma}(s)}{\mathcal{G}_{\gamma}(s)}=\frac{\mathrm{d}}{\mathrm{d}s}\Big[\ln\frac{\mathcal{G}(s)}{\mathcal{G}(\gamma s)}\Big]$ is abs. mon., we conclude that $\mathcal{G}_{\gamma}$ is the p.g.f. of a Poisson stopped sum.

(**3**) By 2.1.1.(2), $\frac{\mathcal{G}(\gamma_2 s)}{\mathcal{G}(\gamma_1 s)}$ is abs. mon., and by property 2 of abs. mon. functions $\frac{\mathcal{G}'_{\gamma_1,\gamma_2}(s)}{\mathcal{G}_{\gamma_1,\gamma_2}(s)}=\gamma_2\frac{\mathcal{G}'(\gamma_2 s)}{\mathcal{G}(\gamma_2 s)}-\gamma_1\frac{\mathcal{G}'(\gamma_1 s)}{\mathcal{G}(\gamma_1 s)}$ is abs. mon. Since $\mathcal{G}_{\gamma_1,\gamma_2}(0)=\frac{\mathcal{G}(\gamma_1)}{\mathcal{G}(\gamma_2)}>0$ and $\mathcal{G}_{\gamma_1,\gamma_2}(1)=1$, it follows that $\mathcal{G}_{\gamma_1,\gamma_2}(s)=\frac{\mathcal{G}(\gamma_1)}{\mathcal{G}(\gamma_2)}\frac{\mathcal{G}(\gamma_2 s)}{\mathcal{G}(\gamma_1 s)}$ is the p.g.f. of a Poisson stopped sum.

(**4**) As we have seen, $\mathcal{G}$ with $\mathcal{G}(0) > 0$ is the p.g.f. of a discrete geometric stopped sum iff $\mathcal{G}(s) = \frac{\mathcal{G}(0)}{1 - s\,\mathcal{H}_{N_{\alpha,0}}(s)}$, where $\mathcal{H}_{N_{\alpha,0}}(s) < 1$ for $s \in [0,\,1)$ is abs. mon. As

$$\frac{\mathcal{G}'(s)}{\mathcal{G}(s)} = \frac{\frac{\mathcal{G}(0)\,\frac{\mathrm{d}}{\mathrm{d}s}\left[s\,\mathcal{H}_{N_{\alpha,0}}(s)\right]}{\left(1 - s\,\mathcal{H}_{N_{\alpha,0}}(s)\right)^2}}{\frac{\mathcal{G}(0)}{1 - s\,\mathcal{H}_{N_{\alpha,0}}(s)}} = \frac{\frac{\mathrm{d}}{\mathrm{d}s}\left[s\,\mathcal{H}_{N_{\alpha,0}}(s)\right]}{1 - s\,\mathcal{H}_{N_{\alpha,0}}(s)}\ ,$$

and from property 1 of abs. mon. functions $\frac{\mathrm{d}}{\mathrm{d}s}\left[s\,\mathcal{H}_{N_{\alpha,0}}(s)\right]$ is abs. mon., from property 3(b) we know that $\frac{1}{1 - s\,\mathcal{H}_{N_{\alpha,0}}(s)}$ is abs. mon., and the product of abs. mon. functions is abs. mon., it follows that $\frac{\mathcal{G}'(s)}{\mathcal{G}(s)}$ is abs. mon. From part (1) of Corollary 2.1.1., it follows that $\mathcal{G}$ is the p.g.f. of a Poisson stopped sum. □

If the probability generating function $\mathcal{G}_Y$ of $Y \frown F_Y$ depends on the parameter θ so that $\mathcal{G}_Y(s\,|\,k\theta) = [\mathcal{G}_Y(s\,|\,\theta)]^k$, then $F_X \vee F_Y = F_Y \underset{K}{\wedge} F_X$, where $\vee$ denotes the stopped sum of Y independent copies of X, and $\underset{K}{\wedge}$ denotes the mixture of $Y\,|\,K$, with mixing distribution F_X (Gurland, 1957). Therefore the class of discrete Poisson stopped sums coincides with the class of discrete mixtures of Poisson random variables. In what mixtures of geometric random variables and geometric stopped sums, the former is stricly included in the later.

3. EXTENSIONS

We now investigate the nondegenerate discrete random variables $N_{\alpha,\beta,\gamma}$ whose probability mass function $\{p_n\}_{n\in\mathbb{N}}$ satisfies

$$\frac{p_{n+1}}{p_n} = \alpha + \beta\,\frac{\mathbb{E}(U_0^n)}{\mathbb{E}(U_\gamma^n)} = \alpha + \beta\,\frac{1-\gamma}{1-\gamma^{n+1}} \qquad \text{for } n = 0, 1, ..., \quad \alpha, \beta \in \mathbb{R}\ , \tag{3.1}$$

where $U_\gamma \frown \mathit{Uniform}(\gamma,\,1)$, $\gamma \in (-1,\,1)$, with $p_0 > 0$. If $\gamma = 0$, all possible solutions are geometric random variables, and when $\gamma \to 1$ we get Panjer's class of counting distributions.

For a nondegenerate solution of (3.1) with infinite support to exist, we must have

$$\alpha + \beta\,\frac{\mathbb{E}(U_0^n)}{\mathbb{E}(U_\gamma^n)} = \alpha + \beta\,\frac{1-\gamma}{1-\gamma^{n+1}} > 0$$

for every integer n. According to the signs of β and γ, the infimum of this factor is either $\alpha + \beta$ (for $n = 0$), $\alpha + \frac{\beta}{1+\gamma}$ (for $n = 1$), or $\alpha + \beta(1-\gamma)$ (when $n \to \infty$), so we must have $\alpha + \beta > 0$, $\alpha + \frac{\beta}{1+\gamma} > 0$, and $\alpha + \beta(1-\gamma) \geq 0$. Then, applying

the ratio test to the sum

$$\sum_{k\geq 0} p_k \,=\, p_0 \sum_{k=0}^{\infty} \prod_{n=0}^{k-1} \left(\alpha + \beta \, \frac{1-\gamma}{1-\gamma^{n+1}} \right)$$

we see that it converges iff $0 \leq \alpha + \beta(1-\gamma) < 1$. Thus a necessary and sufficient condition for a solution of (3.1) with infinite support (random variable with finite support cannot be infinitely divisible) to exist is that

$$\min\left\{ \alpha+\beta,\, \alpha + \frac{\beta}{1+\gamma} \right\} > 0 \qquad \text{and} \qquad 0 \leq \alpha + \beta(1-\gamma) < 1 \ .$$

Rewriting (3.1) as

$$(3.2) \quad (1-\gamma^{n+1})\, f_{N_{\alpha,\,\beta,\,\gamma}}(n+1) \,=\, \left[\alpha+\beta(1-\gamma)\right] f_{N_{\alpha,\,\beta,\,\gamma}}(n) - \alpha\,\gamma^{n+1} f_{N_{\alpha,\,\beta,\,\gamma}}(n) \ ,$$

for $\gamma \in (-1,\, 1)$, $n = 0, 1, ...$, multiplying both sides by s^{n+1} and summing we get

$$(3.3) \qquad \left[1 - \left(\alpha + \beta(1-\gamma)\right) s\right] \mathcal{G}_{\alpha,\,\beta,\,\gamma}(s) \,=\, (1 - \alpha\,\gamma\, s)\, \mathcal{G}_{\alpha,\,\beta,\,\gamma}(\gamma s) \ ,$$

where $\mathcal{G}_{\alpha,\,\beta,\,\gamma}(s) = \sum\limits_{n=0}^{\infty} f_{N_{\alpha,\,\beta,\,\gamma}}(n)\, s^n$ denotes the probability generating function of the probability mass function $\left\{ f_{N_{\alpha,\,\beta,\,\gamma}}(n) \right\}_{n=0}^{\infty}$, and from that

$$(3.4) \qquad \mathcal{G}_{\alpha,\,\beta,\,\gamma}(s) \,=\, \mathcal{G}_{\alpha,\,\beta,\,\gamma}(\gamma^{n+1} s) \prod_{k=0}^{n} \frac{1 - \alpha\,\gamma^{k+1} s}{1 - \left[\alpha+\beta(1-\gamma)\right]\gamma^k s} \ .$$

Observing that

$$(3.5) \qquad \frac{\mathcal{G}_{\alpha,\,\beta,\,\gamma}(s)}{\mathcal{G}_{\alpha,\,\beta,\,\gamma}(1)} \,=\, \frac{\mathcal{G}_{\alpha,\,\beta,\,\gamma}(\gamma^{n+1} s)}{\mathcal{G}_{\alpha,\,\beta,\,\gamma}(\gamma^{n+1})} \prod_{k=0}^{n} \frac{\frac{1 - \alpha\,\gamma^{k+1} s}{1 - \left[\alpha+\beta(1-\gamma)\right]\gamma^k s}}{\frac{1 - \alpha\,\gamma^{k+1}}{1 - \left[\alpha+\beta(1-\gamma)\right]\gamma^k}}$$

and letting $n \to \infty$,

$$(3.6) \qquad \mathcal{G}_{\alpha,\,\beta,\,\gamma}(s) \,=\, \prod_{k=0}^{\infty} \frac{1 - \alpha\,\gamma^{k+1} s}{1 - \alpha\,\gamma^{k+1}} \, \frac{1 - \left[\alpha+\beta(1-\gamma)\right]\gamma^k}{1 - \left[\alpha+\beta(1-\gamma)\right]\gamma^k s} \ .$$

If $\gamma \in [0, 1)$, $\alpha < 0$ and $\beta \in (-\frac{\alpha}{1-\gamma},\, \frac{1-\alpha}{1-\gamma})$, we recognize in

$$\mathcal{G}_{\alpha,\,\beta,\,\gamma}(s) \,=\, \prod_{k=0}^{\infty} \frac{1 - \alpha\,\gamma^{k+1} s}{1 - \alpha\,\gamma^{k+1}} \, \frac{1 - \left[\alpha+\beta(1-\gamma)\right]\gamma^k}{1 - \left[\alpha+\beta(1-\gamma)\right]\gamma^k s} \ ,$$

the probability generating function of an infinite sum of independent random variables, the k-th summand being the result of randomly adding 1, with probability $\frac{\alpha\,\gamma^{k+1}}{\alpha\,\gamma^{k+1}-1}$, to an independent *Geometric*$(1 - [\alpha + \beta(1-\gamma)]\gamma^k)$ random variable.

The limiting case $\gamma = 1$ may be approached as follows: rewriting (3.3) as

$$\frac{\mathcal{G}_{\alpha,\beta,\gamma}(s) - \mathcal{G}_{\alpha,\beta,\gamma}(\gamma s)}{\alpha\, s\Big[\mathcal{G}_{\alpha,\beta,\gamma}(s) - \mathcal{G}_{\alpha,\beta,\gamma}(\gamma s)\Big] + (1-\gamma)\, s\Big[\beta\, \mathcal{G}_{\alpha,\beta,\gamma}(s) + \alpha\, \mathcal{G}_{\alpha,\beta,\gamma}(\gamma s)\Big]} = 1\ ,$$

dividing the numerator and the denominator by $(1-\gamma)s$ and letting $\gamma \to 1$, we get

$$\frac{\mathcal{G}'_{\alpha,\beta,1}(s)}{\alpha\, s\, \mathcal{G}'_{\alpha,\beta,1}(s) + \beta\, \mathcal{G}_{\alpha,\beta,1}(s) + \alpha\, \mathcal{G}_{\alpha,\beta,1}(s)} = 1 \iff \frac{\mathcal{G}'_{\alpha,\beta,1}(s)}{\mathcal{G}_{\alpha,\beta,1}(s)} = \frac{\alpha+\beta}{1-\alpha\, s}\ ,$$

the expression we obtain working out the probability generating function in Panjer's iterative expression $p_{\alpha,\beta}(n+1) = (\alpha + \frac{\beta}{n+1})\, p_{\alpha,\beta}(n)$, $\alpha, \beta \in \mathbb{R}$, $n = 0, 1, \ldots$.

We now focus on the case $\alpha = 0$, for which $\beta \in (0, \frac{1}{1-\gamma})$, and

$$\mathcal{G}_{0,\beta,\gamma}(s) = \prod_{k=0}^{\infty} \frac{1 - \beta(1-\gamma)\,\gamma^k}{1 - \beta(1-\gamma)\,\gamma^k s} = \prod_{k=0}^{\infty} \frac{1-w_k}{1-w_k s}\ , \tag{3.7}$$

where $w_k = \beta(1-\gamma)\gamma^k$. If $\gamma \in [0,1)$, we get that have $N_{0,\beta,\gamma} = \sum_{k=0}^{\infty} W_k$, with $W_k \frown$ *Geometric*$(1 - \beta(1-\gamma)\,\gamma^k)$ independent summands. If $\gamma = 0$, the above expression simplifies to $\mathcal{G}_{0,\beta,0}(s) = \frac{1-\beta}{1-\beta s}$. Therefore we conclude that $N_{0,\beta,0} = N_{\beta,0} \frown$ *Geometric*$(1-\beta)$, $\beta \in (0,1)$.

Let us point out that the probability mass function of a random variable $N_{0,\beta,\gamma}$, $\gamma \in (-1,1)$, trivially satisfies

$$\frac{1-\gamma^{n+1}}{1-\gamma}\, p_{n+1} = \sum_{k=0}^{n} p_k\, r_{n-k}\ ,$$

with $r_0 = \beta$ and $r_1 = r_2 = \cdots = r_n = 0$, provided that

$$0 < \beta = \sum_{n=0}^{\infty} r_n\, s^n = \mathcal{H}(s) < \frac{1}{1-\gamma}\ ,$$

a point which will be of relevance in the following section.

4. DISCRETE INFINITELY DIVISIBLE DISTRIBUTIONS AND $\mathcal{C}_\gamma$ CLASSES

In what follows we investigate the classes $\mathcal{C}_\gamma$, $\gamma \in (-1,1)$, of nondegenerate counting random variables (distributions, p.g.f.) whose probability mass function satisfies $\tilde{p}_0 > 0$ and the general recursive relation

$$\frac{1-\gamma^{n+1}}{1-\gamma}\, \tilde{p}_{n+1} = \sum_{k=0}^{n} \tilde{p}_k\, r_{n-k}\ , \qquad n = 0, 1, \ldots\ , \tag{4.1}$$

with $r_k \geq 0$, which extends Panjer's recursive expression for the probability mass function of the classes of Poisson stopped sums ($\mathcal{C}_1$) and of geometric stopped sums ($\mathcal{C}_0$). It is well known that any geometric infinitely divisible lattice distribution is infinitely divisible in the classical sense, a result that follows from the fact that $\frac{1-p}{1-ps} = \exp\big\{\ln(1-p)\,[\mathcal{P}(s)-1]\big\}$, where $\mathcal{P}(s) = -\frac{1}{\ln(1-p)} \sum_{k=0}^{\infty} \frac{(ps)^k}{k}$ is the p.g.f. of a logarithmic random variable.

As before, multiplying both members of (4.1) by s^{n+1} and summing for $n \geq 0$, we obtain

$$\frac{\mathcal{G}(s) - \mathcal{G}(\gamma s)}{1-\gamma} \;=\; s\,\mathcal{G}(s)\,\mathcal{H}_\gamma(s)\;, \tag{4.2}$$

where $\mathcal{G}(s) = \sum_{n=0}^{\infty} \tilde{p}_n s^n$ and $\mathcal{H}_\gamma(s) = \sum_{n=0}^{\infty} r_n s^n$ converges at least for $|s| \leq 1$. Thus $\mathcal{H}_\gamma$ is by definition abs. mon. Since we have excluded degenerate solutions to (4.1), we must have $\mathcal{H}_\gamma(0) = r_0 = \frac{\tilde{p}_1}{\tilde{p}_0} > 0$.

If $\gamma \in [0,\,1)$, we have

$$\begin{aligned} 1 \;&\geq\; \sum_{n=0}^{\infty} \tilde{p}_{n+1} \;=\; \sum_{n=0}^{\infty} \frac{1-\gamma}{1-\gamma^{n+1}} \sum_{k=0}^{n} \tilde{p}_k r_{n-k} \\ &=\; \sum_{k=0}^{\infty} \tilde{p}_k \sum_{n=0}^{\infty} \frac{(1-\gamma)\,r_n}{1-\gamma^{n+k+1}} \\ &>\; \sum_{k=0}^{\infty} \tilde{p}_k \sum_{n=0}^{\infty} (1-\gamma)\,r_n \;=\; (1-\gamma) \sum_{n=0}^{\infty} r_n\;, \end{aligned}$$

and therefore $|\mathcal{H}_\gamma(s)| \leq \mathcal{H}_\gamma(1) = \sum_{n=0}^{\infty} r_n < \frac{1}{1-\gamma}$ for $|s| \leq 1$.

If $\gamma \in (-1,\,0)$, then $\frac{1-\gamma^{i+1}}{1-\gamma} \leq 1$ for $i = 0, 1, ...$, and by a similar reasoning we conclude that in this case $|\mathcal{H}_\gamma(s)| < 1$ for $|s| \leq 1$.

As was seen in the previous section, the p.m.f. of $N_{0,\,\beta,\,\gamma}$ verifies recursion (4.1) with $r_0 = \beta$ and $r_1 = r_2 = ... = 0$, with $0 < \beta < \frac{1}{1-\gamma}$.

We have the following result:

Theorem 4.1. *Let W be a random variable with p.g.f. $\mathcal{G}$, and $\gamma \in (-1, 1)$.*

$$W \in \mathcal{C}_\gamma \qquad \textit{iff}\;\; \mathcal{G}(s) = \prod_{k=0}^{\infty} \frac{1-(1-\gamma)\,\gamma^k\,\mathcal{H}_\gamma(\gamma^k)}{1-(1-\gamma)\,\gamma^k\,s\;\mathcal{H}_\gamma(\gamma^k s)}\;,$$

where $\mathcal{H}_\gamma$ is a unique abs. mon. function such that $\mathcal{H}_\gamma(0) > 0$ and $\mathcal{H}_\gamma(1) < \max\{1, \frac{1}{1-\gamma}\}$.

Thus, if $\gamma \in [0,1)$ the elements of $\mathcal{C}_\gamma$ are infinite sums $W = \sum_{k=0}^{\infty} X_k$ of independent geometric stopped sums $X_k = \sum_{i=1}^{N_k} Y_{ki}$, whose subordinators are $N_k \frown \mathit{Geometric}(1-(1-\gamma)\,\gamma^k\,\mathcal{H}_\gamma(\gamma^k))$ random variables, and whose i.i.d. summands $Y_{ki} \stackrel{\mathrm{d}}{=} Y_k$ have the p.g.f. $\mathcal{P}_k(s) = \frac{s\,\mathcal{H}_\gamma(\gamma^k s)}{\mathcal{H}_\gamma(\gamma^k)}$.

Proof: We have established that

$$\frac{\mathcal{G}(s)-\mathcal{G}(\gamma s)}{1-\gamma} = s\,\mathcal{G}(s)\,\mathcal{H}_\gamma(s) \quad \Longleftrightarrow \quad \frac{\mathcal{G}(s)}{\mathcal{G}(\gamma s)} = \frac{1}{1-(1-\gamma)\,s\;\mathcal{H}_\gamma(s)} \;.$$

Iterating the above expression, similarly to what we have done to obtain (3.6), we finally get

$$\mathcal{G}(s) \;=\; \prod_{k=0}^{\infty} \frac{1-(1-\gamma)\,\gamma^k \mathcal{H}_\gamma(\gamma^k)}{1-(1-\gamma)\,\gamma^k s\,\mathcal{H}_\gamma(\gamma^k s)} \;. \tag{4.3}$$

If $\gamma \in [0,1)$, we further have

$$\mathcal{G}(s) \;=\; \prod_{k=0}^{\infty} \frac{1-w_k}{1-w_k \frac{s\,\mathcal{H}_\gamma(\gamma^k s)}{\mathcal{H}_\gamma(\gamma^k)}} \;=\; \prod_{k=0}^{\infty} \frac{1-w_k}{1-w_k \mathcal{P}_k(s)}$$

where $w_k = (1-\gamma)\,\gamma^k \mathcal{H}_\gamma(\gamma^k)$, and the $\mathcal{P}_k(s) = \frac{s\,\mathcal{H}_\gamma(\gamma^k s)}{\mathcal{H}_\gamma(\gamma^k)}$ are (unique) probability generating functions such that $\mathcal{P}_k(0) = 0$. □

Theorem 4.2. *Let W be a counting random variable with p.g.f. $\mathcal{G}$, and $\gamma \in (-1,1)$. $W \in \mathcal{C}_\gamma$ iff $\mathcal{H}_\gamma(s) = \frac{\mathcal{G}(s)-\mathcal{G}(\gamma s)}{(1-\gamma)\,s\,\mathcal{G}(s)}$ is abs. mon.*

We can use this result to show that the geometric distribution verifies (4.1) for nonnegative γ. In fact, if $X_\theta \frown \mathit{Geometric}(1-\theta)$, with $0<\theta<1$, we have $r_k = \gamma^k \theta^{k+1} \geq 0$ and $\mathcal{H}_\gamma(1) = \frac{\theta}{1-\gamma\theta} < \frac{1}{1-\gamma}$. Given the uniqueness of the coefficients of $\mathcal{H}_\gamma$, we may also conclude that the geometric distribution does not belong to $\mathcal{C}_\gamma$ when $\gamma \in (-1,0)$.

The truncated geometric distribution with support on the even integers, Y_θ, given by the p.m.f.

$$p_n = \begin{cases} (1-\theta^2)\,\theta^n & \text{if } n = 2k \text{ even} \\ 0 & \text{if } n = 2k+1 \text{ odd} \end{cases} \;, \qquad 0<\theta<1 \;,$$

is an element of $\mathcal{C}_\gamma$ for all $\gamma \in (-1,1]$, since it verifies (4.1) with $r_{2k} = 0$, $r_{2k+1} = (1+\gamma)\,\gamma^{2k}\theta^{2k+2}$, and $\mathcal{H}_\gamma(1) = \frac{(1+\gamma)\theta^2}{1-(\gamma\theta)^2} < \frac{1}{1-\gamma}$.

It's interesting to note that the p.g.f. of Y_θ is $\mathcal{G}_{Y_\theta}(s) = \frac{1-\theta^2}{1-\theta^2 s^2} = \mathcal{G}_{X_{\theta^2}}(s^2)$. It is not difficult to show that if $X \in \mathcal{C}_0$ has the p.g.f. $\mathcal{G}$, then $\mathcal{G}(s^2)$ is the p.g.f. of an element of $\mathcal{C}_\gamma$, for every $\gamma \in (-1,1]$.

Corolary 4.2.1. Let W be a counting random variable with p.g.f. $\mathcal{G}$, and $\gamma \in (-1,1)$. If $W \in \mathcal{C}_\gamma$, $\frac{\mathcal{G}(s)}{\mathcal{G}(\gamma s)}$ is absolutely monotone.

Proof: From the proof of Theorem 4.1, $\frac{\mathcal{G}(s)}{\mathcal{G}(\gamma s)} = \frac{1}{1-(1-\gamma)\, s\, \mathcal{H}_\gamma(s)}$. If $\gamma \in [0,1)$, we have $s\,\mathcal{H}_\gamma(s) \le \mathcal{H}_\gamma(1) < \frac{1}{1-\gamma}$ for $0 \le s \le 1$; on the other hand, if $\gamma \in (-1,0)$ we have $(1-\gamma)s\,\mathcal{H}_\gamma(s) \le (1-\gamma)\,\mathcal{H}_\gamma(1) < 1$ for $0 \le s \le \frac{1}{1-\gamma}$. Thus, it follows from property 3(b) of abs. mon. functions that $\frac{\mathcal{G}(s)}{\mathcal{G}(\gamma s)}$ is abs. mon. (in $[0,1]$ for nonnegative γ, and in $[0, \frac{1}{1-\gamma}]$ for negative γ). □

Corolary 4.2.2. For $\gamma \in (-1,1)$, $\mathcal{C}_\gamma \subset \mathcal{C}_1$.

Proof: Taking derivatives on both sides of $1 - \frac{\mathcal{G}(\gamma s)}{\mathcal{G}(s)} = (1-\gamma)\, s\, \mathcal{H}_\gamma(s)$, we obtain

$$\frac{\mathcal{G}'(s)\,\mathcal{G}(\gamma s) - \gamma\,\mathcal{G}'(\gamma s)\,\mathcal{G}(s)}{\mathcal{G}^2(s)} \;=\; (1-\gamma)\,\frac{\mathrm{d}}{\mathrm{d}s}\big[s\,\mathcal{H}_\gamma(s)\big]\;,$$

equivalent to

$$\frac{\mathcal{G}'(s)}{\mathcal{G}(s)} - \gamma\,\frac{\mathcal{G}'(\gamma s)}{\mathcal{G}(\gamma s)} \;=\; (1-\gamma)\,\frac{\mathcal{G}(s)}{\mathcal{G}(\gamma s)}\,\frac{\mathrm{d}}{\mathrm{d}s}\big[s\,\mathcal{H}_\gamma(s)\big]\;. \tag{4.4}$$

Therefore, in view of Corollary 4.2.1 and of property 1 of abs. mon. functions $\frac{\mathcal{G}'(s)}{\mathcal{G}(s)} - \gamma\,\frac{\mathcal{G}'(\gamma s)}{\mathcal{G}(\gamma s)}$ is abs. mon. which in turn (property 2 of abs. mon. functions) implies that $\frac{\mathcal{G}'(s)}{\mathcal{G}(s)}$ is abs. mon.

The result follows from Corollary 2.1.1. □

The inclusion is strict: the *Poisson*(μ) distribution belongs to $\mathcal{C}_1$ for all $\mu > 0$, but does not belong to $\mathcal{C}_\gamma$ when $\gamma \in (-1,1)$, since from Theorem 4.2 we have $r_k = (-1)^k (1-\gamma)^k \frac{\mu^{k+1}}{(k+1)!}$, so that $\mathcal{H}_\gamma$ is not abs. mon.

Corolary 4.2.3. For $|\gamma_1| \le \gamma_2 < 1$, $\mathcal{C}_{\gamma_1} \subset \mathcal{C}_{\gamma_2}$.

Proof: Let $\mathcal{G}$ be the p.g.f. of a random variable $W \in \mathcal{C}_{\gamma_1} \subset \mathcal{C}_1$.

$$\frac{\mathcal{H}_{\gamma_2}(s) - \gamma_1 \mathcal{H}_{\gamma_2}(\gamma_1 s)}{\mathcal{H}_{\gamma_1}(s) - \gamma_2 \mathcal{H}_{\gamma_1}(\gamma_2 s)} = \frac{1-\gamma_1}{1-\gamma_2}\,\frac{\frac{\mathcal{G}(\gamma_1\gamma_2\, s)}{\mathcal{G}(\gamma_1 s)} - \frac{\mathcal{G}(\gamma_2\, s)}{\mathcal{G}(s)}}{\frac{\mathcal{G}(\gamma_1\gamma_2\, s)}{\mathcal{G}(\gamma_2\, s)} - \frac{\mathcal{G}(\gamma_1\, s)}{\mathcal{G}(s)}} = \frac{1-\gamma_1}{1-\gamma_2}\,\frac{\mathcal{G}(\gamma_2\, s)}{\mathcal{G}(\gamma_1 s)}\;.$$

From Corollary 2.1.1.(3), $\frac{\mathcal{G}(\gamma_2\, s)}{\mathcal{G}(\gamma_1\, s)}$ is abs. mon, and from property 2 of abs. mon. functions $\mathcal{H}_{\gamma_1}(s)-\gamma_2\,\mathcal{H}_{\gamma_1}(\gamma_2\, s)$ is abs. mon. Then $\mathcal{H}_{\gamma_2}(s)-\gamma_1\mathcal{H}_{\gamma_2}(\gamma_1\, s)$, and therefore $\mathcal{H}_{\gamma_2}$, are also abs. mon., which proves that $W\in\mathcal{C}_{\gamma_2}$. □

We can see that the inclusion is strict directly from (4.1). Suppose that $-1<\gamma<\eta<1$ and $0<\beta<\frac{1}{1-\eta}$. We know that $N_{0,\,\beta,\,\eta}\in\mathcal{C}_\eta$, since its p.m.f. satisfies $\frac{1-\eta^{n+1}}{1-\eta}\,p_{n+1}=\beta\,p_n$. Assume that $N_{0,\,\beta,\,\eta}\in\mathcal{C}_\gamma$, that is, $\frac{1-\gamma^{n+1}}{1-\gamma}\,p_{n+1}=\sum_{k=0}^{n}p_k r_{n-k}$. Then $p_1=p_0r_0=\beta\,p_0$ implies $r_0=\beta$, and

$$(1+\gamma)\;p_2\;=\;\frac{\beta}{1+\eta}\;p_1(1+\eta+\gamma-\eta)\;=\;p_1r_0+p_0r_1 \tag{4.5}$$

implies $r_1=-\frac{\eta-\gamma}{1+\eta}\,\beta^2$. But this is negative, therefore $N_{0,\,\beta,\,\eta}\notin\mathcal{C}_\gamma$.

Corolary 4.2.4. Let W be a counting random variable with p.g.f. $\mathcal{G}$, W_γ the random variable with p.g.f. $\mathcal{G}_\gamma(s)=\frac{\mathcal{G}(\gamma)\,\mathcal{G}(s)}{\mathcal{G}(\gamma\, s)}$, and $\gamma\in(-1,1)$. $W\in\mathcal{C}_\gamma$ iff $W_\gamma\in\mathcal{C}_0$.

Proof: As $W\in\mathcal{C}_\gamma\implies W\in\mathcal{C}_1$, from Corollary 2.1.1 we know that $\mathcal{G}_\gamma(s)=\frac{\mathcal{G}(\gamma)\,\mathcal{G}(s)}{\mathcal{G}(\gamma\, s)}$ is a p.g.f.

From the proof of Theorem 2.2 (or simply by taking $\gamma=0$ in Theorem 4.1), in what concerns this p.g.f. $\mathcal{G}_\gamma$ we obtain, with self-explaining notations,

$$\mathcal{H}_0^{(\mathcal{G}_\gamma)}(s)\;=\;\frac{\mathcal{G}_\gamma(s)-\mathcal{G}_\gamma(0)}{s\,\mathcal{G}_\gamma(s)}\;=\;\frac{\mathcal{G}(s)-\mathcal{G}(\gamma s)}{s\,\mathcal{G}(s)}\;=\;(1-\gamma)\,\mathcal{H}_\gamma^{(\mathcal{G})}(s) \tag{4.6}$$

and therefore $\mathcal{H}_0^{(\mathcal{G}_\gamma)}$ is abs. mon. iff $\mathcal{H}_\gamma^{(\mathcal{G})}$ is abs. mon. □

5. FURTHER COMMENTS

1. Geometric infinite divisibility arose from Kovalenko's (1965) extensions of Rényi's (1956) work on random rarefaction, with the general characterization of geometric stable laws given in Kozubowski (1994). This led to a general definition of $\mathcal{N}$-summation schemes, the classical summation scheme being the special case $N_p=\frac{1}{p}$ (degenerate random variables, and therefore a non-random sum of random variables). It is well known that for some families $\mathcal{N}=\{N_p,\ p\in(0,1),\ \mathbb{E}(N_p)=\frac{1}{p}\}$ there exists $\mathcal{N}$-Gaussian laws (for instance for $N_p\frown\mathit{Geometric}(p)$, the corresponding $\mathcal{N}$-Gaussian random variables being the Laplace random variables), while other N_p, for instance $N_p\frown\mathit{Poisson}(\frac{1}{p})$,

do not admit $\mathcal{N}$-Gaussian laws. Although it is easy to prove that in more general branching setings $\mathcal{N}$-Gaussian laws do exist, only the usual Gaussian law and the Laplace geometric–Gaussian law are explicitly exhibited in the references we know.

This research arose from the observation that $\mathcal{C}_0 \subset \mathcal{C}_1$ and that, more generaly, $0 < \gamma_1 < \gamma_2 < 1 \Longrightarrow \mathcal{C}_0 \subset \mathcal{C}_{\gamma_1} \subset \mathcal{C}_{\gamma_2} \subset \mathcal{C}_1$.

Our aim was either to prove that there exist $\gamma \in (0,1)$ such that for $\gamma_1 \leq \gamma$ we could exhibit a $\mathcal{N}$-Gaussian law in $\mathcal{C}_{\gamma_1}$ — which we couldn't — or else to extend $\mathcal{C}_\gamma$ classes for $\gamma < 0$ — which we did — and show that for those it was possible to construct $\mathcal{N}$-Gaussian random variables. Unfortunately for $-1 < \gamma_1 < \gamma_2 < 0$ the chain of inclusions $\mathcal{C}_{\gamma_1} \subset \mathcal{C}_{\gamma_2} \subset \mathcal{C}_0$ is no longer valid.

2. The extension of Katz–Panjer's iterative relation

$$\frac{f(n+1)}{f(n)} = \alpha + \frac{\beta}{n+1} = \alpha + \beta\, \mathbb{E}(U_0^n)\ , \qquad n = 0,1,..., \quad \alpha, \beta \in \mathbb{R}\ , \tag{5.1}$$

by

$$\frac{f(n+1)}{f(n)} = \alpha + \beta\, \frac{\mathbb{E}(U_0^n)}{\mathbb{E}(U_\gamma^n)}\ , \qquad n = 0,1,..., \quad \alpha, \beta \in \mathbb{R}\ , \tag{5.2}$$

where $U_\gamma \frown \mathit{Uniform}(\gamma, 1)$, $\gamma \in (-1, 1]$ may seem arbitrary at this stage, unless it is considered as a first step in extending (5.1) by using more general *Beta*, of which the *Uniform* in (5.2) isn't but a special case, or even more general random variables. Naturaly $\{f(n)\}_{n\in\mathbb{N}}$ is not a p.m.f. unless the restrictions in the parameters are very strong.

3. Panjer's class $\Pi = \Pi^{(0)}$ has been generalized by Sundt and Jewell (1981), who considered the class $\Pi^{(1)}$ of discrete random variables whose probability mass function satisfies

$$f_{\alpha,\beta}(n+1) = \left(\alpha + \frac{\beta}{n+1}\right) f_{\alpha,\beta}(n)\ , \qquad \alpha, \beta \in \mathbb{R}, \quad n = 1,2,... \tag{5.3}$$

Willmot (1987) published the definitive characterization of $\Pi^{(1)}$: the probability mass function of a discrete random variable N, with support $\mathcal{S} = \{1,2,...\}$, satisfies the above expression if N is either a zero-truncated Binomial, Poisson or Negative Binomial random variable, or a Logarithmic (when $\alpha \in (0,1)$ and the index $\frac{\alpha}{\alpha+\beta} \to 0$) or an Engen (1974) Extended Negative Binomial random variable (index $\frac{\alpha}{\alpha+\beta} \in (-1,0)$, where $\alpha \in (0,1]$), and general solutions N^*, with support $\mathcal{S} = \{0,1,2,...\}$, arise from a *hurdle process* (Cameron and Trivedi, 1998, pp. 123–125) $N^* = \begin{cases} 0 & N \\ p_0 & 1-p_0 \end{cases}$, where N is one of the above variables. Klugman *et al.* (1998) describe the solutions as *zero modified* N variables.

Hess, Lewald and Schmidt (2002) considered the even more general setting $\Pi^{(k)}$, $k = 0, 1, ...$, in which the probability mass functions satisfy

$$f_{\alpha,\beta}(n+1) = \left(\alpha + \frac{\beta}{n+1}\right) f_{\alpha,\beta}(n) \ , \qquad \alpha, \beta \in \mathbb{R}, \quad n = k, k+1, ... \ , \tag{5.4}$$

giving a complete description of $\Pi^{(k)}$, $k = 0, 1, ...$ in terms of $\{0, 1..., k-1\}$ modified basic claim number distributions, i.e., the left k-truncated binomial, Poisson, and negative binomial distributions, the other basic claim number distributions are the left truncated *Logarithmic*(k, θ) distribution, and the left truncated *Engen*(k, β, θ) distribution. The extension

$$\frac{f_{\alpha,\beta}(n+1)}{f_{\alpha,\beta}(n)} = \alpha + \beta \, \frac{1-\gamma}{1-\gamma^{n+1}} \ , \qquad \alpha, \beta \in \mathbb{R}, \quad n = k, k+1, ... \ , \tag{5.5}$$

of (3.1) may investigated along similar lines, but with very cumbersome results.

REFERENCES

[1] Bernstein, S. (1928). Sur les fonctions absolument monotones, *Acta Mathematica*, **51**, 1–66.

[2] Cameron, A.C. and Trivedi, P.K. (1998). *Regression Analysis of Count Data*, Cambridge University Press, Cambridge.

[3] Dhaene, J. and Sundt, B. (1998). On approximating distributions by approximating their De Pril transforms, *Scand. Actuar. J.*, 1–23.

[4] Engen, S. (1974). On species frequency models, *Biometrika*, **61**, 263–270.

[5] Feller, W. (1968). *An Introduction to Probability Theory and its Applications*, vol. I, Wiley, New York.

[6] Gurland, J. (1957). Some interrelations among compound and generalized distributions, *Biometrika*, **44**, 265–268.

[7] Hess, K.Th.; Lewald, A. and Schmidt, K.D. (2002). An extension of Panjer's recursion, *ASTIN Bulletin*, **32**, 283–297.

[8] Johnson, N.L.; Kotz, S. and Kemp, A.W. (1992). *Univariate Discrete Distributions*, Wiley, New York.

[9] Katz, L. (1965). *Unified treatment of a broad class of discrete probability distributions*. In "Classical and Contagious Discrete Distributions", Pergamon Press, Oxford, 175–182.

[10] Klugman, S.A.; Panjer, H.H. and Willmot, G.E. (1998). *Loss Models: From Data to Decisions*, Wiley, New York.

[11] Kovalenko, I.N. (1965). On a class of limit distributions for rarefied flows of homogeneous events, *Lit. Mat. Sbornik*, **5**, 569–573. (*Selected Transl. Math. Statist. and Prob.* **9**, Providence, Rhode Island, 1971, 75–81.)

[12] Kozubowski, T.J. (1994). Representation and properties of geometric stable laws, *Approximation, Probability, and Related Fields*, Plenum, New York, 321–337.

[13] Ospina, A.V. and Gerber, H.U. (1987). A simple proof of Feller's characterization of the compound Poisson distribution, *Insurance: Mathematics and Economics*, **6**, 63–64.

[14] Panjer, H.H. (1981). Recursive evaluation of a family of compound distributions, *ASTIN Bulletin*, **12**, 22–26.

[15] Rényi, A. (1956). A characterization of the Poisson process, *MTA Mat. Kut. Int. Közl.*, **1**, 519–527. (English translation: *Selected Papers of Alfred Rényi,* **1***, 1948–1956*, P. Turán, ed., 622–279, Akadémiai Kiadó, Budapest, with a note by D. Szász on ulterior developments up to 1976).

[16] Rólski, T.; Schmidli, H.; Schmidt, V. and Teugels, J. (1999). *Stochastic Processes for Insurance and Finance*, Wiley, New York.

[17] Skellam, J.G. and Shenton, L.R. (1957). Distributions associated with random walks and recurrent events (with discussion), *J. Roy. Statist. Soc.*, **B 19**, 64–118.

[18] Srivastava, H.M. and Manocha, H.L. (1984). *A Treatise on Generating Functions*, Horwood, Chichester.

[19] Sundt, B. and Jewell, W.S. (1981). Further results on recursive evaluation of compound distributions, *ASTIN Bulletin*, **12**, 27–39.

[20] Widder, D.V. (1946). *The Laplace Transform*, Princeton Univ. Press, Princeton.

[21] Willmot, G.E. (1987). Sundt and Jewell's family of discrete distributions, *ASTIN Bulletin*, **18**, 17–29.

REVSTAT – Statistical Journal
Volume 2, Number 2, November 2004

EXTREMAL BEHAVIOUR IN MODELS OF SUPERPOSITION OF RANDOM VARIABLES *

Authors: Luísa Pereira
– Department of Mathematics, University of Beira Interior, 6200 Covilhã, Portugal (lpereira@noe.ubi.pt)

Received: February 2004 Revised: October 2004 Accepted: October 2004

Abstract:

- Let $\mathbf{X}^{(i)} = \{X_{g_i(n)}\}_{n\geq 1}$, $i = 1, 2$, be sequences of random variables, where $\{g_i(n)\}_{n\geq 1}$ are disjoint and strictly increasing sequences of integer numbers such that $\{g_1(n)\}_{n\geq 1} \cup \{g_2(n)\}_{n\geq 1} = \mathbb{N}$. Using superposition of point processes, we study the extremal behaviour of a superposed sequence

$$\{X_n\}_{n\geq 1} \; = \; \{X_{g_1(n)}\}_{n\geq 1} \cup \{X_{g_2(n)}\}_{n\geq 1} \; ,$$

  where we consider the proportion of variables superposed from each sequence asymptotically constant and $\{X_n\}_{n\geq 1}$ verifying some dependence conditions. We apply the obtained results in the computation of the bivariate extremal index.

Key-Words:

- *extreme value; nonstationarity; extremal index; superposition of point processes.*

AMS Subject Classification:

- 60G55, 60G70.

*Research partially suported by FCT/POCTI/FEDER.

1. INTRODUCTION

Let $\mathbf{X}^{(i)} = \{X_{g_i(n)}\}_{n\geq 1}$ be stationary sequences of random variables on the same probability space $(\Omega, \Im, P)$ with common distribution function $F^{(i)}$, $i=1,2$, respectively. Let us suppose that $\{g_i(n)\}_{n\geq 1}$, $i=1,2$, are disjoint and strictly increasing sequences of integer numbers, such that

$$\{g_1(n)\}_{n\geq 1} \cup \{g_2(n)\}_{n\geq 1} \,=\, \mathbb{N}\ .$$

In this paper we consider sequences that arise from the superposition of the variables of the sequences $\mathbf{X}^{(i)}$, $i=1,2$, when considering asymptotically constant the proportion of variables to superpose from each of the sequences, that is,

$$\frac{s_i(n)}{n} \xrightarrow[n\to\infty]{} L_i\ , \qquad i=1,2,\quad L_1+L_2=1\ , \tag{1.1}$$

where $s_i(n) = \#\big\{g_i(j)\colon 1\leq j\leq n \,\wedge\, 1\leq g_i(j)\leq n\big\}$, $i=1,2$.

We study the extremal limiting behaviour of the superposed sequence

$$\{X_n\}_{n\geq 1} \,=\, \{X_{g_1(n)}\}_{n\geq 1}\,\{X_{g_2(n)}\}_{n\geq 1}\ ,$$

usually a nonstationary sequence. Such a behaviour is derived from the convergence in distribution of the sequence, $\{S_n\}_{n\geq 1}$, of the point processes of exceedances of real numbers u_n, $n\geq 1$, generated by the sequence $\{X_n\}_{n\geq 1}$, defined by

$$S_n(B) \,=\, S_n[X_i,u_n](B) \,=\, \sum_{i=1}^{n} 1_{\{X_i>u_n\}}\,\delta_{\frac{i}{n}}(B)\ , \qquad n\geq 1\ ,$$

where B is a Borel subset of $[0,1]$, $\delta_x(\cdot)$ denotes the Dirac measure at $x\in\mathbb{R}$ and 1_A the indicator function of the event A.

By considering, for each $i=1,2$,

$$S_n^{(i)}(B) \,=\, S_n[X_{g_i(j)},u_n](B) \,=\, \sum_{j=1}^{n} 1_{\{X_{g_i(j)}>u_n\}}\,\delta_{\frac{g_i(j)}{n}}(B)\ ,$$

then

$$S_n(B) = S_n^{(1)}(B) + S_n^{(2)}(B)\ ,$$

that is, the sequence of point processes $\{S_n\}_{n\geq 1}$ is the superposition of the point processes $\{S_n^{(i)}\}_{n\geq 1}$, $i=1,2$.

We briefly present, in what follows, some important results concerning the theory of exceedances point processes generated by dependent sequences, both stationary and nonstationary.

Recall that the type of long range dependence condition appropriate for studying the convergence in distribution of $\{S_n\}_{n\geq 1}$ is the condition $\Delta(u_n)$ defined by Hsing *et al.* (1988), in the following way.

Definition 1.1. Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables and $\{u_n\}_{n\geq 1}$ a sequence of real numbers. For each $1 \leq i \leq j$, set $\mathcal{B}_i^j(u_n)$ as the σ-field generated by the events $\{X_s \leq u_n\}$, $i \leq s \leq j$, and, for $1 \leq l \leq n-1$,

$$\alpha_{n,l} = \sup_{1\leq k\leq n-l} \Big\{ \big| P(A\cap B) - P(A)\,P(B) \big| :\ A \in \mathcal{B}_1^k(u_n),\ B \in \mathcal{B}_{k+l}^n(u_n) \Big\} \ . \tag{1.2}$$

The condition $\Delta(u_n)$ is said to hold if there exists a sequence $l_n = o(n)$, as $n \to \infty$, such that

$$\alpha_{n,l_n} \xrightarrow[n\to\infty]{} 0 \ .$$

Note that by taking in (1.2) only events of the form $A = \{X_{i_1} < u_n,\ ...,\ X_{i_p} < u_n\}$ and $B = \{X_{j_1} < u_n,\ ...,\ X_{j_q} < u_n\}$ with

$$1 \leq i_1 < \cdots < i_p < i_p + l < j_1 < \cdots < j_q \leq n \ ,$$

we obtain Leadbetter's $D(u_n)$ condition.

Under condition $\Delta(u_n)$ and additional assumptions of equicontinuity and asymptotic negligibility, Nandagopalan (1990) characterized the possible distributional limits for $\{S_n\}_{n\geq 1}$, as stated in Proposition 1.1.

Let $J_1, ..., J_{k_n}$, $n \geq 1$, be a sequence of partitions of $[0,1]$ such that for each $i = 1, 2, ..., k_n$, $P(S_n(J_i) > 0) > 0$, after certain order n_0. For each $n \geq n_0$ define the following sequences of measures:

$$\nu_n(B) = \sum_{i=1}^{k_n} P\big(S_n(J_i) > 0\big)\, \frac{m(B\cap J_i)}{m(J_i)} \ , \qquad B \in \mathcal{B}([0,1]) \ ,$$

where m denotes the Lebesgue measure,

$$\Pi_{n,x}(k) = \sum_{i=1}^{k_n} \Pi_{n,i}(k)\, \delta_x(J_i) \ , \qquad k \in \mathbb{N},\quad x \in [0,1] \ ,$$

where

$$\Pi_{n,i}(k) = P\Big(S_n(J_i) = k \mid\ S_n(J_i) > 0\Big) \ , \qquad k \in \mathbb{N} \ .$$

Finally for each $a \in \mathbb{R}_+$ define the functions

$$g_{n,a}(x) \ = \int_{\mathbb{N}} \big(1 - \exp(ak)\big)\, d\Pi_{n,x}(k) \ .$$

Proposition 1.1. *Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables verifying condition $\Delta(u_n)$,*

$$g(\epsilon_n) = \sup\Big\{P\big(S_n(I)>0\big): I\subset[0,1],\ m(I)\leq\epsilon_n\Big\} \xrightarrow[n\to\infty]{} 0 \quad \text{if } \epsilon_n \xrightarrow[n\to\infty]{} 0 \tag{1.3}$$

and

$$\liminf_{n\to\infty} P\big(S_n([0,1])=0\big) > 0\ . \tag{1.4}$$

If $\{k_n\}_{n\geq 1}$ is a sequence of integer numbers such that

$$k_n\left(\alpha_{n,l_n} + g\left(\frac{l_n}{n}\right)\right) \xrightarrow[n\to\infty]{} 0 \tag{1.5}$$

and $J_1, ..., J_{k_n}$, is a partition of $[0,1]$ satisfying

$$\max\Big\{m(J_i):\ i=1,2,...,k_n\Big\} \xrightarrow[n\to\infty]{} 0$$

and

> *for each $a\in G$, where G is some nonempty open subset of $\mathbb{R}_+$,*
> *the sequence $\{g_{n,a}\}_{n\geq 1}$ is equicontinuous,*

then the following propositions are equivalent

(1) *The sequence of point processes $\{S_n\}_{n\geq 1}$ converges in distribution to some point process S with Laplace transform, L_S, given by*

$$L_S(f) = \exp\left(-\int_{[0,1]}\int_{\mathbb{N}}\Big(1-\exp\big(-kf(x)\big)\Big)\,d\Pi_x(k)\ d\mu(x)\right)\ , \tag{1.6}$$

for each non-negative measurable function, f, on $[0,1]$, where μ is a finite measure on $[0,1]$ and Π_x is a probability measure on $\mathbb{N}$.

(2) *ν_n converges weakly to a finite measure μ and $\Pi_{n,x}$ converges weakly to a probability measure Π_x on $\mathbb{N}$, for each $x\in[0,1]$.*

Furthermore, Nandagopalan (1990) proves that under conditions (1.3), (1.4) and (1.5) for some partition $J_1, ..., J_{k_n}$ of $[0,1]$ such that $\max\{m(J_i): i=1,2,...,k_n\} \xrightarrow[n\to\infty]{} 0$, if $S_n \xrightarrow[n\to\infty]{d} S$, the Laplace Transform L_S is given by (1.6).

The result of Hsing *et al.* (1988) which gives the convergence in distribution of exceedances point processes of a stationary random sequence is contained in the preceeding proposition. In fact, in the case of stationary sequences for normalized levels and sequences of integer numbers $\{k_n\}_{n\geq 1}$, such that

$$k_n \xrightarrow[n\to\infty]{} \infty\ , \qquad k_n\alpha_{n,l_n} \xrightarrow[n\to\infty]{} 0\ , \qquad \frac{k_n l_n}{n} \xrightarrow[n\to\infty]{} 0\ , \tag{1.7}$$

the assumptions established in the above proposition are verified and, furthermore the multiplicity distribution does not depend on the position of the atom, $\Pi_x = \Pi$, for each $x \in [0,1]$, and the intensity measure μ is equal to a constant times the Lebesgue measure, $\mu(\cdot) = \nu\, m(\cdot)$.

For the class of stationary sequences verifying condition $\Delta(u_n)$, if there exists the extremal index $\theta \in [0,1]$ (Leabetter (1974)), then such a parameter is given by the inverse of the limiting mean cluster size of exceedances. Indeed, if

$$P\big(S_n([0,1])=0\big) \xrightarrow[n\to\infty]{} e^{-\nu}$$

and

$$ES_n([0,1]) = nP(X_1 > u_n) \xrightarrow[n\to\infty]{} \tau > 0$$

then

$$\begin{aligned} \theta &= \Big(\lim_{n\to\infty} E\,\Pi_n\Big)^{-1} \\ &= \lim_{n\to\infty} \frac{P\Big(S_n\big([0,k_n^{-1}]\big) > 0\Big)}{E\,S_n([0,1])} \\ &= \frac{\nu}{\tau}\ . \end{aligned}$$

In section 2 we introduce a condition that guarantees, locally, the asymptotic independence among the maxima of the variables of the sequences, $X^{(i)}$, $i = 1, 2,$ to superpose. Under this condition, for each non-negative integer k, the probability of occurrence of k exceedances of the level u_n by the variables of the superposed sequence $\{X_n\}_{n\geq 1}$, in intervals of length $[\frac{n}{k_n}]$, is asymptotically equal.

For each sequence of this class we can apply the results stated in Proposition 1.1, obtaining a compound Poisson limit $S[\nu, \Pi]$ to $\{S_n\}_{n\geq 1}$. The sequence $\{S_n\}_{n\geq 1}$ behaves asymptotically as though the sequence $\{X_n\}_{n\geq 1}$ is stationary, that is, the multiplicity distribution does not depend on x, $\Pi_x = \Pi$, for each $x \in [0,1]$, and the intensity measure is equal a constant times the Lebesgue measure, $\mu(\cdot) = \nu\, m(\cdot)$.

The relations between the intensity measure $\nu\, m(\cdot)$, the distribution of multiplicities $\Pi(\cdot)$ and the corresponding measures $\nu^{(i)} m(\cdot)$ and $\Pi^{(i)}(\cdot)$, for each of the sequences to superpose, will be analyzed in section 3. We prove that $\nu = \nu^{(1)} + \nu^{(2)}$ and $\Pi(k) = \sum_{i=1}^{2} \frac{\nu^{(i)}}{\nu}\, \Pi^{(i)}(k)$, with $\nu^{(i)} = \theta^{(i)}\, \tau^{(i)} L_i$, $\tau^{(i)} = \lim\limits_{n\to\infty} nP(X_{g_i(1)} > u_n)$ and L_i is given in (1.1), $i = 1, 2$.

In section 4 we will apply the results in the computation of the bivariate extremal index.

2. LIMIT DISTRIBUTION OF THE NUMBER OF EXCEEDANCES IN THE SUPERPOSED SEQUENCE

We define a new condition that guarantees locally, that the maxima of the random variables of the sequences to superpose are asymptotically independent. This condition will be essential to obtain the results in this section.

Definition 2.1. The sequence $\{X_n\}_{n\geq 1}$ verifies the condition $\overset{\bullet}{D}(u_n)$ if

$$k_n\,\beta_n \xrightarrow[n\to\infty]{} 0$$

where

$$\beta_n = \sup\left\{ \left| P\Big(M_n^{(1)}(J)\leq u_n,\, M_n^{(2)}(J)\leq u_n\Big) - P\Big(M_n^{(1)}(J)\leq u_n\Big)P\Big(M_n^{(2)}(J)\leq u_n\Big)\right| : J\subset[0,+\infty[,\ \ m(J)=\left[\frac{n}{k_n}\right]\right\},$$

$M_n^{(i)}(J) = \max\big\{X_{g_i(j)} : 1\leq g_i(j)\leq n,\ g_i(j)\in J\big\}$ and $\{k_n\}_{n\geq 1}$ is a sequence of integer numbers that verifies (1.7).

Under condition $\overset{\bullet}{D}(u_n)$ for the superposed sequence $\{X_n\}_{n\geq 1}$ we can, for each non-negative integer k, approach

$$P\big(S_n(J_j)=k\big) \quad \text{by} \quad P\big(S_n(J_l)=k\big)$$

where $J_i=\big[(i-1)[\frac{n}{k_n}]\frac{1}{n},\, i[\frac{n}{k_n}]\frac{1}{n}\big]$, $j,l\in\{1,2,...,k_n\}$ and $j\neq l$.

Proposition 2.1. *Suppose that the sequence $\{X_n\}_{n\geq 1}$ resulting from the superposition of the variables of the stationary sequences $\{X_{g_i(n)}\}_{n\geq 1}$, $i=1,2$, verifies condition $\overset{\bullet}{D}(u_n)$, where $\{u_n\}_{n\geq 1}$ is a sequence of real numbers such that*

$$nP\big(X_{g_i(1)}>u_n\big) \xrightarrow[n\to\infty]{} \tau^{(i)}\ , \qquad i=1,2\ . \tag{2.1}$$

Then, for each non-negative integer k, we have

$$k_nP\big(S_n(J_i)=k\big) \;=\; k_nP\big(S_n(J_1)=k\big)+o(1)\ .$$

Proof: Since $\{X_{g_i(n)}\}_{n\geq 1}$, $i=1,2$, are stationary sequences

$$\begin{aligned} k_nP\big(S_n(J_i)=k\big) \;&=\; k_nP\big(S_n^{(1)}(J_i)=k\big)+k_nP\big(S_n^{(2)}(J_i)=k\big) \\ &\quad + k_n\sum_{\substack{s_1+s_2=k\\ s_1>0,\,s_2>0}} P\Big(S_n^{(1)}(J_i)=s_1,\, S_n^{(2)}(J_i)=s_2\Big) \\ &= k_nP\big(S_n^{(1)}(J_1)=k\big)+k_nP\big(S_n^{(2)}(J_1)=k\big) \\ &\quad + k_n\sum_{\substack{s_1+s_2=k\\ s_1>0,\,s_2>0}} P\Big(S_n^{(1)}(J_i)=s_1,\, S_n^{(2)}(J_i)=s_2\Big)\ . \end{aligned}$$

Attending now to condition $\overset{\bullet}{D}(u_n)$ we can write

$$\begin{aligned} k_n \sum_{\substack{s_1+s_2=k \\ s_1>0,\, s_2>0}} P\Big(S_n^{(1)}(J_i)=s_1,\, S_n^{(2)}(J_i)=s_2\Big) \;\leq \\ \leq\; k_n P\Big(S_n^{(1)}(J_i)>0,\, S_n^{(2)}(J_i)>0\Big) \\ \leq\; k_n\beta_n + k_n P\big(S_n^{(1)}(J_i)>0\big)\, P\big(S_n^{(2)}(J_i)>0\big) \\ =\; o(1)\ . \end{aligned}$$

So

$$k_n P\big(S_n(J_i)=k\big) \;=\; k_n P\big(S_n(J_1)=k\big) + o(1)\ . \qquad \square$$

We will prove next that when the superposed sequence verifies conditions $\overset{\bullet}{D}(u_n)$ and $\Delta(u_n)$ we can apply to it the results stated in Proposition 1.1. Furthermore, and as said before, the sequence $\{S_n\}_{n\geq 1}$ behaves asymptotically as though the sequence $\{X_n\}_{n\geq 1}$ is stationary, that is, the multiplicity distribution does not depend on x, $\Pi_x = \Pi$, for each $x \in [0,1]$, and the intensity measure is equal a constant times the Lebesgue measure, $\mu(\cdot) = \nu\, m(\cdot)$.

Proposition 2.2. *Suppose that the superposed sequence $\{X_n\}_{n\geq 1}$ verifies conditions $\Delta(u_n)$ and $\overset{\bullet}{D}(u_n)$, where $\{u_n\}_{n\geq 1}$ is a sequence of real numbers verifying (2.1). If the sequence $\{S_n\}_{n\geq 1}$ converges, then we have $S_n \xrightarrow[n\to\infty]{d} S[\nu,\Pi]$, with $\nu = \lim_{n\to\infty} k_n P\big(S_n([0,k_n^{-1}])>0\big)$ and Π is a probability measure such that $\Pi(k) = \lim_{n\to\infty} P\big(S_n([0,k_n^{-1}])=k \mid S_n([0,k_n^{-1}])>0\big)$, $k \in \mathbb{N}$.*

Proof: We are going to prove that the superposed sequence satisfies the assumptions of Proposition 1.1 with $J_i = \big((i-1)[\frac{n}{k_n}]\frac{1}{n},\, i[\frac{n}{k_n}]\frac{1}{n}\big]$, $i=1,2,...,k_n$.

For $I \subset [0,1]$ with $m(I) \leq \epsilon_n$ and $\epsilon_n \xrightarrow[n\to\infty]{} 0$ we have

$$P\big(S_n(I)>0\big) \;\leq\; n\,\epsilon_n \max\Big(P(X_{g_1(1)}>u_n),\, P(X_{g_2(1)}>u_n)\Big) \;=\; o(1)\ ,$$

since, for each $i=1,2$, the sequence $\{X_{g_i(n)}\}_{n\geq 1}$ verifies (2.1).

For each set $I \subset [0,1]$ with Lebesgue measure not greater than $\frac{l_n}{n}$ we also have

$$k_n P\big(S_n(I)>0\big) \;\leq\; k_n \frac{l_n}{n}\, n \max\Big(P(X_{g_1(1)}>u_n),\, P(X_{g_2(1)}>u_n)\Big) \;=\; o(1)\ ,$$

because $\{k_n\}_{n\geq 1}$ is a sequence of integer numbers verifying (1.7).

Since

$$\liminf_{n\to\infty} P\big(S_n([0,1])=0\big) \;=\; 1 - \limsup_{n\to\infty} P\big(S_n([0,1])>0\big)$$

and

$$\begin{aligned}\limsup_{n\to\infty} P\big(S_n([0,1])>0\big) &\le \limsup_{n\to\infty} \max\Big(P\big(S_n^{(1)}([0,1])>0\big),\, P\big(S_n^{(2)}([0,1])>0\big)\Big)\\ &\le \max\big(e^{-\nu^{(1)}}, e^{-\nu^{(2)}}\big)\\ &< 1\ ,\end{aligned}$$

we obtain (1.4).

For each $a\in\mathbb{R}_+$, the sequence $\{g_{n,a}\}_{n\ge1}$ is equicontinuous since if $|x-x'|<\varepsilon$ we have from a certain order

$$\big|g_{n,a}(x) - g_{n,a}(x')\big| \le \sum_{k\ge1} \big|\Pi_{n,i}(k) - \Pi_{n,j}(k)\big|\ ,$$

for some pair of indexes i and j in $\{1,...,k_n\}$ such that J_i and J_j are separated by a length not greater than ε.

By Proposition 2.1, for each Borel subset B of $[0,1]$, we have

$$\begin{aligned}\nu_n(B) &= \sum_{i=1}^{k_n} P\big(S_n(J_i)>0\big)\,\frac{m(B\cap J_i)}{m(J_i)}\\ &= \sum_{i=1}^{k_n}\Big(P\big(S_n(J_1)>0\big) + o(k_n^{-1})\Big)\,\frac{m(B\cap J_i)}{m(J_i)}\\ &= k_n\, m(B)\, P\big(S_n(J_1)>0\big) + o(k_n^{-1})\, k_n\, m(B)\\ &= k_n\, m(B)\, P\big(S_n(J_1)>0\big) + o(1)\ .\end{aligned}$$

Under condition $\Delta(u_n)$ it follows, by the Lemma of asymptotic independence of maxima over disjoint intervals (Leadbetter (1974)), that

$$\begin{aligned}\exp(-\nu) &= \lim_n P\big(\max_{1\le i\le n} X_i \le u_n\big)\\ &= \lim_n \prod_{i=1}^{k_n} P\big(S_n(J_i)=0\big)\ ,\end{aligned}$$

so, by Proposition 2.1,

$$\exp(-\nu) = \lim_n P^{k_n}\big(S_n(J_1)=0\big)$$

and consequently

$$\nu = \lim_n k_n P\big(S_n(J_1)>0\big)\ .$$

Thus, $\nu_n \xrightarrow[n\to\infty]{w} \nu\, m$.

Finally, we observe that for each $i=1,2,...,k_n$,

$$\Pi_{n,i}(k) = \frac{P\big(S_n(J_i)=k\big)}{P\big(S_n(J_i)>0\big)} = \frac{k_n P\big(S_n(J_1)=k\big) + o(1)}{k_n P\big(S_n(J_1)>0\big) + o(1)}$$

and as a consequence,

$$\begin{aligned}\Pi_{n,x}(k) &= \sum_{i=1}^{k_n} \Pi_{n,i}(k)\,\delta_x(J_i) \\ &= \Pi_{n,1}(k)\sum_{i=1}^{k_n}\delta_x(J_i) + o(1)\sum_{i=1}^{k_n}\delta_x(J_i) \\ &= \Pi_{n,1}(k) + o(1)\ , \qquad \text{independent of } x\ .\end{aligned}$$

By Proposition 1.1 we can conclude that if the sequence $\{S_n\}_{n\geq 1}$ converges in distribution then the limit point process, S, has Laplace Transform given by

$$L_S(f) = \exp\left(-\nu\int_{[0,1]}\int_{\mathbb{N}}\left(1-e^{-kf(x)}\right)d\Pi(k)\,dx\right)$$

that is, $\{S_n\}_{n\geq 1}$ converges to a compound Poisson process with intensity measure ν and multiplicity distribution Π. □

It must be noted that under condition $\Delta(u_n)$ for the superposed sequence $\{X_n\}_{n\geq 1}$ we also have the validation of such condition for the sequences to superpose, $\{X_{g_i(n)}\}_{n\geq 1}$, $i=1,2$, and so, if the sequence $\{S_n^{(i)}\}_{n\geq 1}$ converges then the limit point process is a compound Poisson process, $S[\nu^{(i)},\Pi^{(i)}]$, $i=1,2$.

3. DESCRIPTION OF THE ASYMPTOTIC BEHAVIOUR OF THE SEQUENCE $\{S_n\}_{n\geq 1}$ FROM $\{S_n^{(i)}\}_{n\geq 1}$, $i=1,2$

The condition $\overset{\bullet}{D}(u_n)$ allow us to describe the asymptotic behaviour of $\{S_n\}_{n\geq 1}$ from $\{S_n^{(i)}\}_{n\geq 1}$, $i=1,2$, as presented in the next result.

Proposition 3.1. *Suppose that the conditions of Proposition 2.2 hold, the sequences $\{X_{g_i(n)}\}_{n\geq 1}$, $i=1,2$, have extremal indexes $\theta^{(i)}$, $i=1,2$, respectively, and the proportion of variables to superpose from each of these sequences is asymptotically constant as established in (1.1).*

If, for each $i=1,2$, we have

$$S_n^{(i)} \xrightarrow[n\to\infty]{d} S[\nu^{(i)},\Pi^{(i)}]$$

then

$$k_n P\big(S_n([0,k_n^{-1}])>0\big) \xrightarrow[n\to\infty]{} \nu = \nu^{(1)}+\nu^{(2)}$$

and

$$\Pi_n(k) = P\Big(S_n([0,k_n^{-1}])=k \mid S_n([0,k_n^{-1}])>0\Big) \xrightarrow[n\to\infty]{} \Pi(k) = \sum_{i=1}^{2}\frac{\nu^{(i)}}{\nu}\,\Pi^{(i)}(k)\ ,$$

with $\nu^{(i)}=\theta^{(i)}\,\tau^{(i)}L_i$ and $\tau^{(i)}$ given in (2.1), $i=1,2$.

Proof: By using analogous arguments to the ones used in the proof of Proposition 2.1, we obtain

$$\begin{aligned}\lim_{n\to\infty} k_n P\big(S_n([0,k_n^{-1}])>0\big) \;&=\\ &= \lim_{n\to\infty} k_n P\big(S_n^{(1)}([0,k_n^{-1}])>0\big) + \lim_{n\to\infty} k_n P\big(S_n^{(2)}([0,k_n^{-1}])>0\big)\\ &\quad - \lim_{n\to\infty} k_n P\Big(S_n^{(1)}([0,k_n^{-1}])>0,\, S_n^{(2)}([0,k_n^{-1}])>0\Big)\\ &= \nu^{(1)}+\nu^{(2)}\ .\end{aligned}$$

Relatively to the cluster size of exceedances distribution we have

$$\begin{aligned}\Pi_n(k) &= P\Big(S_n([0,k_n^{-1}])=k \mid S_n([0,k_n^{-1}])>0\Big)\\ &= \frac{1}{P\big(S_n([0,k_n^{-1}])>0\big)}\Big(P\big(S_n^{(1)}([0,k_n^{-1}])=k\big)+P\big(S_n^{(2)}([0,k_n^{-1}])=k\big)\Big)\\ &\quad + \frac{1}{P\big(S_n([0,k_n^{-1}])>0\big)}\,P\left(\bigcup_{\substack{s_1\ge1,\,s_2\ge1\\ s_1+s_2=k}}\Big(S_n^{(1)}([0,k_n^{-1}])=s_1,\,S_n^{(2)}([0,k_n^{-1}])=s_2\Big)\right).\end{aligned}$$

Since, for each $i=1,2$, we have

$$\frac{P\big(S_n^{(i)}([0,k_n^{-1}])=k\big)}{P\big(S_n([0,k_n^{-1}])>0\big)} = \Pi_n^{(i)}(k)\,\frac{k_n P\big(S_n^{(i)}([0,k_n^{-1}])>0\big)}{k_n P\big(S_n([0,k_n^{-1}])>0\big)} \;\xrightarrow[n\to\infty]{}\; \frac{1}{\nu}\,\Pi^{(i)}(k)\,\nu^{(i)}\ ,$$

and, under condition $\overset{\bullet}{D}(u_n)$

$$\begin{aligned}&\frac{1}{P\big(S_n([0,k_n^{-1}])>0\big)}\,P\left(\bigcup_{\substack{s_1\ge1,\,s_2\ge1\\ s_1+s_2=k}}\Big(S_n^{(1)}([0,k_n^{-1}])=s_1,\,S_n^{(2)}([0,k_n^{-1}])=s_2\Big)\right) \le\\ &\qquad\le \frac{1}{P\big(S_n([0,k_n^{-1}])>0\big)}\,P\Big(S_n^{(1)}([0,k_n^{-1}])>0,\,S_n^{(2)}([0,k_n^{-1}])>0\Big)\\ &\qquad= o(1)\ ,\end{aligned}$$

the result follows. □

Corollary 3.1. *Under the conditions of the Proposition 3.1, the extremal index of the superposed sequence verifies*

$$\theta \;=\; \frac{\theta^{(1)}\tau^{(1)}L_1+\theta^{(2)}\tau^{(2)}L_2}{\tau^{(1)}L_1+\tau^{(2)}L_2}\ .$$

Note that the extremal index of the superposed sequence depends on $\lim_n \sum_{i=1}^n P(X_i>u_n)=\tau^{(1)}L_1+\tau^{(2)}L_2$, as expected since $\{X_n\}_{n\ge1}$ is a non-stationary sequence.

We finish this section with some remarks about the results obtained previously.

Remark 3.1. Let us suppose that $\theta^{(1)}=\theta^{(2)}$, $F_1 \neq F_2$ and for each $i=1,2$, F_i belongs to the domain of attraction of an extreme value distribution, G. So we have

$$P\left(\max_{1\leq i\leq n} X_i \leq u_n(x)\right) \xrightarrow[n\to\infty]{} G^{\theta^{(1)}}(x) \ .$$

We have in this way found a class of non-stationary sequences for which the so-called Extremal Types Theorem of Leadbetter is still valid.

Remark 3.2. Suppose that the superposed sequence is stationary. Then, under the long range dependence condition $\Delta(u_n)$ if $\{S_n\}_{n\geq 1}$ converges in distribution to some point process S then S is necessarily a compound Poisson process and the extremal index $\theta = \lim_{n\to\infty} \frac{P\left(S_n([0,k_n^{-1}])>0\right)}{ES_n([0,1])}$.

So it seems natural to ask: When the superposed sequence is stationary are there any advantages in the application of the results established in this section? We shall find an affirmative answer.

In the stationary case, the introduction of local dependence conditions (Leadbetter (1983), Leadbetter and Nandagopalan (1989), Ferreira (1994)) enables us to obtain processes with practical interest to compute the extremal index, θ.

By assuming that each sequence $X^{(i)}$, $i=1,2$, does not oscillate rapidly near high extremes in the sense of the usual local dependence conditions we have not, in general, the validition of these conditions by the superposed sequence and consequently we can not apply directly to $\{X_n\}_{n\geq 1}$ the avaiable results.

In this case, the application of Proposition 3.1 facilitates the computation of the extremal index θ of the superposed sequence since we can apply the results under local dependence conditions to each one of the sequences superposed.

4. APPLICATIONS

As an application of the results established previously we point out the computation of the extremal index of a stationary sequence of random vectors $\mathbf{X} = \{(X_n^{(1)}, X_n^{(2)})\}_{n\geq 1}$ with common distribution function, F, belonging to the domain of attraction of a bivariate extreme value distribution, G.

Let us denote by $\widehat{\mathbf{X}}$ the independent sequence associated with $\mathbf{X}$ and by $\max_{1\leq j\leq n} \widehat{X}_j^{(i)}$, $n\geq 1$, $i=1,2$, the corresponding sequences of partial maxima.

We remember the definition of bivariate extremal index introduced by Nandagopalan (1990) and that is a generalization of Leadbetter's definition for unidimensional sequences.

Definition 4.1. The sequence $\mathbf{X}=\{(X_n^{(1)},X_n^{(2)})\}_{n\geq 1}$ has an extremal index $\theta(\tau^{(1)},\tau^{(2)})\in[0,1]$, $\tau=(\tau^{(1)},\tau^{(2)})\in\mathbb{R}_+^2$, when for each $\tau\in\mathbb{R}_+^2$, there are $u_n^{(\tau)}=(u_n^{(\tau^{(1)})},u_n^{(\tau^{(2)})})$, $n\geq 1$, verifying

$$nP\Big(X_1^{(i)}>u_n^{(\tau^{(i)})}\Big)\xrightarrow[n\to\infty]{}\tau^{(i)}\ ,\qquad i=1,2\ ,$$

$$P\left(\max_{1\leq j\leq n}\widehat{X}_j^{(1)}\leq u_n^{(\tau^{(1)})},\ \max_{1\leq j\leq n}\widehat{X}_j^{(2)}\leq u_n^{(\tau^{(2)})}\right)\xrightarrow[n\to\infty]{}G(\tau)$$

and

$$P\left(\max_{1\leq j\leq n}X_j^{(1)}\leq u_n^{(\tau^{(1)})},\ \max_{1\leq j\leq n}X_j^{(2)}\leq u_n^{(\tau^{(2)})}\right)\xrightarrow[n\to\infty]{}G(\tau)^{\theta(\tau)}\ .$$

If $\mathbf{X}$ has extremal index $\theta(\tau)$ then, for each $i=1,2$, $\{X_n^{(i)}\}_{n\geq 1}$ has extremal index $\theta^{(i)}=\lim\limits_{\substack{\tau^{(j)}\to 0^+\\ j\neq i}}\theta(\tau^{(1)},\tau^{(2)})$.

We shall assume, without loss of generality, that the common distribution F of the vectors of the stationary sequence $\mathbf{X}=\{(X_n^{(1)},X_n^{(2)})\}_{n\geq 1}$ has unit Fréchet margins, *id est*,

$$F_1(x)=F_2(x)=\exp(-x^{-1})\ ,\qquad x>0\ .$$

For fixed $\tau^{(1)}$ and $\tau^{(2)}$ and normalized levels $u_n^{(\tau^{(i)})}=\frac{n}{\tau^{(i)}}$ for $\{X_n^{(i)}\}_{n\geq 1}$, $i=1,2$, we have

$$\begin{aligned}(4.1)\qquad P\left(\max_{1\leq j\leq n}X_j^{(1)}\leq u_n^{(\tau^{(1)})},\ \max_{1\leq j\leq n}X_j^{(2)}\leq u_n^{(\tau^{(2)})}\right)&=\\ &=P\left(\max_{1\leq j\leq n}X_j^{(1)}\leq\frac{n}{\tau^{(1)}},\ \max_{1\leq j\leq n}X_j^{(2)}\leq\frac{n}{\tau^{(2)}}\right)\\ &=P\left(\max_{1\leq j\leq n}\tau^{(1)}X_j^{(1)}\leq n,\ \max_{1\leq j\leq n}\tau^{(2)}X_j^{(2)}\leq n\right)\ .\end{aligned}$$

Let us consider the stationary sequences $\{X_{g_i(n)}=\tau^{(i)}X_n^{(i)}\}_{n\geq 1}$, $i=1,2$. By superposing the variables of these sequences we can form different sequences $\{X_n\}_{n\geq 1}$ but the limiting behaviour of $\{S_n\}_{n\geq 1}$ is only affected by the asymptotic proportion of variables to superpose from each one of these sequences and not by the order of the variables of the superposed sequence.

By considering, for example,

$$\{X_n\}_{n\geq 1}=\left\{\tau^{(1)}X_1^{(1)},\tau^{(2)}X_1^{(2)},\tau^{(1)}X_2^{(1)},\tau^{(2)}X_2^{(2)},...,\tau^{(1)}X_n^{(1)},\tau^{(2)}X_n^{(2)}\right\}_{n\geq 1}\ ,$$

we can rewrite (4.1) in the following way

$$P\Big(\max_{1\le j\le 2n} X_j \le n\Big)\ . \tag{4.2}$$

Since, for each $i=1,2$,

$$\frac{s_i(n)}{n} = \frac{\frac{n}{2}}{n} \xrightarrow[n\to\infty]{} L_i = \frac{1}{2}\ ,$$

then, under the conditions established in Proposition 3.1, we have

$$\lim_{n\to\infty} P\Big(\max_{1\le j\le 2n} X_j \le n\Big) = \exp\Big[-\frac{1}{2}\Big(\theta^{(1)}\tau'^{(1)} + \theta^{(2)}\tau'^{(2)}\Big)\Big] \tag{4.3}$$

with

$$\tau'^{(i)} = \lim_{n\to\infty} nP\Big(\tau^{(1)}X_1^{(i)} > n\Big) = \lim_{n\to\infty} n\Big(1-e^{-\frac{\tau^{(i)}}{n}}\Big) = \tau^{(i)}\ . \tag{4.4}$$

By paying attention to (4.1), (4.2), (4.3) and (4.4) we can write

$$\begin{aligned}\lim_{n\to\infty} P\Big(\max_{1\le j\le n} X_j^{(1)} \le u_n^{(\tau^{(1)})},\ \max_{1\le j\le n} X_j^{(2)} \le u_n^{(\tau^{(2)})}\Big) = \\ = \exp\Big[-\frac{1}{2}\Big(\theta^{(1)}\tau^{(1)} + \theta^{(2)}\tau^{(2)}\Big)\Big]\ .\end{aligned} \tag{4.5}$$

On the other hand, from the definition of extremal index $\theta(\tau^{(1)},\tau^{(2)})$,

$$\begin{aligned}\lim_{n\to\infty} P\Big(\max_{1\le j\le n} \tau^{(1)}X_j^{(1)} \le n,\ \max_{1\le j\le n} \tau^{(2)}X_j^{(2)} \le n\Big) = \\ = \Big(\lim_{n\to\infty} P\Big(\max_{1\le j\le n} \tau^{(1)}\widehat{X}_j^{(1)} \le n,\ \max_{1\le j\le n} \tau^{(2)}\widehat{X}_j^{(2)} \le n\Big)\Big)^{\theta\left(\tau^{(1)},\tau^{(2)}\right)} \\ = \Big[\exp\Big[-\frac{1}{2}\Big(\tau^{(1)}+\tau^{(2)}\Big)\Big]\Big]^{\theta\left(\tau^{(1)},\tau^{(2)}\right)}\ ,\end{aligned} \tag{4.6}$$

since under condition $\overset{\bullet}{D}(u_n)$ for $\{X_n\}_{n\ge1}$, the sequence $\{\widehat{X}_n\}_{n\ge1}$ also satisfies $\overset{\bullet}{D}(u_n)$ and, for each $i=1,2$, $\widehat{S}_n^{(i)}([0,1]) = S_n[\widehat{X}_n^{(i)}, u_n^{(\tau_i)}]([0,1])$ converges in distribution to a random variable with Poisson distribution with parameter $\tau^{(i)}$.

By attending to a (4.5) and (4.6) it follows that

$$\exp\Big[-\frac{1}{2}\Big(\theta^{(1)}\tau^{(1)} + \theta^{(2)}\tau^{(2)}\Big)\Big] = \Big[\exp\Big(-\frac{1}{2}\Big(\tau^{(1)}+\tau^{(2)}\Big)\Big)\Big]^{\theta\left(\tau^{(1)},\tau^{(2)}\right)}$$

and so

$$\theta\Big(\tau^{(1)},\tau^{(2)}\Big) = \theta^{(1)}\frac{\tau^{(1)}}{\tau^{(1)}+\tau^{(2)}} + \theta^{(2)}\frac{\tau^{(2)}}{\tau^{(1)}+\tau^{(2)}}\ . \tag{4.7}$$

This result is not surprising since under the condition $\overset{\bullet}{D}(u_n)$ for $\{X_n\}_{n\geq 1}$ we have the asymptotic independence of the maxima of the vector margins and Nandagopalan (1994) proves that in this case the bivariate extremal index is a convex linear combination of the marginal extremal indexes as in (4.7).

We finish this section by exhibiting a nonstationary sequence that verifies condition $\overset{\bullet}{D}(u_n)$.

Example 4.1. Let $\left\{Z_n^{(1)}\right\}_{n\geq 1}$ and $\left\{Z_n^{(2)}\right\}_{n\geq 1}$ be independent sequences of random variables. For $0<\lambda\leq 1$ constant, consider the autoregressive sequences of maxima defined as

$$X_n = \lambda \max\left(X_{n-1}, Z_n^{(1)}\right)$$

and

$$Y_n = \lambda \max\left(Y_{n-1}, Z_n^{(2)}\right)$$

where $X_0 = Y_0$ is independent of $\left\{Z_n^{(1)}\right\}_{n\geq 1}$.

For each $n \geq 1$ it follows that

$$X_n = \max\left(\max_{1\leq j\leq n} \lambda^j Z_{n-j+1}^{(1)},\, \lambda^n X_0\right)$$

and

$$Y_n = \max\left(\max_{1\leq j\leq n} \lambda^j Z_{n-j+1}^{(2)},\, \lambda^n Y_0\right) .$$

So, for each $J \subset [0,+\infty[$ such that $m(J) = r_n = [\frac{n}{k_n}]$, we have

$$\begin{aligned} P\left(M_n^{(1)}(J)\leq u_n,\, M_n^{(2)}(J)\leq u_n\right) &= \\ &= P\left(\bigcap_{s\in J}\bigcap_{j=1}^{s} Z_{s-j+1}^{(1)} \leq \frac{u_n}{\lambda^j}\right) P\left(X_0 \leq \frac{u_n}{\lambda^n}\right) P\left(\bigcap_{s\in J}\bigcap_{j=1}^{s} Z_{s-j+1}^{(2)} \leq \frac{u_n}{\lambda^j}\right) \end{aligned}$$

and

$$\begin{aligned} P\left(M_n^{(1)}(J)\leq u_n\right) P\left(M_n^{(2)}(J) \leq u_n\right) &= \\ &= P\left(\bigcap_{s\in J}\bigcap_{j=1}^{s} Z_{s-j+1}^{(1)} \leq \frac{u_n}{\lambda^j}\right) P^2\left(X_0 \leq \frac{u_n}{\lambda^n}\right) P\left(\bigcap_{s\in J}\bigcap_{j=1}^{s} Z_{s-j+1}^{(2)} \leq \frac{u_n}{\lambda^j}\right) \end{aligned}$$

and consequently,

$$\begin{aligned} \left|P\left(M_n^{(1)}(J)\leq u_n,\, M_n^{(2)}(J)\leq u_n\right) - P\left(M_n^{(1)}(J)\leq u_n\right) P\left(M_n^{(2)}(J)\leq u_n\right)\right| &\leq \\ \leq \left|P\left(X_0\leq \frac{u_n}{\lambda^n}\right) - P^2\left(X_0\leq \frac{u_n}{\lambda^n}\right)\right| &\xrightarrow[n\to\infty]{} 0 . \end{aligned}$$

ACKNOWLEDGMENTS

I am grateful to the referee for his rigorous report, suggestions and corrections which helped in improving the final form of this paper and pointed out new challenges.

I also acknowledge the valuable suggestions from Prof. Helena Ferreira.

REFERENCES

[1] Ferreira, H. (1994). *Condições de dependência local em teoria de valores extremos*, Tese de doutoramento. Universidade de Coimbra.

[2] Hsing, T.; Hüsler, J. and Leadbetter, M.R. (1988). On the exceedance point process for stationary sequence, *Probab. Theory Rel. Fields*, **78**, 97–112.

[3] Leadbetter, M.R. (1974). On extreme values in stationary sequences, *Zeitschrift fur Wahrschein. verw. Gebiete*, **28**, 289–303.

[4] Leadbetter, M.R. (1983). Extremes and local dependence in stationary sequences, *Zeitschrift fur Wahrschein. verw. Gebiete*, **65**, 291–306.

[5] Leadbetter, M.R. and Nandagopalan, S. (1989). *On exceedances point processes for stationary sequences under mild oscillation restrictions.* In "Extremes Values" (J. Hüsler and R.-D. Reiss, Eds.), Springer-Verlag, 69–80.

[6] Nandagopalan, S. (1990). *Multivariate extremes and estimation of the extremal index*, Ph. D. Thesis, University of North Carolina at Chapel Hill.

[7] Nandagopalan, S. (1994). On the multivariate extremal index, *J. of Research of National Inst. of Standards and Technology*, **99**, 543–550.

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

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

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

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